UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Facoltà di Scienze

Corso di Laurea Magistrale in Matematica

Regolarizzazione di problemi mal posti con applicazione alla tomografia computerizzata

Relatore
Prof. Giuseppe Rodriguez

Tesi di Laurea di
Federica Pes

Anno Accademico 2017-2018

Ai miei genitori e a mia sorella

Ringraziamenti

Prima di tutto vorrei ringraziare il prof. Giuseppe Rodriguez per la disponibilità e la presenza costante durante tutto il periodo di preparazione e stesura di questa tesi.
Ringrazio la mia famiglia, che ha condiviso con me i momenti di gioia e mi ha spronata nei momenti di difficoltà. Senza il loro sostegno morale ed economico non avrei mai raggiunto questo traguardo.
Ringrazio le colleghe e i colleghi incontrati in questi anni, in particolare Silvia, Chiara, Maria Lucia, Alessandra, Jessica, Melania, Paola.

Introduzione

I problemi inversi derivano dalla necessità di interpretare misure indirette e incomplete. Come area della matematica, il campo dei problemi inversi è fortemente legato ad applicazioni in diverse aree tra cui l'ingegneria, la geofisica, la medicina, la biologia e la fisica. Quest'area della matematica è cresciuta costantemente negli ultimi decenni. Tale crescita è stata favorita sia dai progressi nel calcolo sia dalle scoperte teoriche.

I problemi inversi sono mal posti, quindi necessitano di un metodo per ricavare una soluzione che approssimi la soluzione vera. Questi metodi sono detti metodi di regolarizzazione.

Questa tesi è un'introduzione ai problemi inversi e ai metodi di regolarizzazione [2]. La tesi è strutturata nel seguente modo.

Il capitolo 1 è un capitolo introduttivo. Si fa una distinzione tra problemi a rango non pieno e problemi a rango mal determinato, ponendo l'accento sul fatto che sia necessario conoscere il concetto di decomposizione ai valori singolari, argomento affrontato nello stesso capitolo seguito dalla decomposizione ai valori singolari generalizzata. Viene considerato il classico esempio di problema inverso mal posto, l'equazione integrale di Fredholm del primo tipo, osservando il motivo della mal posizione. Vengono descritti brevemente i principali approcci alla regolarizzazione, osservando che esistono problemi di regolarizzazione in forma standard e in forma generale e approfondendo nell'ultima parte del capitolo il passaggio dalla forma generale alla forma standard.

Il capitolo 2 riguarda i metodi diretti di regolarizzazione differenziati tra problemi a rango non pieno e problemi a rango mal determinato. Nei primi è utile il concetto di rango numerico e tra i metodi troviamo la SVD troncata e la GSVD troncata, mentre nei secondi non è possibile determinare un rango numerico, quindi sono necessarie delle alternative, tra cui i fattori filtro, e tra i metodi troviamo la regolarizzazione di Tikhonov.

Il capitolo 3 riguarda i metodi iterativi di regolarizzazione, tra cui le iterazioni di Landweber, che è un metodo iterativo classico, e le iterazioni basate sul metodo del gradiente coniugato.

Nel capitolo 4 sono spiegati alcuni criteri di scelta per il parametro di regolarizzazione: il principio di discrepanza, la generalized cross-validation e la curva L. Questi e altri metodi si trovano in [7]. Viene inoltre introdotto il metodo COSE [4, 6].

Il capitolo 5 è un'introduzione alla tomografia computerizzata a raggi X [5]. È spiegata l'attenuazione dei raggi X che attraversano un corpo. Un semplice esempio di problema inverso nel caso di fascio di raggi paralleli mostra la mal posizione. In questo capitolo non può mancare la trasformata di Radon, il cui nome è dovuto a Johann Radon: è una trasformata integrale la cui inversa, detta antitrasformata di Radon, è utilizzata per ricostruire immagini bidimensionali a partire dai dati raccolti nel processo di diagnostica medica detto tomografia assiale computerizzata (TAC).

Il capitolo 6 è dedicato ai test numerici. Si confrontano i metodi TG-SVD e di Tikhonov su problemi test unidimensionali e bidimensionali per determinare una soluzione regolarizzata [3].

Indice

Capitolo 1

Introduzione alla regolarizzazione

Sia $Ax = b$ un sistema lineare di equazioni.

Definizione 1.1. Si definisce **numero di condizionamento** di una matrice, relativamente alla risoluzione di un sistema lineare, la quantità

$$\kappa(A) = \|A\|\|A^{-1}\|. \tag{1.1}$$

Indicato con δb il vettore perturbazione del termine noto e con δx la corrispondente perturbazione della soluzione, il numero di condizionamento misura il massimo fattore di amplificazione dell'errore relativo sulla soluzione rispetto all'errore relativo sui dati [8]

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{1.2}$$

Un numero di condizionamento molto grande della matrice dei coefficienti A in un sistema lineare di equazioni $Ax = b$ implica che alcune righe o colonne della matrice A sono linearmente dipendenti. A volte il numero di condizionamento grande è causato da un modello matematico errato che dovrebbe essere modificato prima di un tentativo di calcolare una soluzione. Gli strumenti numerici, come la decomposizione a valori singolari (SVD), possono identificare le dipendenze lineari e quindi aiutare a migliorare il modello e portare ad un sistema modificato con una matrice meglio-condizionata. Questo sistema modificato può quindi essere risolto con tecniche numeriche standard.

1.1 Problemi con matrici mal condizionate

Dato un problema con una matrice dei coefficienti mal condizionata, bisogna essere in grado di selezionare gli algoritmi migliori per quel determinato problema. Nessun metodo di regolarizzazione è superiore agli altri metodi. Piuttosto, ogni metodo ha i suoi vantaggi, a seconda dell'applicazione in cui viene utilizzato.

In presenza di matrici mal condizionate è importante conoscere la SVD della matrice A. In particolare, il numero di condizionamento di A risulta essere uguale al rapporto tra il più grande e il più piccolo valore singolare di A. Il trattamento numerico di sistemi di equazioni con una matrice dei coefficienti mal condizionata dipende dal tipo di mal condizionamento di A. Ci sono due importanti classi di problemi da considerare e molti problemi pratici appartengono a una di queste due classi.

I **problemi a rango non pieno** sono caratterizzati dalla matrice A che ha un ammasso di piccoli valori singolari e vi è un gap ben definito tra valori singolari grandi e piccoli. Ciò implica che una o più righe e colonne di A sono quasi combinazione lineare di alcune o tutte le righe e le colonne rimanenti. Pertanto, la matrice A contiene informazioni ridondanti e la chiave per il trattamento numerico di tali problemi è estrarre la parte linearmente indipendente in A, per arrivare a un altro problema con una matrice ben condizionata.

I **problemi discreti mal-posti** derivano dalla discretizzazione di problemi mal posti come le equazioni integrali di Fredholm del primo tipo. Qui tutti i valori singolari di A, così come le componenti SVD della soluzione, decadono gradualmente a zero, e diciamo che una condizione discreta di Picard è soddisfatta; vedi §2.2.4. Poiché non vi è un gap nell'insieme dei valori singolari, non vi è alcuna nozione di rango numerico per queste matrici.

1.2 Problemi inversi e mal posti

Il concetto di problemi ben posti e mal posti risale ad Hadamard all'inizio del secolo scorso. Hadamard definisce un *problema* come *ben posto* se:

1. Esiste una soluzione del problema (esistenza).
2. La soluzione è unica (unicità).
3. La soluzione dipende con continuità dai dati (stabilità).

Viceversa, un *problema* si dice *mal posto* quando non verifica una di queste condizioni. È mal posto un problema che non ha soluzioni, oppure che ne ha

più di una, o che ha soluzione instabile, cioè se una piccola perturbazione dei dati può causare una grande perturbazione della soluzione. Hadamard riteneva che i problemi mal posti fossero "artificiali" in quanto non descrivevano sistemi fisici.

Si sbagliava, infatti oggi sorgono problemi mal posti sotto forma di **problemi inversi** in molte aree della scienza e dell'ingegneria. Ad esempio, se si è interessati a determinare la struttura interna di un sistema fisico a partire dal comportamento misurato all'esterno del sistema, o nel determinare l'input sconosciuto che dà origine a un segnale di output misurato (in contrasto con problemi diretti dove l'interesse è nel comportamento del sistema dato l'input o la struttura interna). Alcuni esempi sono l'acustica, la tomografia computerizzata, la geofisica, la ricostruzione di immagini, il telerilevamento, l'elaborazione dei segnali e la statistica.

I problemi inversi lineari possono essere formulati nella seguente forma molto generale:

$$\int_\Omega \text{input} \times \text{system}\, d\Omega = \text{output}\,.$$

In questa formulazione, il problema diretto è calcolare l'output, dato l'input e la descrizione matematica del sistema. L'obiettivo del problema inverso è determinare l'input o il sistema che dà origine alle misurazioni (rumorose) dell'output. Un esempio è la tomografia computerizzata, in cui l' "input" è una sorgente di raggi X, il "sistema" è l'oggetto che viene scansionato (spesso il cervello) e l' "output" è lo smorzamento misurato dei raggi X . L'obiettivo è ricostruire il "sistema", cioè l'oggetto scansionato, dalle informazioni sulla posizione delle sorgenti di raggi X e le misure del loro smorzamento.

Il classico esempio di problema lineare mal posto è l'equazione integrale di Fredholm del primo tipo con nucleo K integrabile al quadrato

$$\int_0^1 K(s,t)f(t)dt = g(s), \qquad 0 \le s \le 1, \tag{1.3}$$

dove g e K sono funzioni note, mentre f è incognita (la soluzione da cercare). In molte applicazioni pratiche di (1.3) il nucleo K è dato, mentre g tipicamente consiste di quantità misurate, cioè g è conosciuta solo con una certa accuratezza e solo in un insieme finito di punti $s_1, ..., s_m$.

Un caso particolare della (1.3) è l'equazione integrale di Fredholm del primo tipo con termine noto discreto

$$\int_0^1 k_i(t)f(t)dt = b_i, \qquad i = 1, ..., m. \tag{1.4}$$

Qui, sono dati m funzionali (o nuclei) k_i su una funzione sconosciuta f, e la (1.4) può essere ottenuta dalla (1.3) con $k_i(t) = K(s_i, t)$ e $b_i = g(s_i)$. Il

problema (1.4) è quindi continuo in una sola variabile t. Entrambe le forme (1.3) e (1.4) danno luogo a sistemi mal condizionati di equazioni algebriche lineari.

1.2.1 L'espansione ai valori singolari

Lo strumento per analizzare le equazioni (1.3) integrali di Fredholm del primo tipo con nucleo integrabile al quadrato è la *singular value expansion* (SVE) del nucleo. Un nucleo K è integrabile al quadrato se la norma

$$\|K\|^2 = \int_0^1 \int_0^1 K(s,t)^2 \, ds \, dt \tag{1.5}$$

è limitata. Per mezzo della SVE, ogni nucleo K integrabile al quadrato può essere scritto come somma infinita

$$K(s,t) = \sum_{i=1}^{\infty} \mu_i u_i(s) v_i(t). \tag{1.6}$$

Le funzioni u_i e v_i sono dette *funzioni singolari* di K. Sono ortonormali, cioè

$$\langle u_i, u_j \rangle = \langle v_i, v_j \rangle = \begin{cases} 1 & \text{se } i = j, \\ 0 & \text{se } i \neq j, \end{cases} \tag{1.7}$$

dove il prodotto scalare $\langle \cdot, \cdot \rangle$ è definito da

$$\langle \phi, \psi \rangle = \int_0^1 \phi(t)\psi(t) dt. \tag{1.8}$$

I numeri μ_i sono i *valori singolari* di K; sono non negativi e hanno ordine non crescente $\mu_1 \geq \mu_2 \geq \mu_3 \geq ... \geq 0$. Soddisfano la relazione $\sum_{i=1}^{\infty} \mu_i^2 = \|K\|^2$.

Una relazione importante tra valori singolari e funzioni singolari è la seguente:

$$\int_0^1 K(s,t) v_i(t) dt = \mu_i u_i(s), \qquad i = 1, 2, ..., \tag{1.9}$$

che mostra come ogni funzione singolare v_i viene mappata nella corrispondente u_i con amplificazione data dal valore singolare μ_i. Dal momento che le $u_i(s)$ e $v_i(t)$ sono ortonormali, possiamo scrivere

$$f(t) = \sum_{i=1}^{\infty} \langle v_i, f \rangle v_i(t) \, ,$$

$$g(s) = \sum_{i=1}^{\infty} \langle u_i, g \rangle u_i(s) \, .$$

Partendo da (1.3) e sfruttando le relazioni appena menzionate si ottiene

$$g(s) = \int_0^1 K(s,t)\left(\sum_{i=1}^{\infty}\langle v_i, f\rangle v_i(t)\right)dt = \sum_{i=1}^{\infty}\int_0^1 \langle v_i, f\rangle K(s,t)v_i(t)dt$$
$$= \sum_{i=1}^{\infty}\mu_i\langle v_i, f\rangle u_i(s)\,. \tag{1.10}$$

Quindi

$$\sum_{i=1}^{\infty}\mu_i\langle v_i, f\rangle u_i(s) = \sum_{i=1}^{\infty}\langle u_i, g\rangle u_i(s), \tag{1.11}$$

che, confrontando i termini delle sommatorie, porta alla soluzione della (1.3):

$$f(t) = \sum_{i=1}^{\infty}\frac{\langle u_i, g\rangle}{\mu_i}v_i(t). \tag{1.12}$$

Sottolineiamo che f esiste solo se il termine destro di (1.12) converge, che equivale a richiedere che g appartenga a $\mathcal{R}(K)$, il range di K.

Il comportamento dei μ_i e delle u_i e v_i non è arbitrario; è legato alle proprietà del nucleo K.

- Più il nucleo K è regolare, più velocemente i μ_i decadono a zero ("smoothness" è misurato dal numero di derivate parziali continue di K).
- Più è piccolo μ_i, più oscillazioni ci saranno nelle u_i e v_i. Questa proprietà è forse impossibile da dimostrare in generale, ma è spesso osservata nella pratica.

1.2.2 La condizione di Picard e l'instabilità della soluzione

La condizione di Picard. Affinché esista una soluzione f integrabile al quadrato dell'equazione (1.3), g deve soddisfare

$$\sum_{i=1}^{\infty}\left(\frac{\langle u_i, g\rangle}{\mu_i}\right)^2 < \infty. \tag{1.13}$$

La condizione di Picard dice che da un certo punto nella sommatoria (1.12), il valore assoluto dei $\langle u_i, g\rangle$ deve decadere più velocemente dei corrispondenti μ_i affinché una soluzione integrabile al quadrato esista.

La condizione (1.13) è identica alla richiesta che g appartenga ad $\mathcal{R}(K)$. Se g ha una componente arbitrariamente piccola al di fuori di $\mathcal{R}(K)$, allora non esiste alcuna soluzione integrabile al quadrato.

Sia $g \notin \mathcal{R}(K)$, sia g_k l'approssimazione di g ottenuta troncando la SVE dopo k termini

$$g_k(s) = \sum_{i=1}^{k} \langle u_i, g \rangle u_i(s) \, .$$

g_k soddisfa la condizione di Picard (1.13) per $k = 1, 2, ...$, e la corrispondente soluzione approssimata f_k è data da

$$f_k(t) = \sum_{i=1}^{k} \frac{\langle u_i, g \rangle}{\mu_i} v_i(t) \, .$$

Come $k \to \infty$, si ha che $g_k \to g$, ma $\|f_k\|_2 = \sqrt{\langle f_k, f_k \rangle} \to \infty$. È questa mancanza di stabilità di f che rende l'equazione integrale di Fredholm mal posta.

Purtroppo in situazioni pratiche abbiamo solo un'approssimazione di g che è contaminata da errori inevitabili

$$g = g^{\text{exact}} + \eta \, , \quad g^{\text{exact}} \in \mathcal{R}(K) \, , \quad \|\eta\|_2 \lesssim \|g^{\text{exact}}\|_2 \, .$$

Idealmente vogliamo calcolare $f^{\text{exact}} = \sum_{i=1}^{\infty} \mu_i^{-1} \langle u_i, g^{\text{exact}} \rangle v_i$. Non possiamo aspettarci che gli errori η soddisfino la condizione di Picard e quindi $g \notin \mathcal{R}(K)$. Qualsiasi approccio naive per calcolare f^{exact} tramite la somma infinita di solito diverge o restituisce un risultato inutile con norma estremamente ampia, non importa quanto piccola sia la perturbazione η. Invece è necessario un metodo di regolarizzazione che sostituisca il problema originale (1.3) con un problema regolarizzato che abbia una soluzione stabile che approssimi f^{exact}. Se η è molto grande, rispetto a g^{exact}, è impossibile calcolare un'approssimazione di f^{exact}; quindi si suppone $\|\eta\|_2 \lesssim \|g^{\text{exact}}\|_2$.

1.2.3 Introduzione alla regolarizzazione

Come abbiamo visto nella sezione precedente, le difficoltà nei problemi mal posti sono legate alla presenza dei piccoli valori singolari di K. Quindi, è necessario incorporare ulteriori informazioni sulla soluzione desiderata per stabilizzare il problema e individuare una soluzione utile e stabile. Questo è lo scopo della regolarizzazione.

L'approccio dominante alla regolarizzazione è di considerare un determinato residuo associato alla soluzione regolarizzata, con norma residua

$$\rho(f) = \left\| \int_0^1 K(s,t) f(t) dt - g(s) \right\|_2$$

e usare uno dei seguenti schemi:

1. Minimizzare $\rho(f)$ con vincolo che $f \in \mathcal{S}_f$ specifico sottoinsieme.
2. Minimizzare $\rho(f)$ con vincolo $\omega(f) \leq \delta$, dove $\omega(f)$ è la "dimensione" di f, δ è un limite superiore specificato.
3. Minimizzare $\omega(f)$ con vincolo $\rho(f) \leq \alpha$.
4. Minimizzare una combinazione lineare di $\rho(f)^2$ e $\omega(f)^2$,
$$\min\{\rho(f)^2 + \lambda^2\omega(f)^2\}$$
dove λ è un specificato fattore di peso.

Qui α, δ e λ sono noti come *parametri di regolarizzazione* e la funzione ω è definita come "smoothing norm". L'idea di base dei quattro schemi è che una soluzione regolarizzata che abbia una norma residua adeguatamente piccola e che soddisfi il vincolo si spera che sia non troppo lontana dalla desiderata e sconosciuta soluzione del problema non perturbato.

Dobbiamo discretizzare il problema di regolarizzazione per risolverlo numericamente. Se la discretizzazione porta a un sistema quadrato

$$Ax = b\,, \qquad A \in \mathbb{R}^{n\times n}\,,$$

o ad un sistema sovradeterminato

$$\min \|Ax - b\|_2\,, \qquad A \in \mathbb{R}^{m\times n}\,, \qquad m > n\,,$$

dove il vettore x rappresenta la funzione f, allora avremo bisogno di un qualche sottoinsieme $\mathcal{S}_x$ o qualche misura di "dimensione" $\Omega(x) \approx \omega(f)$ (in analogia con i 4 schemi). Per esempio, la regolarizzazione di Tikhonov discreta porta al problema di minimizzazione

$$\min\{\|Ax - b\|_2^2 + \lambda^2\Omega(x)^2\}\,.$$

La funzione Ω è definita la *discrete smoothing norm*, ed è spesso, ma non sempre, della forma

$$\Omega(x) = \|Lx\|_2\,, \tag{1.14}$$

dove la matrice L, detta *matrice di regolarizzazione*, è tipicamente la matrice identità, o una matrice diagonale, o un'approssimazione discreta $p \times n$ di un operatore derivata (es. $\Omega(x) \approx \omega(f) = \|f''\|_2$), nel cui caso L è una matrice a banda a rango pieno. Per esempio le matrici

$$L = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n} \tag{1.15}$$

e

$$L = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n} \tag{1.16}$$

sono approssimazioni degli operatori derivata prima e derivata seconda.

Quando $p < n$ allora $\|L \cdot\|_2$ è detta *seminorma*, cioè esiste un vettore $x \neq 0$ (nel nucleo di L) tale che $\|Lx\|_2 = 0$.

Se è disponibile a priori una stima x^* della soluzione desiderata regolarizzata, allora questa informazione può essere presa in considerazione includendo x^* nella norma smoothing discreta $\Omega(x)$, che assume la forma

$$\Omega(x) = \|L(x - x^*)\|_2 .$$

1.3 SVD e sua generalizzazione

Gli strumenti numerici per l'analisi dei problemi a rango non pieno e dei problemi discreti mal posti sono la *Singular Value Decomposition* (SVD) di A e la sua generalizzazione a due matrici, la *Generalized Singular Value Decomposition* (GSVD) della coppia di matrici (A, L).

1.3.1 La SVD

Sia $A \in \mathbb{R}^{m\times n}$ una matrice rettangolare o quadrata, supponiamo $m \geq n$. La SVD di A è una decomposizione della forma

$$A = U\Sigma V^T = \sum_{i=1}^{n} u_i \sigma_i v_i^T , \tag{1.17}$$

dove $U = (u_1, ..., u_n) \in \mathbb{R}^{m\times n}$ e $V = (v_1, ..., v_n) \in \mathbb{R}^{n\times n}$ sono matrici con colonne ortonormali, $U^TU = V^TV = I_n$, e dove la matrice diagonale $\Sigma = \text{diag}(\sigma_1, ..., \sigma_n)$ ha elementi non negativi che si presentano in ordine non crescente $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. I numeri σ_i sono detti *valori singolari* di A mentre i vettori u_i e v_i sono detti *vettori singolari sinistri* e *destri* di A, rispettivamente. La SVD è definita per ogni m e n.

Geometricamente parlando, la SVD di A fornisce due insiemi di vettori di base ortonormali, vale a dire le colonne di U e V, in modo tale che la matrice diventi diagonale quando viene trasformata tramite queste due basi.

I valori singolari sono sempre ben condizionati rispetto alle perturbazioni: se A è perturbata da una matrice E allora la norma $\|E\|_2$ è un limite superiore

per la perturbazione assoluta di ciascun valore singolare

$$|\sigma_i(A+E) - \sigma_i(A)| \leq \|E\|_2 \,.$$

Dalle relazioni $A^TA = V\Sigma^2V^T$ e $AA^T = U\Sigma^2U^T$ vediamo che la SVD di A è strettamente legata alla decomposizione agli autovalori delle matrici simmetriche A^TA e AA^T. Questo mostra che i valori singolari sono univocamente determinati per una data matrice A, quindi possiamo dire che la SVD è essenzialmente unica, a meno di un cambio di segno nella coppia (u_i, v_i), ad eccezione dei vettori singolari associati a valori singolari multipli, in cui solo gli spazi generati dai vettori sono unici. In relazione a problemi discreti mal posti, si riscontrano spesso due caratteristiche peculiari della SVD.

- I valori singolari σ_i decadono gradualmente a zero senza un particolare gap. Un aumento della dimensione di A aumenterà il numero di valori singolari piccoli.
- I vettori singolari sinistri e destri u_i e v_i tendono ad avere più cambi di segno nei loro elementi man mano che l'indice i aumenta, cioè come σ_i diminuisce.

Per vedere come la SVD dà un'idea del mal condizionamento di A, consideriamo le relazioni

$$\left.\begin{array}{rll} Av_i = \sigma_i u_i \,, & \|Av_i\|_2 & = \sigma_i \\ A^Tu_i = \sigma_i v_i \,, & \|A^Tu_i\|_2 & = \sigma_i \end{array}\right\} \qquad i = 1, ..., n \,. \tag{1.18}$$

Vediamo che un piccolo σ_i, rispetto a $\sigma_1 = \|A\|_2$, significa che esiste una certa combinazione lineare delle colonne di A, caratterizzata dagli elementi del vettore singolare destro v_i, tale che $\|Av_i\|_2 = \sigma_i$ è piccolo. Lo stesso vale per u_i e le righe di A. Una situazione con uno o più σ_i piccoli implica che A è quasi a rango non pieno, e i vettori u_i e v_i associati al piccolo σ_i sono vettori che appartengono al nucleo di A^T e A rispettivamente. Da questa proprietà e dalle altre due menzionate sopra, concludiamo che in un problema discreto mal posto la matrice è sempre altamente mal condizionata.

La SVD dà anche informazioni sull'effetto smoothing tipicamente associato a un nucleo K integrabile al quadrato. Al decrescere dei σ_i, u_i e v_i diventano sempre più oscillatori. Consideriamo

$$x = \sum_{i=1}^{n} (v_i^Tx)v_i \qquad \text{e} \qquad Ax = \sum_{i=1}^{n} \sigma_i(v_i^Tx)u_i \,.$$

Queste relazioni mostrano che a causa della moltiplicazione con σ_i, le componenti ad alta frequenza di x sono più smorzate in Ax rispetto alle componenti a bassa frequenza. Inoltre il problema inverso (calcolare x da $Ax = b$ o

$\min \|Ax - b\|_2$) deve avere l'effetto opposto: amplifica le oscillazioni ad alta frequenza in b.

Un altro uso della SVD è legato ai problemi ai minimi quadrati, a rango non pieno. Se A è invertibile, allora $A^{-1} = \sum_{i=1}^{n} v_i \sigma_i^{-1} u_i^T$, così la soluzione di $Ax = b$ è $x = \sum_{i=1}^{n} \sigma_i^{-1}(u_i^T b) v_i$. Se A non è invertibile, allora la pseudoinversa è data da

$$A^\dagger = \sum_{i=1}^{\mathrm{rank}(A)} v_i \sigma_i^{-1} u_i^T \,, \tag{1.19}$$

così se $\mathrm{rank}(A) < n$, la soluzione di $\min \|Ax - b\|_2$ con minima norma è

$$x_{LS} = A^\dagger b = \sum_{i=1}^{\mathrm{rank}(A)} \frac{u_i^T b}{\sigma_i} v_i \,. \tag{1.20}$$

È la divisione per valori singolari piccoli nelle espressioni delle soluzioni x e x_{LS} che amplifica le componenti ad alta frequenza in b.

La sensibilità delle soluzioni x e x_{LS} a perturbazioni di A e b può essere misurata dal *numero di condizionamento* di A

$$\mathrm{cond}(A) = \|A\|_2 \|A^\dagger\|_2 = \frac{\sigma_1}{\sigma_{\mathrm{rank}(A)}} \,. \tag{1.21}$$

1.3.2 La GSVD

La GSVD della coppia di matrici (A, L) è una generalizzazione della SVD di A nel senso che i valori singolari generalizzati di (A, L) sono essenzialmente le radici quadrate degli autovalori generalizzati della coppia di matrici $(A^T A, L^T L)$.

Sia $A \in \mathbb{R}^{m \times n}$, $L \in \mathbb{R}^{p \times n}$ a rango pieno, $m \geq n \geq p$, $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$, dove $\mathcal{N}$ indica il nucleo. La GSVD è una decomposizione di A e di L della forma

$$A = U \begin{pmatrix} \Sigma & 0 \\ 0 & I_{n-p} \end{pmatrix} X^{-1} \,, \qquad L = V(M \,, 0) X^{-1} \,. \tag{1.22}$$

Le colonne di $U \in \mathbb{R}^{m \times n}$ e $V \in \mathbb{R}^{p \times p}$ sono ortonormali, $U^T U = I_n$ e $V^T V = I_p$; $X \in \mathbb{R}^{n \times n}$ è non-singolare con colonne che sono $A^T A$-ortogonali; Σ e M sono matrici $p \times p$ diagonali: $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_p)$, $M = \mathrm{diag}(\mu_1, ..., \mu_p)$, i cui elementi sono non negativi e ordinati in modo tale che

$$0 \leq \sigma_1 \leq \cdots \leq \sigma_p \leq 1 \,, \qquad 1 \geq \mu_1 \geq \cdots \geq \mu_p > 0 \,,$$

e normalizzati tale che $\sigma_i^2 + \mu_i^2 = 1$, $i = 1, ..., p$. Allora i *valori singolari generalizzati* γ_i di (A, L) sono definiti come i rapporti

$$\gamma_i = \frac{\sigma_i}{\mu_i} \,, \qquad i = 1, ..., p \,, \tag{1.23}$$

e appaiono ovviamente in ordine non decrescente. Essendo

$$X^T A^T A X = \begin{pmatrix} \Sigma^2 & 0 \\ 0 & I_{n-p} \end{pmatrix} \qquad \text{e} \qquad X^T L^T L X = \begin{pmatrix} M^2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{1.24}$$

vediamo che (γ_i^2, x_i) sono le soluzioni del problema autovalori-autovettori generalizzati[1] della coppia $(A^T A, L^T L)$ associati a p autovalori generalizzati.

In modo analogo alla SVD, le coppie (σ_i, μ_i) sono ben condizionate rispetto alle perturbazioni di A e L.

Analogamente alla SVD, la GSVD fornisce tre nuovi insiemi di vettori di base linearmente indipendenti (le colonne di U, V e X) tali che le matrici A e L diventano simultaneamente diagonali quando vengono trasformate tramite queste basi. Le due basi associate alle colonne di U e V sono ortonormali.

Per $p < n$ la matrice $L \in \mathbb{R}^{p \times n}$ ha sempre nucleo $\mathcal{N}(L)$ non banale. Le ultime $n - p$ colonne x_i di X soddisfano

$$L x_i = 0\,, \qquad i = p+1, ..., n\,, \tag{1.25}$$

ed esse sono così vettori di base per $\mathcal{N}(L)$.

Quando L è la matrice identità I_n, allora U e V della GSVD sono identiche alla U e V della SVD, e i valori singolari generalizzati di (A, L) sono identici ai valori singolari di A, eccetto per l'ordine inverso.

Non c'è un semplice legame tra i valori/vettori singolari generalizzati e valori/vettori singolari ordinari.

Teorema 1.1. *Siano $\psi_i(A)$ e $\psi_i(L)$ i valori singolari ordinari di A e L rispettivamente, e siano σ_i e μ_i gli elementi diagonali nella GSVD di (A, L). Allora, per ogni $\sigma_i \neq 0$ e ogni μ_i*

$$\left\| \begin{pmatrix} A \\ L \end{pmatrix}^{\dagger} \right\|_2^{-1} \leq \frac{\psi_{n-i+1}(A)}{\sigma_i} \leq \left\| \begin{pmatrix} A \\ L \end{pmatrix} \right\|_2 \tag{1.26}$$

e

$$\left\| \begin{pmatrix} A \\ L \end{pmatrix}^{\dagger} \right\|_2^{-1} \leq \frac{\psi_i(L)}{\mu_i} \leq \left\| \begin{pmatrix} A \\ L \end{pmatrix} \right\|_2 . \tag{1.27}$$

Poiché $\gamma_i = \sigma_i (1 - \sigma_i^2)^{-1/2} \approx \sigma_i$ per piccoli σ_i, i valori singolari generalizzati devono quindi decadere gradualmente fino a zero come fanno i valori singolari ordinari. Di conseguenza, per un problema discreto, si trovano solitamente le seguenti tre caratteristiche della GSVD, simili a quelle della SVD.

[1] Vuol dire $A^T A x_i = \gamma_i^2 L^T L x_i$

- I valori singolari generalizzati γ_i decadono a zero senza alcun gap. Un aumento delle dimensioni di A aumenta il numero di piccoli valori singolari generalizzati.
- I vettori singolari u_i, v_i e x_i hanno più cambi di segno nei loro elementi quando il corrispondente γ_i decresce.
- Se L approssima un operatore derivata, allora le ultime $n - p$ colonne x_i hanno pochissimi cambi di segno, poiché sono i vettori del nucleo di L.

Sottolineiamo nuovamente che la seconda di queste caratteristiche è molto difficile - forse impossibile - da dimostrare in generale, ma che viene osservata in molti problemi discreti mal posti derivanti da applicazioni.

1.4 Trasformazione alla forma standard

Un problema di regolarizzazione con $\Omega(x) = \|L(x - x^*)\|_2$ norma smoothing discreta è detto essere in *forma standard* se la matrice L è la matrice identità I_n. In molte applicazioni la regolarizzazione in forma standard non è la scelta migliore, vale a dire si dovrebbe usare una $L \neq I_n$. Supponiamo che la matrice L di dimensione $p \times n$ abbia rango pieno (p).

Da un punto di vista numerico è molto più semplice trattare problemi in forma standard, fondamentalmente perché è coinvolta solo una matrice, A, invece delle due matrici A e L. Quindi, è conveniente essere in grado di "assorbire" la matrice L nella matrice A, vale a dire, trasformare un dato problema di regolarizzazione con residuo $\|Ax - b\|_2$ e norma smoothing $\Omega(x) = \|L(x - x^*)\|_2$ in uno in forma standard con una nuova variabile $\bar{x}$, un nuovo residuo $\|\bar{A}\bar{x} - \bar{b}\|_2$ e una nuova norma smoothing $\bar{\Omega}(\bar{x}) = \|\bar{x} - \bar{x}^*\|_2$. Ad esempio, per la regolarizzazione di Tikhonov (§2.2.2) vogliamo trasformare il problema in forma generale

$$\min\{\|Ax - b\|_2^2 + \lambda^2 \|L(x - x^*)\|_2^2\}$$

in un problema in forma standard

$$\min\{\|\bar{A}\bar{x} - \bar{b}\|_2^2 + \lambda^2 \|\bar{x} - \bar{x}^*\|_2^2\}\,.$$

Dobbiamo calcolare la nuova matrice $\bar{A}$, il nuovo termine destro $\bar{b}$, e il nuovo vettore $\bar{x}^*$ a partire dalle quantità originali A, L, b e x^*. Dobbiamo poi trasformare la soluzione dalla forma standard $\bar{x}_{\text{reg}}$ nella forma generale, in modo che la soluzione trasformata risolva il problema in forma generale.

Nel caso semplice in cui L sia quadrata e invertibile, la trasformazione è ovvia: $\bar{A} = AL^{-1}$, $\bar{b} = b$, $\bar{x}^* = Lx^*$, e la trasformazione all'indietro diventa $x_\lambda = L^{-1}\bar{x}_\lambda$.

In molte applicazioni, comunque, la matrice L non è quadrata, e la trasformazione diventa più complicata di una semplice inversione di matrice. Ciò che ora è necessario è l'*inversa generalizzata A-pesata di L* definita come

$$L_A^\dagger \equiv \left(I_n - \left(A(I_n - L^\dagger L) \right)^\dagger A \right) L^\dagger \,. \tag{1.28}$$

In generale $L_A^\dagger$ è diversa dalla pseudoinversa $L^\dagger$ (se $p < n$). Se $p \geq n$ allora $L_A^\dagger = L^\dagger$. Inoltre, abbiamo bisogno della componente x_0 della soluzione regolarizzata in $\mathcal{N}(L)$, data da

$$x_0 \equiv \left(A(I_n - L^\dagger L) \right)^\dagger b \,. \tag{1.29}$$

Data la GSVD di (A, L), $L_A^\dagger$ e x_0 possono essere espresse come

$$L_A^\dagger = X \begin{pmatrix} M^{-1} \\ 0 \end{pmatrix} V^T \,, \qquad x_0 = \sum_{i=p+1}^{n} u_i^T b x_i \,. \tag{1.30}$$

Allora le quantità in forma standard assumono la forma

$$\bar{A} = AL_A^\dagger \,, \qquad \bar{b} = b - Ax_0 \,, \qquad \bar{x}^* = Lx^* \,, \tag{1.31}$$

mentre la trasformazione all'indietro alla forma generale diventa

$$x = L_A^\dagger \bar{x} + x_0 \,. \tag{1.32}$$

La componente x_0 è la componente di x che non è affetta dallo schema di regolarizzazione.

Uno dei molti vantaggi della tecnica di trasformazione alla forma standard definita in (1.31) è che esiste una semplice relazione tra la GSVD di (A, L) e la SVD di $\bar{A}$. Se la matrice U_p è costituita dalle prime p colonne di U, cioè $U_p = (u_1, ..., u_p)$, quindi

$$AL_A^\dagger = U_p \Sigma M^{-1} V^T \,, \tag{1.33}$$

mostra che i valori singolari generalizzati γ_i sono i valori singolari di $AL_A^\dagger$, tranne che per l'ordine inverso. Inoltre, i vettori u_i e v_i, $i = 1, ..., p$, sono i vettori singolari sinistri e destri di $AL_A^\dagger$, rispettivamente.

Un altro vantaggio della trasformazione nella forma standard in (1.31) è la semplice relazione $Lx = \bar{x}$ (dovuta a $LL_A^\dagger = I_p$ e $Lx_0 = 0$) e $Ax - b = \bar{A}\bar{x} - \bar{b}$ che portano immediatamente alle equazioni

$$\|Lx\|_2 = \|\bar{x}\|_2 \,, \qquad \|Ax - b\|_2 = \|\bar{A}\bar{x} - \bar{b}\|_2 \,. \tag{1.34}$$

Queste relazioni sono importanti in relazione ai metodi per la scelta del parametro di regolarizzazione (Cap. 4).

Quando viene implementata la trasformazione alla forma standard, è spesso una buona idea distinguere tra i metodi diretti e iterativi di regolarizzazione, (Capitoli 2 e 3). Per i metodi diretti, dobbiamo essere in grado di calcolare esplicitamente la matrice $\bar{A}$, preferibilmente mediante trasformazioni ortogonali, per ragioni di stabilità. Per i metodi iterativi, dobbiamo semplicemente essere in grado di calcolare in modo efficiente i prodotti matrice-vettore $\bar{A}\bar{x}$ e $\bar{A}^T z$. Di seguito, descriviamo due metodi per la trasformazione in forma standard che sono adatti rispettivamente per i metodi diretti e iterativi.

1.4.1 Trasformazione esplicita

La trasformazione *esplicita* alla forma standard per i metodi diretti si basa su due fattorizzazioni QR. Siano p, o, q gli indici che denotano le matrici con p, $n-p$, $m-(n-p)$ colonne rispettivamente. Innanzitutto, calcoliamo una fattorizzazione QR di L^T,

$$L^T = KR = (K_p, K_o) \begin{pmatrix} R_p \\ 0 \end{pmatrix} . \tag{1.35}$$

Osserviamo che essendo L a rango pieno, la sua pseudoinversa è semplicemente $L^\dagger = K_p R_p^{-T}$. Inoltre, le colonne di K_o sono una base ortonormale per il nucleo di L. Successivamente, formiamo la matrice "magra" $m \times (n-p)$ AK_o e calcoliamo la sua fattorizzazione QR,

$$AK_o = HT = (H_o, H_q) \begin{pmatrix} T_o \\ 0 \end{pmatrix} . \tag{1.36}$$

Date queste due fattorizzazioni QR, abbiamo $A(I_n - L^\dagger L) = AK_o K_o^T = H_o T_o K_o^T \Rightarrow (A(I_n - L^\dagger L))^\dagger = K_o T_o^{-1} H_o$, e quindi, secondo le (1.28) e (1.29),

$$L_A^\dagger = (I_n - K_o T_o^{-1} H_o^T A) L^\dagger \,, \qquad x_0 = K_o T_o^{-1} H_o^T b \,. \tag{1.37}$$

Quindi, essendo $AK_o T_o^{-1} H_o^T = H_o H_o^T$, otteniamo

$$\bar{A} = A(I_n - K_o T_o^{-1} H_o^T A) L^\dagger = (I_m - H_o H_o^T) A L^\dagger = H_q H_q^T A L^\dagger \,,$$

$$\bar{b} = b - AK_o T_o^{-1} H_o^T b = (I_m - H_o H_o^T) b = H_q H_q^T b \,.$$

Quando inseriamo queste quantità nella norma residua $\|\bar{A}\bar{x} - \bar{b}\|_2$ vediamo che il fattore H_q più a sinistra non contribuisce alle norme-2. Pertanto, è conveniente lavorare con versioni leggermente ridefinite di $\bar{A}$ e $\bar{b}$ dove questo fattore viene omesso; cioè, le quantità in forma standard diventano

$$\bar{A}' = H_q^T \bar{A} = H_q^T A L^\dagger = H_q^T A K_p R_p^{-T} \,, \qquad \bar{b}' = H_q^T \bar{b} = H_q^T b \,. \tag{1.38}$$

Il modo più efficiente per calcolare $\bar{A}'$ e $\bar{b}'$ è applicare le trasformazioni ortogonali che compongono K e H "al volo" ad A e b quando vengono calcolate le fattorizzazioni QR in (1.35) e (1.36). Quando il problema in forma standard è stato risolto per $\bar{x}_{\text{reg}}$, la trasformazione alla forma generale assume la forma

$$x_{\text{reg}} = L^\dagger \bar{x}_{\text{reg}} + K_o T_o^{-1} H_o^T (b - A L^\dagger \bar{x}_{\text{reg}}) \,. \tag{1.39}$$

Sottolineiamo anche che le quantità ridefinite $\bar{A}'$ e $\bar{b}'$ soddisfano ancora

$$\|Ax - b\|_2 = \|\bar{A}'\bar{x} - \bar{b}'\|_2 \,. \tag{1.40}$$

1.4.2 Trasformazione implicita

Per i metodi iterativi, dove A è accessibile solo tramite i prodotti matrice-vettore con A e A^T, non è pratico formare $\bar{A}$ o $\bar{A}'$ esplicitamente. Invece, si dovrebbero sfruttare le moltiplicazioni di matrice in (1.31) o (1.38) e usare una trasformazione *implicita* alla forma standard. Entrambi gli approcci sono numericamente molto stabili ed entrambi sono adatti per l'uso con l'algoritmo del gradiente coniugato.

Se usiamo l'approccio basato su (1.31), allora abbiamo bisogno di calcolare x_0 così come i prodotti matrice-vettore $L_A^\dagger \bar{x}$ e $(L_A^\dagger)^T x$ in modo efficiente. Data una base W per $\mathcal{N}(L)$, la definizione (1.29) porta all'espressione

$$x_0 = W(AW)^\dagger b \,. \tag{1.41}$$

Per calcolare $L_A^\dagger \bar{x}$ e $(L_A^\dagger)^T x$ in modo efficiente, abbiamo bisogno di calcolare la matrice magra $(n-p) \times n$

$$T = (AW)^\dagger A \,, \tag{1.42}$$

dove $(AW)^\dagger$ proviene da (1.41). Abbiamo anche bisogno di partizionare L, T e x come

$$L = (L_{11}, L_{12}) \,, \qquad T = (T_{11}, T_{12}) \,, \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{1.43}$$

dove $L_{11} \in \mathbb{R}^{p\times p}$, $T_{11} \in \mathbb{R}^{(n-p)\times p}$, $x_1 \in \mathbb{R}^p$. Allora $L_A^\dagger$ è data da

$$L_A^\dagger = (I_n - WT)\begin{pmatrix} L_{11}^{-1} \\ 0 \end{pmatrix} = \left(\begin{pmatrix} I_p \\ 0 \end{pmatrix} - WT_{11}\right) L_{11}^{-1}\,.$$

Quindi, i due vettori y e $\bar{y}$ sono dati dalle relazioni

$$y = L_A^\dagger \bar{x} = \left(\begin{pmatrix} I_p \\ 0 \end{pmatrix} - WT_{11}\right) L_{11}^{-1}\bar{x}$$

e

$$\bar{y} = (L_A^\dagger)^T x = (L_{11}^{-1})^T\Big((I_p, 0) - T_{11}^T W^T\Big)x\,,$$

i quali portano ai seguenti algoritmi per calcolare y e $\bar{y}$:

$$\hat{x} \leftarrow L_{11}^{-1}\bar{x}\,, \qquad y \leftarrow \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix} - WT_{11}\hat{x}\,, \tag{1.44}$$

$$\hat{x} \leftarrow x_1 - T_{11}^T W^T x\,, \qquad \bar{y} \leftarrow L_{11}^{-T}\hat{x}\,. \tag{1.45}$$

Consideriamo ora l'approccio basato su (1.38). Allora x_0 è calcolato per mezzo di (1.37), e se la matrice WT in (1.44) e (1.45) è sostituita da $K_o K_o^T$, dove K_o è una base ortonormale per $\mathcal{N}(L)$, allora i due algoritmi calcolano $y = L^\dagger\bar{x}$ e $\bar{y} = (L^\dagger)^T x$.

Capitolo 2

Metodi diretti di regolarizzazione

Si può scegliere tra due classi di metodi di regolarizzazione: quelli che sono basati su una sorta di "decomposizione canonica" come la decomposizione ai valori singolari (SVD) e quelli che evitano tali scomposizioni. Nella prima classe di metodi troviamo i metodi diretti trattati in questo capitolo, mentre i metodi iterativi (cap. 3) appartengono alla seconda classe di algoritmi.

2.1 Problemi a rango non pieno

Vediamo metodi numerici che sono utili per la regolarizzazione di problemi con matrice dei coefficienti A a rango numerico non pieno, vale a dire, problemi per cui c'è un ben determinato gap tra valori singolari di A grandi e piccoli.

2.1.1 Rango numerico

Il rango di una matrice A è definito come il numero di colonne di A linearmente indipendenti. Inoltre il rango è uguale al numero di valori singolari strettamente positivi di A. In presenza di errori (errori di misurazione, errori di approssimazione e discretizzazione, nonché errori di arrotondamento), questa definizione non è utile in quanto le colonne di A che, da un punto di vista matematico, sono strettamente linearmente indipendenti, possono essere considerate quasi linearmente dipendenti da un punto di vista pratico. Quindi è utile introdurre il concetto di *rango numerico*: è il numero di colonne di A che, rispetto ad un prefissato livello di errore, sono praticamente linearmente indipendenti.

Definizione 2.1. L' ϵ-*rango numerico* r_ϵ di una matrice A, rispetto ad una tolleranza ϵ, è definito da

$$r_\epsilon = r_\epsilon(A, \epsilon) \equiv \min_{\|E\|_2 \leq \epsilon} \text{rank}(A + E)\,. \tag{2.1}$$

In altre parole, l'ϵ-rango di A è uguale al numero di colonne di A che sono linearmente indipendenti per ogni perturbazione di A con una norma minore o uguale alla tolleranza ϵ. In termini dei valori singolari di A, l'ϵ-rango numerico r_ϵ soddisfa

$$\sigma_{r_\epsilon} > \epsilon \geq \sigma_{r_\epsilon+1}\,. \tag{2.2}$$

Sottolineiamo che l'ϵ-rango numerico ha senso solo quando c'è un *gap ben definito* tra σ_{r_ϵ} e $\sigma_{r_\epsilon+1}$. Il numero r_ϵ deve essere robusto per piccole perturbazioni della soglia ϵ e dei valori singolari. Altrimenti, si dovrebbe evitare la nozione di rango numerico e utilizzare invece i metodi di regolarizzazione progettati per problemi senza gap nell'insieme dei valori singolari.

Osserviamo che la SVD di A è una scomposizione rank-revealing, dal momento che r_ϵ può essere immediatamente trovato dall'ispezione dei valori singolari di A. Una matrice A è a rango non pieno, secondo la definizione di rango numerico, se A ha almeno un valore singolare molto piccolo, cioè se $\sigma_n \ll \sigma_1$.

Ci sono due relazioni chiave che portano a maggiori informazioni sull'ϵ-rango numerico definito sopra. La prima è la relazione tra i valori singolari e i vettori singolari,

$$\|Av_i\|_2 = \sigma_i\,, \qquad i = 1, ..., n\,.$$

La seconda è legata alla perturbazione dei valori singolari quando una matrice A^{exact} è perturbata:

$$|\sigma_i(A^{\text{exact}}) - \sigma_i(A^{\text{exact}} + E)| \leq \|E\|_2\,, \qquad i = 1, ..., n\,.$$

La prima relazione afferma che se σ_i è "piccolo" rispetto alla norma $\sigma_1 = \|A\|_2$, allora il corrispondente vettore singolare destro v_i è "quasi" un vettore del nucleo di A. La seconda relazione quantifica la nozione di "piccolo": qualsiasi valore singolare $\sigma_i(A^{\text{exact}} + E)$ della matrice perturbata non può essere distinto da un reale valore singolare nullo $\sigma_i(A^{\text{exact}})$ della matrice non perturbata se $\sigma_i(A^{\text{exact}} + E) \leq \|E\|_2$.

Sottolineiamo che nei problemi di regolarizzazione che coinvolgono una matrice di regolarizzazione $L \neq I_n$ e una trasformazione in forma standard (§1.4), è il rango numerico della matrice $AL_A^\dagger$ che è importante, e non quello di A.

2.1.2 SVD troncata e GSVD troncata

Usiamo queste decomposizioni in problemi di regolarizzazione a rango numerico non pieno.

Sistemi di equazioni a rango non pieno

Nell'impostazione ideale, senza perturbazioni ed errori, il trattamento dei problemi a rango non pieno $Ax = b$ e $\min \|Ax - b\|_2$ è semplice: basta ignorare le componenti SVD associate ai valori singolari nulli e calcolare la soluzione per mezzo di (1.20):

$$x_{LS} = \sum_{i=1}^{\text{rank}(A)} \frac{u_i^T b}{\sigma_i} v_i \, .$$

In pratica però, A non è mai esattamente a rango non pieno, ma al contrario numericamente a rango non pieno; cioè, ha uno o più valori singolari piccoli ma non nulli tali che $r_\epsilon < \text{rank}(A) = n$. I piccoli valori singolari generano difficoltà, essendo al denominatore. Per capire perché, ricordiamo che la norma di x_{LS} è data da

$$\|x_{LS}\|_2^2 = \sum_{i=1}^{n} \left(\frac{u_i^T b}{\sigma_i} \right)^2 \, .$$

Quindi, $\|x_{LS}\|_2$ è molto grande a causa del piccolo σ_i, a meno che b si trovi quasi nel range di A, cioè, a meno che gli ultimi $n - r_\epsilon$ coefficienti $u_i^T b$ soddisfino

$$|u_i^T b| < \sigma_i \, , \qquad i = r_\epsilon + 1, ..., n \, . \tag{2.3}$$

Ogni volta che sono presenti errori in b, è improbabile che questo requisito sia soddisfatto e la soluzione x_{LS} è quindi dominata dalle ultime $n - r_\epsilon$ componenti SVD.

L'approccio più comune alla regolarizzazione di problemi numericamente a rango non pieno è considerare la matrice A come una rappresentazione rumorosa di una matrice matematicamente a rango non pieno, e sostituire A con una matrice vicina ad A e matematicamente a rango non pieno. La scelta standard è la matrice A_k che ha rango k definita come

$$A_k \equiv \sum_{i=1}^{k} u_i \sigma_i v_i^T \, , \tag{2.4}$$

cioè, sostituiamo i piccoli valori singolari non nulli $\sigma_{k+1}, ..., \sigma_n$ con zeri esatti. Tra tutte le matrici Z_k con rango k, la matrice A_k minimizza sia la norma-2 che la norma di Frobenius della differenza $A - Z_k$.

È naturale scegliere il rango k di A_k come l'ϵ-rango numerico di A, cioè $k = r_\epsilon$, perché $k < r_\epsilon$ porta alla perdita di informazione associata a grandi valori singolari, mentre $k > r_\epsilon$ porta a una soluzione con norma grande.

Quando A è sostituita da A_k, allora otteniamo un nuovo problema dei minimi quadrati $\min \|A_k x - b\|_2$. La soluzione x_k di minima norma-2 $\|x\|_2$ di questo problema è unica ed è data da

$$x_k = A_k^\dagger b = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i \,. \tag{2.5}$$

La soluzione x_k è indicata come la soluzione *SVD troncata*. Il metodo è denominato *SVD troncata* (TSVD), e la matrice A_k in (2.4) è chiamata matrice TSVD.

Riassumendo, la soluzione TSVD regolarizzata x_k si ottiene sostituendo prima la matrice A mal condizionata con la matrice A_k che ha rango k, poi calcolando la soluzione dei minimi quadrati di minima norma

$$\min \|x\|_2 \quad \text{subject to} \quad \min \|A_k x - b\|_2 \,. \tag{2.6}$$

La norma di x_k è $\|x_k\|_2 = (\sum_{i=1}^{k} (u_i^T b)^2 \sigma_i^{-2})^{1/2}$, che può ovviamente essere molto più piccola della norma $\|x_{LS}\|_2$ della soluzione dei minimi quadrati. Si noti che, come in tutti i problemi di regolarizzazione, otteniamo questa riduzione della norma della soluzione consentendo una norma del residuo più ampia.

Poiché la soluzione TSVD x_k è una soluzione regolarizzata con minima norma-2, essa è connessa con la regolarizzazione in forma standard, cioè con il vincolo $\Omega(x) = \|x\|_2$. Tuttavia, è comune nei problemi di regolarizzazione utilizzare un vincolo più generale $\Omega(x) = \|Lx\|_2$.

Per affrontare tali problemi possiamo usare una trasformazione in forma standard (§1.4) per calcolare la matrice $\bar{A}$ e il corrispondente $\bar{b}$, e quindi applicare il metodo TSVD a $\bar{A}$ e $\bar{b}$. Quindi, se k indica il numero di valori singolari mantenuti di $\bar{A}$, allora in termini di GSVD della coppia di matrici (A, L), calcoliamo la matrice TSVD intermedia $Z_k = \sum_{i=p-k+1}^{p} u_i \gamma_i v_i^T$, che è l'approssimazione a rango k più vicina a $AL_A^\dagger$. Quindi la soluzione è data da $x_{L,k} = L_A^\dagger Z_k^\dagger (b - Ax_0) + x_0$ (da (1.32)), che porta all'espressione

$$x_{L,k} = \sum_{i=p-k+1}^{p} \frac{u_i^T b}{\sigma_i} x_i + \sum_{i=p+1}^{n} u_i^T b x_i \,, \tag{2.7}$$

e $x_{L,k}$ è indicata come la soluzione *GSVD troncata* (TGSVD). Si noti che l'ultimo termine in (2.7) è la componente di $x_{L,k}$ che sta nel nucleo di L.

Approssimazioni di una matrice

Come abbiamo visto, le approssimazioni di matrice svolgono un ruolo importante nella regolarizzazione dei sistemi di equazioni lineari a rango non pieno tramite la scelta dell'approssimazione a rango k di A. In particolare, la matrice TSVD A_k (2.4) è la matrice con rango k più vicina ad A rispetto alla norma-2 e alla norma di Frobenius, e dalla SVD otteniamo immediatamente

$$\|A - A_k\|_2 = \sigma_{k+1}\,, \qquad \|A - A_k\|_F = (\sigma_{k+1}^2 + \cdots + \sigma_n^2)^{1/2}\,. \tag{2.8}$$

Lo stesso problema di approssimazione di matrice si presenta nel metodo TGSVD quando si calcola l'approssimazione Z_k a rango k più vicina alla matrice $AL_A^\dagger$.

2.2 Problemi a rango mal determinato

I problemi discreti mal posti sono sistemi di equazioni derivati dalla discretizzazione di problemi mal posti. La caratteristica principale di questi problemi è che tutti i valori singolari della matrice dei coefficienti decadono gradualmente fino a zero, senza gap. Qualunque soglia ϵ venga utilizzata in (2.2), l'ϵ-rango numerico è altamente mal determinato, e quindi il concetto di "rango numerico" non è utile per questi problemi.

Di conseguenza, la regolarizzazione di problemi discreti mal posti è più complicata del semplice filtraggio di un gruppo di piccoli valori singolari.

2.2.1 Caratteristiche dei problemi discreti mal posti

Nella trattazione pratica di problemi discreti mal posti, a causa dell'enorme numero di condizionamento della matrice dei coefficienti, è necessario incorporare un qualche tipo di regolarizzazione nella procedura di soluzione per il sistema discretizzato $Ax = b$ o $\min \|Ax - b\|_2$, al fine di calcolare una soluzione utile. È anche conveniente introdurre il concetto di condizione di Picard discreta.

In *assenza di errori*, un problema discreto mal posto è caratterizzato da una matrice dei coefficienti A^{exact} i cui valori singolari σ_i^{exact} decadono gradualmente a zero e i cui vettori singolari u_i^{exact} e v_i^{exact} tendono ad avere più cambi di segno nei loro elementi man mano che l'indice i aumenta, cioè quando il corrispondente σ_i^{exact} decresce. Inoltre, i coefficienti $|(u_i^{\text{exact}})^T b^{\text{exact}}|$ decadono a zero almeno altrettanto velocemente come i valori singolari σ_i^{exact}.

In pratica, ci troviamo di fronte a vari tipi di errori in A e b. Una fonte di errori è il processo di discretizzazione coinvolto nella configurazione

del sistema lineare, e gli errori di approssimazione influenzano sia A che b. Un'altra fonte comune di errori sono gli errori di misurazione; questi errori sono presenti in b. Infine, non possiamo evitare gli errori di arrotondamento coinvolti nei calcoli con A e b e questi errori di arrotondamento possono essere interpretati come perturbazioni dei dati in input A e b.

L'effetto di tutti questi errori è che i valori singolari σ_i e i coefficienti di Fourier $u_i^T b$ non si comportano esattamente come descritto sopra. I valori singolari σ_i decrescono monotonicamente (per definizione) fino a quando tendono a stabilirsi ad un livello τ_A determinato dagli errori in A. Anche i coefficienti $|u_i^T b|$ decadono fino a quando si stabiliscono ad un livello τ_b, determinato dagli errori in b.

I due livelli di errore τ_A e τ_b, rispettivamente per i valori singolari e i coefficienti di Fourier, determinano quante informazioni sul sistema esatto (con A^{exact} e b^{exact}) possono essere estratte dal sistema dato (con A e b). Supponiamo che i valori singolari σ_i si livellano su τ_A per $i \geq i_A$, e che i coefficienti di Fourier $|u_i^T b|$ si livellano su τ_b per $i \geq i_b$. Allora possiamo solo aspettarci di recuperare quelle componenti SVD della soluzione per la quale gli errori in σ_i e $u_i^T b$ non dominano, cioè le componenti $u_i^T b/\sigma_i$ per $i \leq \min(i_A, i_b)$.

La situazione tipica è che gli errori di misura in b sono maggiori degli altri tipi di errori in A e b, e che gli errori relativi nel termine a destra $\|b^{\text{exact}} - b\|_2/\|b^{\text{exact}}\|_2$ sono quindi più grandi rispetto agli errori relativi nella matrice $\|A^{\text{exact}} - A\|_2/\|A^{\text{exact}}\|_2$. In questa situazione, i coefficienti $|u_i^T b|$ si livellano su τ_b per $i \geq i_b$ prima che i σ_i si livellino su τ_A, e quindi possiamo - grossolanamente - recuperare le prime i_b componenti SVD della soluzione. Le restanti $n - i_b$ componenti SVD sono dominate dagli errori. Queste componenti dominano la soluzione ordinaria (ai minimi quadrati) indesiderata (1.20) e pertanto dovrebbero essere filtrate nella soluzione regolarizzata.

2.2.2 Regolarizzazione di Tikhonov

L'idea chiave nel metodo di Tikhonov è di incorporare nel modello di partenza ipotesi a priori sulla dimensione e sul livello di smoothness della soluzione desiderata, nella forma della funzione smoothing $\omega(f)$ nel caso continuo, o nella (semi)norma $\|Lx\|_2$ nel caso discreto.

Per problemi discreti mal posti, la regolarizzazione di Tikhonov in forma generale porta al problema di minimizzazione

$$\min \left\{ \|Ax - b\|_2^2 + \lambda^2 \|Lx\|_2^2 \right\}, \tag{2.9}$$

dove il parametro di regolarizzazione λ controlla il peso dato alla minimizzazione del termine di regolarizzazione, in corrispondenza alla minimizzazione della norma del residuo.

Il problema di Tikhonov (2.9) ha due importanti formulazioni alternative:

$$(A^TA + \lambda^2 L^TL)x = A^Tb \qquad \text{e} \qquad \min \left\| \begin{pmatrix} A \\ \lambda L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2 .$$

Da queste formulazioni vediamo che se $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$ (la matrice dei coefficienti $\left(\begin{smallmatrix} A \\ \lambda L \end{smallmatrix}\right)$ ha rango pieno), allora la soluzione di Tikhonov $x_{L,\lambda}$ è unica ed è data da

$$x_{L,\lambda} = A_\lambda^\# b \qquad \text{con} \qquad A_\lambda^\# = (A^TA + \lambda^2 L^TL)^{-1}A^T , \tag{2.10}$$

dove $A_\lambda^\#$ è l'inversa regolarizzata di Tikhonov.

L'*algoritmo di bidiagonalizzazione di Eldén* è il modo più efficiente e numericamente stabile per calcolare la soluzione al problema di Tikhonov. Se il problema è dato in forma generale ($L \neq I_n$), allora si dovrebbe prima trasformarlo in forma standard con $\bar{A}$ e $\bar{b}$ per mezzo dell'algoritmo descritto in §1.4.1. Quindi dovrebbe essere trattato come un problema dei minimi quadrati della forma

$$\min \left\| \begin{pmatrix} \bar{A} \\ \lambda I_p \end{pmatrix} \bar{x} - \begin{pmatrix} \bar{b} \\ 0 \end{pmatrix} \right\|_2 .$$

Questo problema può essere ridotto a un problema equivalente sparso e altamente strutturato. L'idea chiave è di trasformare $\bar{A}$ in una matrice $p \times p$ bidiagonale superiore $\bar{B}$ per mezzo di trasformazioni ortogonali sinistra e destra,

$$\bar{A} = \bar{U}\bar{B}\bar{V}^T . \tag{2.11}$$

Una volta che $\bar{A}$ è stata ridotta a una matrice bidiagonale $\bar{B}$, eseguiamo la sostituzione $\bar{x} = \bar{V}\bar{\xi}$ e otteniamo il problema

$$\min \left\| \begin{pmatrix} \bar{B} \\ \lambda I_p \end{pmatrix} \bar{\xi} - \begin{pmatrix} \bar{U}^Tb \\ 0 \end{pmatrix} \right\|_2 , \tag{2.12}$$

che può essere risolto per $\bar{\xi}_\lambda = \bar{V}^T\bar{x}_\lambda$. Le rotazioni di Givens sono utilizzate per annullare la sottomatrice diagonale λI_p. Allora $\bar{x}_\lambda$ è dato da $\bar{x}_\lambda = \bar{V}\bar{\xi}_\lambda$.

Sottolineiamo che la (semi)norma della soluzione e la norma del residuo soddisfano le relazioni

$$\|Lx_{L,\lambda}\|_2 = \|\bar{\xi}_\lambda\|_2 , \qquad \|Ax_{L,\lambda} - b\|_2 = \|\bar{B}\bar{\xi}_\lambda - \bar{U}^Tb\|_2 .$$

Ciò significa che non è necessario eseguire le trasformazioni all'indietro $\bar{x}_\lambda = \bar{V}\bar{\xi}_\lambda$ o $x_{L,\lambda} = L_A^\dagger \bar{x}_\lambda + x_0$ per calcolare queste norme.

Un'osservazione fondamentale sulla regolarizzazione di Tikhonov è che il mal condizionamento di A viene aggirato introducendo un nuovo problema con una nuova matrice dei coefficienti ben condizionata $\binom{A}{\lambda L}$ con rango pieno. Un modo diverso di trattare il mal condizionamento di A è quello di derivare un nuovo problema con una matrice dei coefficienti ben condizionata a rango non pieno. Questa è la filosofia alla base dei metodi TSVD e TGSVD, che abbiamo introdotto in §2.1.2 in relazione a problemi a rango numerico non pieno. Invece per problemi con rango numerico mal determinato, non è ovvio che il troncamento della SVD/GSVD usando la "forza bruta" porti a una soluzione regolarizzata.

Vincoli di disuguaglianza

A volte è conveniente aggiungere alcuni vincoli alla soluzione di Tikhonov, come la non-negatività, la monotonia o la convessità. Tutti e tre i vincoli possono essere formulati come vincoli di disuguaglianza della forma

$$x \geq 0 \qquad \text{non-negatività}, \tag{2.13}$$

$$L_1 x \geq 0 \qquad \text{monotonia}, \tag{2.14}$$

$$L_2 x \geq 0 \qquad \text{convessità}, \tag{2.15}$$

dove L_1 e L_2 approssimano gli operatori derivata prima e seconda, rispettivamente. I vincoli possono anche essere incorporati in altri metodi di regolarizzazione diretta, come ad esempio nel metodo TSVD.

2.2.3 Fattori filtro

Alla luce della discussione di cui sopra possiamo dire che nella regolarizzazione di un problema discreto mal posto è importante scoprire *quali* componenti SVD errate filtrare e *come* filtrarle. Anche la *quantità* di regolarizzazione o di filtraggio è estremamente importante.

Per comprendere questi aspetti in modo più dettagliato, è conveniente introdurre il concetto di fattori di filtro. Se A ha rango pieno, possiamo sempre scrivere la soluzione regolarizzata x_{reg} nella forma

$$x_{\text{reg}} = V\Theta\Sigma^\dagger U^T b \qquad \text{o} \qquad x_{\text{reg}} = X\Theta \begin{pmatrix} \Sigma^\dagger & 0 \\ 0 & I_{n-p} \end{pmatrix} U^T b$$

usando la SVD o la GSVD, rispettivamente. Se $\Theta \in \mathbb{R}^{n\times n}$ è una matrice diagonale, $\Theta = \text{diag}(f_i)$, allora gli elementi diagonali f_i sono chiamati i *fattori di filtro* per il metodo di regolarizzazione.

In particolare, per molti metodi di regolarizzazione in forma standard con $L = I_n$, come la regolarizzazione di Tikhonov

$$\min \left\{ \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2 \right\} ,$$

possiamo scrivere la soluzione regolarizzata x_{reg} e il corrispondente vettore residuo $b - Ax_{\text{reg}}$ in termini della SVD di A nelle forme

$$x_{\text{reg}} = \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i \tag{2.16}$$

e

$$b - Ax_{\text{reg}} = \sum_{i=1}^{n} (1 - f_i) u_i^T b u_i + \sum_{i=n+1}^{m} u_i^T b u_i \, .$$

Allo stesso modo, per molti metodi di regolarizzazione in forma generale con $L \neq I_n$, come la regolarizzazione di Tikhonov in forma generale,

$$\min \left\{ \|Ax - b\|_2^2 + \lambda^2 \|Lx\|_2^2 \right\} , \tag{2.17}$$

possiamo scrivere la soluzione regolarizzata x_{reg}, così come i corrispondenti vettori Lx_{reg} e $b - Ax_{\text{reg}}$, in termini della GSVD di (A, L) nelle forme

$$x_{\text{reg}} = \sum_{i=1}^{p} f_i \frac{u_i^T b}{\sigma_i} x_i + \sum_{i=p+1}^{n} u_i^T b x_i \, , \tag{2.18}$$

$$Lx_{\text{reg}} = \sum_{i=1}^{p} f_i \frac{u_i^T b}{\gamma_i} v_i \, , \tag{2.19}$$

e

$$b - Ax_{\text{reg}} = \sum_{i=1}^{p} (1 - f_i) u_i^T b u_i + (I_m - UU^T) b \, . \tag{2.20}$$

Ricordiamo che $x_0 = \sum_{i=p+1}^{n} u_i^T b x_i$ è la componente non regolarizzata di x_{reg}. Il vettore $(I_m - UU^T)b$ in (2.20) è la componente incompatibile di b che si trova al di fuori del range di A.

Nelle equazioni precedenti, i fattori di filtro f_i per i particolari metodi di regolarizzazione caratterizzano lo smorzamento o il filtraggio delle componenti SVD/GSVD.[1] Per alcuni metodi esistono formule esplicite per i fattori di filtro, per altri metodi non ci sono espressioni note.

[1] Se A è esattamente a rango non pieno, allora semplicemente esclude i termini con $\sigma_i = 0$ nelle sommatorie pertinenti.

Le formulazioni in (2.16) e (2.18) sono valide anche se il metodo di regolarizzazione è della forma

$$\min \|Ax - b\|_2 \qquad \text{subject to} \qquad x \in \mathcal{S}_x \,, \tag{2.21}$$

dove $\mathcal{S}_x$ è un sottospazio k-dimensionale e k è il parametro di regolarizzazione. Ad esempio, le iterazioni regolarizzanti del gradiente coniugato (CG) (cap. 3) è un metodo di regolarizzazione della forma (2.21) per il quale possiamo derivare i corrispondenti fattori di filtro. Tuttavia, sottolineiamo che esistono anche metodi di regolarizzazione per i quali Θ non è diagonale (e le formulazioni (2.16) e (2.18) non sono valide).

I fattori di filtro sono in genere vicini a 1 per i σ_i grandi e molto più piccoli di 1 per i σ_i piccoli. In questo modo, i contributi alla soluzione regolarizzata corrispondenti ai σ_i più piccoli vengono effettivamente filtrati. La differenza tra i vari metodi di regolarizzazione con la stessa matrice L sta nel modo in cui i fattori di filtro f_i sono definiti.

Ad esempio, se introduciamo la GSVD di (A, L), i fattori filtro per la regolarizzazione di Tikhonov in forma standard e in forma generale sono dati da

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}, \;\; L = I_n \qquad \text{e} \qquad f_i = \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2}, \;\; L \neq I_n \,, \tag{2.22}$$

per $i = 1, ..., n$ e $i = 1, ..., p$ rispettivamente. Per il caso $L = I_n$, vediamo che se $\sigma_i > \lambda$, allora $f_i \approx 1$, mentre $f_i \approx \sigma_i^2/\lambda^2$ quando $\sigma_i < \lambda$; e allo stesso modo per il caso $L \neq I_n$ con σ_i sostituito da γ_i, quindi il filtraggio agisce per quelle componenti SVD/GSVD per le quali $\sigma_i < \lambda$ e $\gamma_i < \lambda$, rispettivamente.

Invece i fattori filtro per i metodi TSVD e TGSVD risultano più semplici rispetto a quelli del metodo di Tikhonov, in quanto consistono semplicemente di zero e uno:

$$\text{TSVD:} \;\; f_i = \begin{cases} 1 \,, & i \leq k \,, \\ 0 \,, & i > k \,, \end{cases} \qquad\qquad \text{TGSVD:} \;\; f_i = \begin{cases} 0 \,, & i \leq n - k \,, \\ 1 \,, & i > n - k \,, \end{cases} \tag{2.23}$$

dove k è il parametro di troncamento.

2.2.4 Condizione di Picard discreta

Sia b^{exact} il termine destro non perturbato, e sia τ_A il livello al quale i valori singolari di A tendono ad annullarsi. La condizione di Picard discreta è soddisfatta se, per tutti i valori singolari generalizzati $\gamma_i > \tau_A \|L^\dagger\|_2$, i corrispondenti coefficienti $|u_i^T b^{\text{exact}}|$ decadono a zero più velocemente dei γ_i.

Capitolo 3

Metodi iterativi di regolarizzazione

I metodi iterativi per i sistemi lineari di equazioni e i problemi lineari ai minimi quadrati si basano su schemi di iterazione che accedono alla matrice dei coefficienti A solo tramite moltiplicazioni matrice-vettore con A e A^T e producono una sequenza di vettori di iterazione $x^{(k)}$, $k = 1, 2, ...$, che convergono alla soluzione desiderata. I metodi iterativi sono preferibili ai metodi diretti quando la matrice dei coefficienti è così grande da richiedere troppo tempo o troppa memoria per lavorare con una decomposizione esplicita di A.

Siamo interessati ai metodi di regolarizzazione iterativa in cui ogni vettore di iterazione $x^{(k)}$ può essere considerato come una soluzione regolarizzata, con il numero di iterazione k che gioca il ruolo del parametro di regolarizzazione. Quindi, abbiamo bisogno di schemi di iterazione con la proprietà intrinseca che essi, quando applicati a problemi discreti mal posti, inizialmente prelevano quelle componenti SVD $(u_i^T b/\sigma_i)v_i$ corrispondenti ai valori singolari più grandi (che sono quelli desiderati in una soluzione regolarizzata), in modo tale che il numero di iterazioni k possa essere considerato come un parametro di regolarizzazione. Questo fenomeno viene a volte definito *semiconvergenza*, perché il vettore di iterazione $x^{(k)}$ inizialmente si avvicina a una soluzione regolarizzata e poi, negli stadi successivi delle iterazioni, converge a qualche altro vettore indesiderato, spesso la soluzione dei minimi quadrati $x_{LS} = A^\dagger b$.

3.1 Alcuni aspetti pratici

Per problemi a larga scala, i metodi di regolarizzazione iterativa possono essere alternative favorevoli ai metodi diretti per i seguenti motivi.

- La matrice A non viene mai alterata, ma solo "toccata" tramite i prodotti matrice-vettore con A e A^T, mentre le fattorizzazioni matriciali - e in particolare le fattorizzazioni ortogonali come la fattorizzazione QR e l'SVD - distruggono qualsiasi sparsità o struttura di A. Quindi i metodi iterativi sono adatti ogni volta che è possibile sfruttare la sparsità o la struttura di A nelle moltiplicazioni matrice-vettore.

- Poiché il numero di iterazioni k svolge il ruolo del parametro di regolarizzazione, i metodi di regolarizzazione iterativa producono una sequenza di soluzioni regolarizzate $x^{(k)}$ e i residui corrispondenti $r^{(k)} = b - Ax^{(k)}$ le cui proprietà possono essere monitorate come k cresce. Ciò è utile, ad esempio, quando si decide quando interrompere le iterazioni.

In connessione con l'uso di metodi iterativi per risolvere i sistemi simmetrici definiti positivi di equazioni lineari $Ax = b$ è comune usare una qualche forma di precondizionamento $MAx = Mb$ che migliora la convergenza del metodo. In particolare, è importante scegliere il precondizionatore M tale che il numero di condizionamento di MA sia inferiore a quello di A.

La situazione è diversa per i problemi discreti mal posti. Qui, non ha senso migliorare il condizionamento di A (e quindi il condizionamento di A^TA) perché siamo interessati a una soluzione regolarizzata che consiste essenzialmente in una frazione di tutte le componenti SVD. Preferiremmo utilizzare un precondizionatore che migliora una parte dello spettro dei valori singolari di A - vale a dire quei valori singolari da σ_1 a σ_k che contribuiscono maggiormente alla soluzione regolarizzata - e lascia invariati i restanti valori singolari. Una possibilità è di usare un metodo iterativo stazionario classico (§3.2) come precondizionatore in un metodo CG (§3.3). I precondizionatori per problemi discreti sono oggetto di ricerca attuale.

In §1.4.2 abbiamo visto come un problema di regolarizzazione in forma generale può sempre essere trasformato in un problema in forma standard. Per semplificare la presentazione, in questo capitolo assumeremo senza perdita di generalità che qualsiasi trasformazione in forma standard necessaria sia "incorporata" nella matrice dei coefficienti data, tenendo presente la semplice relazione (1.33) tra la GSVD di (A, L) e la SVD di $AL_A^\dagger$.

Quando le moltiplicazioni con $L_A^\dagger$ e $(L_A^\dagger)^T$ sono "incorporate" negli schemi iterativi, agiscono come una sorta di "precondizionatore", e possiamo lavorare direttamente con $x^{(k)}$ ed evitare la trasformazione dal vettore in forma standard $\bar{x}^{(k)}$ a $x^{(k)}$. Per vedere questo, notiamo che se $x^{(0)} = 0$ allora $\bar{x}^{(k)}$ può sempre essere scritto come un polinomio $\mathcal{P}_k$ in $\bar{A}^T\bar{A}$ di grado $k-1$ per il vettore $\bar{A}^T\bar{b}$:

$$\bar{x}^{(k)} = \mathcal{P}_k(\bar{A}^T\bar{A})\bar{A}^T\bar{b}\,.$$

Se inseriamo $\bar{A} = AL_A^\dagger$ e $\bar{b} = b - Ax_0$ in questa espressione otteniamo

$$\bar{x}^{(k)} = \mathcal{P}_k\Big((L_A^\dagger)^T A^T A(L_A^\dagger)\Big)(L_A^\dagger)^T A^T(b - Ax_0)\,.$$

Usando le eq. (1.30) e (1.31) per $L_A^\dagger$ e x_0 insieme con la GSVD, è semplice dimostrare che $(L_A^\dagger)^T A^T A x_0 = 0$. Quindi, inserendo le espressioni di cui sopra per $\bar{x}^{(k)}$ nella trasformazione all'indietro $x^{(k)} = L_A^\dagger \bar{x}^{(k)} + x_0$, otteniamo

$$x^{(k)} = \mathcal{P}_k\Big(L_A^\dagger(L_A^\dagger)^T A^T A\Big)L_A^\dagger(L_A^\dagger)^T A^T b + x_0\,. \tag{3.1}$$

Da questa relazione vediamo che possiamo considerare la matrice $L_A^\dagger(L_A^\dagger)^T$ un "precondizionatore" per i metodi iterativi, e sottolineiamo che lo scopo di questo "precondizionatore" non è migliorare il condizionamento della matrice di iterazione ma piuttosto garantire che il vettore di iterazione "precondizionata" $x^{(k)}$ si trovi nel sottospazio corretto e quindi minimizzi $\|Lx^{(k)}\|_2$.

3.2 Metodo iterativo stazionario classico

Uno dei metodi iterativi classici è l'*iterazione di Landweber*, che assume la forma

$$x^{(k)} = x^{(k-1)} + \omega A^T r^{(k-1)}\,, \qquad k = 1, 2, ..., \tag{3.2}$$

dove $x^{(0)}$ è il vettore iniziale (spesso $x^{(0)} = 0$), ω è un parametro reale che soddisfa $0 < \omega < 2\|A^T A\|_2^{-1}$, e $r^{(k)} = b - Ax^{(k)}$ è il vettore residuo corrispondente a $x^{(k)}$. Questo metodo è stato generalizzato alla forma

$$x^{(k)} = x^{(k-1)} + \mathcal{F}(A^T A)A^T r^{(k-1)}\,, \qquad k = 1, 2, ..., \tag{3.3}$$

dove $\mathcal{F}$ è una funzione razionale di $A^T A$. L'iterazione classica di Landweber corrisponde a $\mathcal{F}(A^T A) = \omega$. Se $A = \sum_{i=1}^n u_i \sigma_i v_i^T$ è la SVD di A, allora la decomposizione agli autovalori di $\mathcal{F}(A^T A)$ è data da

$$\mathcal{F}(A^T A) = \sum_{i=1}^n v_i \mathcal{F}(\sigma_i^2) v_i^T\,,$$

ed è facile mostrare che i fattori di filtro $f_i^{(k)}$ per il vettore all'iterazione k-esima $x^{(k)}$ in (3.3) sono dati da

$$f_i^{(k)} = 1 - \Big(1 - \sigma_i^2 \mathcal{F}(\sigma_i^2)\Big)^k\,, \qquad i = 1, ..., n\,. \tag{3.4}$$

In questo modo, il numero di iterazione k determina i fattori di filtro e quindi svolge il ruolo di un parametro di regolarizzazione. In particolare, i fattori di filtro per (3.2) diventano

$$f_i^{(k)} = 1 - (1 - \omega\sigma_i^2)^k \,, \qquad i = 1, ..., n \,,$$

nel qual caso $f_i^{(k)} \approx k\omega\sigma_i^2$ per $\sigma_i \ll \omega^{-1/2}$ mentre $f_i^{(k)} \approx 1$ per grandi σ_i.

Una particolare scelta di $\mathcal{F}$, vale a dire

$$\mathcal{F}(A^TA) = (A^TA + \lambda^2 I_n)^{-1} \,, \tag{3.5}$$

porta allo schema iterativo $x^{(k)} = (A^TA + \lambda^2 I_n)^{-1}(A^Tb + \lambda^2 x^{(k-1)})$ spesso definito come *regolarizzazione di Tikhonov iterata* quando $x^{(0)} = 0$. In particolare, la prima iterazione $x^{(1)}$ è la soluzione di Tikhonov x_λ. Con questa scelta di $\mathcal{F}$, ogni step di iterazione comporta il calcolo di una soluzione regolarizzata di Tikhonov in un sistema con una matrice dei coefficiente fissa A e un termine destro $r^{(k)}$ e i fattori di filtro assumono la forma

$$f_i^{(k)} = 1 - \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}\right)^k \,, \qquad i = 1, ..., n \,. \tag{3.6}$$

3.3 Iterazioni CG regolarizzanti

Attualmente, vi è un grande interesse nei metodi di regolarizzazione iterativa basati sul metodo del gradiente coniugato (CG). Questo metodo è stato originariamente progettato per risolvere grandi sistemi sparsi di equazioni con una matrice dei coefficienti definita positiva e simmetrica.

In relazione ai problemi dei minimi quadrati e ai problemi di regolarizzazione, il metodo CG viene applicato alle equazioni normali $A^TAx = A^Tb$ la cui matrice dei coefficienti A^TA è simmetrica e semidefinita positiva.

Una proprietà essenziale delle iterate CG $x^{(k)}$ con vettori residui $r^{(k)} = b - Ax^{(k)}$ è che i corrispondenti vettori residui $A^Tr^{(k)} = A^Tb - A^TAx^{(k)}$ per le equazioni normali sono ortogonali. Una conseguenza importante di questo è che se il vettore iniziale $x^{(0)}$ è zero, allora la norma della soluzione $\|x^{(k)}\|_2$ cresce monotonicamente con k. La norma del residuo $\|r^{(k)}\|_2$, d'altra parte, decresce monotonicamente con k. Il comportamento monotono di entrambi $\|x^{(k)}\|_2$ e $\|r^{(k)}\|_2$ è importante in connessione con l'uso del criterio della curva L (§4.4) come regola di stop per le iterazioni CG regolarizzanti.

Il metodo CG produce spesso vettori di iterazione in cui le componenti spettrali associate ai grandi autovalori tendono a convergere più velocemente rispetto alle restanti componenti. In relazione a problemi discreti mal posti,

lo stesso comportamento si osserva quando l'algoritmo CG viene applicato alle equazioni normali $A^TAx = A^Tb$. Poiché gli autovalori di A^TA sono semplicemente σ_i^2, ciò significa che le componenti SVD associate ai grandi valori singolari tendono a convergere più velocemente delle restanti componenti SVD, nel qual caso l'algoritmo CG ha un effetto di regolarizzazione intrinseco quando viene fermato molto prima della convergenza alla soluzione dei minimi quadrati $x_{LS} = A^\dagger b$.

3.3.1 Implementazione

In un'implementazione dell'algoritmo CG standard per matrici definite positive simmetriche, il ciclo consiste di sole cinque istruzioni. Esistono diverse implementazioni matematicamente equivalenti di questo algoritmo quando è applicato al sistema $A^TAx = A^Tb$. Sperimentalmente si è trovato che l'implementazione più stabile, in media, è quella che spesso viene chiamata CGLS. Il cuore di questo algoritmo[1] comprende cinque istruzioni, da eseguire per $k = 1, 2, \ldots$.

[1]Per l'algoritmo vedere [2] pag. 143.

Capitolo 4

Criteri di scelta del parametro

Finora abbiamo discusso vari algoritmi per calcolare una soluzione regolarizzata. Tuttavia, nessun metodo di regolarizzazione è completo senza un metodo per scegliere il parametro di regolarizzazione (parametro continuo λ o parametro discreto k). In questo capitolo discutiamo diversi metodi di scelta del parametro.

La maggior parte dei metodi di scelta del parametro sono basati su norme residue e, nel caso della curva L, anche sulla "dimensione" Ω della soluzione, tipicamente la sua seminorma. Quando si risolvono problemi in forma generale attraverso una trasformazione alla forma standard, è importante richiamare le relazioni

$$\|Lx\|_2 = \|\bar{x}\|_2 \qquad \text{e} \qquad \|Ax - b\|_2 = \|\bar{A}\bar{x} - \bar{b}\|_2 \,.$$

Queste relazioni assicurano che l'applicazione di una regola di scelta del parametro basata sulla norma al problema originale con A e b, o al problema in forma standard con $\bar{A}$ e $\bar{b}$, produca esattamente lo stesso parametro di regolarizzazione.

Per prima cosa descriviamo alcuni metodi di scelta del parametro che fanno uso esplicito della norma dell'errore nel termine destro. Quindi passiamo ai metodi che non richiedono questa informazione, in particolare il metodo della curva L.

4.1 Introduzione alla scelta del parametro

In letteratura sui metodi di scelta del parametro, l'attenzione è spesso rivolta alla convergenza della soluzione regolarizzata poiché gli errori nel termine noto tendono a zero. Cioè, dato uno schema di regolarizzazione per calcolare $x_{\text{reg}}(\lambda)$ e un metodo di scelta del parametro per calcolare λ, spesso basato

esplicitamente sulla norma dell'errore $\|e\|_2 = \|b - b^{\text{exact}}\|_2$, si esamina quanto velocemente $x_{\text{reg}}(\lambda)$ converge a x^{exact} al tendere di $\|e\|_2 \to 0$ e $\lambda \to 0$. In particolare, è desiderabile sviluppare metodi il cui tasso di convergenza sia il più veloce possibile, il che porta allo studio della "ottimalità dell'ordine" dei metodi di regolarizzazione.

Un altro aspetto dei metodi di scelta dei parametri è altrettanto importante. Poiché gli errori sono "fissi" nella maggior parte degli esperimenti, e poiché spesso accade che gli esperimenti non possano essere ripetuti - sia per motivi pratici che per il costo di preparazione dell'esperimento - siamo interessati a estrarre quante più informazioni possibile dai dati forniti. Di conseguenza, vogliamo determinare il parametro di regolarizzazione che bilancia l'errore di regolarizzazione e l'errore di perturbazione nella soluzione calcolata per il problema dato, al fine di minimizzare l'errore totale.

Per essere precisi, scriviamo l'errore nella soluzione regolarizzata $x_{\text{reg}} = A^{\#}b$ come

$$x^{\text{exact}} - x_{\text{reg}} = A^{\dagger}b^{\text{exact}} - A^{\#}b = (A^{\dagger} - A^{\#})b^{\text{exact}} - A^{\#}e\,. \tag{4.1}$$

Qui, $(A^{\dagger} - A^{\#})b^{\text{exact}}$ è l'*errore di regolarizzazione* e $A^{\#}e$ è l'*errore di perturbazione*. Quando la condizione discreta di Picard è soddisfatta, in media, l'errore di regolarizzazione diminuisce e l'errore di perturbazione aumenta man mano che viene introdotta una minore regolarizzazione. In corrispondenza del parametro di regolarizzazione ottimale, le due componenti di errore si bilanciano a vicenda.

Sia λ_{opt} il parametro di regolarizzazione ottimale che corrisponde alla minimizzazione dell'errore $x^{\text{exact}} - x_{\text{reg}}$ (4.1). L'obiettivo è derivare algoritmi di scelta del parametro che approssimano questo λ_{opt} nel modo più accurato e affidabile possibile.

Per problemi con rango numerico ben determinato, abbiamo visto che la scelta del parametro di regolarizzazione è fortemente connessa all'ϵ-rango numerico r_ϵ definito in §2.1.1. Se scegliamo il parametro di troncamento $k < r_\epsilon$, allora ovviamente tralasciamo troppe informazioni e domina l'errore di regolarizzazione. D'altra parte, se $k > r_\epsilon$, allora l'errore di perturbazione può dominare, a causa della divisione per i piccoli valori singolari. Pertanto, è naturale scegliere il parametro di regolarizzazione in modo tale che le r_ϵ componenti SVD o GSVD più grandi vengano conservate nella soluzione regolarizzata. Si noti che r_ϵ, e quindi il parametro di regolarizzazione, per questi problemi è indipendente dal termine destro.

Per i problemi discreti mal posti, la scelta del parametro di regolarizzazione è più complicata. All'interno di un certo intervallo di parametri di regolarizzazione, di solito non esiste una scelta particolare che si distingua come "naturale" rispetto alle altre scelte.

I metodi di scelta del parametro possono essere approssimativamente divisi in due classi a seconda delle loro ipotesi sulla norma dell'errore $\|e\|_2$. Le due classi possono essere caratterizzate come segue.

1. Metodi basati sulla conoscenza, o una buona stima, di $\|e\|_2$.
2. Metodi che non richiedono $\|e\|_2$, ma cercano invece di estrarre queste informazioni dal termine destro dato.

Quando sono disponibili informazioni affidabili su $\|e\|_2$, è fondamentale utilizzare queste informazioni, e questo è il cuore del principio di discrepanza e dei metodi correlati. Quando non sono disponibili informazioni particolari su $\|e\|_2$, è più difficile escogitare un metodo affidabile di scelta del parametro.

4.2 Il principio della discrepanza

Il metodo basato sulla conoscenza di $\|e\|_2$ più diffuso è il *principio di discrepanza*. Se il problema mal posto $Ax = b$ è consistente, cioè tale per cui $Ax^{\text{exact}} = b^{\text{exact}}$ vale esattamente, allora l'idea è scegliere il parametro di regolarizzazione λ tale che la norma del residuo sia uguale a priori a un limite superiore δ_e per $\|e\|_2$, cioè,

$$\|Ax_\lambda - b\|_2 = \delta_e\,, \qquad \text{dove} \qquad \|e\|_2 \leq \delta_e\,. \tag{4.2}$$

La soluzione regolarizzata calcolata per mezzo di (4.2) corrisponde a quel punto sulla curva L dato dall'intersezione con la linea verticale data da (4.2). Per un parametro di regolarizzazione discreto k, si dovrebbe usare il k più piccolo per il quale $\|Ax_k - b\|_2 \leq \delta_e$.

4.3 Generalized Cross-Validation

La *Generalized Cross-Validation* (GCV) è un metodo di scelta del parametro di regolarizzazione che non richiede la conoscenza di $\|e\|_2$. Il metodo GCV si basa su considerazioni statistiche, vale a dire che un buon valore del parametro di regolarizzazione dovrebbe prevedere i valori di dati mancanti. Più precisamente, se un elemento arbitrario b_i di b è mancante, la soluzione regolarizzata dovrebbe prevedere questo dato mancante tra le osservazioni.

Il metodo GCV è un metodo predittivo che cerca di minimizzare il presunto errore ai minimi quadrati $\|Ax_\lambda - b^{\text{exact}}\|_2$. Poiché b^{exact} è sconosciuto, si considera la funzione GCV

$$\mathcal{G}(\lambda) = \frac{\|Ax_\lambda - b\|_2^2}{\text{traccia}(I_m - AA^{\#})^2}\,. \tag{4.3}$$

Il metodo GCV restituisce il parametro λ_{GCV} che minimizza la funzione $\mathcal{G}(\lambda)$.

Si può dimostrare che se la condizione discreta di Picard è soddisfatta e il rumore è bianco, allora questo λ_{GCV} è molto vicino al λ_{opt} che minimizza la norma $\|Ax_\lambda - b^{\text{exact}}\|_2$.

4.4 Curva L

In questa sezione discutiamo un criterio di scelta del parametro che non richiede $\|e\|_2$. Si basa sulla *curva L* definita come grafico della norma smoothing discreta $\Omega(x_{\text{reg}})$ della soluzione regolarizzata, ad esempio, la (semi)norma $\|Lx_{\text{reg}}\|_2$, rispetto alla corrispondente norma del residuo $\|Ax_{\text{reg}} - b\|_2$. La curva L mostra chiaramente il compromesso tra la minimizzazione di queste due quantità, che è il cuore di qualsiasi metodo di regolarizzazione. La curva L è una curva continua quando il parametro di regolarizzazione è continuo, come nella regolarizzazione di Tikhonov, mentre se il parametro di regolarizzazione è discreto, come nella TSVD, la curva L consiste in un insieme discreto di punti.

La nostra discussione è limitata al caso $\Omega(x_{\text{reg}}) = \|Lx_{\text{reg}}\|_2$. Tuttavia, notiamo che norme, seminorme e altre misure diverse della "dimensione" della soluzione regolarizzata definiscono diversi algoritmi di regolarizzazione e talvolta può essere vantaggioso tracciare la curva L utilizzando queste norme diverse.

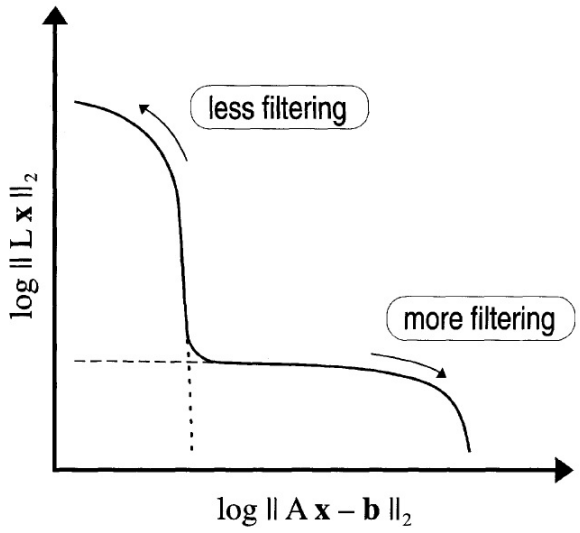

Figura 4.1: Generica forma della curva L. Notare la scala log-log.

Quando la curva è tracciata in *scala log-log*, ha un aspetto a forma di L (quindi il suo nome) con un angolo che separa le parti verticale e orizzontale della curva (Fig 4.1).

La curva L per la regolarizzazione di Tikhonov gioca un ruolo centrale in connessione con i metodi di regolarizzazione per i problemi discreti mal posti perché divide il primo quadrante in due regioni. È impossibile costruire qualsiasi soluzione che corrisponda ad un punto al di sotto della curva L di Tikhonov; qualsiasi soluzione regolarizzata deve trovarsi sulla curva o sopra questa curva. La curva L di Tikhonov ha le seguenti proprietà.

Teorema 4.1. *Sia $x_{L,\lambda}$ la soluzione di Tikhonov regolarizzata. Allora $\|Lx_{L,\lambda}\|_2$ è una funzione convessa monotona decrescente di $\|Ax_{L,\lambda} - b\|_2$, e se introduciamo le norme residue estreme corrispondenti a regolarizzazione zero e infinita,*

$$\delta_0 = \|(I_m - UU^T)b\|_2\,, \qquad \delta_\infty = (\delta_0^2 + \|U_p U_p^T b\|_2^2)^{1/2}\,,$$

dove $U_p = (u_1, ..., u_p)$, allora

$$\delta_0 \le \|Ax_{L,\lambda} - b\|_2 \le \delta_\infty\,, \qquad 0 \le \|Lx_{L,\lambda}\|_2 \le \|Lx_{LS}\|_2\,.$$

Inoltre, ogni punto (δ, η) della curva L è una soluzione dei due seguenti problemi ai minimi quadrati con vincolo di disuguaglianza:

$$\delta = \min \|Ax - b\|_2 \qquad \textit{subject to} \qquad \|Lx\|_2 \le \eta\,, \quad 0 \le \eta \le \|Lx_{LS}\|_2\,,$$

$$\eta = \min \|Lx\|_2 \qquad \textit{subject to} \qquad \|Ax - b\|_2 \le \delta\,, \quad \delta_0 \le \delta \le \delta_\infty\,.$$

Per qualsiasi metodo di regolarizzazione lineare esiste sempre una matrice $A^\#$, che chiamiamo l'*inversa regolarizzata*, in modo tale che la soluzione regolarizzata x_{reg} possa essere scritta come

$$x_{\text{reg}} = A^\# b\,. \tag{4.4}$$

Scrivendo $b = b^{\text{exact}} + e$, dove e è la perturbazione e $b^{\text{exact}} = Ax^{\text{exact}}$, la soluzione regolarizzata prende la forma $x_{\text{reg}} = A^\# b^{\text{exact}} + A^\# e$. Quindi, nel caso generale ($L \ne I_n$), l'errore è dato da

$$\begin{aligned} x^{\text{exact}} - x_{\text{reg}} &= (x^{\text{exact}} - A^\# b^{\text{exact}}) - A^\# e \\ &= \sum_{i=1}^{p} (1 - f_i) \frac{u_i^T b^{\text{exact}}}{\sigma_i} x_i \\ &\quad - \left(\sum_{i=1}^{p} f_i \frac{u_i^T e}{\sigma_i} x_i + \sum_{i=p+1}^{n} u_i^T e x_i \right)\,. \end{aligned} \tag{4.5}$$

L'errore consiste di due componenti: l'errore di perturbazione $A^{\#}e$ proveniente dall'errore e del termine destro dato, e l'errore di regolarizzazione $x^{\text{exact}} - A^{\#}b^{\text{exact}}$ dovuto alla regolarizzazione della componente priva di errori b^{exact}. Quando viene introdotta pochissima regolarizzazione, la maggior parte dei fattori di filtro f_i sono approssimativamente 1, e l'errore $x^{\text{exact}} - x_{\text{reg}}$ è dominato dall'errore di perturbazione $A^{\#}e$. Questa situazione è chiamata *undersmoothing*, e corrisponde alla parte più alta della curva L sopra l'angolo centrale. Quando viene introdotta una grande quantità di regolarizzazione, la maggior parte dei fattori di filtro sono piccoli, $f_i \ll 1$ e l'errore $x^{\text{exact}} - x_{\text{reg}}$ è dominato dall'errore di regolarizzazione $x^{\text{exact}} - A^{\#}b^{\text{exact}}$. Questa situazione è chiamata *oversmoothing*, e corrisponde alla parte più a destra della curva L alla destra dell'angolo.

Ricapitolando, la parte orizzontale corrisponde a soluzioni in cui il parametro di regolarizzazione è troppo grande e la soluzione è dominata da errori di regolarizzazione, mentre la parte verticale corrisponde a soluzioni in cui il parametro di regolarizzazione è troppo piccolo e la soluzione è dominata da errori di perturbazione.

Per un dato termine destro $b = b^{\text{exact}} + e$, c'è un parametro di regolarizzazione ottimale che bilancia l'errore di perturbazione e l'errore di regolarizzazione in x_{reg}. Una caratteristica essenziale della curva L è che questo parametro di regolarizzazione ottimale non è lontano dal parametro di regolarizzazione che corrisponde all'angolo della curva L. Tale caratteristica è la base di questo criterio. In altre parole, localizzando l'angolo della curva L si può calcolare un'approssimazione del parametro di regolarizzazione ottimale e quindi, a sua volta, calcolare una soluzione regolarizzata con un buon bilanciamento tra i due tipi di errori.

Per avere una definizione *operativa* di "angolo" definiamo l'angolo della curva L come quel punto sulla curva

$$(\zeta(\lambda), \eta(\lambda)) = (\log \|Ax_{\text{reg}} - b\|_2, \log \Omega(x_{\text{reg}}))$$

che ha massima curvatura. La curvatura κ è definita come

$$\kappa(\lambda) = \frac{\zeta'\eta'' - \zeta''\eta'}{((\zeta')^2 + (\eta')^2)^{3/2}}\,, \tag{4.6}$$

dove la derivata è rispetto a λ. Quindi il *criterio della curva L* è equivalente al calcolo del parametro di regolarizzazione che massimizza la curvatura.

Aspetti computazionali

Sebbene la curva L sia definita facilmente, calcolare il punto di massima curvatura in modo numericamente affidabile non è forse così facile come potrebbe sembrare.

Se le funzioni $\zeta(\lambda) = \log \|Ax_{\text{reg}} - b\|_2$ e $\eta(\lambda) = \log \Omega(x_{\text{reg}})$ sono definite da alcune formule computabili e se la curva L è due volte differenziabile, allora è semplice calcolare la curvatura mediante (4.6) e individuare il valore di λ che corrisponde alla massima curvatura, situazione che si verifica ad esempio quando si utilizza la regolarizzazione di Tikhonov su un problema per il quale SVD o GSVD è nota.

In molte situazioni siamo limitati a conoscere solo un numero finito di punti (ζ_i, η_i) sulla curva L. Questo è il caso, ad esempio, per i metodi TSVD e CG, e in questi e in altri casi la curva non è differenziabile. In senso computazionale, la curva L consiste quindi in un numero di punti discreti corrispondenti a diversi valori del parametro di regolarizzazione al quale abbiamo valutato $\zeta(\lambda)$ e $\eta(\lambda)$.

Abbiamo riscontrato che in molti casi, i punti su una curva L discreta sono raggruppati, fornendo dettagli della curva L che non sono rilevanti per le nostre considerazioni. Ad esempio, se c'è un ammasso di piccoli valori singolari, allora la curva L per la TSVD avrà un gruppo di punti per i corrispondenti valori del parametro di troncamento k. Questa situazione non si verifica per la regolarizzazione di Tikhonov perché tutti i componenti della soluzione entrano gradualmente man mano che i fattori di filtro cambiano da zero a uno.

Per ragioni computazionali, dobbiamo definire una curva liscia differenziabile associata ai punti discreti in modo tale che tutti i dettagli vengano scartati mentre venga mantenuta la forma complessiva della curva L. Si consiglia di adattare una *curva spline cubica* ai punti discreti della curva L. Tale curva ha diverse caratteristiche favorevoli in relazione al nostro problema: è due volte differenziabile, può essere differenziata in modo numericamente stabile e ha caratteristiche locali di conservazione della forma. Tuttavia, dobbiamo fare attenzione a non approssimare troppo bene i dettagli dei gruppi di punti.

Successivamente si può calcolare facilmente l'angolo della curva spline per mezzo di (4.6) e, se il parametro di regolarizzazione è discreto, determinare il punto sull'originale curva L discreta più vicino all'angolo della curva spline.

4.5 Comparison of solution estimator

È una procedura per confrontare le soluzioni di Tikhonov (2.10) con le soluzioni TSVD (2.5) e in questo modo determinare i parametri di regolarizza-

zione adatti per questi metodi. Ad ogni soluzione TSVD

$$x_k = A_k^\dagger b = \sum_{j=1}^{k} \frac{u_j^T b}{\sigma_j} v_j \,,$$

possiamo associare una soluzione Tikhonov

$$x_\lambda = (A^T A + \lambda^2 I)^{-1} A^T b$$

come segue. Sia

$$\rho_k = \|b - Ax_k\| \tag{4.7}$$

la norma del residuo corrispondente alla soluzione TSVD. Calcoleremo la soluzione di Tikhonov che fornisce la stessa norma del residuo.

Proposizione 4.2. *Sia r l'indice dell'ultimo valore singolare σ_r non nullo. Per ogni $1 \leq k < r$, esiste un parametro di regolarizzazione di Tikhonov $\lambda = \lambda_k$ tale che*

$$\|b - Ax_{\lambda_k}\| = \rho_k \,. \tag{4.8}$$

Sia $k_{\min}$ l'intero per il quale la norma della differenza

$$\delta_k = \|x_{\lambda_k} - x_k\| \tag{4.9}$$

ha il suo primo minimo locale all'aumentare di k. Scegliamo $\lambda = \lambda_{k_{\min}}$ come parametro di regolarizzazione di Tikhonov. Questo parametro è in accordo con il principio di discrepanza se il rumore in b è dell'ordine $\rho_{k_{\min}}$. Pertanto, possiamo usare $\rho_{k_{\min}}$ come stima per la norma del rumore in b.

I calcoli dell'algoritmo sono semplici e poco costosi quando è disponibile la SVD di A. La soluzione di Tikhonov è data da

$$x_\mu = \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_j^2 + \lambda^2} (u_j^T b) v_j \,. \tag{4.10}$$

Per i dettagli sul calcolo di λ_k e sul perché la scelta dell'indice k è appropriata si veda [4].

Quest'algoritmo calcola la SVD di A. Per problemi a larga scala non è pratico. Invece, riduciamo i problemi dei minimi quadrati a larga scala a piccole dimensioni eseguendo alcuni passaggi della bidiagonalizzazione di Golub-Kahan. Per i dettagli si veda [4].

Il metodo COSE appena descritto è adatto per problemi di regolarizzazione in forma standard. Può essere esteso a problemi di regolarizzazione in forma generale utilizzando la GSVD di (A, L). [6]

Capitolo 5

La Tomografia Computerizzata

Nel primo capitolo abbiamo introdotto tra gli esempi di problema inverso la tomografia computerizzata a raggi X. In questo capitolo approfondiamo quest'applicazione. Il problema consiste nel ricostruire la struttura interna sconosciuta di un corpo fisico a partire dalla conoscenza delle immagini a raggi X prese da direzioni diverse.

Il modello matematico generale è della forma

$$\mathbf{m} = \mathcal{A}f + \epsilon\,, \tag{5.1}$$

dove f è una funzione continua a tratti definita su un sottoinsieme di $\mathbb{R}^d$, $\mathbf{m} \in \mathbb{R}^k$ è un vettore di numeri dati da un dispositivo di misurazione, $\mathcal{A}$ è un operatore lineare che può essere correlato, ad esempio, ad un'equazione integrale. Il vettore $\epsilon \in \mathbb{R}^k$ modella gli errori sconosciuti derivanti dal rumore di misura, che è inevitabile nelle situazioni pratiche. L'informazione che abbiamo su ϵ è una disuguaglianza $\|\epsilon\| \leq \delta$ con $\delta > 0$ costante nota. Tale numero δ può essere spesso trovato mediante la calibrazione del dispositivo di misurazione.

Il problema inverso è: *"Date delle misurazioni rumorose* $\mathbf{m} = \mathcal{A}f + \epsilon$ *e* $\delta > 0$ *con* $\|\epsilon\| \leq \delta$*, estrarre informazioni su* f*"*.

In caso di problemi inversi lineari discreti consideriamo le misure della forma

$$\mathbf{m} = A\mathbf{f} + \epsilon\,, \tag{5.2}$$

dove $\mathbf{m} \in \mathbb{R}^k$, $\mathbf{f} \in \mathbb{R}^n$ e $A \in \mathbb{R}^{k\times n}$.

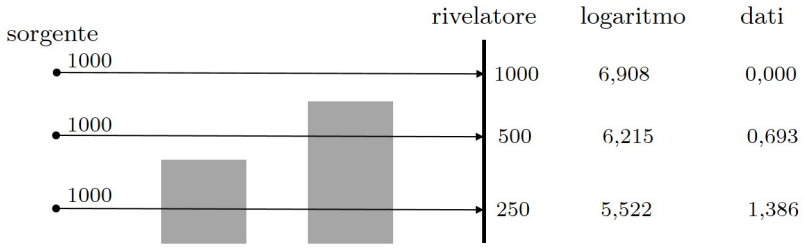

Figura 5.1: Semplice esempio che illustra l'attenuazione dei raggi X.

5.1 Un semplice esempio: due lastre di alluminio

Per prima cosa vediamo la legge esponenziale di attenuazione dei raggi X usando un esempio molto semplice, con due lastre di alluminio (figura 5.1). Tipicamente, i raggi X emanano da una posizione approssimativamente puntiforme all'interno di un tubo a raggi X. Quel punto è chiamato *sorgente di raggi X*. Tre raggi X vengono inviati verso un rivelatore, ciascuno composto inizialmente da 1000 fotoni. Il rivelatore è in grado di contare quanti fotoni arrivano in ogni punto. Uno dei raggi arriva al rivelatore attraverso lo spazio vuoto, consegnando tutti i 1000 fotoni. Un altro raggio viaggia attraverso una lastra di alluminio la cui larghezza attenua della metà il numero dei fotoni che raggiungono il rivelatore. Ciò significa che metà dei fotoni che entrano nella lastra vengono assorbiti all'interno della lastra. Il terzo raggio incontra due lastre di alluminio.

I dati del conteggio dei fotoni possono ora essere trasformati in dati integrali di linea tramite due semplici passaggi. Innanzitutto, si considera il logaritmo naturale di ogni numero di fotoni. Quindi, dal momento che l'integrale del raggio dello spazio vuoto deve essere zero, si sottrae ciascun logaritmo dal logaritmo corrispondente al raggio dello spazio vuoto. Nel semplice esempio mostrato in fig. 5.1, i dati di attenuazione risultanti sono nulli per il raggio che non attraversa le lastre, un numero positivo (0,693) per il raggio che attraversa una lastra e il doppio di quello (1,386) per il raggio che attraversa due lastre.

5.2 Dai dati del conteggio dei fotoni ai dati integrali

L'esempio delle due lastre è abbastanza semplice in quanto è coinvolto solo materiale omogeneo. Consideriamo ora una radiografia che attraversa lungo una linea retta un *phantom* che rappresenta una sezione bidimensionale della testa di un paziente (figura. 5.2). Consideriamo il *phantom* all'interno del quadrato unitario definito da $0 \leq x_1 \leq 1$ e $0 \leq x_2 \leq 1$. Supponiamo che il raggio X viaggi lungo il percorso orizzontale definito da $0 \leq x_1 \leq 1$ e $x_2 = \frac{1}{2}$.

L'interazione tra radiazioni e materia riduce l'intensità del raggio. Indichiamo con $I_0 = I(0)$ l'intensità iniziale del raggio X prima che attraversi il *phantom* e con $I_1 = I(1)$ l'intensità dopo l'attraversamento. Inoltre, denotiamo con $I(x_1)$ l'intensità nel punto $(x_1, \frac{1}{2})$ durante il cammino dalla sorgente al rivelatore.

In contrasto con il semplice esempio delle lastre omogenee, la sezione trasversale del *phantom*, come la sezione traversale di una testa contiene vari tessuti con differenti proprietà di attenuazione dei raggi X. Modelliamo questa situazione utilizzando una funzione coefficiente di attenuazione non negativa $f(x_1, x_2)$, il cui valore fornisce la perdita di intensità relativa dei

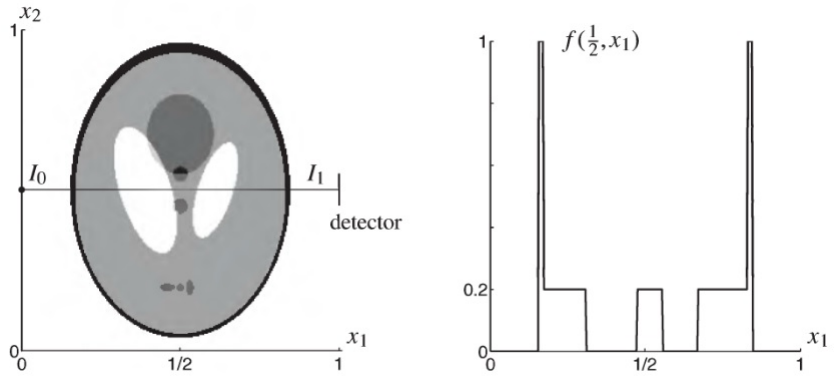

Figura 5.2: A sinistra: un raggio X che attraversa il famoso *Shepp-Logan phantom*, che rappresenta in modo stilizzato la sezione di una testa umana; le alte attenuazioni sono rappresentate con sfumature di grigio più scure, mentre le basse attenuazioni con sfumature più chiare. A destra: grafico del coefficiente di attenuazione lungo il cammino del raggio X. Le figure sono tratte da [5].

raggi X in una piccola distanza dx:

$$\frac{dI(x_1)}{I(x_1)} = -f\left(x_1, \frac{1}{2}\right) dx_1 \,.$$

Ad esempio, le ossa hanno un coefficiente di attenuazione più alto del tessuto cerebrale e il fluido cerebrospinale (ellissi bianche nel *phantom*) fornisce praticamente un'attenuazione nulla.

L'integrazione lungo i raggi X dalla sorgente al rivelatore dà

$$\int_0^1 f\left(x_1, \frac{1}{2}\right) dx_1 = -\int_0^1 \frac{I'(x_1)}{I(x_1)} dx_1 = \log I_0 - \log I_1 \,. \tag{5.3}$$

Ora è noto il termine destro di (5.3): I_0 dalla calibrazione e I_1 dalla misurazione. Il termine sinistro di (5.3) consiste in un integrale della funzione incognita f su una linea retta.

Per quanto riguarda il rumore, la quantità I_1 è un multiplo costante di una variabile casuale con distribuzione di Poisson. In genere viene campionato nella pratica utilizzando un convertitore analogico-digitale che produce output intero contenente errori di troncamento e rumore elettronico aggiuntivo. Il logaritmo di I_1 porta a una variabile casuale con statistiche notevolmente complicate. Tuttavia, di solito è abbastanza plausibile modellare la misura come

$$\log I_0 - \log I_1 = \int_0^1 f\left(x_1, \frac{1}{2}\right) dx_1 + \epsilon \,, \tag{5.4}$$

dove $\epsilon \sim \mathcal{N}(0, \sigma^2)$ è una variabile casuale distribuita normalmente.

5.3 Dati tomografici continui: trasformata di Radon

Nella sezione precedente abbiamo descritto come trasformare i dati di attenuazione da un singolo raggio X in dati integrali riguardanti un coefficiente di attenuazione non negativo $f : \mathbb{R}^2 \to \mathbb{R}$. Lo scopo dell'imaging tomografico è quello di raccogliere informazioni su f utilizzando diversi angoli di vista.

Definiamo la *trasformata di Radon*, indicata con $\Re$, in questo modo. Sia $\theta \in \mathbb{R}$ un angolo misurato in radianti, e indichiamo con

$$\vec{\theta} := \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \in \mathbb{R}^2$$

il vettore unitario che ha angolo θ rispetto all'asse x_1. La trasformata di Radon della funzione f dipende dal parametro angolare θ e dal parametro

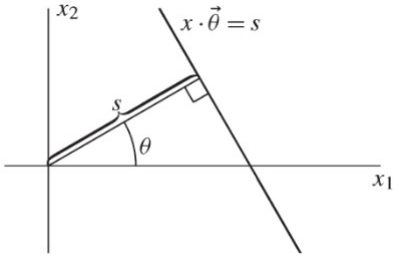

Figura 5.3: Illustrazione della definizione della trasformata di Radon.

lineare $s \in \mathbb{R}$:

$$\Re f(s,\theta) = \int_{x\cdot\vec{\theta}=s} f(x)dx^{\perp}\,, \tag{5.5}$$

dove $dx^{\perp}$ indica la misura unidimensionale di Lebesgue lungo la retta definita da $\{x \in \mathbb{R}^2 : x \cdot \vec{\theta} = s\}$. Osserviamo che la parametrizzazione dei dati tomografici fornita dalla formula (5.5) è correlata alla cosiddetta geometria a raggi paralleli utilizzata negli scanner di prima generazione per tomografia computerizzata (TAC) negli anni '70.

La trasformata di Fourier e la trasformata di Radon sono legate in modo semplice. Questo risultato è noto come teorema della sezione centrale.

Definizione 5.1. La trasformata di Fourier di una funzione definita su $\mathbb{R}$ è data da

$$\mathcal{F}(x)(\xi) = \widehat{f}(\xi) = \int_{\mathbb{R}} f(x)e^{-ix\xi}dx\,.$$

Teorema 5.1. *Sia f una funzione assolutamente integrabile definita sulla retta reale. Per ogni numero reale r e vettore unitario $\vec{\theta}$, abbiamo l'identità*

$$\int_{-\infty}^{\infty} \Re f(s,\vec{\theta})e^{-isr}ds = \widehat{f}(r\vec{\theta})\,. \tag{5.6}$$

Dimostrazione. Dalla definizione di trasformata di Radon

$$\begin{aligned}\int_{-\infty}^{\infty} \Re f(s,\vec{\theta})e^{-isr}ds &= \int_{-\infty}^{\infty}\int_{x\cdot\vec{\theta}=s} f(x)e^{-isr}dx^{\perp}ds \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)e^{-ix\cdot(r\vec{\theta})}dx_1dx_2 \\ &= \widehat{f}(r\vec{\theta}).\end{aligned}$$

□

Se introduciamo la notazione $\widetilde{h}(s,\vec{\theta}) = \int_{-\infty}^{\infty} h(t,\vec{\theta})e^{-its}dt$ per indicare la trasformata di Fourier unidimensionale di una funzione nel parametro scalare,

allora il teorema della sezione centrale dice

$$\widetilde{\Re f}(r,\vec{\theta}) = \widehat{f}(r\vec{\theta}).$$

La formula di inversione di Radon fornisce un modo per ottenere f dalla sua trasformata di Radon nel caso ideale.

Teorema 5.2. *Se f è una funzione assolutamente integrabile definita sulla retta reale e $\widehat{f}$ è assolutamente integrabile, allora*

$$f(x) = \frac{1}{(2\pi)^2}\int_0^{\pi}\int_{-\infty}^{\infty} e^{isx\cdot\vec{\theta}}\widetilde{\Re f}(s,\vec{\theta})|s|dsd\theta \tag{5.7}$$

Dimostrazione. Osserviamo che la trasformata di Radon soddisfa $\Re f(-s,-\vec{\theta}) = \Re f(s,\vec{\theta})$,

$$\begin{aligned}
\widetilde{\Re f}(-s,-\vec{\theta}) &= \int_{-\infty}^{\infty} \Re f(t,-\vec{\theta})e^{-it(-s)}dt \\
&= \int_{-\infty}^{\infty} \Re f(t,-\vec{\theta})e^{-i(-t)s)}dt \\
&= \int_{-\infty}^{\infty} \Re f(-t,-\vec{\theta})e^{-its}dt \\
&= \int_{-\infty}^{\infty} \Re f(t,\vec{\theta})e^{-its}dt \\
&= \widetilde{\Re f}(s,\vec{\theta})
\end{aligned}$$

Ora dalla formula di inversione di Fourier, con $\xi = (r\cos\theta, r\sin\theta)$,

$$\begin{aligned}
f(x) &= \frac{1}{(2\pi)^2}\int_{\mathcal{R}^2} \widehat{f}(\xi)e^{ix\cdot\xi}d\xi \\
&= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{\infty} \widehat{f}(r\vec{\theta})e^{irx\cdot\vec{\theta}}rdrd\theta \\
&= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{\infty} \widetilde{\Re f}(r,\vec{\theta})e^{irx\cdot\vec{\theta}}rdrd\theta \\
&= \frac{1}{(2\pi)^2}\int_0^{\pi}\int_{-\infty}^{\infty} \widetilde{\Re f}(r,\vec{\theta})e^{irx\cdot\vec{\theta}}|r|drd\theta\,,
\end{aligned}$$

dove l'ultima uguaglianza segue dal fatto che $\widetilde{\Re f}(-s,-\vec{\theta}) = \widetilde{\Re f}(s,\vec{\theta})$. □

Riassumendo, ciò si traduce nel seguente algoritmo di ricostruzione idealizzato per il CT imaging a raggi X:

- Sia f il coefficiente di attenuazione di una sezione bidimensionale di un oggetto tridimensionale. L'intensità del raggio $I_{(s,\vec{\theta})}$ soddisfa l'equazione differenziale

$$\frac{dI_{(s,\vec{\theta})}}{I_{(s,\vec{\theta})}} = -f(s,\vec{\theta})ds.$$

- Misuriamo la trasformata di Radon di f,

$$\Re f(s, \vec{\theta}) = \log\left(\frac{I_0}{I_d}\right),$$

  dove I_0 è l'intensità del raggio nella sorgente e I_d è l'intensità nel rivelatore.

- Ricostruiamo f dalla formula di inversione di Radon (5.7).

Per l'algoritmo filtered back-projection, consideriamo l'integrale radiale nella formula di inversione di Radon come un filtro. Indichiamo l'output del filtro con $\mathcal{G}\Re f(t, \vec{\theta})$, dove

$$\mathcal{G}\Re f(t, \vec{\theta}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\Re f}(r, \vec{\theta}) e^{irt}|r|dr.$$

Poi, con $t = x \cdot \vec{\theta}$,

$$f(x) = \frac{1}{2\pi}\int_0^{\pi} \mathcal{G}\Re f(x \cdot \vec{\theta}, \vec{\theta})d\theta.$$

Si noti che da questa formula si vede che le componenti a bassa frequenza vengono soppresse da $|r|$ e le componenti ad alta frequenza sono amplificate.

Ricordiamo che la trasformata di Fourier di $g'(t)$ è

$$\mathcal{F}(\partial_t g)(\xi) = i\xi\widehat{g}(\xi).$$

Quindi se avessimo r invece di $|r|$ nella formula di inversione di Radon, avremmo avuto la

$$\text{“inversion formula”} = \frac{1}{2\pi i}\int_0^{\pi} \partial_r \Re f(r, \theta)d\theta.$$

Se r è un valore reale, questa quantità è puramente immaginaria! Pertanto, $|r|$ è molto importante!

5.4 Dati tomografici discreti

Modelliamo i dati tomografici a raggi X mediante un insieme limitato $\Omega \subset \mathbb{R}^2$, un coefficiente di attenuazione non negativo f supportato in $\bar{\Omega}$, e un gruppo finito $\{L_j\}_{j=1}^k$ di rette $L_j \subset \mathbb{R}^2$ che interseca Ω.

Consideriamo un esempio semplice di geometria a fascio di raggi paralleli illustrato in figura 5.4. La variabile angolare è campionata con passi equidistanti sul semicerchio:

$$\theta_j = \theta_1 + \left(\frac{j-1}{J}\right)\pi, \qquad 1 \le j \le J, \tag{5.8}$$

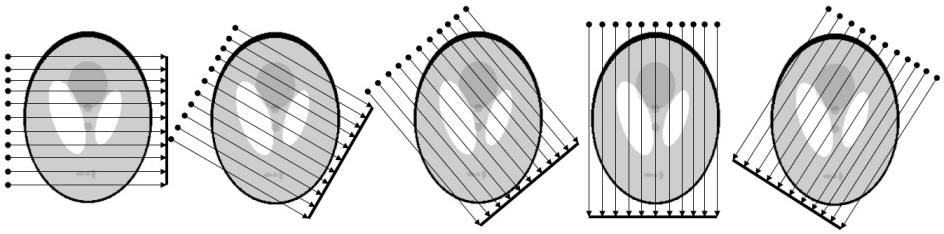

Figura 5.4: Geometria a fascio di raggi paralleli. $J = 5$ e $N = 11$.

dove $\theta_1 \in \mathbb{R}$ è una costante appropriata, un angolo di riferimento. Anche il parametro lineare s è campionato uniformemente su un intervallo:

$$s_\nu = -S + 2\left(\frac{\nu - 1}{N}\right) S, \qquad 1 \leq \nu \leq N, \tag{5.9}$$

dove $S > 0$.

Definendo $k = JN$, otteniamo

$$\mathbf{m} = \mathcal{A}f + \epsilon = \begin{bmatrix} \int_{L_1} f(x_1, x_2) ds_1 \\ \vdots \\ \int_{L_k} f(x_1, x_2) ds_k \end{bmatrix} + \epsilon, \tag{5.10}$$

dove ds_j indica la misura unidimensionale di Lebesgue lungo la retta L_j.

Per la soluzione computazionale abbiamo bisogno di costruire un modello di dimensione finita della forma (5.2). Discretizziamo il problema tomografico dividendo l'area sconosciuta in n pixel e assumendo che i valori di attenuazione siano costanti all'interno di ciascun pixel. Numeriamo i pixel da 1 a n e chiamiamo i corrispondenti valori di attenuazione $\mathbf{f}_j \geq 0$ per $j = 1, ..., n$.

La misura $\mathbf{m}_i$ dell'integrale di linea di f sulla retta L_i è approssimata da

$$\mathbf{m}_i = \int_{L_i} f(x_1, x_2) ds \approx \sum_{j=1}^{n} a_{ij} \mathbf{f}_j, \tag{5.11}$$

dove a_{ij} è la distanza che L_i attraversa nel j-esimo pixel. Si noti che solo i pixel che intersecano il raggio L_i sono inclusi in questa somma. Inoltre, se abbiamo k misurazioni nel vettore $\mathbf{m} \in \mathbb{R}^k$, allora l'ultima formula produce un'equazione matriciale $\mathbf{m} = A\mathbf{f}$, dove la matrice è definita da $A = (a_{ij})$.

Per renderci conto della mal posizione, consideriamo la seguente discretizzazione (fig. 5.5), dove $J = 2$, $N = 3$, $k = 6$ e il numero totale di pixel è $N^2 = 9$. Abbiamo diviso il dominio quadrato $\Omega \subset \mathbb{R}^2$ in 9 pixel. La lunghezza del lato di ogni pixel è 1. All'interno dei pixel c'è un valore costante $\mathbf{f}_j$ di attenuazione. Le sei frecce rappresentano i raggi X utilizzati per

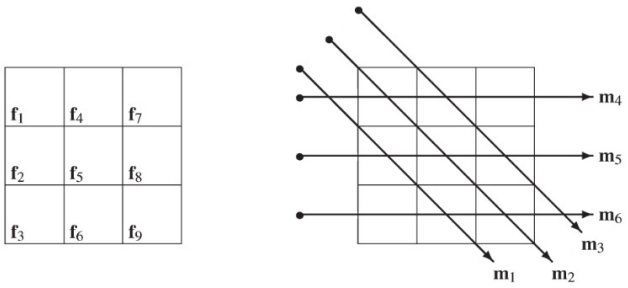

Figura 5.5: Numero di angoli $J = 2$, numero di raggi $N = 3$.

sondare la struttura interna di Ω. I dati misurati costituiscono il vettore $\mathbf{m} = [\mathbf{m}_1, ..., \mathbf{m}_6]^T$. Il modello risultante è

$$\begin{bmatrix} 0 & \sqrt{2} & 0 & 0 & 0 & \sqrt{2} & 0 & 0 & 0 \\ \sqrt{2} & 0 & 0 & 0 & \sqrt{2} & 0 & 0 & 0 & \sqrt{2} \\ 0 & 0 & 0 & \sqrt{2} & 0 & 0 & 0 & \sqrt{2} & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \\ \mathbf{f}_4 \\ \mathbf{f}_5 \\ \mathbf{f}_6 \\ \mathbf{f}_7 \\ \mathbf{f}_8 \\ \mathbf{f}_9 \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \\ \mathbf{m}_4 \\ \mathbf{m}_5 \\ \mathbf{m}_6 \end{bmatrix}. \tag{5.12}$$

Questo modello è semplice e di dimensione piccola. Tuttavia, già mostra una caratteristica tipica dei problemi inversi: la non unicità della soluzione. Infatti si riescono a trovare diversi vettori $\mathbf{f} = [\mathbf{f}_1, ..., \mathbf{f}_9]^T$ che producono lo stesso vettore di dati $\mathbf{m} = [\mathbf{m}_1, ..., \mathbf{m}_6]^T$. Ad esempio $\mathbf{f} = [4, 1, 1, 4, 3, 0, 5, 4, 2]^T$ e $\mathbf{f} = [5, 1, 4, 6, 5, 0, 2, 2, -1]^T$. Così il problema inverso non può essere risolto usando solo le informazioni provenienti dai dati misurati.

Capitolo 6

Test numerici

In questo capitolo vengono trattati degli esperimenti numerici effettuati con il software MATLAB su problemi unidimensionali e bidimensionali. Abbiamo utilizzato il *Regularization Tools* (`regtools`) [3], un pacchetto Matlab per l'analisi e la risoluzione di problemi discreti mal posti (`http://www.imm.dtu.dk/~pcha/Regutools/`).

6.1 Caso unidimensionale

Abbiamo utilizzato il problema test `shaw`, disponibile nel pacchetto `regtools`. È la discretizzazione di un'equazione integrale di Fredholm del primo tipo (1.3). La funzione `shaw` genera la matrice A e il vettore soluzione x che rappresentano, rispettivamente, la discretizzazione del nucleo K e della soluzione f. Il termine b è ottenuto come prodotto $b = Ax$.

Il primo test numerico riguarda il confronto tra due metodi di calcolo della soluzione regolarizzata: la GSVD troncata e il metodo di Tikhonov. Prima di tutto abbiamo deciso la dimensione della matrice A: abbiamo fissato $m = 32$ per ottenere una matrice A di dimensione 32×32. Abbiamo considerato tre diverse matrici di regolarizzazione L: la matrice identità I_m, la matrice che approssima l'operatore derivata prima D_1 e quella dell'operatore derivata seconda D_2. Abbiamo stabilito tre diversi livelli di noise: 10^{-6}, 10^{-4} e 10^{-2}, per ottenere il vettore $\tilde{b}$ affetto da errore. Per ogni L e per ogni noise abbiamo fatto $z = 10$ prove, cioè abbiamo generato 10 realizzazioni del vettore $\tilde{b}$ per ognuno dei tre livelli di noise e per ogni matrice di regolarizzazione. Abbiamo calcolato sia per la TGSVD sia per Tikhonov la soluzione che presenta il minor errore (`e min`) rispetto alla soluzione vera e la soluzione regolarizzata in corrispondenza dell'angolo della curva L, considerando l'errore rispetto

Tabella 6.1: Problema test: shaw, m=32, z=10

noise	matrice	e min tgsvd	e corn tgsvd	e min tikh	e corn tikh
1.00e-06	I	1.08e-01	8.45e-01	1.10e-01	6.60e-01
	D_1	9.05e-02	1.15e-01	8.15e-02	1.04e-01
	D_2	5.66e-02	5.67e-02	5.00e-02	6.25e-02
1.00e-04	I	2.24e-01	2.49e-01	1.76e-01	2.81e-01
	D_1	1.96e-01	2.60e-01	1.66e-01	2.41e-01
	D_2	2.20e-01	4.43e-01	2.01e-01	6.50e-01
1.00e-02	I	7.58e-01	1.38e+00	6.17e-01	1.06e+00
	D_1	7.54e-01	3.21e+00	7.22e-01	3.05e+00
	D_2	1.40e+00	3.80e+00	1.07e+00	3.69e+00

Tabella 6.2: Problema test: baart, m=32, z=10

noise	matrice	e min tgsvd	e corn tgsvd	e min tikh	e corn tikh
1.00e-06	I	6.35e-02	7.13e-02	5.65e-02	8.01e-02
	D_1	5.85e-02	6.64e-02	6.98e-02	6.31e-02
	D_2	4.84e-03	4.84e-03	5.29e-03	5.79e-03
1.00e-04	I	1.27e-01	1.82e-01	9.69e-02	1.44e-01
	D_1	1.13e-01	1.27e-01	7.30e-02	1.21e-01
	D_2	2.15e-02	2.15e-02	1.32e-02	2.04e-02
1.00e-02	I	2.12e-01	2.20e-01	1.79e-01	2.09e-01
	D_1	1.65e-01	1.65e-01	1.49e-01	4.26e-01
	D_2	8.75e-02	1.82e+12	7.34e-02	9.23e-02

alla soluzione vera (`e corn`). Abbiamo riportato nella tabella 6.1 la media aritmetica degli errori sulle 10 prove.

Dalla tabella 6.1 vediamo che in alcuni casi è migliore la TGSVD e in altri casi Tikhonov. Questo vuol dire che la soluzione regolarizzata della TGSVD e la soluzione regolarizzata di Tikhonov approssimano la soluzione desiderata circa allo stesso modo. Osserviamo che, a parità di noise, i risultati ottenuti con matrice di regolarizzazione diversa da I sono in genere migliori, questo perché la soluzione è smooth.

Abbiamo fatto gli stessi calcoli utilizzando il problema test `baart`, disponibile nel pacchetto `regtools`. È la discretizzazione di un'equazione integrale di Fredholm del primo tipo (1.3). Anche in questo caso, la funzione `baart` genera la matrice A e il vettore soluzione x, da cui calcoliamo il termine $b = Ax$. Abbiamo riportato i risultati nella tabella 6.2.

Dalla tabella 6.2 e dalla figura 6.3 osserviamo un errore molto grande della soluzione TGSVD ottenuta in corrispondenza del corner nel caso noise=10^{-2}

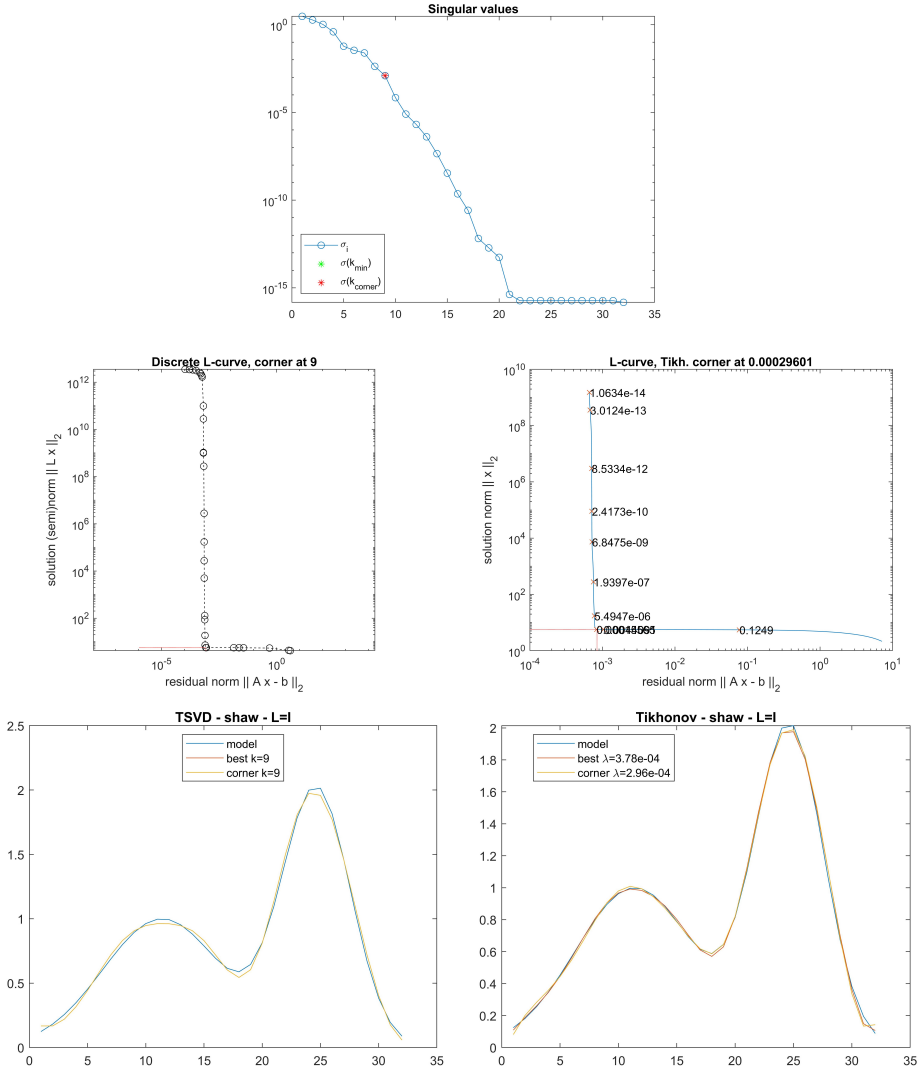

Figura 6.1: Problema test: shaw, N=32, matrice di regolarizzazione I, noise= 10^{-4}. In alto: valori singolari, in rosso il valore singolare in corrispondenza del parametro di troncamento. Al centro e in basso: curva L e soluzione in una delle dieci prove.

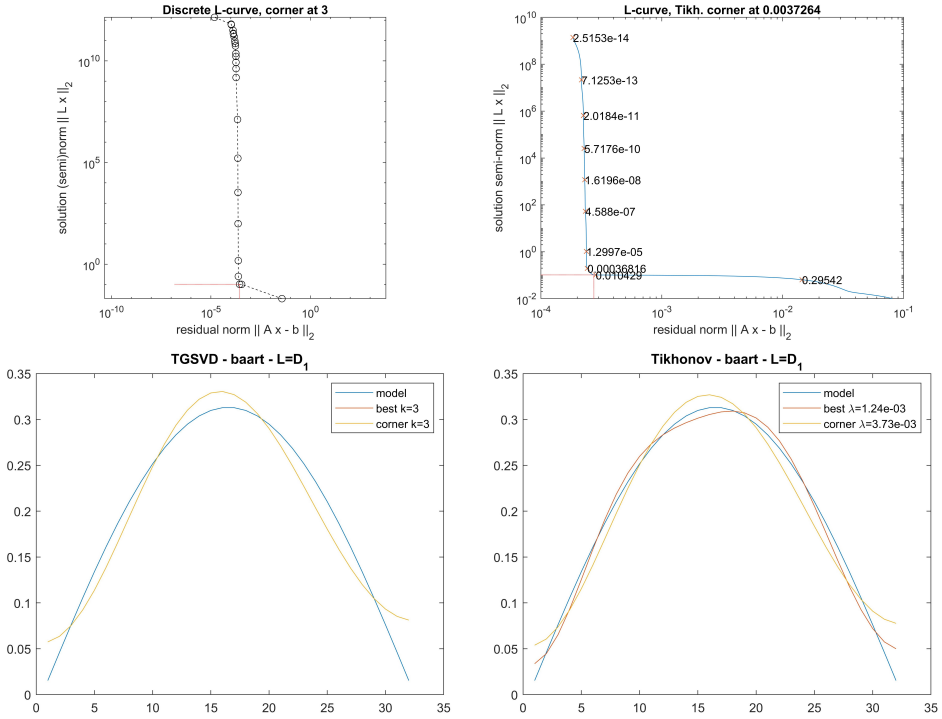

Figura 6.2: Problema test: baart, N=32, matrice di regolarizzazione D_1, noise= 10^{-4}. Curva L e soluzione in una delle dieci prove.

con matrice di regolarizzazione D_2. Un limite della curva L è che può capitare che non assuma la forma di L, il che implica l'impossibilità di determinare il corner e quindi non è possibile determinare una buona soluzione regolarizzata. Dalla fig. 6.3 vediamo che la curva L (in alto a sinistra) non ha la forma L.

6.2 Caso bidimensionale

Abbiamo utilizzato il problema test `tomo`, disponibile nel pacchetto `regtools`. Questa funzione crea un semplice problema test della tomografia bidimensionale (fig 6.4). Il dominio $[0, N] \times [0, N]$ è diviso in N^2 celle di dimensione unitaria e un totale di N^2 raggi in direzioni casuali penetrano in questo dominio. Abbiamo fissato $N = 32$ e stabilito un noiselevel $= 10^{-4}$. La funzione `tomo` crea una matrice A di dimensione 1024×1024. Abbiamo confrontato i metodi TGSVD e Tikhonov con matrici di regolarizzazione I_N, D_1, D_2 e $L = \begin{bmatrix} I_N \otimes D_1 \\ D_1 \otimes I_N \end{bmatrix}$, dove $\otimes$ indica il prodotto tensoriale.

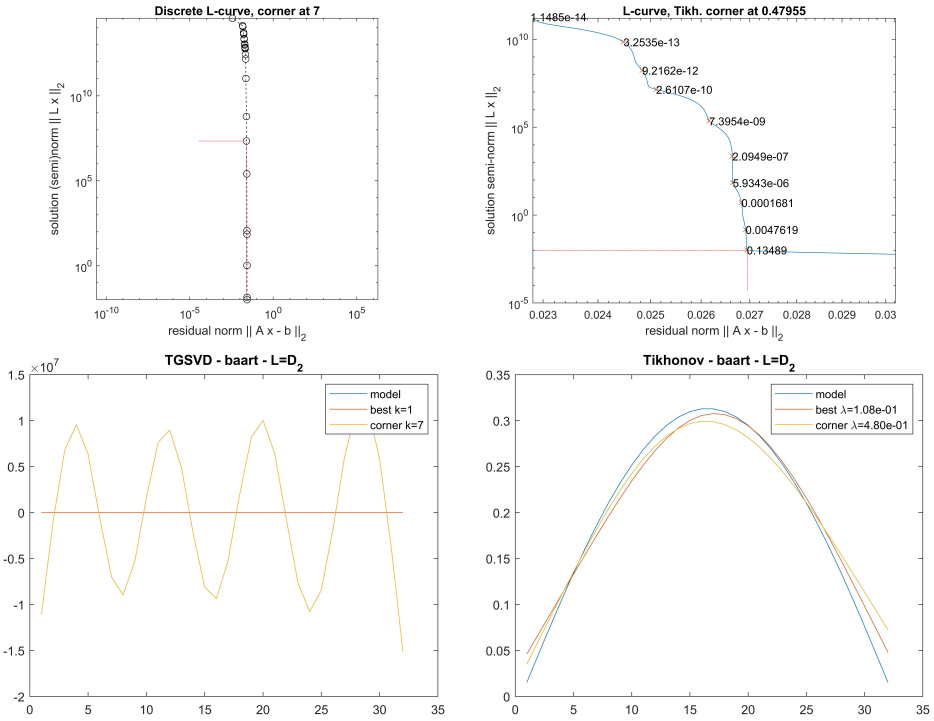

Figura 6.3: Problema test: baart, N=32, matrice di regolarizzazione D_2, noise= 10^{-2}. Curva L e soluzione in una delle dieci prove.

Abbiamo calcolato sia per la TGSVD sia per Tikhonov la soluzione che presenta il minor errore rispetto alla soluzione vera e la soluzione regolarizzata in corrispondenza dell'angolo della curva L, considerando l'errore rispetto alla soluzione vera. Le tabelle 6.3, 6.4, 6.5 e 6.6 riportano i risultati ottenuti e le figure 6.5, 6.6, 6.7 e 6.8 rappresentano le soluzioni. Osserviamo dalle figure che i due metodi ricostruiscono l'immagine in modo soddisfacente, a parte nel caso con matrice di regolarizzazione L, in cui il metodo di Tikhonov fallisce quando cerca di costruire la soluzione in corrispondenza del corner della curva L.

Dato che la dimensione delle matrici A ed L è molto elevata, questo può essere considerato un problema a larga scala. Calcolare la GSVD della coppia (A, L) è costoso. Prima di calcolare una soluzione regolarizzata, il problema di grandi dimensioni deve essere ridotto ad un problema di piccole dimensioni tramite il metodo Golub-Kahan bidiagonalization. Un approccio di questo tipo è utilizzato nel metodo denominato *comparison of solution estimator* (COSE) [4, 6], in cui le soluzioni regolarizzate di Tikhonov vengono confrontate con le iterazioni del metodo LSQR [1]. Abbiamo utilizzato la funzione `cosegkgt`, disponibile nel pacchetto `cosereg` (`http://bugs.unica.it/~gppe/soft/#cosereg`), al caso di matrice di regolarizzazione L. La funzione `cosegkgt` calcola la soluzione del problema ai minimi quadrati $\min \|Lx\|$ soggetto al vincolo $\min \|Ax - b\|$. È possibile fissare un numero massimo di step per il processo di Golub-Kahan: abbiamo fatto esperimenti imponendo nmax=400 e nmax=1100. Dai risultati (fig. 6.8) constatiamo che si ottiene una soluzione di qualità inferiore rispetto ai metodi TGSVD e di Tikhonov (nel caso di determinazione di una soluzione con errore minimo rispetto alla soluzione vera), ma comunque si tratta di una

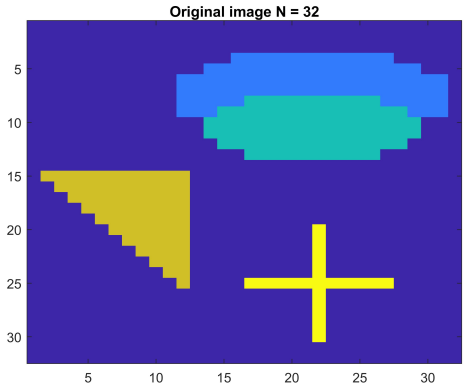

Figura 6.4: Problema test tomo: soluzione esatta.

Tabella 6.3: Problema test: tomo, N=32, matrice di regolarizzazione I

		best	corner
tsvd	parametro	1022	1022
	errore	9.75e-01	9.75e-01
tikhonov	parametro	2.02e-09	4.96e-07
	errore	9.19e-01	9.75e-01

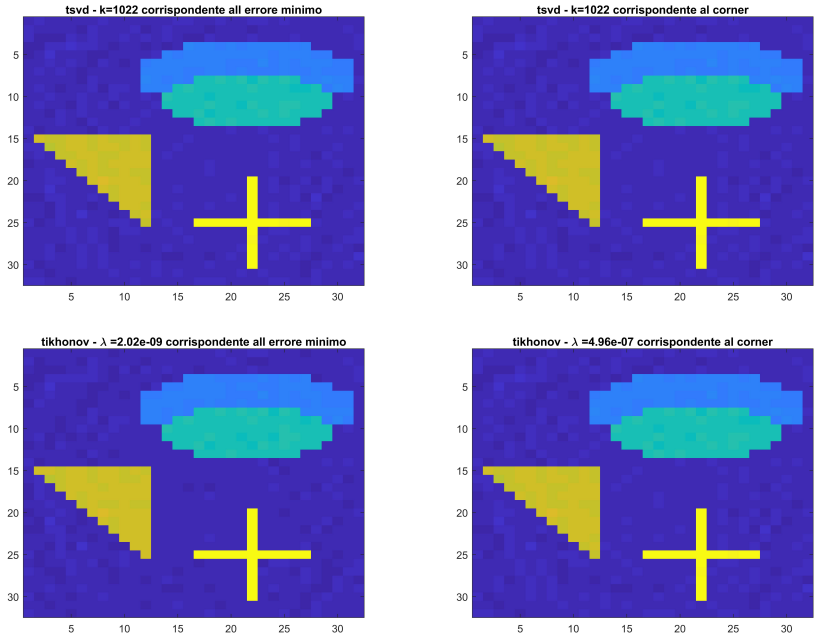

Figura 6.5: Soluzione regolarizzata, problema test: tomo, N=32, matrice di regolarizzazione I.

soluzione visivamente buona già nel caso nmax=400. Come già accennato prima, il metodo di Tikhonov fallisce nel caso di soluzione regolarizzata con parametro in corrispondenza del corner. Mentre il metodo COSE determina sempre una soluzione e inoltre può essere applicato anche a problemi di dimensioni molto grandi.

Infine abbiamo considerato un modello di simulazione dei dati più realistico, il cosiddetto *Shepp-Logan phantom* (fig. 6.9). È un modello costante a tratti di una sezione trasversale di una testa umana (§5.2). Il phantom viene definito utilizzando ellissi e può essere realizzato in qualsiasi risoluzione discreta desiderata. La prima immagine nella fig. 6.9 sembra liscia, ma è stata ottenuta tramite la risoluzione discreta 512×512. Mentre la seconda immagine è stata ottenuta con risoluzione 80×80 ed è quella che abbiamo

Tabella 6.4: Problema test: tomo, N=32, matrice di regolarizzazione D_1

		best	corner
tgsvd	parametro	1021	1021
	errore	1.35e+00	1.35e+00
tikhonov	parametro	5.30e-11	1.47e-07
	errore	9.25e-01	1.35e+00

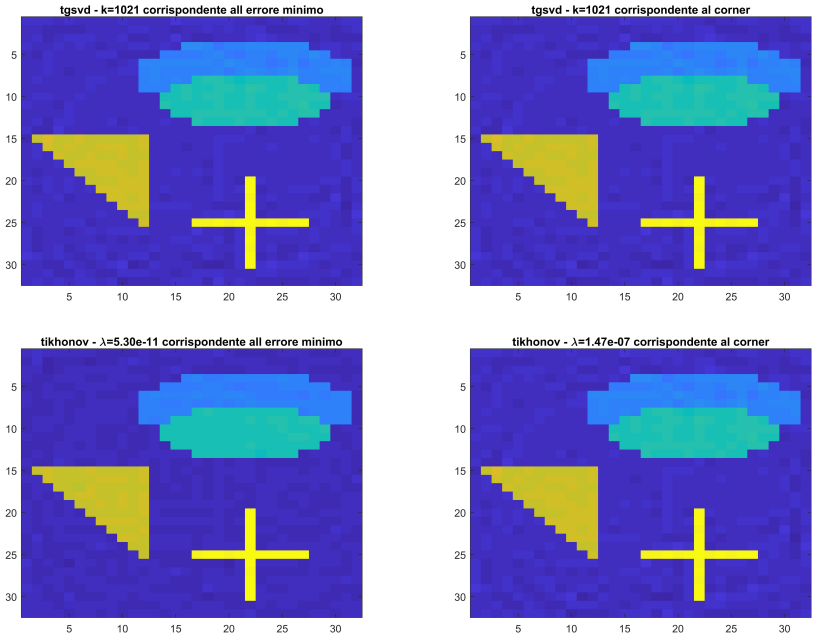

Figura 6.6: Soluzione regolarizzata, problema test: tomo, N=32, matrice di regolarizzazione D_1.

Tabella 6.5: Problema test: tomo, N=32, matrice di regolarizzazione D_2

		best	corner
tgsvd	parametro	1020	1020
	errore	1.13e+00	1.13e+00
tikhonov	parametro	5.34e-11	1.09e-07
	errore	5.51e-01	1.13e+00

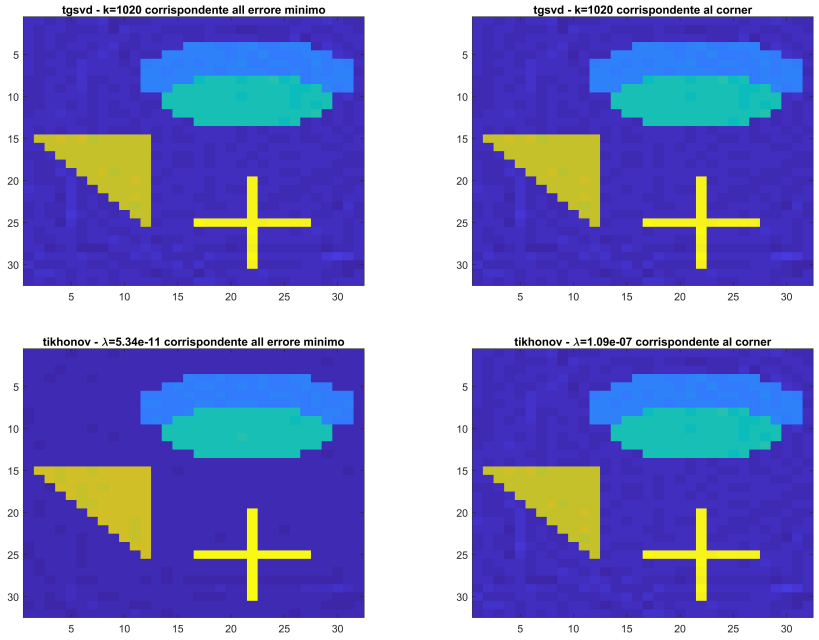

Figura 6.7: Soluzione regolarizzata, problema test: tomo, N=32, matrice di regolarizzazione D_2.

Tabella 6.6: Problema test: tomo, N=32, matrice di regolarizzazione L

		best	corner
tgsvd	parametro	1019	1019
	errore	1.17e+00	1.17e+00
tikhonov	parametro	2.29e-10	1.39e+13
	errore	9.58e-01	3.28e+01
cose	parametro	360	/
nmax 400	errore	6.55e+00	/
cose	parametro	919	/
nmax 1100	errore	4.05e+00	/

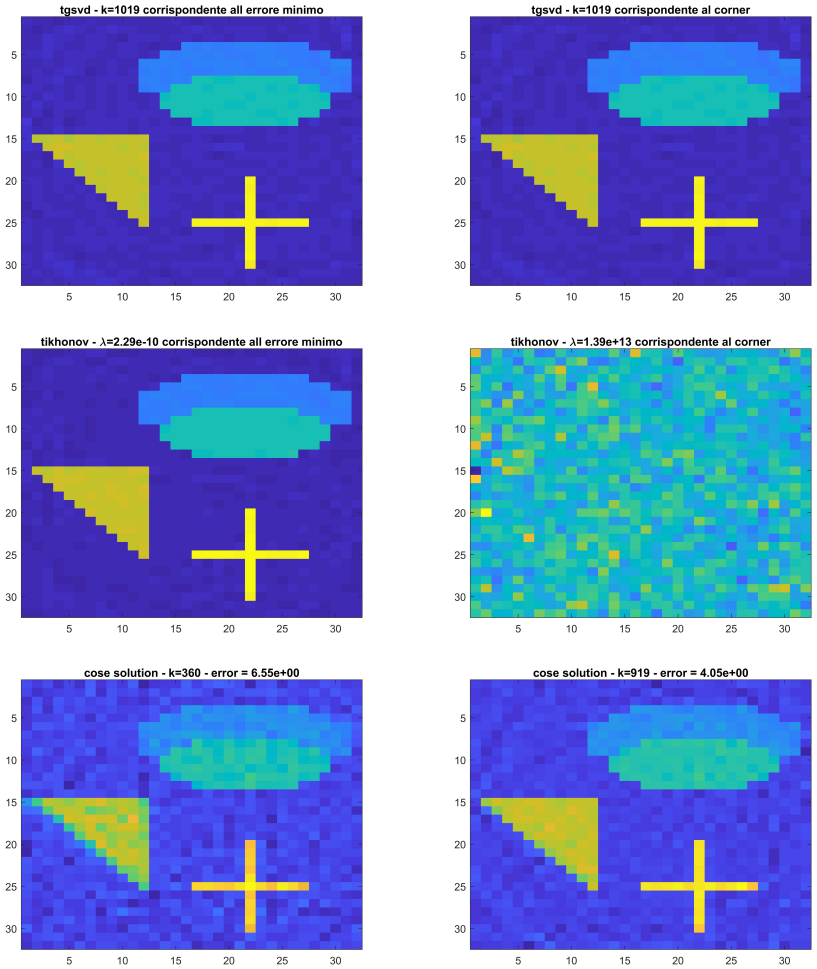

Figura 6.8: Soluzione regolarizzata, problema test: tomo, N=32, matrice di regolarizzazione L. In basso: soluzioni ottenute con il procedimento cose, nmax=400 a destra, nmax=1100 a sinistra.

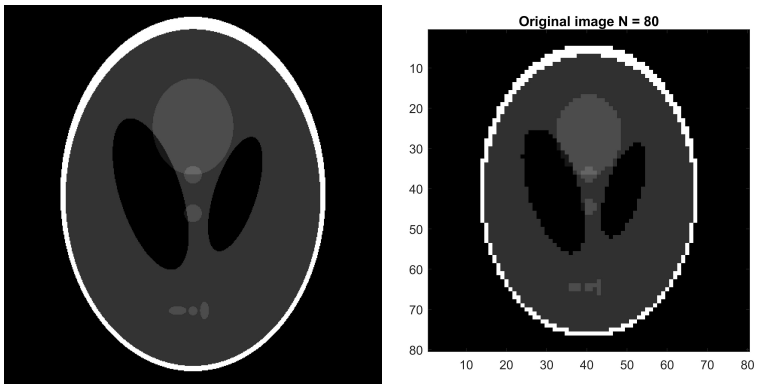

Figura 6.9: Shepp-Logan phantom. A sinistra: N=512. A destra: N=80.

Tabella 6.7: Problema test: shepp-logan, N=80

noise	matrice	e min tgsvd	e corn tgsvd	e min tikh	e corn tikh
1.00e-04	I	6.07e+00	7.04e+00	6.03e+00	6.57e+00
	D_1	7.83e+00	1.61e+01	7.77e+00	1.68e+01
	D_2	1.28e+01	1.60e+01	2.46e+01	1.62e+01
	L	6.06e+00	6.19e+00	6.03e+00	1.71e+01

utilizzato per i nostri esperimenti.

Abbiamo considerato quattro diverse matrici di regolarizzazione: I, D_1, D_2 e L definite come negli esempi precedenti. Abbiamo fissato il livello di noise: 10^{-4}. Abbiamo calcolato sia per la TGSVD sia per Tikhonov la soluzione che presenta il minor errore (`e min`) rispetto alla soluzione vera e la soluzione regolarizzata in corrispondenza dell'angolo della curva L, considerando l'errore rispetto alla soluzione vera (`e corn`). Abbiamo costruito la matrice A tramite la funzione `radon` di Matlab. Avendo fissato N= 80 e il vettore degli angoli da 0^o a 180^o con passo 5, `radon` restituisce la matrice A di dimensione 4329×6400. La tabella 6.7 e le figure 6.10, 6.11, 6.12, 6.13 riportano i risultati.

Dalle figure vediamo che la determinazione del corner della curva L e il conseguente calcolo di una soluzione regolarizzata con parametro in corrispondenza del corner fallisce completamente. Il metodo del corner della curva L è tutt'altro che infallibile. Osserviamo che nei casi di matrice di regolarizzazione D_1 e D_2 l'immagine risulta rigata verticalmente, questo perché gli operatori derivata agiscono su colonne. Questo effetto non lo osserviamo nei casi di matrice I ed L, perché per quanto riguarda L, essa agisce su righe e su colonne.

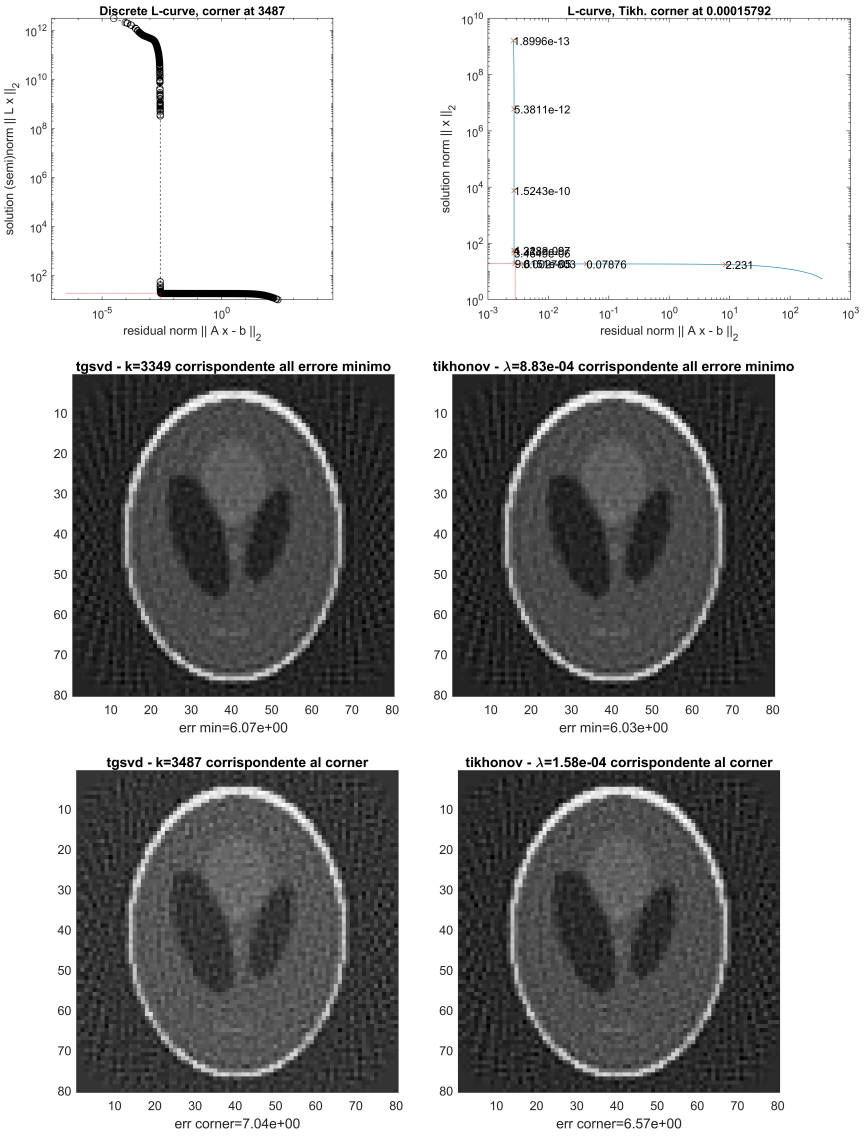

Figura 6.10: Problema test: shepp-logan, N=80, matrice di regolarizzazione I, noise= 10^{-4}. Curva L e soluzioni.

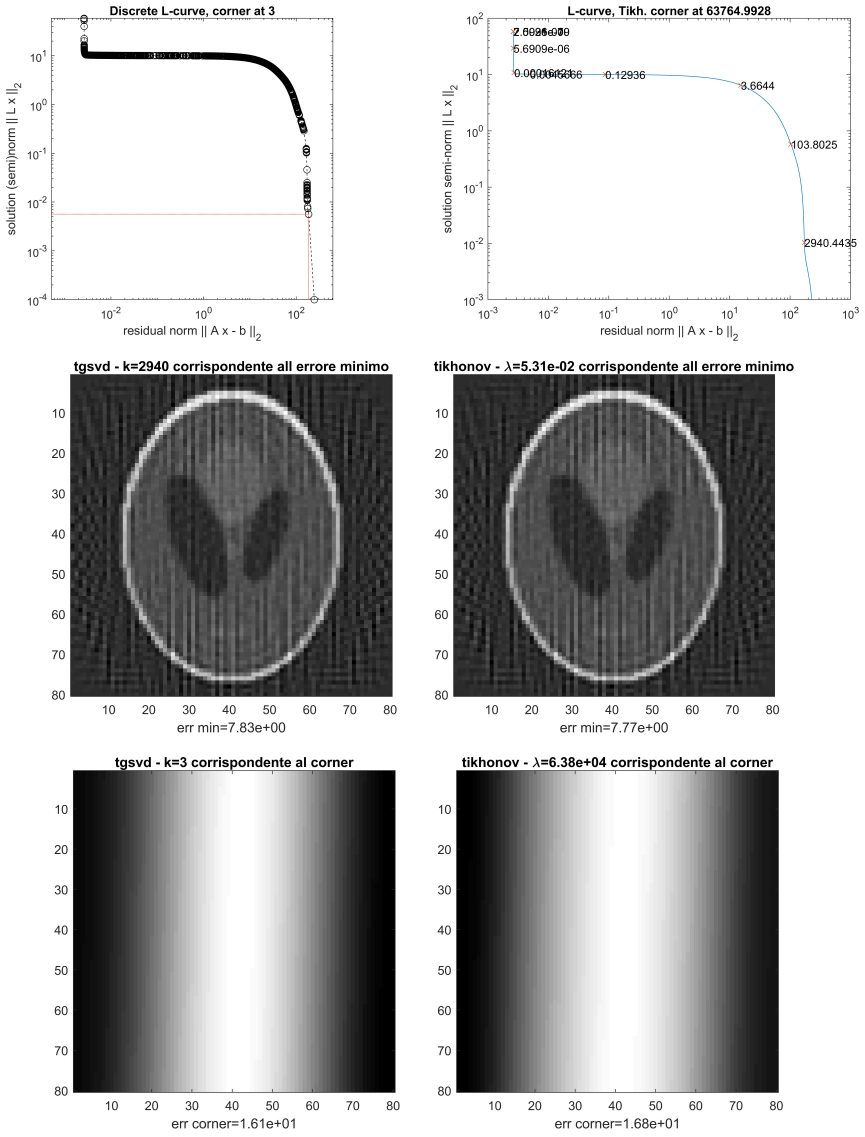

Figura 6.11: Problema test: shepp-logan, N=80, matrice di regolarizzazione D_1, noise= 10^{-4}. Curva L e soluzioni.

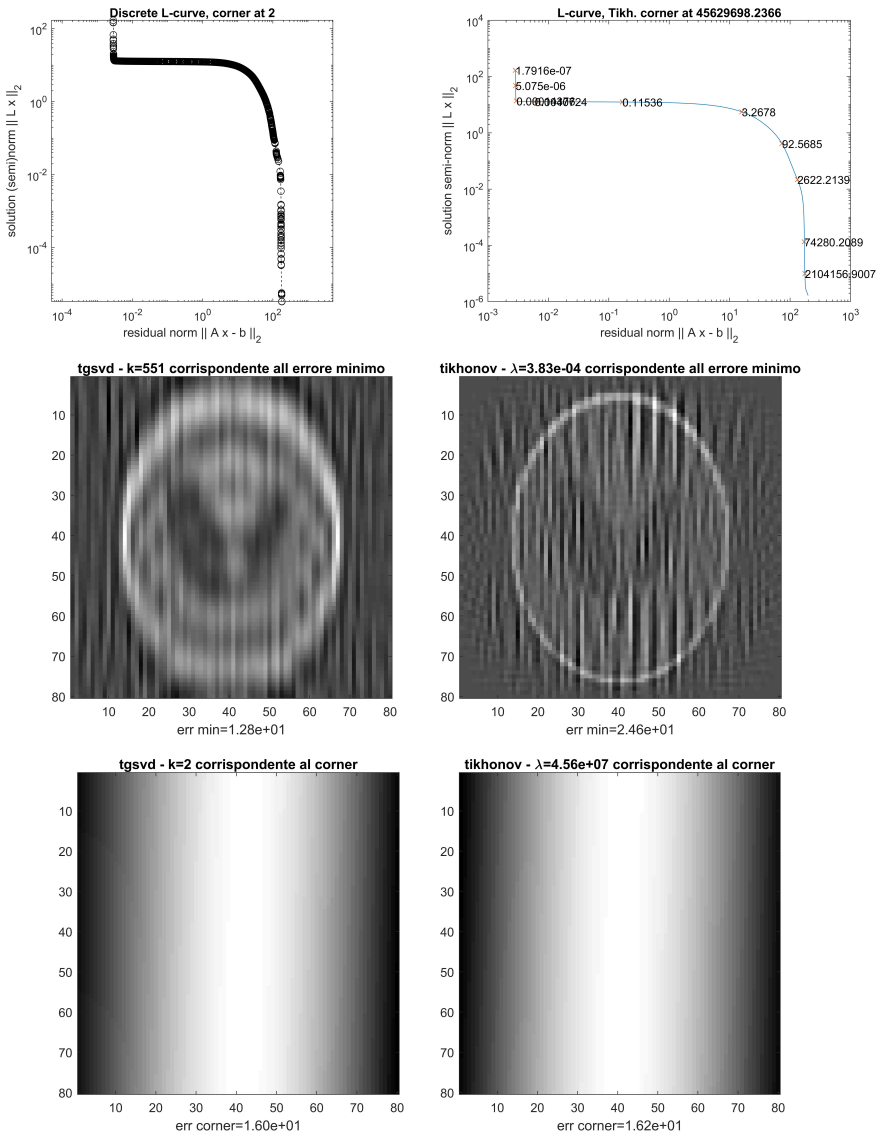

Figura 6.12: Problema test: shepp-logan, N=80, matrice di regolarizzazione D_2, noise= 10^{-4}. Curva L e soluzioni.

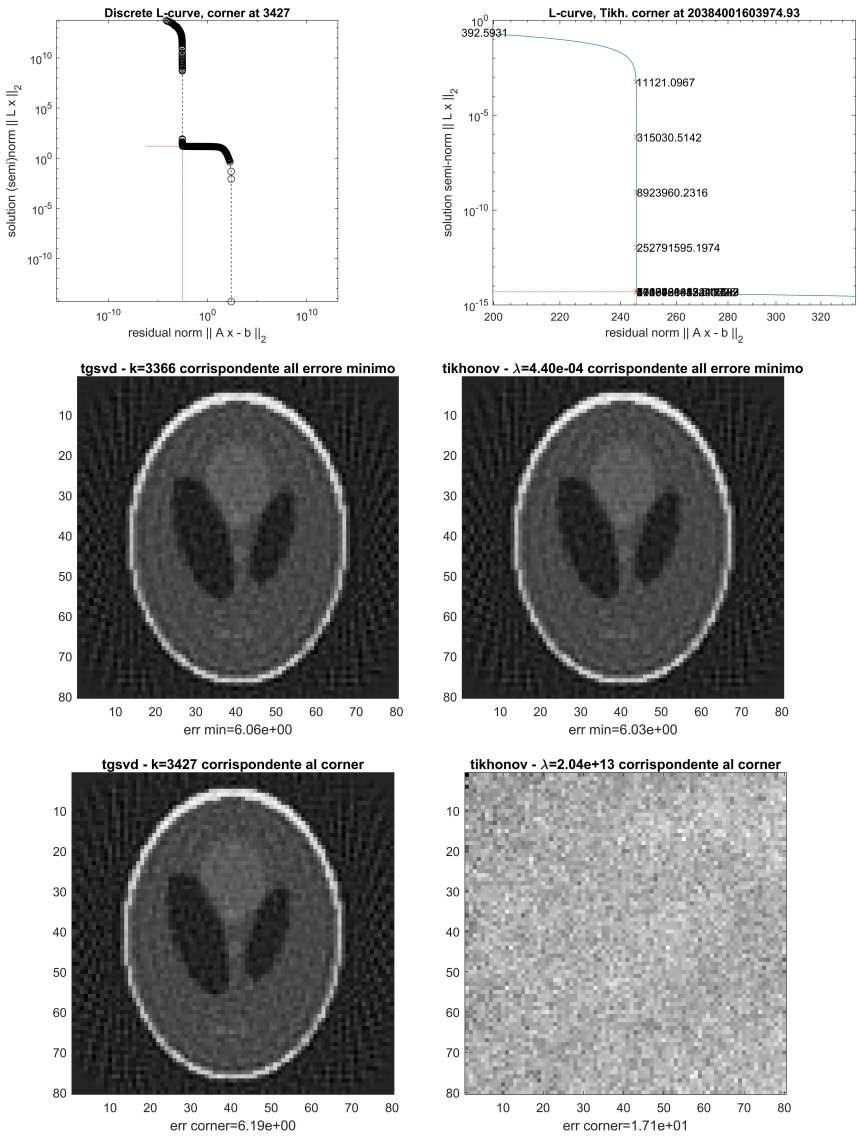

Figura 6.13: Problema test: shepp-logan, N=80, matrice di regolarizzazione L, noise= 10^{-4}. Curva L e soluzioni.

Capitolo 7

Conclusioni

In questa tesi sono stati illustrati dei metodi di regolarizzazione per problemi mal posti. Ai primi quattro capitoli teorici segue un capitolo in cui è stata approfondita l'applicazione alla tomografia computerizzata. Infine sono stati fatti degli esperimenti numerici per verificare l'efficacia di alcuni dei metodi di regolarizzazione e dei criteri di scelta del parametro, in particolare la curva L. Non sempre la curva L ha dato dei risultati soddisfacenti. Invece sono stati ottenuti dei buoni risultati applicando il metodo COSE.

Bibliografia

[1] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[2] P. C. Hansen. *Rank–Deficient and Discrete Ill–Posed Problems*. SIAM, Philadelphia, 1998.

[3] P. C. Hansen. Regularization Tools: version 4.0 for Matlab 7.3. *Numerical Algorithms*, 46:189–194, 2007.

[4] M. E. Hochstenbach, L. Reichel, and G. Rodriguez. Regularization parameter determination for discrete ill–posed problems. *Journal of Computational and Applied Mathematics*, 273:132–149, 2015.

[5] Jennifer L. Mueller and Samuli Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*, volume 10 of *Computational science and engineering*. SIAM, 2012.

[6] Y. Park, L. Reichel, G. Rodriguez, and X. Yu. Parameter determination for Tikhonov regularization problems in general form. *J. Comput. Appl. Math.*, 343:12–25, 2018. DOI 10.1016/j.cam.2018.04.049. Available online.

[7] L. Reichel and G. Rodriguez. Old and new parameter choice rules for discrete ill–posed problems. *Numerical Algorithms*, 63(1):65–87, 2013.

[8] G. Rodriguez. *Algoritmi Numerici*. Pitagora Editrice, Bologna, 2008.