

# Analisi di un metodo iterativo per sistemi lineari con matrice Toeplitz non simmetrica

## Calcolo Scientifico

Nikita Deniskin

Università di Pisa

22 Marzo 2021

- Si presenta l'articolo di Jennifer Pestana, Andrew Wathen “*A preconditioned MINRES method for nonsymmetric Toeplitz matrices*” (SIAM J. Matrix Anal. Appl. 2015).
- Si vuole analizzare un nuovo metodo iterativo per matrici di Toeplitz non simmetriche. L'idea è di “capovolgere” la matrice, in modo che diventi simmetrica, per poi risolvere il sistema con MINRES.
- Chiameremo il nuovo algoritmo YMINRES. Nella sperimentazione lo vedremo confrontato con GMRES e LSQR per diverse tipologie di matrici, confrontando il numero di iterazioni necessarie per la convergenza.
- Infine si analizzeranno risultati di convergenza del metodo, evidenziando alcune formule errate presenti nell'articolo e calcolando numericamente la stima ottenuta.

## Matrici Toeplitz

$A \in \mathbb{R}^{n \times n}$  è una matrice di Toeplitz se su ogni diagonale le entrate sono uguali:  $A_{ij} = a_{i-j}$  per ogni  $1 \leq i, j \leq n$ .

$$A = \begin{pmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-(n-2)} & a_{-(n-1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_1 & \cdot & \cdot & \cdot & \cdot & a_{-(n-2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & a_{-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n-2} & \cdot & \cdot & \cdot & \cdot & a_{-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n-1} & a_{n-2} & \cdots & a_2 & a_1 & a_0 \end{pmatrix}$$

## Matrice circolante

Una matrice circolante  $C$  di taglia  $n$  è una matrice di Toeplitz, con la proprietà aggiuntiva che la  $k$ -esima sottodiagonale ha le stesse entrate della  $n - k$ -esima sopradiagonale.

$$C = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_2 & c_1 \\ c_1 & \ddots & \ddots & \ddots & \ddots & c_2 \\ c_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & c_{n-2} \\ c_{n-2} & \ddots & \ddots & \ddots & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & \cdots & c_2 & c_1 & c_0 \end{pmatrix}$$

Sia  $\omega_n$  la radice  $n$ -esima dell'unità. Definiamo le seguenti matrici:

- $F$ :  $F_{ij} = \omega_n^{-ij}$ . Fare  $Fx$  corrisponde a `fft(x)` in MATLAB.
- $F^{-1}$ :  $F_{ij} = \frac{1}{n}\omega_n^{ij}$ . Fare  $F^{-1}x$  corrisponde a `ifft(x)` in MATLAB.
- $U = \sqrt{n}F^{-1}$  e  $U^H = \frac{1}{\sqrt{n}}F$  sono matrici unitarie,  $UU^H = I$ .

Sia  $Q$  la seguente matrice circolante:

$$Q = \begin{pmatrix} 0 & & & & 1 \\ 1 & 0 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 & \\ & & & & & & 0 \end{pmatrix} \quad Q = U^H S U = F S F^{-1} \quad (1)$$

dove  $S = \text{diag}(1, \omega_n, \omega_n^2, \dots, \omega_n^{n-1})$ .

Una generica matrice circolante  $C$  si può scrivere come

$$C = \sum_{k=0}^{n-1} c_k Q^k = U^H \left( \sum_{k=0}^{n-1} c_k S^k \right) U \quad (2)$$

## Precondizionatore Strang

Sia  $A$  una matrice Toeplitz, con entrate  $A_{ij} = a_{i-j}$ . Il preconditionatore di Strang  $C^{str}(A)$  relativo ad  $A$  si ottiene mantenendo la diagonali “centrali” di  $A$ , e completando le altre in modo da ottenere una matrice circolante.

## Precondizionatore Strang

Sia  $A$  una matrice Toeplitz, con entrate  $A_{ij} = a_{i-j}$ . Il preconditionatore di Strang  $C^{str}(A)$  relativo ad  $A$  si ottiene mantenendo la diagonali "centrali" di  $A$ , e completando le altre in modo da ottenere una matrice circolante.

$$A = \begin{pmatrix} 1 & 16 & 17 & 18 & 19 \\ 2 & 1 & 16 & 17 & 18 \\ 3 & 2 & 1 & 16 & 17 \\ 4 & 3 & 2 & 1 & 16 \\ 5 & 4 & 3 & 2 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 16 & 17 & 3 & 2 \\ 2 & 1 & 16 & 17 & 3 \\ 3 & 2 & 1 & 16 & 17 \\ 17 & 3 & 2 & 1 & 16 \\ 16 & 17 & 3 & 2 & 1 \end{pmatrix}$$



## Precondizionatore Strang

Sia  $A$  una matrice Toeplitz, con entrate  $A_{ij} = a_{i-j}$ . Il preconditionatore di Strang  $C^{str}(A)$  relativo ad  $A$  si ottiene mantenendo la diagonali “centrali” di  $A$ , e completando le altre in modo da ottenere una matrice circolante.

Nel caso dispari,  $n = 2k + 1$ . Le diagonali che vengono mantenute e poi ricopiate sono quella centrale, le  $k$  sottodiagonali e le  $k$  sopradiagonali.

Nel caso pari,  $n = 2k$ . La diagonale centrale è mantenuta, assieme alle  $k - 1$  sottodiagonali e le  $k - 1$  sopradiagonali. L'ultima diagonale rimasta, la  $k$ -esima, può essere scelta in vari modi, per esempio  $c_k = \frac{1}{2}(a_k + a_{-k})$ .

## Precondizionatore Ottimale

Sia  $A$  matrice generica, e  $U$  la matrice unitaria di Fourier.  
 Definiamo  $\mathcal{M}_U = \{U^H \Lambda U \mid \Lambda \text{ è diagonale}\}$ . Il preconditionatore  
 ottimale per  $A$  è la matrice  $C_U^{opt}(A) \in \mathcal{M}_U$  che risolve il problema  
 di minimo  $\min_{W \in \mathcal{M}_U} \|A - W\|_F$ .

## Precondizionatore Ottimale

Sia  $A$  matrice generica, e  $U$  la matrice unitaria di Fourier. Definiamo  $\mathcal{M}_U = \{U^H \Lambda U \mid \Lambda \text{ è diagonale}\}$ . Il preconditionatore ottimale per  $A$  è la matrice  $C_U^{opt}(A) \in \mathcal{M}_U$  che risolve il problema di minimo  $\min_{W \in \mathcal{M}_U} \|A - W\|_F$ .

### Proprietà:

- $\mathcal{M}_U$  è l'insieme delle matrici circolanti, ma la definizione si può generalizzare per  $U$  unitaria qualsiasi.
- $C_U^{opt}(A) = U^H \text{diag}(UAU^H) U$
- Se  $A$  è Toeplitz, allora  $c_k = \frac{(n-k)a_k + ka_{k-n}}{n}$ , ovvero:

$$C_U^{opt}(A) = \sum_{k=0}^{n-1} \left( \frac{(n-k)a_k + ka_{k-n}}{n} \right) Q^k$$

## Precondizionatore Superottimale

Sia  $A$  matrice generica, e  $U$  la matrice unitaria di Fourier.  
 Definiamo  $\mathcal{M}_U = \{U^H \Lambda U \mid \Lambda \text{ è diagonale}\}$ . Il preconditionatore Superottimale per  $A$  è la matrice  $C_U^{opt}(A) \in \mathcal{M}_U$  che risolve il problema di minimo  $\min_{W \in \mathcal{M}_U} \|I - W^{-1}A\|_F$ .

## Precondizionatore Superottimale

Sia  $A$  matrice generica, e  $U$  la matrice unitaria di Fourier. Definiamo  $\mathcal{M}_U = \{U^H \Lambda U \mid \Lambda \text{ è diagonale}\}$ . Il preconditionatore Superottimale per  $A$  è la matrice  $C_U^{opt}(A) \in \mathcal{M}_U$  che risolve il problema di minimo  $\min_{W \in \mathcal{M}_U} \|I - W^{-1}A\|_F$ .

**Proprietà:** Se  $A$  e  $C_U^{opt}(A)$  non sono singolari, il preconditionatore superottimale esiste ed è unico.

$$\begin{aligned} C_U^{sup}(A) &= C_U^{opt}(AA^H) \left( C_U^{opt}(A^H) \right)^{-1} = \\ &= U^H \text{diag}(UAA^H U^H) \left( \text{diag}(UA^H U^H) \right)^{-1} U \end{aligned} \quad (3)$$

## Algoritmo YMINRES

Sia  $A \in M(n, \mathbb{R})$  matrice di Toeplitz in generale non simmetrica.  
 Si vuole risolvere il sistema  $Ax = b$ .

Sia  $Y$  la matrice:

$$Y = \begin{pmatrix} & & & & 1 \\ & & & \dots & \\ & & & \dots & \\ & & & \dots & \\ 1 & \dots & \dots & \dots & \end{pmatrix}$$

Allora fare prodotto  $Yv$  equivale a capovolgere il vettore  $v$ , non bisogna fare conti. Similmente, anche il prodotto per matrice  $YA$  equivale a capovolgere la matrice.

## Algoritmo YMINRES

$$Ax = b \implies YAx = Yb$$

Se  $A$  è Toeplitz,  $YA$  è Hankel, dunque è simmetrica e possiamo risolvere il sistema applicando MINRES.

## Algoritmo YMINRES

$$Ax = b \implies YAx = Yb$$

Se  $A$  è Toeplitz,  $YA$  è Hankel, dunque è simmetrica e possiamo risolvere il sistema applicando MINRES.

MINRES è un metodo iterativo basato sui sottospazi di Krylov. Non richiede di avere a disposizione la matrice  $A$ , ma è sufficiente una funzione che restituisca il prodotto  $Ax$ .

Anche per l'algoritmo YMINRES, non è necessario formare esplicitamente la matrice  $YA$ . Per calcolare  $YAx$ , si calcola  $Ax$  e poi lo si capovolge.



## Algoritmo YMINRES

$$Ax = b \implies YAx = Yb$$

Se  $A$  è Toeplitz,  $YA$  è Hankel, dunque è simmetrica e possiamo risolvere il sistema applicando MINRES.

MINRES è un metodo iterativo basato sui sottospazi di Krylov. Non richiede di avere a disposizione la matrice  $A$ , ma è sufficiente una funzione che restituisca il prodotto  $Ax$ .

Anche per l'algoritmo YMINRES, non è necessario formare esplicitamente la matrice  $YA$ . Per calcolare  $YAx$ , si calcola  $Ax$  e poi lo si capovolge.

Vedremo che questa implementazione di base non è molto efficiente. Si può migliorare utilizzando dei preconditionatori.

Useremo preconditionatori circolanti simmetrici: Strang, Ottimale, Superottimale.

## Precondizionamento

Dato  $C$  matrice circolante, si può diagonalizzare  $C = U^H D U$ .  
 Definiamo  $|C| = U^H |D| U$ , che è una matrice simmetrica  
 (semi)definita positiva.

## Precondizionamento

Dato  $C$  matrice circolante, si può diagonalizzare  $C = U^H D U$ .

Definiamo  $|C| = U^H |D| U$ , che è una matrice simmetrica (semi)definita positiva.

Si fattorizzi  $|C| = L L^H$  con  $L$  reale, es:  $L = U^H |D|^{\frac{1}{2}} U$ .

$$(L^{-1} Y A L^{-H}) (L^H x) = L^{-1} Y b$$

$$\tilde{A} \tilde{x} = \tilde{b}.$$

## Precondizionamento

Dato  $C$  matrice circolante, si può diagonalizzare  $C = U^H D U$ .  
Definiamo  $|C| = U^H |D| U$ , che è una matrice simmetrica  
(semi)definita positiva.

Si fattorizzi  $|C| = L L^H$  con  $L$  reale, es:  $L = U^H |D|^{\frac{1}{2}} U$ .

$$(L^{-1} Y A L^{-H}) (L^H x) = L^{-1} Y b$$

$$\tilde{A} \tilde{x} = \tilde{b}.$$

$\tilde{A} = (L^{-1} Y A L^{-H})$  resta una matrice simmetrica, per via  
dell'operazione di congruenza.

Utilizzeremo quindi solo preconditionatori circolanti definiti positivi  
per garantire l'esistenza di  $L^{-1}$  e  $L^{-H}$ .

## Sottospazi di Krylov e convergenza di MINRES

$$\tilde{A}\tilde{x} = \tilde{b}.$$

$$\mathcal{K}_k(\tilde{A}, r) = \text{span}(r, \tilde{A}r, \tilde{A}^2r, \dots, \tilde{A}^{k-1}r)$$

$$\mathcal{K}_k^{+1}(\tilde{A}, r) = \mathcal{K}_k(\tilde{A}, \tilde{A}r) = \text{span}(\tilde{A}r, \tilde{A}^2r, \tilde{A}^3r, \dots, \tilde{A}^k r)$$

## Sottospazi di Krylov e convergenza di MINRES

$$\tilde{A}\tilde{x} = \tilde{b}.$$

$$\mathcal{K}_k(\tilde{A}, r) = \text{span}(r, \tilde{A}r, \tilde{A}^2r, \dots, \tilde{A}^{k-1}r)$$

$$\mathcal{K}_k^{+1}(\tilde{A}, r) = \mathcal{K}_k(\tilde{A}, \tilde{A}r) = \text{span}(\tilde{A}r, \tilde{A}^2r, \tilde{A}^3r, \dots, \tilde{A}^k r)$$

MINRES genera una sequenza di iterate  $\tilde{x}_k$  e residui  $\tilde{r}_k$  che soddisfano

$$\tilde{x}_k \in \tilde{x}_0 + \mathcal{K}_k(\tilde{A}, r_0)$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k \quad \tilde{r}_k \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0)$$

$$\tilde{r}_k = \operatorname{argmin} \{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k \quad \tilde{r}_k = \operatorname{argmin}\{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

$$|C| = LL^H \quad \tilde{A} = L^{-1}YAL^{-H} \quad \tilde{x}_k = L^H x_k$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k \quad \tilde{r}_k = \operatorname{argmin}\{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

$$|C| = LL^H \quad \tilde{A} = L^{-1}YAL^{-H} \quad \tilde{x}_k = L^H x_k$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k = L^{-1}Y(b - Ax_k) = L^{-1}Y r_k$$



$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k \quad \tilde{r}_k = \operatorname{argmin}\{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

$$|C| = LL^H \quad \tilde{A} = L^{-1}YAL^{-H} \quad \tilde{x}_k = L^H x_k$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k = L^{-1}Y(b - Ax_k) = L^{-1}Y r_k$$

$$\begin{aligned} \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) &= L^{-1}Y r_0 + \mathcal{K}_k^{+1}(L^{-1}YAL^{-H}, L^{-1}r_0) = \\ &= L^{-1}Y (r_0 + \mathcal{K}_k^{+1}(A|C|^{-1}Y, r_0)) \end{aligned}$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k \quad \tilde{r}_k = \operatorname{argmin}\{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

$$|C| = LL^H \quad \tilde{A} = L^{-1}YAL^{-H} \quad \tilde{x}_k = L^H x_k$$

$$\tilde{r}_k = \tilde{b} - \tilde{A}\tilde{x}_k = L^{-1}Y(b - Ax_k) = L^{-1}Y r_k$$

$$\begin{aligned} \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) &= L^{-1}Y r_0 + \mathcal{K}_k^{+1}(L^{-1}YAL^{-H}, L^{-1}r_0) = \\ &= L^{-1}Y (r_0 + \mathcal{K}_k^{+1}(A|C|^{-1}Y, r_0)) \end{aligned}$$

$$\begin{aligned} r_k = YL\tilde{r}_k &= YL \cdot \operatorname{argmin}\{ \|v\|_2 \mid v \in L^{-1}Y (r_0 + \mathcal{K}_k^{+1}(A|C|^{-1}Y, r_0)) \} \\ &= \operatorname{argmin}\{ \|L^{-1}Yw\|_2 \mid w \in r_0 + \mathcal{K}_k^{+1}(A|C|^{-1}Y, r_0) \} \end{aligned}$$

$r_k$  è il vettore che minimizza

$$\|L^{-1} Y r_k\|_2 = (r_k^H Y L^{-H} L^{-1} Y r_k)^{\frac{1}{2}} = (r_k^H Y |C|^{-1} Y r_k)^{\frac{1}{2}} =$$

$r_k$  è il vettore che minimizza

$$\|L^{-1}Yr_k\|_2 = (r_k^H Y L^{-H} L^{-1} Y r_k)^{\frac{1}{2}} = (r_k^H Y |C|^{-1} Y r_k)^{\frac{1}{2}} =$$

$|C|^{-1}$  è circolante e simmetrica, dunque  $Y|C|^{-1} = |C|^{-1}Y$ .

$$= (r_k^H |C|^{-1} Y^2 r_k)^{\frac{1}{2}} = (r_k^H |C|^{-1} r_k)^{\frac{1}{2}} = \left\| |C|^{-\frac{1}{2}} r_k \right\|_2$$

$r_k$  è il vettore che minimizza

$$\|L^{-1}Yr_k\|_2 = (r_k^H Y L^{-H} L^{-1} Y r_k)^{\frac{1}{2}} = (r_k^H Y |C|^{-1} Y r_k)^{\frac{1}{2}} =$$

$|C|^{-1}$  è circolante e simmetrica, dunque  $Y|C|^{-1} = |C|^{-1}Y$ .

$$= (r_k^H |C|^{-1} Y^2 r_k)^{\frac{1}{2}} = (r_k^H |C|^{-1} r_k)^{\frac{1}{2}} = \left\| |C|^{-\frac{1}{2}} r_k \right\|_2$$

Mentre nell'articolo di riferimento c'è scritto che viene minimizzata

$$\left\| |C|^{-1} r_k \right\|_2 = (r_k^H |C|^{-2} r_k)^{\frac{1}{2}}$$

$$\tilde{r}_k = \operatorname{argmin} \{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

Come si può dare una stima su  $\|\tilde{r}_k\|$ ?

$$\tilde{r}_k = \operatorname{argmin} \{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

Come si può dare una stima su  $\|\tilde{r}_k\|$ ?

$$\|\tilde{r}_k\|_2 = \min_{p \in \Pi_k} \left\| p(\tilde{A})\tilde{r}_0 \right\|_2 \quad \Pi_k = \{p \text{ polinomio, } p(0) = 1, \deg(p) \leq k\}$$

$$\tilde{r}_k = \operatorname{argmin} \{ \|v\|_2 \mid v \in \tilde{r}_0 + \mathcal{K}_k^{+1}(\tilde{A}, \tilde{r}_0) \}$$

Come si può dare una stima su  $\|\tilde{r}_k\|$ ?

$$\|\tilde{r}_k\|_2 = \min_{p \in \Pi_k} \left\| p(\tilde{A})\tilde{r}_0 \right\|_2 \quad \Pi_k = \{p \text{ polinomio, } p(0) = 1, \deg(p) \leq k\}$$

Usando che  $\tilde{A}$  è simmetrica:

$$\|\tilde{r}_k\|_2 \leq \min_{p \in \Pi_k} \left( \max_{\lambda \in \operatorname{Sp}(\tilde{A})} |p(\lambda)| \right) \|\tilde{r}_0\|_2$$



## Campbell, Ipsen, Keller, Meyer (1996)

Data  $A$  matrice,  $\varepsilon > 0$ , e dei punti  $\gamma_p \neq 0$ ,  $1 \leq p \leq P$ , dividiamo gli autovalori  $\lambda$  in  $A$  in gruppi  $\Lambda_p = \text{Sp}(A) \cap \{z \mid |z - \gamma_p| < \varepsilon \gamma_p\}$ . Gli autovalori restanti sono “outliers” e li indichiamo con  $\Lambda_{out}$ . Applicando l’algoritmo GMRES al sistema  $Ax = b$ , la sequenza di residui  $r_k$  ottenuta soddisfa:

$$\frac{\|r_{d+Pk}\|_2}{\|r_0\|_2} \leq M(\sigma^{P-1}\varepsilon)^k$$

$$\delta = \max_p \max_{|z-\gamma_p|=\varepsilon\gamma_p} \max_{\lambda \in \Lambda_{out}} \frac{|\lambda - z|}{|\lambda|}; \quad \sigma = \max_p \max_{|z-\gamma_p|=\varepsilon\gamma_p} \max_{q \neq p} \frac{|\gamma_q - z|}{|\gamma_q|}$$

$$M = P \varepsilon \delta^d \max_p \max_{|z-\gamma_p|=\varepsilon\gamma_p} \|(zI - A_p)^{-1}\|_2$$

$A_p$  è la proiezione di  $A$  sugli autospazi relativi agli autovalori di  $\Lambda_p$ .

- I  $\gamma_p$  sono i centri dei cluster, il raggio è  $\varepsilon\gamma_p$ .
- $\sigma$  è la massima “distanza” tra i vari cluster.
- $\delta$  è la massima “distanza” tra gli outliers e i cluster.
- La stima è valida solo per le iterazioni dopo la  $d$ -esima.
- Il fattore di riduzione è  $\sigma^{P-1}\varepsilon$  ogni  $P$  passi, ovvero  $\sigma^{1-\frac{1}{P}}\varepsilon^{\frac{1}{P}}$  ad ogni passo.
- Per via del termine  $\|(zI - A_p)^{-1}\|_2$ , i cluster non devono essere troppo “stretti” attorno agli autovalori interni  $\Lambda_p$ .
- La presenza di  $|\lambda|$  al denominatore in  $\delta$  indica che gli autovalori non devono essere troppo vicini a 0

## Campbell, Ipsen, Keller, Meyer (1996) - Versione specializzata

Data  $A$  Toeplitz,  $C$  preconditionatore circolante non singolare e  $\varepsilon > 0$ , dividiamo gli autovalori di  $A|C|^{-1}Y$ . Ci sono  $p$  autovalori  $\lambda$  tali che  $|\lambda - 1| < \varepsilon$  e  $q$  tali che  $|\lambda + 1| < \varepsilon$ . Chiamiamo i restanti  $d = n - p - q$  "outliers", e li denotiamo come  $\Lambda_{out}$ .

Applicando l'algoritmo YMINRES al sistema  $Ax = b$  con preconditionatore circolante  $|C|$ , la sequenza di residui  $r_k$  soddisfa:

$$\frac{\| |C|^{-\frac{1}{2}} r_{d+2k} \|_2}{\| |C|^{-\frac{1}{2}} r_0 \|_2} \leq M (2\varepsilon + \varepsilon^2)^k$$

$$\delta = \max_{\gamma=\pm 1} \max_{|z-\gamma|=\varepsilon} \max_{\lambda \in \Lambda_{out}} \frac{|\lambda - z|}{|\lambda|}$$

$$M = 2\varepsilon \delta^d \max_{\gamma=\pm 1} \max_{\lambda \in \Lambda_p} \frac{1}{\varepsilon - |\lambda - \gamma|}$$

- Per passare dalla formulazione particolare a quella generale, si considera che YMINRES applicato ad  $A$  con preconditionatore  $|C|$  è equivalente a MINRES/GMRES applicato a  $L^{-1}YAL^{-H}$ .
- Nell'articolo di riferimento presenta questa stima con  $\| |C|^{-1}r_{d+2k} \|_2$  al posto di  $\| |C|^{-\frac{1}{2}}r_{d+2k} \|_2$
- $\delta$  diventa grande se ci sono autovalori  $\lambda$  di  $A|C|^{-1}Y$  prossimi a 0.
- Gli autovalori di  $A|C|^{-1}Y$  sono raggruppati molto vicino a  $\pm 1$ , che euristicamente garantisce il buon funzionamento delle stime.
- La stima ci garantisce una convergenza lineare. Sperimentalmente però è un superlineare.
- La stima è utile dal punto di vista teorico, ma numericamente la stima è molto larga, come vedremo nell'ultimo esperimento.

## Sperimentazione

- Sono stati fatti sette esperimenti, ognuno con una tipologia di matrice Toeplitz per diversi valori di  $n$ , da 10 a 10000. Vedremo i risultati di tre esperimenti.
- Per ciascun esperimento, si mettono a confronto i risultati ottenuti con diversi algoritmi (YMINRES, GMRES e LSQR) e diversi preconditionatori.
- Utilizziamo gli stessi valori dell'articolo di riferimento: tolleranza  $10^{-8}$ , MAXIT=1000.
- La prima matrice è sparsa e ben condizionata, la seconda è sparsa e mal condizionata, la terza è piena.
- Sono indicati con un trattino "-" i casi in cui non si raggiunge la tolleranza entro MAXIT iterazioni, mentre sono evidenziati in rosso casi con instabilità numerica.

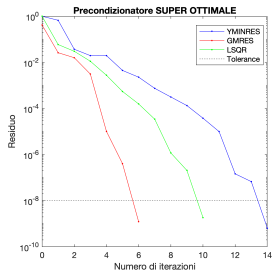
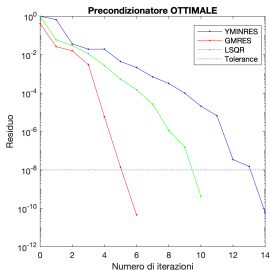
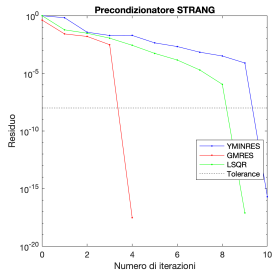
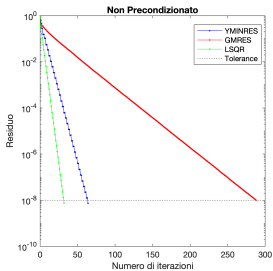


Numero di iterazioni per metodo:

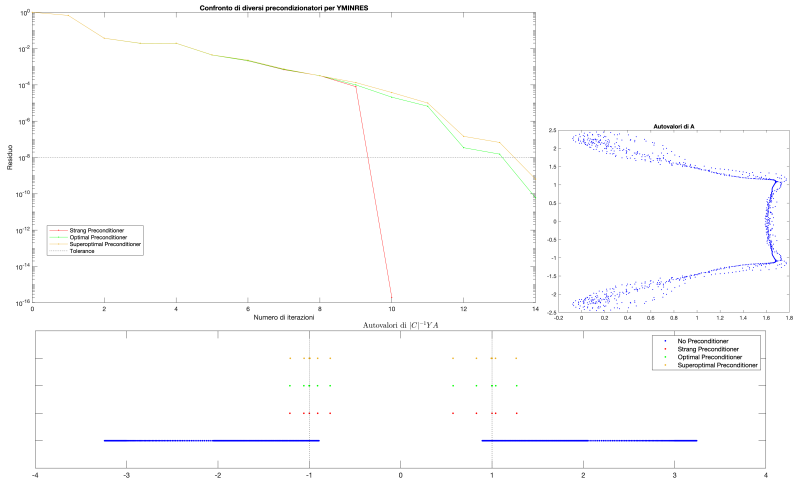
Non preconditionato, Strang				Ottimale, Superottimale			
$n$	YMINRES	GMRES	LSQR	$n$	YMINRES	GMRES	LSQR
10	10	10	20	10	10	10	20
100	58	81	62	100	16	7	22
500	64	255	64	500	14	6	20
1000	64	288	64	1000	14	6	20
2000	64	289	64	2000	12	6	20
5000	64	290	64	5000	12	5	20
10000	64	291	64	10000	12	5	20
10	10	5	18	10	10	10	20
100	10	4	18	100	17	8	22
500	10	4	18	500	14	7	20
1000	10	4	18	1000	14	6	20
2000	10	4	18	2000	14	6	20
5000	10	4	18	5000	12	6	20
10000	10	4	18	10000	12	6	20

Numero di condizionamento di  $A_n$ :

$n$	10	100	500	1000	2000	5000	10000
$\kappa(A_n)$	2.89	3.59	3.63	3.63	3.63	3.63	3.63









Numero di iterazioni per metodo:

Non preconditionato, Strang			
$n$	YMINRES	GMRES	LSQR
10	10	10	20
100	144	80	316
500	870	415	1772
1000	-	861	-
2000	-	-	-
5000	-	-	-
10000	-	-	-

10	10	5	14
100	10	5	16
500	10	5	20
1000	10	5	22
2000	12	5	26
5000	32	10	8
10000	-	6	8

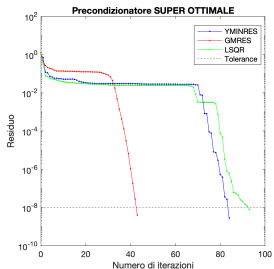
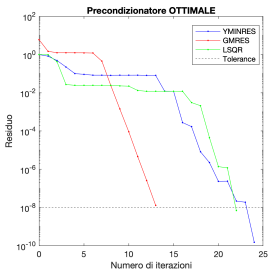
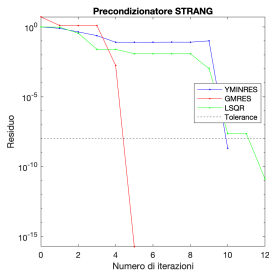
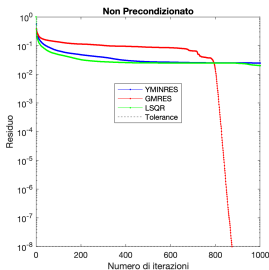
Ottimale, Superottimale			
$n$	YMINRES	GMRES	LSQR
10	10	10	18
100	12	11	22
500	18	12	36
1000	24	13	44
2000	28	15	60
5000	82	19	16
10000	-	22	16

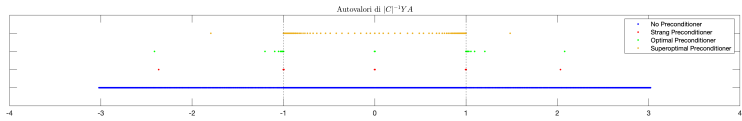
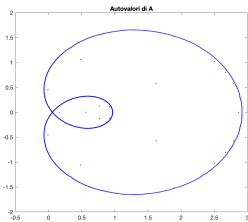
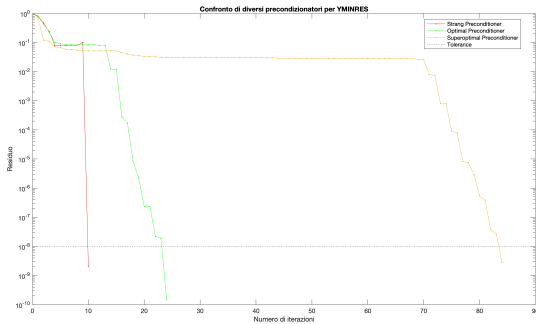
  

10	10	10	20
100	34	20	62
500	65	34	126
1000	84	43	194
2000	112	55	362
5000	258	77	130
10000	26	81	82

 Numero di condizionamento di  $A_n$ :

$n$	10	100	500	1000	2000	5000	10000
$\kappa(A_n)$	12.4	112	537	$1.10 \cdot 10^3$	$2.21 \cdot 10^3$	$5.52 \cdot 10^3$	$1.10 \cdot 10^4$





## Esperimento 3

$A$  ha simbolo  $f(x) = |x|e^{ix}$  con  $x \in [-\pi, \pi]$ . L'entrata  $a_k$  è il  $k$ -esimo coefficiente di Fourier di  $f(x)$ , relativo a  $e^{ikx}$ :  
È una matrice piena.

$$a_k = \begin{cases} \frac{-2}{\pi(k-1)^2} & \text{per } k \text{ pari} \\ 0 & \text{per } k \neq 1 \text{ dispari} \\ \frac{\pi}{2} & \text{per } k = 1 \end{cases}$$

Numero di iterazioni per metodo:

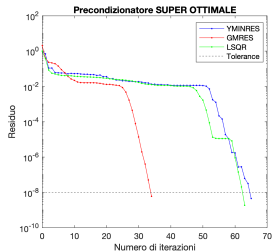
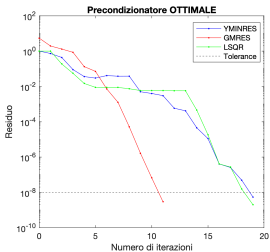
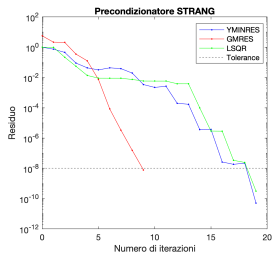
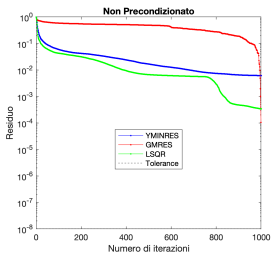
Non preconditionato, Strang				Ottimale, Superottimale			
$n$	YMINRES	GMRES	LSQR	$n$	YMINRES	GMRES	LSQR
10	10	10	20	10	10	9	20
100	124	100	272	100	18	11	32
500	721	500	1470	500	19	10	38
1000	-	-	-	1000	19	10	38
2000	-	-	-	2000	19	11	36
5000	-	-	-	5000	19	10	42
10000	-	-	-	10000	19	10	42

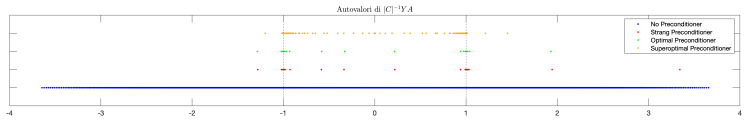
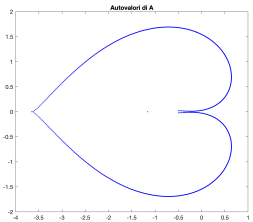
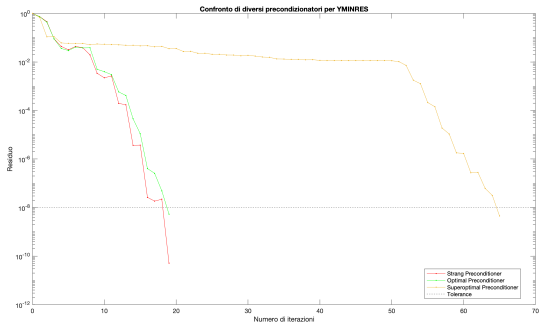
10	10	7	20	10	10	10	20
100	16	9	30	100	28	17	48
500	17	9	36	500	51	28	96
1000	19	9	38	1000	65	34	126
2000	19	9	38	2000	83	42	178
5000	19	9	42	5000	121	60	358
10000	19	9	42	10000	173	77	538

 Numero di condizionamento di  $A_n$ :

$n$	10	100	500	1000	2000	5000	10000
$\kappa(A_n)$	21.3	113	499	991	$1.98 \cdot 10^3$	$4.92 \cdot 10^3$	$9.81 \cdot 10^3$







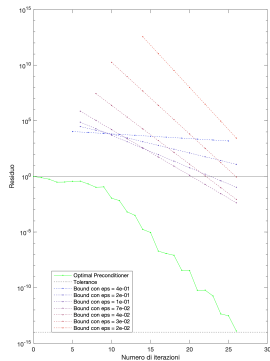
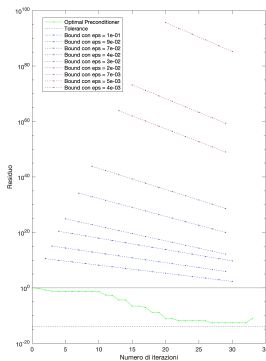
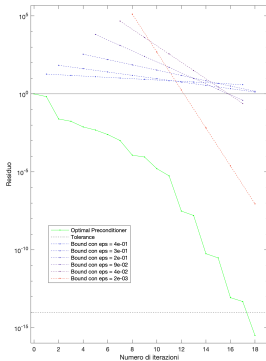
## Commenti generali

- Le versioni non preconditionate sono molto lente nella convergenza, mentre per quelle preconditionate basta qualche decina di iterazioni
- Ripetendo uno stesso esperimento 10 volte variando il termine noto  $b$ , abbiamo visto che le versioni preconditionate hanno varianza bassa, richiedendo in circa metà dei casi sempre lo stesso numero di iterazioni, e nell'altra metà variando solo di 1-2 iterazioni.
- Nel caso preconditionato, il numero di iterazioni richieste viene circa uguale a quello presentato nell'articolo. Curiosamente, per GMRES spesso la differenza era di esattamente un'unità.
- L'algoritmo GMRES è quello che da risultati migliori.

## Commenti generali

- Il preconditionatore migliore è Strang, per i sei esperimenti (tra cui 1 e 2) in cui la matrice è sparsa. Nell'esperimento 3, ha lo stesso comportamento del preconditionatore Ottimale. In uno degli esperimenti però, per  $n$  dispari preconditionando con Strang convergeva velocemente, invece per  $n$  pari il preconditionatore era singolare e il sistema non convergeva.
- Gli autovalori di  $A|C|^{-1}Y$  sono raggruppati attorno a  $\pm 1$  per  $C$  Ottimale e Strang. Per superottimale sono più distribuiti in  $(-1, 1)$ .
- Negli esperimenti 2 e 3 notiamo un andamento a scalini: tratti orizzontali di convergenza lenta, seguiti da tratti verticali di convergenza veloce.

## Esperimento 4



Rappresentazioni della stima  $\frac{\|r_{d+2k}\|_2}{\|r_0\|_2} \leq M(2\varepsilon + \varepsilon^2)^k$  per tre tipologie di matrici ( $n = 1000$ ) e preconditionatore ottimale, al variare di  $\varepsilon$ .

- Le stime sono tutte lineari, mentre la convergenza è leggermente superlineare.
- $\varepsilon$  non può essere scelto troppo piccolo, altrimenti il numero di outliers diventa enorme.
- La costante  $M$  è troppo grande per avere una buona stima della convergenza.
- Tuttavia, per ogni  $\varepsilon$  e relativo  $d$ , sembra che la riduzione al passo  $r_d$  sia circa di un fattore  $\varepsilon$ : le rette sono approssimativamente parallele alla tangente al grafico in  $d$ , e la loro inclinazione aumenta gradualmente.

Grazie per l'attenzione!

## Bibliografia

- 1 J. PESTANA, A.J. WATHEN, *A preconditioned MINRES method for nonsymmetric Toeplitz matrices*, SIAM J. Matrix Anal. Appl. Vol. 36, No. 1, pp. 273–288 (2015)
- 2 S. L. CAMPBELL, I. C. IPSEN, C. T. KELLEY, C. D. MEYER,  $\beta$  *GMRES and the minimal polynomial*, BIT Numerical Mathematics Vol. 36, No. 4, pp. 664-675. (1996)
- 3 H. ELMAN, D. SILVESTER, A. WATHEN, *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, Oxford University Press, New York, 2005.
- 4 R. H.-F. CHAN, X.-Q. JIN, *An Introduction to Iterative Toeplitz Solvers*, SIAM (2007)