ARGOMENTI DEL CORSO CALCOLO NUMERICO

A.A. 2009/10

$$A = L \cdot U$$

$$L = \begin{pmatrix} 1 & & & \\ \ell_{2,1} & 1 & & \\ \vdots & & \ddots & \\ \ell_{n,1} & \dots & \ell_{n,n-1} & 1 \end{pmatrix} \quad U = \begin{pmatrix} u_{1,1} & \dots & & u_{1,n} \\ & u_{2,2} & & \\ & & \ddots & \vdots \\ & & & u_{n,n} \end{pmatrix}$$

Algebra Lineare Numerica

Giulio Casciola

(novembre 2004, rivista e corretta novembre 2009)

Indice

Capitolo 1

Sistemi Lineari: metodi diretti

Uno dei problemi più frequenti nel calcolo scientifico è la soluzione di un sistema lineare; se questo è composto da tante equazioni quante incognite si dice sistema quadrato o normale. In forma matriciale può essere scritto come

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

dove A è una data matrice di ordine $n \times n$, $\mathbf{b}$ è un dato vettore colonna con n elementi ed $\mathbf{x}$ è il vettore delle incognite.
In algebra lineare si studiano metodi per risolvere sistemi lineari non singolari. Un metodo noto è quello di Cramer (od anche regola di Cramer) nel quale ogni componente della soluzione è espressa come quoziente di un determinante sempre diverso a numeratore e di uno stesso determinante a denominatore, così che la soluzione del sistema lineare si riduce al calcolo di $n + 1$ determinanti di ordine n. Se si cerca di risolvere un sistema di 20 equazioni con la regola di Cramer, sarebbe necessario calcolare 21 determinanti di ordine 20. Come è noto per calcolare il determinante di una matrice $n \times n$, per esempio con la formula di Laplace o quella di Leibniz, servono $(n-1)n!$ moltiplicazioni; nel nostro esempio numerico sarà $19 \cdot 20!$ e quindi l'intero sistema lineare comporterà $19 \cdot 20! \cdot 21$ moltiplicazioni, più un ugual numero di addizioni. Su un normale PC, oggi si possono fare circa 10^9 moltiplicazioni al secondo (Giga FLOPS FLoating point OPeration per Second), così che solo le moltiplicazioni richiederanno circa 30782 anni, sempre che non manchi la corrente durante il calcolo.
In algebra lineare si insegna anche che la soluzione di $A\mathbf{x} = \mathbf{b}$ può essere scritta come $\mathbf{x} = A^{-1}\mathbf{b}$, dove A^{-1} è l'inversa di A. Comunque nella maggioranza dei problemi pratici, ciò non è necessario ed eventualmente risulta inopportuno calcolare A^{-1}. Con un esempio estremo, ma illustrativo,

consideriamo un sistema di appena una equazione, come

$$7x = 21.$$

Il miglior modo per risolvere tale sistema è con la divisione

$$x = \frac{21}{7} = 3.$$

L'uso della matrice inversa porterebbe invece a

$$x = 7^{-1} \cdot 21 = 0.142857\ldots \cdot 21 = 2.99997\ldots.$$

L'inversa richiede più operazioni, una divisione ed una moltiplicazione invece di appena una divisione. È il maggior numero di operazioni che ci induce ad evitare il calcolo dell'inversa. Ci concentreremo quindi sulla soluzione diretta di sistemi piuttosto che sul calcolo dell'inversa.

Vedremo essenzialmente due metodi diretti per risolvere un sistema lineare $A\,\mathbf{x} = \mathbf{b}$. Entrambi i metodi si basano su una *fattorizzazione* della matrice A, ossia costruiscono due matrici il cui **prodotto** sia uguale ad A:

1. Fattorizzazione LU (L (Low) matrice triangolare inferiore e diagonale unitaria, U (Up) matrice triangolare superiore)

$$A = L\ U$$

2. Fattorizzazione QR (Q matrice ortogonale, R (Right) matrice triangolare superiore)

$$A = Q\ R$$

Lo scopo è quello di ricondurre la soluzione del sistema (pieno e/o senza struttura particolare) $A\mathbf{x} = \mathbf{b}$ alla soluzione di uno o due sistemi (sparsi e/o con particolare struttura); la soluzione del sistema si ottiene in due fasi:

1. fattorizzazione di A;

2. soluzione di sistemi lineari di forma triangolare, le cui caratteristiche di soluzione sono più semplici e meno costose (rispetto a quelle del sistema originario).

1.1 Fattorizzazione LU

In questa sezione si studia la possibilità di fattorizzare la matrice A nel prodotto di una matrice L triangolare inferiore per una matrice U triangolare superiore:

$$A_{n\times n} = L_{n\times n} \cdot U_{n\times n}$$

con

$$L = \begin{pmatrix} 1 & & & & \\ \ell_{2,1} & 1 & & & \\ \ell_{3,1} & & 1 & & \\ \vdots & & & \ddots & \\ \ell_{n,1} & \dots & & \ell_{n,n-1} & 1 \end{pmatrix} \quad U = \begin{pmatrix} u_{1,1} & \dots & & & u_{1,n} \\ & u_{2,2} & & & \\ & & u_{3,3} & & \\ & & & \ddots & \vdots \\ & & & & u_{n,n} \end{pmatrix}.$$

L'esistenza di una tale fattorizzazione renderebbe facile la determinazione della soluzione del sistema normale non singolare, infatti si avrebbe:

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

e ponendo $U\mathbf{x} = \mathbf{y}$ si potrebbe risolvere il sistema

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

per sostituzione in avanti (forward-substitution); determinato $\mathbf{y}$ si risolve il sistema

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

per sostituzione all'indietro (backward-substitution).

1.1.1 Sostituzione in Avanti

Il primo sistema da risolvere è

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

in cui L è una matrice triangolare inferiore.

- La soluzione si ottiene con la sostituzione in avanti, mediante l'algoritmo che segue

  $y(1) = b(1)/L(1,1)$

  per $i = 2, ..., n$

  $y(i) = b(i) - L(i, 1 : i-1) \ * \ y(1 : i-1)$

$y(i) = y(i)/L(i,i)$

La struttura dell'algoritmo è triangolare. Per generare $y(i)$ si compiono $2i-1$ operazioni e per la precisione $i-1$ moltiplicazioni, 1 divisione e $i-1$ sottrazioni.

- Il costo computazionale della sostituzione in avanti è dato da

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

- Se L è una matrice triangolare inferiore con elementi diagonali unitari, allora nell'algoritmo non ci sono divisioni:

  $y(1) = b(1)$

  per $i = 2, ..., n$

  $y(i) = b(i) - L(i, 1:i-1) \ * \ y(1:i-1)$

  ed il costo computazionale diventa

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

Esempio 1.1 *Sostituzione in avanti.*

$$\begin{pmatrix} \ell_{1,1} & 0 & 0 \\ \ell_{2,1} & \ell_{2,2} & 0 \\ \ell_{3,1} & \ell_{3,2} & \ell_{3,3} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

$y_1 = b_1/\ell_{1,1}$ — 1 *operazione*

$\downarrow$

$y_2 = (b_2 - \ell_{2,1} * y_1)/\ell_{2,2}$ — 3 *operazioni*

$\downarrow$

$y_3 = (b_3 - \ell_{3,1} * y_1 - \ell_{3,2} * y_2)/\ell_{3,3}$ — 5 *operazioni*

Costo computazionale: $n^2 = 3^2 = 9$ *operazioni.*

1.1.2 Sostituzione all'indietro

Il secondo sistema da risolvere è

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

in cui U è una matrice triangolare superiore.

- La soluzione si ottiene con la sostituzione all'indietro, mediante l'algoritmo che segue

  $x(n) = y(n)/U(n,n)$

  per $i = n-1, ..., 1$

  $x(i) = y(i) - U(i, i+1:n) \ * \ x(i+1:n)$

  $x(i) = x(i)/U(i,i)$

  La struttura dell'algoritmo è triangolare. Per generare $x(i)$ si compiono $2i-1$ operazioni e per la precisione $i-1$ moltiplicazioni, 1 divisione e $i-1$ sottrazioni.

- Il costo computazionale della sostituzione all'indietro è dato da

$$1 + \sum_{i=2}^{n}(2i-1) = ... = n^2$$

Esempio 1.2 *Sostituzione all'indietro*

$$\begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ 0 & u_{2,2} & u_{2,3} \\ 0 & 0 & u_{3,3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$x_3 = y_3/u_{3,3}$ — 1 *operazione*

$\downarrow$

$x_2 = (y_2 - u_{2,3} * x_3)/u_{2,2}$ — 3 *operazioni*

$\downarrow$

$x_1 = (y_1 - u_{1,2} * x_2 - u_{1,3} * x_3)/u_{1,1}$ — 5 *operazioni*

Costo computazionale: $n^2 = 3^2 = 9$ *operazioni.*

Riassumendo: determinata una fattorizzazione LU di A, il vettore soluzione $\mathbf{x}$ si può ottenere facilmente risolvendo due sistemi triangolari, uno inferiore e l'altro superiore.

Ci si concentrerà allora sulla fattorizzazione LU di A; si osserva che la sola ipotesi che A non sia singolare non garantisce l'esistenza di una fattorizzazione LU, infatti vale il seguente teorema.

Teorema 1.1 *Se i minori principali di ordine k di A per $k = 1, \ldots, n-1$ sono diversi da zero, allora esiste una ed una sola fattorizzazione LU di A.*

1.1.3 Metodo di Gauss

Procediamo in un caso semplice ad illustrare come funziona il metodo di Gauss per fattorizzare LU una matrice $A_{n\times n}$ (il procedimento in oggetto può essere applicato anche a matrici rettangolari).
Sia

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix}$$

con

$$a_{1,1} \neq 0 \quad \text{e} \quad det \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1} \neq 0.$$

Consideriamo la matrice L_1 siffatta:

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{a_{2,1}}{a_{1,1}} & 1 & 0 \\ -\frac{a_{3,1}}{a_{1,1}} & 0 & 1 \end{pmatrix}$$

(si noti la necessità che $a_{1,1} \neq 0$; $a_{1,1}$ è detto perno o pivot) allora sarà:

$$L_1 A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & a^{(1)}_{2,2} & a^{(1)}_{2,3} \\ 0 & a^{(1)}_{3,2} & a^{(1)}_{3,3} \end{pmatrix}$$

con $a^{(1)}_{i,j} = a_{i,j} - \frac{a_{i,1}}{a_{1,1}} a_{1,j},\ i,j = 2,3.$
Consideriamo ora la matrice L_2 siffatta:

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a^{(1)}_{3,2}}{a^{(1)}_{2,2}} & 1 \end{pmatrix}$$

(si noti la necessità che $a^{(1)}_{2,2} = \frac{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}{a_{1,1}} \neq 0$, ossia che $a_{1,1}a_{2,2} - a_{1,2}a_{2,1} \neq 0$; $a^{(1)}_{2,2}$ è detto perno o pivot) allora sarà:

$$L_2(L_1 A) = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & a^{(1)}_{2,2} & a^{(1)}_{2,3} \\ 0 & 0 & a^{(2)}_{3,3} \end{pmatrix}$$

con $a^{(2)}_{3,3} = a^{(1)}_{3,3} - \dfrac{a^{(1)}_{3,2}}{a^{(1)}_{2,2}} a^{(1)}_{2,3}$.
Chiamiamo U la $L_2 L_1 A$ e notiamo che la U è del tipo cercato. Si noti che la L_1 e la L_2 sono non singolari, allora

$$A = L_1^{-1} L_2^{-1} U.$$

Chi sono L_1^{-1} e L_2^{-1} e chi è $L_1^{-1} L_2^{-1}$? Saranno:

$$L_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{a_{2,1}}{a_{1,1}} & 1 & 0 \\ \frac{a_{3,1}}{a_{1,1}} & 0 & 1 \end{pmatrix} \quad L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{a^{(1)}_{3,2}}{a^{(1)}_{2,2}} & 1 \end{pmatrix}$$

$$L_1^{-1} L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{a_{2,1}}{a_{1,1}} & 1 & 0 \\ \frac{a_{3,1}}{a_{1,1}} & \frac{a^{(1)}_{3,2}}{a^{(1)}_{2,2}} & 1 \end{pmatrix}$$

e se chiamiamo quest'ultima L avremo che $A = L\ U$ come volevamo.
$U \longleftarrow A; \qquad L \longleftarrow I_n;$
per $k = 1, .., n-1$
per $j = k+1, .., n$
$\ell(j,k) = u(j,k)/u(k,k)$
$u(j, k+1:n) \ = \ u(j, k+1:n) - \ell(j,k) \ * \ u(k, k+1:n)$

Per determinare la fattorizzazione suddetta sono necessarie $n(n+1)/2$ moltiplicazioni/divisioni floating point per il calcolo degli elementi di L e $(n-1)^2+(n-2)^2+\ldots+2^2+1 = (n-1)n(2n-1)/6$ moltiplicazioni/divisioni floating point per gli elemnti di U. La complessità computazionale, in termini di operazioni di moltiplicazione e divisione, è quindi:

$$\frac{(n-1)n(n+1)}{3} = \frac{n^3 - n^2}{3} \approx \frac{1}{3} n^3.$$

Esempio 1.3 *Fattorizzazione $A = L\ U$ di Gauss.*

Sia $A\mathbf{x} = \mathbf{b}$ *con* $A = \begin{pmatrix} 2 & 1 & 0 \\ 4 & 5 & 2 \\ 6 & 15 & 12 \end{pmatrix}$ *e* $\mathbf{b} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$.

L_1 deve essere triangolare inferiore e tale da rendere nulli gli elementi della prima colonna di A sotto l'elemento $a_{11} = 2$.

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}$$

$$L_1\ A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 0 \\ 4 & 5 & 2 \\ 6 & 15 & 12 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 12 & 12 \end{pmatrix}$$

L_2 deve essere triangolare inferiore e tale da rendere nulli gli elementi della seconda colonna di $L_1\ A$ sotto ad $a^{(1)}_{2,2} = 3$.

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{pmatrix}$$

$$L_2\ (L_1\ A) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 12 & 12 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{pmatrix} = U$$

Allora

$$L_2\ L_1\ A = U,$$

$$A = L_1^{-1}\ L_2^{-1}\ U = L\ U$$

e risulta

$$L = L_1^{-1}\ L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix}.$$

Fattorizzata la matrice, si procede alla soluzione dei sistemi triangolari

$$L\mathbf{y} = \mathbf{b} \quad e \quad U\mathbf{x} = \mathbf{y}.$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \quad y_1 = 2;$$

$$\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \end{pmatrix} \quad y_2 = -3;$$

$$(1)(y_3) = (8) \quad y_3 = 8;$$

$$\begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \\ 8 \end{pmatrix} \quad x_3 = 2;$$

$$\begin{pmatrix} 2 & 1 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -7 \end{pmatrix} \quad x_2 = -7/3;$$

$$(2)(x_1) = (13/3) \quad x_1 = 13/6.$$

1.1.4 Calcolo di A^{-1}

Sia $A \in \mathbb{R}^{n\times n}$, invertibile; vogliamo calcolare A^{-1}, risolvendo il sistema

$$A\ X = I_n, \qquad X = [\mathbf{x}_1 \mid \mathbf{x}_2 \mid ..... \mid \mathbf{x}_n], \qquad I_n = [\mathbf{e}_1 \mid \mathbf{e}_2 \mid ..... \mid \mathbf{e}_n]$$

I_n è la matrice identità di dimensione n; i vettori $\mathbf{e}_1,\ ...\ \mathbf{e}_n$ sono la base canonica di $\mathbb{R}^n$. $X \in \mathbb{R}^{n\times n}$ è la matrice incognita.
Risolvere $A\ X = I_n$, per ottenere $X = A^{-1}\ I_n$, è equivalente a risolvere n sistemi lineari, tutti con la stessa matrice A dei coefficienti

$$\begin{cases} A\ \mathbf{x}_1 = \mathbf{e}_1 \\ A\ \mathbf{x}_2 = \mathbf{e}_2 \\ \quad \vdots \\ A\ \mathbf{x}_n = \mathbf{e}_n \end{cases}$$

Si fattorizzi $A = L\ U$ una sola volta, e poi si risolvano 2 n sistemi lineari di forma triangolare

$$\begin{cases} \textbf{(i)} \quad A = L\ U & \frac{1}{3}n^3 - \frac{1}{3}n^2 \\ \textbf{(ii)} \quad \begin{cases} L\mathbf{y}_1 = \mathbf{e}_1 & n^2 - n \\ U\mathbf{x}_1 = \mathbf{y}_1 & n^2 \end{cases} \\ \quad \vdots \\ \quad \vdots \\ \quad \begin{cases} L\mathbf{y}_n = \mathbf{e}_n & n^2 - n \\ U\mathbf{x}_n = \mathbf{y}_n & n^2 \end{cases} \end{cases}$$

1.1.5 Calcolo del $det(A)$

Fattorizzare $A = L\ U$ ci permette di calcolare con costo computazionale basso il determinante della matrice A; vale

$$det(A) = det(L)\ det(U) = det(U) = \prod_{i=1}^{n} u_{i,i}$$

Abbiamo utilizzato il Teorema di Binet, che $det(L) = 1$ e che il determinante di una matrice triangolare è dato dal prodotto degli elementi della diagonale.

1.1.6 Metodo di Gauss con scambio delle righe

Sia dato il sistema lineare

$$\begin{pmatrix} 0 & 3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \end{pmatrix};$$

essendo $a_{1,1} = 0$, non si può applicare l'algoritmo di fattorizzazione di Gauss alla matrice dei coefficienti, ma si osserva che questo sistema è equivalente a

$$\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \end{pmatrix};$$

ed ora $a_{1,1} \neq 0$.
Allora se A è non singolare è sempre possibile trovare una matrice di permutazione P per cui $P\ A$ sia fattorizzabile $L\ U$, cioè

$$P\ A = L\ U.$$

Matrice di Permutazione: si ottiene dalla matrice identità I_n permutando le righe. Applichiamo P ad una matrice $A \in \mathbb{R}^{n \times n}$:

$$P\ A \quad \longrightarrow \quad \text{permuta righe di } A$$

$$A\ P \quad \longrightarrow \quad \text{permuta colonne di } A$$

Ad esempio:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \qquad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

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

Ovviamente non è pensabile determinare in qualche modo una matrice P opportuna, premoltiplicarla per A (scambio di righe e quindi ordine delle equazioni) e poi applicare l'algoritmo visto, bensì si dovrà determinare P costruttivamente. Questo è ciò che fa l'algoritmo di eliminazione di Gauss con scambio delle righe (pivoting parziale). Si illusta il metodo con un esempio. Sia data la matrice

$$A = A_0 = \begin{pmatrix} 4 & -8 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{pmatrix};$$

il pivot $a_{i,1}^{(0)}$ è 4 e quindi $P_1 = I$ e

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/4 & 0 & 1 \end{pmatrix};$$

si costruisce

$$A_1 = L_1\ P_1\ A_0 = \begin{pmatrix} 4 & -8 & 2 \\ 0 & 0 & 5 \\ 0 & 1 & 5/2 \end{pmatrix}.$$

Il pivot $a_{i,2}^{(1)}$ è 1; le matrici P_2 ed L_2 saranno:

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Si costruisce

$$A_2 = L_2\ P_2\ A_1 = \begin{pmatrix} 4 & -8 & 2 \\ 0 & 1 & 5/2 \\ 0 & 0 & 5 \end{pmatrix} = U.$$

In definitiva è

$$L_2\ P_2\ L_1\ P_1\ A = U.$$

Si noti che

$$L_2\ P_2\ L_1\ P_1 = L_2\ P_2\ L_1\ P_2^{-1}\ P_2 P_1$$

e se poniamo $T = L_2\ P_2\ L_1\ P_2^{-1}$ si ha che T è triangolare inferiore e quindi

$$T\ P_2\ P_1\ A = U$$

e ponendo

$$L^{-1} = T \quad \text{e} \quad P = P_2\ P_1$$

si ha

$$P\ A = L\ U.$$

Completando l'esempio si ha

$$L = P_2\ L_1^{-1}\ P_2^{-1}\ L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

ed infatti

$$L\ U = \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & -8 & 2 \\ 0 & 1 & 5/2 \\ 0 & 0 & 5 \end{pmatrix} = \begin{pmatrix} 4 & -8 & 2 \\ 1 & -1 & 3 \\ 2 & -4 & 6 \end{pmatrix} = P\ A.$$

Nell'implementare tale metodo ci si accorge immediatamente che non è possibile perdere l'informazione P, infatti se fattorizziamo $P\ A$ il sistema da risolvere sarà

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

1.1.7 Esistenza della Fattorizzazione $P\ A = L\ U$

Che cosa possiamo fare se dobbiamo considerare una matrice A per la quale il metodo di fattorizzazione semplice di Gauss fallisce? Arrichiamo il metodo con un procedimento di pivotaggio. Vale infatti il seguente teorema.

Teorema 1.2 (Esistenza di $P\ A = L\ U$) *Per ogni matrice $A \in \mathbb{R}^{n\times n}$ non singolare, esiste una matrice di permutazione P tale che*

$$P\ A = L\ U$$

con L triangolare inferiore con elementi unitari e U triangolare superiore.

Il teorema enunciato assicura che, se la fattorizzazione di Gauss di una matrice A non singolare sembra fallire (perché ad un certo passo k si ottiene un perno $a_{k,k}^{(k)} \approx 0$), è possibile superare questo passo andando a cercare (nella colonna $k-$esima della matrice A_k corrente) un nuovo perno $a_{i,k}^{(k)} \neq 0$ ed introducendo una permutazione di righe in modo che

$a_{i,k}^{(k)}$ occupi la posizione di $a_{kk}^{(k)}$

$$L_k \ ... \ L_1 \ A = \begin{pmatrix} \ddots & \vdots & & & \\ \\ \dots & a_{k,k}^{(k)} = 0 & \dots & \leftarrow & \text{riga } k \\ & a_{k+1,k}^{(k)} & & & \\ & \vdots & & & \\ & a_{i,k}^{(k)} \neq 0 & \dots & \leftarrow & \text{riga } i \\ & \vdots & & & \\ & a_{n,k}^{(k)} & \dots & & \\ \\ & \uparrow & & & \\ & \text{colonna } k & & & \end{pmatrix}.$$

1.1.8 Stabilità della Fattorizzazione $L\ U$

Poiché le operazioni aritmetiche che intervengono nell'algoritmo di fattorizzazione sono effettuate in aritmetica finita, segue che questi algoritmi, anziché generare i fattori L ed U per A, generano dei fattori non esatti

$$\tilde{L} = L + \delta L \quad \text{e} \quad \tilde{U} = U + \delta U.$$

Posto $A + \delta A = \tilde{L}\ \tilde{U}$, cioè la matrice di cui $\tilde{L}\ \tilde{U}$ è effettivamente una fattorizzazione, si ha

$$A + \delta A = \tilde{L}\ \tilde{U} = (L + \delta L)(U + \delta U) =$$

$$= L\ U + U\ \delta L + L\ \delta U + \delta L\ \delta U$$

da cui

$$\delta A = U\ \delta L + L\ \delta U + \delta L\ \delta U$$

da cui segue che se gli elementi di L ed U sono grandi, gli elementi di δA sono grandi e quindi, gli errori di arrotondamento si amplificano (analisi dell'errore all'indietro). Perciò diremo che la fattorizzazione $A = L\ U$ è **stabile** numericamente se gli elementi di L ed U non sono troppo grandi rispetto agli elementi di A. In particolare se esistono delle costanti a e b indipendenti dagli elementi e dall'ordine di A tali che $|\ell_{i,j}| \leq a$ e $|u_{i,j}| < b$, allora si dice che la fattorizzazione $L\ U$ è **stabile in senso forte**; se le costanti a e b dipendono dall'ordine di A, allora si dice che la fattorizzazione $L\ U$ è **stabile in senso debole**.
Sebbene si sia visto che per ogni matrice, a meno di una permutazione

di righe, esiste sempre la fattorizzazione $L\ U$, in generale questa non è stabile. Infatti gli elementi

$$\ell_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}$$

possono assumere valori grandi e quindi gli elementi di L possono crescere oltre ogni limite.
Questo inconveniente può essere eliminato facendo uno scambio di righe ad ogni passo dell'algoritmo, così da scegliere i pivot $a_{k,k}^{(k)}$ in modo che

$$|a_{k,k}^{(k)}| \geq \{|a_{i,k}^{(k)}|\}_{i=k,\ldots,n}.$$

Questa strategia è indicata con **pivoting parziale** e limita ad 1 il valore degli elementi della matrice L, mentre non riesce a limitare gli elementi di U che possono crescere esponenzialmente con l'ordine n di A. Infatti si ha che:

$$\max |u_{i,j}| \leq 2^{n-1} \max |a_{i,j}|.$$

Perciò l'algoritmo di Gauss con pivoting parziale genera una fattorizzazione **stabile in senso debole** infatti si ha $a = 1$ e $b = 2^{n-1}$.

1.1.9 Metodo di Gauss con scambio delle righe e perno massimo

Le considerazioni sopra esposte portano a modificare l'algoritmo di Gauss introducendo lo scambio delle righe non solo per individuare un pivot non nullo, ma quello massimo. Questo si realizza al passo $k-$esimo scegliendo come pivot fra gli $a_{i,k}^{(k)}$ con $i = k, \ldots, n$ il più grande in valore assoluto, cioè

$$|a_{k,k}^{(k)}| \geq \max_{i=k+1,\ldots,n} \{|a_{i,k}^{(k)}|\}.$$

Quando si usa questo criterio per la scelta dei pivot, il metodo di Gauss si chiama metodo di eliminazione di Gauss con scambio delle righe e perno (o pivot) massimo.

1.2 Condizionamento del Problema $A\mathbf{x} = \mathbf{b}$

In questa sezione si vuole esaminare come perturbazioni sugli elementi della matrice A e sugli elementi del termine noto $\mathbf{b}$ influenzano la soluzione $\mathbf{x}$ del sistema lineare. Queste perturbazioni sono tipicamente dovute agli errori di approssimazione quando la matrice A ed il termine noto $\mathbf{b}$

vengono rappresentati con numeri finiti. Per essere in grado di stimare gli errori, dobbiamo introdurre una misura delle *dimensioni* di un vettore o *distanza fra vettori*

1.2.1 Richiami sul concetto di norma

Siano $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Una **norma vettoriale** $\|\cdot\|$ è una funzione $\mathbb{R}^n \to \mathbb{R}$ che associa ad un vettore di $\mathbb{R}^n$ un valore reale (la lunghezza di quel vettore). Tale funzione, per essere una norma vettoriale deve soddisfare le tre seguenti proprietà:

1. $\|\mathbf{x}\| \geq 0 \quad \forall \mathbf{x} \quad$ e $\quad \|\mathbf{x}\| = 0 \quad$ se e solo se $\quad \mathbf{x} = \mathbf{0}$
2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (disuguaglianza triangolare)
3. $\|\alpha\mathbf{x}\| = |\alpha| \ \|\mathbf{x}\| \quad$ con $\quad \alpha \in \mathbb{R}$.

Le norme vettoriali più importanti sono:

- $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- $\|\mathbf{x}\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}} = \sqrt{\mathbf{x}^T\mathbf{x}} = (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}$
- $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Questi sono tutti casi particolari della norma p

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}.$$

Osservazione 1.1 *La $\|\cdot\|_2$ è una generalizzazione ad $\mathbb{R}^n$ dell'usuale distanza in $\mathbb{R}^2$ o $\mathbb{R}^3$ ed è detta norma Euclidea. La $\|\cdot\|_\infty$ è detta norma infinito o norma del massimo.*

Con le norme, possiamo introdurre i concetti di **distanza** e **continuità** in $\mathbb{R}^n$. Sia $\tilde{\mathbf{x}}$ un vettore approssimazione di un vettore $\mathbf{x}$ non nullo. Per una data norma vettoriale $\|\cdot\|$, si definisce
errore assoluto

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|$$

ed **errore relativo**

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|}.$$

Teorema 1.3 (di equivalenza delle norme) *Siano* $\|\cdot\|'$ *e* $\|\cdot\|''$ *due norme vettoriali. Allora le due norme sono equivalenti, cioè esistono* α *e* $\beta \in \mathbb{R}$ *con* $0 < \alpha \leq \beta$*, tali che per ogni* $\mathbf{x} \in \mathbb{R}^n$ *è*

$$\alpha \ \|\cdot\|'' \leq \|\cdot\|' \leq \beta \ \|\cdot\|''.$$

Teorema 1.4 *Per ogni* $\mathbf{x} \in \mathbb{R}^n$ *si ha:*

- $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \ \|\mathbf{x}\|_\infty$
- $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \ \|\mathbf{x}\|_2$
- $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n \ \|\mathbf{x}\|_\infty$

Sia A una matrice $m \times n$ cioè $A \in \mathbb{R}^{m\times n}$. Possiamo pensare di definire una norma di matrici per misurare la dimensione di una matrice e la distanza fra matrici.
Una **norma matriciale** $\|\cdot\|$ è una funzione $\mathbb{R}^{m\times n} \to \mathbb{R}$ che associa ad una matrice di $\mathbb{R}^{m\times n}$ un valore reale. Tale funzione, per essere una norma matriciale deve soddisfare le tre seguenti proprietà:

1. $\|A\| \geq 0 \quad \forall A \quad$ e $\quad \|A\| = 0 \quad$ se e solo se $\quad A = 0$
2. $\|A + B\| \leq \|A\| + \|B\|$ (disuguaglianza triangolare)
3. $\|\alpha A\| = |\alpha| \ \|A\| \quad$ con $\quad \alpha \in \mathbb{R}$.

Definizione 1.1 (Norma indotta di matrice) *Per ogni norma vettoriale, possiamo definire una corrispondente norma matriciale come*

$$\|A\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A \ \mathbf{x}\|}{\|\mathbf{x}\|}.$$

Osservazione 1.2 *Si noti che se* $A \in \mathbb{R}^{m\times n}$*, deve essere* $\mathbf{x} \in \mathbb{R}^n$ *da cui si tratta del max di una norma in* $\mathbb{R}^m$ *su una norma in* $\mathbb{R}^n$*.*

Così definita, $\|A\|$ è una norma di matrice, cioè soddisfa le tre proprietà date.

Risultato 1.1 *Se* $\|\cdot\|$ *denota una norma vettoriale e la corrispondente norma matriciale indotta, allora*

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

$$\|A \ B\| \leq \|A\| \cdot \|B\|$$

Le norme matriciali più importanti sono:

- $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}|$
- $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{i,j}|$
- $\|A\|_2 = \left(\lambda_{max}(A^T\ A)\right)^{1/2}$ (norma spettrale)

Risultato 1.2 (Relazioni fra le norme matriciali) *Per le norme che sono state introdotte valgono le seguenti relazioni, che possono essere dimostrate dalle relazioni fre le norme vettoriali e la definizione di norma indotta:*

- $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\ \|A\|_\infty$
- $\frac{1}{\sqrt{n}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\ \|A\|_1$
- $\max_{i,j} |a_{i,j}| \leq \|A\|_2 \leq n\ \max_{i,j} |a_{i,j}|$
- $\|A\|_2 \leq \sqrt{\|A\|_1\ \|A\|_\infty}$

1.2.2 Errore Inerente

Affrontiamo lo studio dell'errore inerente del problema $A\ \mathbf{x} = \mathbf{b}$ considerando separatamente eventuali perturbazioni sulla matrice A e sul vettore dei termini noti $\mathbf{b}$.

Introduciamo un vettore di perturbazione $\delta\mathbf{b} \in \mathbb{R}^n$ sul termine noto; cerchiamo $\mathbf{x} + \delta\mathbf{x} \in \mathbb{R}^n$ soluzione del sistema perturbato

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

poiché è $A\ \mathbf{x} = \mathbf{b}$ risulterà

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

da cui

$$\delta\mathbf{x} = A^{-1}\ \delta\mathbf{b}.$$

Passando alle norme

$$\|\delta\mathbf{x}\| = \|A^{-1}\ \delta\mathbf{b}\| \leq \|A^{-1}\| \cdot \|\delta\mathbf{b}\|$$

inoltre vale

$$\|\mathbf{b}\| = \|A\ \mathbf{x}\| \leq \|A\| \cdot \|\mathbf{x}\|.$$

Allora

$$\begin{aligned}\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} &\leq \frac{\|A^{-1}\| \cdot \|\delta\mathbf{b}\|}{\|\mathbf{x}\|} \\ &\leq \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.\end{aligned}$$

Dunque il condizionamento di $A\ \mathbf{x} = \mathbf{b}$ dipende dalla costante

$$K = \|A\| \cdot \|A^{-1}\|$$

detto **numero di condizione** di A (rispetto a perturbazioni sul termine noto).

Introduciamo ora una matrice di perturbazione $\delta A \in \mathbb{R}^{n\times n}$ sulla matrice dei coefficienti, in modo che $A + \delta A$ sia ancora invertibile (si può dimostrare che per essere sicuri che $A + \delta A$ sia invertibile è sufficiente che $r = \|A^{-1}\| \cdot \|\delta A\| < 1$). Cerchiamo $\mathbf{x} + \delta\mathbf{x} \in \mathbb{R}^n$ soluzione del sistema lineare perturbato

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

Sviluppando si ha

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

$$\delta A\ \mathbf{x} + A\ \delta\mathbf{x} + \delta A\ \delta\mathbf{x} = \mathbf{0}$$

$$A\ \delta\mathbf{x} = -\delta A\ \mathbf{x} - \delta A\ \delta\mathbf{x}$$

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

Passando alle norme

$$\begin{aligned}\|\delta\mathbf{x}\| &= \|A^{-1} \cdot \delta A(\mathbf{x} + \delta\mathbf{x})\| \\ &\leq \|A^{-1}\| \cdot \|\delta A\| \cdot (\|\mathbf{x}\| + \|\delta\mathbf{x}\|)\end{aligned}$$

da cui

$$\|\delta\mathbf{x}\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|\mathbf{x}\| + \|A^{-1}\| \cdot \|\delta A\| \cdot \|\delta\mathbf{x}\|$$

$$(1 - \|A^{-1}\| \cdot \|\delta A\|)\|\delta\mathbf{x}\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|\mathbf{x}\|$$

e ricordando che $\|A^{-1}\| \cdot \|\delta A\| < 1$ e $\|A^{-1}\| \cdot \|\delta A\| = K\|\delta A\|/\|A\|$, si ottiene

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{K}{1-r}\frac{\|\delta A\|}{\|A\|};$$

dunque ancora il condizionamento di $A\ \mathbf{x} = \mathbf{b}$ dipende dalla costante $K = \|A\| \cdot \|A^{-1}\|$ (ora rispetto alla perturbazione sulla matrice dei coefficienti).

Teorema 1.5 *Sia A non sigolare e sia $r = \|A^{-1}\| \cdot \|\delta A\| < 1$; allora la matrice $A + \delta A$ è non singolare, e*

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

La soluzione del sistema perturbato

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

soddisfa

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

1.3 Fattorizzazione QR

Data una matrice A di dimensione $n \times n$, esiste sempre una fattorizzazione Q R con Q matrice ortogonale ed R matrice triangolare superiore:

$$A_{n\times n} = Q_{n\times n} \cdot R_{n\times n}$$

con

$Q^T\ Q\ =\ I$
od anche
$Q^T\ =\ Q^{-1}$

$$R = \begin{pmatrix} r_{1,1} & \cdots & & r_{1,n} \\ & r_{2,2} & & \\ & & r_{3,3} & \\ & & & \ddots \quad \vdots \\ 0 & & & r_{n,n} \end{pmatrix}.$$

Osservazione 1.3 *Sia $Q_{n\times n}$ ortogonale ed* $\mathbf{a}$ *un vettore $n \times 1$, allora $\|Q\mathbf{a}\|_2 = \|\mathbf{a}\|_2$, infatti*

$$\|Q\mathbf{a}\|_2^2 = (Q\mathbf{a})^T(Q\mathbf{a}) = (\mathbf{a}^T Q^T)(Q\mathbf{a}) =$$
$$= \mathbf{a}^T(Q^T Q)\mathbf{a} = \mathbf{a}^T I\mathbf{a} = \mathbf{a}^T\mathbf{a} = \|\mathbf{a}\|_2^2.$$

L'esistenza di una tale fattorizzazione rende facile la determinazione della soluzione del sistema normale non singolare $A\mathbf{x} = \mathbf{b}$, infatti si avrebbe:

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

e ponendo $R\mathbf{x} = \mathbf{y}$ si potrebbe risolvere il sistema

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

sfruttando la definizione di Q ortogonale cioè $\mathbf{y} = Q^T\mathbf{b}$; determinato $\mathbf{y}$ si risolve il sistema

$$R\mathbf{x} = \mathbf{y}$$

per sostituzione all'indietro (backward-substitution).

1.3.1 Matrici elementari di Householder

Per matrice elementare di Householder si intende una matrice H tale che

$$H\ \mathbf{a} = \pm\|\mathbf{a}\|_2\ \mathbf{e}_1 = \begin{pmatrix} \pm\|\mathbf{a}\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Limitandoci al caso reale, cioè a vettori ad elementi reali si può dire che le matrici elementari di Householder sono simmetriche ($H = H^T$) ed ortogonali ($H^T H = I$ cioè $H^{-1} = H^T$).

Assegnato $\mathbf{a}$, l'opportuna trasformazione di Householder si può costruire così; si definisce vettore di Householder $\mathbf{v}$ come

$$\mathbf{v} = \begin{pmatrix} a_1 \pm \|\mathbf{a}\|_2 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

quindi

$$H = I - \beta\mathbf{v}\mathbf{v}^T \quad \text{con} \quad \beta = \frac{2}{\|\mathbf{v}\|_2^2}.$$

Si può dimostrare che H è ortogonale [1], mentre la simmetria è ovvia dalla forma della H. Inoltre se applicata al vettore $\mathbf{a}$ si ha:

$$\begin{aligned} H\ \mathbf{a} &= (I - \beta\mathbf{v}\mathbf{v}^T)\mathbf{a} \\ &= \mathbf{a} - \beta\mathbf{v}\mathbf{v}^T\mathbf{a} \\ &= \mathbf{a} - \mathbf{v} \qquad \text{infatti} \qquad \beta\mathbf{v}^T\mathbf{a} = 1 \\ &= \mp\|\mathbf{a}\|_2\mathbf{e}_1. \end{aligned}$$

Verifichiamo $\beta\mathbf{v}^T\mathbf{a} = 1$:

$$\begin{aligned} \mathbf{v}^T\mathbf{a} &= (a_1 \pm \|\mathbf{a}\|_2, a_2, \ldots, a_n) \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \\ &= a_1^2 \pm a_1\|\mathbf{a}\|_2 + a_2^2 + \ldots + a_n^2 \\ &= \|\mathbf{a}\|_2^2 \pm a_1\|\mathbf{a}\|_2 \\ &= \|\mathbf{a}\|_2(\|\mathbf{a}\|_2 \pm a_1). \end{aligned}$$

[1] $H^T H = (I - \beta\mathbf{v}\mathbf{v}^T)(I - \beta\mathbf{v}\mathbf{v}^T) = I - 2\beta\mathbf{v}\mathbf{v}^T + \beta^2\mathbf{v}\mathbf{v}^T\mathbf{v}\mathbf{v}^T = I - 2\beta\mathbf{v}\mathbf{v}^T + \beta\frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\|\mathbf{v}\|_2^2\mathbf{v}^T = I$ *c.v.d.*

inoltre $\beta = \dfrac{2}{\|\mathbf{v}\|_2^2}$ e

$$\begin{aligned}\|\mathbf{v}\|_2^2 &= (a_1 \pm \|\mathbf{a}\|_2)^2 + a_2^2 + \ldots + a_n^2) \\ &= a_1^2 \pm 2a_1\|\mathbf{a}\|_2 + \|\mathbf{a}\|_2^2 + a_2^2 + \ldots + a_n^2 \\ &= 2\|\mathbf{a}\|_2^2 \pm 2a_1\|\mathbf{a}\|_2 \\ &= 2\|\mathbf{a}\|_2(\|\mathbf{a}\|_2 \pm a_1).\end{aligned}$$

e quindi

$$\beta \mathbf{v}^T \mathbf{a} = 1.$$

Osservazione 1.4 *Dato il vettore* $\mathbf{a} = (1,1,1)^T$*, si determini la matrice di Householder che azzera la seconda e terza componente.*

$$\mathbf{v} = \begin{pmatrix} a_1 - \|\mathbf{a}\|_2 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1-\sqrt{3} \\ 1 \\ 1 \end{pmatrix} \qquad \beta = \frac{1}{\|\mathbf{a}\|_2(\|\mathbf{a}\|_2 - a_1)} = \frac{1}{\sqrt{3}(\sqrt{3}-1)}.$$

$$\begin{aligned} H = I - \beta\mathbf{v}\mathbf{v}^T &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{\sqrt{3}(\sqrt{3}-1)} \begin{pmatrix} 1-\sqrt{3} \\ 1 \\ 1 \end{pmatrix} (1-\sqrt{3}, 1, 1) \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{\sqrt{3}(\sqrt{3}-1)} \begin{pmatrix} (1-\sqrt{3})^2 & (1-\sqrt{3}) & (1-\sqrt{3}) \\ (1-\sqrt{3}) & 1 & 1 \\ (1-\sqrt{3}) & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{\sqrt{3}-1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}(\sqrt{3}-1)} & \frac{1}{\sqrt{3}(\sqrt{3}-1)} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}(\sqrt{3}-1)} & \frac{1}{\sqrt{3}(\sqrt{3}-1)} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1-\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} & \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{2}+\frac{\sqrt{3}}{6} \\ -\frac{\sqrt{3}}{3} & \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{2}+\frac{\sqrt{3}}{6} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} & \frac{1}{2}-\frac{\sqrt{3}}{6} & -\frac{1}{2}-\frac{\sqrt{3}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{2}-\frac{\sqrt{3}}{6} \end{pmatrix} \end{aligned}$$

Ovviamente $H\mathbf{a} = (\sqrt{3}, 0, 0)^T$.

1.3.2 Metodo di Householder

Procediamo ad illustrare come funziona l'algoritmo per fattorizzare $Q\ R$ una matrice $A_{n\times n} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ (il procedimento in oggetto può essere applicato anche a matrici rettangolari $A_{m\times n}$ con $m \geq n$ e $rango(A) = n$).

Si pone $A_1 := A$ e si costruisce una matrice elementare di Householder H_1 tale che

$$H_1\ A_1 = \begin{pmatrix} \|\mathbf{a}_1^{(1)}\| & \times & \ldots & \times \\ 0 & \vdots & & \vdots \\ \vdots & & & \\ 0 & \times & \ldots & \times \end{pmatrix} = A_2$$

quindi una H_2 tale che

$$H_2\ A_2 = \begin{pmatrix} \|\mathbf{a}_1^{(1)}\| & \times & \times & \ldots & \times \\ 0 & \|\mathbf{a}_2^{(2)}\| & \times & \ldots & \times \\ \vdots & 0 & \vdots & & \\ & \vdots & & & \\ 0 & 0 & \times & \ldots & \times \end{pmatrix} = A_3$$

e così via fino ad una H_{n-1} tale che

$$H_{n-1}\ A_{n-1} = \begin{pmatrix} \|\mathbf{a}_1^{(1)}\| & \times & \times & \ldots & \times \\ 0 & \|\mathbf{a}_2^{(2)}\| & \times & \ldots & \times \\ \vdots & 0 & \ddots & & \vdots \\ & \vdots & & \ddots & \times \\ 0 & 0 & \ldots & & \|\mathbf{a}_n^{(n-1)}\| \end{pmatrix} = R$$

così che

$$H_{n-1}\ H_{n-2}\ \ldots H_1\ A = R.$$

Ma le H_k sono tutte ortogonali ed il prodotto di matrici ortogonali è ancora una matrice ortogonale, da cui

$$A = H_1^T\ H_2^T\ \ldots H_{n-1}^T\ R = Q\ R.$$

Vediamo alcuni dettagli sulla costruzione delle matrici H_k:
al primo passo, sia $\mathbf{a}_1^{(1)}$ il vettore formato dagli elementi della prima colonna di $A_1 = A$ e sia

$$\theta_1 = \begin{cases} +1 & \text{se} \quad a_{1,1}^{(1)} \geq 0 \\ -1 & \text{se} \quad a_{1,1}^{(1)} < 0 \end{cases}$$

posto

$$\beta_1 = \frac{1}{\|\mathbf{a}_1^{(1)}\|_2(\|\mathbf{a}_1^{(1)}\|_2 + |a_{1,1}^{(1)}|)}$$

e

$$\mathbf{v}_1 = \begin{pmatrix} \theta_1(\|\mathbf{a}_1^{(1)}\|_2 + |a_{1,1}^{(1)}|) \\ a_{2,1}^{(1)} \\ a_{3,1}^{(1)} \\ \vdots \\ a_{n,1}^{(1)} \end{pmatrix}$$

allora la prima matrice elementare di Householder è data da

$$H_1 = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T;$$

al $k-$esimo passo, sia $\mathbf{a}_k^{(k)}$ il vettore di ordine $n-k+1$ formato dagli elementi della $k-$esima colonna di A_k con indice di riga maggiore e uguale a k, e sia

$$\theta_k = \begin{cases} +1 & \text{se} \quad a_{k,k}^{(k)} \geq 0 \\ -1 & \text{se} \quad a_{k,k}^{(k)} < 0 \end{cases}$$

posto

$$\beta_k = \frac{1}{\|\mathbf{a}_k^{(k)}\|_2(\|\mathbf{a}_k^{(k)}\|_2 + |a_{k,k}^{(k)}|)}$$

e

$$\mathbf{v}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \theta_k(\|\mathbf{a}_k^{(k)}\|_2 + |a_{k,k}^{(k)}|) \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{n,k}^{(k)} \end{pmatrix}$$

la $k-$esima matrice elementare di Householder è data da

$$H_k = I - \beta_k \mathbf{v}_k \mathbf{v}_k^T.$$

Osservazione 1.5 *Se al $k-$esimo passo si ha $\mathbf{a}_k^{(k)} = \mathbf{0}$, si pone $H_k = I$, cioè il $k-$esimo passo non comporta alcuna operazione.*

Esempio 1.4 *Si calcoli la fattorizzazione Q R della matrice*

$$A_1 = A = \begin{pmatrix} 72 & -144 & -144 \\ -144 & -36 & -360 \\ -144 & -360 & 450 \end{pmatrix}.$$

Al primo passo si ha

$$\beta_1 = \frac{1}{62208}, \qquad \mathbf{v}_1 = \begin{pmatrix} 288 \\ -144 \\ -144 \end{pmatrix}$$

per cui

$$H_1 = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T = \frac{1}{6} = \begin{pmatrix} -2 & 4 & 4 \\ 4 & 4 & -2 \\ 4 & -2 & 4 \end{pmatrix}$$

e

$$A_2 = \begin{pmatrix} -216 & -216 & 108 \\ 0 & 0 & 486 \\ 0 & -324 & 324 \end{pmatrix}.$$

Al secondo passo si ha

$$\beta_2 = \frac{1}{104976}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 324 \\ -324 \end{pmatrix}$$

per cui

$$H_2 = I - \beta_2 \mathbf{v}_2 \mathbf{v}_2^T == \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

e

$$R = A_3 = \begin{pmatrix} -216 & -216 & 108 \\ 0 & -324 & 324 \\ 0 & 0 & -486 \end{pmatrix}.$$

Inoltre

$$Q = H_1 \; H_2 = \frac{1}{6} \begin{pmatrix} -2 & 4 & 4 \\ 4 & -2 & 4 \\ 4 & 4 & -2 \end{pmatrix}.$$

1.3.3 Implementazione del metodo di Householder

Il metodo di Householder, per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$, può essere implementato senza calcolare effettivamente le matrici H_k. Si procede nel seguente modo: si considera la matrice

$$T_1 = [A_1|\mathbf{b}_1] = [A|\mathbf{b}]$$

e si costruiscono β_1, $\mathbf{v}_1$ ed il vettore di $n+1$ componenti

$$\mathbf{y}_1^T = \mathbf{v}_1^T T_1 = (\theta_1(\|\mathbf{a}_1^{(1)}\|_2 + |a_{1,1}^{(1)}|), a_{2,1}^{(1)}, \ldots, a_{n,1}^{(1)}) \left(\begin{array}{cccc|c} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \ldots & a_{1,n}^{(1)} & b_1^{(1)} \\ a_{2,1}^{(1)} & & \ldots & & \vdots \\ \vdots & & \ldots & & \vdots \\ a_{n,1}^{(1)} & \ldots & & a_{n,n}^{(1)} & b_n^{(1)} \end{array}\right).$$

Allora è

$$\begin{aligned} T_2 = &\ H_1\ T_1 \\ = &\ (I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T)\ T_1 \\ = &\ T_1 - \beta_1 \mathbf{v}_1 \mathbf{y}_1^T \end{aligned}$$

Al $k-$esimo passo si costruisce

$$\mathbf{y}_k^T = \mathbf{v}_k^T T_k$$

e

$$T_{k+1} = T_k - \beta_k \mathbf{v}_k \mathbf{y}_k^T.$$

Dopo $n-1$ passi si ottiene la matrice

$$T_n = [A_n|\mathbf{b}_n] = [R|\mathbf{b}_n]$$

e quindi il sistema $R\mathbf{x} = \mathbf{b}_n$ con matrice dei coefficienti triangolare superiore equivalente al sistema $A\mathbf{x} = \mathbf{b}$.

1.3.4 Costo Computazionale per risolvere $A\mathbf{x} = \mathbf{b}$ tramite $Q\ R$

Si esamina il costo del metodo di soluzione proposto che non comporta la determinazione esplicita delle metrici H_k e della matrice ortogonale Q. Al $k-$esimo passo, per determinare $\mathbf{v}_k$ e β_k, assunte le prime $k-1$ componenti

di $\mathbf{v}_k$ nulle e quindi con al più $n-k+1$ componenti non nulle, sarà:

$$\mathbf{v}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \theta_k(\|\mathbf{a}_k^{(k)}\|_2 + |a_{k,k}^{(k)}|) \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{n,k}^{(k)} \end{pmatrix} \qquad \beta_k = \frac{2}{\|\mathbf{v}_k\|_2^2}$$

e servono: $n-k+1$ moltiplicazioni ed una radice quadrata per calcolare $\|\mathbf{a}_k^{(k)}\|_2$, una moltiplicazione per $\|\mathbf{v}_k\|_2^2$ ed una divisione per β_k per un totale di $n-k+3$ molt./div.
Per $\mathbf{y}_k$ servono $(n-k+1)^2+n-k+1$ moltiplicazioni, mentre per $\mathbf{v}_k\mathbf{y}_k^T$ ne servono $(n-k+1)^2$. In totale

$$n-k+3-2(n-k+1)^2+(n-k+1) =$$

$$2(n-k)^2+6(n-k)+6.$$

Poiché i passi sono $k=1,\ldots,(n-1)$ si ha

$$\frac{2(n-1)n(2n-1)}{6}+6\frac{(n-1)n}{2}+4(n-1) =$$

$$\frac{2}{3}(n-1)n(n+4)+4(n-1) \approx \frac{2}{3}n^3.$$

1.3.5 Stabilità Numerica della Fattorizzazione $Q\ R$

Procedendo con l'analisi all'indietro, esattamente come già fatto nel caso della fattorizzazione $L\ U$ si ha che

$$\delta A \sim R \cdot \delta Q + Q \cdot \delta R$$

da cui si cerca una limitazione superiore per gli elementi delle matrici Q ed R. Si trova che, essendo Q unitaria, vale

$$\max_{i,j} |q_{i,j}| \leq \|Q\|_2 = 1;$$

inoltre si ha che

$$\max_{i,j} |r_{i,j}| \leq \sqrt{n}\ \max_{i,j} |a_{i,j}|.$$

Da cui risulta che l'algoritmo di fattorizzazione con matrici elementari di Householder ($Q\ R$) è stabile in senso debole con estremo $\sqrt{n}$, che risulta comunque più stabile dell'algoritmo $L\ U$ essendo $\sqrt{n} \ll 2^{n-1}$.

Capitolo 2

Autovalori e Autovettori

Definizione 2.1 *data una matrice* $A \in \mathbb{R}^{m\times n}$*, viene definito autovalore di* A *un numero* $\lambda \in \mathbb{C}$ *per cui valga la relazione*

$$A\,\mathbf{x} = \lambda\mathbf{x} \qquad con \quad \mathbf{x} \neq \mathbf{0}. \tag{2.1}$$

$\mathbf{x}$ *è detto autovettore corrispondente a* λ*.*

L'insieme degli autovalori di una matrice A costituisce lo **spettro** di A e l'autovalore massimo in modulo è detto **raggio spettrale** di A ed è indicato con $\rho(A)$.

La relazione (2.1) data può essere vista come un sistema lineare omogeneo

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

che ammette soluzioni non nulle se e solo se

$$det(A - \lambda I) = 0. \tag{2.2}$$

Sviluppando la (2.2) risulta

$$det(A - \lambda I) = p(\lambda) = a_0 + a_1\lambda + \ldots + a_n\lambda^n$$

in cui

$$a_n = (-1)^n \qquad a_0 = det(A) \qquad a_{n-1} = (-1)^{n-1}tr(A)$$

dove con $tr(A)$ si indica la somma degli elementi diagonali di A detta traccia di A.
Dalle relazioni che legano i coefficienti e le radici di una equazione algebrica risulta che

$$\sum_{i=1}^{n} \lambda_i = tr(A) \qquad \text{e} \qquad \prod_{i=1}^{n} \lambda_i = det(A).$$

Il polinomio $p(\lambda)$ è detto **polinomio caratteristico** di A e l'equazione $p(\lambda) = 0$ è detta **equazione caratteristica** di A.
Per il teorema fondamentale dell'algebra l'equazione caratteristica ha, in campo complesso, n radici. Quindi una matrice di ordine n ha n autovalori nel campo complesso.
Poiché gli autovalori sono soluzioni non nulle del sistema lineare omogeneo visto, un autovettore corrispondente ad un autovalore λ risulta determinato a meno di una costante moltiplicativa $\alpha \neq 0$, cioè se $\mathbf{x}$ è un autovettore di A, anche $\alpha\mathbf{x}$ è un autovettore di A, corrispondente allo stesso autovalore.

Esempio 2.1 *Il polinomio caratteristico della matrice*

$$A = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$$

si ricava dal determinante

$$det(A - \lambda I) = det \begin{pmatrix} 1-\lambda & 3 \\ 3 & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - 9 = \lambda^2 - 2\lambda - 8.$$

L'equazione caratteristica corrispondente è:

$$\lambda^2 - 2\lambda - 8 = 0$$

ed ha come radici $\lambda_1 = -2$ *e* $\lambda_2 = 4$ *che sono gli autovalori della matrice A. L'autovettore corrispondente a* $\lambda_1 = -2$ *si calcola risolvendo il sistema*

$$\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{0};$$

dalla prima equazione si ottiene

$$x_1 + x_2 = 0$$

da cui $x_1 = -x_2$ *e qualunque vettore*

$$\mathbf{x} = \alpha \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

con $\alpha \neq 0$*, è autovettore corrispondente all'autovalore* $\lambda_1 = -2$*.*
L'autovettore corrispondente a $\lambda_2 = 4$ *si determina risolvendo il sistema*

$$\begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{0};$$

dalla prima equazione si ottiene

$$-x_1 + x_2 = 0$$

da cui $x_1 = x_2$ *e qualunque vettore*

$$\mathbf{x} = \alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

con $\alpha \neq 0$*, è autovettore corrispondente all'autovalore* $\lambda_2 = 4$*.*

Osservazione 2.1 *Dato un polinomio si può costruire una matrice i cui autovalori sono gli zeri del polinomio; tale matrice si dice* **companion** *o matrice di* **Frobenius***. Dato*

$$p(\lambda) = \sum_{i=0}^{n} a_i \lambda^i$$

la matrice in questione è così definita:

$$F = \begin{pmatrix} 0 & \dots & & & 0 & -\frac{a_0}{a_n} \\ 1 & \ddots & & & \vdots & -\frac{a_1}{a_n} \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & & \ddots & 0 & \\ 0 & \dots & & 0 & 1 & -\frac{a_{n-1}}{a_n} \end{pmatrix}$$

Esempio 2.2 *Sia* $p(\lambda) = \lambda^2 - 2\lambda - 8$*, allora*

$$F = \begin{pmatrix} 0 & 8 \\ 1 & 2 \end{pmatrix}$$

verifichiamolo:

$$det(F - \lambda I) = det \begin{pmatrix} -\lambda & -8 \\ -1 & 2 - \lambda \end{pmatrix} = \lambda(\lambda - 2) - 8 = \lambda^2 - 2\lambda - 8.$$

2.1 Proprietà degli Autovalori

- Gli autovalori di una matrice diagonale o triangolare sono uguali agli elementi diagonali.

- Se λ è un autovalore di una matrice A non singolare e $\mathbf{x}$ un autovettore corrispondente, allora risulta $\lambda \neq 0$ e $1/\lambda$ è un autovalore di A^{-1} con $\mathbf{x}$ autovettore corrispondente. Infatti da $A\mathbf{x} = \lambda\mathbf{x}$ si ha $\mathbf{x} = \lambda A^{-1}\mathbf{x}$ e quindi $\lambda \neq 0$ e $A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}$.

- Se λ è autovalore di una matrice ortogonale A, cioè $A^T A = AA^T = I$, allora risulta $|\lambda| = 1$. Infatti dalla relazione $A\mathbf{x} = \lambda\mathbf{x}$ si ha $(A\mathbf{x})^T = (\lambda\mathbf{x})^T$ e quindi $\mathbf{x}^T A^T = \lambda\mathbf{x}^T$ da cui si ha

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

  e poiché A è ortogonale risulta

  $$\mathbf{x}^T\mathbf{x} = \lambda^2\mathbf{x}^T\mathbf{x} \qquad \text{e} \quad |\lambda| = 1.$$

2.2 Proprietà degli Autovettori

Teorema 2.1 *Autovettori corrispondenti ad autovalori distinti sono linearmente indipendenti.*

Osservazione 2.2 *Dal precedente teorema risulta che se una matrice A di ordine n ha n autovalori tutti distinti, allora A ha n autovettori linearmente indipendenti. Segue che se A non ha autovalori distinti, A può avere, ma anche non avere, n autovettori linearmente indipendenti.*

Definizione 2.2 *La molteplicità di un autovalore λ come radice dell'equazione caratteristica, è indicata con $\sigma(\lambda)$, ed è detta* **molteplicità algebrica** *di λ. Il massimo numero di autovettori linearmente indipendenti corrispondenti a λ è indicato con $\tau(\lambda)$ ed è detta* **molteplicità geometrica** *di λ.*

Osservazione 2.3 *È evidente che*

$$1 \leq \sigma(\lambda) \leq n \qquad e \qquad 1 \leq \tau(\lambda) \leq n.$$

Teorema 2.2 *Vale la seguente disuguaglianza*

$$\tau(\lambda) \leq \sigma(\lambda).$$

2.3 Similitudine fra Matrici

Definizione 2.3 *Due matrici A e $B \in \mathbb{R}^{n\times n}$ si dicono* **simili** *se esiste una matrice non singolare S per cui*

$$A = S\ B\ S^{-1}.$$

Osservazione 2.4 *La trasformazione che associa la matrice A alla matrice B viene detta trasformazione per similitudine; se la matrice S è ortogonale, la trasformazione viene detta trasformazione per similitudine ortogonale.*

Teorema 2.3 *Due matrici simili hanno gli stessi autovalori con le stesse molteplicità algebriche e geometriche.*

dim. Siano A e B simili, cioè tali che

$$A = S\ B\ S^{-1}.$$

Si ha che

$$\begin{aligned} det(A-\lambda I) &= det(S\ B\ S^{-1} - \lambda S\ S^{-1}) \\ &= det(S(B-\lambda I)S^{-1}) \\ &= det(S)\ det(B-\lambda I)\ det(S^{-1}) \\ &= det(B-\lambda I) \end{aligned}$$

per cui le due matrici hanno lo stesso polinomio caratteristico e quindi hanno gli stessi autovalori con le stesse molteplicità algebriche.
Se $\mathbf{x}$ è autovettore di A corrispondente all'autovalore λ, risulta

$$S\ B\ S^{-1}\mathbf{x} = \lambda\mathbf{x}$$

e quindi

$$B\ S^{-1}\mathbf{x} = \lambda S^{-1}\mathbf{x}$$

perciò il vettore $\mathbf{y} = S^{-1}\mathbf{x}$ è autovettore di B corrispondente a λ.
Inoltre, essendo S^{-1} non singolare, se $\mathbf{x}_i, i = 1,\ldots,\tau(\lambda)$ sono autovettori linearmente indipendenti di A, anche $\mathbf{y}_i = S^{-1}\mathbf{x}_i, i = 1,\ldots,\tau(\lambda)$ sono linearmente indipendenti.

Definizione 2.4 *Una matrice A simile ad una matrice diagonale D si dice* **diagonalizzabile.**

Teorema 2.4 *Una matrice A di ordine n è diagonalizzabile se e solo se ha n autovettori linearmente indipendenti. Inoltre le colonne della matrice S, per cui $S^{-1}\ A\ S$ è diagonale, sono gli autovettori di A.*

dim. Siano $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ n autovettori linearmente indipendenti di A corrispondenti agli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_n$. Siano D la matrice diagonale avente λ_i come $i-$esimo elemento diagonale, ed S la matrice la cui $i-$esima colonna è uguale a $\mathbf{x}_i$. Dalla relazione

$$A\mathbf{x}_i = \lambda_i \mathbf{x}_i \qquad i = 1, \ldots n$$

si ha che

$$A\ S = S\ D.$$

Essendo S non singolare, perché formata da colonne linearmente indipendenti, esiste S^{-1}; quindi si ha

$$A = S\ D\ S^{-1}.$$

Viceversa, sia $A = S\ D\ S^{-1}$, con D matrice diagonale con gli autovalori di A come elementi diagonali. Allora risulta

$$A\ S = S\ D.$$

Indicando con $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ le colonne di S, si ha

$$A[\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n] = [\lambda_1 \mathbf{s}_1, \lambda_2 \mathbf{s}_2, \ldots, \lambda_n \mathbf{s}_n].$$

Perciò le colonne di S sono n autovettori di A che risultano linearmente indipendenti.

Esempio 2.3 *Sia*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Questa ha come autovalori

$$\lambda_1 = 1 - \sqrt{2} \qquad \lambda_2 = 1 \qquad \lambda_3 = 1 + \sqrt{2}$$

e i corrispondenti autovettori sono

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}.$$

Allora A è diagonalizzabile, cioè esiste una matrice S, non singolare, tale che

$$A = S\ D\ S^{-1}.$$

Infatti per quanto visto sarà:

$$S = \begin{pmatrix} 1 & 1 & 0 \\ -\sqrt{2} & 0 & \sqrt{2} \\ 0 & 1 & 1 \end{pmatrix}$$

da cui risulta

$$S^{-1} = \begin{pmatrix} 1/4 & -\sqrt{2}/4 & 1/4 \\ 1/2 & 0 & -1/2 \\ 1/4 & \sqrt{2}/4 & 1/4 \end{pmatrix}$$

e

$$A = S \begin{pmatrix} 1-\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+\sqrt{2} \end{pmatrix} S^{-1}.$$

Osservazione 2.5 *Fra le trasformazioni per similitudine che associano alla matrice B la matrice $A = S\ B\ S^{-1}$, hanno particolare importanza quelle per cui S è ortogonale, cioè $S^T\ S = S\ S^T = I$. Il teorema che segue mostra come sia possibile, mediante una trasformazione per similitudine ortogonale, ricondurre una qualsiasi matrice ad una forma triangolare superiore.*

Teorema 2.5 (Forma Normale di Schur) *Sia $A \in \mathbb{C}^n$ e siano $\lambda_1, \ldots, \lambda_n$ i suoi autovalori. Allora esiste una matrice ortogonale/unitaria Q e una matrice triangolare superiore T i cui elementi diagonali sono i λ_i, tali che*

$$A = Q\ T\ Q^T.$$

se la matrice A ha elementi reali esiste la forma normale reale di Schur.

Teorema 2.6 (Forma Normale Reale di Schur) *se $A \in \mathbb{R}^{n\times n}$, esiste una matrice ortogonale $Q \in \mathbb{R}^{n\times n}$ e una matrice $T \in \mathbb{R}^{n\times n}$ triangolare superiore a blocchi della forma*

$$T = \begin{pmatrix} R_{1,1} & R_{1,2} & \ldots & R_{1,m} \\ & R_{2,2} & & \vdots \\ & & \ddots & \\ 0 & & & R_{m,m} \end{pmatrix}.$$

dove i blocchi $R_{j,j}$ per $j = 1, \ldots, m$ hanno ordine 1 o 2. Se λ_j è autovalore reale di A, allora $R_{j,j}$ ha ordine 1 e coincide con λ_j, se invece è complesso, allora il blocco $R_{j,j}$ ha ordine 2 ed ha come autovalori λ_j e $\bar{\lambda}_j$ (il complesso coniugato). La somma delle dimensioni dei blocchi $R_{j,j}$ $j = 1, \ldots, m$ è pari ad n.

Osservazione 2.6 *Una classe particolarmente importante di matrici è quella delle matrici normali, cioè tali che*

$$A\ A^H = A^H\ A$$

dove H sta per Hermitiana ed equivale ad una matrice simmetrica in campo complesso. Questa classe è importante perché comprende tutte e sole le matrici diagonalizzabili con trasformazioni per similitudine unitarie. Vale infatti il seguente teorema.

Teorema 2.7 *Una matrice $A \in \mathbb{C}^{n\times n}$ è normale (cioè $A\ A^H = A^H\ A$) se e solo se esiste una matrice unitaria Q tale che*

$$A = Q \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} Q^T$$

in cui $\lambda_1, \lambda_2, \ldots, \lambda_n$ sono gli autovalori di A. La matrice Q ha per colonne gli autovettori di A, che quindi sono a due a due ortonormali.

Capitolo 3

Sistemi lineari: metodi iterativi

Per risolvere un sistema lineare $A\mathbf{x} = \mathbf{b}$, oltre ai metodi diretti, si possono utilizzare anche i metodi iterativi, che risultano particolarmente convenienti se la matrice è sparsa, cioè se il numero degli elementi non nulli di A è dell'ordine della dimensione della matrice. Infatti quando si usa un metodo diretto può accadere che nelle matrici intermedie vengano generati molti elementi diversi da zero in corrispondenza di elementi nulli della matrice iniziale. Poiché i metodi diretti non sfruttano la sparsità della matrice, soprattutto poi se A è anche di grandi dimensioni, può essere conveniente utilizzare un metodo iterativo. Esistono comunque dei casi nei quali la matrice A è sparsa, ma è più conveniente usare dei metodi diretti che sfruttano specifiche proprietà di struttura della matrice. Cominciamo con alcuni richiami su successioni convergenti di vettori e matrici.

Definizione 3.1 *Una successione* $\{\mathbf{x}^{(k)}\}$ *di vettori di* $\mathbb{R}^n$ *si dice convergente al vettore* $\mathbf{x}^* \in \mathbb{R}^n$ *se esiste una norma vettoriale per cui*

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

in tal caso si pone

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*.$$

Per il teorema di equivalenza delle norme, la definizione appena data non dipende da una particolare norma. La condizione di convergenza data si traduce in una condizione di convergenza delle successioni formate dalle singole componenti. Infatti, considerando la norma ∞ o del max

$$|x_i^{(k)} - x_i^*| \leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|_\infty \qquad i = 1, \ldots, n$$

e quindi

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

da cui

$$\lim_{k\to\infty} x_i^{(k)} = x_i^*.$$

Viceversa se vale quest'ultima è ovviamente verificata la condizione di convergenza data nella definizione, per la norma ∞.

Per le successioni di matrici $\{A^{(k)}\}$ si può dare una definizione di convergenza analoga a quella data per vettori. Il seguente teorema è di fondamentale importanza nello studio dei metodi iterativi per la soluzione di sistemi lineari.

Teorema 3.1 *Sia $A \in \mathbb{R}^{n\times n}$, allora*

$$\lim_{k\to\infty} A^k = 0 \quad \textit{se e solo se} \quad \rho(A) < 1$$

con $\rho(A)$ il raggio spettrale della matrice A.

Teorema 3.2 *Sia $A \in \mathbb{R}^{n\times n}$, allora*

$$det(I - A) \neq 0$$

e

$$\lim_{k\to\infty} \sum_{i=0}^{k} A^i = (I - A)^{-1} \quad \textit{se e solo se} \quad \rho(P) < 1.$$

Osservazione 3.1 *Come per le serie numeriche, si usa scrivere*

$$\sum_{i=0}^{\infty} A^i = (I - A)^{-1}.$$

3.1 Decomposizione della matrice

Sia $A \in \mathbb{R}^{n\times n}$ una matrice non singolare e si consideri la decomposizione di A nella forma

$$A = M \; - \; N$$

dove M è una matrice non singolare. Sostituendo tale decomposizione nel sistema $A\mathbf{x} = \mathbf{b}$ si ha

$$(M \; - \; N)\mathbf{x} = \mathbf{b}$$

$$M\mathbf{x} \; - \; N\mathbf{x} = \mathbf{b}$$

ed essendo M non singolare

$$\mathbf{x} = M^{-1}\ N\mathbf{x}\ +\ M^{-1}\mathbf{b}.$$

Posto $P := M^{-1}\ N$ e $\mathbf{q} := M^{-1}\mathbf{b}$ si ottiene il seguente sistema

$$\mathbf{x} = P\mathbf{x} + \mathbf{q}$$

equivalente ad $A\mathbf{x} = \mathbf{b}$.
Dato un vettore iniziale $\mathbf{x}^{(0)}$, si considera la successione $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ così definita

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q} \quad k = 1, 2, \ldots \tag{3.1}$$

Se la successione $\{\mathbf{x}^{(k)}\}$ è convergente, cioè

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$$

allora passando al limite nella (3.1) risulta

$$\mathbf{x}^* = P\mathbf{x}^* + \mathbf{q} \tag{3.2}$$

cioè $\mathbf{x}^*$ è la soluzione del sistema $\mathbf{x} = P\mathbf{x}+\mathbf{q}$ e quindi del sistema $A\mathbf{x} = \mathbf{b}$.
La relazione (3.1) individua un metodo iterativo in cui, partendo da un vettore iniziale $\mathbf{x}^{(0)}$, la soluzione viene approssimata utilizzando una successione $\mathbf{x}^{(k)}$ di vettori. La matrice P si dice **matrice di iterazione** del metodo.

Osservazione 3.2 *Al variare del vettore iniziale $\mathbf{x}^{(0)}$ si ottengono dalla (3.1) diverse successioni $\{\mathbf{x}^{(k)}\}$, alcune delle quali possono essere convergenti ed altre no. Un metodo iterativo è detto convergente se, qualunque sia il vettore iniziale $\mathbf{x}^{(0)}$, la successione $\{\mathbf{x}^{(k)}\}$ è convergente.*

Teorema 3.3 *Il metodo iterativo (3.1) risulta convergente se e solo se $\rho(A) < 1$.*

dim. Sottraendo dalla (3.2) la (3.1) si ha

$$\mathbf{x}^* - \mathbf{x}^{(k)} = P(\mathbf{x}^* - \mathbf{x}^{(k-1)}) \quad k = 1, 2, \ldots$$

Indicato con

$$\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$$

il vettore **errore** alla $k-$esima iterazione, si ha

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)} \quad k = 1, 2, \ldots$$

e quindi

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)} = P^2\mathbf{e}^{(k-2)} = \ldots = P^k\mathbf{e}^{(0)}.$$

(Condizione sufficiente) Se $\rho(P) < 1$, per il teorema visto, risulta

$$\lim_{k\to\infty} P^k = 0$$

e segue che per ogni vettore $\mathbf{e}^{(0)}$ si ha

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}.$$

(Condizione necessaria) Se il metodo è convergente il $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}$ vale per ogni $\mathbf{x}^{(0)}$, e in particolare deve valere se $\mathbf{x}^{(0)}$ è tale che il vettote $\mathbf{e}^{(0)} = \mathbf{x}^* - \mathbf{x}^{(0)}$ è un autovettore di P corrispondente ad un autovalore λ di modulo massimo, cioè $|\lambda| = \rho(P)$.
In questo caso risulta

$$P\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$$

e quindi

$$\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)} = \lambda^k\mathbf{e}^{(0)}.$$

Segue che

$$\lim_{k\to\infty} [\rho(P)]^k = 0$$

e quindi $\rho(P) =< 1$.

La condizione $\rho(P) < 1$, necessaria e sufficiente per la convergenza della (3.1), non è in generale di agevole verifica. Conviene utilizzare, quando è possibile, delle condizioni sufficienti di convergenza di più facile verifica. Una tale condizione è data dal seguente teorema.

Teorema 3.4 *Se esiste una norma matriciale indotta $\|\cdot\|$, per cui $\|P\| < 1$, il metodo iterativo (3.1) è convergente.*

dim. La tesi segue dal teorema di convergenza visto e dalla proprietà che

$$\rho(P) < \|P\|.$$

Osservazione 3.3 *Il raggio spettrale $\rho(P)$ di una matrice viene a volte anche definito come l'estremo inferiore di tutte le norme di P.*

Osservazione 3.4 *Poiché il determinante di una matrice è uguale al prodotto degli autovalori, se $|det(P)| \geq 1$, almeno uno degli autovalori di P è in modulo magiore o uguale ad 1 e quindi il metodo (3.1) non è convergente.*

Poiché la traccia di una matrice è uguale alla somma degli autovalori, se $|tr(P)| > n$*, almeno uno degli autovalori di* P *è in modulo maggiore o uguale ad* 1 *e il metodo (3.1) non è convergente.*
Quindi le condizioni:

$$|det(P)| < 1 \qquad e \qquad |tr(P)| < n$$

sono condizioni necessarie affinché il metodo (3.1) sia convergente.

3.2 Controllo della Convergenza

Se $\mathbf{e}^{(k-1)} \neq 0$, la quantità

$$\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(k-1)}\|}$$

esprime la **riduzione dell'errore al** $k-$**esimo passo**.
la media geometrica delle riduzioni dell'errore sui primi k passi

$$\sigma_k = \left(\frac{\|\mathbf{e}^{(1)}\|}{\|\mathbf{e}^{(0)}\|} \cdot \frac{\|\mathbf{e}^{(2)}\|}{\|\mathbf{e}^{(1)}\|} \cdots \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(k-1)}\|}\right)^{\frac{1}{k}} = \left(\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|}\right)^{\frac{1}{k}}$$

esprime la **riduzione media per passo** dell'errore relativo ai primi k passi. Dalla

$$\|\mathbf{e}^{(k)}\| \leq \|P^k\| \cdot \|\mathbf{e}^{(0)}\|$$

risulta

$$\sigma_k \leq (\|P^k\|)^{\frac{1}{k}}.$$

La quantità che si ottiene facendo tendere k all'infinito esprime la **riduzione asintotica media per passo** e, come risulta dal seguente teorema, è indipendente dalla particolare norma utilizzata.

Teorema 3.5 *Sia* $A \in \mathbb{R}^{n\times n}$ *e sia* $\|\cdot\|$ *una qualunque norma matriciale indotta. Allora*

$$\lim_{k\to\infty} (\|P^k\|)^{\frac{1}{k}} = \rho(P).$$

Osservazione 3.5 *La quantità* $\rho(P)$*, indipendente dalla norma utilizzata e dall'indice di iterazione* k*, viene quindi assunta come* **misura della velocità di convergenza** *del metodo (3.1).*

3.3 Test di Arresto

Poiché con un metodo iterativo non è possibile calcolare la soluzione con un numero finito di iterazioni, occorre individuare dei criteri per l'arresto del procedimento.
I criteri più comunemente usati, fissata una tolleranza Tol, che tiene conto anche della precisione utilizzata nei calcoli, sono i seguenti

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq Tol$$

oppure, se $\mathbf{x}^{(k)} \neq \mathbf{0}$

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} \leq Tol.$$

Si noti che queste due condizioni non garantiscono che la soluzione sia stata approssimata con la precisione Tol. Infatti da

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)}$$

è

$$\begin{aligned} \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} &= (\mathbf{x}^* - \mathbf{x}^{(k-1)}) - (\mathbf{x}^* - \mathbf{x}^{(k)}) \\ &= \mathbf{e}^{(k-1)} - \mathbf{e}^{(k)} \\ &= (I - P)\mathbf{e}^{(k-1)} \end{aligned}$$

e passando alle norme, se $\|P\| < 1$, per il teorema seguente si ha

$$\begin{aligned} \|\mathbf{e}^{(k-1)}\| &\leq \|(I - P)^{-1}\| \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \\ &\leq \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{1 - \|P\|}. \end{aligned}$$

Per cui può accadere che $\|\mathbf{e}^{(k-1)}\|$ sia elevata anche se le condizioni di arresto date sono verificate.

Teorema 3.6 *Sia $A \in \mathbb{R}^{n\times n}$ e sia $\|\cdot\|$ una qualunque norma matriciale indotta tale che $\|A\| < 1$. Allora la matrice $I \pm A$ è non singolare e vale la disuguaglianza*

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

- In un programma che implementa un tale metodo iterativo deve essere comunque previsto un controllo per interrompere l'esecuzione quando il numero delle iterazioni diventa troppo elevato.

- Può anche accadere che un metodo iterativo la cui matrice di iterazione P è tale che $\rho(P) < 1$, per gli effetti indotti dagli errori di arrotondamento non converga in pratica. Questo accade, in particolare, quando la matrice A è fortemente mal condizionata e $\rho(P)$ è molto vicino ad 1.

- Si deve rilevare che un metodo iterativo, rispetto ad un metodo diretto è in generale meno sensibile alla propagazione degli errori. Infatti il vettore $\mathbf{x}^{(k)}$ può essere considerato come il vettore generato con una sola iterazione a partire dal vettore iniziale $\mathbf{x}^{(k-1)}$, e quindi risulta affetto dagli errori di arrotondamento generati dalla sola ultima iterazione.

- In un metodo iterativo, ad ogni iterazione, il costo computazionale è principalmente determinato dalle operazioni di moltiplicazione della matrice P per un vettore, che richiede n^2 operazioni moltiplicative. Se A è sparsa, il numero di moltiplicazioni è dell'ordine di n. In questo caso i metodi iterativi possono risultare competitivi con quelli diretti.

3.4 Metodi di Jacobi e Gauss-Seidel

Tra i metodi iterativi individuati da una particolare scelta della decomposizione $A = M - N$ sono particolarmente importanti il metodo di Jacobi e il metodo di Gauss-Seidel, per i quali è possibile dare delle condizioni sufficienti di convergenza. Si consideri la decomposizione della matrice A

$$A = D - L - U$$

con

$$D = \begin{pmatrix} a_{1,1} & & & \\ & a_{2,2} & & \\ & & \ddots & \\ & & & a_{n,n} \end{pmatrix},$$

$$L = \begin{pmatrix} 0 & \dots & & 0 \\ -a_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & & \\ -a_{n,1} & \dots & -a_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -a_{1,2} & \dots & -a_{1,n} \\ \vdots & \ddots & \ddots & \\ & & & -a_{n-1,n} \\ 0 & \dots & & 0 \end{pmatrix}.$$

Scegliendo

$$M = D \qquad N = L + U,$$

si ottiene il metodo di Jacobi, mentre scegliendo

$$M = D - L \qquad N = U,$$

si ottiene il metodo di Gauss-Seidel.
Per queste decomposizioni risulta $det(M) \neq 0$ se e solo se tutti gli elementi diagonali di A sono non nulli.
Indicando con J la matrice di iterazione del metodo di Jacobi, dalla $P = M^{-1}N$ si ha

$$J = D^{-1}(L + U)$$

per cui la (3.1) diviene

$$\mathbf{x}^{(k)} = J\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b} = D^{-1}[\mathbf{b} + (L + U)\mathbf{x}^{(k-1)}]$$

e in termini di componenti

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

Il metodo di Jacobi è detto anche metodo degli spostamenti simultanei, in quanto le componenti del vettore $\mathbf{x}^{(k)}$ sostituiscono simultaneamente, al termine dell'iterazione, le componenti di $\mathbf{x}^{(k-1)}$.

Indicando con G la matrice di iterazione di Gauss-Seidel, dalla $P = M^{-1}N$ si ha

$$G = (D - L)^{-1}U$$

per cui la (3.1) diviene

$$\mathbf{x}^{(k)} = G\mathbf{x}^{(k-1)} + (D - L)^{-1}\mathbf{b}.$$

Per descrivere questo metodo in termini di componenti, conviene prima trasformarla nel modo seguente:

$$(D - L)\mathbf{x}^{(k)} = U\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$D\mathbf{x}^{(k)} = L\mathbf{x}^{(k)} + U\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$\mathbf{x}^{(k)} = D^{-1}\left(L\mathbf{x}^{(k)} + U\mathbf{x}^{(k-1)} + \mathbf{b}\right)$$

ottenendo quindi

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

Confrontando questa con quella di Jacobi, risulta che per calcolare le componenti del vettore $\mathbf{x}^{(k)}$ sono utilizzate componenti già calcolate dello stesso vettore. Per questo motivo il metodo di Gauss-Seidel è detto anche metodo degli spostamenti successivi.

Nell'implementare Jacobi è necessario disporre, contemporaneamente di entrambe i vettori $\mathbf{x}^{(k-1)}$ e $\mathbf{x}^{(k)}$, mentre per Gauss-Seidel è sufficiente disporre di un solo vettore.

Osservazione 3.6 *In molte applicazioni il metodo di Gauss-Seidel, che utilizza immediatamente i valori calcolati nella iterazione corrente, risulta più veloce del metodo di Jacobi. Però esistono casi in cui risulta non solo che il metodo di Jacobi sia più veloce di Gauss-Seidel, ma anche che Jacobi sia convergente e Gauss-seidel no.*

Per i metodi di Jacobi e Gauss-Seidel si possono ricavare delle condizioni sufficienti di convergenza di facile verifica sul sistema lineare.

Teorema 3.7 *Se la matrice A è a predominanza diagonale in senso stretto, allora i metodi di Jacobi e di Gauss-Seidel sono convergenti.*

Osservazione 3.7 *Una matrice $A \in \mathbb{R}^{n\times n}$ si dice a predominanza diagonale in senso stretto, se vale*

$$|a_{i,j}| > \sum_{j=1\ j\neq i}^{n} |a_{i,j}| \qquad i = 1,\ldots,n \quad \textit{per righe}$$

$$|a_{i,j}| > \sum_{i=1\ i\neq j}^{n} |a_{i,j}| \qquad j = 1,\ldots,n \quad \textit{per colonne}$$

Esempio 3.1 *Applichiamo il metodo di Jacobi al sistema lineare*

$$A\mathbf{x} = \mathbf{b} \qquad \textit{con} \qquad A = \begin{pmatrix} 20 & 2 & -1 \\ 2 & 13 & -2 \\ 1 & 1 & 1 \end{pmatrix} \qquad \textit{e} \qquad \mathbf{b} = \begin{pmatrix} 25 \\ 30 \\ 2 \end{pmatrix}.$$

La decomposizione di Jacobi è

$$M = \begin{pmatrix} 20 & 0 & 0 \\ 0 & 13 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -2 & 1 \\ -2 & 0 & 2 \\ -1 & -1 & 0 \end{pmatrix}$$

La matrice di iterazione J ed il vettore $\mathbf{q}$ sono

$$J = \begin{pmatrix} 0 & -1/10 & 1/20 \\ -2/13 & 0 & 2/13 \\ -1 & -1 & 0 \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} 5/4 \\ 30/13 \\ 2 \end{pmatrix}$$

Risulta $\rho(J) = 0.44896 < 1$*, da cui il metodo converge e converge alla soluzione* $\mathbf{x} = (1, 2, -1)^T$*. Sia* $\mathbf{x}^{(0)} = (0, 0, 0)^T$*, allora:*

$$\begin{aligned}
\mathbf{x}^{(1)} &= (1.25, 2.3\ldots, 2)^T & \|\mathbf{e}^{(1)}\|_2 &\approx 3.026\\
\mathbf{x}^{(2)} &= (1.12\ldots, 2.42\ldots, -1.55\ldots)^T & \|\mathbf{e}^{(2)}\|_2 &\approx 0.71\\
\mathbf{x}^{(3)} &= (0.93\ldots, 1.89\ldots, -1.54\ldots)^T & \|\mathbf{e}^{(3)}\|_2 &\approx 0.55\\
\mathbf{x}^{(4)} &= (0.98\ldots, 1.92\ldots, -0.82\ldots)^T & \|\mathbf{e}^{(4)}\|_2 &\approx 0.19\\
\mathbf{x}^{(5)} &= (1.01\ldots, 2.02\ldots, -0.91\ldots)^T & \|\mathbf{e}^{(5)}\|_2 &\approx 0.09\\
\mathbf{x}^{(6)} &= (1.00\ldots, 2.01\ldots, -1.04\ldots)^T & \|\mathbf{e}^{(6)}\|_2 &\approx 0.046
\end{aligned}$$

Esempio 3.2 *Applichiamo il metodo di Gauss-Seidel al sistema lineare*

$$A\mathbf{x} = \mathbf{b} \qquad \textit{con} \qquad A = \begin{pmatrix} 20 & 2 & -1 \\ 2 & 13 & -2 \\ 1 & 1 & 1 \end{pmatrix} \qquad \textit{e} \qquad \mathbf{b} = \begin{pmatrix} 25 \\ 30 \\ 2 \end{pmatrix}.$$

La decomposizione di Gauss-Seidel è

$$M = \begin{pmatrix} 20 & 0 & 0 \\ 2 & 13 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -2 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

La matrice di iterazione G *ed il vettore* $\mathbf{q}$ *sono*

$$G = \begin{pmatrix} 0 & -1/10 & 1/20 \\ 0 & 1/65 & 9/130 \\ 0 & 11/130 & -51/260 \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} 5/4 \\ 55/26 \\ -71/52 \end{pmatrix}$$

Risulta $\rho(G) = 0.243858 < 1$*, da cui il metodo converge e converge alla soluzione* $\mathbf{x} = (1, 2, -1)^T$*. Sia* $\mathbf{x}^{(0)} = (0, 0, 0)^T$*, allora:*

$$\begin{aligned}
\mathbf{x}^{(1)} &= (1.25, 2.11\ldots, -1.36)^T & \|\mathbf{e}^{(1)}\|_2 &\approx 0.45\\
\mathbf{x}^{(2)} &= (0.97\ldots, 2.05\ldots, -0.91\ldots)^T & \|\mathbf{e}^{(2)}\|_2 &\approx 0.101\\
\mathbf{x}^{(3)} &= (0.998\ldots, 2.08\ldots, -1.011\ldots)^T & \|\mathbf{e}^{(3)}\|_2 &\approx 0.084\\
\mathbf{x}^{(4)} &= (0.991\ldots, 2.07\ldots, -0.99\ldots)^T & \|\mathbf{e}^{(4)}\|_2 &\approx 0.078
\end{aligned}$$

Capitolo 4

Metodi per Autovalori e Autovettori

Tra i diversi metodi numerici esistenti per calcolare gli autovalori e gli autovettori di una matrice, alcuni sono di tipo generale e sono applicabili a matrici dense e senza struttura, altri utilizzano in modo specifico eventuali proprietà di struttura o sparsità della matrice, permettendo di trattare problemi anche con dimensioni molto grandi. Alcuni metodi possono essere utilizzati per calcolare tutti gli autovalori e autovettori di una matrice, altri invece servono per calcolare solo alcuni autovalori, per esempio quelli che si trovano all'estremità dello spettro, ed i corrispondenti autovettori, come è richiesto in molte applicazioni. Tutti i metodi possono essere divisi in due classi:

1. Metodi in cui il calcolo viene effettuato in due fasi

   - riduzione con metodi diretti della matrice A in una matrice simile B, di cui sia più agevole calcolare gli autovalori;
   - calcolo degli autovalori di B con un metodo iterativo.

   Questi metodi si applicano in generale a problemi di piccole dimensioni, per i quali tutti i dati su cui si opera possono essere contenuti nella memoria centrale del calcolatore.

2. Metodi completamente iterativi, che richiedono ad ogni passo la moltiplicazione di una matrice per un vettore, o la risoluzione di un sistema lineare. Questi metodi si applicano in generale a problemi di grandi dimensioni, anche nel caso in cui non sia possibile contenere tutti i dati nella memoria del calcolatore.

Nei metodi della prima classe, per la riduzione della matrice A nella matrice B, si utilizzano metodi diretti analoghi a quelli descritti per la fattorizzazione delle matrici. Nel caso più generale la matrice B che si ottiene è tale che

$$b_{i,j} = 0 \quad \text{per} \quad i > j+1 \quad i,j = 1,\ldots,n.$$

Una matrice B con questa proprietà è detta essere in **forma di Hessenberg superiore**. La trasformazione per similitudine della matrice A nella matrice B è fatta per passi successivi

$$A_{k+1} = S_k \; A_k \; S_k^{-1} \quad k = 1,\ldots,m-1$$

dove

$$A_1 := A \quad \text{e alla fine} \quad B := A_m$$

per cui posto $S := S_{m-1} \; S_{m-2} \; \ldots S_1$, risulta

$$B = S \; A \; S^{-1}$$

e se $\mathbf{x}$ è autovettore di B, $S\mathbf{x}$ è autovettore di A.
Le matrici S_k sono di solito matrici elementari di Gauss o di Householder. Se S_k è una matrice di Householder, risulta

$$\|S_k\|_2 \cdot \|S_k^{-1}\|_2 = 1$$

I metodi iterativi per il calcolo degli autovalori di B potrebbero essere anche applicati direttamente alla matrice A. Trasformando però la matrice A nella matrice B, si abbassa notevolmente la complessità computazionale (per esempio per il metodo QR che vedremo si passa da $O(n^3)$ a $O(n^2)$). Per il calcolo degli autovalori della matrice B, due sono le tecniche più usate:

- se sono richiesti solo pochi autovalori rispetto alla dimensione della matrice (non più del 25%), conviene usare un metodo iterativo che calcoli un singolo autovalore per volta, come ad esempio un metodo di iterazione funzionale applicato all'equazione caratteristica o il metodo delle potenze inverse.

- se sono richiesti tutti o molti degli autovalori, il metodo migliore è in generale il metodo QR che vedremo.

Osservazione 4.1 *Se la matrice A è hermitiana, e la trasformazione viene eseguita con matrici unitarie, la matrice B risulta hermitiana tridiagonale.*

4.1 Riduzione di una matrice in forma di Hessenberg

Se si applicano ad una matrice $A_{n\times n}$ matrici elementari di Householder con l'obiettivo di azzerare gli elementi della colonna $j-$esima di indici $i = j+2, \ldots, n$ si ottiene una matrice di Hessenberg superiore ad un costo computazionale di $\frac{5}{3}n^3$ operazioni moltiplicative.

Esempio 4.1 *Si consideri la matrice* $A \in \mathbb{R}^{4\times 4}$

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

Applicando il metodo di Householder, al primo passo si ottiene

$$\mathbf{v}_1 = (0.0, 2.732051, 1.0, 1.0)^T \qquad \beta_1 = 0.2113248$$

e quindi $H_1 = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T$ *e* $A_2 = H_1 A H_1^T$

$$A_2 = \begin{pmatrix} 4 & -3.464098 & -0.366024 & -1.366024 \\ -1.73205 & 7.666641 & -0.5446615 & -1.12209 \\ 0 & 0.08931351 & 1.877991 & 1.032692 \\ 0 & 1.244013 & -0.6993591 & 2.455341 \end{pmatrix}.$$

Al secondo passo si ottiene

$$\mathbf{v}_2 = (0.0, 0.0, 1.336528, 1.244013)^T \qquad \beta_2 = 0.5999027$$

e quindi $H_2 = I - \beta_2 \mathbf{v}_2 \mathbf{v}_2^T$ *e* $A_3 = H_2 A_2 H_2^T$

$$A_3 = \begin{pmatrix} 4 & -3.464098 & 1.388726 & 0.2672625 \\ -1.73205 & 7.666641 & 1.158131 & 0.4629154 \\ 0 & -1.247213 & 2.476189 & -0.7423077 \\ 0 & 0 & 0.9897442 & 1.857142 \end{pmatrix}$$

che è in forma di Hessenberg superiore.

4.2 Metodo Q R per il calcolo degli autovalori

Il metodo QR è il metodo più usato per calcolare tutti gli autovalori di una matrice, in quanto il più efficiente.

Il metodo è molto complicato, sia come descrizione che come implementazione, anche se il principio su cui si basa è semplice. Il metodo richiede tutta una serie di accorgimenti, senza i quali non potrebbe essere efficiente: riduzione preliminare della matrice in forma di Hessenberg superiore (o triangolare nel caso di matrici hermitiana) per ridurre il costo computazionale ad ogni iterazione, utilizzazione di una tecnica di traslazione per aumentare la velocità di convergenza; riduzione dell'ordine della matrice quando un autovalore è stato approssimato con sufficiente precisione, per calcolare un altro autovalore.
Il metodo QR, che è stato descritto da Francis nel 1961, utilizza la fattorizzazione QR di una matrice; esso deriva da un precedente metodo, detto metodo LR, proposto da Rutishauser nel 1958, che utilizza la fattorizzazione LU di una matrice.

4.2.1 Algoritmo di base

Nel metodo QR viene generata una successione $\{A_k\}$ di matrici nel modo seguente: posto

$$A_1 := A$$

per $k = 1, 2, \ldots$ si calcola una fattorizzazione QR di A_k

$$A_k = Q_k R_k$$

dove Q_k è unitaria e R_k è triangolare superiore, e si definisce la matrice A_{k+1} per mezzo della relazione

$$A_{k+1} := R_k Q_k.$$

per le precedenti risulta che

$$A_{k+1} = Q_k^T A_k Q_k$$

e quindi le matrici della successione $\{A_k\}$ sono tutte simili fra loro. Sotto opportune ipotesi la successione converge ad una matrice triangolare superiore (diagonale se A è hermitiana) che ha come elementi diagonali gli autovalori di A.

4.2.2 Risultati di convergenza

Il seguente teorema di convergenza, viene dato in ipotesi piuttosto restrittive in cui è più facile la sua dimostrazione, ma la sua validità può essere provata anche in ipotesi più deboli.

Teorema 4.1 *Sia $A \in \mathbb{C}^{n\times n}$ tale che i suoi autovalori $\lambda_i,\ i = 1,\ldots,n$ abbiano moduli tutti distinti, cioè*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0.$$

Indicata con X la matrice degli autovettori di A, tale che

$$A = XDX^{-1},$$

in cui D è la matrice diagonale degli autovalori, si supponga che la matrice X^{-1} ammetta la fattorizzazione LU. Allora esistono delle matrici S_k tali che

$$\lim_{k\to\infty} S_k^H R_k S_{k-1} = \lim_{k\to\infty} S_{k-1}^H A_k S_{k-1} = T$$

e

$$\lim_{k\to\infty} S_{k-1}^H Q_k S_k = I$$

dove T è triangolare superiore con gli elementi diagonali uguali a $\lambda_1, \lambda_2, \ldots, \lambda_n$. Quindi gli elementi diagonali di A_k tendono agli autovalori di A.

Se X^{-1} non ammette fattorizzazione LU, si può dimostrare che il metodo QR è ancora convergente. In questo caso gli elementi diagonali di T coincidono ancora con i λ_i, ma non sono più in ordine di modulo decrescente. Se l'ipotesi che tutti gli autovalori abbiano modulo distinto, non è verificata, la successione formata dagli elementi diagonali di A_k non converge. Questa ipotesi è troppo restrittiva, e non consente di utilizzare il metodo QR in casi particolarmente importanti nelle applicazioni, come quelli in cui la matrice A ha elementi reali e autovalori non reali. Però anche in questo caso il metodo QR può essere applicato con opportune varianti. Sia ad esempio

$$|\lambda_1| > \ldots > |\lambda_r| = |\lambda_{r+1}| > \ldots > |\lambda_n| > 0$$

dove λ_r e λ_{r+1} sono due numeri complessi coniugati, oppure due numeri reali. Allora la successione delle $\{A_k\}$ o meglio degli elementi diagonali non converge agli autovalori di stesso modulo, ma gli autovalori dei blocchi diagonali convergono a λ_r e λ_{r+1}. Come detto, gli elementi diagonali di A_k di indice diverso da r ed $r+1$ invece convergono agli altri autovalori. Situazioni analoghe si presentano quando la matrice A ha più autovalori di modulo uguale e in questo caso il metodo QR genera matrici R_k con struttura triangolare a blocchi, in cui gli autovalori dei blocchi diagonali convergono ad autovalori di A.

4.2.3 Costo computazionale e stabilità

Il metodo QR applicato a una matrice di ordine n ha ad ogni passo un costo computazionale dell'ordine di n^3 operazioni moltiplicative (per calcolare la fattorizzazione $A_k = Q_k R_k$ e per moltiplicare la matrice triangolare R_k per le matrici elementari di Householder).
Per abbassare il costo computazionale globale conviene prima trasformare la matrice A in forma di Hessenberg superiore. Questa trasformazione viene eseguita una sola volta perché il metodo QR, applicato a matrici in forma di Hessenberg superiore produce matrici A_k in forma di Hessenberg superiore.
Infatti se A_k è in forma di Hessenberg superiore, la matrice Q_k è data dal prodotto di $n-1$ matrici elementari di Hoseholder che sono in forma di Hessenberg superiore e quindi la matrice A_{k+1}, prodotto di una matrice triangolare superiore R_k per una matrice Q_k in forma di Hessenberg superiore, risulta ancora in forma di Hessenberg superiore.
Se la matrice A è hermitiana, la matrice in forma di Hessenberg superiore, ottenuta applicando ad A il metodo di Householder, è ancora hermitiana, e quindi risulta tridiagonale. Inoltre anche tutte le matrici A_k generate dal metodo QR sono hermitiane e quindi tridiagonali.
Il metodo QR applicato ad una matrice A in forma di Hessenberg superiore ha ad ogni passo un costo di $2n^2$ operazioni moltiplicative (che è il costo computazionale per calcolare la fattorizzazone $A_k = Q_k R_k$).
Per quel che riguarda la stablità si può dimostrare che il metodo QR per il calcolo degli autovalori gode delle stesse proprietà di stabilità di cui gode la fattorizzazione QR di una matrice.

4.3 Condizionamento del calcolo degli autovalori

Una matriice può avere alcuni autovalori molto sensibili alle perturbazioni dei suoi elementi mentre altri insensibili. È conveniente avere un numero che definisce il condizionamento di una matrice rispetto al problema degli autovalori; esso sarà il **numero di condizione spettrale**.
È evidente che tale numero può fornire una informazione insoddisfacente. Infatti se qualche autovalore è molto sensibile allora il numero di condizione spettrale è necessariamente grande, anche se altri autovalori sono insensibili. Si può ovviare a questo inconveniente introducendo numeri di condizione che evidenziano la sensibilità dei singoli autovalori e che si

chiamano **numeri di condizione della matrice** rispetto al problema degli autovalori.

Definizione 4.1 *Si dice norma assoluta $\|\cdot\|$ una norma matriciale indotta che verifichi le seguenti proprietà*

$$\|D\| = \max_{i=1,\dots,n} |d_{i,i}|$$

per ogni matrice diagonale $D \in \mathbb{C}^{n\times n}$.

Osservazione 4.2 *Le norme $\|\cdot\|_1, \|\cdot\|_2$ e $\|\cdot\|_\infty$ sono assolute.*

Teorema 4.2 (di Bauer-Fike) *Sia $A \in \mathbb{C}^{n\times n}$ una matrice diagonalizzabile, cioè tale che*

$$A = TDT^{-1}$$

con D diagonale e T non singolare. Se $\delta A \in \mathbb{C}^{n\times n}$ e η è un autovalore di $A + \delta A$, allora esiste almeno un autovalore λ di A tale che

$$|\lambda - \eta| \leq \mu(T)\|\delta A\|$$

dove $\mu(T) = \|T\| \cdot \|T^{-1}\|$ per una norma assoluta $\|\cdot\|$.

Introduciamo un numero di condizione che dia una misura dell'effetto della perturbazione su un solo autovalore di A.

Teorema 4.3 *Sia $A \in \mathbb{C}^{n\times n}$, λ un autovalore di A di molteplicità algebrica 1, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{n\times n}$, $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, tali che*

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

$$\mathbf{y}^H A = \lambda\mathbf{y}^H.$$

Allora è $\mathbf{y}^H\mathbf{x} \neq 0$ ed inoltre esiste nel piano complesso un intorno V dello zero e una funzione $\lambda(\epsilon) : V \to \mathbb{C}$, analitica, tale che

1. *$\lambda(\epsilon)$ è autovalore con molteplicità 1 di $A + \epsilon F$, $F \in \mathbb{C}^{n\times n}$*
2. *$\lambda(0) = \lambda$*
3. *$\lambda'(0) = \frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}$*
4. *a meno di termini di ordine superiore in ϵ è*

$$\lambda(\epsilon) - \lambda = \epsilon\frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}.$$

da questo risultato si ha che la variazione nell'autovalore dovuta alla perturbazione ϵF di A è proporzionale ad ϵ.
Inoltre il condizionamento del problema dipende dalla quantità

$$\left| \frac{\mathbf{y}^H F \mathbf{x}}{\mathbf{y}^H \mathbf{x}} \right|$$

che, data F, è tanto più grande quanto più piccolo è $|\mathbf{y}^H \mathbf{x}|$.
Nel caso di autovalore λ di molteplicità algebrica $\sigma(\lambda) > 1$ e molteplicità geometrica $\tau(\lambda)$ si può arrivare ad una relazione del tipo

$$|\lambda_i(\epsilon) - \lambda| \leq \gamma |\epsilon|^{1/\mu}$$

per $i = 1, \ldots, \sigma(\lambda)$, γ costante positiva; se $\mu > 1$ il problema del calcolo di λ può essere fortemente malcondizionato.

Capitolo 5

Il problema dei Minimi Quadrati

5.1 Le equazioni normali

Sia

$$A\mathbf{x} = \mathbf{b} \tag{5.1}$$

un sistema lineare con $A \in \mathbb{R}^{m\times n}$ tale che $m \geq n$. Se $m > n$, il sistema (5.1) ha più equazioni che incognite e si dice **sovradeterminato**. Se il sistema (5.1) non ha soluzione, fissata una norma vettoriale $\|\cdot\|$, si cercano i vettori $\mathbf{x} \in \mathbb{R}^n$ che minimizzano la quantità $\|A\mathbf{x} - \mathbf{b}\|$. In norma 2, il problema diventa quello di determinare un vettore $\mathbf{x} \in \mathbb{R}^n$ tale che

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{y}\in\mathbb{R}^n} \|A\mathbf{y} - \mathbf{b}\|_2 = \mu. \tag{5.2}$$

Questo problema viene detto **problema dei minimi quadrati**.

Teorema 5.1 *Se A ha rango massimo la soluzione del problema (5.2) esiste, è unica e coincide con la soluzione del sistema lineare*

$$A^T A\mathbf{x} = A^T\mathbf{b}. \tag{5.3}$$

Tale sistema viene detto sistema delle **equazioni normali**.

dim. Siano

$$S(A) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n\}$$

e

$$S(A)^\perp = \{\mathbf{z} \in \mathbb{R}^m : \mathbf{z}^T\mathbf{y} = 0,\ \forall \mathbf{y} \in S(A)\}$$

il sottospazio di $\mathbb{R}^m$ immagine di A, e il sottospazio ortogonale a $S(A)$. Il vettore $\mathbf{b}$ può essere così decomposto

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2, \quad \text{dove} \quad \mathbf{b}_1 \in S(A) \quad \text{e} \quad \mathbf{b}_2 \in S(A)^\perp$$

per cui per il residuo

$$\mathbf{r} = \mathbf{b}_1 - A\mathbf{x} + \mathbf{b}_2 = \mathbf{y} + \mathbf{b}_2,$$

dove

$$\mathbf{y} = \mathbf{b}_1 - A\mathbf{x} \in S(A) \quad \text{e} \quad \mathbf{b}_2 \in S(A)^\perp$$

vale

$$\|\mathbf{r}\|_2^2 = (\mathbf{y} + \mathbf{b}_2)^T(\mathbf{y} + \mathbf{b}_2) = \|\mathbf{y}\|_2^2 + \|\mathbf{b}_2\|_2^2,$$

in quanto $\mathbf{y}^T\mathbf{b}_2 = \mathbf{b}_2^T\mathbf{y} = 0$. Poiché solo $\mathbf{y}$ dipende da $\mathbf{x}$, si ha che $\|\mathbf{r}\|_2^2$ è minimo se e solo se $\mathbf{b}_1 = A\mathbf{x}$, cioè se e solo se il vettore $\mathbf{r} \in S(A)^\perp$ ed è quindi ortogonale alle colonne di A, cioè

$$A^T\mathbf{r} = A^T(\mathbf{b} - A\mathbf{x}) = \mathbf{0}.$$

Ne segue quindi che $\mathbf{x}$ è soluzione di (5.2) se e solo se è soluzione di (5.3).

5.2 Metodo QR per i minimi quadrati

Sia A di rango massimo e si applichi il metodo di Householder per ottenere la fattorizzazione $Q\ R$ di A; sarà $Q \in \mathbb{R}^{m\times m}$ ortogonale ed $R \in \mathbb{R}^{m\times n}$ con la seguente particolare struttura

$$R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

con $R_1 \in \mathbb{R}^{n\times n}$ triangolare superiore e non singolare essendo H di rango massimo.
Dalla fattorizzazione $A = Q\ R$ si ha

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|_2 &= \|QR\mathbf{x} - \mathbf{b}\|_2 \\ &= \|Q(R\mathbf{x} - Q^T\mathbf{b})\|_2 \\ &= \|R\mathbf{x} - Q^T\mathbf{b}\|_2 \\ &= \|R\mathbf{x} - \mathbf{c}\|_2 \end{aligned}$$

dove $\mathbf{c} = Q^T\mathbf{b}$.
Partizionando il vettore $\mathbf{c}$ nel modo seguente

$$\mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}$$

con $\mathbf{c}_1 \in \mathbb{R}^n$ e $\mathbf{c}_2 \in \mathbb{R}^{m-n}$, si ha che

$$R\mathbf{x} - \mathbf{c} = \begin{pmatrix} R_1\mathbf{x} - \mathbf{c}_1 \\ -\mathbf{c}_2 \end{pmatrix}$$

da cui

$$\begin{aligned} \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 &= \min_{\mathbf{x}} \|R\mathbf{x} - \mathbf{c}\|_2^2 \\ &= \min_{\mathbf{x}} \left(\|R_1\mathbf{x} - \mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2\right) \\ &= \|\mathbf{c}_2\|_2^2 + \min_{\mathbf{x}} \left(\|R_1\mathbf{x} - \mathbf{c}_1\|_2^2\right) \end{aligned}$$

Poicé R_1 è non singolare, esiste una ed una sola soluzione per il sistema

$$R_1\mathbf{x} = \mathbf{c}_1,$$

sia questa $\mathbf{x}^*$ e sarà

$$\min_{\mathbf{x}} \left(\|R_1\mathbf{x} - \mathbf{c}_1\|_2^2\right) = \|R_1\mathbf{x}^* - \mathbf{c}_1\|_2^2 = 0$$

Segue che $\mathbf{x}^*$ è la soluzione del problema dei minimi quadrati e

$$\|\mathbf{c}_2\|_2 = \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2.$$

Bibliografia

[Bau88] D. Bau, N. Trefethen. *Numerical Linear Algebra*. SIAM, 1988.

[Bini88] D. Bini, M. Capovani, O. Menchi. *Metodi Numerici per l'Algebra Lineare*. Zanichelli, 1988.

[Farin88] G. Farin, D. Hansford. *The Geometry Toolbox for Graphics Modeling*. A.K.Peters, 1988.