Metodi diretti

Definizione numero di condizionamento: data una matrice non singolare $A \in \mathbb{R}^{nxn}$, il numero di condizionamento di A è il prodotto tra la norma della matrice e la norma della sua inversa:

$$K_p(A) = \|A\|_p \|A^{-1}\|_p \text{ con } 1 \leq p \leq +\infty$$

$$K_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sqrt{\frac{\lambda_{max}(A^T A)}{\lambda_{min}(A^T A)}}$$

Definizione numero di condizionamento spettrale: data una matrice non singolare $A \in \mathbb{R}^{nxn}$, il numero di condizionamento spettrale di A è il prodotto tra il raggio spettrale della matrice e il raggio spettrale della sua inversa:

$$K(A) = \rho(A)\rho(A^{-1}) = \frac{max|\lambda_i(A)|}{min|\lambda_i(A)|}$$

Corollario: $\overrightarrow{e_{rel}} = \frac{\|\vec{e}\|}{\|\vec{x}\|} \leq K_2(A)\overrightarrow{r_{rel}} = (A)\frac{\|\vec{r}\|}{\|\vec{b}\|}$ $\quad \vec{r} = \vec{b} - A\hat{\vec{x}}$ $\quad \vec{e} = \vec{x} - \hat{\vec{x}}$

Matrice triangolare inferiore

$$x_1 = \frac{b_1}{l_{11}}$$

$$x_i = \frac{1}{l_{ii}}\left(b_i - \sum_{j=1}^{i-1} l_{ij}x_j\right) \qquad \forall i = 2, \dots, n$$

n° operazioni: n^2
Risoluzione con metodo della sostituzione in avanti

```
function [ y ] = sostavanti( L,b )

n = length(b);
y = zeros(n,1);
y(1) = b(1)/L(1,1);

for i = 2:n
    y(i) = (b(i)-L(i,1:i-1)*y(1:i-1))/L(i,i);
end
```

Matrice triangolare superiore

$$x_n = \frac{b_n}{u_{nn}}$$

$$x_i = \frac{1}{u_{ii}}\left(b_i - \sum_{j=i+1}^{n} u_{ij}x_j\right) \qquad \forall i = n-1, \dots, 1$$

n° operazioni: n^2
Risoluzione con metodo della sostituzione indietro

```
function x=sost_indietro(U,y)

n=length(y);
x = zeros(n,1);
x(n) = y(n) / U(n,n);

for i=n-1:-1:1
    x(i) = (y(i) - U(i,i+1:n) * x(i+1:n)) / U(i,i);
end
```

Matrici generali	n° operazioni: $\frac{2}{3}n^3$ Risoluzione con metodo della fattorizzazione LU con o senza pivoting <u>Condizione necessaria e sufficiente per l'esistenza della fattorizzazione LU</u>: la matrice $A \in \mathbb{R}^{nxn}$ ammette un'unica fattoriazione LU se e solo se le sottomatrici principali di A sono non singolari, ovvero $\det(A_i) \neq 0 \quad \forall i = 1, \dots, n-1$ <u>Condizioni sufficienti per l'esistenza della fattorizzazione LU</u>: la matrice $A \in \mathbb{R}^{nxn}$ ammette un'unica fattoriazione LU se è verificata una delle seguenti condizioni: • A è simmetrica definita positiva • A è dominanza diagonale stretta per righe • A è a dominanza diagonale stretta per colonne
	`[L,U,P] = lu(A);`

Metodi iterativi

<u>Idea di base:</u> presa una matrice $A \in \mathbb{R}^{nxn}$ non singolare e dato il vettore iniziale $\vec{x}^{(0)} \in \mathbb{R}^n$ si ha che per $k = 0,1, \dots (crit.arresto)$ $\vec{x}^{(k+1)} = B\vec{x}^{(k)} + \vec{g}$

<u>Condizione necessaria e sufficiente per la convergenza dei metodi iterativi</u>: il metodo iterativo converge alla soluzione esatta $\vec{x}$ del sistema lineare $A\vec{x} = \vec{b}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ se e solo se $\rho(B) < 1$. Inoltre la convergenza è tanto più rapida tanto più $\rho(B)$ è piccolo

<u>Criteri d'arresto</u>:
1. Residuo: $\tilde{e}^{(k)} = \|\vec{r}^k\|$
2. Residuo relativo: $\widetilde{e_{rel}}^{(k)} = \frac{\|\vec{r}^k\|}{\|\vec{b}\|}$
3. Differenza tra iterate successive: $\tilde{e}^{(k)} = \vec{x}^{(k)} - \vec{x}^{(k-1)}$

Metodo di Richardson stazionario	<u>Parametro</u>: $\alpha_k = \alpha = cost$ <u>Matrice delle iterate</u>: $B_\alpha = I - \alpha P^{-1}A$ <u>Convergenza metodo di Richardson stazionario</u>: siano $P, A \in \mathbb{R}^{nxn}$ matrici simmetriche e definite positive allora il metodo di Richardson stazionario converge alla soluzione esatta $\vec{x}$ del sistema lineare $A\vec{x} = \vec{b}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ se e solo se il parametro scelto α è compreso tra: $$0 < \alpha < \frac{2}{\lambda_{max}(P^{-1}A)}$$

<table>
<tr><td rowspan="2">Metodo di splitting standard</td><td>Parametro: $\alpha_k = \alpha = 1$

Matrice delle iterate: $B = I - P^{-1}A$

Scelta di P: secondo Jacobi la matrice P è una matrice diagonale che ha come elementi quelli della diagonale della matrice A. secondo Gauss-Seidel invece la matrice P è la matrice triangolare inferiore estratta da A

Condizioni sufficienti per la convergenza del metodo:
Se A è simmetrica definita positiva allora il metodo di Gauss-Seidel converge per ogni $\vec{x}^{(0)} \in \mathbb{R}^n$
Se A è dominanza diagonale stretta per righe allora i due metodi convergono per ogni $\vec{x}^{(0)} \in \mathbb{R}^n$
Se A è tridiagonale e ha elementi sulla diagonale diversi da 0 allora entrambi i metodi convergono o entrambi non convergono. Se convergono Gauss-Seidel converge più rapidamente di Jacobi perché $\rho(B_{GS}) = [\rho(B_J)]^2$</td></tr>
<tr><td>Dato $\vec{x}^{(0)} \in \mathbb{R}^n$
Calcolo $\vec{r}^{(0)} = \vec{b} - A\vec{x}^{(0)}$
Per $k = 0,1,\dots,(crit.arresto)$
$P\vec{z}^{(k)} = \vec{r}^{(k)}$
$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \vec{z}^{(k)}$
$\vec{r}^{(k+1)} = \vec{r}^{(k)} - A\vec{z}^{(k)}$</td></tr>
<tr><td rowspan="3">Metodo del gradiente</td><td>Parametro: $\alpha_k = \frac{{\vec{r}^{(k)}}^T \vec{r}^{(k)}}{{\vec{r}^{(k)}}^T A\vec{r}^{(k)}}$ $\quad P = I$

Matrice delle iterate: $B_\alpha = I - \alpha_k A$

Convergenza metodo del gradiente: se A è simmetrica e definita positiva il metodo del gradiente converge a $\vec{x}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ e:
$$\left\|\vec{e}^{(k)}\right\|_A \leq d^k \left\|\vec{e}^{(0)}\right\|_A$$
$$d^k = \frac{K(A) - 1}{K(A) + 1}$$</td></tr>
<tr><td>Dato $\vec{x}^{(0)} \in \mathbb{R}^n$
Calcolo $\vec{r}^{(0)} = \vec{b} - A\vec{x}^{(0)}$
Per $k = 0,1,\dots,(crit.arresto)$
$\alpha_k = \frac{{\vec{r}^{(k)}}^T \vec{r}^{(k)}}{{\vec{r}^{(k)}}^T A\vec{r}^{(k)}}$
$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \vec{r}^{(k)}$
$\vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k A\vec{r}^{(k)}$</td></tr>
<tr><td><pre>function [x, res, niter, incr] = grad(x0, A, b, maxiter, toll)

res = zeros(maxiter,1);
incr = zeros(maxiter,1);
x = x0;
r = b-A*X;</pre></td></tr>
</table>

```
for k = 1:maxiter
    xold = x;
    alphak = (r'*r)/(r'*A*r);
    x = xold + alphak * r;
    r = b - A*x;
    res(k) = norm(r)/norm(b);
    incr(k) = norm(x-xold)/norm(xold);
    if res(k)>toll
        break;
    end
end

niter = k;
res   = res(1:k);
incr  = incr(1:k);

if niter<maxiter
    fprintf ('Si è raggiunta la convergenza')
    niter
else
    fprintf ('Non si è raggiunta la convergenza')
end
```

Metodo del gradiente precondizionato

Parametro: $\alpha_k = \frac{\vec{z}^{(k)^T}\vec{r}^{(k)}}{\vec{z}^{(k)^T}A\vec{z}^{(k)}}$ $\quad P \neq I$ è precondizionatore

Matrice delle iterate: $B = I - \alpha_k P^{-1}A$

Convergenza metodo del gradiente precondizionato: se A è simmetrica e definita positiva il metodo del gradiente precondizionato converge a $\vec{x}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ e:

$$\left\|\vec{e}^{(k)}\right\|_A \leq {d_p}^k \left\|\vec{e}^{(0)}\right\|_A$$

$${d_p}^k = \frac{K(P^{-1}A) - 1}{K(P^{-1}A) + 1}$$

Dato $\quad \vec{x}^{(0)} \in \mathbb{R}^n$

Calcolo $\quad \vec{r}^{(0)} = \vec{b} - A\vec{x}^{(0)}$

Per $\quad k = 0,1,\dots,(crit.arresto)$

$$P\vec{z}^{(k)} = \vec{r}^{(k)}$$
$$\alpha_k = \frac{\vec{z}^{(k)^T}\vec{r}^{(k)}}{\vec{z}^{(k)^T}A\vec{z}^{(k)}}$$
$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k\vec{z}^{(k)}$$
$$\vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k A\vec{z}^{(k)}$$

Metodo del gradiente coniugato

Parametro: $\alpha_k = \frac{p^{(k)^T}\vec{r}^{(k)}}{\vec{p}^{(k)^T}A\vec{p}^{(k)}}$ $\qquad \beta_k = \frac{\vec{p}^{(k)^T}\vec{p}^{(k)}}{\vec{p}^{(k)^T}A\vec{p}^{(k)}}$ $\qquad P \neq I$

Convergenza metodo del gradiente coniugato: se A è simmetrica e definita positiva il metodo del gradiente coniugato converge a $\vec{x}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ in n iterazioni (in aritmetica esatta):

$$\left\|\vec{e}^{(k)}\right\|_A \leq 2c^k \left\|\vec{e}^{(0)}\right\|_A$$

	$$c^k = \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}$$
	Dato $\vec{x}^{(0)} \in \mathbb{R}^n$ Calcolo $\vec{r}^{(0)} = \vec{b} - A\vec{x}^{(0)}$ $\vec{p}^{(0)} = \vec{r}^{(0)}$ Per $k = 0,1, \dots, (crit.arresto)$ $\alpha_k = \frac{{p^{(k)}}^T \vec{r}^{(k)}}{{\vec{p}^{(k)}}^T A\vec{p}^{(k)}}$ $\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \vec{p}^{(k)}$ $\vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k A\vec{p}^{(k)}$ $\beta_k = \frac{{\vec{p}^{(k)}}^T \vec{p}^{(k)}}{{\vec{p}^{(k)}}^T A\vec{p}^{(k)}}$ $\vec{p}^{(k+1)} = \vec{r}^{(k+1)} - \beta_k \vec{p}^{(k)}$
Metodo del gradiente coniugato precondizionato	<u>Parametro</u>: $\alpha_k = \frac{{\vec{z}^{(k)}}^T \vec{r}^{(k)}}{{\vec{z}^{(k)}}^T A\vec{z}^{(k)}}$ $\beta_k = \frac{{\vec{P}^{(k)}}^T \vec{r}^{(k)}}{{\vec{P}^{(k)}}^T A\vec{P}^{(k)}}$ $P \neq I$ <u>Convergenza metodo del gradiente coniugato precondizionato</u>: se A è simmetrica e definita positiva il metodo del gradiente coniugato precondizionato converge a $\vec{x}$ per ogni scelta del vettore iniziare $\vec{x}^{(0)} \in \mathbb{R}^n$ in n iterazioni (in aritmetica esatta): $$\left\|\vec{e}^{(k)}\right\|_A \leq 2{c_p}^k \left\|\vec{e}^{(0)}\right\|_A$$ $${c_p}^k = \frac{\sqrt{K(P^{-1}A)} - 1}{\sqrt{K(P^{-1}A)} + 1}$$

Equivalenza tra minimo energia e soluzione del sistema

Enunciato: risolvere il sistema lineare $A\vec{x} = \vec{b}$ equivale a minimizzare l'energia $\Phi(\vec{y})$ (ovvero una funzione di stato che ha la proprietà di essere quadratica e non negativa e di decrescere nel tempo qualora il sistema si asintoticamente stabile ed evolva liberamente) definita nel seguente modo:

$$\Phi(\vec{y}) = \frac{1}{2}\vec{y}^T A\vec{y} - \vec{y}^T\vec{b}$$

Quindi si ha che:

$$x \text{ è } soluzione\ di\ A\vec{x} = \vec{b} \quad \Leftrightarrow \quad \nabla\Phi(\vec{y}) = 0$$

Dimostrazione: per dimostrare questo teorema occorre dimostrare le due implicazioni:

- Si suppone che x sia soluzione del sistema $A\vec{x} = \vec{b}$ e si dimostra che sia il minimo della funzione Φ

$$\Phi(\vec{x} + \vec{v}) = \frac{1}{2}(\vec{x} + \vec{v})^T A(\vec{x} + \vec{v}) - (\vec{x} + \vec{v})^T\vec{b} =$$

$$= \frac{1}{2}\vec{x}^T A\vec{x} + \frac{1}{2}\vec{x}^T A\vec{v} + \frac{1}{2}\vec{v}^T A\vec{x} + \frac{1}{2}\vec{v}^T A\vec{v} - \vec{x}^T\vec{b} - \vec{v}^T\vec{b} =$$

$$= \Phi(\vec{x}) + \frac{1}{2}\vec{x}^T A\vec{v} + \frac{1}{2}\vec{v}^T A\vec{x} + \frac{1}{2}\vec{v}^T A\vec{v} - \vec{v}^T\vec{b} =$$

$\vec{x}^T A\vec{v} = (\vec{x}^T A\vec{v})^T = \vec{v}^T A\vec{x} = \vec{v}^T\vec{b}$

$$= \Phi(\vec{x}) + \frac{1}{2}\vec{v}^T\vec{b} + \frac{1}{2}\vec{v}^T\vec{b} + \frac{1}{2}\vec{v}^T A\vec{v} - \vec{v}^T\vec{b} =$$

$$= \Phi(\vec{x}) + \frac{1}{2}\vec{v}^T A\vec{v} > \Phi(\vec{x})$$

Quindi per ogni $\vec{v} \neq 0$ vale che $\Phi(\vec{x} + \vec{v}) > \Phi(\vec{x})$ e quindi $\vec{x}$ è punto di minimo della funzione Φ.

- Si suppone che x sia punto minimo della funzione Φ e si dimostra che è soluzione del sistema $A\vec{x} = \vec{b}$
  Se $\vec{x}$ è il punto di minimo della funzione Φ, allora per il teorema di Fermat vale che $\nabla\Phi(\vec{x}) = \overline{0}$. Scriviamo $\Phi(\vec{y})$ nel seguente modo e poi calcoliamo le derivate parziali così da poter compilare $\nabla\Phi(\vec{y})$:

$$\Phi(\vec{y}) = \frac{1}{2}\sum_{i,j=1}^{n} a_{i,j}y_i y_j - \sum_{i=1}^{n} b_i y_i$$

$$\frac{\partial\Phi}{\partial y_k} = \frac{1}{2}\sum_{j=1}^{n} a_{k,j}y_j + \frac{1}{2}\sum_{i=1}^{n} a_{k,i}y_i - b_k = \sum_{i=1}^{n} a_{k,i}y_i - b_k = (A\vec{y} - \vec{b})_k$$

Da cui si determina:

$$\nabla\Phi(\vec{y}) = A\vec{y} - \vec{b} = -\vec{r}$$

$$\nabla\Phi(\vec{x}) = -\vec{r} = \overline{0} = -(A\vec{x} - \vec{b})$$

Per il teorema di Fermat $\nabla\Phi(\vec{x}) = \overline{0}$

Adeguatezza criterio della differenza tra iterate successive per metodi iterativi

Enunciato: per valutare l'errore nei metodi iterativi occorre utilizzare lo stimatore dell'errore. Tale stimatore può essere determinato in diversi modi, uno dei quali è proprio la differenza tra iterate successive ovvero:

$$\tilde{e}^{(k)} = \left|\left|\vec{\delta}^{(k)}\right|\right| = \left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|$$

Si vuole ora dimostrare quando questo criterio è soddisfacente, ovvero si vuole trovare una relazione che leghi lo stimatore dell'errore definito in questo modo con l'errore stesso.

Dimostrazione: per definizione la norme dell'errore al passo k viene definito nel seguente modo:

$$\left|\left|\vec{e}^{(k)}\right|\right| = \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right| =$$

$$= \left|\left|\vec{x} - \vec{x}^{(k+1)} - \vec{x}^{(k)} + \vec{x}^{(k+1)}\right|\right| =$$

Si aggiunge e sottrae $\vec{x}^{(k+1)}$

$$= \left|\left|\vec{x} - \vec{x}^{(k+1)}\right|\right| + \left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right| \leq$$

$$\leq \left|\left|\vec{e}^{(k+1)}\right|\right| + \left|\left|\vec{\delta}^{(k)}\right|\right|$$

Sapendo che: $\left|\left|\vec{e}^{(k+1)}\right|\right| \leq \rho(B) \left|\left|\vec{e}^{(k)}\right|\right|$

$$\left|\left|\vec{e}^{(k)}\right|\right| \leq \rho(B) \left|\left|\vec{e}^{(k)}\right|\right| + \left|\left|\vec{\delta}^{(k)}\right|\right|$$

$$\left|\left|\vec{e}^{(k)}\right|\right| - \rho(B) \left|\left|\vec{e}^{(k)}\right|\right| \leq \left|\left|\vec{\delta}^{(k)}\right|\right|$$

$$\left|\left|\vec{e}^{(k)}\right|\right| \left(1 - \rho(B)\right) \leq \left|\left|\vec{\delta}^{(k)}\right|\right|$$

$$\left|\left|\vec{e}^{(k)}\right|\right| \leq \frac{1}{1 - \rho(B)} \left|\left|\vec{\delta}^{(k)}\right|\right|$$

Ma per il criterio che stiamo usando $\left|\left|\vec{\delta}^{(k)}\right|\right| = \tilde{e}^{(k)}$ quindi:

$$\left|\left|\vec{e}^{(k)}\right|\right| \leq \frac{1}{1 - \rho(B)} \tilde{e}^{(k)}$$

Quindi:

- Se $\rho(B) \lesssim 1 \quad \rightarrow \quad \vec{e}^{(k)} \gg \tilde{e}^{(k)} \quad \rightarrow$ criterio insoddisfacente
- Se $\rho(B) \gtrsim 0 \quad \rightarrow \quad \vec{e}^{(k)} \cong \tilde{e}^{(k)} \quad \rightarrow$ criterio soddisfacente