SISTEMI LINEARI

Sistema di m equazioni in n incognite:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad i = 1,\ldots,m$$

$$\updownarrow \qquad (1)$$

$$Ax = b$$

Soluzione del sistema: n-upla che soddisfi tali equazioni. Trattiamo solo sistemi quadrati m = n in cui $A \in R^{nxn}$, $b \in R^n$.

In tal caso $\exists_1\ x \in R^n$ soluzione di (1) se e solo se:

1) $\exists\ A^{-1}$ oppure 2) ranK(A) = n oppure 3) $A\underline{x} = 0 \Rightarrow \underline{x} = 0$.

Teorema di Cramer

Se $\det(A) \neq 0$ $\exists_1$ soluzione del sistema data da:

$$x_i = \frac{\det(\Delta_i)}{\det(A)} \qquad (2)$$

con $\Delta_i = \begin{vmatrix} . & . & b_1 & . & . \\ . & & & & \\ . & & b_n & & \end{vmatrix}$

↓

i-esima colonna

Costo computazionale di (2): (n+1)! Flops.

Se n = 50, 10^9 flops $\Rightarrow$ time = $9 \cdot 6 \cdot 10^{47}$ anni!

ANALISI DI STABILITA' PER SISTEMI LINEARI

Numero di condizionamento di una matrice $A \in C^{nxn}$:

$$\exists\ A^{-1} : \qquad k(A) = \|A\|\|A^{-1}\|$$

con $\|\cdot\|$ norma matriciale indotta (*)

Poiché: $1 = \|AA^{-1}\| \leq \|A\| \cdot \|A^{-1}\| = k(A)$

più k(A) è grande, maggiore è la sensibilità della soluzione di Ax = b alle perturbazioni nei dati.

(*) Norma matriciale indotta: $\exists\ x \in R^n : \|Ax\| = \|A\| \cdot \|x\|$

Definendo: $\text{dist}_p(A) = \min \left\{ \frac{\|\delta A\|_p}{\|A\|_p} : A + \delta A \quad \text{singolare} \right\}$

Si può dimostrare che: $\text{dist}_p(A) = \frac{1}{k_p(A)}$

Quindi, maggiore è k(A), più vicino è il comportamento di A+δA ad una matrice singolare.

NB: il determinante di una matrice non è un indice di condizionamento. Si possono infatti trovare matrici con determinante piccolo e numero di condizionamento grande e viceversa.

$B \in R^{nxn}$ $\quad b_{ii} = 1, \quad b_{ij} = -1 \quad i < j, \quad b_{ij} = 0 \quad i > j$

Esempio:

$$B = \begin{bmatrix} 1 & -1 & -1 & -1 \\ & 1 & & \vdots \\ & 0 & \ddots & -1 \\ & & & 1 \end{bmatrix}$$

$$\det B = 1, \quad k(B) = n2^{n-1}$$

Teorema 1.

Sia A una matrice qualunque. Per ogni $\|\cdot\|$ indotta: $\rho(A) \le \|A\|$.

Dimostrazione: Sia $|\lambda| = \rho(A)$ e $\underline{x}$ autovettore ad esso associato, e sia $\|x\| = 1$

$\rho(A) = |\lambda| = |\lambda| \|x\| = \|\lambda x\| = \|Ax\| \le \|A\| \|x\| = \|A\|$

Da tale teorema si ha pure: $k(A) = \|A\| \|A^{-1}\| \ge \frac{\lambda_{max}}{\lambda_{min}}$

Teorema 2.

Per una matrice quadrata si ha: $\rho(A) < 1 \Leftrightarrow \|A\| < 1$ per qualche $\| \ \|$ indotta.

Teorema 3.

Sia A qualunque e la norma sia tale che $\|A\| < 1$. Allora I+A è non singolare e:

$$\left\| (I + A)^{-1} \right\| \le \frac{1}{1 - \|A\|}$$

Dimostrazione.

Poiché $\|A\| < 1 \rightarrow \rho(A) < 1$. Pertanto $I+A$ è non singolare poiché i suoi autovalori non possono essere nulli. Quindi:

$$(I + A)(I + A)^{-1} = I$$

$$(I + A)^{-1} + A(I+A)^{-1} = I$$

Da cui:

$$(I + A)^{-1} = I - A(I + A)^{-1}$$

$$\left\|(I + A)^{-1}\right\| \le 1 + \|A\|\left\|(I + A)^{-1}\right\|$$

$$\left\|(I + A)^{-1}\right\|(1 - \|A\|) \le 1$$

e poiché per ipotesi $\|A\| < 1$ si ha che $(1 - \|A\|) > 0$ e quindi:

$$\left\|(I + A)^{-1}\right\| \le \frac{1}{1 - \|A\|}$$

In modo analogo, si puo' dimostrare che:

$$\left\|(I - A)^{-1}\right\| \le \frac{1}{1 - \|A\|}.$$

Vediamo ora la relazione di k(A) con le perturbazioni sui dati.

Indichiamo con δA, δx, δb le perturbazioni su A, x, b, rispettivamente. Allora il sistema da risolvere e':

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

e supponiamo che esso sia risolto esattamente.

Diamo una stima dell'errore relativo su x in funzione di δA, δb.

Teorema 4.

Sia $A \in R^{nxn}$, supponiamo che $\exists\, A^{-1}$, $\delta A \in R^{nxn}$ tali che: $\|\delta A\| < \frac{1}{\|A^{-1}\|}$

Allora $x \in R^n$, soluzione di $Ax = b$, $\underline{b} \in R^n$, $\underline{b} \neq 0$ e' tale che:

$$\frac{\|\delta x\|}{\|x\|} \le \frac{k(A)}{1 - k(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right)$$

Dimostrazione.

Per ipotesi: $\|\delta A\| < \frac{1}{\|A^{-1}\|} \Rightarrow \|A^{-1}\delta A\| \le \|A^{-1}\|\|\delta A\| < 1.$

Pertanto $(I + A^{-1}\delta A)$ è invertibile poiche' sono soddisfatte le ipotesi del teorema 3. Si ha:

$$\left\|\left(I + A^{-1}\delta A\right)^{-1}\right\| \le \frac{1}{1-\left\|A^{-1}\delta A\right\|} \le \frac{1}{1-\left\|A^{-1}\right\|\left\|\delta A\right\|}$$

Da $(A + \delta A)(x + \delta x) = b + \delta b$ si ha:

$$\cancel{Ax} + A\delta x + x\,\delta A + \delta A\delta x = \cancel{b} + \delta b$$

$$\delta x(A + \delta A) = \delta b - x\,\delta A$$

$$\delta x = (A + \delta A)^{-1}(\delta b - x\delta A) = (I + A^{-1}\delta A)^{-1}A^{-1}(\delta b - x\,\delta A)$$

Passando alle norme:

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

Dividiamo per $\|x\|$:

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

Ma : $\|A\|\|x\| \ge \|Ax\| = \|b\|$ quindi:

$$\frac{\|\delta x\|}{\|x\|} \le \frac{k(A)}{1-k(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right) \qquad \bullet$$

Teorema 5.

Con le ipotesi del teorema 4 si supponga che sia anche $\delta A = 0$. Allora:

$$\frac{1}{k(A)}\frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|} \le k(A)\frac{\|\delta b\|}{\|b\|}$$

Dimostrazione.

Si dimostra solo la prima disuguglianza perché la seconda è conseguenza del teorema 4.

$\delta x = A^{-1}\delta b \Rightarrow \|\delta b\| \le \|A\|\|\delta x\| \Rightarrow \|x\|\|\delta b\| \le \|x\|\|A\|\|\delta x\|$ e poiché $\|x\| \le \left\|A^{-1}\right\|\|b\|$ si ha: $\|x\|\|\delta b\| \le k(A)\|\delta x\|\|b\|$ che da' la tesi.

Vediamo ora un altro modo, di tipo analitico, per ricavare k(A).

Siano $A, F \in \Re^{nxn}$, $b, f \in \Re^n$, $\varepsilon \in \Re^+$, $\det(A) \neq 0$.

$$\begin{cases}\left(A + \varepsilon\ F\right)\ x(\varepsilon) = b + \varepsilon\ f \\ x(0) = x\end{cases} \qquad (3)$$

Sia ε piccolo, $\det(A + \varepsilon F) \neq 0$. La soluzione della (3) e' data da:

$$x(\varepsilon) = (A + \varepsilon F)^{-1}(b + \varepsilon f)$$

Deriviamo la (3) rispetto ad ε nell' intorno dello zero:

$$Fx(\varepsilon) + (A + \varepsilon F)\,\dot{x}(\varepsilon) = f$$

$$\text{Per } \varepsilon = 0 \text{ si ha:} \quad Fx(0) + A\,\dot{x}(0) = f$$

$$\text{Da cui:} \quad \dot{x}(0) = A^{-1}(f - Fx(0))$$

$$x(\varepsilon) \sim x(0) + \varepsilon\,\dot{x}(0)$$

$$\frac{\|x(\varepsilon) - x(0)\|}{\|x(0)\|} \sim \frac{\|\varepsilon\,\dot{x}(0)\|}{\|x(0)\|} = \frac{\|\varepsilon\, A^{-1}(f - Fx(0))\|}{\|x(0)\|} \leq \varepsilon\|A^{-1}\|\left(\frac{\|f\|}{\|x(0)\|} + \|F\|\right) = \varepsilon\|A^{-1}\|\|A\|\left(\frac{\|f\|}{\|A\|\|x\|} + \frac{\|F\|}{\|A\|}\right) \leq$$

$$\leq k(A)\left(\frac{\|\varepsilon f\|}{\|b\|} + \frac{\|\varepsilon F\|}{\|A\|}\right)$$

Quindi, il numero di condizionamento $k(A) = \|A\|\|A^{-1}\|$ e' correlato all'errore da:

$$k(A) \geq \frac{\text{errore sui risultati}}{\text{errore sui dati}}$$

Notiamo che: $k(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1$

Quanto più k(A) è prossimo ad 1 tanto piu' A è ben condizionata. Pero' la conoscenza di $\|A^{-1}\|$ non è facile da ottenere.

Modo empirico (analisi a posteriori)

Perturbare i dati e vederne l'influenza sui risultati. Se la matrice non è mal condizionata si può risolvere il sistema.

Esempio di matrice mal condizionata: la matrice di Hilbert.

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & & \frac{1}{n+1} \\ \vdots & & & & \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & & \frac{1}{2n-1} \end{bmatrix}$$

n	$k(H_n)$
3	$5 \cdot 10^2$
4	$1 \cdot 10^4$
5	$4 \cdot 10^5$
6	$1 \cdot 10^7$
…	…
10	$1 \cdot 10^{13}$

Correlazione tra k(A) e ρ(A):

$$k(A) \geq \rho(A) \cdot \rho(A^{-1})$$

$$k(A) \geq \frac{\max_{\lambda \in \sigma} |\lambda|}{\min_{\lambda \in \sigma} |\lambda|}$$

Pertanto, prima di risolvere numericamente un sistema lineare, ci si deve accertare che la matrice dei coefficienti sia non singolare e ben condizionata. La risoluzione numerica prevede due possibili strategie: quelle basate sui *metodi diretti* e quelle basate sui *metodi iterativi*. La scelta del tipo di metodo si basa essenzialmente sul tipo di matrice e sulle risorse a disposizione: tempo di calcolo e spazio di memoria. Infatti, mentre i metodi diretti sono adatti ai sistemi con matrici piene, i metodi iterativi sono adatti ai sistemi con matrici *sparse*, contenenti cioe' molti zeri.
Occupiamoci prima dei metodi diretti. Poiche', come vedremo, il risultato di tali metodi e' sempre un sistema triangolare, occupiamoci prima di risolvere un tale sistema.

Risoluzione di sistemi triangolari

Sia dato il seguente sistema lineare 3x3 non degenere:

$$\begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$Lx = b$$

Poiché $\det(L) \neq 0 \quad \ell_{ii} \neq 0$, la soluzione è quindi data da:

$$\begin{cases} x_1 = b_1/\ell_{11} \\ x_2 = (b_2 - \ell_{21}x_1)/\ell_{22} \\ x_3 = (b_3 - \ell_{31}x_1 - \ell_{32}x_2)/\ell_{33} \end{cases}$$

In generale si ha quindi (metodo delle sostituzioni in avanti):

$$x_1 = b_1 / \ell_{11}$$

$$x_i = \left(b_i - \sum_{j=1}^{i-1} \ell_{ij} x_j \right) \Big/ \ell_{ii} \quad i = 2, \ldots, n$$

Costo computazionale: numero di moltiplicazioni e divisioni = n(n+1)/2

numero di addizioni e sottrazioni = n(n - 1)/2

per un totale di n^2 flops.

Metodo delle sostituzioni indietro. Si deve risolvere il sistema: $Ux = b$ ovvero:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$\Rightarrow \quad x_n = b_n / u_{nn}$

$$x_i = \left(b_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \Big/ u_{ii} \quad i = n-1, \ldots, 1$$

Metodi diretti

La soluzione è ottenuta con un numero finito di passi.

Metodo di eliminazione di Gauss

Sia Ax = b con det(A) ≠ 0:

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_n = b_1 \\ \vdots \\ a_{n1}x_1 + \ldots + a_{nn}x_n = b_n \end{cases}.$$

Sia $a_{11} \neq 0$. Se ciò non si ha si scambia la prima riga con una delle successive in cui il coefficiente di x_1 sia diverso da zero.

Sia $m_{i1} = -\frac{a_{i1}}{a_{11}}$ per i = 2,…,n e aggiungiamo alla i-esima equazione la prima moltiplicata per m_{i1}. Si ha:

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

dove: $a_{ij}^{(2)} = a_{ij} + m_{i1}a_{1j}$ i, j = 2,…,n

$b_i^{(2)} = b_i + m_{i1}b_1$ i = 2,…,n

Operiamo allo stesso modo dalla seconda equazione moltiplicando per $m_{i2} = -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$.

Al passo n-1 si ottiene un sistema triangolare che si risolve con il metodo della sostituzione all'indietro.

Il costo computazionale del metodo di Gauss e' $\approx \frac{4}{3}n^3$.

Perché il metodo di Gauss funzioni è necessario che gli elementi a_{ii} siano diversi da zero. Ciò non è comunque sufficiente a garantire che nei passi successivi gli elementi diagonali non si annullino. Infatti sia:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix} \quad a_{ii} \neq 0\,,\ i=1,2,3$$

Eppure: $$A^{(2)} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{bmatrix} \quad \text{da cui } a_{22}^{(2)} = 0$$

Abbiamo quindi bisogno di condizioni più restrittive su A. Vedremo più avanti che se tutti i minori principali di A sono non nulli allora anche gli elementi diagonali in tutti i passi di eliminazione saranno non nulli.
Notiamo che la matrice A ha il secondo minore principale uguale a zero. Scambiando in $A^{(2)}$ la seconda e la terza riga il metodo funziona.
Per evitare inoltre problemi di arrotondamento si usano le tecniche del *pivot parziale* e del *pivot totale.*

Pivot parziale. Al j-esimo passo si cerca la riga I contenente il massimo elemento della j-esima colonna: $a_{Ij} = \max_{i \le j \le n} |a_{ij}|$ e si scambia la riga i con la riga I. Pertanto al primo passo: $a_{I1} = \max_{\substack{i \ge 1 \\ j \le n}} |a_{i1}|$.

Pivot totale. Si trova il massimo elemento della matrice: $a_{IJ} = \max_{i,j} |a_{ij}|$ e si scambiano la riga i con la riga I e la colonna j con la colonna J.
Il metodo del pivot totale è più preciso ma bisogna memorizzare l'ordine di eliminazione delle variabili e quindi si occupa molta memoria.

Per far vedere la necessità del pivoting abbiamo il seguente esempio:

$$\begin{cases} 0.0001x + 1.00y = 1.00 \\ 1.00x + 1.00y = 2.00 \end{cases}$$

$x = 1.00010 \quad y = 0.99990 \quad$ soluz. analitica

$x_G = 0.00 \quad y_G = 1.00 \quad$ soluz. con Gauss

Riscrivendo il sistema:

$$\begin{cases} 1.00x + 1.00y = 2.00 \\ 0.0001x + 1.00y = 1.00 \end{cases}$$

$x_G = 1.00, \quad y_G = 1.00$

Metodi di fattorizzazione. Sono una riformulazione matriciale del metodo di Gauss. Consistono nel trovare una matrice S non singolare e formare un sistema equivalente a quello originale.

$$Ax = b \;\Rightarrow\; SAx = Sb, \quad SA = U$$

U = matrice triangolare superiore.

Se S è triangolare inferiore lo è pure S^{-1}:

$$A = S^{-1}U = LU$$

Fattorizzazione **LU** : L triangolare inferiore, $\ell_{ii} = 1 \Rightarrow$ Gauss

Fattorizzazione $\mathbf{LL^T}$: L con elementi diagonali positivi $\Rightarrow$ Cholesky

Fattorizzazione **QR** : Q ortogonale, R triangolare superiore $\Rightarrow$ Householder

Riformulazione matriciale del metodo di Gauss

I vantaggi di fattorizzare A nel prodotto LU derivano dal fatto che L ed U non dipendono dal termine noto. Poiché il costo computazionale della procedura di eliminazione è $\sim\frac{2}{3}n^3$flops si ha un risparmio di operazioni se si devono risolvere più sistemi lineari che hanno la stessa matrice.

Sia : $A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$

e: $L_1 = \begin{bmatrix} 1 & & & 0 \\ m_{21} & 1 & & \\ \vdots & & \ddots & \\ m_{n1} & 0 & & 1 \end{bmatrix}$ con $m_{i1} = -\frac{a_{i1}}{a_{11}}$ $\quad i = 2,\dots n$

Il prodotto L_1A equivale al primo passo di Gauss.

In generale, il passo i-esimo e' $L_i A$, dove:

$$L_i = \begin{bmatrix} 1 & & & & \\ & \ddots & & 0 & \\ & m_{ji} & 1 & & \\ 0 & \vdots & 0 & \ddots & \\ & m_{ni} & & & 1 \end{bmatrix} \quad \text{con } m_{ji} = -\frac{a_{ji}}{a_{ii}} \quad j = i+1,\dots n$$

Alla fine si ha: $U = L_{n-1}L_{n-2}\dots L_2L_1A$

Poniamo: $\tilde{L} = L_{n-1}\dots L_1 \;\Rightarrow\; U = \tilde{L}A$; $A = \tilde{L}^{-1}U$ e ponendo $L = \tilde{L}^{-1}$ si ha: $A = LU$.

La soluzione di

$$Ax = b \quad \Leftrightarrow \quad LUx = b$$

si trova in due passi:

i) si pone: $Ly = b$ e si risolve per y

ii) da: $Ux = y$ si trova x.

La fattorizzazione LU può essere combinata con il pivoting parziale e con lo scaling dei fattori mediante la premoltiplicazione di matrici di permutazione.

Non c'è comunque unicità nella scelta di L ed U se L ed U sono generiche. Ciò si può vedere in due modi:

I. $$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} \ell_{11} & & 0 \\ \vdots & \ddots & \\ \ell_{n1} & \cdots & \ell_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \\ 0 & & u_{nn} \end{bmatrix}$$

Uguagliando i termini si hanno n^2 equazioni che però contengono ognuna $\frac{n(n+1)}{2}$ incognite per un totale di $n^2 + n$ incognite; n di esse vanno quindi determinate arbitrariamente.

II. Siano L_1U_1 ed L_2U_2 due fattorizzazioni di A:

$$A = L_1U_1 = L_2U_2 \quad \Rightarrow \quad L_2^{-1}L_1 = U_2U_1^{-1}$$

Poiché la matrice a sinistra è triangolare inferiore e quella a destra è triangolare superiore, perche' esse siano uguali devono necessariamente essere diagonali. Indicando tale matrice diagonale con D, si ha: $L_1 = L_2D$, $U_1 = D^{-1}U_2$

Scegliendo come costanti arbitrarie del punto I. :

$$\ell_{11} = \ell_{22} = \ldots = \ell_{nn} = 1$$

si ha il metodo di **Doolittle** che è il metodo di fattorizzazione equivalente all'eliminazione gaussiana.

Scegliendo invece:

$$u_{11} = u_{22} = \ldots = u_{nn} = 1$$

si ha il metodo di **Crout.**

Da un punto di vista computazionale, è possibile memorizzare le matrici L ed U nella stessa area di memoria di A.

Comunque, non sempre esiste una fattorizzazione LU di A.

Esempio: $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Sebbene esista A^{-1} non è possibile fattorizzare A.

Invece la matrice I + A, che è singolare, ha una fattorizzazione LU.

Teorema.

$A \in C^{nxn}$, $\det(A_k) \neq 0$ $k = 1, \ldots, n-1 \Rightarrow \exists_1 L, U : A = LU$

Corollario. Sotto le stesse ipotesi del teorema, esiste un'unica fattorizzazione di Doolittle e si ha che: $\det(A) = \prod_{i=1}^{n} u_{ii}$.

Dimostrazione del corollario.

Si dimostra per induzione. Per k = 1 si ha: $A_1 = L_1U_1$ ovvero $a_{11} = \ell_{11}u_{11} = u_{11} \Rightarrow \exists_1 u_{11}$

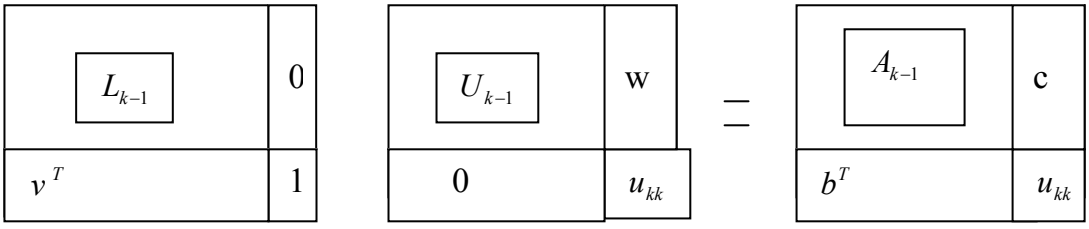
Sia vera la tesi per k-1, cioè $\exists_1 L_{k-1}, U_{k-1}$ per i quali si abbia:

$A_{k-1} = L_{k-1}U_{k-1}$, $\det(A_{k-1}) = \prod_{i=1}^{k-1} u_{ii}$

e dimostriamo che:

$A_k = L_kU_k$ $\det(A_k) = \prod_{i=1}^{k} u_{ii}$.

Effettuiamo il prodotto a blocchi:



L . U = A

$L_{k-1}U_{k-1} = A_{k-1}$ vero per le ipotesi di induzione.

$L_{k-1}w = c$ $\det(L_{k-1}) = 1 \Rightarrow \exists_1 w : L_{k-1}w = c$

$v^T U_{k-1} = b^T$ $\det(A_{k-1}) = \det(L_{k-1})\det(U_{k-1}) = \det(U_{k-1}) \Rightarrow \exists_1 v : U_{k-1}^T v = b$

$v^T w + u_{kk} = a_{kk}$, $u_{kk} = a_{kk} - v^T w$

Ma v, w, a_{kk} sono unici $\Rightarrow$ anche u_{kk} è unico.

$\Rightarrow A_k = L_k U_k$, $\det(A_k) = \det(L_k)\det(U_k) = \prod_{i=1}^{k} u_{ii}$. •

Le ipotesi del teorema pero' non sono facili da verificare.

Se A e' tale che $\det(A) \neq 0$ $\Rightarrow$ $\exists$ P matrice di permutazione :

$$PA = LU$$

Per due tipi di matrici non è necessario uno scambio di righe o di colonne per aversi la fattorizzazione LU: diagonalmente dominanti, simmetriche definite positive.

Metodo di Cholesky.

Teorema.

Sia $A \in Mat(n,n)$, $A = A^T$, $x^T A x > 0$ $\Rightarrow$ per $\forall\, x \neq 0$ esiste almeno una L triangolare inferiore :

$$A = LL^T$$

Se si impone che $\ell_{ii} > 0$ la fattorizzazione è unica.

Dimostrazione.

Per il criterio di Sylvester: $\det(A_k) > 0$ $\forall$ k.

Per il teorema precedente esiste un'unica fattorizzazione LU. Ponendo:

$$\begin{bmatrix} u_{11} & & 0 \\ \vdots & \ddots & \\ 0 & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \\ 0 & & u_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

si ha: $a_{kk} = \sum_{p=1}^{k} u_{pk}^2 = u_{kk}^2 + \sum_{p=1}^{k-1} u_{pk}^2 \Rightarrow u_{kk}^2 = a_{kk} - \sum_{p=1}^{k-1} u_{pk}^2$

$$a_{kj} = \sum_{i=1}^{k} u_{ki} u_{ij} = u_{kk} u_{kj} + \sum_{i=1}^{k-1} u_{ki} u_{ij} \Rightarrow u_{kj} = \left(a_{kj} - \sum_{i=1}^{k-1} u_{ki} u_{ij} \right) \Big/ u_{kk} \quad k > j$$

da cui si ha il metodo di Cholesky:

$$u_{11} = \sqrt{a_{11}}$$

$$u_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} u_{ik} u_{kj} \right) \Big/ u_{ii} \qquad i = 2,\ldots,n$$

$$u_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} u_{ik}^2 \right)^{1/2}$$

Il costo computazionale di questo metodo e' meta' di quello di Gauss e cioe' $\frac{2}{3}n^3$.

Tale metodo è il migliore per le matrici simmetriche definite positive poiché non distrugge la simmetria della matrice.

I metodi fin qui visti utilizzano un numero di operazioni $\propto n^3$ se n è l'ordine della matrice. Mostriamo pero' che in alcuni casi speciali, come quello delle matrici tridiagonali, tale costo e' di ordine *n* se il metodo e' utilizzato ad hoc.

Sistemi tridiagonali

$a_{ij} = 0 \quad : |i - j| > 1$. Scriviamo la matrice, che ha $3n$-2 elementi, come prodotto di due matrici particolari le cui incognite sono α_i, *i=1,…,n* e γ_i, *i=1,…,n-1.*

$$\begin{bmatrix} a_1 & c_1 & 0 & & & & \\ b_2 & a_2 & c_2 & 0 & & 0 & \\ & b_3 & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & 0 & & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & & & b_n & a_n \end{bmatrix} = \begin{bmatrix} \alpha_1 & & & & \\ b_2 & \alpha_2 & & 0 & \\ & b_3 & \alpha_3 & & \\ & 0 & \ddots & \ddots & \\ & & & b_n & \alpha_n \end{bmatrix} \begin{bmatrix} 1 & \gamma_1 & & & \\ & 1 & \gamma_2 & 0 & \\ & & \ddots & & \\ & 0 & & \ddots & \gamma_{n-1} \\ & & & & 1 \end{bmatrix}$$

$a_1 = \alpha_1$ $\quad\quad$ $\alpha_1\gamma_1 = c_1$

$a_i = \alpha_i + b_i\gamma_{i-1}$ $\quad\quad$ $i = 2,\ldots,n$

$\alpha_i\gamma_i = c_i$ $\quad\quad$ $i = 2,\ldots,n-1$

$\Rightarrow$

$\alpha_1 = a_1$ $\quad\quad$ $\gamma_1 = c_1/\alpha_1$

$\alpha_i = a_i - b_i\gamma_{i-1}$ $\quad\quad$ $i = 2,\ldots,n$

$\gamma_i = c_i/\alpha_i$ $\quad\quad$ $i = 2,\ldots,n-1$

Costo computazionale: 8n - 7 flops.

I metodi di fattorizzazione modificano la matrice iniziale e, se questa è *sparsa*, cioè ha molti zeri, si hanno problemi di memoria. In tali casi e' piu' conveniente utilizzare i metodi iterativi.

Metodi iterativi

I metodi iterativi generano una successione di vettori $\{x^{(k)}\}_{k\in N}$ che si spera converga alla soluzione di $A\underline{x} = \underline{b}$. La matrice A non viene modificata.

Sia $A \in Mat(n,n)$, $\det(A) \neq 0$. Poniamo:

$$Ax = b$$

$$A = M - N$$

$$(M - N)x = b$$

$$Mx_{n+1} = Nx_n + b$$

$$x_{n+1} = M^{-1}Nx_n + M^{-1}b$$

Una decomposizione o *splitting* di A si dice *regolare* se: $\det(M) \neq 0$, $M^{-1} \geq 0$, $N \geq 0$.

Un metodo iterativo è detto *convergente* se per qualunque vettore iniziale x_0 la successione $\{x^{(k)}\}_{k\in N}$ è convergente.

Teorema.

Sia $A = M - N$ uno splitting regolare di A e sia: $||M^{-1}N|| \leq \lambda < 1$. Allora:

I) A è non singolare

II) Il metodo iterativo associato a tale splitting è convergente

III) $||x_n - x|| \leq \lambda^n ||x_0 - x||$ che dà un limite all'errore commesso.

Dim:

I) per assurdo sia A singolare. $Ay = 0$ ha almeno una soluzione non banale:

$$y \neq 0 : (M - N)y = 0, \quad y = M^{-1}Ny$$

$$||y|| \leq ||M^{-1}Ny|| \leq \lambda ||y|| \quad \rightarrow \lambda \geq 1 : \text{ assurdo}$$

III) Poniamo: $P = M^{-1}N$, $e_n = x_n - x$

$$e_n = Pe_{n-1} = \ldots = P^n e_0$$

$$||e_n|| \leq ||P^n|| \, ||e_0|| \leq \lambda^n ||e_0||$$

II) $\lim_{n\to\infty} \|e_n\| = \lim_{n\to\infty} \|x_n - x\| = \lim_{n\to\infty} \lambda^n \|e_0\| = 0$ poiché $\lambda < 1$. •

Condizioni necessarie per la convergenza di un metodo iterativo di facile verifica:

- poiché il determinante di una matrice è il prodotto degli autovalori, allora se $|\det(M^{-1}N)| \geq 1$ almeno uno degli autovalori è ≥ 1 e quindi il metodo non può convergere.
- Poiché la traccia(*) di una matrice è la somma degli autovalori, allora se $|t_r(M^{-1}N)| \geq n$ almeno uno degli autovalori è ≥ 1 e quindi il metodo non può convergere.

  Quindi: $|\det(M^{-1}N)| < 1$, $|t_r(M^{-1}N)| < n$ sono condizioni necessarie per la convergenza del metodo.

  (*) ricordiamo che: $t_a(A) = \sum_{i=1}^{n} a_{ii}$.

Teorema.

Condizione necessaria e sufficiente perché un metodo iterativo sia convergente è che $\rho(M^{-1}N) < 1$.

Metodo di Jacobi

Sia dato un sistema lineare di ordine 3.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \end{cases} \quad \text{con } a_{11}, a_{22}, a_{33} \neq 0.$$

Ricaviamo le componenti:

$$\begin{cases} x_1 = (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 = (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \\ x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \end{cases}$$

Partendo da un vettore iniziale arbitrario $x^{(0)} \in R^3$ si genera la successione $x^{(k)}$ dalle relazioni:

$$\begin{cases} x_1^{(k+1)} = (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)})/a_{11} \\ x_2^{(k+1)} = (b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)})/a_{22} \\ x_3^{(k+1)} = (b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)})/a_{33} \end{cases}$$

Per un sistema generale, il metodo di Jacobi è:

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \qquad i = 1,\ldots,n$$

Metodo di Gauss-Seidel

Poiché nella prima sommatoria si usano le componenti "vecchie" si può usare una variante che tiene conto delle "nuove" componenti e ciò dà luogo al metodo di Gauss-Seidel che in generale è più veloce del metodo di Jacobi.

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \qquad i = 1,\ldots,n$$

Formulazione matriciale dei metodi di Jacobi e Gauβ-Seidel.

Decomponiamo A: $A = D - E - F$

dove D è la diagonale di A, E la sua parte inferiore ed F quella superiore.

$$(D - E - F)x = b$$

$$Dx^{(k+1)} = (E + F)x^{(k)} + b$$

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b$$

La matrice: $M_J = D^{-1}(E + F)$ e' la *matrice di Jacobi*.

Il metodo di Jacobi converge se A è strettamente diagonalmente dominante (c.s.) ovvero se:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad i = 1, \ldots, n$$

si ha:

$$(D - E)x^{(k+1)} = Fx^{(k)} + b$$

$$x^{(k+1)} = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b \qquad M_{GS} = (D - E)^{-1}F$$

che dà il metodo di Gauβ - Seidel.

Converge se A è simmetrica definita positiva (condizione sufficiente):

$$a_{ij} = a_{ji}$$

$$x^T Ax > 0 \quad \forall\, x \neq 0$$

e converge anche se A è strettamente diagonalmente dominante.

Tali metodi sono molto lenti se $\rho(M^{-1}N) \sim 1$, dove:

$$M = D, \qquad N = E + F \text{ in Jacobi}$$

$$M = D - E, \qquad N = F \text{ in Gauss-Seidel}$$

Per accelerare la convergenza si usano i *metodi di rilassamento*.

Metodo SOR (Successive Over-Relaxation)

Tale metodo consiste nel calcolare una iterata di Gauss-Seidel ed effettuare una correzione dipendente da un parametro ω:

$$x^{(k+1)} = \omega\ \hat{x}^{(k+1)} + (1 - \omega)x^{(k)}$$

dove $\hat{x}^{(k+1)}$ è il passo (k+1) di G.S.

Ricaviamo tale schema:

$$Ax = b \rightarrow \omega Ax = \omega b$$

$$Dx + \omega(D - E - F)x = \omega b + Dx$$

$$Dx - \omega Ex = Dx + \omega(F - D)x + \omega b$$

Se $\omega = 1$ si ha G.S. . Se $\omega \neq 0$ la parte sinistra è triangolare inferiore. Introduciamo L ed R:

$$L = D^{-1}E, \quad R = D^{-1}F$$

$$x^{(k+1)} = H(\omega)x^{(k)} + \omega(D - \omega E)^{-1}b$$

dove:

$H(\omega) = (D - \omega E)^{-1}[D(1-\omega) + \omega F] = [D(I - \omega L)]^{-1}D[(1-\omega)I + \omega R] = (I - \omega L)^{-1}D^{-1}D[(1-\omega)I + \omega R] =$
$= (I - \omega L)^{-1}[(1-\omega)I + \omega R]$

Convergenza per SOR

Teorema.

$\rho(H(\omega)) \geq |\omega-1| \quad \forall\ \omega \in R.$

Pertanto SOR diverge se $\omega \leq 0$ oppure $\omega \geq 2$ e si ha convergenza per: $0 < \omega < 2$

Dim: Siano λ_i gli autovalori di $H(\omega)$. Si ha:

$$\left|\prod_{i=1}^{n}\lambda_i\right| = \det(H(\omega)) = \det[(I - \omega L)^{-1}]\ \det[(1-\omega)I + \omega R] = |1 - \omega|^n$$

Pertanto deve esistere almeno un λ_i tale che $|\lambda_i| \geq |1 - \omega|$ e perché ci sia convergenza deve essere $|1 - \omega| < 1$ cioè $0 < \omega < 2$.

Se A è simmetrica definita positiva, $0 < \omega < 2$ è condizione necessaria e sufficiente.

Se A è strettamente diagonalmente dominante, $0 < \omega \leq 1$ è condizione necessaria e sufficiente.

Criterio di arresto per i metodi iterativi.

Data una tolleranza ε, un metodo iterativo si deve fermare quando:

$$\frac{\left\| x^{(k+1)} - x^{(k)} \right\|_\infty}{\left\| x^{(k)} \right\|_\infty} < \varepsilon$$

Poiché ciò potrebbe non verificarsi mai, bisogna introdurre un altro criterio di arresto dato dal numero massimo di iterazioni da eseguire.

Velocità di convergenza

Stimiamo il numero k di iterazioni necessarie per ridurre l'errore iniziale di un fattore 10^{-m} o più.

$$||e^{(k)}|| \leq 10^{-m} ||e^{(0)}||$$

$$||e^{(k)}|| \leq ||(M^{-1}N)^k|| \, ||e_0|| \leq 10^{-m} ||e^{(0)}||$$

$$||(M^{-1}N)^k|| \leq 10^{-m}$$

$$\rho((M^{-1}N)^k) = (\rho(M^{-1}N))^k$$

$$(\rho(M^{-1}N))^k \leq 10^{-m}$$

$$k \log(\rho(M^{-1}N)) \leq -m$$

perche' si abbia convergenza, deve essere $\rho(M^{-1}N) < 1$ e quindi:

$$k \geq \frac{m}{-\log \rho\left(M^{-1}N\right)}$$

$R = -\log(\rho(M^{-1}N))$ velocità di convergenza

$$k \geq \frac{m}{R}$$

Maggiore è R (e quindi più piccolo è ρ) minore è k.

<u>Metodo del gradiente</u>

Per matrici simmetriche definite positive, la risoluzione del sistema lineare:

$$Ax = b$$

è equivalente a trovare il punto di minimo $\underline{x} \in R^n$ della forma quadratica:

$$\phi(\underline{y}) \equiv \frac{1}{2}\underline{y}^T A\underline{y} - \underline{y}^T b$$

calcolando infatti il gradiente di ϕ, che ha componenti: $\frac{\partial \phi}{\partial y_i}$ $i = 1, \ldots, n$ si ha:

$$\nabla \phi(\underline{y}) = \frac{1}{2}(A^T + A)\underline{y} - \underline{b} = A\underline{y} - \underline{b}$$

poiché $A^T = A$. Pertanto: $Ax = b \Leftrightarrow \nabla \phi(\underline{y}) = 0$

Problema: determinare x minimo di ϕ partendo da $\underline{x}^{(0)} \in R^n$ e quindi scegliere opportune direzioni lungo le quali avvicinarsi ad x. Tale direzione non è nota a priori. Sia:

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k \underline{d}^{(k)}$$

α_k = lunghezza del passo lungo la direzione $\underline{d}^{(k)}$.

Una delle scelte per tale direzione e' direzione di discesa piu' rapida: metodo *steepest descent*.

$$\nabla \phi(\underline{x}^{(k)}) = A\underline{x}^{(k)} - \underline{b} = - r^{(k)}$$

$$\underline{d}^{(k)} = \nabla \phi(\underline{x}^{(k)})$$

α_k si calcola minimizzando ϕ:

$$\phi(\underline{x}^{(k+1)}) = \frac{1}{2}(\underline{x}^{(k)} + \alpha_k \underline{r}^{(k)})^T A(\underline{x}^{(k)} + \alpha_k \underline{r}^{(k)}) - (\underline{x}^{(k)} + \alpha_k \underline{r}^{(k)})^T b$$

$$\frac{\partial \phi}{\partial \alpha_k} = 0 \Rightarrow \alpha_k = \frac{\underline{r}^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}}$$

Ciò ha una semplice interpretazione geometrica nel caso n = 2.

Sia $A = \mathrm{diag}(\lambda_1, \lambda_2)$, $0 < \lambda_1 \leq \lambda_2$, $\underline{b} = (b_1, b_2)^T$

Le curve $\phi(x_1, x_2) = c$ descrivono una successione di ellissi.

Se $\lambda_1 = \lambda_2$ si hanno dei cerchi e il metodo converge in una sola iterazione poiché la direzione del gradiente passa per il centro. Se invece $\lambda_2 >> \lambda_1$ il metodo converge lentamente.

N.B. Se la matrice A non è simmetrica il metodo è applicato alla matrice A^TA che è simmetrica e si risolve il sistema equivalente:

$$A^TAx = A^Tb$$

La convergenza del metodo è migliorata se come direzione di discesa non si sceglie quella più ripida, determinata dal gradiente, ma si sceglie la direzione coniugata. Si ha quindi il metodo dei gradienti coniugati.