Algoritmo per risoluzione di equazioni razionali per matrici

Nikita Deniskin

Università di Pisa

Metodi di Approssimazione
prof. Federico Poloni

16 Dicembre 2021

Sia $f : \Omega \to \mathbb{C}$ una funzione analitica, con dominio $\Omega \subseteq \mathbb{C}$. Sia A una matrice di dimensione N.
Vogliamo trovare una soluzione X all'equazione:

$$f(X) = A$$

Sia $f : \Omega \to \mathbb{C}$ una funzione analitica, con dominio $\Omega \subseteq \mathbb{C}$. Sia A una matrice di dimensione N.
Vogliamo trovare una soluzione X all'equazione:

$$f(X) = A$$

Mostriamo algoritmo per il caso f funzione razionale:

$$r(X) = A, \qquad r(x) = \frac{p(x)}{q(x)}$$

$p(x), q(x)$ sono polinomi di grado $m-1$ e $n-1$, rispettivamente.

Definizione (Soluzione Primaria)

Sia X una soluzione all'equazione $f(X) = A$. Essa è *primaria* se è soddisfatta una delle seguenti condizioni equivalenti:

1. Esiste un polinomio p tale che $p(A) = X$.
2. Esiste f^{-1} un'inversa di f, definita sugli autovalori di A e analitica in un intorno degli autovalori con indice ≥ 1, tale che $f^{-1}(A) = X$.
3. Valgono le due condizioni
   a Per ogni coppia di autovalori ξ_1, ξ_2 di X, si ha $f(\xi_1) \neq f(\xi_2)$.
   b Ogni autovalore ξ di X tale che $f'(\xi) = 0$, è un autovalore *semisemplice*.

Definizione (Soluzione Primaria)

Sia X una soluzione all'equazione $f(X) = A$. Essa è *primaria* se è soddisfatta una delle seguenti condizioni equivalenti:

1. Esiste un polinomio p tale che $p(A) = X$.
2. Esiste f^{-1} un'inversa di f, definita sugli autovalori di A e analitica in un intorno degli autovalori con indice ≥ 1, tale che $f^{-1}(A) = X$.
3. Valgono le due condizioni
   - a Per ogni coppia di autovalori ξ_1, ξ_2 di X, si ha $f(\xi_1) \neq f(\xi_2)$.
   - b Ogni autovalore ξ di X tale che $f'(\xi) = 0$, è un autovalore *semisemplice*.

$1 \Rightarrow 3$ La condizione 3a) è necessaria per l'esistenza dell'inversa $f^{-1}(A)$. La condizione 3b) serve per evitare che un blocco di Jordan di X si spezzi.
Per le altre implicazioni, basta scegliere come p un polinomio interpolatore appropiato.

Per una qualsiasi g derivabile, definiamo la differenza divisa

$$g[x,y] = \begin{cases} \dfrac{g(x)-g(y)}{x-y} & \text{se } x \neq y; \\ g'(x) & \text{se } x = y; \end{cases}$$

Per una qualsiasi g derivabile, definiamo la differenza divisa

$$g[x,y] = \begin{cases} \dfrac{g(x)-g(y)}{x-y} & \text{se } x \neq y; \\ g'(x) & \text{se } x = y; \end{cases}$$

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1,0,\dots,0,0)^\top$ e $e_N = (0,0,\dots,0,1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A+\gamma e_1 e_N^\top) = f(A) + f'(\lambda)\,\gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A+\gamma e_1 e_N^\top) = f(A) + f[\lambda,\mu]\,\gamma e_1 e_N^\top$.

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1, 0, \dots, 0, 0)^\top$ e $e_N = (0, 0, \dots, 0, 1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f'(\lambda)\,\gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f[\lambda, \mu]\,\gamma e_1 e_N^\top$.

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1, 0, \dots, 0, 0)^\top$ e $e_N = (0, 0, \dots, 0, 1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f'(\lambda)\, \gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f[\lambda, \mu]\, \gamma e_1 e_N^\top$.

$$
\overset{A}{\begin{pmatrix}
\lambda & 1 & & & & \\
& \ddots & 1 & & & \\
& & \lambda & 0 & & \\
& & & \mu & 1 & \\
& & & & \ddots & 1 \\
& & & & & \mu
\end{pmatrix}}
\qquad
\overset{f(A)}{\begin{pmatrix}
f(\lambda) & * & * & & & \\
& \ddots & * & & & \\
& & f(\lambda) & & & \\
& & & f(\mu) & * & * \\
& & & & \ddots & * \\
& & & & & f(\mu)
\end{pmatrix}}
$$

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1, 0, \ldots, 0, 0)^\top$ e $e_N = (0, 0, \ldots, 0, 1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f'(\lambda)\,\gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f[\lambda, \mu]\,\gamma e_1 e_N^\top$.

$$
\overset{A+\gamma e_1 e_N^\top}{\begin{pmatrix}
\lambda & 1 & & & & {\color{red}\gamma} \\
 & \ddots & 1 & & & \\
 & & \lambda & 0 & & \\
 & & & \mu & 1 & \\
 & & & & \ddots & 1 \\
 & & & & & \mu
\end{pmatrix}}
\qquad
\overset{f(A+\gamma e_1 e_N^\top)}{\begin{pmatrix}
f(\lambda) & * & * & & & {\color{red}\gamma\, f[\lambda,\mu]} \\
 & \ddots & * & & & \\
 & & f(\lambda) & & & \\
 & & & f(\mu) & * & * \\
 & & & & \ddots & * \\
 & & & & & f(\mu)
\end{pmatrix}}
$$

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1, 0, \dots, 0, 0)^\top$ e $e_N = (0, 0, \dots, 0, 1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f'(\lambda)\,\gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f[\lambda, \mu]\,\gamma e_1 e_N^\top$.

Dimostrazione: Dividendo A in blocchi di taglia $N-1$ e 1, si ha:

$$A = \begin{pmatrix} B & 0 \\ 0 & a_{nn} \end{pmatrix} \qquad f(A) = \begin{pmatrix} f(B) & 0 \\ 0 & f(a_{nn}) \end{pmatrix}$$

$$A + \gamma e_1 e_N^\top = \begin{pmatrix} B & v \\ 0 & a_{nn} \end{pmatrix} \qquad f(A + \gamma e_1 e_N^\top)) = \begin{pmatrix} f(B) & * \\ 0 & f(a_{nn}) \end{pmatrix}$$

Lemma

Sia A una matrice di taglia N e f funzione. Siano $e_1 = (1, 0, \dots, 0, 0)^\top$ e $e_N = (0, 0, \dots, 0, 1)^\top$ e $\gamma \in \mathbb{C}$.

- Se A è un blocco di Jordan con autovalore λ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f'(\lambda)\,\gamma e_1 e_N^\top$.
- Se A è formata da due blocchi di Jordan con autovalori λ, μ, allora $f(A + \gamma e_1 e_N^\top) = f(A) + f[\lambda, \mu]\,\gamma e_1 e_N^\top$.

Dimostrazione: Dividendo A in blocchi di taglia $N-1$ e 1, si ha:

$$A = \begin{pmatrix} B & 0 \\ 0 & a_{nn} \end{pmatrix} \qquad f(A) = \begin{pmatrix} f(B) & 0 \\ 0 & f(a_{nn}) \end{pmatrix}$$

$$A + \gamma e_1 e_N^\top = \begin{pmatrix} B & v \\ 0 & a_{nn} \end{pmatrix} \qquad f(A + \gamma e_1 e_N^\top)) = \begin{pmatrix} f(B) & * \\ 0 & f(a_{nn}) \end{pmatrix}$$

Le prime $N-1$ colonne sono uguali. Dividendo in blocchi di 1 e $N-1$, si ottiene che l'unico elemento diverso è in posizione $(1, N)$. Il suo valore può essere calcolato scomponendo a blocchi.

Osservazione. Se $f[\lambda, \mu] = 0$, allora $f(A) = f(A + \gamma e_1 e_N^\top)$.

$$
\begin{array}{c}
A + \gamma e_1 e_N^\top \\
\begin{pmatrix}
\lambda & 1 & & & & \color{red}{\gamma} \\
 & \ddots & 1 & & & \\
 & & \lambda & 0 & & \\
 & & & \mu & 1 & \\
 & & & & \ddots & 1 \\
 & & & & & \mu
\end{pmatrix}
\end{array}
\qquad
\begin{array}{c}
f(A + \gamma e_1 e_N^\top) \\
\begin{pmatrix}
f(\lambda) & * & * & & & \color{red}{\gamma\, f[\lambda, \mu]} \\
 & \ddots & * & & & \\
 & & f(\lambda) & & & \\
 & & & f(\mu) & * & * \\
 & & & & \ddots & * \\
 & & & & & f(\mu)
\end{pmatrix}
\end{array}
$$

Definizione

Una soluzione X all'equazione $f(X) = A$ si dice *isolata*, se esiste un intorno di $\mathbb{C}^{n^2}$ in cui X è l'unica soluzione dell'equazione.

Definizione

Una soluzione X all'equazione $f(X) = A$ si dice *isolata*, se esiste un intorno di $\mathbb{C}^{n^2}$ in cui X è l'unica soluzione dell'equazione.

Proposizione

Sia X una soluzione non primaria all'equazione $f(X) = A$. Allora X non è isolata.

Definizione

Una soluzione X all'equazione $f(X) = A$ si dice *isolata*, se esiste un intorno di $\mathbb{C}^{n^2}$ in cui X è l'unica soluzione dell'equazione.

Proposizione

Sia X una soluzione non primaria all'equazione $f(X) = A$. Allora X non è isolata.

Dimostrazione: Se X non è primaria, allora esistono $\xi_1 \neq \xi_2$ con $f(\xi_1) = f(\xi_2)$, oppure esiste ξ_0 con $f'(\xi_0) = 0$.

Definizione

Una soluzione X all'equazione $f(X) = A$ si dice *isolata*, se esiste un intorno di $\mathbb{C}^{n^2}$ in cui X è l'unica soluzione dell'equazione.

Proposizione

Sia X una soluzione non primaria all'equazione $f(X) = A$. Allora X non è isolata.

Dimostrazione: Se X non è primaria, allora esistono $\xi_1 \neq \xi_2$ con $f(\xi_1) = f(\xi_2)$, oppure esiste ξ_0 con $f'(\xi_0) = 0$.
Nel primo caso, siano J_1, J_2 i relativi blocchi di Jordan. Sia $X(\gamma)$ la matrice ottenuta sostituendo il blocco $T =$diag(J_1, J_2) con $T + \gamma e_1 e_k^\top$, si ha $f(T) = f(T + \gamma e_1 e^\top)$ per il lemma. Dunque $f(X) = f(X(\gamma))$ per ogni $\gamma \in \mathbb{C}$.
Il secondo caso si fa similmente.

Definizione

Una soluzione X all'equazione $f(X) = A$ si dice *isolata*, se esiste un intorno di $\mathbb{C}^{n^2}$ in cui X è l'unica soluzione dell'equazione.

Proposizione

Sia X una soluzione non primaria all'equazione $f(X) = A$. Allora X non è isolata.

Dimostrazione: Se X non è primaria, allora esistono $\xi_1 \neq \xi_2$ con $f(\xi_1) = f(\xi_2)$, oppure esiste ξ_0 con $f'(\xi_0) = 0$.
Nel primo caso, siano J_1, J_2 i relativi blocchi di Jordan. Sia $X(\gamma)$ la matrice ottenuta sostituendo il blocco $T =$diag(J_1, J_2) con $T + \gamma e_1 e_k^\top$, si ha $f(T) = f(T + \gamma e_1 e^\top)$ per il lemma. Dunque $f(X) = f(X(\gamma))$ per ogni $\gamma \in \mathbb{C}$.
Il secondo caso si fa similmente.

Corollario. Una soluzione isolata è primaria.

Teorema

Sia X una soluzione all'equazione $f(X) = A$. Siano $\xi_1, \ldots, \xi_N$ gli autovalori di X. Sono equivalenti:

1. X è una soluzione isolata.
2. X è primaria e vale una condizione più forte:
   - a Per ogni coppia di autovalori ξ_i, ξ_j di X, si ha $f(\xi_i) \neq f(\xi_j)$.
   - b Ogni autovalore ξ di X tale che $f'(\xi) = 0$, è un autovalore *semplice*.
3. X è l'unica soluzione con autovalori $\xi_1, \ldots, \xi_N$.
4. $f[\xi_i, \xi_j] \neq 0$ per ogni $i \neq j$.

Risolvere $r(X) = A$ è equivalente a $p(X) = A\,q(X)$.

Risolvere $r(X) = A$ è equivalente a $p(X) = A\,q(X)$.
Scriviamo A in forma di Schur, con T triangolare superiore e U unitaria

$$A = U\,T\,U^{-1}$$

Risolvere $r(X) = A$ è equivalente a $p(X) = A\, q(X)$.
Scriviamo A in forma di Schur, con T triangolare superiore e U unitaria

$$A = U\, T\, U^{-1}$$

Trasformo X con il cambio di base dato da U

$$X = U\, Y\, U^{-1}$$

Risolvere $r(X) = A$ è equivalente a $p(X) = A\,q(X)$.
Scriviamo A in forma di Schur, con T triangolare superiore e U unitaria

$$A = U\,T\,U^{-1}$$

Trasformo X con il cambio di base dato da U

$$X = U\,Y\,U^{-1}$$

Anche T, Y soddisfano l'equazione:

$$p(Y) = T\,q(Y)$$

Cerchiamo Y con la stessa suddivisione triangolare superiore a blocchi di T. Anche $q(Y)$ e $p(Y)$ avranno questa suddivisione.

$$p(Y) = c_m Y^{m-1} + c_{m-1} Y^{m-2} + \ldots + c_3 Y^2 + c_2 Y + c_1 I$$

È possibile calcolare $p(Y)$ in modo ricorsivo, equivalentemente alla decomposizione di Horner:

$$p(Y) = c_m Y^{m-1} + c_{m-1} Y^{m-2} + \ldots + c_3 Y^2 + c_2 Y + c_1 I$$

È possibile calcolare $p(Y)$ in modo ricorsivo, equivalentemente alla decomposizione di Horner:

$$\begin{cases} P^{[m]} = c_m\, I \\ P^{[m-1]} = c_{m-1}\, I + Y\, P^{[m]} \\ \qquad \ldots \\ P^{[2]} = c_2\, I + Y\, P^{[3]} \\ P^{[1]} = c_1\, I + Y\, P^{[2]} \end{cases}$$

$P^{[k]}$ è una matrice di taglia N, che dipende da Y.
(L'indicizzazione è stata scelta per essere la stessa del codice Matlab)

Similmente per il polinomio $q(x)$, di grado $n-1$:

$$q(Y) = d_n Y^{n-1} + d_{n-1} Y^{n-2} + \ldots + d_3 Y^2 + d_2 Y + d_1 I$$

$$\begin{cases} Q^{[n]} = d_n\, I \\ Q^{[n-1]} = d_{n-1}\, I + Y\, Q^{[n]} \\ \qquad \ldots \\ Q^{[2]} = d_2\, I + Y\, Q^{[3]} \\ Q^{[1]} = d_1\, I + Y\, Q^{[2]} \end{cases}$$

Guardiamo la ricorrenza sui blocchi diagonali di $P^{[u]}$:

$$P^{[u]} = c_u \, I + Y \, P^{[u+1]}$$

$$P_{ii}^{[u]} = c_u + (Y \cdot P^{[u+1]}) = c_u + \sum_{h=1}^{N} Y_{ih} \, P_{hi}^{[u+1]} = c_u + Y_{ii} \, P_{ii}^{[u+1]}$$

Guardiamo la ricorrenza sui blocchi diagonali di $P^{[u]}$:

$$P^{[u]} = c_u I + Y P^{[u+1]}$$

$$P_{ii}^{[u]} = c_u + (Y \cdot P^{[u+1]}) = c_u + \sum_{h=1}^{N} Y_{ih} P_{hi}^{[u+1]} = c_u + Y_{ii} P_{ii}^{[u+1]}$$

Questa dipende unicamente dal blocco (i, i). Per cui determiniamo Y_{ii} come una soluzione al sistema

$$p(Y_{ii}) = T_{ii} q(Y_{ii})$$

per ogni i, e da esse ricaviamo i blocchi Y_{ij} con $i \neq j$.

$$P_{ij}^{[u]} = (c_u\, I + Y\, P^{[u+1]})_{ij} = (Y\, P^{[u+1]})_{ij}$$

$$P_{ij}^{[u]} = \sum_{k=i}^{j+1} Y_{ik}\, P_{kj}^{[u+1]} = Y_{ii}\, P_{ij}^{[u+1]} + \sum_{k=i+1}^{j-1} Y_{ik}\, P_{kj}^{[u+1]} + Y_{ij}\, P_{jj}^{[u+1]}$$

$$P_{ij}^{[u]} = (c_u\, I + Y\, P^{[u+1]})_{ij} = (Y\, P^{[u+1]})_{ij}$$

$$P_{ij}^{[u]} = \sum_{k=i}^{j+1} Y_{ik}\, P_{kj}^{[u+1]} = Y_{ii}\, P_{ij}^{[u+1]} + \underbrace{\sum_{k=i+1}^{j-1} Y_{ik}\, P_{kj}^{[u+1]}}_{C_{ij}^{[u+1]}} + Y_{ij}\, P_{jj}^{[u+1]}$$

$$P_{ij}^{[u]} = (c_u\, I + Y\, P^{[u+1]})_{ij} = (Y\, P^{[u+1]})_{ij}$$

$$P_{ij}^{[u]} = \sum_{k=i}^{j+1} Y_{ik}\, P_{kj}^{[u+1]} = Y_{ii}\, P_{ij}^{[u+1]} + \underbrace{\sum_{k=i+1}^{j-1} Y_{ik}\, P_{kj}^{[u+1]}}_{C_{ij}^{[u+1]}} + Y_{ij}\, P_{jj}^{[u+1]}$$

$$P_{ij}^{[u]} = Y_{ii}\, P_{ij}^{[u+1]} + C_{ij}^{[u+1]} + Y_{ij}\, P_{jj}^{[u+1]}$$

$$P_{ij}^{[u]} = (c_u\, I + Y\, P^{[u+1]})_{ij} = (Y\, P^{[u+1]})_{ij}$$

$$P_{ij}^{[u]} = \sum_{k=i}^{j+1} Y_{ik}\, P_{kj}^{[u+1]} = Y_{ii}\, P_{ij}^{[u+1]} + \underbrace{\sum_{k=i+1}^{j-1} Y_{ik}\, P_{kj}^{[u+1]}}_{C_{ij}^{[u+1]}} + Y_{ij}\, P_{jj}^{[u+1]}$$

$$P_{ij}^{[u]} = Y_{ii}\, P_{ij}^{[u+1]} + C_{ij}^{[u+1]} + Y_{ij}\, P_{jj}^{[u+1]}$$

Valori per m grande

$$C_{ij}^{[m]} = 0, \qquad P_{ij}^{[m]} = [c_m I]_{ij} = 0 \quad \text{perché } i \neq j,$$

$$P_{ij}^{[m-1]} = 0 + 0 + Y_{ij}\, P_{jj}^{[m]} = c_m P_{jj}^{[m]},$$

Invece $C_{ij}^{[m-1]}$, $P_{ij}^{[m-2]}$ e successivi sono non nulli.

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} C_{ij}^{[u+f]}$$

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} \, P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} \, C_{ij}^{[u+f]}$$

Dimostrazione:

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \, P_{ij}^{[u+1]}$$

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} \, P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} \, C_{ij}^{[u+f]}$$

Dimostrazione:

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \, P_{ij}^{[u+1]}$$

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} \, P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} \, C_{ij}^{[u+f]}$$

Dimostrazione:

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \, P_{ij}^{[u+1]}$$

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \left(C_{ij}^{[u+2]} + Y_{ij} \, P_{jj}^{[u+2]} + Y_{ii} \, P_{ij}^{[u+2]} \right)$$

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} \, P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} \, C_{ij}^{[u+f]}$$

Dimostrazione:

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \, P_{ij}^{[u+1]}$$

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \left(C_{ij}^{[u+2]} + Y_{ij} \, P_{jj}^{[u+2]} + Y_{ii} \, P_{ij}^{[u+2]} \right)$$

Per $i \neq j$, il blocco $P_{ij}^{[u]}$ si può scrivere in funzione dei blocchi con u maggiore

$$P_{ij}^{[u]} = \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} \, P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} \, C_{ij}^{[u+f]}$$

Dimostrazione:

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \, P_{ij}^{[u+1]}$$

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \left(C_{ij}^{[u+2]} + Y_{ij} \, P_{jj}^{[u+2]} + Y_{ii} \, P_{ij}^{[u+2]} \right)$$

$$P_{ij}^{[u]} = C_{ij}^{[u+1]} + Y_{ij} \, P_{jj}^{[u+1]} + Y_{ii} \left(C_{ij}^{[u+2]} + Y_{ij} \, P_{jj}^{[u+2]} + Y_{ii} \, (\dots) \right)$$

L'equazione $p(Y) = T\,q(Y)$, ovvero $P^{[1]} = T\,Q^{[1]}$, ha in (i,j):

$$P_{ij}^{[1]} = \left(T\,Q^{[1]}\right)_{ij} = T_{ii}\,Q_{ij}^{[1]} + \sum_{k=i+1}^{j} T_{ik}\,Q_{kj}^{[1]}$$

L'equazione $p(Y) = T\,q(Y)$, ovvero $P^{[1]} = T\,Q^{[1]}$, ha in (i,j):

$$P_{ij}^{[1]} = \left(T\,Q^{[1]}\right)_{ij} = T_{ii}\,Q_{ij}^{[1]} + \sum_{k=i+1}^{j} T_{ik}\,Q_{kj}^{[1]}$$

$$P_{ij}^{[1]} - T_{ii}\,Q_{ij}^{[1]} = \sum_{k=i+1}^{j} T_{ik}\,Q_{kj}^{[1]}$$

$$\begin{aligned} P_{ij}^{[1]} - T_{ii}Q_{ij}^{[1]} = \sum_{e=0}^{m-2} Y_{ii}^{e}\,Y_{ij}\,P_{jj}^{[e+2]} + \sum_{f=0}^{m-3} Y_{ii}^{f}\,C_{ij}^{[f+2]} - \\ - T_{ii}\left(\sum_{g=0}^{n-2} Y_{ii}^{g}\,Y_{ij}\,Q_{jj}^{[g+2]} + \sum_{h=0}^{n-3} Y_{ii}^{h}\,D_{ij}^{[h+2]}\right) \end{aligned}$$

$$P_{ij}^{[1]} - T_{ii} Q_{ij}^{[1]} = \sum_{e=0}^{m-2} Y_{ii}^{e} Y_{ij} P_{jj}^{[e+2]} + \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} - \\ - T_{ii} \left(\sum_{g=0}^{n-2} Y_{ii}^{g} Y_{ij} Q_{jj}^{[g+2]} + \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]} \right)$$

$$\sum_{e=0}^{m-2} Y_{ii}^{e} Y_{ij} P_{jj}^{[e+2]} - T_{ii} \sum_{g=0}^{m-2} Y_{ii}^{g} Y_{ij} Q_{jj}^{[g+2]} = \\ = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[1]} - \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} + T_{ii} \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]}$$

$$P_{ij}^{[1]} - T_{ii} Q_{ij}^{[1]} = \sum_{e=0}^{m-2} Y_{ii}^{e} Y_{ij} P_{jj}^{[e+2]} + \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} - \\ - T_{ii} \left(\sum_{g=0}^{n-2} Y_{ii}^{g} Y_{ij} Q_{jj}^{[g+2]} + \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]} \right)$$

$$\sum_{e=0}^{m-2} Y_{ii}^{e} Y_{ij} P_{jj}^{[e+2]} - T_{ii} \sum_{g=0}^{m-2} Y_{ii}^{g} Y_{ij} Q_{jj}^{[g+2]} = \\ = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[1]} - \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} + T_{ii} \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]}$$

Supponiamo che tutti i blocchi diagonali di T siano 1×1. Allora l'equazione è scalare, e posso esplicitare Y_{ij}:

$$\left(\sum_{e=0}^{m-2} Y_{ii}^{e} P_{jj}^{[e+2]} - T_{ii} \sum_{g=0}^{m-2} Y_{ii}^{g} Q_{jj}^{[g+2]} \right) Y_{ij}$$
$$= \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[1]} - \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} + T_{ii} \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]}$$

Supponiamo che tutti i blocchi diagonali di T siano 1×1. Allora l'equazione è scalare, e posso esplicitare Y_{ij}:

$$\left(\sum_{e=0}^{m-2} Y_{ii}^{e} P_{jj}^{[e+2]} - T_{ii} \sum_{g=0}^{m-2} Y_{ii}^{g} Q_{jj}^{[g+2]} \right) Y_{ij} = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[1]} - \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} + T_{ii} \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]}$$

$$\psi_{ij} Y_{ij} = \varphi_{ij}$$

$$\begin{cases} \psi_{ij} = \sum_{e=0}^{m-2} Y_{ii}^{e} P_{jj}^{[e+2]} - T_{ii} \sum_{g=0}^{m-2} Y_{ii}^{g} Q_{jj}^{[g+2]} \\ \varphi_{ij} = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[1]} - \sum_{f=0}^{m-3} Y_{ii}^{f} C_{ij}^{[f+2]} + T_{ii} \sum_{h=0}^{n-3} Y_{ii}^{h} D_{ij}^{[h+2]} \end{cases}$$

Algoritmo invrat

Setting: caso in cui ogni blocco T_{ii} ha dimensione 1, sono scalari
Algoritmo: sia $l = j - i$, lavoriamo via via sulla l-esima diagonale

- Per $l = 0$, trovo ogni y_{ii} come soluzione di $p(y_{ii}) = t_{ii}\, q(y_{ii})$.
- Calcolo $p_{ii}^{[u]}, q_{ii}^{[v]}$ per ogni i, u, v.
- Per $l = 1, \dots, N-1$:
  1. Calcolo $C_{ij}^{[u]} = \sum_{k=i+1}^{j-1} Y_{ik}\, p_{kj}^{[u]}$ e $D_{ij}^{[v]} = \sum_{k=i+1}^{j-1} Y_{ik}\, q_{kj}^{[v]}$ per ogni u, v.
  2. Calcolo ψ_{ij} e φ_{ij}
  3. Calcolo $Y_{ij} = \varphi_{ij}/\psi_{ij}$.
  4. Aggiorno $p_{ij}^{[u]}, q_{ij}^{[v]}$.

Bisogna garantire che $\psi_{ij} \neq 0$.

Bisogna garantire che $\psi_{ij} \neq 0$.

Siano $p(x), q(x), r(x)$ come prima, e $a, b \in \mathbb{C}$.
Sia $p^{[k]}(x)$ il calcolo parziale di $p(x)$ utilizzando Horner.

$$\psi(a, b) = \sum_{e=0}^{m-2} a^e \, p^{[e+2]}(b) - r(a) \sum_{g=0}^{m-2} a^g \, q^{[g+2]}(b)$$

Bisogna garantire che $\psi_{ij} \neq 0$.

Siano $p(x), q(x), r(x)$ come prima, e $a, b \in \mathbb{C}$.
Sia $p^{[k]}(x)$ il calcolo parziale di $p(x)$ utilizzando Horner.

$$\psi(a, b) = \sum_{e=0}^{m-2} a^e \, p^{[e+2]}(b) - r(a) \sum_{g=0}^{m-2} a^g \, q^{[g+2]}(b)$$

Proposizione

Supponiamo $q(a), q(b) \neq 0$. Allora vale la seguente relazione:

$$\psi(a, b) = p[a, b] - r(a)\, q[a, b] = r[a, b]\, q(b)$$

$$\psi(a,b) = \sum_{e=0}^{m-2} a^e\, p^{[e+2]}(b) - r(a) \sum_{g=0}^{m-2} a^g\, q^{[g+2]}(b)$$

$$\psi_{ij} = \psi(Y_{ii}, Y_{jj}) = \sum_{e=0}^{m-2} Y_{ii}^e\, p^{[e+2]}(Y_{jj}) - r(Y_{ii}) \sum_{g=0}^{m-2} Y_{ii}^g\, q^{[g+2]}(Y_{jj}) =$$

$$\psi_{ij} = r[Y_{ii}, Y_{jj}]\, q(Y_{jj})$$

Proposizione

Si ha $\psi_{ij} \neq 0$ se e solo se

- Per $Y_{ii} \neq Y_{jj}$, se $r(Y_{ii}) \neq r(Y_{jj})$ e $q(Y_{jj}) \neq 0$,
- Per $Y_{ii} = Y_{jj}$ se $r'(Y_{ii}) \neq 0$ e $q(Y_{jj}) \neq 0$.

Teorema

Siano $p(x), q(x)$ coprimi, $r(x) = p(x)/q(x)$. Sia T triangolare superiore di taglia n, e Y una soluzione a $r(Y) = T$. Sono equivalenti:

- Y è una soluzione isolata
- Scegliendo Y_{ii} soluzione di $r(X) = T_{ii}$ per ogni i, l'algoritmo calcola univocamente Y_{ij}, ovvero la soluzione esiste ed è unica.

Teorema

Siano $p(x), q(x)$ coprimi, $r(x) = p(x)/q(x)$. Sia T triangolare superiore di taglia n, e Y una soluzione a $r(Y) = T$. Sono equivalenti:

- Y è una soluzione isolata
- Scegliendo Y_{ii} soluzione di $r(X) = T_{ii}$ per ogni i, l'algoritmo calcola univocamente Y_{ij}, ovvero la soluzione esiste ed è unica.

Dimostrazione: Ogni Y_{ij} è soluzione di $\psi_{ij} Y_{ij} = \varphi_{ij}$; basta mostrare che $\psi_{ij} \neq 0$, equivalente a $r[Y_{ii}, Y_{jj}]\, q(Y_{jj}) \neq 0$.
La condizione di coprimalità tra p, q implica $q(Y_{jj}) \neq 0$.
Per il Teorema, chiedere che $r[Y_{ii}, Y_{jj}] \neq 0$ per ogni i, j è equivalente a Y soluzione isolata.

Sperimentazione

Vogliamo risolvere il sistema $r(X) = A$, con $r(z)$ funzione razionale assegnata.
Confrontiamo tre algoritmi:

- INVRAT: l'algoritmo descritto, con forma triangolare di Schur
- DIAG: diagonalizza $A = UDU^{-1}$ e calcola $f(A) = Uf(D)U^{-1}$.
- APPROX_DIAG: perturba $A \to A + B$ con B dell'ordine di $\varepsilon \approx 10^{-16}$. Fattorizza $A + B = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$ e approssima $f(A)$ con $f(A + B) = \widetilde{U}f(\widetilde{D})\widetilde{U}^{-1}$.

Esperimento 1

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 39

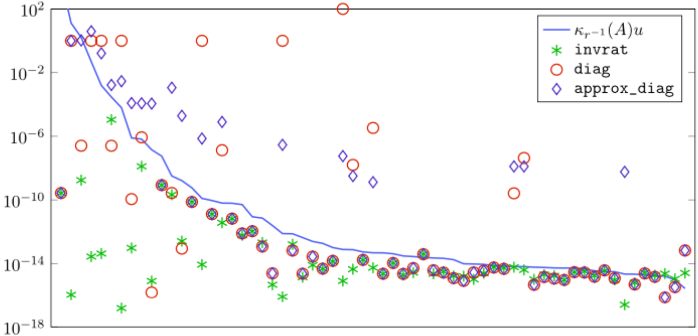


Fig. 1. Relative forward errors of **invrat**, **diag**, and **approx_diag** on the test set.

$$r(z) = \left(\frac{z^3}{120} + \frac{z^2}{10} + \frac{z}{2} + 1\right)\left(-\frac{z^3}{120} + \frac{z^2}{10} - \frac{z}{2} + 1\right)^{-1}$$

Sono scelte 63 matrici A, con N che varia da 2 a 10, prese da GALLERY di Matlab.

Esperimento 1

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 39

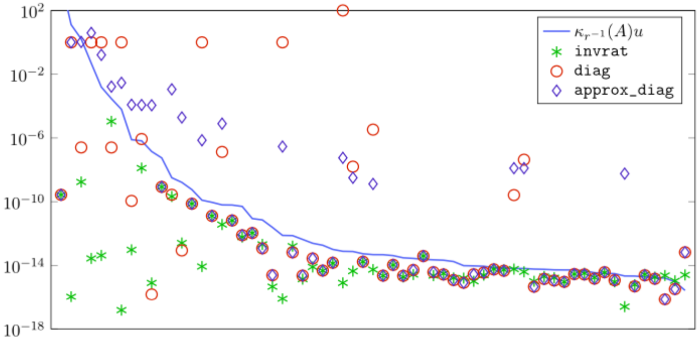


Fig. 1. Relative forward errors of **invrat**, **diag**, and **approx_diag** on the test set.

Viene calcolata la soluzione $X = r^{-1}(A)$ tramite un pacchetto ad alta precisione (Advanpix Multiprecision Computing Toolbox), con 512 cifre decimali.

Esperimento 1

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 39

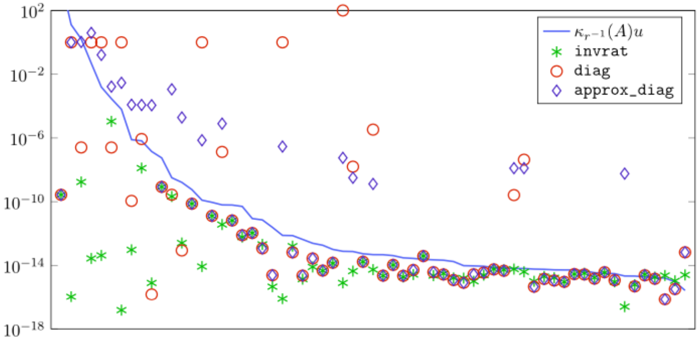


Fig. 1. Relative forward errors of **invrat**, **diag**, and **approx_diag** on the test set.

Per ciascuna A, sono calcolate $X_1 = \text{invrat}(A)$, $X_2 = \text{diag}(A)$, $X_3 = \text{approx\_diag}(A)$. Nel grafico vengono rappresentati le norme degli errori $X - X_1$, $X - X_2$, $X - X_3$.

Esperimento 1

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 39

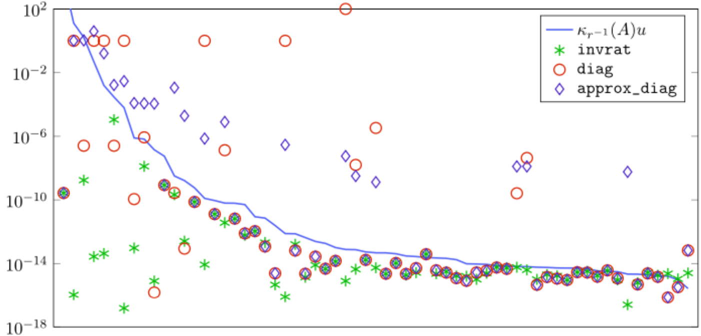
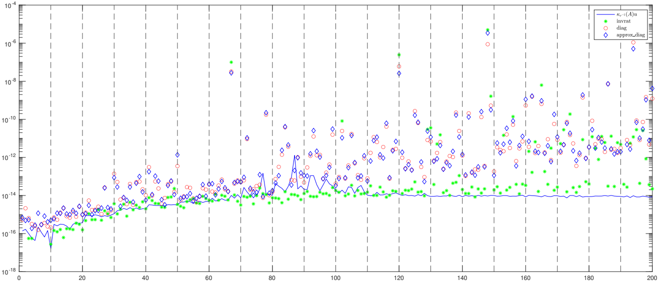


Fig. 1. Relative forward errors of **invrat**, **diag**, and **approx_diag** on the test set.

La linea blu rappresenta il condizionamento in norma 1 della funzione di matrice $r^{-1}(A)$, stimata con FUNM_CONDEST di Higham (Matrix Function Toolbox).

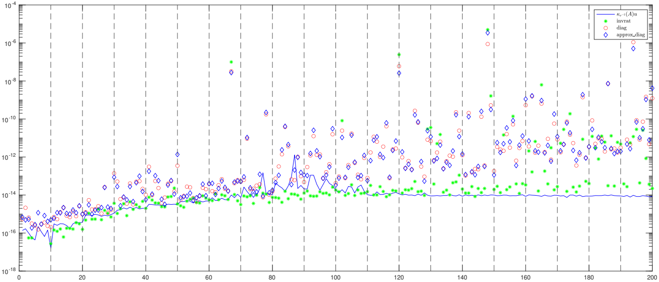
Esperimento 1



$$r(z) = \left(\frac{z^3}{120} + \frac{z^2}{10} + \frac{z}{2} + 1\right)\left(-\frac{z^3}{120} + \frac{z^2}{10} - \frac{z}{2} + 1\right)^{-1}$$

Per ogni N tra 1 e 20, sono generate 10 matrici A con distribuzione uniforme su $[0, 1]$ per ciascuna entrata.

Esperimento 1



Per ciascuna A, sono calcolate $X_1 = \text{invrat}(A)$, $X_2 = \text{diag}(A)$, $X_3 = \text{approx\_diag}(A)$. Nel grafico vengono rappresentati le norme degli errori di $A - r(X_1)$, $A - r(X_2)$, $A - r(X_3)$.

Esperimento 2

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 41

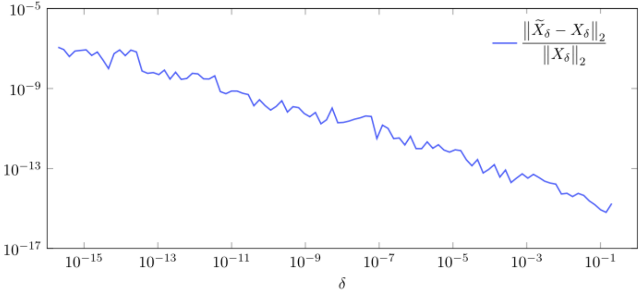


Fig. 2. Relative error of the Schur algorithm for computing the solution of the matrix equation $r(X_\delta) = A_\delta$ in Test 2 with spectrum $\{r_1^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$.

$r(z) = \frac{-z}{z^2+1}$
La funzione $r(z)$ ha valori critici per $r(z) = \pm 0.5$.
Si ha $N = 3$, e la matrice di test è $A = MDM^{-1}$ dove M è matrice casuale, e D è una perturbazione di δ dei valori critici.

Esperimento 2

M. Fasi, B. Iannazzo / Linear Algebra and its Applications 560 (2019) 17–42 41

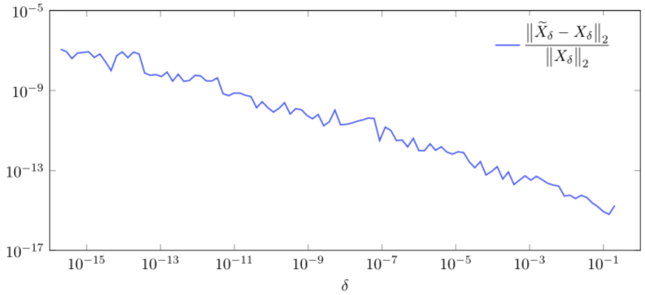
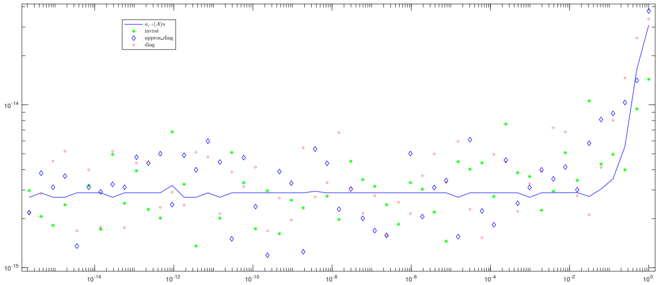


Fig. 2. Relative error of the Schur algorithm for computing the solution of the matrix equation $r(X_\delta) = A_\delta$ in Test 2 with spectrum $\{r_1^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$.

In questo grafico viene rappresentato l'errore relativo di X_δ, confrontando la soluzione calcolata con invrat in aritmetica double, e con precisione di 512 cifre decimali. L'errore aumenta quando gli autovalori sono vicini ai valori critici di r.

Esperimento 2

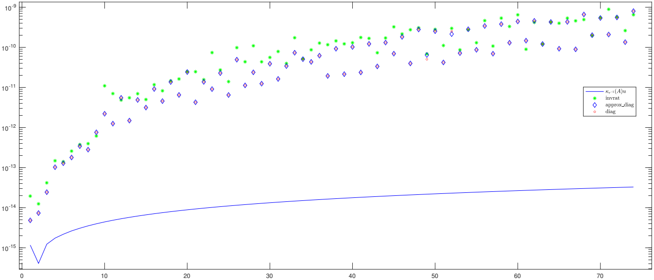


$r(z) = \frac{-z}{z^2+1}$

Si ha $N = 3$, e la matrice di test è $A = MDM^{-1}$ dove M è matrice casuale, e D è una perturbazione di δ dei valori critici.

Il grafico sono rappresentate le norme degli errori $A - r(X_1)$, $A - r(X_2)$, $A - r(X_3)$.

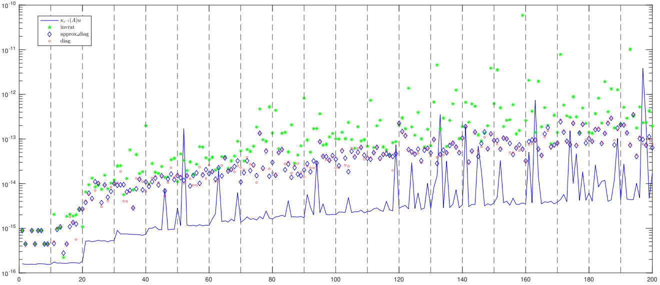
Esperimento 3



$$r(z) = \frac{3z^4 + z^2 - 5z + 1}{z^2 - 3z + 6}$$

Matrice circolante di taglia N da 1 a 74, con le prime cifre di π.
È una matrice normale, e i metodi danno risultati vicini.

Esperimento 4



$r(z) = \frac{z^2+3z-1}{z}; \quad r^{-1}(x) = \frac{1}{2}\left(x - 3 + \sqrt{x^2 - 6x + 13}\right)$

Il ramo scelto per r^{-1} è quello destro. Per ogni N tra 1 e 20, sono generate 10 matrici X con autovalori $>= 3$, viene calcolato $A = r(X)$ e successivamente vengono applicati gli algoritmi per calcolare $X_i = r^{-1}(A)$.

Nel grafico sono rappresentate le norme degli errori $X - X_i$.

Grazie per l'attenzione!

Bibliografia

1. MASSIMILIANO FASI, BRUNO IANNAZZO, *Computing primary solutions of equations involving primary matrix functions*, Linear Algebra Appl. 560 (2019), pp. 17–42.