Corso di Analisi Numerica - AN410

Parte 2: metodi diretti per sistemi lineari

Roberto Ferretti

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DMF

- Richiami sulle norme e sui sistemi lineari
- Il Metodo di Eliminazione di Gauss
- Il Metodo di Eliminazione con pivoting
- Le fattorizzazioni LU e di Cholesky
- La fattorizzazione QR
- La stabilità dei metodi diretti per sistemi lineari

Richiami sulle norme e sui sistemi lineari

Una *norma* $\|\cdot\|$ è una applicazione da uno spazio vettoriale X in $\mathbb{R}$ che soddisfa le seguenti proprietà:

$$\|x\| \geq 0\ , \quad \|x\| = 0 \ \text{ se e solo se } \ x = 0;$$

$$\|cx\| = |c|\ \|x\| \quad (c \in \mathbb{R});$$

$$\|x + y\| \leq \|x\| + \|y\|.$$

Le tre norme di uso più comune in Analisi Numerica sono le norme euclidea, $\|\cdot\|_\infty$ e $\|\cdot\|_1$ definite da:

$$\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}, \quad \|x\|_\infty = \max_i |x_i|\,, \quad \|x\|_1 = \sum_i |x_i|$$

Se X è lo spazio degli operatori lineari limitati su uno spazio vettoriale Y, allora si richiede di regola che la norma soddisfi anche le ulteriori proprietà

$$\|AB\| \leq \|A\| \ \|B\| \quad \text{(submoltiplicatività);}$$
$$\|Ax\|_Y \leq \|A\| \ \|x\|_Y \quad \text{(compatibilità)}$$

e si indica come *norma naturale* (associata ad una certa norma su Y) la seguente norma su X:

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

In particolare, le tre norme matriciali naturali associate rispettivamente alle norme vettoriali euclidea, $\|\cdot\|_\infty$ e $\|\cdot\|_1$ sono:

$$\|A\|_2 = \rho(A^tA)^{1/2}\,, \quad \|A\|_\infty = \max_i \sum_j |a_{ij}|\,, \quad \|A\|_1 = \max_j \sum_i |a_{ij}|$$

dove si è indicato con $\rho(B) = \max_j |\lambda_j(B)|$ il *raggio spettrale* di una matrice. Un'altra norma matriciale compatibile con la norma euclidea è la norma di Frobenius

$$\|A\|_F = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$$

che però non è una norma naturale ($\|A\|_F \geq \|A\|_2$).

Sistema lineare: in forma compatta $Ax = b$, in forma estesa

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{1}$$

• E' noto che il sistema ammette soluzione unica a patto che la matrice A sia nonsingolare, ed esistono algoritmi risolutivi

• La complessità del metodo di Cramer è fattoriale, mentre eliminazione e sue varianti hanno complessità polinomiale

Esempio: numero di operazioni necessarie nei vari algoritmi per la soluzione di un sistema lineare 3×3:

- Con il metodo di Cramer la soluzione viene calcolata come

$$x_k = \frac{\Delta_k}{\Delta} \quad (k = 1, 2, 3)$$

dove Δ, Δ_k sono determinanti 3×3. Si effettuano in totale 3 quozienti + (4 determinanti)$\times$(6 termini)$\times$(2 moltiplicazioni + 1 somma) = 75 operazioni

• Con il metodo di eliminazione per diagonalizzazione (o di Gauss–Jordan) si porta il sistema nella forma

$$\begin{cases} \alpha_1 x_1 = \beta_1 \\ \alpha_2 x_2 = \beta_2 \\ \alpha_3 x_3 = \beta_3 \end{cases}$$

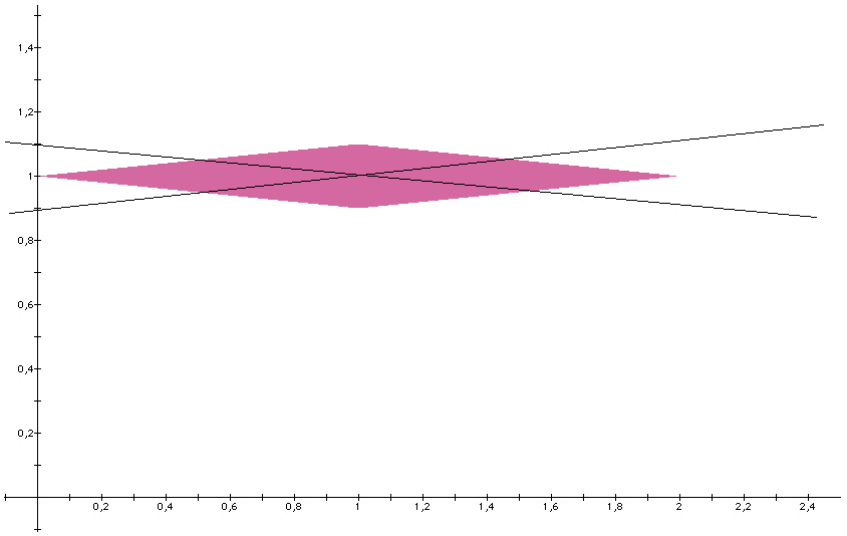
Ogni variabile deve essere eliminata, tramite combinazione lineare, da due equazioni, per un totale di (6 eliminazioni)$\times$(1 divisione + 3 prodotti + 3 somme) + 3 divisioni = 45 operazioni

• Con il metodo di eliminazione per triangolarizzazione si porta il sistema nella forma

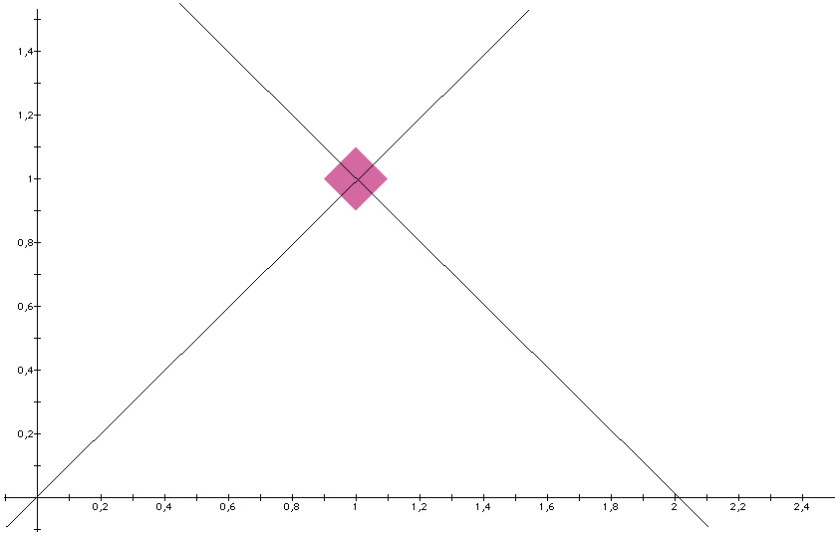
$$\begin{cases} \alpha_{11}x_1 + \alpha_{12}x_2 + \alpha_{13}x_3 = \beta_1 \\ \alpha_{22}x_2 + \alpha_{23}x_3 = \beta_2 \\ \alpha_{33}x_3 = \beta_3 \end{cases}$$

e si risolve il sistema triangolare partendo da x_3, per un totale di (3 eliminazioni)$\times$(1 divisione + 3 prodotti + 3 somme) + 1 divisione + (1 prodotto + 1 somma + 1 divisione) + (2 prodotti + 2 somme + 1 divisione) = 30 operazioni. Questa è la strada tipica che si segue numericamente.

La stabilità della soluzione di un sistema lineare rispetto alle perturbazioni sarà analizzata più nel dettaglio in seguito, ma ci si può aspettare che il condizionamento non sia buono se le righe della matrice A sono "quasi linearmente dipendenti" (cosa che in due dimensioni equivale geometricamente a cercare l'intersezione di due rette con coefficienti angolari simili).



cattivo condizionamento



buon condizionamento

Il Metodo di Eliminazione di Gauss

Si basa sul principio di generare mediante opportune combinazioni lineari di righe una sequenza di sistemi equivalenti

$$A^{(1)}x = b^{(1)} \rightarrow A^{(2)}x = b^{(2)} \rightarrow \cdots \rightarrow A^{(n)}x = b^{(n)}$$

l'ultimo dei quali è un sistema in forma triangolare

• E' l'algoritmo che presenta la minore complessità computazionale per matrici generiche, piene.

• In aritmetica finita la propagazione degli errori di arrotondamento può però diventare proibitiva in dimensione alta.

Il punto di arrivo del processo di eliminazione di variabili è un sistema triangolare

$$\begin{cases} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n = \beta_1 \\ \qquad\quad \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n = \beta_2 \\ \qquad\qquad\qquad\quad \vdots \\ \qquad\qquad\qquad\quad \alpha_{nn}x_n = \beta_n \end{cases} \tag{2}$$

La sua soluzione si puo' calcolare tramite le cosiddette *sostituzioni all'indietro* come

$$x_n = \frac{\beta_n}{\alpha_{nn}}\,, \qquad x_k = \frac{1}{\alpha_{kk}}\left(\beta_k - \sum_{j=k+1}^{n} \alpha_{kj}x_j\right) \qquad (k = n-1, \ldots, 1)$$

in cui il valore di una incognita viene ottenuto sulla base di quelle (successive) già calcolate.

Per arrivare alla forma triangolare si parte dal sistema iniziale, che riscriviamo come $A^{(1)}x = b^{(1)}$, o per esteso:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \qquad \vdots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)}. \end{cases} \tag{3}$$

Supponendo che $a_{11}^{(1)} \neq 0$, la eliminazione della variabile x_1 si effettua sommando alla riga k–esima la prima moltiplicata per $-a_{k1}^{(1)}/a_{11}^{(1)}$. Dopo $n-1$ combinazioni lineari così costruite, la variabile x_1 sarà presente solo nella prima equazione.

Al secondo passo di eliminazione, il sistema sarà nella forma

$$\begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ \qquad\qquad a_{22}^{(2)} x_2 + \cdots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \qquad\qquad\qquad \vdots \\ \qquad\qquad a_{n2}^{(2)} x_2 + \cdots + a_{nn}^{(2)} x_n = b_n^{(2)}. \end{cases} \tag{4}$$

Supponendo che $a_{22}^{(2)} \neq 0$, per eliminare la variabile x_2 dalle ultime $n-2$ equazioni si somma alla riga k–esima la seconda moltiplicata per $-a_{k2}^{(2)}/a_{22}^{(2)}$. Dopo questa operazione la variabile x_2 comparirà solo nelle prime due equazioni (e così via...).

Naturalmente, non è sempre vero che l'elemento (detto *pivot*) $a_{kk}^{(k)}$ sia non nullo.

• Ciò è vero se e solo se il minore principale di ordine k è nonsingolare (ad esempio, se $A > 0$)

• Succede però che se $\det A \neq 0$, almeno uno tra gli elementi $a_{ik}^{(k)}$ per $i > k$ sarà non nullo, quindi per proseguire l'algoritmo basta scambiare la riga k–esima con la riga su cui compare il pivot non nullo

• In aritmetica finita non basta scegliere un pivot non nullo, e la sua scelta ha forti conseguenze sulla accuratezza del risultato

Esempio: il punto $(10, 1)$ è soluzione del sistema

$$\begin{cases} 70x_1 + 700x_2 = 1400 \\ 3x_1 + 31x_2 = 61 \end{cases}$$

Risolvendo il sistema per eliminazione con aritmetica a tre cifre significative, il moltiplicatore relativo ad x_1 vale $3/70 = 0.0429$ in aritmetica finita, ed il sistema triangolarizzato diviene

$$\begin{cases} 70x_1 + 700x_2 = 1400 \\ x_2 = 1 \end{cases}$$

che ha la soluzione corretta.

Se invece si scambiano le righe il sistema ha la forma

$$\begin{cases} 3x_1 + 31x_2 = 61 \\ 70x_1 + 700x_2 = 1400 \end{cases}$$

Il moltiplicatore relativo ad x_1 vale ora $70/3 = 23.3$ in aritmetica finita, ed il sistema triangolarizzato diviene

$$\begin{cases} 3x_1 + 31x_2 = 61 \\ \quad -22x_2 = -20 \end{cases}$$

che ha soluzione (sempre nell'aritmetica a tre cifre significative) $x_1 = 10.9$, $x_2 = 0.909$ (cioè un errore dell'ordine del 9%).

Effetti che riducono la precisione del metodo di eliminazione:

• Il sistema di arrivo è solo approssimativamente triangolare (ad esempio, nel secondo caso si ha in realtà $a_{21}^{(2)} = 0.01$)

• Nella somma di numeri di ordini di grandezza diversi, il numero più piccolo perde cifre significative a causa dell'aumento dell'esponente

• Coefficienti di una equazione ottenuti come differenza di numeri "grandi" presentano un aumento dell'errore relativo rispetto ai coefficienti da cui provengono (queste ultime due situazioni sono più probabili con moltiplicatori grandi)

Complessità del metodo di eliminazione:

• La soluzione del sistema triangolare richiede per la k–esima variabile $n-k$ somme ed altrettanti prodotti. Il numero di operazioni è quindi

$$2+4+6+\cdots+2(n-1)=2O\left(\frac{n^2}{2}\right)=O(n^2)$$

• La fase di triangolarizzazione del sistema richiede, per eliminare la variabile k–esima, $(n-k)^2$ prodotti ed altrettante somme. La complessità di questa fase (è quella prevalente) è quindi di

$$2(n-1)^2+2(n-2)^2+\cdots+8+2=2O\left(\frac{n^3}{3}\right)=O\left(\frac{2n^3}{3}\right)$$

indice

Il Metodo di Eliminazione con pivoting

Nella strategia di *pivoting parziale*, che è basata su permutazioni delle sole righe, al passo i–esimo di eliminazione viene portata in i–esima posizione l'equazione j–esima (con $j \geq i$), dove

$$|a_{ji}^{(i)}| = \max_{k \geq i} |a_{ki}^{(i)}|.$$

• La complessità di questa operazione è lineare per ogni passo di eliminazione, perciò (poiché un passo di eliminazione opera su tutta una sottomatrice ed ha quindi complessità quadratica) non è la complessità rilevante asintoticamente.

Nella strategia di *pivoting totale*, basata su permutazioni sia di righe che di colonne, al passo i–esimo di eliminazione viene portato in posizione di pivot l'elemento $a_{jl}^{(i)}$ (con $j, l \geq i$) per cui si abbia

$$|a_{jl}^{(i)}| = \max_{k,h \geq i} |a_{kh}^{(i)}|.$$

- Occorre tenere memoria della operazione di scambio di variabili

- La complessità del pivoting totale è quadratica per ogni passo di eliminazione, è quindi asintoticamente confrontabile con la complessità della operazione di eliminazione

indice

Le fattorizzazioni LU e di Cholesky

L'eliminazione di una generica variabile x_i equivale a ottenere $A^{(i+1)} = T_i A^{(i)}$ moltiplicando a sinistra $A^{(i)}$ per una matrice di trasformazione

$$T_i = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{i+1,i} & & \\ & & \vdots & \ddots & \\ & & -m_{ni} & & 1 \end{pmatrix}$$

dove gli elementi m_{ki} per $k > i$ sono i moltiplicatori definiti da

$$m_{ki} = \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}}$$

Supponendo che non siano necessarie permutazioni di righe, la matrice triangolare superiore $A^{(n)} = U$ si ottiene quindi come

$$A^{(n)} = T_{n-1}A^{(n-1)} = T_{n-1}T_{n-2}A^{(n-2)} = \cdots =$$
$$= T_{n-1}T_{n-2}\cdots T_1 A^{(1)} = \Lambda A$$

dove la matrice $\Lambda = T_{n-1}T_{n-2}\cdots T_1$ è triangolare inferiore come prodotto di matrici t.i.. Ne segue che $\Lambda A = U$ e quindi, ponendo $\Lambda^{-1} = L$:

$$A = LU$$

con L ancora triangolare inferiore in quanto inversa di una matrice t.i. (inoltre, sia Λ che L hanno elementi unitari sulla diagonale).

Poiché $\Lambda = T_{n-1}T_{n-2}\cdots T_1$, allora $L = \Lambda^{-1} = T_1^{-1}\cdots T_{n-1}^{-1}$. Ponendo

$$m_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{k+1,k} \\ \vdots \\ m_{nk} \end{pmatrix}$$

e quindi $T_k = I - m_k e_k^t$, si verifica facilmente che

- La inversa della trasformazione T_k è $T_k^{-1} = I + m_k e_k^t$

- Il prodotto $T_j^{-1}T_k^{-1}$ con $j < k$ è $T_j^{-1}T_k^{-1} = I + m_j e_j^t + m_k e_k^t$, e da qui per induzione $L = T_1^{-1}\cdots T_{n-1}^{-1} = I + m_1 e_1^t + m_2 e_2^t + \cdots + m_{n-1} e_{n-1}^t$

Altrimenti, utilizzando la formula di prodotto $a_{ij} = \sum_k l_{ik} u_{kj}$ e partendo dalla prima riga di A si ottiene (poiché $l_{11} = 1$):

$$a_{1j} = l_{11} u_{1j} = u_{1j}$$

da cui si ottiene $u_{1j} = a_{1j}$ e quindi tutta la prima riga di U. Passando poi alla prima colonna di A:

$$a_{i1} = l_{i1} u_{11}$$

e poiché l'elemento u_{11} è stato già calcolato in precedenza, si ottiene per $i \geq 2$:

$$l_{i1} = \frac{a_{i1}}{u_{11}}.$$

Dalla seconda riga di A si ha:

$$a_{2j} = l_{21}u_{1j} + l_{22}u_{2j}$$

e quindi, per $j \geq 2$,

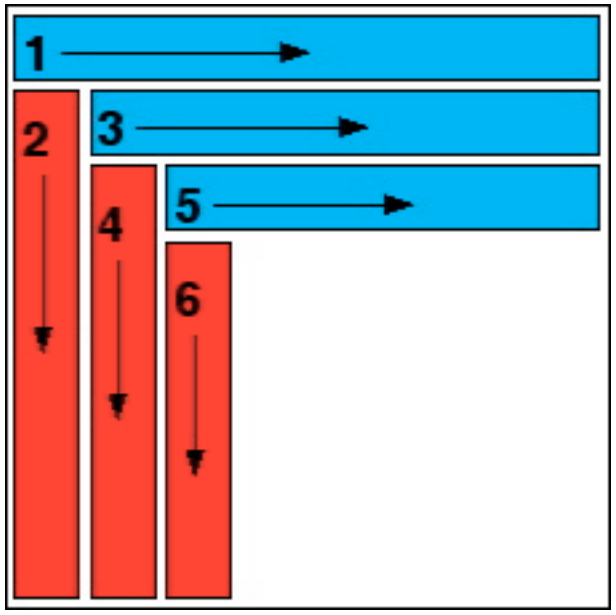
$$u_{2j} = a_{2j} - l_{21}u_{1j},$$

mentre considerando la seconda colonna di A si ha analogamente

$$a_{i2} = l_{i1}u_{12} + l_{i2}u_{22}$$

da cui si ottiene per $i \geq 3$:

$$l_{i2} = \frac{1}{u_{22}}(a_{i2} - l_{i1}u_{12}) \quad ...$$



per $j \geq p$, dalla p–esima riga di A:

$$u_{pj} = a_{pj} - \sum_{k<p} l_{pk} u_{kj},$$

per $i > q$ dalla q–esima colonna:

$$l_{iq} = \frac{1}{u_{qq}} \left(a_{iq} - \sum_{k<q} l_{ik} u_{kq} \right).$$

• La necessità di permutazioni di righe si evidenzia con la comparsa di un pivot $u_{pp} = 0$. In questo caso l'algoritmo di fattorizzazione pivotata deve scambiare le righe corrispondenti di A e della parte già calcolata di L (infatti nella fattorizzazione qualsiasi operazione sulle righe di A, $TA = TLU = (TL)U$ è come se fosse fatta su L) e tenere memoria della permutazione per effettuarla sul termine noto (infatti il sistema si riscrive $PAx = LUx = Pb$)

• La pivotazione non è necessaria per matrici definite positive o a diagonale dominante (in questo caso tutti i minori principali sono nonsingolari)

• La soluzione del sistema lineare $Ax = b$, introducendo la variabile ausiliaria z, si ottiene dalla successiva soluzione dei due sistemi triangolari $Lz = b$ e $Ux = z$

• La complessità di questo metodo di soluzione è di un ordine inferiore al metodo di eliminazione se la matrice è già fattorizzata (cioè se si risolvono più sistemi lineari con la stessa matrice ma diversi termini noti)

• La stabilità della fattorizzazione LU non è migliore di quella del metodo di eliminazione (gli elementi di U possono crescere molto)

Se la matrice A è definita positiva, si dimostra che si può porre $U = L^t$, cioè $A = LL^t$ (fattorizzazione di Cholesky), e quindi applicando lo stesso procedimento a partire dalla formula di prodotto $a_{ij} = \sum_k l_{ik} l_{jk}$ si ottiene per gli elementi sulla diagonale

$$l_{pp} = \left(a_{pp} - \sum_{k<p} l_{pk}^2 \right)^{\frac{1}{2}},$$

e per $i > p$:

$$l_{ip} = \frac{1}{l_{pp}} \left(a_{ip} - \sum_{k<p} l_{ik} l_{pk} \right).$$

• La pivotazione non è necessaria nella fattorizzazione di Cholesky, operando su una matrice A definita positiva

• La complessità di questa fattorizzazione è la metà di quella della fattorizzazione LU o del metodo di eliminazione, dovendosi calcolare un solo fattore triangolare

• La stabilità della fattorizzazione di Cholesky è decisamente migliore di quella del metodo LU (la maggiorazione $l_{pp} = \sqrt{a_{pp} - \sum_{k<p} l_{pk}^2} \leq \sqrt{a_{pp}}$ mostra che gli elementi di L crescono "più lentamente")

Altri problemi legati alla soluzione di sistemi lineari per fattorizzazione:

- Calcolo dell'inversa: in questo caso le colonne x_i della matrice inversa A^{-1} si ottengono come soluzione dei sistemi lineari

$$Ax_i = e_i$$

e A si fattorizza una volta per tutte all'inizio del calcolo risolvendo solo due sistemi triangolari per ogni termine noto

- Calcolo del determinante: poiché $\det L = 1$, si ha

$$\det A = \det L \det U = \det U = \prod_i u_{ii}$$

indice

La fattorizzazione QR

Un altro modo di fattorizzare una matrice A $n \times m$ (con $n \geq m$, e quindi non necessariamente quadrata) è nel prodotto di una matrice Q ortogonale e di una matrice R triangolare superiore.

• Dal punto di vista teorico, la possibilità di fattorizzare $A = QR$ discende dal procedimento di ortogonalizzazione di Gram–Schmidt applicato alle colonne della matrice A

• Dal punto di vista numerico la strada che si segue è invece di triangolarizzare la matrice A con trasformazioni ortogonali (a sinistra)

• La soluzione del sistema $Ax = b$ viene ottenuta risolvendo nell'ordine i sistemi lineari $Qz = b$ (che ha soluzione $z = Q^t b$) e $Rx = z$ (che è triangolare)

• La complessità di questo metodo di fattorizzazione è più alta degli altri metodi, mentre la stabilità è migliore. Il metodo QR si utilizza quindi in situazioni di condizionamento particolarmente sfavorevole

indice

La stabilità dei metodi diretti per sistemi lineari

Il condizionamento intrinseco del problema, se valutato in termini di errore relativo, è legato al cosiddetto numero di condizionamento $K_*(A) = \|A\|_* \, \|A^{-1}\|_*$ della matrice A rispetto alla norma $\|\cdot\|_*$

- Se si perturba solo il termine noto b, la soluzione $x + \delta x$ del sistema $A(x + \delta x) = b + \delta b$ è affetta da una perturbazione relativa

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

- Nel caso più generale, si ha una espressione più complessa ma conclusioni qualitativamente simili

La stabilità dei metodi diretti tipo MEG o fattorizzazione LU aumenta decisamente con la pivotazione, ed è migliore per metodi quali le fattorizzazioni QR e di Cholesky. E' comunque difficile risolvere accuratamente sistemi malcondizionati in dimensione alta

• La valutazione a posteriori della accuratezza tramite il residuo dipende anch'essa dal condizionamento del sistema

• Sistemi di grandi dimensioni, specie se malcondizionati, si risolvono spesso in modo più efficiente ed accurato con metodi iterativi

indice