Decadimento esponenziale delle inverse di matrici a banda

Silvio Martinico

Sommario

In questa relazione analizzeremo, con dei test, il decadimento esponenziale delle inverse di particolari matrici a banda; vedremo due diversi metodi, uno che sfrutta l'iterazione omografica e che lavora con matrici tridiagonali a blocchi ed un altro che invece sfrutta la teoria spettrale ed un risultato di Chebyshev sull'approssimazione di $(x-a)^{-1}$ tramite polinomi e che lavora con matrici a banda.

1 Iterazione omografica e bounds per le inverse di matrici a banda

Seguendo la presentazione data da Dedieu in [1], consideriamo una matrice di dimensione $mn \times mn$ tridiagonale a blocchi:

$$S_n = \begin{bmatrix} A_1 & C & & & & \\ B & A & C & & & \\ & B & A & C & & \\ & & \ddots & \ddots & \ddots & \\ & & & B & A & C \\ & & & & B & A_\infty \end{bmatrix},$$

dove A_1, A_∞, A, B, C sono matrici $m \times m$ invertibili.

Vogliamo mostrare che esistono delle costanti $K > 0$ ed $r \in (0,1)$ indipendenti da n tali che:

$$||S^{-1}_{n,i,j}|| \leq Kr^{|i-j|} \qquad i,j = 1,...,n \ ,$$

dove $S_{n,i,j}$ indica il blocco (i,j) di S_n.
La dimostrazione di questo fatto si articola in tre step.

1.1 Convergenza dell'iterazione omografica

Definiamo intanto l'iterazione omografica:

$$\begin{cases} U_{i+1} = A - BU_i^{-1}C \\ U_1 = A_1 \end{cases} \ . \tag{1}$$

Siano $S_{k,n}(A_1)$ il minore principale $k \times k$ di S_n, per $k = 1,...,n-1$, $||\cdot||$ una norma di matrici che coincide con il raggio spettrale per le matrici diagonali e $P(\lambda) := \lambda^2 B - \lambda A + C$, $Q(\lambda) := \lambda^2 C^T - \lambda A^T + B^T$ dei polinomi matriciali.

Teorema 1.1. *Supponiamo valgano le seguenti condizioni:*

i) Per ogni n e per ogni $k < n$, $S_{k,n}(A_1)$ è invertibile;

ii) L'equazione matriciale $BX^2 - AX + C = 0$ ammette due soluzioni Y, Z invertibili e tali che:

$$\max\{|\lambda| : \lambda \in Spec(Y)\} \ < \ \min\{|\lambda| : \lambda \in Spec(Z)\} \ ;$$

iii) Le matrici $CY^{-1} - A_1$ e $CZ^{-1} - A_1$ sono invertibili.

Allora l'iterazione (1) converge a CY^{-1} ed esistono $K > 0$ ed $r \in (0,1)$, indipendenti da i, tali che:

$$||U_i - CY^{-1}|| \leq Kr^i \qquad \textit{per ogni } i \ .$$

Osservazione 1.1. Supponiamo che l'equazione $\det(P(\lambda)) = 0$ abbia $2m$ soluzioni λ_i con:

$$0 < |\lambda_1| \leq ... \leq |\lambda_m| < |\lambda_{m+1}| \leq ... \leq |\lambda_{2m}| \ .$$

Quindi, per ogni i, $det(P(\lambda_i)) = 0$, il che implica che esistano $2m$ vettori u_i tali che per ogni i $u_i \neq 0$ e $P(\lambda_i)u_i = 0$. Supponiamo che $u_1, ..., u_m$ e $u_{m+1}, ..., u_{2m}$ siano linearmente indipendenti. Allora le matrici Y e Z, definite rispettivamente dalle coppie di autovalori ed autovettori $(\lambda_1, u_1), ..., (\lambda_m, u_m)$ e $(\lambda_{m+1}, u_{m+1}), ..., (\lambda_{2m}, u_{2m})$, soddisfano la condizione ii).

Osservazione 1.2. Se $A_1 = A$ vale la condizione iii). Infatti, ricordando che Y e Z sono soluzioni di $BX^2 - AX + C = 0$, moltiplicando per X^{-1} a destra si ha $CX^{-1} = A - BX$ e quindi $CX^{-1} - A_1 = -BX[I + X^{-1}B^{-1}(A_1 - A)]$ che è invertibile se $A_1 = A$ e, più in generale, se il raggio spettrale di $X^{-1}B^{-1}(A_1 - A)$ è minore di 1.

Queste osservazioni saranno importanti per l'implementazione.

1.2 Fattorizzazione di S_n

Vediamo adesso come fattorizzare S_n come prodotto di tre matrici; questa fattorizzazione ci servirà poi per trarre delle conclusioni sui blocchi di S_n^{-1}.

Siano:

a) $U_1 = A_1$;

b) $U_i = A - BU_{i-1}^{-1}C$ per $i = 2, ..., n-1$;

c) $U_n = A_\infty - BU_{n-1}^{-1}C$;

d) $V_i = BU_i^{-1}$;

e) $W_i = U_i^{-1}C$.

Consideriamo adesso le matrici $mn \times mn$ a blocchi U, V e W e siano U_{ij}, V_{ij} e W_{ij} i rispettivi blocchi di dimensione $m \times m$ e poniamo:

$$V_{ij} = \begin{cases} I & \text{se } i = j \\ V_j & \text{se } i = j+1 \\ 0 & \text{altrimenti} \end{cases} \quad W_{ij} = \begin{cases} I & \text{se } i = j \\ W_i & \text{se } j = i+1 \\ 0 & \text{altrimenti} \end{cases} \quad U_{ij} = \begin{cases} U_i & \text{se } i = j \\ 0 & \text{altrimenti} \end{cases}$$

Allora abbiamo il seguente teorema:

Teorema 1.2. *Se per ogni $k = 1, ..., n-1$ la sottomatrice $S_{k,n}(A_1)$ è invertibile, allora:*

$$S_n = VUW \ .$$

ed il corollario:

Corollario. *Sotto le ipotesi del teorema abbiamo:*

$$S^{-1}_{n,i,j} = \sum_{\max(i,j)\leq k} (-1)^{i+j} W_i W_{i+1} \cdots W_{k-1} U_k^{-1} V_{k-1} V_{k-2} \cdots V_j \ ,$$

con la convenzione che $V_{k-1}V_{k-2}\cdots V_j = I$ per $k = j$ e $W_iW_{i+1}\cdots W_{k-1} = I$ per $k = i$.

Sia la dimostrazione del teorema che quella del corollario sono dei semplici calcoli con matrici a blocchi.

Adesso, grazie a questi risultati preliminari, vedremo il risultato principale:

1.3 Bound per l'inversa di S_n

Notiamo innanzitutto che $P(\lambda) = 0 \iff Q(\lambda^{-1}) = 0$, infatti:

$$\lambda^2 Q(\lambda^{-1}) = \lambda^{-2} C^T - \lambda^{-1} A^T + B^T = P(\lambda)^T \ .$$

Teorema 1.3. *Supponiamo che valgano le seguenti condizioni:*

i) Per ogni $k < n$, $S_{k,n}(A_1)$ è invertibile;

ii) Le equazioni matriciali $P(X) = 0$ e $Q(X) = 0$ abbiano due soluzioni, rispettivamente Y, Z e Y_1, Z_1, tali che:

$$\lambda_m = \max\{|\lambda| : \lambda \in Spec(Y)\} < \min\{|\lambda| : \lambda \in Spec(Z)\} \, ,$$

$$\mu_m = \max\{|\lambda| : \lambda \in Spec(Y_1)\} < \min\{|\lambda| : \lambda \in Spec(Z_1)\} \, ,$$

e vale $\lambda_m, \mu_m < 1$;

iii) $CY^{-1} - A_1$, $CZ^{-1} - A_1$ e $A_\infty - BY$ sono invertibili.

Allora $\forall\ \alpha > \max\{\lambda_m, \mu_m\}\ \exists\ K > 0$ *indipendente da* n *tale che:*

$$||S^{-1}_{n,i,j}|| \leq K\alpha^{|i-j|}, \qquad i,j = 1,...,n\,. \tag{2}$$

Inoltre, se Y *e* Y_1 *sono diagonalizzabili, possiamo prendere* $\alpha = \max\{\lambda_m, \mu_m\}$.

1.4 Implementazione

Notiamo che, conoscendo A, B e C, non è difficile verificare se siamo nelle ipotesi dell'Osservazione 1.1, la quale, come già detto, ci garantisce l'ipotesi *ii*) del teorema. Inoltre, se S_n è una matrice di Toepliz a blocchi, cioè dove $A_1 = A_\infty = A$, grazie all'Osservazione 1.2 abbiamo anche l'ipotesi *iii*). Infine, se S_n è fortemente dominante diagonale, il *primo Teorema di Gerschgorin* ci garantisce che $S_{k,n}$ è invertibile per ogni k e quindi abbiamo anche l'ipotesi *i*). Vediamo quindi una possibile implementazione che sfrutta quanto appena detto:

```
function K = homographic (A_1, A_inf, A, B, C, r, n)

    m = size(A,1);
    S = full(blktridiag(A,B,C,n));
    S(1:m,1:m) = A_1;
    S((n*m-m+1):end,(n*m-m+1):end) = A_inf;

    syms l;
    p = B*l^2 - A*l + C;
    q = (C')*l^2 - (A')*l + B';
    p = det(p);
    q = det(q);
    p1 = sym2poly(p);
    q1 = sym2poly(q);

    l = roots(p1);
    mu = roots(q1);
```

```
    [~,idx] = sort(abs(l)); %ordino gli autovalori
    l = l(idx);

    [~,idx] = sort(abs(mu)); %ordino gli autovalori
    mu = mu(idx);

    K = 0;

    if (abs(l(1)) > 0 && abs(l(m)) < abs(l(m+1)) &&
        abs(mu(1)) > 0 && abs(mu(m)) < abs(mu(m+1)))

        lambda_m = abs(l(m)); %max autovalore di Y
        mu_m = abs(mu(m));    %max autovalore di Y1

        mx = max(lambda_m, mu_m);
        mx = mx + r;

        I = inv(S);
        X = zeros(n);

        for i = 1:n
            for j = 1:n
                bij = I(((i-1)*m+1):i*m, ((j-1)*m+1):j*m);
                X(i,j) = norm(bij)/((mx)^(abs(i-j)));
            end
        end

        K = max(X(:));
    end
end
```

Il primo test è stato eseguito sulla matrice dell'Esempio 4.4 dell'articolo [1]. La matrice S_n utilizzata è una matrice di Toepliz a blocchi, dove:

$$A = A_1 = A_\infty = \begin{bmatrix} -1152 & 720 & -160 \\ 720 & -480 & 120 \\ -160 & 120 & -40 \end{bmatrix}, \qquad B = C^T = \begin{bmatrix} -1 & 12 & -60 \\ 0 & -3 & 30 \\ 0 & 0 & -5 \end{bmatrix}.$$

Il codice di sopra trova il K più grande tale che valga l'equazione (2). Il valore di α è ottenuto grazie ad r che viene sommato al massimo tra λ_m e μ_m per ottenere $\alpha > \max\{\lambda_m, \mu_m\}$ ($= 0.5353$ in questo caso). Eseguendo il codice per diversi valori di n possiamo notare che K si stabilizza e possiamo quindi determinare il suo valore tramite la sperimentazione; riportiamo i risultati in tabella:

r = 0.01	
n	**K**
5	4.47544
8	5.6028
10	5.8693
50	5.9786
100	5.9786
500	5.9786
1000	5.9786

r = 0.1	
n	**K**
5	3.8414
8	4.7324
10	4.8611
50	4.9135
100	4.9135
500	4.9135
1000	4.9135

r = 0.2	
n	**K**
5	3.6320
8	4.2285
10	4.3091
50	4.3419
100	4.3419
500	4.3419
1000	4.3419

r = 0.4	
n	**K**
5	3.632
8	4.2285
10	4.3091
50	4.3419
100	4.3419
500	4.3419
1000	4.3419

Per il secondo tentativo sperimentale usiamo una matrice con blocchi di dimensione 1, quindi una matrice tridiagonale, che sia anche una matrice di Toepliz. Prendiamo:

$$A = A_1 = A_\infty = 12, \qquad B = 4, \qquad C = 5\,.$$

Abbiamo che $P(\lambda) = 4\lambda^2 - 12\lambda + 5$ e, dato che la dimensione dei blocchi è 1, $\det(P(\lambda)) = P(\lambda)$ e $P(\lambda) = 0$ ha soluzioni $\lambda_{1,2} = \frac{1}{2}, \frac{5}{2}$. Infine S_n è fortemente dominante diagonale, quindi abbiamo tutte le ipotesi del Teorema 1.3. Vediamo i risultati della sperimentazione:

r = 0.1	
n	**K**
3	0.1154
5	0.1230
8	0.1248
10	0.1250
50	0.1250
100	0.1250
500	0.1250
1000	0.1250

r = 0.2	
n	**K**
3	0.1154
5	0.1230
8	0.1248
10	0.1250
50	0.1250
100	0.1250
500	0.1250
1000	0.1250

1.5 Considerazioni

Dai risultati in tabella possiamo osservare intanto che, fissato α, esiste davvero un K indipendente da n per il quale vale (2) e sembra che possiamo prendere il K che si stabilizza per $n \geq 50$.

Possiamo inoltre notare che, per certi valori di r (nel primo caso per $r \geq 2$, nel secondo per ogni r provato), i valori di K sono sempre gli stessi e non dipendono quindi nemmeno da α. Questo accade perché il K massimo si ottiene in corrispondenza dei blocchi sulla diagonale principale, quindi quando $i = j$ e quindi quando in (2) abbiamo $\alpha^{|i-j|} = \alpha^0 = 1$, cioè la disuguaglianza (2) non dipende da α (e quindi nemmeno da r).

2 Decadimento delle inverse di matrici a banda

Anche nell'articolo [2] vedremo il decadimento degli elementi di inverse di matrici a banda man mano che ci allontaniamo dalla diagonale principale, usando però un approccio diverso.

Definizione 2.1. Data una matrice quadrata A di dimensione n diremo che è una matrice a banda di ampiezza m se vale:

$$|i-j| > m \implies A_{ij} = 0\,.$$

Chiaramente chiediamo che $m < n-1$, altrimenti qualsiasi matrice potrebbe essere considerata una matrice a banda di ampiezza $\geq n-1$.

L'idea è quella di utilizzare intanto la teoria spettrale per scrivere:

$$||A^{-1} - p(A)|| = \max_{x \in \sigma(A)} |1/x - p(x)|\,,$$

dove A è una matrice definita positiva e p è un polinomio a coefficienti reali, e successivamente utilizzare la teoria dell'approssimazione per stimare l'errore. In realtà possiamo studiare anche il caso in cui A non è definita positiva, usando al suo posto gli autovalori di AA^H che saranno positivi.

La norma utilizzata è quella euclidea $||\cdot||_2$.

Vediamo direttamente il risultato che ci interessa:

Teorema 2.1. *Siano A una matrice invertibile a banda di ampiezza m ed $[a,b]$ il più piccolo intervallo contenente lo spettro di AA^H. Posti $r := b/a$, $q := \frac{\sqrt{r}-1}{\sqrt{r}+1}$ e $\lambda_1 := q^{\frac{1}{m}}$, esiste una costante C_1 dipendente da A tale che:*

$$|A^{-1}(i,j)| \leq C_1 \lambda_1^{|i-j|}\,.$$

Possiamo prendere $C_1 = (m+1)\cdot||A||_2 \cdot q^{-1} \cdot C(a,r)$, dove $C(a,r) := \max\{a^{-1}, C_0\}$ e $C_0 := \frac{(1+\sqrt{r})^2}{2ar}$.

2.1 Implementazione

Vediamo una possibile implementazione in cui calcoliamo il bound C_1 dato dal Teorema 2.1 e lo confrontiamo con il minimo bound possibile:

```
function [K, C_1] = decay (A, m)
    %K è il bound effettivo, C_1 quello calcolato con
    % la formula data dal Teorema 2.1
    n = size(A,1);

    l = eig(A*A');

    a = min(l);
    b = max(l);

    r = b/a;

    q = (sqrt(r)-1)/(sqrt(r)+1);

    lambda_1 = q^(1/m);

    C_0 = ((1 + sqrt(r))^2)/(2*a*r);

    C = max(1/a, C_0);

    C_1 = (m+1)*norm(A)*(1/q)*C;

    I = inv(A);
    X = zeros(n);

    for i = 1:n
        for j = 1:n
            X(i,j) = I(i,j)/(lambda_1^(abs(i-j)));
        end
    end
```

```
    K = max(X(:));

end
```

Vediamo i risultati ottenuti dalla sperimentazione con le matrici utilizzate nella sperimentazione precedente, le quali rientrano nelle ipotesi del Teorema 2.1. La matrice S_n i cui blocchi sono:

$$A = A_1 = A_\infty = \begin{bmatrix} -1152 & 720 & -160 \\ 720 & -480 & 120 \\ -160 & 120 & -40 \end{bmatrix}, \qquad B = C^T = \begin{bmatrix} -1 & 12 & -60 \\ 0 & -3 & 30 \\ 0 & 0 & -5 \end{bmatrix},$$

è una matrice a banda con banda di ampiezza $m = 6$.

Vediamo cosa succede per vari valori di n, dove n questa volta indica la dimensione di S_n:

n	$\mathbf{C_1}$	**K**
6	$2.6074 \cdot 10^4$	0.8977
12	$3.7128 \cdot 10^5$	1.9550
24	$1.7654 \cdot 10^6$	2.5812
48	$3.1318 \cdot 10^6$	2.6449
96	$3.6752 \cdot 10^6$	2.6453
498	$3.8723 \cdot 10^6$	2.6453
1002	$3.8782 \cdot 10^6$	2.6453

Considerando invece la matrice S_n i cui blocchi sono:

$$A = A_1 = A_\infty = 12, \qquad B = 4, \qquad C = 5\,,$$

andiamo a vedere i risultati ottenuti dalla sperimentazione:

n	C_1	K
5	5.1431	0.1230
10	7.5705	0.1250
20	8.7805	0.1250
50	9.2350	0.1250
100	9.3081	0.1250
500	9.3323	0.1250
1000	9.3331	0.1250

2.2 Considerazioni e confronto con i bounds forniti dal metodo precedente

Salta subito all'occhio dalle tabelle, soprattutto dalla prima, la differenza (anche di diversi ordini di grandezza) tra il parametro C_1 calcolato dall'algoritmo ed il valore davvero richiesto. Il prezzo da pagare per conoscere a priori questo valore è quindi la precisione di esso.

Le costanti moltiplicative tuttavia non ci dicono esattamente come si comportano i due algoritmi in generale, in quanto la velocità di convergenza, soprattutto per matrici grandi, è data principalmente dai valori α e λ_1. Vediamo infatti adesso con altri test di analizzare più a fondo la situazione e di confontare i due algoritmi.

2.3 Ulteriori sperimentazioni e confronti

Vediamo qualche altra possibile sperimentazione. Una cosa interessante da provare è quella di utilizzare l'algoritmo di $Demko-Moss-Smith$ senza conoscere però gli autovalori di AA^H a priori. Infatti, in applicazioni reali, dove solitamente si lavora con matrici di grandi dimensioni, il calcolo esplicito degli autovalori è molto dispendioso e renderebbe poco utile l'algoritmo. Per questo proveremo adesso ad utilizzare il Primo Teorema di Gershgorin per dare dei bound agli autovalori di AA^H e confronteremo i risultati ottenuti con questo metodo con i risultati ottenuti con il calcolo effettivo degli autovalori. C'è da osservare che,

dovendo calcolare $\sqrt{r}$, dove $r = \frac{b}{a}$ e $[a, b]$ è un intervallo che contiene gli autovalori di AA^H, vorremmo che b ed a fossero di segno concorde, perché alla fine avremo una disuguaglianza che ha al RHS $C_1 \cdot \lambda^{|i-j|}$ e, sia C_1 che λ dipendono da $\sqrt{r}$, quindi vogliamo che questo valore sia reale, altrimenti ha poco senso parlare di disuguaglianze sul campo complesso (potremmo passare ai moduli ma bisognerebbe comunque rivedere l'intero teorema e non è detto che si riesca ad ottenere la disuguaglianza voluta). Nel caso del calcolo esplicito degli autovalori non avevamo problemi perché AA^H è una matrice simmetrica e definita positiva (in generale è semidefinita positiva, inoltre noi sappiamo per ipotesi che A è invertibile) e quindi i suoi autovalori sono reali e positivi. Nel caso della stima con il Teorema di Gershgorin invece, dobbiamo assicurarci che AA^H sia fortemente dominante diagonale, in modo che le stime degli autovalori non prendano valori di segno opposto tra loro. Dato che AA^H è definita positiva, sappiamo già che $a > 0$, però, non riuscendo a dare un bound positivo dal basso a priori, non possiamo nemmeno usare $a = 0$ perché dobbiamo poi dividere b per a e quindi ci serve $a > 0$, quindi l'unico modo per garantirci un a positivo tramite le stime con il Teorema di Gershgorin è prendere A in modo che AA^H sia fortemente dominante diagonale.

Passiamo a vedere i risultati della sperimentazione. La prima matrice utilizzata è una matrice S tridiagonale 20×20, quindi, vedendola con la solita rappresentazione, avremo:

$$A_1 = A_\infty = A = 50, \quad B = 4, \quad C = 5\,.$$

Andremo ad analizzare il decadimento degli elementi sulla parte destra della seconda riga su un grafico che ha sull'asse delle ascisse la differenza $|i-j|$, cioè la distanza dalla diagonale, e sull'asse delle ordinate il valore, in scala logaritmica, $C_1 \cdot \lambda_1^{|i-j|}$ calcolato sia con i valori approssimati che con i valori esatti degli autovalori.

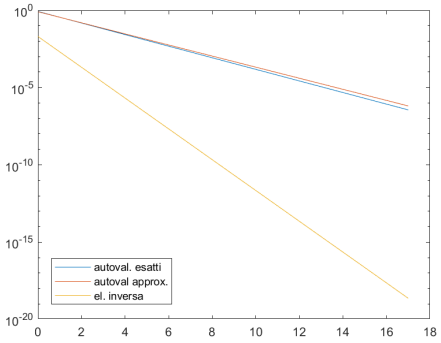


Figura 1: Terza semi-riga destra

In questo caso i valori di a e b sono: $a = 1689.273$, $b = 3469.169$ e $a = 1601$, $b = 3481$ rispettivamente nel caso del calcolo esatto e nel caso dell'approssimazione e i valori di C_1 e λ_1 ottenuti sono: $C_1 = 0.847$, $\lambda_1 = 0.4219$ e $C_1 = 0.8104$, $\lambda_1 = 0.4379$.

Prima di commentare, vediamo un altro test con una matrice diversa. Questa volta prendiamo una matrice 40×40 tridiagonale a blocchi, con blocchi di dimensione 2. I blocchi della matrice sono:

$$A = A_1 = A_\infty = \begin{bmatrix} 110 & 3 \\ 5 & 70 \end{bmatrix}, \qquad B = C^T = \begin{bmatrix} 8 & 4 \\ 0 & 2 \end{bmatrix}.$$

Andiamo a vedere in un grafico, esattamente come prima, il risultato dei test:

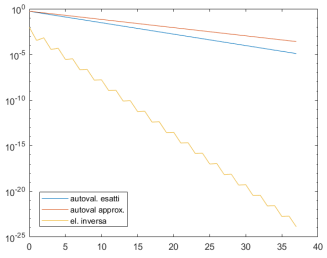


Figura 2: Terza semi-riga destra

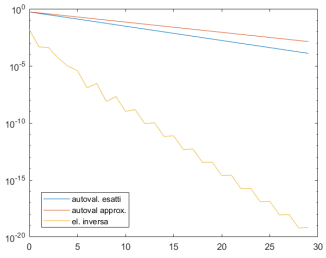


Figura 3: Decima semi-riga destra

In questo secondo caso i valori di a e b sono $a = 4361.24$, $b = 16135.35$ e $a = 2689$, $b = 17577$ rispettivamente nel caso del calcolo esatto e dell'approssimazione, mentre i valori di C_1 e λ_1 sono rispettivamente $C_1 = 0.5325$, $\lambda_1 = 0.7497$ e $C_1 = 0.5397$, $\lambda_1 = 0.8134$. Notiamo che, rispetto alla matrice precedente, questa volta i valori approssimati sono un po' meno fedeli ai valori esatti e questo si vede anche graficamente.

Come possiamo verificare nei tre grafici appena visti, l'approssimazione data dal Primo Teorema di Gershgorin è abbastanza buona, si ha un decadimento esponenziale non molto più lento rispetto a quello ottenuto nel caso del calcolo effettivo degli autovalori, la differenza in tutti i casi è inferiore ai due ordini di grandezza. Tuttavia sembra che le stime ottenute non siano molto aderenti ai valori effettivi della matrice inversa.

Ovviamente quanto appena osservato dipenderà dalle matrici in questione, magari per alcune matrici il Teorema di Gershgorin potrebbe darci delle approssimazioni molto larghe e potrebbe quindi minare la velocità di convergenza.

In questo caso l'algoritmo "decay" è stato modificato nel seguente modo:

```
function [X, I] = decay2 (A, m)

    n = size(A,1);

    %APPLICO IL PRIMO TEOREMA DI GERSHGORIN

    G = A*A';

    sx = zeros(1,n);
    dx = zeros(1,n);

    for i = 1:n
        sx(i) = G(i,i) - sumabs(G(i,1:n)) + abs(G(i,i));
        dx(i) = G(i,i) + sumabs(G(i,1:n)) - abs(G(i,i));
```

```
    end

    a = min(sx);
    b = max(dx);

    r = abs(b/a);
    q = (sqrt(r)-1)/(sqrt(r)+1);
    lambda_1 = (abs(q))^(1/m);
    C_0 = ((1 + sqrt(r))^2)/(2*a*r);
    C = max(1/a, C_0);
    C_1 = (m+1)*norm(A)*(1/q)*C;

    I = inv(A);
    X = zeros(n);

    for i = 1:n
        for j = 1:n
            X(i,j) = C_1*(lambda_1)^(abs(i-j));
        end
    end
end
```

dove i valori che vengono ritornati sono rispettivamente la matrice che ha per componenti il bound ottenuto per ogni componente dell'inversa e, appunto, l'inversa.

Vediamo adesso un confronto grafico tra i due metodi. Non avendo una formula per il parametro K nel caso dell'articolo di Dedieu, ci limiteremo a confrontare solamente i due fattori esponenziali ottenuti con i due metodi, cioè α e λ_1.
Eseguiremo il confronto per due matrici tridiagonali, la prima con componenti:

$$A_1 = A_\infty = A = 12, \quad B = 4, \quad C = 5\,,$$

e la seconda con componenti:

$$A_1 = A_\infty = A = -23, \quad B = 3, \quad C = -8\,.$$

Per l'algoritmo di Dedieu abbiamo scelto il parametro $r = 0.01$. Nei seguenti grafici sull'asse delle ascisse troviamo la distanza dalla diagonale $|i - j|$ e sull'asse delle ordinate rispettivamente $\alpha^{|i-j|}$ per l'algoritmo di Demko e $\lambda_1^{|i-j|}$ per l'algoritmo di Demko-Moss-Smith:

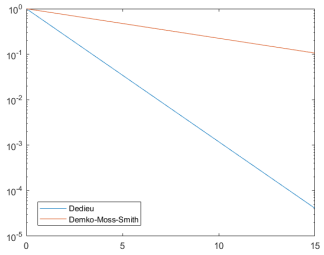


Figura 4: $A = 12$, $B = 4$, $C = 5$

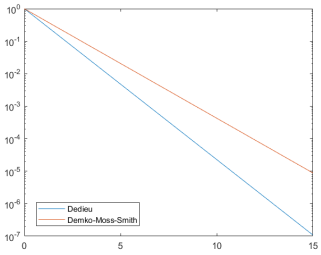


Figura 5: $A = -23$, $B = 3$, $C = -8$

Come possiamo osservare dai grafici, l'algoritmo di Dedieu ha un decadimento esponenziale più veloce, questo fatto è sicuramente dovuto alla maggiore specificità dell'algoritmo che riesce quindi a garantire un bound più aderente alle vere caratteristiche della matrice, mentre l'algoritmo di Demko-Moss-Smith da un lato richiede ipotesi meno stringenti e dall'altro però fornisce un bound più largo.

Anche con matrici simmetriche il comportamento dei due algoritmi non cambia, proviamo infatti a prendere delle matrici tridiagonali in cui $B = C$. La prima matrice avrà i seguenti blocchi di dimensione 1:

$$A_1 = A_\infty = A = 65, \quad B = 8, \quad C = 8\,,$$

mentre la seconda:

$$A_1 = A_\infty = A = 26, \quad B = 5, \quad C = 5\,.$$

Consideriamo questa volta delle matrici più grandi, di dimensione 200×200. Vediamo graficamente cosa succede:

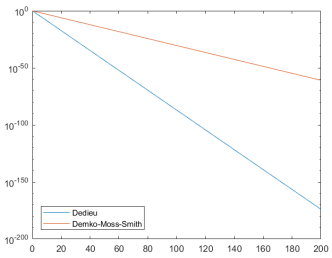


Figura 6: $A = 65$, $B = 8$ $C = 8$

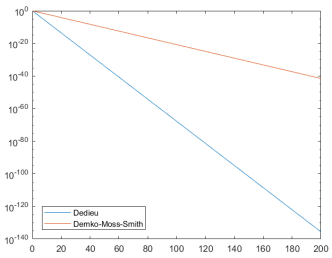


Figura 7: $A = 26$, $B = 5$ $C = 5$

Anche in questo caso sembra che l'algoritmo proposto da Dedieu si comporti meglio.

Un caso molto interessante da studiare è invece quello di matrici definite positive, in quanto l'articolo di Demko-Moss-Smith propone una proposizione apposita per questo caso e quindi sicuramente ci può dare un bound migliore per questo tipo di matrici.
Enunciamo questa proposizione e vediamo poi cosa succede a livello sperimentale.

Proposizione 2.2. *Sia S una matrice a banda di ampiezza m, invertibile e definita positiva. Sia $[a, b]$ il più piccolo intervallo che contiene lo spettro di S. Siano inoltre:*

$$r = \frac{b}{a}, \qquad q = \frac{\sqrt{r}-1}{\sqrt{r}+1}, \qquad C_0 = \frac{(1+\sqrt{r})^2}{2ar}, \qquad \lambda = q^{2m}\,.$$

Allora si ha:

$$|A^{-1}(i,j)| \leq C \cdot \lambda^{|i-j|}\,,$$

dove $C = \max\{a^{-1}, C_0\}$.

Le ultime matrici utilizzate sono definite positive, possiamo quindi usarle e vedere i risultati dei test sotto forma di grafico:

Possiamo vedere come, nel caso di matrici definite positive, la situazione si ribalta. Questo succede perché nella Proposizione 2.2 usiamo gli autovalori di S piuttosto che gli autovalori di SS^T che vengono amplificati dal prodotto.

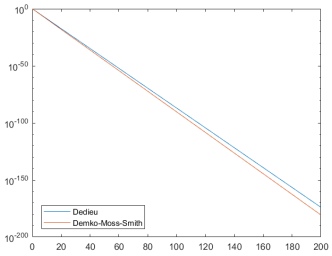


Figura 8: $A = 65$, $B = 8$, $C = 8$

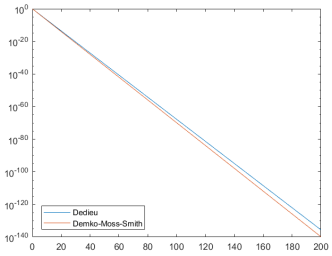


Figura 9: $A = 26$, $B = 5$, $C = 5$

Dato che l'algoritmo proposto da Dedieu ha avuto la peggio, proviamo a prendere un r ancora più piccolo (finora abbiamo utilizzato $r = 0.01$). Proviamo con $r = 0.0001$ nel caso della seconda matrice:

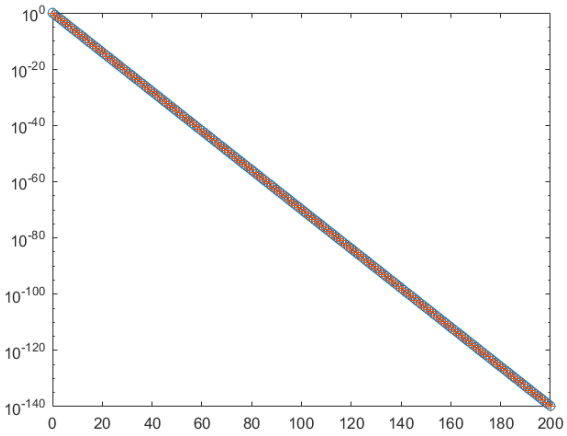


Figura 10: $A = 26$, $B = 5$, $C = 5$, $r = 0.0001$

Vediamo dal grafico che l'algoritmo di Dedieu migliora prendendo valori di r più piccoli, fino a sovrapporsi all'algoritmo di Demko-Moss-Smith. Nel grafico sono stati cambiati i marker delle linee poiché altrimenti non riusciamo a distinguerle perché sono praticamente sovrapposte. Infatti il valore di α ottenuto è 0.2010 mentre $\lambda = 0.2000$.

Per concludere possiamo quindi dire che l'algoritmo di Dedieu si è mostrato net-

tamente superiore nel caso generale, mentre è di poco inferiore nel caso in cui le matrici trattate sono definite positive. Inoltre l'algoritmo di Demko-Moss-Smith ci fornisce anche il parametro C, quindi in caso di prestazioni simili, forse è l'algoritmo più conveniente. Ricordiamo anche però che richiede il calcolo esplicito degli autovalori e che, in generale, l'utilizzo di un loro bound ottenuto tramite il Primo Teorema di Gershgorin mina di molto la velocità di decadimento.

Grazie a questi test ci siamo fatti un'idea del comportamento dei due algoritmi nei vari casi, soprattutto a livello asintotico (quindi concentrandoci principalmente sul parametro che ci dà la convergenza esponenziale piuttosto che sulla costante che lo moltiplica). Come abbiamo già detto però l'algoritmo di Dedieu non ci fornisce un valore del parametro K e non ci fornisce nemmeno un valore esatto per il parametro esponenziale α, ma ci dà solamente un suo limite inferiore. In generale abbiamo visto che conviene prendere unvalore di r più piccolo possibile in modulo, poiché così, al crescere di $|i-j|$ il valore $\alpha^{|i-j|}$ decresce più velocemente. Sorge allora una domanda: cosa succede per matrici piccole? Converrà ancora prendere un valore di r il più vicino a 0? O forse sarà meglio prendendo un valore di r più grande e giocarsela con il parametro K? Vediamo con una sperimentazione cosa succede.

Consideriamo tre diverse matrici: la prima, con blocchi di dimensione 1, è la matrice 8×8 dove:

$$A_1 = A_\infty = A = 12, \quad B = 4, \quad C = 5\,.$$

Riportiamo in tabella i valori di r testati ed i rispettivi K trovati:

r	**K**
$0,5$	$0,0417$
$0,2$	$0,0417$
$0,1$	$0,0417$
$0,01$	$0,0417$

Come possiamo notare dalla tabella, al variare di r, K rimane costante; questo accade perché il valore massimo di K è richiesto in corrispondenza di un ele-

mento diagonale e quindi quando $i = j$, cioè $|i - j| = 0$. Per questo motivo in questo caso la scelta di r non influisce sulla determinazione del parametro K.

Lo stesso vale per la matrice con:

$$A_1 = A_\infty = A = 26, \quad B = C = 5\,.$$

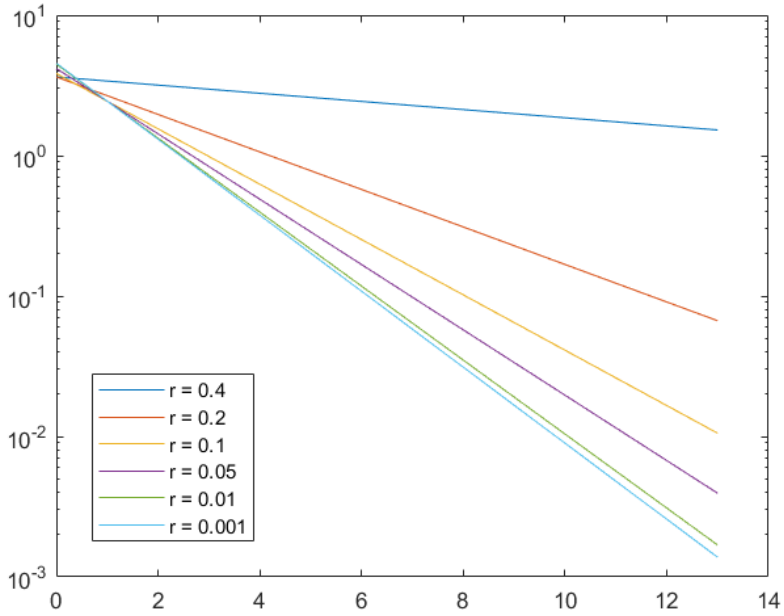
Le cose cambiano invece per la matrice con blocchi di dimensione 3 dell'esempio dell'articolo di Dedieu. Infatti, prendendo questa matrice di dimensione 15×15, abbiamo i seguenti risultati:

r	**K**
$0,4$	$3,632$
$0,2$	$3,632$
$0,1$	$3,8414$
$0,05$	$4,1695$
$0,01$	$4,4754$
$0,001$	$4,55057$

Vediamo adesso cosa succede a livello grafico. Mettiamo sull'asse delle ascisse il valore $|i - j|$ e sull'asse delle ordinate la funzione $K \cdot \alpha^{|i-j|}$ per i vari valori di r utilizzati (dove ricordiamo che α dipende da r).

Come possiamo vedere, anche se in questo caso la scelta di r influisce sul parametro K tanto da fare in modo che per piccoli valori di $|i - j|$ si ribalti la situazione (cioè la convergenza sia più veloce per valori di r più grandi), questo vale solamente per valori di $|i - j|$ davvero piccoli, cioè al più $|i - j| = 3$ e quindi non abbastanza da garantire un vantaggio nella scelta di un r più grande, anche perché la nostra matrice ha dei blocchi di dimensione 3 e quindi la sua dimensione dovrà essere almeno 9 affinchè abbia senso parlare di matrici a banda. Anche in questo caso la scelta migliore sembra quindi quella di prendere r molto piccolo.

In definitiva possiamo dire che per matrici molto piccole questo approccio non sembra vincente poiché o il valore di K viene determinato dagli elementi sulla



diagonale e quindi si può giocare solamente sul parametro r, oppure comunque non sembra particolarmente vantaggioso.

Riferimenti bibliografici

[1] Jean-Pierre Dedieu. "Matrix homographic iterations and bounds for the inverses of certain band matrices". In: *Linear Algebra and its Applications* 111 (1988), pp. 29–42. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(88)90049-3. URL: https://www.sciencedirect.com/science/article/pii/0024379588900493.

[2] Stephen Demko, William F. Moss e Philip W. Smith. "Decay Rates for Inverses of Band Matrices". In: *Mathematics of Computation* 43.168 (1984), pp. 491–499. ISSN: 00255718, 10886842. URL: http://www.jstor.org/stable/2008290.