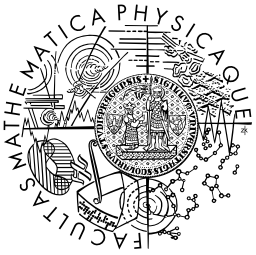


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Kateřina Pokorná

Aproximace metodou TLS: lineární fitování dat pro problémy s nepřesným modelem

Katedra numerické matematiky

Vedoucí bakalářské práce: doc. RNDr. Iveta Hnětynková, Ph.D.

Studijní program: Obecná matematika

Studijní obor: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Ráda bych poděkovala doc. RNDr. Ivetě Hnětynkové, Ph.D., vedoucí mé bakalářské práce, za čas, který mi věnovala, za ochotu, trpělivost, pochopení a mnoho cenných rad, nejen co se této práce týče.

Název práce: Aproximace metodou TLS: lineární fitování dat pro problémy s nepřesným modelem

Autor: Kateřina Pokorná

Katedra: Katedra numerické matematiky

Vedoucí bakalářské práce: doc. RNDr. Iveta Hnětynková, Ph.D., Katedra numerické matematiky

Abstrakt: V předložené práci se budeme zabývat lineární aproximační úlohou, kde pozorování i model jsou zatíženy chybami, a zaměříme se na problém *úplných nejmenších čtverců (TLS)*, jímž lze takové úlohy řešit. Shrneme klasickou teorii existence a jednoznačnosti TLS řešení, uvedeme klasický TLS algoritmus a podíváme se na komplikace, které mohou při jeho implementaci nastat. Dále budeme studovat *singulární rozklad (SVD)* matice, jež se využívá při konstrukci TLS řešení. Podrobně popíšeme metodu jeho výpočtu. Protože je výpočet SVD poměrně náročný, soustředíme se dále na možnost aproximace jeho části potřebné ke konstrukci TLS řešení, tzv. *singulárních tripletů*, založené na *Golub-Kahanově iterační bidiagonalizaci*. Nakonec budeme v numerických experimentech testovat vliv kvality aproximace nejmenších singulárních tripletů na spočtené TLS řešení.

Klíčová slova: lineární aproximační problém, chyby v datech, úplné nejmenší čtverce, singulární rozklad

Title: Approximation by the TLS method: linear data fitting for problems with unprecise models

Author: Kateřina Pokorná

Department: Department of Numerical Mathematics

Supervisor: doc. RNDr. Iveta Hnětynková, Ph.D., Department of Numerical Mathematics

Abstract: In this thesis, we concern ourselves with the linear approximation problem, where errors in both the observation and the data are considered. We focus on the *total least squares* problem *(TLS)*, which may be used in solving such tasks. We summarise basic theory of the existence and uniqueness of the TLS solution, present the classic TLS algorithm and examine some possible complications, which may appear during its implementation. Furthermore, we shall study the *singular value decomposition (SVD)*, which is used in constructing the TLS solution. As the SVD is rather difficult to compute, we discuss one of the possible methods of approximating only its part necessary for the construction of the TLS solution, the so called *singular triplets*. This method is based on *Golub-Kahan iterative bidiagonalization*. Finally, we shall test how the quality of the approximation of the smallest singular triplets influences the computed TLS solution.

Keywords: linear approximation problem, data errors, total least squares, singular value decomposition

Obsah

Seznam použitého značení

$\mathbb{N}$	množina přirozených čísel
$\mathbb{R}$	množina reálných čísel
$\mathbb{C}$	množina komplexních čísel
x	vektor
e_i	i-tý jednotkový vektor
A	matice
$A^\top$	transponovaná matice A
A^*	hermitovsky sdružená matice A
$\mathcal{R}(A)$	obor hodnot matice A
$Ker(A)$	jádro matice A
$rank(A)$	hodnost matice A
$diag(a_1, \ldots, a_n)$	diagonální matice s prvky $a_1, \ldots, a_n$ na diagonále
$span\{v_1, \ldots, v_n\}$	lineární obal vektorů $v_1, \ldots, v_n$
u_i	i-tý levý singulární vektor
v_i	i-tý pravý singulární vektor
σ_i	i-té singulární číslo
(σ_i, u_i, v_i)	singulární triplet
U	matice levých singulárních vektorů
V	matice pravých singulárních vektorů
Σ	diagonální matice se singulárními čísly na diagonále
$A^{(k)}$	aproximace matice A hodnosti k
$\mathcal{K}_k(A,v) \equiv$ $span\{v, Av, \ldots, A^{k-1}v\}$	k-tý Krylovův podprostor generovaný maticí A a vektorem v

$\|a\|$	absolutní hodnota čísla a
$\|\|x\|\|$	Euklidovská norma vektoru x
$\|\|A\|\|$	spektrální norma matice A
$\|\|A\|\|_F$	Frobeniova norma matice A

LS	problém nejmenších čtverců
TLS	problém úplných nejmenších čtverců
SVD	singulární rozklad
GKB	Golub-Kahanova iterační bidiagonalizace

Úvod

V praxi se často setkáváme s potřebou řešit lineární aproximační problémy $Ax \approx b$, kde $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$. Zároveň musíme počítat s tím, že jak matice modelu A, tak vektor pozorování b mohou být zatíženy chybami různého původu. Ty mohou být např. diskretizační, zaokrouhlovací, chyby modelu nebo i lidské pochybení. Tyto perturbace pak způsobují, že úloha nemá přesné řešení a je tedy třeba hledat v určitém smyslu vhodnou opravu matice A nebo vektoru b tak, aby opravená soustava již přesné řešení měla. Toto řešení pak budeme považovat za aproximaci neznámého vektoru x. Pro konkrétní aplikace viz Van Huffel a Vandewalle (1991), Kapitola 1.

Předpokládáme-li, že chyby obsahuje pouze vektor pozorování b, jedním z nástrojů, jakými lze takové problémy řešit, je problém *nejmenších čtverců (LS)* (Golub a Van Loan (1980)). Ovšem tento předpoklad bývá často nerealistický z důvodu výskytu chyb zmíněných výše a je tedy třeba uvažovat i chyby v modelu A. Předložená práce se zabývá problémem *úplných nejmenších čtverců (TLS)* (Golub a Van Loan (1980)), což je jeden ze způsobů fitování dat pro lineární aproximační problémy s nepřesným modelem. Nejprve shrneme základní teorii problému úplných nejmenších čtverců, zaměříme se na analýzu existence a jednoznačnosti TLS řešení a uvedeme tzv. klasický TLS algoritmus (Van Huffel a Vandewalle (1991), Sekce 3.6.1) pro jeho výpočet. Podrobně byla metoda TLS včetně jejích rozšíření pro vícenásobné vektory pravých stran analyzována například v původním článku Golub a Van Loan (1980), v knize Van Huffel a Vandewalle (1991), nebo v Hnětynková a kol. (2011).

Při řešení lineárního aproximačního problému metodou úplných nejmenších čtverců se využívá *singulárního rozkladu (SVD)* (Duintjer Tebbens a kol. (2012), Kapitola 5) tzv. rozšířené matice dat $[b,A]$. Jeho výpočet, který v práci podrobně popíšeme, ale bývá obecně náročný. Zároveň pro spočtení TLS řešení není třeba znát celý SVD matice dat, ale stačí mít k dispozici pouze tzv. nejmenší *singulární triplety*, tj. trojice singulárních čísel a příslušných pravých a levých singulárních vektorů. V této práci se tedy dále budeme zabývat možným způsobem aproximace částečného SVD, který je založen na metodě převodu matice na bidiagonální tvar, tzv. *Golub-Kahanově iterační bidiagonalizaci (GKB)* (Golub a Kahan (1965)).

Nakonec budeme pomocí numerických experimentů v prostředí MATLAB testovat vliv kvality aproximace nejmenších singulárních tripletů na spočtené TLS řešení. K tomu budeme využívat třech metod aproximace - funkce 'svd' a 'svds', jež jsou obě implementované v MATLABu a ta druhá je založena právě na metodě zmíněné výše, a funkci 'lobpcg' (Knyazev (2001)).

V první kapitole shrneme klasickou teorii problému úplných nejmenších čtverců, analýzu existence a jednoznačnosti TLS řešení a nakonec uvedeme tzv. klasický TLS algoritmus. Ve druhé kapitole zmíníme některé problémy, které mohou při implementaci TLS v konečné aritmetice nastat, a možné způsoby, jakými k nim přistupovat. Dále popíšeme výpočet singulárního rozkladu a zaměříme se na metody aproximace jeho části. Ve třetí kapitole pak budeme provádět numerické experimenty.

1. Problém úplných nejmenších čtverců

V první kapitole se budeme zabývat analýzou problému úplných nejmenších čtverců (TLS), což je jedna z metod pro řešení lineární aproximační úlohy. Rozebereme, v jakých případech řešení ve smyslu TLS existuje a za jakých podmínek je jednoznačné. Na konci kapitoly uvedeme také tzv. klasický TLS algoritmus.

1.1 Motivace, základní pojmy a definice

V této kapitole čerpáme z knihy Van Huffel a Vandewalle (1991), učebnice Duintjer Tebbens a kol. (2012) a článku Golub a Van Loan (1980).

Mějme matici modelu $A \in \mathbb{R}^{n\times m}$ a vektor pozorování $b \in \mathbb{R}^n$ a uvažujme problém nalezení vektoru $x \in \mathbb{R}^m$ tak, aby $Ax = b$. Zároveň uvažujme případ, kdy je tato úloha tzv. přeurčená, tj. $n > m$. Nemá-li tato soustava přesné řešení, tedy pokud $b \notin \mathcal{R}(A)$, je třeba hledat v nějakém smyslu nejlepší opravu dat A nebo b tak, aby již pozorování korelovalo s modelem. Neboli uvažujeme lineární aproximační problém

$$Ax \approx b \qquad A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n. \tag{1.1}$$

Nejprve předpokládejme, že pouze vektor pozorování b je zatížen chybami a model A je dán přesně. Potom jedním z možných nástrojů k řešení úlohy (1.1) je minimalizace ve smyslu nejmenších čtverců, kterou můžeme definovat následujícím způsobem.

Definice 1 (Problém nejmenších čtverců (LS))**.**
Nechť $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$. Problém nejmenších čtverců je úloha nalezení vektorů $x \in \mathbb{R}^m$ a $f \in \mathbb{R}^n$, splňujících

$$\min_{f\in\mathbb{R}^n} ||f|| \quad \textit{tak, aby} \quad Ax = b + f. \tag{1.2}$$

Lze snadno nahlédnout, že řešení ve smyslu LS lze vyjádřit jako vektor minimalizující normu rezidua $||b - Ax||$ přes $x \in \mathbb{R}^m$. Neboli problém nejmenších čtverců má vždy řešení, což je jednou jeho z nezpochybnitelných výhod. Ovšem v praxi bývá předpoklad, že matice modelu A je dána přesně, často nerealistický, neboť chyby například v měření nebo modelování ovlivňují i matici A. Jedním ze způsobů, jak zohlednit i tyto perturbace modelu A, je v určitém smyslu zobecnění problému nejmenších čtverců, a to problém úplných nejmenších čtverců. Tedy nadále budeme předpokládat, že jak vektor b, tak matice A, jsou zatíženy chybami. Budeme hledat řešení aproximační úlohy (1.1) ve smyslu úplných nejmenších čtverců popsané v následující definici.

Definice 2 (Problém úplných nejmenších čtverců (TLS))**.**
Nechť $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$. Problém úplných nejmenších čtverců je úloha nalezení vektorů $x \in \mathbb{R}^m$, $f \in \mathbb{R}^n$ a matice $E \in \mathbb{R}^{n\times m}$, splňujících

$$\min_{f\in\mathbb{R}^n, E\in\mathbb{R}^{n\times m}} |||[f,E]|||_F \quad \textit{tak, aby} \quad (A+E)x = b+f. \tag{1.3}$$

Vektor f nazýváme oprava pozorování, matici E oprava modelu a matici $[f,E]$ nazýváme matice korekce dat.

Příklad. Mějme aproximační problém $Ax \approx b$, kde

$$A = \begin{pmatrix}0\\1\end{pmatrix}, b = \begin{pmatrix}2\\0\end{pmatrix}.$$

Pak minimální oprava dat je tvaru $[f,E] = \begin{pmatrix}0 & 0\\0 & -1\end{pmatrix}, |||[f,E]|||_F = 1$. Ta ale nepřevádí úlohu na kompatibilní problém. Vezmeme-li opravu dat tvaru $[\tilde{f},\tilde{E}] = \begin{pmatrix}0 & \epsilon\\0 & -1\end{pmatrix}$, kde $\epsilon > 0$, pak již opravená soustava má řešení a to je tvaru $x = \frac{2}{\epsilon}$. Pro tuto opravu platí $|||[\tilde{f},\tilde{E}]|||_F = \sqrt{1+\epsilon^2}$. Volba ϵ byla ale libovolná, můžeme tedy vzít ϵ libovolně malé a nalézt tak vždy menší opravu dat, která bude stále dávat kompatibilní problém.
Neexistuje tedy oprava dat minimální v normě, jak požaduje (1.3), tato úloha tedy nemá řešení ve smyslu TLS. To není překvapivé, neboť zde $\mathcal{R}(A) \perp b$, pozorování b není korelováno s modelem A a jediná smysluplná aproximace řešení je $x = 0$. TLS minimalizací však nelze získat.

Z příkladu výše vidíme, že problém úplných nejmenších čtverců nemusí mít vždy řešení, a to ani pro problém malých rozměrů s maticí plné sloupcové hodnosti. Abychom mohli analyzovat, kdy TLS řešení existuje a kdy je dokonce jednoznačné, připomeňme si nyní definici singulárního rozkladu matice (Duintjer Tebbens a kol. (2012), Kapitola 5; Golub a Van Loan (1996), Podkapitola 2.5).

Definice 3 (Singulární rozklad (SVD))**.**
Nechť $A \in \mathbb{R}^{n\times m}$ je matice hodnosti r, $r \in \mathbb{N}$. Singulárním rozkladem matice A nazveme rozklad

$$A = U\Sigma V^\top,$$

kde $U \in \mathbb{R}^{n\times n}$ a $V \in \mathbb{R}^{m\times m}$ jsou unitární matice, $\Sigma \in \mathbb{R}^{n\times m}$ je diagonální matice s čísly $\sigma_1,\ldots,\sigma_{min(n,m)}$ na diagonále, pro něž platí

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{min(n,m)} = 0.$$

Čísla $\sigma_1,\ldots,\sigma_{min(n,m)}$ se nazývají singulární čísla matice A. Vektory $u_1,\ldots,u_n$ jsou levé singulární vektory a $v_1,\ldots,v_m$ jsou pravé singulární vektory matice A.

Definice 4 (Dyadický rozvoj)**.**
Nechť $A \in \mathbb{R}^{n\times m}$ je matice hodnosti r, $r \in \mathbb{N}$, a nechť $A = U\Sigma V^\top$ je její singulární rozklad. Potom

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

nazveme dyadickým rozvojem matice A.
Trojice $(\sigma_i,\ u_i,\ v_i)$ se nazývá singulární triplet matice A.

Nakonec uveďme větu, kterou budeme taktéž potřebovat k analýze existence a jednoznačnosti TLS řešení. Její důkaz je k nalezení v Duintjer Tebbens a kol. (2012), str. 133-134.

Věta 1 (Eckart, Young, Mirsky)**.**
Nechť $A \in \mathbb{R}^{n\times m}$ je matice hodnosti r, $k < r$, $r, k \in \mathbb{N}$, a nechť $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ je její dyadický rozvoj. Potom nejlepší aproximací matice A maticí hodnosti k je matice

$$A^{(k)} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top.$$

Navíc platí

$$\min_{B\in\mathbb{R}^{n\times m}, rank(B)=k} ||A - B||_F = ||A - A^{(k)}||_F = \sqrt{\sum_{i=k+1}^{r} {\sigma_i}^2}.$$

1.2 Existence a jednoznačnost řešení

Rovnici (1.3) z definice problému úplných nejmenších čtverců lze ekvivalentně zapsat ve tvaru

$$([f,E] + [b,A]) \begin{pmatrix} -1 \\ x \end{pmatrix} = 0.$$

Toto vyjádření je ekvivalentní tomu, že existuje vektor v jádru matice $([f,E] + [b,A])$, který má nenulovou první složku. Tedy problém úplných nejmenších čtverců lze ekvivalentně zformulovat jako hledání minimální opravy dat tak, aby bylo jádro $([f,E] + [b,A])$ netriviální a existoval zde vektor s nenulovou první složkou.

Je-li totiž $z = (z_1, \ldots, z_{m+1})^\top \in Ker([f,E] + [b,A])$, splňující $z_1 \neq 0$, pak lze řešení aproximační úlohy (1.1) ve smyslu TLS vyjádřit jako

$$x^{TLS} = -\frac{1}{z_1}(z_2, \ldots, z_{m+1})^\top. \tag{1.4}$$

Analyzujme nyní existenci a jednoznačnost TLS řešení v několika případech zvlášť.

1) $\boldsymbol{rank(A) = m}$

Nejprve předpokládejme, že matice A má plnou sloupcovou hodnost. Potom matice $[b,A]$ má také plnou sloupcovou hodnost, neboť předpokládáme, že $b \notin \mathcal{R}(A)$. Neboli $Ker([b,A]) = span\{0\}$. Z odvození výše plyne, že hledáme nejmenší opravu dat $[f,E]$ tak, aby opravená matice dat $([f,E] + [b,A])$ již neměla plnou sloupcovou hodnost.

Vezmeme-li $[b,A] = \sum_{i=1}^{m+1} \sigma_i u_i v_i^\top$ dyadický rozvoj rozšířené matice dat, potom z Věty 1 plyne, že minimální oprava dat je tvaru

$$[f,E] = -\sigma_{m+1} u_{m+1} v_{m+1}^\top. \tag{1.5}$$

Potom je jádro matice $([f,E] + [b,A])$ netriviální. Zbývá analyzovat, zda pomocí této opravy dostaneme kompatibilní problém, jak požaduje (1.3).

$\boldsymbol{a)}$ $\boldsymbol{\sigma_m > \sigma_{m+1} > 0}$

Nyní předpokládejme, že nejmenší singulární číslo rozšířené matice dat je jednonásobné. Potom vezmeme-li opravu dat tvaru (1.5), bude jádro opravené matice dat tvaru $Ker([f,E] + [b,A]) = span\{v_{m+1}\}$, kde v_{m+1} je pravý singulární vektor příslušný nejmenšímu singulárnímu číslu rozšířené matice dat. Potom rozlišme dva případy.

- $e_1^\top v_{m+1} \neq 0$
  Nyní je-li první složka vektoru v_{m+1} nenulová, je TLS řešení tvaru popsaném v (1.4). Neboli v tomto případě existuje právě jedno TLS řešení úlohy (1.1). Podrobnější rozbor včetně důkazu existence jednoznačného TLS řešení pro tento případ lze najít ve Van Huffel a Vandewalle (1991), str. 34-35.

- $e_1^\top v_{m+1} = 0$
  Je-li první složka vektoru v_{m+1} nulová, nelze výše popsanou konstrukcí sestrojit opravu dat splňující (1.3). Dokonce platí, že pak řešení úlohy (1.1) ve smyslu TLS neexistuje. Pro detailní rozbor viz Golub a Van Loan (1980), Paige a Strakoš (2005).

$\boldsymbol{b)}$ $\boldsymbol{\sigma_l > \sigma_{l+1} = \ldots = \sigma_{m+1} > 0}$

Nyní předpokládejme, že nejmenší singulární číslo rozšířené matice dat je vícenásobné s násobností $(m-l+1)$, $l < m$. Vezměme $\mathcal{V}_{m+1} = span\{v_{l+1}, \ldots, v_{m+1}\}$, kde $v_{l+1}, \ldots, v_{m+1}$ jsou pravé singulární vektory příslušné σ_{m+1}. Tedy $\mathcal{V}_{m+1}$ je pravý singulární prostor příslušný k σ_{m+1}. Potom pro každý vektor $v \in \mathcal{V}_{m+1}$ a k němu příslušný levý singulární vektor u je oprava vyjádřená vztahem (1.5) možnou minimální opravou dat. Dále označme $V_{m+1} = [v_{l+1}, \ldots, v_{m+1}]$ matici pravého singulárního prostoru k σ_{m+1} a opět rozlišme dva případy.

- $e_1^\top V_{m+1} \neq 0$
  Nyní má-li matice V_{m+1} alespoň jeden prvek v prvním řádku nenulový, pak pro každý vektor z $\mathcal{V}_{m+1}$ s nenulovou první složkou je vektor x vyjádřený vzorcem (1.4) řešením problému (1.1) ve smyslu TLS. Neboli v tomto případě existuje nekonečně mnoho TLS řešení úlohy (1.1).

Je ale možné sestrojit právě jedno TLS řešení s nejmenší normou. K tomu použijeme unitární transformaci matice V_{m+1}, v našem případě je vhodná Householderova reflexe (Duintjer Tebbens a kol. (2012), Podkapitola 3.2).
Z vyjádření (1.4) plyne, že minimalizace normy řešení dosáhneme maximalizací první složky vektoru $v \in \mathcal{V}_{m+1}$ v absolutní hodnotě. Budeme tedy chtít matici V_{m+1} pomocí vhodné matice Householderovy reflexe H převést na matici

$$V_{m+1}H = \begin{bmatrix} \bullet & 0 & \dots & 0 \\ * & * & * & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & * & * \end{bmatrix}, \tag{1.6}$$

kde $\bullet = \pm||e_1^\top V_{m+1}||$. Neboli chceme nalézt matici reflexe prvního řádku matice V_{m+1}, tj. vektoru $y_1 = (e_1^\top V_{m+1})^\top$, na vektor $y_2 = \pm||y_1||e_1$. Potom jednotkovým normálovým vektorem nadroviny zrcadlení je $q = \frac{y_1 - y_2}{||y_1 - y_2||}$ a tedy matice Householderovy reflexe je tvaru $H = I - 2qq^\top$.
Pak TLS řešení úlohy (1.1) minimální v normě vyjádříme tvarem (1.4) při volbě pravého singulárního vektoru $\tilde{v} = V_{m+1}He_1$.
Pro detaily odvození včetně důkazu existence právě jednoho TLS řešení minimálního v normě v tomto případě viz Van Huffel a Vandewalle (1991), str. 57-58.

- $e_1^\top V_{m+1} = 0$
  Je-li první řádek matice V_{m+1} nulový, opět nelze výše popsanou konstrukcí sestavit minimální opravu dat, která by převedla (1.1) na kompatibilní soustavu. Stejně tak se dá dokázat, že pro tento případ řešení úlohy (1.1) ve smyslu TLS neexistuje. Pro podrobnosti viz Golub a Van Loan (1980), Paige a Strakoš (2005).

2) $\boldsymbol{rank(A) < m}$

Nakonec nechť matice A nemá plnou sloupcovou hodnost. Pak TLS řešení úlohy (1.1) opět neexistuje (Golub a Van Loan (1980), Paige a Strakoš (2005)).

V případech, kdy TLS řešení úlohy (1.1) neexistuje, je možné hledat tzv. negenerické řešení (Van Huffel a Vandewalle (1991), Sekce 3.4.1). To spočívá v nalezení nejmenšího singulárního čísla, pro něž již existuje pravý singulární vektor s nenulovou první složkou, označme je $\sigma_p > \sigma_{m+1}$ a $\mathcal{V}_p$ příslušný pravý singulární prostor. Pak lze zkonstruovat negenerické řešení jako ve (1.4), které je pro jednonásobné σ_p jednoznačné a pro vícenásobné je opět možné najít to s minimální normou. Příslušná oprava dat je pak tvaru jako ve (1.5), ta už ale není minimální, jak požaduje (1.3), tedy negenerické řešení už není řešení ve smyslu TLS.

Na závěr této sekce uveďme větu, která formuluje postačující podmínku pro existenci a jednoznačnost TLS řešení. Její tvrzení včetně důkazu je k nalezení například v původním článku Golub a Van Loan (1980).

Věta 2 (Golub, Van Loan)**.**
Nechť $A \in \mathbb{R}^{n\times m}$ je matice s plnou sloupcovou hodností a $b \in \mathbb{R}^n$ je vektor splňující $b \notin \mathcal{R}(A)$. Dále uvažujme dyadické rozvoje matice A a rozšířené matice dat $[b,A]$ tvaru

$$A = \sum_{i=1}^{m} \sigma'_i u'_i v'^{\top}_i \quad a \quad [b,A] = \sum_{i=1}^{m+1} \sigma_i u_i v_i^{\top}.$$

Pokud platí $\sigma'_m > \sigma_{m+1}$, pak existuje právě jedno TLS řešení úlohy $Ax \approx b$.

Jak jsme již uvedli, tato věta formuluje pouze postačující podmínku pro existenci jednoznačného TLS řešení, ne však nutnou. Její předpoklady totiž implikují, že nejmenší singulární číslo rozšířené matice dat je jednonásobné a příslušný pravý singulární vektor má nenulovou první složku, tedy první případ v 1a) z analýzy existence TLS řešení výše. Jak jsme ale ukázali, TLS řešení existuje za určitých podmínek i v případu 1b).

1.3 Klasický TLS algoritmus

Při výpočtu TLS řešení lze postupovat jako při odvozování jeho existence v minulé sekci. Nejprve spočteme singulární rozklad rozšířené matice dat $[b,A]$, z něj určíme nejmenší kladné singulární číslo σ_{min} včetně jeho násobnosti k a k němu matici příslušného pravého singulárního prostoru V_{min}. Následně zjistíme, jestli pravý singulární prostor obsahuje vektor s nenulovou první složkou. Jestli ano, vyjádříme TLS řešení vztahem (1.4), případně pomocí Householderovy reflexe nalezneme to minimální v normě, pokud $k > 1$. V opačném případě přistoupíme k dalšímu nejmenšímu singulárnímu číslu a celý postup opakujeme, potom je takto nalezené řešení pouze negenerické.

Algoritmus shrnující tento postup nazýváme klasickým TLS algoritmem. V předložené práci uvedeme pouze jeho základní verzi, pro podrobnosti a rozšíření viz Van Huffel a Vandewalle (1991), Sekce 3.6.1 a Kapitola 4.

Nyní uveďme základní verzi klasického TLS algoritmu. Mějme tedy matici $A \in \mathbb{R}^{n\times m}$ a vektor $b \in \mathbb{R}^n$ jako vstupní data. Předpokládejme navíc, že $b \notin \mathcal{R}(A)$. Výstupem algoritmu pak bude TLS řešení aproximační úlohy (1.1) nebo její negenerické řešení.

Algoritmus 1 Klasický TLS algoritmus

Vstup: A, b
Výstup: x^{TLS}

$U\Sigma V^\top$ ← singulární rozklad matice [b,A]
$\sigma_1, \ldots, \sigma_{m+1}$ ← prvky na diagonále matice Σ
$v_1, \ldots, v_{m+1}$ ← sloupce matice V
p ← index nejmenšího kladného singulárního čísla

loop
  k ← násobnost σ_p
  if $\exists i \in \{p-k+1, \ldots, p\} : e_1^\top v_i \neq 0$ **then**
    if k>1 **then**
      H ← matice Householderovy reflexe k $[v_{p-k+1}, \ldots, v_p]$ jako v (1.6)
      v_p ← $[v_{p-k+1}, \ldots, v_p]He_1$
    end if
    x^{TLS} ← $-\frac{1}{v_p(1)} v_p(2:end)$
    break
  else
    p ← $p-k$
  end if
end loop

return x^{TLS}

2. Výpočetní aspekty TLS

Ve druhé kapitole se budeme zabývat problémy, které mohou nastat při implementaci klasického TLS algoritmu představeného v předešlé podkapitole, a uvedeme způsoby, jakými je k nim možno přistupovat. Dále podrobně popíšeme metodu výpočtu singulárního rozkladu matice a podíváme se také na možné benefity při nahrazení výpočtu celého SVD pouze aproximací jeho části. Nakonec se podrobněji zaměříme na konkrétní způsoby aproximace nejmenších singulárních tripletů.

2.1 Problémy při implementaci TLS

V minulé kapitole jsme se zaměřili na klasickou teorii výpočtu a analýzy existence a jednoznačnosti TLS řešení a uvedli jsme základní teoretickou verzi klasického TLS algoritmu. Při jeho implementaci ale musíme být opatrnější. V praxi totiž nepočítáme v přesné aritmetice, ale v aritmetice konečné (FPA - Floating-point arithmetics). Tudíž všechna naše data jsou zaokrouhlována a výsledky počítány s konečnou přesností, tedy zatíženy šumem. To představuje určité komplikace při implementaci TLS.

První problém nastává hned při určení hodnosti matice, neboli nulových singulárních čísel ve spočteném SVD rozšířené matice dat $[b,A]$. Jelikož počítáme v konečné aritmetice, spočtený SVD je zatížen chybami a tudíž se nabízí otázka, která singulární čísla považovat za nulová. Jeden z možných způsobů je zavedení parametru R, kdy všechna jemu rovna nebo menší singulární čísla se považují za nulová. Možná volba parametru byla uvedena v Van Huffel a Vandewalle (1991), str. 89. Ta, za předpokladu, že chyby obsažené v rozšířené matici dat jsou nezávislé, stejně rozdělené s nulovou střední hodnotou a rozptylem σ, uvažuje parametr $R = \sqrt{2max\{n,m+1\}}\sigma$.

Obdobný problém nastává při určování násobnosti singulárního čísla, kdy opět kvůli šumu v datech musíme uvažovat způsoby, jimiž určíme, která singulární čísla budeme považovat za stejná. Jednou z možností, uvedené v Van Huffel a Vandewalle (1991), str. 89, je pokládat za stejná všechna singulární čísla z intervalu $[\sigma_p, \sqrt{\sigma_p^2 + \delta}]$, kde $\delta > 0$ může být zadáno uživatelem, nebo v případě jako výše je uvažováno $\delta = 2max\{n,m+1\}\sigma^2$.

Jiný přístup byl prezentován v Dvořák (2021), a to zavedení parametru tolerance $tol > 0$. Potom singulární čísla σ_i, $i < p$, považujeme za stejná se σ_p, jestliže splňují

$$\frac{\sigma_i - \sigma_p}{\sigma_p} < tol.$$

V této práci ve vlastní implementaci TLS algoritmu volíme parametr $R = 10^{-10}$ k identifikaci nulových singulárních čísel. Pro určení násobnosti singulárních čísel pak používáme druhý zmíněný přístup a parametr tol klademe taktéž roven 10^{-10}.

2.2 Výpočet singulárního rozkladu

V první kapitole jsme uvedli postup při výpočtu TLS řešení pomocí singulárního rozkladu rozšířené matice dat [b,A]. V této podkapitole se zaměříme na výpočet samotného SVD. Metoda výpočtu singulárního rozkladu byla podrobně popsána a analyzována v původním článku Golub a Kahan (1965). Dále čerpáme z Golub a Van Loan (1996), Podkapitola 8.6, a Duintjer Tebbens a kol. (2012), Kapitola 5.

V této podkapitole tedy uvažujeme matici $A \in \mathbb{R}^{n\times m}$, $n > m$, pro níž chceme spočíst její singulární rozklad. Tento předpoklad není omezující, v opačném případě bychom mohli počítat SVD transponované matice.

2.2.1 Vztah mezi singulárním a spektrálním rozkladem

Nejprve se zaměřme na vztah mezi singulárním a spektrálním rozkladem, který můžeme definovat následujícím způsobem (viz například Duintjer Tebbens a kol. (2012), str. 44).

Definice 5 (Spektrální rozklad pro symetrické matice)**.**
Nechť $A \in \mathbb{R}^{n\times n}$ je symetrická matice. Pak existuje její rozklad tvaru

$$A = X\Lambda X^\top,$$

kde $\Lambda \in \mathbb{R}^{n\times n}$ je diagonální matice s vlastními čísly A na diagonále a $X \in \mathbb{R}^{n\times n}$ je unitární matice příslušných vlastních vektorů. Tento rozklad nazýváme spektrálním rozkladem matice A.

Uvažujme $A = U\Sigma V^\top$ singulární rozklad matice A dle Definice 3. Potom, uvažujeme-li matici $A^\top A \in \mathbb{R}^{m\times m}$, pak je její spektrální rozklad zřejmě tvaru

$$A^\top A = V\ diag(\sigma_1^2, \ldots, \sigma_m^2)\ V^\top.$$

Stejně tak, uvažujeme-li matici $AA^\top \in \mathbb{R}^{n\times n}$, její spektrální rozklad je tvaru

$$AA^\top = U\ diag(\sigma_1^2, \ldots, \sigma_m^2, 0, \ldots, 0)\ U^\top.$$

Neboli na singulární rozklad matice A lze nahlížet jako na spektrální rozklad symetrické pozitivně semidefinitní matice $A^\top A$ nebo $AA^\top$, podle toho, která volba redukuje rozměry problému. Je-li navíc $rank(A) = m$, pak $A^\top A$ je regulární. Ze spektrálního rozkladu pozitivně semidefinitní matice rovnou dostaneme pravé nebo levé singulární vektory a singulární čísla matice A vyjádříme jako odmocniny z vlastních čísel. Zbývající singulární vektory dopočteme dle vztahu

$$Av_i = \sigma_i u_i, \qquad i = 1, \ldots, rank(A).$$

Výpočet singulárního rozkladu tímto způsobem může být výhodný v případech, kdy jeden z rozměrů matice je malý. Značnou nevýhodou však je, že podmíněnost zmíněných pozitivně semidefinitních matic je rovna druhé mocnině podmíněnosti původní matice A, kde uvažujeme tzv. zobecněné číslo podmíněnosti. To je definováno vztahem $\kappa(A) = \frac{\sigma_1}{\sigma_r}$, kde r je hodnost matice A. Neboli, je-li matice

A špatně podmíněna, převedli jsme už tak špatně podmíněný problém na ještě mnohem hůře podmíněný a tímto způsobem spočtená malá singulární čísla mohou být určena nepřesně.

Na singulární rozklad matice A se ale dá nahlížet ještě jedním způsobem, formulovaným v následujícím lemmatu (Larsen (1998), Tvrzení 1).

Lemma 3. *Nechť $A \in \mathbb{R}^{n\times m}$, $n > m$, je matice, $A = U\Sigma V^\top$ její singulární rozklad a $U = [U_1, U_2]$, kde $U_1 \in \mathbb{R}^{n\times m}, U_2 \in \mathbb{R}^{n\times(n-m)}$.*

Označme $Q = \frac{1}{\sqrt{2}}\begin{bmatrix} U_1 & U_1 & \sqrt{2}U_2 \\ V & -V & 0 \end{bmatrix}$ a uvažujme matici $\begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$.

Pak její spektrální rozklad je tvaru

$$\begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} = Q\ diag(\sigma_1, \ldots, \sigma_m, -\sigma_1, \ldots, -\sigma_m, 0, \ldots, 0)\ Q^\top.$$

Z lemmatu je zřejmé, že spektrální rozklad matice $\begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$ nám rovnou dává singulární čísla a pravé i levé singulární vektory matice A, není tedy nutné nic dopočítávat. Tento přístup sice zvětšuje dimenzi problému, avšak nezhoršuje takovým způsobem jeho podmíněnost, jako konstrukce s maticemi $A^\top A$ a $AA^\top$.

2.2.2 Transformace na bidiagonální tvar

Výpočet singulárního rozkladu matice A spočívá ve dvou krocích. Prvním krokem je převedení matice A na horní bidiagonální tvar. Důvodem je maximální zjednodušení struktury matice pro výpočetní úsporu v následujících krocích. V případě, že je matice A řídká, lze tuto transformaci provést tzv. Golub-Kahanovou iterační bidiagonalizací (Golub a Kahan (1965)). Jinak se standardně využívají unitární transformace, tj. Givensovy rotace (Duintjer Tebbens a kol. (2012), Podkapitola 3.1) nebo Householderovy reflexe (Duintjer Tebbens a kol. (2012), Podkapitola 3.2). Ty zaručují numerickou stabilitu, není však vhodné je aplikovat na velké řídké matice, neboť hned při prvním kroku může dojít k jejich výraznému zaplnění.

My uvedeme transformaci matice A na horní bidiagonální tvar pomocí Householderových reflexí. Připomeňme, že tyto matice jsou tvaru $H = I - 2qq^\top$, kde q je jednotkový normálový vektor nadroviny zrcadlení, viz podkapitola 1.2. Budeme konstruovat matice $H_1, \ldots, H_{m-1} \in \mathbb{R}^{n\times n}$, které budou nulovat poddiagonální prvky ve sloupcích, a matice $\tilde{H}_1, \ldots, \tilde{H}_{m-2} \in \mathbb{R}^{m\times m}$, které vynulují prvky mimo diagonálu a naddiagonálu v řádcích, dle následujícího schématu.

$$A = \begin{bmatrix} \bullet & \bullet & \bullet & \ldots & \bullet \\ \bullet & \bullet & \bullet & \ldots & \bullet \\ \bullet & \bullet & \bullet & \ldots & \bullet \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bullet & \bullet & \bullet & \ldots & \bullet \end{bmatrix} \xrightarrow{H_1} \begin{bmatrix} \bullet & \bullet & \bullet & \ldots & \bullet \\ 0 & \bullet & \bullet & \ldots & \bullet \\ 0 & \bullet & \bullet & \ldots & \bullet \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \bullet & \bullet & \ldots & \bullet \end{bmatrix} \xrightarrow{\tilde{H}_1} \begin{bmatrix} \bullet & \bullet & 0 & \ldots & 0 \\ 0 & \bullet & \bullet & \ldots & \bullet \\ 0 & \bullet & \bullet & \ldots & \bullet \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \bullet & \bullet & \ldots & \bullet \end{bmatrix}$$

$$\xrightarrow{H_2} \begin{bmatrix} \bullet & \bullet & 0 & \dots & 0 \\ 0 & \bullet & \bullet & \dots & \bullet \\ 0 & 0 & \bullet & \dots & \bullet \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \bullet & \dots & \bullet \end{bmatrix} \xrightarrow{\tilde{H}_2} \quad \dots \quad \xrightarrow{H_{m-1}} \begin{bmatrix} \bullet & \bullet & & & \\ & \ddots & \ddots & & \\ & & \ddots & \bullet & \\ & & & \bullet & \end{bmatrix}$$

Označme nyní

$$(H_{m-1} \dots H_1) A (\tilde{H}_1 \dots \tilde{H}_{m-2}) = \begin{bmatrix} B \\ 0 \end{bmatrix},$$

kde $B \in \mathbb{R}^{m \times m}$ je čtvercová matice horního bidiagonálního tvaru. Převedli jsme tedy problém nalezení singulárního rozkladu matice A na nalezení rozkladu matice B. Pro jednoduchost dále předpokládejme, že matice B má plnou sloupcovou hodnost. Jinak lze matici B převést na blokově diagonální tvar a tedy zredukovat problém na hledání SVD dvou matic menších rozměrů (Golub a Van Loan (1996), str. 454).

Při této transformaci jsme konstruovali $(m-2)$ matic Householderových reflexí k vynulování řádků a $(m-1)$ k vynulování sloupců. Řádové výpočetní náklady se rovnají $4nm^2 - \frac{4}{3}m^3$ operacím, viz Golub a Van Loan (1996), Sekce 5.4.3.

2.2.3 Modifikace implicitního QR algoritmu

V sekci 2.2.1 jsme ukázali, že možnými metodami pro výpočet singulárního rozkladu matice B je výpočet spektrálního rozkladu matic $B^\top B$ nebo $\begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}$. My uvedeme metodu založenou na rozkladu matice $B^\top B$, aniž bychom ji museli explicitně konstruovat. Tím se vyhneme zvyšování podmíněnosti problému a s tím spojeným nepříjemnostem. Stejně tak lze tuto metodu použít pro výpočet rozkladu druhé matice bez její explicitní konstrukce.

Mějme tedy matici $B^\top B$, ta je tridiagonální symetrická pozitivně definitní. Pro výpočet jejího spektrálního rozkladu se využívá tzv. shiftovaného implicitního QR algoritmu pro symetrické matice (Watkins (2010), Podkapitola 5.6, Golub a Van Loan (1996), Sekce 8.3.5), nazývaného také Francisův algoritmus. Tento vychází z jednoduše vypadajícího explicitního QR algoritmu (Golub a Van Loan (1996), Podkapitola 7.5). Explicitní QR algoritmus nyní uvedeme v obecném tvaru pro matici A_0 nesymetrickou, která byla v rámci preprocessingu převedena na horní Hessenbergův tvar.

Algoritmus 2 Explicitní QR algoritmus

Vstup: A_0 v horním Hessenberově tvaru
Výstup: $A_k, Q_1, \ldots, Q_k$

for $k = 1, 2, \ldots$ **do**
  $Q_k R_k \quad \leftarrow$ QR rozklad matice A_{k-1}
  $A_k \quad \leftarrow R_k Q_k$
end for

Pro definici a rozbor QR rozkladu viz například Duintjer Tebbens a kol. (2012), Podkapitola 3.5. QR rozklad matice A_{k-1} v Algoritmu 2 provádíme pomocí Givensových rotací. Protože matice A_{k-1} je horní Hessenbergova pro všechna $k = 1, 2, \ldots$, což plyne z konstrukce, postačí k tomu $(m-1)$ matic Givensových rotací $G_1^{(k)}, \ldots, G_{m-1}^{(k)}$ nulujících její poddiagonální prvky.

Uveďme nyní větu, kterou budeme potřebovat pro následující diskuzi. Její tvrzení včetně důkazu je k nalezení v Duintjer Tebbens a kol. (2012), Kapitola 2.

Věta 4 (Schurova)**.**
Nechť $A \in \mathbb{R}^{n\times n}$ je matice. Pak existuje obecně komplexní unitární matice $X \in \mathbb{C}^{n\times n}$ taková, že platí

$$A = XRX^*,$$

kde $R \in \mathbb{C}^{n\times n}$ je horní trojúhelníková matice s vlastními čísly matice A na diagonále. Tento rozklad nazýváme Schurovým rozkladem matice A.
Je-li navíc matice A normální, pak je matice R diagonální a matice X obsahuje ve sloupcích vlastní vektory matice A. Je-li matice A symetrická, jsou matice R a X reálné.

Všimněme si, že pro matice A_k z explicitního QR algoritmu platí

$$A_k = Q_k^\top A_{k-1} Q_k = \cdots = Q_k^\top \ldots Q_1^\top A_0 Q_1 \ldots Q_k.$$

Neboli všechny matice A_0, A_{k-1}, A_k jsou si unitárně podobné a mají tedy stejná vlastní čísla. Označíme-li navíc matici $\tilde{Q}_k = Q_1 \ldots Q_k$, pak platí

$$A_0 = \tilde{Q}_k A_k \tilde{Q}_k^\top. \tag{2.1}$$

Potom za předpokladu, že neexistují dvě různá vlastní čísla matice A_0 stejná v absolutní hodnotě, lze dokázat, že matice A_k konverguje k horní trojúhelníkové matici (Watkins (2002), Podkapitola 6.2, Golub a Van Loan (1996), Sekce 7.3.3). Neboli výraz (2.1) konverguje k Schurovu rozkladu matice A_0.

Matice A_k z explicitního QR algoritmu tedy aproximuje horní trojúhelníkovou matici a matice $\tilde{Q}_k$ unitární matici ze Schurova rozkladu matice A_0. Je-li navíc A_0 normální, speciálně symetrická, aproximuje matice A_k diagonální matici s vlastními čísly A_0 na diagonále a matice $\tilde{Q}_k$ aproximuje unitární matici vlastních vektorů.

Tato verze QR algoritmu klade ale velmi silný předpoklad na vlastní čísla matice A_0, aby byla zajištěna konvergence. Tomu se lze vyhnout aplikací tzv. shiftu (Golub a Van Loan (1996), Sekce 7.5.2 - 7.5.4), který zároveň urychluje konvergenci algoritmu. Označme $\rho_k \in \mathbb{R}, \rho_k \neq 0$ zvolený shift. V shiftovaném explicitním algoritmu se v každé iteraci místo QR rozkladu matice A_{k-1} počítá QR rozklad shiftované matice $A_{k-1} - \rho_k I$, stejně tak se pak shift přičítá zpět k matici A_k, aby byla zajištěna podobnost matic A_0, A_{k-1}, A_k. Tedy uvnitř for cyklu počítáme

$$\begin{aligned} Q_k R_k \quad &\leftarrow \text{ QR rozklad matice } A_{k-1} - \rho_k I \\ A_k \quad &\leftarrow R_k Q_k + \rho_k I \end{aligned}$$

Možnými volbami shiftu ρ_k jsou například Reyileighův shift nebo Wilkinsonův dvojitý double shift, které v každé iteraci aproximují nejmenší nebo komplexně sdružená vlastní čísla matice A_0.

Tato implementace ale představuje problém v momentě, kdy je shift dobrou aproximací vlastního čísla, pak je totiž shiftovaná matice téměř singulární, což má vliv na spolehlivost výpočtu. Řešení představuje tzv. implicitní QR algoritmus (Golub a Van Loan (1996), Sekce 7.5.5), který umožňuje přechod k nové aproximační matici A_k bez explicitní konstrukce problematické shiftované matice. Verzi implicitního QR algoritmu s Wilkinsonovým shiftem pro symetrické matice používáme při výpočtu spektrálního rozkladu matice $B^\top B$. Základní konstrukci si nyní přiblížíme.

Kdybychom aplikovali implicitní QR algoritmus s Wilkinsonovým shiftem pro symetrické matice na matici $A_0 = B^\top B$, pak bychom v každé iteraci spočetli aktuální shift ρ_k, matici Givensovy rotace $G_1^{(k)}$, která nuluje první poddiagonální prvek shiftované matice $A_{k-1} - \rho_k I$ a tu aplikovali na matici A_{k-1} zleva a následně i transponovanou zprava. To by vytvořilo obecně nenulový prvek, tzv. bulge, na pozici (3,1) matice

$$G_1^{(k)} A_{k-1} (G_1^{(k)})^\top.$$

Následně bychom zkonstruovali matici Givensovy rotace $G_2^{(k)}$, která by při aplikaci zleva vynulovala bulge na pozici (3,1) a při aplikaci transponované matice zprava by vytvořila nový na pozici (4,2). Analogicky bychom postupovali dále, než bychom bulge vysunuli ven z matice. Potom pro výslednou matici platí

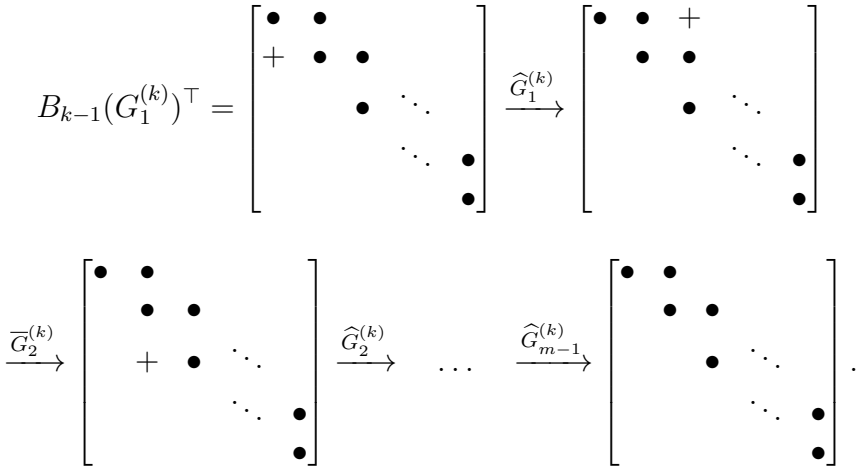
$$(G_{m-1}^{(k)} \dots G_1^{(k)}) A_{k-1} ((G_1^{(k)})^\top \dots (G_{m-1}^{(k)})^\top) = Q_k^\top A_{k-1} Q_k = A_k. \tag{2.2}$$

Pro podrobný popis implicitního QR algoritmu viz Watkins (2010), Podkapitola 5.6.

My se ale z dříve uvedených důvodů chceme vyhnout konstrukci matice $B^\top B$. Aplikujme tedy matici rotace $G_1^{(k)}$ přímo na matici B_{k-1} při značení $B_0 = B$,

$$B_{k-1} = \begin{bmatrix} \bullet & \bullet & & & \\ & \bullet & \bullet & & \\ & & \bullet & \ddots & \\ & & & \ddots & \bullet \\ & & & & \bullet \end{bmatrix} \xrightarrow{(G_1^{(k)})^\top} \begin{bmatrix} \bullet & \bullet & & & \\ + & \bullet & \bullet & & \\ & & \bullet & \ddots & \\ & & & \ddots & \bullet \\ & & & & \bullet \end{bmatrix} = B_{k-1} (G_1^{(k)})^\top.$$

Přenásobení matice B_{k-1} transponovanou maticí $G_1^{(k)}$ zprava nyní vytvořilo bulge na pozici (2,1). Dále konstruujeme matice Givensových rotací $\widehat{G}_1^{(k)}, \ldots, \widehat{G}_{m-1}^{(k)}$ a $\overline{G}_2^{(k)}, \ldots, \overline{G}_{m-1}^{(k)}$ analogicky jako v minulém odstavci, abychom vysunuli nechtěný bulge ven z matice.



Všimněme si, že při variantě QR algoritmu pro čtvercovou matici B není nutné zachovávat podobnost matic B_0, B_{k-1}, B_k. Rotace konstruované zprava a zleva tedy nejsou stejné.

Označme výslednou matici

$$(\widehat{G}_{m-1}^{(k)} \ldots \widehat{G}_1^{(k)}) B_{k-1} ((G_1^{(k)})^\top \overline{G}_2^{(k)}, \ldots, \overline{G}_{m-1}^{(k)}) = U_k^\top B_{k-1} V_k = B_k. \tag{2.3}$$

Potom lze dokázat, že matice Q_k zkonstruovaná při aplikaci implicitního QR algoritmu na matici $B^\top B$, viz (2.2), se v každé iteraci rovná matici V_k zkonstruované při aplikaci Givensových rotací přímo na matici B z rovnosti (2.3). Pro podrobnosti viz Golub a Van Loan (1996), str. 347 a 454. Neboli podle diskuze výše matice $\tilde{V}_k = V_1 \ldots V_k$ aproximuje matici vlastních vektorů matice $B^\top B$, tedy matici pravých singulárních vektorů matice B. Z toho plyne, že matice $\tilde{U}_k = U_1 \ldots U_k$ aproximuje matici levých singulárních vektorů matice B a zároveň zřejmě matice B_k aproximuje diagonální matici se singulárními čísly matice B na diagonále.

Celkem jsme tedy modifikovali implicitní QR algoritmus pro výpočet spektrálního rozkladu matice $B^\top B$ tak, abychom nemuseli tuto matici explicitně konstruovat a zároveň rovnou spočetli požadovaný singulární rozklad matice B.

Nakonec si shrňme celý postup výpočtu SVD matice A. Nejprve jsme zadanou matici A převedli do horního bidiagonálního tvaru a tím zredukovali problém nalezení jejího SVD na nalezení SVD menší čtvercové bidiagonální matice B. Následně jsme využili souvislosti mezi SVD matice B a spektrálním rozkladem matice $B^\top B$ a modifikací známé metody pro problém vlastních čísel jsme spočetli singulární rozklad matice B.

V každé iteraci QR algoritmu jsme konstruovali $(2m-2)$ matic Givensových rotací, kdy jedna vyžaduje řádově $O(m)$ operací. Celkové náklady výpočtu singulárního rozkladu tedy tvoří náklady za preprocessing matice a $O(m^2)$ za každou iteraci QR algoritmu. Pro podrobný rozbor viz Golub a Van Loan (1996), Sekce 5.4.5.

Je dokázáno, že tímto způsobem lze singulární čísla spočíst velmi přesně, s relativní přesností na úrovni strojové přesnosti (Barlow (2001), Demmel a Kahan (1990)). Kompletní SVD algoritmus je k nalezení v Watkins (2010), Podkapitola 5.8, nebo Golub a Van Loan (1996), Algoritmy 5.4.2, 8.6.1 a 8.6.2.

2.3 Metody aproximace singulárních tripletů

V první kapitole jsme uvedli postup při výpočtu TLS řešení pomocí singulárního rozkladu rozšířené matice dat $[b,A]$. Jak je z odvození zřejmé, pro jeho výpočet ale není třeba znát celý SVD matice dat, nýbrž jen jeho část. Potřebujeme mít k dispozici pouze nejmenší singulární čísla a k nim příslušné pravé singulární prostory, tedy nám stačí znát jen nejmenší singulární triplety matice dat. Zároveň výpočet singulárního rozkladu je obecně náročný a nákladný, často tedy přistupujeme pouze k aproximaci částečného SVD. V této podkapitole se zaměříme na některé metody aproximace nejmenších singulárních tripletů.

Existuje řada publikací zabývajících se vývojem nových a vylepšováním již existujících metod aproximace částečného singulárního rozkladu matice. My zde uvedeme iterační metodu pro převedení matice na bidiagonální tvar, tzv. Golub-Kahanovu iterační bidiagonalizaci (GKB), a vysvětlíme, jak lze na jejím základě získat aproximaci částečného SVD. Ta byla podrobně popsána v původním článku Golub a Kahan (1965), nebo v Larsen (1998) a Baglama a Reichel (2005). Metoda GKB úzce souvisí s tzv. Lanczosovou metodou uvedenou v Lanczos (1950), dále viz Duintjer Tebbens a kol. (2012), Podkapitola 7.2.

Mějme matici $A \in \mathbb{R}^{n\times m}$. GKB převádí matici A na základě projekcí do dvou Krylovových podprostorů, pro vysvětlení viz Duintjer Tebbens a kol. (2012), Kapitola 9, na (částečný) dolní bidiagonální tvar. Vstupní data GKB algoritmu jsou matice A a startovní vektor $x_0 \in \mathbb{R}^n$. V každé iteraci spočteme tzv. levé a pravé bidiagonalizační vektory $s_{i+1} \in \mathbb{R}^n$ a $w_i \in \mathbb{R}^m$ a skaláry α_i a β_{i+1}, prvky kýžené bidiagonální matice B_k.

Algoritmus 3 Golub-Kahanova iterační bidiagonalizace

Vstup: A, x_0
Výstup: $s_1, \ldots, s_{k+1}, w_1, \ldots, w_k, \alpha_1, \ldots, \alpha_k, \beta_2, \ldots, \beta_{k+1}$

$\beta_1 \leftarrow ||x_0||$
$s_1 \leftarrow \frac{x_0}{\beta_1}$
$w_0 \leftarrow 0$
for $i = 1, \ldots, k$ **do**
  $y_i \leftarrow A^\top s_i - \beta_i w_{i-1}$
  $\alpha_i \leftarrow ||y_i||$
  $w_i \leftarrow \frac{y_i}{\alpha_i}$
  $x_i \leftarrow A w_i - \alpha_i s_i$
  $\beta_{i+1} \leftarrow ||x_i||$
  $s_{i+1} \leftarrow \frac{x_i}{\beta_{i+1}}$
end for

Protože čísla α_i a β_{i+1} pochází z normalizace, platí $\alpha_i \geq 0$ a $\beta_{i+1} \geq 0$ pro všechny relevantní indexy i. Pokud by v některé iteraci $\alpha_i = 0$ nebo $\beta_{i+1} = 0$, pak by se výpočet zastavil. Pro jednoduchost předpokládejme, že k tomu pro náš počet iterací k nedošlo. Pak výsledná bidiagonální matice je tvaru

$$B_k = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{bmatrix} \in \mathbb{R}^{(k+1)\times k}.$$

Označme $S_{k+1} = [s_1, \ldots, s_{k+1}]$, $W_k = [w_1, \ldots, w_k]$ matice levých a pravých bidiagonalizačních vektorů. Z konstrukce vektorů x_i a y_i v algoritmu pak okamžitě plynou dva rekurenční vztahy, které lze v maticovém tvaru formulovat následovně

$$AW_k = S_{k+1}B_k,$$

$$A^\top S_{k+1} = W_k B_k^\top + \alpha_{k+1} w_{k+1} e_{k+1}^\top.$$

Zároveň jsou z konstrukce levé a pravé bidiagonalizační vektory v přesné aritmetice ortonormální a platí následující lemma, jehož důkaz je čistě konstrukční a proto ho zde neuvádíme.

Lemma 5. *Nechť $A \in \mathbb{R}^{n\times m}$ je matice a $s_i, s_{k+1} \in \mathbb{R}^n$ a $w_i \in \mathbb{R}^m$, $i = 1, \ldots, k$, jsou bidiagonalizační vektory spočtené Algoritmem 3. Pak pro následující Krylovovy prostory platí*

$$\mathcal{K}_{k+1}(AA^\top, s_1) \equiv span\{s_1, AA^\top s_1, \ldots, (AA^\top)^k s_1\} = span\{s_1, \ldots, s_{k+1}\},$$

$$\mathcal{K}_k(A^\top A, w_1) \equiv span\{w_1, A^\top A w_1, \ldots, (A^\top A)^{k-1} w_1\} = span\{w_1, \ldots, w_k\}.$$

Neboli levé a pravé bidiagonalizační vektory tvoří ortonormální báze těchto Krylovových podprostorů. Pokud by α_i nebo β_{i+1} byly v nějaké iteraci i rovny nule, znamená to, že příslušný Krylovův prostor je již A-invariantní.

Lze dokázat, že GKB je velmi úzce spjatá s Lanczosovou metodou (Golub a Kahan (1965), viz též Larsen (1998)). Konkrétně GKB matice A je ekvivalentní Lanczosově metodě aplikované na matici $\overline{A} = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$ se startovním vektorem $\begin{pmatrix} s_1 \\ 0 \end{pmatrix} \in \mathbb{R}^{(n+m)}$. Ta převádí matici iteračně na (částečný) tridiagonální tvar $T_k \in \mathbb{R}^{k\times k}$ pomocí projekcí do sekvence Krylovových prostorů. Označme $L_k \in \mathbb{R}^{(n+m)\times k}$ matici s navzájem ortonormálními sloupci tvořícími bázi příslušného Krylovova prostoru, která je výstupem Lanczosovy metody. Dále označme (μ, z) vlastní pár tridiagonální matice T_k. V případě, že se k rovná maximální možné dimenzi zmíněného Krylovova prostoru, tj. dimenzi, kdy už je prostor invariantní na násobení maticí $\overline{A}$, tvoří vlastní čísla matice T_k podmnožinu vlastních čísel matice $\overline{A}$. Platí totiž

$$\overline{A}L_k = L_k T_k \qquad \text{a tedy} \qquad \overline{A}(L_k z) = L_k T_k z = \mu(L_k z).$$

Neboli μ je vlastní číslo matice $\overline{A}$ a $L_k z$ je příslušný vlastní vektor. Pokud je ale k menší, platí pouze vztah

$$\overline{A}L_k = L_k T_k + t_{k+1,k} l_{k+1} e_k^\top,$$

kde $t_{k+1,k}$ značí prvek matice T_{k+1} na pozici $(k+1,k)$ a l_{k+1} je poslední sloupcový vektor matice L_{k+1}. Pak vlastní čísla matice T_k pouze aproximují k (typicky největších) vlastních čísel matice $\overline{A}$. V tomto případě nazýváme μ Ritzova čísla a vektory $L_k z$ Ritzovy vektory matice $\overline{A}$.

Podle souvislosti mezi SVD matice A a spektrálním rozkladem matice $\overline{A}$ uvedené v sekci 2.2.1 a podle diskuze výše, lze Ritzova čísla a vektory matice $\overline{A}$ nalézt spočtením SVD matice B_k z GKB. Neboli, označíme-li

$$B_k = \tilde{U}\ diag(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_k)\ \tilde{V}^\top \quad \text{a} \quad A = U\ diag(\sigma_1, \ldots, \sigma_m)\ V^\top$$

singulární rozklady matic B_k a A, pak za určitých předpokladů platí

$$\begin{aligned} \tilde{\sigma}_i &\approx \sigma_i, \\ S_k \tilde{u}_i &\approx u_i, \\ W_k \tilde{v}_i &\approx v_i, \end{aligned} \qquad i = 1, \ldots, k.$$

Pro podrobnou analýzu viz například Larsen (1998).

Neboli singulární čísla matice B_k aproximují k singulárních čísel matice A a singulární vektory A lze aproximovat pomocí singulárních vektorů B_k po přenásobení maticí příslušných bidiagonalizačních vektorů. Navíc SVD matice B_k umíme rychle a spolehlivě spočítat pomocí metody uvedené v sekci 2.2.3.

Lze nahlédnout, že je pro výpočet SVD výhodnější použít GKB namísto Lanczosovy metody, neboť využívá speciální struktury matice $\overline{A}$ a neoperuje zbytečně s jejími nulovými bloky.

Pro shrnutí analýzy chyby GKB viz Larsen (1998) a zde uvedené reference. Zároveň jsme prezentovali pouze elementární verzi GKB, existuje mnoho sofistikovanějších implementací a úprav. Například GKB s částečnou nebo úplnou reortogonalizací (Larsen (1998)), neboť stejně jako u Lanczosovy metody i u GKB způsobují zaokrouhlovací chyby v konečné aritmetice ztrátu ortogonality, nebo implicitně restartovaná GKB (Baglama a Reichel (2005)).

V numerických experimentech budeme pro výpočet celého SVD matice dat $[b,A]$ používat funkci 'svd' implementovanou v MATLABu, jejímž základem je metoda popsaná v podkapitole 2.2. K aproximaci singulárních tripletů pak budeme využívat funkci 'svds' implementovanou v MATLABu, která je založena na výše popsaném přístupu, jehož jádrem je Golub-Kahanova iterační bidiagonalizace. Pro srovnání pak použijeme dobře známou funkci 'lobpcg' staženou z https://www.mathworks.com/matlabcentral/fileexchange/48-locally-optimal-block-preconditioned-conjugate-gradient. Ta je založena na metodě předpodmíněných sdružených gradientů (Paige a Saunders (1975)) a slouží k aproximaci nejmenších vlastních párů hermitovských matic. Proto ji podle sekce 2.2.1 budeme používat pro aproximaci vlastních párů matice $[b,A]^\top[b,A]$. Tuto metodu z důvodu její komplexnosti zde již podrobně popisovat nebudeme. Pro detaily viz Knyazev (2001) a dále Hetmaniuk a Lehoucq (2006), Duersch a kol. (2018) aj.

3. Numerické experimenty

V minulých kapitolách jsme podrobně rozebrali problém úplných nejmenších čtverců a způsob využití celého singulárního rozkladu matice dat $[b,A]$, případně možných aproximací jeho části, při konstrukci TLS řešení aproximační úlohy (1.1). V této kapitole se zaměříme na numerické experimenty. Zejména budeme zkoumat citlivost konstruovaného TLS řešení na perturbace v datech. Porovnáme tři možné metody aproximace nejmenších singulárních tripletů - funkce 'svd', 'svds' a 'lobpcg'. Budeme testovat vliv kvality této aproximace na spočtené TLS řešení.

Experimenty provádíme na počítači s 1,4 GHz čtyřjádrovým Intel Core i5 procesorem, ve verzi R2021b prostředí MATLAB. Používáme zejména vlastní programy, z nichž nejvýznamnější je implementace TLS algoritmu z podkapitoly 1.3, funkce 'tls', s možností volby jedné ze tří aproximací částečného SVD matice dat. Tady využíváme některých technik diskutovaných ve druhé kapitole a také v MATLABu zabudovaných funkcí 'svd' a 'svds', stejně jako funkce 'lobpcg'.

3.1 Testovací problémy

Pro účely experimentů jsme generovali testovací matice dat $[b,A] \in \mathbb{R}^{n\times(m+1)}$ s předem známými singulárními čísly způsobem, který byl prezentován v Dax (2019). Tedy vzali jsme diagonální matici $\Sigma \in \mathbb{R}^{n\times(m+1)}$ se zvolenými singulárními čísly na diagonále. Pomocí příkazu 'randn' jsme vygenerovali náhodné vektory $h_1 \in \mathbb{R}^n$ a $h_2 \in \mathbb{R}^{(m+1)}$, s jejichž využitím jsme pak položili matice pravých a levých singulárních vektorů rovny Householderovým maticím tvaru

$$U = I - 2\frac{h_1 h_1^\top}{h_1^\top h_1} \in \mathbb{R}^{n\times n} \quad \text{a} \quad V = \left(I - 2\frac{h_2 h_2^\top}{h_2^\top h_2}\right)^\top \in \mathbb{R}^{(m+1)\times(m+1)}.$$

To jsou jistě unitární matice, tedy rozklad tvaru

$$C = U\Sigma V^\top$$

je singulárním rozkladem matice, kterou označme C. V experimentech jsme pak první sloupcový vektor takto generovaných matic považovali za vektor pozorování z aproximačního problému (1.1) a zbylou matici za danou matici modelu, tedy $C = [b,A]$.

Pro účely následující diskuze jsme zvolili dvě testovací matice dat generované popsaným způsobem, obě s rozměry (2000×1000). První matice C má rovnoměrně rozdělená singulární čísla tvaru

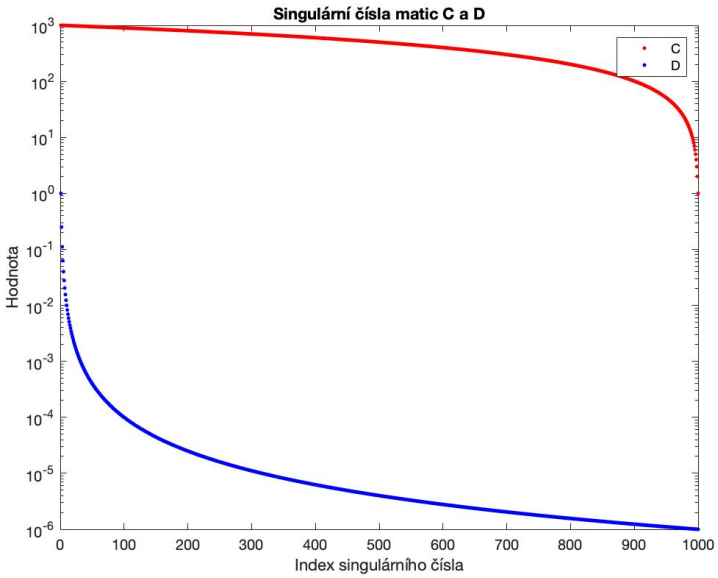
$$\tilde{\sigma}_i = 1000 - i + 1, \qquad i = 1,\ldots,1000.$$

Matice je zřejmě plné sloupcové hodnosti, má jednonásobná singulární čísla a její číslo podmíněnosti je rovno $\kappa(C) = 10^3$.

Druhou testovací matici generujeme analogicky a pro rozlišení ji značíme D. Ta má rychle klesající singulární čísla daná vztahem

$$\overline{\sigma}_i = \frac{1}{i^2}, \qquad i = 1, \ldots, 1000.$$

Tato má taktéž plnou sloupcovou hodnost, jednonásobná singulární čísla a její číslo podmíněnosti je rovno $\kappa(D) = 10^6$.



Obrázek 3.1: Singulární čísla matic C a D vykreslená v logaritmické škále.

3.2 Experiment 1

Nejdříve se zaměříme na testování citlivosti TLS řešení na přesnost singulárního rozkladu matice dat. Proto budeme uměle vnášet perturbaci do singulárních čísel matice dat a budeme pozorovat změnu normy rozdílu TLS řešení přesného a perturbovaného problému v závislosti na její velikosti. Budeme sledovat vliv perturbací o velikostech 10^{-17} až 10^5.

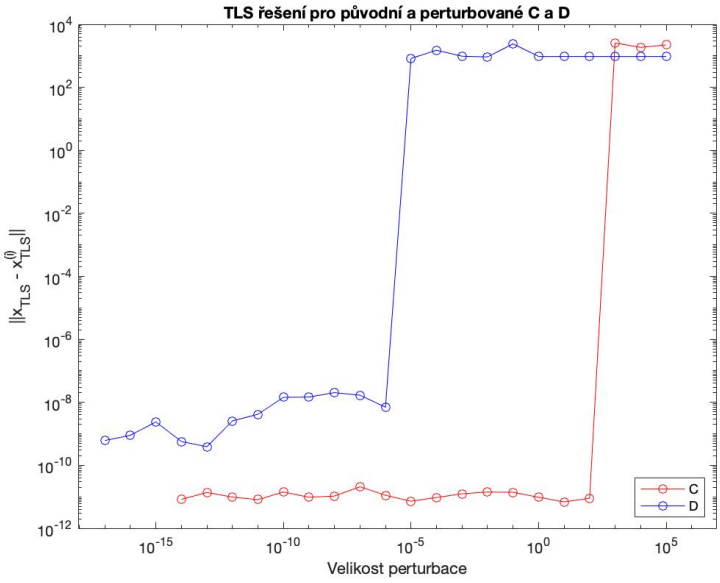
Nejdříve jsme vygenerovali diagonální matici perturbace E stejných rozměrů jako C s náhodnými kladnými čísly na diagonále (využitím příkazu 'rand'), tu jsme znormovali. Následně jsme perturbovali singulární čísla matice C a generovali tak perturbované matice C_i dle vzorce

$$C_i = U(\Sigma + 10^i \cdot E)V^\top, \qquad i = -17, \ldots, 5,$$

kde $U\Sigma V^\top$ značí singulární rozklad matice C.

Označme nyní x_{TLS} přesné TLS řešení problému s maticí dat $C = [b,A]$, tedy řešení $Ax \approx b$, spočtené vlastní funkcí 'tls', která SVD matice C spočetla pomocí funkce 'svd'. Označme dále $x_{TLS}^{(i)}$ TLS řešení problému s perturbovanou maticí dat $C_i = [b_i, A_i]$, tedy řešení $A_i x \approx b_i$, spočtené stejným způsobem. V experimentu sledujeme chybu TLS řešení měřenou jako $||x_{TLS} - x_{TLS}^{(i)}||$ v závislosti na perturbaci singulárních čísel matice C, která je rovna 10^i, pro $i = -17, \ldots, 5$.

Matici D perturbujeme analogicky a vykreslujeme stejné normy rozdílů TLS řešení.



Obrázek 3.2: Chyba TLS řešení $||x_{TLS} - x_{TLS}^{(i)}||$ pro matice C a D v závislosti na velikosti perturbace jejich singulárních čísel.

Z Obrázku 3.2 vidíme, že perturbace singulárních čísel matice C do řádu 10^{-15} nezpůsobily žádné změny v TLS řešení. Je tomu tak, neboť poslední diagonální prvek matice E, který perturboval nejmenší singulární číslo matice C označené $\tilde{\sigma}_{1000}$, je roven 0.01312. Tedy pro perturbace těchto řádů je výsledná chyba $\tilde{\sigma}_{1000}$ pod úrovní strojové přesnosti a neovlivnila výpočet. Zároveň lze nahlédnout, že singulární čísla této matice nejsou příliš citlivá na perturbace velikosti až 10^2, neboť ty se projevují chybou v TLS řešení $||x_{TLS} - x_{TLS}^{(i)}||$ řádu 10^{-11}. Při perturbaci velikosti 10^3 a vyšší pak pozorujeme výrazný skok v chybě TLS řešení, a to až na řád 10^3. Pro vysvětlení se zaměřme na pravý singulární prostor, ze kterého se pro každou perturbaci spočetlo TLS řešení. Z výstupů funkce 'tls' víme, že pro každou perturbaci bylo TLS řešení spočteno z jednonásobného nejmenšího singulárního čísla a příslušného pravého singulárního vektoru. Označíme-li tedy v_{1000}, resp. $v_{1000}^{(i)}$, pravý singulární vektor příslušný nejmenšímu singulárnímu číslu matice C, resp. C_i, pak chyba $||v_{1000} - v_{1000}^{(i)}||$ byla pro $i = -14, \ldots, 2$ řádově rovna 10^{-14}. Pro větší perturbace ale byla řádu jednotek. Tudíž skok v chybě TLS řešení, který pozorujeme pro perturbaci velikosti 10^3, je způsoben změnou v pravém singulárním prostoru, ve kterém se TLS řešení hledá. Neboli perturbace této velikosti způsobila změnu v pořadí singulárních čísel matice C a tím i v pravém singulárním prostoru.

Naopak můžeme vidět, že TLS řešení je citlivější na malé změny singulárních čísel matice dat D. Tyto perturbace se totiž projevují jako chyby v TLS řešení o dva až tři řády větší než u matice C. Zároveň pozorujeme skok na řádově stejnou chybu v řešení jako u předchozí matice, ale to už při perturbaci velikosti 10^{-5}. Tento skok v chybě řešení opět koresponduje se skokem v chybě pravého singulárního vektoru, ze kterého bylo spočteno TLS řešení. Tedy pro matici D způsobila změnu v pořadí singulárních čísel již perturbace velikosti 10^{-5}.

Z prvního experimentu tedy můžeme usoudit, že TLS řešení je citlivější na perturbace nejmenších singulárních čísel matice dat, která od sebe nejsou dobře oddělena a tvoří tzv. clustery. Jak jsme rozebrali výše, toto není překvapivé, neboť u takových singulárních čísel způsobí i malé perturbace změny v pořadí a tím i v singulárních prostorech, ve kterých se pak TLS řešení hledá.

3.3 Experiment 2

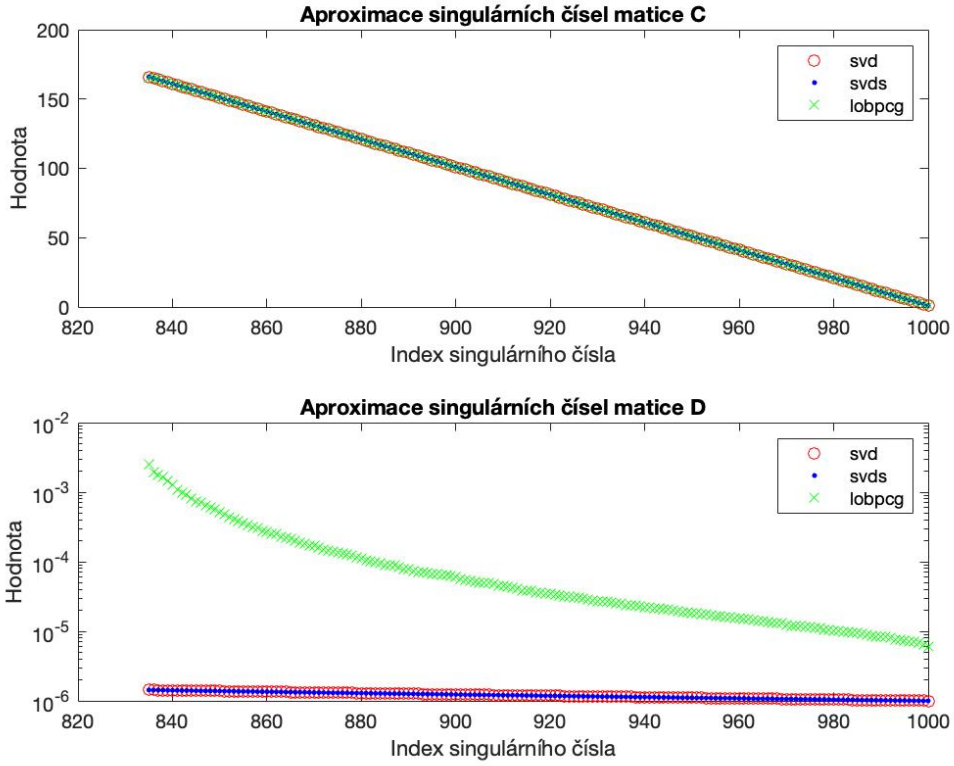
Ve druhém experimentu se budeme soustředit na funkce aproximující SVD (případně jeho část) diskutované v podkapitole 2.3, a to 'svd', 'svds' a 'lobpcg'. Budeme porovnávat kvalitu, s jakou tyto funkce aproximují nejmenší singulární čísla testovacích matic C a D.

Jak jsme diskutovali na konci podkapitoly 2.3, funkci 'lobpcg' aplikujeme na matici $C^\top C$. Jelikož počet nejmenších vlastních párů, které tato funkce aproximuje, nesmí přesáhnout pětinu rozměru matice, soustředíme se na aproximaci $\lfloor \frac{1000}{6} \rfloor = 166$ nejmenších singulárních čísel, kde $\lfloor\ \rfloor$ značí dolní celou část. Vstupními parametry funkce 'lobpcg' jsme volili náhodnou matici jako počáteční aproximaci vlastních vektorů, defaultní parametr tolerance a maximální počet iterací roven 1000. Tedy funkci jsme volali příkazem '$lobpcg(randn(1000{,}166), C^\top C, 1000 \cdot \sqrt{eps}, 1000)$', kde '$eps$' značí strojovou přesnost. Stejně tak postupujeme u aproximace singulárních čísel matice D.

Označme nyní $S = diag(\Sigma)$ vektor přesných 166 nejmenších singulárních čísel matice C, resp. D, S_{svd} vektor singulárních čísel spočtených funkcí 'svd' a $||S - S_{svd}||$ nazývejme chybou aproximace singulárních čísel funkcí 'svd'. Analogicky označme S_{svds} a S_{lobpcg}. V následujícím grafu pak vykreslujeme jednotlivé vektory aproximace singulárních čísel a v tabulce uvádíme chybu této aproximace.

	$\|\|S - S_{svd}\|\|$	$\|\|S - S_{svds}\|\|$	$\|\|S - S_{lobpcg}\|\|$
C	$1.1997 \cdot 10^{-12}$	$9.2771 \cdot 10^{-13}$	$3.1076 \cdot 10^{-12}$
D	$4.0299 \cdot 10^{-20}$	$5.0685 \cdot 10^{-20}$	$5.1678 \cdot 10^{-3}$

Tabulka 3.1: Chyba aproximace singulárních čísel matic C a D funkcemi 'svd', 'svds' a 'lobpcg'.



Obrázek 3.3: Singulární čísla matic C a D aproximovaná funkcemi 'svd', 'svds' a 'lobpcg'.

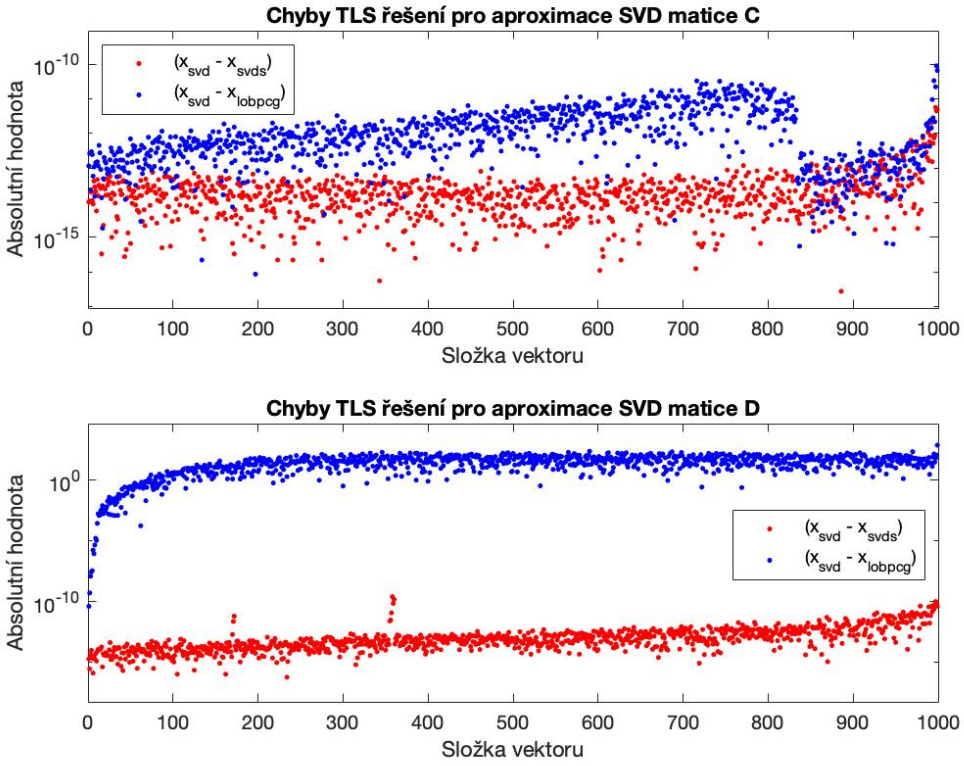
Jak lze vidět z Obrázku 3.3, všechny tři funkce poskytují srovnatelnou aproximaci nejmenších singulárních čísel matice C. To potvrzují i data z Tabulky 3.1, neboť vidíme, že chyby každé této aproximace jsou relativně malé a vzájemně srovnatelné, pohybující se na rozmezí řádu 10^{-13} a 10^{-12}.

U aproximací singulárních čísel matice D pozorujeme odlišné chování. Již z Obrázku 3.3 je zřejmé, že aproximace funkcí 'lobpcg' se výrazně liší od těch spočtených ostatními dvěma funkcemi. Z Tabulky 3.1 pak plyne, že funkce 'svd' a 'svds' aproximují singulární čísla matice D velmi dobře, neboť chyby se pohybují dokonce několik řádů pod úrovní strojové přesnosti. Naopak funkce 'lobpcg' poskytuje velmi špatnou aproximaci, respektive spočtená singulární čísla se i řádově liší od těch přesných, ačkoliv funkce zkonvergovala. To je dáno špatnou podmíněností matice D pro náš problém, kde nejmenší singulární čísla, která aproximujeme, jsou všechna téměř řádu 10^{-6}. Připomeňme, že číslo podmíněnosti této matice je rovno $\kappa(D) = 10^6$. Jak jsme diskutovali v sekci 2.2.1, pro takovou matici je nevhodné aproximovat malá singulární čísla matice D vlastními čísly matice $D^\top D$, jak to děláme při aplikaci funkce 'lobpcg'. Číslo podmíněnosti matice $D^\top D$ je potom rovno 10^{12} a to ovlivňuje spolehlivost výpočtu jejích nejmenších vlastních čísel.

3.4 Experiment 3

V posledním experimentu se zaměříme na vliv kvality aproximace nejmenších singulárních tripletů testovacích matic dat C a D pomocí funkcí 'svd', 'svds' a 'lobcg' na spočtené TLS řešení. K tomu budeme používat vlastní implementovanou funkci 'tls', která pro aproximaci SVD matice dat používá každou ze tří uvedených funkcí.

Označme x_{svd} TLS řešení problému s maticí dat C, resp. D, spočtené funkcí 'tls', která pro aproximaci singulárních tripletů matice C, resp. D, využila funkci 'svd'. Toto považujme za přesné TLS řešení. Stejně tak označme x_{svds} a x_{lobpcg} TLS řešení stejného problému spočtené s využitím funkcí 'svds' a 'lobpcg'. V následujícím grafu pak vykreslujeme chyby TLS řešení ve smyslu absolutních hodnot složek chybových vektorů $(x_{svd} - x_{svds})$ a $(x_{svd} - x_{lobpcg})$ pro obě testovací matice. V tabulce pak uvádíme celkovou normu této chyby.



Obrázek 3.4: Chyby TLS řešení měřené jako chybové vektory $(x_{svd} - x_{svds})$ a $(x_{svd} - x_{lobpcg})$ pro matice C a D vykreslené po složkách v absolutní hodnotě.

	$\|\|x_{svd} - x_{svds}\|\|$	$\|\|x_{svd} - x_{lobpcg}\|\|$
C	$9.2956 \cdot 10^{-12}$	$2.8884 \cdot 10^{-10}$
D	$3.6518 \cdot 10^{-10}$	$2.1639 \cdot 10^{3}$

Tabulka 3.2: Normy chyby TLS řešení $||x_{svd} - x_{svds}||$ a $||x_{svd} - x_{lobpcg}||$ pro matice C a D.

Z Obrázku 3.4 můžeme nahlédnout, že TLS řešení problému s maticí dat C se výrazně neliší pro všechny aproximace nejmenších singulárních tripletů. Vidíme, že jednotlivé složky chybových vektorů obou TLS řešení $(x_{svd} - x_{svds})$ a $(x_{svd} - x_{lobpcg})$ se v absolutní hodnotě pohybují od úrovně strojové přesnosti do řádu 10^{-10}, tedy nejsou nijak velké. Zároveň z Tabulky 3.2 jsou normy těchto chyb také relativně malé, řádů 10^{-12} a 10^{-10}.

Naopak u matice D pozorujeme rozdíl mezi TLS řešením za využití funkce 'svds' a tím za využití 'lobpcg'. Absolutní hodnoty složek $(x_{svd} - x_{svds})$ se pohybují ve stejném rozmezí jako u matice dat C, stejně tak norma chyby je řádu 10^{-10}. Za to u řešení x_{lobpcg} se jednotlivé složky chybového vektoru $(x_{svd} - x_{lobpcg})$ v absolutní hodnotě pohybují až v řádu jednotek a celková norma chyby je velikosti 10^3.

Nic z výše pozorovaného není překvapující, neboť z minulého experimentu víme, jak vypadají nejmenší singulární čísla matic C a D aproximovaná jednotlivými funkcemi. Když víme, že funkce 'lobpcg' aproximuje singulární čísla matice D velmi špatně, nemůžeme očekávat ani přesnost TLS řešení x_{lobpcg}, jehož konstrukce je na této aproximaci založena.

Závěr

V této práci jsme se soustředili na problém úplných nejmenších čtverců, což je jeden z možných nástrojů pro řešení lineárního aproximačního problému, kdy pozorování i model jsou zatíženy chybami. Tento jsme podrobně analyzovali, zaměřili jsme se na konstrukci jeho řešení a popsali jsme s tím spojený výpočet singulárního rozkladu matice a možné způsoby aproximace jeho části.

V numerických experimentech jsme se zaměřili na citlivost konstruovaného TLS řešení na perturbace v datech a porovnali jsme tři metody aproximace nejmenších singulárních tripletů. Pro ilustraci výsledků jsme zvolili dvě matice dat, první dobře podmíněnou s rovnoměrně rozloženými singulárními čísly, druhou hůře podmíněnou s rychle klesajícími a k sobě se přibližujícími singulárními čísly.

V prvním experimentu jsme uměle perturbovali singulární čísla matice dat a sledovali vliv velikosti této perturbace na spočtené TLS řešení. Dle očekávání jsme pozorovali, že TLS řešení je citlivější na perturbace u matic, jejichž nejmenší singulární čísla jsou si navzájem blízká.

Ve druhém experimentu jsme porovnávali tři metody aproximace nejmenších singulárních tripletů - pomocí funkcí 'svd', 'svds' a 'lobpcg'. U funkcí 'svd' a 'svds' jsme zpravidla nepozorovali velký rozdíl a tyto funkce poskytovali velmi dobrou aproximaci. Nicméně funkce 'lobpcg' se ukázala být nevhodná pro aproximaci nejmenších singulárních čísel špatně podmíněných matic. To je zřejmě dáno naším přístupem, který je založen na konstrukci ještě hůře podmíněné matice $D^\top D$.

Nakonec jsme v rámci spojení předchozích experimentů testovali vliv kvality aproximace nejmenších singulárních tripletů na spočtené TLS řešení. Dle očekávání se TLS řešení lišilo pro problémy s maticí dat, u které se lišily jednotlivé aproximace. Tj. zejména u špatně podmíněných matic bylo TLS řešení spočtené využitím funkce 'lobpcg' velmi nepřesné.

Dále by bylo možné testovat dané fenomény pro problémy větších dimenzí a porovnat také časovou a výpočetní náročnost TLS algoritmu při počítání celého SVD matice dat versus pouze při aproximaci několika nejmenších singulárních tripletů. Zároveň by bylo vhodné více se zaměřit na možnost předpodmínění a přidání dalších volitelných parametrů při aplikaci funkce 'lobpcg'.

Seznam použité literatury

Baglama, J. a Reichel, L. (2005). Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, **27**(1), 19–42.

Barlow, J. (2001). More accurate bidiagonal reduction for computing the singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, **23**, 761–798.

Dax, A. (2019). Computing the smallest singular triplets of a large matrix. *Results in Applied Mathematics*, **3**, 100006.

Demmel, J. a Kahan, W. (1990). Accurate singular values of bidiagonal matrices. *SIAM Journal on Scientific and Statistical Computing*, **11**, 873–912.

Duersch, J. A., Shao, M., Yang, C. a Gu, M. (2018). A robust and efficient implementation of lobpcg. *SIAM Journal on Scientific Computing*, **40**(5), C655–C676.

Duintjer Tebbens, J., Hnětynková, I., Plešinger, M., Strakoš, Z. a Tichý, P. (2012). *Analýza metod pro maticové výpočty: základní metody*. Matfyzpress.

Dvořák, J. (2021). Vlastnosti a konstrukce core problému v úlohách fitování dat s násobným pozorováním. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, Katedra numerické matematiky.

Golub, G. H. a Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *SIAM Journal on Numerical Analysis*, **2**(2), 205–224.

Golub, G. H. a Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, **17**(6), 883–893.

Golub, G. H. a Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, third edition.

Hetmaniuk, U. a Lehoucq, R. (2006). Basic selection in lobpcg. *Journal of Computational Physics*, **218**(1), 324–332.

Hnětynková, I., Plešinger, M., Sima, D. M., Strakoš, Z. a Van Huffel, S. (2011). The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works. *SIAM Journal on Matrix Analysis and Applications*, **32**(3), 748–770.

Knyazev, A. V. (2001). Toward the optimal preconditioned eigensolver: Locally optimal block preconjugate conjugate gradient method. *SIAM Journal on Scientific Computing*, **23**(2), 517–541.

Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, **45**(4).

Larsen, R. M. (1998). Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, **27**.

Paige, C. C. a Saunders, M. A. (1975). Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, **12**(4), 617–629.

Paige, C. C. a Strakoš, Z. (2005). Core problems in linear algebraic systems. *SIAM Journal on Matrix Analysis and Applications*, **27**(3), 861–875.

Van Huffel, S. a Vandewalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. Society for Industrial and Applied Mathematics.

Watkins, D. S. (2002). *Fundamentals of Matrix Computations*. John Wiley & Sons, Ltd, second edition.

Watkins, D. S. (2010). *Fundamentals of Matrix Computations*. John Wiley & Sons, Ltd, third edition.