Corso della scuola di dottorato:

NUMERICAL METHODS FOR INVERSE PROBLEMS

Metodi iterativi per la risoluzione di sistemi lineari

Dottorandi:

Mario Cascetta

Efisio Casti

Nicola Cau

Introduzione

Oltre ai metodi diretti per la risoluzione di un sistema lineare, si possono usare i metodi iterativi che si dimostrano efficienti e stabili anche quando si hanno delle matrici di grandi dimensioni. Scopo di questo lavoro è valutare la potenzialità di alcuni di questi metodi al variare dei parametri che caratterizzano il sistema attraverso delle sperimentazioni numeriche.

Metodi iterativi

I metodi iterativi per la risoluzione di un sistema lineare generano, partendo da un vettore iniziale $x(0)$, una successione di vettori definiti $x(k)$ con $k = 1,2,.....$ n, che, fissate determinate ipotesi, converge alla soluzione del problema.
A differenza dei metodi diretti, i metodi iterativi non richiedono nè la modifica della matrice del sistema nè la sua effettiva memorizzazione, ma è sufficiente poter accedere in qualche modo ai suoi elementi, risultando particolarmente convenienti nel caso di matrici di grandi dimensioni, strutturate o sparse.
Agli errori sperimentali e di arrotondamento, si aggiungono gli errori di troncamento, derivanti dal fatto che la soluzione cercata deve essere necessariamente approssimata troncando la successione per un indice sufficientemente grande.
Infine, è possibile ridurre di molto il tempo di elaborazione, eseguendo un minor numero di iterazioni in quei casi in cui non sia richiesta un'elevata accuratezza modificando il criterio di arresto.
I metodi che verranno adoperati in questa sede sono i seguenti:

- Metodo di Jacobi;
- Metodo di Gauss – Seidel;
- Metodo del gradiente.

Per la risoluzione di sistemi lineari del tipo $Ax=b$, si utilizzano dei metodi iterativi lineari e stazionari del primo ordine come quelli appena citati ed assumono la forma del tipo:

$$x^{(k+1)} = Bx^{(k)} + f$$

Il sistema considerato è lineare perchè lo è la relazione che lo esprime, stazionario perchè la matrice B non cambia all'avanzare dell'iterazione e del primo ordine poiché il vettore $x^{(k+1)}$ dipende solo da quello precedente $x^{(k)}$. E' opportuno, però, dare alcune definizioni riguardanti la convergenza e la consistenza del metodo in esame.

Un metodo si dice *globalmente convergente* se dato un vettore iniziale $x^{(0)}$ per ogni $x \in R^n$ si ha che:

$$\lim_{k\to\infty} \left\| x^{(k)} - x \right\| = 0$$

Mentre si dice *consistente* se:

$$x^{(k)} = x \Rightarrow x^{(k+1)} = x$$

Inoltre, va sottolineato che la consistenza è una *condizione necessaria* per la convergenza e che un metodo può essere consistente e non convergente, ma non il suo contrario.

La convergenza del metodo considerato dipende essenzialmente dalla matrice B; vi sono, infatti, due teoremi che stabiliscono le condizioni per la convergenza.

1. Un metodo iterativo lineare è convergente se esiste una norma consistente di B per la quale si abbia $\|B\| < 1$ che, per le proprietà matriciali, comporterebbe $\|e^k\| \to 0$ dove con e si intende l'errore sulla soluzione.
2. Un metodo iterativo lineare è convergente se e solo se il raggio spettrale della matrice di iterazione B è inferiore all'unità $\rho(B) < 1$.

In linea di principio, partendo dall'equazione Ax=b, è opprtuno scrivere la matrice A come differenza di due matrici:

$$A = P - N$$

in cui $\det(P) \neq 0$. Sostituendo all'equazione che diventa:

$$(P - N)x = b \quad \rightarrow \quad Px = Nx + b$$

E infine si trasforma arbitrariamente in un metodo iterativo:

$$x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b$$

Che è una forma analoga a quella indicata in precedenza ($x^{(k+1)} = Bx^{(k)} + f$).

Metodo di Jacobi

Per prima cosa è necessario scomporre la matrice A nella maniera seguente (*splitting additivo*):

$$A = D - E - F = \begin{pmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & a_{nn} \end{pmatrix} - \begin{pmatrix} 0 & 0 & \cdots & 0 \\ -a_{11} & 0 & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \ddots & -a_{n,n-1} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \vdots & \vdots \\ \vdots & \ddots & \ddots & -a_{n-1,n} \\ 0 & \ddots & 0 & 0 \end{pmatrix}$$

Nel caso particolare del metodo di Jacobi si considerano le seguenti matrici:

$$\begin{cases} P = D \\ N = E + F \end{cases}$$

Che portano poi alla forma canonica:

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

Se invece lo si vuole esprimere in forma di coordinate diventa:

$$x_j^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right] \qquad \text{con i=1,2……n}$$

Come ultima osservazione, si può dire che, poiché la comvergenza è globale, non dipende dal vettore iniziale $x^{(0)}$.

Metodo di Gauss-Seidel

Partendo sempre dallo splitting additivo effettuato in precedenza, il metodo di Gauss- Seidel pone la seguente situazione:

$$\begin{cases} P = D - E \\ N = F \end{cases}$$

Sostituendo nell'equazione principale porta ad avere:

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

che permette di ricavare $x^{(k+1)}$ risolvendo un sistema triangolare inferiore; esprimendo l'equazione in forma di coordinate si ha:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] \qquad \text{i=1,2…….n;}$$

Il metodo di Gauss Seidel non è parallelizzabile, in quanto la componente di $x^{(k+1)}$ dipende dai componenti aventi indice compreso tra 1 e i-1; nonostante ciò, però, in molte occasioni converge più velocemente del metodo di Jacobi e con numero di iterazioni decisamente inferiore. Non è detto che i metodi di Gauss Seidel e Jacobi siano direttamente applicabili a una matrice A, perché potrebbe avere zeri in diagonale, pertanto sarebbe opportuno effettuare prima qualche cambio di riga, inoltre, va aggiunto che la convergenza è globale, perciò non dipende dal vettore iniziale $x^{(0)}$.

Metodo del gradiente coniugato

Il metodo del gradiente coniugato è una variante del metodo del gradiente semplice da cui differisce per una scelta più accurata delle cosiddette direzioni di discesa, consentendo di convergere a soluzione tramite un numero di iterazioni molto inferiore alla dimensione del sistema, utile soprattutto se si opera con sistemi lineari di dimensione estremamente elevata. Per comprenderne la teoria, però è opportuno fare qualche passo indietro e capire, più in generale, da dove nasca il metodo del gradiente.

Sia A una matrice simmetrica definita positiva e si consideri la seguente forma quadratica:

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

Tale funzione raggiunge il suo valore minimo nel punto in cui si annulla il suo gradiente:

$$\nabla\phi(y) = \frac{1}{2}\left(A + A^T\right) y - b = Ay - b = 0$$

Si osserva, dunque, che il problema della minimizzazione appena mostrato equivale a trovare la soluzione del sistema lineare Ax = b. Nel caso del gradiente semplice, la soluzione si trova minimizzando la funzione $\phi(y)$ con un metodo iterativo non stazionario del tipo:

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

a partire da un vettore iniziale $x^{(0)}$, lungo le direzioni di decrescita $d^{(k)}$, con passi di lunghezza α_k. In questo caso, la direzione $d^{(k)}$ è quella di massima discesa (*steepest descent*), ossia quella opposta alla direzione del gradiente della funzione $\phi(y)$ nel punto $x^{(k)}$. Nel metodo del gradiente coniugato, invece, si parte dall'assunto per cui un vettore $x^{(k)}$ risulta ottimale rispetto ad una determinata direzione p se è rispettata la seguente condizione:

$$\phi\left(x^{(k)}\right) \leq \phi\left(x^{(k)} + \lambda p\right) \qquad \forall \lambda \in R$$

Il vettore $x^{(k)}$ è ottimale rispetto a p se e solo se la direzione p è ortogonale al residuo $r^{(k)}$, cioè:

$$p^T r^{(k)} = 0$$

Sostituendo nell'equazione principale della funzione $\phi(y)$ si avrà la situzione seguente:

$$\phi\left(x^{(k)} + \lambda p\right) = \frac{1}{2}\left(x^{(k)} + \lambda p\right)^T A\left(x^{(k)} + \lambda p\right) - \left(x^{(k)} + \lambda p\right)^T b$$

Eseguendo i calcoli si arriva ad una forma più semplificata:

$$\phi\left(x^{(k)} + \lambda p\right) = \phi\left(x^{(k)}\right) + \frac{1}{2}(p)^T A p^T \lambda^2 - (p)^T r^{(k)} \lambda$$

Derivando rispetto a λ e annullando la derivata si avrà che:

$$\frac{d}{d\lambda}\phi\left(x^{(k)}+\lambda p\right)= p^T Ap\lambda - p^T r^{(k)} = 0$$

Poiché $x^{(k)}$ è ottimale rispetto a p, $\phi(x^{(k)})$ deve avere un minimo per $\lambda = 0$, che sarebbe una soluzione banale, mentre imponendo $p^T r(k) = 0$ si avrebbe effettivamente la soluzione $\lambda = 0$ non banale che dimostrerebbe la tesi. Il problema del metodo del gradiente semplice è che l'ottimalità del vettore rispetto ad una certa direzione rimane solamente per il primo passo dell'iterazione; con il gradiente coniugato si vuole fare in modo che l'ottimalità del vettore k-esimo, quindi per i passi successivi al primo, rispetto ad una certa direzione venga ereditata da tutti i successivi elementi della successione.

Si suppone perciò, che $x^{(k)}$ sia ottimale rispetto a p, e quindi che $pT\ r^{(k)} = 0$, e si ponga:

$$x^{(k+1)} = x^{(k)} + q$$

Perchè anche $x^{(k+1)}$ sia ottimale rispetto a p è necessario che

$$0 = p^T r^{(k+1)} = p^T\left(r^{(k)} - Aq\right)= -p^T Aq$$

Che significa che le direzioni p e q devono essere A-ortogonali o A-coniugate; partendo con $p^{(0)}=r^{(0)}$ e considerando direzioni del tipo

$$p^{(k+1)} = r^{(k+1)} - \beta_k p^{(k)}$$

La condizione necessaria è che $p^{(k+1)}$ e $p^{(k)}$ siano due direzioni A-coniugate, cioè che:

$$\left(p^{(k)}\right)^T Ap^{(k+1)} = 0$$

si traduce nella condizione β_k data da:

$$\beta_k = \frac{\left(p^{(k)}\right)^T Ar^{(k+1)}}{\left(p^{(k)}\right)^T Ap^{(k)}}$$

Effettuando una scelta del parametro β_k di questo tipo si ha che:

$$\left(p^{(i)}\right)^T Ap^{(k+1)} = 0 \quad \text{i= 1,2…..,k-1}$$

cioè che la direzione $p^{(k+1)}$ è A-coniugata con tutte le direzioni generate precedentemente. Il fatto che il vettore $x^{(k+1)}$ seguente:

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

con α_k dato dalla seguente espressione:

$$\alpha_k = \frac{p^{(k)^T} r^{(k)}}{p^{(k)^T} Ap^{(k)}}$$

sia ottimale rispetto alle direzioni p(i), con i = 0, . . . , k, significa che lungo tali direzioni non è possibile far diminuire ulteriormente il valore della funzione obiettivo $\phi(y)$, e giustifica il fatto che l'algoritmo converga alla soluzione esatta in un numero finito di iterazioni.

Risultati

I risultati riportati nei grafici seguenti vogliono porre a confronto i tre metodi adoperati, che si ricordano essere il metodo di Jacobi, di Gauss- Seidel, e del gradiente coniugato. Le matrici di prova (simmetriche, definite positive) sono generate dal seguente comando presente in Matlab:

A = sprandsym(n,density,rc,kind)

Dove con *n* si indica la dimensione della matrice, con *density* la densità (ossia la quantità di valori diversi da zero presenti all'interno della matrice), il fattore *rc*, che rappresenta il reciproco del numero di condizionamento (indica, in pratica, quanto la matrice sia diagonale) ed, infine, il parametro *kind*, che posto uguale ad 1 consente di generare una matrice simmetrica definita positiva.

Fatto ciò, si sono costruiti gli algoritmi caratterizzati da un ciclo *while*, al cui interno è stato implementato un criterio di arresto per ottenere la convergenza che si verifica quando la differenza della norma della soluzione e del residuo con quelle ottenute nell'iterazione precedente risulta inferiore ad un determinato valore.

Nel caso dei metodi iterativi per la risoluzione di sistemi lineari, al passo k è possibile ottenere una maggiorazione per l'errore $e^{(k)} = x^{(k)} - x$ in termini del vettore residuo $r^{(k)}=b-Ax^{(k)}$. Eseguendo alcuni passaggi matematici, si arriva ad un criterio di arresto espresso come segue:

$$\frac{\left\| r^{(k)} \right\|}{\left\| b \right\|} \leq \tau$$

Poiché potrebbe verificare una non convergenza del metodo, si è deciso di fissare un numero massimo di iterazioni, al fine di evitare un loop infinito inserendo un ciclo *if* dove si impone l'arresto immediato al superamento di un determinato numero di cicli.

Passando alle sperimentazioni numeriche, si è partiti fissando la dimensione della matrice (n=100), ed il fattore *rc* pari a 0,9 (matrice fortemente diagonale) e si sono confrontati i tre metodi (fig. 1).

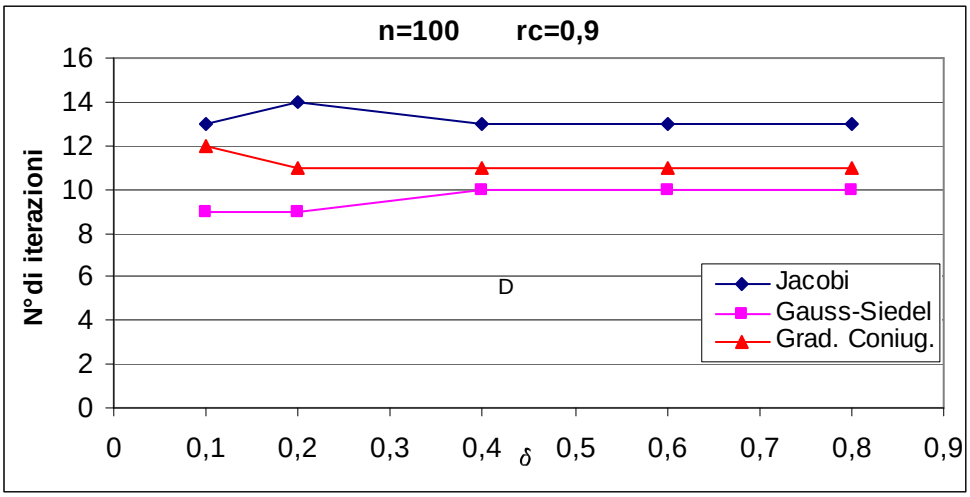

Figura 1: Numero di iterazioni necessarie al variare della densità.

Nella figura 1 si è osservato che il numero di iterazioni non cambia sostanzialmente con l'aumentare della densità della matrice che spazia nell'intervallo [0,1], con il metodo di Gauss-Seidel che richiede un numero di iterazioni inferiore rispetto agli altri due.

Nella figura 2 si è fatto variare il parametro *rc* (che in pratica rappresenta il reciproco del numero di condizionamento), mentre si sono mantenute costanti la densità e la dimensione della matrice. Si è notato che all'aumentare di *rc*, il numero di iterazioni decresce sensibilmente per tutti i metodi, ma con il metodo di Jacobi che, per valori bassi, non arriva a convergenza e per valori maggiori richiede sempre un numero di iterazioni superiore.

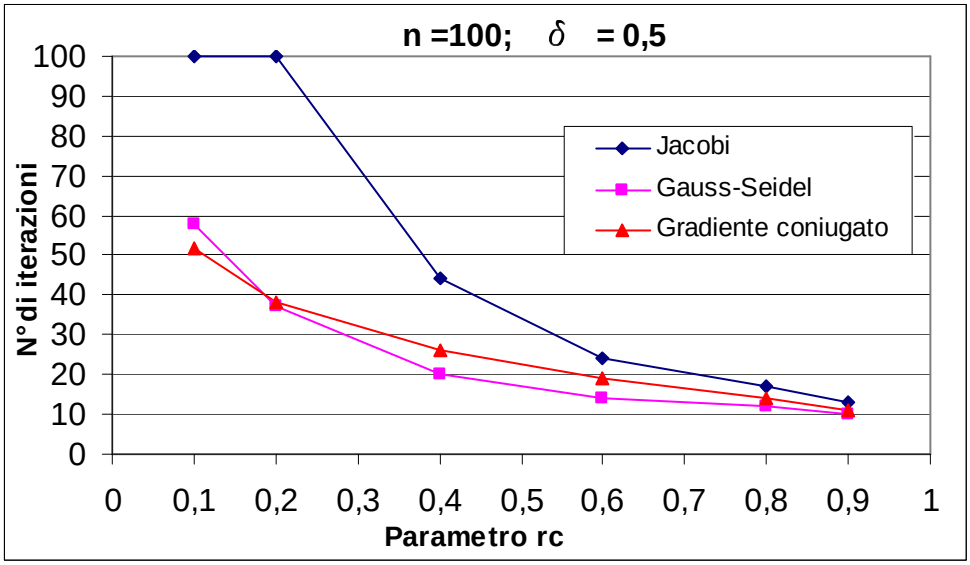

Figura 2

Nella figura 3 si è modificata la dimensione della matrice di partenza, osservando come il numero di iterazioni tenda a salire fino ad appiattirsi.

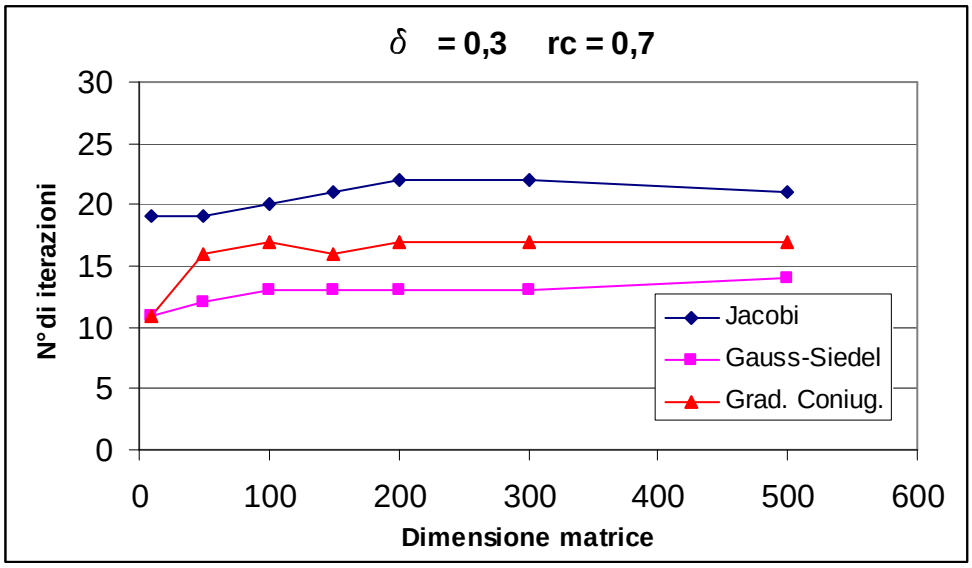

Figura 3

Infine si sono provati i tre metodi ipotizzando una situazione particolare: da un lato un valore molto basso della densità (quindi alta sparsità della matrice) e, contemporaneamente, un basso valore di rc che comporta una diagonalizzazione minore della matrice stessa (figura 4).

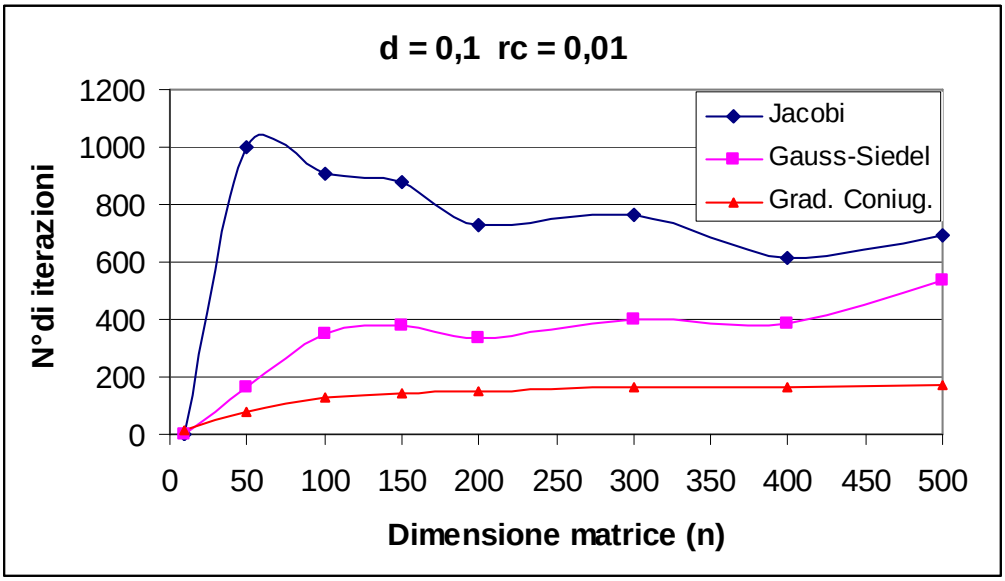

Figura 4

I risultati sono mostrati nella figura 3 sopra, dove si nota come il metodo di Jacobi abbia un picco di iterazioni richieste per la risoluzione del problema per matrici con dimensione intorno a 50, mentre per valori più elevati il numero di iterazioni tende a decrescere lentamente. Nel caso degli altri due, si parte con un numero di iterazioni limitato per matrici piccole, mentre, al crescere della dimensione delle stesse, cresce numero di iterazioni fino a stabilizzarsi con valori più bassi nel caso del metodo del gradiente coniugato.