Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática Pura e Aplicada

Métodos de Quadrados Mínimos Totais Regularizados

Jonathan Ruiz Quiroz
Orientador: Prof. Dr. Fermín Sinforiano Viloche Bazán

Florianópolis
Fevereiro de 2014

Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática Pura e Aplicada

Métodos de Quadrados Mínimos Totais Regularizados

Dissertação apresentada ao Curso de Pós-Graduação em Matemática Pura e Aplicada, do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina, para a obtenção do grau de Mestre em Matemática, com Área de Concentração em Matemática Aplicada.

Jonathan Ruiz Quiroz
Florianópolis, Fevereiro de 2014

Métodos de Quadrados Mínimos Totais Regularizados

por

Jonathan Ruiz Quiroz

Esta Dissertação foi julgada para a obtenção do Título de Mestre, en Matemática, Área de Concentração em Matemática Aplicada, e aprovada em sua forma final pelo Curso de Pós-Graduação em Matemática Pura e Aplicada.

Prof. Dr. Daniel Gonçalves
Coordenador

Comissão Examinadora

Prof. Dr. Fermín S. Viloche Bazán
(Orientador - UFSC)

Prof. Dr. Marcelo Vitor Wüst Zibetti (UTFPR)

Prof. Dr. Juliano de Bem Francisco (UFSC)

Prof. Dra. Melissa Weber Mendonça (UFSC)

Florianópolis, Fevereiro de 2014.

Agradecimentos

Primeiro quero agradecer a Deus por tudo o que ele tem feito por mim.

À Meus pais Irma e José para todo o seu amor, exemplo e apoio nesta fase da minha vida.

Ao Meu amor Lila por ter sido minha companheira incondicional durante estes anos de estudo.

Eu também quero agradecer ao Professor Dr. Fermín por ter aceio ser o meu orientador. Suas recomendações foram muito importantes na preparação deste trabalho.

Aos Professores Marcelo, Melissa e Juliano por terem aceito participar da banca, pela leitura e recomendações que melhoraram este trabalho.

Ao Programa de Pós-Graduação em Matemática Pura e Aplicada da UFSC pela oportunidade. À Elisa por sua ajuda.

Ao programa de bolsas Capes pelo auxilio financeiro nestes dois anos.

Resumo

Neste trabalho estudamos métodos de regularização para o problema de Quadrados Mínimos Totais (RTLS) baseado em técnicas da Álgebra Linear Numérica e teoria de regularização.

O foco principal do trabalho é o estudo da regularização de Tikhonov para o método de Quadrados Mínimos Totais (TLS) e de uma técnica de truncamento que atua como regularizador. No primeiro caso, abordamos um método desenvolvido por Renaut e Guo baseado na resolução de um sistema não linear através de um problema de autovalores lineares e sobre o tamanho da solução.

Resultados numéricos mostram que este método pode não funcionar em alguns problemas. Então, estudamos o método TLS truncado (T-TLS) e introduzimos um critério de escolha do parâmetro de truncamento baseado no trabalho de Bazán, Cunha e Borges que não requer informação prévia sobre a solução. Ambos os métodos são ilustrados numericamente e comparados com respeito à qualidade das soluções.

Os resultados numéricos mostram que o método de truncamento é uma boa alternativa para resolver o problema RTLS.

Palavras-chave: SVD, Regularização de Tikhonov, Quadrados mínimos totais, Métodos iterativos.

Abstract

In this paper we study regularization methods for Total Least Squares problems (RTLS) based on Numerical Linear Algebra tools and regularization theory.

The focus of the work is to study the Tikhonov regularization method for Total Least Square (TLS) and a truncation technique which acts as regularization. First, we study a method developed by Renaut and Guo based on linear eigenvalue problems and on a priori information about the size of the solution.

Numerical results show that this method may not work in some problems. Then, we study the truncated TLS method (T-TLS) and introduce a criterion for choosing the truncation parameter based on work by Bazán, Borges and Cunha that does not require any a priori information about the solution. Both methods are illustrated numerically and compared in terms of efficiency and accuracy.

The numerical results show that the truncation method is a good alternative to solve the RTLS problem.

Keywords: SVD, Tikhonov regularization, Total Least Squares, Iterative methods.

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

LS	Quadrados Mínimos
SVD	Decomposição em Valores Singulares
GSVD	Decomposição em Valores Singulares Generalizados
TSVD	Decomposição em Valores Singulares Truncada
TLS	Quadrados Mínimos Totais
MRS	Espectroscopia de Ressonância Magnética
RTLS	Quadrados Mínimos Totais Regularizados
RTLSEVP	Quadrados Mínimos Totais Regularizados via Problema de Autovalores
T-TLS	Mínimos Quadrados Totais Truncados

Notações

As seguintes notações serão usadas neste trabalho.

- O espaço vetorial n-dimensional é denotado por $\mathbb{R}^n$.
- $\mathcal{R}(S)$ denota o espaço coluna, $\mathcal{R}(S^T)$ é o espaço das linhas e $\mathcal{N}(S)$ denota o espaço nulo ou núcleo de S.
- Para as matrizes diagonais, fazemos a seguinte notação. Se $A \in \mathbb{R}^{m\times n}$ escrevemos

  $$A = \text{diag}(\alpha_1, \ldots, \alpha_p), \quad p = \min\{m, n\},$$

  então $a_{ij} = 0$, se $i \neq j$ e $a_{ii} = \alpha_i$ para $i = 1, \ldots, p$.
- A matriz identidade $m \times m$ é denotada por I_m.
- A norma Frobenius de uma matriz $m \times n$ M é definida por

  $$\|M\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} m_{ij}^2}.$$

- A norma 2 de um vetor n-dimensional é definida por

  $$\|y\|_2 = \sqrt{\sum_{i=1}^{n} y_i^2}.$$

- Definimos a norma 2 de uma matriz $m \times n$ M por

  $$\|M\|_2 = \sup_{y\neq 0} \frac{\|My\|_2}{\|y\|_2}.$$

  e é igual ao maior valor singular de M.

- Posto(A) denota o posto da matriz A, definida como o número de valores singulares não nulos.
- O menor valor singular da matriz A é denotado por $\sigma_{min}(A)$.
- A_k denota uma matriz de posto k. A matriz $[C_k \quad D_k]$ denota uma matriz aumentada de posto k.
- $A^{\dagger}$ denota a pseudo-inversa de A.
- O conjunto $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$, denota uma sequência de números reais, e denotamos por $\{x^{(k)}\}_{k\in\mathbb{N}}$ a sequência de vetores em $\mathbb{R}^n$.

Sumário

Introdução

No modelo clássico de quadrados mínimos (LS)

$$\min \|Ax - b\|_2$$

onde $A \in \mathbb{R}^{m\times n}$, $m \geq n$, e $b \in \mathbb{R}^m$, é assumido frequentemente que a matriz A é exata e o vetor b é contaminado por erros. Esta hipótese não é sempre realista, pois erros de medição, de modelagem, de instrumentação, também acrescentam incertezas na matriz de dados A.

Uma maneira de lidar com problemas desta natureza é através do método de Quadrados Mínimos Totais (TLS) que é um método apropriado quando existem perturbações na matriz de dados A e no vetor de observações b. Neste trabalho fazemos um estudo da regularização de Tikhonov para o método TLS.

O problema foi estudado por Golub e Van Loan [11], teoricamente e algoritmicamente, e fortemente baseado na decomposição em valores singulares. Nos últimos anos o interesse no método TLS é mantido devido ao desenvolvimento da eficiência computacional e algoritmos confiáveis para resolver este problema.

É necessário comentar que o TLS é uma das muitas técnicas de ajuste de estimativas dos dados, quando as variáveis estão sujeitas a erros. Muitas outras aproximações gerais para este problema guiaram a outras técnicas de ajuste para problemas lineares, assim como não lineares.

Na prática também aparecem problemas nos quais a matriz A é mal condicionada e a solução tem que satisfazer uma condição quadrática $\|Lx\|_2 \leq \delta$, onde $L \in \mathbb{R}^{p\times n}$ e $\delta > 0$. Estes casos aparecem naturalmente nos problemas inversos, onde queremos, por exemplo, estudar a estrutura de um problema físico a partir de seu comportamento. Um dos métodos importantes para resolver problemas mal condicionados é a regularização de Tikhonov [36], o qual incorpora hipóteses adicionais sobre o tamanho e a suavidade da solução desejada que ajuda a

contornar a sensibilidade da matriz A sobre a solução. Para problemas discretos, a regularização de Tikhonov na forma geral leva ao problema de minimização

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 + \lambda\|Lx\|_2^2$$

onde o parâmetro de regularização $\lambda > 0$ controla o peso, dado pelo termo de regularização $\|Lx\|_2$ relativo à minimização da norma residual [36].

Em nosso trabalho estudamos o caso no qual a matriz A e b estão sujeitos a perturbações, introduzindo técnicas que foram desenvolvidas previamente ao problema TLS bem como aplicações do método de quadrados mínimos totais. O presente trabalho está organizado como segue.

No primeiro capítulo estudamos o método de quadrados mínimos e a decomposição em valores singulares, junto com as propriedades e os resultados teóricos importantes desta decomposição, tal como o Teorema de Eckart-Young sobre aproximações duma matriz, que será muito usado neste trabalho.

No capítulo 2 definimos o problema TLS, introduzindo seus fundamentos teóricos tais como teoremas de existência e unicidade da solução. Além disso, também estudamos a relação entre as soluções do método LS e o método TLS usando o ângulo entre subespaços e finalmente comparamos os métodos LS e TLS para o problema de ressonância magnética.

No capítulo 3 introduzimos a regularização de Tikhonov para o método TLS baseado em [14], estudando as propriedades mais importantes do método, para depois usá-las em conexão com o método TLS regularizado (RTLS), incluindo os casos $L = I$ e $L \neq I$. O capítulo também considera uma outra forma de regularização baseada na idéia de truncamento do método SVD truncada (TSVD) e inclui um estudo do método de bidiagonalização de Lanczos que é útil para problemas com dimensões muito grandes.

No capítulo 4 estudaremos um método iterativo para calcular a solução aproximada do problema RTLS baseado em autovalores e denotado por RTLSEVP. Este algoritmo será usado para resolver alguns problemas teste extraídos de [17] e os resultados serão comparados com o método de truncamento baseado na teoria desenvolvida em [7], o qual não requer informações adicionais sobre o tamanho e a suavidade da solução desejada.

Capítulo 1

Preliminares

Neste capítulo apresentamos algumas idéias básicas que servem como motivação e suporte teórico para o tema central deste trabalho.

1.1 Problema de Quadrados Mínimos (LS)

A solução numérica de equações integrais de primeira espécie

$$\int_a^b K(x,y)f(y)dy = g(x), \quad c \leq x \leq d \tag{1.1.1}$$

requer o uso de métodos de discretização, que em geral, resultam em problemas de minimização

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2 \tag{1.1.2}$$

onde $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ onde $m \geq n$. Neste capítulo apresentamos os princípios do problema (1.1.2), chamado neste trabalho de problema LS. Van Huffel e Vandewalle [37] definiram um problema LS como básico quando são satisfeitas as condições abaixo.

- O lado direito b é um único vetor em $\mathbb{R}^m$.
- O problema tem solução.
- A solução é única.

Em certos problemas, ambos a matriz A e o lado direito b estão sujeitos a incertezas e uma maneira de lidar com estes problemas é através

do método de Quadrados Mínimos Totais (TLS). Neste capítulo vamos desenvolver a teoria necessária para o estudo desse método. Considere o problema de encontrar um vetor $x \in \mathbb{R}^n$ tal que $Ax = b$, onde $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^m$ é o vetor de observações. Quando existem mais equações do que incógnitas, ou seja, $m > n$, o sistema é sobredeterminado e o problema não tem solução se $b \notin \mathcal{R}(A)$. Neste caso usamos a notacão $Ax \approx b$.

A solução x do problema (1.1.2) é caracterizada pelo seguinte Teorema

Teorema 1.1.1. *O vetor x é solução do problema (1.1.2) se e somente se*

$$A^T(b - Ax) = 0.$$

Demonstração. Ver [3]. □

Este Teorema diz que o resíduo $r = b - Ax$ associado à solução x é ortogonal ao $\mathcal{R}(A)$, como mostra a Figura 1.1.

Então o lado direito é decomposto em duas componentes ortogonais.

$$b = b' + r = Ax + r, \quad r \perp Ax,$$

onde b' é a projeção ortogonal de b sobre $\mathcal{R}(A)$.

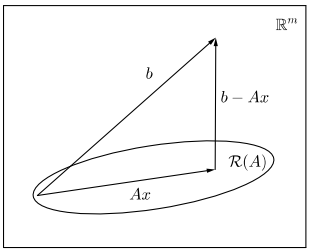

Figura 1.1: Descrição geométrica do Teorema 1.1.1.

Corolário 1.1.1. *(Solução LS e Resíduo) Se* posto$(A) = n$*, então (1.1.2) tem solução LS única dada por*

$$x = (A^T A)^{-1} A^T b \tag{1.1.3}$$

e a correspondente correção LS é dado pelo resíduo.

$$r = b - Ax = b - b', \quad b' = P_A(b)$$

onde $P_A = A(A^TA)^{-1}A^T$ *é a projeção ortogonal sobre* $\mathcal{R}(A)$.

Se posto$(A) < n$, o problema LS (1.1.2) tem infinitas soluções, pois se x é solução do (1.1.2) e $z \in \mathcal{N}(A)$ então $x+z$ é também uma solução. Denotamos por x_{LS} a solução do problema (1.1.2). Note que se A tem posto completo, existe uma única solução.

Na seguinte seção vamos encontrar expressões para a solução x_{LS} e o resíduo $\|Ax_{LS} - b\|_2$. em termos da SVD.

1.2 A Decomposição em Valores Singulares

A decomposição em valores singulares (SVD) de uma matriz $A \in \mathbb{R}^{m\times n}$ é de muita importância teórica e prática em Álgebra Linear Numérica como vamos verificar neste trabalho.

Teorema 1.2.1. *(SVD) Seja* $A \in \mathbb{R}^{m\times n}$ *com* $m \geq n$ *então existem matrizes ortonormais* $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$ *e* $V = [v_1, \ldots, v_m] \in \mathbb{R}^{n\times n}$ *tais que*

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T \tag{1.2.1}$$

onde $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ *possui elementos não negativos tais que*

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0.$$

Demonstração. O Teorema pode ser encontrado em [16]. □

Os números σ_i são chamados valores singulares de A, e os vetores u_i e v_i são chamados vetores singulares à esquerda e à direita de A, respectivamente.

Geometricamente, a SVD de A fornece duas bases de vetores ortonormais (as colunas de U e V) tais que a matriz A é diagonal quando é transformada nessas bases.

É fácil verificar pela comparação das colunas nas equações $AV = U\Sigma$ e $A^TU = \Sigma^TV$ que

$$Av_i = \sigma_i u_i \quad \text{e} \quad A^Tu_i = \sigma_i v_i, \quad i = 1, \ldots, n. \tag{1.2.2}$$

A SVD também mostra informação sobre a estrutura de A. Se a SVD de A é dada pelo Teorema 1.2.1, e

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0,$$

então

$$\begin{aligned} \text{posto}(A) &= r, \\ \mathcal{R}(A) &= \mathcal{R}([u_1, \ldots, u_r]), \\ \mathcal{N}(A) &= \mathcal{R}([v_{r+1}, \ldots, v_n]), \\ \mathcal{R}(A^T) &= \mathcal{R}([v_1, \ldots, v_r]), \\ \mathcal{N}(A^T) &= \mathcal{R}([u_{r+1}, \ldots, u_m]). \end{aligned}$$

A norma 2 e a norma Frobenius da matriz A são caracterizadas em termos da SVD:

$$\|A\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2 = \sigma_1^2 + \ldots + \sigma_n^2,$$

$$\|A\|_2 = \sup_{y \neq 0} \frac{\|Ay\|_2}{\|y\|_2} = \sigma_1.$$

A SVD permite definir o número de condição da matriz A.

Definição 1.2.1. *Seja $A \in \mathbb{R}^{m\times n}$ com posto r, consideremos a SVD de A dada por (1.2.1). O número de condição de A é definido como*

$$\kappa(A) = \|A\|_2\|A^\dagger\|_2 = \frac{\sigma_1}{\sigma_r}, \tag{1.2.3}$$

onde $A^\dagger$ denota a pseudo-inversa geralizada de Moore-Penrose de A.

Note que se A é inversível a inversa é dada pela expressão

$$A^{-1} = \sum_{i=1}^{n} \sigma_i^{-1} v_i u_i^T,$$

caso contrário, a pseudo-inversa $A^\dagger$ é definida por

$$A^\dagger = V\Sigma^\dagger U^T = \sum_{i=1}^{r} \sigma_i^{-1} v_i u_i^T,$$

em que $\Sigma^\dagger = \text{diag}(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0)$ com $\Sigma^\dagger \in \mathbb{R}^{n\times m}$.

A pseudo-inversa da matriz A satisfaz as quatro condições de Moore-Penrose:

(i) $AA^{\dagger}A = A$.

(ii) $A^{\dagger}AA^{\dagger} = A^{\dagger}$.

(iii) $(AA^{\dagger})^T = AA^{\dagger}$.

(iv) $(A^{\dagger}A)^T = A^{\dagger}A$.

1.2.1 Problemas Discretos Mal Postos

O problema (1.1.2) é dito problema discreto mal posto quando os valores singulares da matriz decaem para zero, i.e., apresentam um número de condição elevado. Outra característica deste tipo de problemas é que os elementos dos vetores singulares à esquerda e à direita u_i e v_i têm grandes mudanças de sinal quando o índice i cresce, com mostra a Figura 1.2.

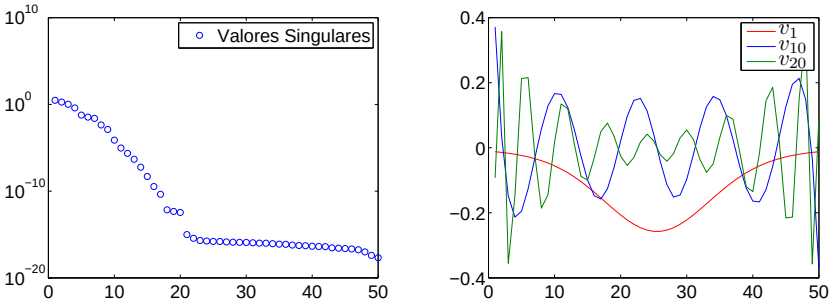

Figura 1.2: Valores e vetores singulares.

Quando $b \in \mathbb{R}^m$ está contaminado por erros, ou seja, $b = b_{exato} + e$, com b_{exato} sendo o vetor sem perturbações desejado e desconhecido, a solução de quadrados mínimos, $x_{LS} = A^{\dagger}b$, não tem nenhuma relação com a solução exata do problema e nem tem utilidade prática por estar completamente dominada pelos erros.

A SVD é uma poderosa ferramenta computacional para resolver problemas LS. A razão disso é que as matrizes ortogonais que transformam a matriz A numa matriz diagonal (1.2.1) não mudam a norma 2 dos vetores. O seguinte Teorema expressa a solução LS usando a decomposição SVD.

Teorema 1.2.2. *(Solução de norma mínima LS de $Ax \approx b$). Seja (1.2.1) a SVD de $A \in \mathbb{R}^{m\times n}$, i.e., $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$, e assumimos que*

posto$(A) = r$. *Se* $b \in \mathbb{R}^m$, *então o vetor*

$$x_{LS} = \sum_{i=1}^{r} \sigma_i^{-1} v_i u_i^T b \tag{1.2.4}$$

minimiza $\|Ax - b\|_2$ *e é o minimizador de norma mínima. Além disso*

$$\rho^2 = \|Ax_{LS} - b\|_2^2 = \sum_{i=r+1}^{m} (u_i^T b)^2. \tag{1.2.5}$$

Demonstração. Ver [13]. □

No problema de minimização (1.1.2) se o vetor de dados é da forma $b = b_{exato} + e$, com b_{exato} o vetor sem pertubações e e o vetor de incertezas, usando a SVD obtemos

$$x_{LS} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i, \ r = \text{posto}(A). \tag{1.2.6}$$

a solução do problema (1.1.2). Sendo $b = b_{exato} + e$ temos

$$x_{LS} = \sum_{i=1}^{r} \left(\frac{u_i^T b_{exato}}{\sigma_i} v_i + \frac{u_i^T e}{\sigma_i} v_i \right). \tag{1.2.7}$$

Devido à divisão por pequenos valores singulares, os coeficientes $\frac{u_i^T e}{\sigma_i}$ são grandes, fazendo com que a parcela do erro seja dominante, tornando assim esta abordagem inútil. Portanto, é necessário estabilizar a solução. Uma maneira de amenizar o efeito da influência do erro na solução é truncando a soma em (1.2.6) para $s < r$ termos:

$$x_{LS}^s = \sum_{i=1}^{s} \frac{u_i^T b}{\sigma_i} v_i, \tag{1.2.8}$$

onde s, chamado índice de truncamento, é escolhido de modo que exista um balanço apropriado entre a qualidade da informação do problema que é capturada e a quantidade de erro que é incluída na solução. Este método é conhecido como o método da SVD Truncada (TSVD) [16].

Para ilustrar o método TSVD, na Figura 1.3 consideramos duas soluções numéricas de um sistema $Ax = b$, 64×64, que provém da discretização de uma equação integral de Fredholm de primeira espécie, chamada de shaw [16]. A parte esquerda mostra a solução calculada

usando a função inversa de MATLAB `x=inv(A)*b`. O lado direito mostra a solução TSVD (linha verde) obtida pela retenção de 7 componentes associadas aos maiores valores singulares, junto com a solução exata (linha azul). Vemos que a solução dada pela inversão de A tem muitas oscilações de grande amplitude.

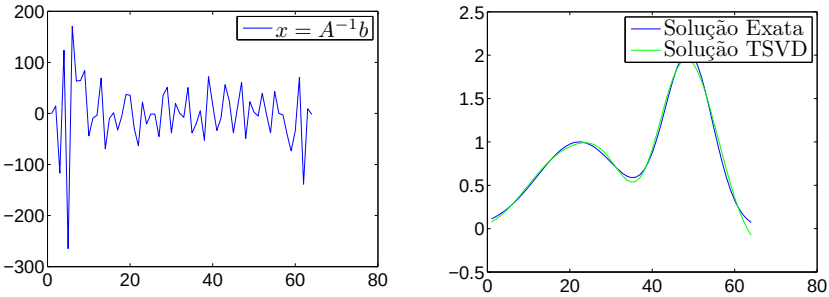

Figura 1.3: Soluções de um sistema discreto $Ax = b$.

Por outro lado a solução TSVD apresenta um bom balanço entre o erro relativo $\|x_{exata} - x_{TSVD}\|_2/\|x_{exata}\|_2 = 0.0475$ e a norma residual relativa $\|Ax_{TSVD} - b\|_2/\|b\|_2 = 2.2042 \times 10^{-5}$, pois em algúns casos temos um resíduo relativo pequeno e um erro relativo grande. O número de condição da matriz A da Figura 1.3 é $\kappa(A) = 4.0583 \times 10^{20}$ o qual indica que a matriz A tem alta sensibilidade a erros no vetor de dados.

Matriz com número de condição grande é chamada mal condicionada. No problema de equações lineares, o número de condição (1.2.3) mede a sensibilidade da solução aos erros da matriz A e o lado direito b. Isto mostra que pequenas mudanças nas entradas de A, podem produzir grandes variações na solução x de $Ax \approx b$ quando as colunas são quase dependentes, i.e. $\sigma_n \approx 0$.

A expressão (1.2.8) pode ser escrita como

$$x_{LS}^s = \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i$$

onde

$$f_i = \begin{cases} 1, & \text{se } 1 \leq i \leq s \\ 0, & \text{se } s < i \leq n \end{cases}$$

Os coeficientes f_i são chamados fatores de filtro da solução (1.2.8). Mais adiante veremos que existem outras fórmulas para esses valores.

A SVD cumpre um papel importante em problemas de aproximação de matrizes, como mostra o seguinte Teorema.

Teorema 1.2.3. *(Eckart-Young-Mirsky) Seja a matriz* $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ *com* $\text{posto}(A) = r$. *Considere a SVD de* A

$$A = U\Sigma V^T.$$

Então se $k < r$

$$\min_{\text{posto}(B)=k} \|A - B\|_2$$

é atingido em A_k *onde*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \quad \text{e} \quad \|A - A_k\|_2 = \sigma_{k+1}. \tag{1.2.9}$$

Demonstração. Desde que $U^T A_k V = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ temos $\text{posto}(A_k) = k$ e também

$$U^T(A - A_k)V = \text{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r)$$

portanto $\|A - A_k\|_2 = \sigma_{k+1}$.

Suponha que $\text{posto}(B) = k$, para algum $B \in \mathbb{R}^{m\times n}$. Então podemos encontrar uma base ortonormal de vetores $x_1, \ldots, x_{n-k}$ tal que $\mathcal{N}(B) = \text{span}\{x_1, \ldots, x_{n-k}\}$. Usando a dimensão tem-se

$$\text{span}\{x_1, \ldots, x_{n-k}\} \cap \text{span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}.$$

Seja z um vetor unitário na norma 2 nesta interseção. Como $Bz = 0$ e

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$$

temos

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2$$

o que completa a prova do Teorema. □

Observação: O Teorema 1.2.3 foi provado inicialmente para a norma Frobenius [9], para esta norma temos

$$\min_{\text{posto}(B)=k} \|A - B\|_F = \|A - A_k\|_F = (\sigma_{k+1} + \ldots + \sigma_r)^{1/2}.$$

Este Teorema será de muita utilidade nas seguintes seções, pois permite trocar um sistema $Ax \approx b$ por um sistema aproximado $A_k x \approx b_k$.

O seguinte Teorema mostra a sensibilidade dos valores singulares.

Teorema 1.2.4. *Sejam A e $\widetilde{A} = A+E \in \mathbb{R}^{m\times n}$, $m \geq n$ com os valores singulares $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ e $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \ldots \geq \widetilde{\sigma}_n$, respectivamente. Então*

$$|\sigma_i - \widetilde{\sigma}_i| \leq \|E\|_2,$$

$$\sum_{i=1}^{n} |\sigma_i - \widetilde{\sigma}_i| \leq \|E\|_F^2.$$

Demonstração. Ver [3]. □

O seguinte Teorema apresenta a propriedade de interlacing para valores singulares e mostra o que acontece quando são removidas algumas linhas ou colunas da matriz.

Teorema 1.2.5. *(Propriedade de interlacing dos valores singulares) Seja $C \in \mathbb{R}^{m\times n}$ com valores singulares $\bar{\sigma}_1 \geq \ldots \geq \bar{\sigma}_{\min\{m,n\}}$. Seja $D \in \mathbb{R}^{p\times q}$ uma submatriz de C, com valores singulares $\sigma_1 \geq \ldots \geq \sigma_{\min\{p,q\}}$, e definimos por conveniência $\bar{\sigma}_t = 0$ para $\min\{m,n\} < t \leq \max\{m,n\}$ e $\sigma_t = 0$ para $\min\{p,q\} < t \leq \max\{p,q\}$. Então*

(1) $\bar{\sigma}_i \geq \sigma_i$ para $i = 1, \ldots, \min\{p,q\}$.

(2) $\sigma_i \geq \bar{\sigma}_{i+(m-p)+(n-q)}$ para $i \leq \min\{p+q-m, p+q-n\}$.

Demonstração. Ver [35]. □

Se D é o resultado de remover uma coluna de C então o Teorema 1.2.5 diz que

(1) Se $m \geq n$: $\bar{\sigma}_1 \geq \sigma_1 \geq \bar{\sigma}_2 \geq \sigma_2 \geq \ldots \geq \sigma_{n-1} \geq \bar{\sigma}_n \geq 0$.

(2) Se $m < n$: $\bar{\sigma}_1 \geq \sigma_1 \geq \bar{\sigma}_2 \geq \sigma_2 \geq \ldots \geq \bar{\sigma}_m \geq \sigma_m \geq 0$.

Teorema 1.2.6. *(Teorema de Interlacing de Cauchy). Seja A uma matriz Hermitiana de dimensão n, e seja B a submatriz principal de ordem $n-1$. Se $\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$ são os autovalores de A e $\mu_{n-1} \leq \ldots \leq \mu_2 \leq \mu_1$ são os autovalores de B, então*

$$\lambda_n \leq \mu_{n-1} \leq \lambda_{n-1} \leq \mu_{n-2} \leq \ldots \leq \lambda_2 \leq \mu_1 \leq \lambda_1.$$

Demonstração. Ver [39]. □

Capítulo 2

Método de Quadrados Mínimos Totais

O problema de quadrados mínimos totais foi introduzido na literatura por Golub e Van Loan [11] e Van Huffel e Vandewalle [37], como uma alternativa para o problema de quadrados mínimos, no caso em que ambos a matriz A e o vetor b contém incertezas. Uma forma de contornar os erros em A é introduzindo um termo corretor $\Delta\widetilde{A}$.

2.1 Princípios do Método TLS

Definição 2.1.1. *Seja o sistema linear de m equações lineares $Ax \approx b$ com n variáveis x. Considere o problema*

$$\min_{[\widetilde{A}\ \widetilde{b}]\in\mathbb{R}^{m\times(n+1)}} \|[A \quad b] - [\widetilde{A} \quad \widetilde{b}]\|_F^2 \tag{2.1.1}$$

$$\text{sujeito a} \quad \widetilde{b} \in \mathcal{R}(\widetilde{A}). \tag{2.1.2}$$

Quando a solução $[\widetilde{A} \quad \widetilde{b}]$ é encontrada, então qualquer x que satisfaz

$$\widetilde{A}x = \widetilde{b} \tag{2.1.3}$$

é chamado solução TLS e é denotado por x_{TLS}. $[\Delta\widetilde{A} \quad \Delta\widetilde{b}] = [A \quad b] - [\widetilde{A} \quad \widetilde{b}]$ é o correspondente corretor TLS.

A análise do problema TLS depende fortemente do uso da SVD da matriz aumentada $[A \quad b]$ e começa com a observação que o sistema

$Ax = b$ pode ser reescrito como

$$[A \quad b] \begin{bmatrix} x \\ -1 \end{bmatrix} = 0. \tag{2.1.4}$$

Sejam

$$A = U\ \Sigma\ V^{\mathbf{T}}, \tag{2.1.5}$$

e

$$[A \quad b] = \overline{U}\ \overline{\Sigma}\ \overline{V}^{\mathbf{T}}, \tag{2.1.6}$$

as decomposições em valores singulares das matrizes A e $[A \quad b]$ respectivamente. Uma consequência imediata é que se $\overline{\sigma}_{n+1} \neq 0$, então posto$([A \quad b]) = n + 1$ e o espaço nulo da matriz ampliada é trivial, logo o conjunto de equações (2.1.4) não é compatível. Daí, para obter uma solução, o posto da matriz aumentada $[A \quad b]$ deve ser reduzido de $n + 1$ para n. Para isso, decomponha a matriz $\overline{V}$ como

$$\overline{V} = [\overline{v}_1, \ldots, \overline{v}_{n+1}] = \begin{bmatrix} \overline{V}_{11} & \overline{v}_{12} \\ \overline{v}_{21}^T & \overline{v}_{22} \end{bmatrix}, \quad \overline{\Sigma} = \begin{bmatrix} \overline{\Sigma}_1 & 0 \\ 0 & \overline{\sigma}_{n+1} \\ 0 & 0 \end{bmatrix}, \tag{2.1.7}$$

em que $\overline{V}_{11}, \overline{\Sigma}_1 \in \mathbb{R}^{n\times n}$ e $\overline{v}_{12}, \overline{v}_{21} \in \mathbb{R}^n$. Usando o Teorema 1.2.3, a melhor aproximação de posto n no sentido da norma Frobenius da matriz $[A \quad b]$ é

$$[\widetilde{A}_n \quad \widetilde{b}_n] = \sum_{i=1}^{n} \overline{\sigma}_i \overline{u}_i \overline{v}_i^T.$$

O corretor TLS minimal é

$$\overline{\sigma}_{n+1} = \min_{posto([\widetilde{A}_n \ \widetilde{b}_n])=n} \|[A \quad b] - [\widetilde{A}_n \quad \widetilde{b}_n]\|_F$$

e é atingido quando

$$[A \quad b] - [\widetilde{A}_n \quad \widetilde{b}_n] = [\Delta\widetilde{A} \quad \Delta\widetilde{b}] = \overline{\sigma}_{n+1}\overline{u}_{n+1}\overline{v}_{n+1}^T.$$

Note que a matriz corretora TLS tem posto um. É claro que o sistema aproximado

$$[\widetilde{A}_n \quad \widetilde{b}_n] \begin{bmatrix} x \\ -1 \end{bmatrix} = 0.$$

é compatível, além disso usando o fato de que $\overline{v}_{n+1} \in \mathcal{N}([\widetilde{A}_n \quad \widetilde{b}_n])$, se $[\overline{v}_{n+1}]_{n+1} \neq 0$, temos que

$$\begin{bmatrix} x \\ -1 \end{bmatrix} = -\frac{1}{[\overline{v}_{n+1}]_{n+1}} \overline{v}_{n+1}. \tag{2.1.8}$$

Portanto, usando (2.1.7) temos que a solução TLS pode ser expressa como:

$$x_{TLS} = -\frac{\overline{v}_{12}}{\overline{v}_{22}}. \tag{2.1.9}$$

Note que se $\overline{\sigma}_{n+1} = 0$, então $[A \quad b]$ tem posto n. Neste caso o sistema é compatível e não precisamos aproximar a matriz ampliada. Também, se $\overline{\sigma}_{n+1}$ é um valor singular simples (não repetido), temos que $\mathcal{N}([\widetilde{A}_n \quad \widetilde{b}_n]) = span\{\overline{v}_{n+1}\}$ e a solução TLS é única. O seguinte Teorema descreve as condições para a existência e unicidade da solução TLS.

Teorema 2.1.1. *(Solução do problema TLS básico $Ax \approx b$) Sejam (2.1.5) a SVD de A e (2.1.6) a SVD de $[A \quad b]$ respectivamente. Se $\sigma_n > \overline{\sigma}_{n+1} > 0$, então*

$$[\widetilde{A}_n \quad \widetilde{b}_n] = \sum_{i=1}^{n} \overline{\sigma}_i \overline{u}_i \overline{v}_i^T \tag{2.1.10}$$

é uma solução do problema (2.1.1) e a solução x_{TLS} é única.

Demonstração. O Teorema de interlacing (1.2.5) para valores singulares implica que

$$\bar{\sigma}_1 \geq \sigma_1 \geq \ldots \bar{\sigma}_n \geq \sigma_n \geq \bar{\sigma}_{n+1}.$$

A hipótese $\sigma_n > \bar{\sigma}_{n+1}$ garante que $\bar{\sigma}_{n+1}$ não é um valor singular repetido de $[A \quad b]$. Agora note que se $[A \quad b]^T[A \quad b][y^T \; 0]^T = \bar{\sigma}_{n+1}^2[y^T \; 0]^T$ e $0 \neq y \in \mathbb{R}^n$, segue que $A^TAy = \bar{\sigma}_{n+1}^2 y$ ou seja $\bar{\sigma}_{n+1}^2$ é autovalor de A^TA. Isto é uma contradição pois σ_n^2 é o menor autovalor de A^TA.

Portanto, $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$ contém um vetor cuja $(n+1)$-ésima componente é não nula, logo o problema TLS tem solução. Como $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$ tem dimensão um, esta solução é unica. As igualdades (2.1.9), (2.1.10) seguem diretamente da aplicação do Teorema de Eckart-Young-Mirsky 1.2.3 como foi provado acima. □

É interessante notar a equivalência das condições

$$\sigma_n > \bar{\sigma}_{n+1} \Leftrightarrow \bar{\sigma}_n > \bar{\sigma}_{n+1} \quad \text{e} \quad [\bar{v}_{n+1}]_{n+1} \neq 0.$$

Uma útil e conhecida caracterização da solução x_{TLS} e do corretor minimal TLS é dada no seguinte Teorema.

Teorema 2.1.2. *(Expressão Fechada da Solução Básica TLS) Sejam (2.1.5) a SVD de A e (2.1.6) a SVD de $[A \quad b]$ respectivamente. Se $\sigma_n > \overline{\sigma}_{n+1}$, então*

$$x_{TLS} = (A^TA - \bar{\sigma}_{n+1}^2 I)^{-1} A^T b \tag{2.1.11}$$

e

$$\bar{\sigma}_{n+1}^2\left(1+\sum_{i=1}^{n}\frac{(u_i^Tb)^2}{\sigma_i^2-\overline{\sigma}_{n+1}^2}\right)=\rho^2=\min_{x\in\mathbb{R}^n}\|Ax-b\|_2^2. \tag{2.1.12}$$

Demonstração. A condição $\sigma_n > \overline{\sigma}_{n+1}$ garante que x_{TLS} existe e é única dada por (2.1.9). Como os vetores singulares v_i são autovetores de $[A \quad b]^T[A \quad b]$, x_{TLS} também satisfaz a seguinte equação de autovetores:

$$\begin{aligned}[A \quad b]^T[A \quad b]\begin{bmatrix}x_{TLS}\\-1\end{bmatrix}&=\begin{bmatrix}A^TA & A^Tb\\ b^TA & b^Tb\end{bmatrix}\begin{bmatrix}x_{TLS}\\-1\end{bmatrix}\\&=\bar{\sigma}_{n+1}^2\begin{bmatrix}x_{TLS}\\-1\end{bmatrix}.\end{aligned} \tag{2.1.13}$$

A igualdade (2.1.11) segue da parte superior de (2.1.13). Para obter (2.1.12) usamos a SVD de A, assim temos

$$\begin{bmatrix}\Sigma^T\Sigma & g\\ g^T & \|b\|_2^2\end{bmatrix}\begin{bmatrix}z\\-1\end{bmatrix}=\bar{\sigma}_{n+1}^2\begin{bmatrix}z\\-1\end{bmatrix},$$

onde $g=\Sigma^TU^Tb$, $z=V^Tx_{TLS}$. Desta equação vemos que

$$(\Sigma^T\Sigma-\bar{\sigma}_{n+1}^2I)z=g \quad \text{e} \quad \bar{\sigma}_{n+1}^2+g^Tz=\|b\|_2^2.$$

Substituindo z na última expressão temos

$$\bar{\sigma}_{n+1}^2+g^T(\Sigma^T\Sigma-\bar{\sigma}_{n+1}^2I)^{-1}g=\|b\|_2^2.$$

Isto pode ser reescrito como

$$\bar{\sigma}_{n+1}^2+\sum_{i=1}^{n}\frac{\sigma_i^2(u_i^Tb)^2}{\sigma_i^2-\bar{\sigma}_{n+1}^2}=\sum_{i=1}^{m}(u_i^Tb)^2$$

ou

$$\bar{\sigma}_{n+1}^2\left[1+\sum_{i=1}^{n}\frac{(u_i^Tb)^2}{\sigma_i^2-\bar{\sigma}_{n+1}^2}\right]=\sum_{i=n+1}^{m}(u_i^Tb)^2.$$

A igualdade (2.1.13) segue já que

$$\min_x\|b-Ax\|_2=\min_w\|U^Tb-\Sigma w\|_2^2=\sum_{i=n+1}^{m}(u_i^Tb)^2.$$

□

Corolário 2.1.1. *Sejam (2.1.5) a SVD de A e (2.1.6) a SVD de* $[A \quad b]$ *respectivamente. Se* $\sigma_n > \overline{\sigma}_{n+1}$*, então*

$$x_{TLS} = (I - \bar{\sigma}_{n+1}^2 (A^T A)^{-1}) x_{LS} = (I + \bar{\sigma}_{n+1}^2 (A^T A - \bar{\sigma}_{n+1}^2 I)^{-1}) x_{LS}$$

O seguinte resultado mostra uma relação das soluções LS e TLS.

Corolário 2.1.2. *Sejam (2.1.5) o SVD de A e (2.1.6) o SVD de* $[A \quad b]$ *respectivamente. Seja* b' *a projeção ortogonal de b sobre* $\mathcal{R}(A)$*. Se* $\sigma_n > \overline{\sigma}_{n+1}$*, então*

$$\begin{aligned} \|x_{TLS} - x_{LS}\|_2 &= \bar{\sigma}_{n+1}^2 \|(A^T A - \bar{\sigma}_{n+1}^2 I)^{-1} x_{LS}\|_2 \\ &\leq \bar{\sigma}_{n+1}^2 \|b'\|_2 \sigma_n^{-1} (\sigma_n^2 - \bar{\sigma}_{n+1}^2)^{-1} \\ &\leq \bar{\sigma}_{n+1}^2 \|x_{LS}\|_2 (\sigma_n^2 - \bar{\sigma}_{n+1}^2)^{-1}, \\ \|x_{TLS}\|_2 &\geq \|x_{LS}\|_2 \end{aligned}$$

Vamos considerar o número de condição do problema TLS e sua relação com o problema de quadrados mínimos. Para assegurar a unicidade das soluções LS e TLS, assumimos que $\sigma_n > \bar{\sigma}_{n+1}$. O vetor $[x_{TLS}, -1]^T$ é um autovetor de $[A \quad b]^T [A \quad b]$ com $\bar{\sigma}_{n+1}^2$ como autovalor associado, i.e.,

$$\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix} = \bar{\sigma}_{n+1}^2 \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix}.$$

A primeira linha desta equação pode ser escrita como

$$(A^T A - \bar{\sigma}_{n+1}^2 I) x_{TLS} = A^T b.$$

Aqui um número positivo da matriz identidade é subtraído de $A^T A$ e o problema TLS é uma deregularização do problema de quadrados mínimos. Como

$$\kappa(A^T A - \bar{\sigma}_{n+1}^2 I) = \frac{\sigma_1^2 - \bar{\sigma}_{n+1}^2}{\sigma_n^2 - \bar{\sigma}_{n+1}^2} > \frac{\sigma_1^2}{\sigma_n^2} = \kappa(A^T A)$$

segue que problema TLS é sempre pior condicionado do que o problema LS.

2.2 Análise do Método TLS Via Projeções Ortogonais

Nos métodos LS e TLS, temos construído soluções aproximadas baseadas em sistemas com matrizes de posto incompleto obtidas projetando o problema original num subespaço de pequena dimensão. Por

exemplo, no caso da técnica TLS truncada, o problema a ser resolvido, $\widetilde{A}_k x = \widetilde{b}_k$, resulta da aproximação de posto k, $[\widetilde{A}_k \quad \widetilde{b}_k] = U_1 U_1^T [A \quad b]$, onde $U_1 U_1^T$ é a matriz de projeção ortogonal sobre o subespaço span$\{\overline{u}_1, \ldots, \overline{u}_k\}$. Ou seja, o sistema resultante é obtido via projeção ortogonal. A pergunta natural é: o que acontece com a solução do "problema projetado" em relação a solução exata do problema original se em de lugar $U_1 U_1^T$ usamos outra matriz de projeção? Nesta seção vamos estudar essas relações e tentar responder a essa pergunta.

2.2.1 Método Geral de Projeções Ortogonais e análise de perturbações

Começamos introduzindo formalmente o conceito de operador projeção que vamos usar nesta seção.

Definição 2.2.1. *Seja $D \in \mathbb{R}^{n\times q}$. $P_D \in \mathbb{R}^{n\times n}$ é uma matriz de projeção ortogonal sobre $\mathcal{R}(D)$ se $\mathcal{R}(P_D) = \mathcal{R}(D)$, $P_D^2 = P_D$, e $P_D^T = P_D$.*

Definição 2.2.2. *Sejam $C, D \in \mathbb{R}^{n\times q}$, dizemos que C é uma perturbação aguda de D se $\|P_C - P_D\|_2 < 1$ e $\|P_{C^T} - P_{D^T}\|_2 < 1$. Neste caso dizemos que C e D são agudas.*

A seguinte definição permite comparar o ângulo entre dois subespaços com a mesma dimensão.

Definição 2.2.3. *Suponha que $\mathcal{R}(C)$ e $\mathcal{R}(D)$ são subespaços equidimensionais de $\mathbb{R}^q$. Definimos o ângulo entre estes subespaços por*

$$\sin\phi \equiv \|P_C - P_D\|_2,$$

onde ϕ é chamado de ângulo entre os subespaços [13].

Seja M um método geral de projeção (como por exemplo os métodos LS ou TLS). Denotamos por x_M a solução de norma mínima de

$$\widetilde{A}x = \widetilde{b} \tag{2.2.1}$$

onde $[\widetilde{A} \quad \widetilde{b}]$ é a matriz de posto k que aproxima $[A \quad b]$, baseada no método de projeção M. Denotemos por $\widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ a decomposição em valores singulares de $\widetilde{A}$ e por $[\Delta\widetilde{A} \quad \Delta\widetilde{b}] = [A \quad b] - [\widetilde{A} \quad \widetilde{b}]$ a matriz corretora. Consideremos também a matriz aproximada A_k de posto k de A com

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Precisamos saber quando o sistema (2.2.1) tem solução. O seguinte resultado dá condições suficientes da existência da solução x_M.

Teorema 2.2.1. *Seja $P \in \mathbb{R}^{m\times m}$ uma matriz projeção ortogonal, $[\widetilde{A} \quad \widetilde{b}] = P[A \quad b]$, e $\Delta\widetilde{A} = A - \widetilde{A}$. Então* posto$(\widetilde{A}) = k$ *e $\widetilde{A}x = \widetilde{b}$ é compatível, sempre que $\|\Delta\widetilde{A}\|_2 < \sigma_k$.*

Demonstração. É claro que posto$(\widetilde{A}) \leq$ posto$([\widetilde{A} \quad \widetilde{b}]) \leq k$. Precisamos mostrar que $\widetilde{b} \in \mathcal{R}(\widetilde{A})$, para tanto vamos mostrar que posto$(\widetilde{A}) = k$. De fato, como $\widetilde{A} = PA$, usando [34, p. 34] segue que

$$\sigma_i \geq \widetilde{\sigma}_i \text{ para } i = 1, \ldots, k. \tag{2.2.2}$$

Agora, como $A - \widetilde{A} = \Delta\widetilde{A}$, segue do Teorema de perturbação de valores singulares (1.2.4) e (2.2.2) que

$$\sigma_i - \widetilde{\sigma}_i \leq \|\Delta\widetilde{A}\|_2, \text{ para } i = 1, \ldots, k.$$

Usando a hipótese $\|\Delta\widetilde{A}\|_2 < \sigma_k$, temos que $0 < \sigma_k - \|\Delta\widetilde{A}\|_2 \leq \widetilde{\sigma}_k$. Logo $0 < \widetilde{\sigma}_k \leq \ldots \leq \widetilde{\sigma}_1$ implica que posto$(\widetilde{A}) = k$ e portanto $\widetilde{b} \in \mathcal{R}(\widetilde{A})$. □

Esta condição sugere que $\widetilde{A}$ seja uma perturbação aguda de A_k, no Teorema 1.2.3. Note do Corolário 1.1.1 que para o problema LS, a matriz projeção ortogonal é $P = P_A$ e a matriz projetada é

$$[\widetilde{A} \quad \widetilde{b}] = P_A[A \quad b] = [P_A A \quad P_A b] = [A \quad b'].$$

Seja $[\overline{A} \quad \overline{b}] = [A \quad b] + [\Delta A \quad \Delta b]$ uma perturbação da matriz $[A \quad b]$, denotamos por $[\overline{A}_k \quad \overline{b}_k]$ a matriz aproximação de posto k da matriz $[\overline{A} \quad \overline{b}]$ baseada no método M. Definimos a matriz corretora $[\Delta\overline{A} \quad \Delta\overline{b}] = [\overline{A} \quad \overline{b}] - [\overline{A}_k \quad \overline{b}_k]$ e assumimos que $\|\Delta\overline{A}\|_2 + \|\Delta A\|_2 < \sigma_k$. Pelo Teorema 2.2.1, $\overline{A}_k x = \overline{b}_k$ é compatível e denotamos por $\bar{x}_M$ a solução de norma mínima. Por outro lado de (2.2.1) temos

$$[\widetilde{A} \quad \widetilde{b}] \begin{bmatrix} x_M \\ -1 \end{bmatrix} = 0.$$

Sejam as colunas

$$Y = \begin{bmatrix} Y_1 \\ y_2^T \end{bmatrix} \text{ com } Y_1 \in \mathbb{R}^{n\times(n-k+1)} \text{ e } y_2 \in \mathbb{R}^{n-k+1}$$

uma base ortonormal do núcleo de $[\widetilde{A} \quad \widetilde{b}]$, $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$.

Como $[x_M^T \ -1]^T \in \mathcal{N}([\widetilde{A} \quad \widetilde{b}])$ segue que

$$\begin{bmatrix} x_M \\ -1 \end{bmatrix} = \begin{bmatrix} Y_1 \\ y_2^T \end{bmatrix} s_Y \quad \text{para algum } s_Y \in \mathbb{R}^{n-k+1}.$$

Assim temos um sistema compatível

$$y_2^T s_Y = -1.$$

Como x_M é a solução de norma mínima temos que

$$s_Y = -(y_2^T)^\dagger = -y_2(y_2^T y_2)^{-1} = -\frac{y_2}{\|y_2\|_2^2}.$$

Então

$$x_M = -Y_1 \frac{y_2}{\|y_2\|_2^2}.$$

Além disso, se $Q \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ é uma matriz ortonormal tal que

$$YQ = \begin{array}{c} \begin{matrix} \text{n-k} & 1 \end{matrix} \\ \left[\begin{array}{cc} D & v \\ 0 & \gamma \end{array}\right] \end{array} \begin{matrix} n \\ 1 \end{matrix}$$

onde $\gamma \neq 0$, segue que

$$\begin{aligned} x_M = -Y_1(y_2^T)^\dagger &= -(Y_1Q)(y_2^TQ)^\dagger \\ &= -[D \quad v][0 \quad \gamma]^\dagger \\ &= -[D \quad v]\begin{bmatrix} 0 \\ \gamma^{-1} \end{bmatrix} \\ &= -\gamma^{-1}v. \end{aligned}$$

Daí só precisamos multiplicar um vetor por uma constante γ^{-1} para encontrar a solução x_M.

Se $Z = [Z_1^T \ z_2]^T$ é outra base ortonormal de $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$, existe uma matriz ortogonal $Q \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ tal que $Z = QY$. Então

$$x_M = -Y_1(y_2^T)^\dagger = -Y_1QQ^T(y_2^T)^\dagger = (-Y_1Q)(y_2^TQ)^\dagger = -Z_1(z_2^T)^\dagger.$$

O que significa que a solução não depende da escolha da base do núcleo de $[\widetilde{A} \quad \widetilde{b}]$.

De forma análoga suponha que a condição $\max(\|\Delta\widetilde{A}\|_2, \|\Delta\overline{A}\|_2 + \|\Delta A\|_2) < \sigma_k$ é satisfeita, e sejam as colunas de $\overline{Y}$ a base ortonormal

de $\mathcal{N}([\overline{A}_k \quad \overline{b}_k])$. Seja a matriz $\overline{Q} \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ uma matriz ortogonal tal que

$$\overline{Y}\,\overline{Q} = \begin{array}{c} \begin{matrix} n\text{-}k & 1 \end{matrix} \\ \begin{bmatrix} \overline{D} & \overline{v} \\ 0 & \overline{\gamma} \end{bmatrix} \end{array} \begin{matrix} n \\ 1 \end{matrix}$$

Então $\overline{x}_M = -\overline{Y}_1(\overline{y}_2^T)^\dagger = -\bar{\gamma}\overline{v}^{-1}$. Logo

$$\begin{bmatrix} \overline{x}_M \\ -1 \end{bmatrix} - \begin{bmatrix} x_M \\ -1 \end{bmatrix} = \begin{bmatrix} v \\ \gamma \end{bmatrix} \gamma^{-1} - \begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \overline{\gamma}^{-1}. \tag{2.2.3}$$

Seja $W \in \mathbb{R}^{(n+1)\times n}$ uma matriz ortonormal com a partição

$$W = \begin{array}{c} \begin{matrix} n \end{matrix} \\ \begin{bmatrix} W_1 \\ w_2^T \end{bmatrix} \end{array} \begin{matrix} n \\ 1 \end{matrix}$$

tal que $W^T[v^T \ \gamma]^T = 0$ (i.e. as colunas de W completam o espaço $\mathbb{R}^{n+1}$). Da equação (2.2.3) segue que

$$W^T \left(\begin{bmatrix} \overline{x}_M \\ -1 \end{bmatrix} - \begin{bmatrix} x_M \\ -1 \end{bmatrix} \right) = -W^T \begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \overline{\gamma}^{-1}, \tag{2.2.4}$$

e consequentemente pela partição de W temos

$$W_1^T(\overline{x}_M - x_M) = -W^T \begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \overline{\gamma}^{-1}. \tag{2.2.5}$$

A expressão

$$\sin\phi_M \equiv \left\| W^T \begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \right\|_2$$

denota o seno do ângulo entre subespaços.

$$\mathcal{R}\left(\begin{bmatrix} v \\ \gamma \end{bmatrix} \right) \quad \text{e} \quad \mathcal{R}\left(\begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \right).$$

Agora podemos apresentar o seguinte resultado.

Teorema 2.2.2. *Seja* $[\overline{A} \quad \overline{b}] = [A \quad b] + [\Delta A \ \Delta b]$. *Sejam* x_M *e* $\overline{x}_M$ *as soluções de norma mínima dos sistemas compatíveis* $\widetilde{A}x = \widetilde{b}$ *e* $\overline{A}_k x =$

$\overline{b}_k$ *obtidas pelo método de projeção* M*, respectivamente. Sempre que* $\max(\|\Delta\widetilde{A}\|_2, \|\Delta\overline{A}\|_2 + \|\Delta A\|_2) < \sigma_k$*, temos*

$$\sin\phi_M \leq \|x_M - \overline{x}_M\|_2 \leq \sin\phi_M \sqrt{1+\|x_M\|_2^2}\sqrt{1+\|\overline{x}_M\|_2^2} \quad (2.2.6)$$

onde ϕ_M *é o ângulo entre os subespaços* $\mathcal{R}\left(\begin{bmatrix} v \\ \gamma \end{bmatrix}\right)$ *e* $\mathcal{R}\left(\begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix}\right)$.

Demonstração. Pelo Teorema CS [26] a matriz

$$\begin{bmatrix} W_1 & v \\ w_2^T & \gamma \end{bmatrix}$$

é uma matriz ortogonal, onde W_1 é uma matriz quadrada; sabemos que $\sigma_{min}^{-1}(W_1) = |\gamma^{-1}|$. De (2.2.5) temos

$$\begin{aligned} \sigma_{min}(W_1)\|\bar{x}_M - x_M\|_2 &\leq \|W_1^T(\bar{x}_M - x_M)\|_2 \\ &= \left\|W^T \begin{bmatrix} \overline{v} \\ \overline{\gamma} \end{bmatrix} \overline{\gamma}^{-1}\right\|_2 \\ &\leq \sin\phi_M |\bar{\gamma}^{-1}|, \end{aligned}$$

ou

$$\|\bar{x}_M - x_M\|_2 \leq \sin\phi_M |\bar{\gamma}^{-1}| \sigma_{min}^{-1}(W_1).$$

Logo, segue que

$$\begin{aligned} \|\bar{x}_M - x_M\|_2 &\leq \sin\phi_M |\bar{\gamma}^{-1}||\gamma^{-1}| \\ &= \sin\phi_M \sqrt{1+\|\bar{x}_M\|_2^2}\sqrt{1+\|x_M\|_2^2}. \end{aligned} \quad (2.2.7)$$

Isto prova a desigualdade a direita em (2.2.6). Para a outra desigualdade,

$$\begin{aligned} \|\bar{x}_M - x_M\|_2 &\geq \|W_1^T(\bar{x}_M - x_M)\|_2 \\ &= \left\|W^T \begin{bmatrix} \overline{v} \\ \bar{\gamma} \end{bmatrix} \bar{\gamma}^{-1}\right\|_2 \\ &\geq \sin\phi_M \bar{\gamma}^{-1} \\ &\geq \sin\phi_M, \end{aligned}$$

pois $|\bar{\gamma}^{-1}| \geq 1$. □

O Teorema 2.2.2 mostra que para o método M, toda perturbação $[\Delta A \ \Delta b]$ que iguala $\mathcal{R}\left(\begin{bmatrix} v \\ \gamma \end{bmatrix}\right)$ com $\mathcal{R}\left(\begin{bmatrix} \overline{v} \\ \bar{\gamma} \end{bmatrix}\right)$ faz o sistema $\bar{A}_k x = \bar{b}_k$

compatível e gera a mesma solução x_M. As perturbações $[\Delta A\ \Delta b]$ que conseguem este resultado estão caracterizados por

$$[\Delta A\ \Delta b] = -[\Delta\widetilde{A} \quad \Delta\widetilde{b}] + H^T,$$

onde $\mathcal{R}(H) \perp \mathcal{R}\left(\begin{bmatrix} v \\ \gamma \end{bmatrix}\right)$. Assim temos várias perturbações que produzem a mesma solução. No caso em que $[\Delta A \quad \Delta b]$ forem arbitrários concluimos que para o método M os efeitos da perturbação dependem do ruído presente em $\mathcal{R}\left(\begin{bmatrix} \bar{v} \\ \bar{\gamma} \end{bmatrix}\right)$. Van Huffel e Vandewalle [37] chegaram a esta conclusão quando estudaram os efeitos de perturbação no problema TLS de posto completo.

O Teorema 2.2.2 mostra que quando $\mathcal{R}\left(\begin{bmatrix} v \\ \gamma \end{bmatrix}\right)$ é perturbado de modo que $0 < \sin\phi_M$, a solução $\overline{x}_M$ não coincide com x_M. Na presença do ruído $[\Delta A \quad \Delta b]$ a melhor precisão que o método M pode atingir é medido por $\sin\phi_M$. Neste caso, o produto das raízes quadradas em (2.2.6) representa um número de condição para o método de projeção. Seja

$$\sin\theta_M = \|YY^T - \overline{Y}\ \overline{Y}^T\|_2$$

o seno do ângulo entre $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$ e $\mathcal{N}([\overline{A}_k \quad \overline{b}_k])$.

Um cálculo direto para encontrar estimativas para $\sin\phi_M$ em termos de parâmetros conhecidos não é fácil. Portanto, vamos atacar o problema fazendo uma análise do $\sin\theta_M$. Existem várias razoẽs para fazer isto.

(1) $\sin\phi_M = \sin\theta_M$ quando $k = n$. Este caso acontece em muitas aplicações.

(2) As estimativas para $\sin\theta_M$ em termos de parâmetros conhecidos é possível. Estes resultados fornecem resultados adicionais quando A é de posto incompleto e o sistema é compatível.

(3) Podemos limitar $\sin\phi_M$ usando $\sin\theta_M$ mais um termo, como segue:

Definamos as seguintes matrizes de projeção

$$P_Y = Y\ Y^T, \quad P_{v,\gamma} = [v^T \quad \gamma]^T[v^T \quad \gamma],$$
$$P_{\overline{Y}} = \overline{Y}\ \overline{Y}^T, \quad P_{\overline{v},\overline{\gamma}} = [\overline{v}^T \quad \overline{\gamma}]^T[\overline{v}^T \quad \overline{\gamma}]$$
$$P_D = [D^T \quad 0]^T[D^T \quad 0], \quad \text{e} \quad P_{\overline{D}} = [\overline{D} \quad 0]^T[\overline{D} \quad 0].$$

Usando o fato de que $P_Y = P_D + P_{v,\gamma}$, $P_{\overline{Y}} = P_{\overline{D}} + P_{\overline{v},\overline{\gamma}}$, $P_D P_{v,\gamma} = 0$, $P_{\overline{D}} P_{\overline{v},\overline{\gamma}} = 0$, para qualquer vetor $z \in \mathbb{R}^{n+1}$ temos

$$\begin{aligned}\|(P_Y - P_{\overline{Y}})z\|_2^2 =& \|(P_D - P_{\overline{D}})z\|_2^2 + \|(P_{v,\gamma} - P_{\overline{v},\overline{\gamma}})z\|_2^2 \\ & - z^T(P_D P_{\overline{v},\overline{\gamma}} + P_{v,\gamma} + P_{v,\gamma} P_{\overline{D}} + P_{\overline{v},\overline{\gamma}} P_D)z.\end{aligned}$$

Em particular, se z_γ é um vetor unitário tal que

$$\|P_{v,\gamma} - P_{\overline{v},\overline{\gamma}}\|_2 = \|(P_{v,\gamma} - P_{\overline{v},\overline{\gamma}})z_\gamma\|_2,$$

então segue que

$$\sin\phi_M \leq \sin\theta_M + \epsilon_M, \tag{2.2.8}$$

onde

$$\epsilon_M = |z_\gamma^T(P_D P_{\overline{v},\overline{\gamma}} + P_{v,\gamma} + P_{v,\gamma} P_{\overline{D}} + P_{\overline{v},\overline{\gamma}} P_D)z_\gamma - \|(P_D - P_{\overline{D}})z_\gamma\|_2^2|^{1/2}.$$

Os pontos mencionados acima são motivação para encontrar cotas superiores para $\sin\theta_M$ em termos do ajuste das matrizes $[\widetilde{A} \quad \widetilde{b}]$ e $[\overline{A}_k \quad \overline{b}_k]$, correções e perturbações. Para obtermos a estimativa desejada, vamos usar dois lemas auxiliares.

Lema 2.2.1. *Seja* $[C \quad D] = [\widetilde{C} \quad \widetilde{D}] + [\Delta\widetilde{C} \quad \Delta\widetilde{D}]$, *onde* $[\widetilde{C} \quad \widetilde{D}]$ *é obtida usando o método de projeção* M. *Então*

$$[\widetilde{C} \quad \widetilde{D}]^\dagger[\Delta\widetilde{C} \quad \Delta\widetilde{D}] = 0.$$

Demonstração. Ver [1]. □

Lema 2.2.2. *Se* C *é uma perturbação aguda de* D, *com* $C = D + E$, *então*

$$\|C^\dagger - D^\dagger\|_2 \leq \mu\|C^\dagger\|_2\|D^\dagger\|_2\|E\|_2,$$

onde $\mu = (1 + \sqrt{5})/2$.

Demonstração. Ver [38]. □

Teorema 2.2.3. *Seja* $[\overline{A} \quad \overline{b}] = [A \quad b] + [\Delta A \quad \Delta b]$ *com* $[\widetilde{A} \quad \widetilde{b}]$ *e* $[\overline{A}_k \quad \overline{b}_k]$ *obtidas pela projeção ortogonal usando o método* M. *Defina* $\mu = (1 + \sqrt{5})/2$. *Se* $\max(\|\Delta\widetilde{A}\|_2, \|\Delta\overline{A}\|_2 + \|\Delta A\|_2) < \sigma_k$ *então*

$$\begin{aligned}\sin\theta_M \leq & \|[\widetilde{A} \quad \widetilde{b}]^\dagger\|_2\|[\Delta A \quad \Delta b]\|_2 \\ & + \mu\|[\widetilde{A} \quad \widetilde{b}]^\dagger\|_2\|[\overline{A}_k \quad \overline{b}_k]^\dagger\|_2\|[\widetilde{A} \quad \widetilde{b}] - [\overline{A}_k \quad \overline{b}_k]\|_2\|\Delta\overline{A} \quad \Delta\overline{b}\|,\end{aligned}$$

onde θ_M *é o ângulo entre os subespaços* $\mathcal{N}([\widetilde{A} \quad \widetilde{b}])$ *e* $\mathcal{N}([\overline{A}_k \quad \overline{b}_k])$.

Demonstração. Seja $\overline{R}^{\perp} \equiv I - [\overline{A}_k \quad \overline{b}_k]^{\dagger}[\overline{A}_k \quad \overline{b}_k]$. Usando os lemas acima temos

$$\begin{aligned}
\sin\theta_M &= \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}[\widetilde{A} \quad \widetilde{b}] - [\overline{A}_k \quad \overline{b}_k]^{\dagger}[\overline{A}_k \quad \overline{b}_k]\|_2 \\
&= \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}[\widetilde{A} \quad \widetilde{b}]\overline{R}^{\perp}\|_2 \\
&= \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}([\widetilde{A} \quad \widetilde{b}] - [\overline{A}_k \quad \overline{b}_k])\overline{R}^{\perp}\|_2 \\
&\leq \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}([\widetilde{A} \quad \widetilde{b}] - [\overline{A}_k \quad \overline{b}_k])\|_2 \\
&= \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}([\Delta A \quad \Delta b] - [\Delta\overline{A} \quad \Delta\overline{b}])\|_2 \\
&= \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}[\Delta A \quad \Delta b] - ([\widetilde{A} \quad \widetilde{b}]^{\dagger} - [\overline{A}_k \quad \overline{b}_k]^{\dagger})[\Delta\overline{A} \quad \Delta\overline{b}]\|_2 \\
&\leq \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}\|_2\|[\Delta A \quad \Delta b]\|_2 + \|[\widetilde{A} \quad \widetilde{b}]^{\dagger} - [\overline{A}_k \quad \overline{b}_k]^{\dagger}\|_2\|[\Delta\overline{A} \quad \Delta\overline{b}]\|_2 \\
&\leq \|[\widetilde{A} \quad \widetilde{b}]^{\dagger}\|_2\|[\Delta A \quad \Delta b]\|_2 \\
&\quad + \mu\|[\widetilde{A} \quad \widetilde{b}]^{\dagger}\|_2\|[\overline{A}_k \quad \overline{b}_k]^{\dagger}\|_2\|[\widetilde{A} \quad \widetilde{b}] - [\overline{A}_k \quad \overline{b}_k]\|_2\|[\Delta\overline{A} \quad \Delta\overline{b}]\|_2.
\end{aligned}$$

□

2.2.2 Estimativas para os Métodos LS e TLS

Quando estudamos sistemas lineares com perturbações é importante distinguir problemas zero e não zero residuais. A Figura 2.1 ajuda a esclarecer esta relação. Definimos o ângulo β_0 entre os dados exatos b_0 e $\mathcal{R}(A_0)$ como sendo

$$\sin\beta_0 = \frac{\|b_0 - A_0x\|_2}{\|b_0\|_2}.$$

O problema é chamado zero residual se $b_0 \in \mathcal{R}(A_0)$, i.e., β_0 é zero. Em outras palavras, quando o ruído não está presente nos dados, existe uma relação exata mas não observável $A_0x_0 = b_0$. Os problemas TLS se aplicam a estes casos. A sensibilidade da solução depende linearmente do número de condição.

Em problemas não zero residuais, $b_0 \notin \mathcal{R}(A_0)$ e então o ângulo β_0 é diferente de zero. Isto significa que, ainda sem a presença de ruído nos dados não é possível encontrar uma relação linear exata $A_0x_0 = b_0$. Estes problemas aparecem, por exemplo, em modelos de ajuste e predição quando desejamos aproximar dados não lineares por meio de modelos lineares ou predizer a resposta de sistemas por modelos de sistemas simplificados.

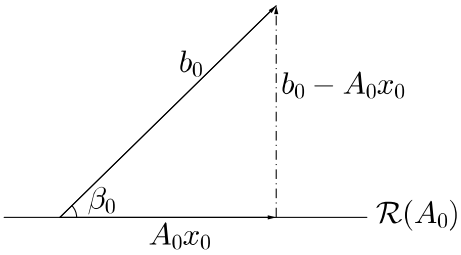

Figura 2.1: Ilustração geométrica do ângulo β_0 com b_0 e $\mathcal{R}(A_0)$.

Nesta parte vamos considerar o problema zero residual e vamos usar resultados anteriores para investigar a presição dos métodos LS e TLS na presença de perturbações nos dados. Vamos considerar o caso Posto$(A) = k$ e $Ax = b$.

Em problemas de cálculo de frequências a matriz coeficiente A tem posto $k \leq n$ e $Ax = b$ é compatível quando os dados são exatos. Porém na presença de ruído temos que lidar com um problema perturbado $\bar{A}x \approx \bar{b}$.

Vamos assumir que Posto$(A) = k$ e $Ax = b$ é compatível; portanto as soluções LS e TLS coincidem: $x_0 = x_{LS} = x_{TLS}$. Seja

$$[\overline{A} \quad \overline{b}] = [A \quad b] + [\Delta A \quad \Delta b],$$

denotamos por $\overline{A}_k$ a matriz aproximação de posto k de $\overline{A} \equiv \overline{A}_k + \Delta\overline{A}_k$. Vamos resolver o problema LS truncado $\overline{A}_k x = \overline{b}$. Este sistema é equivalente a encontrar a solução de norma mínima $\overline{x}_{LS}$ do sistema compatível

$$\overline{A}_k x = \overline{b}_k \tag{2.2.9}$$

onde $\overline{b}_k = \overline{A}_k \overline{A}_k^{\dagger} \overline{b}$ é a projeção ortogonal de $\overline{b}$ sobre $\mathcal{R}(\overline{A}_k)$. Pelos resultados anteriores, para calcular a cota de $\sin\phi_{LS}$, precisamos calcular uma cota superior para $\sin\theta_{LS}$

$$\sin\theta_{LS} = \text{dist}(\mathcal{N}([A \quad b]), \mathcal{N}([\overline{A}_k \quad \overline{b}_k])).$$

Também estamos interessados em aplicar o método TLS truncado ao sistema $\overline{A}x = \overline{b}$. Seja

$$[\widetilde{A} \quad \widetilde{b}] = \sum_{i=1}^{k} \bar{\sigma}_i \bar{u}_i \bar{v}_i^T$$

a matriz aproximação de posto k da matriz $[\overline{A} \quad \overline{b}] = \sum_{i=1}^{n+1} \bar{\sigma}_i \bar{u}_i \bar{v}_i^T$ e $[\Delta\widetilde{A} \quad \Delta\widetilde{b}] = [\overline{A} \quad \overline{b}] - [\widetilde{A} \quad \widetilde{b}]$. Então o método TLS calcula a solução de norma mínima $\widetilde{x}_{TLS}$ do sistema compatível

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

De forma análoga que acima temos que calcular cotas para $\sin\phi_{TLS}$ usando as cotas de $\sin\theta_{TLS}$

$$\sin\theta_{TLS} = \text{dist}(\mathcal{N}([A \quad b]), \mathcal{N}([\widetilde{A} \quad \widetilde{b}])).$$

Teorema 2.2.4. *Assuma que o posto(A)=k, e que x_0 é a solução de norma mínima do sistema compatível $Ax = b$. Seja $[\overline{A} \quad \overline{b}] = [A \quad b] + [\Delta A \quad \Delta b]$ e denote por $\overline{x}_{LS}$ e $\widetilde{x}_{TLS}$ as soluções de norma mínima dos sistemas perturbados como acima, com $\|[\Delta A \quad \Delta b]\|_2 < \sigma_k$. Então*

$$\sin\phi_{LS} \leq \|x_0 - \overline{x}_{LS}\|_2 \leq \sin\phi_{LS}\sqrt{1 + \|x_0\|_2^2}\sqrt{1 + \|\overline{x}_{LS}\|_2^2}$$

onde

$$\sin\phi_{LS} \leq \frac{\|[\Delta A \quad \Delta b]\|_2}{\sigma_k - \|\Delta A\|_2} + \epsilon_{LS};$$

$$\sin\phi_{TLS} \leq \|x_0 - \widetilde{x}_{TLS}\|_2 \leq \sin\phi_{TLS}\sqrt{1 + \|x_0\|_2^2}\sqrt{1 + \|\widetilde{x}_{TLS}\|_2^2}$$

onde

$$\sin\phi_{TLS} \leq \frac{\|[\Delta A \quad \Delta b]\|_2}{\overline{\sigma}_k - \|[\Delta A \quad \Delta b]\|_2} + \epsilon_{TLS}.$$

Demonstração. Ver [8]. □

Desprezando ϵ_{LS} e ϵ_{TLS}, vemos que a cota superior para o ângulo TLS é menor que o ângulo LS, desde que $\sigma_k \leq \bar{\sigma}_k$. De fato, concluimos que o método TLS melhora quando $\bar{\sigma}_k/\sigma_k$ cresce.

2.2.3 Relação entre as Soluções LS e TLS

Finalmemte temos o resultado que relaciona as soluções LS e TLS

Teorema 2.2.5. *Sejam A e $[A \quad b]$ com a decomposição usual SVD. Seja $R_k = b - Ax_{LS}$, $x_{LS} = -\gamma'^{-1}v'$, $x_{TLS} = -\widetilde{\gamma}^{-1}\widetilde{v}$. Se $\overline{\sigma}_{k+1} < \sigma_k$ então*

$$\sin\phi \leq \|x_{LS} - x_{TLS}\|_2 \leq \sin\phi\sqrt{1 + \|x_{LS}\|_2^2}\sqrt{1 + \|x_{TLS}\|_2^2}$$

onde ϕ é o ângulo entre os subespaço $\mathcal{R}([v'^T \ \gamma']^T)$ e $\mathcal{R}([\widetilde{v}^T \ \widetilde{\gamma}]^T)$,

$$\sin\phi \leq (\mu\bar{\sigma}_{k+1}(2\bar{\sigma}_{k+1} + \|R_k\|_2)/\sigma_k^2 + \epsilon), \text{ e } \mu = (1 + \sqrt{5})/2.$$

Demonstração. Ver [8]. □

2.3 Resultados Numéricos Preliminares

Nesta seção apresentamos um estudo comparativo da eficiência dos métodos LS e TLS quando aplicados a um problema da área de espectroscopia de ressonância magnética (MRS). Neste caso a matriz de dados A tem estrutura Hankel, i.e., as entradas são definidas como $a_{i,j} = h_{i+j-1}$

$$A = \begin{bmatrix} h_1 & h_2 & \dots & h_n \\ h_2 & h_3 & \dots & h_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ h_n & h_{n+1} & \dots & h_{n+m} \end{bmatrix}$$

e $b = [h_0, \dots, h_{n-1}]^T$, onde $n + m \leq q$, e h_k é um sinal modelado por

$$h_k = \sum_{j=1}^{p} c_j e^{\iota\phi j} e^{(\alpha_j + \iota\omega_j)k\Delta t}, \quad \iota = \sqrt{-1}, \quad k = 0, 1, \dots, q.$$

O problema consiste em estimar os parâmetros α_j, β_j, ϕ_j, e c_j, a partir de medidas experimentais do sinal h_k.

Se h_k é livre de erros, é conhecido que posto$(A) = p$ e a solução de norma mínima do problema $\min \|Ax - b\|_2$ pode ser usada para estimar as constantes de interesse através de técnicas de predição linear [5].

A principal dificuldade do problema é que, como o sinal experimental é da forma $\overline{h}_k = h_k^{\text{exato}} + e_k$, então a matriz é da forma $\overline{A} = A + E$, com posto completo e vetor de dados é $\overline{b} = b^{\text{exato}} + e$. Maiores informações podem ser encontradas na referência [5].

Para nosso exemplo consideramos os valores da Tabela 2.1, com $p = 11$, $m = n = 256$, $q = 600$, $\Delta t = 0.000333$ e $\phi_j = \xi_j \pi/180$.

j	c_j	ξ_j (graus)	α_j	$\omega_j/2\pi$ (Hz)
1	75	135	50	-86
2	150	135	50	-70
3	75	135	50	-54
4	150	135	50	152
5	150	135	50	168
6	150	135	50	292
7	150	135	50	308
8	150	135	25	360
9	1400	135	285	440
10	60	135	25	490
11	500	135	200	530

Tabela 2.1: Valores exatos para o sinal MRS.

Consideramos matriz perturbada $\overline{A}$, sendo $\overline{A} = A + \sigma\mathcal{E}$, onde $\mathcal{E}$ é uma matriz de ruído Gaussiano, σ é o desvio padrão nas partes real e imaginária. Note que a matriz exata A tem posto 11.

A Figura 2.2 mostra a parte real do sinal usado no experimento.

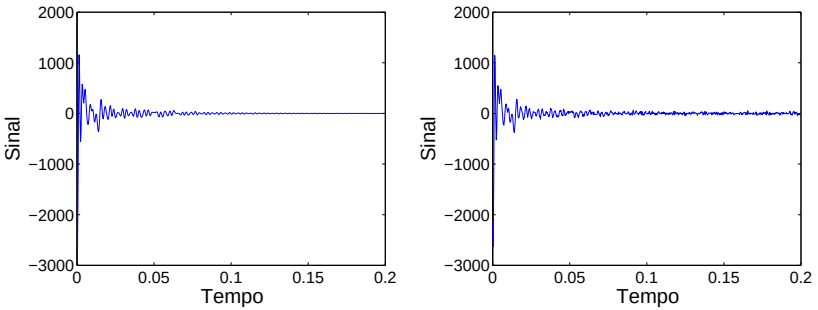

Figura 2.2: Sinais puro e perturbado.

Apresentamos resultados obtidos com a técnica LS truncada e TLS truncada usando em ambos os casos o índice de truncamento $k = 11$, considerando vários valores do desvio padrão.

A qualidade das soluções x_{LS}, x_{TLS} em termos de erro relativo e resíduos relativos com respeito à solução original x_0 são mostrados na Tabela 2.2 e ilustrados graficamente na Figura 2.3.

σ	$\|x_{LS}-x_0\|_2/\|x_0\|_2$	$\|x_{TLS}-x_0\|_2/\|x_0\|_2$	$\|Ax_{LS}-b\|_2/\|b\|_2$	$\|Ax_{TLS}-b\|_2/\|b\|_2$
2	0.01627	0.01631	0.00203	0.00205
4	0.03266	0.03279	0.00397	0.00403
6	0.04928	0.04950	0.00587	0.00594
8	0.06622	0.06650	0.00780	0.00779
10	0.08358	0.08386	0.00984	0.00961
12	0.10143	0.10167	0.01208	0.01142
14	0.11989	0.12005	0.01459	0.01325
16	0.13910	0.13929	0.01742	0.01515
18	0.15936	0.16033	0.02067	0.01716

Tabela 2.2: Erro relativo e Resíduo relativo das soluções LS e TLS.

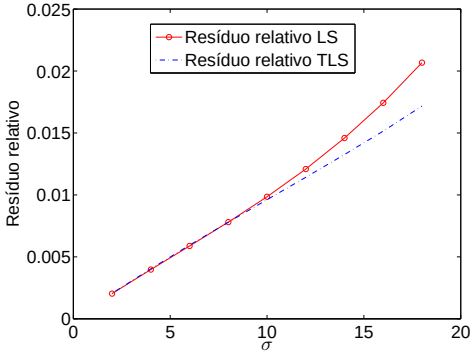

Figura 2.3: Resíduo relativo.

Vemos que os erros relativos associados às soluções LS e TLS são muito próximos. No caso do resíduo relativo, vemos que para valores grandes de σ a solução TLS é melhor.

As estimativas dos valores dos parâmetros α e ω são apresentados na Tabela 2.3 e graficamente nas Figuras 2.4 e 2.5.

σ	$\|\widetilde{\alpha}_{LS}-\alpha\|_2/\|\alpha\|_2$	$\|\widetilde{\alpha}_{TLS}-\alpha\|_2/\|\alpha\|_2$	$\|\widetilde{\omega}_{LS}-\omega\|_2/\|\omega\|_2$	$\|\widetilde{\omega}_{TLS}-\omega\|_2/\|\omega\|_2$
2	0.00571	0.00618	0.00039	0.00040
4	0.01015	0.01172	0.00073	0.00078
6	0.01406	0.01662	0.00102	0.00111
8	0.01845	0.02098	0.00127	0.00142
10	0.02448	0.02495	0.00150	0.00169
12	0.03305	0.02884	0.00172	0.00192
14	0.04473	0.03309	0.00196	0.00213
16	0.05989	0.03833	0.00225	0.00230
18	0.07904	0.04567	0.00262	0.00246

Tabela 2.3: Erros Relativos dos parâmetros α e ω.

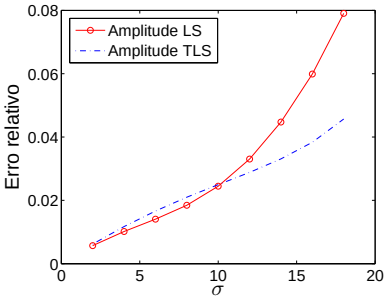

Figura 2.4: Amplitude.

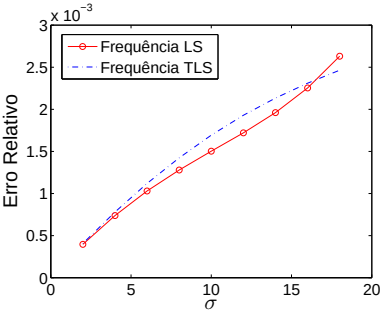

Figura 2.5: Frequência.

Vemos que ambos métodos fornecem boas estimativas das frequências ω independente do valor de σ. A mesma observação vale para os parâmetros α quando σ é pequeno. Porém, para valores maiores de σ, vemos que o método TLS proporciona uma melhor precisão, confirmando que o método TLS é uma excelente alternativa para problemas onde o ruído afeta a matriz e o vetor de dados.

Capítulo 3

Método de Quadrados Mínimos Totais Regularizados

Neste capítulo apresentamos a teoria básica do método de Regularização de Tikhonov para o problema TLS e um método de regularização via truncamento do método TLS.

3.1 Regularização de Tikhonov do problema LS e a GSVD

A idéia básica da regularização é incorporar alguma informação adicional ao problema que permita estabilizá-lo, e determinar uma solução aproximada e compatível com os dados de entrada; um dos métodos mais usuais é o método de regularização de Tikhonov.

Nesta seção apresentamos um estudo introdutório à regularização de Tikhonov bem como a Decomposição em Valores Singulares Generalizada (GSVD) que será usada para calcular a solução regularizada.

No método de Tikhonov para o problema LS, substituimos o problema (1.1.2) por

$$x_\lambda = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\{\|Ax-b\|_2^2 + \lambda\|L(x-x_0)\|_2^2\} \tag{3.1.1}$$

onde λ é uma constante positiva, chamada de parâmtro de regularização, escolhida de modo a controlar o tamanho do vetor solução, e L

é uma matriz $\mathbb{R}^{p\times n}$ que define uma seminorma sobre a solução. Geralmente, L representa o operador primeira derivada ou segunda derivada. x_0 é uma aproximação inicial para a solução caso esteja disponível, caso contrário define-se $x_0 = 0$.

O desafio deste problema é escolher um parâmetro λ tal que x_λ aproxime satisfatoriamente a solução exata x_{exata} de (1.1.2). A escolha da matriz L determina o tipo de problema de Tikhonov:

$$L = I_n\text{: Problema de Tikhonov na Forma Padrão.}$$
$$L \neq I_n\text{: Problema de Tikhonov na Forma Geral.}$$

Esquemas equivalentes à regularização de Tikhonov foram propostos por J. Riley [32] e D. L. Phillips [28], mas foi G. H. Golub o primeiro autor a propôr uma maneira apropriada de resolver o problema (3.1.1). A idéia é tratar este problema como um problema de quadrados mínimos

$$x_\lambda = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\| \begin{bmatrix} b \\ \sqrt{\lambda}Lx_0 \end{bmatrix} - \begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x \right\|_2^2 \tag{3.1.2}$$

cujas equações normais são

$$(A^TA + \lambda L^TL)x_\lambda = A^Tb + \lambda L^TLx_0. \tag{3.1.3}$$

A solução da equação (3.1.3) pode ser escrita, para o caso frequente $x_0 = 0$, como

$$(A^TA + \lambda L^TL)x_\lambda = A^Tb \tag{3.1.4}$$

sendo a unicidade obtida se $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$.

Considerando a regularização de Tikhonov na forma padrão, x_λ é dado por

$$x_\lambda = \sum_{i=1}^{r} f_i \frac{u_i^Tb}{\sigma_i} v_i \tag{3.1.5}$$

onde $r = \text{posto}(A)$ e

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \cong \begin{cases} 1, & \sigma_i \gg \sqrt{\lambda} \\ \frac{\sigma_i^2}{\lambda}, & \sigma_i \ll \sqrt{\lambda} \end{cases} \tag{3.1.6}$$

são chamados fatores de filtro para a regularização de Tikhonov.

Os fatores f_i filtram as componentes de erro da solução. Assim, se em (3.1.5) λ for muito grande, a solução calculada pode não ter incorporado informações da solução do problema. Em contrapartida, se λ for muito pequeno, pouco ruído pode ter sido filtrado e a solução encontrada não é relevante.

O resíduo associado pode ser escrito como

$$r_\lambda = b - Ax_\lambda = \sum_{i=1}^{r}(1 - f_i)u_i^T b u_i + b_\perp$$

onde $b_\perp = b - \sum_{i=1}^{r} u_i^T b u_i = \sum_{i=r+1}^{m} u_i^T b u_i$ é a componente do vetor b que não pertence ao espaço coluna da matriz A. Como $\{u_i\}$ é uma base ortonormal em $\mathbb{R}^m$, então $b = \sum_{i=1}^{m}(u_i^T b)u_i$. Asim,

$$\begin{aligned} \|x_\lambda\|_2^2 &= \sum_{i=1}^{r}\left(f_i \frac{u_i^T b}{\sigma_i}\right)^2 \\ \|r_\lambda\|_2^2 &= \sum_{i=1}^{r}\left((1 - f_i)u_i^T b\right)^2 + \|b_\perp\|_2^2. \end{aligned} \tag{3.1.7}$$

O caso geral pode ser abordado eficientemente através da Decomposição em Valores Singulares Generalizados (GSVD).

Teorema 3.1.1. *(GSVD). Seja o par matricial (A, L) em que $A \in \mathbb{R}^{m\times n}$ e $L \in \mathbb{R}^{p\times n}$ com $m \geq n \geq p$ com* posto$(L) = p$. *Então existem matrizes $U \in \mathbb{R}^{m\times m}$ e $V \in \mathbb{R}^{p\times p}$ com colunas ortonormais e X uma matriz não singular tais que*

$$A = U \begin{bmatrix} \Sigma & 0 \\ 0 & I_{n-p} \end{bmatrix} X^{-1}, \quad L = V[M \quad 0]X^{-1} \tag{3.1.8}$$

com $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ e $M = \mathrm{diag}(\mu_1, \ldots, \mu_p)$. Os coeficientes σ_i e μ_i satisfazem

$$0 \leq \sigma_1 \leq \ldots \leq \sigma_p \leq 1 \quad \text{e} \quad 1 \geq \mu_1 \geq \ldots \geq \mu_p \geq 0. \tag{3.1.9}$$

Também, $\Sigma^2 + M^2 = I_p$ e os valores singulares generalizados de (A, L) são definidos como $\gamma_i := \frac{\sigma_i}{\mu_i}$.

Demonstração. A prova pode ser encontrada em [13]. □

Uma característica relevante dos valores singulares generalizados de um par matricial (A, L) é que crescem à medida que i aumenta como mostra a Figura 3.1.

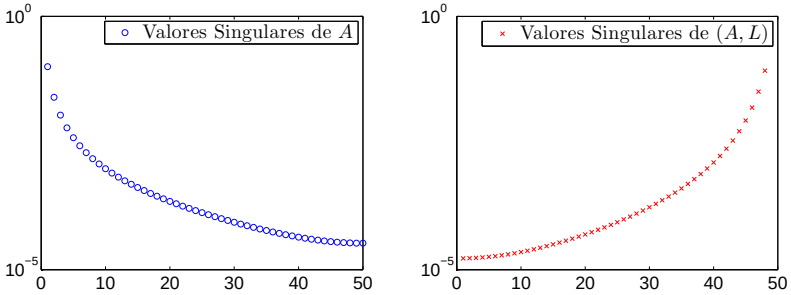

Figura 3.1: Valores singulares e valores singulares generalizados.

Na prática, para problemas onde a GSVD de (A, L) pode ser calculada facilmente, a solução do método de regularização de Tikhonov dada na equação (3.1.4) pode ser escrito como

$$x_\lambda = \sum_{i=1}^{p} f_i \frac{u_i^T b}{\sigma_i} z_i + \sum_{i=p+1}^{n} (u_i^T b) z_i \tag{3.1.10}$$

onde $f_i = \frac{\gamma_i^2}{\gamma_i^2+\lambda}$ e z_i são as colunas da matriz X dada pela GSVD.

Neste caso temos o resíduo associado dado por

$$r_\lambda = b - Ax_\lambda = \sum_{i=1}^{p}(1-f_i)u_i^T b u_i + \sum_{i=n+1}^{m}(u_i^T b)u_i,$$

usando o fato que $Az_i = \sigma_i u_i$, para $i = 1, \ldots, p$ e $Az_i = u_i$ para $i = p+1, \ldots, n$. Assim

$$\|r_\lambda\|_2^2 = \sum_{i=1}^{p}\left((1-f_i)u_i^T b\right)^2 + \sum_{i=n+1}^{m}(u_i^T b)^2.$$

Note que $\|Lx_\lambda\|_2 = \|x_\lambda\|_2$ quando $L = I$. Para o caso em que $L \neq I$, a norma $\|x_\lambda\|_2$ é substituída pela seminorma $\|Lx_\lambda\|_2$. Usando a GSVD temos

$$\|Lx_\lambda\|_2^2 = \sum_{i=1}^{p}\left(f_i \frac{u_i^T b}{\gamma_i}\right)^2.$$

Vemos que se λ cresce, a seminorma $\|Lx\|_2$ do vetor solução decresce de forma monótona, enquanto o resíduo $\|Ax - b\|_2$ cresce de forma monótona.

3.1.1 Métodos de Escolha do Parâmetro de Regularização

Na literatura existem vários métodos de escolha deste parâmetro, vamos mencionar quatro deles que têm sido usados mais frequentemente.

(1) O método de Discrepância atribuido a Morozov [25], escolhe λ tal que a norma do resíduo seja igual a uma cota superior δ para $\|e\|_2$, i.e., escolhemos o parâmetro λ que satisfaz a equação não linear

$$\|b - Ax_\lambda\|_2 = \delta, \quad \|e\|_2 \leq \delta.$$

Este é um dos poucos métodos que determina o parâmetro λ como função da norma do erro $\|e\|_2$ presente no vetor de dados b.

(2) Outro método importante é o método de Validação Cruzada Generalizada (GCV), desenvolvido por G. H Golub, M. T. Heath e G. Wahba [12]. Neste método o parâmetro de regularização é o valor que minimiza a função

$$G_{A,b}(\lambda) = \frac{n\left\|(I - AA_\lambda^\dagger b)\right\|_2^2}{\left(tr(I - AA_\lambda^\dagger)\right)}$$

onde $A_\lambda^\dagger = (A^TA + \lambda I)^{-1}A^T$.

(3) O método da curva L, introduzido por P. C. Hansen [18], escolhe como parâmetro de regularização o valor que maximiza a curvatura da curva parametrizada por $\lambda \geq 0$.

$$\mathcal{L}(\lambda) = \{(a, b);\ a = \log(\|r_\lambda\|_2^2),\ b = \log(\|x_\lambda\|_2^2)\}$$

onde x_λ e r_λ são dados por (3.1.7).

(4) O algoritmo de Ponto Fixo foi proposto por Bazán [6] baseado no trabalho de Regińska [29], este método minimiza a função

$$\psi_\mu = x(\lambda)y(\lambda)^\mu, \quad \mu > 0.$$

em que $y(\lambda) = \|x_\lambda\|_2^2$ e $x(\lambda) = \|r_\lambda\|_2^2$. Bazán provou que o valor $\lambda = \lambda^*$ que minimiza a função ψ_μ é ponto fixo da função

$$\phi_\mu = \sqrt{\mu}\frac{\|b - Ax_\lambda\|_2}{\|x_\lambda\|_2}.$$

A diferença entre os métodos acima é que os três últimos não dependem do conhecimento de qualquer cota para a norma do erro e. Na prática eles funcionam bem numa variedade grande de problemas, mas podem falhar ocasionalmente, e, por isso, são conhecidos como heurísticos.

3.2 Regularização de Tikhonov e TLS

A regularização do problema TLS é baseado nas idéias do método de regularização de Tikhonov para o problema LS. Então, lembramos que a versão geral do método de Tikhonov tem a forma

$$\min_x \{\|Ax - b\|_2^2 + \lambda \|Lx\|_2^2\} \tag{3.2.1}$$

e que a solução é

$$\left(A^T A + \lambda L^T L\right) x = A^T b. \tag{3.2.2}$$

Agora observamos que a regularização de Tikhonov para o problema LS tem uma importante formulação equivalente

$$\begin{aligned} &\min_x \|Ax - b\|_2 \\ &\textit{sujeito a } \|Lx\|_2 \leq \delta, \end{aligned} \tag{3.2.3}$$

onde δ é uma constante positiva. O problema de quadrados mínimos sujeito (3.2.3) pode ser resolvido usando o método dos multiplicadores de Lagrange, ver apêndice A. De fato, consideremos a função Lagrangeana

$$\mathcal{L}(x, \lambda) = \|Ax - b\|_2^2 + \lambda \left(\|Lx\|_2^2 - \delta^2\right), \tag{3.2.4}$$

pode ser demonstrado que se $\delta \leq \|Lx_{LS}\|_2$, onde x_{LS} é a solução de quadrados mínimos do sistema (1.1.2), então a solução x_δ de (3.2.3) é idêntica à solução de Tikhonov x_λ de (3.2.1) para um λ apropriado, e existe uma relação monótona entre os parâmetros δ e λ.

Para levar esta idéia ao problema TLS, adicionamos a desigualdade $\|Lx\|_2 \leq \delta$ do vetor solução x no problema (2.1.1). Portanto, a formulação do problema regularizado TLS (RTLS) pode ser escrita como

$$\begin{aligned} &\min_{[\widetilde{A}\ \widetilde{b}] \in \mathbb{R}^{m \times (n+1)}} \left\| [\ A \quad b\ ] - [\ \widetilde{A} \quad \widetilde{b}\ ] \right\|_F^2 \\ &\textit{sujeito a } \widetilde{A}x = \widetilde{b} \textit{ e } \|Lx\|_2 \leq \delta. \end{aligned} \tag{3.2.5}$$

A função Lagrangeana correspondente é

$$\hat{\mathcal{L}}(\widetilde{A}, x, \mu) = \left\| [\ A \quad b\ ] - [\ \widetilde{A} \quad \widetilde{A}x\ ] \right\|_F^2 + \mu \left(\|Lx\|_2^2 - \delta^2\right) \tag{3.2.6}$$

O seguinte teorema caracteriza completamente solução x_δ do problema regularizado (3.2.5).

Teorema 3.2.1. *A solução regularizada TLS $\bar{x}_\delta$ de (3.2.5), com a desigualdade como igualdade, é a solução do problema*

$$\left(A^T A + \lambda_I I_n + \lambda_L L^T L\right) x = A^T b \tag{3.2.7}$$

onde os parâmetros λ_I e λ_L são

$$\lambda_I = -\frac{\|Ax - b\|_2^2}{1 + \|x\|_2^2} \tag{3.2.8}$$

$$\lambda_L = \mu\left(1 + \|x\|_2^2\right) \tag{3.2.9}$$

onde μ é o multiplicador de Lagrange em (3.2.6). Os dois parâmetros estão relacionados pela equação

$$\lambda_L \delta^2 = b^T(b - Ax) + \lambda_I$$

Ainda mais, o resíduo TLS satisfaz

$$\|[A \quad b] - [\widetilde{A} \quad \widetilde{b}]\|_F^2 = -\lambda_I. \tag{3.2.10}$$

Demonstração. Para caracterizar $\bar{x}_\delta$, igualamos as derivadas parciais da função Lagrangeana a zero. Derivando em relação às entradas de $\widetilde{A}$, temos

$$\widetilde{A} - A - rx^T = 0, \tag{3.2.11}$$

onde $r = b - \widetilde{A}x$. Por outro lado, derivando em relação às entradas de x resulta

$$-\widetilde{A}^T r + \mu L^T L x = 0. \tag{3.2.12}$$

Derivando em relação a μ

$$\|Lx\|_2^2 = \delta^2. \tag{3.2.13}$$

Substituindo r, na expressão (3.2.12), temos a igualdade

$$\left(\widetilde{A}^T \widetilde{A} + \mu L^T L\right) x = \widetilde{A}^T b. \tag{3.2.14}$$

Usando (3.2.11) e (3.2.12), temos $A = \widetilde{A} - rx^T$ e $\widetilde{A}^T r = L^T L x$. Portanto

$$A^T A = \widetilde{A}^T \widetilde{A} - \mu x x^T L^T L + \|r\|_2^2 x x^T - \mu L^T L x x^T \tag{3.2.15}$$

e

$$A^T b = \widetilde{A}^T b - \left(r^T b\right) x. \tag{3.2.16}$$

Inserindo (3.2.15) e (3.2.16) em (3.2.14) e usando a identidade (3.2.13)

$$\begin{aligned} xx^T L^T Lx &= \delta^2 x \\ \|r\|_2^2 xx^T x &= \|r\|_2^2 \|x\|_2^2 x, \end{aligned}$$

chegamos à seguinte expressão

$$\left(A^T A + \lambda_I I_n + \lambda_L L^T L\right) x = A^T b,$$

com

$$\lambda_I = \mu\delta^2 - \|r\|_2^2 \|x\|_2^2 - r^T b \tag{3.2.17}$$

e

$$\lambda_L = \mu\left(1 + \|x\|_2^2\right). \tag{3.2.18}$$

O passo seguinte é a eliminação do multiplicador de Lagrange μ, nas expressões λ_I e λ_L. Primeiro, usamos a relação (3.2.11)

$$r = b - \widetilde{A}x = b - Ax - \|x\|_2^2 r,$$

para obter

$$\left(1 + \|x\|_2^2\right) r = b - Ax. \tag{3.2.19}$$

Tomando norma e elevando ao quadrado

$$\left(1 + \|x\|_2^2\right) \|r\|_2^2 = \frac{\|b - Ax\|_2^2}{1 + \|x\|_2^2}. \tag{3.2.20}$$

Por outro lado, multiplicando (3.2.12) por x e isolando μ temos

$$\mu = \frac{x^T \widetilde{A}^T r}{\|Lx\|_2^2} = \frac{1}{\delta^2}\left(r^T b - \|r\|_2^2\right). \tag{3.2.21}$$

Inserindo (3.2.20) e (3.2.21) em (3.2.17) temos

$$\lambda_I = -\frac{\|b - Ax\|_2^2}{1 + \|x\|_2^2}. \tag{3.2.22}$$

Passando agora para o parâmetro λ_L, usamos (3.2.20) e (3.2.21) para obter

$$\lambda_L = \mu\left(1 + \|x\|_2^2\right) = \frac{1}{\delta^2}\left(r^T b - \|r\|_2^2\right)\left(1 + \|x\|_2^2\right) \tag{3.2.23}$$

Finalmente, a relação entre λ_L e λ_I pode ser derivada como segue: De (3.2.20) e (3.2.22), temos $\lambda_I = -\|r\|_2^2 \left(1 + \|x\|_2^2\right)$, daí usando (3.2.19), (3.2.23) obtemos

$$\lambda_L = \frac{1}{\delta^2} \left[b^T (b - Ax) + \lambda_I \right]. \tag{3.2.24}$$

Para calcular o erro de aproximação, usamos a relação (3.2.11)

$$[\ A \quad b\ ] - [\ \widetilde{A} \quad \widetilde{A}x\ ] = [\ A - \widetilde{A} \quad r\ ] = [\ -rx^T \quad r\ ] = -r \begin{bmatrix} x \\ -1 \end{bmatrix}^T,$$

junto com (3.2.20) e (3.2.22). Assim temos:

$$\left\| [\ A \quad b\ ] - [\ \widetilde{A} \quad \widetilde{b}\ ] \right\|_F^2 = \left(1 + \|x\|_2^2\right) \|r\|_2^2 = -\lambda_I. \tag{3.2.25}$$

□

Com estes resultados concluimos que $\bar{x}_\delta$ é solução da equação (3.2.7) com λ_I e λ_L dados por (3.2.22) e (3.2.24) respectivamente. Agora vamos discutir as implicações deste Teorema para a forma padrão $L = I_n$ (matriz identidade).

3.2.1 O Caso da Forma Padrão

Na forma padrão, a equação (3.2.7) fica na forma

$$(A^T A + \lambda_{IL} I_n) x = A^T b$$

com $\lambda_{IL} = \lambda_I + \lambda_L$. A solução RTLS $\bar{x}_\delta$ de (3.2.5) tem a mesma forma do que a solução de Tikhonov x_δ de (3.2.3). As duas soluções têm a seguinte relação

Teorema 3.2.2. *Seja $L = I_n$ e seja $\bar{\sigma}_{n+1}$ o menor valor singular da matriz $[A \quad b]$. Então para qualquer valor de δ, as soluções $\bar{x}_\delta$ e x_δ estão relacionados como segue*

δ	soluções	λ_{IL}
$\delta < \|x_{LS}\|_2$	$\bar{x}_\delta = x_\delta$	$\lambda_{IL} > 0$
$\delta = \|x_{LS}\|_2$	$\bar{x}_\delta = x_\delta = x_{LS}$	$\lambda_{IL} = 0$
$\|x_{LS}\|_2 < \delta < \|x_{TLS}\|_2$	$\bar{x}_\delta \neq x_\delta = x_{LS}$	$0 > \lambda_{IL} > -\bar{\sigma}_{n+1}^2$
$\delta \geq \|x_{TLS}\|_2$	$\bar{x}_\delta = x_{TLS},\ x_\delta = x_{LS}$	$\lambda_{IL} = -\bar{\sigma}_{n+1}^2$

Demonstração. Precisamos determinar o sinal de λ_{IL} como uma função de δ. Para tanto, usamos o fato de que os vetores correpondentes x são soluções da formulação Lagrangeano (3.2.4) do problema de regularização de Tikhonov (3.2.3). Vamos estudar cada caso presentado na Tabela 3.2.2. Na demostração vamos supor que

$$u_j^T b \neq 0, \quad j = 1, \ldots, n. \tag{3.2.26}$$

- Se $\delta < \|x_{LS}\|_2$, usando o Teorema KKT temos que

$$\mathcal{L}'_x(x, \lambda) = 2A^T(Ax - b) + 2\lambda x = 0 \tag{3.2.27}$$

então

$$(A^T A + \lambda I)x = A^T b.$$

Seja $A = U\Sigma V^T$ a decomposição SVD de A. Então a solução x é

$$x = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda}(u_i^T b)v_i$$

onde $\lambda \neq -\sigma_i^2, \ \forall i = 1, \ldots, n$.

Por outro lado, temos que $\|x\|_2^2 \leq \delta^2 < \|x_{LS}\|_2^2$. Usando o fato que V é uma matriz ortonormal segue que

$$\sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2}(u_i^T b)^2 < \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(u_i^T b)^2 \tag{3.2.28}$$

então

$$0 < \sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2} - \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} \right) (u_i^T b)^2.$$

Vejamos os casos, $\lambda = 0$, ou $\lambda < 0$.

Se $\lambda = 0$ então $\mathcal{L}(x, \lambda) = \|Ax - b\|_2^2$ e a solução é $x = x_{LS}$. Mas pela condição da restrição $\|x_{LS}\|_2 = \|x\|_2 = \delta < \|x_{LS}\|_2$ é uma contradição.

Se $\lambda < 0$ segue que

$$\frac{1}{\sigma_i^2} < \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2}, \ \forall i = 1, \ldots, n. \quad \text{e} \quad \lambda \neq -\sigma_i^2.$$

Então

$$\|x_{LS}\|_2 < \|x\|_2 < \|x_{LS}\|_2$$

contradição. Portanto, $\lambda > 0$ e temos que $\bar{x}_\delta = x_\delta$.

- Se $\delta = \|x_{LS}\|_2$ temos

$$\sum_{i=1}^{n} \left(\frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} - \frac{1}{\sigma_i^2} \right) (u_i^T b)^2 = 0.$$

  Assim $\lambda = 0$ é uma solução.

- Se $\delta > \|x_{LS}\|_2$, temos que λ é negativo usando uma idéia análoga ao primeiro caso.

- Se $\delta < \|x_{TLS}\|_2$, então usando os fatores de filtro, temos que

$$\sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} (u_i^T b)^2 < \sum_{i=1}^{n} \frac{\sigma_i^2}{(\bar{\sigma}_{n+1}^2 - \sigma_i^2)^2} (u_i^T b)^2.$$

  Vamos supor, por absurdo, que $\lambda \leq -\bar{\sigma}_{n+1}^2$, então temos que

$$\frac{1}{\sigma_i^2 + \lambda} - \frac{1}{\sigma_i^2 - \bar{\sigma}_{n+1}^2} \geq 0, \ \forall i = 1, \ldots, n$$

  leva a uma contradição com a expressão acima. Portanto $\lambda > -\bar{\sigma}_{n+1}^2$.

- Se $\delta = \|x_{TLS}\|_2$, usando o Teorema 2.1.2 temos que x_{TLS} é uma solução e $\lambda = -\bar{\sigma}_{n+1}^2$, onde $\bar{\sigma}_{n+1}$ é o menor valor singular de $[A \quad b]$.

□

Observação: Se $u_i^T b = 0, \ \forall i = 1, \ldots, n$ então a solução do problema $Ax = b$ é $x = 0$. Lembre que usamos a hipótese (3.2.26) pois estamos procurando a solução não trivial.

As seguintes conclusões podem ser feitas a partir do Teorema acima:

(1) Enquanto $\delta \leq \|x_{LS}\|_2$, o método RTLS produz soluções que são similares às soluções do método de Tikhonov. Ou seja, substituir o resíduo dos quadrados mínimos com o resíduo da formulação TLS não produz novos resultados quando $L = I_n$ e $\delta \leq \|x_{LS}\|_2$.

(2) Já que pelo Corolário 2.1.2 $\|x_{TLS}\|_2 \geq \|x_{LS}\|_2$, vemos que existe uma quantidade grande de valores de δ para o qual o multiplicador λ_{IL} é negativo.

(3) A solução RTLS $\bar{x}_\delta$ é diferente da solução de Tikhonov, e pode ser mais dominada por erros que a solução x_{LS} [14].

3.2.2 Caso da Forma Geral

Em muitas aplicações a matriz L é usada para incorporar restrições de suavidade sobre a solução. Para isso é preciso escolher uma matriz L diferente da matriz identidade. Escolhas frequentes de L incluem

$$L_1 = \begin{bmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}$$

$$L_2 = \begin{bmatrix} -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n}$$

as quais são os operadores discretos da primeira e da segunda derivada respectivamente. Neste caso, a solução RTLS $\bar{x}_\delta$ é diferente da solução de Tikhonov quando o resíduo $Ax - b$ é diferente de zero, já que ambos λ_I e λ_L são não nulos. Pelo Teorema 3.2.2 percebemos que λ_L é sempre positivo quando $\delta < \|x_{TLS}\|_2$, pois o multiplicador de Lagrange μ é positivo para estes valores de λ.

Por outro lado, λ_I é sempre negativo, e adiciona portanto certa deregularização à solucão. Dado δ, existem muitos valores para λ_I e λ_L e portanto muitas soluções x, que satisfazem (3.2.7)-(3.2.9), porém só uma delas resolve o problema de otimização (3.2.5). Segundo (3.2.10) esta solução corresponde ao menor valor de $|\lambda_I|$.

Teorema 3.2.3. *Dado δ, a solução $\bar{x}_\delta$ é relacionada com a solução TLS como segue*

δ	solution	λ_I	λ_L
$\delta < \|Lx_{TLS}\|_2$	$\bar{x}_\delta \neq x_{TLS}$	$\lambda_I < 0 \quad \partial\lambda_I/\partial\delta > 0$	$\lambda_L > 0$
$\delta \geq \|Lx_{TLS}\|_2$	$\bar{x}_\delta = x_{TLS}$	$\lambda_I = -\bar{\sigma}_{n+1}^2$	$\lambda_L = 0$

Demonstração. Ver [14]. □

Note que se a matriz $\lambda_I I_n + \lambda_L L^T L$ é definida positiva, então a solução RTLS corresponde à solução de Tikhonov com termo de penalização $\lambda_I \|x\|_2^2 + \lambda_L \|Lx\|_2^2$. Se a matriz $\lambda_I I_n + \lambda_L L^T L$ é definida negativa ou indefinida, não existe interpretação equivalente.

(1) Se $\delta < \|Lx_{TLS}\|_2$, onde x_{TLS} é a solução (2.1.9), a restrição de desigualdade é obrigatória. O multiplicador de Lagrange μ é positivo e por (3.2.23), segue que $\lambda_L > 0$. De (3.2.22), temos que λ_I é sempre negativo, isto adiciona um pouco de deregularização na solução

(2) O resíduo (3.2.25) é uma função monótona decrescente de δ, e portanto, λ_I é uma função monótona crescente de δ. Se $\delta = \|Lx_{TLS}\|_2$, o multiplicador de Lagrange μ é zero e a solução TLS regularizada $\bar{x}_\delta$ coincide com a solução x_{TLS}; para um δ grande, a desigualdade não é obrigatoria e portanto, a solução naõ muda.

3.3 Regularização TLS via Truncamento

O método de truncamento do problema TLS é usado para problemas mal condicionados. A técnica é similar ao SVD truncado onde os valores singulares pequenos de A são considerados nulos. Em ambos os métodos a informação redundante em A e $[A \quad b]$, respectivamente, associados aos valores singulares pequenos, é descartada e o problema original é substituido por um sistema aproximado no sentido da norma Frobenius. No caso do método TLS, vamos aproximar a matriz $[A \quad b]$, pela matriz de posto k

$$[\widetilde{A}_k \quad \widetilde{b}_k] = \sum_{i=1}^{k} \overline{\sigma}_i \overline{u}_i \overline{v}_i^T$$

e vamos considerar o problema $\widetilde{A}_k x = \widetilde{b}_k$. Quando este sistema é compatível, a solução de norma mínima é chamada de k-ésima solução TLS e é denotado por $x_{\mathrm{TLS}}^{(k)}$. Para construir tal solução procuramos por soluções no espaço nulo da matriz aproximação:

$$\mathcal{N}([\widetilde{A}_k \quad \widetilde{b}_k]) = span\{\overline{v}_{k+1}, \ldots, \overline{v}_{n+1}\}.$$

Isto é, procuramos por soluções da forma

$$\begin{bmatrix} x \\ -1 \end{bmatrix} = \sum_{i=k+1}^{n+1} a_i \overline{v}_i = \begin{bmatrix} \overline{V}_{12} \\ \overline{v}_{22}^T \end{bmatrix} a, \tag{3.3.1}$$

onde $a = [a_{k+1}, \ldots, a_{n+1}]^T \in \mathbb{R}^{n-k+1}$, considerando a partição

$$\overline{V} = [\overline{v}_1, \ldots, \overline{v}_{n+1}] = \begin{bmatrix} \overline{V}_{11} & \overline{V}_{12} \\ \overline{v}_{21}^T & \overline{v}_{22}^T \end{bmatrix}, \tag{3.3.2}$$

onde $\overline{V}_{11} \in \mathbb{R}^{n\times k}$, $\overline{V}_{12} \in \mathbb{R}^{n\times(n-k+1)}$, $\overline{v}_{21} = [[\overline{v}_1]_{n+1}, \ldots, [\overline{v}_k]_{n+1}]^T \in \mathbb{R}^k$, e $\overline{v}_{22} = [[\overline{v}_{k+1}]_{n+1}, \ldots, [\overline{v}_{n+1}]_{n+1}]^T \in \mathbb{R}^{n-k+1}$. Assim, para determinar a solução de norma mínima, note que da ultima equação em (3.3.1) temos

$$\overline{v}_{22}^T a = -1 \Leftrightarrow \sum_{i=k+1}^{n+1} a_i [\overline{v}_i]_{n+1} = -1.$$

Mas de (3.3.1)

$$\left\| \begin{bmatrix} x \\ -1 \end{bmatrix} \right\|_2^2 = 1 + \|x\|_2^2 = \sum_{i=k+1}^{n+1} a_i^2, \tag{3.3.3}$$

vemos que para minimizar a norma da solução, precisamos minimizar o valor de $\sum_{i=k+1}^{n+1} a_i^2$. Isto pode ser feito resolvendo o problema de minimização

$$\min_{a_i} \sum_{i=k+1}^{n+1} a_i^2 \quad \text{sujeito a} \quad \sum_{i=k+1}^{n+1} a_i [\overline{v}_i]_{n+1} = -1.$$

Usando o método dos multiplicadores de Lagrange, a primeira condição de otimalidade para a função Lagrangeana

$$\mathcal{L}(a, \lambda) = \frac{1}{2} \sum_{i=k+1}^{n+1} a_i^2 + \lambda \left(\sum_{i=k+1}^{n+1} a_i [\overline{v}_i]_{n+1} + 1 \right)$$

produz

$$\begin{aligned} a_i + \lambda [\overline{v}_i]_{n+1} &= 0, \ i = k+1, \ldots, n+1, \\ \sum_{i=k+1}^{n+1} a_i [\overline{v}_i]_{n+1} &= -1, \end{aligned}$$

e assim obtemos

$$a = -\frac{1}{\|\overline{v}_{22}\|_2^2} \overline{v}_{22}. \tag{3.3.4}$$

Então, de (3.3.1) e (3.3.4), a solução com norma mínima é dada por:

$$x_{\text{TLS}}^{(k)} = -\frac{1}{\|\overline{v}_{22}\|_2^2} \overline{V}_{12} \overline{v}_{22}. \tag{3.3.5}$$

Note que de (3.3.3), (3.3.4) e o Teorema Eckart-Young-Mirsky, temos:

$$\|x_{\mathrm{TLS}}^{(k)}\|_2^2 = \frac{1}{\|\overline{v}_{22}\|_2^2} - 1$$

e

$$\|R_k\|_F^2 = \|[A \quad b] - [\widetilde{A}_k \quad \widetilde{b}_k]\|_F^2 = \overline{\sigma}_{k+1}^2 + \ldots + \overline{\sigma}_{n+1}^2,$$

mostrando que a norma da k-ésima solução TLS, $\|x_{\mathrm{TLS}}^{(k)}\|_2$ cresce em relação a k, enquanto a norma residual $\|R_k\|_F$ decresce. Para ilustrar estas propriedades usamos o problema teste `shaw` de [17], onde A e b têm 5% de ruído nos dados, e $n = 50$, ver Figura 3.3.

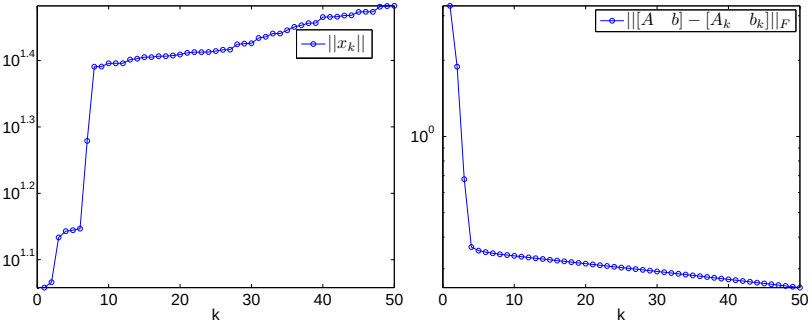

Figura 3.2: Norma da solução aproximada e norma do Resíduo na norma Frobenius.

Como neste problema teste a solução exata é conhecida, para ilustrar o comportamento da qualidade das soluções aproximadas via TLS truncado, em Figura 3.3 mostramos o erro relativo como função de k.

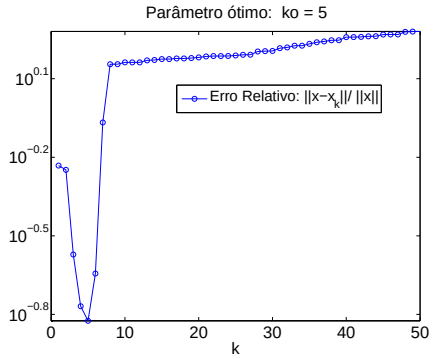

Figura 3.3: Erro relativo na solução TLS truncado.

Vemos que o erro relativo decresce até atingir um valor mínimo e depois começa a crescer. A Figura 3.3 mostra que o metodo TLS truncado atua como método de regularização e que a solução apropriada deve ser construida truncando no índice em que o erro atinge o mínimo. Na prática, a solução exata é desconhecida, logo esse erro não pode ser calculado e portanto o mínimo também não. Ou seja, para que o mẽtodo TLS truncado seja de utilidade, precisamos de métodos que permitam estimar o índice que minimiza o erro. Este assunto será discutido posteriormente.

3.3.1 Fatores de Filtro para o TLS Truncado

Nesta seção vamos buscar uma expressão para os fatores de filtro do método TLS, usando a análise de Fierro (1997) [8]. As decomposições de A e $[A \quad b]$, dadas em (2.1.5) e (2.1.6), respectivamente, serão usadas frequentemente. Para este fim, primeiro vamos obter representações gerais para os valores singulares $\bar{\sigma}_j$ e os vetores singulares a direita $\bar{v}_j$ de $[\ A \quad b\ ]$ em termos do sistema singular $\{(\sigma_i, v_i, u_i)\}$ de A. Vamos assumir que Posto$(A) = n$, Posto$([A \quad b]) = n+1$, e que os valores singulares de A são simples. Também vamos supor que

$$u_j^T b \neq 0, \quad j = 1, \ldots, n.$$

Usando (2.1.5) temos

$$[\ A \quad b\ ]^T [\ A \quad b\ ] = \begin{bmatrix} A^T A & A^T b \\ b^T A & \|b\|_2^2 \end{bmatrix} = \bar{\bar{V}} S \bar{\bar{V}}^T,$$

onde

$$\bar{\bar{V}} = \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix} \tag{3.3.6}$$

e

$$S = \begin{bmatrix} \Sigma^T \Sigma & \Sigma^T U^T b \\ b^T U \Sigma & \|b\|_2^2 \end{bmatrix}.$$

Note que S é uma matriz definida positiva, pois Posto$([A \quad b]) = n+1$. Escrevendo a decomposição em valores singulares de S como

$$S = V_s \Sigma_s^T \Sigma_s V_s^T, \quad \Sigma_s^T \Sigma_s = [\mathrm{diag}(\sigma_{sj}^2)_{(n+1)\times(n+1)}], \tag{3.3.7}$$

temos

$$[\ A \quad b\ ]^T [\ A \quad b\ ] = \bar{\bar{V}} V_s \Sigma_s^T \Sigma_s V_s^T \bar{\bar{V}}^T.$$

Pela equação (2.1.6) obtemos, $\bar{\Sigma}^T\bar{\Sigma} = \Sigma_s^T\Sigma_s$ e $\bar{V} = \bar{\bar{V}}V_s$. Explicitamente, temos

$$\bar{\sigma_j} = \sigma_{sj}, \quad j = 1, \ldots, n+1, \tag{3.3.8}$$

e

$$\bar{v}_j = \bar{\bar{V}}v_{sj}, \quad j = 1, \ldots, n+1, \tag{3.3.9}$$

onde os v_{sj} são os vetores coluna de V_s. No seguinte passo da nossa análise, escrevemos S como

$$S = \begin{bmatrix} \sigma_1^2 & \ldots & 0 & \sigma_1 u_1^T b \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & \sigma_n^2 & \sigma_n u_n^T b \\ \sigma_1 u_1^T b & \ldots & \sigma_n u_n^T b & \|b\|_2^2 \end{bmatrix} \tag{3.3.10}$$

e expressamos (3.3.7) como

$$Sv_{sj} = \sigma_{sj}^2 v_{sj}, \quad j = 1, \ldots, n+1. \tag{3.3.11}$$

Então, de (3.3.10) e (3.3.11) obtemos

$$(\sigma_{sj}^2 - \sigma_i^2)[v_{sj}]_i = \sigma_i(u_i^T b)[v_{sj}]_{n+1}, \quad i = 1, \ldots, n \tag{3.3.12}$$

$$\sum_{i=1}^{n} \sigma_i(u_i^T b)[v_{sj}]_i = \left(\sigma_{sj}^2 - \|b\|_2^2\right)[v_{sj}]_{n+1}. \tag{3.3.13}$$

Para $j = 1, \ldots, n+1$, o sistema singular da matriz S tem duas características interessantes:

(1) $[v_{sj}]_{n+1} \neq 0$;

(2) σ_{sj} não coincide com os valores singulares de A.

Vamos provar estas afirmações. Suponha que $[v_{sj}]_{n+1} = 0$. Neste caso temos duas situações

(1) Vamos supor que existem i_1 e i_2 tal que $[v_{sj}]_{i_1} \neq 0$ e $[v_{sj}]_{i_2} \neq 0$. Então, de (3.3.12), segue que $\sigma_{i_1} = \sigma_{i_2} = \sigma_{sj}$, que é uma contradição, pois os valores singulares de A são simples.

(2) Vamos supor que existe um i tal que $[v_{sj}]_i \neq 0$ e $[v_{sj}]_l = 0$ para todo $l \neq i$. Pela equação (3.3.13) e a hipótese $u_i^T b \neq 0$, segue que $\sigma_i = 0$ que é uma contradição pois temos por hipótese que Posto$(A) = n$.

Portanto, $[v_{sj}]_{n+1} \neq 0$. Voltando agora à segunda afirmação, vamos supor que existe i tal que $\sigma_i = \sigma_{sj}$. Por (3.3.12) temos que $\sigma_i(u_i^T b)[v_{sj}]_{n+1} = 0$. Como $u_i^T b \neq 0$ e $[v_{sj}]_{n+1} \neq 0$, obtemos um resultado contraditório $\sigma_i = 0$. Portanto σ_{sj} não coincide com algum valor singular σ_i de A, e temos

$$[v_{sj}]_i = \frac{\sigma_i}{\sigma_{sj}^2 - \sigma_i^2}(u_i^T b)[v_{sj}]_{n+1}, \quad i = 1, \ldots, n.$$

A segunda afirmação junto com (3.3.8) implica que as desigualdades de interlacing, para os valores singulares de A e $[A \quad b]$ são estritas, i.e.,

$$\bar{\sigma}_1 > \sigma_1 > \ldots > \bar{\sigma}_k > \sigma_k > \bar{\sigma}_{k+1} > \ldots > \sigma_n > \bar{\sigma}_{n+1}, \tag{3.3.14}$$

onde k é o índice de truncamento do método TLS truncado. Agora podemos derivar uma expressão final para os vetores singulares a direita de $[\ A \quad b\ ]$. Por (3.3.6) e (3.3.9), temos

$$\bar{v}_j = \bar{V} v_{sj} = \begin{bmatrix} V \begin{bmatrix} [v_{sj}]_1 \\ \vdots \\ [v_{sj}]_n \end{bmatrix} \\ [v_{sj}]_{n+1} \end{bmatrix},$$

logo as entradas do vetor singular a direita $\bar{v}_j$ são dados por

$$\begin{bmatrix} [\bar{v}_j]_1 \\ \vdots \\ [\bar{v}_j]_n \end{bmatrix} = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_{sj}^2 - \sigma_i^2}(u_i^T b)[v_{sj}]_{n+1} v_i \tag{3.3.15}$$

e

$$[\bar{v}_j]_{n+1} = [v_{sj}]_{n+1}, \tag{3.3.16}$$

para $j = 1, \ldots, n+1$. Resumimos o resultado acima no seguinte Teorema.

Teorema 3.3.1. *Seja (2.1.5) a decomposição em valores singulares da matriz A e assuma que* $\text{Posto}(A) = n$ *e* $\text{Posto}([A \quad b]) = n + 1$. *Além disso, assuma (3.2.26). Se (3.3.7) é a decomposição em valores singulares da matriz S definida por (3.3.10), então os valores singulares da matriz $[A \quad b]$ são dados por (3.3.8), enquanto as entradas dos vetores singulares à direita são dados por (3.3.15) e (3.3.16).*

O seguinte Teorema descreve os fatores de filtro para a solução TLS truncada

$$x_{TLS}^{(k)} = -\frac{1}{\|\bar{v}_{22}\|_2^2} \bar{V}_{12} \bar{v}_{22}, \tag{3.3.17}$$

Teorema 3.3.2. *Com as mesmas hipóteses do Teorema 3.3.1, os fatores de filtro para a solução TLS truncada, são dados por*

$$f_i = -\frac{1}{\|\bar{v}_{22}\|_2^2}\sum_{j=k+1}^{n+1}\frac{\sigma_i^2}{\bar{\sigma}_j^2-\sigma_i^2}[\bar{v}_j]_{n+1}^2 = \frac{1}{\|\bar{v}_{22}\|_2^2}\sum_{j=1}^{k}\frac{\sigma_i^2}{\bar{\sigma}_j^2-\sigma_i^2}[\bar{v}_j]_{n+1}^2 \tag{3.3.18}$$

para $i = 1\ldots, n$.

Demonstração. Usando (3.3.15) junto com (3.3.8) e (3.3.16), expressamos a solução TLS truncada (3.3.17) como

$$\begin{aligned} x_{TLS}^{(k)} &= -\frac{1}{\|\bar{v}_{22}\|_2^2}\bar{V}_{12}\bar{v}_{22} \\ &= -\frac{1}{\|\bar{v}_{22}\|_2^2}\sum_{j=k+1}^{n+1}[\bar{v}_j]_{n+1}\begin{bmatrix}[\bar{v}_j]_1\\ \vdots \\ [\bar{v}_j]_n\end{bmatrix} \\ &= -\frac{1}{\|\bar{v}_{22}\|_2^2}\sum_{i=1}^{n}\left(\sum_{j=k+1}^{n+1}\frac{\sigma_i^2}{\bar{\sigma}_j^2-\sigma_i^2}[\bar{v}_j]_{n+1}^2\right)\frac{1}{\sigma_i}(u_i^T b)v_i. \end{aligned}$$

Para obter a segunda representação em (3.3.18), usamos a condição de ortogonalidade $V_s V_s^T = I_{n+1}$. Então

$$\sum_{j=1}^{n+1} v_{sj}v_{sj}^T = I_{n+1}$$

implica

$$\sum_{j=1}^{n+1}[v_{sj}]_i[v_{sj}]_{n+1} = 0, \quad i = 1,\ldots,n \tag{3.3.19}$$

e

$$\sum_{j=1}^{n+1}[v_{sj}]_{n+1}^2 = 1. \tag{3.3.20}$$

Por outro lado, de (3.3.12) junto com (3.3.8), temos

$$\sigma_i(u_i^T b)\frac{[v_{sj}]_{n+1}^2}{\bar{\sigma}_j^2-\sigma_i^2} = [v_{sj}]_i\ [v_{sj}]_{n+1}, \quad i = 1,\ldots,n. \tag{3.3.21}$$

Agora, (3.3.19) e (3.3.21) junto com (3.2.26) e (3.3.16) produzem

$$\sum_{j=1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{v}_j]_{n+1}^2 = 0, \quad i = 1, \ldots, n,$$

e portanto,

$$\sum_{j=k+1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{v}_j]_{n+1}^2 = -\sum_{j=1}^{k} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{v}_j]_{n+1}^2, \quad i = 1, \ldots, n.$$

A demonstração está completa. □

Os fatores de filtro do TLS truncado podem ser estimados como seguem.

Teorema 3.3.3. *Com as mesmas hipóteses do Teorema 3.3.1, os fatores de filtro satisfazem*

$$0 < f_i - 1 \leq \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2 - \bar{\sigma}_{k+1}^2}, \quad i = 1, \ldots, k \tag{3.3.22}$$

e

$$0 < f_i \leq \frac{1 - \|\bar{v}_{22}\|_2^2}{\|\bar{v}_{22}\|_2^2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2}, \quad i = k+1, \ldots, n. \tag{3.3.23}$$

Demonstração. Para $i = 1, \ldots, k$, usamos a primeira representação em (3.3.18) e o resultado

$$\|\bar{v}_{22}\|_2^2 = \sum_{j=k+1}^{n+1} [\bar{v}_j]_{n+1}^2 \tag{3.3.24}$$

para obter

$$\begin{aligned} f_i &= \frac{1}{\|\bar{v}_{22}\|_2^2} \sum_{j=k+1}^{n+1} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{v}_j]_{n+1}^2 \\ &= \frac{1}{\|\bar{v}_{22}\|_2^2} \sum_{j=k+1}^{n+1} \frac{\sigma_i^2 - \bar{\sigma}_j^2 + \bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{v}_j]_{n+1}^2 \\ &= 1 + \frac{1}{\|\bar{v}_{22}\|_2^2} \sum_{j=k+1}^{n+1} \frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{v}_j]_{n+1}^2. \end{aligned} \tag{3.3.25}$$

Pela desigualdade de interlacing para os valores singulares de A e $[A \quad b]$ dados por (3.3.14), vemos que, para $i = 1, \ldots, k$, temos

$$\sigma_i > \bar{\sigma}_{k+1} = \max_{j=k+1,\ldots,n+1}(\bar{\sigma}_j).$$

Portanto, o segundo termo de (3.3.25) é positivo. Ainda mais, de $\bar{\sigma}_j \leq \bar{\sigma}_{k+1}$, para $j = k+1, \ldots, n+1$ deduzimos que

$$\frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} \leq \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2 - \bar{\sigma}_{k+1}^2},$$

e portanto,

$$\sum_{j=k+1}^{n+1} \frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2}[\bar{v}_j]_{n+1}^2 \leq \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2 - \bar{\sigma}_{k+1}^2} \sum_{j=k+1}^{n+1} [\bar{v}_j]_{n+1}^2.$$

Este resultado junto com (3.3.24) implica a desigualdade (3.3.22).

Para $i = k+1, \ldots, n$, consideremos a segunda representação em (3.3.18), isto é

$$f_i = \frac{1}{\|\bar{v}_{22}\|_2^2} \sum_{j=1}^{k} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2}[\bar{v}_j]_{n+1}^2.$$

De (3.3.14), vemos que, para $i = k+1, \ldots, n$, temos

$$\sigma_i < \bar{\sigma}_k = \min_{j=1,\ldots,k}(\bar{\sigma}_j).$$

Como $\bar{\sigma}_j \geq \bar{\sigma}_k$ para $j = 1, \ldots, k$, temos

$$f_i = \frac{1}{\|\bar{v}_{22}\|_2^2} \sum_{j=1}^{k} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2}[\bar{v}_j]_{n+1}^2 \leq \frac{1}{\|\bar{v}_{22}\|_2^2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2} \sum_{j=1}^{k} [\bar{v}_j]_{n+1}^2.$$

Finalmente, de (3.3.16) e (3.3.20) obtemos

$$\begin{aligned} \sum_{j=1}^{k} [\bar{v}_j]_{n+1}^2 &= 1 - \sum_{j=k+1}^{n} [\bar{v}_j]_{n+1}^2 \\ &= 1 - \|\bar{v}_{22}\|_2^2, \end{aligned}$$

o que mostra a desigualdade (3.3.23). □

3.3.2 Bidiagonalização de Lanczos para o TLS truncado

Quando a dimensão da matriz A não é muito grande, a decomposição em valores singulares da matriz aumentada $[A \quad b]$ pode ser calculada diretamente e, com isso, a solução x_{TLS} pode ser calculada eficientemente. Porém a SVD não é viável para problemas de grande porte devido ao alto custo computacional. Neste caso vamos usar o algoritmo baseado na bidiagonalização de Lanczos [13], chamado algoritmo TLS truncado de Lanczos.

Este algoritmo usa a bidiagonalização da matriz A para obter, depois de k iterações, a fatoração

$$A\widehat{V}_k = \widehat{U}_{k+1}\widehat{B}_k, \tag{3.3.26}$$

onde $\widehat{U}_{k+1} \in \mathbb{R}^{m\times(k+1)}$, $\widehat{V}_k \in \mathbb{R}^{n\times k}$ são matrizes com colunas ortonormais e $\widehat{B}_k \in \mathbb{R}^{(k+1)\times k}$ tem a forma

$$\widehat{B}_k = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix}.$$

Esta fatoração projeta o problema TLS sobre o espaço gerado pelas colunas $\widehat{U}_{k+1} \in \mathbb{R}^{m\times(k+1)}$. A projeção é consequencia de assumir que para um k suficientemente grande, os valores singulares de A que contribuem à solução regularizada já foram capturados. O problema projetado é dado por

$$\begin{aligned} &\min_{[\widetilde{A}_k \;\; \widetilde{b}_k]\in\mathbb{R}^{m\times(n+1)}} \left\| \widehat{U}_{k+1}^T \left([A \quad b] - [\widetilde{A}_k \quad \widetilde{b}_k] \right) \begin{bmatrix} \widehat{V}_k & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \right\|_F^2 \\ &\textit{sujeito a } \widehat{U}_{k+1}^T \widetilde{A}_k \widehat{V}_k z_k = \widehat{U}_{k+1}^T \widetilde{b}_k, \end{aligned}$$

onde $x = \widehat{V}_k z_k$ para algum $z_k \in \mathbb{R}^k$. Usando (3.3.26) temos que

$$\begin{aligned} \widehat{U}_{k+1}^T [A \quad b] \begin{bmatrix} \widehat{V}_k & 0 \\ 0 & 1 \end{bmatrix} &= [\; \widehat{U}_{k+1}^T A \widehat{V}_k \quad \widehat{U}_{k+1}^T b \;] \\ &= [\; B_k \quad \beta_1 e_1^{(k+1)} \;] \end{aligned} \tag{3.3.27}$$

assim o problema de minimização (2.1.1), pode ser escrito como

$$\min_{[\widetilde{B}_k \ \widetilde{e}_k]\in\mathbb{R}^{(k+1)\times(k+1)}} \left\| [\ B_k \quad \beta_1 e_1^{(k+1)}\ ] - [\ \widetilde{B}_k \quad \widetilde{e}_k\ ] \right\|_F^2 \qquad (3.3.28)$$
$$\textit{sujeito a } \widetilde{B}_k z_k = \widetilde{e}_k,$$

onde $\widetilde{B}_k = \widehat{U}_{k+1}^T \widetilde{A}_k \widehat{V}_k$, $\widetilde{e}_k = \widehat{U}_{k+1}^T \widetilde{b}_k$ e $e_1^{(k+1)} \in \mathbb{R}^{k+1}$ é o primeiro vetor canónico. Então em cada passo de Lanczos, usamos o algoritmo TLS para problemas de pequena dimensão (3.3.28) para calcular a solução TLS truncada. De maneira precisa, assumindo a decomposição em valores singulares

$$[\ B_k \quad \beta_1 e_1^{(k+1)}\ ] = \bar{\bar{U}}^{(k)} \bar{\bar{\Sigma}}^{(k)} \bar{\bar{V}}^{(k)T},$$

onde

$$\bar{\bar{V}}^{(k)} = \begin{bmatrix} \bar{\bar{V}}_{11}^{(k)} & \bar{\bar{V}}_{12}^{(k)} \\ \bar{\bar{V}}_{21}^{(k)T} & \bar{\bar{v}}_{22}^{(k)} \end{bmatrix},\ \bar{\bar{V}}_{11}^{(k)} \in \mathbb{R}^{k\times k},\ \bar{\bar{V}}_{12}^{(k)}, \bar{\bar{V}}_{21}^{(k)} \in \mathbb{R}^k,$$

a solução TLS de (3.3.28) é

$$z_k = -\frac{1}{\bar{\bar{v}}_{22}^{(k)}} \bar{\bar{V}}_{12}^{(k)},$$

e a matriz de aproximação é

$$[\widetilde{B}_k \quad \widetilde{e}_k] = \sum_{i=1}^{k} \bar{\bar{\sigma}}_i^{(k)} \bar{\bar{u}}_i^{(k)} \bar{\bar{v}}_i^{(k)T}.$$

Usando o método do Capítulo 2 o resíduo é dado por

$$\|[B_k \quad \beta_1 e^{(k+1)}] - [\widetilde{B}_k \quad \widetilde{e}_k]\|_F = \bar{\bar{\sigma}}_{k+1}^{(k)}.$$

Além disso, pelo Teorema 1.2.5 aplicado nas matrizes $[B_k \quad \beta_1 e_1^{(k+1)}]$ e $[B_{k+1} \quad \beta_1 e_1^{(k+2)}]$, o resíduo é uma função decrescente em k. A solução truncada TLS é da forma

$$x_k = \widehat{V}_k z_k = -\frac{1}{\bar{\bar{v}}_{22}} \widehat{V}_k \bar{\bar{V}}_{12}.$$

e

$$[\widetilde{A}_k \quad \widetilde{b}_k] = [\widehat{U}_{k+1} \widetilde{B}_k \widehat{V}_k^T \quad \widehat{U}_{k+1} \widetilde{e}_k].$$

Usando o fato de que $\widehat{U}_{k+1}$, $\widehat{V}_k$ são ortonormais, de (3.3.27) o resíduo é

$$\begin{aligned}
\left\|R_k\right\|_F &= \left\|[A \quad b] - [\widetilde{A}_k \quad \widetilde{b}_k]\right\|_F \\
&= \left\|[A \quad b] - [\widehat{U}_{k+1}\widetilde{B}_k\widehat{V}_k^T \quad \widehat{U}_{k+1}\widetilde{e}_k]\right\|_F \\
&= \left\|\widehat{U}_{k+1}^T\left([A \quad b] - [\widehat{U}_{k+1}\widetilde{B}_k\widehat{V}_k^T \quad \widehat{U}_{k+1}\widetilde{e}_k]\right)\begin{bmatrix}\widehat{V}_k & 0\\ 0 & 1\end{bmatrix}\right\|_F \\
&= \left\|[\widehat{U}_{k+1}^T A\widehat{V}_k \quad \widehat{U}_{k+1}^T b] - [\widetilde{B}_k \quad \widetilde{e}_k]\right\|_F \\
&= \left\|[B_k \quad \beta_1 e_1^{(k+1)}] - [\widetilde{B}_k \quad \widetilde{e}_k]\right\|_F
\end{aligned}$$

Portanto, no algoritmo TLS truncado de Lanczos, a norma da solução aproximada $\|x_k\|$ também cresce com k e a norma residual $\|R_k\|_F$ também decresce com k.

Para este tipo de problemas é importante ter alguma estimativa do valor de k adequado. O assunto será abordado no capítulo 5.

Capítulo 4

Métodos para Calcular Soluções TLS Regularizadas

O objetivo deste capítulo é discutir um método para calcular soluções TLS regularizadas (RTLS) baseado num trabalho de Guo e Renault [30], bem como um método de regularização baseado em truncamento e na escolha de um índice apropriado de truncamento introduzido aqui.

4.1 RTLS Via Problema de Autovalores segundo Renault e Guo

Para descrever o método, é importante lembrar que, devido ao Teorema 3.2.1, como os parâmetros λ_I e λ_L dependem de x (que denotamos aqui por $\bar{x}_\delta$), o sistema de equações (3.2.7) é muito difícil de resolver. Para contornar esta dificultade, o conjunto de equações (3.2.7), (3.2.8) e (3.2.24) tem sido usado de duas maneiras diferentes para resolver o problema RTLS. Em [14, 22, 31] λ_I é escolhido como parâmetro livre e, para λ_L fixado, o problema (3.2.7) é resolvido fornecendo um par (x, λ_I), e a seguir o parâmetro λ_L é atualizado usando (3.2.24). O processo continua até ter a convergência da sequência. Reciprocamente, em [21, 33] para λ_I escolhido, a solução de (3.2.7) fornece um par (x, λ_L) e o método gera uma sequência convergente de parâmetros λ_I. Em ambos os casos, o sistema (3.2.7) é resolvido através de um

problema de autovalores linear e quadrático respectivamente.

O método está baseado na observação que o cálculo de $\bar{x}_\delta$ requer a solução de um problema de autovalores que provém das equações (3.2.7), (3.2.8) e (3.2.24). De fato, reescrevendo essas equações temos

$$B(\lambda_L)\begin{bmatrix} x \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} x \\ -1 \end{bmatrix}, \tag{4.1.1}$$

onde

$$B(\lambda_L) = \begin{bmatrix} A^TA + \lambda_L(x)L^TL & A^Tb \\ b^TA & -\lambda_L(x)\delta^2 + b^Tb \end{bmatrix}$$

é uma matriz $(n+1)\times(n+1)$, $\lambda = -\lambda_I$, com λ_I e λ_L dados por (3.2.22) e (3.2.24), e x satisfazendo a restrição $\|Lx\|_2 = \delta$.

Para resolver estes problemas de autovalores, os métodos iterativos reutilizam a maior quantidade de informação possível da iteração prévia. Guo e Renault [30] usaram o método da potência inversa para obter o auto-par $\left(\lambda, \begin{bmatrix} x \\ -1 \end{bmatrix}\right)$. Para este método vamos assumir que

$$\sigma_{min}([AK \quad b]) < \sigma_{min}(AK),$$

onde K é uma base ortonormal do núcleo de L. Em [2] foi provado que esta condição é suficiente para que o algoritmo possa atingir a solução de (4.1.1). No seguinte Teorema mostramos que a solução x_{RTLS} é caracterizada por um sistema autovalor autovetor. Renaut e Guo [31] deram condições de otimalidade de primeira ordem para este problema.

Teorema 4.1.1. *A solução x_{RTLS} do problema RTLS (3.2.5) sujeito a constantes ativas satisfaz o problema de autovalor aumentado*

$$B(\lambda_L)\begin{bmatrix} x_{RTS} \\ -1 \end{bmatrix} = -\lambda_I \begin{bmatrix} x_{RTS} \\ -1 \end{bmatrix}, \tag{4.1.2}$$

com

$$B(\lambda_L) = M + \lambda_L N, \quad M := [A \quad b]^T[A \quad b], \quad N := \begin{bmatrix} L^TL & 0 \\ 0 & -\delta^2 \end{bmatrix}$$

e λ_I e λ_L são dados por (3.2.22) e (3.2.23) respectivamente. Reciprocamente, se $\left([\tilde{x}^T - 1]^T, -\tilde{\lambda}\right)$ é o auto par de $B(\lambda_L)$, onde $\lambda_L(\tilde{x})$ satisfaz (3.2.23), então $\tilde{x}$ satisfaz (3.2.7), e $\tilde{\lambda} = -\frac{\|b-A\tilde{x}\|_2}{1+\|\tilde{x}\|_2^2}$.

Demonstração. Ver [31]. □

A escolha do menor autovalor é motivado por (3.2.25), pois precisamos minimizar o resíduo na norma Frobenius. Para tanto usamos o método da potência inversa, que calcula em poucos passos o menor autovalor em valor absoluto e o autovetor correspondente.

É conhecido, pela equação (3.2.25) que a solução TLS minimiza o problema

$$x_{TLS} = \underset{x}{\operatorname{argmin}}\, \phi(x) = \underset{x}{\operatorname{argmin}}\, \frac{\|Ax - b\|_2^2}{1 + \|x\|_2^2}, \tag{4.1.3}$$

onde ϕ é chamado o quociente de Rayleigh da matriz M. Isto sugere uma formulação alternativa para o método RTLS.

$$\min_x \phi(x) \quad \text{s.a.} \quad \|Lx\|_2 \leq \delta.$$

4.1.1 Suporte Teórico do Método Iterativo

Pela igualdade (4.1.3) e o Teorema 4.1.1, a solução RTLS é obtida calculando estimativas do menor $|\lambda_I| = \phi(x)$ que resolve o problema de autovalores. Quando o sistema (4.1.2) é satisfeito as condições ativas também são. Assim, para descrevermos o algoritmo de Renault e Guo, vamos introduzir matriz dependente de um parâmetro, $\mathbf{B}(\theta) = M + \theta N$, $\theta > 0$, e vamos denotar o menor autovalor correspondente ao autovetor $[x_\theta^T \ -1]^T$ de $\mathbf{B}(\theta)$ por ϱ_{n+1}. Também vamos introduzir a função

$$g(x) = (\|Lx\|_2^2 - \delta^2)/(1 + \|x\|_2^2). \tag{4.1.4}$$

Tendo introduzido essas notações, dada uma constante δ, o objetivo é determinar θ tal que $g(x_\theta) = 0$. Este é um problema de cálculo de raízes que pode ser resolvido de diferentes maneiras, como por exemplo pelo método da bisseção. Os seguintes resultados formam a base do algoritmo para o problema.

Lema 4.1.1. *Suponha que $b^T A \neq 0$ e $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$, então o menor valor singular da matriz $\boldsymbol{B}(\theta)$ é simples.*

Demonstração. Pela equação autovalor- autovetor

$$\mathbf{B}(\theta) \begin{bmatrix} x_\theta \\ -1 \end{bmatrix} = \varrho_\theta \begin{bmatrix} x_\theta \\ -1 \end{bmatrix}$$

temos

$$(A^T A + \theta L^T L - \varrho_\theta I) x_\theta = A^T b.$$

Pela hipótese $A^T b \neq 0$, segue então que ϱ_θ não é um autovalor de $A^T A + \theta L^T L$. Como $A^T A + \theta L^T L$ é uma submatriz de $\mathbf{B}(\theta)$ pelo Teorema de

interlacing de autovalores segue que ϱ_θ é estritamente menor que o mínimos dos autovalores de $A^TA + \theta L^TL$, e portanto é simple. □

Lema 4.1.2. *Se* $[A \quad b]$ *é uma matriz de posto completo, existe um único valor positivo, denotado por* θ^c*, tal que* $\boldsymbol{B}(\theta^c)$ *é singular, e*

1. *O autovalor nulo de* $\boldsymbol{B}(\theta^c)$ *é simples.*
2. *Quando* $0 \leq \theta < \theta^c$*,* $\boldsymbol{B}(\theta)$ *é definida positiva.*
3. *Quando* $\theta > \theta^c$*,* $\boldsymbol{B}(\theta)$ *tem só um autovalor negativo; e os outros são positivos.*

Demonstração. Ver [31]. □

Lema 4.1.3. *Se* $b^TA \neq 0$*, e* $[A \quad b]$ *é uma matriz e posto completo, então*

1. *Existe* $\lambda_L^* \in [0, \theta^c]$ *que resolve o problema* $\|Lx_{RTLS}\|_2^2 = \delta^2$*,*
2. *esta solução é única,*
3. *quando* $\lambda_L \in (0, \lambda_L^*),\ g(x_{\lambda_L}) > 0$ *e* $\lambda_L \in (\lambda_L^*, \infty),\ g(x_{\lambda_L}) < 0.$

Demonstração. Ver [31]. □

Este resultado diz que existe uma única solução de nosso problema e que o algoritmo que encontra esta solução depende de atualizar o parâmetro λ_L e do sinal de $g(x_{\lambda_L})$. De (3.2.23) é imediato que x_θ esta relacionado com θ por

$$\theta = \frac{1}{\delta^2}(b^T(b - Ax_\theta) - \phi(x_\theta)).$$

Isto sugere um método iterativo para calcular θ,

$$\theta^{(k+1)} = \frac{1}{\delta^2}(b^T(b - Ax_{\theta^{(k)}}) - \phi(x_{\theta^{(k)}})),$$

onde no passo k, $[x_{\theta^{(k)}}^T \quad -1]^T$ é o autovetor de $\varrho_{n+1}^{(k)}$. Multiplicando a equação (4.1.1) por $[x^T \quad -1]$ temos a seguinte relação

$$\lambda = -\frac{1}{\|x\|_2^2 + 1}(\|Ax - b\|_2^2 + \lambda_L(x)(\|Lx\|_2^2 - \delta^2)).$$

Assim temos que a iteração para $\varrho_{n+1}^{(k)}$ é

$$\varrho_{n+1}^{(k)} = \phi(x_{\theta^{(k)}}) + \theta^{(k)} g(x_{\theta^{(k)}}).$$

Também usando a relação

$$b^T A x_{\theta^{(k)}} - b^T b + \delta^2 \theta^{(k)} = -\varrho_{n+1}^{(k)}$$

temos uma equação que atualiza $\theta^{(k)}$ a través de:

$$\theta^{(k+1)} = \theta^{(k)} + \frac{\theta^{(k)}}{\delta^2} g(x_{\theta^{(k)}}). \tag{4.1.5}$$

Resta demonstrar que esta iteração vai gerar o valor apropriado θ que resolve o problema

$$\|Lx_{RTLS}\|_2^2 = \delta^2. \tag{4.1.6}$$

Para investigar as propriedades de convergência da equação (4.1.5), vamos usar o parâmetro λ_L em lugar de θ. Os resultados do Lema 4.1.3 sugerem o uso de um parâmetro dependente $0 < \iota^{(k)} \leq 1$ escolhido tal que $g(x_{\lambda_L^{(k+1)}})$ tem o mesmo sinal do que $g(x_{\lambda_L^{(0)}})$.

$$\lambda_L^{(k+1)} = \lambda_L^{(k)} + \iota^{(k)} \frac{\lambda_L^{(k)}}{\delta^2} g(x_{\lambda_L^{(k)}}), \quad 0 < \iota^{(k)} \leq 1. \tag{4.1.7}$$

Lema 4.1.4. *Suponha que $\lambda_L^{(0)} > 0$. Sejam as sequências $\{x_{\lambda_L^{(k)}}\}$ e $\lambda_L^{(k)}$, $k = 1, 2, \ldots$, gerados por (4.1.7), com a sequência e parâmetros $0 < \iota^{(k)} \leq 1$ tal que $g(x_{\lambda_L^{(k+1)}}) g(x_{\lambda_L^{(0)}}) > 0$. Então*

1. *$\lambda_L^{(k)} > 0$ para qualquer inteiro positivo k.*
2. *Se $g(x_{\lambda_L^{(0)}}) < 0$, então as sequências $\{\lambda_L^{(k)}\}$ e $\{\phi(x_{\lambda_L^{(k)}})\}$ decrescem de forma monótona, mas $\{\varrho_{n+1}^{(k)}\}$ e $\{\phi(x_{\lambda_L^{(k)}})\}$ crescem de forma monótona.*
3. *Se $g(x_{\lambda_L^{(0)}}) > 0$, então as sequências $\{\lambda_L^{(k)}\}$ e $\{\phi(x_{\lambda_L^{(k)}})\}$ crescem de forma monótona, mas $\{\varrho_{n+1}^{(k)}\}$ e $\{\phi(x_{\lambda_L^{(k)}})\}$ decrescem de forma monótona.*
4. *Se $g(x_{\lambda_L^{(0)}}) = 0$, $\lambda_l^{(0)}$ resolve o problema 4.1.6.*

Demonstração. Ver [31]. □

Note que para un dado inicial $0 < \lambda_L^{(0)} < \theta^c$, a tendência da sequência monótona gerada depende se $\lambda_L^{(0)} < \lambda_L^*$ ou $\lambda_L^{(0)} > \lambda_L^*$, mas em qualquer caso $\mathbf{B}(\lambda_L^{(k)})$ é sempre definida positiva.

Teorema 4.1.2. *A iteração (4.1.7) com um dado inicial* $\lambda_L^{(0)} > 0$ *converge à unica solução* λ_L^* *do problema 4.1.6.*

Demonstração. Ver [31]. □

Com estes resultados teóricos, escrevemos o algoritmo básico que calcula a solução RTLS (3.2.5) via problema de autovalores (RTLSEVP) o qual usa o menor autovalor em módulo para cada parâmetro de Lagrange λ_L.

Algoritmo RTLSEVP:

Dados de entrada : δ, $\lambda_L^{(0)} > 0$ e $\varepsilon > 0$.

1. Calcular o autopar $(\varrho_{n+1}^{(0)}, [x^{(0)^T}, -1]^T)$ de $B(\lambda_L^{(0)})$.
   Faça $k = 0$, $\iota = 1$.
2. No passo k
   Se $k > 0$ e $g(x^{(k+1)})g(x^{(k)}) < 0$, então $\iota = \iota/2$
   senão $\iota = 1$, $k = k + 1$.
3. Atualize λ_L
   $\lambda_L^{(k)} = \lambda_L^{(k-1)} \left(1 + \frac{\iota}{\delta^2} g(x^{(k-1)})\right)$
4. Atualize x
   Calcule o menor autovalor ϱ_{n+1}^{k+1} e o correspondente autovetor $[x^{(k)^T}, -1]^T$ de $B(\lambda_L^{(k)})$
5. Critério de parada
   Se $|g(x^{(k)})| < \varepsilon$, então Fim
   senão $k = k + 1$
   $x_{RTLS} = x^{(k)}$

4.1.2 Experimentos Numéricos

Apresentamos resultados numéricos usando problemas do pacote Regularization Tools de P. C. Hansen [17]. O pacote implementa um conjunto de métodos através de rotinas na linguagem `MATLAB` destinadas ao tratamento de problemas discretos mal postos (criação, análise e resolução), que inclui vários problemas teste com uma rotina para cada um deles.

Os problemas utilizados provém da discretização de equações integrais de Fredholm de primeira espécie da forma

$$\int_a^b K(x,y)f(y)dy = g(x), \quad c \le x \le d.$$

Em todos os casos, a Toolbox fornece a tripla $\{\bar{A}_{ex},\ \bar{b}_{ex},\ \bar{x}_{ex}\}$ tal que

$$\bar{A}_{ex}\bar{x}_{ex} = \bar{b}_{ex}.$$

Como o pacote do Hansen trabalha com matrizes quadradas e nosso interesse é lidar com problemas sobredeterminados ($m \geq n$), em nossos exemplos numéricos consideramos problemas com A_{ex} e b_{ex} definidos por

$$A_{ex} = \begin{bmatrix} \bar{A}_{ex} \\ \bar{A}_{ex} \end{bmatrix}, \quad b_{ex} = \begin{bmatrix} \bar{b}_{ex} \\ \bar{b}_{ex} \end{bmatrix}.$$

Note que a igualdade $A_{ex}x_{ex} = b_{ex}$ é ainda satisfeita, onde $x_{ex} = \bar{x}_{ex}$.

Para testar o método RTLS, adicionamos ruído aos dados exatos para obter um sistema $Ax \approx b$, como segue

$$A = A_{ex} + \|A_{ex}\|_2 0.01(NR_A)(E/\|E\|_2)$$

e

$$b = b_{ex} + \|b_{ex}\|_2 0.01(NR_b)(e/\|e\|_2),$$

onde NR_A e NR_b são os níveis de ruído relativo da matriz A_{ex} e o vetor b_{ex} respectivamente. Por exemplo para $NR_A = 5$, a matriz resultante A tem 5% de ruído nos dados. A matriz E e o vetor e de perturbações são gerados pela rotina *randn* do `MATLAB`. Neste trabalho consideramos o caso $NR_A = NR_b = NR$ e denotamos por $\sigma = 0.01NR$, e para cada problema analisamos a solução com diferentes níveis de ruído, $\sigma = 0.001,\ 0.01,\ 0.05$, o que corresponde a 0.1%, 1% e 5%, respectivamente. Além disso, em todos nossos exemplos a solução regularizada correspondente ao problema LS (RLS) também é calculada e sua qualidade é comparada com aquela da solução RTLS. O parâmetro de regularização λ no caso LS é determinado pelo método do ponto-fixo como descrito em [6].

Em nossos exemplos consideramos a matriz $A \in \mathbb{R}^{400\times 200}$, a matriz L é o operador primeira derivada, e $\delta = 0.95\|Lx_{ex}\|_2$. Com a finalidade de usar eficientemente o programa `MATLAB` para o algoritmo RTLS e usar só iterações que adicionam novas informações na sequência, consideramos o critério e parada

$$|g(x^{(k)})| < \varepsilon, \tag{4.1.8}$$

onde $\varepsilon > 0$ é um numéro próximo de zero e a função g é definida em (4.1.4). Nos experimentos numéricos usamos $\epsilon = 10^{-4}$ e o parâmetro inicial $\lambda_L^{(0)} = 0,5$. Também, para evitar um número grande de passos

para certos problemas onde o critério acima fica muito exigente, vamos considerar como critério alternativo de parada o primeiro k tal que

$$\left| \frac{\lambda_I^{(k+1)} - \lambda_I^{(k)}}{\lambda_I^{(k)}} \right| < 5 \times 10^{-3} \tag{4.1.9}$$

Ou seja, o processo iterativo é interrompido quando algum dos critérios acima é satisfeito e o índice de parada é denotado por k_m.

Heat

O problema heat envolve uma equação integral de Volterra de primeira espécie com intervalo de integração $[0, 1]$. O núcleo é $K(s,t) = k(s-t)$ onde

$$k(t) = \frac{t^{-3/2}}{2\kappa\sqrt{\pi}} \exp\left(-\frac{1}{4\kappa^2 t}\right).$$

O parâmetro κ controla o mal condicionamento da matriz A, para nosso trabalho consideramos $\kappa = 1$ o qual produz uma matriz mal condicionada. A Figura 4.1 mostra a solução deste problema.

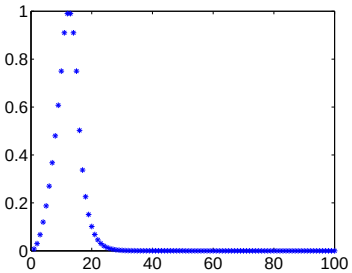

Figura 4.1: Solução do problema heat.

A Figura 4.2 mostra resultados obtidos para o caso $\sigma = 0.001$. Neste caso, o critério de parada atingido no algoritmo RTLSEVP é atingido em $k = 334$ e a solução RTLS produz um erro relativo 0.0669 (aprox. 7%). Já a solução RLS produz um erro relativo 0.0387 (aprox. 4%). Embora isso, visualmente ambas as soluções parecem idênticas, ver Figura 4.2 (linha inferior).

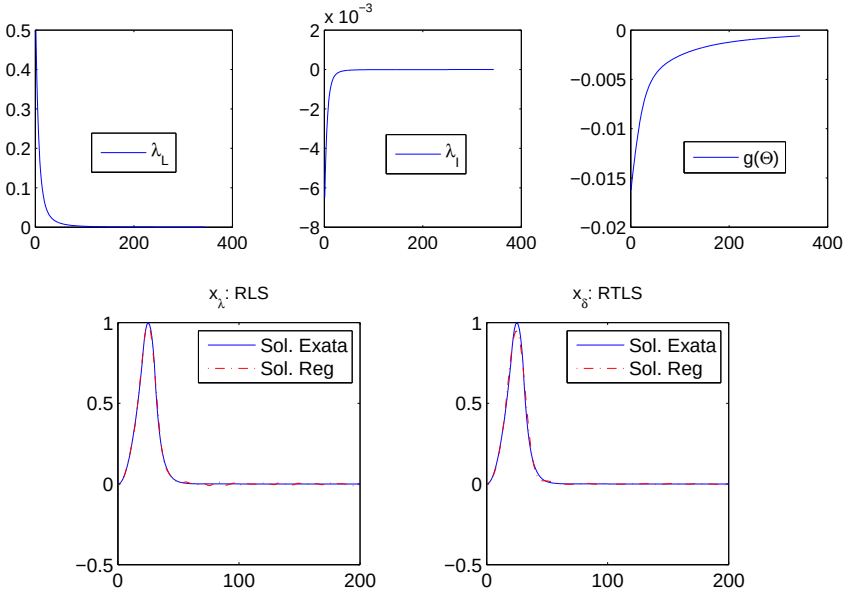

Figura 4.2: Linha superior: Parâmetros λ_I, λ_L e função $g(\theta)$ para o problema heat usando dados com ruído de 0.1%. Linha inferior: Soluções regularizadas LS e TLS.

Na Figura acima vemos que as sequências $\lambda_I^{(k)}$ e $\lambda_L^{(k)}$ convergem, portanto o critério (4.1.9) permite usar uma quantidade adequada de iterações. Também podemos observar que a função g fica perto da sua raiz. Um resumo dos resultados numéricos para os três níveis de ruído, no que diz respeito ao número de iterações e a qualidade das soluções em termos de erro relativo é apresentado na Tabela 4.1.

Método RLS			Método RTLS		
$\|x_{ex} - x_\lambda\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
$\lambda = 0.0038$	$\lambda = 0.053$	$\lambda = 2.726$	$k_m = 344$	k_m=116	$k_m = 42$
0.0342	0.1538	0.670	0.0642	0.1303	0.6150

Tabela 4.1: Resultados das soluções x_λ e x_δ para três níveis de ruído.

As Tabelas acima mostram que na presença de pouco ruído nos dados os métodos RLS e RTLS produzem resultados parecidos, porém, quando o ruído cresce ambos os métodos fornecem soluções de qualidade

muito inferior. Além disso, para estudar a sensibilidade da solução RTLS a mudanças na constante δ, o mesmo problema foi resolvido com $\delta = 2.5\|Lx_{ex}\|$. Neste experimento, para o caso $\sigma = 0.001$, obtivemos um erro relativo $\|x_{ex} - x_\delta\|_2 / \|x_{ex}\|_2 = 1,2825$, mostrando que a solução x_δ pode ser muito sensível a variações em δ. Esta observação, para o caso da forma padrão, também foi feita em [24].

Phillips

Discretizando a equação integral de primeira espécie (1.1.1). Definimos a função

$$\phi(x) = \begin{cases} 1 + \cos\left(\frac{\pi x}{3}\right), & |x| < 3 \\ 0, & |x| \geq 3 \end{cases}$$

Então o núcleo K, a solução f e o lado direito g são dados por:

$$\begin{aligned} K(s,t) &= \phi(s-t) \\ f(t) &= \phi(t) \\ g(s) &= (6 - |s|)\left(1 + \frac{1}{2}\cos\left(\frac{\pi s}{3}\right)\right) + \frac{9}{2\pi}\sin\left(\frac{\pi |s|}{3}\right). \end{aligned}$$

Os intervalos de integração são $[-6, 6]$. A Figura 4.3 mostra a solução exata deste problema.

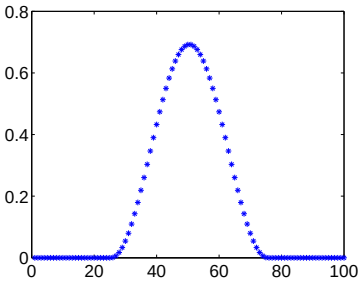

Figura 4.3: Solução do problema Phillips.

A Figura 4.4 mostra as soluções usando os métodos RLS e RTLS para dados com 5% de ruído. Neste caso a solução RTLS é atingida em $k = 5$ e tem um erro relativo 0.0917 (aprox. 9%). A solução RLS produz um erro relativo 0.0656 (aprox. 7%). As imagens mostram algumas diferenças das soluções.

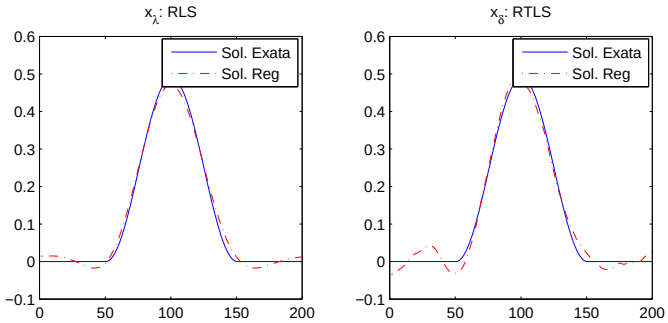

Figura 4.4: Soluções RLS e RTLS para o problema phillips com 5% de ruído.

A Tabela 4.2 mostra para diferentes níveis de ruído o erro relativo, o valor do parâmetro λ, o número de iterações e a qualidade das soluções x_λ e x_δ.

Método RLS			Método RTLS		
$\|x_{ex} - x_\lambda\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
λ= 0.240	$\lambda = 2.494$	$\lambda = 12.85$	k_m= 2	k_m=2	k_m= 5
0.0107	0.0195	0.0656	0.0110	0.0546	0.0917

Tabela 4.2: Resultados da solução x_λ e x_δ para três níveis de ruído.

Podemos observar para este caso que o método RLS aproxima um pouco melhor a solução do que o método RTLS para os diferentes níveis de ruído. Também vemos que o método RTLS converge em poucos passos e a qualidade da solução é bastante boa.

Shaw

Este problema provém da discretização da equação integral de Fredholm de primeira espécie (1.1.1) com o intervalo de integração $[-\pi/2, \pi/2]$. A equação integral é um modelo undimensional de um problema de reconstrução de imagens. O núcleo K e a função f são dados por

$$K(s,t) = (\cos(s) + \cos(t))^2 \left(\frac{\sin(u)}{u}\right)^2$$
$$u = \pi(\sin(s) + \sin(t))$$

$$f(t) = a_1 \exp(-c_1(t - t_1)^2) + a_2 \exp(-c_2(t - t_2)^2).$$

Os parâmetros a_1, a_2, etc. são constantes que determinan a forma da solução f, para este problema vamos usar $a_1 = 2,\ a_2 = 1,\ c_1 = 6,\ c_2 = 2,\ t_1 = 0.8,\ t_2 = -0.5$. A Figura 4.5 mostra a solução exata deste problema

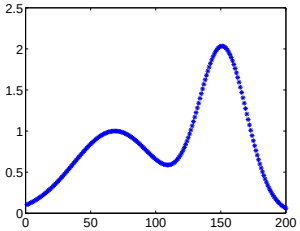

Figura 4.5: Solução do problema shaw.

A Figura 4.6 mostra as soluções usando os métodos RLS e RTLS para dados com 0.1% de ruído. Neste caso a solução RTLS é atingida em $k = 500$ e tem um erro relativo 0.1240 (aprox. 13%). A solução RLS produz um erro relativo 0.0427 (aprox. 5%). As imagens mostram algumas diferenças entre as soluções.

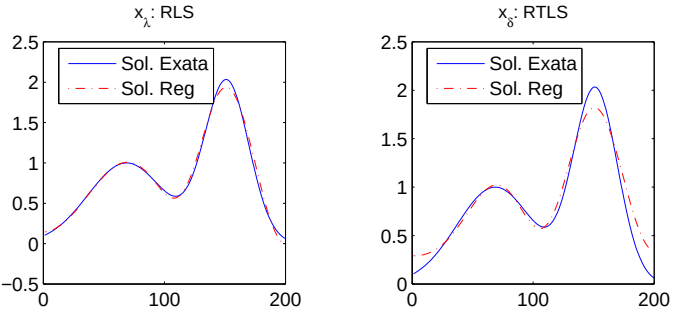

Figura 4.6: Soluções RLS e RTLS para o problema shaw com 0.1% de ruído.

Os resultados numéricos para os três níveis de ruído são apresentados na Tabela 4.3. Como podemos ver, para este problema os métodos RLS e RTLS produzem boas soluções quando o nível de ruído é 0.1%,

Método LS			Método RTLS		
$\|x_{ex} - x_\lambda\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
$\lambda = 0.1366$	$\lambda = 7.415$	$\lambda = 26.221$	$k_m = 500$	$k_m = 2$	$k_m = 2$
0.0427	0.5480	0.5226	0.1240	0.2692	22.1887

Tabela 4.3: Erro relativo da solução RTLS para o problema shaw.

mas a qualidade deteriora significativamente quando o nível de ruído aumenta.

Com o intuito de melhorar a qualidade, o problema também foi resolvido mudando a tolerância do critério de parada (4.1.9) para 10^{-5}. Feito isso foi verificado que a solução melhora mas não muito significatimente. Isto indica que o problema shaw é muito sensível a perturbações nos dados.

4.2 Método de Truncamento para o Problema TLS

Na seção anterior vimos que o método iterativo nem sempre consegue calcular uma solução apropriada para o problema RTLS. Nesta seção vamos propor um método alternativo para construir soluções regularizadas TLS baseado num critério de escolha do índice de truncamento para o método TLS truncado.

4.2.1 O Critério do Produto Mínimo

Iniciamos com uma descrição breve do critério do Produto Mínimo para o método TSVD usado como método de regularização para problemas LS (1.1.2) onde apenas o vetor b sofre perturbações. Baseado na hipótese do lado direito exato b_{ex} satisfazer a condição discreta de Picard (CDP) [23] e o vetor de ruídos ter entradas aleatórias, com distribuição normal e média zero, pode ser mostrado [7] que o erro na k-ésima solução TSVD é

$$\|x_{\rm ex} - x_k\|_2 \leq \|x_{\rm ex} - A_k^\dagger b_{\rm ex}\|_2 + \|A_k^\dagger e\|_2 \equiv E_1(k) + E_2(k), \quad (4.2.1)$$

em que

$$E_1(k) = \sqrt{\sum_{j=k+1}^{n} \frac{|u_j^T b_{\rm ex}|^2}{\sigma_j^2}}, \quad E_2(k) = \sqrt{\sum_{j=1}^{k} \frac{|u_j^T e|^2}{\sigma_j^2}}. \quad (4.2.2)$$

O primeiro termo denota o erro devido à regularização, este diminui com k e pode se tornar pequeno para k grande. O segundo termo, referido como erro de perturbação, mede a intensidade do erro e aumenta com k, sendo extremamente grande quando valores singulares pequenos entram no somatório. A escolha do parâmetro de regularização requer, então, que haja um balanço entre os dois termos de modo a minimizar a estimativa do erro na solução. Uma análise detalhada desenvolvida em [7] sugere que a estimativa do erro deve ser minimizada em $k = k^*$ onde k^* indica a transição dos coeficientes de Fourier $|u_i^T b|$ determinada pela condição discreta de Picard: para $i < k^*$ esses coeficientes aproximam $|u_i^T b_{ex}|$ (e portanto decaem como os valores singulares), enquanto para $i \geq k^*$ $|u_i^T b_{ex}|$ ficam quase constantes e muito próximos de $|u_i^T e|$. A consequência mais importante dessa análise é que a estimativa do erro é mínima quando $k = k^*$. Ou seja, o menor erro associado ao método TSVD deve ocorrer quando truncamos a k^* termos.

No entanto, na prática não temos o vetor e e a estimativa (4.2.1) não pode ser minimizada para encontrar k^*, então o que podemos fazer é minimizar alguma função que tenha comportamento semelhante. Para tanto, notamos que da SVD de A temos

$$\|x_k\|_2^2 = \sum_{j=1}^{k} \frac{|u_j^T b|^2}{\sigma_j^2}, \quad \|\mathsf{r}_k\|_2^2 = \|b - Ax_k\|_2^2 = \sum_{j=k+1}^{n} |\mathsf{u}_j^T b|^2 + \delta_0^2, \tag{4.2.3}$$

em que δ_0 denota a norma da componente de b que não pertence ao espaço coluna de A. Agora para $k \geq 1$, seja $\Psi_k = \|x_k\|_2 \|r_k\|_2$. Como $\|x_k\|_2$ é crescente e $\|r_k\|_2$ é decrescente [18], minimizar $\log(\Psi_k)$ é equivalente a minimizar a soma de dois termos, um crescente e outro decrescente. Então, em princípio, deve existir um inteiro k no qual Ψ_k é minimizada. Portanto, o critério do produto mínimo para o método TSVD seleciona o parâmetro $\widehat{k}$ tal que [7]

$$\widehat{k} = \operatorname{argmin} \Psi_k, \quad \Psi_k = \|x_k\|_2 \|b - Ax_k\|_2. \tag{4.2.4}$$

A motivação deste critério vem da observação de que para k pequeno temos $\|x_k\|_2$ pequeno e $\|b - Ax_k\|_2$ grande e, portanto, Ψ_k não é minimizada. Por outro lado, para k grande temos $\|x_k\|_2$ grande e $\|b - Ax_k\|_2$ pequeno e, mais uma vez, Ψ_k não é minimizada. Isto sugere que o minimizador de Ψ_k corresponde a um bom balanço entre o tamanho de ambas as normas. Uma análise desenvolvida em [7] mostra que um bom parâmetro de regularização para a TSVD é o minimizador de Ψ_k (veja a Figura 4.7).

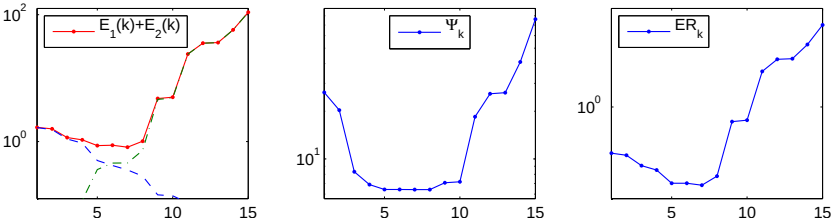

Figura 4.7: Estimativa (4.2.2), função Ψ_k e erros relativos em x_k para o problema gravity com $n = 32$ e NR = 0,02. Erro relativo mínimo atingido em $k^* = 7 = \operatorname{argmin} \Psi_k$ e $\|x_7 - x_{\text{ex}}\| = 0{,}0778\|x_{\text{ex}}\|_2$.

Aplicação do Critério do Produto Mínimo ao Problema TLS

Pelos resultados da seção 4.3 temos que o Resíduo $\|R_k\|_F$ é uma função decrescente e a norma das iterações $\|x_k\|_2$ é uma função crescente. Assim podemos usar o critério do produto mínimo para determinar um índice de truncamento adequado para o método TLS, chamado Método TLS Truncado (T-TLS). Ou seja, tomamos como índice de truncamento o primeiro inteiro k_t tal que

$$k_t = \operatorname{argmin} \Psi_k, \quad \Psi_k = \|x_k\|_2 \|R_k\|_F. \tag{4.2.5}$$

Ou seja k_t é o primeiro índice no qual Ψ_k atinge um mínimo local. Além disso, para certos problemas onde Ψ_k tem o primeiro mínimo local numa região quase plana, usamos como critério de escolha o primeiro índice tal que

$$\frac{\Psi_{k+1} - \Psi_k}{\Psi_1} < \varepsilon \tag{4.2.6}$$

onde ε é um número positivo suficientemente pequeno.

4.2.2 Experimentos Numéricos

Como na seção anterior, trabalharemos com um sistema $Ax \approx b$ onde $A \in \mathbb{R}^{400\times 200}$, $b \in \mathbb{R}^{400}$, usando os mesmos níveis de ruído nos dados e os mesmos problemas já analisados anteriormente. O parâmetro que minimiza a sequência do erro relativo associado das iteradas será chamado de parâmetro ótimo e denotado por k_o. Ele depende da solução exata x_{ex} e na prática é desconhecido. Denotamos por $x_{TTLS} = x_{k_t}$ a solução atingida no passo k_t.

Heat

Para este problema fazemos a comparação da qualidade das soluções do método RTLS e o método T-TLS para dados com ruído de 0.1% junto os parâmetros ótimo k_o e k_t respectivamente.

A Figura 4.8 mostra que os parâmetros k_o e k_t têm valores próximos, também que o método T-TLS atinge a solução em $k = 34$ com um erro relativo de 0.0395 (aprox. 4%).

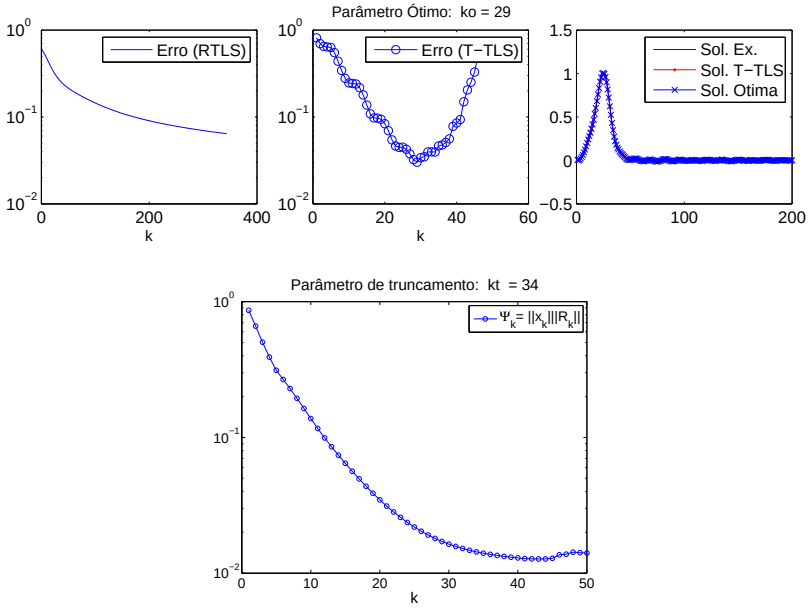

Figura 4.8: Linha superior: Erros relativos RLS e T-TLS para o problema heat usando dados com ruído de 0.1% e solução x_{TTLS}. Linha inferior: Função Ψ_k.

Note que neste problema o índice k_t é escolhido segundo o critério (4.2.6). A Tabela 4.4 mostra que para diferentes níveis de ruído as soluções x_{TTLS} e x_δ não diferem muito, mas podemos observar que o método T-TLS não consegue uma boa aproximação na presença de ruído $\sigma = 5\%$ nos dados.

Método T-TLS			Método RTLS		
$\|x_{ex} - x_{TTLS}\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
k_t=34	k_t=23	k_t=5	k_m=344	k_m=116	k_m= 42
0.0395	0.1666	0.6345	0.0642	0.1303	0.6150

Tabela 4.4: Erro relativo da solução T-TLS e RTLS com diferentes níveis de ruído para o problema heat.

Phillips

Mostramos a comparação do erro do problema RTLS com o problema T-TLS para o ruído $\sigma = 0.05$ junto parâmetro de truncamento k_t.

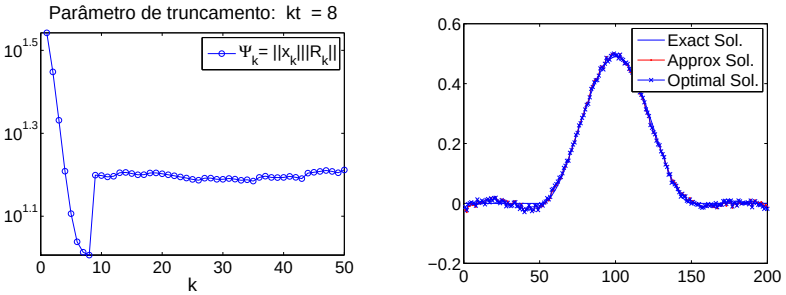

Figura 4.9: Função ψ_k e solução T-TLS para o problema phillips.

Na imagem esquerda da Figura 4.9 mostra que a função ψ_k alcança o valor mínimo em $k_t = 8$, também vemos que a solução x_{TTLS} é uma boa aproximação da solução x_{ex}, porém a solução x_{TTLS} perde suavidade pois este método não considera a restrição $\|Lx\|_2 \leq \delta$.

A Tabela 4.5 mostra o valor do erro relativo no parâmetro k_t e k_m das soluções x_{TTLS} e x_δ respectivamente, para diferentes níveis de ruído.

Método T-TLS			Método RTLS		
$\|x_{ex} - x_{TTLS}\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
k_t= 15	k_t= 11	k_t= 8	k_m= 2	k_m=2	k_m= 5
0.0119	0.0868	0.0477	0.0110	0.0546	0.0917

Tabela 4.5: Erro relativo dos métodos T-TLS e RTLS para o problema phillips.

Esta Tabela mostra que as soluções usando os métodos T-TLS e RTLS são parecidas e fornecem uma boa aproximação da solução exata.

Shaw

Mostramos a comparação do erro do problema RTLS com o problema T-TLS para o ruído $\sigma = 0.001$ junto com o parâmetro ótimo k_o.

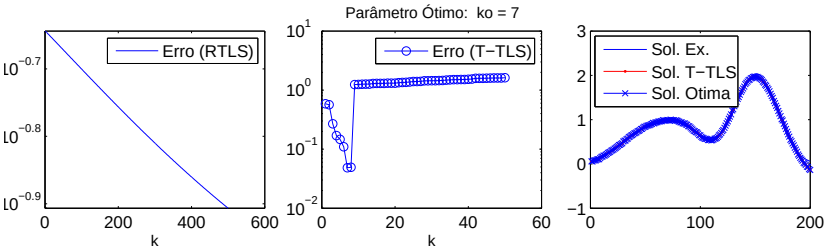

Figura 4.10: Erros relativos dos métodos T-TLS e ETLS aplicado ao problema shaw com 0.1% de ruído.

A Tabela 4.6 mostra o valor do erro relativo no parâmetro k_t para diferentes níveis de ruído.

Método T-TLS			Método RTLS		
$\|x_{ex} - x_{TTLS}\|_2/\|x_{ex}\|_2$			$\|x_{ex} - x_\delta\|_2/\|x_{ex}\|_2$		
$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$	$\sigma = 0.1\%$	$\sigma = 1\%$	$\sigma = 5\%$
k_t= 7	k_t= 5	k_t= 4	k_m = 500	k_m= 2	k_m=2
0.0485	0.1657	0.1729	0.1240	0.2692	22.1887

Tabela 4.6: Erro relativo da solução T-TLS para o problema shaw.

Da Tabela 4.6 vemos que o erro relativo calculado pelo método T-TLS é melhor melhor do que o método RTLS, porém para este método a solução x_{TTLS} não possui suavidade quando o ruído $\sigma = 0.05$ como mostra a Figura 4.11.

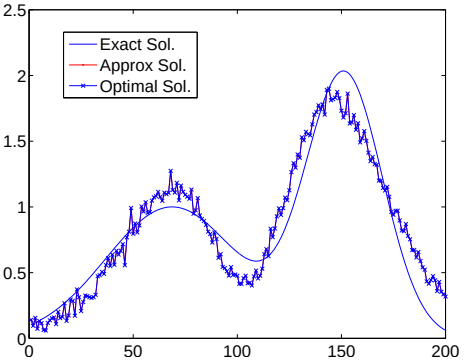

Figura 4.11: T-TLS aplicado ao problema Shaw com 5% de ruído.

Capítulo 5

Conclusões

O objetivo principal deste trabalho foi o estudo de métodos de regularização para o método TLS, devido a necessidade de resolver problemas mal condicionados onde a matriz e o vetor de dados sofrem perturbações. Para tanto, baseado na potencialidade teórica da SVD e do conceito de projeções ortogonais, estudamos a técnica TLS e sua relação com a técnica LS, incluindo estimativas em relação a solução exata do problema linear e vimos que a técnica TLS fornece boas estimativas de solução como a técnica LS. Os resultados teóricos da comparação foram verificados no modelo de ressonância magnética onde comprovamos a eficiência do método TLS.

O estudo do método de Tikhonov TLS regularizado (RTLS) nos levou a um problema de autovalores

$$B(\lambda_L)\begin{bmatrix} x \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} x \\ -1 \end{bmatrix},$$

onde B, λ_L e λ dependem do vetor x. Este sistema de equações é difícil de se resolver analiticamente, assim, a alternativa foi usar um método iterativo baseao no problema de autovalores chamado de RTLSEVP.

O método RTLSEVP tem aspectos teóricos e computacionais que são simples de se processar, mas ele depende do conhecimento de boas estimativas do número $\delta = \|Lx_{ex}\|_2$ que na prática não é conhecido. Daí, a solução pode não ser satisfatória se δ não é estimado corretamente ou o processo iterativo pode ser muito demorado e portanto pouco atrativo. O método alternativo baseado nos autovalores quadráticos requer o uso da teoria de autovalores generalizados e outros métodos computacionais que não foram estudados por falta de tempo. Esta

opção pode ser objeto de estudo num trabalho posterior. Por outro lado, temos visto que o método iterativo nem sempre consegue alcançar uma solução razoável quando δ não é estimado corretamente.

Nestre trabalho apresentamos o uso do método TLS como alternativa de regularização, incluindo um critério de escolha do índice de truncamento. A escolha do índice de truncamento foi baseada num trabalho recente da literatura [7] e motivada pelo fato de que as sequências $\|x_k\|_2$ e $\|R_k\|_F$ se comportam de maneira análoga as normas da solução e do resíduo no método SVD truncado. Como vimos este método requer poucas iterações e consegue soluções com qualidade semelhante aquelas obtidas pela técnica RTLS, mas sem a informação a priori do número δ.

Apêndice A

Problema de Minimização com Restrições de Igualdade e Desigualdade

Agora vamos desenvolver a teoria necessária para problemas de minimização com restrições. De modo geral estudaremos o problema

$$\min_{\substack{h(x)=0 \\ g(x)\leq 0}} f(x) \tag{A.0.1}$$

onde $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^l$ com $h = (h_1, \ldots, h_l)$ e $g : \mathbb{R}^n \to \mathbb{R}^m$ com $g = (g_1, \ldots, g_m)$.

Observação: Dizemos que $g(x) \leq 0$ se $g_i(x) \leq 0,\ \forall i = 1, \ldots, m$.

Definição A.0.1. *Seja $\bar{x} \in D = \{x \in \mathbb{R}^n;\ h(x) = 0 \text{ e } g(x) \leq 0\}$, definimos o conjunto de restrições ativas em $\bar{x}$ como*

$$I(\bar{x}) = \{i \in \{1, \ldots, m\};\ g_i(\bar{x}) = 0\}$$

Definição A.0.2. *Suponha que g e h são diferenciáveis em $\bar{x} \in D$*

$$H(\bar{x}) = \{d \in \mathbb{R}^n;\ d \in \mathcal{N}\left(h'(x)\right)\ e\ \langle g_i'(\bar{x}), d\rangle \leq 0,\ \forall i \in I(\bar{x})\}$$

A.0.3 Condições de Regularidade/Qualificação

Vamos estabelecer as condições necessárias para nosso Teorema principal nesta seção

(1) Dizemos que as funções h e g são afins, se

$$h(x) = Ax - a, \text{ com } A \in \mathbb{R}^{n\times l} \text{ e } a \in \mathbb{R}^l.$$
$$g(x) = Bx - b, \text{ com } B \in \mathbb{R}^{n\times m} \text{ e } b \in \mathbb{R}^m.$$

(2) Condição de Mangasarian-Fromovitz (MFCQ)

(i) $h'(\bar{x})$ é uma função sobrejetora.

(ii) Existe $\bar{d} \in \mathcal{N}(h'(\bar{x}))$ tal que $\langle g_i'(\bar{x}), \bar{d}\rangle < 0, \ \forall i \in I(\bar{x})$.

(3) Condição de Slater

(i) A função h é afim.

(ii) A função g é convexa e existe um vetor $\tilde{x} \in \mathbb{R}^n$ tal que $h(\tilde{x}) = 0$ e $g_i(\tilde{x}) < 0, \ \forall i \in \{1, \ldots, m\}$.

Observação: A funcão $f : \mathbb{R}^n \to \mathbb{R}$ é convexa se

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y), \ \forall x, y \in \mathbb{R}^n, \ \forall \lambda \in [0,1].$$

Teorema A.0.1. *(Karush-Kuhn-Tucker) Considere el problema A.0.1 onde as funções $f : \mathbb{R}^n \to \mathbb{R}$ e $g : \mathbb{R}^n \to \mathbb{R}^m$ diferenciáveis em $\bar{x}$ e $h : \mathbb{R}^n \to \mathbb{R}^l$ continuamente diferenciável em $\bar{x}$. Suponha que*

$$D = \{x \in \mathbb{R}^n; \ h(x) = 0 \text{ e } g(x) \leq 0\}.$$

Se g e h satisfazem qualquer um das condições de regularidade como visto acima (linearidade das restrições, MFCQ ou Slater) então existem constantes $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_l) \in \mathbb{R}^l$ e $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_m) \in \mathbb{R}^m$ tais que:

$$f'(\bar{x}) + \sum_{i=1}^{l} \bar{\lambda}_i h_i'(\bar{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i'(\bar{x}) = 0. \tag{A.0.2}$$
$$h_i(\bar{x}) = 0, \ i = 1, \ldots, l. \tag{A.0.3}$$
$$g_i(\bar{x}) \leq 0, \ i = 1, \ldots, m. \tag{A.0.4}$$
$$\bar{\mu}_i \geq 0, \ i = 1, \ldots, m \tag{A.0.5}$$
$$\bar{\mu}_i . g_i(\bar{x}) = 0, \ i = 1, \ldots, m. \tag{A.0.6}$$

Este resultado permite definir

Definição A.0.3. *A função definida por*

$$\begin{aligned}\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m &\longrightarrow \mathbb{R} \\ \mathcal{L}(x, \lambda, \mu) &= f(x) + \langle \lambda, h(x) \rangle + \langle \mu, g(x) \rangle\end{aligned}$$

é chamada a função Lagrangeano, e λ e μ que satisfazem os resultados do Teorema A.0.1, são chamados os multiplicadores de Lagrange.

Observação: Note que

$$\mathcal{L}'_x(x, \lambda, \mu) = f'(x) + h'(x)^T \lambda + g'(x)^T \mu$$

e a primeira condição do Teorema A.0.1 (KKT) diz que

$$\mathcal{L}'_x(x, \lambda, \mu) = 0$$

A.1 Mínimos Quadrados com Restrição de Igualdade

Nesta seção vamos estudar o problema

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{s.a} \quad \|Lx\|_2^2 = \delta^2 \tag{A.1.1}$$

para obter algumas propriedades dos valores λ_I e λ_L na seguinte seção. Vamos mencionar algumas condições que garantem a existência da solução da equação (A.1.1) e desenvolver un pouco de teoria sobre este problema. Estes resultados foram desenvolvidos em [10].

(1) O conjunto $F = \{x \in \mathbb{R}^n;\ \|Lx\|_2 = \delta\}$ é não vazio e $\delta > \min_x \|Lx\|_2$. Se δ é muito pequeno a solução pode não existir.

(2) Posto $\left(\begin{bmatrix} A \\ L \end{bmatrix} \right) = n$.

Se não assumirmos a condição sobre o posto, caso existe uma solução, ela poderia não ser única, pois podemos adicionar os elementos da interseção dos espaços nulos de A e L.

A.1.1 A Função de Lagrange e as Equações Normais

As hipóteses acima garantem que a solução do (3.2.5) é um ponto estacionário da função de Lagrange com multiplicador λ.

$$\mathcal{L}(x, \lambda) = \|Ax - b\|_2^2 + \lambda(\|Lx\|_2^2 - \delta^2).$$

Usando as propriedades de $\mathcal{L}$ temos as equações normais.

$$(A^T A + \lambda L^T L)x = A^T b \tag{A.1.2}$$

$$\|Lx\|_2^2 = \delta^2 \tag{A.1.3}$$

Note que na primeira condição excluimos o caso que a solução minimiza $\|Lx\|_2$, pois só é possível se $\delta = \min_x \|Lx\|_2$. Note que a equação (A.1.2) pode ser escrita como

$$\nabla_x \|Ax - b\|_2^2 = -\lambda \nabla_x \|Lx\|_2^2 \tag{A.1.4}$$

a qual é uma condição necessária para a solução de (3.2.5). As equações normais podem ter muitas soluções, uma delas é a solução de (3.2.5). A primeira questão é decidir qual das soluções das equações normais resolve o problema.

O seguinte Teorema compara duas soluções das equações normais, e serve para esclarecer sobre a escolha da solução que minimiza $\|Ax-b\|_2$

Teorema A.1.1. *Se (x_1, λ_1),e (x_2, λ_2) são soluções das equações normais (A.1.2) e (A.1.3), então*

$$\|Ax_2 - b\|_2^2 - \|Ax_1 - b\|_2^2 = \frac{\lambda_1 - \lambda_2}{2} \|L(x_1 - x_2)\|_2^2. \tag{A.1.5}$$

Demonstração. Ver [10]. □

Pelo Teorema temos que

$$\|Ax_2 - b\|_2 > \|Ax_1 - b\|_2, \quad \text{se } \lambda_1 > \lambda_2.$$

Se podemos provar que o lado direito da equação (A.1.5) não é zero, então a solução do problema original é solução das equações normais com o maior λ. O seguinte Teorema é um resultado similar ao Teorema anterior.

Teorema A.1.2. *Se (x_1, λ_1) e (x_2, λ_2) são soluções das equações normais, então*

$$(\lambda_1 + \lambda_2)\{\|Ax_2 - b\|_2^2 - \|Ax_1 - b\|_2^2\} = (\lambda_2 - \lambda_1)\|A(x_1 - x_2)\|_2^2. \tag{A.1.6}$$

Demonstração. Ver [10]. □

Analisaremos o que acontece quando o lado direito da equação (A.1.5) é zero. Sejam $(x_1, \lambda_1) \neq (x_2, \lambda_2)$ duas soluções das equações normais e assuma que

$$\frac{\lambda_1 - \lambda_2}{2} \|L(x_1 - x_2)\|_2^2 = 0. \tag{A.1.7}$$

Isto é possivel se

(i) $\lambda_1 = \lambda_2 := \lambda$ mas $x_1 \neq x_2$. Como x_1 e x_2 são soluções de (A.1.2) temos

$$(A^T A + \lambda L^T L)x_1 = A^T b$$
$$(A^T A + \lambda L^T L)x_2 = A^T b.$$

Substraindo a segunda equação da primeira temos

$$(A^T A + \lambda L^T L)(x_1 - x_2) = 0. \tag{A.1.8}$$

Equação (A.1.8) mostra que neste caso $\lambda = -\mu$, onde μ é um autovalor do problema de autovalor generalizado

$$A^T Ax = \mu L^T Lx \tag{A.1.9}$$

e $x_1 = x_2 + v_\mu$ onde v_μ é um autovetor associado a μ. Portanto se $\lambda_1 = \lambda_2 = -\mu$ então as soluções x_1 e x_2 têm o mesmo valor $\|Ax - b\|_2$.

(ii) Se

$$\lambda_1 \neq \lambda_2 \quad \text{mais } \|L(x_1 - x_2)\|_2^2 = 0. \tag{A.1.10}$$

Pelo Teorema A.1.1, temos que $\|Ax_2 - b\|_2^2 - \|Ax_1 - b\|_2^2 = 0$. Logo do Teorema A.1.2

$$\|A(x_2 - x_1)\|_2^2 = 0. \tag{A.1.11}$$

Porém (A.1.10) e (A.1.11) é equivalente à

$$\begin{bmatrix} A \\ L \end{bmatrix} (x_2 - x_1) = 0. \tag{A.1.12}$$

Pela segunda hipótese temos que $x_1 = x_2 = x'$. Usando a equação (A.1.4) neste caso temos

$$\nabla_x \|Ax' - b\|_2^2 = \nabla_x \|Lx'\|_2^2 = 0$$

e portanto

$$\delta = \|Lx'\|_2 = \min_x \|Lx\|_2.$$

Isto contradiz a primeira hipótese. Assim este caso não é possivel.

Consequentemente temos:

Lema A.1.1. *Com as hipóteses (1) e (2) sejam (x_1, λ_1) e (x_2, λ_2) soluções das equações normais. Se $\lambda_1 \neq \lambda_2$ então*

$$Lx_1 \neq Lx_2 \quad \text{e} \quad Ax_1 \neq Ax_2.$$

A.1.2 Caracterização da Solução

Usando as equações (A.1.5) e (A.1.6) obtemos

Corolário A.1.1. *Sejam* (x_1, λ_1) *e* (x_2, λ_2) *soluções das equações normais junto com as hipóteses (1) e (2). Então*

$$-\frac{\lambda_1 + \lambda_2}{2}\|L(x_1 - x_2)\|_2^2 = \|A(x_1 - x_2)\|_2^2. \tag{A.1.13}$$

A equação (A.1.13) tem uma importante consequência: Para qualquer par de soluções (x_1, λ_1), (x_2, λ_2) com $\lambda_1 \neq \lambda_2$ segue que $\lambda_1 + \lambda_2 < 0$. Isto significa que ambos λ_i são negativos ou se, por exemplo, $\lambda_1 < 0$ então $\lambda_2 < -\lambda_1 < 0$. Logo

Corolário A.1.2. *As equações normais (A.1.2) e (A.1.3) têm no máximo uma solução* $(\tilde{x}, \tilde{\lambda})$ *com* $\tilde{\lambda} > 0$. *Para qualquer outra solução* (x, λ) *temos que* $\lambda < -\tilde{\lambda}$.

A.2 Quociente de Rayleigh

A solução do problema (3.2.5) resolve o problema

$$x_{TLS} = \underset{x}{\operatorname{argmin}}\, \phi(x) = \underset{x}{\operatorname{argmin}} \frac{\|Ax - b\|_2^2}{1 + \|x\|_2^2}, \tag{A.2.1}$$

onde ϕ é chamado o quociente de Rayleigh da matriz $M = [A \quad b]^T[A \quad b]$. A formulação alternativa para o método RTLS é

$$\min_x \phi(x) \quad \text{s.a.} \quad \|Lx\|_2 \leq \delta.$$

Assim temos o funcional Lagrangeano

$$\mathcal{L}(x, \mu) = \phi(x) + \mu(\|Lx\|_2^2 - \delta^2).$$

Embora $\phi(x)$ não é côncavo, seus pontos estacionários podem ser caracterizados pelo seguinte Teorema.

Teorema A.2.1. *O quociente de Rayleigh de uma matriz simétrica A é estacionário só nos autovetores da matriz.*

Demonstração. Ver [27]. □

Lema A.2.1. *Se os extremos dos valores singulares da matriz* $[A \quad b]$ *são simples, então a função* $\phi(x)$ *tem um único ponto máximo, um ponto mínimo e* $n - 1$ *pontos de sela.*

Demonstração. Ver [31]. □

Teorema A.2.2. *Suponha que* $[A \quad b]$ *satisfazem as hipóteses do Lema A.2.1,* $\bar{\sigma}_n > \bar{\sigma}_{n+1}$ *e que a constante* δ *é conhecida. Se a igualdade é ativa, então*

$$\|Lx_{RTLS}\|_2^2 = \delta^2 \quad \text{e} \quad \mu > 0.$$

Demonstração. Ver [31]. □

Referências Bibliográficas

[1] K. S. Arun, A unitarily constrained total least square problem in signal processing, SIAM J: Matrix Anal. Appl., 13 (1992), pp. 729-745.

[2] A. Beck and A. Ben-Tal. On the solution of the Tikhonov regularization of the total least squares problem. SIAM J. Optim., 17, 98 - 118, 2006.

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

[4] A. Doicu, T. Trautmann, and F. Schreier, "Numerical Reguralization for Atmospheric Inverse Problems", Springer-Verlag Berlin Heidelberg, pp. 254-256, 2010.

[5] F. S. Bazán V., CGLS-GCV: a hybrid algorithm for low-rank-deficient problems, Applied Numerical Mathematics, Vol. 47, pp. 91-108, 2003.

[6] F. S. Bazán V., Fixed-point iterations in determining the Tikhonov regularization parameter, Inverse problems, 24, 2008.

[7] Fermín S. Bazán Viloche, Maria C. C. Cunha, and Leonardo S. Borges, Extension of GKB-FP algorithm to large-scale general-form Tikhonov regularization, Numerical Linear Algebra with Applications, John Wiley & Sons, Ltd, 2013.

[8] Ricardo D. Fierro and R. Bunch, Ortogonal Projection and Total Least Square, Linear Algebra Appl. 2:135-153 (1995).

[9] C. Eckart and G. Young, The approximation of one matriz by another of lower rank, Psychometrica, John Wiley, New York, 1951.

[10] Walter Gander, Least squares with a quadratic constraint, Numerische Mathematik, Vol. 36, No. 3.(1980), pp. 291-307.

[11] G. H. Golub, and C. F. Van Loan, An analysis of the total least square problem, SIAM J. Number. 6, Anal., Vol. 17, pp. 883-893, 1980.

[12] G. H. Golub, M. T. Heath and G. Wahba, Generalized cross-validation as a metohd for choosing a godd ridge parameter, Technometrics, 21, pp, 215-223, 1979.

[13] G. H. Golub, AND C. F. Van Loan, "Matriz Computations". 3.ed Maryland: Jhon Hopkins University Press, 1996.

[14] G. H. Golub, P. C. Hansen, and D. P. O'Leary, Tikhonov Regularization and Total Least Squares. SIAM J. Matrix Anal. Appl., 21:185-194, 1999.

[15] P. C. Hansen, "Discrete inverse problems: Insight and algorithms", Philadelphia: SIAM, 2010.

[16] P. C. Hansen, "Rank-deficient and discrete ill-possed problems", Philadelphia: SIAM, 1998.

[17] P. C. Hansen, "Regularization Tools - A Matlab Package for Analysis ans Solution of Discrete Ill-Posed Problems", Numerical Algorithms 6, pp.1-35, 1994.

[18] P. C. Hansen, Analysis of discrete ill-possed problems by means of the L-curve, SIAM Review, 34, pp. 561-580, 1992.

[19] P. C. Hansen, O'Leary D. P. The use of the L-curve in the regularization of discrete ill-posed problems. SIAM Journal on Scientific Computing 1993; 14 :1487-1503.

[20] Hochstenback ME, Reichel L. An Iterative Method For Tikhonov Regularization With a General Linear Regularization Operator. Journal of Integral Equations and Applications 2010; 22:463-480.

[21] J. Lampe and H. Voss, On a quadratic eigenproblem occuring in regularized total least squares. Comput. Stat. Data Anal., 52/2:1090 - 1102, 2007.

[22] J. Lampe and H. Voss, Solving Regularized Total Least Squares Problems Based on Eigenproblems. Technical report, Institute of Numerical Simulation, Hamburg Universuty of Technology, 2008. Submitted to Taiwwanese Journal of Mathematics.

[23] Kirsch, A.. An Introduction to the Mathematical Theory of Inverse Problems, Applied Mathematical Sciences Vol. 120, New York: Springer, 1996.

[24] Shuai Lu, Sergei v. Pereversev, Ulrich Tautenhahn, Regularized Total Least Squares: Computational aspects and error bounds. Siam J. Matrix Anal. Appl. Vol. 31, No. 3, pp. 918-941.

[25] V. A. Morozov, On the solution of functional equations bye the method of regularization, Soviet Math. Dokl, Vol, 7, pp 414-417, 1996.

[26] C.C. Paige and M. A. Saunders, Towards a generalized singular value decomposition, SIAM J. Numer. Anal. 18:398-405 (1981).

[27] B. N. Parlett, The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[28] D. L. Phillips, A Technique for the Numerical Solution of Certain Integral Equations of the First Kind, J. ACM, Vol. 9, 1962.

[29] T. Regińska, A regularization parameter i discrete ill-possed problems, SIAM J. Sci. Comput. Vol. 17, No 3, pp. 740-749, 1996.

[30] R. A. Renaut and H. Guo, A regularized total least square algorithm, in Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications, S. Van Huffel and P. Lemmerling, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 57-66.

[31] R. A. Renaut and H. Guo. Efficient algorithms for solution of regularized total least squares. SIAM J. Matrix Anal. Appl., 26:457-476, 2005.

[32] J. D. Riley, Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix, Math. Tables Aids Comput., 9 (1955), pp. 96 - 101.

[33] D. M. Sima. Regularization Techniques in Model Fitting and Parameter Estimation. PhD tesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.

[34] G. W. Stewart and Ji-guang Sun. Matrix Perturbation Theory. Academic Press, 1990.

[35] R. C. Thompson, Principal submatrices IX: interlacing inequalities for singular values os submatrices, Linear Algebra Appl., 5 (1972), pp.1-12.

[36] Tikhonov AN. *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4, p. 1035-1038, 1963.

[37] S. Van Huffel. and J. Vandewalle, "The Total Least Square Problems - Computational Aspects and Analysis", SIAM, Philadelphia, PA, 1991.

[38] Per Åke Wedin, Perturbation theory for pseudoinverses, BIT 13:217-232 (1973).

[39] J. H. Wilkinson, "Convergence of the LR, QR, and Related Algorithms," Comp. J. 8, 77-84, 1965.