Corso della scuola di dottorato:

NUMERICAL LINEAR ALGEBRA: TOOLS AND METHODS

Metodi diretti e regolarizzazione

Dottorandi:

Mario Cascetta

Efisio Casti

INTRODUZIONE

L'utilizzo di un calcolatore per la risoluzione di un problema matematico comporta intrinsecamente che i dati siano sempre affetti da errori. La sola introduzione di un dato comporta, delle volte, delle approssimazioni perché la rappresentazione deve avvenire con un numero limitato di cifre decimali. Non solo, anche i calcoli elementari avvengono con un'aritmetica discreta e quindi, a loro volta, i risultati saranno affetti da errori e così via fino al risultato finale.

Se si considera il problema della risoluzione di un sistema lineare del tipo:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots\ldots\ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots\ldots\ldots + a_{2n}x_n = b_2 \\ \ldots\ldots \\ \ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots\ldots\ldots a_{mn}x_n = b_m \end{cases} \quad (1)$$

Si può rappresentare nella forma più compatta

$$\mathbf{Ax=b} \quad (2)$$

dove A è la matrice dei coefficienti, x il vettore soluzione e b il vettore dei termini noti. Se la matrice A è mal condizionata allora anche piccoli errori di rappresentazione dei dati potranno portare ad una soluzione che contiene dei grossi errori e che quindi si discosterà molto da quella che è la soluzione esatta. Questa "sensibilità" del problema agli errori sui dati può essere misurata attraverso un parametro K(A), detto fattore di condizionamento.

$$K(A) = \frac{\|\mathbf{A}\|}{\|\mathbf{A}^{-1}\|} \quad (3)$$

Si ha che per K elevati possono corrispondere grossi errori sulla soluzione nonostante si abbiano piccoli errori sui dati.

Scopo di questo lavoro è valutare come sia possibile utilizzare i metodi di risoluzione dei sistemi lineari, che già funzionano bene per i sistemi dove la matrice è ben condizionata, nei casi in cui il sistema abbia un matrice molto mal condizionata. In particolare si faranno delle sperimentazioni numeriche utilizzando dei metodi diretti utilizzando la matrice di Hilbert, che possiede la proprietà di essere molto mal condizionata.

Se oltre alla matrice di prova fisso anche la soluzione si può calcolare a ritroso il termine noto:

$$\mathbf{b = Ae}$$

in questo modo un buon parametro di confronto può essere la norma dell'errore tra la soluzione trovata con i diversi metodi di risoluzione e la soluzione esatta fissata in principio.

$$||\mathbf{x}-\mathbf{e}|| \tag{4}$$

I metodi utilizzati saranno quelli:

- Di Gauss con la fattorizzazione A=LU
- QR
- SVD e TSVD

In principio, si farà vedere come i diversi metodi si comportano al variare della dimensione della matrice. Infatti, all'aumentare della dimensione della matrice, K aumenta facendo aumentare di conseguenza gli errori. In seguito, si confronteranno due strumenti per la scelta di una giusta regolarizzazione TSVD:

- GCV
- QUASI OPTIMAL

METODI DIRETTI PER LA RISOLUZIONE DEI SISTEMI LINEARI

Per la risoluzione dei sistemi lineari sono disponibili numerosi metodi diretti che, pur essendo diversi fra loro, condividono la stessa idea di base, ossia trasformare, mediante un algoritmo con un numero finito di operazioni, un sistema lineare generico in un sistema equivalente che ne renda più semplice la risoluzione. Esistono, come noto, particolari sistemi di equazioni lineari che necessitano di un carico computazionale notevolmente inferiore rispetto al caso generale. Sono i sistemi diagonali, ortogonali e triangolari.

Metodo di Gauss (fattorizzazione A=LU)

Il metodo si basa sull'assunzione che se operiamo delle operazioni elementari sui termini della matrice possiamo trasformare il problema in uno equivalente. Funziona solo per matrici con gli elementi diagonali non nulli, quindi anche per matrici simmetriche definite positive. L'algoritmo di Gauss opera la fattorizzazione:

$$\mathbf{A}=\mathbf{LU} \tag{5}$$

cioè trasforma la matrice d'origine nel prodotto di due matrici triangolari.

Quindi il nostro sistema di partenza si può assumere equivalente al:

$$\mathbf{LUx}=\mathbf{b} \tag{6}$$

Se operiamo la sostituzione:

$$\mathbf{Ux} = \mathbf{y} \tag{7}$$

si possono risolvere facilmente i due sistemi triangolari:

$$\begin{cases} \mathbf{Ly} = \mathbf{b} \\ \mathbf{Ux} = \mathbf{y} \end{cases} \tag{8}$$

Se qualche elemento della diagonale della matrice A è molto piccolo, si possono generare degli overflow e il condizionamento del sistema finale può aumentare in maniera significativa. Quest'ultimo problema può essere ovviato tramite il pivoting parziale, ma che non utilizzeremo per il nostro caso. In Matlab per effettuare la fattorizzazione LU si usa il comando:

```
[ L U ] = lu(A)
```

Fattorizzazione QR

Attraverso una serie di operazioni si può trasformare il sistema come:

$$\mathbf{A} = \mathbf{QR} \tag{9}$$

Dove Q è una matrice ortogonale ed R una matrice triangolare superiore. Possiamo, dopo avere effettuato le sostituzioni:

$$\mathbf{QRx} = \mathbf{b} \text{ e } \mathbf{Rx} = \mathbf{y} \tag{10}$$

risolvere il sistema finale

$$\begin{cases} \mathbf{Qy} = \mathbf{b} \\ \mathbf{Rx} = \mathbf{y} \end{cases} \tag{11}$$

Il principale vantaggio è quel di poter utilizzare la matrice trasposta, infatti per l'ortogonalità di Q questa coincide con quella inversa. La matrice Q ha anche la proprietà di non peggiorare il condizionamento, che si mantiene costante e pari a uno.

In matlab il comando da usare per effettuare la fattorizzazione è:

```
[ Q R ] = qr(A)
```

SVD (singular value decomposition)

Il teorema fondamentale di esistenza della decomposizione ai valori singolari di una qualsiasi matrice rettangolare con coefficienti appartenenti al campo complesso ci assicura la possibilità di scomporre la nostra matrice quadrata definita nel campo reale. Infatti, in generale, se $\mathbf{A} \in \mathbb{C}^{m \times n}$ allora esistono una matrice unitaria $\mathbf{U} \in \mathbb{C}^{m \times m}$ e una matrice unitaria $\mathbf{V} \in \mathbb{C}^{n \times n}$ tali che:

$$\mathbf{A} = \mathbf{USV}^T \tag{12}$$

Dove la matrice $\mathbf{S} \in \mathbb{R}^{m \times n}$ ha elementi σ_{ij} nulli per $i \neq j$ e per $i = j$ ha elementi $\sigma_{ii} = \sigma_i$ con

$\sigma_1 \geq \sigma_2 \geq .....\geq \sigma_p \geq 0$, con p che assume il valore minimo tra m ed n (p = min(m,n)), detti valori singolari di A. Inoltre, indicate con $\mathbf{u}_i$, i=1,…,m e $\mathbf{v}_i$, i=1,…,n le colonne di U e V vengono detti vettori rispettivamente singolari sinistri e vettori singolari destri di A. Quindi, dal punto di vista geometrico, la SVD di A fornisce due basi di vettori ortogonali, le colonne di U e V, tali che la matrice diventa diagonale quando trasformata rispetto a queste basi.

Mentre, se vogliamo applicare questa scomposizione al caso della risoluzione del sistema lineare (1) si ha che la soluzione può essere espressa come:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (\mathbf{USV}^T)^{-1}\mathbf{b} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i \tag{13}$$

Da quest'ultima relazione si può facilmente prevedere come piccoli cambiamenti in A o in b possano indurre grandi cambiamenti sulla soluzione effettivamente calcolata nel caso di valori singolari piccoli. Si può anche dimostrare che il calcolo dei valori singolari di una matrice risulta sempre ben condizionato, ossia piccole perturbazioni degli elementi della matrice di partenza inducono nei risultati perturbazioni non superiori ad esse.

Usando le proprietà della SVD si può avere una misura del condizionamento:

$$K(A) = \frac{\sigma_1}{\sigma_n} \tag{14}$$

In matlab esiste una funzione che implementa un algoritmo usato per la scomposizione ai valori singolari.

```
[U S V] = svd(A),
```

Una volta ottenute le matrici U, S, V, si risolve con semplicità il seguente sistema:

$$\begin{cases} \mathbf{Uz} = \mathbf{b} \\ \mathbf{Sy} = \mathbf{z} \\ \mathbf{V}^T\mathbf{x} = \mathbf{y} \end{cases} \tag{15}$$

RISULTATI DEL TEST

La sperimentazione numerica si è svolta prendendo in esame una matrice di Hilbert, fissando il vettore soluzione ***e*** e ricavando, di conseguenza, il termine noto **b**; successivamente, si sono trovate le soluzioni adoperando i tre metodi considerati e si è confrontata la norma-2 della differenza con la

soluzione esatta nota, facendo variare la dimensione della matrice del sistema da un minimo di 2 ad un massimo di 30.

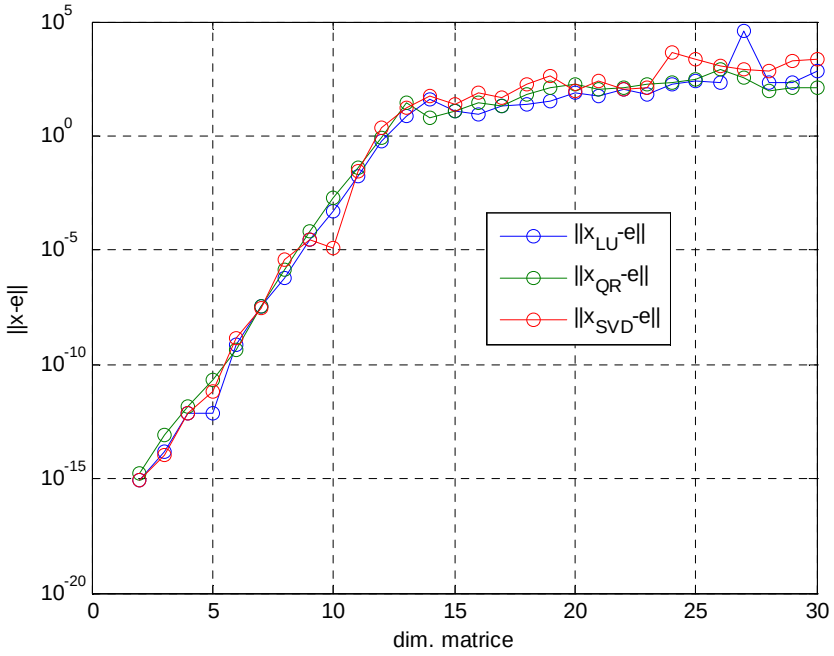

Fig. 1 - Grafico dell'andamento del valore della norma-2 dell'errore al variare della dimensione della matrice

Il comportamento dei tre metodi risulta molto simile e sembra seguire il valore del condizionamento della matrice di partenza. Nel metodo di Gauss ciò può essere spiegato con il fatto che le matrici triangolari hanno, per definizione, problemi di condizionamento, mentre nel caso della fattorizzazione QR, pur avendo la metrice Q un condizionamento uguale ad 1, prevale quello della matrice R, che, in quanto triangolare, presenta analoghi problemi descritti per il metodo di Gauss.
Anche la risoluzione con la SVD segue l'andamento del condizionamento, ma si può aggiungere un'ulteriore considerazione per spiegare meglio quale possa essere il suo utilizzo nel caso delle matrici malcondizionate. Analizzando le relazioni seguenti derivate direttamente dalle definizioni della SVD, si nota che se si considerano vettori singolari destri associati a valori singolari molto piccoli, questi ultimi sono caratterizzati da combinazioni lineari aventi i componenti di maggior peso appartenenti al nucleo di A. Ciò implica che la scomposizione della matrice A potrebbe essere fatta considerandola quasi a rango non pieno.

$$\mathbf{A} = \sum_{i=1}^{n} \mathbf{u}_i \sigma_i \mathbf{v}_i^T$$

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i \qquad (16)$$

$$\left\| \mathbf{A}\mathbf{v}_i \right\| = \sigma_i , \qquad \text{per i=1,…,n}$$

LA REGOLARIZZAZIONE TSVD PER MATRICI MAL CONDIZIONATE

Per porre rimedio ai problemi generati dall'uso della SVD si può pensare semplicemente di trascurare i valori singolari caratterizzati da valori molto piccoli. La soluzione del sistema lineare diventa allora:

$$\mathbf{x}_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i \quad \text{, dove } k \leq n \tag{17}$$

Si ottiene dunque una soluzione approssimata troncando l'espansione della SVD al k-esimo termine; il metodo prende il nome di TSVD e può essere utilizzata attraverso la funzione presente nel pacchetto Regularization tools per Matlab:

```
[x_tsvd,rho,eta] = tsvd(U,s,V,b,k)
```

Si devono fornire U, s e V ottenuti dalla funzione per il calcolo della svd, mentre b è noto e k è un parametro che va scelto seguendo determinate metodologie. La funzione restituisce la soluzione (x_tsvd), le norme dei residui (rho) e della soluzione (eta).

Per illustrare meglio quanto succede, ci si serve di due matrici di Hilbert rispettivamente con n=5 e n=15 e per ciascuna di esse si calcola la soluzione e si osserva il comportamento della norma al variare del parametro k fino al valore k=n, che coincide con la SVD a rango completo.

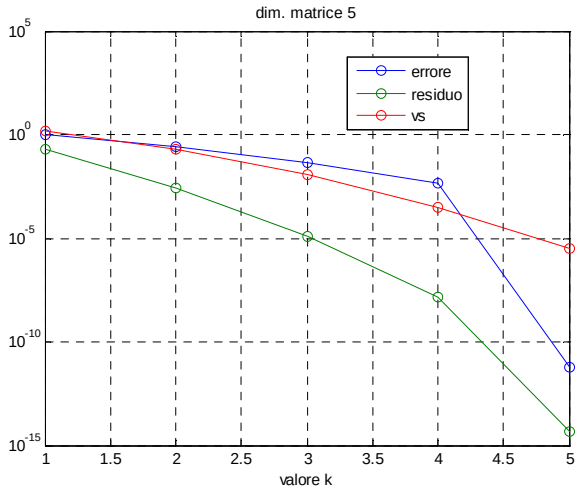

Fig. 2 - Andamento del valore della norma-2 per l'errore della soluzione e per il residuo al variare di k per una matrice di n=5.

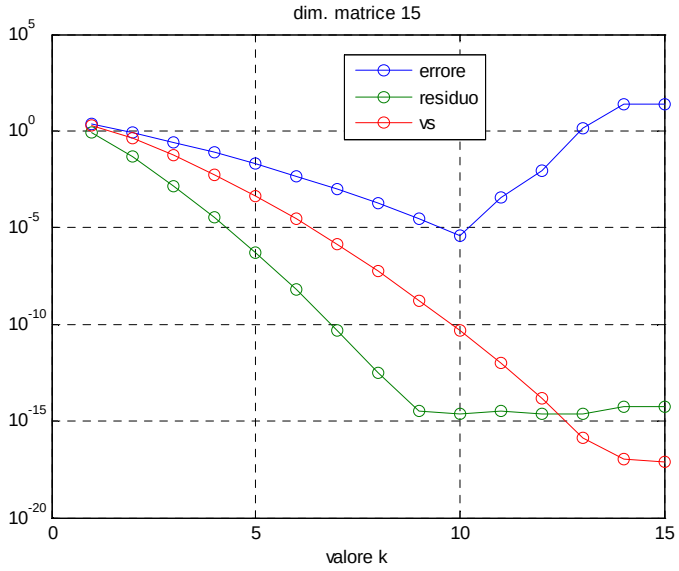

Fig. 3 - Andamento del valore della norma dell'errore, del residuo e del valore singolare k-esimo, al variare di k, per una matrice di dimensione 15

Nel caso della matrice con n=5, il valore migliore di k corrisponde al rango di A che è pari a 5, dal momento che non si ha ancora un valore del condizionamento tale per cui si manifestino degli errori elevati.

IL grafico relativo alla matrice con n=15 porta alla scelta di k=10, che corrisponde ad un valore minimo sulla norma sull'errore. Inizialmente, all'aumentare di k, la soluzione migliora, in quanto meglio approssimata per via dell'adozione di un numero sempre maggiore di valori singolari per descriverla. Successivamente, l'aggiunta di ulteriori valori singolari porta, viceversa, l'errore ad assumere valori inaccettabili, poiché questi valori singolari sono molto piccoli.

Metodi per la scelta del parametro di regolarizzazione

Un criterio di regolarizzazione è completo solo dopo aver definito un metodo per la scelta del parametro di regolarizzazione; questi metodi si possono dividere in due classi:

- i metodi che si basano sulla conoscenza, o una buona stima, della norma dell'errore $\|x-e\|_2$;
- i metodi che non richiedono la norma dell'errore $\|x-e\|_2$.

Nel primo caso, normalmente, essendo presenti informazioni affidabili sulla $\|x-e\|_2$, si adopera il principio di discrepanza che non sarà, però, oggetto di analisi.

Quando, invece, non si conosce alcuna informazione sulla norma dell'errore si ricorre ai quei metodi liberi dal vincolo della conoscenza di $\|x-e\|_2$, chiamati qualche volta metodi euristici.

Un'ulteriore classificazione si può fare a seconda si usi un parametro di regolarizzazione continuo

come, ad esempio il λ utilizzato nel metodo di Tikhonov, oppure il parametro discreto k usato in precedenza per la tsvd.

Il criterio GCV

Nel pacchetto Regularization Tools sono presenti alcune routines per la stima del paramentro k da fornire alla funzione tsvd in modo da ottenere una soluzione approssimata il più possibile vicina a quella esatta. Uno di questi è la GCV (Generalized Cross Validation) che si basa sulla ricerca del minimo della seguente funzione:

$$G = \frac{\left\|\mathbf{Ax}-\mathbf{b}\right\|^2}{\left(trace(\mathbf{I}_n - \mathbf{AA}^I)\right)^2} \tag{18}$$

Dove A^I è la matrice che produce la soluzione regolarizzata. Per una matrice quadrata si può dimostrare che l'equazione precedente, riferita alla tsvd, diventa:

$$G(k) = \frac{\left\|\mathbf{b}-\mathbf{Ax}_k\right\|^2}{\left(n-k\right)^2} = \frac{1}{\left(n-k\right)^2}\sum_{i=k+1}^{n}\left(\mathbf{u}_i^T\mathbf{b}\right)^2 \tag{19}$$

Questo metodo ha il grosso vantaggio di non richiedere alcuna informazione sulla norma dell'errore. Infatti, si cerca di minimizzare lo scarto quadratico medio dei residui, dividendolo con la funzione a denominatore. In matlab si richiama la funzione GCV come segue:

```
k_gcv = gcv(U,s,b,'tsvd')
```

Fornendo come parametri la matrice U, i vettori s, precedentemente ricavati dalla SVD e il vettore del termine noto b, la funzione restituisce il parametro da utilizzare sulla tsvd ed il grafico della G(k) che mette in evidenza il minimo.

Il criterio quasi optimal

Al fine di confrontare il metodo della GCV, si è preso come riferimento il tool per la scelta del parametro di regolarizzazione QUASI OPTIMAL. Questo metodo è stato definito principalmente per la regolarizzazione del parametro continuo λ, ma può essere usato anche nel caso discreto facendo delle opportune assunzioni. Si tratta di trovare il minimo della funzione Q definita come:

$$Q \equiv \left\|\frac{d\mathbf{x}_\lambda}{d\lambda}\right\| \tag{20}$$

Come si può dimostrare, questo approccio consente di trovare un buon compromesso tra l'effetto delle perturbazioni e la regolarizzazione degli errori della soluzione. Per una regolarizzazione con il parametro discreto k, si pone $\lambda = \gamma_k$ e si può considerare l'approssimazione:

$$\left\| \frac{d\mathbf{x}_\lambda}{d\lambda} \right\| \approx \frac{\|\Delta \mathbf{x}_k\|}{|\Delta\lambda|}, \tag{21}$$

Con:

$|\Delta\lambda| = \gamma_{k+1} - \gamma_k \approx \gamma_k$ si ha:

$$\|\Delta \mathbf{x}_k\| = \frac{\mathbf{u}_k^T \mathbf{b}}{\gamma_k}$$

E sostituendo si ottiene la funzione nella forma generale:

$$Q(k) \approx \frac{\mathbf{u}_k^T \mathbf{b}}{\gamma_k}$$

Mentre, per il caso della TSVD, si può porre:

$$Q(k) \approx \frac{\mathbf{u}_k^T \mathbf{b}}{\sigma_k} \ . \tag{22}$$

La funzione presente nel pacchetto in matlab, analogamente al caso precedente, restituisce il valore del parametro k da utilizzare per la regolarizzazione e il grafico della funzione Q.

```
k_qopt = quasiopt(U,s,b,'tsvd');
```

RISULTATI DEL TEST DEL METODO TSVD

Con le stesse modalità dei test precedenti si sono regolarizzati i problemi costruiti con la matrice di Hilbert per dimensioni che vanno da 2 a 30.

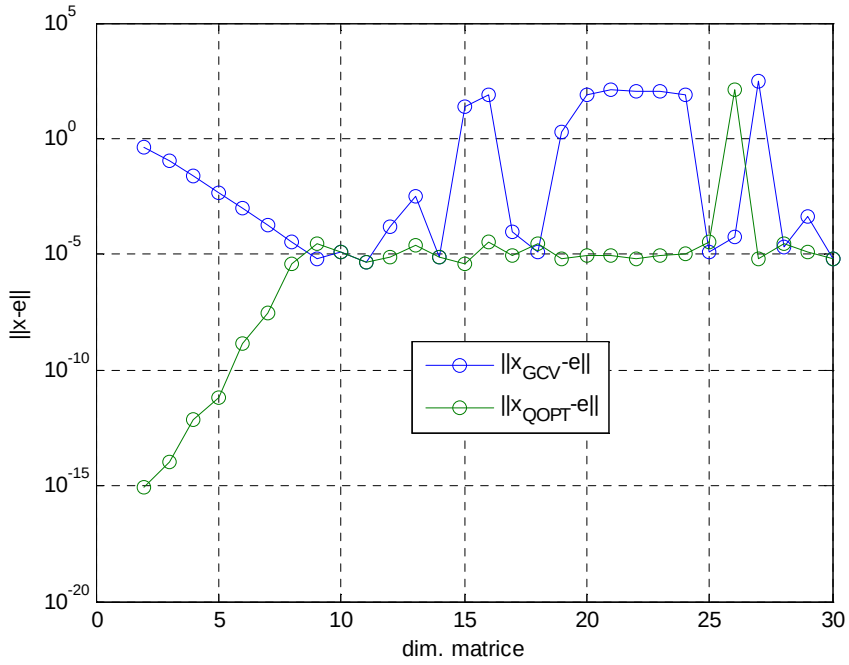

Fig. 4 – Confronto del valore della norma dell'errore per i due metodi della scelta del parametro di regolazione al variare della dimensione della matrice.

Come si può notare, la GVC porta a degli errori minori dell'unità per dimensioni della matrice minori di 12, mentre, per valori maggiori, l'errore assume un comportamento irregolare. Nel metodo quasi optimal, invece, l'errore è quasi sempre minore dell'unità. Prendiamo in considerazione ora il caso della matrice A con n=15, dove si analizza il comportamento della funzione G(k) relativa alla GCV e della funzione Q(k) del metodo quasi optimal.

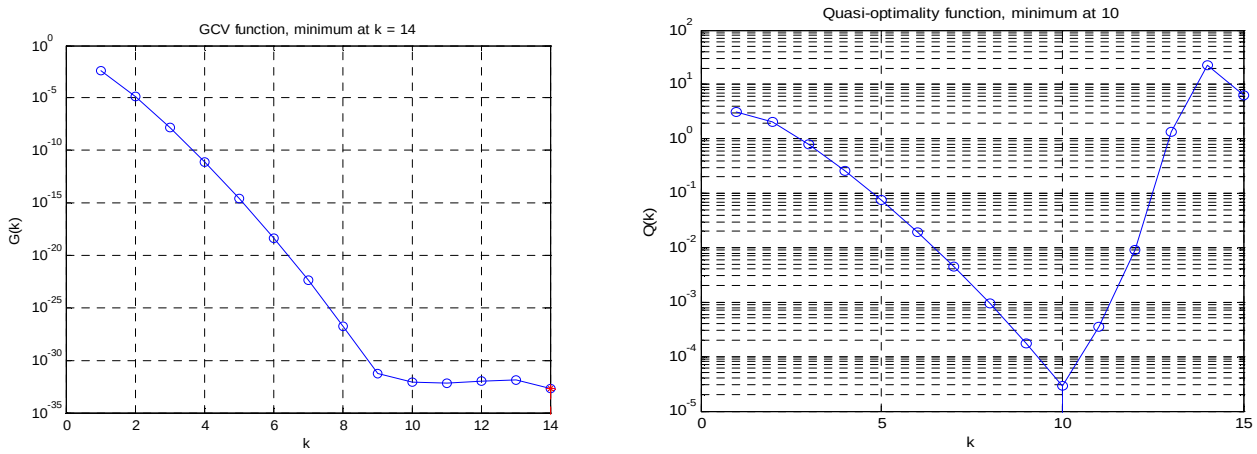

Fig. 5 – Grafici relativi alle funzioni G(k) e Q(k) per una matrice n=15.

Nel grafico relativo alla funzione G(k) il punto di minimo si trova in corrispondenza di k=14; si osserva che la funzione decresca fino a d un valore di k pari a 9 per poi proseguire in maniera quasi piatta fino al minimo descritto poc'anzi. Il minimo, quindi, viene a trovarsi in corrispondenza di valori quasi uguali a quello del rango della svd completa. La funzione GCV porta alla scelta di un k errato, in quanto si prenderebbero in considerazione dei valori singolari troppo piccoli che generano un mal-condizionamento del sistema con errori sulla soluzione molto grandi, senza, perciò, alcun vantaggio. La funzione Q(k), che dipende anche dalla stima dell'errore, riesce a fornire un valore di k che genera degli errori minori.

Per quel che riguarda il metodo quasi-optimal si è osservato un comportamento anomalo relativo al caso della matrice con n=26; infatti, osservando il grafico, il minimo viene raggiunto apparentemente in corrispondenza del punto k = 11, mentre la scelta fatta dall'algoritmo ricade sul punto k=23 con un valore di Q(k) pari a 0.

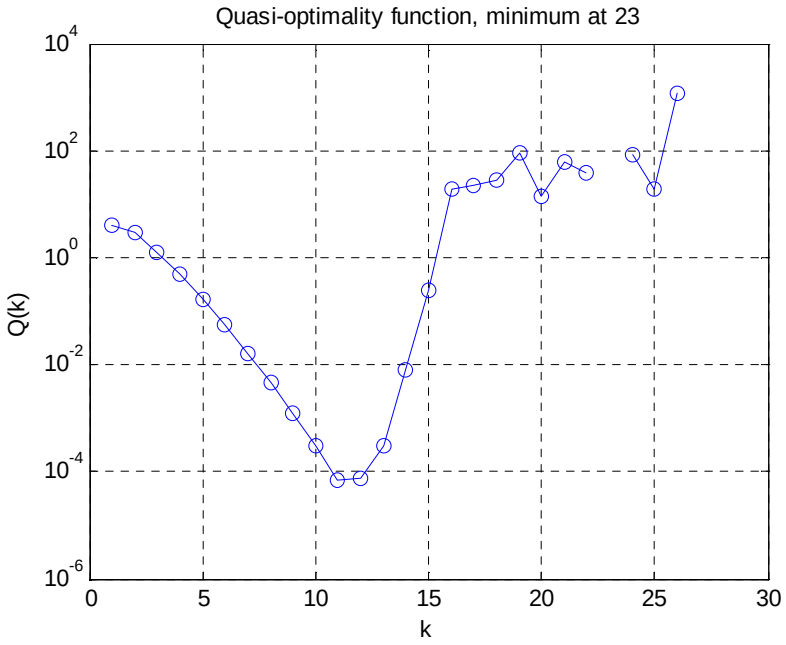

Fig. 5 – Grafici relativi alla funzione Q(k) per una matrice n=26.

Da un punto di vista numerico, la causa va ricercata nel numeratore della funzione di Q(k) che assume valore nullo.