Algebra lineare numerica

Alvise Sommariva

Università degli Studi di Padova

10 maggio 2023

Definizione (Norma vettoriale (spazio vettoriale sopra $\mathbb{R}$))

Dato uno spazio vettoriale V (sopra il campo $\mathbb{R}$), la funzione $\|\cdot\| : V \to \mathbb{R}$ è una norma vettoriale se

1. $\|x\| \geq 0$ *per ogni* $x \in V$ *e* $\|x\| = 0$ *se e solo se* $x = 0_V$,
2. $\|\alpha x\| = |\alpha|_{\mathbb{R}} \|x\|$ *per ogni* $x \in V$ *e scalare* $\alpha \in \mathbb{R}$,
3. *vale la cosidetta disuguaglianza triangolare*

$$\|x + y\| \leq \|x\| + \|y\|$$

per ogni $x, y \in V$.

Definizione (Norma vettoriale (spazio vettoriale sopra $\mathbb{C}$))

Dato uno spazio vettoriale V (sopra il campo $\mathbb{C}$), la funzione $\|\cdot\| : V \to \mathbb{R}$ è una norma vettoriale se

1. $\|x\| \geq 0$ *per ogni* $x \in V$ *e* $\|x\| = 0$ *se e solo se* $x = 0_V$,
2. $\|\alpha x\| = |\alpha|_{\mathbb{C}} \|x\|$ *per ogni* $x \in V$ *e scalare* $\alpha \in \mathbb{C}$,
3. *vale la cosidetta disuguaglianza triangolare*

$$\|x + y\| \leq \|x\| + \|y\|$$

per ogni $x, y \in V$.

Nota. (Norma in $\mathbb{C}$)

Si ricorda che se $z = a + i \cdot b$ *allora* $|z|_{\mathbb{C}} = \sqrt{a^2 + b^2}$.

Esempio (Alcune norme vettoriali)

Se $V = \mathbb{C}^n$ o $V = \mathbb{R}^n$ e $x = (x_1, x_2, \ldots, x_n) \in V$ alcuni esempi di norme sono

- *per $p \in [1, +\infty)$, la norma p è definita come*

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

- *la norma del massimo è definita per $p = \infty$ come*

$$\|x\|_\infty = \max_{j=1,\ldots,n} |x_j|$$

Nota.

Di seguito, non distingueremo se lo spazio vettoriale V, sia uguale $\mathbb{C}^n$ o $V = \mathbb{R}^n$. Qualora necessario, per semplicità, si ponga $V = \mathbb{R}^n$.

Nota. (Alcuni casi notevoli della norma p)

Si osservi che essendo $\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}$

- *per* $p = 1$, *si definisce la importante norma* 1

$$\|x\|_1 = \sum_{j=1}^{n} |x_j|$$

- *per* $p = 2$, *si definisce la importante norma* 2, *detta anche euclidea*

$$\|x\|_2 = \left(\sum_{j=1}^{n} |x_j|^2\right)^{1/2}$$

Esempio

Si consideri il vettore $v = (4, -3) \in \mathbb{R}^2$. *Allora*

- $\|v\|_1 = |4| + |-3| = 7$,
- $\|v\|_2 = \sqrt{|4|^2 + |-3|^2} = 5$,
- $\|v\|_\infty = \max(|4|, |-3|) = 4$.

Esempio

Si considerino i vettori $u = (4, -3)$, $v = (4.001, -2.999) \in \mathbb{R}^2$.
I due vettori sono numericamente vicini. Per verificarlo calcoliamo la loro distanza in termine di norme ovvero

$$dist(u, v) := \|u - v\|_p, \quad p = [1, \infty].$$

Allora, visto che $u - v = [-0.001, -0.001]$

- $\|u - v\|_1 = 0.002$,
- $\|u - v\|_2 = 0.0014$,
- $\|u - v\|_\infty = \max(|-0.001|, |-0.001|) = 0.001$,

risultati che sottolineano che u e v sono vicini rispetto alle distanze definite dalle norme.

Definizione

Sia $\|\cdot\|$ una norma vettoriale. Si definisce norma vettoriale indotta di una matrice in $\mathbb{C}^{n\times n}$ la funzione che mappa ogni matrice $A \in \mathbb{C}^{n\times n}$ nel valore

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

Nota.

Tali norme si definiscono anche nel caso particolare che $A \in \mathbb{R}^{n\times n}$.

Definizione (Raggio spettrale)

Si definisce raggio spettrale di A la quantità

$$\rho(A) = \max_{k=1,\dots,n} (|\lambda_k|)$$

dove λ_k è un autovalore di A, ovvero esiste $x_k \neq 0$ detto autovettore di A, tale che $Ax_k = \lambda_k x_k$.

Esempio (Alcune norme matriciali)

Posto $A = (a_{i,j})$, alcuni esempi di norme matriciali risultano:

- *per $p = 1$, partendo dalla norma vettoriale $\|\cdot\|_1$, si dimostra che la norma 1 matriciale risulta uguale al massimo della somma dei moduli delle componenti di ogni colonna ovvero*

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{i,j}|;$$

- *per $p = 2$, partendo dalla norma vettoriale $\|\cdot\|_2$, si dimostra che la norma 2 matriciale risulta uguale a*

$$\|A\|_2 = \sqrt{\rho(A^* A)};$$

dove

  - *se $A = (a_{i,j}) \in \mathbb{R}^{n \times n}$ allora A^* è la trasposta A^T di A, ovvero $a^*_{i,j} = a_{j,i}$,*
  - *se $A \in \mathbb{C}^{n \times n}$ allora A^* è la trasposta coniugata A^H di A ovvero $a^*_{i,j} = \overline{a_{j,i}}$;*

- *per $p = \infty$, partendo dalla norma vettoriale $\|\cdot\|_\infty$, si dimostra che la norma ∞ matriciale risulta uguale al massimo della somma dei moduli delle componenti di ogni riga ovvero*

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{i,j}|$$

Esempio (1)

Si consideri la matrice $A \in \mathbb{R}^{2\times 2}$

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

Allora

- $\|A\|_1 = \max(|1| + |-3|, |-2| + |4|) = 6$;
- *osservato che* $A^* = A^T$ *abbiamo*

$$A^T A = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} 10 & -14 \\ -14 & 20 \end{bmatrix}$$

i cui autovalori valgono rispettivamente

$$\lambda_1 \approx 0.1339312526814949, \quad \lambda_2 \approx 29.86606874731851$$

e di conseguenza

$$\|A\|_2 = \sqrt{\max(|\lambda_1|, |\lambda_2|)} \approx 5.464985704219044;$$

- $\|A\|_\infty = \max(|1| + |-2|, |-3| + |4|) = 7$.

Esempio (2)

Si considerino le matrici

$$A = \begin{bmatrix} 1.001 & -2.999 \\ -2.002 & 4.005 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$$

Le due matrici sono numericamente vicine. Per verificarlo calcoliamo la loro distanza in termine di norme ovvero $dist(A, B) := \|A - B\|_p, \quad p = [1, \infty]$.
Allora, visto che

$$A - B = \begin{bmatrix} 0.0010 & 0.0010 \\ -0.0020 & 0.0050 \end{bmatrix}$$

ricaviamo

- $\|A - B\|_1 = \max(0.003, 0.006) = 0.006$,
- $\|A - B\|_2 = 0.00541565477905833 2$,
- $\|A - B\|_\infty = \max(0.002, 0.007) = 0.007$,

risultati che sottolineano che A e B sono matrici vicine rispetto alle distanze definite dalle norme.

Teorema (Disuguaglianze norme indotte)

Per le norme matriciali indotte, se $A, B \in \mathbb{C}^{n\times n}$, $x \in \mathbb{C}^{n\times 1}$,

$$\|Ax\| \leq \|A\|\|x\|, \tag{1}$$

$$\|AB\| \leq \|A\|\|B\|. \tag{2}$$

Dimostrazione. (Facoltativa)

1. *La prima disuguaglianza é banalmente verificata per* $x = 0$. *Altrimenti, per* $x \neq 0$,

$$\|A\| = \sup_{y\neq 0} \frac{\|Ay\|}{\|y\|} \geq \frac{\|Ax\|}{\|x\|} \Rightarrow \|Ax\| \leq \|A\|\|x\|.$$

2. *Per quanto riguarda la seconda disuguaglianza, dal primo punto, se* $y \neq 0$, *allora*

$$\|ABy\| = \|A(By)\| \leq \|A\|\|By\| \leq \|A\|\|B\|\|y\|$$

da cui $\|ABy\|/\|y\| \leq \|A\|\|B\|$ *e quindi*

$$\|AB\| = \sup_{y\neq 0} \frac{\|ABy\|}{\|y\|} \leq \|A\|\|B\|.$$

Definizione (Indice di condizionamento)

Sia $A \in \mathbb{C}^{n\times n}$ una matrice invertibile ossia tale che $\det(A) \neq 0$. *Si dice indice di condizionamento della matrice A, relativamente alla norma indotta* $\|\cdot\|$,

$$k(A) := \|A\|\|A^{-1}\|.$$

Nota.

Si osservi che dalla definizione, se I è la matrice identica di $\mathbb{C}^{n\times n}$

$$\|I\| = \sup_{x\neq 0} \frac{\|Ix\|}{\|x\|} = \sup_{x\neq 0} \frac{\|x\|}{\|x\|} = 1$$

e quindi, se la norma $\|\cdot\|$ *è indotta, essendo* $\|ST\| \leq \|S\|\|T\|$ *per ogni* $S, T \in \mathbb{C}^{n\times n}$, *ricaviamo*

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = k(A)$$

ovvero $k(A) \geq 1$.

Valgono i seguenti risultati

Teorema (Stime dal basso dell'indice di condizionamento)

1. *Qualsiasi sia la norma matriciale indotta, e A invertibile,*

$$1 \leq \frac{|\lambda_{max}|}{|\lambda_{min}|} \leq k(A)$$

   dove λ_{min} e λ_{max} sono rispettivamente gli autovalori di A di minimo e massimo modulo.

2. *Se A è simmetrica, ovvero $A = A^T$,*

$$k_2(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|}.$$

Esempio (Matrice di Hilbert)

La matrice $H = (h_{i,j})_{i,j=1,\dots,n}$ *con*

$$h_{i,j} = \frac{1}{i+j-1}$$

è nota come matrice di Hilbert di ordine n. Per queste matrici gli indici di condizionamento crescono molto rapidamente, come si può vedere dalla tabella.

n	c_n	n	c_n	n	c_n
1	$1.00e+00$	5	$4.77e+05$	9	$4.93e+11$
2	$1.93e+01$	6	$1.50e+07$	10	$1.60e+13$
3	$5.24e+02$	7	$4.75e+08$	11	$5.22e+14$
4	$1.55e+04$	8	$1.53e+10$	12	$1.62e+16$

Tabella: *Indice di condizionamento* $c_n = cond_2(H_n)$ *della matrici di Hilbert* H_n *di ordine n*

Vale la seguente risposta della soluzione di un sistema lineare agli errori del vettore termine noto.

Teorema (Condizionamento ed errore nel termine noto)

Se

1. $A \in \mathbb{C}^{n \times n}$ *è una matrice invertibile,*
2. $b \in \mathbb{C}^n$, $b \neq 0$,
3. $Ax = b$,
4. $A(x + \delta x) = b + \delta b$,
5. $\| \cdot \|$ *è una norma indotta,*

allora vale la stima

$$\frac{\|\delta x\|}{\|x\|} \leq k(A) \frac{\|\delta b\|}{\|b\|}$$

dove $k(A) = \|A\|\|A^{-1}\|$.

Dimostrazione. (Facoltativa)

Da

$$A(x+\delta x) = b+\delta b, \quad Ax = b$$

abbiamo che

$$A\delta x = A((x+\delta x)-x) = A(x+\delta x) - Ax = (b+\delta b) - b = \delta b \Leftrightarrow \delta x = A^{-1}\delta b.$$

Da (1), $\|Su\| \leq \|S\|\|u\|$*, per ogni* $S \in \mathbb{C}^{n\times n}$*,* $u \in C^n$ *e quindi per* $S = A^{-1}$*,* $u = \delta b$ *si ha* $\|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|$ *da cui*

$$0 \leq \|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|. \tag{3}$$

D'altra parte, essendo $Ax = b$*, necessariamente per (1)*

$$\|b\| = \|Ax\| \leq \|A\|\|x\|$$

e quindi, essendo $x \neq 0$ *in quanto* $b \neq 0$*, moltiplicando i membri di* $0 \leq \|b\| \leq \|A\|\|x\|$ *per* $\frac{1}{\|x\|\|b\|}$

$$0 \leq \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{4}$$

Da (3), (4), visto che se $0 \leq s \leq u$*,* $0 \leq t \leq v$ *allora* $0 \leq st \leq uv$ *e* $k(A) = \|A^{-1}\|\|A\|$*, ricaviamo*

$$\frac{\|\delta x\|}{\|x\|} = \|\delta x\| \cdot \frac{1}{\|x\|} \leq \|A^{-1}\|\|\delta b\| \frac{\|A\|}{\|b\|} = \|A^{-1}\|\|A\| \frac{\|\delta b\|}{\|b\|} = k(A)\frac{\|\delta b\|}{\|b\|}.$$

Commento (Relazione tra $\|\delta x\|/\|x\|$ e $\|\delta b\|/\|b\|$)

- *Questo teorema mostra che si fa un errore relativo sul termine noto pari a* $\frac{\|\delta b\|}{\|b\|}$ *allora si compie un errore relativo tra soluzione* x *e soluzione* $x + \delta x$ *del problema perturbato*

$$\frac{\|x - (x + \delta x)\|}{\|x\|} = \frac{\|\delta x\|}{\|x\|}$$

tale che

$$\frac{\|\delta x\|}{\|x\|} \leq k(A) \frac{\|\delta b\|}{\|b\|}.$$

- *Quindi, se* $k(A) \gg 1$, *può accadere che la soluzione* $x + \delta x$ *del problema perturbato sia molto distante da* x *e di conseguenza la soluzione del sistema lineare* $Ax = b$ *molto suscettibile a piccoli errori sul termine noto.*
- *In altri termini più grande è* $k(A)$ *più i sistemi lineari* $Ax = b$ *sono difficili da trattare nel senso che piccoli errori sui dati* b *possono portare a grossi errori sulla soluzione* x.

Esempio

Siano

$$A = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0.7 \end{bmatrix}, \quad b + \delta b = \begin{bmatrix} 1.01 \\ 0.69 \end{bmatrix}.$$

Il sistema $Ax = b$ *ha soluzione*

$$x = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}$$

mentre $A(x + \delta x) = b + \delta b$ *ha soluzione*

$$x + \delta x = \begin{bmatrix} -0.17 \\ 0.22 \end{bmatrix}.$$

Di conseguenza seppure $\frac{\|\delta b\|_\infty}{\|b\|_\infty} = 0.01$ *abbiamo* $\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 1.7$.

Questo significa che a un piccolo errore relativo sui dati abbiamo riscontrato un eccessivo errore relativo sulla soluzione.

Osserviamo che in questo caso $k(A) = \|A\|_\infty \|A^{-1}\|_\infty = 289$ *e quindi non troppo vicino a* 1.

Teorema (Condizionamento e errore sui dati della matrice)

Se

1. $A \in \mathbb{C}^{n\times n}$ *è una matrice invertibile,*
2. $b \in \mathbb{C}^n$, $b \neq 0$,
3. $Ax = b$,
4. $(A + \delta A)(x + \delta x) = b$, *con* $\det(A + \delta A) \neq 0$,

allora per una qualsiasi norma indotta $\|\cdot\|$ *vale la stima*

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq k(A)\frac{\|\delta A\|}{\|A\|}.$$

dove $k(A) = \|A\|\|A^{-1}\|$.

Commento

Il precedente teorema dice che in presenza di un forte indice di condizionamento, piccoli errori relativi sulla matrice possono indurre grandi errori relativi nella soluzione. Si noti che usualmente $\|x\| \approx \|x + \delta x\|$.

Dimostrazione. (Facoltativa)

Essendo

- $(A+\delta A)(x+\delta x)=b$,
- $Ax=b$,

necessariamente

$$\begin{aligned} 0=(A+\delta A)(x+\delta x)-b &= Ax+A\delta x+\delta Ax+\delta A\delta x-b \\ &= A\delta x+\delta Ax+\delta A\delta x = A\delta x+\delta A(x+\delta x) \end{aligned}$$

da cui $A\delta x=-\delta A(x+\delta x)$ *ovvero, moltiplicando ambo i membri per* A^{-1},

$$\delta x=-A^{-1}\delta A(x+\delta x)$$

e per (1), (2), da $\|Tu\|\le\|T\|\|u\|$, $\|ST\|\le\|S\|\|T\|$ *per ogni* $S,T\in\mathbb{C}^{n\times n}$, $u\in\mathbb{C}^n$,

$$\|\delta x\|=\|-A^{-1}\delta A(x+\delta x)\|\le\|A^{-1}\delta A\|\|x+\delta x\|\le\|A^{-1}\|\|\delta A\|\|x+\delta x\|. \quad (5)$$

Essendo $\|A\|\neq 0$, *dividendo ambo i membri di (5) per* $\|x+\delta x\|$, *da* $k(A)=\|A\|\|A^{-1}\|$

$$\frac{\|\delta x\|}{\|x+\delta x\|}\le\|A^{-1}\|\|\delta A\|=\frac{\|A^{-1}\|\|A\|}{\|A\|}\|\delta A\|=k(A)\frac{\|\delta A\|}{\|A\|}.$$

Esempio

Siano

$$A = \begin{bmatrix} 1.0000 & 0.5000 & 0.3333 \\ 0.5000 & 0.3333 & 0.2500 \\ 0.3333 & 0.2500 & 0.2000 \end{bmatrix}, A + \delta A = \begin{bmatrix} 1.0010 & 0.5030 & 0.3403 \\ 0.5040 & 0.3383 & 0.2520 \\ 0.3423 & 0.2520 & 0.2050 \end{bmatrix},$$

Il sistema $Ax = b$, *per* $b = [1, 2, 3]^T$ *ha soluzione approssimativamente*

$$x \approx \begin{bmatrix} 27.3232 \\ -193.6572 \\ 211.5374 \end{bmatrix}$$

mentre $(A + \delta A)(x + \delta x) = b$ *ha soluzione*

$$x + \delta x \approx \begin{bmatrix} 5.5489 \\ -75.3599 \\ 98.0063 \end{bmatrix}.$$

Di conseguenza seppure $\frac{\|\delta A\|_2}{\|A\|_2} = 9.35310e-03$ *abbiamo* $\frac{\|\delta x\|_2}{\|x+\delta x\|_2} = 1.33653e+00$. *Questo significa che a un piccolo errore relativo sui dati abbiamo riscontrato un eccessivo errore relativo sulla soluzione.*

Osserviamo che in questo caso $k(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 528.5409$ *e quindi non troppo vicino a* 1.

Vogliamo ora studiare il caso in cui tanto la matrice quanto il termine noto siano soggetti ad errori.

Teorema (Facoltativo)

Se

1. $A \in \mathbb{C}^{n\times n}$ *è una matrice invertibile,*
2. $b \in \mathbb{C}^n$, $b \neq 0$,
3. $Ax = b$,
4. $(A+\delta A)(x+\delta x) = b+\delta b$, *con* $\det(A+\delta A) \neq 0$,
5. $\|\cdot\|$ *è una norma indotta,*
6. $k(A)\|\delta A\| < \|A\|$,

allora vale la stima

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

Commento

Il teorema stima l'errore relativo $\frac{\|\delta x\|}{\|x\|}$ *relativamente alle soluzioni, in funzione degli errori relativi sulla perturbazione della matrice* $\frac{\|\delta A\|}{\|A\|}$ *e sul termine noto* $\frac{\|\delta b\|}{\|b\|}$.

Il valore

$$\mu(A) := \frac{k(A)}{1 - k(A)\|\delta A\|/\|A\|}$$

è la costante di amplificazione degli errori.

Visto che il denominatore è minore di 1 *(in effetti* $k(A)\|\delta A\| < \|A\|$ *porge* $k(A)\|\delta A\|/\|A\| < 1$*), deduciamo che se* $k(A) \gg 1$*, allora* $\mu(A) \gg 1$ *e quindi puó essere* $\frac{\|\delta x\|}{\|x\|} \gg 1$*, ovvero* $\|\delta x\| \gg \|x\|$ *per cui la soluzione del sistema perturbato*

$$(A + \delta A)(x + \delta x) = b + \delta b$$

può essere molto distante da quella di $Ax = b$.

Teorema (Facoltativo)

Se

1. $\|\cdot\|$ *è una norma matriciale indotta da una vettoriale,*
2. $\|A\| < 1$,

allora "$I - A$" e "$I + A$" sono invertibili e si ha

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}, \ \|(I+A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Dimostrazione. (Facoltativo)

Essendo $(A+\delta A)(x+\delta x) = b + \delta x$ *e* $Ax = b$ *deduciamo*

$$b+\delta b \quad = \quad (A+\delta A)(x+\delta x) = Ax + A\delta x + \delta Ax + \delta A\delta x = b + A\delta x + \delta Ax + \delta A\delta x$$

da cui sottraendo b ad ambo i membri $\delta b = A\delta x + \delta Ax + \delta A\delta x = \delta Ax + (A+\delta A)\delta x$ *ovvero* $(A+\delta A)\delta x = \delta b - \delta Ax$.
Posto $B = A^{-1}\delta A$, *ricaviamo* $A + \delta A = A + AA^{-1}\delta A = A(I + A^{-1}\delta A) = A(I+B)$ *e da* $k(A)\|\delta A\| < \|A\|$

$$\|B\| \quad = \quad \|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| = \frac{\|A^{-1}\|\|A\|}{\|A\|}\|\delta A\| = k(A)\frac{\|\delta A\|}{\|A\|} \leq \frac{\|A\|}{\|A\|} = 1$$

ovvero $\|B\| < 1$.
Osserviamo che la matrice $A + \delta A$ *è invertibile in quanto lo è A, lo è* $I + B$ *per il teorema 0.6 e da*

$$A + \delta A = A(I + A^{-1}\delta A) = A(I+B)$$

lo è pure $A + \delta A$ *perchè il prodotto di matrici invertibili è invertibile (ricordare che* $\det(MN) = \det(M)\det(N)$, *e quindi se* $\det(M), \det(N) \neq 0$ *allora* $\det(MN) \neq 0$*).*

Da un teorema precedente, moltiplicando ambo i membri per $(A+\delta A)^{-1}$, *visto che* $(MN)^{-1} = N^{-1}M^{-1}$

$$\delta x = (A+\delta A)^{-1}(\delta b - \delta A x) = (A(I+B))^{-1}(\delta b - \delta A x) = (I+B)^{-1}A^{-1}(\delta b - \delta A x)$$

da cui, essendo $\|(I+B)^{-1}\| \leq 1/(1-\|B\|)$, $B = A^{-1}\delta A$

$$\begin{aligned}\|\delta x\| &= \|(I+B)^{-1}A^{-1}(\delta b - \delta A x)\| \leq \|(I+B)^{-1}\|\|A^{-1}\|\|\delta b - \delta A x\| \\ &\leq \frac{1}{1-\|B\|}\|A^{-1}\|\|\delta b - \delta A x\| = \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|}\|\delta b - \delta A x\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|)\end{aligned}$$

e quindi riassumendo

$$\|\delta x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|). \tag{6}$$

Essendo $b = Ax$ *abbiamo* $\|b\| = \|Ax\| \leq \|A\|\|x\|$ *da cui*

$$\frac{1}{\|A\|\|x\|} \leq \frac{1}{\|b\|} \tag{7}$$

e moltiplicando membro a membro i risultati di (6), (7),

$$\begin{aligned}\frac{\|\delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\|x\|\frac{1}{\|x\|}\right) \leq \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\frac{\|1\|}{\|A\|}\left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\right) \\ &\leq \frac{k(A)}{1-\|A^{-1}\|\|A\|\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|A\|\|x\|} + \frac{\|\delta A\|}{\|A\|}\right) \leq \frac{k(A)}{1-k(A)\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right). \end{aligned} \tag{8}$$

Esempio (Facoltativo)

Il sistema $Ax = b$, *dove*

$$A = \begin{bmatrix} 1 & 2 \\ 0.499 & 1.001 \end{bmatrix}, b = \begin{bmatrix} 3 \\ 1.5 \end{bmatrix},$$

ha soluzione $x = (1, 1)^T$. *Se consideriamo*

$$A + \delta A = \begin{bmatrix} 1 & 2 \\ 0.5 & 1.0015 \end{bmatrix}, \quad b + \delta b = \begin{bmatrix} 3 \\ 1.4985 \end{bmatrix}.$$

- $(A + \delta A)u = b$ *ha soluzione* $u = (3, 0)^T$ *(perturbata matrice A),*
- $Av = b + \delta b$ *ha soluzione* $v = (2, 0.5)^T$ *(perturbato termine noto b),*
- $(A + \delta A)w = (b + \delta b)$ *ha soluzione* $w = (5, -1)^T$ *(perturbati A e b).*

Ne risulta che per possibili piccole perturbazioni alla matrice e/o al termine noto abbiamo ottenuto nei diversi casi una soluzione molto distante dall'originaria. Si ha

- $\det(A) \approx 0.003$, $\det(\delta A) \approx 0.0015$, $k_2(A) \approx 2083.666853410356$,
- *abbiamo* $k_2(A)/(1 - k_2(A)||\,||\delta A||_2/||A||_2) \approx 3.05387e + 04$;
- $\frac{\|\delta x\|_2}{\|x\|_2} \approx 3.16228e + 00$, $\frac{\|\delta b\|_2}{\|b\|_2} \approx 4.47214e - 04$ *e* $\frac{\|\delta A\|_2}{\|A\|_2} \approx 4.47178e - 04$.

La stima asserisce in effetti che

$$\begin{aligned} 3.16228e + 00 &\leq 3.05387e + 04 \cdot (4.47214e - 04 + 4.47178e - 04) \\ &\approx 2.73136e + 01. \end{aligned} \tag{9}$$

Proposito.

Il proposito di questa sezione è di introdurre il metodo dell'eliminazione gaussiana per risolvere sistemi lineari $Ax = b$ dove supponiamo A invertibile, cioè tale che $\det(A) \neq 0$.

Per semplificare l'esposizione cominciamo con un esempio.

Esempio

Si risolva il sistema lineare

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 6x_1 - 2x_2 + 2x_3 = 6 \\ -9x_1 + 7x_2 + x_3 = -1 \end{cases}$$

mediante eliminazione gaussiana.

Sappiamo che

- se moltiplichiamo ambo i membri di una stessa equazione per una costante non nulla, il nuovo problema ha le stesse soluzioni del precedente;
- se scambiamo due equazioni, il nuovo problema ha le stesse soluzioni del precedente;
- in generale se moltiplichiamo la i-sima equazione per una costante non nulla e sottraiamo il risultato membro a membro alla j-sima, ricaviamo una equazione che sostituita alla j-sima, propone un nuovo problema lineare che ha le stesse soluzioni del precedente.

Svolgimento.

Diremo che due sistemi sono equivalenti, se hanno le stesse soluzioni. Da:

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 6x_1 - 2x_2 + 2x_3 = 6 \\ -9x_1 + 7x_2 + x_3 = -1 \end{cases}$$

- *moltiplichiamo la prima equazione per* $m_{2,1} = 6/3 = 2$ *e la sottraiamo alla seconda, ottenendo il sistema lineare equivalente*

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 0 \cdot x_1 + 2x_2 + 4x_3 = 6 \\ -9x_1 + 7x_2 + x_3 = -1 \end{cases}$$

- *moltiplichiamo la prima equazione per* $m_{3,1} = -9/3 = -3$ *e la sottraiamo alla terza, ottenendo il sistema lineare equivalente*

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 0x_1 + 2x_2 + 4x_3 = 6 \\ 0x_1 + 1x_2 - 2x_3 = -1 \end{cases}$$

  in cui le componenti nulle della prima colonna, precedentemente ottenute, non vengono modificate.

- *moltiplichiamo la seconda equazione per* $m_{3,2} = 1/2$ *e la sottraiamo alla terza, ottenendo il sistema lineare equivalente*

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 0x_1 + 2x_2 + 4x_3 = 6 \\ 0x_1 + 0x_2 - 4x_3 = -4 \end{cases}$$

  in cui le componenti nulle della colonna precedente non vengono modificate.

Il sistema finale

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 0x_1 + 2x_2 + 4x_3 = 6 \\ 0x_1 + 0x_2 - 4x_3 = -4 \end{cases}$$

è facile da risolvere.

- *L'ultima equazione* $-4x_3 = -4$ *ha una sola variabile e una sola incognita e comporta che*

$$x_3 = 1.$$

- *Inseriamo* $x_3 = 1$ *nella penultima equazione* $2x_2 + 4x_3 = 6$ *e otteniamo*

$$2x_2 + 4 \cdot 1 = 6$$

  da cui $2x_2 = 2$ *e quindi* $x_2 = 1$.

- *Inseriamo* $x_2 = x_3 = 1$ *nella prima equazione* $3x_1 - 2x_2 - x_3 = 0$ *e otteniamo*

$$3x_1 - 2 \cdot 1 - 1 = 0$$

  da cui

$$x_1 = 1.$$

Quindi il sistema ha soluzioni $x_1 = 1$, $x_2 = 1$, $x_3 = 1$.

Nota. (Sistema matriciale)

Il sistema può essere scritto matricialmente come $Ax = b$ dove

$$A = \begin{bmatrix} 3 & -2 & -1 \\ 6 & -2 & 2 \\ -9 & 7 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 6 \\ -1 \end{bmatrix},$$

e abbiamo mostrato essere equivalente a $Ux = c$ dove

$$U = \begin{bmatrix} 3 & -2 & -1 \\ 0 & 2 & 4 \\ 0 & 0 & -4 \end{bmatrix}, \qquad c = \begin{bmatrix} 0 \\ 6 \\ -4 \end{bmatrix}.$$

Dal punto di vista matriciale, abbiamo generato una sequenza di sistemi lineari equivalenti, ma via via più semplici.

L'ultimo è un sistema $Ux = c$ con U triangolare superiore, ovvero tale che $U_{i,j} = 0$ qualora $i > j$, che abbiamo risolto facilmente mediante sostituzione all'indietro.

Definizione

Una matrice $A = (a_{i,j}) \in \mathbb{C}^{n\times n}$ si dice

- *triangolare inferiore se $a_{i,j} = 0$ qualora $i < j$;*
- *triangolare superiore se $a_{i,j} = 0$ qualora $i > j$.*

Esempio

La matrice

$$L = \begin{bmatrix} 3 & 0 & 0 \\ 4 & 2 & 0 \\ 1 & 0 & -4 \end{bmatrix}$$

è triangolare inferiore ($a_{i,j} = 0$ se l'indice di riga strettamente minore di quello di colonna).

Esempio

La matrice

$$U = \begin{bmatrix} 3 & -2 & -1 \\ 0 & 2 & 4 \\ 0 & 0 & -4 \end{bmatrix}$$

è triangolare superiore ($a_{i,j} = 0$ se l'indice di riga strettamente maggiore di quello di colonna).

Algoritmo. (Sostituzione all'indietro e in avanti)

Sia $b = (b_j) \in \mathbb{C}^n$ e supponiamo di dover risolvere il sistema $Ax = b$, con $A_{k,k} \neq 0$ per $k = 1, \ldots, n$.

In tal caso,

- *se A è triangolare superiore e invertibile, si usa l'algoritmo della sostituzione all'indietro ovvero*

$$\begin{cases} x_n = \frac{b_n}{a_{n,n}} \\ x_k = \frac{b_k - (a_{k,k+1}x_{k+1} + \ldots + a_{k,n}x_n)}{a_{k,k}}, \quad k = n-1, \ldots, 1. \end{cases} \tag{10}$$

- *se A è triangolare inferiore e invertibile, si usa l'algoritmo della sostituzione in avanti ovvero*

$$\begin{cases} x_1 = b_1/a_{1,1}, \\ x_k = \frac{b_k - (a_{k,1}x_1 + \ldots + a_{k,k-1}x_{k-1})}{a_{k,k}}, \quad k = 2, \ldots, n. \end{cases} \tag{11}$$

Nota. (Complessitá risoluzione sistemi triangolari con eliminazione gaussiana)

Si osservi che la risoluzione di sistemi lineari triangolari necessita approssimativamente $n^2/2$ operazioni moltiplicative, dove n è l'ordine della matrice.

Infatti il numero di operazioni moltiplicative, nel caso della sostituzione all'indietro risulta

$$1 + 2 + \ldots + n = n(n+1)/2 \approx n^2/2.$$

In questa sezione trattiamo la risoluzione di sistemi lineari $Ax = b$ dove in generale si suppone

- $A \in \mathbb{C}^{n\times n}$, invertibile e non triangolare,
- $b \in \mathbb{C}^n$.

Posti

- $A = A^{(1)} = (a^{(1)}_{i,j})$,
- $b = b^{(1)} = (b^{(1)}_j)$,

possiamo immagazzinare le componenti che lo descrivono nella matrice aumentata

$$[A^{(1)}|b^{(1)}] = \left[\begin{array}{cccc|c} a^{(1)}_{1,1} & a^{(1)}_{1,2} & \dots & a^{(1)}_{1,n} & b^{(1)}_1 \\ a^{(1)}_{2,1} & a^{(1)}_{2,2} & \dots & a^{(1)}_{2,n} & b^{(1)}_2 \\ \dots & \dots & \dots & \dots & \dots \\ a^{(1)}_{n,1} & a^{(1)}_{n,2} & \dots & a^{(1)}_{n,n} & b^{(1)}_n \end{array}\right].$$

Primo passo. Supponiamo sia $a_{1,1}^{(1)} \neq 0$. Intendiamo definire un sistema $[A^{(2)}|b^{(2)}]$ tale che $A^{(2)}x = b^{(2)}$ sia equivalente a $A^{(1)}x = b^{(1)}$, ovvero con la stessa soluzione, ma in cui $A^{(2)} = (a_{i,j}^{(2)})$ sia tale che

$$a_{i,1}^{(2)} = 0, \quad i = 2, \ldots, n.$$

A tal proposito definiamo il moltiplicatore

$$m_{i,1} = \frac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}}, \, i = 2, \ldots, n$$

e in successione moltiplichiamo la prima riga per $-m_{i,1}$ e la sommiamo l'i-sima,

$$a_{i,j}^{(2)} = a_{i,j}^{(1)} - m_{i,1}a_{1,j}^{(1)}, \, j = 2, \ldots, n,$$

$$b_i^{(2)} = b_i^{(1)} - m_{i,1}b_1^{(1)}, \, i = 2, \ldots, n.$$

Si vede che si raggiunge lo scopo prefisso visto che per definizione dei moltiplicatori

$$a_{i,1}^{(2)} = a_{i,1}^{(1)} - m_{i,1}a_{1,1}^{(1)} = 0, \quad i = 2, \ldots, n.$$

Di conseguenza la matrice aumentata $[A^{(2)}|b^{(2)}]$ avrà la forma

$$[A^{(2)}|b^{(2)}] = \left[\begin{array}{cccc|c} a_{1,1}^{(2)} & a_{1,2}^{(2)} & \ldots & a_{1,n}^{(2)} & b_1^{(2)} \\ 0 & a_{2,2}^{(2)} & \ldots & a_{2,n}^{(2)} & b_2^{(2)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & a_{n,2}^{(2)} & \ldots & a_{n,n}^{(2)} & b_n^{(2)} \end{array}\right].$$

Passo k-simo. Supponiamo $k = 2, \ldots, n-1$. Assumiamo che sia

$$[A^{(k)}|b^{(k)}] = \left[\begin{array}{cccccccc|c} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \ldots & \ldots & \ldots & \ldots & a_{1,n}^{(k)} & b_1^{(k)} \\ 0 & a_{2,2}^{(k)} & \ldots & \ldots & \ldots & \ldots & a_{2,n}^{(k)} & b_2^{(k)} \\ & \ddots & \ddots & & & & & \\ \ldots & \ddots & \ddots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{k,k}^{(k)} & \ldots & \ldots & a_{k,n}^{(k)} & b_k^{(k)} \\ 0 & \ldots & 0 & a_{k+1,k}^{(k)} & \ldots & \ldots & a_{k+1,n}^{(k)} & b_{k+1}^{(k)} \\ 0 & \ldots & 0 & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{n,k}^{(k)} & \ldots & \ldots & a_{n,n}^{(k)} & b_n^{(k)} \end{array}\right].$$

Se l'elemento pivotale $a_{k,k}^{(k)} \neq 0$ definiamo il moltiplicatore

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad i = k+1, \ldots, n.$$

moltiplichiamo la k-sima riga per $-m_{i,k}$ e la sommiamo alla i-sima, con $i = k+1, \ldots, n$, ovvero

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - m_{i,k} a_{k,j}^{(k)}, \quad j = k+1, \ldots, n,$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{i,k} b_k^{(k)}, \quad j = k+1, \ldots, n.$$

Nuovamente, come nel primo passo, dalla definizione dei moltiplicatori, otteniamo

$$a_{i,k}^{(k+1)} = 0, \quad i = k+1, \ldots, n,$$

da cui

$$[A^{(k+1)}|b^{(k+1)}] = \left[\begin{array}{ccccccc|c} a_{1,1}^{(k+1)} & a_{1,2}^{(k+1)} & \ldots & \ldots & \ldots & \ldots & a_{1,n}^{(k+1)} & b_1^{(k+1)} \\ 0 & a_{2,2}^{(k+1)} & \ldots & \ldots & \ldots & \ldots & a_{2,n}^{(k+1)} & b_2^{(k+1)} \\ \ldots & \ddots & \ddots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{k,k}^{(k+1)} & \ldots & \ldots & a_{k,n}^{(k+1)} & b_k^{(k)} \\ 0 & \ldots & 0 & 0 & a_{k+1,k+1}^{(k+1)} & \ldots & a_{k+1,n}^{(k)} & b_{k+1}^{(k)} \\ 0 & \ldots & 0 & 0 & a_{n-1,k+1}^{(k+1)} & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & a_{n,k+1}^{(k+1)} & \ldots & a_{n,n}^{(k)} & b_n^{(k)} \end{array}\right].$$

Con questo procedimento, dopo $n-1$ passi otteniamo un sistema $A^{(n)}x = b^{(n)}$ in cui la matrice $A^{(n)}$ è triangolare superiore con elementi diagonali $A_{k,k}^{(n)} \neq 0$, $k = 1, \ldots, n$.

Di conseguenza il sistema $A^{(n)}x = b^{(n)}$ può essere facilmente risolto con l'algoritmo di sostituzione all'indietro.

Definizione (Fattorizzazione LU)

Una matrice $A \in \mathbb{C}^{n\times n}$ è fattorizzabile LU se esistono

- $L = (l_{i,j}) \in \mathbb{C}^{n\times n}$ *triangolare inferiore con elementi diagonali* $l_{i,i} = 1$,
- $U = (u_{i,j}) \in \mathbb{C}^{n\times n}$ *triangolare superiore,*

tali che $A = LU$ (cf. [9]).

Importante. (Pivot non nulli)

Dall'applicazione della procedura precedente per la risoluzione di un sistema lineare, qualora tutti i pivot $a_{k,k}^{(k)}$ siano non nulli, otteniamo un sistema $A^{(n)}x = b^{(n)}$ equivalente all'originario e durante il calcolo introduciamo i moltiplicatori $m_{i,k}$, $k = 1, \dots, n-1$, $i = k+1, \dots, n$. Siano

- $U = A^{(n)}$,
- *L la matrice*

$$L = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ m_{2,1} & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \\ \vdots & \vdots & \vdots & \dots & \vdots & 0 \\ m_{n,1} & m_{n,2} & \dots & \dots & m_{n,n-1} & 1 \end{bmatrix}.$$

Allora $A = LU$ (cf. [1, p.511], [3, p.143, p.158]).

Questo risultato è rilevante, perchè tale fattorizzazione viene molto utilizzata, come vedremo in seguito con qualche esempio, indipendentemente dalla risoluzione di sistemi lineari.

Esempio

Sia, come nell'esempio iniziale,

$$A = \begin{bmatrix} 3 & -2 & -1 \\ 6 & -2 & 2 \\ -9 & 7 & 1 \end{bmatrix},$$

Alla fine dell'eliminazione gaussiana senza pivoting abbiamo determinato $U = A^{(n)}$ *uguale a*

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

Tenendo conto dei moltiplicatori $m_{2,1} = 2$, $m_{3,1} = -3$, $m_{2,2} = 0.5$,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ m_{2,1} & 1 & 0 \\ m_{3,1} & m_{3,2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0.5 & 1 \end{bmatrix},$$

e si verifica facilmente che $A = LU$, *con* L, U *triangolari e della forma richiesta. In Matlab:*

```
>> A=[3 -2 -1; 6 -2 2; -9 7 1];
>> U=[3 -2 -1; 0 2 4; 0 0 -4]; L=[1 0 0; 2 1 0; -3 0.5 1];
>> L*U  % verifichiamo che A=L*U
ans =
     3    -2    -1
     6    -2     2
    -9     7     1
>> norm(A-L*U) % la norma di una matrice C nulla se e solo se C e' nulla
ans =
     0
>>
```

Commento

Nella precedente sezione, uno dei punti cruciali è risultato che i pivot $a_{k,k}^{(k)}$ fossero non nulli. Questo ci ha permesso di calcolare i moltiplicatori

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad k = 1, \ldots, n-1, \quad i = k+1, \ldots, n$$

e di seguito ridurre il problema iniziale alla risoluzione di un sistema $A^{(n)}x = b^{(n)}$, con $A^{(n)}$ triangolare superiore.

Possono sorgere i seguenti problemi:

- *qualche $a_{k,k}^{(k)}$ risulta nullo e quindi il processo precedente è inapplicabile;*
- *qualche $a_{k,k}^{(k)}$ risulta molto piccolo in modulo, questione che rende il metodo soggetto a una cattiva propagazione degli errori (cf. [1, p.516], [3, p.174]).*

La soluzione risulta la seguente, detta del pivoting (parziale) o del massimo pivot.

Supponiamo di voler azzerare le componenti sotto la diagonale della colonna k-sima e determiniamo

$$c_k = \max_{k \le i \le n} |a_{i,k}^{(k)}|.$$

Certamente $c_k \neq 0$, altrimenti, dopo qualche conto, si vedrebbe che la matrice originaria A di cui studiamo $Ax = b$ sarebbe non invertibile, diversamente da quanto richiesto.

- *Sia $i_k := \phi(k) \geq k$ il più piccolo indice di riga, per cui il massimo c_k è ottenuto.*
- *Se $i_k > k$, scambiamo la riga i_k-sima con la k-sima della matrice aumentata $[A^{(k)}|b^{(k)}]$ ottenendo una nuova matrice aumentata, diciamo $[\hat{A}^{(k)}|\hat{b}^{(k)}]$, su cui possiamo applicare la tecnica introdotta al k-simo passo dell'eliminazione gaussiana senza pivoting.*

  Con questa tecnica, la nuova matrice aumentata avrà il pivot più grande possibile in modulo, con vantaggi per la propagazione degli errori.

Definizione (Matrice di permutazione)

Una matrice $P = (p_{i,j})$ si dice di permutazione se le sue componenti sono 0 o 1, e ogni riga e colonna contengono una sola componente uguale a 1.

Se calcoliamo $y = Px$ il risultato sara' un vettore con le stesse componenti di x opportunamente *scambiate* (in matematica si usa dire *permutate*).

Esempio

Consideriamo la matrice di permutazione

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

e il vettore $x = (x_1, x_2, x_3)^T$. Allora

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 \\ 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 \\ 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}.$$

Metodo di eliminazione gaussiana con pivoting (parziale)

Importante.

Dall'applicazione della procedura dell'eliminazione gaussiana con pivoting per la risoluzione di un sistema lineare, abbiamo ottenuto un sistema $A^{(n)}x = b^{(n)}$ equivalente all'originario. Poniamo

- *P matrice di permutazione ottenuta come $P = P_{n-1} \cdot \ldots \cdot P_1$ dove P_k è la matrice di permutazione che la scambia la riga i_k con la k-sima, ovvero la matrice identica eccetto per le righe i_k, k in cui*

$$P_{i_k,i_k} = 0, P_{i_k,k} = 1, P_{k,i_k} = 1, P_{k,k} = 0;$$

- *$U = A^{(n)}$ di forma triangolare superiore,*
- *$L = (l_{i,j})$ la matrice*

$$L = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ m_{2,1} & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ m_{n,1} & m_{n,2} & \ldots & \ldots & m_{n,n-1} & 1 \end{bmatrix}.$$

di forma triangolare inferiore, con componenti diagonali $l_{k,k} = 1$, $k = 1, \ldots, n$.

Allora $PA = LU$ (cf. [1, p.511], [3, p.143, p.158]).

Questo risultato è rilevante, perchè tale fattorizzazione viene molto utilizzata indipendentemente dalla risoluzione di sistemi lineari.

Nota. (storica)

L'algoritmo di eliminazione gaussiana è stato scoperto in Cina circa 2000 anni fa, con più recenti contributi in lavori di Lagrange e Jacobi. La descrizione moderna è stata introdotta solo nel 1930. [7, p.607].

Per capire perché si eseguono queste operazioni matriciali, producendo una sequenza di sistemi lineari equivalenti tra loro, mostriamo le idee su un esempio. Si userá un linguaggio matriciale, seppure si analizzi un sistema di equazioni. Per "riga" si intenderá "equazione", per "colonna" si intenderá "variabile".

Esempio

Si risolva il sistema lineare

$$\begin{cases} 3x_1 - 2x_2 - x_3 = 0 \\ 6x_1 - 2x_2 + 2x_3 = 6 \\ -9x_1 + 7x_2 + x_3 = -1 \end{cases}$$

mediante il metodo di eliminazione gaussiana con pivoting (parziale).

Il termine maggiore in modulo della prima colonna è il terzo e vale $|-9| = 9$. Scriviamo $\phi(1) = 3$, intendendo che nella prima colonna, la componente di massimo modulo é la terza. Scambiamo la prima riga, ovvero equazione, con la terza e ricaviamo il sistema equivalente

$$\begin{cases} -9x_1 + 7x_2 + x_3 = -1 \\ 6x_1 - 2x_2 + 2x_3 = 6 \\ 3x_1 - 2x_2 - x_3 = 0 \end{cases}$$

Applichiamo ora i primi passi dell'eliminazione gaussiana relativamente a

$$\begin{cases} -9x_1 + 7x_2 + x_3 = -1 \\ 6x_1 - 2x_2 + 2x_3 = 6 \\ 3x_1 - 2x_2 - x_3 = 0 \end{cases}$$

- Moltiplichiamo la prima riga, ovvero equazione, per $m_{2,1} = (-6/9) = -2/3$ e la sottraiamo alla seconda, ottenendo

$$\begin{cases} -9x_1 + 7x_2 + x_3 = -1 \\ 0x_1 + (8/3)x_2 + (8/3)x_3 = 16/3 \\ 3x_1 - 2x_2 - x_3 = 0 \end{cases}$$

- Moltiplichiamo la prima riga, ovvero equazione, per $m_{3,1} = (-1/3) = -1/3$ e la sottraiamo alla terza, ottenendo

$$\begin{cases} -9x_1 + 7x_2 + x_3 = -1 \\ 0x_1 + (8/3)x_2 + (8/3)x_3 = 16/3 \\ 0x_1 + (1/3)x_2 - (2/3)x_3 = -1/3 \end{cases}$$

- Osserviamo che il termine maggiore in modulo della seconda colonna, di indice maggiore o uguale di quello di colonna, é quello della seconda riga visto che $\max(8/3, 1/3) = 8/3$. Con le notazioni introdotte, $\phi(2) = 2$ e quindi non si richiedono scambi di equazioni (si dovrebbe scambiare la seconda equazione con la seconda equazione).

  Moltiplichiamo la seconda riga per $m_{3,2} = (1/3)/(8/3) = 1/8$ e la sottraiamo alla terza, ottenendo

$$\begin{cases} -9x_1 + 7x_2 + x_3 = -1 \\ 0x_1 + (8/3)x_2 + (8/3)x_3 = 16/3 \\ 0x_1 + 0x_2 - x_3 = -1 \end{cases}$$

Il sistema finale, è facile da risolvere mediante sostituzione all'indietro.

- L'ultima equazione ha una sola variabile e una sola incognita e comporta che $x_3 = 1$.
- Inseriamo questo risultato nella penultima equazione e otteniamo

$$(8/3)x_2 + (8/3) \cdot 1 = 16/3$$

  da cui $x_2 = 1$.

- Inseriamo questi risultati nella prima equazione e otteniamo

$$-9x_1 + 7 \cdot 1 + 1 = -1$$

  da cui $x_1 = 1$.

Nota.

L'esempio precedente è stato scritto in forma di sistema di equazioni, ma può essere facilmente riscritto direttamente in forma matriciale.

Passiamo alla fattorizzazione LU della matrice

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

utilizzata nell'esempio precedente per risolvere $Ax = b$, con $b = (0, 6, -1)^T$. Per quanto visto, analizzando il sistema finale del processo, $U = A^{(n)}$ e dunque

$$U = \begin{bmatrix} -9 & 7 & 1 \\ 0 & 8/3 & 8/3 \\ 0 & 0 & -1 \end{bmatrix}.$$

Ricordando i moltiplicatori $m_{2,1} = -2/3$, $m_{3,1} = -1/3$, $m_{3,2} = 1/8$, abbiamo

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -2/3 & 1 & 0 \\ -1/3 & 1/8 & 1 \end{bmatrix}.$$

Nel passo relativo alla prima colonna, avendo scambiato la prima con la terza riga (era $\phi(1) = 3$), la matrice P_1, detta di permutazione, risulta la matrice identica eccetto per le righe 1, 3 che sono nulle ad eccezione di $(P_1)_{1,3} = (P_1)_{3,1} = 1$, da cui

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

mentre nel passo relativo alla seconda colonna, visto che non abbiamo effettuato scambi (in effetti era $\phi(2) = 2$)

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(ovvero la matrice identica). Quindi $P = P_2 \cdot P_1 = I \cdot P_1 = P_1$ e $PA = LU$.

Avendo calcolato la fattorizzazione $PA = LU$, verifichiamo i risultati in Matlab.

```
>> P=[0 0 1; 0 1 0; 1 0 0];
>> A=[3 -2 -1; 6 -2 2; -9 7 1];
>> PA=P*A;
>> L=[1 0 0; -2/3 1 0; -1/3 1/8 1];
>> U=[-9 7 1; 0 8/3 8/3; 0 0 -1];
>> LU=L*U;
>> norm(PA-LU,2)
ans =
    4.9651e-16
>>
```

Si afferma che $||P * A - L * U||_2 \approx 4.9651e - 16$ ovvero che anche numericamente $P * A = L * U$ (la distanza euclidea matriciale tra PA e LU é molto piccola).

Definizione (Predominanza diagonale)

La matrice $A \in \mathbb{C}^{n\times n}$, $A = (a_{i,j})$, è a predominanza diagonale per righe se per ogni $i = 1, \ldots, n$ risulta

$$|a_{i,i}| \geq \sum_{j=1, j\neq i}^{n} |a_{i,j}|$$

e per almeno un indice s si abbia

$$|a_{s,s}| > \sum_{j=1, j\neq s}^{n} |a_{s,j}|.$$

La matrice $A \in \mathbb{C}^{n\times n}$ è a predominanza diagonale stretta per righe se per ogni $i = 1, \ldots, n$ risulta

$$|a_{i,i}| > \sum_{j=1, j\neq i}^{n} |a_{i,j}|.$$

La matrice $A \in \mathbb{C}^{n\times n}$ è a predominanza diagonale (stretta) per colonne se A^T è a predominanza diagonale (stretta) per righe.

Nota. (Matrici che hanno fattorizzazione $A = LU$)

Esistono classi di matrici come quelle

1. *a predominanza diagonale stretta per righe (cf. [11]), o*
2. *a predominanza diagonale stretta per colonne (cf. [11]), o*
3. *simmetriche e definite positive,*

per cui è sempre possibile la fattorizzazione $A = LU$.

Nota. (Fattorizzazione $PA = LU$ anche per matrici non invertibili)

Abbiamo visto che una qualsiasi matrice invertibile possiede una fattorizzazione del tipo $PA = LU$.

In realtà una matrice arbitraria possiede una fattorizzazione di questo tipo (cf. [3, p.145]).

Esempio

La matrice

$$\begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}.$$

è

- *a predominanza diagonale stretta sia per righe che per colonne.*
- *simmetrica in quanto $A = A^T$ e definita positiva avendo per autovalori*

$$\lambda_1 \approx 5.4142, \lambda_2 \approx 4.0000, \lambda_3 \approx 2.5858$$

A tal proposito ricordiamo che una matrice é definita positiva se e solo se ha tutti gli autovalori positivi.

Pseudocodice

Le seguenti righe di codice implementano la prima parte dell'algoritmo di eliminazione gaussiana senza pivoting, cui deve seguire la risoluzione di un sistema triangolare superiore mediante sostituzione all'indietro.

```
for k=1:n-1
  for i=k+1:n
       m(i,k)=a(i,k)/a(k,k);
       for j=k+1:n
         a(i,j)=a(i,j)-m(i,k)*a(k,j);
       end
       b(i)=b(i)-m(i,k)*b(k);
  end
end
```

Notiamo che al termine del processo la matrice A coincide con L e che il codice produce errore se durante il processo qualche `a(k,k)` *é nullo.*

Il costo della eliminazione gaussiana con pivoting è essenzialmente lo stesso di quello senza pivoting, visto che a parte il calcolo di un massimo, si tratta esclusivamente di scambiare al più due righe per ogni indice $k = 1, \ldots, n-1$.

Teorema (Complessità fattorizzazione PA=LU)

Il numero di operazioni moltiplicative richiesto per effettuare la fattorizzazione PA=LU è al più dell'ordine di $n^3/3$.

Dimostrazione. (Facoltativa)

Dal pseudocodice, si vede che il numero di operazioni moltiplicatorie necessarie per effettuare la parte di eliminazione gaussiana, precedente alla sostituzione, è

$$\sum_{p=1}^{n-1}\sum_{k=p+1}^{n}(n-p+2) = \sum_{p=1}^{n-1}\sum_{k=p+1}^{n}2+\sum_{p=1}^{n-1}\sum_{k=p+1}^{n}(n-p) = 2\sum_{p=1}^{n-1}(n-p)+\sum_{p=1}^{n-1}(n-p)^2 = 2\sum_{k=1}^{n-1}k+\sum_{k=1}^{n-1}k^2$$

Nell'ultima riga, abbiamo osservato che posto $k = n-p$,

a) $\sum_{p=1}^{n-1}(n-p) = \sum_{k=1}^{n-1}k$,

b) $\sum_{p=1}^{n-1}(n-p)^2 = \sum_{k=1}^{n-1}k^2$.

Inoltre

- *si ha* $2\sum_{k=1}^{n-1}k = 2\frac{(n-1)n}{2} \approx n^2$
- *per quanto riguarda* $\sum_{k=1}^{n-1}k^2$ *abbiamo che*

$$\frac{(n-1)^3}{3} = \int_0^{n-1}x^2dx \le \sum_{k=1}^{n-1}k^2 \le \int_1^n x^2dx = \frac{n^3-1}{3} \tag{12}$$

da cui, utilizzando la definizione di asintotico e il teorema del confronto, $\sum_{k=1}^{n-1}k^2 \sim \frac{n^3}{3}$.

In definitiva,

$$2\sum_{k=1}^{n-1}k+\sum_{k=1}^{n-1}k^2 \approx n^2+\frac{n^3}{3} \approx \frac{n^3}{3}.$$

n	cputime
1	$2.28081e-04$
2	$1.14614e-03$
4	$1.14304e-04$
8	$4.85808e-05$
16	$4.79512e-04$
32	$1.15395e-04$
64	$2.23282e-03$
128	$2.85307e-03$
256	$2.95903e-03$
512	$1.71150e-02$
1024	$4.20015e-02$
2048	$2.45661e-01$
4096	$1.55579e+00$
8192	$9.08754e+00$
16384	$6.18414e+01$

Tabella: Secondi necessari per risolvere un sistema $Ax = b$, con una generica $A \in \mathbb{R}^{n \times n}$, mediante eliminazione gaussiana su un Mac Book Pro, con processore 2,7 GHz Intel Core i5 e 16 GB di memoria.

Nota. (Importante)

1 Visto che risolvere ognuno dei sistemi triangolari costa circa $n^2/2$ mentre la parte precedente necessita di circa $n^3/3$ operazioni moltiplicative, deduciamo che risolvere il sistema $Ax = b$ con l'eliminazione gaussiana necessita di circa $n^3/3$ operazioni moltiplicative (si ricordi che $O(n^3/3) + 2O(n^2/2) = O(n^3/3)$).

2 La complessità della fattorizzazione $PA = LU$ è essenzialmente la stessa della eliminazione gaussiana, ovvero $\frac{n^3}{3}$.

3 In generale l'algoritmo di eliminazione gaussiana per la risoluzione di un sistema lineare $Ax = b$ con A invertibile, puó essere implementato come segue

1. *Calcola la fattorizzazione $PA = LU$. Osserviamo che essendo P invertibile (si dimostra che $|\det(P)| = 1$)*

$$Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb \Leftrightarrow Ly = \tilde{b} \text{ con } Ux = y, \tilde{b} = Pb.$$

2. *Calcola $\tilde{b} = Pb$.*
3. *Risolvi il sistema triangolare inferiore $Ly = \tilde{b}$ con la sostituzione in avanti.*
4. *Risolvi il sistema triangolare superiore $Ux = y$ con la sostituzione all'indietro.*

La complessità dell'algoritmo così implementato è di $\frac{n^3}{3}$ operazioni moltiplicative, ma offre il vantaggio di avere a disposizione la fattorizzazione $PA = LU$.

Nota. (Decomposizione di Cholesky)

Se la A matrice è simmetrica e definita positiva, ovvero

- $A = A^T$,
- *tutti gli autovalori di A sono positivi,*

è più conveniente utilizzare la decomposizione di Cholesky [1, p.524], in cui

$$A = LL^T$$

con $L = (l_{i,j})$ triangolare inferiore con elementi diagonali $l_{k,k} > 0$, $k = 1, \ldots, n$.

La determinazione di questa fattorizzazione richiede approssimativamente $\frac{n^3}{6}$ operazioni moltiplicative.

Di conseguenza, per calcolare la soluzione di $Ax = b$, basta risolvere i due sistemi triangolari $Ly = b$, $L^T x = y$. Quindi in generale la soluzione del sistema lineare si può ottenere in

$$O\left(\frac{n^3}{6}\right) + O\left(\frac{n^2}{2}\right) + O\left(\frac{n^2}{2}\right) = O\left(\frac{n^3}{6}\right)$$

operazioni moltiplicative.

Esempio

La matrice simmetrica

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

è definita positiva, in quanto l'autovalore minimo è $\lambda_{min} = 0.271\ldots > 0$.
Di conseguenza possiede una decomposizione di Cholesky $A = LL^T$. *In questo caso particolare risulta:*

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

In Matlab la routine `chol` *calcola la trasposta di tale matrice L. Vediamo in effetti che i risultati sono corretti.*

```
>> A=gallery('minij',5);
>> LT=chol(A); L=LT';
>> L
L =
     1     0     0     0     0
     1     1     0     0     0
     1     1     1     0     0
     1     1     1     1     0
     1     1     1     1     1
>> norm(A-L*L',inf)
ans =
     0
```

Problema.

Di seguito mostriamo alcune applicazione della fattorizzazione LU, ovvero

- *il calcolo del determinante di una matrice,*
- *la determinazione dell'inversa di una matrice quadrata non singolare,*

paragonandoli con altri metodi.

A tal proposito ricordiamo che

- *Il metodo di eliminazione di Gauss può essere usato per calcolare il determinante di una matrice quadrata.*
- *Il costo computazionale della formula ricorsiva di Laplace per il determinante è di circa* $2n!$ *flops (cf. [17]).*

Osserviamo che

- *se* $PA = LU$, *visto che* $P^{-1} = P^T$ *abbiamo che* $A = P^T LU$ *e quindi per il teorema di Binet per cui* $\det(M * N) = \det(M) \cdot \det(N)$ *qualora* $M, N \in \mathbb{R}^{n \times n}$

$$\det(A) = \det(P^T LU) = \det(P^T) \cdot \det(L) \cdot \det(U).$$

- *se* s *é il numero di scambi effettuati dal pivoting, allora* $\det(P^T) = (-1)^s$;
- *visto che il determinante di una matrice triangolare* $T = (t_{i,j}) \in \mathbb{C}^{n \times n}$ *è*

$$\det(T) = \prod_{k=1}^{n} t_{k,k}$$

abbiamo che

- *da* $l_{k,k} = 1$, $k = 1, \ldots, n$, *si ricava* $\det(L) = 1$,
- $\det(U) = \prod_{k=1}^{n} u_{k,k}$,

e quindi

$$\det(A) = (-1)^s \cdot \prod_{k=1}^{n} u_{k,k}.$$

In definitiva, visto che

- *il calcolo di* $\prod_{k=1}^{n} u_{k,k}$ *necessita solo di* $n-1$ *operazioni moltiplicative,*
- *la fattorizzazione* $PA = LU$ *ha complessità dell'ordine di* $n^3/3$,

il calcolo del determinante può essere effettuato con circa $n^3/3$ *operazioni moltiplicative.*

Nella tabella paragoniamo il tempo di calcolo necessario a un supercomputer [15] come Summit che eroga 200 petaflops, ovvero $2 \cdot 10^{17}$ *operazioni al secondo, per calcolare il determinante di una matrice generica di ordine n. Si tenga conto che si presume che l'età dell'universo sia di circa "*$13.82e+9$*" anni.*

n	CPU_L	CPU_{LU}	n	CPU_L	CPU_{LU}
5	$6.0e-16$ *sec*	$2.1e-16$ *sec*	75	$1.2e+92$ *anni*	$7.0e-13$ *sec*
10	$1.8e-11$ *sec*	$1.7e-15$ *sec*	100	$4.7e+140$ *anni*	$1.7e-12$ *sec*
25	$9.0e+02$ *anni*	$2.6e-14$ *sec*	125	$9.4e+191$ *anni*	$3.3e-12$ *sec*
50	$1.8e+42$ *anni*	$2.1e-13$ *sec*	150	$2.9e+245$ *anni*	$5.6e-12$ *sec*

Tabella: *In questa tabella paragoniamo il tempo di calcolo* CPU_L, CPU_{LU} *necessario a un supercomputer come Summit che eroga 200 petaflops, ovvero* $2 \cdot 10^{17}$ *operazioni al secondo, per calcolare il determinante di una matrice generica di ordine n, rispettivamente con la regola di Laplace e via fattorizzazione* $PA = LU$.

Per calcolare l'inversa di una matrice $A \in \mathbb{C}^{n \times n}$ con il metodo dei cofattori [13] serve calcolare $n^2 + 1$ determinanti di matrici estratte da A eliminando la i-sima riga e j-sima colonna, con $i, j = 1, \ldots, n$, più il determinante di A.

Se brutalmente per ognuna di queste $n^2 + 1$ matrici di dimensione $(n-1) \times (n-1)$ si usa la fattorizzazione LU, avente complessitá $\frac{(n-1)^3}{3}$, necessitano

$$C_{COF} = (n^2 + 1) \cdot \frac{(n-1)^3}{3} \approx \frac{n^5}{3}$$

operazioni moltiplicative.

Alcune applicazioni della fattorizzazione LU: calcolo dell'inversa di una matrice

Si supponga per semplicità sia $A = LU$ e si risolvano i sistemi

$$Ax^{(j)} = e^{(j)}, \quad j = 1, \ldots, n, \quad e^{(j)} = (0, \ldots, 0, 1, 0 \ldots, 0).$$

Allora A^{-1} è la matrice la cui j-sima colonna corrisponde a $x^{(j)}$.

Per eseguire i calcoli necessitano

- *una fattorizzazione LU per cui servono circa $n^3/3$ operazioni,*
- *risolvere n sistemi lineari $Ax^{(j)} = e^{(j)}$, ognuno richiedente una sostituzione in avanti $Ly^{(j)} = e^{(j)}$ e una all'indietro $Ux^{(j)} = y^{(j)}$.*

  (a) Nel caso di $Ly^{(j)} = e^{(j)}$, si vede facilmente che le prime $j-1$ componenti di $y^{(j)}$ sono nulle (si veda ad esempio (11) ponendo $b_1, \ldots, b_{j-1} = 0$) e quindi per risolvere $Ly^{(j)} = e^{(j)}$ servono circa $(n-j+1)^2/2$ operazioni moltiplicative.

  Quindi la risoluzione di tutti i sistemi $Ly^{(j)} = e^{(j)}$, $j = 1, \ldots, n$ ha complessità

$$\sum_{j=1}^{n} (n-j+1)^2/2 = \sum_{j=1}^{n} j^2/2 = (1/2) \sum_{j=1}^{n} j^2 \approx \frac{n^3}{6}$$

  (b) Nel caso di $Ux^{(j)} = y^{(j)}$, il costo computazionale è invece di circa $n^2/2$ operazioni moltiplicative per ogni $j = 1, \ldots, n$, ovvero in totale $n^3/2$ operazioni.

Quindi, per calcolare l'inversa di A con fattorizzazione LU servono circa

$$\frac{n^3}{3} + \frac{n^3}{6} + \frac{n^3}{2} = \frac{(2+1+3)n^3}{6} = n^3$$

operazioni moltiplicative, e quindi con complessità inferiore a quella richiesta da una implementazione brutale del metodo dei cofattori che è risultata essere $n^5/3$.

Esempio

Si calcoli l'inversa della matrice

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

Essendo

- la soluzione x_1 di $Ax = [1, 0, 0]^T$ è

$$x_1 = \begin{bmatrix} 1.185185185185185e+00 \\ -2.222222222222222e-01 \\ -3.703703703703703e-02 \end{bmatrix}.$$

- la soluzione x_2 di $Ax = [0, 1, 0]^T$ è

$$x_2 = \begin{bmatrix} -4.444444444444446e-01 \\ 3.333333333333334e-01 \\ -1.111111111111111e-01 \end{bmatrix}.$$

- la soluzione x_3 di $Ax = [0, 0, 1]^T$ è

$$x_3 = \begin{bmatrix} -9.259259259259259e-01 \\ 1.111111111111111e-01 \\ 1.851851851851852e-01 \end{bmatrix}.$$

e quindi essendo $A^{-1} = [x_1, \ x_2, \ x_3]$

$$A^{-1} = \begin{bmatrix} 1.185185185185185e+00 & -4.444444444444446e-01 & -9.259259259259259e-01 \\ -2.222222222222222e-01 & 3.333333333333334e-01 & 1.111111111111111e-01 \\ -3.703703703703703e-02 & -1.111111111111111e-01 & 1.851851851851852e-01 \end{bmatrix}.$$

In Matlab

```
>> A=[2 5 7; 1 5 2; 1 4 8]; B=inv(A); format long e
>> B
B =
     1.185185185185185e+00    -4.444444444444445e-01    -9.259259259259258e-01
    -2.222222222222222e-01     3.333333333333334e-01     1.111111111111111e-01
    -3.703703703703703e-02    -1.111111111111111e-01     1.851851851851852e-01
>>
```

Commento

I metodi come l'eliminazione gaussiana calcolano in aritmetica esatta la soluzione x^* di $Ax = b$, $A \in \mathbb{C}^{n\times n}$ in un numero finito di passi, sono detti metodi diretti.

D'altra parte

- *la matrice n è di grandi dimensioni e quindi ci sono sia problemi di immagazzinamento dati che di complessità computazionale (ovvero $n^3/3$ può essere proibitivo in termini di costo);*
- *la matrice A può possedere molti zeri, che si verifica spesso non aiutare a migliorare la performance del metodo diretto.*

Una alternativa consiste nei metodi iterativi che tipicamente non è detto determinino la soluzione esatta, ma possono approssimarla arbitrariamente bene. Usualmente ogni iterazione calcola un prodotto matrice-vettore, e quindi ha complessità di circa n^2 operazioni moltiplicative.

Se i metodi iterativi forniscono l'approssimazione della soluzione richiesta in un $k \ll n/3$ iterazione, essendo $kn^2 \ll n^3/3$ risultano convenienti rispetto alla generale fattorizzazione LU.

Sia $A \in \mathbb{C}^{n\times n}$ e definiamo lo *splitting*

$$A = D - E - F$$

dove

- D è una matrice diagonale, non singolare
- E è triangolare inferiore con elementi diagonali nulli,
- F è triangolare superiore con elementi diagonali nulli.

Esempio

Sia A la matrice

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

allora puó essere descritta in termini di D, E, F con

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, E = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -5 & 0 & 0 & 0 \\ -8 & -9 & 0 & 0 \\ -3 & -4 & -5 & 0 \end{bmatrix}, F = \begin{bmatrix} 0 & -2 & -3 & -4 \\ 0 & 0 & -7 & -2 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Se $A \in \mathbb{C}^{n\times n}$ è una matrice quadrata, risulta importante anche lo splitting

$$A = P - N, \ \det(P) \neq 0. \qquad (13)$$

Allora

$$Ax = b \Leftrightarrow (P-N)x = b \Leftrightarrow Px - Nx = b \Leftrightarrow Px = Nx + b$$

da cui, per l'invertibilità di P, moltiplicando ambo i membri per P^{-1}, ricaviamo

$$x = P^{-1}Px = P^{-1}(Nx+b) = \underbrace{P^{-1}N}_{B}x + \underbrace{P^{-1}b}_{c} = Bx + c := \phi(x). \qquad (14)$$

Quindi, posti $B := P^{-1}N$, $c := P^{-1}b$, la soluzione x^* di $Ax = b$ risolve pure $x = Bx + c$ (e viceversa).

Conseguentemente, vista la struttura di punto fisso $x = \phi(x)$ con $\phi(x) = Bx + c$, si considera il metodo iterativo (stazionario)

$$x^{(k+1)} = \phi(x^{(k)}) = Bx^{(k)} + c \qquad (15)$$

(Metodi di Jacobi e Gauss-Seidel)

Sia $A = D - E - F$ con

1. *D matrice diagonale,*
2. *E triangolare inferiore,*
3. *F triangolare superiore.*

Allora

- *il metodo di Jacobi corrisponde a scegliere, se D è invertibile,*

$$P = D,\ N = E + F$$

- *il metodo di Gauss-Seidel corrisponde a scegliere, se D è invertibile,*

$$P = D - E,\ N = F$$

Nota.

Il metodo di Jacobi fu pubblicato dall'autore nel 1845 in un giornale di astronomia, mentre quello che conosciamo come metodo di Gauss-Seidel fu il risultato di una collaborazione di Seidel con Jacobi nel 1874. Gauss aveva già descritto un metodo analogo nel 1845 [4, p.466].

Commento (Osservazioni sul metodo di Jacobi e Gauss-Seidel)

Se $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}$, $det(A) \neq 0$ *ed* $Ax = b$, *visto che* $(Ax)_i = \sum_{j=1}^{n} a_{i,j}x_j$, *allora*

$$\sum_{j=1}^{n} a_{i,j}x_j = b_i, \ \ i = 1, \dots, n$$

e quindi, per $i = 1, \dots, n$, *evidenziando il termine* i*-simo,*

$$\sum_{j=1}^{i-1} a_{i,j}x_j + a_{i,i}x_i + \sum_{j=i+1}^{n} a_{i,j}x_j = \sum_{i=1}^{n} a_{i,j}x_j = b_i,$$

da cui portando a secondo membro $\sum_{j=1}^{i-1} a_{i,j}x_j + \sum_{j=i+1}^{n} a_{i,j}x_j$

$$a_{i,i}x_i = b_i - \sum_{j=1}^{i-1} a_{i,j}x_j - \sum_{j=i+1}^{n} a_{i,j}x_j,$$

e quindi nel caso $a_{i,i} \neq 0$ *per* $i = 1, \dots, n$

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j - \sum_{j=i+1}^{n} a_{i,j}x_j}{a_{i,i}}, \ \ i = 1, \dots, n.$$

Da

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j - \sum_{j=i+1}^{n} a_{i,j}x_j}{a_{i,i}}, \quad i = 1, \ldots, n.$$

il metodo di Jacobi si ricava ponendo nelle occorrenze del membro di

- *sinistra* $x_i^{(k+1)}$ *al posto di* x_i,
- *destra* $x_j^{(k)}$ *al posto di ogni occorrenza di* x_j *(con* $j = 1, \ldots, n$*),*

ricavando

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)})/a_{ii}, \quad i = 1, \ldots, n, \tag{16}$$

Nota.

Il metodo ottenuto e' lo stesso descritto matricialmente come

$$x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b$$

con $P = D$ *e* $N = E + F$.

Esempio

Applicando

$$x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i - \left(\sum_{j<i} a_{i,j} x_j^{(k)} + \sum_{j>i} a_{i,j} x_j^{(k)} \right) \right), \quad i = 1, \ldots, n$$

al caso di $A \in \mathbb{C}^{3\times 3}$*, otteniamo, supponendo disponibile* $x^{(k)}$*:*

1. $x_1^{(k+1)} = \frac{1}{a_{1,1}}(b_1 - a_{1,2}x_2^{(k)} - a_{1,3}x_3^{(k)})$;
2. $x_2^{(k+1)} = \frac{1}{a_{2,2}}(b_2 - a_{2,1}x_1^{(k)} - a_{2,3}x_3^{(k)})$;
3. $x_3^{(k+1)} = \frac{1}{a_{3,3}}(b_3 - a_{3,1}x_1^{(k)} - a_{3,2}x_2^{(k)})$.

Si nota che

- *dopo aver calcolato* $x_1^{(k+1)}$ *tale nuova approssimazione della prima componente* x_1^* *della soluzione* x^* *non é sfruttata nel determinare* $x_2^{(k+1)}$*;*
- *dopo aver calcolato* $x_2^{(k+1)}$ *tale nuova approssimazione della prima componente* x_2^* *della soluzione* x^* *non é sfruttata nel determinare* $x_3^{(k+1)}$*.*

Ogni $x_j^{(k+1)}$*,* $j = 1, 2, 3$ *é calcolato direttamente dall'iterazione precedente* $x^{(k)} = (x_i^{(k)})_i$*, senza richiedere la conoscenza di qualche* $x_i^{(k+1)}$ *con* $i = 1, \ldots, j-1$*.*

Da

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j - \sum_{j=i+1}^{n} a_{i,j}x_j}{a_{i,i}}, \quad i = 1, \ldots, n.$$

il metodo di Gauss-Seidel si ricava ponendo nelle occorrenze del membro di

- *sinistra* $x_i^{(k+1)}$ *al posto di* x_i,
- *destra nella prima sommatoria* $x_i^{(k+1)}$ *al posto di* x_i *e nella seconda sommatoria* $x_i^{(k)}$

ricavando

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)})/a_{ii}, \quad i = 1, \ldots, n, \qquad (17)$$

Nota.

Il metodo ottenuto e' lo stesso descritto matricialmente come

$$x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b, \quad P = D - E, \quad N = F.$$

Una sua generalizzazione porta ai metodi SOR (cf. [14]).

Esempio

Applicando

$$x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i - \left(\sum_{j<i} a_{i,j} x_j^{(k+1)} + \sum_{j>i} a_{i,j} x_j^{(k)} \right) \right), \quad i = 1, \ldots, n$$

al caso di $A \in \mathbb{C}^{3\times 3}$*, otteniamo:*

1. $x_1^{(k+1)} = \frac{1}{a_{1,1}}(b_1 - a_{1,2}x_2^{(k)} - a_{1,3}x_3^{(k)})$;
2. $x_2^{(k+1)} = \frac{1}{a_{2,2}}(b_2 - a_{2,1}x_1^{(k+1)} - a_{2,3}x_3^{(k)})$;
3. $x_3^{(k+1)} = \frac{1}{a_{3,3}}(b_3 - a_{3,1}x_1^{(k+1)} - a_{3,2}x_2^{(k+1)})$.

Si nota che

- *dopo aver calcolato* $x_1^{(k+1)}$ *tale nuova approssimazione della prima componente* x_1^* *della soluzione* x^* *é sfruttata nel determinare* $x_2^{(k+1)}$;
- *dopo aver calcolato* $x_2^{(k+1)}$ *tale nuova approssimazione della prima componente* x_2^* *della soluzione* x^* *é sfruttata nel determinare* $x_3^{(k+1)}$.

Nota. (Consistenza)

Consideriamo un metodo iterativo stazionario

$$x^{(k+1)} = Bx^{(k)} + c, \quad k = 0, 1, 2, \ldots$$

per l'approssimazione della soluzione x^* del sistema lineare $Ax = b$, con A invertibile, ovvero $\det(A) \neq 0$.

1. *Vogliamo che il metodo iterativo stazionario converga a x^*, ovvero che*

$$\lim_{k \to +\infty} \|x^{(k)} - x^*\| = 0$$

per qualche norma indotta, e che tale x^* risolva $Ax = b$.

2. *Osserviamo che se $x^{(k)} \to v^*$ allora pure $x^{(k+1)} \to v^*$ da cui,*

$$\begin{aligned} v^* &= \lim_{k \to +\infty} x^{(k+1)} = \lim_{k \to +\infty} (Bx^{(k)} + c) = \lim_{k \to +\infty} Bx^{(k)} + \lim_{k \to +\infty} c \\ &= B \lim_{k \to +\infty} x^{(k)} + c = Bv^* + c, \end{aligned}$$

ovvero v^* risolve il problema di punto fisso $x = \phi(x)$ con

$$\phi(x) = Bx + c.$$

Siccome desideriamo che $x^{(k)} \to x^*$, con x^* soluzione di $Ax = b$, necessariamente si chiede che tale x^* sia l'unica soluzione del problema di punto fisso $x = Bx + c$. Qualora $\|B\| < 1$, tale proprietá detta di consistenza é garantita tanto dal metodo di Jacobi che da quello di Gauss-Seidel.

Teorema (Convergenza globale dei metodi iterativi)

Siano

- $\|\cdot\|$ *una norma matriciale indotta da una norma vettoriale,*
- $B \in \mathbb{C}^{n\times n}$, *con* $\|B\| < 1$,
- *si supponga che la soluzione* x^* *di* $Ax = b$ *sia l'unica soluzione di* $x = Bx + c$.

Allora qualsiasi sia $x_0 \in \mathbb{C}^n$, *la successione definita da* $x^{(k+1)} = Bx^{(k)} + c$, *con* $k = 0, 1, 2, \ldots$, *converge a* x^*.

Dimostrazione. (Facoltativa)

Da $\|B\| < 1$, *abbiamo che la matrice* $I - B$ *è invertibile e quindi* x^* *è l'unica soluzione di* $(I - B)x = c$ *ovvero di* $x = Bx + c$. *Da*

$$x^{(k+1)} = Bx^{(k)} + c, \quad k = 0, 1, 2, \ldots, \quad x^* = Bx^* + c, \quad k = 0, 1, 2, \ldots$$

necessariamente, sottraendo membro a membro, per la linearità di B,

$$x^{(k+1)} - x^* = (Bx^{(k)} + c) - (Bx^* + c) = B(x^{(k)} - x^*).$$

Quindi, da $\|My\| \leq \|M\|\|y\|$, $\|MN\| \leq \|M\|\|N\|$, *per M, N, y arbitrari,*

$$\begin{aligned} 0 \leq \|x^{(k)} - x^*\| &= \|B(x^{(k-1)} - x^*)\| \leq \|B\|\|x^{(k-1)} - x^*\| \leq \|B\|^2\|x^{(k-2)} - x^*\| \\ &\leq \|B\|^3\|x^{(k-3)} - x^*\| \leq \ldots \leq \|B\|^{k+1}\|x^{(0)} - x^*\| \end{aligned}$$

e per il teorema del confronto, visto che $\|B\|^k \to 0$, $\lim_k \|x^{(k)} - x^*\| = 0$ *ovvero* $x^{(k)} \to x^*$.

Teorema (Convergenza globale dei metodi iterativi)

Siano

- $B \in \mathbb{C}^{n\times n}$,
- *si supponga che la soluzione* x^* *di* $Ax = b$ *sia l'unica soluzione di* $x = Bx + c$.
- *sia* $\rho(B) = \max_{k=1,\dots,n} |\lambda_k|$ *il raggio spettrale di* B *(dove* $\lambda_1, \dots, \lambda_n$ *sono gli autovalori di* B*).*

Allora qualsiasi sia $x^{(0)} \in \mathbb{C}^n$, *la successione* $\{x^{(k)}\}$ *definita da*

$$x^{(k+1)} = Bx^{(k)} + c, \quad k = 0, 1, 2, \dots$$

converge a x^* *se e solo se* $\rho(B) < 1$.

Nota.

1. *Si noti che in questo asserto compare, a differenza del precedente un se e solo se.*
2. *Si noti compare il raggio spettrale di B (e non di A).*
3. *Si noti che richiede il calcolo dell'autovalore di modulo massimo di B, un problema potenzialmente complicato. Di solito, noto la struttura della matrice A ci sono dei teoremi che offrono risultati sul raggio spettrale di B.*

A partire dai precedenti teoremi si è potuto dimostrare (non facile!)

Teorema (Convergenza Jacobi per matrici a predominanza diagonale stretta)

Se la matrice A è a predominanza diagonale (stretta) per righe, allora il metodo di Jacobi converge.

Teorema (Convergenza Gauss-Seidel per matrici a predominanza diagonale stretta)

Se la matrice A è a predominanza diagonale (stretta) per righe, allora il metodo di Gauss-Seidel converge.

Teorema (Convergenza G.-S. per matrici simmetriche e definite positive)

Se la matrice A è simmetrica e definita positiva allora il metodo di Gauss-Seidel converge.

Nota. (Grandezza raggio spettrale)

In generale, con stime quantitative, si vede che sono da preferire metodi con raggio spettrale $\rho(B)$ il piú piccolo possibile.

Esempio (Matrice a predominanza diagonale stretta)

Consideriamo il sistema lineare $Ax = b$ dove

$$A = \begin{bmatrix} 11 & -5 & -5 \\ 5 & 12 & 6 \\ 6 & -4 & 11 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 23 \\ 13 \end{bmatrix},$$

che ha soluzione $x^* = [1, 1, 1]^T$ (cf. [3, p.252]). Si dica se convergono i metodi di Jacobi e Gauss-Seidel.

- La matrice A è a predominanza diagonale in senso stretto e quindi tanto il metodo di Jacobi, quanto quello di Gauss-Seidel, risultano convergenti. In effetti $\rho(B_J) \approx 0.79$, $\rho(B_{GS}) \approx 0.83$ e questo si dimostra implicare che asintoticamente il metodo di Jacobi converge piú rapidamente di quello di Gauss-Seidel.
- Gli errori assoluti $\|x^{(k)} - x^*\|_\infty$ compiuti rispettivamente dal metodo di Jacobi e Gauss-Seidel sono esposti nella tabella che segue, da cui si evince la convergenza dei metodi.

n	E_J	E_{GS}
10	$1.6e-01$	$2.2e-01$
30	$1.5e-03$	$6.2e-03$
50	$1.4e-05$	$1.7e-04$
70	$1.3e-07$	$4.8e-06$
90	$1.2e-09$	$1.3e-07$
110	$1.1e-11$	$3.8e-09$
130	$1.0e-13$	$1.1e-10$
150	$1.1e-15$	$2.9e-12$
170	$-$	$8.2e-14$
190	$-$	$2.2e-15$

Tabella: Errori assoluti $\|x^{(k)} - x^*\|_\infty$ compiuti dalle n-sime iterazioni di Jacobi e Gauss-Seidel, nel risolvere un certo problema $Ax = b$ con A a predominanza diagonale stretta, con vettore iniziale $x_0 = 0 \in \mathbb{R}^3$.

Esempio (Matrice definita positiva)

Si studi la convergenza dei metodi di Jacobi e Gauss-Seidel, relativamente al sistema lineare $Ax = b$ dove

$$A = \left[\begin{array}{ccc|ccc|ccc} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ \hline -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{array}\right], \quad b = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix},$$

che ha soluzione $x^* = [1, \ldots, 1]^T \in \mathbb{R}^9$.

La matrice A è un esempio di matrice di Poisson ed è definita positiva, in quanto il suo autovalore più piccolo vale

$$\lambda_{\min} \approx 1.171572875253810e+00 > 0.$$

Le linee mostrano come $A \in \mathbb{R}^{9\times 9}$ segua *pattern* particolari con *blocchi* pari a

$$B = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}, \quad -I = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad 0_{3\times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

Gli errori assoluti $\|x^{(k)} - x^*\|_\infty$ compiuti rispettivamente dal metodo di Jacobi e Gauss-Seidel partendo da $x_0 = [0, \ldots, 0]^T \in \mathbb{R}^9$. Entrambi i metodi convergono, e per quanto concerne il metodo di Gauss-Seidel, ciò era garantito dal teorema di convergenza, in quanto A é simmetrica e definita positiva.

Si vede che $\rho(B_J) \approx 0.7071$, mentre $\rho(B_{GS}) \approx 0.5$ e quindi ci si aspetta che asintoticamente il metodo di Gauss-Seidel converga in meno iterazioni di quello di Jacobi.

n	E_J	E_{GS}
10	$4.7e-02$	$1.8e-03$
30	$4.6e-05$	$1.7e-09$
50	$4.5e-08$	$1.8e-15$
70	$4.4e-11$	–
90	$4.3e-14$	–

Tabella: Errori assoluti $\|x^{(k)} - x^*\|_\infty$ compiuti dalle n-sime iterazioni di Jacobi e Gauss-Seidel, nel risolvere un certo problema $Ax = b$ con A matrice di Poisson, con vettore iniziale $x_0 = 0 \in \mathbb{R}^9$. convergenti.

Esempio (Mancata convergenza dei metodi di Jacobi e Gauss-Seidel)

Consideriamo il sistema $Ax = b$ con

$$A = \begin{bmatrix} 3 & -2 & -1 \\ 6 & -2 & 2 \\ -9 & 7 & 1 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 6 \\ -1 \end{bmatrix},$$

avente soluzione $x^* = (1,1,1)^T$. Si verifichi la mancata convergenza globale dei metodi di Jacobi e Gauss-Seidel partendo da $x^{(0)} = (1-\epsilon, 1-\epsilon, 1-\epsilon)^T$ con $\epsilon = 10^{-14}$.

- Utilizziamo i metodi di Jacobi e Gauss-Seidel con dato iniziale

$$x^{(0)} = (1 - 10^{-14}, 1 - 10^{-14}, 1 - 10^{-14})^T$$

  e quindi molto prossimo al vettore soluzione x^* in quanto $\|x^* - x^{(0)}\|_\infty = 10^{-14}$.
- Entrambi i metodi citati, pur partendo da un tal $x^{(0)}$, non convergono, come si può vedere dalla relativa tabella. In effetti, si verifica che $1 < \rho(B_J) \approx 8.66$, $1 < \rho(B_{GS}) \approx 9.62$, e quindi entrambi i metodi iterativi non sono globalmente convergenti.

n	E_J	E_{GS}
10	$2.6e-12$	$1.4e-05$
30	$6.7e-09$	$6.3e+14$
50	$1.8e-05$	$2.9e+34$
70	$4.5e-02$	$1.4e+54$
90	$1.2e+02$	$6.3e+73$
110	$2.9e+05$	$2.9e+93$
130	$7.3e+08$	$1.4e+113$
150	$1.8e+12$	$6.3e+132$
170	$4.4e+15$	$2.9e+152$

Tabella: Errori assoluti $\|x^{(k)} - x^*\|_\infty$ compiuti dalle n-sime iterazioni di Jacobi e Gauss-Seidel, nel risolvere un certo problema $Ax = b$ con A matrice utilizzata come esempio per l'eliminazione gaussiana, con vettore iniziale $x^{(0)} = (1 - 10^{-14}, 1 - 10^{-14}, 1 - 10^{-14})^T \in \mathbb{R}^3$.

Sia $\epsilon > 0$ una tolleranza fissata dall'utente. Il metodo iterativo

$$x^{(k+1)} = Bx^{(k)} + c$$

viene arrestato utilizzando

- *il criterio dello step, ovvero si interrompe il processo quando*

$$\|x^{(k+1)} - x^{(k)}\| \leq \epsilon,$$

- *criterio del residuo, ovvero si interrompe il processo quando*

$$\|b - Ax^{(k)}\| \leq \epsilon,$$

- *criterio del residuo relativo, ovvero si interrompe il processo quando*

$$\frac{\|b - Ax^{(k)}\|}{\|b\|} \leq \epsilon, \quad b \neq 0_{\mathbb{C}^n}$$

Si consideri il sistema lineare di 3 equazioni in due incognite

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 - x_2 = 0 \\ x_1 + 3x_2 = 0 \end{cases}$$

Risulta chiaro che il sistema non ha soluzione. Infatti l'unica soluzione delle prime due equazioni è $x_1 = x_2 = 1/2$, che però non verifica $x_1 + 3x_2 = 0$.

In termini matriciali, il problema $Ax = b$, non ha soluzione, dove

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 3 \end{pmatrix}, \; b = (1, 0, 0)^T.$$

In alternativa, risulta quindi ragionevole calcolare $x^* = (x_1^*, x_2^*)$ tale che

$$\gamma := \min_{x \in \mathbb{R}^n} \|b - Ax\|_2 = \|b - Ax^*\|_2 \tag{18}$$

dove

$$\|u\|_2 = \sqrt{\sum_i u_i^2}.$$

La prima questione riguarda l'esistenza e unicità di tale minimo x^*.

Teorema (cf. [3], p. 432)

Sia X l'insieme dei vettori di $\mathbb{R}^n$ tali che $x^* \in X$ se e solo se

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2 = \|b - Ax^*\|_2 \tag{19}$$

Supponiamo $A \in \mathbb{R}^{m \times n}$ con $m \geq n$ e $b \in \mathbb{R}^m$. Valgono le seguenti proprietà

1. *$x \in X$ se e solo se*
$$A^T Ax = A^T b, \tag{20}$$
cioé x risolve il sistema delle equazioni normali.
2. *X è un insieme non vuoto, chiuso e convesso (ovvero se $u, v \in X$ allora $\theta u + (1-\theta)v \in X$, $\theta \in [0,1]$.*
3. *Esiste $x^* \in X$ tale che*
$$\|x^*\|_2 = \min_{x \in X} \|x\|_2.$$
Tale x^* è detto soluzione di minima norma.
4. *L'insieme X si riduce ad un solo elemento x^* se e solo se la matrice A ha rango massimo.*

Si osservi che tale teorema vale pure sostituendo $\mathbb{C}$ a ogni occorrenza di $\mathbb{R}$.

Di conseguenza se $A \in \mathbb{R}^{m\times n}$, con $m \geq n$ è una matrice di rango massimo n (ovvero, le colonne di A sono n vettori linearmente indipendenti di $\mathbb{R}^m$ e quindi si puó *estrarre*, eliminando righe, una matrice $n \times n$ non singolare), allora la soluzione del sistema sovradeterminato $Ax = b$ ai minimi quadrati,

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2 = \min_{x\in\mathbb{R}^n} \sqrt{\sum_{i=1}^{m}(b_i - \sum_{j=1}^{n} a_{i,j}x_j)^2} \tag{21}$$

è equivalente a risolvere il sistema lineare detto delle equazioni normali

$$A^T Ax = A^T b$$

dove la matrice $A^T A$ è simmetrica e definita positiva.

Desideriamo calcolare il polinomio di miglior approssimazione $p_n^* \in \mathbb{P}_n$ tale che date m coppie (s_i, t_i), $i = 1, \ldots, m$, $m \geq n+1$, sia minima

$$\min_{p_n \in \mathbb{P}_n} \sqrt{\sum_{i=1}^{m} (t_i - p_n(s_i))^2}. \tag{22}$$

Visto che $p_n(s) = \sum_{j=1}^{n+1} \gamma_j s^{j-1} = s_1 + s_2 s + s_3 s^2 + \ldots + s_{n+1} s^n$, da (22) abbiamo che i coefficienti incogniti $\gamma = (\gamma_j)_{j=1,\ldots,n+1}$ sono tali che sono soluzione di

$$\min_{\gamma \in \mathbb{R}^{n+1}} \sqrt{\sum_{i=1}^{m} (t_i - \sum_{j=1}^{n+1} \gamma_j s_i^{j-1})^2} \tag{23}$$

ovvero γ è soluzione ai minimi quadrati del sistema sovradeterminato $A\gamma = b$, cioè

$$\min_{\gamma \in \mathbb{R}^{n+1}} \|b - A\gamma\|_2 = \min_{\gamma \in \mathbb{R}^{n+1}} \sqrt{\sum_{i=1}^{m} (b_i - \sum_{j=1}^{n} a_{i,j}\gamma_j)^2} \tag{24}$$

in cui

- $A = (a_{i,j}) \in \mathbb{R}^{m \times n}$ con $a_{i,j} = s_i^{j-1}$,
- $b = (t_i)_{i=1,\ldots,m}$,
- $\gamma = (\gamma_i)_{i=1,\ldots,n+1}$.

Quindi per risolvere (22) si determina la soluzione $\gamma = (\gamma_i)$ del sistema sovradeterminato (24) e con tali coefficienti si ottiene il polinomio di miglior approssimazione $p_n(s) = \sum_{j=1}^{n+1} \gamma_j s^{j-1}$.

Se la matrice $A \in \mathbb{R}^{m \times n}$, $m \geq n$, ha rango massimo (ossia pari a n), si dimostra che la matrice $A^T A$ è simmetrica e definita positiva e quindi possiede una fattorizzazione di Cholesky, ossia esiste una matrice L triangolare inferiore con elementi diagonali positivi tale che $A^T A = LL^T$.

Di conseguenza, la soluzione x^* del problema ai minimi quadrati deve risolvere le equazioni normali $A^T Ax = A^T b$ e posto $y = L^T x$ é tale che

$$A^T Ax = A^T b \Leftrightarrow LL^T x = A^T b \Leftrightarrow Ly = A^T b$$

Quindi si risolve prima $Ly = A^T b$ e, detta y^* la soluzione, di seguito $L^T x = y^*$.
Il costo computazionale per la risoluzione del sistema è

- $n^2 m/2$ operazioni moltiplicative per il calcolo di $A^T A$ (si usa il fatto che tale matrice è simmetrica!);
- nm operazioni moltiplicative per il calcolo di $A^T b$;
- $n^3/6$ operazioni per la risoluzione del sistema lineare $A^T Ax = A^T b$ col metodo di Cholesky.

e visto che usualmente $\frac{n^2 m}{2} \gg nm$, il numero di operazioni moltiplicative risulta

$$\frac{n^2 m}{2} + nm + \frac{n^3}{6} \approx \frac{n^2 m}{2} + \frac{n^3}{6}.$$

Esempio

Risolvere il problema ai minimi quadrati $Ax = b$ dove

$$A = \begin{bmatrix} 2 & 5 \\ 1 & 1 \\ 3 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

mediante fattorizzazione di Cholesky.

Svolgiamo i conti in Matlab:

```
>> A=[2 5; 1 1 ; 3 2];
>> b=ones(3,1);
>> R=chol(A'*A); % triangolare sup. tale che R'*R=A
>> L=R'; % triangolare inf. tale che R'*R=A
>> y=L\A'*b; % risolve Ly=A^Tb
>> x=L'\y % risolve L'x=y.
x =
     3.358778625954199e-01
     7.633587786259541e-02
>>
```

Definizione (Fattorizzazione QR)

Data una matrice $A \in \mathbb{R}^{m\times n}$, $m \geq n$, avente rango massimo n, si dice che A è fattorizzabile QR se e solo se esistono

- *la matrice quadrata $Q \in \mathbb{R}^{m\times m}$ unitaria ovvero $QQ^T = Q^TQ = I_m$,*
- *la matrice rettangolare $R \in \mathbb{R}^{m\times n}$ triangolare superiore, ovvero $r_{i,j} > 0$ se $i > j$,*

tali che $A = QR$.

Importante.

Sotto queste ipotesi, si mostra che la fattorizzazione QR esiste ed e' unica.

Esempio

Sia

$$A = \begin{bmatrix} 72 & -144 & -144 \\ -144 & -36 & -360 \\ -144 & -360 & 450 \end{bmatrix}$$

Il determinante della matrice è $\det(A) = -34012224$ *e quindi la matrice A ha rango massimo, ovvero* $rk(A) = 3$. *Si vede che (cf. [3, p. 184])*

$$Q = \frac{1}{6} \cdot \begin{bmatrix} -2 & 4 & 4 \\ 4 & -2 & 4 \\ 4 & 4 & -2 \end{bmatrix}, \quad R = \begin{bmatrix} -216 & -216 & 108 \\ 0 & -324 & 324 \\ 0 & 0 & -486 \end{bmatrix}.$$

Esempio

Si determini, se esistente, la fattorizzazione QR di

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

La matrice $A \in \mathbb{R}^{3\times 2}$ ha rango 2 perchè la sottomatrice

$$A_{1,2} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

composta dalle prime due righe è non singolare (si ha $\det(A_{1,2}) = 1$).

Se fattorizziamo $A = QR$ in Matlab otteniamo

$$Q = \begin{bmatrix} -7.071067811865472e-01 & 4.082482904638630e-01 & 5.773502691896258e-01 \\ -7.071067811865475e-01 & -4.082482904638630e-01 & -5.773502691896258e-01 \\ 0 & -8.164965809277261e-01 & 5.773502691896256e-01 \end{bmatrix}$$

$$R = \begin{bmatrix} -1.414213562373095e+00 & -7.071067811865475e-01 \\ 0 & -1.224744871391589e+00 \\ 0 & 0 \end{bmatrix}$$

In particolare, si ha che

- $\|Q'Q - I_3\|_\infty \approx 6.3642e-16$,
- $\|QQ' - I_3\|_\infty \approx 7.3882e-16$,
- $\|A - QR\|_\infty \approx 7.3882e-16$.

dove $I_3 \in \mathbb{R}^{3\times 3}$ é la matrice identica.

Il fatto che le norme infinito matriciali siano vicino a 0 implica che il loro argomento e' circa la matrice nulla, ovvero che numericamente

- $Q'Q - I_3 \approx 0$ (cioé $Q'Q \approx I_3$),
- $QQ' - I_3 \approx 0$ (cioé $QQ' \approx I_3$),
- $A - QR \approx 0$ (cioé $A \approx QR$).

Vediamo un ulteriore esempio in Matlab.

```
>> A=[1 2 3; 5 3 6; 2 8 3; 1 5 9] % matrice 4 x 3 (m=4, n=3)
A =
     1     2     3
     5     3     6
     2     8     3
     1     5     9
>> [Q,R]=qr(A) % display in "format short"
Q = % matrice tale che Q'*Q=I
    -0.1796    0.1040    0.1548   -0.9659
    -0.8980   -0.4203    0.0297    0.1265
    -0.3592    0.7453   -0.5587    0.0575
    -0.1796    0.5070    0.8142    0.2185
R = % matrice rettangolare e "triangolare"
    -5.5678   -6.8250   -8.6211
          0    7.4444    4.5888
          0         0    6.2944
          0         0         0
>> norm(A-Q*R) % numericamente A=Q*R se  norm(A-Q*R) e' piccola
ans =
   4.9370e-15
>> Q'*Q % verifichiamo che Q'*Q=I
ans =
    1.0000   -0.0000   -0.0000    0.0000
   -0.0000    1.0000         0   -0.0000
   -0.0000         0    1.0000   -0.0000
    0.0000   -0.0000   -0.0000    1.0000
>>
```

Nota.

Il calcolo della fattorizzazione QR è in sostanza equivalente all'ortogonalizzazione di Gram-Schmidt delle colonne di A (cf. [3, p.9].

n	*cputime*	n	*cputime*
2	$1e-05$	128	$3e-03$
4	$2e-05$	256	$5e-03$
8	$5e-05$	512	$2e-02$
16	$7e-05$	1024	$1e-01$
32	$1e-04$	2048	$6e-01$
64	$2e-03$	4096	$4e+00$

Tabella: *Secondi necessari per calcolare la fattorizzazione QR di una generica $A \in \mathbb{R}^{n\times n}$, mediante eliminazione gaussiana su un Mac Book Pro, con processore 2,7 GHz Intel Core i5 e 16 GB di memoria.*

- Sia $A \in \mathbb{R}^{m \times n}$, $m \geq n$, una matrice avente rango massimo, e sia $A = QR$,
- $b \in \mathbb{R}^n$.

Supponiamo di dover calcolare la soluzione x ai minimi quadrati di $Ax = b$.

Per il teorema precedente x^* è pure soluzione delle equazioni normali $A^T Ax = A^T b$.
Essendo

1. $A = QR$,
2. $Q^T Q = I$,
3. $(QR)^T = R^T Q^T$,

$$A^T A \overset{(1)}{=} (QR)^T QR \overset{(3)}{=} R^T Q^T QR \overset{(2)}{=} R^T R$$

e quindi $A^T Ax = A^T b$ se e solo se

$$R^T Rx = A^T b.$$

Ricordando che $R \in \mathbb{R}^{m \times n}$ è una matrice triangolare, si risolve

- $R^T y = A^T b$ (n equazioni e m incognite),
- $Rx = y$ (m equazioni e n incognite, ma solo n sono significative).

Stadio 1. Analizziamo il primo sistema $R^T y = A^T b$ con $R \in \mathbb{R}^{m\times n}$, $A^T b \in \mathbb{R}^{n\times 1}$.
Se per una certa matrice $L \in \mathbb{R}^{n\times n}$ triangolare inferiore

$$R^T = [\underbrace{L}_{n} \underbrace{\mathbf{0}}_{m-n}]\}n, \; y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, y_1 \in \mathbb{R}^n, y_2 \in \mathbb{R}^{m-n}$$

si vede facilmente che essendo

$$A^T b = R^T y = [L \quad \mathbf{0}] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Ly_1 + \mathbf{0}y_2 = Ly_1$$

si ha

$$R^T y = A^T b \Leftrightarrow Ly_1 = A^T b$$

che richiede la risoluzione di un sistema triangolare inferiore $Ly_1 = A^T b$ di dimensione $n \times n$.

Si osservi che qualsiasi sia $y_2^* \in \mathbb{R}^{m-n}$, il vettore $y = [y_1^*, y_2^*]^T$ risolve $R^T y = A^T b$.

Stadio 2. Una volta ottenuto $y_1^* \in \mathbb{R}^{n\times 1}$ soluzione di $Ly_1 = A^T b$, e posto $y_2^* \in \mathbb{R}^{(m-n)\times 1}$ arbitrario, visto che $R \in \mathbb{R}^{m\times n}$, dobbiamo risolvere il sistema sovradeterminato $Rx = y^*$ con $y^* = [y_1^*; y_2^*] \in \mathbb{R}^{m\times 1}$.
A tal proposito, visto che $x^* \in \mathbb{R}^{n\times 1}$ e

$$R = \begin{bmatrix} L^T \\ \mathbf{0} \end{bmatrix},\ y^* = \begin{bmatrix} y_1^* \\ y_2^* \end{bmatrix} \Rightarrow \begin{bmatrix} y_1^* \\ y_2^* \end{bmatrix} = y^* = Rx = \begin{bmatrix} L^T \\ \mathbf{0} \end{bmatrix} x = \begin{bmatrix} L^T x \\ 0_{\mathbb{R}^{m-n}} \end{bmatrix},$$

e quindi la soluzione $x^* \in \mathbb{R}^{n\times 1}$ del problema sovradeterminato $Ax = b$ puó essere ricavata calcolando la soluzione del sistema quadrato triangolare superiore e nonsingolare $L^T x = y_1^*$.

Poiché la soluzione ai minimi quadrati richiede la soluzione di

- $R^T y = A^T b$ (n equazioni e m incognite),
- $Rx = y^*$ (m equazioni e n incognite, ma solo n sono significative).

e considerato per A generica

- il costo della fattorizzazione QR (si utilizza un metodo dovuto ad Householder o a Givens);
- la valutazione di $A^T b$ che richiede in generale circa mn operazioni moltiplicative, visto che $A \in \mathbb{R}^{m \times n}$;
- la risoluzione di $R^T y = A^T b$ che richiede di determinare le prime n componenti di y (le altre vengono poste uguali a 0) e quindi richiede circa $n^2/2$ operazioni, visto che R è triangolare;
- la risoluzione di $Rx = y^*$ che richiede circa $n^2/2$ operazioni, visto che $R \in \mathbb{R}^{m \times n}$ è una matrice triangolare e le ultime $m - n$ componenti di y sono nulle.

Si dimostra che il numero di moltiplicazioni necessarie per la risoluzione di sistemi sovradeterminati, con il metodo QR, risulta circa $2mn^2 - 2n^3/3$ utilizzando il metodo di Householder e $3mn^2 - n^3$ utilizzando il metodo di Givens (cf. [5, p.263]).

Esempio

Risolvere, mediante fattorizzazione QR, il problema ai minimi quadrati $Ax = b$ dove

$$A = \begin{bmatrix} 2 & 5 \\ 1 & 1 \\ 3 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

```
>> format long; A=[2 5; 1 1 ; 3 2]; b=ones(3,1); [Q,R]=qr(A);
>> Q
Q =
   -0.534522483824849   0.840626032489748   0.087370405666104
   -0.267261241912424  -0.070052169374146  -0.961074462327142
   -0.801783725737273  -0.537066631868450   0.262111216998311
>> R
R =
   -3.741657386773941  -4.543441112511214
                    0   3.058944729337695
                    0                   0
>> Rtrasposta=R';
>> L=Rtrasposta(1:2,1:2)
L =
   -3.741657386773941                   0
   -4.543441112511214   3.058944729337695
>> y=L\(A'*b)
y =
   -1.603567451474546
    0.233507231247152
>> x=L'\y
x =
    0.335877862595420
    0.076335877862595
>>
```

Definizione (Decomposizione SVD (ovvero ai valori singolari, cf. [3, p.444]))

Sia $A \in \mathbb{R}^{m\times n}$. Allora esistono

- *una matrice $U \in \mathbb{R}^{m\times m}$ unitaria (ovvero $U^T U = UU^T = I$);*
- *una matrice $V \in \mathbb{R}^{n\times n}$ unitaria (ovvero $V^T V = VV^T = I$);*
- *una matrice $\Sigma \in \mathbb{R}^{m\times n}$ che ha elementi $\sigma_{i,j} = 0$ se $i \neq j$ e se $i = j$ allora $\sigma_{i,i} = \sigma_i$ con*

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq \sigma_{k+1} = \ldots = \sigma_p = 0, \ p = \min\{m, n\}$$

tali che

$$A = U\Sigma V^T.$$

Gli elementi σ_j sono detti valori singolari di A.

- I valori singolari $\{\sigma_k\}_{k=1,\ldots,p}$ sono le radici quadrate degli autovalori di $A^T A$.
- Se la matrice A é simmetrica allora i valori singolari corrispondono con gli autovalori di A in modulo.
- L'ultimo k per cui $\sigma_k > 0$ é il rango di A (cf. [3, p.448]).

Definizione (Decomposizione SVD (ovvero ai valori singolari, cf. [3, p.454]))

Sia $A \in \mathbb{R}^{m\times n}$ con $m \geq n \geq k$ di rango k e $A = U\Sigma V^T$. la decomposizione in valori singolari di A. Allora, se

- *u_i è l'i-sima colonna di U;*
- *v_i è l'i-sima colonna di V;*

la soluzione di minima norma risulta

$$x^* = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i. \tag{25}$$

Se $k = n$ allora x^* é l'unica soluzione nel senso dei minimi quadrati del problema $Ax = b$.

- Si osservi che per calcolare x^* non bisogna risolvere sistemi lineari.
- Esistono vari algoritmi per calcolare la soluzione mediante fattorizzazione SVD. Col metodo di Golub-Reinsch, la determinazione di U, Σ, V costa $4m^2n + 8n^3$ operazioni moltiplicative, mentre con il metodo R-SVD la complessitá risulta $2m^2n + 11n^3$ ([5, p.263]).

Esempio

Risolvere, mediante fattorizzazione SVD, il problema ai minimi quadrati $Ax = b$ dove

$$A = \begin{bmatrix} 2 & 5 \\ 1 & 1 \\ 3 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Con tale metodo, ricaviamo come in precedenza la soluzione x^* di $Ax = b$.

```
>> A=[2 5; 1 1 ; 3 2];
>> [U,S,V] = svd(A) % decomposizione SVD (display matrici in "format short")
U =
   -0.8288    0.5527    0.0874
   -0.2161   -0.1722   -0.9611
   -0.5161   -0.8154    0.2621
S =
    6.3866         0
         0    1.7921
         0         0
V =
   -0.5358   -0.8443
   -0.8443    0.5358
>> b=ones(3,1);
>> format long e;
>> x=sum(diag((U(:,1:size(A,2)))'*b./diag(S))*V); % soluzione calcolata con SVD.
>> x'
ans =
     3.358778625954198e-01
     7.633587786259541e-02
>>
```

Si noti che `x=sum(diag((U(:,1:size(A,2)))'*b./diag(S))*V);` é equivalente a (25).

Bibliografia I

[1] K. Atkinson, *Introduction to Numerical Analysis*, Wiley, 1989.

[2] K. Atkinson, W. Han, *Elementary Numerical Analysis*, Wiley, 2003.

[3] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, 1988.

[4] J.F. Epperson, *Introduzione all'analisi numerica. Teoria, metodi, algoritmi.*, Mc Graw-Hill, 2009.

[5] G.H.Golub and C.F. Van Loan, *Matrix Computations*, 3rd edition, John Hopkins University Press, (1996).

[6] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM Frontiers in Applied Mathematics. SIAM, Philadelphia, (1995).

[7] L.N. Trefethen, *Numerical Analysis*, Princeton Companion to Mathematics, 2008.

[8] Wikipedia, Decomposizione di Cholesky.

[9] Wikipedia, Decomposizione LU.

[10] Wikipedia, Decomposizione QR.

[11] Wikipedia, Matrice a diagonale dominante.

[12] Wikipedia, Matrice definita positiva.

[13] Wikipedia, Matrice invertibile: Metodo della matrice dei cofattori.

[14] Wikipedia, SuccessiveOverRelaxation.

[15] Wikipedia, Supercomputer.

[16] Wikipedia, Teorema di Binet.

[17] Wikipedia, Teorema di Laplace.