Capitolo 7

IL PROBLEMA LINEARE DEI MINIMI QUADRATI

1. Le equazioni normali

Sia

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

un sistema lineare in cui la matrice $A \in \mathbf{C}^{m\times n}$ dei coefficienti è tale che $m \geq n$. Se $m > n$, il sistema (1) ha più equazioni che incognite e si dice *sovradeterminato*. Se il sistema (1) non ha soluzione, fissata una norma vettoriale $\| \ . \ \|$, si ricercano i vettori $\mathbf{x} \in \mathbf{C}^n$ che minimizzano la quantità $\|A\mathbf{x} - \mathbf{b}\|$. In norma 2, il problema diventa quello di determinare un vettore $\mathbf{x} \in \mathbf{C}^n$ tale che

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{y}\in\mathbf{C}^n} \|A\mathbf{y} - \mathbf{b}\|_2 = \gamma. \tag{2}$$

Tale problema viene detto *problema dei minimi quadrati*.

Il seguente teorema caratterizza l'insieme X dei vettori $\mathbf{x} \in \mathbf{C}^n$ che soddisfano alla (2).

7.1 Teorema. *Valgono le seguenti proprietà:*

a) $\mathbf{x} \in X$ *se e solo se*

$$A^H A\mathbf{x} = A^H \mathbf{b}. \tag{3}$$

Il sistema (3) viene detto sistema delle equazioni normali o sistema normale.

b) X *è un insieme non vuoto, chiuso e convesso.*

c) L'insieme X *si riduce ad un solo elemento* $\mathbf{x}^*$ *se e solo se la matrice* A *ha rango massimo.*

d) Esiste $\mathbf{x}^* \in X$ *tale che*

$$\|\mathbf{x}^*\|_2 = \min_{\mathbf{x}\in X} \|\mathbf{x}\|_2. \tag{4}$$

Il vettore $\mathbf{x}^*$ *è l'unico vettore di* X *che appartiene a* $N(A^H A)^\perp$ *ed è detto soluzione di minima norma.*

Dim. a) Siano

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

e

$$S(A)^{\perp} = \{\ \mathbf{z} \in \mathbf{C}^m\ :\ \mathbf{z}^H\mathbf{y} = 0,\ \text{per ogni } \mathbf{y} \in S(A)\ \}$$

il sottospazio di $\mathbf{C}^m$ immagine di A, e il sottospazio ortogonale a $S(A)$ (si vedano i paragrafi 2 e 6 del capitolo 1). Il vettore $\mathbf{b}$ può essere così decomposto

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

per cui per il residuo

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

vale

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

in quanto $\mathbf{y}^H\mathbf{b}_2 = \mathbf{b}_2^H\mathbf{y} = 0$. 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 $\mathbf{r}$ appartiene a $S(A)^{\perp}$ ed è quindi ortogonale alle colonne di A, cioè

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

Ne segue quindi che $\mathbf{x} \in X$ se e solo se $\mathbf{x}$ è soluzione di (3). Inoltre risulta $\gamma^2 = \|\mathbf{b}_2\|_2^2$.

Nel caso di $\mathbf{R}^2$ con una matrice A di rango 1 si può dare la seguente interpretazione geometrica, illustrata nella figura 7.1. Il vettore $\mathbf{b} = \mathbf{r} - A\mathbf{x}$ risulta decomposto in un sol modo nel vettore $\mathbf{b}_2 = \mathbf{r} \in S(A)^{\perp}$ e nel vettore $\mathbf{b}_1 = A\mathbf{x} \in S(A)$. Il vettore $A\mathbf{x}$ è quindi la proiezione ortogonale del vettore $\mathbf{b}$ sul sottospazio generato dalle colonne di A.

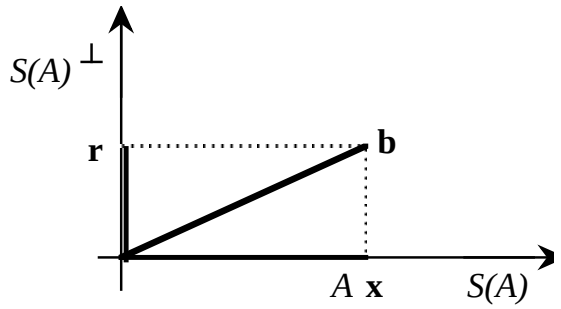

Fig. 7.1 - Proiezione ortogonale del vettore $\mathbf{b}$.

b) Da quanto detto precedentemente, segue che l'insieme X è non vuoto. Se $\mathbf{x}_0$ è tale che

$$A^HA\mathbf{x}_0 = A^H\mathbf{b}$$

allora risulta

$$X = \{\ \mathbf{x} \in \mathbf{C}^n\ :\ \mathbf{x} = \mathbf{x}_0 + \mathbf{v},\ \mathbf{v} \in N(A^H A)\ \}.$$

Quindi X è una *varietà lineare affine*, parallela ad $N(A^H A)$, passante per $\mathbf{x}_0$, e poiché $N(A^H A)$ è chiuso e convesso, X è un insieme chiuso e convesso.

c) La matrice A ha rango massimo se e solo se la matrice $A^H A$ è non singolare (si veda il paragrafo 6 del capitolo 1) e quindi se e solo se il sistema (3) ha una e una sola soluzione $\mathbf{x}^*$. Perciò l'insieme X è costituito dal solo elemento $\mathbf{x}^*$ se e solo se la matrice A ha rango massimo. In tal caso l'insieme $N(A^H A)$ è costituito dal solo elemento nullo.

d) l'esistenza della soluzione di minima norma è ovvia nel caso in cui X si riduce al solo elemento $\mathbf{x}^*$. Se X non si riduce al solo elemento $\mathbf{x}^*$, sia $\mathbf{x}_0 \in X$ e si consideri l'insieme

$$B = \{\ \mathbf{x} \in \mathbf{C}^n\ :\ \|\mathbf{x}\|_2 \leq \|\mathbf{x}_0\|_2\ \}.$$

Poiché, se $\mathbf{x} \in X$, ma $\mathbf{x} \notin B$, risulta $\|\mathbf{x}\|_2 > \|\mathbf{x}_0\|_2$, allora

$$\min_{\mathbf{x} \in X} \|\mathbf{x}\|_2 = \min_{\mathbf{x} \in X \cap B} \|\mathbf{x}\|_2.$$

L'insieme $X \cap B$ è un insieme non vuoto, limitato e chiuso, in quanto intersezione di insiemi chiusi, e quindi compatto; essendo la norma una funzione continua, esiste un $\mathbf{x}^* \in X \cap B$ per cui vale la (4).

Inoltre $\mathbf{x}^*$ è l'unico vettore di X appartenente a $N(A^H A)^\perp$. Infatti esistono e sono unici $\mathbf{y} \in N(A^H A)$ e $\mathbf{z} \in N(A^H A)^\perp$ tali che $\mathbf{x}^* = \mathbf{y} + \mathbf{z}$. Poiché $\mathbf{x}^*$ è soluzione di (3), è $A^H A(\mathbf{y} + \mathbf{z}) = A^H \mathbf{b}$, da cui $A^H A\mathbf{z} = A^H \mathbf{b}$ e quindi $\mathbf{z}$ è soluzione del problema (2). Se $\mathbf{y}$ non fosse uguale a $\mathbf{0}$, $\mathbf{z}$ avrebbe norma 2 minore di $\|\mathbf{x}^*\|_2$, ciò che è assurdo perché $\mathbf{x}^*$ è la soluzione di minima norma: ne segue che $\mathbf{x}^* = \mathbf{z} \in N(A^H A)^\perp$.

Nel caso di $\mathbf{R}^2$ con una matrice A di rango 1, si può dare l'interpretazione geometrica illustrata nella figura 7.2, in cui è riportata la varietà X, parallela alla varietà $N(A^H A)$. Il punto $\mathbf{x}_0$ è un qualunque punto di X. Il punto $\mathbf{x}^*$ è quello di minima norma, e quindi quello più vicino all'origine O dello spazio $\mathbf{C}^n$. ∎

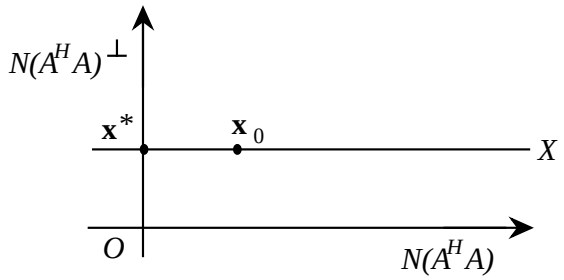

Fig. 7.2 - $\mathbf{x}^*$ è la soluzione di minima norma.

Se la matrice A ha rango massimo, allora la soluzione del problema dei minimi quadrati può essere ottenuta risolvendo il sistema (3). In tal caso, poiché la matrice $A^H A$ è definita positiva, si può utilizzare per la risoluzione il metodo di Cholesky. Determinata la matrice L triangolare inferiore tale che

$$LL^H = A^H A,$$

la soluzione $\mathbf{x}^*$ di (3) viene calcolata risolvendo successivamente i due sistemi di ordine n con matrice dei coefficienti triangolare

$$L\mathbf{y} = A^H \mathbf{b}$$
$$L^H \mathbf{x} = \mathbf{y}.$$

Il costo computazionale è di $n^2 m/2$ operazioni moltiplicative per la costruzione della matrice hermitiana $A^H A$ e di $n^3/6$ operazioni moltiplicative per il calcolo della soluzione del sistema (3). Quindi in totale il calcolo della soluzione del problema dei minimi quadrati per mezzo della risoluzione del sistema (3) con il metodo di Cholesky ha un costo computazionale di

$$f_1(n,m) = \frac{n^2}{2}(m + \frac{n}{3}) \quad \text{operazioni moltiplicative.} \tag{5}$$

7.2 Esempio. Si consideri il problema dei minimi quadrati (2) con

$$A = \frac{1}{45}\begin{bmatrix} 14 & 32 & -38 \\ -44 & 58 & 8 \\ -18 & 96 & 51 \\ 63 & -36 & 54 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Per determinare una soluzione di (2) si costruiscono

$$A^T A = \frac{1}{81}\begin{bmatrix} 257 & -244 & 64 \\ -244 & 596 & 88 \\ 64 & 88 & 281 \end{bmatrix}, \quad A^T \mathbf{b} = \frac{1}{3}\begin{bmatrix} 1 \\ 10 \\ 5 \end{bmatrix}.$$

Poiché la matrice $A^T A$ è non singolare, si applica il metodo di Cholesky, ottenendo la fattorizzazione LL^T, dove

$$L = \frac{1}{9\sqrt{257}}\begin{bmatrix} 257 & 0 & 0 \\ -244 & 306 & 0 \\ 64 & \frac{2124}{17} & \frac{243}{17}\sqrt{257} \end{bmatrix}.$$

La soluzione $\mathbf{x}^*$ risulta

$$\mathbf{x}^* = \frac{1}{54}\begin{bmatrix} 46 \\ 43 \\ 2 \end{bmatrix}.$$

Sostituendo nella (2) si ha che

$$A\mathbf{x}^* - \mathbf{b} = \frac{1}{5}\,[\text{-}1,\ \text{-}4,\ 2,\ \text{-}2]^T,$$

e quindi $\gamma = \|A\mathbf{x}^* - \mathbf{b}\|_2 = 1.$ ∎

Se la matrice A non ha rango massimo, non si può risolvere il sistema (3) con il metodo di Cholesky, ma si può applicare il metodo di Gauss con la variante del massimo pivot.

7.3 Esempio. Si consideri il problema dei minimi quadrati (2) con

$$A = \frac{1}{45}\begin{bmatrix} 6 & 12 & \text{-}72 \\ \text{-}16 & \text{-}7 & \text{-}8 \\ 58 & 16 & 104 \\ 87 & 24 & 156 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Si ha

$$A^TA = \frac{1}{81}\begin{bmatrix} 449 & 128 & 772 \\ 128 & 41 & 184 \\ 772 & 184 & 1616 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix},$$

e la matrice A^TA è singolare. Calcolando la fattorizzazione LU della matrice A^TA con il metodo di Gauss si ottiene

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{128}{449} & 1 & 0 \\ \dfrac{772}{449} & \text{-}8 & 1 \end{bmatrix}, \quad U = \frac{1}{81}\begin{bmatrix} 449 & 128 & 772 \\ 0 & \dfrac{2025}{449} & \text{-}\dfrac{16200}{449} \\ 0 & 0 & 0 \end{bmatrix}.$$

Ne segue che la matrice A ha rango 2. L'insieme X risulta formato dai vettori

$$\mathbf{x} = \begin{bmatrix} -\dfrac{1}{5} - 4h \\ \dfrac{13}{5} + 8h \\ h \end{bmatrix}, \quad h \in \mathbf{C}.$$

Il vettore di minima norma $\mathbf{x}^*$ può essere ricavato calcolando il valore di h per cui la funzione

$$f(h) = \|\mathbf{x}\|_2^2$$

è minima. Tale valore è $h = -\dfrac{4}{15}$, a cui corrisponde il vettore

$$\mathbf{x}^* = \frac{1}{15}\,[13,\ 7,\ -4]^T.$$

Sostituendo nella (2) si ha che

$$A\mathbf{x}^* - \mathbf{b} = \frac{1}{3}\,[-1,\ -4,\ -1,\ 0]^T,$$

e quindi $\gamma = \|A\mathbf{x}^* - \mathbf{b}\|_2 = \sqrt{2}$. ∎

Operando in aritmetica finita il sistema (3) può risultare non risolubile, e in tal caso non può essere utilizzato per risolvere il problema (2).

7.4 Esempio. Sia u la precisione di macchina con cui si eseguono i calcoli. Si consideri il problema (2) con

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

dove α è un numero tale che $u \le \alpha < 1$ e $\alpha^2 < u$. Si ha

$$A^T A = \begin{bmatrix} 25 & 25 \\ 25 & 25 + \alpha^2 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 7 \\ 7 + \alpha \end{bmatrix}.$$

Applicando il metodo di Cholesky, si calcola la matrice L della fattorizzazione LL^T di $A^T A$:

$$L = \begin{bmatrix} 5 & 0 \\ 5 & \alpha \end{bmatrix},$$

da cui si ricava la soluzione del problema (2) :

$$\mathbf{x}^* = \left[\frac{7}{25} - \frac{1}{\alpha}\,,\ \frac{1}{\alpha}\right]^T.$$

Però operando con precisione u, poiché $\alpha^2 < u$, la matrice effettivamente calcolata al posto della $A^T A$ è

$$\begin{bmatrix} 25 & 25 \\ 25 & 25 \end{bmatrix}$$

e ha rango 1. Con tale matrice al posto della A^TA il sistema (3) non sarebbe risolubile. ∎

2. Metodo QR per il calcolo della soluzione del problema dei minimi quadrati

Si esamina ora un altro procedimento, detto metodo QR, che opera direttamente sulla matrice A fattorizzandola nella forma QR.

Si supponga dapprima che la matrice $A \in \mathbf{C}^{m\times n}$ abbia rango massimo $k = n \leq m$. Si applica il metodo di Householder alla matrice A, ottenendo una successione di matrici di Householder

$$P^{(k)} \in \mathbf{C}^{m\times m},\ k = 1, \ldots, n.$$

Posto $Q^H = P^{(1)}P^{(2)}\ldots P^{(n)}$, risulta

$$A = QR, \tag{6}$$

dove la matrice $R \in \mathbf{C}^{m\times n}$ ha la forma

$$R = \begin{bmatrix} R_1 \\ \\ O \end{bmatrix} \quad \begin{matrix} \} & n \text{ righe} \\ \\ \} & m-n \text{ righe} \end{matrix} \tag{7}$$

ed R_1 è una matrice triangolare superiore non singolare in quanto A ha rango massimo. Dalla (6) si ha

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|_2 &= \|QR\mathbf{x} - \mathbf{b}\|_2 = \|Q(R\mathbf{x} - Q^H\mathbf{b})\|_2 = \|R\mathbf{x} - Q^H\mathbf{b}\|_2 \\ &= \|R\mathbf{x} - \mathbf{c}\|_2. \end{aligned} \tag{8}$$

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

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \\ \mathbf{c}_2 \end{bmatrix} \quad \begin{matrix} \} & n \text{ componenti} \\ \\ \} & m-n \text{ componenti} \end{matrix}$$

per la (7) vale

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

Per la (8) risulta

$$\begin{aligned} \min_{\mathbf{x}\in\mathbf{C}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 &= \min_{\mathbf{x}\in\mathbf{C}^n} \|R\mathbf{x} - \mathbf{c}\|_2^2 = \min_{\mathbf{x}\in\mathbf{C}^n} \left[\|R_1\mathbf{x} - \mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2\right] \\ &= \|\mathbf{c}_2\|_2^2 + \min_{\mathbf{x}\in\mathbf{C}^n} \|R_1\mathbf{x} - \mathbf{c}_1\|_2^2. \end{aligned}$$

Poiché R_1 è non singolare, la soluzione $\mathbf{x}^*$ del sistema lineare

$$R_1\mathbf{x} = \mathbf{c}_1 \tag{9}$$

è tale che

$$\min_{\mathbf{x}\in\mathbf{C}^n} \|R_1\mathbf{x} - \mathbf{c}_1\|_2 = \|R_1\mathbf{x}^* - \mathbf{c}_1\|_2 = 0.$$

Ne segue che $\mathbf{x}^*$ è la soluzione del problema (2) e

$$\gamma = \|\mathbf{c}_2\|_2.$$

Analogamente al caso della risoluzione dei sistemi lineari, il metodo può essere applicato senza calcolare effettivamente né le matrici $P^{(k)}, k = 1, \ldots, n$, né la matrice Q. Si può procedere infatti nel modo seguente: sia

$$P^{(k)} = I - \beta_k \mathbf{v}_k \mathbf{v}_k^H, \quad k = 1, \ldots, n,$$

secondo la notazione del paragrafo 12 del capitolo 4. Si considera la matrice

$$T^{(1)} = [A \mid \mathbf{b}]$$

e si costruisce le successione delle matrici $T^{(k)}$ tali che

$$T^{(k+1)} = P^{(k)}T^{(k)} = T^{(k)} - \beta_k \mathbf{v}_k \mathbf{v}_k^H T^{(k)}.$$

Al termine dopo n passi si ottiene la matrice

$$T^{(n+1)} = [R \mid \mathbf{c}].$$

Il costo computazionale di questa fattorizzazione è di $n^2(m - n/3)$ operazioni moltiplicative (si veda il paragrafo 13 del capitolo 4); il costo computazionale della risoluzione del sistema triangolare (9) è di $n^2/2$ operazioni moltiplicative. Quindi il costo computazionale del calcolo della soluzione del problema dei minimi quadrati è di

$$f_2(n, m) = n^2(m - \frac{n}{3}) \qquad \text{operazioni moltiplicative.} \tag{10}$$

Confrontando questo costo computazionale con quello riportato nella (5) risulta che

$$f_2(n, m) \geq f_1(n, m) \qquad \text{per} \quad m \geq n,$$

per cui il metodo QR richiede in generale più operazioni di quante sono richieste risolvendo con il metodo di Cholesky il sistema normale (3). Se $m = n$ i due metodi hanno lo stesso costo computazionale.

La matrice R_1 ottenuta con il metodo QR e la matrice L^H ottenuta applicando il metodo di Cholesky alla matrice $A^H A$ sono uguali a meno della moltiplicazione per una matrice di fase. Infatti dalla (6) si ha

$$LL^H = A^H A = R^H Q^H QR = R^H R = R_1^H R_1,$$

dove sia la L^H che la R_1 sono triangolari superiori. Quindi

$$R_1 = DL^H,$$

dove $D \in \mathbf{C}^{n\times n}$ è una matrice diagonale unitaria.

7.5 Esempio. Per calcolare la soluzione del problema dei minimi quadrati (2) dell'esempio 7.2, si applica il metodo QR. Si considera la matrice

$$T^{(1)} = [A \mid \mathbf{b}] = \frac{1}{45}\begin{bmatrix} 14 & 32 & -38 & 45 \\ -44 & 58 & 8 & 45 \\ -18 & 96 & 51 & 45 \\ 63 & -36 & 54 & 45 \end{bmatrix},$$

e dopo 3 passi si ottiene la matrice

$$T^{(4)} = \begin{bmatrix} R_1 & \mathbf{c}_1 \\ O & \mathbf{c}_2 \end{bmatrix} = \frac{1}{9\sqrt{257}}\begin{bmatrix} -257 & 244 & -64 & -27 \\ 0 & -306 & -\frac{2124}{17} & -\frac{4221}{17} \\ 0 & 0 & \frac{243}{17}\sqrt{257} & -\frac{9}{17}\sqrt{257} \\ 0 & 0 & 0 & 9\sqrt{257} \end{bmatrix}.$$

Risolvendo il sistema (9) si ricava la soluzione $\mathbf{x}^*$ già trovata nell'esempio 7.2. Inoltre è

$$\mathbf{c}_2 = [1],$$

e quindi $\gamma = \|\mathbf{c}_2\|_2 = 1$. ∎

Se la matrice A non ha rango massimo, la matrice R_1 ottenuta ha almeno un elemento diagonale nullo e quindi non è possibile calcolare la soluzione del sistema (9). Questa difficoltà viene superata applicando il metodo QR con la tecnica del *massimo pivot per colonne* nel modo seguente: al k-esimo passo, costruita la matrice $A^{(k)}$ della forma

$$A^{(k)} = \begin{bmatrix} C^{(k)} & D^{(k)} \\ O & B^{(k)} \end{bmatrix} \quad \begin{matrix} \} & k-1 \text{ righe} \\ \} & m-k+1 \text{ righe} \end{matrix}$$

si determina la colonna di $B^{(k)}$ la cui norma 2 è massima. Sia j, $1 \leq j \leq n-k+1$, l'indice di tale colonna. Se $j \neq 1$, si scambiano fra loro la k-esima e la $(k+j-1)$-esima colonna della matrice $A^{(k)}$. Quindi si applica la matrice elementare $P^{(k)}$ alla matrice con le colonne così permutate. Se il rango di A è $r < m$, questo procedimento termina dopo r passi, e si ottiene una decomposizione del tipo

$$A\Pi = QR, \tag{11}$$

dove $\Pi \in \mathbf{R}^{n\times n}$ è una matrice di permutazione, $Q \in \mathbf{C}^{m\times m}$ è una matrice unitaria ed R è della forma

$$R = \begin{bmatrix} R_1 & S \\ O & O \end{bmatrix} \begin{matrix} \} & r \text{ righe} \\ \} & m-r \text{ righe} \end{matrix} \tag{12}$$

in cui $R_1 \in \mathbf{C}^{r\times r}$ è triangolare superiore non singolare e $S \in \mathbf{C}^{r\times(n-r)}$. Gli elementi diagonali di R_1 risultano positivi e ordinati in ordine di modulo non crescente. Dalla (11) si ottiene

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|R\Pi^T\mathbf{x} - \mathbf{c}\|_2, \quad \text{dove} \quad \mathbf{c} = Q^H\mathbf{b}.$$

Partizionando i vettori $\mathbf{y} = \Pi^T\mathbf{x}$ e $\mathbf{c}$ nel modo seguente

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \begin{matrix} \} & r \text{ componenti} \\ \} & n-r \text{ componenti} \end{matrix} \qquad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} \begin{matrix} \} & r \text{ componenti} \\ \} & m-r \text{ componenti} \end{matrix}$$

per la (12) risulta

$$R\Pi^T\mathbf{x} - \mathbf{c} = \begin{bmatrix} R_1\mathbf{y}_1 + S\mathbf{y}_2 - \mathbf{c}_1 \\ -\mathbf{c}_2 \end{bmatrix}.$$

Poiché per ogni vettore $\mathbf{y}_2 \in \mathbf{C}^{n-r}$, esiste un vettore $\mathbf{y}_1 \in \mathbf{C}^r$, tale che

$$R_1\mathbf{y}_1 = \mathbf{c}_1 - S\mathbf{y}_2,$$

la soluzione del problema

$$\min_{\substack{\mathbf{y}_1\in\mathbf{C}^r \\ \mathbf{y}_2\in\mathbf{C}^{n-r}}} \|R_1\mathbf{y}_1 + S\mathbf{y}_2 - \mathbf{c}_1\|$$

non è unica e risulta

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} R_1^{-1}(\mathbf{c}_1 - S\mathbf{y}_2) \\ \mathbf{y}_2 \end{bmatrix},$$

e

$$\mathbf{x} = \Pi \begin{bmatrix} \mathbf{y}_1 \\ \\ \mathbf{y}_2 \end{bmatrix}.$$

Si osservi che se la matrice S è nulla, la soluzione $\mathbf{x}^*$ di minima norma si ottiene ponendo $\mathbf{y}_2 = \mathbf{0}$, mentre ciò non è vero nel caso generale.

Dal punto di vista pratico si fissa una costante ϵ, che dipende dalla precisione con cui si eseguono i calcoli, e il procedimento termina quando tutte le colonne di $B^{(k)}$ hanno norma minore di ϵ.

7.6 Esempio. Applicando il metodo QR al problema dell'esempio 7.3, si ha

$$T^{(1)} = [A \mid \mathbf{b}] = \frac{1}{45} \begin{bmatrix} 6 & 12 & -72 & 45 \\ -16 & -7 & -8 & 45 \\ 58 & 16 & 104 & 45 \\ 87 & 24 & 156 & 45 \end{bmatrix},$$

e dopo 3 passi si ottiene la matrice

$$T^{(4)} = \begin{bmatrix} R_1 & S & \mathbf{c}_1 \\ O & O & \mathbf{c}_2 \end{bmatrix} = \frac{1}{9\sqrt{449}} \begin{bmatrix} -449 & -128 & -772 & -243 \\ 0 & -45 & 360 & -117 \\ 0 & 0 & 0 & 9\sqrt{449} \\ 0 & 0 & 0 & 9\sqrt{449} \end{bmatrix}.$$

Quindi la soluzione $\mathbf{x}$ non è unica e dipende da un parametro h. Se si pone uguale ad h la terza componente x_3 di $\mathbf{x}$, le altre due si ottengono risolvendo il sistema

$$\begin{bmatrix} -449 & -128 \\ 0 & -45 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -243 \\ -117 \end{bmatrix} - h \begin{bmatrix} -772 \\ 360 \end{bmatrix}.$$

Inoltre è

$$\mathbf{c}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e quindi $\gamma = \|\mathbf{c}_2\|_2 = \sqrt{2}$. ∎

La fattorizzazione QR può essere ricavata anche utilizzando il metodo di Givens, che può essere conveniente quando la matrice A è sparsa.

Il metodo QR consente infine di risolvere alcuni problemi come quello dell'esempio 7.4, in cui il sistema normale non può essere effettivamente risolto, a causa degli errori di rappresentazione degli elementi della matrice $A^H A$.

7.7 Esempio. Applicando il metodo QR al problema dell'esempio 7.4 si ha

$$T^{(1)} = \begin{bmatrix} 3 & 3 & 1 \\ 4 & 4 & 1 \\ 0 & \alpha & 1 \end{bmatrix},$$

e dopo 3 passi, operando con la precisione di macchina u, si ottiene

$$T^{(4)} = \begin{bmatrix} -5 & -5 & -\frac{7}{5} \\ 0 & -\alpha & -1 \\ 0 & 0 & \frac{1}{5} \end{bmatrix},$$

da cui si ricava

$$\mathbf{x}^* = \left[\frac{7}{25} - \frac{1}{\alpha}\,,\ \frac{1}{\alpha}\right]^T. \qquad \blacksquare$$

3. Norme di matrici non quadrate

Il concetto di norma può essere esteso anche a matrici non quadrate. In particolare, se $A \in \mathbf{C}^{m\times n}$, dove m e n sono interi qualsiasi, le norme matriciali indotte, considerate nel capitolo 3, vengono definite per mezzo della relazione

$$\|A\| = \max_{\|\mathbf{x}\|'=1} \|A\mathbf{x}\|'',$$

dove $\|\ .\ \|'$ e $\|\ .\ \|''$ sono norme vettoriali rispettivamente su $\mathbf{C}^n$ e su $\mathbf{C}^m$.

Si può dimostrare, in modo analogo a quanto fatto nel caso delle matrici quadrate, che nel caso in cui le due norme vettoriali coincidano con la norma 2, la norma matriciale indotta che si ottiene è data da

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

Anche per matrici non quadrate si definisce la norma di Frobenius di A nel modo seguente

$$\|A\|_F = \sqrt{\text{tr }(A^H A)}.$$

Inoltre se $U \in \mathbf{C}^{m\times m}$ e $V \in \mathbf{C}^{n\times n}$ sono matrici unitarie, poiché

$$(U^H AV)^H (U^H AV) = V^H A^H AV,$$

risulta

$$\|U^H AV\|_2 = \sqrt{\rho(A^H A)} = \|A\|_2$$

e

$$\|U^H AV\|_F = \sqrt{\text{tr } (A^H A)} = \|A\|_F.$$

4. Decomposizione ai valori singolari di una matrice

Lo studio della soluzione del problema dei minimi quadrati può essere condotto anche utilizzando la decomposizione ai valori singolari di una matrice.

7.8 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$. Allora esistono una matrice unitaria $U \in \mathbf{C}^{m\times m}$ e una matrice unitaria $V \in \mathbf{C}^{n\times n}$ tali che*

$$A = U\Sigma V^H, \tag{13}$$

dove la matrice $\Sigma \in \mathbf{R}^{m\times n}$ ha elementi σ_{ij} nulli per $i \neq j$ e per $i = j$ ha elementi $\sigma_{ii} = \sigma_i$, con

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0, \quad p = \min\{m, n\}.$$

La decomposizione (13) è detta *decomposizione ai valori singolari* della matrice A, mentre i valori σ_i, per $i = 1, \ldots, p$, sono detti i *valori singolari* di A. Indicate con $\mathbf{u}_i$, $i = 1, \ldots, m$, e $\mathbf{v}_i$, $i = 1, \ldots, n$, le colonne rispettivamente di U e di V, i vettori $\mathbf{u}_i$ e $\mathbf{v}_i$, $i = 1, \ldots, p$, sono detti rispettivamente *vettori singolari sinistri e vettori singolari destri* di A. La matrice Σ è univocamente determinata, anche se le matrici U e V non lo sono.

Dim. Si considera per semplicità il caso $m \geq n$ (se $m < n$, si sostituisce A con A^H). Si procede dimostrando per induzione su n che la tesi vale per ogni $m \geq n$.

Per $n = 1$ è $A = \mathbf{a} \in \mathbf{C}^m$. Si pone $\sigma_1 = \|\mathbf{a}\|_2$ e si considera come matrice U la matrice di Householder tale che $U\mathbf{a} = \sigma_1 \mathbf{e}_1$. La matrice V è la matrice $V = [1]$.

Per $n > 1$ si dimostra che se la tesi vale per le matrici di $\mathbf{C}^{k\times(n-1)}$, con $k \geq n-1$, allora vale per le matrici di $\mathbf{C}^{m\times n}$, con $m \geq n$. Sia $\mathbf{x} \in \mathbf{C}^n$, tale che $\|\mathbf{x}\|_2 = 1$ e $\|A\|_2 = \|A\mathbf{x}\|_2$. Si consideri il vettore

$$\mathbf{y} = \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} \in \mathbf{C}^m.$$

Allora $\|\mathbf{y}\|_2 = 1$ e $A\mathbf{x} = \sigma_1 \mathbf{y}$, con $\sigma_1 = \|A\|_2$. Siano poi $V_1 \in \mathbf{C}^{n\times n}$ e $U_1 \in \mathbf{C}^{m\times m}$ matrici unitarie le cui prime colonne sono uguali rispettivamente a $\mathbf{x}$ e $\mathbf{y}$. Poiché

$$U_1^H AV_1 \mathbf{e}_1 = U_1^H A\mathbf{x} = U_1^H \sigma_1 \mathbf{y} = \sigma_1 [1, 0, \ldots, 0]^T,$$

è

$$A_1 = U_1^H A V_1 = \begin{bmatrix} \sigma_1 & \mathbf{w}^H \\ \mathbf{0} & B \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & m-1 \text{ righe} \end{matrix}$$

in cui $\mathbf{w} \in \mathbf{C}^{n-1}$, $B \in \mathbf{C}^{(m-1)\times(n-1)}$ e $\mathbf{0} \in \mathbf{C}^{m-1}$. Si dimostra ora che $\mathbf{w} = \mathbf{0}$. Si supponga per assurdo che $\mathbf{w} \neq \mathbf{0}$ e si consideri il vettore $\mathbf{z} = \begin{bmatrix} \sigma_1 \\ \mathbf{w} \end{bmatrix} \neq \mathbf{0}$, per cui

$$A_1 \mathbf{z} = \begin{bmatrix} \sigma_1^2 + \mathbf{w}^H \mathbf{w} \\ B\mathbf{w} \end{bmatrix} = \begin{bmatrix} \|\mathbf{z}\|_2^2 \\ B\mathbf{w} \end{bmatrix}.$$

' Si ha

$$\|A_1 \mathbf{z}\|_2^2 = \|\mathbf{z}\|_2^4 + \|B\mathbf{w}\|_2^2 \geq \|\mathbf{z}\|_2^4,$$

da cui, dividendo per $\|\mathbf{z}\|_2^2$ si ottiene

$$\frac{\|A_1 \mathbf{z}\|_2^2}{\|\mathbf{z}\|_2^2} \geq \|\mathbf{z}\|_2^2,$$

e quindi

$$\|A_1\|_2^2 \geq \sigma_1^2 + \mathbf{w}^H \mathbf{w}. \tag{14}$$

D'altra parte è $\|A_1\|_2 = \|A\|_2$, in quanto A_1 è ottenuta da A con trasformazioni unitarie e quindi

$$\|A_1\|_2 = \sigma_1. \tag{15}$$

Dal confronto fra la (14) e la (15) segue l'assurdo. Quindi

$$A_1 = \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & B \end{bmatrix}.$$

Dalla (15) segue che

$$\sigma_1 \geq \|B\|_2. \tag{16}$$

Infatti

$$\sigma_1^2 = \|A_1\|_2^2 = \rho(A_1^H A_1) = \rho\Big(\begin{bmatrix} \sigma_1^2 & \mathbf{0}^H \\ \mathbf{0} & B^H B \end{bmatrix} \Big) = \max\left[\sigma_1^2, \rho(B^H B)\right]$$
$$\geq \rho(B^H B) = \|B\|_2^2.$$

Poiché $B \in \mathbf{C}^{(m-1)\times(n-1)}$ e $m-1 \geq n-1$, per l'ipotesi induttiva si ha

$$U_2^H B V_2 = \Sigma_2,$$

dove le matrici $U_2 \in \mathbf{C}^{(m-1)\times(m-1)}$ e $V_2 \in \mathbf{C}^{(n-1)\times(n-1)}$ sono unitarie e $\Sigma_2 \in \mathbf{R}^{(m-1)\times(n-1)}$ ha elementi $\sigma_2 \geq \ldots \geq \sigma_p$. Poiché $\sigma_2 = \|B\|_2 \leq \sigma_1$ per la (16), la tesi segue con

$$U = U_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & U_2 \end{bmatrix}, \quad V = V_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & V_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & \Sigma_2 \end{bmatrix}.$$

Le matrici U e V non sono univocamente determinate: infatti se

$$A = U \Sigma V^H$$

è una decomposizione ai valori singolari di A, e se $S \in \mathbf{C}^{n\times n}$ è una matrice di fase e $Z \in \mathbf{C}^{(m-n)\times(m-n)}$ è una matrice unitaria, anche

$$A = U \begin{bmatrix} S & O \\ O & Z \end{bmatrix} \Sigma S^H V^H$$

è una decomposizione ai valori singolari di A. Inoltre se $\sigma_i = \sigma_{i+1} = \ldots = \sigma_{i+j}$, per $j \geq 1$, detta P una qualunque matrice unitaria di ordine $j+1$ e considerata la matrice diagonale a blocchi

$$Q = \begin{bmatrix} I_{i-1} & O & O \\ O & P & O \\ O & O & I_{n-j-i} \end{bmatrix},$$

si ha che

$$A = U \begin{bmatrix} Q & O \\ O & I_{m-n} \end{bmatrix} \Sigma\, Q^H V^H$$

è una decomposizione ai valori singolari di A. ∎

Dal teorema 7.8 segue che

$$A = U \Sigma V^H = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^H \tag{17}$$

$$\left.\begin{aligned} A\mathbf{v}_i &= \sigma_i \mathbf{u}_i \\ A^H \mathbf{u}_i &= \sigma_i \mathbf{v}_i \end{aligned}\right\}, \quad i = 1, \ldots, p.$$

7.9 Esempio. La matrice A dell'esempio 7.2 ha la decomposizione ai valori singolari

$$A = U\Sigma V^H,$$

dove

$$U = \frac{1}{15}\begin{bmatrix} 2 & -4 & 14 & 3 \\ 8 & -1 & -4 & 12 \\ 11 & 8 & 2 & -6 \\ -6 & 12 & 3 & 6 \end{bmatrix}, \quad V = \frac{1}{9}\begin{bmatrix} -4 & 4 & 7 \\ 8 & 1 & 4 \\ 1 & 8 & -4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

La matrice A dell'esempio 7.3 ha la decomposizione ai valori singolari

$$A = U\Sigma V^H,$$

dove

$$U = \frac{1}{15}\begin{bmatrix} -4 & 14 & 2 & 3 \\ -1 & -4 & 8 & 12 \\ 8 & 2 & 11 & -6 \\ 12 & 3 & -6 & 6 \end{bmatrix}, \quad V = \frac{1}{9}\begin{bmatrix} 4 & 7 & -4 \\ 1 & 4 & 8 \\ 8 & -4 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

■

Dal teorema 7.8 segue il

7.10 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$ e sia*

$$A = U\Sigma V^H$$

la sua decomposizione ai valori singolari, dove

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_p = 0.$$

Allora valgono le seguenti proprietà

a) $$A = U_k \Sigma_k V_k^H = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \quad \textit{dove}$$

$U_k \in \mathbf{C}^{m\times k}$ è la matrice le cui colonne sono $\mathbf{u}_1, \ldots, \mathbf{u}_k$,

$V_k \in \mathbf{C}^{n\times k}$ è la matrice le cui colonne sono $\mathbf{v}_1, \ldots, \mathbf{v}_k$,

$\Sigma_k \in \mathbf{R}^{k\times k}$ è la matrice diagonale i cui elementi principali sono

$$\sigma_1, \ldots, \sigma_k.$$

b) Il nucleo di A è generato dai vettori $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$.

c) L'immagine di A è generata dai vettori $\mathbf{u}_1, \ldots, \mathbf{u}_k$, e quindi

$$\textit{rango di } A = k.$$

d) σ_i^2, $i = 1, \ldots, p$, sono gli autovalori della matrice $A^H A$ (se $m < n$ i restanti autovalori sono nulli) e quindi

$$\|A\|_F^2 = \sum_{i=1}^{k} \sigma_i^2,$$

$$\|A\|_2 = \sigma_1.$$

e) Se $m = n$ e A è normale, allora $\sigma_i = |\lambda_i|$, $i = 1, \ldots, n$, dove i λ_i sono gli autovalori di A, e i vettori singolari destri e sinistri coincidono con gli autovettori di A.

Dim. Si supponga per semplicità che $p = n \leq m$ (se $n > m$ si sostituisce A con A^H).

a) La matrice Σ ha la forma

$$\Sigma = \begin{bmatrix} \Sigma_k & O \\ O & O \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & m-k \text{ righe} \end{matrix}$$

per cui, partizionando le matrici U e V nel modo seguente

$$U = [U_k \mid U'_{m-k}], \qquad V = [V_k \mid V'_{n-k}],$$

dalla (17) risulta che

$$A = U_k \Sigma_k V_k^H. \tag{18}$$

b) Se $\mathbf{x} \in \mathbf{C}^n$, la condizione $A\mathbf{x} = \mathbf{0}$ per la (17) è equivalente alla condizione

$$U \Sigma V^H \mathbf{x} = \mathbf{0},$$

e, poiché U è non singolare, è equivalente a

$$\Sigma V^H \mathbf{x} = \mathbf{0}. \tag{19}$$

Il vettore $\mathbf{z} = \Sigma V^H \mathbf{x}$ può essere partizionato nel modo seguente

$$\mathbf{z} = \begin{bmatrix} \Sigma_k V_k^H \mathbf{x} \\ \mathbf{0} \end{bmatrix} \begin{matrix} \} & k \text{ componenti,} \\ \} & m-k \text{ componenti,} \end{matrix} \tag{20}$$

per cui la (19) può essere scritta come $\Sigma_k V_k^H \mathbf{x} = \mathbf{0}$, ossia $V_k^H \mathbf{x} = \mathbf{0}$. Quindi $A\mathbf{x} = \mathbf{0}$ se e solo se $\mathbf{x}$ è ortogonale alle prime k colonne di V, ed essendo V unitaria, se e solo se $\mathbf{x}$ è generato dalle restanti colonne di V.

c) Dalla (18) si ha

$$\mathbf{y} = A\mathbf{x} = U_k \Sigma_k V_k^H \mathbf{x} = U_k \mathbf{z}, \tag{21}$$

dove $\mathbf{z} = \Sigma_k V_k^H \mathbf{x} \in \mathbf{C}^k$. Quindi $\mathbf{y}$ è generato dalle colonne di U_k. Viceversa dalla (20) si ha che, poiché la matrice $\Sigma_k V_k^H$ è di rango massimo, per ogni $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$, esiste uno $\mathbf{z} \neq \mathbf{0}$ per cui vale la (21).

d) Dalla (17) si ha che

$$A^H A = V \Sigma^H \Sigma V^H,$$

dove $\Sigma^H \Sigma \in \mathbf{R}^{n \times n}$ è la matrice diagonale i cui elementi principali sono $\sigma_1^2, \ldots, \sigma_p^2$. Poiché la traccia e il raggio spettrale di due matrici simili sono uguali, si ha

$$\|A\|_F^2 = \text{tr } (A^H A) = \sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{k} \sigma_i^2$$

e

$$\|A\|_2^2 = \rho(A^H A) = \sigma_1^2,$$

e poiché $\sigma_1 > 0$, risulta $\|A\|_2 = \sigma_1$.

e) Se A è normale, dalla forma normale di Schur di A

$$A = UDU^H,$$

segue che

$$A^H A = UD^H DU^H,$$

perciò gli autovalori σ_i^2 di $A^H A$ sono tali che

$$\sigma_i^2 = \overline{\lambda}_i \lambda_i = |\lambda_i|^2, \quad \text{per } i = 1, \ldots, n.$$

∎

7.11 Esempio. La matrice A dell'esempio 7.2 ha rango 3 (infatti la matrice $A^H A$ è non singolare); come risulta dall'esempio 7.9 i suoi valori singolari sono $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$. Per il punto d) del teorema 7.10 risulta

$$\|A\|_F = \sqrt{14}, \qquad \|A\|_2 = 3.$$

La matrice A dell'esempio 7.3 ha rango 2: infatti $\sigma_1 = 5, \sigma_2 = 1, \sigma_3 = 0$, come risulta dall'esempio 7.9. Per il punto d) del teorema 7.10 risulta

$$\|A\|_F = \sqrt{26}, \qquad \|A\|_2 = 1,$$

e l'insieme degli $\mathbf{x} \in \mathbf{R}^3$ tali che $A\mathbf{x} = \mathbf{0}$ è generato dal vettore

$$\mathbf{v}_3 = \frac{1}{9}\,[\text{-}4,\ 8,\ 1]^T. \qquad \blacksquare$$

7.12 Esempio. La matrice hermitiana

$$A = \frac{1}{81}\begin{bmatrix} \text{-}65 & 76 & 104 \\ 76 & \text{-}206 & 8 \\ 104 & 8 & 109 \end{bmatrix}$$

è tale che $A = UDU^H$, dove U, matrice unitaria, e D, matrice diagonale, sono

$$U = \frac{1}{9}\begin{bmatrix} \text{-}4 & 4 & 7 \\ 8 & 1 & 4 \\ 1 & 8 & \text{-}4 \end{bmatrix}, \quad D = \begin{bmatrix} \text{-}3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \text{-}1 \end{bmatrix}.$$

Poiché gli autovalori di A sono $\lambda_1 = \text{-}3, \lambda_2 = 2, \lambda_3 = \text{-}1$, i valori singolari di A sono $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$ e la decomposizione ai valori singolari di A è data da

$$A = U\Sigma V^H,$$

dove

$$V = \frac{1}{9}\begin{bmatrix} 4 & 4 & \text{-}7 \\ \text{-}8 & 1 & \text{-}4 \\ \text{-}1 & 8 & 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$\blacksquare$

Dal teorema 7.10 si può ricavare anche un procedimento per calcolare i valori e i vettori singolari di A. Per semplicità si suppone $m \geq n$ (se fosse $m < n$ basta riferirsi alla matrice A^H). Questo procedimento si articola nei seguenti passi:

a) si calcolano gli autovalori e gli autovettori, normalizzati in norma 2, della matrice $A^H A$ e si considera la seguente decomposizione in forma normale di Schur della matrice $A^H A$

$$A^H A = QDQ^H, \quad D \in \mathbf{R}^{n\times n},\ Q \in \mathbf{C}^{n\times n}, \tag{22}$$

in cui gli elementi principali di D sono gli autovalori in ordine non crescente di $A^H A$ e Q è la corrispondente matrice degli autovettori (un

metodo stabile per calcolare la decomposizione (22) sarà descritto nel paragrafo 9);

b) si calcola la matrice

$$C = AQ \in \mathbf{C}^{m\times n}, \tag{23}$$

e si determina, utilizzando la tecnica del massimo pivot per colonne, la fattorizzazione QR della matrice

$$C\Pi = UR = U \begin{bmatrix} R_1 \\ \\ O \end{bmatrix}, \tag{24}$$

dove $\Pi \in \mathbf{R}^{n\times n}$ è una matrice di permutazione, $U \in \mathbf{C}^{m\times m}$ è una matrice unitaria ed $R_1 \in \mathbf{C}^{n\times n}$ è una matrice triangolare superiore con gli elementi principali reali non negativi e ordinati in modo non crescente. Le condizioni imposte sull'ordinamento degli elementi principali di R_1 rendono unica questa fattorizzazione se gli elementi principali di R_1 sono tutti distinti.

Da (23) e (24) si ottiene

$$A = UR\Pi^T Q^H, \tag{25}$$

da cui

$$A^H A = Q\Pi R^H U^H UR\Pi^T Q^H = Q\Pi R^H R\Pi^T Q^H = Q\Pi R_1^H R_1 \Pi^T Q^H,$$

e quindi per la (22) risulta

$$R_1^H R_1 = \Pi^T D\Pi.$$

Poiché la matrice $\Pi^T D\Pi$ risulta essere diagonale, ne segue che $R_1^H R_1$ è diagonale, e quindi R_1 non può che essere diagonale. Inoltre, poiché gli elementi principali di R_1 e di D sono ordinati in modo non crescente, se gli autovalori di $A^H A$ sono tutti distinti è $\Pi = I$. Quindi la (25), se si pone $\Sigma = R$ e $V = Q\Pi$, rappresenta la decomposizione ai valori singolari di A.

La decomposizione ai valori singolari di una matrice consente anche di risolvere il seguente problema di minimo: data una matrice $A \in \mathbf{C}^{m\times n}$ di rango k, e fissato un intero $r < k$, qual è la matrice $B \in \mathbf{C}^{m\times n}$ di rango r più "vicina" ad A? Vale infatti il seguente

7.13 Teorema. *Sia* $A \in \mathbf{C}^{m\times n}$ *e sia*

$$A = U\Sigma V^H$$

la decomposizione ai valori singolari di A*, dove*

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_p = 0,$$

e sia r *un intero positivo minore o uguale a* k*. Indicando con*

$$A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H$$

e con

$$S = \{\ B \in \mathbf{C}^{m\times n}\ :\ \textit{rango di } B = r\ \},$$

si ha

$$\min_{B\in S} \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}.$$

Dim. Sia $\Sigma_r \in \mathbf{R}^{m\times n}$ la matrice i cui elementi sono $\sigma_{ii} = \sigma_i$ per $i = 1, \ldots, r$ e $\sigma_{ij} = 0$ altrimenti. Allora vale

$$U^H A_r V = \Sigma_r$$

e quindi per il punto c) del teorema 7.10 è rango di $A_r = r$. Risulta inoltre

$$\|A - A_r\|_2 = \|U^H(A - A_r)V\|_2 = \|\Sigma - \Sigma_r\|_2 = \sigma_{r+1}, \tag{26}$$

in quanto σ_{r+1} è il massimo degli elementi non nulli di $\Sigma - \Sigma_r$. Sia $B \in S$. Il nucleo di B ha dimensione $n - r$ perché B ha rango r. Poiché l'intersezione fra $N(B)$ e il sottospazio T di $\mathbf{C}^n$ generato dai vettori $\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}$ non può ridursi al solo vettore nullo, in quanto dim $T+$ dim $N(B) = n + 1$, esiste un elemento $\mathbf{z} \in N(B) \cap T$, $\mathbf{z} \neq \mathbf{0}$. Si supponga che $\|\mathbf{z}\|_2 = 1$; essendo $\mathbf{z}$ elemento di T, si può scrivere

$$\mathbf{z} = \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_j. \tag{27}$$

Per il punto a) del teorema 7.10 si ha

$$A\mathbf{z} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}) = \sum_{i=1}^{r+1} \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}), \tag{28}$$

in quanto, essendo $\mathbf{z} \in T$ e V unitaria, è $\mathbf{v}_i^H \mathbf{z} = 0$ per $i = r+2, \ldots, k$. Poiché $\mathbf{z} \in N(B)$, è $B\mathbf{z} = \mathbf{0}$ e si ha

$$\|A - B\|_2^2 \geq \|(A - B)\mathbf{z}\|_2^2 = \|A\mathbf{z}\|_2^2. \tag{29}$$

D'altra parte per la (28), poiché i vettori $\mathbf{u}_i$, $i = 1, \ldots, r+1$, sono ortonormali, si ha

$$\|A\mathbf{z}\|_2^2 = \sum_{i=1}^{r+1} \sigma_i^2 |\mathbf{v}_i^H \mathbf{z}|^2.$$

Poiché $\sigma_i^2 \geq \sigma_{r+1}^2$, $i = 1, \ldots, r+1$, si ha

$$\|A\mathbf{z}\|_2^2 \geq \sigma_{r+1}^2 \sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2. \tag{30}$$

Per la (27) è

$$\begin{aligned} \sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2 = \sum_{i=1}^{r+1} \left| \mathbf{v}_i^H \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} \left| \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_i^H \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} |\alpha_i|^2 \\ = \|\mathbf{z}\|_2^2 = 1, \end{aligned}$$

e quindi dalla (30) segue

$$\|A\mathbf{z}\|_2 \geq \sigma_{r+1}. \tag{31}$$

Confrontando la (31) e la (29) si ha che

$$\|A - B\|_2 \geq \sigma_{r+1},$$

e poiché per la (26) è $\|A - A_r\|_2 = \sigma_{r+1}$, ne segue che

$$\min_{B \in S} \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}. \qquad \blacksquare$$

L'importanza del teorema 7.13 risiede nel fatto che esso consente di quantificare esattamente, tramite il valore singolare σ_{r+1}, la distanza in norma 2 della matrice A dalla "più vicina" matrice di rango r, e quindi di stimare l'errore che si commette quando la matrice A, a seguito di operazioni eseguite in aritmetica finita, viene sostituita con una matrice di rango r. Questa stima non è ottenibile facilmente utilizzando le fattorizzazioni LU e QR, che hanno lo scopo di trasformare la matrice A in una matrice della forma

$$\begin{bmatrix} B \\ \\ O \end{bmatrix},$$

in cui B è triangolare superiore: il rango di A viene ricavato mediante il numero di elementi principali di B che sono diversi da zero a meno di una precisione prefissata. Questo modo di procedere non garantisce assolutamente un buon risultato perché è molto difficile stimare bene il rango di una matrice triangolare. Si esamini a questo proposito il seguente esempio dovuto a Wilkinson.

7.14 Esempio. Si consideri la matrice triangolare superiore B di ordine n i cui elementi sono

$$b_{ij} = \begin{cases} 1 & \text{se } i = j, \\ -1 & \text{se } i < j, \\ 0 & \text{se } i > j \end{cases}$$

(la matrice B^T è stata usata nell'esercizio 4.19). La matrice ha rango n. Se però l'elemento di indici $(n, 1)$ viene perturbato della quantità $\epsilon = -2^{2-n}$, la matrice così ottenuta ha rango $n - 1$. Quindi una piccola perturbazione introdotta su un elemento non principale altera il rango della matrice triangolare B. Infatti se si calcolano i valori singolari di B si ottiene:

n	σ_{n-1}	σ_n
5	1.509442	$0.9298509\ 10^{-1}$
10	1.502146	$0.2929687\ 10^{-2}$
15	1.500909	$0.9170198\ 10^{-4}$
20	1.500494	$0.2969867\ 10^{-5}$

Quindi al crescere di n la matrice B è sempre più vicina ad una matrice di rango $n - 1$, anche se questo non appare evidente dagli elementi principali, che sono gli autovalori di B. ∎

5. Risoluzione del problema dei minimi quadrati con i valori singolari

Utilizzando il teorema 7.10 è possibile dare una formulazione esplicita della soluzione $\mathbf{x}^*$ di minima norma del problema dei minimi quadrati e del corrispondente γ, anche nel caso in cui la matrice A non sia di rango massimo.

7.15 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$ di rango k, con $m \geq n \geq k$, e sia*

$$A = U\Sigma V^H$$

la decomposizione ai valori singolari di A. Allora la soluzione di minima norma del problema (2) è data da

$$\mathbf{x}^* = \sum_{i=1}^{k} \frac{\mathbf{u}_i^H \mathbf{b}}{\sigma_i} \mathbf{v}_i$$

e

$$\gamma^2 = \sum_{i=k+1}^{m} |\mathbf{u}_i^H \mathbf{b}|^2.$$

Dim. Poiché la norma 2 è invariante per trasformazioni unitarie, si ha

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|U^H(A\mathbf{x} - \mathbf{b})\|_2^2 = \|U^H AVV^H\mathbf{x} - U^H\mathbf{b}\|_2^2,$$

e posto $\mathbf{y} = V^H\mathbf{x}$, si ha

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|\Sigma\mathbf{y} - U^H\mathbf{b}\|_2^2 = \sum_{i=1}^{n} |\sigma_i y_i - \mathbf{u}_i^H\mathbf{b}|^2 + \sum_{i=n+1}^{m} |\mathbf{u}_i^H\mathbf{b}|^2$$

$$= \sum_{i=1}^{k} |\sigma_i y_i - \mathbf{u}_i^H\mathbf{b}|^2 + \sum_{i=k+1}^{m} |\mathbf{u}_i^H\mathbf{b}|^2, \tag{32}$$

dove y_i , $i = 1, \ldots, n$, sono le componenti di $\mathbf{y}$. Il minimo della (32) viene raggiunto per

$$y_i = \frac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}, \quad i = 1, \ldots, k. \tag{33}$$

Fra tutti i vettori $\mathbf{y} \in \mathbf{C}^n$ per cui vale la (33), il vettore di minima norma $\mathbf{y}^*$ è quello per cui

$$y_i^* = \begin{cases} \dfrac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}, & \text{per } i = 1, \ldots, k, \\ 0, & \text{per } i = k+1, \ldots, n. \end{cases}$$

Poiché $\mathbf{x} = V\mathbf{y}$, è $\|\mathbf{x}^*\|_2 = \|\mathbf{y}^*\|_2$ e quindi

$$\mathbf{x}^* = V\mathbf{y}^* = \sum_{i=1}^{k} y_i^*\mathbf{v}_i = \sum_{i=1}^{k} \frac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}\ \mathbf{v}_i,$$

e dalla (32) risulta

$$\gamma^2 = \|A\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=k+1}^{m} |\mathbf{u}_i^H \mathbf{b}|^2. \qquad \blacksquare$$

7.16 Esempio. Dalla decomposizione ai valori singolari della matrice A dell'esempio 7.2

$$A = U\Sigma V^H,$$

dove U, V e Σ sono quelle riportate nell'esempio 7.9, si ha

$$\mathbf{x}^* = \frac{\mathbf{u}_1^H \mathbf{b}}{\sigma_1}\mathbf{v}_1 + \frac{\mathbf{u}_2^H \mathbf{b}}{\sigma_2}\mathbf{v}_2 + \frac{\mathbf{u}_3^H \mathbf{b}}{\sigma_3}\mathbf{v}_3,$$

e poiché $\mathbf{u}_i^H \mathbf{b} = 1$, per $i = 1, \ldots, 4$, risulta

$$\mathbf{x}^* = \frac{1}{3}\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2 + \mathbf{v}_3 = \frac{1}{54}\,[46,\ 43,\ 2]^T,$$

e $\quad \gamma = |\mathbf{u}_4^H \mathbf{b}| = 1.$

Dalla decomposizione ai valori singolari della matrice A dell'esempio 7.3

$$A = U\Sigma V^H,$$

dove U, V e Σ sono quelle riportate nell'esempio 7.9, si ha

$$\mathbf{x}^* = \frac{\mathbf{u}_1^H \mathbf{b}}{\sigma_1}\mathbf{v}_1 + \frac{\mathbf{u}_2^H \mathbf{b}}{\sigma_2}\mathbf{v}_2,$$

e poiché $\mathbf{u}_i^H \mathbf{b} = 1$, per $i = 1, \ldots, 4$, risulta

$$\mathbf{x}^* = \frac{1}{5}\mathbf{v}_1 + \mathbf{v}_2 = \frac{1}{15}\,[13,\ 7,\ \text{-}4]^T,$$

e $\quad \gamma = \sqrt{|\mathbf{u}_3^H \mathbf{b}|^2 + |\mathbf{u}_4^H \mathbf{b}|^2} = \sqrt{2}.$ $\blacksquare$

6. Pseudoinversa di Moore-Penrose

Se la matrice A è quadrata e non singolare, la soluzione del sistema (1) e del problema dei minimi quadrati (2) coincidono e possono essere espresse nella forma

$$\mathbf{x}^* = A^{-1}\mathbf{b},$$

per mezzo della matrice inversa A^{-1}. Il concetto di matrice inversa può essere esteso anche al caso di matrici A per cui A^{-1} non esiste. In questo caso si definisce una matrice pseudoinversa di A, indicata con il simbolo A^+, che consente di scrivere la soluzione di minima norma del problema (2) nella forma

$$\mathbf{x}^* = A^+\mathbf{b}.$$

7.17 Definizione. Sia $A \in \mathbf{C}^{m\times n}$ una matrice di rango k. La matrice $A^+ \in \mathbf{C}^{n\times m}$ tale che

$$A^+ = V\Sigma^+U^H,$$

dove $\Sigma^+ \in \mathbf{R}^{n\times m}$ è la matrice che ha elementi σ_{ij} nulli per $i \neq j$ e per $i = j$ ha elementi

$$\sigma_{ii}^+ = \begin{cases} \dfrac{1}{\sigma_i}, & \text{per } i = 1, \ldots, k, \\ 0, & \text{per } i = k+1, \ldots, p, \end{cases}$$

è detta *pseudoinversa di Moore-Penrose* di A. ∎

Valgono le seguenti proprietà (si veda l'esercizio 7.11):

a) La matrice $X = A^+$ è l'unica matrice di $\mathbf{C}^{n\times m}$ che soddisfa alle seguenti *equazioni di Moore-Penrose*:
   1) $AXA = A$,
   2) $XAX = X$,
   3) $(AX)^H = AX$,
   4) $(XA)^H = XA$.

b) Se il rango di A è massimo, allora

$$\begin{array}{ll} \text{se } m \geq n, & A^+ = (A^HA)^{-1}A^H, \\ \text{se } m \leq n, & A^+ = A^H(AA^H)^{-1}, \\ \text{se } m = n = \text{ rango di } A, & A^+ = A^{-1}. \end{array}$$

È immediato verificare che per il teorema 7.15 risulta

$$\mathbf{x}^* = A^+\mathbf{b}$$

e

$$\gamma = \|(I - AA^+)\mathbf{b}\|_2.$$

Inoltre A^+ è la soluzione dei seguenti problemi (si veda l'esercizio 7.29)

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_2,$$

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_F.$$

Utilizzando la matrice A^+ è anche possibile estendere il concetto di condizionamento alle matrici quadrate singolari e alle matrici non quadrate.

7.18 Definizione. Sia $A \in \mathbf{C}^{m\times n}$ una matrice di rango k. Si definisce *numero di condizionamento* di A il numero

$$\mu(A) = \|A\| \, \|A^+\|$$

dove $\| \, . \, \|$ è una qualsiasi norma matriciale. ∎

Si osservi che se la norma usata è la norma 2, per il teorema 7.10 d) è

$$\mu_2(A) = \frac{\sigma_1}{\sigma_k}, \tag{34}$$

ed inoltre, poiché la matrice $(A^H A)^+$ ha per valori singolari non nulli le quantità

$$\frac{1}{\sigma_i^2}, \quad i = 1, \ldots, k,$$

è

$$\mu_2(A^H A) = \frac{\sigma_1^2}{\sigma_k^2}, \tag{35}$$

Confrontando le (34) e (35), ne segue che

$$\mu_2(A^H A) = [\mu_2(A)]^2. \tag{36}$$

L'importanza della matrice pseudoinversa A^+ di A è essenzialmente di carattere teorico, in quanto il calcolo della pseudoinversa di una matrice può risultare molto instabile: infatti, se una piccola perturbazione degli elementi di A modifica il rango della matrice, si può generare una grossa perturbazione degli elementi di A^+.

7.19 Esempio. Siano

$$A = \frac{1}{15}\begin{bmatrix} 6 & -8 \\ 6 & -8 \\ 3 & -4 \end{bmatrix}, \quad \delta A = \frac{\epsilon}{15}\begin{bmatrix} 4 & 3 \\ -8 & -6 \\ 8 & 6 \end{bmatrix}, \quad \epsilon \neq 0.$$

La decomposizione ai valori singolari di A è

$$A = U\Sigma V^T, \quad U = \frac{1}{3}\begin{bmatrix} 2 & 1 & -2 \\ 2 & -2 & 1 \\ 1 & 2 & 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad V = \frac{1}{5}\begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix},$$

e quindi la matrice A ha rango 1 e risulta

$$A^+ = V\Sigma^+U^T = V\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}U^T = \frac{1}{15}\begin{bmatrix}6 & 6 & 3\\ -8 & -8 & -4\end{bmatrix}.$$

La decomposizione ai valori singolari di $A+\delta A$ è

$$A+\delta A = U\Sigma' V^T, \quad \Sigma' = \begin{bmatrix}1 & 0\\ 0 & \epsilon\\ 0 & 0\end{bmatrix}, \quad U \text{ e } V \text{ come sopra,}$$

e quindi la matrice $A+\delta A$ ha rango 2 e risulta

$$(A+\delta A)^+ = V\Sigma'^+U^T = V\begin{bmatrix}1 & 0 & 0\\ 0 & 1/\epsilon & 0\end{bmatrix}U^T = A^+ + \frac{1}{15\epsilon}\begin{bmatrix}4 & -8 & 8\\ 3 & -6 & 6\end{bmatrix}.$$

La perturbazione δA è tale che $\|\delta A\|_2 = |\epsilon|$ e genera sugli elementi di A^+ una perturbazione la cui norma 2 è

$$\|(A+\delta A)^+ - A^+\|_2 = \frac{1}{15|\epsilon|}\left\|\begin{bmatrix}4 & -8 & 8\\ 3 & -6 & 6\end{bmatrix}\right\|_2 = \frac{1}{|\epsilon|}.$$

Se per esempio $\epsilon = 10^{-6}$, una perturbazione δA tale che $\|\delta A\|_2 = 10^{-6}$ produce una perturbazione sugli elementi della pseudoinversa tale che $\|(A+\delta A)^+ - A^+\|_2 = 10^6$. ∎

Nel caso in cui il rango della matrice $A+\delta A$ non sia diverso da quello di A, la matrice A^+ risulta affetta da una perturbazione assai minore che nel caso precedente. Vale infatti il seguente teorema, per la cui dimostrazione si veda [12].

7.20 Teorema. *Siano* $A,\ \delta A \in \mathbf{C}^{m\times n}$, *tali che* $\|A^+\|_2\|\delta A\|_2 < 1$ *e rango di* $(A+\delta A) \le$ *rango di* A. *Allora*

$$\textit{rango di } (A+\delta A) = \textit{ rango di } A$$

e

$$\frac{\|(A+\delta A)^+ - A^+\|_2}{\|A^+\|_2} \le \alpha\frac{\mu_2(A)\epsilon_A}{1-\mu_2(A)\epsilon_A},$$

dove $\epsilon_A = \dfrac{\|\delta A\|_2}{\|A\|_2}$ *e* α *è una costante positiva minore di 2.*

∎

7. Condizionamento del problema dei minimi quadrati

Nel caso in cui la matrice A sia di rango massimo, vale il seguente teorema di perturbazione del problema dei minimi quadrati (per la dimostrazione si veda [12]).

7.21 Teorema. *Siano $m \geq n$, $A \in \mathbf{C}^{m\times n}$ una matrice di rango massimo, $\delta A \in \mathbf{C}^{m\times n}$, tale che $\|A^+\|\,\|\delta A\| < 1, \mathbf{b} \in \mathbf{C}^m, \mathbf{b} \neq \mathbf{0}$ e $\delta\mathbf{b} \in \mathbf{C}^m$. Allora la matrice $A+\delta A$ è ancora di rango massimo e, indicata con $\mathbf{x}+\delta\mathbf{x}$ la soluzione del problema dei minimi quadrati perturbato*

$$\min_{\mathbf{y}\in\mathbf{C}^n} \|(A+\delta A)\mathbf{y} - (\mathbf{b}+\delta\mathbf{b})\|_2,$$

risulta

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\mu_2(A)}{1-\epsilon_A\mu_2(A)}\left[\left(1+\mu_2(A)\frac{\gamma}{\|A\|_2\|\mathbf{x}\|_2}\right)\epsilon_A + \frac{\|\mathbf{b}\|_2}{\|A\|_2\|\mathbf{x}\|_2}\,\epsilon_b\right], \quad (37)$$

dove γ è il residuo del problema definito in (2) e

$$\epsilon_A = \frac{\|\delta A\|_2}{\|A\|_2}, \qquad \epsilon_b = \frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}. \qquad \blacksquare$$

Questo teorema consente di trarre alcune indicazioni riguardo al metodo da scegliere per risolvere il problema dei minimi quadrati, in modo da non pregiudicare l'accuratezza della soluzione. Quando la matrice A ha rango massimo con il metodo QR, effettuata la fattorizzazione (6), si risolve il sistema (9), la cui matrice R_1 è tale che

$$\mu_2(R_1) = \mu_2(R) = \mu_2(A),$$

cioè la stabilità del metodo è legata a $\mu_2(A)$. Se invece la soluzione del problema (2) viene ottenuta attraverso il sistema delle equazioni normali, la cui matrice è A^HA, la stabilità del metodo è legata a $\mu_2(A^HA)$, che per la (36) è uguale a $[\mu_2(A)]^2$.

D'altra parte dalla (37), che si può anche assumere come maggiorazione dell'errore inerente, segue che l'errore ϵ_A di perturbazione della matrice A è amplificato dal fattore

$$\frac{c_A}{1-\epsilon_A\mu_2(A)}, \quad \text{dove } c_A = \mu_2(A) + [\mu_2(A)]^2\frac{\gamma}{\|A\|_2\|\mathbf{x}\|_2}.$$

Se il residuo γ è così piccolo, che il secondo termine di c_A risulta minore del primo, allora la maggiorazione dell'errore inerente è dominata da $\mu_2(A)$. In questo caso non è conveniente ricorrere alla risoluzione del sistema normale, il cui condizionamento è $[\mu_2(A)]^2$, e conviene usare il metodo QR. Se invece il residuo γ è tale che il secondo termine di c_A risulta dominante, la maggiorazione dell'errore inerente è dominata dal termine $[\mu_2(A)]^2$. Allora dal punto di vista della stabilità, il metodo QR e i metodi basati sulla risoluzione del sistema normale, sono equivalenti. In questo caso, e nel caso in cui non sia possibile determinare a priori se per un dato problema di

minimi quadrati il residuo γ è piccolo oppure no, per la scelta del metodo occorre tenere presente:

a) il costo computazionale, che in generale è maggiore per il metodo QR;

b) eventuali proprietà di struttura delle matrici A e $A^H A$, per esempio eventuale sparsità della matrice A che non si trasmette alla matrice $A^H A$, per cui il costo computazionale del metodo QR è minore;

c) la disponibilità di memoria centrale del calcolatore che, quando m sia molto più grande di n, può essere sufficiente per contenere la matrice $A^H A$ ma non la matrice A;

d) la precisione di macchina, che può essere sufficiente a rappresentare gli elementi della matrice A ma non quelli della matrice $A^H A$ (si veda l'esempio 7.4, in cui nel calcolo di $A^H A$ si perdono informazioni fondamentali). Inoltre gli elementi di $A^H A$ potrebbero non essere rappresentabili a causa di overflow o di underflow.

7.22 Esempio. Sia $A \in \mathbf{R}^{m\times n}$ la matrice con elementi

$$a_{ij} = \lambda_i^{j-1}, \quad j = 1, \ldots, n,$$

dove i numeri $\lambda_i,\ i = 1, \ldots, m$ sono a due a due distinti. Questa matrice, che interviene nell'interpolazione di funzioni, ha rango massimo e se $m = n$ è detta matrice di *Vandermonde*.

Gli elementi della matrice $C = A^T A$ sono

$$c_{kj} = \sum_{i=1}^{m} \lambda_i^{k+j-2}, \quad k,\ j = 1, \ldots, n.$$

Tale matrice appartiene alla classe delle matrici di *Hankel*, i cui elementi sono definiti da $2n-1$ parametri $\alpha_k,\ k = 1, \ldots, 2n-1$, nel modo seguente

$$h_{ij} = \alpha_{i+j-1},\ i = 1, \ldots, n,\ j = 1, \ldots, n$$

(la matrice dell'esempio 4.21 è una matrice di Hankel). In questo caso per risolvere il problema (2) è conveniente, dal punto di vista del costo computazionale, passare attraverso la risoluzione del sistema normale, perché la costruzione della matrice $A^T A$ e del vettore $A^T \mathbf{b}$ richiede un numero di operazioni moltiplicative dell'ordine di $2mn$, ed inoltre esistono metodi per risolvere sistemi lineari con matrici di Hankel che hanno un costo computazionale dell'ordine di $n \log_2^2 n$.

Se però i numeri λ_i sono troppo piccoli o troppo grandi, è possibile che gli elementi c_{kj} non siano rappresentabili a causa di overflow o di underflow. In tal caso è necessario ricorrere al metodo QR. ∎

Se la matrice A non ha rango massimo, allora la risoluzione del problema (2), e in particolare il calcolo della soluzione di minima norma, è molto più delicata. La difficoltà più grossa si incontra nella determinazione del rango della matrice A. Quando la matrice A è fortemente mal condizionata, solo la determinazione dei valori singolari di A consente un'effettiva comprensione della natura del problema. Infatti:

a) se la matrice A e il vettore $\mathbf{b}$ sono ottenuti da dati sperimentali, per l'effetto congiunto dell'incertezza sui dati e del mal condizionamento della matrice si possono ottenere più soluzioni del problema (2) che, pur essendo molto diverse fra di loro, hanno tutte norma molto vicina alla norma minima e quindi sono tutte accettabili come approssimazioni della soluzione di minima norma.

b) Se il malcondizionamento della matrice A è generato dalla presenza di un certo numero di valori singolari vicini fra di loro e molto più piccoli degli altri, può essere utile considerare il problema approssimato che si ottiene eliminando i valori singolari più piccoli ed esprimere la soluzione utilizzando solo i valori singolari più grossi e i corrispondenti vettori singolari. Il valore del residuo γ che si ottiene operando in questo modo può dare una misura di quanto la soluzione calcolata è accettabile.

c) il calcolo dei valori e dei vettori singolari di una matrice è un problema ben posto, come si vedrà nel prossimo paragrafo, e il metodo per calcolare i valori e i vettori singolari che fa uso dell'algoritmo di Golub e Reinsch, descritto nel paragrafo 9 è stabile [9].

È opportuno rilevare però che il costo computazionale del calcolo dei valori e dei vettori singolari risulta in generale superiore a quello richiesto dai metodi diretti, come il metodo di Cholesky applicato al sistema normale o il metodo QR. Lawson e Hanson [12] stimano che il costo computazionale richiesto per la risoluzione del problema (2) tramite il calcolo dei valori singolari sia circa il doppio di quello del metodo QR quando m è molto più grande di n, e fino a 9 volte quello del metodo QR quando m è poco più grande di n.

8. Teoremi di perturbazione per i valori singolari

A differenza di quanto avviene per il problema del calcolo degli autovalori di una matrice, il calcolo dei valori singolari è un problema sempre ben condizionato, perché, come si vedrà, piccole perturbazioni degli elementi della matrice inducono nei risultati perturbazioni non superiori a quelle dei dati.

Sia $A \in \mathbf{C}^{m\times n}$, con $m \geq n$ (se $m < n$ le considerazioni che seguono si

applicano ad A^H) e si consideri la matrice hermitiana

$$B = \begin{bmatrix} O & A \\ A^H & O \end{bmatrix} \quad \in \mathbf{C}^{(m+n)\times(m+n)}. \tag{38}$$

Sia $A = U\Sigma V^H$ la decomposizione ai valori singolari di A e siano

$$U = [U_1 \mid U_2], \quad \Sigma = \begin{bmatrix} \Sigma_1 \\ O \end{bmatrix},$$

dove $U_1 \in \mathbf{C}^{m\times n}$, $U_2 \in \mathbf{C}^{m\times(m-n)}$ e $\Sigma_1 \in \mathbf{R}^{n\times n}$. La matrice

$$Z = \frac{1}{\sqrt{2}} \begin{bmatrix} U_1 & U_1 & \sqrt{2}\,U_2 \\ V & -V & O \end{bmatrix} \quad \in \mathbf{C}^{(m+n)\times(m+n)},$$

è unitaria e tale che

$$B = \begin{bmatrix} O & A \\ A^H & O \end{bmatrix} = Z \begin{bmatrix} \Sigma_1 & O & O \\ O & -\Sigma_1 & O \\ O & O & O \end{bmatrix} Z^H,$$

e quindi le colonne di Z costituiscono un insieme ortonormale di autovettori di B. La matrice B ha come autovalori i numeri σ_i e $-\sigma_i$, $i = 1, \ldots, n$, dove σ_i sono i valori singolari di A, oltre ad $m-n$ autovalori nulli, poiché $m \geq n$; se il rango k di A è minore di n, la molteplicità algebrica dell'autovalore nullo è $m + n - 2k$. I seguenti teoremi derivano dai corrispondenti teoremi di perturbazione degli autovalori.

7.23 Teorema. *Siano $A \in \mathbf{C}^{m\times n}$ e $\widehat{A} \in \mathbf{C}^{m\times(n-1)}$ la matrice ottenuta da A eliminando una colonna. Se σ_i, $i = 1, \ldots, \min\{m, n\}$, sono i valori singolari di A e τ_i, $i = 1, \ldots, \min\{m, n-1\}$, sono i valori singolari di $\widehat{A}$, risulta*

$$\text{se } m \geq n \qquad \sigma_1 \geq \tau_1 \geq \sigma_2 \geq \ldots \geq \tau_{n-1} \geq \sigma_n \geq 0,$$

$$\text{se } m < n \qquad \sigma_1 \geq \tau_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq \tau_m \geq 0.$$

Dim. Si consideri la matrice $\widehat{B} \in \mathbf{C}^{(m+n-1)\times(m+n-1)}$ ottenuta a partire da $\widehat{A}$, in modo analogo alla matrice B della (38). Quindi $\widehat{B}$ si può ottenere da B eliminando una colonna e la riga corrispondente. Per il teorema 6.10 gli autovalori di $\widehat{B}$ separano quelli di B. ∎

7.24 Teorema. *Siano A e $\delta A \in \mathbf{C}^{m\times n}$. Se σ_i, τ_i e ψ_i, $i = 1, \ldots, n$, sono i valori singolari di A, di δA e di $A + \delta A$, risulta*

$$|\psi_i - \sigma_i| \leq \tau_1 = \|\delta A\|_2, \quad i = 1, \ldots, n.$$

Dim. Se $m \geq n$, la matrice

$$\begin{bmatrix} O & \delta A \\ \delta A^H & O \end{bmatrix},$$

ha autovalori

$$\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n \geq -\tau_n \geq \ldots \geq -\tau_1,$$

oltre a $m - n$ autovalori nulli. Per il teorema 6.14 segue che

$$\sigma_i - \tau_1 \leq \psi_i \leq \sigma_i + \tau_1,$$

da cui

$$|\psi_i - \sigma_i| \leq \tau_1 = \|\delta A\|_2, \quad i = 1, \ldots, n.$$

Se $m < n$, la dimostrazione può essere condotta in modo analogo. ∎

Da questo teorema risulta che la perturbazione generata sui valori singolari da una perturbazione δA sugli elementi della matrice A è limitata superiormente da $\|\delta A\|_2$.

7.25 Esempio. Siano

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \delta A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \epsilon & 0 & 0 & 0 \end{bmatrix}.$$

Per il teorema 7.24 i valori singolari σ_i e ψ_i, $i = 1, \ldots, 4$ rispettivamente delle matrici A e $A + \delta A$ verificano la relazione

$$|\psi_i - \sigma_i| \leq |\epsilon|, \quad i = 1, \ldots, 4.$$

Gli autovalori λ_i e μ_i, $i = 1, \ldots, 4$ di A e di $A + \delta A$ verificano invece la relazione (si veda l'esempio 6.6)

$$|\lambda_i - \mu_i| \leq \sqrt[4]{|\epsilon|}.$$

In effetti i valori singolari di A sono $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\sigma_4 = 0$ e quelli di $A + \delta A$ sono $\psi_1 = \psi_2 = \psi_3 = 1$, $\psi_4 = |\epsilon|$, mentre gli autovalori di A sono nulli e gli autovalori di $A + \delta A$ sono i μ_i tali che $\mu_i^4 = \epsilon$, $i = 1, \ldots, 4$.

Ad esempio, nel caso che $\epsilon = 10^{-8}$, introducendo una perturbazione sull'elemento a_{41} di modulo pari ad ϵ, i valori singolari risultano affetti da un perturbazione uguale, mentre gli autovalori risultano affetti da una perturbazione di modulo pari ad 10^{-2}. ∎

Questo esempio illustra come il problema del calcolo dei valori singolari di una matrice risulti essere ben posto, anche se l'associato problema del calcolo degli autovalori è mal posto, caso che può presentarsi quando la matrice non è diagonalizzabile.

Se la matrice è normale autovalori e valori singolari hanno lo stesso modulo. Se la matrice non è normale la differenza fra i moduli degli autovalori e i valori singolari può essere arbitrariamente grande, come risulta anche dall'esempio 7.25. Vale il seguente teorema.

7.26 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$. Per ogni autovalore λ di A vale*

$$\sigma_n \leq |\lambda| \leq \sigma_1.$$

Dim. Sia $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$. Allora

$$\mathbf{x}^H A^H A\mathbf{x} = |\lambda|^2 \|\mathbf{x}\|_2^2. \tag{39}$$

Dalla decomposizione ai valori singolari di A segue che

$$A^H A = V \Sigma^2 V^H,$$

in cui $V \in \mathbf{C}^{n\times n}$ è unitaria, e quindi posto $\mathbf{y} = V^H\mathbf{x}$, è

$$\mathbf{x}^H A^H A\mathbf{x} = \mathbf{y}^H \Sigma^2 \mathbf{y} = \sum_{i=1}^{n} \sigma_i^2 |y_i|^2,$$

e poiché

$$\sigma_n^2 \|\mathbf{y}\|_2^2 \leq \sum_{i=1}^{n} \sigma_i^2 |y_i|^2 \leq \sigma_1^2 \|\mathbf{y}\|_2^2$$

e $\|\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$, per la (39) risulta

$$\sigma_n \leq |\lambda| \leq \sigma_1.$$ ∎

Se la matrice $A \in \mathbf{C}^{n\times n}$ è non singolare, il suo numero di condizionamento in norma 2 è per la (34)

$$\mu_2(A) = \frac{\sigma_1}{\sigma_n},$$

mentre per il teorema 7.26 risulta in generale

$$\mu_2(A) \geq \frac{\lambda_1}{\lambda_n},$$

con il segno di uguaglianza se A è normale. Quindi una matrice non normale può essere malcondizionata anche se il rapporto fra il massimo e il minimo modulo degli autovalori non è grande.

Queste considerazioni hanno suggerito studi e ricerche per individuare matrici "quasi normali", cioè con autovalori "vicini" in modulo ai valori singolari, e per migliorare quindi il condizionamento del problema del calcolo degli autovalori.

7.27 Esempio. La matrice triangolare B dell'esempio 7.14 ha gli autovalori tutti uguali a 1. Però al cresce di n il minimo valore singolare tende a zero, per cui la matrice risulta mal condizionata, anche per valori non troppo grossi di n. Al variare di n il numero di condizionamento $\mu_2(A) = \sigma_1/\sigma_n$ è riportato nella seguente tabella

n	$\mu_2(A)$
5	$2.942748\ 10^1$
10	$1.918453\ 10^3$
15	$9.512388\ 10^4$
20	$3.996832\ 10^6$

■

9. Calcolo della forma normale di Schur di $A^H A$

Per quanto visto nel paragrafo 4, gli autovalori di $A^H A$ sono i quadrati dei valori singolari $\sigma_1, \ldots, \sigma_n$ di A e il calcolo della decomposizione ai valori singolari di A richiede, come primo passo, che si calcolino gli autovalori e gli autovettori di $A^H A$. Per semplificare questo primo passo conviene trasformare prima la matrice A in una matrice B bidiagonale superiore a elementi reali, con il seguente algoritmo di *Golub e Reinsch*, che utilizza una successione di trasformazioni unitarie. Se la matrice A è sparsa può essere

conveniente utilizzare trasformazioni di Givens al posto di quelle di Householder. Per semplicità si suppone $m \geq n$ (se $m < n$ si applica l'algoritmo alla matrice A^H).

Posto $A^{(1)} = A$, si costruisce la prima matrice elementare di Householder $P^{(1)}$ in modo che la matrice

$$A^{(2)} = P^{(1)} A^{(1)}$$

abbia nulli gli elementi della prima colonna al di sotto di quello principale. La matrice $A^{(2)}$ risulta

$$A^{(2)} = \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B^{(2)} \end{bmatrix},$$

in cui $\mathbf{c} \in \mathbf{C}^{n-1}$. Si costruisce poi la matrice elementare di Householder $K^{(1)} \in \mathbf{C}^{(n-1)\times(n-1)}$ tale che il vettore $K^{(1)}\mathbf{c}$ abbia nulle tutte le componenti di indice maggiore o uguale a 2. Indicata con $H^{(1)}$ la matrice elementare di Householder

$$H^{(1)} = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & K^{(1)} \end{bmatrix},$$

la matrice $A^{(3)} = A^{(2)} H^{(1)}$ ha nulli gli elementi della prima riga che hanno indice di colonna maggiore o uguale a 3 e gli elementi della prima colonna che hanno indice di riga maggiore o uguale a 2. Si ripete il procedimento per $n - 2$ volte, annullando alternativamente gli elementi delle colonne e delle righe, e proseguendo poi sull'ultima colonna se $m > n$, come indicato nella figura 7.3 per il caso particolare $n = 4,\ m > 4$.

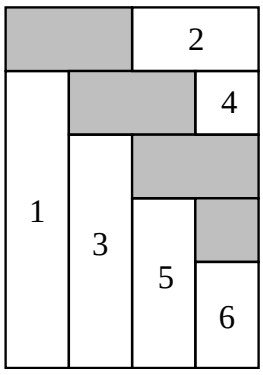

Fig. 7.3 - Riduzione a forma bidiagonale.

La successione $A^{(1)} = A, A^{(2)}, \ldots, A^{(2n-1)}$ viene costruita nel modo seguente:

$$\left.\begin{aligned} A^{(2k)} &= P^{(k)} A^{(2k-1)} \\ A^{(2k+1)} &= A^{(2k)} H^{(k)} \end{aligned}\right\} \quad k = 1, 2, \ldots, n-2$$

$$A^{(2n-2)} = P^{(n-1)} A^{(2n-3)}$$

$$A^{(2n-1)} = P^{(n)} A^{(2n-2)} \quad \text{se } m > n,$$

dove $P^{(k)} \in \mathbf{C}^{m\times m}$, $k = 1, 2, \ldots, n$, è una matrice elementare di Householder che annulla gli elementi della k-esima colonna di $A^{(2k-1)}$ con indice di riga maggiore o uguale a $k+1$ (se $m = n$ è $P^{(n)} = I$), e $H^{(k)} \in \mathbf{C}^{n\times n}$, $k = 1, 2, \ldots, n-2$, è una matrice elementare di Householder che annulla gli elementi della k-esima riga di $A^{(2k)}$ con indice di colonna maggiore o uguale a $k+2$. Dopo $2n-2$ passi risulta quindi

$$P^{(n)} \ldots P^{(2)} P^{(1)} A H^{(1)} H^{(2)} \ldots H^{n-2}$$

$$= \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & \alpha_n \\ & & & & \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} n \text{ righe} \\ \left.\begin{matrix} \end{matrix}\right\} m-n \text{ righe} \end{matrix}$$

in cui α_i, $i = 1, \ldots, n$ e β_i, $i = 1, \ldots, n-1$, sono in generale numeri complessi.

Se gli α_i e i β_i non sono tutti reali, esistono due matrici di fase S e T (si veda l'esercizio 7.35) tali che

$$S \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & \alpha_n \end{bmatrix} T = \begin{bmatrix} |\alpha_1| & |\beta_1| & & & \\ & |\alpha_2| & |\beta_2| & & \\ & & \ddots & \ddots & \\ & & & \ddots & |\beta_{n-1}| \\ & & & & |\alpha_n| \end{bmatrix}.$$

Posto

$$P = \begin{bmatrix} S & O \\ O & I_{m-n} \end{bmatrix} P^{(n)} \ldots P^{(2)} P^{(1)} \in \mathbf{C}^{m\times m}$$

e

$$H = H^{(1)} H^{(2)} \ldots H^{n-2} T \in \mathbf{C}^{n\times n},$$

risulta allora

$$PAH = \begin{bmatrix} B \\ \\ O \end{bmatrix},$$

dove $B \in \mathbf{R}^{n\times n}$ è una matrice bidiagonale superiore.

L'algoritmo di Golub e Reinsch richiede

$$2mn^2 - \frac{2}{3}n^3$$

operazioni moltiplicative [9].

Se $m > \frac{5}{3}n$ si riduce il costo computazionale se, prima di applicare il metodo di Golub e Reinsch, la matrice A viene fattorizzata nella forma QR [3].

7.28 Esempio. Sia

$$A = \frac{1}{100}\begin{bmatrix} \text{-}50 & 230 & 235 \\ 50 & \text{-}142 & 81 \\ 50 & 38 & \text{-}159 \\ 100 & \text{-}4 & 122 \\ \text{-}150 & 126 & \text{-}343 \end{bmatrix}.$$

Posto $A^{(1)} = A$, con l'algoritmo di Golub e Reinsch si ha

$$P^{(1)} = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T, \text{ con } \mathbf{v}_1 = \frac{1}{2}\,[\text{-}5,\ 1,\ 1,\ 2,\ \text{-}3]^T,\ \beta_1 = \frac{1}{5},$$

$$A^{(2)} = P^{(1)}A^{(1)} = \frac{1}{5}\begin{bmatrix} 10 & \text{-}9 & 12 \\ 0 & \text{-}3 & 4 \\ 0 & 6 & \text{-}8 \\ 0 & 8 & 6 \\ 0 & \text{-}6 & \text{-}17 \end{bmatrix},$$

$$H^{(1)} = I - \gamma_1 \mathbf{w}_1 \mathbf{w}_1^T, \text{ con } \mathbf{w}_1 = \frac{12}{5}\,[0,\ \text{-}2,\ 1]^T,\ \gamma_1 = \frac{5}{72},$$

$$A^{(3)} = A^{(2)}H^{(1)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & \text{-}2 & 0 \\ 0 & 0 & 2 \\ 0 & \text{-}2 & \text{-}3 \end{bmatrix},$$

$$P^{(2)} = I - \beta_2 \mathbf{v}_2 \mathbf{v}_2^T, \text{ con } \mathbf{v}_2 = [0,\ 4,\ \text{-}2,\ 0,\ \text{-}2]^T,\ \beta_2 = \frac{1}{12},$$

$$A^{(4)} = P^{(2)}A^{(3)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & -2 \end{bmatrix},$$

$$P^{(3)} = I - \beta_3 \mathbf{v}_3 \mathbf{v}_3^T, \text{ con } \mathbf{v}_3 = [0,\ 0,\ 4,\ 2,\ -2]^T,\ \beta_3 = \frac{1}{12},$$

$$A^{(5)} = P^{(3)}A^{(4)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

da cui si ha

$$A = P^T \begin{bmatrix} B \\ \\ O \end{bmatrix} H^T,$$

dove

$$P = P^{(3)}P^{(2)}P^{(1)} = \frac{1}{180} \begin{bmatrix} -45 & 45 & 45 & 90 & -135 \\ -75 & -45 & 135 & 30 & 75 \\ -145 & 69 & -51 & -62 & 13 \\ -35 & -33 & -93 & 134 & 59 \\ 50 & 150 & 30 & 40 & 70 \end{bmatrix},$$

$$B = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & -3 \end{bmatrix},$$

$$H = H^{(1)} = \frac{1}{5} \begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & 4 \\ 0 & 4 & 3 \end{bmatrix}.$$

■

Dopo aver applicato l'algoritmo di Golub e Reinsch la matrice A risulta quindi della forma

$$A = P^H \begin{bmatrix} B \\ \\ O \end{bmatrix} H^H,$$

in cui P e H sono matrici unitarie e $B \in \mathbf{R}^{n\times n}$ è bidiagonale superiore e poiché

$$A^H A = HB^T BH^H, \tag{40}$$

gli autovalori di B^TB sono ancora $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Per calcolare $\sigma_1, \sigma_2, \ldots,$ σ_n basta quindi calcolare gli autovalori della matrice simmetrica, semidefinita positiva e tridiagonale B^TB.

Con il metodo QR per il calcolo degli autovalori descritto nel capitolo 6 si ottiene la decomposizione

$$B^TB = WDW^T, \tag{41}$$

in cui $D \in \mathbf{R}^{n\times n}$ è la matrice diagonale avente come elementi principali i σ_i^2, $i = 1, \ldots, n$, ordinati in modo non crescente, e $W \in \mathbf{R}^{n\times n}$ è una matrice ortogonale. Da (40) e (41) si ottiene

$$A^HA = HWDW^TH^H,$$

che coincide con la (22) se si pone $Q = HW$.

7.29 Esempio. Per la matrice A dell'esempio 7.28 si ha

$$A^TA = HB^TBH^T,$$

dove

$$H = \frac{1}{5}\begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & 4 \\ 0 & 4 & 3 \end{bmatrix},$$

$$B^TB = \begin{bmatrix} 4 & 6 & 0 \\ 6 & 18 & 6 \\ 0 & 6 & 13 \end{bmatrix}.$$

Gli autovalori di B^TB calcolati con il metodo QR sono

$$\sigma_1^2 = 23.34213, \quad \sigma_2^2 = 10.31179, \quad \sigma_3^2 = 1.346078,$$

e quindi i valori singolari di A sono

$$\sigma_1 = 4.831369, \quad \sigma_2 = 3.211198, \quad \sigma_3 = 1.160206,$$

e la matrice A risulta di rango 3.

La matrice ortogonale degli autovettori di B^TB è

$$H = \begin{bmatrix} 0.2591518 & -0.3622751 & 0.8953192 \\ 0.8354232 & -0.3810986 & -0.3960196 \\ 0.4846731 & 0.8505999 & 0.2038906 \end{bmatrix}.$$

Posto

$$C = AQ = AHW = \begin{bmatrix} 1.863326 & 2.755032 & 0.01697129 \\ 1.067666 & -1.305669 & -0.2788946 \\ -1.438590 & -0.1623822 & 0.9091571 \\ 1.433844 & -0.1479529 & 0.6420239 \\ -3.821606 & 0.9841209 & -0.1710005 \end{bmatrix},$$

si fattorizza la matrice C nella forma QR con il metodo di Householder; risulta

$$C = UR,$$

dove

$$R = \begin{bmatrix} 4.831369 & 0 & 0 \\ 0 & 3.211198 & 0 \\ 0 & 0 & 1.160206 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

e

$$U = \begin{bmatrix} 0.3856735 & 0.8579459 & 0.01462799 & -0.2087997 & 0.2671558 \\ 0.2209868 & -0.4065989 & -0.2403848 & -0.2269533 & 0.8225245 \\ -0.2977611 & -0.05056683 & 0.7836169 & -0.5247251 & 0.1392329 \\ 0.2967787 & -0.04607503 & 0.5533710 & 0.7317119 & 0.2611086 \\ -0.7910007 & 0.3064681 & -0.1473861 & 0.3068481 & 0.4056067 \end{bmatrix}.$$

La decomposizione ai valori singolari di A è allora data da

$$A = URV^T,$$

dove

$$V = Q = HW = \begin{bmatrix} 0.2591518 & -0.3622751 & 0.8953192 \\ -0.1135125 & 0.9091362 & 0.4007223 \\ 0.9591415 & 0.2054818 & -0.1944808 \end{bmatrix}.$$

In questo caso la matrice A ha rango massimo e il problema dei minimi quadrati (2) con

$$\mathbf{b} = [1,\ 1,\ 1,\ 1,\ 1]^T$$

ha un'unica soluzione, che per il teorema 7.15 è data da

$$\mathbf{x}^* = \frac{\mathbf{q}_1^T\mathbf{b}}{\sigma_1}\ \mathbf{t}_1 + \ \frac{\mathbf{q}_2^T\mathbf{b}}{\sigma_2}\ \mathbf{t}_2 + \ \frac{\mathbf{q}_3^T\mathbf{b}}{\sigma_3}\ \mathbf{t}_3,$$

in cui i vettori $\mathbf{q}_j$ e $\mathbf{t}_j, j = 1, 2, 3$ sono rispettivamente le prime tre colonne della matrice U e della matrice V. Risulta allora

$$\mathbf{x}^* = [0.6592600,\ 0.5244430,\ \text{-}0.1560473]^T. \qquad \blacksquare$$

Il metodo QR per il calcolo degli autovalori di B^TB, matrice tridiagonale, richiede ad ogni iterazione un numero di operazioni moltiplicative lineare in n. Poiché B è bidiagonale, anche il calcolo di B^TB richiede un numero di operazioni moltiplicative lineare in n. In [9] è esposta una particolare tecnica che consente di utilizzare il metodo QR per il calcolo degli autovalori di B^TB senza calcolare esplicitamente gli elementi di B^TB. Anche questa tecnica richiede ad ogni iterazione un numero di operazioni moltiplicative lineare in n.

Come è già stato rilevato, uno dei problemi più delicati è quello del calcolo del rango della matrice A. In pratica, fissata una tolleranza ϵ, si assume come rango di A il numero degli elementi principali b_{ii} di B tali che $|b_{ii}| \geq \epsilon$. La scelta di ϵ assume quindi un ruolo molto importante nella risoluzione del problema.

In generale, prima di calcolare gli autovalori e gli autovettori della matrice B^TB, è opportuno esaminare se uno degli elementi α_i o β_i di B è nullo, perché in tal caso il problema del calcolo degli autovalori di B^TB è ricondotto al calcolo degli autovalori di due matrici di ordine inferiore.

a) Se esiste un indice $i,\ 1 \leq i \leq n-1$, per cui $\beta_i = 0$, allora la matrice B ha la forma

$$B = \begin{bmatrix} B_1 & O \\ & \\ O & B_2 \end{bmatrix} \begin{matrix} \} & i \text{ righe} \\ & \\ \} & n-i \text{ righe} \end{matrix}$$

dove B_1 e B_2 sono blocchi quadrati, e risulta

$$B^TB = \begin{bmatrix} B_1^TB_1 & O \\ & \\ O & B_2^TB_2 \end{bmatrix}, \tag{42}$$

in cui le matrici $B_1^TB_1$ e $B_2^TB_2$ hanno rispettivamente ordine i e $n-i$;

b) se esiste un indice $i,\ 2 \leq i \leq n-1$, per cui $\alpha_i = 0$ e $\beta_i \neq 0$, allora la matrice B ha la forma

$$B = \begin{bmatrix} B_1 & O \\ & \\ O & B_2 \end{bmatrix},$$

dove $B_1 \in \mathbf{R}^{(i-1)\times i}$ e $B_2 \in \mathbf{R}^{(n-i+1)\times(n-i)}$, per cui B^TB è ancora della forma (42).

7.30 Esempio. La matrice bidiagonale

$$B = \begin{bmatrix} 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & \frac{1}{6}\sqrt{6} & 2 \\ 0 & 0 & 0 & 0 & \sqrt{3} \end{bmatrix},$$

ha l'elemento $b_{22} = 0$. Il problema del calcolo degli autovalori di $B^T B$ viene ricondotto al calcolo degli autovalori delle due matrici

$$B_1^T B_1 = \begin{bmatrix} 9 & -3 \\ -3 & 1 \end{bmatrix}, \quad \text{e} \quad B_2^T B_2 = \begin{bmatrix} 8 & -2 & 0 \\ -2 & \frac{7}{6} & \frac{1}{3}\sqrt{6} \\ 0 & \frac{1}{3}\sqrt{6} & 7 \end{bmatrix}. \qquad \blacksquare$$

10. Calcolo della soluzione di minima norma con il metodo del gradiente coniugato

Il metodo del gradiente coniugato descritto nel capitolo 5 per la risoluzione dei sistemi lineari può essere utilizzato anche per calcolare la soluzione di minima norma del problema dei minimi quadrati

$$\min_{\mathbf{y} \in \mathbf{R}^n} \|A\mathbf{y} - \mathbf{b}\|_2^2, \quad \text{dove } A \in \mathbf{R}^{m \times n} \text{ e } \mathbf{b} \in \mathbf{R}^m, \quad \text{con} \quad m \geq n.$$

In questo caso il funzionale che si deve minimizzare è dato da

$$\|A\mathbf{y} - \mathbf{b}\|_2^2 = 2(\tfrac{1}{2}\mathbf{y}^T A^T A\mathbf{y} - \mathbf{b}^T A\mathbf{y}) + \mathbf{b}^T\mathbf{b}, \quad \mathbf{y} \in \mathbf{R}^n$$

e soluzione del problema dei minimi quadrati è un vettore $\mathbf{x}$ tale che

$$\Phi(\mathbf{x}) = \min_{\mathbf{y} \in \mathbf{R}^n} \Phi(\mathbf{y}),$$

dove

$$\Phi(\mathbf{y}) = \tfrac{1}{2}\|A\mathbf{y} - \mathbf{b}\|_2^2 - \mathbf{b}^T\mathbf{b} = \tfrac{1}{2}\mathbf{y}^T A^T A\mathbf{y} - \mathbf{b}^T A\mathbf{y}.$$

Si può quindi applicare il metodo del gradiente coniugato utilizzando l'algoritmo esposto nel paragrafo 7 del capitolo 5 con l'ovvia modifica che la direzione del gradiente negativo di $\Phi(\mathbf{x})$ nel punto $\mathbf{x}_k$ è data da

$$-\nabla\Phi(\mathbf{x}_k) = A^T(\mathbf{b} - A\mathbf{x}_k) = A^T\mathbf{r}_k,$$

dove $\mathbf{r}_k$ è il residuo del punto $\mathbf{x}_k$. L'algoritmo risulta allora il seguente:

1. $k = 0,\ \mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

$$\mathbf{s}_k = A^T\mathbf{r}_k,$$

$$\beta_k = \frac{\|\mathbf{s}_k\|_2^2}{\|\mathbf{s}_{k-1}\|_2^2} \quad (\beta_0 = 0, \text{ per } k = 0),$$

$$\mathbf{p}_k = \mathbf{s}_k + \beta_k\mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{s}_0, \text{ per } k = 0),$$

$$\alpha_k = \frac{\|\mathbf{s}_k\|_2^2}{\|A\mathbf{p}_k\|_2^2},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k,$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k,$$

$k = k + 1$ e si vada al punto 2.

Come condizione di arresto si può usare la stessa condizione di arresto usata nel capitolo 5, cioè

$$\|\mathbf{s}_k\|_2 < \epsilon\,\|\mathbf{b}\|_2,$$

dove ϵ è una tolleranza prefissata.

7.31 Esempio. Si applica il metodo del gradiente coniugato al problema di minimi quadrati (2) con

$$A = \frac{1}{100}\begin{bmatrix} \text{-}50 & 230 & 235 \\ 50 & \text{-}142 & 81 \\ 50 & 38 & \text{-}159 \\ 100 & \text{-}4 & 122 \\ \text{-}150 & 126 & \text{-}343 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

(la soluzione è stata calcolata per mezzo dei valori singolari negli esempi 7.28 e 7.29). Fissato $\mathbf{x}_0 = \mathbf{0}$, con l'algoritmo descritto si ottiene la successione

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.	0.2432537	-0.06277519	1.508581
2	0.1100399	0.3457329	0.001886844	1.271655
3	0.6592587	0.5244448	-0.1560494	$0.8422623\ 10^{-5}$

Si assume quindi $\mathbf{x}_3$ come approssimazione di $\mathbf{x}^*$. ∎

Se la matrice A ha rango massimo, con questo algoritmo si ottiene un'approssimazione della soluzione $\mathbf{x}^*$ del problema (2); se A non ha rango massimo si ottiene un'approssimazione di una soluzione $\mathbf{x}$ del problema (2), che dipende dalla scelta del punto iniziale $\mathbf{x}_0$. Scegliendo $\mathbf{x}_0 = \mathbf{0}$ si ottiene un'approssimazione della soluzione $\mathbf{x}^*$ di minima norma. Infatti per ogni $\mathbf{x} \in N(A)$, si ha

$$\mathbf{x}^T \mathbf{s}_k = \mathbf{x}^T A^T \mathbf{r}_k = 0 \quad \text{per ogni } k,$$

e poiché $\mathbf{p}_0 = \mathbf{s}_0$ e $\mathbf{x}_0 = \mathbf{0}$, procedendo per induzione su k, si ha

$$\mathbf{x}^T \mathbf{p}_k = \mathbf{x}^T \mathbf{s}_k + \beta_k \mathbf{x}^T \mathbf{p}_{k-1} = 0,$$

$$\mathbf{x}^T \mathbf{x}_{k+1} = \mathbf{x}^T \mathbf{x}_k + \alpha_k \mathbf{x}^T \mathbf{p}_k = 0,$$

cioè $\mathbf{x}_{k+1}$ appartiene allo spazio $N(A)^\perp$. Poiché in aritmetica esatta la successione $\{\mathbf{x}_k\}$, dopo al più n passi, raggiunge una soluzione del problema (2), questa deve appartenere allo spazio $N(A)^\perp$, ed essendo $N(A)^\perp = N(A^H A)^\perp$ (si veda l'esercizio 1.38), coincide con $\mathbf{x}^*$, che è l'unica soluzione di (2) appartenente a $N(A^H A)^\perp$.

7.32 Esempio. Si applica il metodo del gradiente coniugato al problema dei minimi quadrati (2) con

$$A = \frac{1}{45} \begin{bmatrix} 6 & 12 & -72 \\ -16 & -7 & -8 \\ 58 & 16 & 104 \\ 87 & 24 & 156 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

già studiato nell'esempio 7.3, la cui soluzione di minima norma è stata calcolata ricorrendo ai valori singolari nell'esempio 7.16. Fissato $\mathbf{x}_0 = [1,\ 1,\ 1]^T$, si ottiene la successione

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.4538005	0.8656725	-0.1101770	0.2133383
2	0.6197463	0.9604924	-0.2049520	$0.3888539\ 10^{-3}$
3	0.6197532	0.9604941	-0.2049382	$0.5444772\ 10^{-6}$

e $\mathbf{x}_3$ approssima la soluzione (si veda l'esempio 7.3)

$$\mathbf{x} = \begin{bmatrix} -\dfrac{1}{5} - 4h \\ \dfrac{13}{5} + 8h \\ h \end{bmatrix}, \quad \text{per } h = -0.2049382,$$

che non è di minima norma, con l'errore effettivo

$$\|\mathbf{x}_3 - \mathbf{x}\|_2 = 0.1013279\ 10^{-5}.$$

Fissato $\mathbf{x}_0 = \mathbf{0}$, la successione che si ottiene è

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.1246007	0.04153361	0.1661344	0.9774478
2	0.8666654	0.4666667	-0.2666695	$0.8394349\ 10^{-4}$
3	0.8666668	0.4666670	-0.2666665	$0.1644842\ 10^{-6}$

e $\mathbf{x}_3$ approssima la soluzione di minima norma $\mathbf{x}^*$ con l'errore effettivo

$$\|\mathbf{x}_3 - \mathbf{x}^*\|_2 = 0.4525244\ 10^{-6}.$$ ∎

7.33 Esempio (approssimazione polinomiale). Sia $f(x)$ una funzione reale e siano $x_i, i = 1, 2, \ldots, m,\ m$ punti, a due a due distinti, appartenenti al dominio di $f(x)$. Fra tutti i polinomi

$$p_{n-1}(x) = \alpha_0 + \alpha_1 x + \ldots + \alpha_{n-1} x^{n-1}$$

di grado minore o uguale a $n-1$, quello per cui lo *scarto quadratico*

$$\sum_{i=1}^{m} [p_{n-1}(x_i) - f(x_i)]^2$$

è minimo viene detto *polinomio di approssimazione ai minimi quadrati* di $f(x)$. I coefficienti del polinomio di approssimazione ai minimi quadrati possono essere determinati con una delle tecniche descritte in questo capitolo, risolvendo il problema (2) in cui la matrice $A \in \mathbf{R}^{m\times n}$ e i vettori $\mathbf{y} \in \mathbf{R}^n$ e $\mathbf{b} \in \mathbf{R}^m$ hanno gli elementi

$$a_{ij} = x_i^{j-1}, \quad y_j = \alpha_{j-1}, \quad b_i = f(x_i), \quad i = 1, \ldots, m,\ j = 1, \ldots, n.$$

Poiché i punti x_i sono a due a due distinti, la matrice A (come si è rilevato nell'esempio 7.22) ha rango massimo e quindi il problema dei minimi quadrati (2) ha una e una sola soluzione, che fornisce i coefficienti del polinomio di approssimazione dei minimi quadrati di grado $n-1$.

Un confronto fra i metodi esposti viene ora fatto per il caso particolare $x_i = \frac{1}{i},\ i = 1, \ldots, m,\ f(x) = \sqrt{x}$, per $m = 10$ in cui al variare di $n-1$ il numero di condizionamento $\mu_2(A)$ e il minimo γ del problema sono approssimativamente i seguenti:

$n-1$	$\mu_2(A)$	γ
2	26	0.5366
3	153	1.112
4	969	2.323
5	6561	5.334
6	48520	6.500
7	397678	14.41

I metodi sperimentati sono:

NR risoluzione del sistema normale con il metodo di Cholesky (par. 1)

QR metodo QR (par. 2)

VS calcolo per mezzo dei valori singolari (par. 5)

GC calcolo per mezzo del gradiente coniugato (par. 10).

I risultati ottenuti sono riportati nella seguente tabella, in cui t è il tempo di calcolo impiegato, misurato in millesimi di secondo, ed $\|\mathbf{e}\|_2$ è la norma 2 dell'errore assoluto dei coefficienti determinati.

metodo	$n-1=2$		$n-1=3$		$n-1=4$	
	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$
NR	0.22	$0.16\ 10^{-4}$	0.24	$0.13\ 10^{-2}$	0.37	$0.11\ 10^{0}$
QR	0.18	$0.27\ 10^{-5}$	0.25	$0.11\ 10^{-4}$	0.37	$0.18\ 10^{-3}$
VS	2.34	$0.84\ 10^{-5}$	3.67	$0.62\ 10^{-4}$	5.28	$0.18\ 10^{-3}$
GC	0.59	$0.76\ 10^{-5}$	1.10	$0.98\ 10^{-4}$	2.34	$0.29\ 10^{-3}$

metodo	$n-1=5$		$n-1=6$		$n-1=7$	
	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$
NR	–	–	–	–	–	–
QR	0.46	$0.43\ 10^{-2}$	0.57	$0.36\ 10^{-1}$	0.80	$0.10\ 10^{0}$
VS	7.02	$0.15\ 10^{-2}$	9.13	$0.65\ 10^{-2}$	10.7	$0.58\ 10^{1}$
GC	3.87	$0.92\ 10^{-3}$	8.00	$0.11\ 10^{-1}$	7.18	$0.37\ 10^{2}$

Dai risultati riportati in questa tabella, risulta che, in accordo con quanto esposto nel paragrafo 7, il metodo **NR** è il meno stabile e addirittura non consente di calcolare la soluzione per $n \geq 6$, perché la matrice $A^T A$ è così malcondizionata che, a causa della limitata precisione di calcolo, non viene riconosciuta come definita positiva dal metodo di Cholesky. Il metodo

QR consente di calcolare la soluzione $\mathbf{x}^*$ in un tempo minore del metodo **NR** e nettamente minore dei metodi **VS** e **GC**. Il tempo di esecuzione del metodo **GC** dipende dal valore della tolleranza ϵ fissata per la condizione di arresto: è possibile ridurre il tempo di esecuzione utilizzando una tecnica di precondizionamento. Per quanto riguarda l'errore della soluzione, i metodi **QR**, **VS** e **GC** hanno un comportamento confrontabile, con un errore che è legato a $\mu_2(A)$.

Comunque queste tecniche non sono adatte a risolvere il problema dell'approssimazione polinomiale quando n è grande per l'elevato malcondizionamento della matrice A. Altre tecniche, basate sull'uso di opportuni polinomi ortogonali, consentono una migliore risoluzione di questo problema. ∎

11. Il metodo di Lanczos per il calcolo dei valori e dei vettori singolari

In molte applicazioni la matrice A del problema (2) è di grosse dimensioni e sparsa: in questi casi le tecniche esposte nei paragrafi precedenti possono non essere praticamente applicabili se le matrici intermedie generate non sono sparse. Inoltre talvolta possono essere richiesti solo pochi valori e vettori singolari. Se la matrice A è reale il metodo di Lanczos può essere convenientemente applicato anche in questo caso.

Nel paragrafo 8 si è visto che gli autovalori della matrice B definita nella (38) e i corrispondenti autovettori, consentono di calcolare i valori e i vettori singolari della matrice A. Se $m = n$ risulta $U_1 = U$ e

$$Z = \frac{1}{\sqrt{2}} \begin{bmatrix} U_1 & U_1 \\ V & -V \end{bmatrix},$$

e quindi gli autovettori $\mathbf{z}$ di B sono tutti della forma

$$\mathbf{z} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1.$$

Se $m > n$, allora vi sono degli autovettori $\mathbf{z}$ di B della forma

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{u} \in \mathbf{C}^m, \ \|\mathbf{u}\|_2 = 1, \tag{43}$$

corrispondenti all'autovalore nullo, che, essendo dovuti al fatto che la matrice A è rettangolare, non corrispondono ad alcun valore singolare di A.