CAPITOLO 1

Problema ai minimi quadrati e fattorizzazione QR

1.1. Problema ai minimi quadrati

Siano assegnati $A \in \mathbb{C}^{m\times n}$, con $m \geq n$, e $\mathbf{b} \in \mathbb{C}^m$. Si vogliono determinare $\mathbf{x}^* \in \mathbb{C}^n$ e $\gamma \in \mathbb{R}$ tali che

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{C}^n} \|A\mathbf{x}-\mathbf{b}\|_2, \quad \gamma = \|A\mathbf{x}^*-\mathbf{b}\|_2, \tag{1.1}$$

ovvero si vuole determinare il vettore in $\mathbb{C}^n$ che produce nell'immagine di A, range$(A) = \{\mathbf{y} \in \mathbb{C}^m : \mathbf{y} = A\mathbf{t}$ per $\mathbf{t} \in \mathbb{C}^n\}$, il vettore in $\mathbb{C}^m$ più vicino a $\mathbf{b}$, e si vuole misurare questa distanza.

Assumiamo che il rango k di A sia massimo (ovvero $k = n$). Se $m > n$ e il sistema $A\mathbf{x} = \mathbf{b}$ non ha soluzione, ovvero il il termine noto $\mathbf{b}$ è tale che $\mathbf{b} \notin$ range(A), allora decomponendo come $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$, con $\mathbf{b}_1 \in$ range(A) e $\mathbf{b}_2 \in$ range$(A)^\perp = \{\mathbf{z} \in \mathbb{C}^m : \mathbf{z}^H\mathbf{y} = 0$ per ogni $\mathbf{y} \in$ range$(A)\}$, avremo necessariamente che la componente $\mathbf{b}_2$ è non nulla. Chiedere che il residuo $\mathbf{r}$,

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} = \mathbf{b}_1 - A\mathbf{x} + \mathbf{b}_2, \quad \text{con } \mathbf{b}_1 - A\mathbf{x} \in \text{range}(A),\ \mathbf{b}_2 \in \text{range}(A)^\perp,$$

abbia norma euclidea minima equivale a chiedere che $\mathbf{x}^*$ sia tale che $A\mathbf{x}^* = \mathbf{b}_1$; ciò equivale a chiedere che il residuo sia dato dalla componente $\mathbf{b}_2$ (e di conseguenza ad avere $\gamma = \|\mathbf{b}_2\|_2$), ovvero che il residuo sia ortogonale all'immagine di A; ciò infine equivale a chiedere che il residuo sia nel nucleo di A^H (dato che, come è noto, $\ker(A^H) = \text{range}(A)^\perp$). Siamo dunque arrivati a scrivere il problema di minimo in (1.1) nella forma equivalente: determinare (univocamente) la soluzione $\mathbf{x}^* \in \mathbb{C}^n$ del sistema

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

Il sistema $A^HA\mathbf{x} = A^H\mathbf{b}$ si definisce sistema delle *equazioni normali*. Dato che A ha rango uguale a n, la matrice $A^HA \in \mathbb{C}^{n\times n}$ è invertibile (per la precisione, è hermitiana definita positiva). La soluzione del sistema è data dal vettore $\mathbf{x}^* = (A^HA)^{-1}A^H\mathbf{b}$. Quindi, nell'ipotesi che A sia di rango massimo, il problema ai minimi quadrati ha unica soluzione $\mathbf{x}^*$ (e naturalmente anche γ è univocamente determinato). Il vettore $\mathbf{x}^*$ si dice soluzione di $A\mathbf{x} = \mathbf{b}$ nel senso dei minimi quadrati.

A questo punto introduciamo una generalizzazione della nozione di matrice inversa nel caso in cui la matrice A non sia quadrata e neanche necessariamente di rango massimo (e dunque A^{-1} non esiste). La matrice *pseudo-inversa* di

Moore Penrose, che interviene nella soluzione del sistema $A\mathbf{x} = \mathbf{b}$ nel senso dei minimi quadrati, è denotata $A^\dagger$ e risulta definita univocamente dalle proprietà

$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger, \quad (AA^\dagger)^H = AA^\dagger, \quad (A^\dagger A)^H = A^\dagger A.$$

Nel caso di rango massimo, e $m \geq n$, si ha

$$A^\dagger = (A^H A)^{-1} A^H \in \mathbb{C}^{n\times m},$$

e dunque $\mathbf{x}^* = A^\dagger \mathbf{b}$. È immediato verificare che vale $A^\dagger A = I_n$, dove con con I_n denotiamo la matrice identità di ordine n, ovvero la pseudo-inversa $A^\dagger$ della matrice $A \in \mathbb{C}^{m\times n}$, con $m \geq n$, di rango massimo è una inversa sinistra.
Inoltre, se $m = n$, vale $A^\dagger = A^{-1}$.

Vogliamo però evitare di risolvere, pericolosamente e onerosamente, il sistema delle equazioni normali. Pericolosamente perché vedremo che il numero di condizionamento spettrale del sistema delle equazioni normali, ovvero $\kappa_2(A^H A)$ può essere molto grande. Onerosamente perché si devono quantomeno calcolare esplicitamente la matrice dei coefficienti $A^H A$ e il termine noto $A^H \mathbf{b}$.

1.2. Algoritmo di fattorizzazione QR e riflettori di Householder

II seguente teorema mostra che una matrice rettangolare ammette una fattorizzazione di tipo QR, ovvero può essere fattorizzata nel prodotto di una matrice unitaria e una matrice trapezoidale superiore.

TEOREMA 1.1. *Data $A \in \mathbb{C}^{m\times n}$, esistono una matrice unitaria $Q \in \mathbb{C}^{m\times m}$ e una matrice $R = (r_{ij}) \in \mathbb{C}^{m\times n}$ con elementi $r_{ij} = 0$ per $i > j$ tali che*

$$A = QR.$$

Per dimostrare costruttivamente il precedente risultato, noto come *Teorema di esistenza della fattorizzazione QR,* possiamo utilizzare i riflettori elementari di Householder. Osserviamo che la fattorizzazione QR non è unica. Per convincersene basta considerare una qualsiasi matrice di fase, ovvero una matrice diagonale e unitaria $S \in \mathbb{C}^{m\times m}$

$$S = \begin{pmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_m \end{pmatrix},$$

dove $\theta_i \in \mathbb{C}$ e $|\theta_i| = 1$, per $i = 1, \ldots, m$. Se vale $A = QR$, allora vale anche $A = QS \cdot S^H R$, dove QS è unitaria e $S^H R$ è trapezoidale superiore.

Premettiamo alla dimostrazione del teorema di esistenza della fattorizzazione QR la definizione e alcune proprietà dei riflettori elementari che, come vedremo, possono essere utilizzati per annullare un blocco di componenti di un vettore.

DEFINIZIONE 1.2. *Dato* $\mathbf{u} \in \mathbb{C}^m$*, si definisce riflettore elementare (di Householder) associato ad* $\mathbf{u}$ *la seguente matrice* $U^{(m)} \in \mathbb{C}^{m\times m}$*,*

$$U^{(m)} = I_m - 2\frac{\mathbf{u}\mathbf{u}^H}{\|\mathbf{u}\|_2^2}. \tag{1.2}$$

I riflettori elementari hanno le seguenti proprietà.

PROPOSIZIONE 1.3. *Un riflettore elementare è una matrice unitaria, invertibile e hermitiana.*

DIMOSTRAZIONE. La dimostrazione, che consiste essenzialmente nella verifica dell'unitarietà e dell'hermitianità della matrice in (1.2), è lasciata per esercizio. □

PROPOSIZIONE 1.4. *Se il riflettore elementare in (1.2) viene moltiplicato per un vettore* $\mathbf{x} \in \mathbb{C}^m$*,* $\mathbf{x} \neq \mathbf{0}$*, ne opera la riflessione rispetto all'iperpiano* $\text{span}(\mathbf{u})^\perp = \{\mathbf{v} \in \mathbb{C}^m : \mathbf{v}^H\mathbf{u} = 0\}$.

DIMOSTRAZIONE. Decomponendo il vettore non nullo $\mathbf{x}$ come $\mathbf{x} = \alpha\mathbf{u} + \mathbf{w}$, dove $\alpha \in \mathbb{C}$ e $\mathbf{w} \in \text{span}(\mathbf{u})^\perp$, è immediato verificare che si ha $U^{(m)}\mathbf{x} = -\alpha\mathbf{u} + \mathbf{w}$. Dunque il vettore $U^{(m)}\mathbf{x}$ è il riflesso di $\mathbf{x}$ rispetto all'iperpiano $\text{span}(\mathbf{u})^\perp$. □

PROPOSIZIONE 1.5. *Il riflettore elementare in (1.2), scelto* $\mathbf{u} = \mathbf{x} + \sigma\mathbf{e}_1^{(m)}$ *(dove* $\mathbf{e}_1^{(m)}$ *denota il primo vettore della base canonica di* $\mathbb{C}^m$*,* $\mathbf{e}_1^{(m)} = (1, 0, \ldots, 0)^T$*), con* $\sigma = \|\mathbf{x}\|_2 \, e^{i\theta} \in \mathbb{C}$*,* $\theta \in [0, 2\pi)$*, fa sì che*

$$U^{(m)}\mathbf{x} = -\sigma\mathbf{e}_1^{(m)}.$$

Omettiamo la dimostrazione, che si riduce comunque solo a qualche conto.

Osserviamo esplicitamente che il modulo di σ deve essere obbligatoriamente uguale a $\|\mathbf{x}\|_2$ (in quanto deve valere $\left\|U^{(m)}\mathbf{x}\right\|_2 = \|\mathbf{x}\|_2$ perché $U^{(m)}$ è unitaria).

Per quanto riguarda l'arbitrarietà dell'argomento θ di σ, osserviamo che, nelle applicazioni, θ viene generalmente scelto uguale all'argomento di x_1, prima componente di $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$. Tale scelta permette di ridurre eventuali problemi di instabilità. Per convincersene, si consideri inizialmente il caso reale: se il segno di $\sigma = \pm\|\mathbf{x}\|_2$ viene scelto uguale a quello di x_1, allora il calcolo della prima componente di $\mathbf{u}$, ovvero $x_1 + \sigma$, si riduce a $\pm(|x_1| + \|\mathbf{x}\|_2)$ e dunque non può essere soggetto a errori dovuti alla cancellazione numerica di cifre significative.

Ricordiamo infatti che l'operazione di macchina $a + b$, di due numeri reali x e y molto vicini tra loro in valore assoluto e di segno opposto, può risultare affetta da un ingente errore di *roundoff*, potendo l'errore relativo presente nella rappresentazione di x e y subire la pericolosa amplificazione $(|x| + |y|)/|x + y|$.

È poi immediato estendere tali considerazioni al caso generale, optando per il calcolo di $x_1 + \sigma = (|x_1| + \|\mathbf{x}\|_2)e^{i\arg x_1}$. Infine, nel caso che la prima componente del vettore $\mathbf{x}$ sia nulla, per convenzione si sceglie il segno positivo (ovvero $\theta = 0$).

DIMOSTRAZIONE DEL TEOREMA 1.1. La dimostrazione consiste nella costruzione degli n riflettori elementari $U_1, U_2, \ldots, U_n$ tali che

$$A = U_1 U_2 \cdots U_n R,$$

dove R è una matrice trapezoidale superiore. Definiamo eslicitamente il riflettore elementare U_k che figurerà al k-esimo passo del processo di fattorizzazione QR della matrice A:

$$U_k = \begin{pmatrix} I_{k-1} & \mathbf{0} \\ \mathbf{0} & U^{(m+1-k)} \end{pmatrix} \in \mathbb{C}^{m\times m},$$

dove $U^{(m+1-k)}$ è il riflettore elementare di ordine $m+1-k$ associato al vettore

$$\mathbf{u}_k = (x_k, \ldots, x_m)^T + \sigma_k \mathbf{e}_1^{(m+1-k)} \in \mathbb{C}^{m+1-k}, \quad \sigma_k = \|(x_k, \ldots, x_m)\|_2 \, \mathrm{e}^{\mathrm{i}\theta_k},$$

con $\theta_k \in [0, 2\pi)$; in tal modo, per la Proposizione 1.5, vale

$$U^{(m+1-k)}(x_k, \ldots, x_m)^T = -\sigma_k \mathbf{e}_1^{(m+1-k)}.$$

In conclusione, U_k risulta essere il riflettore elementare associato al vettore

$$\mathbf{u} = \begin{pmatrix} \mathbf{0} \\ \mathbf{u}_k \end{pmatrix} \in \mathbb{C}^m,$$

e per costruzione, se applicato al vettore $\mathbf{x}$ non ne altererà le prime $k-1$ componenti e ne annullerà tutte le componenti da $k+1$ a m.

Il ruolo chiave del vettore $\mathbf{x}$ viene giocato dalla k-esima colonna della matrice $U_{k-1}\cdots U_1 A$ (quindi per U_1 dalla prima colonna della matrice A).

In maggior dettaglio, al k-esimo passo del processo di fattorizzazione della matrice A, il riflettore elementare U_k agisce sulle singole colonne della matrice $U_{k-1}\cdots U_1 A$ lasciandone inalterate le prime $k-1$ componenti e modificandone le successive $m-k+1$. In particolare, a seguito di questa pre-moltiplicazione, la k-esima colonna (che è, come abbiamo detto, il vettore $\mathbf{x}$ associato al riflettore elementare U_k) si troverà nella sua forma definitiva di colonna k-esima della matrice R, ovvero di vettore con le ultime $m-k$ componenti nulle; al contrario, le colonne da 1 a $k-1$ di $U_{k-1}\cdots U_1 A$, non contenendo componenti non nulle dalla k-esima componente in poi, non verranno alterate.

Dovremo costruire n riflettori elementari $U_1, U_2, \ldots, U_n$ in modo tale che

$$U_n \cdots U_2 U_1 A = R,$$

dove R è trapezoidale superiore. Si avrà quindi

$$A = (U_n \cdots U_2 U_1)^{-1} R = U_1^{-1} U_2^{-1} \cdots U_n^{-1} R = U_1^H U_2^H \cdots U_n^H R = U_1 U_2 \cdots U_n R = QR.$$

□

OSSERVAZIONE 1.1. *Si dimostra che il costo della fattorizzazione QR di A con l'algoritmo che abbiamo descritto risulta essere di $2n^2(m-n/3) + \mathcal{O}(m^2)$ operazioni.*

Ricordiamo che, se la matrice assegnata A è quadrata (ovvero se $m = n$), nella fattorizzazione LU - che esiste ed è unica se e solo se le sottomatrici principali di A sono invertibili - si hanno $n-1$ matrici triangolari inferiori $L_1, L_2, \ldots, L_{n-1}$ tali che

$$L_{n-1} \cdots L_2 L_1 A = U.$$

Analogamente nella fattorizzazione QR si dovranno costruire $n-1$ riflettori elementari $U_1, U_2, \ldots, U_{n-1}$ (infatti non ne servono n se A è quadrata) in modo tale che

$$U_{n-1} \cdots U_2 U_1 A = R,$$

dove R è triangolare superiore. Il numero di operazioni che servono per la fattorizzazione QR di A è $4n^3/3 + \mathcal{O}(n^2)$, il doppio quindi di quelle che garantiscono la fattorizzazione LU.

Per quanto riguarda la stabilità della fattorizzazione, per gli elementi di Q e R vale

$$\max_{i,j} |q_{ij}| \leq 1 \qquad \max_{i,j} |r_{ij}| \leq \sqrt{n} \max_{i,j} |a_{ij}|.$$

L'algoritmo di fattorizzazione QR è dunque stabile *in senso debole*, in quanto la maggiorazione di R coinvolge una costante dipendente dalla dimensione di A. D'altro canto, confrontando con i risultati di stabilità relativi alla fattorizzazione LU con pivoting, dato che $\sqrt{n} \ll 2^{n-1}$, si ha per l'algoritmo appena descritto garanzia di maggiore stabilità.

Se A è di rango massimo, a partire dalla fattorizzazione completa si possono ricavare le matrici della fattorizzazione *ridotta*

$$A = Q_n R_n,$$

dove $Q_n \in \mathbb{C}^{m \times n}$ ha colonne ortonormali e $R_n \in \mathbb{C}^{n \times n}$ è una matrice triangolare superiore. Di fatto, usando notazioni MATLAB, $Q_n = Q(:, 1:n)$ e $R_n = R(1:n, :)$. Q_n forma una base ortonormale per l'immagine di A. Più in generale, Q_k forma una base ortonormale per lo spazio generato dalle prime k colonne di A (il fatto che la k-esima colonna di A dipenda solo dalle prime k colonne di Q è in linea con la forma triangolare di R).

OSSERVAZIONE 1.2. *Se A è di rango massimo e addizionalmente si richiede che gli elementi diagonali della matrice R_n (necessariamente non nulli) siano positivi, allora Q_n e R_n sono univocamente determinati - ma non $Q(:, n+1:m)$ - inoltre R_n coincide con il fattore triangolare superiore della fattorizzazione di Cholesky della matrice hermitiana definita positiva $A^H A$, ovvero si ha $A^H A = R_n^H R_n = L L^H$.*

La fattorizzazione QR ridotta di A, che abbiamo assunto di rango massimo, può essere generata in alternativa con l'algoritmo di *ortogonalizzazione di Gram-Schmidt*, che risulta però essere meno stabile del precedente. Questa strategia consiste nel calcolare una base ortonormale per le colonne della matrice A. Denotiamo con $\mathbf{a}_1, \ldots, \mathbf{a}_n$ i vettori colonna di A, che sono una base per l'immagine di A.

Poniamo inizialmente

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}$$

e, procedendo per $k = 1, \dots, n-1$, calcoliamo iterativamente

$$\widetilde{\mathbf{q}}_{k+1} = \mathbf{a}_{k+1} - \sum_{j=1}^{k} \mathbf{q}_j^H \mathbf{a}_{k+1}\, \mathbf{q}_j \quad \Longrightarrow \quad \mathbf{q}_{k+1} = \frac{\widetilde{\mathbf{q}}_{k+1}}{\|\widetilde{\mathbf{q}}_{k+1}\|_2}.$$

Quindi le colonne $\mathbf{a}_1, \dots, \mathbf{a}_n$ potranno essere scritte come combinazione lineare dei vettori $\mathbf{q}_1, \dots, \mathbf{q}_n$ che costituiscono una base ortonormale di range(A). In particolare, vale $\mathbf{a}_k = r_{1k}\mathbf{q}_1 + \cdots + r_{kk}\mathbf{q}_k$ con $k = 1, \dots, n$ - per opportuni coefficienti r_{ij}, $i \le j$:

$$\left(\begin{array}{c|c|c} & & \\ \mathbf{a}_1 & \cdots & \mathbf{a}_n \\ & & \end{array} \right) = \left(\begin{array}{c|c|c} & & \\ \mathbf{q}_1 & \cdots & \mathbf{q}_n \\ & & \end{array} \right) \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}.$$

1.3. Soluzione del problema ai minimi quadrati con la fattorizzazione QR

Osserviamo che, facendo uso della fattorizzazione QR (completa), possiamo scrivere

$$A^H(A\mathbf{x} - \mathbf{b}) = R^H Q^H (A\mathbf{x} - \mathbf{b}) = R^H (Q^H Q R\mathbf{x} - Q^H \mathbf{b}) = R^H (R\mathbf{x} - Q^H \mathbf{b}),$$

dove

$$R = \begin{pmatrix} R_n \\ 0 \end{pmatrix}, \quad Q^H \mathbf{b} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}, \quad \text{con } R_n \in \mathbb{C}^{n\times n},\ \mathbf{c}_1 \in \mathbb{C}^n,\ \mathbf{c}_2 \in \mathbb{C}^{m-n}.$$

Il problema si riduce quindi alla soluzione del sistema triangolare $R_n\mathbf{x} = \mathbf{c}_1$, dove $\mathbf{c}_1 = Q_n^H \mathbf{b}$. Dato che A è di rango massimo, abbiamo la garanzia che la matrice R_n è invertibile. Per il calcolo di $\mathbf{x}^*$ sarebbe sufficiente la fattorizzazione QR ridotta. D'altro canto, la fattorizzazione completa è necessaria per calcolare $\mathbf{c}_2$, che servirà per il calcolo di γ. Infatti, dato che $\|A\mathbf{x}^* - \mathbf{b}\|_2 = \|QR\mathbf{x}^* - \mathbf{b}\|_2 = \|R\mathbf{x}^* - Q^H\mathbf{b}\|_2$, si ha $\gamma = \|\mathbf{c}_2\|_2$.