2- METODI DIRETTI

I metodi diretti per la risoluzione numerica dei sistemi lineari consistono sostanzialmente nell'applicazione del metodo di riduzione di Gauss con il quale, attraverso la sostituzione di ogni riga con opportune combinazioni lineari della stessa riga con altre, si perviene ad un sistema equivalente di forma triangolare e quindi di immediata risoluzione.

Si badi che, sebbene la soluzione esatta del sistema

$$Ax=b$$

si esprima con

$$x=A^{-1}b,$$

c'è una differenza sostanziale tra il risolvere il sistema, cioè trovare x, e il calcolare la matrice inversa A^{-1}. Basti pensare all'equazione lineare

$$7x=21$$

la cui soluzione è

$$x = \frac{21}{7} = 3 \quad \text{(con qualunque precisione di macchina)}$$

e richiede una sola operazione, mentre attraverso il calcolo dell'inversa 1/7 si ottiene la soluzione

$$x = \frac{1}{7}\,21 = 0.142857 \times 21 = 2.99997$$

che richiede due operazioni anzichè una e rivela, di conseguenza, una maggiore propagazione dell'errore.

Il calcolo dell'inversa A^{-1} equivale infatti a risolvere il sistema Ax=b *per tutti i termini noti* b o, più precisamente, data la linearità di R^n, per n termini noti indipendenti. Infatti, poichè l'inversa A^{-1} soddisfa l'equazione matriciale AX=I, le colonne c_i di X sono ottenute risolvendo gli n sistemi

$$Ac_i=e_i \qquad i=1,...,n.$$

Dovendo invece risolvere il sistema Ax=b più di n volte per valori diversi del termine noto b, allora conviene disporre dell'inversa.

Osserveremo comunque che la riduzione a forma triangolare equivale, anche in termini di numero di operazioni, alla fattorizzazione A=LU con L triangolare inferiore ed U triangolare superiore che, una volta ottenuta, può essere utilizzata per la risoluzione di Ax=b per ogni b.

Metodo di riduzione di Gauss e fattorizzazione LU.

Consideriamo il problema Ax=b e definiamo la seguente successione di sistemi equivalenti

$$A^{(i)}x=b^{(i)} \qquad i=1,...,n-1$$

Fissato $A^{(0)}:=A$ e $b^{(0)}:=b$, il sistema di partenza è

$$\begin{aligned}
&a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + \ . \ . \ . \ + a_{1n}^{(0)}x_n = b_1^{(0)}\\
&a_{21}^{(0)}x_1 + a_{22}^{(0)}x_2 + \ . \ . \ . \ + a_{2n}^{(0)}x_n = b_2^{(0)}\\
&. \ . \ .\\
&a_{n1}^{(0)}x_1 + a_{n2}^{(0)}x_2 + \ . \ . \ . \ + a_{nn}^{(0)}x_n = b_n^{(0)}
\end{aligned}$$

dal quale si ottine l'equazione $A^{(1)}x=b^{(1)}$ nel seguente modo.
Con un opportuno scambio di righe nel sistema ci si assicura preliminarmente che $a_{11}^{(0)}\neq 0$ e si definiscono quindi i *moltiplicatori*

$$m_{21}= - \frac{a_{21}^{(0)}}{a_{11}^{(0)}} \ . \ . \ . , \ m_{n1}= - \frac{a_{n1}^{(0)}}{a_{11}^{(0)}} \ .$$

Mentre la prima riga del nuovo sistema rimane inalterata, per ogni riga sottostante si esegue la seguente sostituzione che ha lo scopo di annullare i coefficienti $a_{21},a_{31},...,a_{n1}$ della prima colonna di $A^{(1)}$:

$$a_{ij}^{(1)} = a_{ij}^{(0)} + a_{1j}^{(0)}\, m_{i1} \qquad j=1,2,...,n$$
$$b_i^{(1)} = b_1^{(0)}\, m_{i1} + b_i^{(0)}$$

e ciò per ogni riga i=2,...,n. All'atto pratico le precedenti sostituzioni saranno eseguite solo per j=2,..,n perchè per j=1 già sappiamo che i coefficienti $a_{i1}^{(1)}$ i=2,..,n sono nulli.

Il nuovo sistema $A^{(1)}x=b^{(1)}$ assume quindi la forma

$$\begin{aligned} a_{11}^{(1)}x_1 + \; & a_{12}^{(1)}x_2 + \; . \; . \; . \; + a_{1n}^{(1)}x_n = b_1^{(1)} \\ & a_{22}^{(1)}x_2 + \; . \; . \; . \; + a_{2n}^{(1)}x_n = b_2^{(1)} \end{aligned}$$

$$a_{n2}^{(1)}x_2 + \; . \; . \; . \; + a_{nn}^{(1)}x_n = b_n^{(1)}$$

Osserviamo che il coefficiente $a_{22}^{(1)}$ ed i coefficienti sottostanti $a_{i\,2}^{(1)}$ i=3,..,n non possono essere tutti nulli inquanto il determinante di $A^{(1)}$, che coincide con quello di A, è dato dal prodotto di $a_{11}^{(1)}$ per il determinante del minore $A_{11}^{(1)}$ che, di conseguenza, non può essere nullo.

Siamo di nuovo in grado di effettuare uno scambio delle righe sul sistema $A^{(1)}x=b^{(1)}$ che porti nella posizione di indice (2,2) (che d'ora in poi chiameremo *posizione di pivot* della matrice $A^{(1)}$) un coefficiente non nullo.

Ottenuto il sistema $A^{(k)}x=b^{(k)}$ del tipo

$$\begin{aligned} a_{11}^{(1)}x_1 + \; . \; . \; . \; & a_{1k}^{(1)}x_k + \; . \; . \; . \; . \; . \; . \; + a_{1n}^{(1)}x_n = b_1^{(1)} \\ . \; . \; . \qquad & \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \qquad\quad \vdots \\ & a_{kk}^{(k)}x_k + a_{k;k+1}^{(k)}x_{k+1} + \; . \; . \; + a_{kn}^{(k)}x_n = b_k^{(k)} \\ & \qquad\qquad a_{k+1;k+1}^{(k)}x_{k+1} + \; . \; . \; + a_{k+1;n}^{(k)}x_n = b_{k+1}^{(k)} \\ & \qquad\qquad\qquad\quad \vdots \qquad\qquad\qquad\quad \vdots \\ & \qquad\qquad a_{n;k+1}^{(k)}x_{k+1} + \; . \; . \; + a_{nn}^{(k)}x_n = b_n^{(k)} \end{aligned}$$

si porta in posizione di pivot (k+1,k+1) un coefficiente non nullo (attraverso un eventuale scambio della riga (k+1)-esima con una riga sottostante), e si effettuano le sostituzioni

$$a_{i\,j}^{(k+1)} = a_{i\,j}^{(k)} + a_{k+1;j}^{(k)}\, m_{i,k+1} \qquad j=k+1,...,n$$
$$b_i^{(k+1)} = b_{k+1}^{(k)}\, m_{i,k+1} + b_i^{(k)}$$

per i=k+2,...,n, dove i moltiplicatori sono dati da:

$$m_{i,k+1} = -\frac{a_{i;k+1}^{(k)}}{a_{k+1;k+1}^{(k)}} \qquad i=k+2,...,n$$

Al passo (n-1)-esimo si trova infine il sistema $A^{(n-1)}x=b^{(n-1)}$

$$\begin{aligned} a_{11}^{(n-1)}x_1 \;+\; \cdots\cdots \;+\; a_{1n}^{(n-1)}x_n &= b_i^{(n-1)} \\ \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot & \\ \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot & \\ \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot & \\ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} \end{aligned}$$

la cui matrice $A^{(n-1)}$ è di tipo triangolare superiore e indicheremo con U.
A questo punto è immediato calcolare le componenti di x all'indietro da x_n ad x_1.

Osserviamo che, a parte eventuali scambi di righe, la matrice $A^{(1)}$ è stata ottenuta da $A^{(0)}$ attraverso la moltiplicazione

$$A^{(1)}=M_1A^{(0)}$$

Dove

$$M_1 = \begin{Vmatrix} 1 & & & \\ m_{21} & 1 & & \\ \cdot & \cdot & & \\ & & \cdot & \\ m_{n1} & \cdots & & 1 \end{Vmatrix}$$

e successivamente,

$$A^{(2)}=M_2A^{(1)}=M_2M_1A^{(0)}$$

.....

$$A^{(n-1)}=M_{n-1}A^{(n-2)}=M_{n-1}...M_1A^{(0)}=U$$

dove, in generale,

$$M_k = \begin{Vmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & m_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & m_{n,k} & & 1 \end{Vmatrix}$$

Siccome le matrici M_k sono tutte invertibili, abbiamo fattorizzato la matrice A in $A=(M_{n-1}...M_1)^{-1}U=M^{-1,1}...M^{-1,n-1}U$.
Le matrici del tipo M_k, dette **matrici elementari di Gauss**, sono esprimibili come somma dell'identità e di una matrice di rango 1:

$$M_k=I+m_ke_k^T \quad \text{dove} \quad m_k=(0,....0,m_{k+1,k},...,m_{n,k})^T$$

Si osservi che la matrice $m_ke_k^T$ è nihilpotente $\left((m_ke_k^T)^2=0\right)$ cosicchè la serie di Neumann è $(I+m_ke_k^T)^{-1}=I-m_ke_k^T$. Si ha quindi la seguente espressione per l'inversa M_k^{-1}

$$M_k^{-1} = \begin{Vmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & -m_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -m_{n,k} & & 1 \end{Vmatrix}$$

e per il prodotto $M_k^{-1}\,M_{k+1}^{-1}=(I-m_ke_k^T)(I-m_ke_{k+1}^T)=I-m_ke_k^T-m_ke_{k+1}^T$

$$M_k^{-1} \quad M_{k+1}^{-1} = \begin{Vmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -m_{k+1;k} & 1 & & \\ & & & -m_{k+2,k+1} & & \\ & & \vdots & \vdots & \ddots & \\ & & -m_{n,k} & -m_{k+2,k+1} & & 1 \end{Vmatrix}$$

Cosicchè è facile vedere che la matrice $L_{k+1}:=M_1^{-1}...M_{k+1}^{-1}$ è

$$L_{k+1} = \begin{Vmatrix} 1 & & & & & \\ -m_{21} & 1 & & & & \\ \vdots & & 1 & & & \\ & & \ddots & -m_{k+2,k+1} & \ddots & \\ \vdots & & & \vdots & & \ddots \\ -m_{n1} & \ldots & -m_{k+2,k+1} & & & 1 \end{Vmatrix}$$

ed è ottenuta da L_k semplicemente riempiendo la parte sottodiagonale della colonna (k+1)-esima con i nuovi moltiplicatori cambiati di segno.

Infine la matrice $L:=L_{n-1}=M_1^{-1}...M_{n-1}^{-1}$ è ancora di tipo triangolare inferiore con determinante unitario:

$$L_{n-1} = \begin{Vmatrix} 1 & & & \\ -m_{21} & 1 & \ddots & \\ \vdots & & \ddots & \\ \vdots & & & 1 \\ -m_{n1} & \ldots & -m_{n,n-1} & 1 \end{Vmatrix}$$

Abbiamo dunque ottenuto la fattorizzazione

$$A = L\,U$$

con L ed U triangolari inferiore e superiore e con $\det(A)=\det(U)$.

Supponiamo ora che, una volta ottenuto il sistema $A^{(k)}x=b^{(k)}$, prima di calcolare $A^{(k+1)}$ si esegua lo scambio della riga (k+1)-esima con la i-esima (i>k+1) attraverso la matrice di permutazione $P_{i,k+1}$:

$$A^{(k)} \longrightarrow \quad P_{i,k+1}A^{(k)} \quad = P_{i,k+1}M_k...M_1A$$

cosicchè

$$A^{(k+1)}=M_{k+1}P_{i,k+1}M_k...M_1A$$

e quindi

$$A=M_1^{-1}...M_k^{-1}\ P_{i,k+1}M_{k+1}^{-1}\ A^{(k+1)} \ = L_kP_{i,k+1}M_{k+1}^{-1}\ A^{(k+1)}$$

Applicando la permutazione $P_{i,k+1}$ ad entrambi i membri si trova

$$P_{i,k+1}A=P_{i,k+1}L_k\ P_{i,k+1}M_{k+1}^{-1}\ A^{(k+1)}$$

dove la matrice

$$P_{i,k+1}\ L_k\ P_{i,k+1}$$

differisce da L_k solo per lo scambio dei moltiplicatori delle righe (k+1)-esima ed i-esima. Ciò significa che se la matrice $A^{(k)}$ necessita di una certa permutazione di righe $P_{i,k+1}$, la stessa permutazione va effettuata sulle corrispondenti righe della matrice dei moltiplicatori L_k e la fattorizzazione LU che si ottiene alla fine del processo di triangolarizzazione è relativa ad una permutazione P di A che tenga conto di tutte le permutazioni effettuate.

$$L\ U = P\ A$$

In conclusione il sistema da risolvere sarà

$$LUx=Pb$$

la cui soluzione è direttamente ottenibile dalla risoluzione dei due sistemi triangolari

$$Ly=Pb \quad e \quad Ux=y.$$

Osserviamo ancora che l'ordine con cui vengono costruiti i coefficienti di L e di U è il seguente:

1ª riga di U ; 1ª colonna di L ; 2ª riga di U ; 2ª colonna di L ; ... nª riga di U.

Pertanto le colonne di L e le righe di U possono essere allocate, man mano che si calcolano, al posto delle corrispondenti colonne e righe di A i cui coefficienti non sono più necessari. Ciò significa che la fattorizzazione LU non richiede ulteriore occupazione di memoria oltre quella necessaria per la memorizzazione di A.

Concludiamo questo paragrafo con il seguente teorema di unicità.

TEOREMA 2.9. *Ogni matrice* A *non singolare è fattorizzabile in modo unico con due matrici* L *ed* U *tali che* A=LU *dove* L *è triangolare inferiore a diagonale unitaria ed* U *triangolare superiore*.

Dim. Se fosse A=LU ed A=L'U', sarebbe:

$$LU=L'U'$$
$$U=L^{-1}L'U'$$
$$UU'^{-1}=L^{-1}L'.$$

Si osservi ora che $L^{-1}L'$ è ancora triangolare inferiore con diagonale unitara, mentre UU'^{-1} è triangolare superiore. Essendo tra loro uguali devono coincidere con l'identità, di conseguenza L=L' ed U=U'.

Strategia del pivot.

Abbiamo visto che al passo k-esimo della fattorizzazione il pivot $a_{k,k}^{(k-1)}$ deve essere $\neq 0$. La strategia del **pivot parziale** consiste nell'eseguire in ogni caso uno scambio di righe che porti nella posizione di pivot il coefficiente di modulo massimo della colonna sottostante. La strategia del **pivot totale** consiste invece nell'eventuale scambio anche delle colonne successive in modo da portare nella posizione di pivot il coefficiente di modulo massimo tra tutti quelli del minore definito dagli elementi sottostanti: $a_{i,j}^{(k-1)}$ con $i,j \geq k$.

A differenza del pivot parziale, la strategia del pivot totale richiede, ad ogni scambio di colonne, il riordinamento delle incognite. La strategia del pivot oltre ad evitare il rischio di una divisione per zero, consente di ridurre, come vedremo più avanti, la propagazione dell'errore dovuta al gran numero di operazioni richieste per la fattorizzazione.

In presenza di sistemi ben condizionati, la strategia del pivot non sarebbe strettamente necessaria se non per escludere, come già osservato, il rischio di una divisione per zero. Allora ci si chiede se è possibile stabilire a priori che ad ogni passo del processo di fattorizzazione il pivot sia non nullo. A questo proposito valgono i seguenti teoremi:

TEOREMA 2.10. *Se i minori principali di* A *sono non singolari allora, per ogni matrice* $A^{(k)}$, *è* $a_{k+1;k+1}^{(k)} \neq 0$.

Dim. L'asserto è ovviamente vero per k=0 e supponiamolo vero per k. Poichè i minori principali di A e di $A^{(k)}$ hanno lo stesso determinante, anche il minore principale (k+1)-esimo di $A^{(k)}$ è non singolare e quindi $a_{k+1;k+1}^{(k)} \neq 0$.

$$\begin{Vmatrix} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} & \cdots & \cdots & a_{1n}^{(k)} \\ 0 & \ddots & & & & \vdots \\ & & a_{kk}^{(k)} & a_{k;k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & \cdots & 0 & a_{k+1;k+1}^{(k)} & \cdots & a_{k+1;n}^{(k)} \\ & & & \vdots & & \vdots \\ 0 & & 0 & a_{n;k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{Vmatrix}$$

TEOREMA 2.11 . *Se* A *ha predominanza diagonale stretta oppure è definita positiva, allora i minori principali sono non singolari*.

Dim. Sia A a predominanza diagonale, allora lo è ogni minore principale che, per il teorema di Gerschgorin, risulta pertanto non singolare. Sia invece A definita positiva; allora lo è ogni minore principale. Sia infatti A_i il minore principale i-esimo e sia $y_i:=(x_1,...,x_i)\in R^i$.
Detto inoltre $u=(x_1,...,x_i,0,...,0)\in R^n$ si ha

$$y^{T,i}A_i\, y = u^T A\, u > 0$$

e quindi $\det(A_i)\neq 0$.

Complessità computazionale.

Il confronto di algoritmi diversi per la risoluzione di uno stesso problema è, in generale, un compito assai arduo, ancorchè possibile, in quanto la prestazione di un algoritmo dipende in larga misura dal modo in cui esso viene implementato, dal linguaggio di programmazione usato e dalla macchina sulla quale viene eseguito. Non è negli obbiettivi di questo corso fare una analisi così raffinata degli algoritmi e quindi ci accontenteremo di enumerare semplicemente il numero di operazioni necessarie per realizzare l'algoritmo in oggetto, valuteremo cioè la sua **complessità computazionale**. In ogni caso assumeremo come operazione elementare il complesso di una moltiplicazione (o divisione) ed una successiva addizione sul risultato.

Il passaggio dalla matrice $A^{(0)}$ alla matrice $A^{(1)}$ richiede, per ogni riga dalla 2^a alla n^a, il calcolo del moltiplicatore m_{i1} (n-1 op) e delle quantità

$$a_{ij}^{(1)} = a_{ij}^{(0)} + a_{1j}^{(0)}\, m_{i1} \qquad j=2,...,n \qquad \text{(n-1 op)}$$

per un totale di n op. La matrice L_1 ed $A^{(1)}$ richiedono quindi n(n-1) op. Analogamente per il passaggio da $A^{(1)}$ ad $A^{(2)}$ bisogna manipolare la sottomatrice $B^{(1)}$ di dimensione n-1:

$$A^{(1)} = \left\| \begin{array}{c|ccc} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & & \\ & B^{(1)} & \\ 0 & & \end{array} \right\|$$

per un totale di (n-1)(n-2) op. Infine il passaggio da $A^{(n-2)}$ ad $A^{(n-1)}$ richiede 2 op. L'intera trasformazione, cioè la determinazione di L ed U, richiede

$$\sum_{j=1}^{n-1} j(j+1) = \sum_{j=1}^{n-1} j^2 + \sum_{j=1}^{n-1} j = \frac{(n-1)n(2n-1)}{6} + \frac{n(n-1)}{2}$$
$$= \frac{n^3}{3} - \frac{n}{3} \approx \frac{n^3}{3}$$

Infine per calcolare x devo risolvere i due sistemi Ly=b e Ux=y .

Per Ly=b:

$y_1 = b_1$ 0 op

$y_2 = b_2 - m_{21}y_1$ 1 op

$y_3 = b_3 - m_{31}y_1 + m_{32}y_2$ 2 op

...

$y_n = b_n - m_{n1}y_1 - \ldots - m_{n,n-1}y_{n-1}$. n-1 op

in totale n(n-1)/2 op.

Per Ux=y:

$x_n = y_n/u_{nn}$ 1 op

$x_{n-1} = (y_{n-1} - u_{n-1,n}x_n)/u_{n-1,n-1}$ 2 op

...

$x_1 = (y_1 - u_{1,n}x_n - \ldots - u_{12}x_2)/u_{11}$ n op

in totale n(n+1)/2 op.

I due sistemi triangolari richiedono globalmente n^2 op. Asintoticamente il tempo richiesto per la loro risoluzione è trascurabile rispetto al tempo necessario per la fattorizzazione LU.

Metodo di Choleski.

Nel caso in cui la matrice A sia simmetrica e definita positiva la fattorizzazione con matrici triangolari inferiore e superiore può essere fatta in modo più economico. Vale infatti il seguente teorema:

Teorema 2.12. *Sia* A *hermitiana e definita positiva, allora esiste una ed una sola matrice triangolare inferiore* $\bar{L}$ *tale che* $A=\bar{L}\,\bar{L}^H$.

Dim. Sia LU la fattorizzazione di A e sia $D:=\mathrm{diag}(u_{11}....u_{nn})$. Si ha quindi:

$$A=LDD^{-1}U = LDU' \quad \text{con } U'=D^{-1}U \text{ a diagonale unitaria}$$

$$A=A^H = (U')^H DL^H$$

e, per l'unicità della fattorizzazione (con L a diagonale unitaria) si ha:

$$(U')^H = L$$

e quindi:

$$A= LDL^H.$$

Detto infine

$$D^{1/2}:=\mathrm{diag}(\sqrt{u_{11}},...,\sqrt{u_{nn}}) \quad \text{e} \quad \bar{L}:=LD^{1/2},$$

si ha

$$A = LD^{1/2}D^{1/2}L^H = \bar{L}\,\bar{L}^H.$$

Sul piano pratico, la fattorizzazione $\bar{L}\,\bar{L}^H$ si realizza direttamente uguagliando riga per riga il prodotto $\bar{L}\,\bar{L}^H$ con la matrice A. Si ottine così, per la prima riga di $\bar{L}^H$ (supponiamo, per semplicità di notazione, che $\bar{L}$ sia reale):

$$\ell_{11}(\ell_{11},\ell_{21},...,\ell_{2n})= (a_{11},a_{12},...,a_{1n})$$

dalla quale si ricava la prima riga di $\bar{L}^H$:

$\ell_{11}=\sqrt{a_{11}}$
$\ell_{21}=a_{12}/\ell_{11}$
...
$\ell_{2n}=a_{1n}/\ell_{11}$.

Analogamente per la seconda riga :

$$\ell_{21}(\ell_{11},\ell_{21},...,\ell_{2n})+\ell_{22}(0,\ell_{22},\ell_{32},...,\ell_{3n})=(a_{21},a_{22},...,a_{2n})$$

dalla quale si ricava la seconda riga di L^H:

$\ell_{22}=\sqrt{a_{22}-\ell_{21}^2}$
$\ell_{32}=a_{23}-\ell_{21}\ell_{31}/\ell_{22}$
...
$\ell_{n2}=a_{2n}-\ell_{21}\ell_{n1}/\ell_{22}$.

E così di seguito per le righe successive.

Analisi dell'errore.

L'analisi dell'errore nel metodo di Gauss è basata essenzialmente sull'analisi all'indietro ideata da Wilkinson per questo problema intorno agli anni 60. Si dimostra che la fattorizzazione $\bar{L}\bar{U}$ computata in una aritmetica di precisione eps soddisfa la relazione

$$\bar{L}\bar{U}= A + E \qquad \text{con } ||E|| \leq n^2\, g_n\, ||A||_\infty \; eps$$

dove

$$g_n= \frac{\max_{i\,j\,k}|\bar{a}_{i\,j}^{(k)}|}{\max_{i\,j}|a_{i\,j}|}$$

Inoltre, detta $\bar{y}$ la soluzione computata di $\bar{L}y=b$ ed $\bar{x}$ quella di $\bar{U}x=\bar{y}$, si dimostra che $\bar{x}$ soddisfa l'equazione:

$(A + \delta A)\bar{x} = b$ con $||\delta A|| \leq (n^3+3n^2)\, g_n\, ||A||_\infty$ eps.

Obbiettivo di un buon algoritmo è quello di ottenere una fattorizzazione con una perturbazione E più piccola possibile. Ciò dipende essenzialmente dalla costante g_n per la quale si possono dare le seguenti stime:

strategia del pivot parziale: $g_n \leq 2^{n-1}$
strategia del pivot totale: $g_n \leq 1.8\, n^{0.25 \log n}$.

Vediamo ora in generale che cosa comporta, per la soluzione, una perturbazione sui dati del sistema, cioè sulla matrice e sul termine noto. Cominciamo col chiederci quando una pertubazione E sulla matrice non singolare A conserva la nonsingolarità per A+E. A questo proposito vale il seguente risultato.

Lemma di perturbazione (di Banach). *Sia* A *non singolare ed* E *tale che* $||A^{-1}||||E||<1$. *Allora* A+E *è ancora non singolare ed inoltre:*

$$||(A+E)^{-1}|| \leq \frac{||A^{-1}||}{1-||A^{-1}||\,||E||}$$

Detta x la soluzione del sitema Ax=b, sia y la soluzione del sistema perturbato

$(A + \delta A)y = b + \delta b$

per il quale supponiamo $||A^{-1}||\,||\delta A||<1$. Si avrà quindi:

$(A + \delta A)x = b + \delta Ax$

$(A + \delta A)(x - y) = b + \delta Ax - b - \delta b = \delta Ax - \delta b$

$x - y=(A + \delta A)^{-1}(\delta Ax - \delta b)$

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

Inoltre:

$$Ax=b \Rightarrow ||b|| \le ||A||\,||x|| \Rightarrow \frac{1}{||b||} \le \frac{||A||}{||b||}$$

e quindi:

$$\frac{||x-y||}{||x||} \le \frac{||A^{-1}||}{1-||A^{-1}||\,||\delta A||}\left(||\delta A|| + \frac{||\delta b||\,||A||}{||b||}\right)$$

$$\frac{||x-y||}{||x||} \le \frac{||A||\,||A^{-1}||}{1-||A^{-1}||\,||\delta A||}\left(\frac{||\delta A||}{||A||} + \frac{||\delta b||}{||b||}\right)$$

La stima precedente mostra che l'errore relativo causato dalle perturbazioni è maggiorato da una quantità proporzionale alle perturbazioni relative sui dati, con costante di proporzionalità $\frac{||A||\,||A^{-1}||}{1-||A^{-1}||\,||\delta A||}$ che, in prima approssimazione, vale $||A||\,||A^{-1}||$. In particolare, se $\delta A=0$, cioè se la perturbazione riguarda solo il termine noto b, allora si ha:

$$\frac{||x-y||}{||x||} \le ||A||\,||A^{-1}||\,\frac{||\delta b||}{||b||}$$

Il numero

$$K(A) := ||A||\,||A^{-1}||$$

è detto **indice di condizionamento** della matrice A e rappresenta dunque la *sensitività* della soluzione di Ax=b rispetto alle perturbazioni sui dati.

L'indice di condizionamento dipende dalla norma e, nel caso estremo di matrice singolare, lo definiamo uguale a infinito:

$$K_i(A) = \begin{cases} ||A^{-1}||_i\,||A||_i & |A| \neq 0 \\ \infty & |A| = 0 \end{cases}$$

Per ogni norma naturale si ha $K(A) \ge 1$, infatti:

$$1 = ||I|| = ||A^{-1}A|| \le ||A^{-1}||\,||A||$$

In generale non si riesce a calcolare il valore dell'indice di condizionamento senza disporre dell'inversa A^{-1}, si riesce però a dare la seguente stima:

$$||A^{-1}||\,||A|| \geq \rho(A^{-1})\rho(A) = \frac{|\lambda|_{max}}{|\lambda|_{min}}$$

che nel caso di matrici hermitiane diventa, in norma 2,:

$$||A^{-1}||_2||A||_2 = \frac{|\lambda|_{max}}{|\lambda|_{min}}.$$

La situazione di condizionamento ottimale si ha nel caso di K(A)=1 nel quale la matrice è detta **perfettamente condizionata**. Ciò si verifica, sempre nella norma 2, per le matrici unitarie.

Si osservi infine che nel passaggio dalla matrice A alla matrice A^HA si ha un peggioramento del condizionamento. Infatti , mentre abbiamo visto che per la matrice hermitiana A^HA si ha

$$K_2(A^HA) = \frac{|\lambda_{A^HA}|_{max}}{|\lambda_{A^HA}|_{min}},$$

per la matrice A si ha :

$$||A||_2 = \sqrt{\rho(A^HA)} = \sqrt{|\lambda_{A^HA}|_{max}}$$

e

$$||A^{-1}||_2 = \sqrt{\rho((A^HA)^{-1})} = \sqrt{\frac{1}{|\lambda_{A^HA}|_{min}}}$$

e quindi

$$K_2(A) = \sqrt{\frac{|\lambda_{A^HA}|_{max}}{|\lambda_{A^HA}|_{min}}}\,.$$

In definitiva gli indici di condizionamento delle due matrici sono legati della relazione $K_2(A^TA) = \left(K_2(A)\right)^2$ e risultano uguali solo per matrici prefettamente condizionate.

Esempio:

A titolo di esempio si consideri il sistema lineare

$$x_1 + x_2 = 2$$
$$x_1 + 1.01x_2 = 2.01$$

la cui soluzione è $x_1=1$, $x_2=1$. La matrice è simmetrica con autovalori $\lambda_1 \cong 2.005$ e $\lambda_2 \cong .005$ per cui $K_2(A) \cong 401$ che rivela un cattivo condizionamento. Infatti il seguente sistema, ottenuto dal precedente con una perturbazione relativa non superiore a .01 sui coefficienti:

$$y_1 + y_2 = 2$$
$$1.001y_1 + y_2 = 2.01$$

ha come soluzione $y_1=10$ e $y_2=-8$, con uno scarto relativo $\frac{||x-y||_2}{||x||_2} = 9$ rispetto ad una perturbazione sulla matrice $\frac{||\delta A||_2}{||A||_2} \approx 0.005$, che risulta amplificata di un fattore 1800. Ciò non è in contraddizione con le maggiorazioni precedenti inquanto esse valgono sotto la condizione $||\delta A||_2||A^{-1}||_2<1$ che in questo caso non è verificata.

Come già accennato in precedenza (nel paragrafo sui criteri d'arresto per i metodi iterativi) anche il test sul residuo per valutare la bontà di una soluzione approssimata non è accettabile nel caso di sistemi mal condizionati. Proviamo infatti a calcolare il residuo r del primo sistema rispetto ad una soluzione perturbata $y_1=10$ e $y_2=-8$.

$$r_1 = y_1 + y_2 - 2 = 10 - 8 - 2 = 0$$
$$r_2 = y_1 + 1.01y_2 - 2.01 = 10 - 8.08 - 2.01 = -0.09$$

Il residuo risulta molto "piccolo" rispetto allo scarto relativo sulla soluzione che abbiamo visto essere 9. Ciò è completamente giustificato dall'analisi

precedente quando si osservi che il residuo non è altro che una perturbazione sul termine noto. Infatti

$$Ax=b$$

$$Ay=Ay-b+b=r+b$$

Nel nostro caso si ha $\frac{\|r\|_2}{\|b\|_2} = \frac{0.045}{\sqrt{2}} \cong 0.0318$ a cui corrisponde un errore relativo $\frac{\|x-y\|_2}{\|x\|_2} = 9$ che risulta amplificato di un fattore $\cong 283$.

Raffinamento iterativo

I metodi che abbiamo analizzato in questo paragrafo consistono in un numero finito di operazioni alla fine delle quali si ottiene un risultato che differisce dalla soluzione esatta per il solo effetto degli errori di arrotondamento che si sono propagati durante l'esecuzione dell'intero l'algoritmo. In particolare nel caso della fattorizzazione LU, si ottiene un risultato, che indicheremo ora con x^0, che è la soluzione approssimata del problema

$$\bar{L}\bar{U}x=b$$

con $\bar{L}$ ed $\bar{U}$ a loro volta approssimazioni delle matrici L ed U.

Se la matrice A non è troppo "mal condizionata" si può eseguire il seguente *raffinamento iterativo* della soluzione che consiste nella esecuzione di alcuni passi della iterazione

$$Nx^{i+1}=Px^{i} + b$$

definita dallo splitting $N=\bar{L}\bar{U}$ e $P=\bar{L}\bar{U}-A$, a partire dal valore x^0.

Si ha dunque:

$$\bar{L}\bar{U}x^{i+1}=(\bar{L}\bar{U}-A)x^{i} + b$$

$$\bar{L}\bar{U}(x^{i+1}-x^{i}) = -Ax^{i} + b = r^{i}.$$

Le *correzioni* $x^{i+1}-x^i$ si calcolano, come abbiamo già osservato, attraverso la risoluzione in avanti e all'indietro dei sistemi

$$\bar{L}y = r^i$$

$$\bar{U}(x^{i+1}-x^i) = y$$

nei quali i residui r^i devono essere calcolati in doppia precisione. Ciò è suggerito dal fatto che nella risoluzione del sistema $\bar{L}\bar{U}(x^{i+1}-x^i)=r^i$,una perturbazione relativa sul residuo dell'ordine della precisione di macchina eps, si riflette (se la matrice non è troppo "mal condizionata") in un errore relativo sulla soluzione dello stesso ordine di grandezza. Ora la soluzione x^0 è già stata calcolata con la precisione eps e quindi per avere un miglioramento il residuo deve essere calcolato con una precisione superiore. In certi casi il meccanismo di raffinamento è molto efficace ed un paio di iterazioni sono sufficienti per ottenere una sorprendente precisione.