ALGEBRA LINEARE NUMERICA *

A. SOMMARIVA†

1. Condizionamento di matrici e sistemi.

Definizione. Dato uno spazio vettoriale V, 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. $\|\alpha x\| = |\alpha|\|x\|$ per ogni $x \in V$ e scalare $\alpha \in \mathbb{R}$,
3. vale la cosidetta disuguaglianza triangolare

$$\|x + y\| \geq \|x\| + \|y\|$$

per ogni $x, y \in V$.

Esempio. Se $V = \mathbb{C}^n$ e $x = (x_1, x_2, \ldots, x_n) \in \mathbb{C}^n$ 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 come

$$\|x\|_p = \max_{j=1,\ldots,n} |x_j|;$$

Nota. Lo stesso tipo di norme si definisce nel caso particolare i vettori appartengano a $\mathbb{R}^n$.

Nota. Si osservi che

- 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

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

*Ultima revisione: 9 novembre 2018

†Dipartimento di Matematica, Università degli Studi di Padova, stanza 419, via Trieste 63, 35121 Padova, Italia (alvise@math.unipd.it).

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

$$\text{dist}(u, v) := \|u - v\|_p, \;\; p = [1, \infty].$$

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

- $\|u - v\|_1 = 0.002$,
- $\|u - v\|_2 = 0.0014$,
- $\|u - v\|_\infty = \max |4|, | - 3| = 0.001$,

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 $A \in \mathbb{R}^{n\times n}$ nel valore

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

Nota. Lo stesso tipo di norme si definisce nel caso particolare le matrici siano in $\mathbb{R}^{n\times n}$.

Definizione. Si definisce *raggio spettrale* di $A \in \mathbb{C}^{n\times n}$ 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 $Ax_k = \lambda_k x_k$.

Alcuni esempi di norme matriciali risultano:

- per $p = 1$, partendo dalla norma vettoriale $\|\cdot\|_1$, si dimostra che la *norma* 1 matriciale risulta uguale a

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

  dove $A = (a_{i,j})$,
- per $p = 2$, partendo dalla norma vettoriale $\|\cdot\|_2$, si dimostra che la *norma* 2 matriciale risulta uguale a

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

  dove $A = (a_{i,j})$ e se $A \in \mathbb{C}^{n\times n}$ allora A^* è la trasposta coniugata di A mentre se $A \in \mathbb{R}^{n\times n}$ allora A^* è la trasposta di A;
- per $p = \infty$, partendo dalla norma vettoriale $\| \cdot \|_\infty$, si dimostra che la *norma* ∞ matriciale risulta uguale a

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

  dove $A = (a_{i,j})$.

Esempio. 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. 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

$$\text{dist}(A, B) := \|A - B\|_p, \;\; p = [1, \infty].$$

Allora, visto che

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

ricaviamo

- $\|A - B\|_1 = 0.006$,
- $\|A - B\|_2 = 0.005415654779058332$,
- $\|A - B\|_\infty = \max(|4|, |-3|) = 0.007$,

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

TEOREMA 1.1. *Per le norme matriciali indotte, valgono le seguenti disuguaglianze*

$$\|Ax\| \leq \|A\|\|x\|, \tag{1.1}$$

e

$$\|AB\| \leq \|A\|\|B\|. \tag{1.2}$$

Dimostrazione. Mostriamo la prima disuguaglianza. E' banalmente verificata per $x = 0$. Altrimenti, essendo

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

abbiamo che per $x \neq 0$,

$$\|Ax\| \leq \|A\|\|x\|.$$

Per quanto riguarda la seguente disuguaglianza, essendo per il primo punto, se $y \neq 0$

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

da cui

$$\frac{\|ABy\|}{\|y\|} \leq \|A\|\|B\|$$

e quindi

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

□

Definizione. Sia $A \in \mathbb{R}^{n \times n}$ 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 $\mathbb{R}^{n \times n}$

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

e quindi

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

ovvero $k(A) \geq 1$.

Vale il seguente risultato

TEOREMA 1.2. *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.

Se A è simmetrica

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

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 1.1
Indice di condizionamento $c_n = cond_2(H_n)$ della matrici di Hilbert H_n di ordine n

Esempio. 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.

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

TEOREMA 1.3. *Se*
1. $A \in \mathbb{R}^{n\times n}$ *è una matrice invertibile,*
2. $b \in \mathbb{R}^n$, $b \neq 0$,
3. $Ax = b$,
4. $A(x + \delta x) = b + \delta b$,

allora vale la stima

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

Dimostrazione. Da $A(x + \delta x) = b + \delta b$ e $Ax = b$ abbiamo che

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

ovvero $A\delta x = \delta b$ e quindi $\delta x = A^{-1}\delta b$. Ne consegue che

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|. \tag{1.3}$$

D'altra parte, essendo $Ax = b$

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

e quindi, essendo $x \neq 0$ in quanto $b \neq 0$, ricaviamo

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{1.4}$$

Da (1.3), (1.4), ricaviamo

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

□

Commento. Questo teorema mostra che si fa un errore relativo sul termine noto pari a $\frac{\|\delta b\|}{\|b\|}$ allora si compie un errore relativo sulla soluzione $\frac{\|\delta x\|}{\|x\|}$ tale che

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

Quindi, se $k(A) \gg 1$, può accadere che la soluzione perturbata $x + \delta x$ 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 $Ax = b + \delta b$ ha soluzione

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

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.

Mostriamo ora la risposta della soluzione di un sistema lineare agli errori sulla matrice.

TEOREMA 1.4. *Se*

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

allora vale la stima

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

Dimostrazione. Essendo $(A + \delta A)(x + \delta x) = b$ e $Ax = b$, necessariamente

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

da cui

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

ovvero

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

e di conseguenza

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

Quindi, essendo $\|A\| \neq 0$,

$$\frac{\|\delta x\|}{\|x+\delta x\|} \leq \|A^{-1}\|\|\delta A\| = \frac{\|A^{-1}\|\|A\|}{\|A\|}\|\delta A\|$$

da cui

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

□

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

Per prima cosa introduciamo il seguente importante teorema.

TEOREMA 1.5. *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*

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

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

Finalmente,

TEOREMA 1.6. *Se*
1. $A \in \mathbb{R}^{n\times n}$ *è una matrice invertibile,*
2. $b \in \mathbb{R}^n$, $b \neq 0$,
3. $Ax = b$,
4. $(A+\delta A)(x+\delta x) = b + \delta b$, *con* $\det(A+\delta A) \neq 0$,
5. $k(A)\|\delta A\| < \|A\|$,

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

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

è la costante di amplificazione degli errori.

Visto che il denominatore è maggiore di 1 visto che $k(A)\|\delta A\|/\|A\| < 1$, deduciamo che se $k(A) \gg 1$, la soluzione del sistema perturbato $(A+\delta A)(x+\delta x) = b + \delta b$ può essere molto distante da quella di $Ax = b$.

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

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

da cui sottraendo b ad ambo i membri

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

ovvero

$$(A+\delta A)\delta x = \delta b - \delta A x. \tag{1.5}$$

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 $k(A)\|\delta A\| < \|A\|$

$$\begin{aligned} \|B\| = \|A^{-1}\delta A\| &\leq \|A^{-1}\|\|\delta A\| \\ &= \frac{\|A^{-1}\|\|A\|}{\|A\|}\|\delta A\| = k(A)\frac{\|\delta A\|}{\|A\|} \\ &\leq \frac{\|A\|}{\|A\|} = 1 \end{aligned} \tag{1.6}$$

ovvero $\|B\| < 1$.

Osserviamo che la matrice $A+\delta A$ è invertibile in quanto lo è A, lo è $I+B$ per il teorema 1.5 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 (basta ricordare che $\det(MN) = \det(M)\det(N)$, e quindi se $\det(M), \det(N) \neq 0$ allora $\det(MN) \neq 0$).

Dal teorema 1.5, moltiplicando ambo i membri per $(A+\delta A)^{-1}$, visto che $(MN)^{-1} = N^{-1}M^{-1}$ ricaviamo

$$\begin{aligned} \delta x &= (A+\delta A)^{-1}(\delta b - \delta A x) = (A(I+B))^{-1}(\delta b - \delta A x) \\ &= (I+B)^{-1}A^{-1}(\delta b - \delta A x) \end{aligned} \tag{1.7}$$

da cui, essendo nuovamente per il teorema 1.5 $\|(I+B)^{-1}\| \leq 1/(1-\|B\|)$, $B = A^{-1}\delta A$

$$\begin{aligned} \|\delta x\| &= \|(I+B)^{-1}A^{-1}(\delta b - \delta A x)\| \\ &\leq \|(I+B)^{-1}\|\|A^{-1}\|\|\delta b - \delta A x\| \\ &\leq \frac{1}{1-\|B\|}\|A^{-1}\|\|\delta b - \delta A x\| \\ &= \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|}\|\delta b - \delta A x\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|) \end{aligned} \tag{1.8}$$

e quindi riassumendo

$$\|\delta x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|). \tag{1.9}$$

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

$$\frac{1}{\|A\|\|x\|} \leq \frac{1}{\|b\|} \tag{1.10}$$

e moltiplicando membro a membro i risultati di (1.9), (1.10),

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\|x\|\frac{1}{\|x\|}\right) \\
&\leq \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\frac{\|1\|}{\|A\|}\left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\right) \\
&\leq \frac{k(A)}{1-\|A^{-1}\|\|A\|\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|A\|\|x\|} + \frac{\|\delta A\|}{\|A\|}\right) \\
&\leq \frac{k(A)}{1-k(A)\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right).
\end{aligned} \tag{1.11}$$

□

Esempio. 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

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

il sistema

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

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.

Con riferimento al teorema 1.6

- $\det(A) \approx 0.003$, $\det(\delta A) \approx 0.0015$,
- $k_2(A) \approx 2083.666853410356$,
- abbiamo

$$\frac{k_2(A)}{1-k_2(A)\|\frac{\|\delta A\|_2}{\|A\|_2}} \approx 3.05387e+04,$$

- $\frac{\|\delta x\|_2}{\|x\|_2} \approx 3.16228e+00$,
- $\frac{\|\delta b\|_2}{\|b\|_2} \approx 4.47214e-04$ e $\frac{\|\delta A\|_2}{\|A\|_2} \approx 4.47178e-04$.

La stima asserisce in effetti che

$$\begin{aligned}
3.16228e+00 &\leq 3.05387e+04 \cdot (4.47214e-04 + 4.47178e-04) \\
&\approx 2.73136e+01.
\end{aligned} \tag{1.12}$$

2. Metodo di eliminazione di Gauss e fattorizzazione LU. 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 consideri il sistema lineare

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

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 una equazione per una costante e sommiamo il risultato membro a membro a un'altra equazione, otteniamo un nuovo problema che ha le stesse soluzioni del precedente.

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

- moltiplichiamo la prima riga per -2 e la sommiamo alla seconda, ottenendo

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

- moltiplichiamo la prima riga per 3 e la sommiamo alla terza, ottenendo

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

- moltiplichiamo la seconda riga per $-1/2$ e la sommiamo alla terza, ottenendo

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

Il sistema finale, è facile da risolvere.

- 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

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

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

- Inseriamo questi risultati nella prima equazione e otteniamo

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

da cui

$$x_1 = 1.$$

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

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

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*.

2.1. Sistemi lineari triangolari.

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

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

Sia $b = (b_j) \in \mathbb{R}^n$ e supponiamo di dover risolvere il sistema $Ax = b$.
In tal caso,

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

$$\begin{cases} x_n = \frac{b_n}{a_{n,n}} \\ x_k = \frac{b_k - (a_{k,k+1}x_{k+1} + \ldots + a_{k,n}x_n)}{a_{k,k}}, \quad k = n-1, \ldots, 1. \end{cases} \tag{2.1}$$

  Si noti che essendo A invertibile, necessariamente $a_{k,k} \neq 0$ per $k = 1, \ldots, n$.
- se A è triangolare inferiore e invertibile, si usa l'algoritmo della *sostituzione in avanti* ovvero

$$\begin{cases} x_1 = b_1/a_{1,1}, \\ x_k = \frac{b_k - (a_{k,1}x_1 + \ldots + a_{k,k-1}x_{k-1})}{a_{k,k}}, \quad k = 2, \ldots, n. \end{cases} \tag{2.2}$$

  Si noti che essendo A invertibile, necessariamente $a_{k,k} \neq 0$ per $k = 1, \ldots, n$.

Nota. 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

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

2.2. Metodo di eliminazione gaussiana senza pivoting. In questa sezione trattiamo la risoluzione di sistemi lineari $Ax = b$ dove in generale si suppone

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

Posto $A = A^{(1)} = (a^{(1)}_{i,j})$, $b = b^{(1)} = (b^{(1)}_j)$, possiamo riscrivere il sistema nella formula nota come *aumentata*

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

- **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

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

A tal proposito definiamo il *moltiplicatore*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$b_{i}^{(k+1)} = b_{i}^{(k)} - m_{i,k} b_{k}^{(k)}, \;\; j = k+1, \ldots, n.$$

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

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

da cui otteniamo che $[A^{(k+1)}|b^{(k+1)}]$ è uguale a

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

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

Di conseguenza il sistema $A^{(n-1)}x = b^{(n-1)}$ può essere facilmente risolto con l'algoritmo di sostituzione all'indietro.

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

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

tali che $A = LU$ (cf. [8]).

Importante. 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 $A^{(n-1)}x = b^{(n-1)}$ equivalente all'originario e durante il calcolo introduciamo i moltiplicatori $m_{i,k}$, $k = 1, \ldots, n-1$, $i = k+1, \ldots, n$. Siano

- $U = A^{(n-1)}$,
- L la matrice

$$L = \left[\begin{array}{ccccc}
1 & 0 & 0 & \ldots & 0 \\
m_{2,1} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ldots & \vdots \\
m_{n,1} & m_{n,2} & \ldots & \ldots &
\end{array}\right].$$

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 del processo di eliminazione gaussiana senza pivoting abbiamo determinato $U = A^{(n-1)}$ uguale a

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

Tenendo conto dei moltiplicatori

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

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

2.3. Metodo di eliminazione gaussiana con pivoting (parziale). 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 $A^{(n-1)}x = b^{(n-1)}$, con $A^{(n-1)}$ triangolare superiore.

Il problema è che qualche $a_{k,k}^{(k)}$ potrebbe risultare nullo e quindi il processo precedente risulterebbe inapplicabile.

Ma anche se così non fosse, potrebbe accadere che $a_{k,k}^{(k)}$ sia 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*.

Al passo k-simo dell'eliminazione gaussiana, 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 $i := \phi(k)$ il più piccolo indice di riga, con $i \ge k$, per cui il massimo c_k è ottenuto. Se $i > k$, scambiamo la riga i-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.

Importante. Dall'applicazione della procedura dell'eliminazione gaussiana con pivoting per la risoluzione di un sistema lineare, abbiamo ottenuto un sistema $A^{(n-1)}x = b^{(n-1)}$ equivalente all'originario. Poniamo

- $P = P_{n-1} \cdot \ldots \cdot P_1$ dove $P_k = ((p_k)_{i,j})$ è la matrice identica, ad eccezione delle righe k e $\phi(k)$ in cui sono uguali a 1 esclusivamente le componenti $(p_k)_{k,\phi(k)}$ e $(p_k)_{\phi(k),k}$,
- $U = A^{(n-1)}$ di forma triangolare superiore,
- $L = (l_{i,j})$ la matrice

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

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, come vedremo in seguito con qualche esempio, indipendentemente dalla risoluzione di sistemi lineari.

Esempio. Si consideri il sistema lineare

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

Il termine maggiore in modulo della prima colonna è il terzo e vale 9. Con le notazioni introdotte, $\phi(1) = 3$. Scambiamo la prima riga con la terza e ricaviamo il sistema equivalente

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

- Moltiplichiamo la prima riga per $-m_{2,1} = -(-6/9) = 2/3$ e la sommiamo 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 per $-m_{3,1} = -(-1/3) = 1/3$ e la sommiamo 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 $\max 8/3, 1/3 = 8/3$ Con le notazioni introdotte, $\phi(2) = 2$ e quindi non si richiedono scambi di equazioni.

Moltiplichiamo la seconda riga per $-m_{3,2} = -(1/3)/(8/3) = -1/8$ e la sommiamo 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.

- 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, $U = A(n-1)$ e dunque

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

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

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

Avendo scambiato la prima con la terza riga, la matrice P, detta di permutazione, risulta

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

mentre al secondo passo, visto che non abbiamo effettuato scambi $P_2 = I_n$, con $I_n \in \mathbb{R}^{n \times n}$ matrice identica. Quindi $P = P_2 \cdot P_1 = P_1$.

3. Alcune note sulla eliminazione gaussiana e la fattorizzazione LU.

Nota. Definizione. La matrice $A \in \mathbb{C}^{n\times n}$ 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.

Esistono classi di matrici come quelle

1. a predominanza diagonale stretta per righe (cf. [10]), o
2. a predominanza diagonale stretta per colonne (cf. [10]), o
3. simmetriche e definite positive,

per cui è sempre possibile la fattorizzazione $A = LU$ (cf. [6, p.76]).

Nota. 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]).

Nota. Il seguente pseudocodice implementa la prima parte dell'algoritmo di eliminazione gaussiana senza pivoting, cui segue la risoluzione di un sistema triangolare superiore.

```
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
```

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 $k = 1, \ldots, n-1$.

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

$$\begin{aligned}\sum_{p=1}^{n-1}\sum_{k=p+1}^{n}(n-p+2) &= \sum_{p=1}^{n-1}\sum_{k=p+1}^{n}2+\sum_{p=1}^{n-1}\sum_{k=p+1}^{n}(n-p)\\ &= 2\sum_{p=1}^{n-1}(n-p)+\sum_{p=1}^{n-1}(n-p)^2\\ &= 2\sum_{k=1}^{n-1}k+\sum_{k=1}^{n-1}k^2\end{aligned} \tag{3.1}$$

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

$$\begin{aligned}\frac{(n-1)^3}{3} = \int_0^{n-1} x^2dx &\le \sum_{k=1}^{n-1}k^2\\ &\le \int_1^n x^2dx = \frac{n^3-1}{3}\end{aligned} \tag{3.2}$$

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,

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

- 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 $Ax = b$ necessita di circa $n^3/3$ operazioni moltiplicative;
- La complessità della fattorizzazione $PA = LU$ è essenzialmente lo stesso della eliminazione gaussiana, ovvero $\frac{n^3}{3}$.

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

1. Calcola la fattorizzazione $PA = LU$. A questo punto siccome P è invertibile $Ax = b$ se e solo se $LUx = PAx = Pb$.
2. Calcola $\tilde{b} = Pb$.
3. Risolvi il sistema triangolare inferiore $Ly = \tilde{b}$ con la sostituzione in avanti.
4. Risolvi il sistema triangolare inferiore $Ux = y$ con la sostituzione all'indietro.

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 3.1

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.

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. Se la A matrice è simmetrica e definita positiva [11], ovvero $A = A^T$ e tutti gli autovalori di A sono positivi, è più conveniente utilizzare la cosidetta decomposizione di Cholesky [1, p.524], [7], 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 la soluzione del sistema lineare si può ottenere in circa

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

operazioni moltiplicative.

Esempio. La matrice

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

è simmetrica e definita positiva, in quanto l'autovalore minimo è $\lambda_{\text{max}} = 0.271\ldots > 0$.

Inoltre $A = LL^T$ con

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

3.1. Alcune applicazioni della fattorizzazione LU. Di seguito mostriamo alcune applicazione della fattorizzazione LU, ovvero il calcolo del determinante di una matrice e quello della sua inversa, paragonandoli con altri metodi.

3.1.1. Calcolo del determinante di una matrice. 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. [15].

Se $PA = LU$, visto che $P^{-1} = P^T$ abbiamo che $A = P^T LU$ e quindi per il teorema di Binet

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

Osserviamo che

- a seconda degli scambi effettuati $\det(P^T) = (-1)^s$ con s numero di scambi effettuati dal pivoting,
- visto che il determinante di una matrice triangolare $T = (t_{i,j}) \in \mathbb{R}^{n\times n}$ è

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

abbiamo che da $l_{k,k} = 1$, $k = 1, \ldots, n$, si ricava $\det(L) = 1$, e quindi

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

con $\det(P^T) = (-1)^s$ che è calcolabile senza operazioni moltiplicative, ma solo contando gli scambi effettuati.

In definitiva, visto che $\prod_{k=1}^{n} u_{k,k}$ necessita solo di $n-1$ operazioni moltiplicative, mentre 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 [13] 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 3.2

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 $PA = LU$.

3.1.2. Calcolo dell'inversa di una matrice. Per calcolare l'inversa di una matrice $A \in \mathbb{R}^{n\times n}$ con il *metodo dei cofattori* [12] 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 si usa la fattorizzazione LU per ognuna di esse, necessitano

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

operazioni moltiplicative.

Si supponga sia $PA = LU$ e si risolvano i sistemi

$$Ax^{(j)} = e^{(j)}, \;\; j = 1, \ldots, n$$

dove $e^{(j)} = (0, \ldots, 0, 1, 0 \ldots, 0)$ è il j-simo vettore della base canonica. 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 con una sostituzione in avanti e una all'indietro.
  Nel caso di $Ly^{(k)} = e^{(k)}$, si vede facilmente che le prime k cifre di $y^{(k)}$ sono nulle (si veda ad esempio (2.2) ponendo $b_1, \ldots, b_{k-1} = 0$) e quindi per risolvere $Ly^{(k)} = e^{(k)}$ servono

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

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

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

  Nel caso di $Lz^{(k)} = y^{(k)}$, il costo computazionale è invece di $n^2/2$ operazioni moltiplicative per ogni k, ovvero in totale $n^3/2$ operazioni.

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

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

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. Sia

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

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

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

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

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

e quindi A^{-1} è uguale a

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

4. Sistemi sovradeterminati, minimi quadrati e fattorizzazione QR. TEOREMA 4.1. *Sia $A \in \mathbb{R}^{m\times n}$, con $m > n$, una matrice di rango massimo n (ovvero, le colonne di A sono n vettori linearmente indipendenti di $\mathbb{R}^m$).*

Allora la soluzione del sistema sovradeterminato $Ax = b$ ai minimi quadrati,

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2 = \min_{x\in\mathbb{R}^n} \sqrt{\sum_{i=1}^{m}(b_i - \sum_{j=1}^{n} a_{i,j}x_j)^2} \tag{4.1}$$

è equivalente a risolvere il sistema con matrice simmetrica, non singolare, definita positiva

$$A^TAx = A^Tb$$

detto sistema delle equazioni normali.

Nota. Essendo il sistema sovradeterminato, se $m > n$ allora ha più equazioni che incognite e quindi può non esistere $x \in \mathbb{R}^n$ tale che $Ax = b$ ma comunque (4.1) può avere soluzione (che è appunto quella nel senso dei minimi quadrati).

Dimostrazione. Sia $(u, v) = u^Tv \in \mathbb{R}$ il prodotto scalare in $\mathbb{R}^n$. Ricordiamo che

- $(\alpha u, v) = \alpha u^Tv = \alpha(u, v)$ e similmente $(u, \beta v) = \beta u^Tv = \beta(u, v)$;
- $(u, v) = u^Tv = (v^Tu) = (v, u)$ in quanto il trasposto di un numero è il numero stesso;
- $(u+v, u+v) = (u, u)+(u, v)+(v, u)+(v, v) = (u, u)+(u, v)+(u, v)+(v, v) = (u, u) + 2(u, v) + (v, v)$.

Si osservi che x minimizza il polinomio quadratico

$$\phi(x) = \|b - Ax\|^2 = (b - Ax, b - Ax) = \sum_{i=1}^{m}(b_i - \sum_{j=1}^{n} a_{i,j}x_j)^2$$

se e solo se per ogni $v \in \mathbb{R}^n$ si ha

$$\phi(x) \le \phi(x + v)$$

ovvero $\phi(x + v) - \phi(x) \ge 0$ e sviluppando i calcoli,

$$\begin{aligned} \phi(x + v) &= \|b - A(x + v)\|^2 = \|(b - Ax) - Av\|^2 \\ &= (b - Ax, b - Ax) - 2(b - Ax, Av) + (Av, Av) \\ &= \phi(x) - 2(b - Ax, Av) + (Av, Av) \end{aligned} \tag{4.2}$$

deduciamo che

$$\begin{aligned} 0 \le \phi(x + v) - \phi(x) &= -2(b - Ax, Av) + (Av, Av) \\ &= 2(Ax - b, Av) + (Av, Av). \end{aligned} \tag{4.3}$$

Notiamo che essendo v arbitrario, (4.3) vale pure per $v^* = -v$ e quindi

$$0 \le \phi(x+v^*)-\phi(x) = 2(b-Ax, Av^*)+(Av^*, Av^*) = -2(Ax-b, Av)+(Av, Av). \tag{4.4}$$

Se $v = \epsilon u$ con u versore, ovvero vettore tale che $\|u\|_2 = 1$ allora da (4.3) abbiamo con facili conti

$$0 \leq 2\epsilon(Ax - b, Au) + \epsilon^2(Au, Au)$$

ovvero per $\epsilon > 0$ arbitrario

$$0 \leq 2(Ax - b, Au) + \epsilon(Au, Au)$$

cioè

$$0 \leq 2(Ax - b, Au). \tag{4.5}$$

Similmente da (4.3)

$$0 \leq -2\epsilon(Ax - b, Au) + \epsilon^2(Au, Au)$$

ovvero per $\epsilon > 0$ arbitrario

$$0 \leq -2(Ax - b, Au) + \epsilon(Au, Au)$$

cioè

$$0 \leq -2(Ax - b, Au). \tag{4.6}$$

Di conseguenza, per u versore arbitrario, dovendo valere contemporaneamente (4.5) e (4.6) , ricaviamo $0 = -2(Ax - b, Au)$ ovvero

$$0 = (Ax - b, Au) = x^T A^T Au - b^T Au = (A^T Ax)^T u - (A^T b)^T u = (A^T Ax - A^T b, u) \tag{4.7}$$

da cui essendo u versore arbitrario, necessariamente $A^T Ax - A^T b = 0$ ovvero

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

In effetti, se fosse $A^T Ax - A^T b \neq 0$, posto $u = (A^T Ax - A^T b)/\|A^T Ax - A^T b\|$ avremmo

$$\begin{aligned} 0 &= (A^T Ax - A^T b, (A^T Ax - A^T b)/\|A^T Ax - A^T b\|) \\ &= \|A^T Ax - A^T b\|^2/\|A^T Ax - A^T b\| = \|A^T Ax - A^T b\| \end{aligned}$$

ovvero essendo il vettore nullo l'unico di norma nulla

$$A^T Ax - A^T b = 0$$

il che sarebbe assurdo perchè avevamo supposto $A^T Ax - A^T b \neq 0$.

Per dimostrare che $A^T A$ è non singolare, si osservi che se $A^T Av = 0$ allora

$$0 = (0, v) = (A^T Av, v) = v^T A^T Av = (Av, Av) = \|Av\|_2^2$$

da cui siccome $\|x\| = 0$ se e solo $x = 0$, deduciamo che $Av = 0$. Poichè il rango di A è n, le colonne di A sono linearmente indipendenti e quindi $Av = 0$ implica $v = 0$. Ne comporta che $A^T Av = 0$ se e solo $v = 0$, ovvero $A^T A$ è non singolare (il suo nucleo ha solo il vettore nullo!).

Osserviamo che $A^TA \in \mathbb{R}^{n\times n}$ è simmetrica essendo

$$(A^TA)^T = A^T(A^T)^T = A^TA.$$

Infine visto che A^TA è definita positiva se e solo se

$$x^TA^TAx > 0, \text{ per } x \neq 0$$

essendo $Ax = 0$ se e solo se $x = 0$ e

$$x^TA^TAx = (Ax)^TAx = (Ax, Ax) = \|Ax\|^2 \geq 0,$$

deduciamo che A^TA è definita positiva. □

Nota. L'approssimazione polinomiale di grado k ai minimi quadrati può essere reinterpretata come soluzione ai minimi quadrati del sistema sovradeterminato

$$Va = y$$

dove

- $V = (x_i^j)_{i=1,\dots,m, j=0,\dots,k} \in \mathbb{R}^{m\times(k+1)}$, $m \geq k+1$, è una matrice di Vandermonde
- $y = (y_j)_{j=1,\dots,m}$ è tale che $y_j = f(x_j)$ per $j = 1, \dots, m$.

Il teorema precedente asserisce che $a = (a_j)_{j=0,\dots,k}$ è la soluzione del sistema delle equazioni normali

$$V^TVa = V^Ty.$$

Definizione. 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 $A = QR$ (cf. [9]).

Nota. Il calcolo della fattorizzazione QR è in sostanza equivalente all'ortogonalizzazione di Gram-Schmidt delle colonne di A (cf. [3, p.9], [6, p.86]).

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 4.1

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.

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. Sia

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

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

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

4.0.1. Applicazione della fattorizzazione QR alla soluzione di sistemi sovradeterminati. Sia $A \in \mathbb{R}^{m\times n}$, $m \geq n$, una matrice avente rango massimo, e sia $A = QR$.

Se $b \in \mathbb{R}^n$, e 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 $A = QR$, e $Q^T Q = I$ ricaviamo

$$A^T A = (QR)^T QR = R^T Q^T QR = R^T R$$

e quindi

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

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).

Esempio. Supponiamo di voler risolvere ai minimi quadrati $Ax = b$ dove

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

Visto che $A = QR$ con

$$Q = \begin{bmatrix} -5.345224838248486e-01 & 8.406260324897477e-01 & 8.737040566610366e-02 \\ -2.672612419124244e-01 & -7.005216937414555e-02 & -9.610744623271416e-01 \\ -8.017837257372731e-01 & -5.370666318684501e-01 & 2.621112169983115e-01 \end{bmatrix}$$

$$R = \begin{bmatrix} -3.741657386773941e+00 & -4.543441112511214e+00 \\ 0 & 3.058944729337695e+00 \\ 0 & 0 \end{bmatrix}$$

e

$$A^T b = \begin{bmatrix} 6 \\ 8 \end{bmatrix}$$

per prima cosa risolviamo il sistema $R^T y = A^T b$ ovvero

$$\begin{cases} -3.741657386773941 \cdot y_1 + 0 \cdot y_2 + 0 \cdot y_3 = 6 \\ -4.543441112511214 \cdot y_1 + 3.058944729337695 y_2 + 0 \cdot y_3 = 8 \end{cases}$$

che ha soluzione

$$y = \begin{bmatrix} -1.603567451474547e+00 \\ 2.335072312471520e-01 \\ 0 \end{bmatrix}.$$

e quindi $Rx = y$ ovvero

$$\begin{cases} -3.741657386773941 \cdot x_1 - 4.543441112511214 \cdot x_2 = -1.603567451474547 \\ 0 \cdot x_1 + 3.058944729337695 \cdot x_2 = 0.2335072312471520 \\ 0 \cdot x_1 + 0 \cdot x_2 = 0. \end{cases}$$

Alla fine otteniamo

$$x = \begin{bmatrix} 3.358778625954200e-01 \\ 7.633587786259531e-02 \end{bmatrix}.$$

Facoltativo Dal punto di vista del condizionamento è meglio risolvere $R^T Rx = Q^T b$ invece di $A^T Ax = b$, perchè l'indice di condizionamento di R in norma 2 è molto più piccolo dell'indice di condizionamento di $A^T A$. Si può intuire osservando che se A è quadrata e simmetrica,

- $k_2(A^T A) = k_2(A^2) = (k_2(A))^2$,
- $k_2(R) \leq k_2(A)$ visto che $R = Q^T A$ e $\|Q\|_2 = \|Q^T\|_2 = \|Q^{-1}\| = 1$ da cui $k_2(Q^T) = 1$. In realtà si può mostrare che $k_2(R) = k_2(A)$.

5. Metodi iterativi. I metodi che in aritmetica esatta calcolano la soluzione x^* di $Ax = b$, $A \in \mathbb{R}^{n\times n}$ in un numero finito di passi, sono detti metodi *diretti*.

Un esempio è l'eliminazione gaussiana, che per matrici generali determina la soluzione in circa $n^3/3$ operazioni moltiplicative.

Nei problemi determinati dalle scienze applicate, può accadere che

- 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.

I metodi iterativi tipicamente non è detto determinino la soluzione esatta, ma possono approssimarla *arbitrariamente bene*. Tipicamente, ogni iterazione è basata sull'applicazione di un prodotto matrice-vettore, e quindi ha complessità di circa n^2 operazioni moltiplicative.

Di conseguenza, se raggiungono l'approssimazione della soluzione richiesta in un numero di operazioni $k \ll n^3/3$, essendo $kn^2 \ll n^3/3$ risultano convenienti rispetto la fattorizzazione LU.

TEOREMA 5.1. *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* $Ax = b$ *verifichi* $x^* = Bx^* + c$*.*

Allora qualsiasi sia $x_0 \in \mathbb{C}^n$*, la successione* $\{x^{(k)}\}$ *definita da*

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

converge a x^**.*

Dimostrazione. Essendo $\|B\| < 1$, per quanto visto nel teorema 1.5, abbiamo che la matrice $I - B$ è unica e quindi x^* è l'unica soluzione di $(I - B)x = c$ ovvero di $x = Bx + c$.

Essendo

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

e

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

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

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

Quindi

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

e per il teorema del confronto

$$\lim_k \|x^{(k)} - x^*\| = 0$$

ovvero $x^{(k)} \to x^*$. □

Nota. I metodi della forma

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

per la soluzione del sistema lineare $Ax = b$, sono detti usualmente *iterativi stazionari*, perchè B e c non dipendono dall'indice di iterazione k.

TEOREMA 5.2. *Siano*

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

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

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

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

Dimostrazione. Dimostriamo l'asserto nel caso $B = (b_{i,j}) \in \mathbb{C}^{n\times n}$ sia diagonalizzabile, ovvero in cui i suoi autovettori $\{v_k\}$ formino una base di $\mathbb{C}^n$.

($\Leftarrow$) Poniamo di seguito $e^{(k)} := x^{(k)} - x^*$ e supponiamo $\rho(B) < 1$.

Visto che $x^{(0)}, x^* \in \mathbb{C}^n$, abbiamo

$$e^{(0)} := x^{(0)} - x^{=} \sum_{s=1}^{n} c_s v_s$$

con $Bv_s = \lambda_s v_s$, $s = 1, \ldots, n$.

Ma allora, per la linearità di B,

$$\begin{aligned} e^{(1)} := x^{(1)} - x^* &= (Bx^{(0)} + c) - (Bx^* + c) \\ &= B(x^{(0)} - x^*) = Be^{(0)} = B\sum_{s=1}^{n} c_s v_s \\ &= \sum_{s=1}^{n} c_s Bv_s = \sum_{s=1}^{n} c_s \lambda_s v_s, \end{aligned} \tag{5.1}$$

come pure

$$\begin{aligned} e^{(2)} := x^{(2)} - x^* &= (Bx^{(1)} + c) - (Bx^* + c) \\ &= Be^{(1)} = B\sum_{s=1}^{n} c_s \lambda_s v_s \\ &= \sum_{s=1}^{n} c_s \lambda_s Bv_s = \sum_{s=1}^{n} c_s \lambda_s^2 v_s, \end{aligned} \tag{5.2}$$

e più in generale, per $k = 1, 2, 3, \ldots$

$$e^{(k)} := \sum_{s=1}^{n} c_s \lambda_s^k v_s, \tag{5.3}$$

Ma allora

$$0 \leq \|e^{(k)}\| := \|\sum_{s=1}^{n} c_s \lambda_s^k v_s\| \leq \sum_{s=1}^{n} |c_s||\lambda_s^k|\|v_s\|$$
$$\leq \max_{s=1,\dots,n} |\lambda_s^k| \sum_{s=1}^{n} |c_s|\|v_s\|$$
$$= (\rho(B))^k \sum_{s=1}^{n} |c_s|\|v_s\| \tag{5.4}$$

ed essendo $\rho(B) < 1$ concludiamo per il teorema del confronto che

$$\lim_k \|e^{(k)}\| = 0,$$

e quindi $x^{(k)} \to x^*$.

($\Rightarrow$) Mostriamo il fatto che se il metodo converge per qualsiasi $x^{(0)} \in \mathbb{C}^n$ allora $\rho(B) < 1$. Supponiamo per assurdo sia $\rho(B) \geq 1$. Quindi esiste un autovalore λ con autovettore v tale che $|\lambda| \geq 1$.

Sia $x^{(0)} \in \mathbb{C}^n$ scelto cosicchè sia $e^{(0)} = x^* - x^{(0)} = v$, ovvero $x^{(0)} = x^* - v$. Allora

$$e^{(1)} := x^{(1)} - x^* = (Bx^{(0)} + c) - (Bx^* + c)$$
$$= B(x^{(0)} - x^*) = Be^{(0)} = Bv = \lambda v, \tag{5.5}$$

come pure

$$e^{(2)} := x^{(2)} - x^* = (Bx^{(1)} + c) - (Bx^* + c)$$
$$= B(x^{(1)} - x^*) = Be^{(1)} = B(\lambda v) = \lambda Bv = \lambda^2 v \tag{5.6}$$

e più in generale

$$e^{(k)} = \lambda^k v.$$

Ma allora

$$\|e^{(k)}\| = |\lambda^k|\|v\| = |\lambda|^k\|v\|.$$

Se

- $|\lambda| > 1$ ne consegue che $\|e^{(k)}\| \to +\infty$,
- se $|\lambda| = 1$ abbiamo che $\|e^{(k)}\| = \|v\| \neq 0$ e quindi non è infinitesima, perchè costantemente uguale a una costante strettamente positiva.

□

Sia

- $A \in \mathbb{R}^{n\times n}$ una matrice quadrata,
- $A = P - N$ un cosidetto *splitting della matrice* A, con $\det(P) \neq 0$

Allora il sistema $(P - N)x = Ax = b$ e quindi $Px = Nx + b$ da cui, per l'invertibilità di P ricaviamo

$$x = P^{-1}Nx + P^{-1}b. \tag{5.7}$$

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

Questa formulazione suggerisce di considerare metodi le cui iterazioni siano fornite dalle iterazioni successive

$$x^{(k+1)}) = P^{-1}Nx^{(k)}) + P^{-1}b. \tag{5.8}$$

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].

Esempio. Mostriamo un esempio di come una matrice A può essere riscritta in termini di D, E, F come descritti sopra.

Se

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

allora

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

Nota. I metodi di Jacobi e Gauss-Seidel sono riscrivibili componente per componente rispettivamente come

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)})/a_{ii},\ i = 1, \ldots, n, \tag{5.9}$$

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right) /a_{ii},\ i = 1, \ldots, n. \tag{5.10}$$

TEOREMA 5.3. *Se la matrice A è a predominanza diagonale (stretta) per righe, allora i metodi di Jacobi e Gauss-Seidel convergono.*

Dimostrazione. Dimostriamo esclusivamente la parte relativa al metodo di Jacobi.

Essendo la matrice a predominanza diagonale stretta per righe, si ha

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

ovvero

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

Poichè $P = D$, $N = E + F$, allora le iterazioni sono del tipo $x^{(k+1)} = B_J x^{(k)} + c$ con

$$B_J = P^{-1}N = D^{-1}(E + F).$$

In particolare, visto che $(D^{-1}(E + F)|)_{i,j} = -a_{i,j}/a_{i,i}$ se $i \neq j$, e 0 altrimenti,

$$\begin{aligned} \|B_J\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1, j\neq i}^{n} | - a_{i,j}/a_{i,i}| \\ &= \max_{1\leq i\leq n} \sum_{j=1, j\neq i}^{n} |a_{i,j}|/|a_{i,i}| < 1 \end{aligned} \tag{5.11}$$

e quindi per (5.2), il metodo converge per ogni $x^{(0)} \in \mathbb{C}^n$. □ Inoltre,

TEOREMA 5.4. *Se la matrice A è simmetrica e definita positiva, allora il metodo di Gauss-Seidel converge.*

Per una dimostrazione si veda [6, p.132].

5.0.1. Alcuni esempi.

Esempio. 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]).

n	E_J	E_{GS}
10	$1.1e-01$	$1.8e-01$
30	$1.1e-03$	$5.1e-03$
50	$1.1e-05$	$1.4e-04$
70	$1.0e-07$	$4.0e-06$
90	$9.6e-10$	$1.1e-07$
110	$9.4e-12$	$3.2e-09$
130	$8.7e-14$	$8.8e-11$
150	$8.9e-16$	$2.5e-12$
170	–	$6.9e-14$
190	–	$1.9e-15$

TABELLA 5.1

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 $x_0 = 0 \in \mathbb{R}^3$.

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 $\rho(B_J) \approx 0.79$, $\rho(B_{GS}) \approx 0.83$.

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.

Esempio. Consideriamo il sistema lineare $Ax = b$ dove

$$A = \begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix},$$

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

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

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.

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 5.2

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 $x_0 = 0 \in \mathbb{R}^9$.

Esempio. 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 $x^* = (1, 1, 1)^T$.

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.

5.0.2. Metodi di Richardson stazionari. Da (5.7), essendo $A = P - N$ ovvero $N = P - A$ con P invertibile, deduciamo

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

e quindi il metodo delle approssimazioni successive (5.8) si può riscrivere come

$$\begin{aligned} x^{(k+1)} &= (I - P^{-1}A)x^{(k)} + P^{-1}b = x^{(k)} + P^{-1}(b - Ax^{(k)}) \\ &= x^{(k)} + P^{-1}r(x^{(k)}), \end{aligned} \tag{5.12}$$

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 5.3

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 $x^{(0)} = (1 - 10^{-14}, 1 - 10^{-14}, 1 - 10^{-14})^T \in \mathbb{R}^3$.

dove $r(x^{(k)}) = b - Ax^{(k)}$ è il vettore detto *residuo* al passo k-esimo.

Un metodo le cui iterazioni siano del tipo (5.12), si dice di *Richardson stazionario.*

Osserviamo che essendo $x^* = Bx^* + c$, con

- $B = P^{-1}N$, dove $A = P - N$ ovvero $N = P - A$,
- $c = P^{-1}b$,

allora

$$\begin{aligned} P^{-1}Ax^* &= (I - I + P^{-1}A))x^* = (I - P^{-1}(P - A))x^* \\ &= (I - P^{-1}N)x^* = (I - B)x^* = c = P^{-1}b. \end{aligned}$$

Quindi

$$P^{-1}Ax^* = P^{-1}b, \tag{5.13}$$

ovvero invece di $Ax = b$ si studia il problema equivalente (5.13) in cui ci si *augura* che $k(P^{-1}A) \ll k(A)$. Una tal matrice si chiama di *precondizionamento.* Di conseguenza la matrice P dello splitting $A = P - N$ può essere vista come matrice di precondizionamento.

Si mostra che l'azione di P^{-1}è efficace quando $P^{-1} \approx A^{-1}$, nel senso che gli autovalori di $P^{-1}A$ si accumulano intorno ad 1 e quindi la convergenza diventa più rapida.

Nota. Si noti che per quanto concerne i metodi di Jacobi e di Gauss-Seidel, dato un vettore v, il calcolo di $z = P^1v$, ovvero la soluzione del sistema $Pz = v$, ha un basso costo computazionale, essendo P diagonale o triangolare superiore.

Nota. Vari metodi introducono un parametro α di rilassamento,

$$x^{(k+1)} = x^{(k)} + \alpha P^{-1}r(x^{(k)})$$

che aumenti l'efficacia del precondizionatore, cercando di diminuire o addirittura minimizzare il raggio spettrale di $B_\alpha = I - \alpha P^{-1}A$, che ha l'effetto di aumentare la velocità di convergenza.

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

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

viene arrestato utilizzando

- il *criterio dello step* (cf. [6, p.159]), 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

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

Si dimostra che vale la seguente stima dell'errore relativo

$$\frac{\|x^* - x^{(k)}\|}{x^*} \leq k(A) \frac{\|b - Ax^{(k)}\|}{\|b\|}.$$

Quindi, se $k(A) \gg 1$, nonostante $\frac{\|b-Ax^{(k)}\|}{\|b\|} < \epsilon$, ovvero possa essere piccolo, possiamo avere che l'errore relativo $\frac{\|x^*-x^{(k)}\|}{x^*}$ sia molto grande (cf. [6, p.161]), mentre la stima è buona qualora $k(A) \approx 1$.

RIFERIMENTI BIBLIOGRAFICI

[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] A. Quarteroni e F. Saleri, *Elementi di calcolo numerico*, Progetto Leonardo, 1999.
[6] A. Quarteroni, R. Sacco e F. Saleri, *Matematica Numerica*, Springer Verlag, 1998.
[7] Wikipedia, Decomposizione di Cholesky, https://it.wikipedia.org/wiki/Decomposizione_di_Cholesky
[8] Wikipedia, Decomposizione LU, https://it.wikipedia.org/wiki/Decomposizione_LU
[9] Wikipedia, Decomposizione QR, https://it.wikipedia.org/wiki/Decomposizione_QR
[10] Wikipedia, Matrice a diagonale dominante, https://it.wikipedia.org/wiki/Matrice_a_diagonale_dominante
[11] Wikipedia, Matrice definita positiva, https://it.wikipedia.org/wiki/Matrice_definita_positiva
[12] Wikipedia, Metodo della matrice dei cofattori, Matrice_invertibile#Metodo_della_matrice_dei_cofattori.
[13] Wikipedia, Supercomputer, https://it.wikipedia.org/wiki/Supercomputer.
[14] Wikipedia, Teorema di Binet, https://it.wikipedia.org/wiki/Teorema_di_Binet.
[15] Wikipedia, Teorema di Laplace, https://it.wikipedia.org/wiki/Teorema_di_Laplace.