Algoritmo di Minimizzazione della Traccia per il Problema agli Autovalori Generalizzato

Presentazione e Fondamenti Teorici

Giuseppe Lombardi

Calcolo Scientifico, 2018

Tabella dei Contenuti

Introduzione

In questo lavoro verrà presentato un metodo iterativo che ci permetterà di ottenere buone approssimazioni su alcuni dei più piccoli autovalori e corrispettivi autovettori del problema generalizzato agli autovalori:

$$Ax = \lambda Bx \tag{1}$$

Dove:

- x vettore di lunghezza n
- λ uno scalare
- A e B matrici simmetriche $n \times n$, con B definita positiva
- solitamente A e B sono molto grandi e sparse

Teorema (1 di ortogonalizzazione simultanea)

A e B matrici simmetriche $n \times n$. Se B è definita positiva allora esiste una matrice Z $n \times n$ tale che:

$$Z^t B Z = I_n, \quad Z^t A Z = \Lambda = diag(\lambda_1, ..., \lambda_n)$$

dove $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ sono gli autovalori di $Ax = \lambda Bx$ e le colonne di Z sono i loro autovettori associati.

Dim.

B simmetrica e definita positiva, per il teorema spettrale $\exists Q$ ortogonale tale che $Q^t B Q = D$ matrice diagonale con elementi sulla diagonale tutti > 0.
Moltiplicando A per opportune matrici si ottiene che $D^{-\frac{1}{2}} Q^t A Q D^{-\frac{1}{2}}$ è simmetrica. Quindi per il teorema spettrale $\exists U$ ortogonale tale che $U^t D^{-\frac{1}{2}} Q^t A Q D^{-\frac{1}{2}} U = \Lambda$ con Λ diagonale.
Infine si ha che $U^t D^{-\frac{1}{2}} Q^t B Q D^{-\frac{1}{2}} U = I$ e quindi Z risulta essere semplicemente $Q D^{-\frac{1}{2}} U$.
È facile verificare che gli elementi sulla diagonale di Λ sono gli autovalori del problema e che le colonne di Z sono i rispettivi autovettori. □

Teorema (2)

Siano date A e B matrici simmetriche $n \times n$, B definita positiva e sia Y^* l' insieme di tutte le matrici Y $n \times p$ per cui $Y^t B Y = I_p$, allora:

$$\min_{Y \in Y^*} tr(Y^t A Y) = \sum_{i=1}^{p} \lambda_i$$

In altre parole:

$$\min_{Y \in Y^*} tr(Y^t A Y) = tr(X^t A X)$$

con X tale che $X^t B X = I_p$ e $X^t A X = diag(\lambda_1, ..., \lambda_p)$ che corrisponde alle prime p colonne di Z (matrice degli autovettori).

Dim(Cenni).

La prova del teorema segue direttamente dal teorema di Courant-Fisher e dalle osservazioni sul quoziente di Rayleigh. □

Definizione

Per una data $A \in \mathbb{R}^{n \times n}$ simmetrica e un vettore $x \in \mathbb{R}^n$ non nullo, il quoziente di Rayleigh è il numero reale:

$$R_A(x) = \frac{x^t A x}{x^t x}.$$

Teorema (3 di Courant-Fisher o del min-max)

Sia $A \in \mathbb{R}^{n\times n}$ simmetrica con autovalori $\lambda_1 \leq ... \leq \lambda_n$, allora

$$\lambda_k = \min_{V_k} \max_{x\neq 0, x\in V_k} r_A(x)$$

$$\lambda_{n-k+1} = \max_{V_k} \min_{x\neq 0, x\in V_k} r_A(x)$$

dove V_k é un sottospazio di $\mathbb{R}^n$ di dimensione k.

Osservazione

$\lambda_1 = \min_{x\neq 0} r_A(x), \qquad \lambda_n = \max_{x\neq 0} r_A(x)$

Osservazione

x_i autovettore relativo all' autovalore λ_i della matrice simmetrica A allora $r_A(x_i) = \lambda_i$. Inoltre i punti stazionari di $r_A(x)$ sono esattamente gli autovettori di A.

Teorema (4 di Ostrowski)

Sia A una matrica simmetrica $n \times n$ con autovalori $\lambda_1 \leq ... \leq \lambda_n$, di cui π_A positivi e ν_A negativi. Sia $M = Y^t A Y$ con Y matrice $n \times p$ e $p < n$ e autovalori $\mu_1 \leq ... \leq \mu_p$ di cui π_M positivi e ν_M negativi. Allora $\pi_M < \pi_A$ e $\nu_M < \nu_A$. Inoltre denotati gli autovalori positivi di A e di M rispettivamente con

$$0 < \lambda_1^+ \leq ... \leq \lambda_{\pi_A}^+, \qquad 0 < \mu_1^+ \leq ... \leq \mu_{\pi_M}^+$$

e quelli negativi con

$$\lambda_1^- \leq ... \leq \lambda_{\nu_A}^- < 0, \qquad \mu_1^- \leq ... \leq \mu_{\nu_M}^- < 0$$

allora valgono le seguenti:

$$\frac{\mu_{\pi_M - j}^+}{\lambda_{\pi_A - j}^+} \leq \rho \quad \textit{per} \quad j = 0, ..., \pi_M - 1$$

$$\frac{\mu_{\nu_M - j}^-}{\lambda_{\nu_A - j}^-} \leq \rho \quad \textit{per} \quad j = 0, ..., \nu_M - 1,$$

dove ρ è il massimo autovalore della matrice simmetrica definita positiva $Y^t Y$.

Teorema (5)

Sia $K = S^t H S$ una matrice $p \times p$ con S $n \times p$ di rango p e H $n \times n$ definita positiva. Allora $tr(K) \leq \rho tr(H)$ dove ρ è il massimo autovalore di $S^t S$.

Dim.

Si prova facilmente che K è definita positiva. Scriviamo rispettivamente gli autovalori di H e di K come

$$0 < \lambda_1 \leq ... \leq \lambda_n, \qquad 0 < \mu_1 \leq ... \leq \mu_p$$

rispettivamente. Dal teorema di Ostrowski segue che

$$\mu_{p-j} \leq \rho \lambda_{n-j} \quad \text{per } j = 1, ..., p-1$$

dove ρ è il più grande autovalore di $S^t S$, e il risultato segue facilmente. □

Minimizzazione della traccia

Per il momento richiediamo l' ipotesi aggiuntiva che anche la matrice A sia definita positiva.

Definizione

Una matrice Y $n \times p$ forma una sezione del problema agli autovalori $Ax = \lambda Bx$ se valgono:

$$Y^t AY = \Sigma \qquad e \qquad Y^t BY = I_p$$

con $\Sigma = diag(\sigma_1, ..., \sigma_p)$.

Il metodo di minimizzazione della traccia fu proposto da Sameh e Wisniewski nel 1982 come variante del *metodo di iterazione simultanea*, nel tentativo di evitare le difficoltà dovute alla risoluzione esatta di certi sistemi lineari, le quali compromettevano la convergenza globale del metodo.

Questo metodo calcola i p piú piccoli autovalori e i corrispettivi autovettori sfruttando la proprietà di riduzione della traccia. Il nostro approccio è quello di trovare una sequenza di iterazioni $Y_{k+1} = F(Y_k)$ tale che:

- Y_k e Y_{k+1} sono una *sezione* del problema
- $tr(Y_{k+1}^t A Y_{k+1}) < tr(Y_k^t A Y_k)$
- $F(Y_k)$ è scelto in modo tale che la convergenza globale del processo sia assicurata.

Successivamente tratteremo il problema come un problema di minimizzazione quadratica

$$\text{Minimizza } tr(Y^t A Y) \quad \text{soggetto ai vincoli } Y^t B Y = I_p.$$

Algoritmo di base

Data Y_k una matrice $n \times p$ che approssima i p autovettori corrispondenti ai p più piccoli autovalori del problema cioè tale che:

$$Y_k^t A Y_k = \Sigma_k = diag(\sigma_1^{(k)}, ..., \sigma_p^{(k)})$$

$$Y_k^t B Y_k = I_p$$

vogliamo costruire una matrice Y_{k+1} della forma

$$Y_{k+1} = (Y_k + \Delta_k) S_k$$

dove S_k e Δ_k sono scelte in modo che:

- $Y_{k+1}^t A Y_{k+1} = \Sigma_{k+1} = diag(\sigma_1^{(k+1)}, ..., \sigma_p^{(k+1)})$
- $Y_{k+1}^t B Y_{k+1} = I_p$
- $tr(Y_{k+1}^t A Y_{k+1}) < tr(Y_k^t A Y_k)$

Scelta di S_k e Δ_k

Consideriamo le correzioni di Δ_k nel sottospazio tangente all' insieme dei vincoli $Y_k B Y_k = I_p$, ciò implica

$$\Delta_k^t B Y_k = 0.$$

Poniamo $\hat{Y} = Y_k + \Delta_k$, allora la scelta di S_k è determinata dalle due matrici $p \times p$:

- $\hat{Y}^t A \hat{Y} = \Sigma_k + Y_k^t A \Delta_k + \Delta_k^t A Y_k + \Delta_k^t A \Delta_k$
- $\hat{Y}^t B \hat{Y} = I_p + \Delta_k^t B \Delta_k$

La seconda matrice è definita positiva con autovalori tutti > 1, quindi dalla sua decomposizione spettrale si ha:

$$\hat{Y}^t B \hat{Y} = U D^2 U^t$$

dove U é ortogonale e $D^2 = diag(\delta_1^2, ..., \delta_p^2)$, con $\delta_i^2 > 1$.
Quindi $D^{-1} U^t (\hat{Y}^t B \hat{Y}) U D^{-1} = I_p$.

Dalla decomposizione spettrale di $D^{-1}U^t(\hat{Y}^t A\hat{Y})UD^{-1}$ otteniamo:

$$V^t D^{-1}U^t(\hat{Y}^t A\hat{Y})UD^{-1}V = \Sigma_{k+1}$$

dove V é ortogonale. Scegliamo quindi $S_k = UD^{-1}V$.
$Y_{k+1} = (Y_k + \Delta_k)S_k$, con Δ_k e S_k cosí scelti, è una *sezione* del problema.

Vale, inoltre, che:

$$\begin{aligned} tr(Y_{k+1}^t A Y_{k+1}) &= tr(V^t(D^{-1}U^t\hat{Y}^t A\hat{Y}UD^{-1})V) \\ &= tr(D^{-1}U^t\hat{Y}^t A\hat{Y}UD^{-1}) = \sum_{i=1}^{p} \frac{g_{ii}}{\delta_i^2} \\ &< \sum_{i=1}^{p} g_{ii} = tr(U^t(\hat{Y}^t A\hat{Y})U) = tr(\hat{Y}^t A\hat{Y}), \end{aligned}$$

dove g_{ii} è l' elemento diagonale di $G = U^t\hat{Y}^t A\hat{Y}U$.
La matrice Δ_k gioca un ruolo cruciale. È scelta per migliorare le approssimazioni degli autovettori riducendo $tr(\hat{Y}^t A\hat{Y})$.

Δ_k tramite iterata sottospazio ottimale

Trattiamo il problema come uno di minimizzazione quadratica vincolata:

$$\text{Minimizza } tr(Y^tAY) \quad \text{soggetto ai vincoli } Y^tBY = I_p.$$

Consideriamo un' iterazione del tipo $Y_{k+1} = (Y_k - \Delta_k)S_k$, S_k è scelta come nella sezione precedente e come prima considero Δ_k nel sottospazio ortogonale ai vincoli. Poniamo $\bar{Y} = Y - \Delta$, dove Δ è scelto in modo che

$$\text{minimizzi } tr((Y-\Delta)^tA(Y-\Delta)) \quad \text{soggetto ai vincoli } Y^tB\Delta = 0.$$

Abbiamo che:

$$\begin{aligned} tr((Y-\Delta)^tA(Y-\Delta)) &= \sum_{j=1}^{p} e_j^t(Y-\Delta)^tA(Y-\Delta)e_j \\ &= \sum_{j=1}^{p}(y_j - d_j)^tA(y_j-d_j), \end{aligned}$$

dove $d_j = \Delta e_j$ e $y_j = Ye_j$.

Consideriamo quindi i problemi equivalenti:

$$\text{minimizza } (y_j - d_j)^t A(y_j - d_j) \qquad \text{soggetto ai vincoli } Y^t B d_j = 0 \tag{2}$$

$$\text{per } j = 1, 2, ..., p \quad \text{dove } d_j = \Delta e_j \text{ e } y_j = Y e_j.$$

Sviluppiamo un metodo efficiente per ottenere i d_j:
Utilizzando i moltiplicatori di Lagrange, ognuno dei p sistemi è equivalente al sistema di equazioni lineari:

$$\begin{bmatrix} A & BY \\ Y^t & 0 \end{bmatrix} \begin{bmatrix} d \\ l \end{bmatrix} = \begin{bmatrix} Ay \\ 0 \end{bmatrix} \tag{3}$$

dove l è il vettore di ordine p che rappresenta i moltiplicatori di Lagrange.
Sfruttando la fattorizzazrione QR di BY $(rk(BY) = p)$ si ha che
$BY = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} R$ dove:

- $R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$ con R_1 triangolare superiore di ordine $p \times p$,
- Q ortogonale con Q_1 di ordine $n \times p$ (di conseguenza Q_2 è $n \times n - p$).

Otteniamo, dunque, i sistemi equivalenti:

$$\begin{bmatrix} Q^tAQ & R \\ R^t & 0 \end{bmatrix} \begin{bmatrix} g \\ l \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

dove $g = Q^t d$ e $f = Q^t Ay$.
Scrivendo $g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ dove g_1 è un vettore di ordine p, da $R^t g = 0$ e dalla non singolarietà di R_1, otteniamo $g_1 = 0$.
Quindi il sistema diventa:

$$Q^tAQ \begin{bmatrix} 0 \\ g_2 \end{bmatrix} + \begin{bmatrix} R_1 l \\ 0 \end{bmatrix} = f.$$

Quindi otteniamo:

$$Q_1^tAQ_2g_2 + R_1l = Q_1^tAy \tag{4a}$$
$$Q_2^tAQ_2g_2 = Q_2^tAy \tag{4b}$$

Siccome vogliamo risolvere il sistema 3 solo per d, basta risolvere il sistema 4b per g_2. Così da ricavare d:

$$d = Qg = Q_2g_2.$$

Il sistema 4b è risolto tramite il metodo del Gradiente Coniugato (CG). Vedi appendice...

I p problemi 2 possono essere scritti come:

$$\text{Minimizza } \|b_j - A^{\frac{1}{2}} y_j\| \qquad \text{soggetto ai vincoli } Y^t B A^{-\frac{1}{2}} b_j = 0$$

$$\text{per } j = 1, ..., p \text{ , dove } b_j = A^{\frac{1}{2}} d_j .$$

La soluzione di ognuno di questi problemi lineari ai minimi quadrati è ottenuta imponendo che $A^{\frac{1}{2}} y_j - b_j$ sia uguale alla proiezione ortogonale di $A^{\frac{1}{2}} y_j$ sullo spazio generato da $A^{-\frac{1}{2}} BY$. Quindi:

$$A^{\frac{1}{2}} y_j - b_j = A^{-\frac{1}{2}} BY (Y^t B A^{-1} BY)^{-1} Y^t B y_j$$

equivalentemente:

$$y_j - d_j = A^{-1} BY (Y^t B A^{-1} BY)^{-1} Y^t B y_j$$

Quindi, poichè vale $Y^t BY = I$ ottengo che:

$$\bar{Y} = Y - \Delta = A^{-1} BY (Y^t B A^{-1} BY)^{-1}. \tag{5}$$

Da 5 si osserva che la nostra iterazione (*iterata sottospazio ottimale*) differisce dall' iterazione dell' algoritmo di Rutishauser's solo per la presenza della matrice $(Y^t BA^{-1}BY)^{-1}$.
Dunque la convergenza globale e il tasso di convergenza derivano direttamente da questo.
Le colonne di Y e $\bar{Y}$ possono essere espresse come combinazioni lineari degli autovettori del problema 1.
Quindi:

$$Y = ZG \qquad \text{e} \qquad \bar{Y} = Z\bar{G}$$

con Z matrice di autovettori.
Abbiamo:

$$ZG = Y \longrightarrow X = Z \begin{bmatrix} I_p \\ 0 \end{bmatrix} \qquad \Longrightarrow \qquad G \longrightarrow \begin{bmatrix} I_p \\ 0 \end{bmatrix}$$

Possiamo scrivere quindi:

$$G = \begin{bmatrix} I_p \\ 0 \end{bmatrix} + F = \begin{bmatrix} I_p + F_1 \\ F_2 \end{bmatrix} \qquad \text{e} \qquad \bar{G} = \begin{bmatrix} I_p \\ 0 \end{bmatrix} + \bar{F} = \begin{bmatrix} I_p + \bar{F}_1 \\ \bar{F}_2 \end{bmatrix}$$

dove le colonne di F e $\bar{F}$ sono i vettori che rappresentano gli errori nelle approssimazioni degli autovettori (colonne di Y e $\bar{Y}$).

Teorema (6)

Siano date A e B matrici simmetriche definite positive e assumiamo che gli autovalori del problema $Ax = \lambda Bx$ siano tali che $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p < \lambda_{p+1} \leq ... \leq \lambda_n$. Inoltre sia Y l' iterazione iniziale dell' algoritmo scelta in modo da avere le colonne linearmente indipendenti e tale che la corrispondente matrice F non sia deficiente in qualche componente. Allora, la colonna j di Y, y_j, converge globalmente all' autovettore x_j corrispondente a λ_j per $j = 1, 2, ..., p$ con un tasso di convergenza asintotica minore o uguale a $|\frac{\lambda_j}{\lambda_{p+1}}|$. Quindi

$$\|\bar{F}e_j\| < |\frac{\lambda_j}{\lambda_{p+1}}|\|Fe_j\| + O(\|F\|^2).$$

Aspetti computazionali

L' algoritmo di minimizzazione della traccia, dove Δ_k è scelto tramite iterata sottospazio minimale, può essere migliorato tramite una tecnica di shifting.
Il metodo di minimizzazione consiste nel risolvere

$$(A - \sigma_j^{(k)} B)x_j = \bar{\lambda}_j^{(k)} B x_j \qquad \text{per } j = 1, ..., p.$$

Questi problemi:

- hanno gli stessi autovettori del problema $Ax = \lambda Bx$
- hanno autovalori $\bar{\lambda}_j^{(k)} = \lambda_j - \sigma_j^{(k)}$ dove λ_j è il j-esimo autovalore del problema e $\sigma_j^{(k)}$ è il parametro di spostamento, approssimazione dell' autovalore λ_j al passo k.

La convergenza globale del nostro algoritmo non è preservata.
Rilassiamo, quindi, l' ipotesi che anche la matrice A sia definita positiva.
La strategia di spostamento è motivata dal teorema 7 e dal fatto che $\sigma_j^{(k+1)} < \sigma_j^{(k)}$ per ogni k cioè le approssimazioni si avvicinano agli autovalori del problema dall' alto.

Teorema (7)

Per ogni arbitrario vettore non nullo u e scalare σ, c'è un autovalore λ del problema tale che:

$$|\lambda - \sigma| \leq \frac{\|(A - \sigma B)u\|_{B^{-1}}}{\|Bu\|_{B^{-1}}}.$$

Tale teorema rappresenta una generalizzazione (per il problema generalizzato agli autovalori) del teorema 10 in appendice.

Questa strategia è delineata dai seguenti punti:

1. $\sigma_j^{(k)} = \sigma_1^{(k)}$ per $1 \leq j \leq p$ quando $\sigma_1^{(k)} < 0$
2. il processo del CG è terminato quando non incontriamo un passo di discesa

3. σ_l è usato come spostamento per la colonna j, $l < j$, solo se $\sigma_l < \lambda_j$

L' algoritmo è molto accelerato quando possiamo usare σ_j come spostamento della colonna j.
Il metodo risultante ha un tasso di convergenza cubico ($\|\bar{F}\| = O(\|F^3\|)$) e tale risultato è una conseguenza del seguente teorema:

Teorema (8)

Per $j \leq p$,

$$\lambda_j - \sigma_j^{(k)} = -e_j^t F^t (\Lambda - \lambda_j I) F e_j.$$

Dim.

Vedi appendice 11. □

Osserviamo che, ai fini dell' euristica:

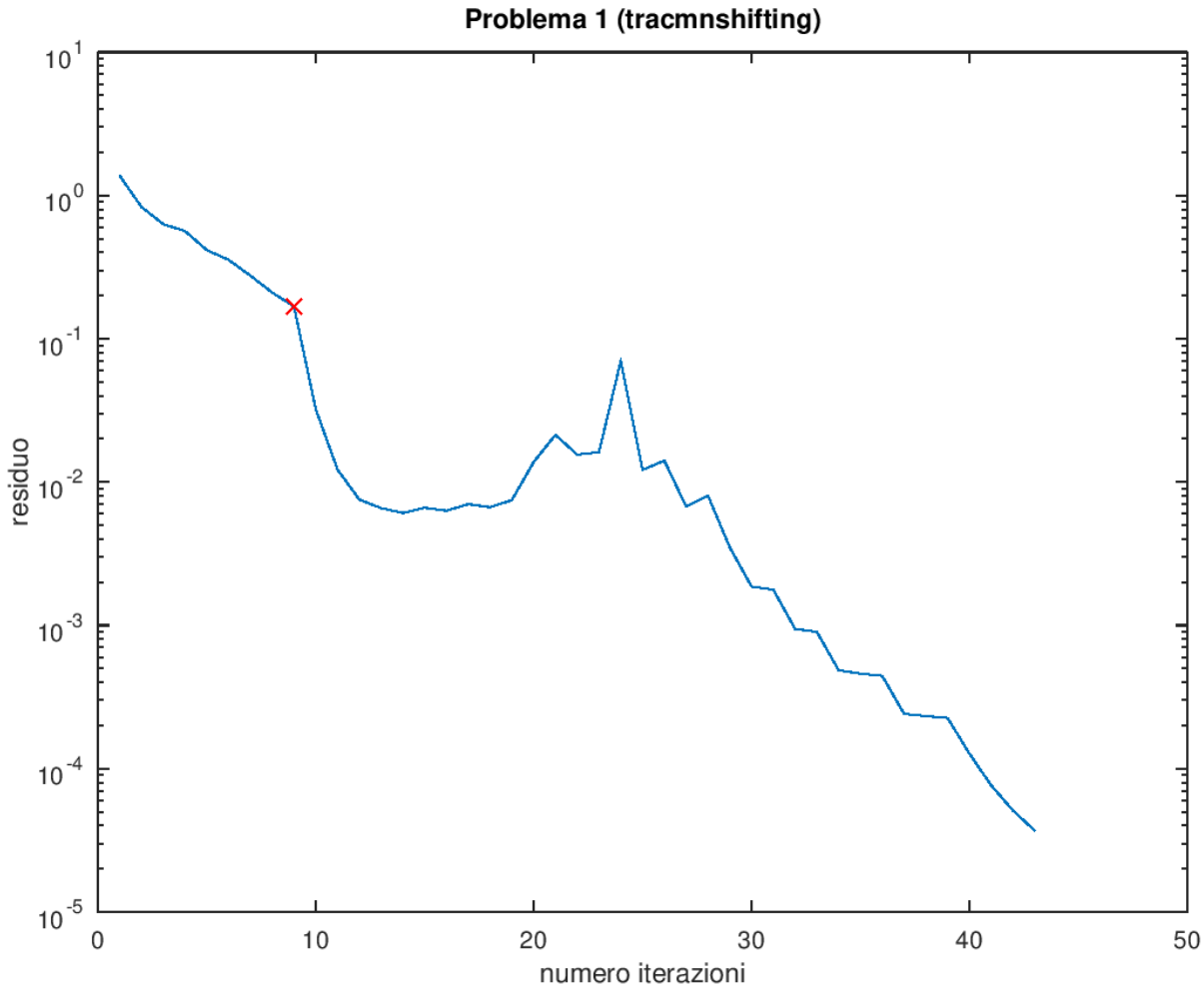
- l' efficenza dell' algoritmo dipende in modo cruciale dalla strategia dello *shifting* impiegata; se spostiamo la colonna j attraverso σ_j troppo tardi l' algoritmo diventa inefficiente nel senso che prendiamo alcuni passi a un tasso di convergenza più lento (lineare) quando invece è possibile un tasso di convergenza cubico. Dall' altra parte se spostiamo troppo presto, la funzione obiettivo con la A aumenta invece di diminuire e la convergenza globale è persa. Dunque il passo 2 della strategia cerca di prevenire tale perdita di convergenza globale mentre il passo 3 evita inutili ritardi di spostamento;
- questa strategia risulta essere ben bilanciata cioè evita eccessive iterazioni del CG e di calcolare troppe sezioni;
- il metodo blocca una colonna di Y quando raggiunge la convergenza;
- la convergenza è ottenuta quando le coppie (autovalore, autovettore) hanno un residuo relativo piccolo e in tal caso terminiamo il processo di iterazione;
- una colonna di Y è accettata come valida approssimazione di un autovettore se i d_j calcolati con il CG sono al livello della precisione di macchina (eps).

Alcune sperimentazioni: Problemi, con matrici sparse, sperimentati al calcolatore

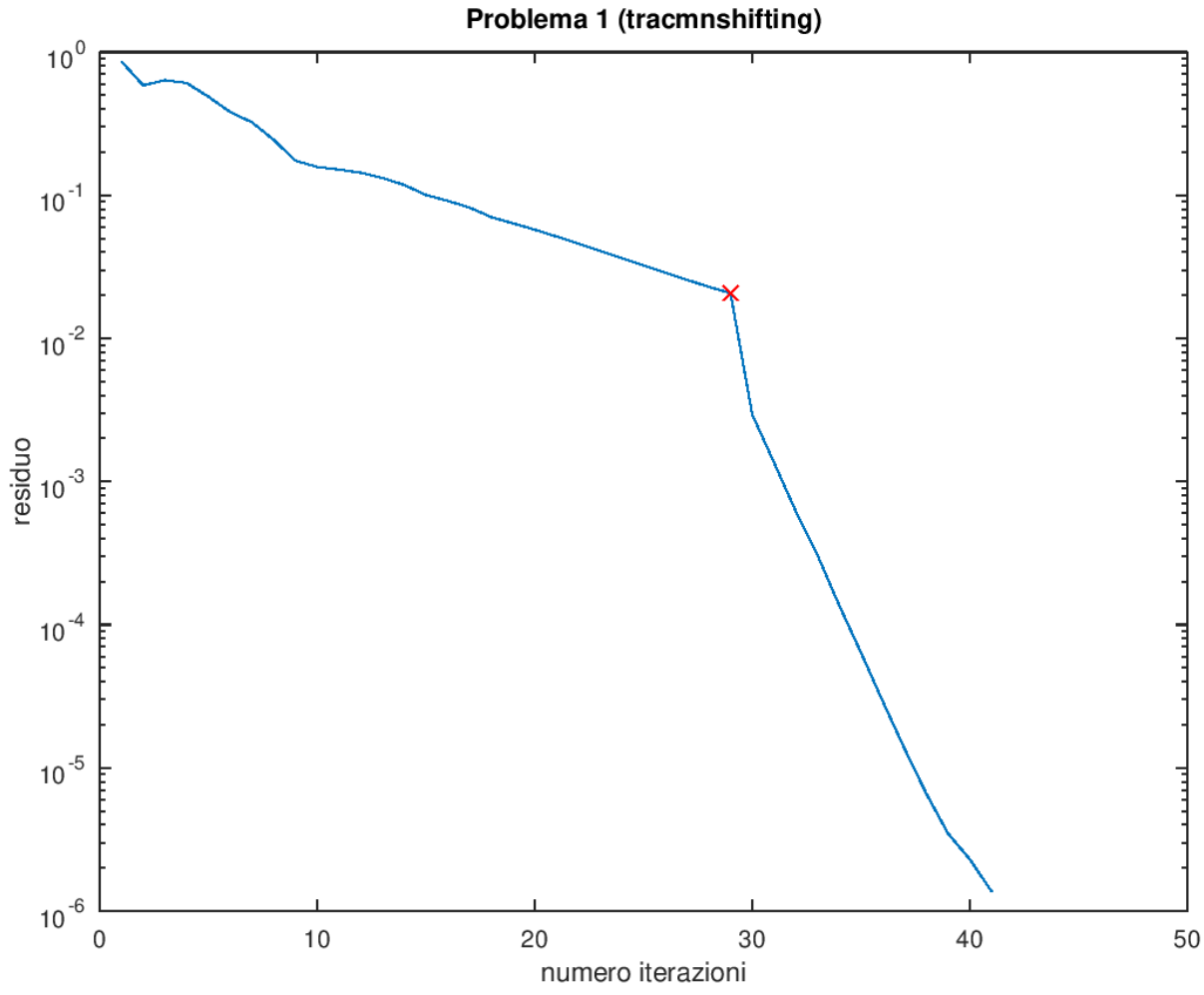
Problema	n	p	d_A	d_B
1	50	5	0.20560	0.20640

Risultati

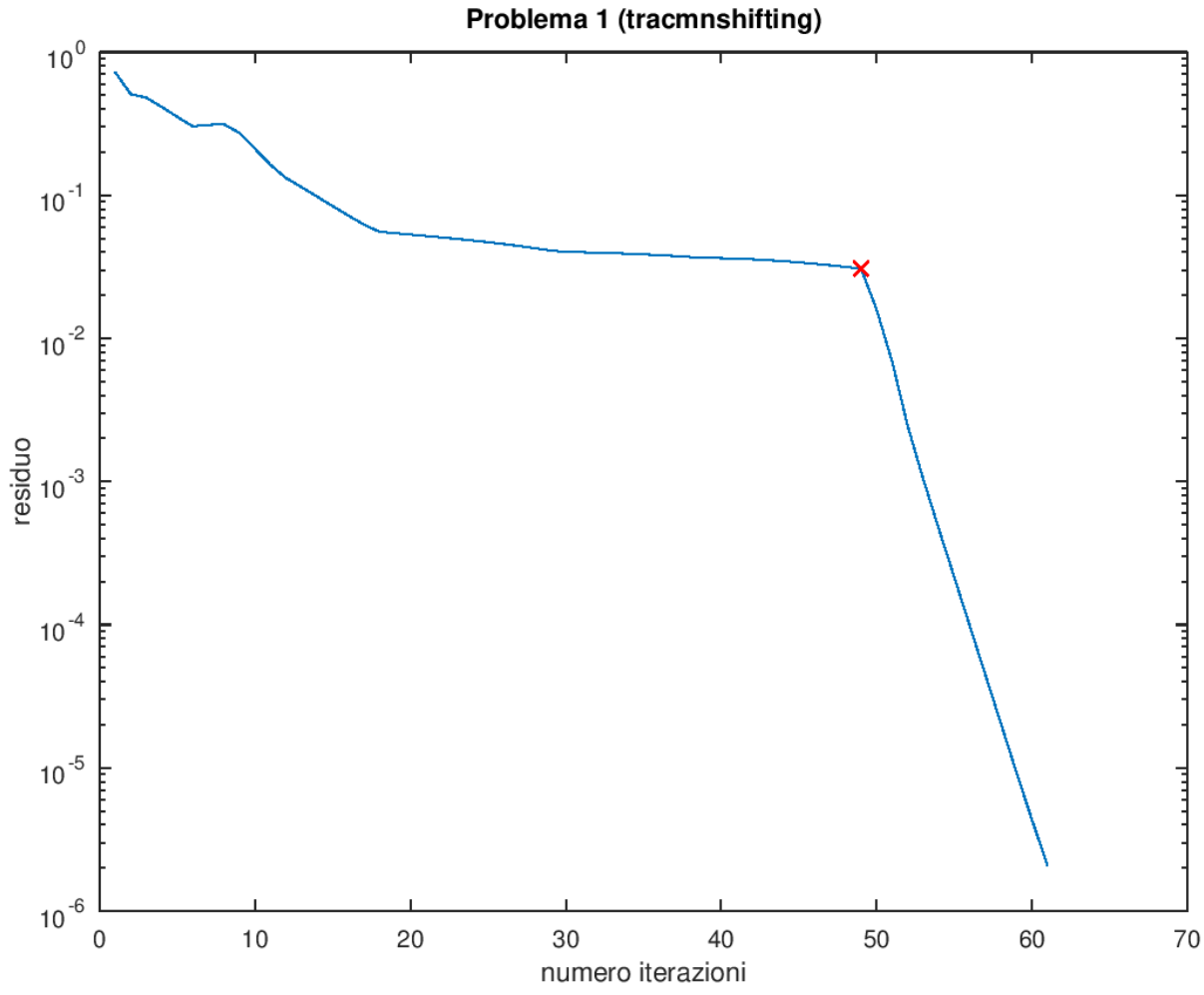
	start	tol	gamma	maxit	iterazioni	tempo(s)
tracmnshifting	10	$10e-6$	eps	1000	43	1.5043
//	30	$10e-6$	eps	1000	41	1.2675
//	50	$10e-6$	eps	1000	61	1.5043
//	100	$10e-6$	eps	1000	102	1.7245



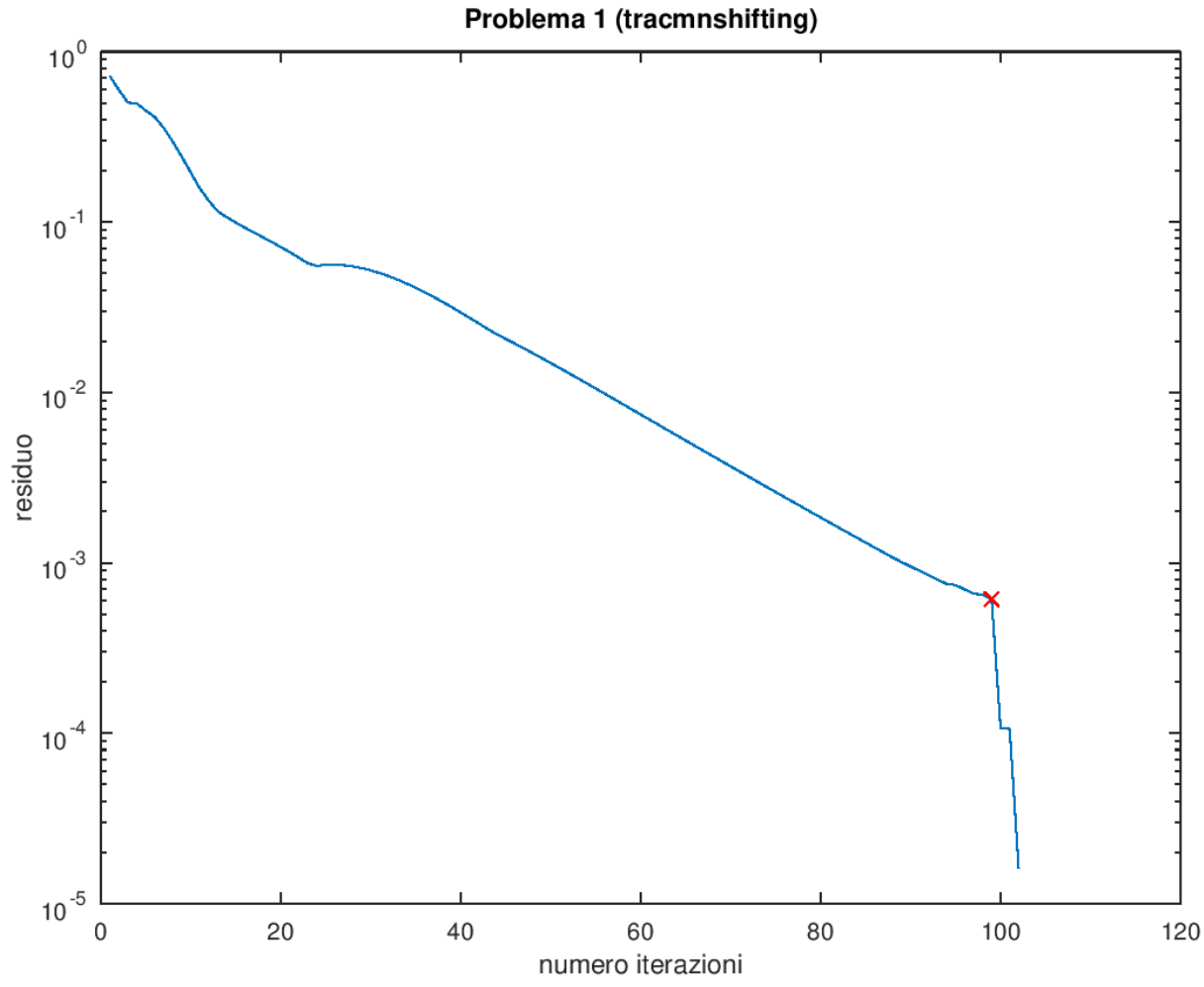
Legenda x = "inizia lo shifting"



Legenda x = "inizia lo shifting"



Legenda x = "inizia lo shifting"

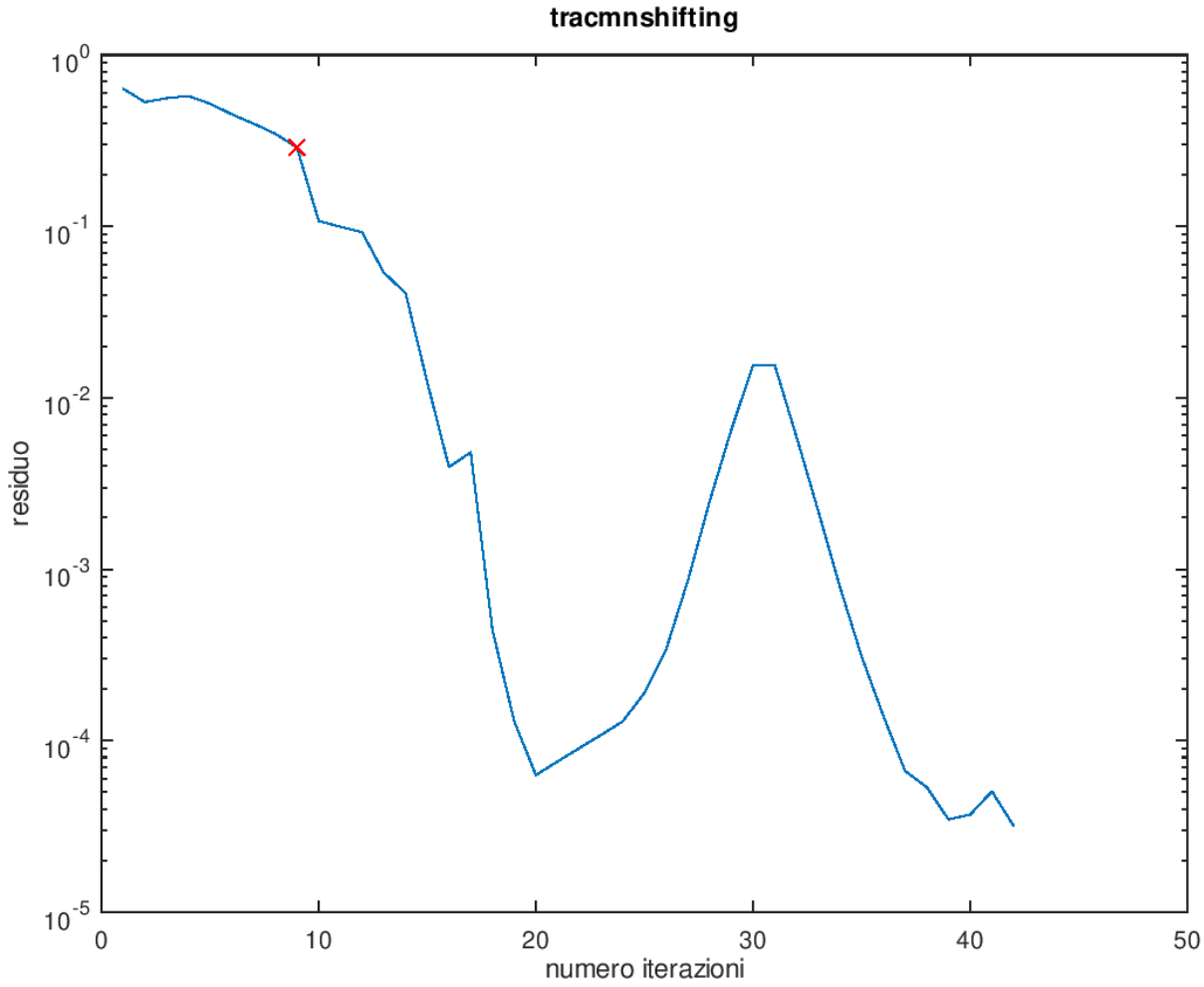


Legenda x = "inizia lo shifting"

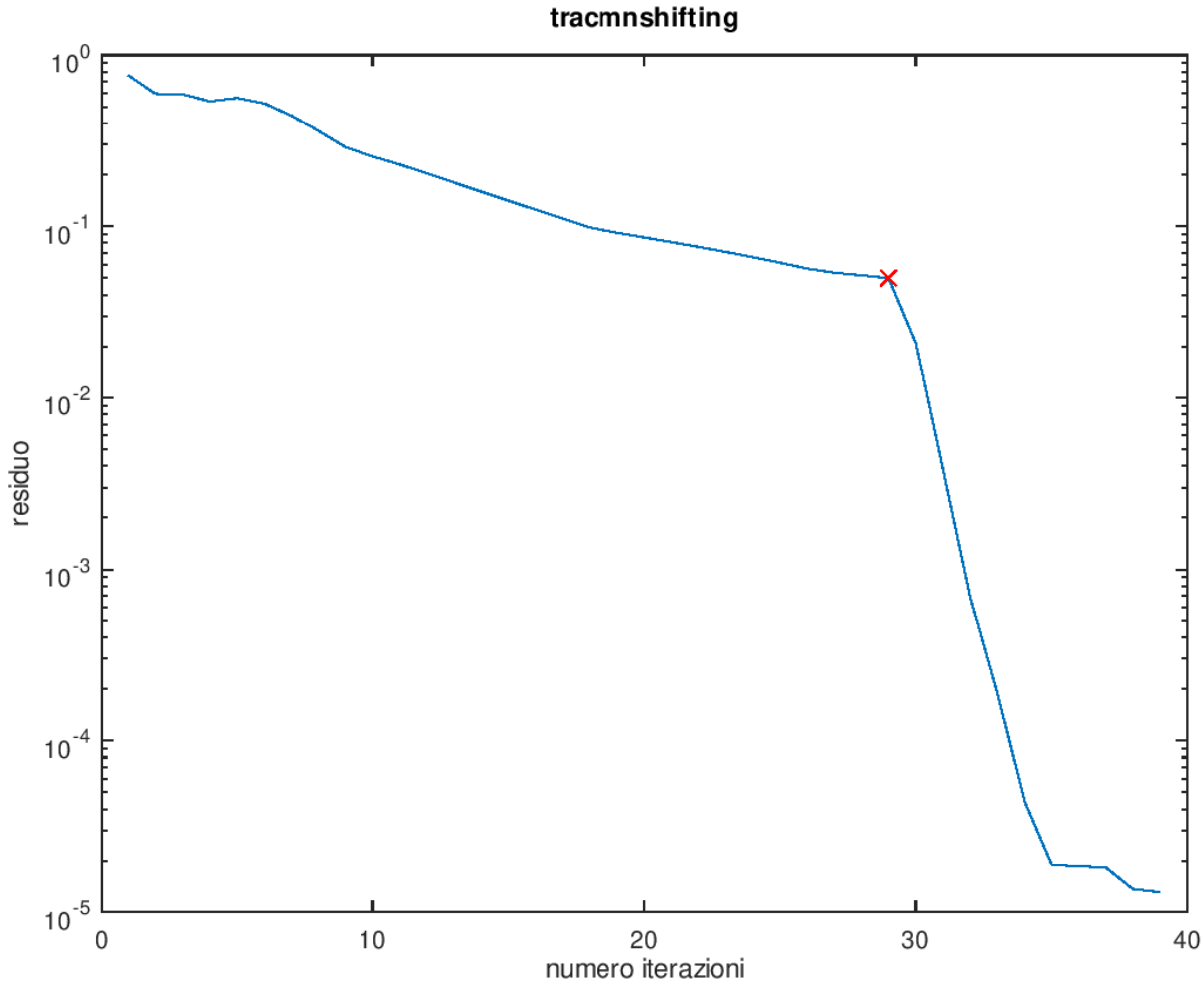
Problema	n	p	d_A	d_B
2	100	5	0.19760	0.19840

Risultati

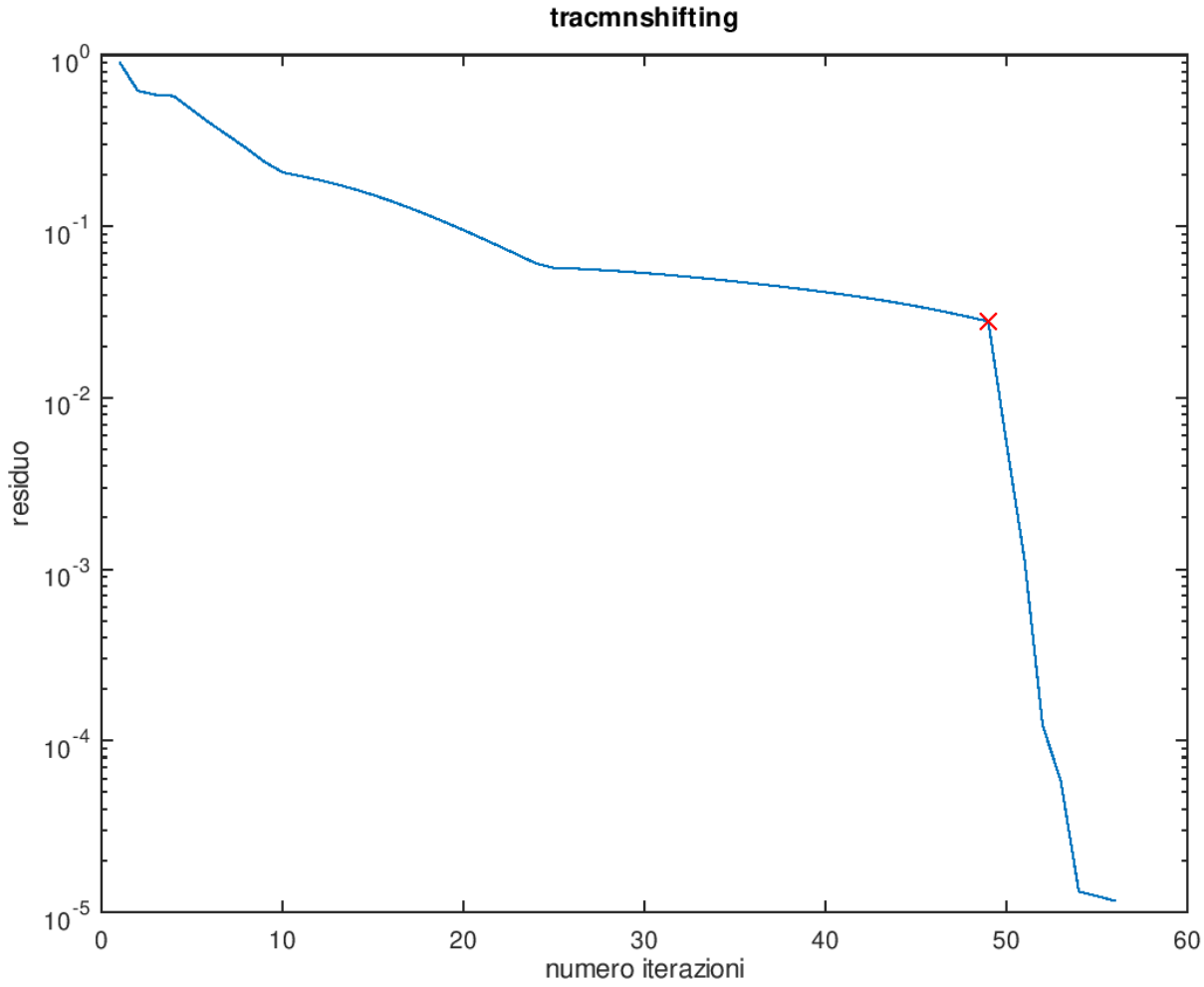
	start	tol	gamma	maxit	iterazioni	tempo(s)
tracmnshifting	10	$10e-6$	eps	1000	42	18.595
//	30	$10e-6$	eps	1000	39	9.9919
//	50	$10e-6$	eps	1000	56	11.692
//	100	$10e-6$	eps	1000	102	17.260



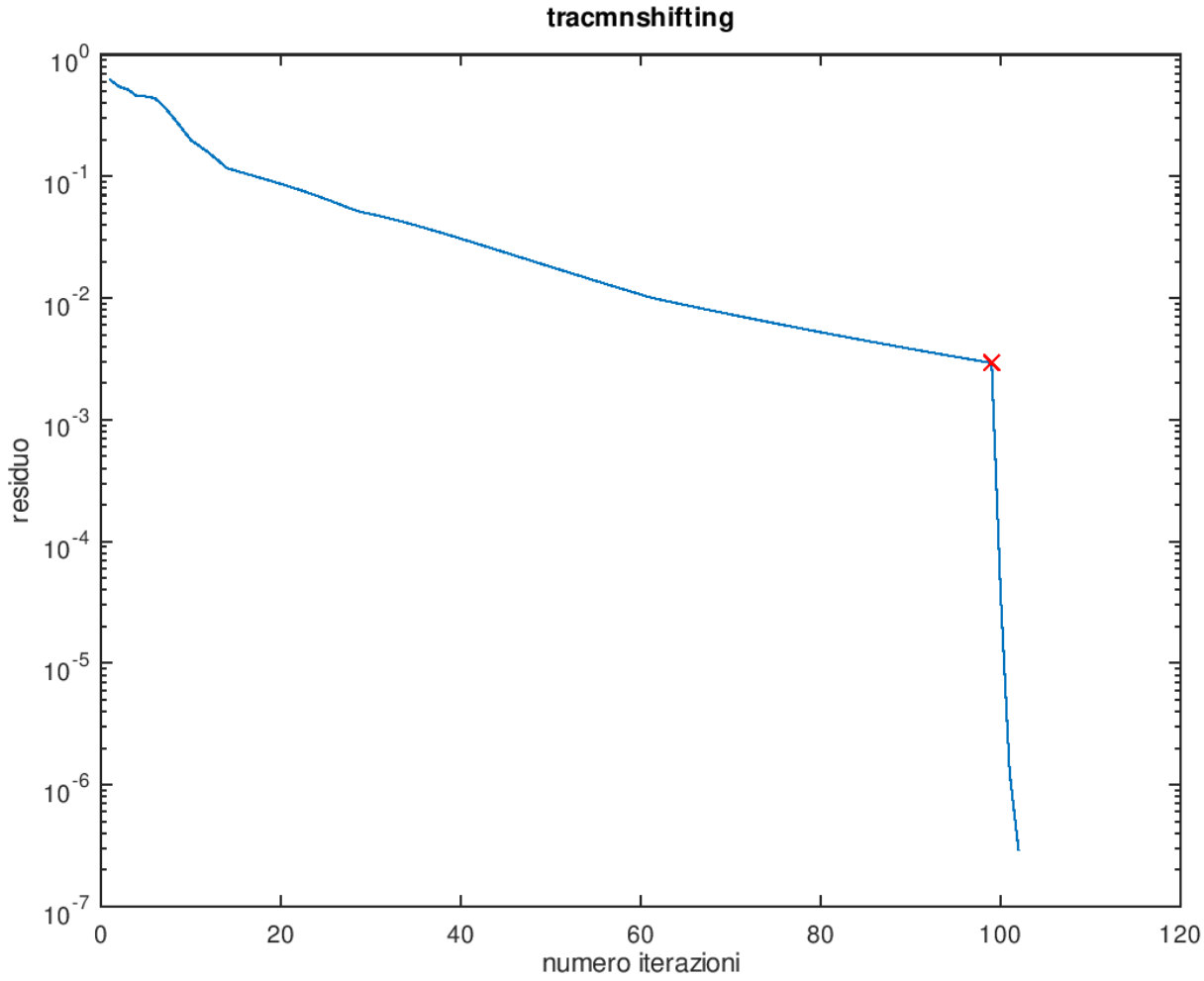
Legenda x = "inizia lo shifting"



Legenda x = "inizia lo shifting"



Legenda x = "inizia lo shifting"



Legenda x = "inizia lo shifting"

Appendice: Δ_k tramite passo di discesa rapido

Consideriamo le correzioni di Δ_k nel sottospazio tangente all' insieme dei vincoli $Y_k B Y_k = I_p$, ciò implica

$$\Delta_k^t B Y_k = 0.$$

Tale condizione risulta soddisfatta se scegliamo $\Delta_k = (I - P_k) Z_k$ dove:

- Z_k è una matrice $n \times p$ che specificheremo in seguito,
- $P_k = B Y_k (Y_k^t B^2 Y_k)^{-1} Y_k^t B$ è la proiezione sul sottospazio generato dalle colonne di BY_k cioè $P_k B Y_k = B Y_k$. P_k è una matrice simmetrica e $P_k^2 = P_k$.

Concentriamoci quindi ora sulla scelta di Δ_k:

$$\Delta_k = (I - P_k)Z_k.$$

Una scelta ragionevole di Z_k per cui vale la disuguaglianza sulle tracce è

$$Z_k = AY_kW_k$$

con $W_k = diag(\omega_1^{(k)}, ..., \omega_p^{(k)})$ da determinare.
Segue che:

$$\begin{aligned} tr(\hat{Y}^t A\hat{Y}) &= tr((Y_k + \Delta_k)^t A(Y_k + \Delta_k)) \\ &= tr(Y_k^t AY_k) + 2tr(Y_k^t A\Delta_k) + tr(\Delta_k^t A\Delta_k) \\ &= tr(\Sigma_k) + 2tr(C_kW_k) + tr(E_kW_k^2) \\ &= tr(\Sigma_k) + tr(2C_kW_k + E_kW_k^2) \end{aligned}$$

dove:

$$C_k = Y_k^t A(I - P_k)AY_k \qquad \text{e} \qquad E_k = Y_k^t A(I - P_k)A(I - P_k)AY_k.$$

Vogliamo determinare $W_k$in modo da minimizzare $tr(\hat{Y}^t A\hat{Y})$, cioè minimizzare

$$\Phi_k = tr(2C_k W_k + E_k W_k^2).$$

Otteniamo il minimo scegliendo:

$$\begin{cases} \omega_i^{(k)} = -\frac{\gamma_{ii}^{(k)}}{\epsilon_{ii}^{(k)}} & \text{se } \epsilon_{ii}^{(k)} \neq 0 \\ \omega_i^{(k)} = 0 & \text{se } \epsilon_{ii}^{(k)} = 0 \end{cases}$$

dove $\gamma_{ii}^{(k)}$ e $\epsilon_{ii}^{(k)}$ sono gli elementi diagonali di C_k e E_k, per $i = 1, ..., p$.
Tale minimo vale

$$\min_{W_k} \Phi_k = -\sum_{i=1, \epsilon_{ii} \neq 0}^{p} \frac{{\gamma_{ii}^{(k)}}^2}{\epsilon_{ii}^{(k)}} < 0.$$

Con tale scelta di Δ_k si ha:

$$tr(Y_{k+1}^t A Y_{k+1}) < tr(\hat{Y}^t A\hat{Y}) < tr(\Sigma_k) = tr(Y_k^t A Y_k).$$

Appendice: metodo del Gradiente Coniugato

Il sistema 4b è risolto tramite il metodo del Gradiente Coniugato (CG).
Viene scelta come iterazione iniziale il vettore $g = 0$ (per semplicità identifichiamo g_2 con g).
Sotto, un algoritmo che implementa tale metodo per la soluzione di 4b.
Termina quando la norma 2 del residuo è ridotta di un fattore $\gamma < 1$.

$g \leftarrow 0$
$r \leftarrow Q_2^t A y - Q_2^t A Q_2 g$ ▷ r residuo relativo
$q \leftarrow r$ ▷ q direzione di decrescita
$res_0 \leftarrow \|r\|$ ▷ res norma due del residuo
while $res > \gamma res_0$ **do**
$\alpha \leftarrow r^t r / q^t Q_2^t A Q_2 q$ ▷ α scalare, scelta ottimale per la direzione di decrescita q
$g \leftarrow g + \alpha q$
$r_{new} \leftarrow r - \alpha Q_2^t A Q_2 q$
$\beta \leftarrow r_{new}^t r_{new} / r^t r$
$q \leftarrow r + \beta q$ ▷ In tal modo $q_0, ..., q_{k+1}$ sono $Q_2^t A Q_2$-coniugati
$r \leftarrow r_{new}$
$res \leftarrow \|r\|$
end while

Appendice: terminare il processo CG

Definizione

Sia $\varepsilon(g_m)$ la funzione dell' errore al passo m dell' algoritmo CG applicato al sistema

$$Q_2^t A Q_2 g = Q_2 A y, \tag{6}$$

esso è dato da:

$$\varepsilon(g_m) = (g_m - g^*)^t Q_2^t A Q_2 (g_m - g^*)$$

dove g^* è l' unica soluzione al sistema 6.

Il problema nelle coordinate originali é

$$\text{minimizza } (y-d)^t A (y-d) \quad \text{soggetto ai vincoli } YBd = 0$$

quindi la funzione errore si può scrivere nel modo equivalente

$$\varepsilon(y_m) = (y_m - y^*)^t A (y_m - y^*)$$

dove $y^* = y - d^*$ e $d^* = Q_2 g^*$.

Teorema (9)

Per l' algoritmo CG

$$\varepsilon(y_m) \leq 4 \left(\frac{1 - k^{-\frac{1}{2}}}{1 + k^{-\frac{1}{2}}} \right)^{2m} \varepsilon(y_0) \tag{7}$$

dove k è il numero di condizionamento di $Q_2^t A Q_2$.

Dal Teorema 6 abbiamo che

$$\bar{f}_j^t A \bar{f}_j \leq \left(\frac{\lambda_j}{\lambda p + 1} \right) f_j^t A f_j + O(\|F\|^3) \tag{8}$$

dove $f_j = F e_j$ e $\bar{f}_j = \bar{F} e_j$ rappresentano gli errori nelle approssimazioni degli autovettori $y_j = Y e_j$ e $\bar{y}_j = \bar{Y} e_j$ per due iterate successive.

Grazie ai risultati precedenti 9 e 39, scegliamo per il passo m del metodo CG il più piccolo intero per cui

$$\varepsilon(y_j^{(m)}) \leq \left(\frac{\lambda_j}{\lambda_{p+1}}\right)^2 \varepsilon(y_j^{(0)})$$

La quantità $\varepsilon(y_j^{(k)})$ può essere stimata da:

$$\hat{\varepsilon}(y_j^{(k)}) = (y_j^{(k)} - y_j^{(k+1)})^t A(y_j^{(k)} - y_j^{(k+1)}).$$

Teorema (10)

Per ogni arbitrario vettore non nullo u e scalare σ, c'è un autovalore λ di A tale che:

$$|\lambda - \sigma| \leq \frac{\|(A - \sigma I)u\|}{\|u\|}.$$

Teorema (11)

Per $j \leq p$,

$$\lambda_j - \sigma_j^{(k)} = -e_j^t F^t (\Lambda - \lambda_j I) F e_j.$$

Dim.

Sia Y_k k-esima sezione: $Y_k^t B Y_k = I_p$ e $Y_k = ZG$ con $G = \begin{bmatrix} I_p \\ 0 \end{bmatrix} + F$; allora per la prima uguaglianza si ha che $G^t G = I_p$. Di conseguenza $e_j^t G^t G e_j = 1$.
Sviluppando si ottiene:

$$e_j^t e_j + 2 e_j^t F e_j + e_j^t F^t F e_j = 1.$$

Quindi $e_j^t F^t F e_j = -2 e_j^t F e_j$.
Inoltre,

$$\begin{aligned} \sigma_j^{(k)} &= y_j^{(k)t} A y_j^{(k)} = e_j^t G^t \Lambda G e_j = \\ &= e_j^t \Lambda e_j + 2\lambda_j e_j^t F e_j + e_j^t F^t \Lambda F e_j = \\ &= \lambda_j - \lambda_j e_j^t F^t F e_j + e_j^t F^t \Lambda F e_j = \\ &= \lambda_j + e_j^t F^t (\Lambda - \lambda_j I) F e_j, \end{aligned}$$

dalla quale segue la tesi. □