UNIVERSITÀ DEGLI STUDI DI ROMA TOR VERGATA

MACROAREA DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA SPECIALISTICA IN MATEMATICA

TESI DI LAUREA

MATRICI DI BASSA COMPLESSITÀ COMPUTAZIONALE ED APPLICAZIONI

Relatore:
Chiar.mo Prof.
DI FIORE CARMINE

Laureando:
VESPAN LUCIAN
matricola: 0137606

Anno Accademico 2024/2025

Indice

Introduzione

Questa tesi si propone di esplorare in profondità la teoria delle matrici strutturate, una classe di matrici che, grazie alla loro particolare conformazione interna, permettono di semplificare notevolmente molte operazioni algebriche e numeriche. In particolare, vengono analizzate le loro proprietà algebriche, il comportamento spettrale e le caratteristiche computazionali, con attenzione alle implicazioni teoriche e alle applicazioni pratiche.

Il punto di partenza dell'indagine è rappresentato dalle matrici di Toeplitz, una delle famiglie più studiate di matrici strutturate, caratterizzate dall'invarianza lungo le diagonali: ogni elemento è determinato esclusivamente dalla differenza tra l'indice di riga e di colonna. Questa proprietà consente una rappresentazione compatta e una drastica riduzione della complessità computazionale, ad esempio nella risoluzione di sistemi lineari o nel calcolo di prodotti matrice-vettore.

Oltre alle matrici di Toeplitz, la trattazione si estende ad altre classi di matrici con struttura ricorrente, tra cui:

- le matrici circolanti, che sono Toeplitz con periodicità ciclica, fondamentali nelle trasformate di Fourier e nei problemi convoluzionali;
- le matrici di Hessenberg, utilizzate nei metodi iterativi e negli algoritmi di riduzione per il calcolo degli autovalori;
- le matrici di Hankel, che presentano simmetria rispetto all'antidiagonale e trovano applicazione nei modelli di processi stocastici e nelle equazioni integrali;
- le matrici di Hilbert, celebri esempi di matrici malcondizionate, rilevanti per l'analisi numerica e l'interpolazione.

Gran parte del lavoro prende spunto e si sviluppa a partire dagli studi condotti dal Prof. Carmine Di Fiore, i cui contributi hanno influenzato in modo decisivo l'approccio adottato in questa tesi, in particolare nella formalizzazione delle algebre di matrici associate a trasformate discrete rapide (come DFT, DST e DHT), e nella costruzione di algoritmi efficienti che sfruttano appieno la struttura delle matrici. L'interesse verso le matrici strutturate nasce non solo dalla loro eleganza teorica,

ma anche dalla loro capacità di modellare fenomeni reali con una ridotta complessità: dalla modellazione di sistemi lineari tempo-invarianti alla compressione dei dati, fino a metodi numerici per la simulazione e il controllo.

Gli obiettivi principali del lavoro sono:

1. Fornire una panoramica esaustiva delle principali classi di matrici strutturate, illustrandone le proprietà algebriche, spettrali e strutturali, con un focus specifico sulle matrici di Toeplitz, grazie alla loro rilevanza teorica e applicativa;

2. Esplorare il ruolo delle algebre di matrici, in particolare quelle associate a trasformate discrete rapide (come la Discrete Fourier Transform e la Discrete Hartley Transform), che ne consentono la diagonalizzazione simultanea e l'implementazione efficiente di operazioni fondamentali;

3. Presentare algoritmi numericamente efficienti per il trattamento delle matrici di Toeplitz, con attenzione a operazioni chiave come la moltiplicazione matrice-vettore e la risoluzione di sistemi lineari, sfruttando la struttura per ridurre la complessità computazionale;

4. Analizzare applicazioni teoriche e computazionali di tali matrici in contesti avanzati, come la decomposizione per dislocamento (displacement decomposition), la forma canonica di Jordan e il problema di Procruste per matrici di tipo Toeplitz.

Struttura del lavoro

Dopo aver introdotto le definizioni delle varie matrici interessate nel primo Capitolo, nel Capitolo 2 si discutono le proprietà delle matrici triangolari di Toeplitz e la loro relazione con le algebre matriciali. Vengono illustrate metodologie per la risoluzione di sistemi lineari e la moltiplicazione matrice-vettore, facendo leva sulla struttura ricorrente e sul concetto di shift matriciale.

Nel Capitolo 3 vengono presentate alcune algebre definite da trasformate discrete rapide, come la DFT e la DST, che consentono la diagonalizzazione simultanea di intere famiglie di matrici strutturate. In tale contesto si introduce la nozione di algebre sdU, con riferimento al ruolo centrale delle trasformate U nella semplificazione computazionale.

Il Capitolo 4 è dedicato alle applicazioni numeriche delle algebre sdU, in particolare nella risoluzione di sistemi Toeplitz e Toeplitz + Hankel (ad esempio con il metodo del Gradiente Coniugato) e nell'ottimizzazione numerica con metodi iterativi quasi-Newton (come BFGS), evidenziando come le proprieta' algebriche e

strutturali consentano una maggiore efficienza dei metodi.

Nel Capitolo 5 sono riportati esempi di applicazioni teoriche delle matrici strutturate, nella definizione dei numeri di Bernoulli, e nella costruzione della forma canonica di Jordan dell'inversa (in quest'ultima, anche la matrice di Tartaglia ha un ruolo fondamentale).

Il Capitolo 6 affronta infine il problema di Procruste per matrici di tipo Toeplitz, esaminando diverse configurazioni (generali, simmetriche, triangolari).

I contenuti raccolti suggeriscono alcune direzioni di approfondimento per studi futuri:

- Estensione teorica a nuove classi di matrici strutturate, incluse quelle ibride o definite da algebre non commutative;
- Ottimizzazione degli algoritmi esistenti, in particolare per il trattamento di matrici di grandi dimensioni;
- Applicazioni in ambiti interdisciplinari, come la signal processing, la compressione dei dati o l'intelligenza artificiale, dove le matrici strutturate si rivelano strumenti fondamentali.

In conclusione, questa tesi ha avuto l'obbiettivo di fornire una sintesi critica e ragionata della letteratura esistente sul tema, valorizzando l'intersezione tra algebra lineare, trasformazioni numeriche rapide e applicazioni computazionali delle matrici strutturate.

Capitolo 1

Definizioni e richiami di algebra lineare

1.1 Matrici, autovalori e autovettori

Iniziamo con la definizione di matrice.

Definizione: Una matrice $A \in \mathbb{C}^{n\times n}$ è un quadrato con n^2 elementi, che sono numeri reali o complessi.

Associati ad una matrice A, ci sono i suoi autovalori, autovettori e gli autospazi degli autovalori.

Definizione: Data una matrice $A \in \mathbb{C}^{n\times n}$, i suoi autovalori sono numeri complessi $\lambda \in \mathbb{C} | Ax = \lambda x$, per qualche vettore $x \in \mathbb{C}^n$ con $x \neq \underline{0}$. (Notiamo che se $\lambda \notin \mathbb{R}$ e $A \in \mathbb{R}^{n\times n} \Rightarrow x$ non è reale. Si dimostra vedendo che se $x \in \mathbb{R}^n \Rightarrow Ax \in \mathbb{R}^n$ ma $\lambda x \in \mathbb{C} \setminus \mathbb{R}$). L'insieme di tutti gli autovettori x di λ, che soddisfano l'uguaglianza di sopra, forma uno spazio vettoriale (sottospazio di $\mathbb{C}^n$) noto come autospazio dell'autovalore λ di A. Si noti che tale autospazio è invariante sotto l'azione di A.

Dati due distinti autovalori di una matrice $A \in \mathbb{C}^{n\times n}$, λ_1 e λ_2, ogni autovettore di λ_1 è linearmente indipendente con ogni autovettore di λ_2. Infatti, se $Ax_1 = \lambda_1 x_1$, $x_1 \neq 0$, $Ax_2 = \lambda_2 x_2$, $x_2 \neq 0$ e $\alpha_1 x_1 + \alpha_2 x_2 = 0 \Rightarrow 0 = p(A)(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 p(\lambda_1) x_1 + \alpha_2 p(\lambda_2) x_2$ per tutti i polinomi p. ($Ax = \lambda x \Rightarrow A^k x = \lambda^k x \Rightarrow \sum_k \alpha_k A^k x = \sum_k \alpha_k \lambda^k x \Rightarrow p(A)x = p(\lambda)x$ con $p(t) = \sum_k \alpha_k t^k$). In particolare, per $p(x) = (x - \lambda_2)/(\lambda_1 - \lambda_2)$ e $p(x) = (x - \lambda_1)/(\lambda_2 - \lambda_1)$ si ottiengono le condizioni: $\alpha_1 x_1 = 0$ e $\alpha_2 x_2 = 0$ che implicano $\alpha_1 = 0$ e $\alpha_2 = 0$ rispettivamente.

Nell'autospazio corrispondente ad un autovalore di A si possono scegliere un insieme di vettori linearmente indipendenti che generano l'autospazio. Questo si può ripetere per ogni autovalore distinto di A. Raccogliendo tutti questi insiemi si può formare una matrice rettangolare $R \in \mathbb{C}^{n\times m}, n \geq m$, le cui colonne sono linearmente indipendenti tale che $AR = RD$, con $D \in \mathbb{C}^{m\times m}$ matrice diagonale $m \times m$, $D_{ii} =$ autovalori di A.

Se m risulta essere uguale a n, allora R è una matrice $n \times n$ quadrata e invertibile (le sue colonne formano una base per $\mathbb{C}^n$), D è una matrice $n \times n$ con tutti gli autovalori di A come elementi sulla diagonale, e l'identità $AR = RD$ può essere riscritta come $R^{-1}AR = D$, in altre parole, A risulta diagonalizzabile mediante una trasformazione di similitudine.

Se $m < n$, non è possibile diagonalizzare A con una tale trasformazione, comunque $R \in \mathbb{C}^{n\times m}$ può essere completata, coinvolgendo opportuni vettori come nuove colonne, in modo da diventare una matrice quadrata invertibile $\chi \in \mathbb{C}^{n\times n}$ tale che $\chi^{-1}A\chi = J$ con J diagonale a blocchi con blocchi diagonali del tipo:

$$\mu_k I_{s_k} + Z_{s_k}^T = \begin{bmatrix} \mu_k & 1 & 0 & \dots & 0 \\ 0 & \mu_k & 1 & \dots. & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 1 \\ 0 & \dots & \dots & 0 & \mu_k \end{bmatrix},$$

$\mu_k \in \sigma(A) = \{$ autovalori di $A\}$, dove almeno uno di tali blocchi è di ordine almeno 2. Poiché l'equazione $Ax = \lambda x$ è equivalente al sistema lineare omogeneo $(\lambda I - A)x = 0$, gli autovalori di A si possono ottenere risolvendo l'equazione caratteristica $det(\lambda I - A) = 0$, essendo quest'ultima una condizione necessaria e sufficiente su λ per l'esistenza di soluzioni non nulle x di $(\lambda I - A)x = 0$.

Se si scrive la matrice $\lambda I - A$, allora è chiaro che $det(\lambda I - A)$ è un polinomio monico in λ di grado n i cui coefficienti sono funzioni degli elementi a_{ij} di A e sono reali se gli a_{ij} sono reali (sono reali anche in altri casi, ad esempio se A è Hermitiana). Quindi l'insieme $\sigma(A)$ dei autovalori λ di A coincide con l'insieme delle n radici $\lambda_1, \lambda_2, \dots, \lambda_n$ della seguente equazione algebrica

$$det(\lambda I - A) = \lambda^n - (\sum_i a_{ii})\lambda^{n-1} + \dots + (-1)^n det(A) = 0 \tag{1.1}$$

(La dimostrazione delle espressioni dei coefficienti di λ^0 e di λ^{n-1} è ommessa.)
Come conseguenza, se a_{ij} sono reali, allora $\lambda \in \sigma(A) \Rightarrow \bar{\lambda} \in \sigma(A)$, cioè l'insieme $\sigma(A)$ è chiuso rispetto alla coniugazione. Si noti che la rappresentazione del polinomio caratteristico $p_A(\lambda) = det(\lambda I - A)$ in termini dei suoi zeri, $det(\lambda I - A) = \Pi_{i=1}^n(\lambda - \lambda_i)$, implica le seguenti identità importanti, che mettono in relazione gli autovalori di A con i suoi elementi:

1. $\sum_i a_{ii} = \sum_i \lambda_i$
2. $det(A) = \Pi_i \lambda_i$

Ad esempio, la seconda relazione implica che una matrice A è singolare (ha determinante zero) se e solo se almeno uno dei suoi autovalori è zero.

Abbiamo visto che associato a qualsiasi matrice A c'è un polinomio di grado n, cioè il polinomio caratteristico $p(\lambda)$. Viceversa, consideriamo qualsiasi polinomio di grado n:

$$x^n + a_{n-1}x^{n-1} + \ldots + a_2x^2 + a_1x + a_0 \tag{1.2}$$

Domanda: esiste una matrice A $n \times n$ tale che il polinomio sopra è proprio il suo polinomio caratteristico? La risposta è si, basta prendere la cosidetta matrice di Frobenius associata al polinomio.

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \ldots & 0 & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 \\ -a_o & -a_1 & -a_2 & \ldots & -a_{n-1} \end{bmatrix}$$

scrivere

$$\lambda I - A = \begin{bmatrix} \lambda & -1 & 0 & \ldots & 0 \\ 0 & \lambda & -1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & -1 & 0 \\ 0 & 0 & \ldots & \lambda & -1 \\ a_o & a_1 & a_2 & \ldots & \lambda + a_{n-1} \end{bmatrix}$$

calcolare $det(\lambda I - A)$ con la regola di Laplace sulla prima colonna di $\lambda I - A$. Per esempio per

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_o & -a_1 & -a_2 \end{bmatrix}$$

e

$$\lambda I - A = \begin{bmatrix} \lambda & -1 & 0 \\ 0 & \lambda & -1 \\ a_o & a_1 & \lambda + a_2 \end{bmatrix}$$

$\Rightarrow det(\lambda I - A) = \lambda(\lambda(\lambda + a_2) + a_1) + a_0(-1)(-1) = \lambda^3 + \lambda^2 a_2 + \lambda^1 a_1 + a_0$.

Quindi il calcolo delle radici di un'equazione algebrica è equivalente al calcolo degli autovalori di una matrice.

Invece di calcolare gli autovalori di una matrice A $n \times n$, è spesso sufficiente localizzarli, cioè trovare una regione del campo complesso che li includa tutti o alcuni di essi. In particolare, il seguente teorema di Gershgorin permette di localizzarli tutti nell'unione di n cerchi facilmente definiti dagli elementi di A, infatti si ha:

$$\lambda \in \mathbb{C}, Ax = \lambda x, x \neq 0, \Rightarrow \lambda \in \bigcup_{i=1}^{n} K_i, K_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|\}.$$

Quando A è irriducibile, vale un'affermazione più precisa. Se $\lambda \in \sigma(A)$ allora λ o si trova nella parte interna di almeno uno dei cerchi di Gershgorin, o si trova sulla frontiera di ogni cerchio di Gershgorin. Ricordiamo che una matrice A è riducibile se esiste $\mathcal{I} \subset \{1,2,\ldots,n\}, \mathcal{I} \neq \{1,2,\ldots,n\}, \mathcal{I} \neq \oslash$, tale che $a_{ij} = 0$, per ogni $i \in \mathcal{I}, j \in \{1,2,\ldots,n\} \setminus \mathcal{I}$, o, equivalentemente se esiste una matrice permutazione P per cui $P^T AP$ è una matrice triangolare superiore a blocchi 2×2 con blocchi diagonali quadrati [Bertaccini, Di Fiore, Zellini].

1.2 Matrici di Toeplitz

Le matrici di Toeplitz, chiamate così dal matematico tedesco Otto Toeplitz, rappresentano una classe di matrici strutturate. Sono matrici in cui ogni diagonale parallela alla diagonale principale presenta elementi costanti. Formalmente, una matrice di Toeplitz di dimensioni $n \times n$ è definita dalla condizione $T_{ij} = t_{i-j}$, quindi è determinata da $2n - 1$ parametri liberi:

$$T = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-n+1} \\ t_1 & t_0 & t_{-1} & \cdots & t_{-n+2} \\ t_2 & t_1 & t_0 & \cdots & t_{-n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{n-1} & t_{n-2} & t_{n-3} & \cdots & t_0 \end{pmatrix}$$

Le matrici di Toeplitz si incontrano in una vasta gamma di applicazioni, tra cui la convoluzione discreta nei sistemi lineari tempo-invarianti, il filtraggio digitale dei segnali. La struttura costante lungo le diagonali permette l'uso di algoritmi veloci, come la trasformata di Fourier veloce (FFT), che riduce la complessità computazionale delle operazioni che coinvolgono T, come quelle di convoluzione.

1.3 Matrici Triangolari

Una matrice quadrata si dice triangolare se tutti gli elementi al di sopra o al di sotto della diagonale principale sono nulli. In particolare, una matrice è detta triangolare inferiore se tutti gli elementi al di sopra della diagonale principale sono nulli, mentre è triangolare superiore se gli elementi nulli si trovano al di sotto della diagonale principale. Nel contesto delle matrici di Toeplitz, una matrice triangolare

superiore ha la seguente forma:

$$G = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-n+1} \\ 0 & t_0 & t_{-1} & \cdots & t_{-n+2} \\ 0 & 0 & t_0 & \cdots & t_{-n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & t_0 \end{pmatrix}$$

L'insieme delle matrici triangolari sup (inf) di Toeplitz è chiusa rispetto al prodotto e all'inversione. Il prodotto di due matrici triangolari sup (inf) è ancora triangolare sup (inf). Data una matrice triangolare sup (inf) con elementi diagonali non nulli, l'inversa è ben definita ed è ancora triangolare sup (inf). In generale G non è diagonalizzabile.

Le matrici di Toeplitz triangolari trovano applicazione nell'elaborazione dei segnali, in particolare nella modellazione di sistemi a risposta impulsiva finita e nelle convoluzioni causali, riflettendo la direzionalità temporale dei segnali discreti.

1.4 Matrici di Hessenberg

Una matrice di Hessenberg è una matrice quadrata che ha tutti gli elementi nulli al di sotto della prima sottodiagonale (nel caso di Hessenberg superiore) o al di sopra della prima sovradiagonale (nel caso di Hessenberg inferiore).

Una qualsiasi matrice può essere ridotta ad una matrice di Hessenberg superiore (o inferiore) tramite trasformazioni per similitudine. Ad esempio, una matrice generica A può essere trasformata in una matrice di Hessenberg superiore H attraverso una trasformazione del tipo:

$$H = Q^H A Q$$

dove Q è una matrice unitaria e H è la matrice di Hessenberg superiore. Questa trasformazione è il primo passo da fare se si vogliono calcolare gli autovalori di A. Poi, con il metodo QR, si calcolano gli autovalori di H, che coincidono con quelli di A. L'algoritmo QR per calcolare gli autovalori di H verrà a costare solo $O(n^2)$ per passo. Nel metodo iterativo GMRES per la risoluzione di sistemi lineari, nella costruzione di v_k tale che

$$min||\underline{b} - A(\underline{x}_0 + \underline{v})||_2 = v \in Span\{\underline{r}_0, A\underline{r}_0 \cdots A^{k-1}\underline{r}_0\} = ||\underline{b} - A(\underline{x}_0 + \underline{v}_k)||_2$$

intervengono matrici di Hessenberg superiori.

Una matrice di Toeplitz può assumere la forma di Hessenberg superiore se tutti gli elementi al di sotto della prima sottodiagonale sono nulli:

$$H = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-n+1} \\ t_1 & t_0 & t_{-1} & \cdots & t_{-n+2} \\ 0 & t_1 & t_0 & \cdots & t_{-n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & t_0 \end{pmatrix}.$$

Tali matrici sono studiate ad esempio in [Slowik].

1.5 Matrici simmetriche

Si definisce matrice simmetrica una matrice quadrata che coincide con la sua trasposta; i suoi elementi sono, quindi, simmetrici rispetto alla diagonale principale. Una matrice di Toeplitz simmetrica ha la forma:

$$S = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-n+1} \\ t_{-1} & t_0 & t_{-1} & \cdots & t_{-n+2} \\ t_{-2} & t_{-1} & t_0 & \cdots & t_{-n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{-n+1} & t_{-n+2} & t_{-n+3} & \cdots & t_0 \end{pmatrix}.$$

Queste matrici trovano applicazione nella teoria dei segnali e nella statistica, in quanto rappresentano processi stazionari in cui la covarianza dipende solo dalla distanza tra due punti. Un esempio di matrice simmetrica è l'Hessiano di una funzione $f \in C^2(\mathrm{R}^n)$.

1.6 Matrici Hermitiane

Una matrice A quadrata a coefficienti in campo complesso, si definisce Hermitiana o matrice autoaggiunta se coincide con la sua trasposta coniugata. Indicata con $\overline{A}$ la matrice complessa coniugata associata ad A e con $\overline{A}^t$ la trasposta della matrice coniugata, si può dire che A è Hermitiana se e solo se $A = \overline{A}^t$. Gli elementi di $A = \overline{A}^t$ lungo la diagonale principale sono reali mentre gli elementi fuori la diagonale sono complessi e soddisfano la relazione $t_{ji} = \overline{t_{ij}}$ dove $\overline{t_{ij}}$ indica il complesso coniugato di t_{ij}. Le matrici Hermitiane hanno gli autovalori tutti reali. Un esempio di matrice di Toeplitz Hermitiana T con dimensioni 4×4 è:

$$T = \begin{pmatrix} 2 & 1+i & 3-2i & 4 \\ 1-i & 2 & 1+i & 3-2i \\ 3+2i & 1-i & 2 & 1+i \\ 4 & 3+2i & 1-i & 2 \end{pmatrix}$$

Le matrici di Toeplitz Hermitiane trovano largo impiego nella meccanica quantistica e nella teoria dei segnali complessi, in cui la simmetria Hermitiana garantisce che gli autovalori della matrice siano reali.

1.7 Matrici definite positive

Una matrice $A \in \mathbb{C}^{n\times n}$ è definita positiva se:

$$x^H Ax > 0 \quad \text{per ogni} \quad x \in \mathbb{C}^n \neq 0.$$

In altre parole, una matrice $A \in \mathbb{C}^{n\times n}$ è definita positiva se, per ogni vettore x non nullo, il prodotto quadratico $x^H Ax$ è reale positivo. Una matrice A è definita positiva se e solo se ha autovalori reali positivi ed è Hermitiana. Un esempio di matrice definita positiva è l'Hessiano di $f : \mathrm{R}^n \to \mathrm{R},\ f \in C^2$, nell'intorno di un punto minimo locale per f.

Le matrici di Toeplitz definite positive hanno applicazioni nella teoria dell'ottimizzazione e nei modelli statistici. Due di queste sono le seguenti:

$$\begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \quad e \quad \begin{pmatrix} 1 & t & \cdots & t^{n-1} \\ t & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t \\ t^{n-1} & \cdots & t & 1 \end{pmatrix}, \ |t| < 1\,.$$

1.8 Matrici di Hankel e di Hilbert

Una matrice di Hankel è una matrice quadrata con diagonali a pendenza positiva costanti. Una matrice di Hankel è definita da un vettore c tale che $H(i,j) = c_{i+j-1}$, ovvero gli elementi della matrice dipendono dalla somma degli indici di riga e colonna. Una matrice di Hankel H ha la seguente forma:

$$H = \begin{pmatrix} c_1 & c_2 & c_3 & \cdots & c_n \\ c_2 & c_3 & c_4 & \cdots & c_{n+1} \\ c_3 & c_4 & c_5 & \cdots & c_{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_n & c_{n+1} & c_{n+2} & \cdots & c_{2n-1} \end{pmatrix}.$$

La differenza con una una matrice di Toeplitz è che la seconda ha elementi costanti lungo le diagonali, mentre la matrice di Hankel ha elementi costanti lungo le anti-diagonali. Quindi, sono strutturalmente simili, ma la simmetria è invertita.

Una particolare matrice di Hankel è la matrice di Hilbert i cui elementi sono:

$$H_{ij} = \frac{1}{i+j-1} \quad i,j = 1 \cdots n$$

È semplice vedere che gli

$$\hat{a}_i : \int_0^1 (f(x) - \sum_{i=0}^{n-1} \hat{a}_i x^i)^2 dx = \min_{a_i} \int_0^1 (f(x) - \sum_{i=0}^{n-1} a_i x^i)^2 dx$$

si ottengono risolvendo un sistema lineare di tipo $Hx = b$ [Bini, Capovani, Menchi]. La matrice di Hilbert è definita positiva, infatti

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = |x + \frac{1}{2}y|^2 + \frac{1}{12}|y|^2 > 0$$

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = |x + \frac{1}{2}y + \frac{1}{3}z|^2 + \frac{1}{12}|y+z|^2 + \frac{1}{180}|z|^2 > 0.$$

Più in generale si vede che $H = LDL^T$ con D diagonale tale che $D_{ii} > 0$ e L triangolare inferiore tale che $L_{ii} = 1$, $\forall i$, e quindi

$$\underline{w}^H H \underline{w} = \sum_{i=1}^{n} D_{ii}[(L^T w)_i]^2 > 0, \; se \; \underline{w} \neq 0.$$

La matrice di Hilbert è nota per essere malcondizionata, [Bini, Capovani, Menchi] il che significa che anche piccoli errori nei dati di input possono portare a grandi errori nei risultati, se l'espressione da valutare è in termini di una matrice di Hilbert o della sua inversa. Ricordiamo che una misura del condizionamento di A $n \times n$ è il numero $||A|| \cdot ||A^{-1}||$, essendo $|| \cdot ||$ una norma matriciale.

1.9 Matrici circolanti

Le matrici circolanti sono particolari matrici di Toeplitz della forma:

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

dove ogni riga è uno shift ciclico verso destra della riga sopra di essa. La struttura della matrice circolante può essere caratterizzata dalla notazione

$$C_{k,j} = c_{(j-k)modn}.$$

Le proprietà delle matrici circolanti sono ben conosciute [Davis] e tra l'altro, come vedremo, consentono di ridurre la complessità delle operazioni con le matrici di Toeplitz.
Ad esempio la matrice

$$\begin{pmatrix} 2 & (-1+1/n) & 0 & \cdots & 0 & (-1+1/n) \\ (-1+1/n) & 2 & (-1+1/n) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & (-1+1/n) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 0 & (-1+1/n) & \ddots & (-1+1/n) \\ (-1+1/n) & 0 & \cdots & 0 & (-1+1/n) & 2 \end{pmatrix}$$

è la matrice circolante che meglio approssima

$$\begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$$

in norma di Frobenius.

Capitolo 2

Algebre di matrici, matrici triangolari di Toeplitz

2.1 Definizione

Un insieme $\mathcal{L} \subset \mathbb{C}^{n\times n}$ è un algebra di matrici se $\mathcal{L}$ è un sottospazio vettoriale di $\mathbb{C}^{n\times n}(\alpha, \beta \in \mathbb{C}, A, B \in \mathcal{L} \Rightarrow \alpha A + \beta B \in \mathcal{L})$ e il prodotto di matrici appartenenti a $\mathcal{L}$ è ancora una matrice di $\mathcal{L}$ $(A, B \in \mathcal{L} \Rightarrow AB \in \mathcal{L}.)$

2.2 Algebre di gruppo

Sia $\mathcal{G} = \{1,2, ..., n\}$ un gruppo finito con elemento identico 1. Si può far corrispondere a $\mathcal{G}$ il seguente insieme $\mathcal{L}$ di matrici:

$$\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : a_{i,j} = a_{ki,kj}, \quad i, j, k \in \mathcal{G}\}.$$

Una prima cosa da osservare è che $\mathcal{L}$ ammette la seguente rappresentazione

$$\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : a_{i,j} = a_{1,i^{-1}j}, \quad i, j \in \mathcal{G}\}.$$

dalla quale si deduce che una matrice in $\mathcal{L}$ è in particolare univocamente definita dalla sua prima riga, che può essere arbitraria. È evidente che $\mathcal{L}$ è un sottospazio vettoriale di $\mathbb{C}^{n\times n}$ di ordine n. Verifichiamo che è chiuso rispetto alla moltiplicazione di matrici. Siano $A, B \in \mathcal{L}$ e $i, j, k \in \mathcal{G}$; allora

$$[AB]_{i,j} = \sum_{s\in\mathcal{G}} [A]_{i,s}[B]_{s,j} = \sum_{s\in\mathcal{G}} [A]_{ki,ks}[B]_{ks,kj} = \sum_{r\in\mathcal{G}} [A]_{ki,r}[B]_{r,kj} = [AB]_{ki,kj}.$$

Per applicazioni interessanti delle algebre di gruppo nella teoria delle displacement decompositions, vedi [Gader].

Vediamo un esempio di algebra di gruppo. Sia $\mathcal{G}$ il gruppo ciclico di ordine n, cioè $\mathcal{G} = \{1,2,...,n\}$ con $i \to g^{i-1}, i \in \mathcal{G}$, dove g è un elemento generatore di $\mathcal{G}$ (si noti che $g^n = 1$). Studiamo la struttura dell'algebra di gruppo $\mathcal{L}$ corrispondente. Per definizione, per l'elemento generico di $A \in \mathcal{L}$ si deve avere

$$a_{i,j} = a_{1,i^{-1}j} = a_{1,(g^{i-1})^{-1}(g^{j-1})} = a_{1,g^{n+1-i}g^{j-1}} = \begin{cases} a_{1,g^{j-i}} = a_{1,j-i+1}, & \text{se } j \geq i, \\ a_{1,g^{n+j-i}} = a_{1,n+j-i+1}, & \text{se } j < i. \end{cases}$$

È evidente che se $j - i$ è costante allora rimane invariato l'elemento (i, j) di A, in altre parole A deve essere una matrice di Toeplitz. Più precisamente, dalle uguaglianze ottenute segue che A è una matrice di Toeplitz definita univocamente dalla sua prima riga ed ha la seguente struttura.

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} & \cdots & a_{1,n} \\ a_{1,n} & \ddots & \ddots & \ddots & \ddots \\ \cdots & \cdots & \ddots & \cdots & a_{1,k} \\ a_{1,k} & \ddots & \vdots & \ddots & \vdots \\ \vdots & a_{1,k} & \cdots & \cdots & a_{1,1} \end{pmatrix}$$

cioè ha, per ogni k, nelle posizioni $(1,k),(2,k+1),...,(n+1-k,n),(n+2-k,1),...,(n,k-1)$ sempre lo stesso numero $a_{1,k}$. Ad esempio nel caso $n = 4$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{14} & a_{11} & a_{12} & a_{13} \\ a_{13} & a_{14} & a_{11} & a_{12} \\ a_{12} & a_{13} & a_{14} & a_{11} \end{bmatrix}.$$

Dunque l'algebra di $\mathcal{L}$ corrispondente al gruppo ciclico di ordine n coincide con l'insieme delle matrici circolanti $n \times n$. Tale $\mathcal{L}$ viene chiamata $\mathcal{C}$ [Davis].

2.2.1 Il commutatore di una matrice

Sia X una matrice $n \times n$ a elementi in $\mathbb{C}$. Sia $\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : AX = XA\}$. È semplice mostrare che $\mathcal{L}$ è un algebra di matrici chiusa per inversione. In generale le matrici di $\mathcal{L}$ non commutano tra loro (si può prendere $X = I$ per verificarlo.)

2.2.2 Lo Spazio dei polinomi di una matrice

Sia X una matrice $n\times n$ a elementi in $\mathbb{C}$. Dato un polinomio $p(t) = a_0+a_1t+...+a_kt^k$, con il simbolo $p(X)$ intendiamo la matrice $a_0I+a_1X+...+a_kX^k$. Sia $\mathcal{L} = \{p(X)\} =$

$\{p(X) : p = \text{polinomi di grado k;}, k \in \mathbb{N}\}$. È semplice mostrare che $\mathcal{L}$ è un'algebra di matrici commutativa la cui dimensione è data dal grado del polinomio minimo di X ed è quindi minore o uguale ad n. Si dimostra inoltre che $\mathcal{L}$ è chiusa per inversione, cioè se $A \in \mathcal{L}$ è non singolare, allora $A^{-1} \in \mathcal{L}$ (suggerimento: utilizzare il teorema di Cailey-Hamilton applicato ad A).

Un'algebra di questo tipo è l'algebra di gruppo $\mathcal{C}$ delle matrici circolanti. Dimostriamolo. Sia Π :

$$\Pi = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

la matrice circolante $n \times n$ la cui prima riga è il vettore $[0\,1\,0 \cdots 0]$. È semplice osservare che la matrice $\sum_{k=1}^{n} a_k \Pi^{k-1}$ è circolante per ogni scelta degli a_k ($\{p(\Pi)\} \subset \mathcal{C}$), e che la generica matrice circolante, quella la cui prima riga è il generico vettore $[a_{11}a_{12}...a_{1n}]$, si può scrivere nella forma $\sum_{k=1}^{n} a_{1,k}\Pi^{k-1}$ $(\mathcal{C} \subset \{p(\Pi)\})$. In altre parole, vale l'uguaglianza $\mathcal{C} = \{p(\Pi)\}$.

Analogamente, si può studiare l'algebra $\mathcal{C}_{-1} = \{p(\Pi_{-1})\}$, chiamata anche spazio delle matrici (-1)-circolanti, che però non è un'algebra di gruppo, dove la matrice Π_{-1} è la seguente:

$$\Pi_{-1} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 1 \\ -1 & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

Si può dimostrare che ogni matrice di Toeplitz T è la somma di una matrice circolante C e una matrice (-1)-circolante C_{-1}.

Tornando al caso di una X generica, vale il seguente teorema [Lancaster, Tismenetsky] :

Teorema 2.2.1. *Sia X una matrice $n \times n$ a elementi in $\mathbb{C}$. Allora $\{p(X)\} \subset \{A : AX = XA\}$ e $dim\{p(X)\} \leq n \leq dim\{A : AX = XA\}$. Inoltre, gli spazi $\{p(X)\}$ e $\{A : AX = XA\}$ coincidono se e solo se $dim\{p(X)\} = n = dim\{A : AX = XA\}$, e in tal caso X si dice non derogatoria.*

Ci sono diverse condizioni equivalenti per la non derogatorietà di una matrice X. Ad esempio, una matrice è non derogatoria se e solo se ad ogni suo autovalore corrisponde un solo blocco di Jordan ovvero se e solo se i polinomi minimo e caratteristico di X coincidono, oppure se e solo se $\{p(X)\}$ ha la struttura di uno spazio

di classe V (vedi [Di Fiore, Zellini, 1995]). Più semplicemente, ogni qual volta si ha

$$p \text{ è un polinomio}, \quad p(X) = 0 \Rightarrow \deg(p) \geq n.$$

la matrice X è non derogatoria.

2.2.3 Le matrici triangolari di Toeplitz

Sia Z la seguente matrice lower-shift $n \times n$

$$(1) \quad Z = \begin{bmatrix} 0 & & & & \\ 1 & & & & \\ & 1 & & & \\ & & . & & \\ & & & 1 & 0 \end{bmatrix}$$

Si noti che la moltiplicazione di Z per un vettore $v = [v_0\, v_1 \cdots v_{n-1}]^T \in \mathbb{C}^n$ sposta in giù le sue componenti $Zv = [0\, v_0\, v_1 \cdots v_{n-2}]^T$. Sia $\mathcal{L}$ lo spazio delle matrici che commutano con Z. Studiamo nei dettagli tale spazio, che ovviamente è anche un'algebra. Sia $A \in \mathbb{C}^{n\times n}$ generica. Allora

$$AZ = \begin{bmatrix} a_{12} & . & a_{1n} & 0 \\ \vdots & & \vdots & \vdots \\ a_{n2} & . & a_{nn} & 0 \end{bmatrix}, \; ZA = \begin{bmatrix} 0 & \dots & 0 \\ a_{11} & \dots & a_{1n} \\ . & & . \\ a_{n-1,1} & \dots & a_{n-1,n} \end{bmatrix}$$

Imponendo l'uguaglianza $AZ = ZA$ si ottengono le condizioni $a_{12} = a_{13} = ... = a_{1n} = a_{23} = a_{2n} = ... = a_{n-1n} = 0$ e $a_{i,j+1} = a_{i-1,j}, j \leq i-1$, dalle quali si deduce la struttura di $A \in \mathcal{L}$: A deve essere una matrice triangolare inferiore di Toeplitz del tipo

$$A = \begin{bmatrix} a_{11} & & & & \\ a_{21} & a_{11} & & & \\ a_{31} & a_{21} & a_{11} & & \\ . & . & . & . & \\ a_{n1} & . & a_{31} & a_{21} & a_{11} \end{bmatrix}$$

Ne segue in particolare che $\dim\{A : AZ = ZA\} = n$ e quindi si ha anche l'identità $\{A : AZ = ZA\} = \{p(Z)\}$. Effettivamente, se si esaminano le potenze di Z ci si accorge che la matrice triangolare di Toeplitz A di cui sopra coincide con il polinomio $\sum_{k=1}^{n} a_{k1} Z^{k-1}$.

Dall'identità $\{A : AZ = ZA\} = \{p(Z)\}$ segue in particolare che l'inversa di una matrice triangolare inferiore di Toeplitz è ancora triangolare inferiore di Toeplitz, ed è quindi anch'essa definita dalla sua prima colonna.

2.2.4 Algebre di matrici simultaneamente diagonalizzabili

Sia $M \in \mathcal{C}^{n\times n}$ una matrice non singolare. Sia $\mathcal{L}$ lo spazio delle matrici simultaneamente diagonalizzate da M, cioè

$$\mathcal{L} = sdM := \{MDM^{-1} : D = \text{matrici diagonali}, D_{ii} \in \mathbb{C}, \forall i\}.$$

È evidente che $\mathcal{L}$ è una algebra di matrici commutativa. Questo risultato segue anche dall'osservazione che $\mathcal{L}$ può essere rappresentato come l'insieme dei polinomi in una matrice X; è sufficiente scegliere $X = M\widetilde{D}M^{-1}$ con $\widetilde{D}$ matrice diagonale con elementi diagonali distinti (si lascia al lettore la verifica di questo fatto). Non è vero il contrario, cioè non è vero in generale che uno spazio del tipo $\{p(X)\}$ sia esprimibile nella forma $\{MDM^{-1}\}$ per qualche matrice M non singolare. Ad esempio, non può essere vero per $\{p(Z)\}$, l'insieme delle matrici triangolari inferiori di Toeplitz, perché la matrice Z in (1) non è diagonalizzabile.

Sia $v \in \mathbb{C}^n$ tale che $(M^T v)_i \neq 0, \forall i$. Si osserva che $A \in \mathcal{L}$ è univocamente determinata dal vettore $v^T A$, infatti $A \in \mathcal{L}$ se e solo se $A = Md(M^T A^T v)d(M^T v)^{-1}M^{-1}$, essendo $d(z)$ la matrice diagonale con elementi diagonali le componenti del vettore z [Di Fiore, Zellini, 1995]. In particolare, se $v = e_h$, allora ogni matrice di $\mathcal{L}$ è univocamente determinata dalla sua h-esima riga.

Nel prossimo capitolo descriveremo nei dettagli alcuni esempi di algebre sdM di matrici simultaneamente diagonalizzate da una matrice M. In tali esempi la matrice M è unitaria, $M^H = M^{-1}$, e definisce una trasformata discreta veloce, cioè ogni prodotto matrice-vettore $Mz, M^H z, z \in \mathbb{C}^n$, è calcolabile effettuando non più di $O(n \log n)$ operazioni aritmetiche. Inoltre mostreremo che per A in tali algebre le operazioni prodotto matrice-vettore Af e risoluzione sistema lineare $Ax = f$ possono essere eseguite al basso costo computazionale di $O(n \log n)$ operazioni aritmetiche, utilizzando procedure alternative a quelle standard. Nella prossima sezione si vedrà che queste affermazioni sono vere anche per A triangolari di Toeplitz. Le algebre di matrici di bassa complessità computazionale, come le sdM che vedremo nel prossimo capitolo o come le triangolari di Toeplitz, possono essere utilizzate per rendere più efficiente la risoluzione numerica di diversi problemi matematici, anche non di algebra lineare.

2.3 Complessità dei calcoli con le matrici triangolari di Toeplitz

Moltiplicare una matrice $n \times n$ triangolare inferiore di Toeplitz per un vettore non richiede più di $O(n \log n)$ operazioni aritmetiche. La stessa affermazione vale per la risoluzione di un sistema lineare di n equazioni la cui matrice dei coefficienti è triangolare inferiore di Toeplitz, e questo perchè tale operazione può essere ricondotta al calcolo di $O(\log n)$ prodotti matrice-vettore dove la matrice è triangolare

di Toeplitz e di dimensione ogni volta la metà. Tutto ciò sarà provato in questa sezione.

2.3.1 Moltiplicare una matrice triangolare inferiore di Toeplitz (t.i.T.) per un vettore

Ci sono almeno due modi per calcolare il prodotto di una matrice di Toeplitz $n \times n, T = (t_{i-j})_{i,j=1}^{n}$ per un vettore in al più $O(n \log n)$ operazioni aritmetiche, ed entrambi prevedono l'uso dell'algoritmo FFT . Uno consiste nell'utilizzare la rappresentazione di T come somma di una matrice circolante e una matrice (-1)-circolante (il lettore può trovare facilmente tale rappresentazione). L'altro, descritto nei dettagli qui di seguito, si basa sull'osservazione che ogni matrice di Toeplitz può essere immersa in una matrice circolante.
Si consideri una generica matrice di Toeplitz T 4×4 ed un vettore $\mathbf{v}$ 4×1. Allora T può essere vista come la sottomatrice in alto a sinistra di una matrice circolante C 8×8, e per il vettore $T\mathbf{v}$ vale la seguente rappresentazione:

$$T\mathbf{v} = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & t_{-3} \\ t_1 & t_0 & t_{-1} & t_{-2} \\ t_2 & t_1 & t_0 & t_{-1} \\ t_3 & t_2 & t_1 & t_0 \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{bmatrix} = \{ \begin{bmatrix} t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1 \\ t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2 \\ t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3 \\ t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0 \\ 0 & t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} \\ t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2} \\ t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 & t_{-1} \\ t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \}_4 =$$

$$= \{C \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} \}_4$$

dove col simbolo $\{\mathbf{z}\}_4$ si intende il vettore le cui componenti sono le prime quattro componenti del vettore $\mathbf{z}$.
Se T è $n \times n$ e $\mathbf{v}$ è $n \times 1$, allora l'osservazione vale ancora e può essere generalizzata, infatti

$T\mathbf{v} = \{C \begin{bmatrix} \mathbf{v} \\ \mathbf{0}_{(b-1)n} \end{bmatrix} \}_n$, dove $C = \mathcal{C}(\mathbf{a})$ è la matrice circolante con prima riga $\mathbf{a}^T = [t_0\, t_{-1} \cdots t_{-n+1}\, \mathbf{0}^T_{(b-2)n+1}\, t_{n-1} \cdots t_1]$

Se n è una potenza di b ($b = 2, 3, ...$), sfruttando la rappresentazione spettrale di $\mathcal{C}(\mathbf{a})$, $\mathcal{C}(\mathbf{a}) = \sqrt{bn} \cdot F_{bn} \cdot d(F_{bn}\mathbf{a}) \cdot F^H_{bn}$, e le proprietà di F_{bn} (vedi il prossimo capitolo), si deduce immediatamente una procedura di costo $O(n \log_b n)$ per il calcolo del prodotto di una matrice di Toeplitz $n \times n$ per un vettore. Tale procedura ha come sotto-procedura l'algoritmo FFT (vedi il prossimo capitolo).
Nella prossima sezione vedremo che la risoluzione di un sistema lineare triangolare inferiore di Toeplitz di n equazioni può ricondursi al calcolo di $O(\log n)$ prodotti

matrice-vettore dove la matrice è sempre triangolare inferiore di Toeplitz ed è di dimensione variabile, che si dimezza ogni volta. Ne segue che è opportuno avere a disposizione un metodo che effettui tali prodotti il più efficientemente possibile. Un metodo abbastanza efficiente si ottiene ponendo $t_{-i} = 0, i = 1, ..., n-1$, nella procedura sopra illustrata. Sarebbero benvenuti metodi che sfruttino meglio la triangolarità delle nostre matrici di Toeplitz.

2.3.2 Un algoritmo per la risoluzione di sistemi triangolari di Toeplitz

In questa sezione si illustrerà un algoritmo di costo $O(n \log_2 n)$ per il calcolo di x tale che $Ax = f$, essendo A una matrice triangolare inferiore di Toeplitz $n \times n$ con n potenza di 2 e $[A]_{11} = 1$ [Di Fiore, Zellini, manuscript].

2.3.3 Lemmi preliminari

Dato un vettore $\mathbf{v} = [v_0 \quad v_1 \quad v_2 \quad \ldots]^T, v_i \in \mathbb{C}$ (in breve $\mathbf{v} \in \mathbb{C}^{\mathbb{N}}$), sia $L(\mathbf{v})$ la matrice semi-infinita triangolare inferiore di Toeplitz con prima colonna $\mathbf{v}$, i.e.

$$L(\mathbf{v}) = \sum_{k=0}^{+\infty} v_k Z^k, Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \cdot & \cdot \end{bmatrix}.$$

Lemma 2.3.1. *(Lemma 1) Siano* $\mathbf{a}, \mathbf{b}, \mathbf{c}$ *vettori di* $\mathbb{C}^N$*. Allora* $L(\mathbf{a})L(\mathbf{b}) = L(\mathbf{c})$ *se e soltanto se* $L(\mathbf{a})\mathbf{b} = \mathbf{c}$*.*

Dimostrazione. Se $L(\mathbf{a})L(\mathbf{b}) = L(\mathbf{c})$, allora la prima colonna della matrice $L(\mathbf{a})L(\mathbf{b})$ deve essere uguale alla prima colonna della matrice $L(\mathbf{c})$, e queste sono rispettivamente i vettori $L(\mathbf{a})\mathbf{b}$ e $\mathbf{c}$. Viceversa, supponiamo che $L(\mathbf{a})\mathbf{b} = \mathbf{c}$. Consideriamo la matrice $L(\mathbf{a})L(\mathbf{b})$. Questa, in quanto prodotto di matrici triangolari inferiori di Toeplitz, è una matrice triangolare inferiore di Toeplitz, e, per ipotesi, la sua prima colonna, $L(\mathbf{a})\mathbf{b}$, coincide con il vettore $\mathbf{c}$, che è la prima colonna della matrice triangolare inferiore di Toeplitz $L(\mathbf{c})$. La tesi segue dal fatto che le matrici triangolari inferiori di Toeplitz sono univocamente definite dalla loro prima colonna. □

Dato un vettore $\mathbf{v} = [v_0 \quad v_1 \quad v_2 \quad \ldots]^T$, sia E la matrice semi-infinita di 0 e 1 che manda $\mathbf{v}$ nel vettore $E\mathbf{v} = [v_0 \quad 0 \quad v_1 \quad 0 \quad v_2 \quad 0 \quad \ldots]^T$:

$$E = \begin{bmatrix} 1 & & & \\ 0 & & & \\ 0 & 1 & & \\ 0 & 0 & & \\ 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

. In altre parole, l'azione di E su $\mathbf{v}$ ha l'effetto di inserire uno zero tra due successive componenti di $\mathbf{v}$. Si osserva facilmente che

$$E^2 = \begin{bmatrix} 1 & & & \\ 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & 1 & & \\ 0 & 0 & & \\ 0 & 0 & & \\ 0 & 0 & & \\ 0 & 0 & 1 & \\ . & . & . & . \end{bmatrix}, \quad E^s = \begin{bmatrix} 1 & & & \\ \mathbf{0} & & & \\ 0 & 1 & & \\ \mathbf{0} & \mathbf{0} & & \\ 0 & 0 & 1 & \\ . & . & . & . \end{bmatrix}, \quad \mathbf{0} = \mathbf{0}_{2^s-1}$$

cioè l'azione di E^s su $\mathbf{v}$ ha l'effetto di inserire $2^s - 1$ zeri tra due successive componenti di $\mathbf{v}$.

Lemma 2.3.2. *(Lemma 2) Siano $\mathbf{u}, \mathbf{v}$ vettori di $\mathbb{C}^N$ con $u_0 = v_0 = 1$. Allora $L(E\mathbf{u})E\mathbf{v} = EL(\mathbf{u})\mathbf{v}$, e, più in generale, per ogni $s \in \mathbb{N}$ si ha $L(E^s\mathbf{u})E^s\mathbf{v} = E^sL(\mathbf{u})\mathbf{v}$.*

Dimostrazione. Scrivendo i vettori $L(E\mathbf{u})E\mathbf{v}$ e $EL(\mathbf{u})\mathbf{v}$ si osserva che sono uguali. Moltiplicando a sinistra per E l'identità $L(E\mathbf{u})E\mathbf{v} = EL(\mathbf{u})\mathbf{v}$ ed utilizzando tale stessa identità con i vettori $E\mathbf{u}$ ed $E\mathbf{v}$ al posto, rispettivamente, di $\mathbf{u}$ e $\mathbf{v}$, si osserva che vale anche l'uguaglianza $L(E^2\mathbf{u})E^2\mathbf{v} = E^2L(\mathbf{u})\mathbf{v}\ldots$ □

Algoritmo

Sia A una matrice triangolare inferiore di Toeplitz $n \times n$ con $[A]_{11} = 1$. Si vuole risolvere il sistema $A\mathbf{x} = \mathbf{f}$. L'algoritmo seguente sfrutta l'osservazione che A^{-1} è ancora una matrice triangolare inferiore di Toeplitz $n \times n$.

1. Si calcola la prima colonna della matrice triangolare inferiore di Toeplitz A^{-1}, ovvero si risolve il sistema lineare particolare $A\mathbf{x} = \mathbf{e}_1$ utilizzando l'algoritmo di costo $O(n \log_2 n)$ illustrato nella sezione seguente, basato sulla ripetuta applicazione dei Lemmi 1 e 2.

2. Si calcola il prodotto matrice-vettore $A^{-1}\mathbf{f}$ effettuando non più di $O(n \log_2 n)$ operazioni aritmetiche utilizzando ad esempio la procedura vista nella Sezione 2.3.1.

2.3.4 Il calcolo della prima colonna dell'inversa di una matrice triangolare inferiore di Toeplitz

Per semplicità illustriamo l'algoritmo per il calcolo di $\mathbf{x}$ tale che $A\mathbf{x} = \mathbf{e}_1$ nel caso $n = 8$. Indicheremo, a volte, cos'è che cambia nel caso generale $n = 2^s, s \in \mathbb{N}$;

comunque tale caso è facilmente deducibile da quello considerato. L'algoritmo si divide in due parti. Nella prima parte si introducono e si calcolano matrici triangolari inferiori di Toeplitz che moltiplicate, una dopo l'altra, a sinistra per la matrice A, hanno l'effetto di trasformarla nella matrice identica. Nella seconda parte si moltiplicano tali matrici, di nuovo una dopo l'altra, per il vettore $\mathbf{e}_1$. Come si vedrà, non si fa altro che applicare una specie di eliminazione di Gauss, ma, invece di annullare colonne, si annullano diagonali. Il costo fnale $O(n \log_2 n)$ dell'algoritmo deriva dal fatto che ad ogni passo della prima parte si annullano metà delle diagonali rimaste non nulle, e dal fatto che la seconda parte è semplificabile sfruttando il fatto che il vettore $\mathbf{e}_1$ ha solo una componente non nulla.
Per prima cosa osserviamo che la matrice A 8×8 può essere vista come la sottomatrice in alto a sinistra di una matrice $L(\mathbf{a})$ semi-infinita triangolare inferiore di Toeplitz con prima colonna $[1\, a_1\, a_2 \cdots a_7\, a_8\, .]^T$.

Passo 1. Trovare $\hat{\mathbf{a}}$ tale che

$$L(\mathbf{a})\hat{\mathbf{a}} = \begin{bmatrix} 1 & & & & & & & \\ a_1 & 1 & & & & & & \\ a_2 & a_1 & 1 & & & & & \\ a_3 & a_2 & a_1 & 1 & & & & \\ a_4 & a_3 & a_2 & a_1 & 1 & & & \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & \\ a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \\ a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 \\ . & . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} 1 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ . \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ a_1^{(1)} \\ 0 \\ a_2^{(1)} \\ 0 \\ a_3^{(1)} \\ 0 \\ . \end{bmatrix} = E\mathbf{a}^{(1)}$$

per certi $a_i^{(1)} \in \mathbb{C}$ e calcolare tali $a_i^{(1)}$. Il calcolo degli $a_i^{(1)}$ richiede, una volta noto $\hat{\mathbf{a}}$, il prodotto di una matrice triangolare inferiore di Toeplitz 8×8 ($2^s \times 2^s$) per un vettore - o, più precisamente, due prodotti t.i.T. 4×4 ($2^{s-1} \times 2^{s-1}$) per vettore (perché?). Vedremo che $\hat{\mathbf{a}}$ è disponibile a costo zero. Notiamo che allora, per il Lemma 1, si ha $L(\hat{\mathbf{a}})L(\mathbf{a}) = L(E\mathbf{a}^{(1)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con una nulla.
Passo 2. Trovare $\hat{\mathbf{a}}^{(1)}$ tale che

$$L(E\mathbf{a}^{(1)})E\hat{\mathbf{a}}^{(1)} = \begin{bmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ a_1^{(1)} & 0 & 1 & & & & & \\ 0 & a_1^{(1)} & 0 & 1 & & & & \\ a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & & & \\ 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & & \\ a_3^{(1)} & 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & \\ 0 & a_3^{(1)} & 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 \\ . & . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \hat{a}_1^{(1)} \\ 0 \\ \hat{a}_2^{(1)} \\ 0 \\ \hat{a}_3^{(1)} \\ 0 \\ . \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ a_1^{(2)} \\ 0 \\ 0 \\ 0 \\ . \end{bmatrix} = E^2\mathbf{a}^{(2)}$$

per certi $a_i^{(2)} \in \mathbb{C}$ e calcolare tali $a_i^{(2)}$. Il calcolo degli $a_i^{(2)}$ richiede, una volta noto $\hat{\mathbf{a}}^{(1)}$, il prodotto di una matrice t.i.T. 4×4 ($2^{s-1} \times 2^{s-1}$ per un vettore - o, più precisamente, due prodotti t.i.T 2×2 ($2^{s-2} \times 2^{s-2}$) per vettore. Notiamo che allora, per il Lemma 1, si ha $L(E\hat{\mathbf{a}}^{(1)})L(E\mathbf{a}^{(1)}) = L(E^2 a^{(2)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con tre nulle. Si noti anche che, per il Lemma 2, se $L(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E\mathbf{a}^{(2)}$ allora $L(E\mathbf{a}^{(1)})E\hat{\mathbf{a}}^{(1)} = E^2\mathbf{a}^{(2)}$. Vedremo che $\hat{\mathbf{a}}^{(1)}$ tale che $L(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E\mathbf{a}^{(2)}$ è disponibile a costo zero.

Passo 3. Trovare $\hat{\mathbf{a}}^{(2)}$ tale che

$$L(E^2\mathbf{a}^{(2)})E^2\hat{\mathbf{a}}^{(2)} = \begin{bmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 1 & & & & \\ a_1^{(2)} & 0 & 0 & 0 & 1 & & & \\ 0 & a_1^{(2)} & 0 & 0 & 0 & 1 & & \\ 0 & 0 & a_1^{(2)} & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & a_1^{(2)} & 0 & 0 & 0 & 1 \\ . & . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \hat{a}^{(2)} \\ 0 \\ 0 \\ 0 \\ . \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ . \end{bmatrix} = E^3\mathbf{a}^{(3)}$$

per certi $a_i^{(3)} \in \mathbb{C}$ e calcolare tali $a_i^{(3)}$. Il calcolo degli $a_i^{(3)}$ richiede, una volta noto $\hat{\mathbf{a}}^{(2)}$, il prodotto di una matrice t.i.T. 2×2 ($2^{s-2} \times 2^{s-2}$) per un vettore - o, più precisamente, due prodotti t.i.T. 1×1 ($2^{s-3} \times 2^{s-3}$) per vettore. Cioè, nessuna operazione nel nostro caso $n = 8$, in cui nessun $a_i^{(3)}$, $i \geq 1$, è richiesto.

Notiamo che allora per il Lemma 1 si ha $L(E^2\hat{\mathbf{a}}^{(2)})L(E^2\mathbf{a}^{(2)}) = L(E^3\mathbf{a}^{(3)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con sette nulle. Si noti anche che per il Lemma 2, se $L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E\mathbf{a}^{(3)}$ allora $L(E^2\mathbf{a}^{(2)})E^2\hat{\mathbf{a}}^{(2)} = E^3\mathbf{a}^{(3)}$. Vedremo che $\hat{\mathbf{a}}^{(2)}$ tale che $L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E\mathbf{a}^{(3)}$ è disponibile a costo zero.

Se $n = 2^s > 8$, allora si continua. Altrimenti è completata la prima parte dell'algoritmo. Riassumendo, abbiamo provato che

$$L(E^2\hat{\mathbf{a}}^{(2)})L(E^1\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})L(\mathbf{a}) = L(E^3\mathbf{a}^{(3)}) \tag{2.1}$$

dove, osserviamo, le sottomatrici 8×8 in alto a sinistra di $L(\mathbf{a})$ e $L(E^3\mathbf{a}^{(3)})$ sono, rispettivamente, la matrice triangolare inferiore di Toeplitz data inizialmente A e la matrice identica I,

$$L(\mathbf{a}) = \begin{bmatrix} 1 & & & & & \\ a_1 & 1 & & & & \\ . & . & . & & & \\ a_7 & . & a_1 & 1 & & \\ a_8 & a_7 & . & a_1 & 1 & \\ . & . & . & . & . & . \end{bmatrix}, L(E^3\mathbf{a}^{(3)}) = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ . & . & . & & & \\ 0 & . & 0 & 1 & & \\ a_1^{(3)} & 0 & . & 0 & 1 & \\ . & . & . & . & . & . \end{bmatrix}.$$

Le operazioni che abbiamo dovuto fare finora sono state: 8×8 t.i.T. $\cdot$ vettore $+4 \times 4$ t.i.T. $\cdot$ vettore (se A è $n \times n$ con $n = 2^s$ le operazioni da farsi sarebbero state: $2^s \times 2^s$ t.i.T. $\cdot$ vettore $+ \ldots + 4 \times 4$ t.i.T. $\cdot$ vettore). Veniamo ora al nostro scopo, calcolare la prima colonna di A^{-1}, e illustriamo quindi la seconda parte dell'algoritmo. Consideriamo il seguente sistema lineare semi-infinito:

$$L(\mathbf{a})\mathbf{z} = E^2\mathbf{v}$$

dove v è un generico vettore semi-infinito di $\mathbb{C}^{\mathbb{N}}$ (se A è $n \times n$ con $n = 2^s$, allora la matrice E va elevata a $s-1$ e non a 2). Tale sistema può essere riscritto come segue

$$\begin{bmatrix} A & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ z_8 \\ \cdot \end{bmatrix} = \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ v_1 \\ 0 \\ 0 \\ 0 \\ v_2 \\ \cdot \end{bmatrix}$$

cioè evidenziando la parte superiore del sistema, di sole 8 equazioni. Prima di procedere, si noti che $\{\mathbf{z}\}_8$ è tale che $A\{\mathbf{z}\}_8 = [v_0\, 0\, 0\, 0\, v_1\, 0\, 0\, 0]^T, v_0,\, v_1 \in \mathbb{C}$. Quindi la scelta $v_0 = 1$ e $v_1 = 0$, renderebbe $\{\mathbf{z}\}_8$ uguale al vettore da noi cercato, $A^{-1}e_1$. Si dimostra immediatamente che il sistema $L(\mathbf{a})\mathbf{z} = E^2\mathbf{v}$ è equivalente al seguente sistema :

$$\begin{bmatrix} I_8 & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ \vdots \end{bmatrix} = L(E^3\mathbf{a}^{(3)})\mathbf{z} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})E^2\mathbf{v}.$$

Per il Lemma 2 il secondo membro di quest'ultima uguaglianza può essere riscritto più convenientemente:

$$L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})E^2\mathbf{v} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})E^2L(\hat{\mathbf{a}}^{(2)})\mathbf{v} = L(\hat{\mathbf{a}})EL(\hat{\mathbf{a}}^{(1)})EL(\hat{\mathbf{a}}^{(2)})\mathbf{v}$$

Quindi, vale la seguente identità:

$$\begin{bmatrix} I_8 & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ \vdots \end{bmatrix} = L(\hat{\mathbf{a}})EL(\hat{\mathbf{a}}^{(1)})EL(\hat{\mathbf{a}}^{(2)})\mathbf{v}.$$

Le matrici coinvolte nella rappresentazione a secondo membro sono triangolari inferiori e le sottomatrici quadrate di E in alto a sinistra, $8 \times 8, 4 \times 4$, hanno le colonne

sul lato destro nulle,

$$\{E\}_4 = \left[\begin{array}{cc||cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \{E\}_8 = \left[\begin{array}{cccc||cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Queste due osservazioni ci permettono di ottenere una efficiente rappresentazione di $\{\mathbf{z}\}_8$:

$$\{\mathbf{z}\}_8 == \{L(\hat{\mathbf{a}})\}_8\{E\}_8\{L(\hat{\mathbf{a}}^{(1)})\}_8\{E\}_8\{L(\hat{\mathbf{a}}^{(2)})\}_8\{\mathbf{v}\}_8 =$$

$$= \{L(\hat{\mathbf{a}})\}_8\{E\}_{8,4}\{L(\hat{\mathbf{a}}^{(1)})\}_4\{E\}_{4,2}\{L(\hat{\mathbf{a}}^{(2)})\}_2\{\mathbf{v}\}_2$$

valida quando $v_i = 0$, $i > 1$. Usando tale formula, quando $v_0 = 1$, $v_1 = 0$, il vettore $\{\mathbf{z}\}_8$ può essere calcolato effettuando 4×4 t.i.T. $\cdot$ vettore $+$ 8×8 t.i.T. $\cdot$ vettore (se A è $n \times n$ con $n = 2^s$ le operazioni da farsi sarebbero state: 4×4 t.i.T. $\cdot$ vettore $+ \cdots + 2^s \times 2^s$ t.i.T. $\cdot$ vettore), ovvero allo stesso costo dell'eliminazione Gaussiana, la prima parte dell'algoritmo.

In conclusione, se $cj2^j$ è un limite superiore per il costo dell'operazione $2^j \times 2^j$ t.i.T. $\cdot$ vettore, allora il costo dell'algoritmo illustrato, nel caso di A $n \times n$ con $n = 2^s$, è $c\sum_{j=2}^{s} j2^j = O(s2^s) = O(n \log_2 n)$.

Ci resta da provare che, effettivamente, si ha $\hat{\mathbf{a}}$ tale che $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ a costo zero. A tal fine è sufficiente osservare con un calcolo diretto che

$$\begin{bmatrix} 1 & & & & & & & & & & \\ a_1 & 1 & & & & & & & & & \\ a_2 & a_1 & 1 & & & & & & & & \\ a_3 & a_2 & a_1 & 1 & & & & & & & \\ a_4 & a_3 & a_2 & a_1 & 1 & & & & & & \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & & & \\ a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & & \\ a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & \\ a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & \\ a_9 & a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \\ . & . & . & . & . & . & . & . & . & . & . \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ -a_1 \\ a_2 \\ -a_3 \\ a_4 \\ -a_5 \\ a_6 \\ -a_7 \\ a_8 \\ -a_9 \\ . \end{bmatrix} =$$

$$\begin{bmatrix} 1 \\ 0 \\ 2a_2 - a_1^2 \\ 0 \\ 2a_4 - 2a_1a_3 + a_2^2 \\ 0 \\ 2a_6 - 2a_1a_5 + 2a_2a_4 - a_3^2 \\ 0 \\ 2a_8 - 2a_1a_7 + 2a_2a_6 - 2a_3a_5 + a_4^2 \\ \vdots \end{bmatrix},$$

$$L(\mathbf{a})\left(\mathbf{e}_1 + \sum_{i=1}^{+\infty}(-1)^i a_i \mathbf{e}_{i+1}\right) = \mathbf{e}_1 + \sum_{i=1}^{+\infty}\delta_{imod2,0}\left(2a_i + \sum_{j=1}^{i-1}(-1)^j a_j a_{i-j}\right)\mathbf{e}_{i+1}.$$

2.3.5 Osservazioni finali

Dato il vettore $\mathbf{a}$ il problema del calcolo di un vettore $\hat{\mathbf{a}}$ tale che $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$, per qualche $\mathbf{a}^{(1)}$, è in realtà un problema di aritmetica polinomiale. Infatti, per

il Lemma 1, l'identità $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ è equivalente all'uguaglianza $L(\mathbf{a})L(\hat{\mathbf{a}}) = L(E\mathbf{a}^{(1)})$, i.e.

$$(\sum_{k=0}^{+\infty} a_k Z^k)(\sum_{k=0}^{+\infty} \hat{a}_k Z^k) = \sum_{k=0}^{+\infty} a_k^{(1)} Z^{2k}.$$

Quindi il problema di aritmetica polinomiale in questione è il seguente: dato il polinomio $a(z) = \sum_{k=0}^{+\infty} a_k z^k$, trovare un polinomio $\hat{a}(z) = \sum_{k=0}^{+\infty} \hat{a}_k z^k$ tale che

$$\hat{a}(z)a(z) = a_0^{(1)} + a_1^{(1)} z^2 + a_2^{(1)} z^4 + \cdots =: a^{(1)}(z^2)$$

per certi $a_i^{(1)}$.

Tale problema è un caso particolare di un problema più generale: trasformare un polinomio pieno $a(z)$ in un polinomio sparso $a^{(1)}(z^k) = \sum_{k=0}^{+\infty} a_k^{(1)} z^{rk}$. È possibile scrivere esplicitamente un polinomio $\hat{a}(z)$ che realizza questa trasformazione, infatti vale il seguente teorema. Infatti vale il seguente teorema [Pagano].

Teorema 2.3.3. *Dato $a(z) = \sum_{k=0}^{+\infty} a_k z^k$, posto $\hat{a}(z) = a(zt)a(zt^2)\cdots a(zt^{r-1})$ dove t è una radice r-esima principale dell'unità ($t \in \mathbb{C}, t^r = 1, t^i \neq 1$ per $0 < i < r$), si ha che*

$$\hat{a}(z)a(z) = a_0^{(1)} + a_1^{(1)} z^r + a_2^{(1)} z^{2r} + \cdots =: a^{(1)}(z^r)$$

per certi $a_i^{(1)}$. Inoltre, se i coefficienti di a sono reali allora anche i coefficienti di $\hat{a}$ sono reali.

L'osservazione alla fine della sezione precedente si ottiene immediatamente dal Teorema 2.3.3, ponendo $r = 2 : \hat{a}(z) = a(-z)$. È evidente che $a(-z)a(z) = a_0^{(1)} + a_1^{(1)} z^2 + a_2^{(1)} z^4 + \cdots$, e che in tal caso i coefficienti di $\hat{a}$ sono disponibili a costo zero, occorre calcolare solo gli $a_i^{(1)}$.

Capitolo 3

Algebre sdU, circolanti e τ

3.1 Algebre di matrici simultaneamente diagonalizzate da trasformate discrete veloci unitarie

Ad ogni matrice unitaria M corrisponde l'algebra di matrici $\mathcal{L} = \{MDM^{-1}\}$, chiusa per inversione e per trasposizione coniugata. Se M è reale, le matrici di $\mathcal{L}$ sono simmetriche ed esiste una base per $\mathcal{L}$ costituita da matrici reali. Vedremo presto che quest'ultima affermazione può essere vera anche se M non è reale. Le algebre $\mathcal{L}$ descritte nel seguito sono quelle corrispondenti a trasformate unitarie M non reali di tipo Fourier, e reali di tipo Hartley e trigonometriche. Altre algebre $\mathcal{L}$, corrispondenti ad altre trasformate unitarie, sono di interesse. Menzioniamo in particolare quelle corrispondenti alle trasformazioni di Householder, che tra l'altro sono di complessità computazionale minima $O(n)$.

3.1.1 Le algebre $\mathcal{C}, \mathcal{C}_{-1}, \mathcal{C}_\phi$, e la trasformata discreta di Fourier (DFT)

Sia Π la matrice circolante $n \times n$, con prima riga $[0\,1\,0\,\cdots\,0]$. Sia e il vettore di $\mathbb{C}^n$ le cui componenti sono tutte uguali a 1. Allora $\Pi \boldsymbol{e} = \boldsymbol{e} = 1\boldsymbol{e}$. Inoltre, per $w \in \mathbb{C}$ si ha

$$\Pi \begin{bmatrix} 1 \\ w \\ w^2 \\ . \\ w^{n-1} \end{bmatrix} = \begin{bmatrix} w \\ w^2 \\ . \\ w^{n-1} \\ 1 \end{bmatrix} = w \begin{bmatrix} 1 \\ w \\ w^2 \\ . \\ w^{n-1} \end{bmatrix}$$

dove l'ultima uguaglianza vale se $w^n = 1$. Più in generale, se $w^n = 1$ valgono le seguenti identità vettoriali:

$$\Pi \begin{bmatrix} 1 \\ w^j \\ . \\ w^{(n-1)j} \end{bmatrix} = \begin{bmatrix} w^j \\ . \\ w^{(n-1)j} \\ 1 \end{bmatrix} = w^j \begin{bmatrix} 1 \\ w^j \\ . \\ w^{(n-1)j} \end{bmatrix}, \quad j = 0, 1, ..., n-1$$

che, messe insieme, diventano l'identità matriciale $\Pi W = W D_{1w^{n-1}}$ coinvolgente le matrici $D_{1w^{n-1}}$ e W definite qui di seguito:

$$D_{1w^{n-1}} = \begin{bmatrix} 1 & & & \\ & w & & \\ & & . & \\ & & & w^{n-1} \end{bmatrix},$$

$$W = \begin{bmatrix} 1 & 1 & . & 1 & . & 1 \\ 1 & w & . & w^j & . & w^{n-1} \\ . & . & . & . & . & . \\ 1 & w^{n-1} & . & w^{(n-1)j} & . & w^{(n-1)(n-1)} \end{bmatrix}.$$

Scegliendo w anche tale che $w^j \neq 1, 0 < j < n$ (ovvero w radice n-esima principale di 1), la matrice diagonale $D_{1w^{n-1}}$ viene ad avere sulla diagonale tutti gli autovalori di Π, che risultano dunque distinti, e le colonne della matrice W vengono ad essere corrispondenti autovettori, unitariamente ortogonali.

Un risultato più completo è riportato nella seguente.

Proposizione 3.1.1. *Sia $w \in \mathbb{C}$ tale che $w^n = 1, w^j \neq 1$ per $0 < j < n$, e $W \in \mathbb{C}^{n\times n}$ la matrice $W = (w^{(i-1)(j-1)})_{i,j=1}^n$. Allora $W^H W = nI$.*

Dimostrazione. Poichè $|w| = 1, \overline{w} = w^{-1}$, si ha $[W^H W]_{ij} = [\overline{W} W]_{ij} = \sum_{k=1}^n [\overline{W}]_{ik}[W]_{kj} = \sum_{k=1}^n \overline{w}^{(i-1)(k-1)} w^{(k-1)(j-1)} = \sum_{k=1}^n w^{(k-1)(j-i)} = \sum_{k=1}^n (w^{j-i})^{k-1}$.

Quindi $[W^H W]_{ij} = n$ se $i = j$, e $[W^H W]_{ij} = \frac{1-(w^{j-i})^n}{1-w^{j-i}} = 0$ se $i \neq j$ (si noti che l'ipotesi $w^j \neq 1$ per $0 < j < n$ è essenziale per rendere $1 - w^{j-i} \neq 0$ se $j \neq i$). □

Da ora in poi si suppone w tale che $w^n = 1, w^j \neq 1$ per $0 < j < n$. Per il risultato della Proposizione, possiamo dunque dire che la seguente matrice (simmetrica) *di Fourier*

$$F = \frac{1}{\sqrt{n}} W, \quad W = (w^{(i-1)(j-1)})_{i,j=1}^n, \quad w^n = 1, \quad w^j \neq 1, \quad 0 < j < n \qquad (3.1)$$

è unitaria, i.e. $F^H F = I$. Si può provare che $F^2 = J\Pi$ (J è la matrice di permutazione $Je_k = e_{n+1-k}, k = 1, ...n$. Ne segue che F^H si ottiene da F permutando le sue colonne (righe), infatti $F^H = J\Pi F = FJ\Pi$. Si dimostra anche che $F^4 = I$, quindi se λ è autovalore di F allora $\lambda \in \{1, -1, i, -i\}$, essendo i l'unità immaginaria.

L'identità matriciale soddisfatta da Π e da W può essere ovviamente riscritta in termini di F, cioè si ha che $\Pi F = F D_{1w^{n-1}}$. Quindi otteniamo l'uguaglianza $\Pi = F D_{1w^{n-1}} F^H$ da cui segue che la matrice di Fourier diagonalizza la matrice Π ($F^H \Pi F$ è diagonale), o, più precisamente, che le colonne della matrice di Fourier formano un sistema di n autovettori unitariamente ortonormali per la matrice Π con corrispondenti autovalori $1, w, ...w^{n-1}$, essendo w una radice n-esima principale dell'unità. Ma se F diagonalizza Π, allora diagonalizza tutti i polinomi in Π, ovvero tutte le matrici circolanti $n \times n$. Più precisamente, chiamata $\mathcal{C}(a)$ la matrice circolante con la prima riga a^T , $\mathcal{C}(a) = \sum_{k=1}^{n} a_k \Pi^{k-1}$, si ha che $\mathcal{C}(a) = \sum_{k=1}^{n} a_k (F D_{1w^{n-1}} F^H)^{k-1} = F(\sum_{k=1}^{n} a_k D_{1w^{n-1}}^{k-1}) F^H = F diag(\sum_{k=1}^{n} a_k w^{(j-1)(k-1)}, j = 1, ...n)) F^H$. Quindi, $\mathcal{C}(a) = F d(Wa) F^H = \sqrt{n} F d(Fa) F^H = F d(F^T a) d(F^T e_1)^{-1} F^H$, dove $a = \mathcal{C}(a)^T e_1$ e $d(v) = diag(v_i)$. Allora $F^H \mathcal{C}(a) F = d(F^T a) d(F^T e_1)^{-1}$

Sia Π_{-1} la matrice $n \times n$ (-1)-circolante con prima riga [010...0]:

$$\Pi_{-1} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 1 \\ -1 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

Procedendo analogamente al caso circolante si trova una matrice M unitaria tale che $\mathcal{C}_{-1} = \{p(\Pi_{-1})\} = \{MDM^{-1} : D = \text{diagonali}\}$. (Nota: M si ottiene da F moltiplicando le sue righe, dalla prima alla n-esima, rispettivamente per $1, \rho, ..., \rho^{n-1}$, dove ρ è una radice principale n-esima di -1). Più in generale, posto $\Pi_\phi = Z^T + \phi e_n e_1^T, \phi \in \mathbb{C}$, e considerato l'insieme $\mathcal{C}_\phi$ dei polinomi in Π_ϕ, si dimostra [Bertaccini, Di Fiore, Zellini] che $\mathcal{C}_\phi = \{MDM^{-1} : D = \text{diagonali}\}$ con $M = D_\phi F$ per una opportuna matrice diagonale D_ϕ, e che M è unitaria se e solo se $|\phi| = 1$. Si osserva quindi che $\mathcal{C}_\phi(a) = \sum_{k=1}^{n} a_k \Pi_\phi^{k-1}$, la matrice ϕ-circolante la cui prima riga è a^T, ammette la seguente rappresentazione:

$$\mathcal{C}_\phi(a) = D_\phi F d(F D_\phi a) d(F D_\phi e_1)^{-1} (D_\phi F)^{-1}.$$

Proposizione 3.1.2. *Dato $z \in \mathbb{C}^n$, la complessità del prodotto matrice-vettore Fz è al più $O(nlogn)$. Tale operazione è chiamata trasformata discreta di Fourier (DFT) di z. Come conseguenza, sia il vettore prodotto matrice-vettore $\mathcal{C}(a)z$ che la soluzione del sistema lineare $\mathcal{C}(a)x = f, f \in \mathbb{C}^n$, sono calcolabili effettuando due DFT (dopo il pre-calcolo della DFT Fa), e, quindi, con al più $O(n \log n)$ operazioni aritmetiche. Analoghe affermazioni valgono più in generale per $\mathcal{C}_\phi(a)$.*

Come conseguenza di questa Proposizione e della Sezione 2.3.1, il prodotto matrice di Toeplitz $n \times n$ per vettore è calcolabile con al più $O(n \log n)$ operazioni aritmetiche. Questo risultato ci ha permesso (nel precedente capitolo) di introdurre un metodo di costo $O(n \log n)$ per il calcolo della prima colonna dell'inversa di

una generica matrice triangolare inferiore di Toeplitz, ovvero per la risoluzione dei sistemi triangolari di Toeplitz. Non è invece noto un algoritmo che risolve sistemi di Toeplitz generici $Tx = f$ con al più $O(nlogn)$ operazioni aritmetiche, a meno che nel conto delle operazioni si omettano quelle fatte su T e non su f; questo è vero anche se si suppone T simmetrica (si vedano gli argomenti precondizionamento di matrici di Toeplitz e formule di dislocamento per l'inversa di matrici di Toeplitz, entrambi trattati più avanti). Più precisamente, il problema sta nel fatto che il calcolo del vettore (o dei vettori) che definiscono l'inversa di T richiede in generale più di $O(n \log n)$ operazioni aritmetiche, se T non è triangolare. Si veda ad esempio il caso, pur favorevole, in cui T è definita positiva, in cui un solo vettore (la prima colonna di T^{-1}, come nel caso triangolare!) definisce T^{-1}. In questo caso l'algoritmo migliore conosciuto finora per la risoluzione di $Tx = f$ richiede $O(n \log_2 n)$ operazioni aritmetiche [Ammar, Gragg], se T e' una generica matrice reale definita positiva di Toeplitz.

Dimostrazione. (Proposizione 3.3.2.) Sia n divisibile per 2. Poichè l'elemento (i, k) di W è $w^{(i-1)(k-1)}$ e l'elemento k di $z \in \mathbb{C}^n$ è z_k, si ha
$(Wz)_i = \sum_{k=1}^{n} w^{(i-1)(k-1)} z_k = \sum_{j=1}^{n/2} w^{(i-1)(2j-2)} z_{2j-1} + \sum_{j=1}^{n/2} w^{(i-1)(2j-1)} z_{2j}$
$= \sum_{j=1}^{n/2} (w^2)^{(i-1)(j-1)} z_{2j-1} + \sum_{j=1}^{n/2} w^{(i-1)(2(j-1)+1)} z_{2j}$
$= \sum_{j=1}^{n/2} (w^2)^{(i-1)(j-1)} z_{2j-1} + w^{i-1} \sum_{j=1}^{n/2} (w^2)^{(i-1)(j-1)} z_{2j}, \quad i = 1, \ldots, n.$

Si noti che w è di fatto una funzione di n, cioè la giusta notazione per w dovrebbe essere w_n. Allora $w^2 = w_n^2$ è tale che $(w_n^2)^{n/2} = 1$ e $(w_n^2)^i \neq 1, 0 < i < n/2$; in altre parole $w_n^2 = w_{n/2}$ (ovvero w_n^2 è radice $n/2$-esima principale di 1). Quindi, abbiamo le identità

$$(W_n z)_i = \sum_{j=1}^{n/2} w_{n/2}^{(i-1)(j-1)} z_{2j-1} + w_n^{i-1} \sum_{j=1}^{n/2} w_{n/2}^{(i-1)(j-1)} z_{2j}, \quad i = 1,2,...,n. \tag{3.2}$$

Ne segue che, per $i = 1,2,...n/2$,

$$(W_n z)_i = (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ . \\ z_{n-1} \end{bmatrix})_i + w_n^{i-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ . \\ z_n \end{bmatrix})_i \quad .$$

Inoltre, ponendo $i = n/2+k, k = 1,2,...,n/2$, in (3.2), poiché $w_n^{n/2} = -1$ otteniamo $(W_n z)_{n/2+k} = \sum_{j=1}^{n/2} w_{n/2}^{n/2(j-1)} w_{n/2}^{(k-1)(j-1)} z_{2j-1} + w_n^{n/2} w_n^{k-1} \sum_{j=1}^{n/2} w_{n/2}^{n/2(j-1)} w_{n/2}^{(k-1)(j-1)} z_{2j}$

$= \sum_{j=1}^{n/2} w_{n/2}^{(k-1)(j-1)} z_{2j-1} - w_n^{k-1} \sum_{j=1}^{n/2} w_{n/2}^{(k-1)(j-1)} z_{2j}$

$= (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ . \\ z_{n-1} \end{bmatrix})_k - w_n^{k-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ . \\ z_n \end{bmatrix})_k, \quad k = 1, ..., n/2$

Quindi, si ha il seguente risultato

$$W_n z = \begin{bmatrix} I & D_{1w_n^{n/2-1}} \\ I & -D_{1w_n^{n/2-1}} \end{bmatrix} \begin{bmatrix} W_{n/2} & O \\ O & W_{n/2} \end{bmatrix} Q_n z, \tag{3.3}$$

dove $D_{1w_n^{n/2-1}} = \begin{bmatrix} 1 & & & \\ & w_n & & \\ & & . & \\ & & & w_n^{n/2-1} \end{bmatrix}$ e Q_n è la matrice di permutazione $Q_n z = [z_1 \quad z_3 \quad . \quad z_{n-1} \quad z_2 \quad z_4 \quad . \quad z_n]^T$. Per la formula (3.3), che rappresenta W_n in termini di due matrici $W_{n/2}$, se c_n denota la complessità del prodotto matrice-vettore $F_n z, F_n = \frac{1}{\sqrt{n}} W_n$, allora $c_n \leq 2c_{n/2} + rn$, r costante e questo implica $c_n = O(n \log_2 n)$, se n è una potenza di 2. Nel caso n sia divisibile non per 2 ma per $b > 2$, con un procedimento simile a quello visto sopra si ottiene una rappresentazione di W_n in termini di b matrici $W_{n/b}$. Se n è una potenza di b da tale rappresentazione si deduce un algoritmo per il calcolo di $W_n z$ di costo $O(n \log_b n)$. □

3.2 L'algebra delle matrici τ e la trasformata discreta seno (DST)

Ovviamente, ogni volta che una matrice $n \times n$ M unitaria definita per tutti gli n, soddisfa una identità del tipo

$$M_n = [\text{sparse matrix}] \begin{bmatrix} M_{n/b} & & \\ & . & \\ & & M_{n/b} \end{bmatrix} [\text{permutation matrix}] \tag{3.4}$$

con b divisore di n, si può dire che i prodotti matrice-vettore $M_n z$ e $M_n^{-1} z$ possono essere calcolati con al più $O(n \log_b n)$ operazioni aritmetiche, e, di conseguenza, l'algebra corrispondente ad M, $\mathcal{L} = \{MDM^{-1}\}$, acquista interesse perchè di bassa complessità computazionale. Come abbiamo già visto, l'identità di cui sopra è verificata per $M =$ trasformata di Fourier F, e, quindi, più in generale, per $M = D_\phi F$. Ma essa è verificata anche per altre matrici M.

L'algebra di matrici τ introdotta in [Bevilacqua, Capovani] e studiata nei dettagli in questa sezione, è solo una delle 16 note algebre trigonometriche o di Jacobi. Allo stesso modo, la trasformata seno, che diagonalizza τ, è solo una delle 16 note

trasformate discrete trigonometriche, di tipo seno e di tipo coseno. Tutte tali 16 trasformate hanno complessità $O(n \log_2 n)$.

Definiamo l'algebra τ introducendo una sua base. Si consideri la seguente matrice tridiagonale $n \times n$

$$Z + Z^T = \begin{bmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & 1 & 0 & . & & \\ & & . & . & 1 & \\ & & & 1 & 0 & \end{bmatrix} \quad . \tag{3.5}$$

Posto $J_1 = I$ e $J_2 = Z + Z^T$, si nota che $e_1^T J_1 = e_1^T$, $e_1^T J_2 = e_2^T$. Inoltre, poichè

$$(Z + Z^T)^2 = \begin{bmatrix} 1 & 0 & 1 & & & \\ 0 & 2 & 0 & 1 & & \\ 1 & 0 & 2 & . & . & \\ & 1 & . & . & . & 1 \\ & & & & 2 & 0 \\ & & & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & & & \\ 0 & 1 & 0 & 1 & & \\ 1 & 0 & 1 & . & . & \\ & 1 & . & . & . & 1 \\ & & . & . & 1 & 0 \\ & & & 1 & 0 & 0 \end{bmatrix} + I,$$

abbiamo $e_1^T((Z + Z^T)^2 - I) = [0\ 0\ 1\ 0\ \ldots 0] = e_3^T$. Si ponga allora $J_3 = (Z + Z^T)^2 - I = J_2(Z + Z^T) - J_1$; si ha $e_1^T J_3 = e_3^T$. Più in generale, si ponga $J_{i+1} = J_i(Z + Z^T) - J_{i-1}$, $i = 2, 3, \ldots, n$ - *1.* La matrice J_{i+1} è un polinomio in $Z + Z^T$ di grado i con la proprietà $e_1^T J_{i+1} = e_{i+1}^T$. Per dimostrare tale proprietà delle J_k, supponiamo di sapere che $e_1^T J_j = e_j^T$,
$j = 1, \ldots, i$ (ed effettivamente lo sappiamo per $j = 1, 2$); allora da ciò segue subito che

$$e_1^T J_{i+1} = e_1^T(J_i(Z + Z^T) - J_{i-1}) = (e_i^T(Z + Z^T)) - e_{i-1}^T = (e_{i-1}^T + e_{i+1}^T) - e_{i-1}^T = e_{i+1}^T.$$

Poichè le matrici $J_1, J_2, \ldots J_n$ sono linearmente indipendenti, possiamo dire che esse generano l'insieme $\{p(Z + Z^T)\}$ di tutti i polinomi nella matrice $Z + Z^T$. Tale insieme è l'algebra τ. Inoltre, poichè $e_1^T J_k = e_k^T$, ogni matrice di τ è univocamente definita dalla sua prima riga, e dati $a_k \in \mathbb{C}$ la matrice di τ la cui prima riga è $a^T = [a_1 \ldots a_n]$ è $\tau(a) = \sum_{k=1}^n a_k J_k$. Troviamo una rappresentazione della matrice $\tau(a)$ in termini dei suoi autovalori e autovettori.
Innanzitutto si osserva che valgono le seguenti uguaglianze vettoriali:

$$\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & . & . & \\ & & . & 0 & 1 \\ & & & 1 & 0 \end{bmatrix} \begin{bmatrix} \sin \frac{j\pi}{n+1} \\ \sin \frac{2j\pi}{n+1} \\ . \\ \sin \frac{nj\pi}{n+1} \end{bmatrix} = 2 \cos \frac{j\pi}{n+1} \begin{bmatrix} \sin \frac{j\pi}{n+1} \\ \sin \frac{2j\pi}{n+1} \\ . \\ \sin \frac{nj\pi}{n+1} \end{bmatrix}, \quad j = 1, ..., n.$$

Tali n uguaglianze possono essere riscritte come una unica uguaglianza matriciale $(Z + Z^T)S = SD$ dove S è la matrice

$$S_{ij} = \sqrt{\frac{2}{n+1}} \sin \frac{ij\pi}{n+1}, \quad i, j = 1, ..., n. \tag{3.6}$$

e D è la matrice diagonale con elementi diagonali distinti $D_{jj} = 2\cos\frac{j\pi}{n+1}, j = 1, ..., n$. Si noti che la matrice S, chiamata matrice seno, è reale, simmetrica e unitaria. Inoltre se $F_{2(n+1)}$ è la matrice di Fourier di ordine $2(n+1)$, allora:

$$J(I - F^2_{2(n+1)})F_{2(n+1)} = \begin{bmatrix} 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & S & \mathbf{0} & -SJ \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & -JS & \mathbf{0} & JSJ \end{bmatrix}$$

(si noti che $F^2_{2(n+1)}$ è una matrice di permutazione). Come conseguenza, una trasformata seno può essere calcolata effettuando una trasformata discreta di Fourier, ovvero con al più $O(n \log n)$ operazioni aritmetiche. Le colonne della matrice seno S formano un sistema di autovettori unitariamente ortonormali per la matrice $Z + Z^T$. In altre parole, la matrice unitaria S diagonalizza $Z + Z^T$ e, ovviamente, diagonalizza ogni polinomio in $Z + Z^T$, i.e. ogni matrice di τ. Riassumendo, si ha che

$$Z + Z^T = SDS,\ J_k = p_{k-1}(Z + Z^T),\ \tau = \{S(\sum_{k=1}^{n} a_k p_{k-1}(D))S : a_k \in \mathbb{C}\} = sdS$$

ed è evidente che la matrice di τ con prima riga a^T ammette la seguente rappresentazione

$$\tau(a) = \sum_{k=1}^{n} a_k J_k = Sd(Sa)d(Se_1)^{-1}S.$$

Da questa formula per $\tau(a)$ segue che prodotti matrice-vettore coinvolgenti matrici τ e la risoluzione di sistemi lineari con matrici dei coefficienti in τ hanno complessità al più $O(n \log n)$. Concludiamo con una osservazione utile per capire rapidamente la struttura delle matrici di τ . Per il Teorema 2.2.1 ogni matrice A dello spazio τ deve commutare con la matrice $Z + Z^T$, o, più precisamente

$$\tau = \{A \in \mathbb{C}^{n\times n} : A(Z + Z^T) = (Z + Z^T)A\}.$$

Ma ciò è equivalente a richiedere che gli elementi di A soddisfano le seguenti n^2 condizioni di "somma in croce"

$$a_{i,j-1} + a_{i,j+1} = a_{i-1,j} + a_{i+1,j}, \quad i, j = 1, ..., n,$$

dove si suppone $a_{0,j} = a_{n+1,j} = a_{i,0} = a_{i,n+1} = 0, i, j = 1, ..., n$. Possiamo usare tali condizioni e il fatto che le matrici di τ sono sia simmetriche che persimmetriche per scrivere $\tau(a)$, per un generico vettore $a = [a_1 a_2 ... a_n]^T$. Per esempio, per n = 4 ed n = 5,

$$\tau(a) = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & a_1 + a_3 & a_2 + a_4 & a_3 \\ a_3 & a_2 + a_4 & a_1 + a_3 & a_2 \\ a_4 & a_3 & a_2 & a_1 \end{bmatrix}, \tau(a) = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ a_2 & a_1 + a_3 & a_2 + a_4 & a_3 + a_5 & a_4 \\ a_3 & a_2 + a_4 & a_1 + a_3 + a_5 & a_2 + a_4 & a_3 \\ a_4 & a_3 + a_5 & a_2 + a_4 & a_1 + a_3 & a_2 \\ a_5 & a_4 & a_3 & a_2 & a_1 \end{bmatrix}$$

$$J_1 = I, J_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, J_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, J_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

e così via.

Per usare i risultati ottenuti sull'algebra τ, svolgiamo il seguente esercizio: provare che per n pari la matrice $J_2 = Z + Z^T$ è invertibile, e calcolare l'inversa.

Risoluzione. Sappiamo che se J_2 è invertibile, allora $J_2^{-1} \in \tau$ (dal fatto che J_2 commuta con $Z + Z^T$ segue che anche J_2^{-1} commuta con $Z + Z^T$). Quindi $J_2^{-1} = \tau(z)$ per qualche $z \in \mathbb{C}^n$, e la tesi equivale a far vedere che esiste z per cui $\tau(z)J_2 = I$. Ma l'identità matriciale $\tau(z)J_2 = I$ è equivalente all'identità vettoriale $z^T J_2 = e_1^T$. Così, per esempio, per $n = 4$ abbiamo la condizione

$$[z_1 z_2 z_3 z_4] \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = [1\ 0\ 0\ 0],$$

che implica $z_1 = 0$, $z_2 = 1$, $z_3 = 0$, $z_4 = -1$ e, quindi,

$$J_2^{-1} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix} = J_2 - J_4.$$

Studiando i casi $n = 6, 8, \cdots$ si puó poi dedurre la formula generale per J_2^{-1} (da trovare!).

Altri esercizi utili sull'algebra τ sono i seguenti:

- Sia T una matrice di Toeplitz simmetrica $n \times n$, i.e. $T = (t_{|i-j|})_{i,j=1}^n$, per certi $t_k \in \mathbb{C}$.

Mostrare che $T = A - B$ dove A è una matrice τ di ordine n e

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & 0 \end{bmatrix}, R \in \tau \cap \mathbb{C}^{(n-2)\times(n-2)}.$$

Risoluzione. Basta scegliere A con prima riga $[t_0\ t_1\ \cdots\ t_{n-1}]$ ed R con prima riga $[t_2\ \cdots\ t_{n-1}]$

- Scrivere l'inversa della matrice di τ la cui prima riga è $[4\,1\,0\,0\,\cdots\,0]$.
- Mostrare che la matrice $ee^T = (1)_{i,j=1}^n$ non appartiene a τ . Tuttavia esistono algebre n-dimensionali di bassa complessita' computazionale di tipo $\{p(X)\}$ che contengono la matrice ee^T (vedi [Di Fiore, 2000], [Di Fiore, Zellini, 2001])

Capitolo 4

Algebre sdU nella risoluzione di sistemi di Toeplitz e Toeplitz più Hankel

4.1 Algebre di Hessenberg nella risoluzione di sistemi di Toeplitz

In questa Sezione riprenderemo dei risultati pubblicati in [Di Fiore, Zellini, 1995, 1998]. Una matrice di Hessenberg è una matrice $X \in \mathbb{C}^{n\times n}$ della forma:

$$X = \begin{bmatrix} r_{11} & b_1 & & & \\ r_{21} & r_{22} & b_2 & & \\ . & . & . & . & \\ . & . & . & . & b_{n-1} \\ r_{n1} & r_{n2} & . & . & r_{nn} \end{bmatrix} \text{ dove } b_i \neq 0, \forall i.$$

Notiamo che le seguenti matrici:

$J_1 = X^0 = I, J_2 = \frac{1}{b_1}(X - r_{11}I), J_3 = \frac{1}{b_2}(J_2X - r_{21}I - r_{22}J_2), \ldots, J_n = \frac{1}{b_{n-1}}(J_{n-1}X - r_{n-1,1}I - r_{n-1,2}J_2 - \ldots - r_{n-1,n-1}J_{n-1})$
sono polinomi in X (di grado 0,1,2, ..., $n-1$), e hanno come prima riga, rispettivamente, $e_1^T, e_2^T, \ldots, e_n^T$.
Allora consideriamo l'insieme H_X di tutti i polinomi in X. Notiamo che la sua dimensione è $\leq n$, poichè per il teorema di Cayley-Hamilton X^n deve essere una combinazione lineare delle precedenti potenze di X. Ma esistono n matrici in H_X che sono linearmente indipendenti e queste sono le J_k. Ne segue che la dimensione di H_X è n.
Per il Teorema 2.2.1. vale che $H_X = \{p(X) : p = polinomio\} = Span\{J_1, J_2, \ldots, J_n\} =$

$\{A \in \mathbb{C}^{n\times n} : AX = XA,\}$. Inoltre, ogni matrice di H_X è univocamente determinata dalla prima riga, cioè dato un vettore $a \in \mathbb{C}^n$ esiste una unica matrice di H_X che ha come prima riga a^T:

$$H_X(a) = \sum_{k=1}^{n} a_k J_k, \quad e_1^T H_X(a) = a^T.$$

Lo spazio delle matrici H_X è chiuso rispetto alla moltiplicazione, quindi H_X è un' algebra, inoltre H_X è commutativa. In particolare, abbiamo che $J_k J_s = J_s J_k$ per ogni *k,s* e quindi:

$$e_i^T H_X(a) = e_i^T \sum_{k=1}^{n} a_k J_k = \sum_{k=1}^{n} a_k e_k^T J_i = a^T J_i, \quad \forall i.$$

Moltiplicando tutto per lo scalare v_i e sommando per $i = 1, \ldots, n$ otteniamo l'uguaglianza:

$$v^T H_X(a) = a^T H_X(v), \quad a, v \in \mathbb{C}.$$

Si nota che se la matrice $H_X(a)$ in H_X è invertibile, allora la sua inversa è un polinomio in X perché $AX = XA \Rightarrow A^{-1}X = XA^{-1}$, cioè esiste un vettore $z \in \mathbb{C}^n$ tale che $(H_X(a))^{-1} = H_X(z)$. Dall'uguaglianza $H_X(z)H_X(a) = I$ segue che z può essere determinato risolvendo il sistema lineare $z^T H_X(a) = e_1^T$.

Un esempio di Algebra di Hessenberg è lo spazio $\mathcal{C}_\epsilon$ delle matrici ϵ-circolanti. Un altro importante esempio di algebra di Hessenberg è l'algebra delle matrici τ. Quindi, $\mathcal{C}_\epsilon(a)$ e $\tau(a)$ sono rispettivamente la matrice ϵ-circolante e la matrice τ con prima riga a^T.
Vogliamo vedere ora un esempio di una formula generale, che coinvolge le algebre di Hessenberg, con la quale ogni matrice A si rappresenta come la somma di prodotti del tipo $M_i N_i$ dove M_i e N_i sono matrici in H_X e $H_{X'}$ dove $X' \approx X$. Il numero di addendi in questa somma è $\alpha + 1$ dove α è il rango di $AX - XA$. Tale formula per A puó essere molto utile se α non dipende dall'ordine n della matrice A.

Lemma 4.1.1. *Sia* $A \in \mathbb{C}^{n\times n}$. *Se* $AX - XA = \sum_{m=1}^{\alpha} x_m y_m^T \quad (x_m, y_m \in \mathbb{C}^n)$ *allora* $\sum_{m=1}^{\alpha} x_m^T p(X^T) y_m = 0$, *per ogni polinomio p.*

Dimostrazione. $\sum_{m=1}^{\alpha} x_m^T p(X^T) y_m = \sum_{m=1}^{\alpha} \sum_{i,j=1}^{n} (x_m)_i (p(X^T))_{ij} (y_m)_j =$

$= \sum_{i,j=1}^{n} \big(\sum_{m=1}^{\alpha} (x_m)_i (y_m)_j \big) (p(X^T))_{ij} =$

$= \sum_{i,j=1}^{n} (AX - XA)_{ij} (p(X^T))_{ij} = \sum_{i,j=1}^{n} \sum_k A_{ik} X_{kj} (p(X^T))_{ij} - \sum_{i,j=1}^{n} \sum_k X_{ik} A_{kj} (p(X^T))_{ij} =$

$= \sum_{i,k}^{n} A_{ik} \sum_j (p(X^T))_{ij} (X^T)_{jk} - \sum_{k,j=1}^{n} A_{kj} \sum_i (X^T)_{ki} (p(X^T))_{ij} =$

$= \sum_{i,k=1}^{n} A_{ik} (p(X^T) X^T)_{ik} - \sum_{k,j=1} A_{kj} (X^T p(X^T))_{kj} = 0.$ □

La formula che troveremo nel prossimo teorema coinvolge algebre di Hessenberg persimmetriche. Questa formula insieme al fatto che $rank(AX - XA) = 2$ per una matrice A di Toeplitz e per $X = \Pi_\epsilon$, sarà usato per ottenere una efficiente rappresentazione della matrice inversa di A dovuta ad Ammar e Gader.

Teorema 4.1.2. *Sia X una matrice di Hessenberg inferiore persimmetrica cioè simmetrica rispetto all'antidiagonale ($X^T = JXJ$), e definiamo $X' \in \mathbb{C}^{n\times n}$ e $\beta \in \mathbb{C}$ tale che vale l'identità:*

$$X = X' + (r_{n1} - \beta)e_n e_1^T.$$

Sia $A \in \mathbb{C}^{n\times n}$. Se $AX - XA = \sum_{m=1}^{\alpha} x_m y_m^T \quad (x_m, y_m^T \in \mathbb{C}^n)$, allora

$$(r_{n1} - \beta)A = -\sum_{m=1}^{\alpha} H_X(\hat{x}_m)H_{X'}(y_m) + (r_{n1} - \beta)H_X(JAe_n)$$

(dove $\hat{v} = Jv$ con $v \in \mathbb{C}^n$).

Dimostrazione. Poniamo $(\star) = -\sum_{m=1}^{\alpha} H_X(\hat{x}_m)H_{X'}(y_m)$. Allora

$$(\star)X - X(\star) = -\sum_{m=1}^{\alpha} H_X(\hat{x}_m)[H_{X'}(y_m)X - XH_{X'}(y_m)] =$$

$$= -(r_{n1} - \beta)\sum_{m=1}^{\alpha} H_X(\hat{x}_m)[H_{X'}(y_m)e_n e_1^T - e_n e_1^T H_{X'}(y_m)] =$$

$$= -(r_{n1} - \beta)\sum_{m=1}^{\alpha} H_X(\hat{x}_m)[\hat{y}_m e_1^T - e_n y_m^T] = (r_{n1} - \beta)\sum_{m=1}^{\alpha} x_m y_m^T =$$

$$= (r_{n1} - \beta)(AX - XA).$$

Quindi, $(r_{n1} - \beta)A - (\star)$ deve commutare con X, cioè deve essere un polinomio in X:
$(r_{n1}-\beta)A-(\star) = H_X(z), z \in {}^n$. Ora imponendo che l'ultima colonna nella matrice a sinistra sia uguale all'ultima colonna della matrice a destra, si ottiene che il vettore z è definito dall'identità:
$Jz = (r_{n1} - \beta)Ae_n$.
Si nota che nella dimostrazione abbiamo usato due volte il fatto che
$\sum_{m=1}^{\alpha} H_X(\hat{x}_m)\hat{y}_m = 0$.
In effetti:

$$e_i^T\big(\sum_{m=1}^{\alpha} H_X(\hat{x}_m)\hat{y}_m\big) = \sum_{m=1}^{\alpha} \hat{x}_m^T J_i \hat{y}_m = \sum_{m=1}^{\alpha} x_m^T J_i^T y_m$$

($J_i = H_X(e_i)$) e l'ultima quantità è 0 per il Lemma 4.1.1 poiché J_i è un polinomio in X.

□

Abbiamo visto che la moltiplicazione di T di Toeplitz per un vettore si può effettuare con $O(n \log_2 n)$ operazioni aritmetiche. Rimane da considerare il problema

di risolvere il sistema lineare di Toeplitz $Tx = b$.
Dimostreremo, utilizzando il Teorema 4.1.2, una semplice rappresentazione dell'inversa di T del tipo
$T^{-1} = C_1 C_{-1} + C_1' C_{-1}'$, che coinvolge due matrici circolanti e due matrici (-1)-circolanti. Tale rappresentazione, dovuta ad Ammar e Gader [Ammar, Gader], ci permette di dimostrare che se non contiamo le operazioni che coinvolgono solo gli elementi di T, allora $Tx = b$ si può risolvere con $O(n \log_2 n)$ operazioni aritmetiche.

Notiamo che, per ogni matrice di Toeplitz $T = (t_{i-j})_{i,j=1}^n$ abbiamo che

$$T\Pi_\epsilon - \Pi_\epsilon T = \begin{bmatrix} \epsilon t_{-n+1} - t_1 & 0 & . & 0 \\ . & . & . & . \\ \epsilon t_{-1} - t_{n-1} & 0 & . & 0 \\ \epsilon t_0 - \epsilon t_0 & t_{n-1} - \epsilon t_{-1} & . & t_1 - \epsilon t_{-n+1} \end{bmatrix}$$

ovvero che:

$$T\Pi_\epsilon - \Pi_\epsilon T = u e_1^T - J e_1 u^T J, \quad u = \epsilon T e_n - \Pi_\epsilon T e_1. \qquad \star$$

Sia ora T invertibile. Allora $T^{-1}\Pi_\epsilon - \Pi_\epsilon T^{-1} = Jp(JT^{-1}u)^T - (T^{-1}u)p^T$, $p = (e_1^T T^{-1})^T = JT^{-1}e_n$.
Posto $\delta(\sigma) = T^{-1}[t_1 \cdots t_{n-1}\, \sigma]^T$ e $q = (T^T)^{-1}[\epsilon t_0\, t_{n-1} \cdots t_1]^T = J\delta(\epsilon t_0)$, si osserva che $T^{-1}u = \epsilon e_n - T^{-1}[t_1 \cdots t_{n-1}\, \epsilon t_0]^T$ e quindi che $JT^{-1}u = \epsilon e_1 - q$.

Applicando il Teorema 4.1.2 per $X = \Pi_\epsilon$ e $X' = \Pi_\beta$ si ottiene la formula

$$\begin{aligned}(\epsilon - \beta)T^{-1} &= -C_\epsilon(p)C_\beta(JT^{-1}u) + C_\epsilon(JT^{-1}u)C_\beta(p) + (\epsilon - \beta)C_\epsilon(p) \\ &= C_\epsilon(p)[-\beta I + C_\beta(q)] + [\epsilon I - C_\epsilon(q)]C_\beta(p)\end{aligned}$$

Osservazione:
Si dimostra che $\delta(\sigma)$ può essere sempre espresso in termini di $Jp = T^{-1}e_n$ e di al piú altre due colonne di T^{-1}. Ad esempio, se $s_{11} = [T^{-1}]_{11} \neq 0$, allora

$$T(-\frac{1}{s_{11}} Z^T T^{-1} e_1) = \begin{bmatrix} t_1 \\ \vdots \\ t_{n-1} \\ -\frac{[0\, t_{n-1} \cdots t_1]T^{-1}e_1}{s_{11}} \end{bmatrix}$$

e quindi

$$\delta(\sigma) = -\frac{1}{s_{11}} Z^T T^{-1} e_1 + \left[\sigma - \left(-\frac{[0\, t_{n-1} \cdots t_1]T^{-1}e_1}{s_{11}}\right)\right] T^{-1} e_n$$

cioè $\delta(\sigma)$ è noto se sono note la prima e l'ultima colonna di T^{-1}. $\square$

Le scelte $\epsilon = 1$, $\beta = -1$ conducono alla formula di Ammar-Gader:

$$2T^{-1} = C_1(p)C_{-1}(q + e_1) + C_1(e_1 - q)C_{-1}(p) =$$

$$= nF[d(Fp)F^H DFd(FD(q + e_1)) + d(F^H DFd(FDp)]F^H \overline{D}$$

dove $T(Jp) = e_n$, $T(Jq) = [t_1 \cdots t_{n-1}\, t_0]^T$, $Jq = \delta(t_0)$. Una volta noti i vettori p, q, Fp, Fq, FDp e Fdq, la formula di Ammar-Gader ci permette di calcolare ogni vettore $T^{-1}b$, al variare di b, con sei trasformazioni discrete di Fourier.

Riportiamo infine qui sotto due formule piu' esplicite per $S = T^{-1}$, valide nell'ipotesi $s_{11} \neq 0$ e T simmetrica. La prima segue dalla formula di Ammar-Gader. La seconda, valida per n pari, e' ricavata in [Di Fiore, Zellini, 1998] come caso particolare di una espressione per T^{-1} che generalizza la formula di Ammar-Gader.

$$T^{-1} = \frac{1}{2s_{11}}[C(s_1)C_{-1}(s_1)^T + C(s_1)^T C_{-1}(s_1)], \quad s_1 = T^{-1}e_1\,,$$

$$T^{-1} = \frac{1}{2s_{11}}\{[\begin{pmatrix} C_{-1}(a)^T & 0 \\ 0 & C(b)^T \end{pmatrix} - \lambda I]C_{-1}(s_1) -$$

$$\begin{pmatrix} C_{-1}(b) & 0 \\ 0 & -C(a) \end{pmatrix} C_{-1}(s_1)^T + s_{11}C_{-1}(s_{m+1})\}(I - \Pi_{-1}^m)$$

dove $a = I_m^1 s_1$, $b = I_n^{m+1} s_1$, $s_1 = T^{-1}e_1$, $s_{m+1} = T^{-1}e_{m+1}$, $\lambda = s_{m+1,1}$, $m = n/2$ (si suppone n pari). Si noti che $C_{-1}(s_1)^T = C_{-1}(-J\Pi_{-1}s_1)$, $C(s_1)^T = C(J\Pi s_1)$.

Per brevità non sono state incluse in questa tesi formule di dislocamento $A = \sum_i \mathcal{M}_i\mathcal{N}_i$ che utilizziamo, nelle definizioni di $\mathcal{M}_i$ e $\mathcal{N}_i$, la stessa algebra ma con dimensioni diverse. Queste sono utili nella rappresentazione di matrici di tipo Toeplitz e Hankel. L'algebra in questione è l'algebra τ, ottenuta come corollari di Teoremi tipo il 4.1.2., (vedi [Di Fiore, Zellini, 1995, 1998]).

4.2 Algebre di tipo Hartley nella soluzione di sistemi Toeplitz più Hankel

Algebre di matrici contenenti la matrice $X = \Pi_\beta + \Pi_\beta^T + \epsilon e_1 e_1^T + \phi e_n e_n^T$ possono essere coinvolte in rappresentazioni efficienti dell'inversa di una matrice Toeplitz più Hankel $T + H$ e questo perché il rango di $(T + H)^{-1}X - X(T + H)^{-1}$ (il rango di dislocamento di $(T + H)^{-1}$, rispetto all'operatore commutatore in questo caso) è sempre 4 e quindi è indipendente da n. In [Di Fiore, 2000] si caratterizzano tutte le algebre $\mathcal{L}$ contenenti X (con base J_k tale che $e_1^T J_k = e_k^T$), e si ottengono formule di dislocamento, definite in termini di tali $\mathcal{L}$, molto efficienti nella rapresentazione di matrici di tipo Toeplitz più Hankel.

Qui descriveremo solo otto delle algebre studiate in [Di Fiore, 2000] (vedi anche [Di Fiore, Zellini, 2001], [Di Fiore, Bortoletti]) e useremo due di queste per

ottenere una formula per $(T+H)^{-1}$, dove T è simmetrica e H è persimmetrica, che consente di risolvere sistemi $(T+H)x = b$ in $O(n \log_2 n)$ operazioni aritmetiche.

Sia cas $\phi = \cos\phi + \sin\phi$. Poniamo

$$H_n = \frac{1}{\sqrt{n}}(cas\frac{2ij\pi}{n})_{i,j=0}^{n-1},\ K_n^T = \frac{1}{\sqrt{n}}(cas\frac{(2i+1)j\pi}{n})_{i,j=0}^{n-1},$$

$$G_n = \frac{1}{\sqrt{n}}(cas\frac{(2i+1)(2j+1)\pi}{n})_{i,j=0}^{n-1}$$

Le matrici H_n, K_n^T, G_n sono matrici unitarie reali. Si noti che H_n e G_n sono simmetriche (G_n è anche persimmetrica). Procedendo in maniera analoga al caso della matrice di Fourier, si può dimostrare che le matrici $V_{2n} = \sqrt{2n}H_{2n}, \sqrt{2n}K_{2n}^T, \sqrt{2n}G_{2n}, \sqrt{2n}K_{2n}$ possono tutte essere espresse in termini di due matrici V_n tramite una identità del tipo

$$V_{2n} = S \begin{bmatrix} V_n & O \\ O & V_n \end{bmatrix} Q_{2n}, \quad S \text{ sparsa}, \quad Q_{2n} \text{ di permutazione.}$$

È sufficiente infatti porre

$$S = \begin{bmatrix} I & X \\ I & -X \end{bmatrix}, con \quad \begin{cases} X = R_K & \text{se } V_n = \sqrt{n}H_n \\ X = R_\gamma & \text{se } V_n = \sqrt{n}K_n^T \end{cases}$$

$$S = \begin{bmatrix} X & Y \\ -YN & XN \end{bmatrix}, con \begin{cases} X = R_+, Y = R_-, N = J & \text{se } V_n = \sqrt{n}G_n \\ X = \widetilde{R}_+, Y = \widetilde{R}_-, N = J & \text{se } V_n = \sqrt{n}K_n \end{cases}$$

dove $R_K = d(\mathbf{c}) + d(\mathbf{s})J\Pi, c_h = \cos\frac{h\pi}{n}, s_h = \sin\frac{h\pi}{n}, R_\gamma = d(\mathbf{c}) + d(\mathbf{s})J, c_h = \cos\frac{(2h+1)\Pi}{2n}, s_h = \sin\frac{(2h+1)\Pi}{2n}, R_\pm = d(\mathbf{c}) \pm d(\mathbf{s})J, c_h = \cos\frac{(2h+1)\Pi}{4n}, s_h = \sin\frac{(2h+1)\Pi}{4n}, \widetilde{R}_\pm = d(\mathbf{c}) \pm d(\mathbf{s})J\Pi, c_h = \cos\frac{h\pi}{2n}, s_h = \sin\frac{h\pi}{2n}$, e h varia ogni volta tra 0 e $n-1$.

Si noti che, più in generale, è possibile esprimere V_{mn} in termini di m matrici V_n. Da queste considerazioni segue che le algebre $\mathcal{L} = sd\frac{1}{\sqrt{n}}V_n$ sono costituite da matrici di bassa complessità, ovvero, se $A \in \mathcal{L}$ e $\mathbf{z} \in \mathbb{C}^n$ allora i vettori $A\mathbf{z}, \mathbf{z} \in \mathbb{C}^n$, e $\mathbf{x}$ tale che $A\mathbf{x} = \mathbf{f}, \mathbf{f} \in \mathcal{C}^n$, sono calcolabili con al più $O(n \log n)$ operazioni aritmetiche. Ha per questo interesse studiare la struttura delle matrici di tali algebre. Per farlo si possono utilizzare convenientemente gli spazi $\mathcal{C}_{\pm1}^S = \{A \in \mathcal{C}_{\pm1} : A = A^T\}$ e $\mathcal{C}_{\pm1}^{SK} = \{A \in \mathcal{C}_{\pm1} : A = -A^T\}$, e così si ottengono le prime quattro algebre di tipo Hartley:

$$sdH_n = \mathcal{C}^S + J\Pi\mathcal{C}^{SK}, \quad sdK_n^T = \mathcal{C}^S + J\mathcal{C}^{SK},$$
$$\gamma = sdG_n = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}, \quad sdK_n = \mathcal{C}_{-1}^S + J\Pi_{-1}\mathcal{C}_{-1}^{SK}.$$

Le rimanenti quattro algebre Hartley-type sono associate a trasformate simili a quelle già considerate:

$$sdH_nI_\eta^T = \mathcal{C}^S + J\Pi\mathcal{C}^S, \quad \eta = sdK_n^TI_\eta = \mathcal{C}^S + J\mathcal{C}^S,$$

$$\mu = sdG_nI_\mu = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^S, \quad sdK_nI_\mu^T = \mathcal{C}_{-1}^S + J\Pi_{-1}\mathcal{C}_{-1}^S.$$

$$I_\eta = \begin{bmatrix} \sqrt{2} & & & \\ & I & & J \\ & & \sqrt{2} & \\ & -J & & I \end{bmatrix}, I\mu = \begin{bmatrix} I & & -J \\ & \sqrt{2} & \\ J & & I \end{bmatrix}$$

È importante osservare che le algebre sdH_n, sdK_n^T, $sdH_nI_\eta^T$, $sdK_n^TI_\eta$ contengono la matrice $\Pi_1 + \Pi_1^T$, in quanto quest'ultima matrice è circolante simmetrica, mentre le rimanenti quattro algebre contengono la matrice $\Pi_{-1} + \Pi_{-1}^T$, (-1)-circolante simmetrica.

Studiamo l'algebra $\gamma := sdG_n$ un po' meglio, anche per applicare la teoria vista nello studio delle algebre sdM. Possiamo innanzitutto osservare che per $n = 2+4s$ ciascuna riga di G_n ha almeno un elemento nullo, mentre per tutti gli altri valori di n si ha che $[G]_{1k} \neq 0, \forall k$. Ciò è come dire che le matrici di γ sono (come accade per le ϕ-circolanti) univocamente determinate da una loro riga (e in particolare dalla loro prima riga), ovvero vale la rappresentazione $\gamma = \{Gd(G^T\mathbf{z})d(G^T\mathbf{e_1})^{-1}G^{-1} : \mathbf{z} \in \mathbb{C}^n\}$ se e solo se $n \neq 2 + 4s$. Per $n = 2 + 4s$, invece, una combinazione di più righe di $A \in \gamma$ è necessaria per definire univocamente A. Ad esempio si osserva che la somma delle righe prima e ultima di $A \in \gamma$ definiscono A per ogni n, cioè, $\forall n$, vale la rappresentazione

$$\gamma = \{Gd(G^T\mathbf{z})d(G^T(\mathbf{e_1} + \mathbf{e_n}))^{-1}G^{-1} : \mathbf{z} \in \mathbb{C}^n\}. \tag{4.1}$$

La struttura delle matrici di γ è rivelata più chiaramente dall'identità $\gamma = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}$. Con $\mathcal{C}_{-1}^S$ si intende l'algebra delle matrici (-1)-circolanti simmetriche $n \times n$, e con $\mathcal{C}_{-1}^{SK}$ si intende lo spazio delle matrici (-1)-circolanti antisimmetriche $n \times n$ (una matrice A è antisimmetrica se $A^T = -A$).

Siano $T = (t_{i-j})$ e $H = (h_{i+j-2})$, con $i, j = 1, \ldots, n$, rispettivamente una matrice Toeplitz e una matrice Hankel di dimensione $n \times n$ con elementi complessi. Dalla $(\star)$ segue

$$T(\Pi_\beta + \Pi_\beta^T) - (\Pi_\beta + \Pi_\beta^T)T = (\underline{u}\,\underline{e}_1^T - J\underline{e}_1\,\underline{u}^TJ) + (J\widetilde{\underline{u}}\,\underline{e}_1^TJ - \underline{e}_1\widetilde{\underline{u}}^T),$$

dove

$$\underline{u} = \beta \begin{pmatrix} t_{-n+1} \\ \vdots \\ t_{-1} \\ t_0 \end{pmatrix} - \Pi_\beta \begin{pmatrix} t_0 \\ t_1 \\ \vdots \\ t_{n-1} \end{pmatrix}, \ \widetilde{\underline{u}} = \beta \begin{pmatrix} t_{n-1} \\ \vdots \\ t_1 \\ t_0 \end{pmatrix} - \Pi_\beta \begin{pmatrix} t_0 \\ t_{-1} \\ \vdots \\ t_{-n+1} \end{pmatrix}$$

$(\widetilde{\underline{u}} = -\frac{1}{\beta}\Pi_\beta J\underline{u}$ se $\beta = \pm 1)$, e da quest'ultima segue che

$$H(\Pi_\beta + \Pi_\beta^T) - (\Pi_\beta + \Pi_\beta^T)H = (\underline{v}\,\underline{e}_1^TJ - J\underline{e}_1\,\underline{v}^T) + (J\widetilde{\underline{v}}\,\underline{e}_1^T - \underline{e}_1\widetilde{\underline{v}}^TJ),$$

dove

$$\underline{v} = \beta \begin{pmatrix} h_0 \\ h_1 \\ \vdots \\ h_{n-1} \end{pmatrix} - \Pi_\beta \begin{pmatrix} h_{n-1} \\ h_n \\ \vdots \\ h_{2n-2} \end{pmatrix}, \ \widetilde{\underline{v}} = \beta \begin{pmatrix} h_{2n-2} \\ \vdots \\ h_n \\ h_{n-1} \end{pmatrix} - \Pi_\beta \begin{pmatrix} h_{n-1} \\ \vdots \\ h_1 \\ h_0 \end{pmatrix}$$

($\widetilde{\underline{v}} = -\frac{1}{\beta}\Pi_\beta J\underline{v}$ se $\beta = \pm 1$). Dunque se $X = \Pi_\beta + \Pi_\beta^T$, si ha

$$(T+H)X - X(T+H) = (\underline{u} + J\widetilde{\underline{v}})\underline{e}_1^T - J\underline{e}_1(\underline{u}^T J + \underline{v}^T) + (J\widetilde{\underline{u}} + \underline{v})\underline{e}_1^T J - \underline{e}_1(\widetilde{u}^T + \widetilde{v}^T J)$$

e quindi, nel caso $T + H$ invertibile

$$(T+H)^{-1}X - X(T+H)^{-1} = -(T+H)^{-1}(\underline{u}+J\widetilde{\underline{v}})e_1^T(T+H)^{-1} + (T+H)^{-1}Je_1(\underline{u}^T J + \underline{v}^T)(T+H)^{-1}$$

$$-(T+H)^{-1}(J\widetilde{\underline{u}} + \underline{v})\underline{e}_1^T J(T+H)^{-1} + (T+H)^{-1}e_1(\widetilde{u}^T + \widetilde{v}^T J)(T+H)^{-1}.$$

Quest'ultima identità e formule di dislocamento presenti in [Di Fiore, 2000], in termini di algebre di tipo Hartley (efficienti se rank $(AX - XA)$ è costante rispetto a n), consentono di ottenere espressioni per $(T + H)^{-1}$ tramite quali $\underline{x} : (T + H) = \underline{f}$ può essere calcolato con $O(n \log_2 n)$ operazioni aritmetiche (se si escludono operazioni preliminari su $(T + H)$).

In particolare, nell'ipotesi che T sia simmetrica ($T = T^T$) e H sia persimmetrica ($H = JHJ$), si ottiene la seguente espressione

$$(T+H)^{-1} = \frac{1}{2}\big[\mu(Jx + e_1)\eta(w) - \mu(w)\eta(Jx - e_1)\big]$$

dove $w = (T+H)^{-1}e_1$, $x = (T+H)^{-1}\big[t_1 + h_{-1}|t_2 + h_0|\cdots|t_n + h_{n-2}\big]^T$. Da questa segue che ogni sistema $(T + H)\underline{x} = \underline{b}$ si può risolvere in $O(n \log_2 n)$ operazioni aritmetiche se sono noti i vettori w.

4.3 Proiezione su *sdU* nei metodi iterativi GC e BFGS

Nei lavori [Di Benedetto, Serra], [Tudisco, Di Fiore, Tyrtyshnikov], [R.Chan, Strang], [T.Chan], [Di Fiore, Zellini, 2001], si mostra come matrici negli spazi $sd\,U$ possano essere usate per precondizionare sistemi lineari $Ax = b$ con A definita positiva e, quindi, aumentare la velocita' di convergenza del metodo del Gradiente Coniugato (GC) nella loro risoluzione numerica. Nei lavori [Di Fiore, Fanelli, Lepore, Zellini], [Cipolla, Di Fiore, Tudisco, Zellini], [Cipolla, Di Fiore, Zellini], [Cai, R.Chan, Di Fiore], si mostra come le stesse matrici consentano di introdurre metodi di bassa

complessita' computazionale di tipo BFGS per la minimizzazione di funzioni generiche di classe C^1. Qui ricorderemo brevemente solo alcuni di questi risultati.
Per una matrice unitaria $U \in \mathbb{C}^{n\times n}$, l'algebra associata *sdU* è definita come:

$$L := sdU := \{Ud(z)U^H \mid z \in \mathbb{C}^n\}.$$

Qui, U^H rappresenta la trasposta coniugata di U, e $d(z)$ è una matrice diagonale con $z \in \mathbb{C}^n$ come elementi della diagonale.

Sia L_B la matrice in L per cui $\|L_B - B\| = \min\{\|X - B\| : \ X \in \ L\}$.

Teorema 4.3.1. *Sia $L = sdU$ e sia $B \in \mathbb{C}^{n\times n}$. Allora:*

1. *$L_B = Ud(z_B)U^H$, dove $[z_B]_i = [U^H BU]_{ii}$, $i = 1, \cdots, n$; in particolare, $z_{xy^T} = d(U^H x)U^T y$, dove $x, y \in \mathbb{C}^n$.*

2. *Se $B \in \mathbb{R}^{n\times n}$, allora $L_B \in \mathbb{R}^{n\times n}$, a condizione che L sia generato da matrici reali, o più in generale, quando il coniugato dello spazio L è incluso in L, cioè $\bar{L} \subseteq L$ (ovvero L è chiuso rispetto alla coniugazione).*

3. *Se $B = B^\star$, allora $L_B = (L_B)^\star$ e $min\,\lambda(B) \leq \lambda(L_B) \leq max\,\lambda(B)$, dove $\lambda(X)$ è lo spettro di X. Di conseguenza, L_B è hermitiana definita positiva quando B è hermitiana definita positiva.*

4. *$tr(L_B) = tr(B)$;*

5. *Se B è hermitiana definita positiva, allora $det(B) \leq det(L_B)$, con l'uguaglianza che si verifica se e solo se U diagonalizza B.*

6. *Se B è hermitiana, allora $\lambda(L_B) \prec \lambda(B)$ (la relazione di maggiorazione spettrale);*

7. *Se B è hermitiana e ϕ è convessa, allora $\sum_{i=1}^n \phi(\lambda_i(L_B)) \leq \sum_{i=1}^n \phi(\lambda_i(B))$;*

8. *Se B è hermitiana e ϕ è convessa monotona crescente, allora $\sum_{i=1}^n \phi(\lambda_i(L_B)) \leq \sum_{i=1}^n \phi(\lambda_i(B))$ per tutti $k \in \{1, \cdots, n\}$;*

9. *Si definisce il numero di condizionamento K per B come $K(B) = (\frac{tr(B)}{n})^{n\det(B)}$. Allora*

$$K(L_B^{-1}B) = min\{K(X^{-1}B) : X \in L, \text{ hermitiana definita positiva}\}.$$

GCP

Per risolvere un sistema lineare $Ax = b$ dove A e' $n \times n$ reale definita positiva e $b \in R^n$, si puo' usare il metodo del Gradiente Coniugato Precondizionato (GCP) con una opportuna matrice P reale definita positiva. Una volta definita la matrice P, il

metodo genera una successione x_k convergente a $x = A^{-1}b$ in al piu' m passi dove m e' il numero degli autovalori distinti della matrice $P^{-1}A$. Inoltre, piu' informazioni si hanno sulla distribuzione dello spettro di $P^{-1}A$, piu' fini maggiorazioni per l'errore al passo k si possono trovare, grazie alle teoria minmax di Chebycev. Ad esempio, se non si sa nulla sullo spettro di $P^{-1}A$, allora

$$\|x - x_k\|_A \leq 2\left(\frac{\sqrt{\mu_P} - 1}{\sqrt{\mu_P} + 1}\right)^k \|x - x_0\|_A,\ k = 0,1,2,\ldots$$

dove $\mu_P = \max \lambda_i(P^{-1}A)/\min \lambda_i(P^{-1}A)$. Se invece si sa che la maggior parte degli autovalori di $P^{-1}A$ sono in un certo intervallo $[a, b] \subset R^+$, e che i rimanenti sono al piu' r e sono tutti maggiori di b (a, b, r ovviamente dipendono da P), allora

$$\|x - x_k\|_A \leq 2\left(\frac{\sqrt{b/a} - 1}{\sqrt{b/a} + 1}\right)^{k-r} \|x - x_0\|_A,\ k = r, r+1, \ldots\ .$$

Da questo tipo di maggiorazioni si deduce che se gli autovalori di $P^{-1}A$ sono addensati in clusters, allora GCP ha una convergenza rapida; cioe' e' come se ogni cluster contasse per un solo autovalore.

Riportiamo qui sotto le equazioni del metodo GCP

$x_0 \in R^n$,
$r_0 = b - Ax_0$,
$d_0 = h_0 = P^{-1}r_0$.

For $k = 0,1,2\ldots$:
$\tau_k = \frac{r_k^T h_k}{d_k^T A d_k}$,
$x_{k+1} = x_k + \tau_k d_k$,
$r_{k+1} = r_k + \tau_k A d_k$,
$h_{k+1} = P^{-1}r_{k+1}$, (*)
$\beta_k = \frac{r_{k+1}^T h_{k+1}}{r_k^T h_k}$,
$d_{k+1} = h_{k+1} + \beta_k d_k$.

Come si vede, ogni passo di GCP richiede come operazioni principali il calcolo del prodotto Ad_k e la risoluzione del sistema lineare $Ph_{k+1} = r_{k+1}$. La velocita' di convergenza del metodo dipende da μ_P, se non ci sono clusters di autovalori; oppure da b/a e da r, se solo alcuni degli autovalori non sono raggruppati sulla sinistra. E cosi' via.

Una matrice P e' un buon precondizionatore se ha le seguenti proprieta':

(1) P deve essere innanzitutto una matrice reale definita positiva;

(2) P deve essere tale che il passo (*) sia poco costoso rispetto al costo del prodotto Ad_k; e

(3) Gli autovalori di $P^{-1}A$ devono raggrupparsi piu' degli autovalori di A.

La scelta $P = (sdU)_A$, con U unitaria di bassa complessita' e tale che $\overline{sdU} \subset sdU$, rende soddisfatte le proprieta' (1) e (2) (vedi i punti 2. e 3. del Teorema 4.3.1). La proprieta' (3) risulta soddisfatta almeno per certe classi di matrici A di Toeplitz, scegliendo sdU uguale all'algebra delle matrici circolanti [T.Chan], [R.Chan, Strang], oppure sdU uguale alle η, μ [Di Fiore, Zellini, 2001].

BFGSP

Si vuole minimizzare una funzione $f : R^n \to R$, $f \in C^1$ limitata inferiormente.

La classe dei metodi iterativi di tipo BFGS qui considerata a tale scopo definisce le direzioni di discesa d_k come $d_k = -B_k^{-1}\nabla f(x_k)$, con B_k reale definita positiva costruita a partire da B_{k-1}; quindi imita strutturalmente il metodo di Newton – ove $d_k = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ – e, allo stesso tempo, evita le modifiche necessarie in Newton per renderlo convergente globalmente (la matrice Hessiana e' definita positiva solo quando x_k e' vicino al minimo). Inoltre, il costo per passo e' ridotto da $O(n^3)$ a $O(n^2)$, e non vi sono derivate seconde da calcolare. C'e' un unico svantaggio nei metodi quasi-Newton di tipo BFGS: l'ordine di convergenza si riduce da quadratico a superlineare.

I metodi iterativi di tipo BFGS sono in pratica una generalizzazione a R^n del metodo della secante *modificato* (si noti la presenza di ω_k) per la ricerca degli zeri di $f'(x)$, con $f : R \to R$,

$$x_{k+1} = x_k - \omega_k b_k^{-1} f'(x_k), \quad b_k = \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}}$$

(una possibile definizione di ω_k puo' essere $\omega_k = \max\{\omega \in (0,1] : |f'(x_{k+1})| < |f'(x_k)|\}$). Si ricorda che tale metodo e' molto competititivo con Newton, perche', pur avendo ordine di convergenza $\frac{1+\sqrt{5}}{2}$, e quindi non quadratica, puo' essere implementato a meta' costo per passo (non e' richiesto il calcolo di $f''(x_k)$).

Dati P matrice reale definita positiva e $s, y \in R^n$ vettori tali che $s^T y \neq 0$, sia

$$\phi(P, s, y) = P + \frac{1}{s^T y} y y^T - \frac{1}{s^T P s} P s s^T P.$$

La matrice $\phi(P, s, y)$ e' reale simmetrica e, se $s^T y > 0$, e' anche definita positiva (si lascia al lettore la dimostrazione). Nei metodi di tipo BFGS, definiti qui sotto in due varianti, la matrice B_k e' costruita da B_{k-1} utilizzando l'operatore ϕ:

$x_0 \in R^n$,
$g_0 = \nabla f(x_0)$,
$B_0 =$ reale definita positiva,

- $d_0 = -B_0^{-1} g_0$ (nota: $d_0^T g_0 < 0$)
  P_0 reale definita positiva (P_0 andrebbe costruita a partire da B_0),
  $d_0 = -P_0^{-1} g_0$ (nota: $d_0^T g_0 < 0$)

For $k = 0,1,2,\ldots$:
$\omega_k \in G_k \cup W_k$
$x_{k+1} = x_k + \omega_k d_k$
$g_{k+1} = \nabla f(x_{k+1})$, $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$
(nota: $s_k^T y_k > 0$ perche' $\omega_k \in G_k \cup W_k$)
$B_{k+1} = \phi(P_k, s_k, y_k)$
(nota: B_{k+1} e' reale definita positiva, perche' lo e' P_k e $s_k^T y_k > 0$)
(1) $d_{k+1} = -B_{k+1}^{-1} g_{k+1}$ (nota: $d_{k+1}^T g_{k+1} < 0$)
P_{k+1} = reale definita positiva (P_{k+1} andrebbe costruita a partire da B_{k+1})
(2) $d_{k+1} = -P_{k+1}^{-1} g_{k+1}$ (nota:$d_{k+1}^T g_{k+1} < 0$)

Definizione degli insiemi G_k e W_k: siano c_1, c_2 tali che $0 < c_1 < c_2 < 1$ (di solito si sceglie c_1 piccolo e $c_2 = 1/2$); allora

$$G_k = \{\omega > 0 : \ f(x_k + \omega d_k) \leq f(x_k) - c_1 \omega d_k^T(-g_k)\},$$

$$W_k = \{\omega > 0 : \ \nabla f(x_k + \omega d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k\}.$$

La condizione $\omega_k \in G_k$ implica che $f(x_{k+1})$ e' minore "sostanzialmente" di $f(x_k)$, se ω_k non e' troppo piccolo, e la condizione $\omega_k \in W_k$ assicura che ω_k non sia troppo piccolo. Inoltre, come il lettore puo' verificare facilmente, la condizione $\omega_k \in W_k$ implica la disuguaglianza $s_k^T y_k > 0$, fondamentale per il funzionamento del metodi di tipo BFGS.

Scelta $P_k = B_k$:

Il classico metodo BFGS, dovuto a Broyden, Fletcher, Goldfarb, Shanno, ottenibile ponendo $P_k = B_k$, si e' rivelato, tra i metodi quasi-Newton secanti (cioe' per cui $B_{k+1} s_k = y_k$), quello piu' efficiente; il suo ordine di convergenza e' superlineare, e il suo costo per passo e' $O(n^2)$ operazioni aritmetiche. BFGS e' estremamente competitivo con il metodo di Newton, comunque si modifichi quest'ultimo per renderlo convergente globalmente.

Scelta $P_k = (sd\,U_k)_{B_k}$:

Un modo automatico piu' generale per definire le matrici P_k e' porre $P_k = (sd\,U_k)_{B_k}$, con U_k matrice unitaria tale che $\overline{sd\,U_k} \subset sd\,U_k$. Infatti, per le proprieta' 2. e 3. del Teorema 4.3.1, tale matrice $(sd\,U_k)_{B_k}$ e' reale definita positiva ogni volta che lo e' B_k.

Se U_k diagonalizza B_k allora $(sd\,U_k)_{B_k} = B_k$, cioe' si ritrova BFGS.

Se invece le matrici U_k sono scelte di bassa complessita' computazionale, cioe' definiscono trasformate discrete veloci di costo al piu' $O(nlog_2 n)$, per quanto riguarda il tempo di esecuzione, e $O(n)$, per quanto riguarda lo spazio di memoria occupato, allora si ottengono metodi innovativi, di costo al piu' $O(nlog_2 n)$ per passo, implementabili con $O(n)$ celle di memoria, in grado di affrontare e risolvere problemi di minimo di grandi dimensioni con minime risorse.

Per $P_k = (sd\,U_k)_{B_k}$ e' stato dimostrato che il metodo (2) ha ordine di convergenza lineare [Di Fiore, Fanelli, Lepore, Zellini], [Di Fiore, Lepore, Zellini], e si e' osservato sperimentalmente (su numerosi problemi test) che il metodo (1) e' molto piu' rapido nella convergenza del metodo (2), e, quindi, risulta molto competitivo con (limited memory BFGS) nella risoluzione di problemi di minimo [Bortoletti, Di Fiore, Fanelli, Zellini].

Capitolo 5

Due applicazioni delle matrici triangolari di Toeplitz

5.1 I primi n numeri di Bernoulli risolvono un sistema triangolare di Toeplitz

5.1.1 Polinomi e numeri di Bernoulli

In questa sezione si ricordano le definizioni e alcune proprieta' dei numeri e dei polinomi di Bernoulli [Gross].

Le condizioni

$$B(x+1) - B(x) = nx^{n-1}, \quad \int_0^1 B(x)\,dx = 0,$$

dove $B(x)$ è un polinomio, definiscono univocamente la funzione $B(x)$. Essa è un particolare polinomio monico di grado n, chiamato n-esimo polinomio di Bernoulli, e indicato con il simbolo $B_n(x)$.

Un semplice calcolo permette di ottenere i primi polinomi di Bernoulli:

$$B_1(x) = x - \frac{1}{2}, \quad B_2(x) = x^2 - x + \frac{1}{6}, \quad B_3(x) = x(x - \frac{1}{2})(x-1), \quad \dots$$

Per convenzione, si assume che $B_0(x) = 1$.

Si dimostra che i polinomi di Bernoulli definiscono i coefficienti dello sviluppo in serie di potenze di diverse funzioni; ad esempio, per ciò che segue, è opportuno ricordare che vale la seguente uguaglianza:

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n(x)}{n!} t^n. \tag{5.1}$$

Inoltre, i polinomi di Bernoulli soddisfano diverse identità. Due delle più importanti sono quelle concernenti il valore della loro derivata e le loro proprietà di simmetria/antisimetria rispetto all'asse $x = \frac{1}{2}$:

$$B_n^{'}(x) = nB_{n-1}(x), \quad B_n(1-x) = (-1)^n B_n(x).$$

In particolare, come conseguenza di queste ultime identità e della loro definizione, si vede facilmente che tutti i polinomi di Bernoulli di grado dispari, eccetto il primo, sono nulli in zero. Al contrario, il valore in zero dei polinomi di Bernoulli di grado pari è, oltre che diverso da zero, particolarmente significativo. In particolare, vale la seguente formula di Eulero:

$$\zeta(2j) = |B_{2j}(0)|\frac{(2\pi)^{2j}}{2(2j)!}, \quad \zeta(s) = \sum_{k=1}^{\infty}\frac{1}{k^s},$$

che mette in relazione i valori $B_{2j}(0)$ con i valori della funzione Zeta di Riemann ζ nei numeri interi positivi pari $2j$. Ad esempio, da questa relazione e dal fatto che $\zeta(2j) \to 1$ se $j \to +\infty$, si deduce che i valori $|B_{2j}(0)|$ tendono a $+\infty$ come $\frac{2(2j)!}{(2\pi)^{2j}}$.

Un'altra formula importante coinvolgente i valori $B_{2j}(0)$ è quella di Eulero-Mclaurin, utile per il calcolo di somme. Se f è una funzione sufficientemente regolare in $[m, n]$, con $m, n \in \mathbb{Z}$, allora

$$\sum_{r=m}^{n} f(r) = \frac{1}{2}\left[f(m) + f(n)\right] + \int_m^n f(x)\,dx + \sum_{j=1}^{k}\frac{B_{2j}(0)}{(2j)!}\left[f^{(2j-1)}(n) - f^{(2j-1)}(m)\right] + u_{k+1},$$

dove

$$u_{k+1} = \frac{1}{(2k+1)!}\int_m^n f^{(2k+1)}(x)\overline{B}_{2k+1}(x)\,dx = -\frac{1}{(2k)!}\int_m^n f^{(2k)}(x)\overline{B}_{2k}(x)\,dx =$$

$$\frac{1}{(2k+2)!}\int_m^n f^{(2k+2)}(x)\left[B_{2k+2}(0) - \overline{B}_{2k+2}(x)\right]dx$$

e $\overline{B}_n$ è l'estensione periodica su $\mathbb{R}$ di B_n definito su [0,1). Ricordiamo che la formula di Eulero-Maclaurin conduce anche a una rappresentazione importante dell'errore commesso dalla formula dei trapezi

$$I_h = h\left[\frac{1}{2}g(a) + \sum_{r=1}^{n-1} g(a+rh) + \frac{1}{2}g(b)\right], \quad h = \frac{b-a}{n},$$

nell'approssimazione di un integrale $I = \int_a^b g(x)\,dx$. Tale rappresentazione, valida per funzioni sufficientemente regolari in $[a, b]$, si ottiene ponendo $m = 0$ e $f(t) = g(a + th)$, e dà:

$$I_h = I + \sum_{j=1}^{k}\frac{h^{2j}B_{2j}(0)}{(2j)!}\left[g^{(2j-1)}(b) - g^{(2j-1)}(a)\right] + r_{k+1},$$

dove

$$r_{k+1} = \frac{g^{(2k+2)}(\xi)h^{2k+2}(b-a)B_{2k+2}(0)}{(2k+2)!}, \quad \xi \in (a,b).$$

Quest'ultima, ove l'errore $I_h - I$ è rappresentato in termini di potenze pari di h, giustifica l'efficienza del metodo di estrapolazione di Romberg per la stima di integrali, quando tale metodo è applicato in combinazione con la formula dei trapezi. È evidente infatti che $\tilde{I}_{h/2} := \frac{2^2 I_{h/2} - I_h}{2^2-1}$ approssima I con un errore dell'ordine di $O(h^4)$, mentre l'errore di $I_{h/2}$ e di I_h, nell'approssimazione di I, è dell'ordine di $O(h^2)$.
I valori $B_{2j}(0)$ sono noti come numeri di Bernoulli.

5.1.2 I numeri di Bernoulli risolvono sistemi triangolari di Toeplitz

In questa sezione vedremo che, come dimostrato in [Di Fiore, Tudisco, Zellini], i primi n numeri di Bernoulli si possono ottenere risolvendo un sistema triangolare inferiore di Toeplitz di n equazioni. Dall'identità (5.1) segue che i numeri di Bernoulli soddisfano la seguente uguaglianza:

$$\frac{t}{e^t - 1} = \frac{-1}{2}t + \sum_{k=0}^{+\infty} \frac{B_{2k}(0)}{(2k)!}t^{2k}.$$

Moltiplicando quest'ultima per $e^t - 1$, sviluppando in potenze di t, ed imponendo che i coefficienti di t^i, con $i = 2, 3, 4, \ldots$, del secondo membro siano uguali a zero, si ottengono le seguenti equazioni:

$$-\frac{1}{2}j + \sum_{k=0}^{\lfloor \frac{j-1}{2} \rfloor} \binom{j}{2k} B_{2k}(0) = 0, \quad j = 2,3,4,\ldots.$$

Ora, mettendo insieme le equazioni per j pari e quelle per j dispari, si ottengono due sistemi lineari triangolari inferiori che definiscono univocamente i numeri di Bernoulli:

$$\begin{bmatrix} \binom{2}{0} & & & & \\ \binom{4}{0} & \binom{4}{2} & & & \\ \binom{6}{0} & \binom{6}{2} & \binom{6}{4} & & \\ \binom{8}{0} & \binom{8}{2} & \binom{8}{4} & \binom{8}{6} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \cdot \end{bmatrix},$$

$$\begin{bmatrix} \binom{1}{0} & & & & \\ \binom{3}{0} & \binom{3}{2} & & & \\ \binom{5}{0} & \binom{5}{2} & \binom{5}{4} & & \\ \binom{7}{0} & \binom{7}{2} & \binom{7}{4} & \binom{7}{6} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{5}{2} \\ \frac{7}{2} \\ \cdot \end{bmatrix}.$$

Da questi possiamo ricavare i primi numeri di Bernoulli:

$$1, \frac{1}{6}, -\frac{1}{30}, \frac{1}{42}, -\frac{1}{30}, \frac{5}{66}, -\frac{691}{2730}, \frac{7}{6}, -\frac{47021}{6630}, \dots$$

Vogliamo dare una forma analitica alle matrici dei coefficienti W_e e W_o di tali sistemi lineari. Per farlo è sufficiente osservare che tali matrici sono sottomatrici della matrice di Tartaglia, la quale ammette una forma analitica. Dopo un po' di conti si conclude che:

$$W_e = Z^T \varphi \sum_{k=0}^{+\infty} \frac{1}{(2k+2)!} \varphi^k, \quad W_o = \begin{bmatrix} 1 & & & & \\ & 3 & & & \\ & & 5 & & \\ & & & 7 & \\ & & & & \cdot \end{bmatrix} \sum_{k=0}^{+\infty} \frac{1}{(2k+1)!} \varphi^k,$$

$$\varphi = \begin{bmatrix} 0 & & & & & \\ 2 & 0 & & & & \\ & 12 & 0 & & & \\ & & 30 & 0 & & \\ & & & 56 & 0 & \\ & & & & \cdot & \cdot \end{bmatrix}.$$

Possiamo dunque scrivere i sistemi lineari come segue:

$$\sum_{k=0}^{+\infty} \frac{2}{(2k+2)!} \varphi^k b = q^e, \quad \sum_{k=0}^{+\infty} \frac{1}{(2k+1)!} \varphi^k b = q^0,$$

dove

$$b = \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \vdots \end{bmatrix}, \quad q^e = \begin{bmatrix} 1 \\ \frac{1}{3} \\ \frac{1}{5} \\ \frac{1}{7} \\ \vdots \end{bmatrix}, \quad q^0 = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \vdots \end{bmatrix}.$$

Ora mostriamo che tali sistemi sono equivalenti a due sistemi lineari triangolari inferiori di Toeplitz. Il nostro scopo è sostituire φ, una matrice la cui sottodiagonale contiene elementi tutti diversi tra loro, con una matrice la cui sottodiagonale contiene elementi tutti uguali tra loro.

Sia $D = \mathrm{diag}(d_1, d_2, d_3, \dots)$, con $d_i \neq 0$. Se si scrive la matrice $D\varphi D^{-1}$, si vede che si possono scegliere i d_i in modo che $D\varphi D^{-1} = xZ$; infatti è sufficiente porre $d_k = \frac{x^{k-1}}{(2k-2)!}$ per $k = 1, 2, 3, \dots$. Si ha dunque:

$$D = \begin{bmatrix} 1 & & & & \\ & \frac{x}{2!} & & & \\ & & \frac{x^2}{4!} & & \\ & & & \frac{x^{n-1}}{(2n-2)!} & \\ & & & & \ddots \end{bmatrix}.$$

Moltiplicando i due sistemi a sinistra per D, otteniamo:

$$\sum_{k=0}^{+\infty} \frac{2}{(2k+2)!} D\varphi^k D^{-1} Db = \sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k Db = Dq^e, \quad \text{(pari)},$$

$$\sum_{k=0}^{+\infty} \frac{x^k}{(2k+1)!} Z^k Db = Dq^o, \quad \text{(dispari)}.$$

Riassumendo, il vettore $\{b\}_n$ (ovvero, i primi n numeri di Bernoulli) può essere ottenuto:

1. Calcolando le prime n componenti della soluzione z del seguente sistema lineare triangolare inferiore di Toeplitz:

$$(\sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k) z = Dq^e,$$

2. Risolvendo il sistema lineare $\{D\}_{n\times n}\{b\}_n = \{z\}_n$ nel campo razionale.

Osserviamo che il calcolo in 1. può essere effettuato con l'algoritmo descritto nella sezione 2.3 ad un costo $O(n \log_2 n)$ e che tale algoritmo può essere reso stabile numericamente mediante una scelta opportuna del parametro x. Ad esempio, osservando che $z_n = \frac{x^{n-1}}{(2n-2)!} B_{2n-2}(0)$, la scelta $x = (2\pi)^2$ renderebbe la successione $z_n, n \in \mathbb{N}$, limitata; infatti, in tal caso $|z_n| \to 2$ se $n \to +\infty$, per la formula di Eulero. Effettuato il calcolo in 1. si ottengono numeri macchina che costituiscono una ottima approssimazione in $\mathbb{R}$ delle quantità $x^s B_{2s}(0)/(2s)!$, con $s = 0, 1, \dots, n-1$. Poi, nella fase 2., da questi numeri macchina occorrerà ricavare i numeri razionali di Bernoulli $B_{2s}(0)$, con $s = 0, 1, \dots, n-1$.

Si può dimostrare che i numeri di Bernoulli, di nuovo a meno di fattori noti, risolvono un sistema triangolare inferiore di Toeplitz dove 2/3 delle diagonali della matrice dei coefficienti sono nulle (vedi [Di Fiore, Tudisco, Zellini]). Questo ovviamente comporta una ulteriore riduzione del costo computazionale, soprattutto se si definisce un algoritmo ad hoc per tali sistemi sparsi di Toeplitz.

5.2 Le matrici triangolari di Toeplitz nella forma canonica di Jordan di A^{-1}

Data $A \in \mathbb{C}^{n\times n}$ è possibile trasformarla per similitudine in forme canoniche $J \in \mathbb{C}^{n\times n}$, dove J ha una struttura particolare, ad esempio J è diagonale a blocchi nelle forme di Jordan e di Frobenius. Nella forma di Jordan i blocchi diagonali sono matrici di Toeplitz triangolari superiori con $[\mu\, 1\, 0\, \cdots\, 0]$ con $\mu \in \sigma(A)$ come prima riga. Nella forma canonica di Frobenius i blocchi diagonali sono matrici di Hessenberg inferiore con zeri ovunque tranne negli elementi $(i, i+1)$, dove ci sono degli 1, e negli elementi dell'ultima riga, dove ci sono i coefficienti dei fattori del polinomio caratteristico di A.

Supponendo di avere una forma canonica di A, cioè di avere χ e J tali che $\chi^{-1}A\chi = J$, vogliamo ricavare la forma canonica di A^{-1} cioè trovare $\widetilde{\chi}$ e $\widetilde{J}$ tale che $\widetilde{\chi}^{-1}A^{-1}\widetilde{\chi} = \widetilde{J}$ [Bozzo, Deidda, Di Fiore].

Per esempio, calcolare la forma canonica di Schur di A^{-1} partendo da quella da A, richiede l'inversione di J, perché l'inversa di una matrice triangolare superiore rimane triangolare superiore, e nient'altro: quindi $\widetilde{\chi} = \chi$ e $\widetilde{J} = J^{-1}$. Anche per quanto riguarda le forme canoniche di Jordan e Frobenius di A^{-1}, le matrici $\widetilde{J}$ possono essere scritte imediatamente. Vediamo questo più in dettaglio:

Nel caso della forma di Jordan, i blocchi diagonali di $\widetilde{J}$ sono semplicemente matrici di Toeplitz triangolari superiori con $[\frac{1}{\mu}\, 1\, 0 \cdots 0]$, $\mu \in \sigma(A)$, come prima riga.

Nel caso della forma di Frobenius, i blocchi diagonali di $\widetilde{J}$ sono matrici di Hessenberg inferiori con zeri ovunque tranne che negli elementi $(i, i+1)$ dove sono degli 1, e negli elementi dell'ultima riga dove sono i coefficienti dei fattori del polinomio caratteristico di A^{-1} immediatamente dedotti dai coefficienti dei corrispondenti fattori del polinomio caratteristico della matrice A.

Per esempio se $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a & b & c & d \end{bmatrix}$ è un blocco diagonale di J, allora $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{a} & -\frac{d}{a} & -\frac{c}{a} & -\frac{b}{a} \end{bmatrix}$ è un blocco diagonale di $\widetilde{J}$.

Al contrario, sia nella forma di Jordan che nella forma di Frobenius, è meno immediato vedere chi è $\widetilde{\chi}$ (a meno che A non sia diagonalizzabile, nel quale caso $\widetilde{\chi} =$

χ per Jordan). In entrambi casi è comunque facile osservare che $\widetilde{\chi}$ dovrebbe essere del tipo $\chi\mathcal{M}$, con $\mathcal{M}$ diagonale a blocchi invertibile tale che $\mathcal{M} = J\mathcal{M}\widetilde{J}$ perchè in questo modo avremmo $A\chi = \chi J \Rightarrow A^{-1}\chi = \chi J^{-1} \Rightarrow A^{-1}\chi\mathcal{M} = \chi J^{-1}\mathcal{M} = \chi\mathcal{M}\widetilde{J}$.

Qui studiamo la condizione $\mathcal{M} = J\mathcal{M}\widetilde{J}$ nel caso di Jordan. Si vedra' che le M per cui $M = (\mu I + Z^T)M(\frac{1}{\mu}I + Z^T)$ possono essere definite in termini della matrice di Tartaglia triangolare superiore e delle matrici triangolari superiori di Toeplitz. Prima di cominciare lo studio, osserviamo che la relazione $\widetilde{\chi} = \chi\mathcal{M}$ può essere utile nello studio della velocità di convergenza del metodo delle potenze quando applicato ad A^{-1} (vedi l'Appendice). Per qualche cenno sullo studio della condizione $\mathcal{M} = J\mathcal{M}\widetilde{J}$ nel caso di Frobenius vedi l'Appendice Bis.

5.2.1 La matrice di Tartaglia

Definizione:

Per $i, j = 1, \ldots, s$ poniamo $P_{ij} = \begin{pmatrix} j-1 \\ i-1 \end{pmatrix}$ per $i \leq j$ e $P_{ij} = 0$ altrimenti. Notiamo che la matrice $P = P_s$ contiene nella sua parte triangolare superiore il triangolo di Tartaglia:

$$P = \begin{bmatrix} \binom{0}{0} & \binom{1}{0} & \binom{2}{0} & \binom{3}{0} & . & \binom{s-1}{0} \\ 0 & \binom{1}{1} & \binom{2}{1} & \binom{3}{1} & . & \binom{s-1}{1} \\ 0 & 0 & \binom{2}{2} & \binom{3}{2} & . & \binom{s-1}{2} \\ . & . & . & . & . & . \\ . & . & . & . & . & \binom{s-1}{s-2} \\ 0 & . & . & . & 0 & \binom{s-1}{s-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & . & 1 \\ 0 & 1 & 2 & 3 & 4 & . & s-1 \\ 0 & 0 & 1 & 3 & 6 & . & \frac{(s-1)(s-2)}{2} \\ . & . & 0 & 1 & 4 & . & . \\ . & . & . & . & .1 & . & . \\ . & . & . & . & . & . & s-1 \\ 0 & . & . & . & . & 0 & 1 \end{bmatrix}.$$

Ad esempio:

$$P_1 = [1], P_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, P_3 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}, P_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}, P_\infty = \begin{bmatrix} 1 & 1 & 1 & 1 & . \\ 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ 0 & 0 & 0 & 1 & . \\ . & . & . & . & . \end{bmatrix}.$$

Consideriamo le seguenti matrici $s \times s$ definite, come la matrice di Tartaglia, per tutti i valori di s compreso ∞:

$$Z = Z_s = \begin{bmatrix} 0 & 0 & 0 & . & . & 0 \\ 1 & 0 & 0 & . & . & . \\ 0 & 1 & 0 & . & . & . \\ . & . & . & . & . & 0 \\ . & . & . & . & 0 & 0 \\ 0 & . & . & 0 & 1 & 0 \end{bmatrix},$$

$$Y = Y_s = \text{diag(1,2,3,...,s)} Z^T = \begin{bmatrix} 0 & 1 & 0 & . & .0 & \\ 0 & 0 & 2 & 0 & . & 0 \\ 0 & 0 & 0 & 3 & . & . \\ . & . & . & 0 & . & . \\ . & . & . & . & . & s-1 \\ 0 & . & . & . & 0 & 0 \end{bmatrix}.$$

L'espressione compatta di P in termini di Y permette di scrivere immediatamente l'inversa di P, infatti

$$P = \sum_{k=0}^{+\infty} \frac{1}{k!} Y^k = e^Y \Rightarrow P^{-1} = e^{-Y} = \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!} Y^k.$$

Per esempio

$$P_2^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \; P_4^{-1} = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$P_7^{-1} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -2 & 3 & -4 & 5 & -6 \\ 0 & 0 & 1 & -3 & 6 & -10 & 15 \\ 0 & 0 & 0 & 1 & -4 & 10 & -20 \\ 0 & 0 & 0 & 0 & 1 & -5 & 15 \\ 0 & 0 & 0 & 0 & 0 & 1 & -6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Un'altra osservazione che si puo' fare e' che il prodotto $P_s^T P_s$ genera una matrice di Tartaglia piena

$$P_s^T P_s = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 1 & 2 & 3 & 4 & \cdots & \cdots \\ 1 & 3 & 6 & \cdots & \cdots & \cdots \\ 1 & 4 & \cdots & \cdots & \cdots & \cdots \\ 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

Infine, noteremo che P_k può essere scritta in funzione di P_{k-1} e che di conseguenza anche P_k^{-1} si può scrivere in funzione di P_{k-1}^{-1}. Quest'ultimo risultato permette di costruire un algoritmo per il calcolo del prodotto $P_s^{-1}v$, con $v \in \mathbb{C}^s$, cioé per la risoluzione di un qualsiasi sistema con P_s come matrice dei coefficienti, implementabile con $s(s-1)/2$ operazioni additive, senza operazioni moltiplicative e senza calcolare gli elementi di P_s o P_s^{-1}.

5.2.2 Matrice Tartaglia e la forma canonica di Jordan della matrice inversa

Data $A \in \mathbb{C}^{n\times n}$, è ben noto che esiste $\chi \in \mathbb{C}^{n\times n}$ invertibile tale che $\chi^{-1} A \chi = J$ dove J è diagonale a blocchi con i blocchi diagonali del tipo $\mu I_s + Z_s^T$, con $\mu \in \sigma(A) = \{\text{autovalori di } A\}$ e $s \in \{1, \ldots, n\}$.

La matrice A è diagonalizzabile se e solo se tutti questi blocchi hanno ordine 1, cioé sono tutti del tipo $\mu I_s + Z_s^T = \mu$, con $\mu \in \sigma(A)$.

La matrice A è non derogatoria, cioè A^n è la prima potenza che può essere scritta come combinazione lineare di potenze minori di A, se e solo se diversi blocchi diagonali hanno diversi μ sulla loro diagonale (o equivalentemente la molteplicità geometrica di ogni autovalore di A è 1). Quindi, ad esempio, una matrice $n \times n$ diagonalizzabile è non derogatoria se e solo se ha n autovalori distinti.

Sia $A \in \mathbb{C}^{n\times n}$ invertibile e assumiamo di conoscere una coppia di Jordan per A, cioè matrici χ e J tali che $\chi^{-1}A\chi = J$; vogliamo dedurre da questa una coppia di Jordan per A^{-1}, cioè matrici $\widetilde{\chi}$ e $\widetilde{J}$ tali che $\widetilde{\chi}^{-1}A^{-1}\widetilde{\chi} = \widetilde{J}$. Per quanto riguarda la $\widetilde{J}$ la risposta è ovvia, infatti se $\mu I_s + Z_s^T$ è un blocco diagonale di J, allora $\frac{1}{\mu} I_s + Z_s^T$ deve essere un blocco diagonale di $\widetilde{J}$.

Per quanto riguarda $\widetilde{\chi}$ la risposta non è ovvia, a meno che A non sia diagonalizzabile ed in quel caso $\widetilde{\chi} = \chi$. Tuttavia, è intuivo che in ogni caso dovrebbero esistere formule che permettono di esprimere $\widetilde{\chi}$ in funzione di χ. Vedremo ora che, abbastanza sorprendentemente, la matrice P di Tartaglia è coinvolta in tale formule.

Un risultato preliminare:

1. $(I + Z^T)P(I - Z^T) = P$

Dimostrazione. $Z^TP - PZ^T - Z^TPZ^T =$

$$\begin{bmatrix} 0 & 1 & 0 & . \\ 0 & 0 & 1 & . \\ 0 & 0 & 0 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & . \\ 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ . & . & . & 1 & . \\ . & . & . & . & . \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & . \\ 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ . & . & . & 1 & . \\ . & . & . & . & . \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & . \\ 0 & 0 & 1 & . \\ 0 & 0 & 0 & . \\ . & . & . & . \end{bmatrix} -$$

$$\begin{bmatrix} 0 & 1 & 0 & . \\ 0 & 0 & 1 & . \\ 0 & 0 & 0 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & . \\ 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ . & . & . & 1 & . \\ . & . & . & . & . \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & . \\ 0 & 0 & 1 & . \\ 0 & 0 & 0 & . \\ . & . & . & . \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ 0 & 0 & 0 & 1 & . \\ . & . & . & . & . \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 1 & . \\ 0 & 0 & 1 & 2 & . \\ 0 & 0 & 0 & 1 & . \\ . & . & . & . & . \end{bmatrix} - \begin{bmatrix} 0 & 1 & 2 & 3 & . \\ 0 & 0 & 1 & 3 & . \\ 0 & 0 & 0 & 1 & . \\ . & . & . & . & . \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & . \\ 0 & 0 & 1 & . \\ 0 & 0 & 0 & . \\ . & . & . & . \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 0 & 1 & 2 & . \\ 0 & 0 & 0 & 1 & . \\ 0 & 0 & 0 & 0 & . \\ . & . & . & . & . \end{bmatrix} - \begin{bmatrix} 0 & 0 & 1 & 2 & . \\ 0 & 0 & 0 & 1 & . \\ 0 & 0 & 0 & 0 & . \\ . & . & . & . & . \end{bmatrix} = \bigcirc.$$

□

2. $\mu \neq 0$, $d(\mu^j) = d_s(\mu^j) = \mathrm{diag}(\mu^j, j = 0, \ldots, s-1) \Rightarrow$
$\Rightarrow d(\mu^j)(I + Z^T) = (I + \frac{1}{\mu} Z^T) d(\mu^j)$

Dimostrazione.
$$\begin{bmatrix} 1 & 0 & 0 & . \\ 0 & \mu & 0 & . \\ 0 & 0 & \mu^2 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & . \\ 0 & 1 & 1 & . \\ 0 & 0 & 1 & . \\ . & . & . & . \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . \\ 0 & \mu & \mu & 0 & . \\ 0 & 0 & \mu^2 & \mu^2 & . \\ . & . & 0 & \mu^3 & . \\ . & . & . & . & . \end{bmatrix} =$$
$$\begin{bmatrix} 1 & \frac{1}{\mu} & 0 & . \\ 0 & 1 & \frac{1}{\mu} & . \\ 0 & 0 & 1 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & . \\ 0 & \mu & 0 & . \\ 0 & 0 & \mu^2 & . \\ . & . & . & . \end{bmatrix}$$
□

3. $(I - Z^T) d((-\mu)^j) = d((-\mu)^j)(I + \mu Z^T)$

Dimostrazione.
$$\begin{bmatrix} 1 & -1 & 0 & . \\ 0 & 1 & -1 & . \\ 0 & 0 & 1 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & . \\ 0 & -\mu & 0 & . \\ 0 & 0 & \mu^2 & . \\ . & . & . & . \end{bmatrix} =$$
$$\begin{bmatrix} 1 & \mu & 0 & . & . \\ 0 & -\mu & -\mu^2 & 0 & . \\ 0 & 0 & \mu^2 & \mu^3 & . \\ . & . & 0 & -\mu^3 & . \\ . & . & . & . & . \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & . \\ 0 & -\mu & 0 & . \\ 0 & 0 & \mu^2 & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} 1 & \mu & 0 & . & . \\ 0 & 1 & \mu & 0 & . \\ . & 0 & 1 & \mu & . \\ . & . & . & . & . \end{bmatrix}$$
□

Dalle precedenti tre identità matriciali segue che:

$$d(\mu^j) P d((-\mu)^j) = d(\mu^j)(I + Z^T) P (I - Z^T) d((-\mu)^j) =$$
$$= (I + \frac{1}{\mu} Z^T) d(\mu^j) P d((-\mu)^j)(I + \mu Z^T)$$

cioé:

$$\mu \in \mathbb{C}, \mu \neq 0, M = d(\mu^j) P d((-\mu)^j) \Rightarrow M = (I + \frac{1}{\mu} Z^T) M (I + \mu Z^T). \qquad (5.2)$$

Il risultato (5.2) ci dice che l'ipotesi (5.3) nel seguente Teorema è soddisfatta da almeno una matrice M.

Teorema 5.2.1. *Assumiamo che* $A \in \mathbb{C}^{n \times n}$, $X \in \mathbb{C}^{n \times s}$, $\mu \in \mathbb{C}$, $\mu \neq 0$ *sono tali che* $AX = X(\mu I_s + Z_s^T) = X \begin{bmatrix} \mu & 1 & 0 & . \\ 0 & \mu & . & 0 \\ . & . & . & 1 \\ 0 & . & 0 & \mu \end{bmatrix}$ *dove* s *è il numero di righe e colonne*

dell'ultima matrice scritta.

Allora $A^{-1}X = X(\mu I_s + Z_s^T)^{-1} = X \begin{bmatrix} \mu & 1 & 0 & . \\ 0 & \mu & . & . \\ . & . & . & 1 \\ 0 & . & 0 & \mu \end{bmatrix}^{-1}$

Se M *è una matrice* $s \times s$ *tale che*

$$(\mu I_s + Z_s^T)^{-1} M = M(\frac{1}{\mu} I_s + Z_s^T) \tag{5.3}$$

allora

$A^{-1}\widetilde{X} = \widetilde{X}(\frac{1}{\mu} I_s + Z_s^T) = \widetilde{X} \begin{bmatrix} \frac{1}{\mu} & 1 & 0 & . \\ 0 & \frac{1}{\mu} & . & 0 \\ . & . & . & 1 \\ 0 & . & 0 & \frac{1}{\mu} \end{bmatrix}$, *per* $\widetilde{X} = XM$

Sia $\chi = [X_1, \dots X_m], J =$*diag(*$\mu_k I_{s_k} + Z_{s_k}^T, k = 1, \dots, m$*) una coppia di Jordan per* A*, cioé* $AX_k = X_k(\mu_k I_{s_k} + Z_{s_k}^T), k = 1, \dots, m$ *,* χ *è invertibile, e quindi* $\chi^{-1}A\chi = J$*; se* M_k *sono matrici* $s_k \times s_k$ *invertibili tali che*
$(\mu_k I_{s_k} + Z_{s_k}^T)^{-1} M_k = M_k(\frac{1}{\mu_k} I_{s_k} + Z_{s_k}^T)$ *e* $\widetilde{X}_k = X_k M_k, k = 1, \dots, m,$
allora
$\widetilde{\chi} = [\widetilde{X_1}, \dots \widetilde{X_m}], \widetilde{J} =$ *diag(*$\frac{1}{\mu} I_{s_k} + Z_{s_k}^T, k = 1, \dots m$*) è una coppia di Jordan per* A^{-1}
cioé
$A^{-1}\widetilde{X_k} = \widetilde{X_k}(\frac{1}{\mu_k} I_{s_k} + Z_{s_k}^T), k = 1, \dots m,$ $\widetilde{\chi}$ *è invertibile, e quindi* $\widetilde{\chi}^{-1}A^{-1}\widetilde{\chi} = \widetilde{J}$ $(m \le n, s_1 + s_2 + \dots + s_m = n)$.

Già conosciamo una matrice invertibile M che soddisfa (5.3) che è chiaramente definita in funzione della matrice P di Tartaglia. Ora affermiamo che ogni matrice M nello spazio vettoriale

$$\mathcal{L} = \mathcal{L}_s^\mu = \{M \in \mathbb{C}^{s \times s} : (5.3) \text{ vale}\},$$

può essere scritta in termini della matrice di Tartaglia. Poi un problema potrebbe essere trovare la $M \in \mathcal{L}$ piú economica, cioé quella per la quale il calcolo del prodotto XM richiede il numero minimo di operazioni. Prima notiamo che la condizione (5.3) vale se e solo se

$$M = \begin{bmatrix} \mu & 1 & 0 & . \\ 0 & \mu & . & 0 \\ . & . & . & 1 \\ 0 & . & 0 & \mu \end{bmatrix} M \begin{bmatrix} \frac{1}{\mu} & 1 & 0 & . \\ 0 & \frac{1}{\mu} & . & 0 \\ . & . & . & 1 \\ 0 & . & 0 & \frac{1}{\mu} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{\mu} & 0 & . \\ 0 & 1 & . & 0 \\ . & . & . & \frac{1}{\mu} \\ 0 & . & 0 & 1 \end{bmatrix} M \begin{bmatrix} 1 & \mu & 0 & . \\ 0 & 1 & . & 0 \\ . & . & . & \mu \\ 0 & . & 0 & 1 \end{bmatrix} \tag{5.4}$$

se e solo se

$$\mu M Z^T + \frac{1}{\mu} Z^T M + Z^T M Z^T = 0 \tag{5.5}$$

Osserviamo che se M soddisfa (5.5) e T_1, T_2 sono matrici triangolari superiori di Toeplitz, allora anche le matrici T_1M, MT_2, T_1MT_2 soddisfano (5.5) perché le matrici superiori triangolari di Toeplitz commutano con Z^T (sono le uniche matrici che commutano con Z^T). Inoltre si può anche osservare che se M soddisfa (5.5) e ha una colonna o una riga nulla, diciamo la r-esima, allora anche tutte le precedenti (successive) colonne (righe) devono essere nulle.
Elenchiamo ora alcuni risultati significativi ottenuti studiando $\mathcal{L}$, tra i quali la sua completa caratterizzazione.

(a) Notiamo che:

$\Gamma = \Gamma_s = d_s(\mu^j) P_s d_s((-\mu)^j) \Rightarrow \Gamma \in \mathcal{L}$ vedi (5.2)

$$\Gamma = \Gamma_s = \begin{bmatrix} 1 & -\mu & \mu^2 & -\mu^3 & \mu^4 & . & (-1)^{s-1}\mu^{s-1} \\ 0 & -\mu^2 & 2\mu^3 & -3\mu^4 & 4\mu^5 & . & (-1)^{s-1}(s-1)\mu^s \\ 0 & 0 & \mu^4 & -3\mu^5 & 6\mu^6 & . & . \\ 0 & 0 & 0 & -\mu^6 & 4\mu^7 & . & . \\ 0 & 0 & 0 & 0 & \mu^8 & . & . \\ . & . & . & . & . & . & (-1)^{s-1}(s-1)\mu^{2s-3} \\ 0 & 0 & 0 & 0 & 0 & . & (-1)^{s-1}\mu^{2s-2} \end{bmatrix}.$$

(b) $G = G_s := \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & -\mu^2\Gamma_{s-1} \end{bmatrix} \Rightarrow G \in \mathcal{L}$

Dimostrazione. $\begin{bmatrix} 1 & \frac{1}{\mu}e_1^T \\ \underline{0} & I_{s-1} + \frac{1}{\mu}Z_{s-1}^T \end{bmatrix} \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & -\mu^2\Gamma_{s-1} \end{bmatrix} \begin{bmatrix} 1 & \mu e_1^T \\ \underline{0} & I_{s-1} + \mu Z_{s-1}^T \end{bmatrix} =$

$$= \begin{bmatrix} 1 & -\mu \begin{bmatrix} 1 & -\mu & \mu^2 & . & . & . \end{bmatrix} \\ 0 & -\mu^2(I_{s-1} + \frac{1}{\mu}Z_{s-1}^T)\Gamma_{s-1} \end{bmatrix} \begin{bmatrix} 1 & \mu e_1^T \\ 0 & I_{s-1} + \mu Z_{s-1}^T \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & \mu e_1^T + \begin{bmatrix} -\mu & \mu^2 & -\mu^3 & . & . & . \end{bmatrix} (I_{s-1} + \mu Z_{s-1}^T \\ 0 & -\mu^2(I_{s-1} + \frac{1}{\mu}Z_{s-1}^T)\Gamma_{s-1}(I_{s-1} + \mu Z_{s-1}^T) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & \begin{bmatrix} \mu & 0 & . & . & . & 0 \end{bmatrix} + \begin{bmatrix} -\mu & 0 & . & . & . & 0 \end{bmatrix} \\ 0 & -\mu^2\Gamma_{s-1} \end{bmatrix}$$

In quest'ultima uguaglianza usiamo il fatto che $\Gamma_{s-1} \in \mathcal{L}_{s-1}^{\mu}$ (vedi il punto (a)).
Notare che

$$G = G_s = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & . & 0 \\ 0 & -\mu^2 & \mu^3 & -\mu^4 & \mu^5 & . & (-1)^{s-1}\mu^s \\ 0 & 0 & \mu^4 & -2\mu^5 & 3\mu^6 & . & (-1)^{s-1}(s-2)\mu^{s+1} \\ 0 & 0 & 0 & -\mu^6 & 3\mu^7 & . & . \\ 0 & 0 & 0 & 0 & \mu^8 & . & (-1)^{s-1}(s-2)\mu^{2s-3} \\ 0 & 0 & 0 & 0 & 0 & . & (-1)^{s-1}\mu^{2s-2} \end{bmatrix} \qquad \square$$

(c) $GT \in \mathcal{L}$ per tutte le matrici T $s \times s$ di Toeplitz triangolari superiori.

(d) Per s = 2,3,4,5 può essere facilmente dimostrato che $M \in \mathcal{L} \Rightarrow$ implica, rispettivamente,

$$M = \begin{bmatrix} a & b \\ 0 & -a\mu^2 \end{bmatrix}, M = \begin{bmatrix} a & b & c \\ 0 & -a\mu^2 & -\mu^2 b + \mu^3 a \\ 0 & 0 & \mu^4 a \end{bmatrix},$$

$$M = \begin{bmatrix} a & b & c & d \\ 0 & -a\mu^2 & -\mu^2 b + \mu^3 a & -\mu^2 c + \mu^3 b - \mu^4 a \\ 0 & 0 & \mu^4 a & \mu^4 b - 2\mu^5 a \\ 0 & 0 & 0 & -\mu^6 a \end{bmatrix},$$

$$M = \begin{bmatrix} a & b & c & d & e \\ 0 & -a\mu^2 & -\mu^2 b + \mu^3 a & -\mu^2 c + \mu^3 b - \mu^4 a & -\mu^2 d + \mu^3 c - \mu^4 b + \mu^5 a \\ 0 & 0 & \mu^4 a & \mu^4 b - 2\mu^5 a & c\mu^4 - 2b\mu^5 + 3a\mu^6 \\ 0 & 0 & 0 & -\mu^6 a & -b\mu^6 + 3a\mu^7 \\ 0 & 0 & 0 & 0 & a\mu^8 \end{bmatrix}.$$

In altre parole possiamo congetturare che per s generico M debba essere del tipo $aG + bGZ^T + cG(Z^T)^2 + \ldots + (.)G(Z^T)^{s-1}$ In effetti, come notato in [Di Fiore, talk in Shenzhen], se $M = (I + \frac{1}{\mu}Z^T)M(I + \mu Z^T)$, allora, per ogni matrice invertibile $L \in \mathcal{L}$, abbiamo

$$L^{-1}M = L^{-1}(I + \frac{1}{\mu}Z^T)M(I + \mu Z^T) = (I + \mu Z^T)^{-1}L^{-1}M(I + \mu Z^T)$$

i.e. $(I + \mu Z^T)L^{-1}M = L^{-1}M(I + \mu Z^T)$. Ma quest'ultima uguaglianza implica che $L^{-1}M$ deve essere triangolare superiore di Toeplitz, cioe', M deve essere nello spazio $\{LT : T = \text{triangolare superiore di Toeplitz}\}$. Quindi, per il punto (c),

$\mathcal{L} = \mathcal{L}_s^\mu = \{M \in \mathbb{C}^{s\times s} : (5.3) \text{ vale }\} = \text{span}\{G, GZ^T, G(Z^T)^2, \ldots, G(Z^T)^{s-1}\}$
(si noti che la generica matrice di $\mathcal{L}$ è determinata dalla prima sua riga).

Problema:
Quale è la matrice $M \in \mathcal{L}$ invertibile per cui il calcolo del prodotto XM (vedi il Teorema 5.2.1) richiede il minimo numero di operazioni?

(e) $\Gamma = GT((-\mu)^j)$, dove

$$T((-\mu)^j) = T_s((-\mu)^j) = \begin{bmatrix} 1 & -\mu & \mu^2 & . & (-\mu)^{s-1} \\ 0 & 1 & -\mu & \mu^2 & . \\ 0 & 0 & 1 & . & . \\ . & . & . & . & -\mu \\ 0 & . & . & 0 & 1 \end{bmatrix}$$

Dimostrazione. Γ e $GT((-\mu)^j)$ hanno la stessa prima riga e sono entrambi in $\mathcal{L}$ (dai punti (a),(c)); quindi per il punto (d) devono coincidere.

□

(f) $P_k = \begin{bmatrix} 1 & \begin{bmatrix} 1 & 1 & 1 & . & 1 & 1 \end{bmatrix} \\ 0 & P_{k-1} \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 0 & 1 & 1 & . & 1 \\ . & . & 1 & . & . \\ 0 & . & . & . & 1 \\ 0 & . & . & 0 & 1 \end{bmatrix} \end{bmatrix}$ e così

$$P_k^{-1} = \begin{bmatrix} 1 & -\begin{bmatrix} 1 & 0 & . & . & . & 0 \end{bmatrix} P_{k-1}^{-1} \\ 0 & \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 0 & 1 & 1 & . & 1 \\ . & . & 1 & . & . \\ 0 & . & . & . & 1 \\ 0 & . & . & 0 & 1 \end{bmatrix}^{-1} P_{k-1}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & -\begin{bmatrix} 1 & 0 & . & . & . & 0 \end{bmatrix} P_{k-1}^{-1} \\ \underline{0} & \begin{bmatrix} 1 & -1 & 0 & . & 0 \\ 0 & 1 & -1 & . & 0 \\ . & . & 1 & . & . \\ 0 & . & . & . & -1 \\ 0 & . & . & 0 & 1 \end{bmatrix} P_{k-1}^{-1} \end{bmatrix}.$$

Dimostrazione. Dai punti (a) e (e) abbiamo che

$$\begin{aligned} & P_s = d_s(\mu^j)^{-1} G_s T_s((-\mu)^j) d_s((-\mu)^j)^{-1} = \\ &= \begin{bmatrix} 1 & 0^T \\ 0 & -\mu P_{s-1} d_{s-1}((-\mu)^j) \end{bmatrix} \begin{bmatrix} 1 & \begin{bmatrix} 1 & 1 & 1 & . & 1 & 1 \end{bmatrix} \\ 0 & -T_{s-1}((-\mu)^j) \frac{1}{\mu} d_{s-1}((-\mu)^j)^{-1} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & \begin{bmatrix} 1 & 1 & 1 & . & 1 & 1 \end{bmatrix} \\ 0 & \mu P_{s-1} d_{s-1}((-\mu)^j) T_{s-1}((-\mu)^j) \frac{1}{\mu} d_{s-1}((-\mu)^j)^{-1} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & \begin{bmatrix} 1 & 1 & 1 & . & 1 & 1 \end{bmatrix} \\ 0 & P_{s-1} \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 0 & 1 & 1 & . & 1 \\ . & . & 1 & . & . \\ . & . & . & . & 1 \\ 0 & . & . & 0 & 1 \end{bmatrix} \end{bmatrix} \end{aligned}$$

poi cambiamo s con k. □

L'applicazione ripetuta della rappresentazione (f) di P_k^{-1} per $k = s, s-1, \cdots 3$ genera una procedura per calcolare $P_s^{-1} v, v \in \mathbb{C}^s$, che evita il calcolo degli elementi di P_s^{-1} e richiede solo $s(s-1)/2$ operazioni additive. Infatti, se $c(k) =$ numero di operazioni additive sufficienti per calcolare un prodotto matrice-vettore che coinvolge P_k^{-1}, da (f) abbiamo che
$c(k) \leq c(k-1) + (k-2) + 1$. Ne segue che
$c(s) \leq c(s-1) + (s-2) + 1 \leq (c(s-2) + (s-3) + 1) + (s-2) + 1 \leq \cdots \leq (c(2) + (1) + 1) + (2) + 1 + \ldots + (s-2) + 1 = (s-1) + \frac{(s-2)(s-1)}{2} = s(s-1)/2$.
La procedura dettagliata e' riportata qui sotto

input: $v_1, v_2, \ldots, v_s$
$i = 1$
$\bullet_1 = v_s$
for $i = 1, \ldots, s-1$ do {

$z = \bullet_1$;
$app(1) = \bullet_1$;
$\bullet_1 = v_{s-i} - z$; $\bullet_{i+1} = 0$; (1 additive op)
for $j = 2, \ldots, i+1$ do {
$app(j) = \bullet_j$;
$\bullet_j = app(j-1) - app(j)$; $(i-1$ additive op)
}
}

5.3 Appendice capitolo 5

Come si dimostra in questa Appendice, lo studio della relazione tra le coppie di Jordan di A e A^{-1}, effettuato nella precedente Sezione 5.2, può essere utile nello studio delle velocità di convergenza delle successioni generate dal metodo delle potenze, quando applicato ad A e ad A^{-1}, nell'ipotesi che sia A sia A^{-1} abbiano un autovalore dominante con molteplicità algebrica e geometrica coincidenti.

Data $A \in \mathbb{C}^{n\times n}$ invertibile, con autovalori $\lambda_1 = \ldots = \lambda_t$, $\lambda_i\,(i = t+1, \ldots, n-q)$, $\lambda_{n-q+1} = \ldots = \lambda_n$, dove $t = m_a(\lambda_1) = m_g(\lambda_1), q = m_a(\lambda_n) = m_g(\lambda_n)$, esistono matrici invertibili $\chi = [\chi_1 \ldots \chi_n]$ e $\widetilde{\chi} = [\widetilde{\chi}_1 \ldots \widetilde{\chi}_n]$, $\widetilde{\chi}_i = \chi_i, i = 1, \ldots, t, n-q+1, \ldots, n$ tali che

$$\chi^{-1}A\chi = J = \begin{bmatrix} \lambda_1 I_t & \bigcirc & \bigcirc \\ \bigcirc & K & \bigcirc \\ \bigcirc & \bigcirc & \lambda_n I_q \end{bmatrix}, \ \widetilde{\chi}^{-1}A^{-1}\widetilde{\chi} = \widetilde{J} = \begin{bmatrix} \frac{1}{\lambda_1} I_t & \bigcirc & \bigcirc \\ \bigcirc & \widetilde{K} & \bigcirc \\ \bigcirc & \bigcirc & \frac{1}{\lambda_n} I_q \end{bmatrix}$$

dove K e $\widetilde{K}$ sono matrici diagonalia blocchi con blocchi diagonali $s \times s$ del tipo, rispettivamente

$$\begin{bmatrix} \mu & 1 & 0 & .0 & \\ 0 & \mu & 1 & . & . \\ 0 & . & . & . & 0 \\ 0 & . & . & . & 1 \\ 0 & . & . & 0 & \mu \end{bmatrix}, \text{ e } \begin{bmatrix} \frac{1}{\mu} & 1 & 0 & .0 & \\ 0 & \frac{1}{\mu} & 1 & . & . \\ 0 & . & . & . & 0 \\ 0 & . & . & . & 1 \\ 0 & . & . & 0 & \frac{1}{\mu} \end{bmatrix},$$

con $\mu \in \{\lambda_{t+1}, \ldots, \lambda_{n-q}\}, 1 \le s \le n-t-q$ rispettivamente.

Le matrici J e $\widetilde{J}$ sono note come le forme canoniche di Jordan di A e A^{-1}.

Assumiamo $|\lambda_1| > |\lambda_{t+1}| \ge \ldots \ge |\lambda_{n-q}| > |\lambda_n|$. Sia v con $||v|| = 1$ tale che si può esprimere come combinazione lineare delle colonne di χ, cioè come

$$v = \sum_{j\le t} \alpha_j \chi_j + \sum_{t<j<n-q+1} \alpha_j \chi_j + \sum_{j\ge n-q+1} \alpha_j \chi_j$$

con i vettori $\alpha = \sum_{j\le t} \alpha_j \chi_j$ e $\hat{\alpha} = \sum_{j\ge n-q+1} \alpha_j \chi_j$ entrambi non nulli (si nota che $A\alpha = \lambda_1\alpha$, $A\hat{\alpha} = \lambda_n\hat{\alpha}$, $A^{-1}\hat{\alpha} = \frac{1}{\lambda_n}\hat{\alpha}$).

Poniamo $v_0 = v'_0 = v$, e scegliamo due vettori u e u'.

Se le successioni: $a_k = Av_{k-1}$; $\phi_k = \frac{u^H a_k}{u^H v_{k-1}}$; $v_k = \frac{1}{||a_k||} a_k$; $a'_k = A^{-1} v'_{k-1}$; $\phi'_k = \frac{u'^H a'_k}{u'^H v'_{k-1}}$; con $k = 1,2,\ldots,$ sono ben definite, e se $u^H \alpha \neq 0$, $u'^H \hat{\alpha} \neq 0$, allora

$$v_k = \frac{A^k v}{||A^k v||}, \left(\frac{|\lambda_1|}{\lambda_1}\right)^k v_k \to \frac{\alpha}{||\alpha||}; \; \phi_k = \frac{u^H A^k v}{u^H A^{k-1} v} \to \lambda_1,$$

$$v'_k = \frac{A^{-k} v}{||A^{-k} v||}, \left(\frac{|\lambda_n|}{\lambda_n}\right)^k v'_k \to \frac{\hat{\alpha}}{||\hat{\alpha}||}; \; \phi'_k = \frac{u'^H A^{-k} v}{u'^H A^{-(k-1)} v} \to \frac{1}{\lambda_n}.$$

Vediamo alcuni dettagli della dimostrazione per quanto riguarda le conclusioni su v_k e ϕ_k. Notiamo che

$$A^k v = \lambda_1^k \left(\sum_{j \leq t} \alpha_j \chi_j \right) + \sum_{t < j < n-q+1} \alpha_j A^k \chi_j + \lambda_n^k \left(\sum_{j \geq n-q+1} \alpha_j \chi_j \right).$$

Grazie alla semplice forma di K, si può ottenere un'espressione esplicita per $A^k \chi_j$ per ogni j, $t < j < n - q + 1$, che consente di concludere che $\frac{1}{\lambda_1^k} A^k \chi_j$, $t < j < n - q + 1$, converge a 0 e che $\frac{1}{\lambda_1^k} A^k v \to \alpha$ con tasso di convergenza

$$\bigcirc \left(max_{j:\lambda_j \neq \lambda_1} |p_{\lambda_j}(k)| \left| \frac{\lambda_j}{\lambda_1} \right|^k \right)$$

con p_{λ_j} polinomio di grado uguale all'ordine del blocco di Jordan di λ_j di massimo ordine meno uno (ad esempio, il grado di p_{λ_n} è 0).
Quindi

$\frac{1}{\lambda_1^k} A^k v \to \alpha$, $\frac{1}{|\lambda_1|^k} ||A^k v|| \to ||\alpha||$, $\frac{1}{\lambda_1^k} u^H A^k v \to u^H \alpha$,

$\frac{1}{\lambda_1^{k-1}} u^H A^{k-1} v \to u^H \alpha$, $\frac{u^H A^k v}{u^H A^{k-1} v} \to \lambda_1$.

Vediamo ora alcuni dettagli di dimostrazione, per quanto riguarda le conclusioni su v'_k e ϕ'_k. Osserviamo che

$$A^{-k} v = \frac{1}{\lambda_1^k} \sum_{j \leq t} \alpha_j \chi_j + \sum_{t < j < n-q+1} \alpha_j A^{-k} \chi_j + \frac{1}{\lambda_n^k} \sum_{j \geq n-q+1} \alpha_j \chi_j$$

Qui per ottenere un'espressione esplicita per la successione di vettori $A^{-k} \chi_j$, $t < j < n - q + 1$, vedere se $\lambda_n^k A - k \chi_j \to 0$ e vedere la sua velocità di convergenza, dovremmo trattare K^{-1}, che è ancora diagonale a blocchi, ma i suoi blocchi diagonali sono matrici di Toeplitz triangolari superiori piene (non bidiagonali). Così non sembra facile andare avanti; allora si esprime v in termini delle colonne $\widetilde{\chi_j}$

della matrice $\widetilde{\chi}$ che trasforma A^{-1} nella sua forma di Jordan $\widetilde{J}$ e si sfrutta la forma semplice di $\widetilde{K}$.

Quindi si riscrive v come segue

$$v = \sum_{j \leq t} \alpha_j \chi_j + \sum_{t<j<n-q+1} \widetilde{\alpha}_j \widetilde{\chi}_j + \sum_{j \geq n-q+1} \alpha_j \chi_j \ .$$

Grazie alla forma semplice di $\widetilde{K}$, si possono ottenere facilmente espressioni esplicite per $A^{-k}\widetilde{\chi}_j$, $t < j < n-q+1$, che risultano essere analoghe a quelle di $A^k\chi_j$, $t < j < n-q+1$ (perché $\widetilde{K}$ è come K eccetto per μ che viene sostituito con $\frac{1}{\mu}$), ma si noti che esse sono in termini delle $\widetilde{\chi}_j$ (invece di χ_j) e di particolari piccole potenze di $\frac{1}{\mu}$ (invece di μ).

Quindi, le conclusioni dello studio del tasso di convergenza a 0 di $\lambda_n^k A^{-k}\widetilde{\chi}_j$, $t < j < n-q+1$, sono analoghe, ma con un importante differenza, cioè il coefficiente di $max_{j:\lambda_j \neq \lambda_n} |p_{\lambda_j}(k)| \left|\frac{\lambda_n}{\lambda_j}\right|^k$ può avere una grandezza molto diversa rispetto al coefficiente di $max_{j:\lambda_j \neq \lambda_1} |p_{\lambda_j}(k)| \left|\frac{\lambda_j}{\lambda_1}\right|^k$ perchè coinvolge $\widetilde{\chi}_j$ e alcune particolari piccole potenze di $\frac{1}{\mu}$ in $\widetilde{K}$, invece di χ_j e le stesse piccole potenze di μ in K.

Per capire meglio quanto questi coefficienti possono essere diversi, sarebbe opportuno avere a disposizione le espressioni di $\widetilde{\chi}_j$ in funzione di χ_j. Queste espressioni sono state ottenute nelle precedente Sezione 5.2.

5.4 Appendice Bis

FROBENIUS

Studiare la condizione $\mathcal{M} = J\mathcal{M}\widetilde{J}$ nel caso della forma di Frobenius, equivale a studiare lo spazio vettoriale di tutte le M per cui

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a & b & c & d \end{bmatrix} M \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{a} & -\frac{d}{a} & -\frac{c}{a} & -\frac{b}{a} \end{bmatrix} .$$

Più in generale, la matrice $s \times s$

$$H = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & \dots & -a_{s-1} \end{bmatrix}$$

con $a_0 \neq 0$, ha $\lambda^s + a_{s-1}\lambda^{s-1} + \ldots + a_1\lambda + a_0$ come polinomio caratteristico.

La matrice $s \times s$

$$\widetilde{H} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -\frac{1}{a_0} & -\frac{a_{s-1}}{a_0} & -\frac{a_{s-2}}{a_0} & \dots & -\frac{a_1}{a_0} \end{bmatrix}$$

ha $\lambda^s + \frac{a_1}{a_0}\lambda^{s-1} + \ldots + \frac{a_{s-2}}{a_0}\lambda^2 + \frac{a_{s-1}}{a_0}\lambda + \frac{1}{a_0}$ come polinomio caratteristico.
Si nota che le radici del secondo polinomio sono l'inverso delle radici del primo polinomio.
La matrice H può essere un blocco diagonale della forma canonica di Frobenius J di una matrice $A \in \mathbb{C}^{n\times n}$ e $\widetilde{H}$ il corrispondente blocco diagonale della forma di Frobenius $\widetilde{J}$ di A^{-1}. Se χ e $\widetilde{\chi}$ sono tali che $\chi^{-1}A\chi = J$ e $\widetilde{\chi}^{-1}A^{-1}\widetilde{\chi} = \widetilde{J}$ allora $\widetilde{\chi} = \chi\mathcal{M}$ dove $\mathcal{M}$ è una matrice diagonale a blocchi con blocchi diagonali M tali che M è invertibile e $M = HM\widetilde{H}$.

Studiamo l'insieme delle matrice M, $s \times s$ tali che $M = HM\widetilde{H}$. Per $s = 2$:

$$\begin{bmatrix} x & y \\ w & z \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} x & y \\ w & z \end{bmatrix} \begin{bmatrix} 0 & 1 \\ \frac{1}{a} & -\frac{b}{a} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} x & y \\ y+bx & xa \end{bmatrix}, \begin{bmatrix} \frac{z}{a} & w - \frac{bz}{a} \\ w & z \end{bmatrix}.$$

Notiamo che

$$\begin{bmatrix} y+bx & ax \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} JHJ \text{ dove } J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ e}$$

$$\begin{bmatrix} \frac{z}{a} & w - \frac{bz}{a} \end{bmatrix} = \begin{bmatrix} w & z \end{bmatrix} \widetilde{H}.$$

Per $s = 3$:

$$\begin{bmatrix} x & y & r \\ w & z & s \\ t & u & v \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a & b & c \end{bmatrix} \begin{bmatrix} x & y & r \\ w & z & s \\ t & u & v \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{a} & -\frac{c}{a} & -\frac{b}{a} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} x & y & r \\ y+cx & r+bx & ax \\ r+cy+(c^2+b)x & by+(a+bc)x & ay+acx \end{bmatrix},$$

$$\begin{bmatrix} \frac{u}{a} - \frac{b}{a^2}v & (\frac{1}{a} + \frac{bc}{a^2})v - \frac{c}{a}u & t - \frac{b}{a}u + (\frac{b^2}{a^2} - \frac{c}{a})v \\ \frac{v}{a} & t - \frac{c}{a}v & u - \frac{b}{a}v \\ t & u & v \end{bmatrix}.$$

Vediamo che

$\left[\begin{array}{ccc} y+cx & r+bx & ax \end{array}\right] = \left[\begin{array}{ccc} x & y & r \end{array}\right] JHJ$ dove $J = \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array}\right]$ e

$\left[\begin{array}{ccc} r+cy+(c^2+b)x & by+(a+bc)x & ay+acx \end{array}\right] = \left[\begin{array}{ccc} x & y & r \end{array}\right] (JHJ)^2,$

$\left[\begin{array}{ccc} \frac{v}{a} & t-\frac{c}{a}v & u-\frac{b}{a}v \end{array}\right] = \left[\begin{array}{ccc} t & u & v \end{array}\right] \widetilde{H},$

$\left[\begin{array}{ccc} \frac{u}{a}-\frac{b}{a^2}v & (\frac{1}{a}+\frac{bc}{a^2})v-\frac{c}{a}u & t-\frac{b}{a}u+(\frac{b^2}{a^2}-\frac{c}{a})v \end{array}\right] = \left[\begin{array}{ccc} t & u & v \end{array}\right] \widetilde{H}^2.$

Congettura:

$M \in \mathbb{C}^{s\times s}$ è tale che $M = HM\widetilde{H}$ se e solo se $e_i^T M = e_1^T M(JHJ)^{i-1}$ dove $i = 1, \ldots, s$ se e solo se
$e_i^T M = e_s^T M \widetilde{H}^{s-i}$ con $i = 1, \ldots, s$ ($\Rightarrow v^T JM = (e_s^T M)\mathcal{H}(v)$), dove

$$JHJ = \left[\begin{array}{ccccc} -a_{s-1} & . & . & -a_1 & -a_0 \\ 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & . \\ . & . & . & 0 & . \\ 0 & . & 0 & 1 & 0 \end{array}\right], \quad J = \left[\begin{array}{cccc} 0 & . & 0 & 1 \\ . & . & 1 & 0 \\ 0 & . & . & . \\ 1 & 0 & . & 0 \end{array}\right] \text{ e}$$

$$\widetilde{H} = \left[\begin{array}{ccccc} 0 & 1 & 0 & . & 0 \\ 0 & 0 & 1 & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & 1 \\ -1/a_0 & -a_{s-1}/a_0 & -a_{s-2}/a_0 & . & -a_1/a_0 \end{array}\right]$$

$\widetilde{\mathcal{H}}$ è l'algebra di Hessenberg generata da $\widetilde{H}$;
$\widetilde{H}(\underline{v})$ è matrice in $\widetilde{\mathcal{H}}$ con la prima riga v^T con $v \in \mathbb{C}^s$ generico.

Capitolo 6

Il problema di Procruste Toeplitz

Analizziamo come un problema dei minimi quadrati ove la soluzione è una matrice $n \times n$ vincolata a stare in sottospazi di matrici di Toeplitz si può risolvere utilizzando la decomposizione dei valori singolari[Eberle,1999], [Eberle, Maciel, 2001]. Per il caso simmetrico di Toeplitz viene proposto un algoritmo basato sul metodo della proiezione alternato.

Denotiamo con $||A||_F^2$ la norma di Frobenius di una matrice $A \in \mathbb{R}^{m\times n}$ definita come

$$||A||_F^2 = \text{trace}(A^T A) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2$$

e consideriamo il seguente problema generale dei minimi quadrati vincolati:

$$min||AX - B||_F^2 \quad \text{tale che} \quad X \in \mathbb{P} \tag{6.1}$$

dove $A, B \in \mathbb{R}^{m\times n}$, con $m > n$ e $\mathbb{P} \subset \mathbb{R}^{n\times n}$ é un sottoinsieme di matrici dato.

Negli ultimi decenni è stata condotta una vasta attività di ricerca su questo tipo di problemi che si presentano in molti settori. Un esempio classico, in statistica, è il problema di trovare la matrice simmetrica definita positiva più vicina ad una matrice di covarianza campione[11 dell'articolo]. Questo caso, delle matrice simmetriche semidefinite positive, è stato analizzato da Higham [Higham,1988].
Nei problemi delle strutture elastiche, la determinazione della matrice di deformazione che mette in relazione le forze f_i con lo spostamento d_i da esse causato secondo le identità $Xf_i = d_i$ si ottiene risolvendo la (6.1). Il caso ortogonale, studiato da Schonemann[13 dell'articolo] e da Higham [Higham,1986], è conosciuto in letteratura come problema ortogonale e simmetrico di "Procruste". Quando $\mathbb{P}$ è un sottoinsieme di matrici persimmetriche (cioè simmetriche rispetto alla diagonale NE-SW) i risultati sono studiati da Eberle [Eberle,1999], [Eberle, Maciel, 2001].

6.1 Caso P=τ ={matrici di Toeplitz }

Si vuole risolvere il problema

$$min||AX - B||_F^2 \quad \text{tale che} \quad X \in \tau \tag{6.2}$$

dove $\tau \subset \mathbb{R}^{n\times n}$ è l'insieme delle matrici di Toeplitz. Ricordiamo che una matrice $T = (t_{ij}) \in \mathbb{R}^{n\times n}$ è una matrice di Toeplitz se è costante lungo le diagonali, cioè esistono scalari $r_{-n+1}, \ldots, r_0, \ldots, r_{n-1}$ tali che $t_{ij} = r_{j-i}$ per ogni i, j:

$$T = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 & . & r_{n-2} & r_{n-1} \\ r_{-1} & r_0 & r_1 & r_2 & r_3 & . & r_{n-2} \\ r_{-2} & r_{-1} & r_0 & r_1 & r_2 & . & r_{n-3} \\ . & . & . & . & . & . & . \\ r_{-(n-3)} & r_{-(n-4)} & . & r_{-1} & r_0 & r_1 & r_2 \\ r_{-(n-2)} & r_{-(n-3)} & . & . & . & r_0 & r_1 \\ r_{-(n-1)} & r_{-(n-2)} & . & . & . & r_{-1} & r_0 \end{bmatrix}.$$

Notiamo che l'insieme τ puó essere rappresentato da

$$\mathcal{T} = \left\{ X \in \mathbb{R}^{n\times n} : X = \sum_{p=-(n-1)}^{(n-1)} \alpha_p G_p \right\}$$

dove $\alpha_p \in \mathbb{R}$ e le matrici $G_p \in \mathbb{R}^{n\times n}$ sono matrici definite come segue

$$(G_p)_{ij} = \begin{cases} 1 & \text{se } j = i + p \\ 0 & \text{altrimenti} \end{cases} \tag{6.3}$$

caratterizzate dal fatto che per ogni i, j esiste un unico indice p, $-(n-1) \leq p \leq (n-1)$, tale che $(G_p)_{ij} = 1$. Chiaramente queste matrici formano una base per il sottospazio τ.

Introduciamo un problema equivalente al problema (6.2). Si consideri la decomposizione singolare di A

$$A = P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T,$$

dove $P \in \mathbb{R}^{m\times m}$ e $Q \in \mathbb{R}^{n\times n}$, sono matrici ortogonali e $\sum = \text{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. Poiché la norma di Frobenius è invariante per trasformazioni ortogonali, si ottiene

$$||AX - B||_F^2 = \left|\left| P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right|\right|_F^2 = \left|\left| P^T \left(P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right) \right|\right|_F^2 =$$

$$= \left|\left| \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - P^T B \right|\right|_F^2 = \left|\left| \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - C \right|\right|_F^2 =$$

$= \left|\left|(\sum Q^T)X - C_1\right|\right|_F^2 + ||C_2||_F^2 = ||T - C_1||_F^2 + ||C_2||_F^2$
con $T = \sum Q^T X \in \mathbb{R}^{n\times n}$, $C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = P^T B, C_1 \in \mathbb{R}^{n\times n}$.
Allora il problema (6.2) è equivalente a

$$\min \quad ||T - C_1||_F^2 \quad \text{tale che} \quad T \in \mathcal{T}' \tag{6.4}$$

dove $\mathcal{T}'$ è il sottoinsieme :

$$\mathcal{T}' = \left\{ T \in \mathbb{R}^{n\times n} : T = \Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l \right\}.$$

Introduciamo la notazione $P_{\mathcal{U}}(A)$ che è la proiezione della matrice A sull'insieme $\mathcal{U}$. Il problema (6.4) è risolto da $P_{\mathcal{T}'}(C_1)$.

Teorema 6.1.1. *Se $C_1 \in \mathbb{R}^{n\times n}$ allora l'unica soluzione del problema (6.4) è:*

$$P_{\mathcal{T}'}(C_1) = \Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l^\star G_l, \quad \textit{dove} \quad \alpha_l^\star = \frac{\sum_{i=1}^{n}\sum_{j>l}^{n}(C_1)_{ij} Q_{(j-l)i}\sigma_i}{\sum_{i=1}^{n}\sum_{j>l}^{n}(Q_{(j-l)i})^2\sigma_i^2}$$

con l tale che $-(n-1) \le l \le (n-1)$.

Dimostrazione. Consideriamo la funzione $f : \mathbb{R}^{2n-1} \to \mathbb{R}$ data da

$$f(\alpha) \quad = \quad f(\alpha_{-(n-1)}, \ldots, \alpha_0, \ldots, \alpha_{(n-1)}) =$$

$$= \quad \frac{1}{2}||T - C_1||_F^2 = \quad \frac{1}{2}||\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l - C_1||_F^2 =$$

$$= \quad \frac{1}{2}\sum_{i,j=1}^{n} \left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l \right)_{ij} - (C_1)_{ij} \right)^2 .$$

Si vogliono determinare punti α per cui $f(\alpha)$ è minima. Poiché f è regolare, tali punti dovranno essere punti stazionari per f. Calcoliamo $\frac{\partial f}{\partial \alpha_p}(\alpha)$ per ogni p tale che $-(n-1) \le p \le (n-1)$:

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = \frac{\partial}{\partial \alpha_p}\left(\frac{1}{2}\sum_{i,j=1}^{n}\left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{n-1}\alpha_l G_l\right)_{ij} - (C_1)_{ij}\right)^2\right) =$$

$$= \sum_{ij=1}^{n}\left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij} - (C_1)_{ij}\right)\frac{\partial}{\partial \alpha_p}\left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij}\right). \quad (6.5)$$

Si può scrivere:

$$\frac{\partial}{\partial \alpha_p}\left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij}\right) = \frac{\partial}{\partial \alpha_p}\left(\sum_{k=1}^{n}\left(\Sigma Q^T\right)_{ik}\left(\sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{kj}\right) =$$

$$= \sum_{k=1}^{n}\left(\Sigma Q^T\right)_{ik}\frac{\partial}{\partial \alpha_p}\left(\left(\sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{kj}\right) =$$

$$= \sum_{k=1}^{n}\left(\Sigma Q^T\right)_{ik}(G_p)_{kj} = \left(\Sigma Q^T G_p\right)_{ij}$$

e la (6.5) diventa:

$$\frac{\partial f}{\partial \alpha_p} = \sum_{ij=1}^{n}\left(\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij} - (C_1)_{ij}\right)\left(\Sigma Q^T G_p\right)_{ij}.$$

Avendo in mente che

$$\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij}$$

è il prodotto interno tra la i-esima riga di ΣQ^T e la j-esima colonna di $\sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l$, otteniamo

$$\left(\Sigma Q^T \sum_{l=-(n-1)}^{(n-1)}\alpha_l G_l\right)_{ij} = \sigma_i \sum_{k=(j-n)}^{(j-1)} Q_{(j-k),i}\alpha_k.$$

Quindi

$$\frac{\partial f}{\partial \alpha_p} = \sum_{ij=1}^{n}\left(\sigma_i \sum_{k=(j-n)}^{(j-1)} Q_{(j-k),i}\alpha_k - (C_1)_{ij}\right)\left(\Sigma Q^T G_p\right)_{ij}. \quad (6.6)$$

L'elemento $\left(\Sigma Q^T G_p\right)_{ij}$ è il prodotto interno tra la i-esima riga di ΣQ^T e la j-esima colonna di G_p. Allora abbiamo

$$\left(\Sigma Q^T G_p\right)_{ij} = \sigma_i \begin{cases} Q_{(j-p)i} & se \ -(n-1) \le p \le 0,\ 1 \le j \le n+p,\ 0 \le p \le n-1,\ p+1 \le j \le n \\ 0 & se \ -(n-1) \le p \le 0,\ j > n+p,\ 0 \le p \le n-1,\ 1 \le j \le p \end{cases} .$$

Inoltre, a causa della struttura sparsa delle G_p,
se $p > 0$ allora $(G_p)_{ij} = 0$ se $1 \le j < p+1$, e
se $p \le 0$, allora $(G_p)_{ij} = 0$ se $n+p < j \le n$.
Quindi la (6.6) può essere scritta come per p>0

$$\frac{\partial f}{\partial \alpha_p} = \sum_{i=1}^{n} \sum_{j>p}^{n} \left(\sigma_i \sum_{k=(j-n)}^{(j-1)} Q_{(j-k)i} \alpha_k - (C_1)_{ij} \right) \left(\sigma_i Q_{(j-p)i}\right) =$$

$$= \sum_{j=p+1}^{n} \sum_{k=(j-n)}^{(j-1)} \sum_{i=1}^{n} Q_{(j-p)i} Q_{(j-k)i} \sigma_i^2 \alpha_k - \sum_{i=1}^{n} \sum_{j>p}^{n} (C_1)_{ij} \sigma_i Q_{j-p,i}.$$

Definendo

$$\sum_{i=1}^{n} \left(Q_{(j-k)i} Q_{(j-p)i}\right) \sigma_i^2 = \upsilon^T \theta$$

con

$$\upsilon^T = (Q_{(j-k)1} Q_{(j-p)1}, Q_{(j-k)2} Q_{(j-p)2}, \ldots, Q_{(j-k)n} Q_{(j-p)n})$$
$$\theta = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)^T,$$

abbiamo che la somma delle componenti di υ è il prodotto interno tra due colonne della matrice ortogonale Q, che è zero quando $k \neq p$. Quindi per $k \neq p$ abbiamo:

$$0 = \left(\sum_{i=1}^{n} \upsilon_i \right) \sigma_n^2 \le \upsilon^T \theta = \sum_{i=1}^{n} \upsilon_i \sigma_i^2 \le \left(\sum_{i=1}^{n} \upsilon_i \right) \sigma_1^2 = 0,^1$$

cioè $\upsilon^T \theta = 0$ per tutti $k \neq p$ e

$$\sum_{j>p}^{n} \sum_{k=(j-n)}^{j-1} \Big[\sum_{i=1}^{n} (Q_{(j-k)i} Q_{(j-p)i}) \sigma_i^2 \Big] \alpha_k = \sum_{j>p}^{n} \sum_{k=(j-n)}^{(j-1)} \alpha_k \upsilon^T \theta = \sum_{j>p}^{n} \alpha_p \sum_{i=1}^{n} (Q_{(j-p)i})^2 \sigma_i^2.$$

Allora, per $p > 0$:

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = \alpha_p \sum_{j>p}^{n} \sum_{i=1}^{n} (Q_{(j-p)i})^2 \sigma_i^2 - \sum_{j>p}^{n} \sum_{i=1}^{n} (C_1)_{ij} Q_{(j-p)i} \sigma_i$$

[1] in questi passaggi ci potrebbe essere un errore

per ogni $p = 1, \ldots, n-1$.

Invece per $p \leq 0$, la (6.6) può essere riscritta come

$$\frac{\partial f}{\partial \alpha_p} = \sum_{i=1}^{n} \sum_{j=1}^{n+p} \left(\sigma_i \sum_{k=j-n}^{j-1} Q_{(j-k)i} \alpha_k - (C_1)_{ij} \right) \sigma_i Q_{(j-p)i} =$$

$$= \sum_{j=1}^{n+p} \sum_{k=j-n}^{j-1} \alpha_k \sum_{i=1}^{n} Q_{(j-k)i} Q_{(j-p)i} \sigma_i^2 - \sum_{i=1}^{n} \sum_{j=1}^{n+p} (C_1)_{ij} \sigma_i Q_{(j-p)i},$$

ovvero

$$\frac{\partial f}{\partial \alpha_p} = \alpha_p \sum_{j=1}^{n+p} \sum_{i=1}^{n} \left(Q_{(j-p)i} \right)^2 \sigma_i^2 - \sum_{j=1}^{n+p} \sum_{i=1}^{n} (C_1)_{ij} Q_{(j-p)i} \sigma_i, \; p = -(n-1), \cdots, -1, 0.$$

Quindi dalla condizione necessaria

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = 0, \qquad p = -(n-1), \cdots, -1, 0, 1, \cdots, n-1.$$

segue, nel caso $\sigma_i > 0$ per ogni i, l'espressione delle componenti dell'unico punto stazionario per f:

$$per \; p = 1, \cdots, n-1 \quad \alpha_p^\star = \frac{\sum_{i=1}^{n} \sum_{j>p}^{n} (C_1)_{ij} Q_{(j-p)i} \sigma_i}{\sum_{i=1}^{n} \sum_{j>p}^{n} (Q_{(j-p)i})^2 \sigma_i^2}, \tag{6.7}$$

$$per \; p = -(n-1), \cdots, 0 \quad \alpha_p^\star = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n+p} (C_1)_{ij} Q_{(j-p)i} \sigma_i}{\sum_{i=1}^{n} \sum_{j=1}^{n+p} (Q_{(j-p)i})^2 \sigma_i^2}. \tag{6.8}$$

Osservazione

Sul caso $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = 0$, si osserva che se $A \neq 0$ il coefficiente di α_0 nell'espressione di $\frac{\partial f}{\partial \alpha_0}$ è sempre diverso da zero perché se fosse zero, si avrebbe

$$Q_{ri} \sigma_i = 0, \; \forall r, i \qquad \Sigma Q^T = 0 \qquad \text{e} \qquad P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T = A = 0.$$

Invece il coefficiente di α_p in $\frac{\partial f}{\partial \alpha_p}$, per $p \neq 0$, può essere nullo quando $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = 0$. $\square$

Ora dobbiamo verificare che $\alpha^\star = (\alpha_p^\star)$ è un minimo di f. Calcoliamo

$$\frac{\partial^2 f}{\partial \alpha_p^2}(\alpha) = \sum_{j>p}^{n} \sum_{i=1}^{n} (Q_{(j-p)i}\sigma_i)^2 > 0, \qquad p = 1, \cdots, n-1,$$

$$\frac{\partial^2 f}{\partial \alpha_p^2}(\alpha) = \sum_{j=1}^{n+p} \sum_{i=1}^{n} (Q_{(j-p)i}\sigma_i)^2 > 0, \qquad p = -(n-1), \cdots, 0,$$

e

$$\frac{\partial^2 f}{\partial \alpha_p \partial \alpha_q}(\alpha) = 0 \qquad \text{se } p \neq q.$$

Questo implica che la matrice Hessiana $\nabla^2 f(\alpha)$ è definita positiva e costante al variare di α; quindi f è strettamente convessa e $\alpha^\star$ è un minimo assoluto globale per f.

□

6.2 Caso $\mathbb{P} = \mathcal{T}_u$ ={matrici triangolari di Toeplitz superiori}

Si considera il problema di minimo

$$\left\{\min||AX - B||_F^2 \quad \text{dove} \quad X \in \mathcal{T}_u\right\} \tag{6.9}$$

dove $A, B \in \mathbb{R}^{m\times n}$ e $\mathcal{T}_u$ è l'insieme delle matrici triangolari superiori di Toeplitz.

Se $X \in \mathcal{T}_u$, allora

$$X = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \dots & \dots & \dots & \alpha_n \\ 0 & \alpha_1 & \alpha_2 & \dots & \dots & \dots & \alpha_{n-1} \\ 0 & 0 & \alpha_1 & \alpha_2 & \dots & \dots & \alpha_{n-2} \\ \vdots & \dots & \dots & \dots & & & \vdots \\ \vdots & & & \dots & \dots & \dots & \vdots \\ 0 & & & & \dots & \alpha_1 & \alpha_2 \\ 0 & 0 & \dots & \dots & \dots & \dots & \alpha_1 \end{bmatrix}.$$

Il sottoinsieme $\mathcal{T}_u$ può essere scritto come:

$$\mathcal{T}_u = \left\{ X \in \mathbb{R}^{n\times n} : X = \sum_{p=1}^{n} \alpha_p G_p \right\}$$

dove $\alpha_p \in \mathbb{R}$ e le matrici $G_p \in \mathbb{R}^{n\times n}$ sono definite per ogni $1 \le i, j, p \le n$ come segue

$$(G_p)_{ij} = \begin{cases} 1 & \text{se } j = i + p - 1 \\ 0 & \text{altrimenti.} \end{cases} \tag{6.10}$$

Le matrici G_p sono caratterizzate dalla proprietà che per ogni elemento i, j con $i \le j$, esiste un unico p tale che $(G_p)_{ij} = 1$. L'insieme $\{G_1, G_2, \dots, G_n\}$ é una base per $\mathcal{T}_u$. Analogamente al caso generale di Toeplitz, viene utilizzata la decomposizione dei valori singolari di A per ricondurci a risolvere il problema equivalente

$$\min \quad ||T - C_1||_F^2 \quad \text{dove} \quad T \in \mathcal{T}_u'. \tag{6.11}$$

con

$$A = P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T$$

$P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{n\times n}$, matrici ortogonali, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ con $\sigma_1 \ge \dots \ge \sigma_n \ge 0$, e

$$\mathcal{T}_u' = \left\{ T \in \mathbb{R}^{n\times n} \quad : \quad T = \Sigma Q^T \sum_{l=1}^{n} \alpha_l G_l \right\}.$$

Quest'ultimo problema è risolto nel seguente

Teorema 6.2.1. *Sia $C_1 \in \mathbb{R}^{n\times n}$, allora l'unica soluzione del problema*

$$min||T - C_1||_F^2 \quad dove \quad T \in \mathcal{T}_u' \tag{6.12}$$

è data da

$$P_{\mathcal{T}_u'}(C_1) = \Sigma Q^T \sum_{l=1}^{n} \alpha_l^\star G_l, \quad dove \quad \alpha_l^\star = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j\geq l}^{n}(C_1)_{ij}Q_{(j-l+1)i}\sigma_i}{\displaystyle\sum_{i=1}^{n}\sum_{j\geq l}^{n}(Q_{(j-l+1)i})^2\sigma_i^2}$$

per ogni $1 \leq l \leq n$.

Dimostrazione. Procedendo come prima, si vede che

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = \sum_{i,j=1}^{n}\left(\left(\Sigma Q^T \sum_{l=1}^{n}\alpha_l G_l\right)_{ij} - (C_1)_{ij}\right)\left(\Sigma Q^T G_p\right)_{ij}.$$

L'elemento $\left(\Sigma Q^T \sum_{l=1}^{n}(\alpha_l G_l)\right)_{ij}$ è il prodotto interno tra i-esima riga di (ΣQ^T) e la j-esima colonna di $\sum_{l=1}^{n}(\alpha_l G_l)$. Allora risulta che

$$\left(\Sigma Q^T \sum_{l=1}^{n}(\alpha_l G_l)\right)_{ij} = \sigma_i \sum_{k=1}^{j} Q_{(j-k+1)i}\alpha_k.$$

Analogamente, l'elemento $\left(\Sigma Q^T G_p\right)_{ij}$ è il prodotto interno tra la i-esima riga di ΣQ^T e la j-esima colonna di G_p. Allora

$$(\Sigma Q^T G_p)_{ij} = \begin{cases} \sigma_i Q_{(j-p+1)i} & \text{se } j \geq p \\ 0 & \text{se } j < p \end{cases}.$$

Quindi:

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = \sum_{i=1}^{n}\sum_{j=p}^{n}\left(\sigma_i \sum_{k=1}^{j} Q_{(j-k+1)i}\alpha_k - (C_1)_{ij}\right)\sigma_i Q_{(j-p+1)i} =$$

$$= \sum_{j=p}^{n}\sum_{k=1}^{j}\alpha_k\left(\sum_{i=1}^{n} Q_{(j-k+1)i}Q_{(j-p+1)i}\sigma_i^2\right) - \sum_{j=p}^{n}\sum_{i=1}^{n}(C_1)_{ij}Q_{(j-p+1)i}\sigma_i =$$

$$= \alpha_p \sum_{j=p}^{n}\sum_{i=1}^{n}\left(Q_{(j-p+1)i}\right)^2\sigma_i^2 - \sum_{j=p}^{n}\sum_{i=1}^{n}(C_1)_{ij}Q_{(j-p+1)i}\sigma_i.$$

□

6.3 Caso $\mathbb{P} = \mathcal{T}_l$ ={matrici triangolari di Toeplitz inferiori}

Consideriamo il problema

$$\min||AX - B||_F^2 \quad \text{dove} \quad X \in \mathcal{T}_l \tag{6.13}$$

dove $A, B \in \mathbb{R}^{m\times n}$ e $\mathcal{T}_l$ è l'insieme delle matrici di Toeplitz triangolari inferiori.

Se $X \in \mathcal{T}_l$ allora abbiamo

$$X = \begin{bmatrix} \beta_1 & 0 & 0 & \dots & \dots & \dots & 0 \\ \beta_2 & \beta_1 & 0 & \dots & \dots & \dots & 0 \\ \beta_3 & \beta_2 & \beta_1 & 0 & \dots & \dots & 0 \\ \vdots & \dots & \dots & \dots & & & \vdots \\ \vdots & & & \dots & \dots & \dots & \vdots \\ \beta_{n-1} & & & & \dots & \beta_1 & 0 \\ \beta_n & \beta_{n-1} & \dots & \dots & \beta_3 & \beta_2 & \beta_1 \end{bmatrix}$$

Il sottoinsieme $\mathcal{T}_l$ puó essere espresso come

$$\mathcal{T}_l = \left\{ X \in \mathbb{R}^{n\times n} : X = \sum_{p=1}^{n} \beta_p G_p \right\}$$

dove $\beta_p \in \mathbb{R}$ e le matrici $G_p \in \mathbb{R}^{n\times n}$ sono definite come

$$(G_p)_{ij} = \begin{cases} 1 & \text{se } j = i - p + 1 \\ 0 & \text{altrimenti} \end{cases} \tag{6.14}$$

per ogni $1 \le i, j, p \le n$. L'insieme $\{G_1, G_2, \dots G_n\}$ è una base per il sottoinsieme $\mathcal{T}_l$. Di nuovo, applicando la decomposizione dei valori singolari di A, il problema può essere trasformato in

$$\min \quad ||T - C_1||_F^2 \quad \text{tale che} \quad T \in \mathcal{T}_l' \tag{6.15}$$

con $A = P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T$, dove $P \in \mathbb{R}^{m\times m}$ e $Q \in \mathbb{R}^{n\times n}$, sono matrici ortogonali, $\Sigma = \text{diag}(\sigma_1, \dots \sigma_n), \sigma_1 \ge \dots \ge \sigma_n \ge 0$, e

$$\mathcal{T}_l' = \left\{ T \in \mathbb{R}^{n\times n} : T = \Sigma Q^T \sum_{p=1}^{n} \beta_p G_p \right\}.$$

Analogamente al caso generale di Toeplitz e al caso delle matrici di Toeplitz triangolari superiori, si ottiene il seguente

Teorema 6.3.1. *Se $C_1 \in \mathbb{R}^{n\times n}$ allora l'unica soluzione del problema*

$$min||T - C - 1||_F^2 \quad tale\ che \quad T \in \mathcal{T}_l' \tag{6.16}$$

è data da

$$P_{\mathcal{T}_l'}(C_1) = \Sigma Q^T \sum_{l=1}^{n} \beta_l^\star G_l, \quad dove \quad \beta_l^\star = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n-l+1}(C_1)_{ij}Q_{(j+l-1)i}\sigma_i}{\sum_{i=1}^{n}\sum_{j=1}^{n-l+1}(Q_{(j+l-1)i})^2\sigma_i^2}$$

per ogni $1 \le l \le n$.

6.4 Il problema simmetrico di Toeplitz

In questa sezione ci occupiamo del problema

$$\min||AX - B||_{\mathbb{F}}^2,\ X \in \mathcal{T} \cap S \tag{6.17}$$

dove $A, B \in \mathbb{R}^{m\times n}$ e $\mathcal{T} \cap S$ è il sottoinsieme di $\mathbb{R}^{n\times n}$ delle matrici simmetriche di Toeplitz. Chiaramente, la soluzione è una matrice simmetrica non solo rispetto alla diagonale principale, ma anche rispetto all'altra diagonale. Poiché la regione di ammissibilità è l'intersezione di due sottospazi e siccome sappiamo risolvere il problema di Procruste persimmetrico[5 dell'articolo] ed il problema generale di Toeplitz (vedi la sezione precedente), possiamo sviluppare un algoritmo basato sul metodo di proiezione alternata [Dykstra], [Boyle, Dykstra], [Escalante, Raydan, 1996], [Escalante, Raydan, 1998]. Quindi la soluzione del problema (6.17) è ottenuta proiettando alternamente sui sottospazi $\mathcal{T}$ e S.

Il problema

$$\min||AX - B||_F^2 \quad \text{dove} \quad X \in S, \tag{6.18}$$

in accordo con Escalante e Raydan[Escalante, Raydan, 1998], può essere trasformato, tramite la decomposizione dei valori singolari di A, nel seguente problema equivalente

$$\min||Z - C_1||_F^2 \text{ al variare di } Z \text{ in } S' = \{Z \in \mathbb{R}^{n\times n} : Z = \Sigma Y,\ Y = Y^T\} \tag{6.19}$$

dove $Z = \Sigma Y$, $Y = Q^T X Q$, $C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = P^T B Q$, $C_1 \in \mathbb{R}^{n\times n}$.

Quindi, se la soluzione dell'ultimo problema è:

$$P_{S'}(C_1) = \Sigma Y_\star = \Sigma(Q^T X_\star Q),$$

dove $Y_\star$ si ottiene con i metodi proposti di Higham per il caso simmetrico, [Higham, 1988], la soluzione del primo problema è:

$$X_\star = Q\Sigma^{-1}P_{S'}(C_1)Q^T.$$

D'altra parte, sappiamo come risolvere il problema di Toeplitz generale, vedi la sezione precedente. Combinando questi risultati, risolvere il problema (6.17) è equivalente a risolvere

$$\min \quad ||Z - C_1||_{\mathbb{F}}^2 \quad \text{dove} \quad Z \in \mathcal{T}'' \cap S' \tag{6.20}$$

essendo

$$\mathcal{T}'' = \left\{ Z \in \mathbb{R}^{n\times n} : Z = \Sigma Q^T \sum_{l=-(n-1)}^{n-1} \alpha_l G_l Q \right\}.$$

La definizione dell'insieme $\mathcal{T}''$ deriva dal fatto che il problema simmetrico e quello di Toeplitz hanno la stessa funzione obiettivo.

L'algoritmo della proiezione alternata per il problema simmetrico di Toeplitz è il seguente:

Algoritmo: Date $A, B \in \mathbb{R}^{m\times n}$, $X_0 = C_1 \in \mathbb{R}^{n\times n}$ dove C_1 si ottiene tramite la decomposizione dei valori singolari di A ($C_1 = \{P^T B Q\}_{n\times n}$), calcola per $i =$ 0, 1, 2, ...

$$X_i = P_{S'}(X_i),$$
$$X_{i+1} = P''_{\mathcal{T}}(X_i)$$

finchè si ha la convergenza. L'algoritmo termina quando due proiezioni consecutive sullo stesso sottospazio sono abbastanza vicine. La convergenza è garantita dalla teoria generale stabilita da Neumann [Neumann] per l'algoritmo della proiezione alternata.

Bibliografia

[1] *G. Ammar, P. Gader, A variant of the Gohberg-Semencul formula involving circulant matrices, SIAM J. Matrix Anal. Appl., 12:534-540, 1991*

[2] *G. S. Ammar, W. B. Gragg, Superfast solution of real positive definite Toeplitz systems, SIAM J. Matrix Anal. Appl., 9:61-76, 1988*

[3] *D. Bertaccini, C. Di Fiore, P. Zellini, Complessita' e iterazione, percorsi, matrici e algoritmi veloci nel calcolo numerico, Bollati Boringhieri, 2013*

[4] *R. Bevilacqua, M. Capovani, Sulle proprieta' di una classe di matrici, I.E.I. (CNR) Nota Interna B72-14, Pisa, Settembre 1972*

[5] *D. Bini, M. Capovani, O. Menchi, Metodi Numerici per l'Algebra Lineare, Zanichelli, 1988*

[6] *D. Bini, F. Di Benedetto, A new preconditioner for the parallel solution of positive denite Toeplitz linear systems, nei: Proc. 2nd SPAA conf., Crete (Greece), luglio 1990 pp. 220-223*

[7] *D. Bini, V. Pan, Numerical and Algebraic Computations with Matrices and Polynomials. Birkauser, Boston, 1994*

[8] *A. Bortoletti, C. Di Fiore, On a set of matrix algebras related to discrete Hartley-type trasnsforms, Linear algebra Appl., 366:65-85, 2003*

[9] *A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini, A new class of quasi-Newtonian methods for optimal learning in MLP-networks, IEEE Trans. Neural Networks, 14:263-273, 2003*

[10] *J.P. Boyle and R.L. Dykstra, A method for finding projections onto the intersections of convex sets in Hilbert spaces, Lecture Notes in Statistics, 37 (1986), 28–47.*

[11] *E. Bozzo, P. Deidda, C. Di Fiore, The Jordan and Frobenius pairs of the inverse, Linear and Multilinear Algebra, 71:1730-1735, 2023*

[12] *E. Bozzo, C. Di Fiore, On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decom position. SIAM Journal on Matrix Analysis and Applications 16(1): 312-326, 1995*

[13] *J. Bunch, Stability of methods for solving Toeplitz systems of equations, SIAM J. Sci. Stat. Comp., 6, pp. 349-364 (1985)*

[14] *A. Burg, S. Haene, W. Fichtner and M. Rupp, "Regularized Frequency Domain Equalization Algorithm and its VLSI Implementation," 2007 IEEE International Symposium on Circuits and Systems (ISCAS), New Orleans, LA, USA, 2007, pp. 3530-3533, doi: 10.1109/ISCAS.2007.378444.*

[15] *J. F. Cai, R. H. Chan, C. Di Fiore, Minimization of a detail-preserving regularization functional for impulse noise removal, J. Math. Imaging Vision, 29:79-91, 2007*

[16] *R. Chan, Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions, IMA J. Numer. Anal., 11, pp. 333-345 (1991)*

[17] *R. Chan, J. Nagy, R. Plemmons, FFT-based preconditioning for Toeplitz block least squares problems, SIAM J. Numer. Anal., 30, pp. 1740-1768 (1993)*

[18] *R.H. Chan and M.K. Ng, Conjugate gradient methods for Toeplitz systems, SIAM Review, 38(3) (1996), 427–482.*

[19] *R. Chan, G. Strang, Toeplitz equations by conjugate gradient with circulant preconditioner, SIAM J. Sci. Stat. Comp., 10, pp. 104-119 (1989)*

[20] *R. Chan, P. Tang, Constrained minimax approximation and optimal preconditioners for Toeplitz matrices, Num. Alg., 5, pp. 353-364 (1993)*

[21] *T. F. Chan, An optimal circulant preconditioner for Toeplitz systems, SIAM J. Sci. Statist. Comput., 9:766-771, 1988*

[22] *S. Cipolla, C. Di Fiore, F. Tudisco, P. Zellini, Adaptive matrix algebras in unconstrained minimization, Linear Algebra Appl., 471:544-568, 2015*

[23] *S. Cipolla, C. Di Fiore, P. Zellini, A variation of Bryden class methods using Householder adaptive transforms, Comput. Opt. Appl., 77:433-463, 2020*

[24] *P. J. Davis, Circulant matrices, Wiley, New York, 1979*

[25] *F. Di Benedetto, S. Serra Capizzano, A unifying approach to abstract matrix algebra preconditioning, Numer. Math., 82:57-90, 1999*

[26] *C. Di Fiore, Matrix algebras and displacement decompositions, SIAM J. Matrix Anal. Appl., 21:646-667, 2000*

[27] *C. Di Fiore, slides of the talk held in the School-Conference on Tensor mathods in Mathematics and Data Science, MSU-BIT, Shenzhen, China, Nov.11-20, 2024*

[28] *C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in quasi-Newton methods for unconstrained minimization, Numer. Math., 94:479-500, 2003*

[29] *C. Di Fiore, F. Lepore, P. Zellini, Hartley-type algebras in displacement and optimization strategies, Linear Algebra Appl., 366:215-232, 2003*

[30] *C. Di Fiore, F. Tudisco, P. Zellini, Lower triangular Toeplitz-Ramanujan systems whose solution yields the Bernoulli numbers, Linear Algebra Appl., 496:510-526, 2016*

[31] *C. Di Fiore, P. Zellini, Matrix decompositions using displacement rank and classes of commutative matrix algebras, Linear Algebra Appl., 229:49-99, 1995*

[32] *C. Di Fiore, P. Zellini, Matrix displacement decompositions and applications to Toeplitz linear systems, Linear Algebra Appl., 268:197-225, 1998*

[33] *C. Di Fiore, P. Zellini, Matrix algebras in optimal preconditioning, Linear Algebra Appl., 335:1-54, 2001*

[34] *C. Di Fiore, P. Zellini, A matrix formulation of lower triangular Toeplitz systems solvers, manuscript*

[35] *R.L. Dykstra, An algorithm for restricted least squares regression, Journal of the Amer. Stat. Ass., 78(384) (1983), 837–842.*

[36] *M.G. Eberle, Métodos de proyecciones alternas en problemas de optimización persimétricos, Master's thesis, Dept. de Matemática, Universidad Nacional del Sur, Bahía Blanca, Argentina, (1999).*

[37] *M.G. Eberle and M.C. Maciel, The persymmetric Procrusto problem, Bahía Blanca,Argentina, (2001). In preparation.*

[38] *M. G. Eberle, M.C. Maciel, Finding the closest Toeplitz, Computational and Applied Mathematics, Vol. 22, N.1. pp.1-18, 2003*

[39] *R. Escalante and M. Raydan, Dykstra's algorithm for constrained least-squares matrix problem, Num. Lin. Alg, with Appl., 36 (1996), 459–471.*

[40] *R. Escalante and M. Raydan, Dykstra's algorithm for constrained least-squares rectangular matrix problem, Computers and Mathematics with Applications, 35(6) (1998), 73–79.*

[41] *G. Fiorentino, S. Serra, Multigrid methods for Toeplitz matrices, Calcolo, 28, pp. 283-305 (1992)*

[42] *P. D. Gader, Displacement operator based decompositions of matrices using circulants or other group matrices, Linear Algebra Appl., 139:111-131, 1990*

[43] *L. Gemignani, Fast inversion of Hankel and Toeplitz matrices, In-form. Proc. Lett., 41, pp. 119-123 (1992)*

[44] *I. Gohberg, V. Olshevsky, Circulant, displacements and decompositions of matrices, Integral Equations Operator Theory, 15:730-743, 1992 W. Gross, Analisi Numerica 2, appunti, A.A.1988-89*

[45] *U. Grenander, G. Szego, Toeplitz Forms and Their Applications. Second edition, Chelsea, New York, 1984*

[46] *W. Gross, Analisi Numerica 2, appunti, A.A.1988-89*

[47] *N.J. Higham, Newton's methods for the matrix square root, Mathematics of Computation, 46(174) (1986), 537–549.*

[48] *N.J. Higham, Computing a nearest symmetric positive semidefinite matrix, Linear Algebra and its Applications, 103 (1988), 103–118.*

[49] *N.J. Higham, The symmetric Procrustes problem, BIT, 28 (1988), 133–143.*

[50] *H. Hu and I. Olkin, A numerical procedure for finding the positive definite matrix closest to a patterned matrix, Statistics and Probability Letters, 12 (1991), 511–515.*

[51] *T. Huckle, Circulant and skewcirculant matrices for solving Toeplitz matrix problems, SIAM J. Matrix Anal. Appl., 13:767-777, 1992*

[52] *I.S. Iokhvidov, Hankel and Toeplitz forms: algebraic theory. Birkhauser, Boston, 1982*

[53] *T. Ku, C. Kuo, On the spectrum of a family of preconditioned Toeplitz matrices, SIAM J. Sci. Stat. Comp., 13, pp. 946-966 (1992)*

[54] *J.Von Neumann, Functional operator, vol. ii: The geometry of orthogonal spaces, In Annals of Math. Studies. Princeton University Press, Princeton, (1950).*

[55] *P. Lancaster, M. Tismenetsky, The Theory of Matrices, Academic Press, New York, 1985*

[56] *C. Pagano, comunicazione personale*

[57] *S.V. Parter, On the distribution of singular values of Toeplitz matrices, Lin. Alg. Appl., 80, pp. 115-130 (1986)*

[58] *J. Rissanen, Algorithm for triangular decomposition of block Hankel and Toeplitz matrices with applications to factoring positive matrix polynomials, Math. Comp., 27, pp. 147-154 (1973)*

[59] *S. Serra, Sulle proprieta' spettrali di matrici precondizionate di Toeplitz, Bollettino dell'Unione Matematica Italiana 11(2): 463-483,(1997)*

[60] *P.H. Schonemann, A generalized solution of the orthogonal Procrustes problem, Psychometrica, 31 (1966), 1–10.*

[61] *R. Slowik, Inverse and determinants of Toeplitz-Hessenberg matrices, Taiwanese J. of Math., 22:901-908, 2018*

[62] *G. Strang, A proposal for Toeplitz matrix calculation, Stud. Appl. Math., 74, pp. 171-176 (1986)*

[63] *F. Trench, An algorithm for the inversion of finite Toeplitz matrices, SIAM J. Appl. Math., 12, pp. 515-522 (1964)*

[64] *F. Tudisco, C. Di Fiore, E. Tyrtyshnikov, Optimal rank matrix algebras preconditioners, Linear Algebra Appl., 438:405-427, 2013*

[65] *E. E. Tyrtyshnikov, A Brief Introduction to Numerical Analysis, Birkhauser, Boston-Berlin, 1997*

[66] *H. Widom, "On the singular values of Toeplitz matrices", Zeit. Anal. Anw., 8, pp. 221-229 (1989).*

[67] *https://www.mat.uniroma2.it/dottorato/Theses/2017/Cipolla.pdf*

[68] *https://www.mat.uniroma2.it/ difiore/RomeMoscow2014.pdf*

[69] *Impara LATEX...e mettilo da parte.*