RISOLUZIONE DI SISTEMI LINEARI: METODI ITERATIVI (Il metodo del gradiente e i test di arresto)

Nella lezione precedente abbiamo introdotto il metodo di Richardson, sia nella versione in cui il parametro di accelerazione α è scelto costante per tutte le iterazioni, sia in quella in cui il parametro di accelerazione è variabile ad ogni iterazione k, generalizzando così ulteriormente il metodo mediante l'introduzione di una scelta dinamica del parametro α_k.
In questa lezione esamineremo i criteri di scelta del parametro α_k, che porteranno alla formulazione dei metodi di tipo gradiente. Inoltre, nella parte finale della lezione, introdurremo due criteri di arresto, basati rispettivamente sul residuo e sull'incremento, per decidere quando interrompere il processo iterativo.
Il metodo di Richardson dinamico introduce un dinamismo ulteriore rispetto al classico metodo di Richardson. Mentre Richardson già prevedeva un parametro che permetteva di accelerare la convergenza, qui introduciamo un parametro che possiamo variare ad ogni iterazione. Ovviamente, il problema che sorge è come determinare questo parametro. In altre parole, come fissare in maniera automatica la scelta del parametro α_k per ogni iterazione k. Vediamo quindi i criteri di scelta del parametro α_k.
Se P ed A sono matrici SDP, possiamo scegliere α_k in modo tale che l'errore che generiamo ad ogni iterazione sia minimo

$$\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\right\|_A \quad \text{sia minima.}$$

Più in particolare, vogliamo che la norma A (vedremo subito cos'è la norma A che rappresenta una particolare norma) di questo errore sia la più piccola possibile. Se osserviamo

$$\begin{cases} \boldsymbol{P}\boldsymbol{z}^{(\boldsymbol{k})} = \boldsymbol{r}^{(\boldsymbol{k})} \\ \boldsymbol{x}^{(\boldsymbol{k+1})} = \boldsymbol{x}^{(\boldsymbol{k})} + \alpha_k \boldsymbol{z}^{(\boldsymbol{k})} \\ \boldsymbol{r}^{(\boldsymbol{k+1})} = \boldsymbol{r}^{(\boldsymbol{k})} - \alpha_k A\boldsymbol{z}^{(\boldsymbol{k})} \end{cases},$$

notiamo subito che $\boldsymbol{x}^{(\boldsymbol{k+1})}$ dipende da α_k ed in particolare la differenza

$$\boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})} = \alpha_k \boldsymbol{z}^{(\boldsymbol{k})},$$

dipenderà da α_k. Quindi anche la differenza dalla soluzione esatta $\boldsymbol{x}$ dipenderà da α_k

$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x} \;\rightarrow\; \text{dipende da } \alpha_k.$$

La richiesta che quindi vorremmo soddisfare è quella di scegliere α_k (tra tutti i possibili α_k) come quello che minimizza la distanza $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}$ nella norma A, ovvero che $\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\right\|_A$ sia minimo. La norma A è chiamata norma dell'energia e si definisce nel seguente modo

$$\|\boldsymbol{a}\|_A = (\boldsymbol{a}^T A \boldsymbol{a})^{\frac{1}{2}} \qquad \forall \boldsymbol{a} \in \mathbb{R}^n.$$

Notiamo che, se fosse la norma euclidea, A sarebbe l'identità I. Se A è la matrice del sistema lineare di partenza, simmetrica e definita positiva, questa $\|\cdot\|_A$ è una nuova norma. Corrispondentemente, potremo definire anche una norma di matrice associata alla norma vettoriale $\|\cdot\|_A$.
Inoltre, osserviamo che la norma A al quadrato di $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}$

$$\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\right\|_A^2$$

è il prodotto scalare fra due vettori

$$\left(A\left(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right),\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right),$$

il vettore $A\left(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right)$, ovvero l'errore pre-moltiplicato per A e l'errore $\boldsymbol{x}^{(k+1)}-\boldsymbol{x}$. Definiamo questo prodotto scalare (che è numero reale positivo) come $F(\alpha_k)$

$$F(\alpha_k)=\left(A\left(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right),\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right)=\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right\|_A^2,$$

ovvero il valore assunto da una funzione F in corrispondenza di questo parametro α_k. Ricordiamo che il parametro α_k è "nascosto" in questa definizione di $F(\alpha_k)$: in precedenza abbiamo visto che esso è incluso nella definizione di $\boldsymbol{x}^{(k+1)}$.
Con questa ultima definizione di $F(\alpha_k)$, abbiamo ricondotto la norma dell'errore al valore assunto da una funzione F in corrispondenza di α_k.
Per richiedere che $\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\right\|_A$ sia minima, dovremmo evidentemente richiedere che la funzione $F(\alpha_k)$ abbia il suo minimo nel punto α_k. Pertanto, questa funzione deve essere minima in α_k, il che implica che

$$F'(\alpha_k)=0.$$

Si può vedere che questa funzione è, di fatto, una parabola rispetto al suo argomento. Una parabola è sempre positiva, quindi è ovviamente derivabile, e nell'estremo la sua derivata prima dovrà essere uguale a zero. Pertanto, prendendo l'espressione di prima e derivando rispetto ad α_k, si ottiene

$$\alpha_k=\frac{\left(\boldsymbol{r}^{(k)}\right)^T\boldsymbol{z}^{(k)}}{(\boldsymbol{z}^{(k)})^TA\boldsymbol{z}^{(k)}},$$

Osserviamo che, se è nota la soluzione al passo k, sono noti anche il residuo $\boldsymbol{r}^{(k)}$ e il residuo precondizionato $\boldsymbol{z}^{(k)}$. Pertanto, possiamo calcolare il prodotto scalare tra questi due vettori. Successivamente, dividiamo per il prodotto scalare tra $\boldsymbol{z}^{(k)}$e $A\boldsymbol{z}^{(k)}$, che sono tutte quantità note. In questo modo, possiamo determinare α_k. Il valore α_k sarà quindi il coefficiente di accelerazione che utilizzeremo per il metodo di Richardson dinamico.

Abbiamo così individuato un criterio automatico per la selezione di α_k. Tutte queste quantità sono facilmente calcolabili tramite semplici operazioni di prodotto scalare tra vettori, e il valore di α_k, così trovato, non dipende dalla conoscenza degli autovalori della matrice. Questo rappresenta, dunque, un criterio di scelta completamente automatico.
Il metodo di Richardson dinamico con questa scelta

$$\alpha_k=\frac{\left(\boldsymbol{r}^{(k)}\right)^T\boldsymbol{z}^{(k)}}{(\boldsymbol{z}^{(k)})^TA\boldsymbol{z}^{(k)}},$$

è conosciuto come metodo di discesa più ripida (o discesa verso il valore zero dell'errore) o come metodo del gradiente precondizionato. In questo caso, stiamo utilizzando una matrice P, che funge da precondizionatore. Se il precondizionatore P fosse uguale all'identità I (cioè $P=I$) e quindi non avessimo alcun precondizionamento, si otterrebbe il metodo del gradiente. Pertanto, il metodo del gradiente precondizionato che abbiamo introdotto rappresenta una generalizzazione del metodo del gradiente.
Nel caso del metodo del gradiente precondizionato, l'errore si riduce secondo questa legge

$$\left\|\boldsymbol{e}^{(k)}\right\|_A\leq\left(\frac{K_2(P^{-1}A)-1}{K_2(P^{-1}A)+1}\right)^k\left\|\boldsymbol{e}^{(0)}\right\|_A,$$

ovvero la norma dell'errore al passo k in norma A è minore o uguale a una certa costante elevata alla k per la norma A dell'errore al passo 0. Il metodo, quindi, dovrà convergere se questa costante $\frac{K_2(P^{-1}A)-1}{K_2(P^{-1}A)+1}$ è minore di 1. Ma questo è sempre vero, poiché il numero di condizionamento in norma 2 è un numero reale positivo (sempre maggiore o uguale a uno). Inoltre, al numeratore a K_2 viene sottratto 1, mentre al denominatore viene sommato 1. Quindi questa costante $\frac{K_2(P^{-1}A)-1}{K_2(P^{-1}A)+1}$, che è sempre minore di uno, elevata a k, tenderà a zero per k che tende all'infinito. Pertanto, qualunque sia l'errore iniziale, esso verrà ridotto a zero. Questo significa che l'errore al passo k, per $k \to \infty$, deve tendere a zero.
A questo punto, è importante riprendere il discorso della scelta del precondizionatore. Abbiamo visto, nelle lezioni precedenti, come la scelta del precondizionatore P si ispiri a due criteri: il primo è quello di essere semplice per poter risolvere sistemi lineari più agevol nel precondizionatore P; il secondo è quello di assicurare la convergenza più rapida possibile.
Vediamo effettivamente l'incidenza di questo discorso nella velocità di convergenza. Andiamo quindi a vedere in corrispondenza di questo parametro α_k ottimale qual è la velocità di convergenza del gradiente precondizionato. Ovvero qual è il raggio spettrale della matrice di iterazione. Osservando la formula dell'errore

$$\left\|\boldsymbol{e}^{(k)}\right\|_A \leq \left(\frac{K_2(P^{-1}A)-1}{K_2(P^{-1}A)+1}\right)^k \left\|\boldsymbol{e}^{(0)}\right\|_A,$$

notiamo che più $P^{-1}A$ ha autovalori simili ad 1 più il numeratore sarà vicino a zero e quindi più rapida sarà la convergenza del metodo. Questo, dunque, spiega perché questa P si chiama precondizionatore: P deve precondizionare A, ovvero deve assicurare che il condizionamento di $P^{-1}A$ sia molto più piccolo di quanto non lo fosse il condizionamento della matrice A.

Si può dimostrare che il metodo del gradiente precondizionato assicura la massima riduzione di questo funzionale

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b},$$

che è noto come funzionale energia.
Possiamo fornire un'interpretazione geometrica. Consideriamo un paraboloide come il funzionale energia $\Phi(\boldsymbol{x})$, che, essendo dipendente da un vettore $\boldsymbol{x}$, è un funzionale (e non una funzione). Il funzionale $\Phi(\boldsymbol{x})$ ha un punto di minimo che corrisponde alla soluzione del sistema lineare di partenza

$$A\boldsymbol{x} = \boldsymbol{b}.$$

Pertanto, il metodo del gradiente precondizionato dovrà garantire la convergenza più rapida possibile verso questo punto di minimo. In effetti, la soluzione del sistema è proprio il punto in cui il funzionale energia $\Phi(\boldsymbol{x})$ assume il valore minimo. In effetti, risolvere il sistema lineare equivale a trovare il punto di energia minima del funzionale $\Phi(\boldsymbol{x})$.

L'interpretazione geometrica del metodo del gradiente (si veda Figura 1) è la seguente: dato un punto $\boldsymbol{x}^{(k)}$, calcoliamo il valore della funzionale energia funzionale $\Phi\left(\boldsymbol{x}^{(k)}\right)$ e prendiamo la corrispondente isosuperficie (cioè, la sezione del funzionale $\Phi(\boldsymbol{x}) = \Phi\left(\boldsymbol{x}^{(k)}\right)$), che rappresenta un livello del funzionale (ovvero, stiamo tagliando il funzionare in maniera orizzontale).

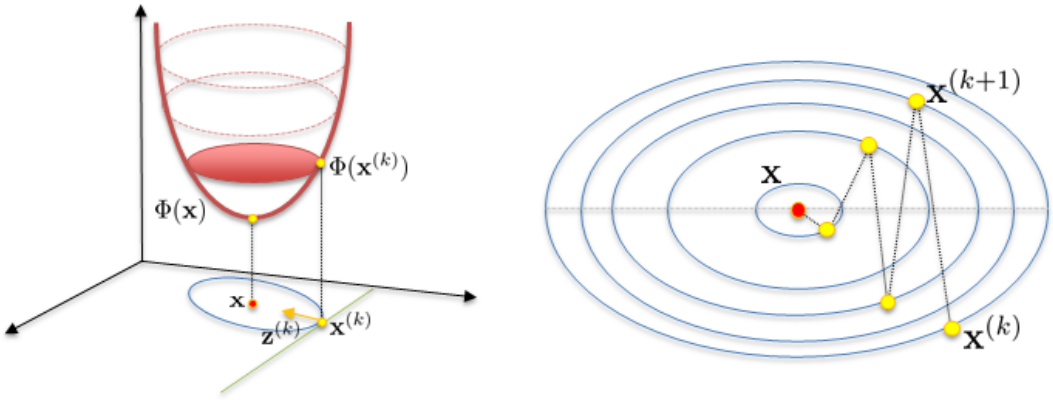

Figura 1: Interpretazione geometrica del metodo del gradiente.

Proiettando questa sezione sullo spazio $\mathbb{R}^n$, otteniamo il bordo di questa proiezione. Su questa proiezione, a partire dal punto $\boldsymbol{x}^{(k)}$, tracciamo la tangente a questa proiezione e ci muoviamo in direzione perpendicolare alla tangente. La direzione perpendicolare alla tangente è rappresentata dal vettore $\boldsymbol{z}^{(k)}$. Pertanto, ci spostiamo lungo direzioni perpendicolari, puntando verso il punto $\boldsymbol{x}$, che è la soluzione del sistema.

Possiamo esaminare più nel dettaglio cosa succede nel piano e determinare i criteri di scelta delle nuove iterate. Considerando il punto $\boldsymbol{x}^{(k)}$, la soluzione $\boldsymbol{x}$ e la curva di proiezione sul piano associata a $\Phi(\boldsymbol{x}^{(k)})$, che sarà una sorta di ellissoide, possiamo determinare il nuovo punto $\boldsymbol{x}^{(k+1)}$utilizzando la formula del metodo di Richardson. Successivamente, tracciamo la curva-proiezione associata a $\Phi(\boldsymbol{x}^{(k+1)})$ e applichiamo di nuovo il metodo per determinare il punto successivo $\boldsymbol{x}^{(k+2)}$. Continuando iterativamente, otteniamo i punti$\boldsymbol{x}^{(k+3)}$, $\boldsymbol{x}^{(k+4)}$, e così via, fino **a** raggiungere la convergenza al punto soluzione $\boldsymbol{x}$.

Questo percorso assicura che, ad ogni iterazione, ci spostiamo verso la soluzione ottimizzando il funzionale energia, riducendo progressivamente il valore di $\Phi(\boldsymbol{x})$.

Vogliamo ora esaminare come introdurre dei criteri per l'arresto dei metodi iterativi. Abbiamo visto ripetutamente che costruire un metodo iterativo equivale a generare una successione di vettori che, al limite, convergerà alla soluzione del problema di partenza, ovvero alla soluzione $\boldsymbol{x}$ del sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$. È pertanto necessario stimare il punto in cui conviene interrompere questa successione di iterate, ossia individuare il minimo n per il quale possiamo fermarci, e a questo punto considerare $\boldsymbol{x}^{(n)}$ come la candidata per rappresentare la soluzione del sistema.

Esaminiamo dunque i cosiddetti test di attesto, ovvero i criteri che possiamo adottare per arrestare il processo iterativo.

Un primo criterio è quello basato sull'analisi del residuo. Supponiamo di essere arrivati al passo $\boldsymbol{x}^{(k)}$, allora possiamo calcolare il residuo $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ e la sua relativa norma

$$\|\boldsymbol{r}^{(k)}\| = \|\boldsymbol{b} - A\boldsymbol{x}^{(k)}\|.$$

La richiesta per il test di arresto sul residuo è che tale norma sia inferiore a ϵ, dove ϵ rappresenta una tolleranza fissata a priori

$$\|\boldsymbol{r}^{(k)}\| < \epsilon.$$

Possiamo quindi arrestare il metodo iterativo non appena troviamo un valore $\boldsymbol{x}^{(k)}$ che permetta di verificare questa relazione sul residuo. Inoltre, abbiamo già visto che il residuo permette di passare alla formula dell'errore attraverso il condizionamento di A. Più precisamente abbiamo mostrato che

$$\frac{\|\boldsymbol{e}^{(k)}\|}{\|\boldsymbol{x}\|} \leq K(A)\frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{b}\|}.$$

Dunque, questo test risulta conveniente quando il numero di condizionamento $K(A)$ è piccolo, ovvero vicino a 1, poiché in tale contesto un residuo ridotto comporta un errore altrettanto contenuto. Di conseguenza, il primo test d'arresto basato sul residuo funziona efficacemente se il numero di condizionamento della matrice è ragionevolmente basso.

Si può considerare l'impiego di ulteriori test di arresto. È fondamentale che tali test siano effettivamente applicabili, ossia che dipendano esclusivamente da quantità già calcolate. Una quantità sicuramente a disposizione, al passo k, è rappresentata dal vettore $\boldsymbol{x}^{(k)}$. Al passo $k+1$, è quindi possibile confrontare la soluzione ottenuta con quella del passo precedente

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

e richiedere che questa differenza in norma sia minore di ϵ, dove ϵ è una tolleranza fissata a priori $(10^{-3}, 10^{-4}, 10^{-6}, \ldots)$. Se questa disuguaglianza è soddisfatta allora si arresta il metodo iterativo e si prende $\boldsymbol{x}^{(k+1)}$ come candidata a rappresentare la soluzione del sistema lineare. Questo test è chiamato test di arresto sull'incremento.

Osserviamo che la relazione che lega fra loro gli errori è

$$\boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)},$$

dove B è la matrice di iterazione. Allora se γ è una norma della matrice B

$$\gamma = \|B\|,$$

risulta che

$$\|\boldsymbol{e}^{(k+1)}\| \leq \gamma\|\boldsymbol{e}^{(k)}\|,$$

ovvero una relazione che lega fra loro l'errore al passo k e l'errore al passo $k+1$. Inoltre, osserviamo che

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = \|(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - (\boldsymbol{x} - \boldsymbol{x}^{(k+1)})\|,$$

dove abbiamo aggiunto e sottratto a secondo membro $\boldsymbol{x}$, per poi raggruppare diversamente queste quantità. Sfruttando a questo punto una proprietà delle norme, ovvero che la norma della differenza di due vettori è maggiore o uguale della differenza delle norme

$$\|(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - (\boldsymbol{x} - \boldsymbol{x}^{(k+1)})\| \geq \|(\boldsymbol{x} - \boldsymbol{x}^{(k)})\| - \|(\boldsymbol{x} - \boldsymbol{x}^{(k+1)})\| = \|\boldsymbol{e}^{(k)}\| - \|\boldsymbol{e}^{(k+1)}\|,$$

e poiché $\|\boldsymbol{e}^{(k+1)}\| \leq \gamma\|\boldsymbol{e}^{(k)}\|$, è possibile affermare che $-\|\boldsymbol{e}^{(k+1)}\| \geq -\gamma\|\boldsymbol{e}^{(k)}\|$ e quindi

$$\|\boldsymbol{e}^{(k)}\| - \|\boldsymbol{e}^{(k+1)}\| \geq \|\boldsymbol{e}^{(k)}\| - \gamma\|\boldsymbol{e}^{(k)}\| = (1-\gamma)\|\boldsymbol{e}^{(k)}\|.$$

Abbiamo quindi trovato che

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| \geq (1-\gamma)\|\boldsymbol{e}^{(k)}\|.$$

Ripartendo quindi ora da questa disuguaglianza

$$\|\boldsymbol{e}^{(k+1)}\| \leq \gamma \|\boldsymbol{e}^{(k)}\|,$$

e sfruttando $(1-\gamma)\|\boldsymbol{e}^{(k)}\| \leq \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|$ da cui $\|\boldsymbol{e}^{(k)}\| \leq \frac{1}{1-\gamma}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|$, si ottiene

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

dato che si era imposto che la quantità $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|$ fosse minore di ϵ. Quindi

$$\|\boldsymbol{e}^{(k+1)}\| \leq \frac{\gamma}{1-\gamma}\epsilon.$$

Se γ è piccolo, cioè molto vicino a zero, il test di arresto sull'incremento risulta efficace, poiché in tal caso il rapporto $\frac{\gamma}{1-\gamma}$ non assume valori elevati e, di conseguenza, se ϵ è piccolo anche l'errore risulterà contenuto. Viceversa, se γ è positivo e prossimo a 1, il test fornisce una stima meno affidabile, perché, nonostante ϵ sia ridotto (ovvero la differenza tra due iterate sia piccola), il rapporto $\frac{\gamma}{1-\gamma}$ può essere elevato, implicando che l'errore potrebbe risultare significativo e, pertanto, il test di arresto in questo caso non sarebbe particolarmente attendibile.
Vediamo adesso a che cosa conduce il test di arresto sull'incremento nel caso del metodo di Richardson. Definendo come γ il valore ρ_{ott}, ovvero il raggio spettrale ottimale della matrice di interazione calcolato in corrispondenza dei α_{ott}, abbiamo che

$$\gamma = \|B\| = \|R_{\alpha_{ott}}\| = \rho_{ott} = \frac{K_2 - 1}{K_2 + 1},$$

dove $K_2 = K_2(P^{-1}A)$ è il numero di condizionamento della matrice precondizionata (ovvero il massimo sul minimo autovalore della matrice $P^{-1}A$). Si ha quindi che

$$\|\boldsymbol{e}^{(k+1)}\| \leq \frac{\gamma}{1-\gamma}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = \frac{\rho_{ott}}{1-\rho_{ott}}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = \frac{K_2 - 1}{2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|,$$

da cui

$$\|\boldsymbol{e}^{(k+1)}\| \leq \frac{K_2 - 1}{2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|.$$

Abbiamo quindi visto che, nell'analisi dei metodi iterativi, per stabilire un criterio di arresto si può considerare sia il residuo al passo k sia la differenza tra le iterate al passo k e al passo $k+1$. Abbiamo evidenziato vantaggi e svantaggi di ciascun approccio.

Concludiamo questa trattazione sui metodi iterativi per la risoluzione di sistemi lineari facendo un accenno ai cosiddetti sistemi sparsi, ovvero a quei sistemi in cui la matrice del sistema lineare contiene un numero di elementi diversi da zero dell'ordine della dimensione della matrice stessa. In una matrice $n \times n$, avente in linea di principio n^2 elementi, solo un numero dell'ordine di n risulta diverso da zero. In tali casi, essendoci una predominanza di elementi nulli rispetto a quelli non nulli, si parla di matrici sparse. Le matrici sparse si incontrano frequentemente nelle applicazioni, in particolare in quelle relative a problemi di equazioni differenziali ordinarie e alle derivate parziali che sono alla base delle applicazioni ingegneristiche.

In queste matrici sparse, la presenza di numerosi zeri rende superflue operazioni di moltiplicazione o addizione quando gli operandi sono nulli, ed è pertanto opportuno sfruttare nel modo più conveniente possibile la struttura di sparsità.

Vediamo alcuni esempi elementari di matrici sparse. Una matrice diagonale è, ovviamente, sparsa; una matrice tridiagonale, avendo solamente $3n$ elementi diversi da zero su n^2 elementi totali, è anch'essa sparsa; analogamente, una matrice pentadiagonale, con cinque diagonali contenenti elementi non nulli, possiede $5n$ elementi diversi da zero su n^2 totali. Per matrici sparse di questo tipo non sussistono particolari criticità, poiché la posizione degli elementi diversi da zero è nota a priori.

I problemi, invece, sorgono quando la matrice è sparsa in maniera arbitraria, ovvero quando non si conosce a priori la disposizione degli elementi non nulli. In generale, non si memorizzano tutti gli elementi delle matrici sparse, ma si ricorre a una memorizzazione compatta che consente di memorizzare esclusivamente gli elementi diversi da zero, in quanto è inutile allocare memoria per gli zeri.

Si osserva, tuttavia, che, nel caso in cui si utilizzino metodi di fattorizzazione per la matrice A, si assiste al fenomeno del "fill-in", o riempimento. Ciò significa che a una matrice sparsa può corrispondere, a seguito del processo di fattorizzazione, una matrice in cui posizioni precedentemente nulle vengono riempite con elementi diversi da zero. Tale fenomeno si verifica, ad esempio, nel metodo di eliminazione di Gauss tramite fattorizzazione LU. Per questo motivo, risulta importante identificare le zone in cui il fill-in avviene e, ancor più, minimizzare il suo effetto.