Tracce di calcolo numerico[1]

Prof. Marco Vianello - Dipartimento di Matematica, Università di Padova
aggiornamento: 19 maggio 2020

5 Elementi di algebra lineare numerica

5.1 Condizionamento di matrici e sistemi

1. partendo dalla definizione di norma matriciale indotta da una norma vettoriale, $\|A\| = \sup_{x\neq 0} \|Ax\|/\|x\|$, si dimostrino le disuguaglianze fondamentali per le norme matriciali indotte: (i) $\|Ax\| \leq \|A\|\,\|x\|$ e (ii) $\|AB\| \leq \|A\|\,\|B\|$ (dove $A, B \in \mathbb{R}^{n\times n}$ e $x \in \mathbb{R}^n$)

2. si può dimostrare che $\|A\|_\infty = \max_{1\leq i\leq n} \{\sum_{j=1}^n |a_{ij}|\}$ (è facile controllare che $\|A\|_\infty \leq \max_{1\leq i\leq n} \{\sum_{j=1}^n |a_{ij}|\}$), mentre $*$ si ha che $\|A\|_2 = \sqrt{\rho(A^tA)}$, dove $\rho(B)$ è il "raggio spettrale" di una matrice B, $\rho(B) = \max_{1\leq i\leq n} \{|\lambda_i|,\ \lambda_i \textit{ autovalore di } B\}$
(dimostrazioni non richieste)

3. stima di condizionamento (risposta della soluzione di un sistema lineare agli errori sul vettore termine noto): dato il sistema quadrato $Ax = b$ con $det(A) \neq 0$, $b \neq 0$ e il sistema perturbato $A(x + \delta x) = b + \delta b$, si mostri che vale la stima
$$\frac{\|\delta x\|}{\|x\|} \leq k(A)\,\frac{\|\delta b\|}{\|b\|}$$
dove $k(A) = \|A\|\,\|A^{-1}\| \geq 1$ è l'*indice di condizionamento* di una matrice invertibile nella norma matriciale indotta dalla norma vettoriale
(traccia: si utilizzi la prima disuguaglianza fondamentale per mostrare che $\|\delta x\| \leq \|A^{-1}\|\,\|\delta b\|$ e $\|x\| \geq \|b\|/\|A\|$, ...; per dimostrare che $k(A) \geq 1$, si osservi che $\|I\| = 1 = \|A\,A^{-1}\| \leq \|A\|\,\|A^{-1}\|$ per la seconda disuguaglianza fondamentale)

4. si ha che $k(A) \geq |\lambda_{max}|/|\lambda_{min}| \geq 1$ con qualsiasi norma matriciale indotta, dove λ_{max} e λ_{min} sono gli autovalori di modulo massimo e minimo della matrice invertibile A; per le matrici simmetriche $k_2(A) = \|A\|_2\,\|A^{-1}\|_2 = |\lambda_{max}|/|\lambda_{min}|$
($*$ traccia (dimostrazione facoltativa): se λ è autovalore di A e $v \neq 0$ autovettore corrispondente, da $Av = \lambda v$, si ricava immediatamente $|\lambda| = \|Av\|/\|v\| \leq \|A\|$; si ricordi poi che gli autovalori dell'inversa sono i reciproci degli autovalori di A, ...; si ricordi che in generale gli autovalori di A^2 sono i quadrati degli autovalori di A, per A simmetrica $A^tA = A^2$, ...)

5. si consideri il seguente esempio:
$$A\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.7 \end{pmatrix},\quad A\begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 1.01 \\ 0.69 \end{pmatrix},\quad A = \begin{pmatrix} 7 & 10 \\ 5 & 7 \end{pmatrix},$$
dove
$$A^{-1} = \begin{pmatrix} -7 & 10 \\ 5 & -7 \end{pmatrix},\quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.1 \end{pmatrix},\quad \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} -0.17 \\ 0.22 \end{pmatrix};$$
si ha che $\|\delta b\|_\infty/\|b\|_\infty = 10^{-2}$ ma $\|\delta x\|_\infty/\|x\|_\infty > 1$; in effetti, $K_\infty(A) = \ldots$

[1] argomenti e quesiti contrassegnati da * sono più impegnativi, se non si è in grado di fare la dimostrazione bisogna comunque sapere (e saper usare) gli enunciati e capire di cosa si sta parlando

6. risposta della soluzione di un sistema lineare agli errori sulla matrice: dato il sistema quadrato $Ax = b$ con $det(A) \neq 0$, $b \neq 0$ e il sistema perturbato $(A + \delta A)(x + \delta x) = b$, con $det(A + \delta A) \neq 0$, si mostri che vale la stima

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq k(A) \, \frac{\|\delta A\|}{\|A\|} \, ,$$

cioé in sostanza il ruolo di $k(A)$ è analogo al caso della risposta agli errori sul vettore termine noto
(traccia: da $(A + \delta A)(x + \delta x) = b$ si ricava $\delta x = -A^{-1}\delta A(x + \delta x)$ e usando la prima disuguglianza fondamentale per una norma indotta ...
$*$ facoltativo: dalla stima scritta sopra si ricava $\|\delta x\| \leq k(A) \, \frac{\|\delta A\|}{\|A\|} \, (\|x\| + \|\delta x\|)$ da cui, se $c(A) = k(A) \, \frac{\|\delta A\|}{\|A\|} < 1$, si ricava immediatamente la stima rigorosa $\frac{\|\delta x\|}{\|x\|} \leq \frac{c(A)}{1-c(A)}$)

7. $*$ *un teorema fondamentale di invertibilità*: se $\|A\| < 1$ in una norma matriciale indotta, allora $I \pm A$ è invertibile e si ha $\|(I \pm A)^{-1}\| \leq \frac{1}{1-\|A\|}$
(traccia della dimostrazione, per matematici: si consideri la serie di potenze matriciali $\sum_{j=0}^{\infty} A^j$; dette $S_n = \sum_{j=0}^{n} A^j$ e $\sigma_n = \sum_{j=0}^{n} \|A\|^j$, si ha che $\|S_m - S_n\| \leq |\sigma_m - \sigma_n|$, quindi la successione $\{S_n\}$ è di Cauchy e di conseguenza convergente a una matrice S, tale che $S(I - A) = \ldots$ e $\|S\| \leq \ldots$; dimostrazione alternativa più semplice: $I - A$ è invertibile, perché i suoi autovalori sono $\mu_i = 1 - \lambda_i$ dove i λ_i sono gli autovalori di A (e quindi $|\lambda_i| \leq \|A\|$), perciò $|\mu_i| = |1 - \lambda_i| \geq 1 - |\lambda_i| > 0$ cioè $I - A$ non ha autovalori nulli; detta $S = (I - A)^{-1}$ da $S(I - A) = I$ si ottiene $1 = \|S - SA\| \geq |\|S\| - \|SA\|| = \|S\| - \|SA\| \geq \|S\| - \|S\| \, \|A\|$ perché $\|SA\| \leq \|S\| \, \|A\| \leq \|S\|$, ...)

8. si può dimostrare che per un sistema in cui anche la matrice sia affetta da errori, $(A + \delta A)(x + \delta x) = b + \delta b$, se $k(A) \, \|\delta A\| < \|A\|$, vale la stima

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

(per $\delta b = 0$ si confronti con la stima rigorosa al punto 6$*$)
($*$ traccia: partendo da $(A+\delta A)\delta x = \delta b - \delta A x$ e osservando che $(A+\delta A) = A(I+B)$ con $B = A^{-1}\delta A$ e $\|B\| \leq \|A^{-1}\| \, \|\delta A\| = k(A) \, \|\delta A\|/\|A\| < 1$, per il teorema fondamentale di invertibilità e le disuguaglianze fondamentali $\delta x = (A+\delta A)^{-1}(\delta b - \delta A x) = (I + B)^{-1}A^{-1}(\delta b - \delta A x)$ e $\|\delta x\| \leq \|(I + B)^{-1}\| \, \|A^{-1}\| \, \|\delta b - \delta A x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|} \, \|\delta b - \delta A x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\| \, \|\delta A\|} \, \|\delta b - \delta A x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\| \, \|\delta A\|} \, (\|\delta b\| + \|\delta A\| \, \|x\|)$, da cui $\frac{\|\delta x\|}{\|x\|} \leq \cdots$)

9. $*$ (cenni di approfondimento facoltativi): la soluzione di sistemi lineari con matrici fortemente mal condizionate (nelle applicazioni non è infrequente avere indici di condizionamento $k(A) \approx 10^{10}, 10^{20}$ o anche superiori) deve essere affrontata con metodi speciali; schematizzando e semplificando, si può dire che molti metodi corrispondono a sostituire il sistema di partenza con una opportuna famiglia parametrizzata di sistemi $A_h x_h = b$ tali che $\varepsilon(h) = \|x - x_h\| \to 0$, $k(A_h) < k(A)$ e $k(A_h) \to k(A)$ per $h \to 0$; dati i sistemi con termine noto perturbato $A_h(x_h + \delta x_h) = b + \delta b$ si ha allora una stima

$$\frac{\|x - (x_h + \delta x_h)\|}{\|x\|} \leq \frac{\|x - x_h\|}{\|x\|} + \frac{\|\delta x_h\|}{\|x\|} \leq \frac{\varepsilon(h)}{\|x\|} + \left(1 + \frac{\varepsilon(h)}{\|x\|}\right) \, k(A_h) \frac{\|\delta b\|}{\|b\|}$$

da cui si vede che conviene minimizzare in h: come nella derivazione numerica con formule alle differenze (si veda la figura alla fine della sezione 4.2), non conviene prendere il parametro h troppo grande (perché si è distanti da x) e neppure troppo piccolo (perché il condizionamento di A_h diventa troppo vicino a quello di A)

5.2 Metodo di eliminazione di Gauss e fattorizzazione LU

1. il metodo di eliminazione di Gauss realizza la sequenza di trasformazioni $A^{(0)} = A \to A^{(1)} \to A^{(2)} \to \ldots \to A^{(i)} \ldots \to A^{(n-1)} = U$ (triangolare superiore), dove $a_{kj}^{(i)} = 0$ per $j+1 \leq k \leq n$, $1 \leq j \leq i$ (si disegnino le matrici $A^{(i)}$ e $A^{(i+1)}$ con la struttura a trapezio di zeri sotto la diagonale), tramite le operazioni vettoriali sulle righe $R_k^{(i+1)} := R_k^{(i)} + \left(-\frac{a_{ki}^{(i)}}{a_{ii}^{(i)}}\right) R_i^{(i)}$, $k = i+1, \ldots, n$ (purché $a_{ii}^{(i)} \neq 0$)

2. cosa succede se al passo i-esimo del metodo di eliminazione $a_{ii}^{(i)} = 0$ e $a_{ki}^{(i)} = 0$ per $i+1 \leq k \leq n$?

3. quali sono gli scopi del pivoting nel metodo di eliminazione di Gauss? (si pensi al caso in cui un elemento diagonale diventa nullo, oppure molto piccolo in modulo; nel secondo caso, si vuole evitare la creazione di numeri approssimati grandi in modulo, che nelle sottrazioni portano a potenziale perdita di precisione)

4. si mostri che il costo computazionale del metodo di eliminazione di Gauss è $\mathcal{O}(n^3)$ flops (floating-point operations), scrivendo lo schema dell'algoritmo
facoltativo: si dimostri che il costo computazionale effettivo è $\sim \frac{2}{3} n^3$ flops

5. si mostri che il costo computazionale della soluzione di un sistema con matrice triangolare superiore o inferiore è $\mathcal{O}(n^2)$ flops, scrivendo lo schema dell'algoritmo di *sostituzione all'indietro* (caso triangolare superiore) e di *sostituzione in avanti* (caso triangolare inferiore)

6. il metodo di eliminazione di Gauss può essere usato per calcolare il determinante di una matrice quadrata (in effetti $det(A) = (-1)^s \, det(U) = (-1)^s \prod_{i=1}^{n} u_{ii}$ dove s è il numero di scambi tra righe); il costo computazionale della formula ricorsiva di Laplace per il determinante è $\sim 2n!$ flops: in tabella gli ordini di grandezza dei tempi di calcolo dei due metodi per diversi valori (relativamente piccoli) di n con un computer da 1Gigaflops (10^9 flops al secondo) e da 1Petaflops (10^{15} flops al secondo)

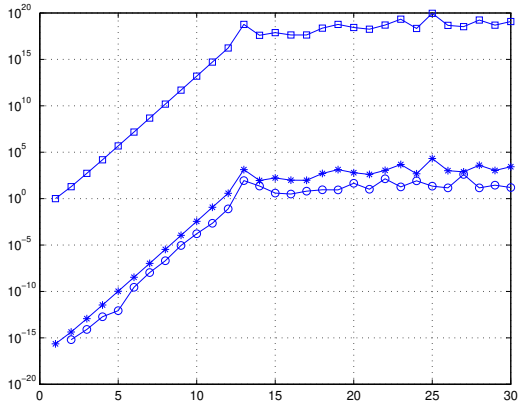
	1Gflops		1Pflops	
n	*Laplace*	*Gauss*	*Laplace*	*Gauss*
10	10^{-3} sec	10^{-6} sec	10^{-9} sec	10^{-12} sec
15	40 min	$3 \cdot 10^{-6}$ sec	2 millisec	$3 \cdot 10^{-12}$ sec
20	10^2 anni	$8 \cdot 10^{-6}$ sec	80 min	$8 \cdot 10^{-12}$ sec
25	10^9 anni	10^{-5} sec	10^3 anni	10^{-11} sec
100	10^{141} anni	10^{-3} sec	10^{135} anni	10^{-9} sec

7. si può dimostrare (dim. non richiesta) che il metodo di eliminazione di Gauss con pivoting per righe produce per matrici non singolari una fattorizzazione $PA = LU$ (dove P è una matrice di permutazione corrispondente agli scambi di righe, L è triangolare

inferiore con 1 sulla diagonale e U è triangolare superiore); la soluzione di un sistema $Ax = b$ con il metodo di eliminazione è equivalente alla fattorizzazione LU seguita dalla soluzione della coppia di sistemi triangolari $Ly = Pb$, $Ux = y$

8. come si può utilizzare la fattorizzazione $PA = LU$ fornita dal metodo di eliminazione con pivoting per righe, per calcolare l'inversa di una matrice? si osservi che il costo computazionale complessivo è $\mathcal{O}(n^3)$ flops
(traccia: le colonne di una matrice si ottengono moltiplicando la matrice per i vettori coordinati $\{e_i\}$, quindi le colonne di A^{-1}, $c_i = A^{-1}e_i$, corrispondono alle soluzioni degli n sistemi $Ac_i = e_i$, ...)

9. perchè la fattorizzazione $PA = LU$ calcolata in aritmetica di macchina, diciamola $\tilde{L}\tilde{U} \approx PA$) è estremamente accurata (cioè, gli elementi di $\tilde{L}\tilde{U}$ coincidono alla precisione di macchina con quelli di PA), ma per certe matrici (vedi matrice di Hilbert $H = (h_{ij})$, $h_{ij} = 1/(i + j - 1)$, $1 \leq i, j \leq n$) la soluzione di sistemi ottenuta con tale fattorizzazione risulta molto poco accurata (se non addirittura completamente sbagliata)?
(traccia: in sostanza è come se si risolvesse il sistema perturbato $\tilde{L}\tilde{U}(x + \delta x) = Pb$, ...; in figura l'errore relativo in $\|\cdot\|_2$ nella soluzione in aritmetica a precisione doppia di $Hx = Hu$, $u = (1, \ldots, 1)^t$ (pallini), le quantità $K_2(H)$ (quadratini) e $K_2(H) \cdot eps$ (asterischi), per $n = 2, \ldots, 30$)



5.3 Sistemi sovradeterminati, minimi quadrati e fattorizzazione QR

1. sia $A \in \mathbb{R}^{m\times n}$, $m > n$, una matrice di rango massimo n (ovvero, le colonne di A sono n vettori linearmente indipendenti di $\mathbb{R}^m$): si mostri che calcolare la soluzione ai minimi quadrati del sistema sovradeterminato $Ax = b$ (che in generale non ha soluzione classica, perché?)

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 = \min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij}x_j \right)^2$$

è equivalente a risolvere il sistema con matrice simmetrica non singolare (e definita positiva) $A^tAx = A^tb$ (detto "sistema delle equazioni normali")

(∗ traccia: si osservi che x minimizza il polinomio quadratico $\phi(x) = \|Ax - b\|_2^2$ se e solo se per ogni $v \in \mathbb{R}^n$ si ha $\phi(x) \leq \phi(x + v)$ ovvero sviluppando i calcoli $2v^t A^t(Ax - b) + v^t A^t Av \geq 0$, e posto $v = \varepsilon u$ con u fissato, per $\varepsilon \to 0$ si ottiene $(A^t(Ax - b), u) \geq 0 \ \forall u \in \mathbb{R}^n$, ma allora prendendo $-u$ al posto di u si ottiene $(A^t(Ax - b), u) \leq 0 \ \forall u$, da cui ...; per dimostrare che $A^t A$ è non singolare, si osservi che se $A^t Av = 0$ allora $(A^t Av, v) = (Av, Av) = 0$ da cui $Av = 0$, ma le colonne di A sono linearmente indipendenti quindi $v = 0$)

2. l'approssimazione polinomiale di grado k ai minimi quadrati può essere reinterpretata come soluzione ai minimi quadrati del sistema $m \times n$ (sovradeterminato) $Va = y$ (si veda la sezione 3.3), dove $m = N$ e $n = k + 1$
   (traccia: per provare che V ha rango massimo, si ricordi che una matrice quadrata di Vandermonde corrispondente a nodi distinti è non singolare, quindi considerando una sottomatrice quadrata della matrice di Vandermonde rettangolare ...)

3. sia $A \in \mathbb{R}^{m \times n}$, $m \geq n$, una matrice di rango massimo: si può dimostrare (dim. non richiesta) che A ammette una fattorizzazione $A = QR$, dove $Q \in \mathbb{R}^{m \times n}$ è una matrice ortogonale (ovvero le colonne di Q sono vettori ortonormali, cioè $Q^t Q = I \in \mathbb{R}^{n \times n}$) e R è una matrice triangolare superiore non singolare (si tratta in sostanza di un procedimento equivalente all'ortogonalizzazione di Gram-Schmidt delle colonne di A: la matrice R^{-1}, che resta triangolare superiore, è la matrice dei coefficienti dell'ortogonalizzazione: si controlli che il prodotto $A\,R^{-1}$ corrisponde a costruire combinazioni lineari ortonormalizzanti delle colonne di A utilizzando i coefficienti delle colonne di R^{-1})

4. come si può applicare la fattorizzazione QR alla soluzione di sistemi sovradeterminati? (traccia: utilizzando $A = QR$, non è necessario calcolare la matrice $A^t A$ perché $A^t A = R^t R$, quindi il sistema delle equazioni normali diventa $R^t Rx = R^t Q^t b$ che è equivalente (perché?) al sistema triangolare $Rx = Q^t b$)

5. dal punto di vista del condizionamento è meglio risolvere $Rx = Q^t b$ invece di $A^t Ax = b$, perché l'indice di condizionamento di R in norma 2 è molto piú piccolo dell'indice di condizionamento di $A^t A$ (dimostrazione non richiesta; si può intuire osservando che se A è quadrata e simmetrica, $k_2(A^t A) = k_2(A^2) = (k_2(A))^2$ e $k_2(R) \leq k_2(A)$ visto che $R = Q^t A$ e $\|Q\|_2 = \|Q^t\|_2 = 1$; in realtà si può dimostrare (non richiesto) che $k_2(R) = k_2(A)$)

5.4 Metodi iterativi

1. cond. suff. per la convergenza delle approssimazioni successive: un sistema quadrato della forma $x = Bx + c$ con $\|B\| < 1$ (in una norma matriciale indotta da una norma vettoriale) ha soluzione unica che si può ottenere come limite della successione di approssimazioni successive vettoriali $\{x_k\}$ definita da

$$x_{k+1} = Bx_k + c\ ,\ \ k = 0, 1, 2, \ldots$$

a partire da un qualsiasi vettore iniziale x_0
(traccia: il sistema ha soluzione unica se e solo se $I - B$ è invertibile, ma $\|B\| < 1$ e per il teorema fondamentale di invertibilità ...; scrivendo $x - x_{k+1} = B(x - x_k)$ si ottiene la stima $\|x - x_{k+1}\| \leq \|B\|\,\|x - x_k\|$, ...)

2. $*$ cond. nec. e suff. per la convergenza delle approssimazioni successive: il metodo delle approssimazioni successive $x_{k+1} = Bx_k + c$, $k \geq 0$, converge alla soluzione di $x = Bx + c$ per qualsiasi scelta dei vettori x_0 e c se e solo se $\rho(B) < 1$ (dove $\rho(B)$ è il "raggio spettrale" della matrice quadrata B, ovvero il max dei moduli degli autovalori)
(traccia della dim. (facoltativa) nel caso B sia diagonalizzabile: scrivendo x_0, c in una base $\{v_i\}$ di autovettori di B, $Bv_i = \lambda_i v_i$, $1 \leq i \leq n$, $x_0 = \sum \alpha_i v_i$, $c = \sum \gamma_i v_i$, si ottiene $x_k = \sum \left(\alpha_i \lambda_i^k + \gamma_i(1 + \lambda_i + \ldots + \lambda_i^{k-1})\right) v_i$, che ha limite $\forall\, x_0, c$ se e solo se $|\lambda_i| < 1$, $1 \leq i \leq n$, con vettore limite $\lim_{k\to\infty} x_k = \sum \frac{\gamma_i}{1-\lambda_i} v_i = (I - B)^{-1} c$)

3. dato uno splitting di una matrice quadrata, $A = P - N$, con $det(P) \neq 0$, il sistema $Ax = b$ si può scrivere nella forma $x = Bx + c$ dove $B = P^{-1}N$ e $c = P^{-1}b$. Esempi di corrispondenti metodi delle approssimazioni successive nell'ipotesi $a_{ii} \neq 0\, \forall i$ sono (posto $A = D - (E + F)$, dove D è la parte diagonale di A ed $-E, -F$ le parti triangolare inferiore e superiore di $A - D$)

   - il metodo di Jacobi: $P = D$, $N = E + F$
   - il metodo di Gauss-Seidel: $P = D - E$, $N = F$

   Si scrivano per componenti tali metodi, e si dimostri che il metodo di Jacobi è convergente per matrici diagonalmente dominanti in senso stretto (cioè $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$, $1 \leq i \leq n$); si può dimostrare (dim. non richiesta) che anche il metodo di Gauss-Seidel converge in tale ipotesi nonché per matrici simmetriche definite positive (traccia: si consideri la norma infinito della matrice di iterazione B del metodo di Jacobi, ...)

4. $*$ il metodo delle approssimazioni successive si può riscrivere come

$$x_{k+1} = (I - P^{-1}A)x_k + P^{-1}b = x_k + P^{-1}r(x_k)$$

(dove $r(x_k) = b - Ax_k$ è il vettore "residuo" al passo k-esimo); il ruolo della matrice P^{-1} può essere visto come quello di "precondizionatore": l'azione di P^{-1} è efficace quando $P^{-1} \approx A^{-1}$, nel senso che gli autovalori di $P^{-1}A$ si accumulano intorno ad 1 e quindi la convergenza diventa più rapida (e nel contempo dato un vettore v, il calcolo di $z = P^{-1}v$, ovvero la soluzione del sistema $Pz = v$, ha basso costo computazionale); vari metodi introducono un parametro di rilassamento α, $x_{k+1} = x_k + \alpha P^{-1} r(x_k)$, che aumenti l'efficacia del precondizionatore (cercando di diminuire o addirittura minimizzare il raggio spettrale di $B(\alpha) = I - \alpha P^{-1}A$)

5. $*$ test di arresto del residuo: dato un qualsiasi metodo iterativo *convergente* per la soluzione di un sistema lineare non singolare $Ax = b$ con $b \neq 0$, si mostri che vale la seguente stima dell'errore relativo (la norma del residuo va pesata da $k(A)/\|b\|$)

$$\frac{\|x - x_k\|}{\|x\|} \leq \frac{k(A)}{\|b\|} \|r(x_k)\|$$

6. una parte spesso preponderante del costo computazionale di un metodo iterativo (si veda la formulazione che coinvolge il vettore residuo) è costituita dal calcolo dei prodotti Ax_k; si osservi che nel caso A sia una matrice "sparsa", cioè con moltissimi zeri e solo una piccola frazione di elementi non nulli, il prodotto matrice-vettore ha un costo che non è più quadratico ma dell'ordine di mn flops, dove m è il numero medio di elementi non nulli per riga (implementando il prodotto in modo da utilizzare solo gli elementi non nulli). Matrici sparse di grande dimensione vengono ad esempio prodotte dai metodi di discretizzazione di equazioni differenziali alle derivate parziali in dimensione 2 e 3, si veda il punto 6.9 del prossimo capitolo.

   In questi casi, detto $\nu(\varepsilon)$ il numero di iterazioni richiesto per un errore non superiore ad ε, il costo del metodo sarà dominato da $mn\nu(\varepsilon)$, da confrontare col costo proporzionale ad n^3 del metodo di eliminazione (con grande vantaggio computazionale del metodo iterativo se $m\nu(\varepsilon) \ll n^2$)