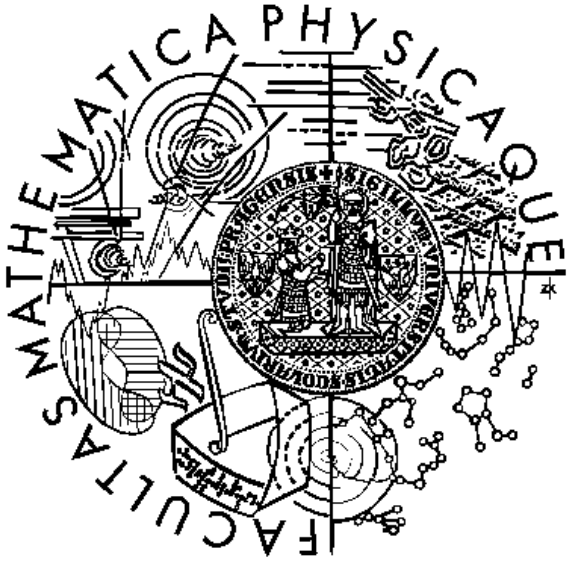
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Ján Eliaš

Problémy spojené s výpočtem největšího společného dělitele

Katedra numerické matematiky

Vedoucí bakalářské práce: doc. RNDr. Jan Zítko, CSc.
Studijní program: Matematika

2009

Poďakovanie

Veľmi rád by som poďakoval všetkým, ktorí prispeli k vypracovaniu tejto bakalárskej práce. Predovšetkým sa chcem poďakovať doc. RNDr. Janovi Zítkovi, CSc. za vedenie bakalárskej práce a Dr. Joabovi Winklerovi, PhD.

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne s použitím citovanej literatúry a uvedených zdrojov. Súhlasím s požičiavaním práce a jej prípadným použitím pre pedagogické, vedecké a prezentačné účely.

V Prahe dňa 02.08.2009 Ján Eliaš

.....

Obsah

Názov práce: Problémy spojené s výpočtom najväčšieho spoločného deliteľa
Autor: Ján Eliaš
Katedra: Katedra numerickej matematiky
Vedúci bakalárskej práce: doc. RNDr. Jan Zítko, CSc.
e-mail vedúceho bakalárskej práce: zitko@karlin.mff.cuni.cz

Anotácia

V mnohých praktických aplikáciach zohráva úlohu výpočet najväčšieho spoločného deliteľa dvoch polynómov (GCD). Ak dva polynómy majú nekonštantný GCD, tak od nich odvodené nepresné polynómy $f(x)$, $g(x)$ sú s pravdepodobnosťou jeden nesúdeliteľné. Avšak každá malá perturbácia koeficientov týchto polynómov môže mať za následok to, že GCD polynómov $f(x) + \delta f(x)$, $g(x) + \delta g(x)$ je opäť netriviálny. Takýto GCD sa nazýva aproximovaný najväčší spoločný deliteľ dvoch nepresných polynómov (AGCD). Existuje niekoľko metód zaoberajúcich sa výpočtom AGCD. V tejto práci je použitá metóda structured total least norm (STNL) aplikovaná na Sylvestrovu rezultantovú maticu.

Kľúčové slová: GCD, STLN, Sylvestrova matica

Title: Problems connected with the calculation of the GCD
Author: Ján Eliaš
Department: Department of Numerical Mathematics
Supervisor: doc. RNDr. Jan Zítko, CSc.
Supervisor's e-mail address: zitko@karlin.mff.cuni.cz

Annotation

The determination of the greatest common divisor (GCD) is available in many aplications. If the polynomials $f(x)$, $g(x)$ have a non-constant GCD, their inexact forms $f(x)+\delta f(x)$, $g(x) + \delta g(x)$ will be coprime. Therefore, it is useful to determine the smallest perturbation, such that the inexact polynomials have a non-constant GCD, so-called an approximate greatest common divisor of two given inexact polynomials. There are some methods dealing with this problems, the method of structured total least norm (STLN) for a Sylvester matrix is used in this work.

Keywords: GCD, STLN, Sylvester matrix

Kapitola 1

Úvod

Výpočet GCD patrí medzi základné problémy výpočtovej matematiky a má význam ako teoretický tak i praktický v teórií riadenia, spracovania signálu a robotiky, teórií sietí, počítačovom dizajne, spracovaní obrazu, šifrovaní a kódovaní informácií a i. V mnohých aplikáciach sa ale pracuje s približnými datami danými s určitou toleranciou (napr. nepresné data získané fyzikálnym meraním, či vplyvom kumulovania zaokrúhľovacích chýb). To ale môže vyústiť do nepríjemných numerických ťažkostí pri výpočte GCD. Navyše výpočet GCD dvoch polynómov je dobrým príkladom tzv. "ill-posed problems". Buď napríklad $f(x) = x^2 + 4x + 4$ a $g(x) = x + 2$. Pri symbolickom výpočte GCD nenastávajú žiadne problémy, určite je $\mathrm{GCD}(f(x), g(x)) = g(x) = x+2$, avšak pre $f(x) = x^2 + 3.999x + 4$ už platí $\mathrm{GCD}(f(x), g(x)) = 1$, pričom malá zmena koeficientov (pridaním 0.001 k druhému koeficientu polynómu $f(x)$) stačí k tomu, aby polynómy boli opäť súdeliteľné. A teda naopak, mierne porušenie koeficientov môže znížiť stupeň GCD až tak, že GCD bude triviálny.

Problém, ktorým sa budeme zaoberať, je nájsť polynómy $\tilde{f}(x)$ a $\tilde{g}(x)$ "blízke" k nepresne zadaným (porušeným) polynómom $f(x)$, resp. $g(x)$ tak, aby mali netriviálny GCD. Pretože GCD "opravených" polynómov sa vo väčšine prípadov nezhoduje s GCD teoreticky presných polynómov, dokonca nemusí mať ani rovnaký stupeň, budeme takýto GCD nazývať aproximovaný GCD (AGCD). V mnohých matematických prácach sú popísané rôzne metódy na nájdenie polynómov $\tilde{f}(x)$ a $\tilde{g}(x)$, resp. AGCD nepresných polynómov. Táto práca vychádza z článkov [9, 10]. Avšak narozdiel od [9, 10], nezameriame sa len na určenie $\tilde{f}(x)$, $\tilde{g}(x)$ a stupňa AGCD metódou STLN, ale na základe [13] ukážeme i možné spôsoby výpočtu samotného AGCD.

Cieľom práce je zhrnúť jeden možný spôsob prístupu k problému výpočtu AGCD. Podáme komplexný súhrn nástrojov, ktorými budeme schopní riešiť daný problém. To znamená, že skôr než pristúpime k metóde STLN, zosumarizujeme niektoré spôsoby riešenia problému najmenších štvorcov s obmedzujúcou podmienkov (tzv. LSE problému). Ukážeme a porovnáme niekoľko metód. Ďalej si zavedieme pojem Sylvestrovej matice a jeho rezultantu, ktorý nám poskytne pevnú datovú štruktúru pre výpočet GCD, pričom vlastný výpočet GCD spočíva v jej vhodných tranformáciách. Podáme dva možné spôsoby, ktoré si numericky otestujeme. Nakoniec pristúpime k metóde STLN, vysvetlíme si princíp metódy a udvedieme niekoľko príkladov.

Všetky programy týkajúce sa práce sú naprogramované v prostredí Matlab R2008a.

Z dôvodu veľkého rozsahu problematiky v práci nie sú, až na výnimky v časti 2.3.3, explicitne uvedené schémy algoritmov, podľa ktorých sme realizovali programy. Tie možno nájsť v dokumentácií k programom na priloženom CD, ktoré je súčasťou bakalárskej práce. CD ďalej obsahuje programy a krátky textový dokument, ktorý uľahčí prácu s CD.

1.1 Použité skratky a značky

Ak nebude explicitne povedané inak, budeme v ďalšom texte používať nasledujúce skratky a označenia:

- $\mathrm{rank}(A)$ ... hodnosť matice A
- A^{-1} ... inverzná matica k matici A
- $\mathrm{diag}(a_1, a_2, \ldots, a_n)$ ... štvorcová diagonálna matica s číslami $a_1, \ldots, a_n$ na hlavnej diagonále
- I_n ... diagonálna $n \times n$ matica s jednotkami na diagonále
- GCD ... najväčší spoločný deliteľ dvoch polynómov
- AGCD ... aproximovaný GCD dvoch polynómov
- $\mathcal{N}(A)$ ... jadro matice A, $\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$
- $\mathrm{Range}(A)$ ... obor hodnôt matice A, $\mathrm{Range}(A) = \{z = Ax : x \in \mathbb{R}^n\}$
- $A = \underset{m}{[A_1,} \ \underset{n}{A_2]}$ ... značí rozdelenie stĺpcov matice A na A_1 (ktorých je m) a A_2 (ktorých je n)
- $B = \begin{matrix}[B_1 \\ B_2]\end{matrix} \begin{matrix} m \\ n \end{matrix}$ ... značí rozdelenie riadkov matice B na B_1 (ktorých je m) a B_2 (ktorých je n)
- $A = [a_1, \ldots, a_n]$ ... označenie stĺpcov matice A, tj. $a_1, \ldots, a_n$ sú stĺpcové vektory matice A
- $z = \pm m\, e^{\pm t}$ ... zápis čísla v pohyblivej radovej čiarke, m je mantisa, e základ (u nás $e = 10$), t exponent, napríklad $eps = 2,22e^{-16}$, zvykneme tiež písať $eps = 2,22e - 16 = 2,22 \times 10^{-16}$)
- eps ... strojová presnosť počítača definovaná ako vzdialenosť čísla 1 od najbližšieho vyšieho čísla v pohyblivej radovej čiarke, ($eps = 2,22e - 16$)

Pre vektor $x \in \mathbb{R}^n$ budeme používať klasické definície noriem:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}, \qquad \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|.$$

Kapitola 2

Riešenie LSE problému

V tejto kapitole sa budeme venovať riešeniu LSE problému. Celkom sa budeme zaoberať niekoľkými postupmi, pričom podľa [3] *spomenieme dva možné prístupy. Bez dôrazu na presnosť formulácie uveďme, že prvý prístup využíva projektovanie riešenia na nulový priestor istej matice a jeho ortogonálny doplnok. V značnej miere sa využíva technika QR rozkladu matíc. Rozsiahly rozbor algoritmov tohto prístupu možno nájsť napríklad v* [3, 8]. *Druhý možný spôsob je metóda založená na priamej eliminácií. Ako príklad priamej eliminácie použijeme algoritmus vyskytujúci sa v práci* [3]*, iné algoritmy sú napríklad v* [1, 4].
Iný postoj k riešeniu LSE problému ponúka metóda váh, ktorej sa budeme venovať v prevažnej časti tejto kapitoly. Túto metódu neskôr použijeme v 4. kapitole.
Pretože časť rozoberaných algoritmov je inšpirovaná jedným spôsobom riešenia problému najmenších štvorcov (LS problému), tak obsahom prvej sekcie je pár poznámok týkajúcich sa práve problému najmenších štvorcov.

2.1 Niekoľko poznámok k LS problému

Existuje niekoľko odlišných spôsobov riešenia problému najmenších štvorcov (značíme LS problém z anglického výrazu "least squares problems"), ktoré možno nájsť napr. v [3, 5]. Jeden možný postup ako nájsť taký vektor $x \in \mathbb{R}^n$, ktorý bude minimalizovať

$$\|Ax - b\|_2, \tag{2.1}$$

je použiť metódu výberu založenú na QR rozklade. O obdĺžnikovej matici $A \in \mathbb{R}^{m\times n}$ budeme predpokladať, že má plnú hodnosť ($m \geq n$, $\mathrm{rank}(A) = n$), teda že jej stĺpce sú lineárne nezávislé [1], $b \in \mathbb{R}^m$. Predpokladajme ďalej, že máme spočítanú ortogonálnu maticu $Q \in \mathbb{R}^{m\times m}$. Vynásobiť data ortogonálnou maticou nezmení geometrický význam úlohy. V tomto prípade hovoríme o ortogonálne invariantnom probléme. Preto platí, že ak vektor x minimalizuje (2.1), tak minimalizuje aj

$$\|Q^T Ax - Q^T b\|_2. \tag{2.2}$$

[1] V prípade, že je $\mathrm{rank}(A) < n$, riešenie nemusí existovať a ak existuje, všeobecne nie je určené jednoznačne. Preto je potrebné nasledujúci postup modifikovať, pozri napríklad [5]. Avšak v našich úvahách sa tento prípad nevyskytuje.

Za danú ortogonálnu maticu sa prirodzene snažíme voliť takú maticu, aby sa problém (2.2) čo najviac zjednodušil. To sa s úspechom podarí pri voľbe ortogonálnej matice z QR rozkladu matice A. Buď teda

$$A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}$$

QR rozklad matice A, kde $R_1 \in \mathbb{R}^{n\times n}$ je horná trojuholníková a $Q \in \mathbb{R}^{m\times m}$ ortogonálna matica. Položme $Q^T b = \begin{bmatrix} r \\ s \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}$. Potom je

$$\|Ax - b\|_2^2 = \|Q^T Ax - Q^T b\|_2^2 = \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} r \\ s \end{bmatrix} \right\|_2^2 = \|R_1 x - r\|_2^2 + \|s\|_2^2.$$

Pretože sme predpokladali, že $\text{rank}(A) = n$, je i $\text{rank}(R_1) = n$ a teda R_1 je regulárna s nenulovými prvkami na diagonále. V dôsledku toho nutne existuje práve jedno riešenie x sústavy rovníc $R_1 x = r$ [2].

Našim cieľom je ale vyriešiť špeciálnejšiu úlohu, a totiž od riešenia problému (2.1) budeme požadovať, aby naviac spĺňalo podmienku $Bx = d$ pre všeobecne obdĺžnikovú maticu B a vektor d. To je obsahom dalšej sekcie.

2.2 LSE problém

Úlohou je nájsť vektor $x \in \mathbb{R}^n$, ktorý rieši problém

$$\min_{Bx=d} \|Ax - b\|_2. \tag{2.3}$$

Úlohu nazývame problémom najmenších štvorcov s obmedzením $Bx = d$, v angličtine sa užíva termín "Linear Least Squares problem with Equality Constraints", resp. "Constrained Linear Least Squares problem", skrátene LSE problém.

Sformulujme najprv všetky predpoklady. Nech $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{p\times n}$ $(n \geq p)$ a $d \in \mathbb{R}^p$. Predpokladajme, že $\text{rank}(B) = p$ a že prienik nulových priestorov matíc A a B je triviálny. Navyše pre druhú z podmienok platí ekvivalencia

$$N(A) \cap N(B) = \{0\} \iff \text{rank} \begin{bmatrix} B \\ A \end{bmatrix} = n.$$

Pre dôkaz implikácie sprava doľava stačí predpokladať, že $\text{rank} \begin{bmatrix} B \\ A \end{bmatrix} < n$. Potom existuje $x \neq 0$ tak, že $\begin{bmatrix} B \\ A \end{bmatrix} x = 0$, z čoho iste plynie platnosť $Bx = 0 \wedge Ax = 0$ a teda dostávame spor. Obrátenou úvahou dostaneme platnosť opačnej implikácie. □

[2] Všimnime si, že pre takto spočítané riešenie x sa veľkosť rezidua $\|Ax - b\|_2$ rovná $\|s\|_2$.

Podľa [3] predpoklady

$$\operatorname{rank}(B) = p \qquad \text{a} \qquad \operatorname{rank}\begin{bmatrix} B \\ A \end{bmatrix} = n \tag{2.4}$$

zaručujú existenciu a jednoznačnosť riešenia (2.3), ktoré si označíme x_{LSE}. Dokonca predstavujú nutnú a postačujúcu podmienku preto, aby riešenie x_{LSE} bolo jednoznačné.

Keby totiž platilo $N(A) \cap N(B) \neq \{0\}$, tak iste existuje $0 \neq z \in N(A) \cap N(B)$ také, že $Az = Bz = 0$. Ale ak x rieši (2.3), tak aj $x + z$ rieši (2.3), čím dostávame dve rôzne riešenia. Postačiteľnosť predpokladov dokážeme konštrukciou algoritmu v sekcií 2.2.1, kde predvedieme istý spôsob ako získať riešenie x problému (2.3) a ukážeme, že platnosť predpokladov stačí pre jednoznačnosť tohto x. □

Prirodzene, požiadavky vyššie zaručia jednoznačnosť riešenia aj pri ostatných preberaných metódach.

V nasledujúcich sekciách predvedieme príklady možných riešení (2.3). V sekcií 2.2.1 spôsob založený na projektovaní riešenia na jadro matice B a jeho ortogonálny doplnok. V sekcií 2.2.2 ukážeme postup priamej eliminácie a na záver bližšie rozoberieme metódu váh a iteračné spresnenie tejto metódy. Všetky spomenuté metódy boli implementované v prostredí Matlab, ich porovnania sú v sekcií 2.3.2, resp. 2.3.4, kde porovnávame výsledky získané metódou váh s jej iteračným spresnením.

2.2.1 Metóda projekcie na jadro matice

Ako už bolo spomenuté v úvode, existujú dva možné spôsoby riešenia (2.3). Jedným z nich sú metódy založené na projekcií riešenia na jadro matice B, ktoré využívajú QR rozklad matice B^T za účelom získania bázy nulového priestoru matice B. Samozrejme, existuje niekoľko rôznych algoritmov, ktoré sa líšia v závislosti na ďalšom postupe. Uvedieme algoritmus, ktorý možno nájsť v [3], resp. v [8].

Namiesto LSE problému (2.3) uvažme LS problém

$$\min_{x \in \mathbb{R}^n} \left\| \begin{bmatrix} B \\ A \end{bmatrix} x - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2, \tag{2.5}$$

pričom máme na pamäti, že x má spĺňať rovnosť $Bx = d$ [3]. Teda, budeme hľadať riešenie LS problému s dôrazom na podmienku $Bx = d$, pričom prejdeme od (2.5) k ortogonálne invariantnému problému s vhodnou ortogonálnou maticou.

Zaveďme $U \in \mathbb{R}^{m \times m}$ a $Q \in \mathbb{R}^{n \times n}$ ortogonálne matice, ktoré určíme neskôr a položme

$$\tilde{U} = \begin{bmatrix} I_p, & 0 \\ 0, & U^T \end{bmatrix} \in \mathbb{R}^{(p+m)\times(p+m)}.$$

Ľahko overíme, že $\tilde{U}$ je štvorcová ortogonálna matica. Stĺpce matice Q rozdeľme na dve časti, totiž označme prvých p stĺpcov matice Q ako Q_1 a zvyšných $n - p$ stĺpcov ako

[3] Na tomto mieste uveďme, že až na výnimky v časti 2.3.1 v práci budeme pracovať so systémom $\begin{bmatrix} B \\ A \end{bmatrix}$. Je to z dôvodu zjednotenia textu a z dôvodu, že systém "B nad A" dáva pri riešení LSE problému metódou váh lepšie numerické výsledky, pozri [8].

Q_2, tj. $Q = \underset{p}{\left[Q_1,\right.} \quad \underset{n-p}{\left.Q_2\right]}$. Konečne, zaveďme substitúciu $y = Q^T x$. Potom je $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Q_1^T x & p \\ Q_2^T x & n-p \end{bmatrix}$ a naopak

$$x = Q\,y = \begin{bmatrix} Q_1, & Q_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Q_1 y_1 + Q_2 y_2 =: x_1 + x_2.$$

Vďaka práve zavedeným označeniam môžeme písať

$$\left\| \begin{bmatrix} B \\ A \end{bmatrix} x - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} B \\ A \end{bmatrix} I x - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} B \\ A \end{bmatrix} Q\,Q^T x - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} B \\ A \end{bmatrix} Q\,y - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2^2 \overset{*}{=}$$

$$\left\| \tilde{U} \begin{bmatrix} B \\ A \end{bmatrix} Q\,y - \tilde{U} \begin{bmatrix} d \\ b \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} B\,Q_1, & B\,Q_2 \\ U^T A\,Q_1, & U^T A\,Q_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} d \\ U^T b \end{bmatrix} \right\|_2^2 =$$

$$\left\| B\,Q_1 y_1 + B\,Q_2 y_2 - d \right\|_2^2 + \left\| U^T A\,Q_1 y_1 + U^T A\,Q_2 y_2 - U^T b \right\|_2^2,$$

kde v rovnosti označenej $*$ sme prešli k ortogonálne invariantnému problému s maticou $\tilde{U}$. Aby sme výrazy v normách čo najviac zjednodušili, použijeme QR rozklady matíc B^T a AQ_2:

- Pretože je $B\,Q_1 y_1 + B\,Q_2 y_2 = B\,Q\,y$, tak za Q voľme ortogonálnu maticu z QR rozkladu B^T. Platí $B\,Q = \begin{bmatrix} R_B^T, & 0 \end{bmatrix}$, kde $Q \in \mathbb{R}^{n \times n}$ a $R_B \in \mathbb{R}^{p \times p}$ je horná trojuholníková matica, ktorá je naviac regulárna, pretože $\mathrm{rank}(B) = p$. Všimnime si, že kvôli predpokladu na hodnosť matice B patrí posledných $n-p$ stĺpcových vektorov Q, ktoré sú lineárne nezávislé a navzájom na seba kolmé, do jadra $\mathcal{N}(B)$. Potom ale je $\mathcal{N}(B) = \mathrm{Range}(Q_2)$, čo znamená, že stĺpcové vektory Q_2 tvoria bázu $\mathcal{N}(B)$.

  Prvý sčítanec môžeme teraz upraviť na tvar

  $$\left\| B\,Q_1 y_1 + B\,Q_2 y_2 - d \right\|_2^2 = \left\| B\,Q\,y - d \right\|_2^2 = \left\| \begin{bmatrix} R_B^T, & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - d \right\|_2^2 =$$

  $$= \left\| R_B^T y_1 - d \right\|_2^2.$$

  Avšak pretože R_B je regulárna, existuje práve jedno riešenie $y_1 \in \mathbb{R}^p$ sústavy $R_B^T y_1 = d$. A teda aj práve jedno $x_1 = Q_1 y_1$. Zostáva spočítať y_2, pričom podľa uvedeného je $x_2 = Q_2 y_2 \in \mathcal{N}(B)$. Tým riešenie x skutočne projektujeme na jadro $\mathcal{N}(B)$ a jeho ortogonálny doplnok.

- Pretože $(p+m) \times n$ matica

  $$\begin{bmatrix} B \\ A \end{bmatrix} \begin{bmatrix} Q_1, & Q_2 \end{bmatrix} = \begin{bmatrix} R_B^T, & 0 \\ AQ_1, & AQ_2 \end{bmatrix} \begin{matrix} p \\ m \end{matrix}$$
  $$\qquad\qquad\quad\; p \qquad n-p$$

má podľa predpokladov (2.4) hodnosť n, tak použitím vzťahu $\operatorname{rank}(R_B^T) = p$ z predchádzajúceho bodu musí byť $\operatorname{rank}(AQ_2) = n - p$. Voľbou ortogonálnej matice U z QR rozkladu $A\,Q_2 \in \mathbb{R}^{m\times(n-p)}$ môžeme pokračovať v úprave druhého sčítanca takto [4]:

$$\left\|U^T A\,Q_1 y_1 + U^T A\,Q_2 y_2 - U^T b\right\|_2^2 = \left\|U^T A\,Q_2 y_2 - U^T(b - A\,x_1)\right\|_2^2 =$$

$$\left\|\begin{bmatrix} R_A \\ 0 \end{bmatrix} y_2 - \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}(b - A\,x_1)\right\|_2^2 = \left\|R_A y_2 - U_1^T(b - A\,x_1)\right\|_2^2 + \left\|U_2^T(b - A\,x_1)\right\|_2^2,$$

kde $U = \underset{n-p}{[U_1,} \ \underset{m-n+p}{U_2}]$. Pretože $U_1^T(b - A\,x_1) \in \mathbb{R}^{n-p}$ a R_A je regulárna matica, vieme spočítať jednoznačne určené riešenie systému $(n-p)\times(n-p)$ rovníc $R_A y_2 = U_1^T(b - A\,x_1)$, čím dostávame vektor y_2 a $x_2 = Q_2 y_2$.

- Riešením LSE problému je vektor

$$x_{LSE} = Qy = \begin{bmatrix} Q_1, & Q_2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Q_1 y_1 + Q_2 y_2 = x_1 + x_2.$$

Uvedomme si, že x_{LSE} spĺňa podmienku $Bx = d$. Totiž platí

$$Bx = BQQ^T x = BQ\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} R_B^T & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = R_B^T y_1 = d.$$

Týmto postupom sme ukázali postačiteľnosť podmienok (2.4), existenciu a jednoznačnosť riešenia x_{LSE}, ktoré minimalizuje (2.5): Vektory y_1 a y_2 dávajú minimálne reziduá príslušných systémov rovníc. V prvom bode z predpokladu $\operatorname{rank}(B) = p$ vyplynula jednoznačnosť x_1, v druhom bode z oboch podmienok v (2.4) zase jednoznačnosť x_2.

2.2.2 Metóda priamej eliminácie

Druhý možný prístup k riešeniu (2.3) ponúkajú eliminačné metódy. Rovnako ako v predchádzajúcej časti existuje niekoľko metód, na ktoré možno nazerať ako na Gaussovu elimináciu. Ukážeme si spôsob popísaný v [1, 3]. Naviac v [1] možno nájsť algoritmus, ktorý používa elimináciu aplikovanú na penalizovaný problém (2.9), s ktorým sa stretneme v časti 2.3. Iný algoritmus je v [4].

Nasledujúca metóda priamej eliminácie využíva na riešenie problému (2.3) QR rozklad matice $B\Pi$, kde Π je $n \times n$ permutačná matica dôležitá z hľadiska stability celého procesu. To znamená, že pri rozklade matice B je potrebná stĺpcová pivotácia. Zostrojením matice Π sa zaoberáme na konci tejto časti.

[4] QR rozklad matice $A\,Q_2 = U\begin{bmatrix} R_A \\ 0 \end{bmatrix}$, kde $U \in \mathbb{R}^{m\times m}$ a $R_A \in \mathbb{R}^{(n-p)\times(n-p)}$ je regulárna horná trojuholníková matica. Regularita R_A plynie zo vzťahu $\operatorname{rank}(AQ_2) = n - p$.

Pripomeňme, že predpokladáme plnú hodnosť matice B, tj. $\text{rank}(B) = p$, $B \in \mathbb{R}^{p\times n}$, $p \le n$. Podľa teórie QR rozkladom $B\Pi$

$$\begin{array}{rcccc} B\Pi & = & Q & \begin{bmatrix} R_1, & R_2 \end{bmatrix} & \begin{matrix} p \\ \end{matrix} \\ & & p & \begin{matrix} p & n-p \end{matrix} & \end{array}$$

získame ortogonálnu maticu $Q \in \mathbb{R}^{p\times p}$, regulárnu hornú trojuholníkovú maticu $R_1 \in \mathbb{R}^{p\times p}$ a maticu $R_2 \in \mathbb{R}^{p\times(n-p)}$. Položme $\Pi^T x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{matrix} p \\ n-p \end{matrix}$. Teraz si môžeme podmienku $Bx = d$ upraviť, totiž platí:

$$Bx = d \Leftrightarrow B\,\Pi\,\Pi^T x = d \Leftrightarrow Q\begin{bmatrix} R_1, & R_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = d \Leftrightarrow R_1 x_1 = Q^T d - R_2 x_2. \tag{2.6}$$

Pretože je R_1 regulárna, tak z poslednej rovnosti vieme spočítať x_1,

$$x_1 = R_1^{-1}\left(Q^T d - R_2 x_2\right).$$

Aby sme určili x_2, vynásobme aj maticu A permutačnou maticou Π. Nech teda je

$$\begin{array}{rcccc} A\Pi & = & \begin{bmatrix} \tilde{A}_1, & \tilde{A}_2 \end{bmatrix} & \begin{matrix} m \\ \end{matrix} \\ & & \begin{matrix} p & n-p \end{matrix} & \end{array}$$

Potom ale

$$\|Ax - b\|_2 = \|A\,\Pi\,\Pi^T x - b\|_2 = \|(\tilde{A}_2 - \tilde{A}_1 R_1^{-1} R_2)x_2 - (b - \tilde{A}_1 R_1^{-1} Q^T d)\|_2 = \|\bar{A}x_2 - \bar{b}\|_2,$$

kde sme použili $\Pi^T x = [x_1,\, x_2]^T$, spočítané x_1 a kde sme zaviedli označenie

$$\begin{aligned} \bar{A} &= \tilde{A}_2 - \tilde{A}_1 R_1^{-1} R_2 \in \mathbb{R}^{m\times(n-p)}, \\ \bar{b} &= b - \tilde{A}_1 R_1^{-1} Q^T d \in \mathbb{R}^m. \end{aligned} \tag{2.7}$$

Ukážme, že z predpokladov (2.4) plynie $\text{rank}(\bar{A}) = n-p$. Predpokladajme pre spor, že $\text{rank}(\bar{A}) < n-p$. Potom existuje $v \neq 0$ tak, že $\bar{A}v = 0$. Avšak z (2.7) dostávame vzťah

$$\tilde{A}_2 v - \tilde{A}_1 R_1^{-1} R_2 v = 0,$$

v ktorom ak položíme $u := -R_1^{-1} R_2 v$, získame rovnice

$$R_1 u + R_2 v = 0, \qquad \tilde{A}_1 u + \tilde{A}_2 v = 0.$$

Ale obe rovnice dávajú pre vektor $w = \Pi\begin{bmatrix} u \\ v \end{bmatrix}$ rovnosti $Aw = 0$ a $Bw = 0$, čo znamená, že $w \neq 0$ leží v $\mathcal{N}(A) \cap \mathcal{N}(B)$ a to je spor. □

Celkom sme problém (2.3) previedli na riešenie LS problému bez obmedzenia

$$\min_{x_2 \in \mathbb{R}^{n-p}} \|\bar{A}x_2 - \bar{b}\|_2$$

s maticou $\bar{A} \in \mathbb{R}^{m\times(n-p)}$, ktorá má podľa predchádzajúcej úvahy plnú hodnosť. Tento problém vieme podľa sekcie 2.1 vyriešiť použitím QR rozkladu matice $\bar{A}$,

$$\bar{A} = Q_{\bar{A}} \begin{bmatrix} R_3 \\ 0 \end{bmatrix},$$

pričom získavame ortogonálnu maticu $Q_{\bar{A}} \in \mathbb{R}^{m\times(n-p)}$ a regulárnu hornú trojuholníkovu maticu $R_3 \in \mathbb{R}^{(n-p)\times(n-p)}$. Podľa sekcie 2.1 teda zostáva vyriešiť sústavu

$$R_3 x_2 = Q_{\bar{A}}^T \bar{b}.$$

Konečne, riešenie LSE problému je $x = \Pi \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Všimnime si, že (2.7) môžeme interpretovať ako p krokov Gaussovej eliminácie. Podľa [4] uvedený proces symbolicky zapisujeme v tvare

$$\begin{bmatrix} B \\ A \end{bmatrix} \xrightarrow{\Pi,\ \text{QR rozklad}} \begin{bmatrix} R_1, & R_2 \\ \tilde{A}_1, & \tilde{A}_2 \end{bmatrix} \xrightarrow{\text{“Gauss. eliminácia”}} \begin{bmatrix} R_1, & R_2 \\ 0, & \tilde{A}_2 - \tilde{A}_1 R_1^{-1} R_2 \end{bmatrix} =$$

$$\begin{bmatrix} R_1, & R_2 \\ 0, & \bar{A} \end{bmatrix} \xrightarrow{\text{QR rozklad}} \begin{bmatrix} R_1, & R_2 \\ 0, & R_3 \\ 0, & 0 \end{bmatrix}.$$

Zastavme sa ešte pri voľbe permutačnej matice Π. Jedným z dôvodov prečo sa matica B pred QR rozkladom upravuje je, že podľa predpokladu má B viac stĺpcov než riadkov a pretože je $\text{rank}(B) = p$, tak $n - p$ jej stĺpcových vektorov je lineárne závislých. Ch. Van Loan v [8] uvádza jednoduchý príklad na LSE problém s maticou

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix},$$

kde pri riešení úlohy bez použitia stĺpcovej pivotácie dostáva neadekvátne riešenia. Preto sa v [8] navrhuje voľba takej matice Π, pri ktorej v súčine $B\Pi$ sú prvé p stĺpce lineárne nezávislé.

Inú možnosť nachádzame v [1]. A totiž maticu Π vytvárame súbežne s QR rozkladom B. Pretože z numerického hľadiska je výhodne pri QR rozklade použiť Givensové rotácie, či Housholderové zrkadlenia, tak v každom kroku pri spracovávaní daného stĺpca matice B sa za tento stĺpec odporúča brať stĺpec s maximálnou normou, resp. stĺpec, ktorého norma je po vynásobení daným parametrom $\alpha \geq 1$ väčšia ako maximum noriem ostatných stĺpcov. V takom prípade stĺpec nemusí byť určený jednoznačne, a tak volíme ľubovoľný z tých stĺpcov, ktoré dané kritérium spĺňajú. Viac informácií možno nájsť v [1], kde je presne popísaný algoritmus, ktorým sa vykonáva QR rozklad matice B a súčasne formuje permutačná matica Π. Pokúsime sa ho priblížiť.

Snažíme sa previesť QR rozklad matice B Givensovými rotáciami s pivotáciou. Pretože ale pre riešenie x LSE problému musí platiť $Bx = d$, tak matice Givensových rotácií budeme rovno aplikovať aj na vektor d. Kvôli jednotnému značeniu položme $B^{(0)} = B$ a $d^{(0)} = d$.

Pre $k = 1, \ldots, p$ opakuj:

- pre $j = k, \ldots, n$ definujme vektory

$$b_j^{k-1} = \left[b_{k,\,j}^{k-1}, b_{k+1,\,j}^{k-1}, \ldots, b_{p,\,j}^{k-1}\right]^T \in \mathbb{R}^{p-k+1}$$

  a položme $N_j^{k-1} = \|b_j^{k-1}\|_2$. Pracujeme s $n-k+1$ stĺpcovými vektormi dĺžky $p-k+1$, čo si môžeme interprétovať formou matice

$$\tilde{B}^{(k-1)} = [b_k^{k-1}, b_{k+1}^{k-1}, \ldots, b_n^{k-1}] \in \mathbb{R}^{(p-k+1)\times(n-k+1)}.$$

- Vyberme index p_k tak, aby pre dané $\alpha \geq 1$ platilo

$$\alpha N_{p_k}^{k-1} \geq \max_{k\leq j\leq n} N_j^{k-1}. \tag{2.8}$$

  Index p_k nie je všeobecne určený jednoznačne. V prípade, že kritériu vyhovuje viac vektorov, vyberieme jeden z nich.

- Označme $\bar{P}_k^T = \bar{P}_k^T(1, p_k - k + 1) \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ permutačnú maticu, ktorá prehodí prvý a $(p_k - k + 1)-$vý stĺpec v $\tilde{B}^{(k-1)}$ a položme

$$P_k^T = \mathrm{diag}(I_{k-1}, \bar{P}_k^T) \in \mathbb{R}^{n\times n}.$$

  Matica $P_k^T = P_k^T(k, p_k)$ prehodí k-tý a p_k-tý stĺpec v $B^{(k-1)}$.

- Nech ďalej je $\bar{G}_k$ ortogonálna matica zloženej Givensovej rotácie, ktorou vynulujeme až na prvú zložku vektor $b_{p_k}^{k-1}$, tzn. $\bar{G}_k$ prevedie $b_{p_k}^{k-1}$ na vektor $N_{p_k}^{k-1}e_1$, kde $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{p-k+1}$. Je $\bar{G}_k \in \mathbb{R}^{(p-k+1)\times(p-k+1)}$, preto definujme

$$G_k = \mathrm{diag}(I_{k-1}, \bar{G}_k) \in \mathbb{R}^{p\times p}.$$

- Položme $B^{(k)} = G_k B^{(k-1)} P_k^T \in \mathbb{R}^{p\times n}$, $d^{(k)} = G_k d^{(k-1)}$.

Graficky pre $k = 1$:

$$B = B^{(0)} = \tilde{B}^{(0)} = \begin{bmatrix} b_{1,1}^{(0)} & b_{1,2}^{(0)} & \cdots & b_{1,p_k}^{(0)} & \cdots & b_{1,n}^{(0)} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{p,1}^{(0)} & b_{p,2}^{(0)} & \cdots & b_{p,p_k}^{(0)} & \cdots & b_{p,n}^{(0)} \\ \uparrow & \uparrow & & \uparrow & & \uparrow \\ b_1^{(0)} & b_2^{(0)} & & b_{p_k}^{(0)} & & b_n^{(0)} \end{bmatrix} \xrightarrow{BP_1^T(1,p_k)}$$

$$\begin{bmatrix} b_{1,p_k}^{(0)} & b_{1,2}^{(0)} & \cdots & b_{1,1}^{(0)} & \cdots & b_{1,n}^{(0)} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{p,p_k}^{(0)} & b_{p,2}^{(0)} & \cdots & b_{p,1}^{(0)} & \cdots & b_{p,n}^{(0)} \\ \uparrow & \uparrow & & \uparrow & & \uparrow \\ b_{p_k}^{(0)} & b_2^{(0)} & & b_1^{(0)} & & b_n^{(0)} \end{bmatrix} \xrightarrow{G_1 BP_1^T(1,p_k)} \left[\begin{array}{c|c} \frac{N_{p_k}}{0} & \begin{array}{ccc} * & * & * \end{array} \\ \vdots & \tilde{B}^{(1)} \\ 0 & \end{array}\right]$$

$= B^{(1)}$.

Pre $k = p+1, \ldots, n-1$ vyberieme maticu $P_k \in \mathbb{R}^{n\times n}$ rovnakým spôsobom ako doteraz, čím si preskúpime aj ostatné $n-p$ stĺpce. To znamená, že v kroku $k \in [p+1,\, n-1]$ uvážime stĺpce matice $B^{(k)}$ počínajúc $k-$tým stĺpcom a na ne užijeme kritérium (2.8). Získame stĺpec s indexom p_k a po definovaní permutačnej matice P_k^T, ktorá prehodí $k-$tý a p_k-tý stĺpec matice $B^{(k)}$, položíme $B^{(k+1)} = B^{(k)} P_{k+1}^T$.
Celkom je

$$\begin{aligned} \Pi &= P_{n-1} \cdot \ldots \cdot P_1, \\ R &= B^{(p)} P_{p+1}^T \cdot \ldots \cdot P_{n-1}^T = \begin{bmatrix} R_1, & R_2 \end{bmatrix}, \\ Q &= G_1^T \cdot \ldots \cdot G_p^T, \end{aligned}$$

čím získavame permutačnú maticu Π a QR rozklad matice $B\Pi = QR = Q\,[R_1,\, R_2]$.

2.3 Riešenie LSE problému metódou váh

Metóda váh je založená na myšlienke, ktorá už bola prezentovaná a totiž ak nás zaujíma riešenie LS, poprípade LSE problému, tak toto riešenie minimalizuje i úlohu ortogonálne invariantnú, tzn. úlohu s datami prenásobenými ortogonálnou maticou. U tejto metódy sa študuje podobný LS problém ako v sekcií 2.2.1, ale narozdiel od sekcie 2.2.1 tu sa zavádza kladný "dostatočne veľký" prirodzený parameter, tzv. váha a to z toho dôvodu, aby sa zdôraznila rovnosť $Bx = d$. O problematike výberu vhodných váh sa zmienime neskôr, podrobnejšie úvahy možno nájsť v [3, 8].

Hľadajme $x \in \mathbb{R}^n$, ktoré rieši penalizovanú úlohu

$$\min_{x\in\mathbb{R}^n} \left\| \begin{bmatrix} \mu B \\ A \end{bmatrix} x - \begin{bmatrix} \mu d \\ b \end{bmatrix} \right\|_2 \tag{2.9}$$

Stále predpokladáme platnosť (2.4), tj.

$$\operatorname{rank}(B) = p \qquad \text{a} \qquad \operatorname{rank} \begin{bmatrix} \mu B \\ A \end{bmatrix} = n, \tag{2.10}$$

požiadavky na ostatné data zostávajú rovnaké. Vidíme, že dostávame klasický LS problém, ktorý vieme vyriešiť použitím sekcie 2.1.

Voľme preto ortogonálnu maticu Q_μ z QR rozkladu $\begin{bmatrix} \mu B \\ A \end{bmatrix}$

$$\begin{array}{c} p \\ m \\ n \end{array} \begin{bmatrix} \mu B \\ A \end{bmatrix} \;=\; \begin{array}{c} \\ p+m \end{array} \begin{bmatrix} Q_\mu \end{bmatrix} \begin{array}{c} n \\ n \end{array} \begin{bmatrix} R_\mu \\ 0 \end{bmatrix} \begin{array}{l} n \\ p+m-n \end{array},$$

kde $Q_\mu \in \mathbb{R}^{(p+m)\times(p+m)}$ je ortogonálna a $R_\mu \in \mathbb{R}^{n\times n}$ horná trojuholníková matica, ktorá je naviac regulárna, pretože predpokladáme $\operatorname{rank} \begin{bmatrix} \mu B \\ A \end{bmatrix} = n$. Ak si maticu Q_μ rozdelíme spôsobom

$$\begin{array}{c} \\ n \end{array} \begin{array}{c} Q_\mu \;=\; \begin{bmatrix} Q_{1,\mu}, & Q_{2,\mu} \end{bmatrix} \\ \begin{array}{cc} n & p+m-n \end{array} \end{array}, \tag{2.11}$$

$Q_{1,\mu} \in \mathbb{R}^{(p+m)\times n}$, $Q_{2,\mu} \in \mathbb{R}^{(p+m)\times p+m-n}$, tak upravovaním ortogonálne invariantného problému k (2.9) s maticou Q_μ zisťujeme, že stačí vyriešiť sústavu $R_\mu x = Q_{1,\mu}^T \begin{bmatrix} \mu d \\ b \end{bmatrix}$. Získané riešenie (2.9) si označme ako $x(\mu)$.

V ďalšej časti sa pokúsime ukázať konvergenciu riešenia $x(\mu)$ k x_{LSE}, pričom precíznu analýzu metódy váh podáme v časti 2.3.1. Najprv ale zopakujme, že v predchádzajúcich sekciách 2.2 a 2.2.1 (resp. 2.2.2) sme za platnosti predpokladu (2.4) dokázali existenciu a jednoznačnosť riešenia x_{LSE} pôvodného problému (2.3). Avšak na x_{LSE} môžeme nazerať aj ako na minimizér funkcionálu $\varphi(x) = \frac{1}{2}\|Ax - b\|_2^2$ vzhľadom k väzbe $Bx = d$, kde funkcionál $\varphi(x)$ je diferencovateľná funkcia. Definujme Lagrangeovu funkciu

$$L(x, \lambda) = \frac{1}{2}(Ax - b)^T(Ax - b) - \lambda^T(Bx - d).$$

Z jednoznačnosti x_{LSE} potom existuje jednoznačne určené $\lambda_{LSE} \in \mathbb{R}^p$, pre ktoré platí $\frac{\partial L}{\partial x}(x_{LSE}, \lambda_{LSE}) = 0$ a $\frac{\partial L}{\partial \lambda}(x_{LSE}, \lambda_{LSE}) = 0$, pričom priamo z definície derivácie máme

$$\frac{\partial L}{\partial \lambda}(x, \lambda) = -(Bx - d)^T,$$

$$\frac{\partial L}{\partial x}(x, \lambda) = (A^T(Ax - b) - B^T\lambda)^T \quad {}^{(5)}.$$

Ak v týchto vzťahoch pre x_{LSE} a λ_{LSE} položíme $r_{LSE} = b - Ax_{LSE}$, dostávame rovnice

$$\begin{aligned} Bx_{LSE} &= d, \\ r_{LSE} + Ax_{LSE} &= b, \\ B^T\lambda_{LSE} + A^T r_{LSE} &= 0, \end{aligned}$$

ktoré odpovedajú sústave

$$\begin{bmatrix} 0 & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda_{LSE} \\ r_{LSE} \\ x_{LSE} \end{bmatrix} = \begin{bmatrix} d \\ b \\ 0 \end{bmatrix}. \tag{2.12}$$

Z predpokladu (2.4) je matica v (2.12) regulárna.

Pretože $x(\mu)$ rieši (2.9), musí nutne vyhovovať aj normálnej rovnici odvodenej od (2.9), tj. musí spĺňať

$$\begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu B \\ A \end{bmatrix} x(\mu) = \begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu d \\ b \end{bmatrix}.$$

Úpravou tejto rovnosti postupne dostaneme:

$$\begin{bmatrix} \mu B^T, & A^T \end{bmatrix} \begin{bmatrix} \mu B \\ A \end{bmatrix} x(\mu) = \begin{bmatrix} \mu B^T, & A^T \end{bmatrix} \begin{bmatrix} \mu d \\ b \end{bmatrix},$$

[5] Pre tento účel stačí položiť $y = Ax - b$ a definovať funkcionál $F(y) = \frac{1}{2}yy^T$, ktorého derivácia podľa y sa rovná y^T. Ďalej stačí použiť vetu o derivácií zloženého zobrazenia.

Jednoznačnosť λ_{LSE} plynie z rovnice $\frac{\partial L}{\partial x}(x, \lambda) = 0$ pre $x = x_{LSE}$, z jednoznačnosti x_{LSE} a zo vzťahu $\operatorname{rank}(B) = p$.

$$(\mu^2 B^T B + A^T A)x(\mu) = \mu^2 B^T d + A^T b,$$
$$B^T \mu^2 (d - Bx(\mu)) + A^T (b - Ax(\mu)) = 0.$$

Položme $\lambda(\mu) = \mu^2(d - Bx(\mu))$ a $r(\mu) = b - Ax(\mu)$, $\lambda(\mu) \in \mathbb{R}^p$, kde $r(\mu) \in \mathbb{R}^m$, čím získame rovnice

$$\begin{aligned} \mu^{-2}\lambda(\mu) + Bx(\mu) &= d, \\ r(\mu) + Ax(\mu) &= b, \\ B^T\lambda(\mu) + A^T r(\mu) &= 0, \end{aligned}$$

ktoré v maticovom zápise odpovedajú rovnosti

$$\begin{bmatrix} \mu^{-2} I_p & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda(\mu) \\ r(\mu) \\ x(\mu) \end{bmatrix} = \begin{bmatrix} d \\ b \\ 0 \end{bmatrix}. \tag{2.13}$$

Porovnaním oboch systémov (2.12) a (2.13) dostávame chcený výsledok

$$\lim_{\mu \to \infty} x(\mu) = x_{LSE}. \tag{2.14}$$

Presné odvôvodnenie posledného výroku sa opiera o vetu zo [7]. Naznačme stručne postup: Označme si

$$X = \begin{bmatrix} 0 & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix}, \qquad Y_\mu = \begin{bmatrix} \mu^{-2} I_p & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix}.$$

Potom pre $\mu \to \infty$ konverguje Y_μ k X a pretože sme vyššie odvodili, že X je regulárna, tak existuje X^{-1} a $\mu_0 > 0$ také, že pre každé $\mu > \mu_0$ existuje matica Y_μ^{-1}. Podľa známej vety z funkcionálnej analýzy, pozri [6], je norma tejto matice rovnomerne ohraničená pre každé $\mu \in [\mu_0, \infty)$. A teda dostávame odhad

$$\left\| \begin{bmatrix} \lambda_{LSE} \\ r_{LSE} \\ x_{LSE} \end{bmatrix} - \begin{bmatrix} \lambda(\mu) \\ r(\mu) \\ x(\mu) \end{bmatrix} \right\| \leq C \left\| X^{-1} \right\| \left\| Y_\mu - X \right\| \left\| Y_\mu^{-1} \right\| \underset{\mu \to \infty}{\longrightarrow} 0,$$

z ktorého hneď plynie (2.14), C je vhodná konštanta.

Aby sme boli schopní zúžitkovať túto informáciu, potrebujeme pre každé μ opakovane riešiť sústavu $R_\mu x = Q^T \begin{bmatrix} \mu d \\ b \end{bmatrix}$ a to znamená, že potrebujeme pre každé μ previesť QR rozklad matice $\begin{bmatrix} \mu B \\ A \end{bmatrix}$. V praktickom počítaní sa tomu vyhýbame. Ukazuje sa, že voľbou dostatočne veľkého μ je riešenie $x(\mu)$ dostatočne blízko k x_{LSE}. Avšak pri veľmi veľkom μ sa môžu vyskytnúť nepríjemné numerické problémy, napríklad pri QR rozklade matice $\begin{bmatrix} \mu B \\ A \end{bmatrix}$. Totiž zavedením parametra μ rastie úmerne s veľkosťou μ aj číslo

podmienenosti matice $\begin{bmatrix}\mu B\\ A\end{bmatrix}$ a teda aj "náchylnosť" na väčšie zaokrúhľovacie chyby. Naopak, pre μ malé môže byť riešenie opäť nepresné, nakoľko nemusí platiť rovnosť $Bx = d$. Príklad, na ktorom je vidieť závislosť riešenia na parametru μ je uvedený v časti 2.3.2 v tabuľke 2.2. Preto sa pokúsime zostrojiť iteračný algoritmus, ktorý spresní riešenie $x(\mu)$ získané voľbou "stredne veľkého" μ [6].

Pretože LSE problém zohraje dôležitú úlohu v poslednej kapitole a v matlabovských programoch z poslednej kapitoly je na riešenie LSE problému použitá práve metóda váh, v ďalšej časti ešte raz podáme analýzu metódy váh.

2.3.1 Analýza metódy váh

V tejto časti podáme analýzu riešenia získaného metódov váh. Cieľom je vyjadriť riešenia x_{LSE} a $x(\mu)$ presnými analytickými formulkami, z ktorých bude opätovne vidieť, že $x(\mu)$ konverguje k x_{LSE} pre μ idúce do nekonečna. Najprv si uvedieme nasledujúcu vetu, ktorú môžeme nájsť napríkad v [3, 5, 8]. Pretože v literatúre sa pri dôkaze vyskytujú nejasnosti, tak si túto vetu dokážeme podrobne znova [7]. Dôkaz založíme na myšlienke CS rozkladu matíc.

Veta 2.3.1. *Nech $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ $(m \geq n \geq p)$, $rank(B) = p$ a $rank\begin{bmatrix}B\\A\end{bmatrix} = n$, potom existujú ortogonálne matice $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{p\times p}$ a regulárna matica $X \in \mathbb{R}^{n\times n}$ tak, že*

$$U^T AX = D_A = \begin{bmatrix} diag(\alpha_1, \ldots, \alpha_n) \\ 0 \end{bmatrix} = \left[\begin{array}{ccc} \alpha_1 & & \\ & \ddots & \\ & & \alpha_n \\ \hline \multicolumn{3}{c}{0} \end{array}\right] \in \mathbb{R}^{m\times n}, \tag{2.15}$$

$$V^T BX = D_B = \begin{bmatrix} diag(\beta_1, \ldots, \beta_p), & 0 \end{bmatrix} = \left[\begin{array}{ccc|c} \beta_1 & & & \\ & \ddots & & 0 \\ & & \beta_p & \end{array}\right] \in \mathbb{R}^{p\times n}. \tag{2.16}$$

Ak naviac $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ sú singulárne čísla matice $\begin{bmatrix}B\\A\end{bmatrix}$, potom platí

$$\|X\|_2 = 1, \tag{2.17}$$

$$\|X^{-1}\|_2 = \sigma_1/\sigma_n, \tag{2.18}$$

$$0 = \alpha_1 = \cdots = \alpha_q < \alpha_{q+1} \leq \cdots \leq \alpha_p \leq \alpha_{p+1} = \cdots = \alpha_n = \sigma_n, \tag{2.19}$$

pre nejaké $q \in [0, p)$.

$$\beta_1 \geq \cdots \geq \beta_p \geq 0, \tag{2.20}$$

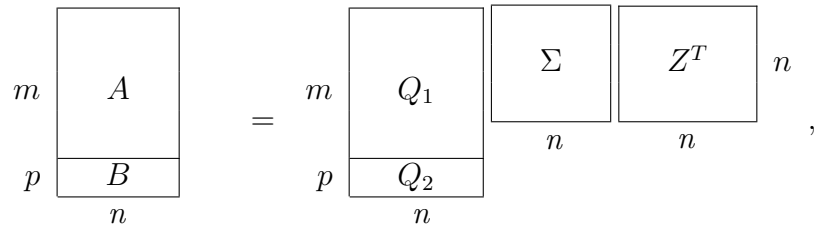
$$\alpha_i^2 + \beta_i^2 = \sigma_n^2, \quad i = 1, \ldots, p. \tag{2.21}$$

[6] Pojmy veľké, stredne veľké a malé μ sú matematicky nepresné. Približný obsah týchto pojmov je objasnený v sekcií 2.3.2.

[7] Napr. v [5] sa rozoberá prípad s rozmermi matíc $m \geq n$ a $p \geq n$. To v našom prípade nenastáva, a preto nie je možné písať niektoré tvrdenia, ktoré sa ďalej v práci objavia, ako dôsledky viet z [5].

Důkaz. Pretože vždy vieme nájsť permutačnú maticu Π, pre ktorú platí $\begin{bmatrix} B \\ A \end{bmatrix} = \Pi \begin{bmatrix} A \\ B \end{bmatrix}$, budeme ďalej pracovať so systémom $\begin{bmatrix} A \\ B \end{bmatrix}$.

Napíšme si singulárny rozklad matice $\begin{bmatrix} A \\ B \end{bmatrix}$:



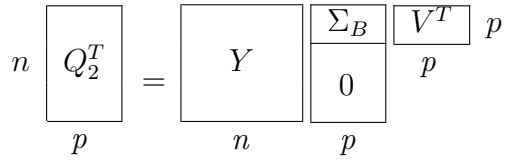
, (i)

kde $Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \in \mathbb{R}^{(m+p)\times n}$, $Z \in \mathbb{R}^{n\times n}$ sú ortogonálne matice a matica $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$ obsahuje singulárne čísla σ_i $(i = 1, 2, \ldots, n)$ matice $\begin{bmatrix} A \\ B \end{bmatrix}$, ktoré si môžeme označiť tak, aby

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

Pretože $\mathrm{rank} \begin{bmatrix} B \\ A \end{bmatrix} = n$, existuje n kladných singulárnych čísel, ktoré sa prirodzene môžu opakovať, avšak nutne je $\sigma_n > 0$ [8].

Z predpokladu $\mathrm{rank}(B) = p$ a z (i) dostávame $\mathrm{rank}(Q_2) = p$ [9]. A tak preveďením singulárneho rozkladu matice Q_2^T, tj.



, (ii)

dostávame p singulárnych čísel $s_1, s_2, \ldots, s_p$, ktoré si znova môžeme označiť tak, aby $s_1 \geq s_2 \geq \cdots \geq s_p$ a teda opäť platí $s_p > 0$. Singulárnym rozkladom získavame ortogonálne matice $Y \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$ a maticu $\Sigma_B = \mathrm{diag}(s_1, s_2, \ldots, s_p) \in \mathbb{R}^{p\times p}$.

Pretože je

$$I_n = Q^T Q = Q_1^T Q_1 + Q_2^T Q_2,$$

tak pre ľubovoľný vektor z jednotkovej sféry S_n platí

$$1 = v^T v = v^T Q_1^T Q_1\, v + v^T Q_2^T Q_2\, v,$$

[8] Samozrejme je

$$\begin{bmatrix} B \\ A \end{bmatrix} = \Pi \begin{bmatrix} A \\ B \end{bmatrix} = \Pi\, Q\, \Sigma\, Z^T = \tilde{Q}\, \Sigma\, Z^T,$$

kde matica $\tilde{Q} = \Pi Q$ je taktiež ortogonálna. A teda vidíme, že oba systémy "A nad B", resp. "B nad A" majú rovnaké singulárne čísla.

[9] Z (i) plynie vzťah $B = Q_2\, \Sigma\, Z^T$.

$$\max_{v \in S_n} v^T Q_2^T Q_2\, v = s_1^2.$$

A pretože $v^T Q_1^T Q_1\, v \geq 0$, $v^T Q_2^T Q_2\, v \geq 0$, tak z týchto vzťahov nutne máme $s_1 \leq 1$ a teda aj $s_i \leq 1$, pre každé $i = 1, \ldots, p$. Z uvedeného teda plynie, že vieme nájsť $q \leq p$ tak, aby platilo

$$1 = s_1 = \cdots = s_q > s_{q+1} \geq \cdots \geq s_p > 0, \tag{iii}$$

čím si Σ_B môžeme upraviť na tvar

$$\Sigma_B = \operatorname{diag}(s_1, s_2, \ldots, s_p) = \begin{bmatrix} I_q & 0 \\ 0 & \Sigma_{\tilde{B}} \end{bmatrix},$$

kde $\Sigma_{\tilde{B}} = \operatorname{diag}(s_{q+1}, s_{q+2}, \ldots, s_p) \in \mathbb{R}^{(p-q)\times(p-q)}$ [10].

Obdĺžnikovú $m \times n$ maticu $Q_1 Y$ rozdeľme na tri časti, a to tak, že prvých q stĺpcov súčinu $Q_1 Y$ označme ako W_1, ďalších $p - q$ ako W_2 a zvyšné W_3, tj.

$$Q_1 Y \;=\; \underset{\text{q}}{[W_1} \;\; \underset{p-q}{W_2} \;\; \underset{n-p}{W_3}]$$

Potom platí

$$\underbrace{\begin{array}{c|c|c|c|c} & m & p & n & \\ \hline m & I_m & 0 & Q_1 & Y \;\; n \\ \hline p & 0 & V^T & Q_2 & \\ \hline & m & p & n & \end{array}}_{\text{K}} \;=\; \underbrace{\begin{array}{c|c|c|c} W_1 & W_2 & W_3 & m \\ \hline I_q & 0 & 0 & q \\ 0 & \Sigma_{\tilde{B}} & 0 & p-q \\ \hline q & p-q & n-p & \end{array}}_{\text{L}} \tag{iv}$$

Vynásobme teraz obe strany rovnosti (iv) transponovanou maticou zapísanou zvlášť v tvare odpovedajúcom matici K a zvlášť matici L. Ľahko sa ukáže, že $K^T K = I_n$, pretože matice Q, Y a V sú ortogonálne. Vynásobme $L^T L$:

$$\begin{bmatrix} W_1^T & I_q & 0 \\ W_2^T & 0 & \Sigma_{\tilde{B}}^T \\ W_3^T & 0 & 0 \end{bmatrix} \begin{bmatrix} W_1 & W_2 & W_3 \\ I_q & 0 & 0 \\ 0 & \Sigma_{\tilde{B}} & 0 \end{bmatrix} = \begin{bmatrix} W_1^T W_1 + I_q & 0 & 0 \\ 0 & W_2^T W_2 + \Sigma_{\tilde{B}}^2 & 0 \\ 0 & 0 & W_3^T W_3 \end{bmatrix}.$$

A pretože musí byť $K^T K = L^T L$, dostávame rovnosť

$$I_n = \begin{bmatrix} I_q & 0 & 0 \\ 0 & I_{p-q} & 0 \\ 0 & 0 & I_{n-p} \end{bmatrix} = \begin{bmatrix} W_1^T W_1 + I_q & 0 & 0 \\ 0 & W_2^T W_2 + \Sigma_{\tilde{B}}^2 & 0 \\ 0 & 0 & W_3^T W_3 \end{bmatrix}.$$

Z toho nám plynie niekoľko pozorovaní:

[10] Prirodzene, môže sa stať, že $q = 0$, avšak na správnosť postupu to nemá žiaden vplyv.

- $W_1^T W_1 = 0 \Rightarrow W_1 = 0$,
- $W_i^T W_j = 0$ pre $i \neq j$,
- W_3 je ortogonálna matica,
- W_2 musí spĺňať rovnosť

$$W_2^T W_2 = I_{p-q} - \Sigma_{\tilde{B}}^2 = \mathrm{diag}(1 - s_{q+1}^2, \ldots, 1 - s_p^2).$$

Položme

$$c_{q+1} = \sqrt{1 - s_{q+1}^2}, \ldots, c_p = \sqrt{1 - s_p^2},$$

$$U_2 = W_2 \,\mathrm{diag}\left(\frac{1}{c_{q+1}}, \ldots, \frac{1}{c_p}\right),$$

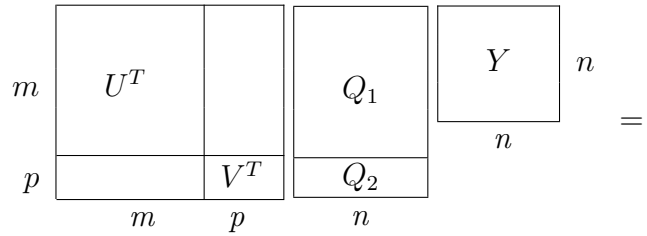
$$U_3 = W_3,$$

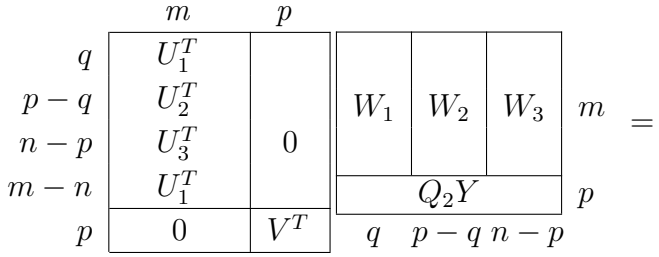
pričom sa ľahko presvedčíme, že U_2 a U_3 sú ortogonálne matice. K vektorom v maticiach U_2 a U_3 doplňme vektory kolmé na $\mathrm{Range}\,\{U_2, U_3\}$, tak aby sme dostali ortogonálnu bázu priestoru $\mathbb{R}^m$. U_1 nech pozostáva z q vektorov, U_4 z $m - n$ vektorov Definujme si teraz ortogonálnu maticu predpisom

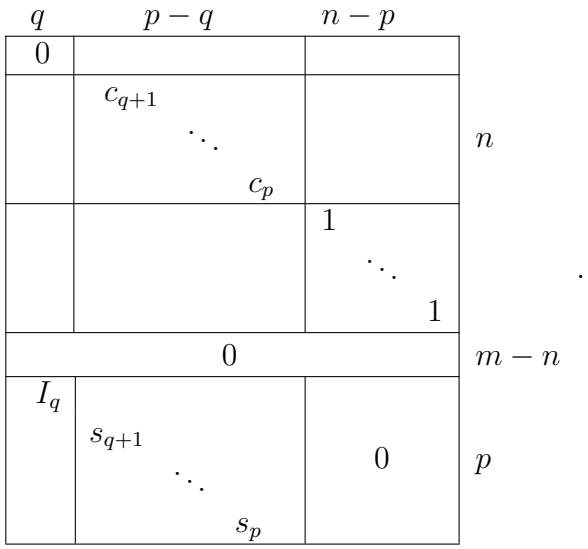
$$U \;=\; \underset{q}{[U_1} \;\; \underset{p-q}{U_2} \;\; \underset{n-p}{U_3} \;\; \underset{m-n}{U_4]} \;\in \mathbb{R}^{m \times m} \quad .$$

Z voľby s_i, $i = 1, \ldots, p$ je $c_1 = \cdots = c_q = 0$. Dodefinujme ešte $c_{p+1} = \cdots = c_n = 1$.

Ukážme, že nasledujúci maticový súčin, ktorý podporuje myšlienku CS rozkladu matíc, vedie k dokončeniu dôkazu. Pri násobení využijeme vlastnosti čísel c_i a s_i, ortogonalitu matice U a predchadzajúce pozorovania. Platí







Takže, k matici Q sme našli ortogonálne matice U, V a Y tak, že súčin

$$\begin{bmatrix} U^T & \\ & V^T \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} Y = \begin{bmatrix} D_c \\ D_s \end{bmatrix}$$

dáva diagonálne matice

$$D_c = \begin{array}{|c|}\hline c_1 \\ \ddots \\ c_n \\ \hline 0 \\ \hline \end{array} \begin{array}{l} n \\ \\ m-n \end{array} \quad a \quad D_s = \begin{array}{|ccc|}\hline s_1 & & \\ & \ddots & \\ & & s_p \\ \hline \end{array} \begin{array}{|c|}\hline \\ \\ \\ \hline \end{array} \begin{array}{l} p \\ \\ \\ \end{array} \begin{array}{c} \\ 0 \\ \\ \end{array}$$

$$\begin{array}{cc} n & \end{array} \qquad \begin{array}{cc} p & n-p \end{array}$$

Pričom vidíme, že dôkaz skutočne rozvýja myšlienku CS rozkladu matíc. Navyše z postupu získavame

- $0 = c_1 = \cdots = c_q < c_{q+1} \leq \cdots \leq c_p \leq c_{p+1} = \cdots = c_n = 1$,
- $1 = s_1 = \cdots = s_q > s_{q+1} \geq \cdots \geq s_p > 0$,
- $c_i^2 + s_i^2 = 1$, pre $i = 1, \ldots, p$.

K dokončeniu dôkazu zostáva dosadiť $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} D_c \\ D_s \end{bmatrix} Y^T$ do (i). Dostávame

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} D_c \\ D_s \end{bmatrix} Y^T \operatorname{diag}(\sigma_1, \ldots, \sigma_n)\, Z^T =$$

$$= \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} D_c \\ D_s \end{bmatrix} \sigma_n \tfrac{1}{\sigma_n} Y^T \operatorname{diag}(\sigma_1, \ldots, \sigma_n)\, Z^T = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} D_A \\ D_B \end{bmatrix} X^{-1},$$

kde sme položili $D_A = \sigma_n D_c$, $D_B = \sigma_n D_s$ a definovali regulárnu maticu X spôsobom

$$X^{-1} = Y^T \frac{1}{\sigma_n} \operatorname{diag}(\sigma_1, \ldots, \sigma_n)\, Z^T = Y^T \begin{bmatrix} \frac{\sigma_1}{\sigma_n} & & \\ & \ddots & \\ & & 1 \end{bmatrix} Z^T$$

a teda

$$X = Z \begin{bmatrix} \frac{\sigma_n}{\sigma_1} & & \\ & \ddots & \\ & & 1 \end{bmatrix} Y.$$

Je zrejme, že $\|X^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$ a $\|X\|_2 = 1$. Ak ešte položíme $\alpha_i = c_i\sigma_n$, $i = 1, \ldots, n$ a $\beta_i = s_i\sigma_n$, $i = 1, \ldots, p$, tak tvrdenie je dokázané. □

Podľa [8] s práve dokázanou vetou ukážeme explicitné vyjadrenie x_{LSE}, $x(\mu)$ a konvergenciu riešenia $x(\mu)$ k riešeniu x_{LSE}. Vetu použijeme k zjednoduchšeniu úlohy (2.3). Predne si uvedomme, že ak si pre $i = 1, \ldots, p$ označíme $\mu_i = \alpha_i/\beta_i$ dostaneme

$$0 = \mu_1 = \cdots = \mu_q < \mu_{q+1} \leq \cdots \leq \mu_p.$$

Zavedením nasledujúceho označenia

$$\begin{aligned} \tilde{b} &= U^T b = \left[u_1^T b, \ldots, u_m^T b\right]^T, \\ \tilde{d} &= V^T d = \left[v_1^T d, \ldots, v_p^T d\right]^T, \\ x &= Xy, \qquad y \in \mathbb{R}^n \end{aligned}$$

si môžeme úlohu (2.3) upraviť, a totiž znova prejdeme k ortogonálne invariantnému problému. Najprv vynásobme sprava rovnicu $Bx = d$ ortogonálnou maticou V^T, je

$$V^T Bx = V^T d \Leftrightarrow V^T BXy = V^T d \Leftrightarrow D_B y = \tilde{d}.$$

A keďže riešenie (2.3) minimalizuje aj $\|U^T Ax - U^T b\|$, kde U^T je ortogonálna matica, môžeme pokračovať v úprave

$$U^T Ax = U^T AXy = D_A y.$$

A teda celkom sa problém (2.3) transformuje na ekvivalentný problém

$$\min_{D_B y = \tilde{d}} \|D_A y - \tilde{b}\|_2, \tag{2.22}$$

pričom riešenia oboch úloh sú vo vzťahu daným maticou X, $x = Xy$.

Ak si uvedomíme, že D_A a D_B sú diagonálne matice definované v (2.15) a (2.16), $\operatorname{rank}(D_B) = p$, $\alpha_1 = \cdots = \alpha_q = 0$, $q < p$, a $\alpha_{p+1} = \cdots = \alpha_n = \sigma_n$, tak riešenie (2.22) musí byť tvaru

$$y_{LSE} = \left[\frac{v_1^T d}{\beta_1}, \quad \cdots \quad, \frac{v_p^T d}{\beta_p}, \quad \frac{u_{p+1}^T b}{\alpha_{p+1}}, \quad \cdots \quad, \frac{u_n^T b}{\alpha_n}\right]^T =$$

$$\left[\frac{v_1^T d}{\beta_1}, \quad \cdots \quad, \frac{v_p^T d}{\beta_p}, \quad \frac{u_{p+1}^T b}{\sigma_n}, \quad \cdots \quad, \frac{u_n^T b}{\sigma_n}\right]^T.$$

Totiž y_{LSE} musí spĺňať rovnosť $D_B\, y_{LSE} = \tilde{d}$. A tak prvých p zložiek vektora y_{LSE} je nutne definovaných práve touto rovnicou. Pretože ale predpokladáme $q < p$, zvyšné

zložky sa dopočítajú zo vzťahu $D_A\, y_{LSE} = \tilde{b}$. Iba v takomto tvare je y_{LSE} riešením (2.22). Potom ale

$x_{LSE} = X y_{LSE} =$

$$= [x_1, \cdots, x_n] \begin{bmatrix} \frac{v_1^T d}{\beta_1} \\ \vdots \\ \frac{v_p^T d}{\beta_p} \\ \frac{u_{p+1}^T b}{\alpha_{p+1}} \\ \vdots \\ \frac{u_n^T b}{\alpha_n} \end{bmatrix} = \sum_{i=1}^{p} \frac{v_i^T d}{\beta_i} x_i + \frac{1}{\sigma_n} \sum_{i=p+1}^{n} (u_i^T b) x_i \quad {}^{(11)}.$$

Riešenie $x(\mu)$, ktoré získame užitím metódy váh, opäť odvodíme úpravou normálnej rovnice príslušnej problému (2.9),

$$\begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu B \\ A \end{bmatrix} x(\mu) = \begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu d \\ b \end{bmatrix}.$$

Roznásobením matíc sa ľahko nahliadne vzťah

$$(A^T A + \mu^2 B^T B) x(\mu) = A^T b + \mu^2 B^T d.$$

Vynásobením zľava maticou X^T a použitím substitúcií (2.15), (2.16) a $x(\mu) = X y(\mu)$ dostaneme

$$(D_A^T D_A + \mu^2 D_B^T D_B) y(\mu) = D_A^T\, \tilde{b} + \mu^2 D_B^T\, \tilde{d} \quad {}^{(12)}.$$

Podobnou úvahou ako pri odvodzovaní tvaru riešenia y_{LSE} je riešením tejto sústavy s diagonálnou maticou vektor

$$y(\mu) = \left[\frac{\alpha_1 u_1^T b + \mu^2 \beta_1 v_1^T d}{\alpha_1^2 + \mu^2 \beta_1^2}, \quad \cdots \quad , \frac{\alpha_p u_p^T b + \mu^2 \beta_p v_p^T d}{\alpha_p^2 + \mu^2 \beta_p^2}, \quad \frac{\alpha_{p+1} u_{p+1}^T b}{\alpha_{p+1}^2}, \quad \cdots \quad , \frac{\alpha_n u_n^T b}{\alpha_n^2} \right]^T,$$

pričom menovateľ každej zložky je vždy nenulový. Podľa predpokladu je totiž μ kladný parameter, ďalej platí (2.19), (2.20) a z dôkazu vety vyplynulo, že $\beta_p > 0$. Konečne je

$$x(\mu) = X y(\mu) = \sum_{i=1}^{p} \frac{\alpha_i u_i^T b + \mu^2 \beta_i v_i^T d}{\alpha_i^2 + \mu^2 \beta_i^2} x_i + \frac{1}{\sigma_n} \sum_{i=p+1}^{n} (u_i^T b) x_i.$$

A teda pre chybu dostávame výraz

$$e(\mu) = x(\mu) - x_{LSE} = \sum_{i=q+1}^{p} \frac{\rho_i \mu_i}{(\mu_i^2 + \mu^2)\beta_i} x_i, \tag{2.23}$$

---

[11] Poznamenajme, že x_{LSE} je súčtom stĺpcových vektorov x_i matice X násobených konštantami (v prípade x_1 až x_p je tým násobkom $\frac{v_i^T d}{\beta_i}$, v prípade x_{p+1} až x_n je to $\frac{u_i^T b}{\sigma_n}$). Podobne tomu bude i naďalej.

[12] Zo substitúcií (2.15), (2.16) hneď plynie $D_A^T = (U^T A X)^T = X^T A^T U$ a $D_B^T = (V^T A X)^T = X^T A^T V$.

kde zavádzame $\rho_i = u_i^T b - \mu_i v_i^T d$ pre $(i = 1, \cdots, p)$. Totiž počítajme:

$$e(\mu) = x(\mu) - x_{LSE} = \sum_{i=1}^{p} \frac{\alpha_i u_i^T b + \mu^2 \beta_i v_i^T d}{\alpha_i^2 + \mu^2 \beta_i^2} x_i - \sum_{i=1}^{p} \frac{v_i^T d}{\beta_i} x_i =$$

$$= \sum_{i=1}^{p} \left(\frac{\alpha_i u_i^T b + \mu^2 \beta_i v_i^T d}{\alpha_i^2 + \mu^2 \beta_i^2} - \frac{v_i^T d}{\beta_i} \right) x_i = \sum_{i=q+1}^{p} \left(\frac{\alpha_i u_i^T b + \mu^2 \beta_i v_i^T d}{\alpha_i^2 + \mu^2 \beta_i^2} - \frac{v_i^T d}{\beta_i} \right) x_i =$$

$$= \sum_{i=q+1}^{p} \frac{\alpha_i \beta_i u_i^T b + \mu^2 \beta_i^2 v_i^T d - v_i^T d(\alpha_i^2 + \mu^2 \beta_i^2)}{(\alpha_i^2 + \mu^2 \beta_i^2)\beta_i} x_i =$$

$$= \sum_{i=q+1}^{p} \frac{\alpha_i \beta_i u_i^T b - \alpha_i^2 v_i^T d}{(\alpha_i^2 + \mu^2 \beta_i^2)\beta_i} x_i = \sum_{i=q+1}^{p} \frac{(u_i^T b - \mu_i v_i^T d)\mu_i}{(\mu_i^2 + \mu^2)\beta_i} x_i =$$

$$= \sum_{i=q+1}^{p} \frac{\rho_i \mu_i}{(\mu_i^2 + \mu^2)\beta_i} x_i.$$

Pričom je opäť vidieť, že pre μ idúce do nekonečna, chyba konverguje k nule.

2.3.2 Numerické porovnanie metód riešiacich LSE problém

V predchádzajúcich častiach sme sa zaoberali teoretickým formulovaním riešení LSE problému. Ukázali sme si tri možné spôsoby riešenia: priamou eliminációu (ďalej PE), projekciou na jadro matice (PJM) a metódov váh (MV). Tiež bolo dokázané, že

$$\text{rank}(B) = p \qquad \text{a} \qquad \text{rank} \begin{bmatrix} B \\ A \end{bmatrix} = n,$$

resp. predpoklad (2.10) je nutnou a postačujúcou podmiekou pre existenciu a jednoznačnosť riešenia, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^m$ a $d \in \mathbb{R}^p$.

Teraz si všetky tri metódy ukážeme na príklade náhodne volených dat z intervalu $[-100,\ 100]$ s rozmermi $m = 7$, $n = 5$, $p = 4$, ktoré spĺňajú uvedené predpoklady (2.4).

Príklad 1.

$$A = \begin{bmatrix} -68 & -17 & 35 & -60 & -71 \\ 83 & -34 & 31 & 21 & -81 \\ 46 & 54 & -88 & 40 & -55 \\ -78 & 75 & 0 & -34 & 19 \\ -73 & -95 & 3 & 32 & 16 \\ -19 & -77 & -62 & 45 & 93 \\ -29 & 38 & 77 & -55 & -43 \end{bmatrix}, \qquad b = \begin{bmatrix} -2 \\ -77 \\ 44 \\ -7 \\ -10 \\ -83 \\ 29 \end{bmatrix},$$

$$B = \begin{bmatrix} -38 & 40 & 38 & -25 & -17 \\ -90 & -89 & -80 & -66 & 66 \\ -94 & -97 & -69 & -21 & 24 \\ -68 & 59 & -85 & 77 & 97 \end{bmatrix}, \qquad d = \begin{bmatrix} -25 \\ 30 \\ -33 \\ 24 \end{bmatrix}.$$

$x(\mu)$	x_{dir}	x_{null}
$0,147494168150341$	$0,147494168150342$	$0,147494168150342$
$0,642081208076150$	$0,642081208076151$	$0,642081208076151$
$-0,228340344397364$	$-0,228340344397364$	$-0,228340344397364$
$-0,827594027726345$	$-0,827594027726346$	$-0,827594027726346$
$0,417140443493639$	$0,417140443493639$	$0,417140443493639$

Tabuľka 2.1: Riešenia LSE problému metódou váh $x(\mu)$, metódou priamej eliminácie x_{dir} a metódou projekcie na jadro matice x_{null}.

Označme si riešenie MV ako $x(\mu)$, riešenie PE ako x_{dir} a konečne riešenie PJM ako x_{null}. Tabuľka 2.1 obsahuje pre dané data riešenia $x(\mu)$ pre voľbu váhy $\mu = 10^8$ a x_{dir}, x_{null}. Vidíme, že hoci metódy MV, PJM a PE sú založené na odlišných ideách, riešenie $x(\mu)$ pre váhu $\mu = 10^8$ a riešenia x_{dir}, x_{null} sú rovnaké až na strojovú presnosť.

Rozdiel riešenia x_{dir} a x_{null} je v norme

$$\|x_{dir} - x_{null}\|_2 = 4,90e-16.$$

Normy rozdielov riešení $x(\mu)$ a x_{dir}, resp. $x(\mu)$ a x_{null} pre všetky uvažované hodnoty $\mu = 10, 10^2, \ldots, 10^{15}$ sú v tabuľke 2.2. Všimnime si, že riešenia $x(\mu)$ a x_{dir}, resp. $x(\mu)$ a x_{null} sú k sebe najbližšie pri voľbe váhy $\mu = 10^8$ a váh väčších. □

Aj pri podobných testovaných príkladoch získavame porovnateľné výsledky ako v uvedenom príklade, ale napríklad už pri datach väčších rozmerov dochádza k zväčšovaniu noriem rozdielov riešení MV od riešení ostatných metód (PE, PJM) a to voľbou váhy $\mu < 10^5$ a $\mu > 10^{11}$. Obrázok 2.1 ukazuje ako sa chová $\|x(\mu) - x_{null}\|_2$ pre voľbu váh $\mu = 10^2, 10^8, 10^{14}$ pri nara-stajúcich rozmeroch dat, ktoré sú dané pomerom $m : n : p = 4k : 3k : 2k$, $k = 1, \ldots, 200$ [13]. Hodnoty dat sú volené náhodne v intervale $[-100, 100]$, tak aby boli splnené predpoklady na hodnosti matíc.

Obrázok ukazuje, že pri volení váhy $\mu = 10^8$ aj s narastajúcimi rozmermi dat je riešenie $x(\mu)$ blízke k riešeniu x_{null} až na malý násobok strojovej presnosti. Pri váhe $\mu = 10^{14}$ je ešte situácia rovnaká s datami veľkosti $m = 32$, $n = 24$ a $p = 16$ (odpovedá koeficientu $k = 8$), ale pre rozmery dané $k \geq 9$ sú riešnia $x(\mu)$ a x_{null} rôzne. Normy $\|x(\mu) - x_{null}\|_2$ sú pri $\mu = 10^2$ rádovo veľkosti 10^{-3}. Podobný graf by sme dostali, keby sme porovnávali riešenia $x(\mu)$ a x_{dir}.

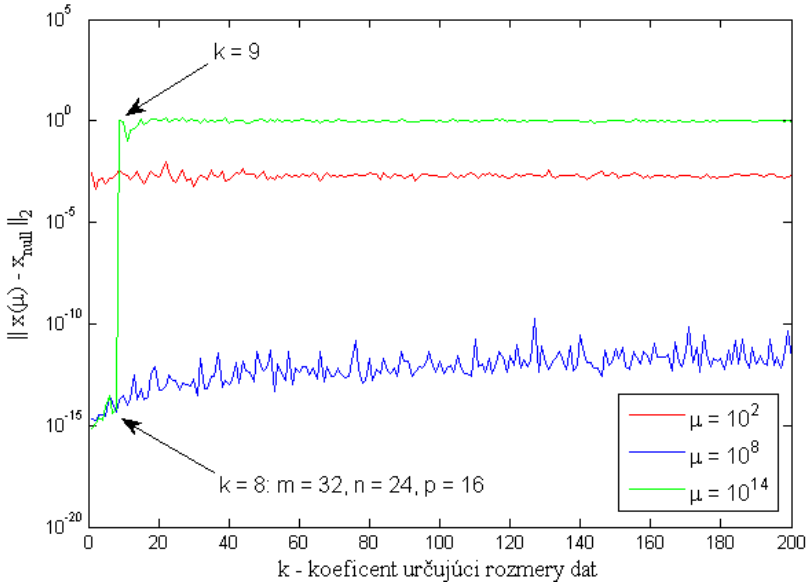
Ch. Van Loan v [8] odporúča pri výpočte LSE riešenia metódou váh používať voľbu $\mu = eps^{-1/2} \approx 10^8$, kde eps značí strojovú presnosť. Z veľkosti noriem rozdielov riešení môžeme usúdiť, že voľbou ktorejkoľvek z metód, pri MV použitím váhy $\mu = 10^8$, získame riešenie LSE problému, ktoré bude až na strojovú presnosť odpovedať skutočnému riešeniu [14].

[13] To znamená, že pre $k = 1$ sú matice A, B a vektory b a d príslušných rozmerov s $m = 4$, $n = 3$ a $p = 2$, pre $k = 200$ zase $m = 800$, $n = 600$ a $p = 400$.

[14] Konečne objasníme nejasnosti okolo pojmov veľké, malé μ. Na základe numerických výpočtov budeme malým μ rozumieť $\mu \in [1, 10^5]$, veľké μ bude $\mu > 10^{11}$, ostatným hodnotám μ hovoríme stredne veľké μ.

μ	$\|x(\mu) - x_{dir}\|_2$	$\|x(\mu) - x_{null}\|_2$
10	$0,0438$	$0,0438$
10^2	$0,0005$	$0,0005$
10^3	$4,6065e-06$	$4,6065e-06$
10^4	$4,6065e-08$	$4,6065e-08$
10^5	$4,6065e-10$	$4,6065e-10$
10^6	$4,6066e-12$	$4,6063e-12$
10^7	$4,5781e-14$	$4,5473e-14$
10^8	$1,6441e-15$	$1,5001e-15$
10^9	$2,0009e-15$	$1,9216e-15$
10^{10}	$6,2125e-16$	$5,6678e-16$
10^{11}	$5,3027e-16$	$7,2431e-16$
10^{12}	$6,9611e-16$	$4,3088e-16$
10^{13}	$9,0280e-16$	$7,5809e-16$
10^{14}	$1,3363e-15$	$1,2332e-15$
10^{15}	$0,3903$	$0,3903$

Tabuľka 2.2: Porovnanie noriem rozdielov riešení LSE problému $x(\mu)$ s x_{dir} a $x(\mu)$ s x_{null} pre hodnoty parametru $\mu = 10^k,\ k = 1, \ldots, 15$.



Obrázok 2.1: Zobrazenie veľkostí noriem rozdielu riešení $x(\mu)$ a x_{null} LSE problému pre náhodne volené data s rozmermi danými vzťahom $m : n : p = 4k : 3k : 2k,\ k = 1, \ldots, 200$. Graf porovnáva závislosť $\|x(\mu) - x_{null}\|_2$ na zväčšujúcich sa rozmeroch dat pre voľbu váhy $\mu = 10^2$ (——), $\mu = 10^8$ (——) a $\mu = 10^{14}$ (——).

2.3.3 Iteračné spresnenie metódy váh

Nasledujúci algoritmus je iteračným spresnením metódy váh. V predchádzajúcej časti sme videli, že metóda váh závisí na voľbe prirodzeného parametra μ, pričom pre príliš malé, resp. veľmi veľké hodnoty μ bolo spočítané riešenie $x(\mu)$ nepresné. Cieľom bude vytvoriť iteračný algoritmus, ktorým získame postupnosť riešení $x^{(k)}$, $k = 1, 2, 3, ...$ takých, že $x^{(k)} \longrightarrow x_{LSE}$ pre k idúce do nekonečna. Pretože sa jedná o algoritmus, ktorý má "vylepšiť" riešenie $x(\mu)$, tak prirodzene volíme $x^{(1)} = x(\mu)$ pre dané nie príliš veľké μ. Numerické výsledky ukazujú, že už pri prvej iterácií dochádza k podstatnému spresneniu riešenia. Viac sa o danom algoritme možno dočítať v [8].

Algoritmus 2.3.1.

1) Pre dané μ spočítaj metódov váh riešenie $x(\mu)$ problému (2.9).
Polož

$$\begin{aligned} x^{(1)} &= x(\mu), \\ r^{(1)} &= b - Ax(\mu), \\ \lambda^{(1)} &= \mu^2(d - Bx(\mu)). \end{aligned}$$

2) Pre $k = 1, 2, 3, \ldots$

2a) Použitím (2.12) spočítaj

$$\begin{bmatrix} \delta_1^{(k)} \\ \delta_2^{(k)} \\ \delta_3^{(k)} \end{bmatrix} = \begin{bmatrix} d \\ b \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda^{(k)} \\ r^{(k)} \\ x^{(k)} \end{bmatrix}.$$

2b) Vyrieš systém (2.13):

$$\begin{bmatrix} \mu^{-2} I_p & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\lambda^{(k)} \\ \Delta r^{(k)} \\ \Delta x^{(k)} \end{bmatrix} = \begin{bmatrix} \delta_1^{(k)} \\ \delta_2^{(k)} \\ \delta_3^{(k)} \end{bmatrix}.$$

2c) Polož

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \Delta x^{(k)}, \\ r^{(k+1)} &= r^{(k)} + \Delta r^{(k)}, \\ \lambda^{(k+1)} &= \lambda^{(k)} + \Delta\lambda^{(k)}. \end{aligned}$$

Idea algoritmu je jasná. Pretože platí (2.14), tak pre pevne zvolené $\mu < \infty$ je $x(\mu) \neq x_{LSE}$. Avšak $x(\mu)$ vyhovuje sústave (2.13), z ktorej vieme spočítať $r^{(1)}$ a $\lambda^{(1)}$. Tie sú dané vzťahmi

$$\begin{aligned} r^{(1)} &= b - Ax(\mu), \\ \lambda^{(1)} &= \mu^2(d - Bx(\mu)). \end{aligned}$$

Pričom nás zaujíma chyba $\|x_{LSE} - x(\mu)\|$, ktorej sme sa dopustili voľbou daného μ. Samozrejme x_{LSE} nepoznáme, avšak vieme, že musí byť riešením (2.12) a teda do (2.12) dosadíme za λ, r a x spočítané hodnoty $\lambda^{(1)}$, $r^{(1)}$ a $x^{(1)} = x(\mu)$ a určíme rozdiel $\left[\delta_1^{(1)}, \delta_2^{(1)}, \delta_3^{(1)}\right]^T$. Pretože je $x(\mu) \neq x_{LSE}$, tak získaný rozdiel je nenulový, ale uvedieme si vetu, ktorá nám hovorí, že $\delta_2^{(1)} = \delta_3^{(1)} = 0$. Tento rozdiel znova dosadíme do (2.13) ako pravú stranu. Pripomeňme, že z regularity systému (2.13) existuje práve jedno riešenie. Označme ho $\left[\Delta\lambda^{(1)}, \Delta r^{(1)}, \Delta x^{(1)}\right]^T$. V poslednom kroku pričítame $\left[\Delta\lambda^{(1)}, \Delta r^{(1)}, \Delta x^{(1)}\right]^T$ k $\left[\lambda^{(1)}, r^{(1)}, x^{(1)}\right]^T$, čím dosiahneme spresnenie riešenia $x(\mu)$. Spresnenie vyplynie z vety 2.3.3. Avšak analýza konvergencie $x^{(k)} \to x_{LSE}$ nie je jednoduchá, preto vetu 2.3.3 dokazovať nebudeme, ale odkážeme na [8]. V ďalšom uvedený postup opakujeme a to dovtedy, kým nie je spĺnená nasledujúca podmienka

$$\|d - Bx^{(k)}\|_2 \leq \delta \|B\|_\infty \|x^{(k)}\|_2, \tag{2.24}$$

kde δ je daná tolerancia. Táto podmienka je heuristikou odvodenou v [8]. Silu podmienky budeme dokumentovať na príklade v časti 2.3.4.

Teraz sa vráťme k analýze predchádzajúceho textu.

Veta 2.3.2. *V algoritme 2.3.1 pre každé k je $\delta_2^{(k)} = 0$ a $\delta_3^{(k)} = 0$.*

Důkaz. Veta sa dokáže indukciou. □

Všimnime si ešte, že $\Delta x^{(k)}$ rieši

$$\min_z \left\| \begin{bmatrix} \mu B \\ A \end{bmatrix} z - \begin{bmatrix} \mu\delta_1^{(k)} \\ 0 \end{bmatrix} \right\|_2, \tag{2.25}$$

čo je opäť penalizovaný LSE problém, riešenie ktorého získame aplikovaním metódy váh zo sekcie 2.3. Tvrdenie nahliadneme z príslušnej normálnej rovnice

$$\begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu B \\ A \end{bmatrix} \Delta x^{(k)} = \begin{bmatrix} \mu B \\ A \end{bmatrix}^T \begin{bmatrix} \mu\delta_1^{(k)} \\ 0 \end{bmatrix}. \tag{2.26}$$

Malou úpravou a položením $\Delta\lambda^{(k)} = \mu^2(\delta_1^{(k)} - B\Delta x^{(k)})$ a $\Delta r^{(k)} = -A\Delta x^{(k)}$ dostávame rovnice odpovedajúce rovnosti

$$\begin{bmatrix} \mu^{-2} I_p & 0 & B \\ 0 & I_m & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\lambda^{(k)} \\ \Delta r^{(k)} \\ \Delta x^{(k)} \end{bmatrix} = \begin{bmatrix} \delta_1^{(k)} \\ 0 \\ 0 \end{bmatrix},$$

ktorá podľa kroku 2b algoritmu 2.3.1 a vety 2.3.2 pre $\Delta x^{(k)}$ platí.

Pomocou vety 2.3.2 nám v algoritme 2.3.1 odpadnú zbytočné výpočty premenných $\lambda^{(k)}$ a $r^{(k)}$, čím sa prirodzene algoritmus zjednodušší. Navyše sa výpočet $\Delta x^{(k)}$ v kroku 2b tohto algoritmu modifikuje na riešenie (2.25). Teda dostávame algoritmus:

Algoritmus 2.3.2.

1) Pre dané μ spočítaj metódov váh riešenie $x(\mu)$ problému (2.9).
   Polož $x^{(1)} = x(\mu)$.

2) Pre $k = 1, 2, 3, \ldots$

   2a) Spočítaj $\delta_1^{(k)} = d - Bx^{(k)}$.

   2b) Rieš $\min \left\| \begin{bmatrix} \mu B \\ A \end{bmatrix} \Delta x^{(k)} - \begin{bmatrix} \mu\delta_1^{(k)} \\ 0 \end{bmatrix} \right\|_2$.

   2c) Polož $x^{(k+1)} = x^{(k)} + \Delta x^{(k)}$.

Nasledujúca veta, ktorá je dokázaná v [8], ukazuje, že $x^{(k)}$ získané algoritmom 2.3.2 konvergujú k x_{LSE}.

Veta 2.3.3. *Pre vektory $x^{(k)}$ vystupujúce v algoritme 2.3.2 platí*

$$x^{(k)} = x_{LSE} + e(\mu,\, k)$$

kde

$$e(\mu,\, k) = \sum_{i=q+1}^{p} \frac{\rho_i}{\alpha_i} \left(\frac{\mu_i^2}{\mu_i^2 + \mu^2} \right)^k x_i.$$

Konštanty ρ_i a μ_i boli odvodené za vetou 2.3.1, q vo vete 2.3.1.

Z tejto vety hneď dostávame

$$\|x_{LSE} - x^{(k)}\| = \|e(\mu,\, k)\| \underset{k\to\infty}{\longrightarrow} 0,$$

pretože platí $\frac{\mu_i^2}{\mu_i^2+\mu^2} < 1$.

2.3.4 Porovnanie metódy váh a iteračného spresnenia

V predchádzajúcej časti sme si ukázali ako možno riešenie $x(\mu)$ získané metódov váh iteračne spresniť. V metóde váh počítame pomocou ortogonálnej matice Q a regulárnej hornej trojuholníkovej matice R_μ z QR rozkladu matice $\begin{bmatrix} \mu B \\ A \end{bmatrix}$

$$\begin{bmatrix} \mu B \\ A \end{bmatrix} = QR = [\, Q_{1,\mu},\, Q_{2,\mu}] \begin{bmatrix} R_\mu \\ 0 \end{bmatrix},$$

zo vzťahov

$$R_\mu x(\mu) = Q_{1,\mu}^T \begin{bmatrix} \mu d \\ b \end{bmatrix},$$

$$R_\mu \Delta x^{(k)} = Q_{1,\mu}^T \begin{bmatrix} \mu\delta_1^{(k)} \\ 0 \end{bmatrix}.$$

postupne pre $k = 1, 2, \ldots$ riešenia $x^{(k)}$ dané vzťahmi v algoritme 2.3.2.

Z programovacieho hľadiska sa v snahe eliminovať zaokrúhľovacie chyby vyhýbame explicitnému vyjadreniu súčinu $Q_{1,\mu}^T \begin{bmatrix} \mu d \\ b \end{bmatrix}$, resp. $Q_{1,\mu}^T \begin{bmatrix} \mu\delta_1^{(k)} \\ 0 \end{bmatrix}$. V praxi sa pri QR rozklade matice prostredníctvom Householderových zrkadlení, či Givensových rotácií transformačné matice priamo aplikujú na pravú stranu sústavy. Tu sa nám ale pravá strana mení, preto by bolo potrebné mať jednotlivé transformácie uložené (v prípade, že QR rozklad prevedieme raz), resp. pre každé k opakovane počítať QR rozklad a transformačné matice aplikovať na pravú stranu. To ale môžeme obísť.

Pretože $\Delta x^{(k)}$ spĺňa (2.26), tak QR rozkladom $\begin{bmatrix} \mu B \\ A \end{bmatrix}$ získavame vzťah

$$R_\mu^T R_\mu \Delta x^{(k)} = \mu^2 B^T \delta_1^{(k)}. \tag{2.27}$$

Vidíme teda, že pri QR rozklade stačí transformačné matice aplikovať na vektor $\begin{bmatrix} \mu d \\ b \end{bmatrix}$, čím získame riešenie $x(\mu)$, ale tieto transformácie nemusíme ukladať, nakoľko pre $k = 1, 2, \ldots$ riešenia $\Delta x^{(k)}$ vieme spočítať z (2.27). Pretože R_μ je podľa predpokladu (2.10) regulárna, tak potom aj súčin $R_\mu^T R_\mu$ je regulárna matica a teda systém (2.27) má jednoznačne určené riešenie $\Delta x^{(k)}$.

Podľa [2] môžeme pri výpočte $\Delta x^{(k)}$ postupovať týmto spôsobom:

$$\begin{aligned}
R_\mu^T R_\mu \Delta z^{(k)} &= \mu^2 B^T \delta_1^{(k)}, \\
r^{(k)} &= \begin{bmatrix} \mu d \\ b \end{bmatrix} - \begin{bmatrix} \mu B \\ A \end{bmatrix} \Delta z^{(k)}, \\
R_\mu^T R_\mu \Delta w^{(k)} &= \begin{bmatrix} \mu B \\ A \end{bmatrix}^T r^{(k)}, \\
\Delta x^{(k)} &= \Delta z^{(k)} + \Delta w^{(k)}.
\end{aligned}$$

To znamená, že v algoritme 2.3.2 nahrádzame krok 2b uvedenými štyrmi operáciami, pričom posledné tri z nich sú ďalším spresnením $\Delta x^{(k)}$.

Nasleduje jednoduchý príklad, na ktorom si ukážeme, že iteračné spresnenie metódy váh skutočne vedie k lepším výsledkom. Riešenie metódy váh bez iteračného spresnenia budeme ako doteraz značiť $x(\mu)$ a riešenie upravené algoritmom 2.3.2 spĺňujúce (2.24) ako X_{iter}.

Príklad 2. Nech sú dané nasledujúce data:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & -1 & 3 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 1 & 1 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 1 \\ 6 \\ 3 \\ 1 \end{bmatrix},$$

$$B = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}.$$

Presné riešenie LSE problému s týmito datami je $x_{exact} = \begin{bmatrix} 0,5, & -0,5, & 1,5 & 0,5 \end{bmatrix}^T$.

Riešenie $x(\mu)$ metódou váh bez iteračného spresnenia získame aplikovaním postupu z časti 2.3. Váhu μ volíme $\mu = 10^8$. Použitím algoritmu 2.3.2 a predchádzajúcich poznámok k programovej realizácií získame riešenie metódou váh s iteračným spresnením x_{iter}. Zopakujme, že iteračný algoritmus je kontrolovaný podmienkou (2.24)

$$\|d - Bx^{(k)}\|_2 \leq \delta \|B\|_\infty \|x^{(k)}\|_2.$$

V našom príklade je pre každú voľbu $\delta < 10^{-16}$ riešenie $x(\mu) = x_{iter}$, čo znamená, že už riešenie $x(\mu)$ vyhovuje kritériu (2.24). V tomto prípade neprebehne iteračný cyklus a platí

$$\|x_{exact} - x(\mu)\|_2 = \|x_{exact} - x_{iter}\|_2 = 4,0030e-16.$$

Pre $\delta = 10^{-16}$ prebehne jeden krok iteračného spresnenia (v algoritme 2.3.2 je $x^{(1)} = x(\mu)$ a $x^{(2)} = x_{iter}$). Platí

$$\|x_{exact} - x(\mu)\|_2 > \|x_{exact} - x_{iter}\|_2.$$

Presne je

$$\|x_{exact} - x(\mu)\|_2 = 4,0030e-16 \qquad \text{a} \qquad \|x_{exact} - x_{iter}\|_2 = 1,5700e-16.$$

Teda, riešenie iteračne upravené je bližšie k skutočnému riešeniu. □

Uveďme ešte príklad, na ktorom je možné vidieť výhodnosť ukončovacieho kritéria (2.24).

Príklad 3. Uvažme data:

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 1 & 2 & -1 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 \\ 2 & 0 & 1 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ -1 \\ -1 \end{bmatrix},$$

$$B = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 0 & -1 & 2 & 0 \end{bmatrix}, \qquad d = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Voľbou $\mu = 10^3$ zistujeme, že prvý iteračný cyklus výrazne vylepšuje riešenie získané metódou váh bez iteračného spresnenia, ktoré sa ako vieme z časti 2.3.2 môže pre malú hodnotu váhy μ podstatne líšiť od presného riešenia. Riešenie bez iteračného spresnenia je označené indexom $k = 1$, $x^{(1)} = x(\mu)$. Nasledujúca tabuľka ukazuje normalizovaný rozdiel počítaného a skutočného riešenia pre dve iterácie a súčasne efektivitu podmienky (2.24).

k	$\frac{\|x^{(k)} - x_{exact}\|_2}{\|x^{(k)}\|_2}$	$\frac{\|d - Bx^{(k)}\|_2}{\|B\|_\infty \|x^{(k)}\|_2}$
1	$3,50e-6$	$1,09e-6$
2	$5,47e-12$	$1,74e-12$
3	$1,21e-15$	$2,79e-17$

Tabuľka 2.3: Ukážka efektivity ukončovacieho kritéria (2.24). Prvý stĺpec tabuľky obsahuje normalizovanú odchýlku od skutočného riešenia pre riešenie získané metódou váh ($k = 1$) a prvé a druhé iteračné spresnenie ($k = 2, 3$). Druhý stĺpec zachytáva správanie sa podmienky (2.24) pre jednotlivé iterácie.

Vidíme, že skutočne pre malú hodnotu váhy $\mu = 10^3$ prvé iteračné spresnenie ($k = 2$) znamená výrazné spresnenie riešenia $x(\mu)$. Druhým iteračným spresnením ($k = 3$) získavame riešenie, ktorého normalizovaný rozdiel od skutočného riešenia sa rovná malému násobku strojovej presnosti *eps*. Z druhého stĺpca tabuľky 2.3 plynie, že ak chceme dosiahnúť presnosť riešenia x_{iter} na úrovni strojovej presnosti k skutočnému riešeniu, potrebujeme voliť v ukončovacej podmienke (2.24) $\delta \approx 10^{-13}$. V takom prípade prebehnú obidva kroky iteračného algoritmu. □

Kapitola 3

Výpočet GCD

Pre presne zadané data je možné na výpočet GCD dvoch polynómov použiť známy Euklidov algoritmus, avšak v pohyblivej radovej čiarke nie je algoritmus stabilný. Navyše v praxi ako výsledky rôznych meraní, či kvôli reprezentácií dat v pamäti počítača, sa pracuje s nepresnými datami s nejakou špecifickou toleranciou. Preto prirodzene vznikla snaha o konštrukciu stabilného algoritmu, ktorý by s úspechom spracovával aj nepresne zadané data. Súhlasne s [13] *ukážeme, že Euklidov algoritmus sa dá implementovať vo forme úprav Sylvestrovej matice, ktorá poskytuje pevnejšiu datovú štruktúru. GCD tak získame transformáciou Sylvestrovej matice na dolný trojuhlníkový tvar, pričom úpravy do istej miery odpovedajú Euklidovmu algoritmu. Ďalej ale rozlíšime dve možné spôsoby úprav. Prvý je založený na Gaussovej eliminácií, druhý využíva idey Givensových rotácií.*

3.1 Sylvestrova matica a jej použitie pri výpočte GCD

Sylvestrovou maticou[1)] zostavenou z koeficientov polynómov

$$f(x) = a_0x^m + a_1x^{m-1} + \cdots + a_{m-1}x + a_m, \tag{3.1}$$

a

$$g(x) = b_0x^n + b_1x^{n-1} + \cdots + b_{n-1}x + b_n, \tag{3.2}$$

[1)] Sylvestrova matica je typ tzv. rezultantových matíc. Do tejto skupiny patrí aj napr. Bézoutova matica, či tzv. companion matica. Zavedenie Sylvestrovej matice a niektoré vlastnosti možno nájsť napr. v [9]

rozumieme maticu

$$S(f,g) = \begin{bmatrix} a_0 & & & & b_0 & & & \\ a_1 & a_0 & & & b_1 & b_0 & & \\ \vdots & a_1 & \ddots & & \vdots & b_1 & \ddots & \\ a_{m-1} & \vdots & \ddots & a_0 & b_{n-1} & \vdots & \ddots & b_0 \\ a_m & a_{m-1} & \ddots & a_1 & b_n & b_{n-1} & \ddots & b_1 \\ & a_m & \ddots & \vdots & & b_n & \ddots & \vdots \\ & & \ddots & a_{m-1} & & & \ddots & b_{n-1} \\ & & & a_m & & & & b_n \end{bmatrix}.$$

Sylvestrova matica $S(f,g) \in \mathbb{R}^{(m+n)\times(m+n)}$ je teda štvorcová matica odstupňovanej štruktúry, ktorej prvých n stĺpcov obsahuje koeficienty polynómu f, zvyšných m stĺpcov koeficienty polynómu g. O polynómoch f a g, $\deg f = m$, $\deg g = n$, budeme predpokladať, že $m \geq n$ a koeficienty a_0, a_m, b_0, b_n sú všetky rôzne od nuly.

Ďalej si zaveďme k-tý Sylvestrov subrezultant $S_k(f,g) \in \mathbb{R}^{(m+n-k+1)\times(m+n-2k+2)}$, ktorý získame z matice $S(f,g)$ odstránením posledných $k-1$ riadkov, posledných $k-1$ stĺpcov z koeficientov polynómu f a posledných $k-1$ stĺpcov z koeficientov polynómu g.

Príklad 4. Napríklad pre polynómy

$$f(x) = a_0x^4 + a_1x^3 + a_2x^2 + a_3x + a_4, \quad g(x) = b_0x^3 + b_1x^2 + b_2x + b_3$$

stupňov 4 a 3 je Sylvestrova matica tvaru

$$S_1 = S(f,g) = \begin{bmatrix} a_0 & & & b_0 & & & \\ a_1 & a_0 & & b_1 & b_0 & & \\ a_2 & a_1 & a_0 & b_2 & b_1 & b_0 & \\ a_3 & a_2 & a_1 & b_3 & b_2 & b_1 & b_0 \\ a_4 & a_3 & a_2 & & b_3 & b_2 & b_1 \\ & a_4 & a_3 & & & b_3 & b_2 \\ & & a_4 & & & & b_3 \end{bmatrix}.$$

Druhý subrezultant

$$S_2 = \begin{bmatrix} a_0 & & & b_0 & & & \\ a_1 & a_0 & & b_1 & b_0 & & \\ a_2 & a_1 & a_0 & b_2 & b_1 & b_0 & \\ a_3 & a_2 & a_1 & b_3 & b_2 & b_1 & b_0 \\ a_4 & a_3 & a_2 & & b_3 & b_2 & b_1 \\ & a_4 & a_3 & & & b_3 & b_2 \\ & & a_4 & & & & b_3 \end{bmatrix} = \begin{bmatrix} a_0 & & b_0 & & \\ a_1 & a_0 & b_1 & b_0 & \\ a_2 & a_1 & b_2 & b_1 & b_0 \\ a_3 & a_2 & b_3 & b_2 & b_1 \\ a_4 & a_3 & & b_3 & b_2 \\ & a_4 & & & b_3 \end{bmatrix}$$

dostaneme z S_1 odstránením posledného riadka, tretieho a posledného stĺpca. □

Zaveďme ešte dôležité označenie a totiž označme prvý stĺpec matice $S_k(f,g)$ ako c_k, zvyšné stĺpce ako A_k. Takže pre každé prípustné k buď

$$S_k(f,g) = [c_k \mid A_k], \tag{3.3}$$

kde $c_k \in \mathbb{R}^{m+n-k+1}$ a $A_k \in \mathbb{R}^{(m+n-k+1)\times(m+n-2k+1)}$.

Jednou z dôležitých vlastností Sylvestrovej matice, o ktorej sa možno dočítať napr. v [9], je vlastnosť, že $\det S(f,g) = 0$ je nutnou a postačujúcou podmienkou k tomu, aby polynómy f a g mali netriviálny spoločný deliteľ. Pričom, ak polynómy sú súdeliteľné, stupeň GCD i samotný GCD môže byť spočítaný transformáciou Sylvestrovej matice. Neskôr sa bližšie zmienime o vzťahu hodnosti Sylvestrovej matice a stupňa GCD dvoch polynómov.

3.2 Výpočet GCD, súvislosť medzi transformáciou Sylvestrovej matice a Euklidovým algoritmom

V tejto sekcií ukážeme spojitosť medzi Euklidovým algoritmom a transformáciou Sylvestrovej matice. Uvidíme, že transformácia Sylvestrovej matice predstavuje iný logický zápis Euklidovho algoritmu. Poznamenajme, že nasledujúce teoretické úvahy platia v presnej aritmetike.

Budeme predpokladať, že máme polynómy $f(x)$ a $g(x)$ definované v (3.1) a (3.2), $\deg f = m$, $\deg g = n$, $m \geq n$ a koeficienty $a_0 \times a_m \neq 0$ a $b_0 \times b_n \neq 0$.

Pripomeňme Euklidov algoritmus.

Kvôli jednotnému značeniu položme $f_0 := f$ a $f_1 = g$. Predpokladajme teraz, že pre $i = 0, 1, 2, \ldots$ máme spočítané polynómy f_i a f_{i+1}, ktoré sú nenulové a

$$\deg f_{i+1} < \deg f_i.$$

Potom existujú polynómy q_i a f_{i+2} spĺňujúce

$$f_i = q_i\, f_{i+1} + f_{i+2}, \tag{3.4}$$

$$\deg f_{i+2} < \deg f_{i+1}$$

A ak platí $f_{i+2} = 0$, tak potom

$$f_{i+1} = \mathrm{GCD}(f,g).$$

V prípade, že f_{i+2} je konštanta rôzna od nuly, tak polynómy f a g sú nesúdeliteľné.

Môžeme teda pristúpiť k popisu transformácie Sylvestrovej matice. Celý postup budeme demonštrovať na jednoduchom príklade s polynómami stupňov $\deg f = 5$ a $\deg g = 2$, pričom precízne predvedieme prvý krok (jednotlivé kroky transformácie budeme nazývať behmi), ktorý v klasickom poňatí Euklidovho algoritmu (3.4) pre $i = 0$ odpovedá nájdeniu polynómu f_2. Podobný príklad i všeobecný postup možno nájsť v [13]. Avšak, hoci sa vyhneme všeobecným zápisom, príklad dá jasnú predstavu o priebehu celého algoritmu [2].

[2] Všeobecnej formulácií sa vyhýbame len kvôli rozsiahlejšiemu zápisu.

Buďte teda

$$f_0(x) = f(x) = a_0x^5 + a_1x^4 + a_2x^3 + a_3x^2 + a_4x + a_5,$$
$$f_1(x) = g(x) = b_0x^2 + b_1x + b_2$$

a im odpovedajúca Sylvestrova matica

$$S = S(f_0, f_1) = \begin{bmatrix} a_0 & & b_0 & & & & \\ a_1 & a_0 & b_1 & b_0 & & & \\ a_2 & a_1 & b_2 & b_1 & b_0 & & \\ a_3 & a_2 & & b_2 & b_1 & b_0 & \\ a_4 & a_3 & & & b_2 & b_1 & b_0 \\ a_5 & a_4 & & & & b_2 & b_1 \\ & a_5 & & & & & b_2 \end{bmatrix}.$$

Prvý beh transformácie spočíva v eliminácií prvkov matice S, ktoré sú príslušné koeficientom polynómu f_0. Eliminácia koeficientov a_0, $a_1, \ldots$ prebieha postupne, pričom vždy zvyšnými koeficientami, ktoré po eliminácií každého z prvkov zostanú v prvom stĺpci, definujeme nový polynóm. V nadväznosti na Euklidov algoritmus (3.4) sa dá tušiť, že beh skončí v momente, keď prvý z vytváraných polynómov bude mať stupeň menší ako je stupeň f_1. Prv než opustíme nepresné úvahy, poznamenajme, že elimináciu možno docieliť niekoľkými spôsobmi. Predvedieme dva z nich, z ktorých ten prvý odpovedá klasickým úpravám v Euklidovom algoritme.

3.2.1 Transformácia Sylvestrovej matice elementárnymi trojuholníkovými maticami

V nasledujúcej časti ukážeme redukciu Sylvestrovej matice pomocou elementárnej trojuholníkovej matice, ktorú pre $n \in \mathbb{N}$ definujeme spôsobom

$$E_{i,j}(\sigma) = I - \sigma e_i e_j^T,$$

kde $I \in \mathbb{R}^{n\times n}$ je jednotková matica a e_i je jej i-tý stĺpec. Pre $i \geq j$ je $E_{i,j}(\sigma)$ dolná trojuholníková, v opačnom prípade horná trojuholníková matica.

Napríklad pre $n = 4$ je

$$E_{3,1}(\sigma) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\sigma & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Pretože je $\deg f_0 > \deg f_1$ a $a_0 \neq 0$, tak prvý krok behu predstavuje vynásobenie

matice $S(f_0, f_1)$ postupne maticami $E_{3,1}(a_0/b_0)$ a $E_{4,2}(a_0/b_0)$, čím dostaneme maticu

$$S^{(1)} = \begin{bmatrix} 0 & & b_0 & & & & \\ a_1^{(1)} & 0 & b_1 & b_0 & & & \\ a_2^{(1)} & a_1^{(1)} & b_2 & b_1 & b_0 & & \\ a_3^{(1)} & a_2^{(1)} & & b_2 & b_1 & b_0 & \\ a_4^{(1)} & a_3^{(1)} & & & b_2 & b_1 & b_0 \\ a_5^{(1)} & a_4^{(1)} & & & & b_2 & b_1 \\ & a_5^{(1)} & & & & & b_2 \end{bmatrix}.$$

To presne odpovedá polynomiálnym operáciam

$$h_4(x) := a_1^{(1)}x^4 + a_2^{(1)}x^3 + a_3^{(1)}x^2 + a_4^{(1)}x + a_5^{(1)} = f_0(x) - f_1(x)\frac{a_0}{b_0}x^3.$$

Prirodzene, môže sa stať, že sa týmto krokom vynuluje aj koeficient $a_1^{(1)}$. V takom prípade sa tomu odpovedajúci krok behu vynechá. Bez újmy na všeobecnosti predpokladajme, že $a_1^{(1)} = 0$ a $a_2^{(1)} \neq 0$. Druhý krok teda vynecháme, resp. lepšie povedané sa matica $S^{(1)}$ prenásobí maticami $E_{4,1} = E_{5,2} = I$, nakoľko $\sigma = a_1^{(1)}/\, b_0 = 0$. Iste potom je

$$h_3(x) = h_4(x) = a_2^{(1)}x^3 + a_3^{(1)}x^2 + a_4^{(1)}x + a_5^{(1)},$$

$\deg h_3 \geq \deg g$. Indexom v h_3 vyjadrujeme stupeň tohto polynómu.

Násobme teraz maticu $S^{(1)}$ maticami $E_{5,1}(a_2^{(1)}/b_0)$ a $E_{6,2}(a_2^{(1)}/b_0)$. Dostaneme

$$S^{(2)} = \begin{bmatrix} 0 & & b_0 & & & & \\ 0 & 0 & b_1 & b_0 & & & \\ 0 & 0 & b_2 & b_1 & b_0 & & \\ a_3^{(2)} & 0 & & b_2 & b_1 & b_0 & \\ a_4^{(2)} & a_3^{(2)} & & & b_2 & b_1 & b_0 \\ a_5^{(2)} & a_4^{(2)} & & & & b_2 & b_1 \\ & a_5^{(2)} & & & & & b_2 \end{bmatrix},$$

čo za dodatočného predpokladu $a_3^{(2)} \neq 0$ odpovedá

$$h_2(x) := a_3^{(2)}x^2 + a_4^{(2)}x + a_5^{(2)} = h_3(x) - f_1(x)\frac{a_2^{(1)}}{b_0}x, \qquad \deg h_2 = \deg f_1 \quad {}^{(3)}.$$

A keďže je $a_3^{(2)} \neq 0$, vynásobme $S^{(2)}$ maticami $E_{6,1}(a_3^{(2)}/b_0)$ a $E_{7,2}(a_3^{(2)}/b_0)$. Získame tak

$$S^{(3)} = \begin{bmatrix} 0 & & b_0 & & & & \\ 0 & 0 & b_1 & b_0 & & & \\ 0 & 0 & b_2 & b_1 & b_0 & & \\ 0 & 0 & & b_2 & b_1 & b_0 & \\ a_4^{(3)} & 0 & & & b_2 & b_1 & b_0 \\ a_5^{(3)} & a_4^{(3)} & & & & b_2 & b_1 \\ & a_5^{(3)} & & & & & b_2 \end{bmatrix}$$

[3] Samozrejme, $a_3^{(2)} \neq 0$ opäť predpokladáme bez újmy na všeobecnosti.

a odpovedajúcu polynomiálnu operáciu

$$h_1(x) := a_4^{(3)}x + a_5^{(3)} = h_2(x) - f_1(x)\frac{a_3^{(2)}}{b_0}$$

s $\deg h_1 < \deg f_1$, čím je prvý beh na konci.

Ak si zosumarizujeme všetky operácie tohto behu, máme

$$f(x) = f_0(x) = h_4(x) + \frac{a_0}{b_0}x^3 f_1(x) = h_3(x) + \frac{a_0}{b_0}x^3 f_1(x) = h_2(x) + \left(\frac{a_0}{b_0}x^3 + \frac{a_2^{(1)}}{b_0}x\right) f_1(x)$$

$$= h_1(x) + \left(\frac{a_0}{b_0}x^3 + \frac{a_2^{(1)}}{b_0}x + \frac{a_3^{(2)}}{b_0}\right) f_1(x) = f_2(x) + q_0(x)g(x),$$

kde sme položili

$$f_2(x) = h_1(x) = a_4^{(3)}x + a_5^{(3)}$$
$$q_0(x) = \frac{a_0}{b_0}x^3 + \frac{a_2^{(1)}}{b_0}x + \frac{a_3^{(2)}}{b_0}.$$

Avšak tieto vzťahy presne kopírujú (3.4) s $i = 0$ ($\deg f_2 = \deg h_1 < \deg f_1 = \deg g$).

Položme

$$T_1 = E_{3,1}(\tfrac{a_0}{b_0})E_{4,2}(\tfrac{a_0}{b_0})E_{4,1}(\tfrac{a_1^{(1)}}{b_0})E_{5,2}(\tfrac{a_1^{(1)}}{b_0})E_{5,1}(\tfrac{a_2^{(1)}}{b_0})E_{6,2}(\tfrac{a_2^{(1)}}{b_0})E_{6,1}(\tfrac{a_3^{(2)}}{b_0})E_{7,2}(\tfrac{a_3^{(2)}}{b_0})$$

a definujme permutačnú maticu $P \in \mathbb{R}^{7\times 7}$ predpisom

$$P_1 = (e_3, e_4, e_5, e_6, e_7, e_1, e_2)\,.$$

Potom

$$S^{(4)} = S^{(3)}P_1 = S(f,g)T_1P_1 = \begin{bmatrix} b_0 & & & & & & \\ b_1 & b_0 & & & & & \\ b_2 & b_1 & b_0 & & & & \\ & b_2 & b_1 & b_0 & & & \\ & & b_2 & b_1 & b_0 & a_4^{(3)} & \\ & & & b_2 & b_1 & a_5^{(3)} & a_4^{(3)} \\ & & & & b_2 & & a_5^{(3)} \end{bmatrix}$$

a vidíme, že vyznačená matica rozmerov 3×3 je opäť Sylvestrova matica polynómov f_1 a f_2. Takže druhý (zároveň posledný) beh transformácie spočíva v úprave $S(f_1, f_2)$ rovnakým postupom ako v popísanom prvom behu a to len vtedy, ak $f_2 \neq 0$. Ak totiž $f_2 = 0$, tak $f_1 = \mathrm{GCD}(f, g)$ a $S^{(4)}$ je dolná trojuholníková matica. Nech teda $f_2 \neq 0$, potom výsledkom druhého behu bude konštanta $f_3 = b_2^{(2)}$, ktorá ak bude nenulová, tak f a g sú nesúdeliteľné a ak rovná nule, tak f_2 je hľadaný $\mathrm{GCD}(f, g)$.

Podobne ako pri prvom behu zaveďme 3×3 matice $\tilde{T}_2$ a $\tilde{P}_2$ a položme $T_2 = \mathrm{diag}(I_4, \tilde{T}_2)$ a $P_2 = \mathrm{diag}(I_4, \tilde{P}_2)$. Celý transformačný proces môžeme zapísať symbolicky v tvare

$$S(f,g) = \begin{bmatrix} a_0 & & b_0 & & & & \\ a_1 & a_0 & b_1 & b_0 & & & \\ a_2 & a_1 & b_2 & b_1 & b_0 & & \\ a_3 & a_2 & & b_2 & b_1 & b_0 & \\ a_4 & a_3 & & & b_2 & b_1 & b_0 \\ a_5 & a_4 & & & & b_2 & b_1 \\ & a_5 & & & & & b_2 \end{bmatrix} \xrightarrow{S\,T_1P_1T_2P_2} \begin{bmatrix} b_0 & & & & & & \\ b_1 & b_0 & & & & & \\ b_2 & b_1 & b_0 & & & & \\ & b_2 & b_1 & b_0 & & & \\ & & b_2 & b_1 & a_4^{(3)} & & \\ & & & b_2 & a_5^{(3)} & a_4^{(3)} & \\ & & & & & a_5^{(3)} & b_2^{(2)} \end{bmatrix},$$

čo je výsledná transformácia matice $S(f,g)$ na dolný trojuholníkový tvar.

Vo všeobecnom prípade postup z prvého behu aplikujeme induktívne na každú Sylvestrovu maticu príslušnú polynómom f_i a f_{i+1}. Získame polynóm f_{i+2}, ktorý ak je nulový, algoritmus končí a $f_{i+1} = \mathrm{GCD}(f,g)$, ak nenulový so stupňom väčším alebo rovným 1, tak pokračuje ďalším behom. V prípade, že f_{i+2} je nenulová konštanta, algoritmus opäť končí a f, g sú nesúdeliteľné.

V prvom behu redukcie $S(f_0, f_1)$ na $S(f_1, f_2)$ platí vzťah dokázaný v [13]

$$\mathrm{rank}\, S(f_0, f_1) = m - \deg f_2 + \mathrm{rank}\, S(f_1, f_2). \tag{3.5}$$

Z (3.5) hneď dostávame, že ak je $f_2 = 0$, čo znamená, že f_1 delí f_0, tak potom $\mathrm{rank}\, S(f_0, f_1) = m$. Ak ďalej transformujeme $S(f_1, f_2)$, $S(f_2, f_3), \ldots$, tak opakovaným použitím (3.5) získame všeobecnú vetu, ktorej dôkaz možno nájsť taktiež v [13].

Veta 3.2.1. *Buďte $f(x)$ a $g(x)$ polynómy definované v (3.1), resp. (3.2),* $\deg(f) = m$, $\deg(g) = n$, $m \geq n$ *a buď* $1 \leq k \leq n$ *nezáporné číslo. Potom platí ekvivalencia*

$$rank\, S(f,g) = m + n - k \Leftrightarrow \deg GCD(f,g) = k.$$

Ak ďalej S_k je k-tý Sylvestrov subrezultant, platí

$$\mathrm{rank}\, S_k(f,g) = m + n - 2k + 1 \Leftrightarrow \deg \mathrm{GCD}(f,g) = k.$$

Z tejto vety hneď plynie, že pre nesúdeliteľné polynómy f a g definované v (3.1), resp. (3.2) je $S(f,g)$ regulárna matica.

3.2.2 c-s transformácia Sylvestrovej matice

Ako už bolo povedané, existuje niekoľko spôsobov ako eliminovať prvky v Sylvestrovej matici. Rovnaký výsledok ako použitím elemntárnych trojuholníkových matíc $E_{i,j}(\sigma)$ docielime i spôsobom pripominajúcim Givensové rotácie. Zavedieme matice G_i, ktoré budú závisieť na istých prvkoch c a s takých, že $c^2 + s^2 = 1$. Nimi budeme prevádzať redukciu Sylvestrovej matice na dolný trojuholníkový tvar a preto túto elimináciu nazývame c-s transformáciou Sylvestrovej matice. Pretože by sme ale v značnej miere

kopírovali predchádzajúci príklad predvedieme hlavnú myšlienku. Pripomeňme predpoklady $\deg f = m \geq \deg g = n$ a $a_0 \neq 0$, $b_0 \neq 0$.

Pretože sa podobne ako v prvom behu eliminácie elementárnymi trojuholníkovými maticami chceme dopracovať k polynómu stupňa menšieho než $\deg g$ potrebujeme odstrániť v

$$S = S(f_0, f_1) = \begin{bmatrix} a_0 & & b_0 & & & & \\ a_1 & a_0 & b_1 & b_0 & & & \\ a_2 & a_1 & b_2 & b_1 & b_0 & & \\ a_3 & a_2 & & b_2 & b_1 & b_0 & \\ a_4 & a_3 & & & b_2 & b_1 & b_0 \\ a_5 & a_4 & & & & b_2 & b_1 \\ & a_5 & & & & & b_2 \end{bmatrix}.$$

koeficienty a_0 až a_3. Vynásobme preto S najprv maticou

$$G_1(c, s) = \begin{bmatrix} c & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & 0 & 0 & 0 \\ s & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & s & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

kde

$$c = \frac{b_0}{\sqrt{a_0^2 + b_0^2}} \qquad \text{a} \qquad s = -\frac{a_0}{\sqrt{a_0^2 + b_0^2}}.$$

Dostaneme maticu

$$S^{(1)} = \begin{bmatrix} 0 & & b_0 & & & & \\ a_1^{(1)} & 0 & b_1 & b_0 & & & \\ a_2^{(1)} & a_1^{(1)} & b_2 & b_1 & b_0 & & \\ a_3^{(1)} & a_2^{(1)} & & b_2 & b_1 & b_0 & \\ a_4^{(1)} & a_3^{(1)} & & & b_2 & b_1 & b_0 \\ a_5^{(1)} & a_4^{(1)} & & & & b_2 & b_1 \\ & a_5^{(1)} & & & & & b_2 \end{bmatrix}.$$

Rovnakými úvahami ako v sekcií (2.2.1) pokračujeme ďalej. To znamená, že si definujme polynóm

$$h_4(x) := a_1^{(1)}x^4 + a_2^{(1)}x^3 + a_3^{(1)}x^2 + a_4^{(1)}x + a_5^{(1)} = cf_0(x) + sf_1(x).$$

Ak je $a_1^{(1)} \neq 0$, $\deg h_4 > \deg g$, ďalej násobíme $S^{(1)}$ maticou

$$G_2(c_1, s_1) = \begin{bmatrix} c_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ s_1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & s_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

kde

$$c_1 = \frac{b_0}{\sqrt{(a_1^{(1)})^2 + b_0^2}} \qquad \text{a} \qquad s_1 = -\frac{a_1^{(1)}}{\sqrt{(a_1^{(1)})^2 + b_0^2}}.$$

Ak ale nastane prípad $a_1^{(1)} = 0$, tak pre práve definované c_1 a s_1 platí $c_1 = 1$ a $s_1 = 0$ a teda $G_2(c_1, s_1) = I$. Prvý beh pokračuje dovtedy, kým pre nejaké h_k nenastane $\deg h_k < \deg g$, potom položíme $f_2 = h_k$.

Na záver sekcie o výpočte GCD sa vráťme k otázke stability transformácie Sylvestrovej matice pri použití elementárnej dolnej trojuholníkovej matice E a matice G c-s transformácie.

3.2.3 Numerický výpočet GCD transformáciami Sylvestrovej matice

V tejto časti budeme prezentovať výsledky programovej realizácie transformácie Sylvestrovej matice popísanými metódami, pričom budeme používať značenie z predchádzajúcich sekcií (3.2.1) a (3.2.2). Problém výpočtu GCD je teda riešený metódou eliminácie pomocou elementárnych trojuholníkových matíc (ďalej len ETM) a c-s transformáciou (CST).

Príklad 5. Uvažujme polynómy stupňov 22 a 13:

$$f(x) = (x - 0.5)^5(x + 0.4)^6(x - 2)^8(x + 2)^3,$$
$$g(x) = (x - 0.5)^3(x + 0.4)^3(x - 2)^3(x + 3)^3(x - 3).$$

Vidíme, že

$$\mathrm{GCD}(f, g) = (x - 0.5)^3(x + 0.4)^3(x - 2)^3.$$

Aplikovaním oboch metód na tieto polynómy získavame riešenie, ktoré si v prípade ETM označíme GCD_{ETM} a pre ktoré platí

$$\|\mathrm{GCD} - GCD_{ETM}\|_2 = 2,60e - 11.$$

V prípade CST budeme $\mathrm{GCD}(f, g)$ značiť GCD_{CST}. Platí

$$\|\mathrm{GCD} - GCD_{CST}\|_2 = 6,87e - 11.$$

Avšak aj napriek podobným výsledkom sme pri testovaní na príkladoch došli k záveru, že CST je stabilnejšia. Totiž pri výpočtoch narážame na problém zaokrúhľovacích chýb. Pretože narozdiel od teoretických formulácií, keď v poslednom behu pri výpočte GCD správne má byť nasledujúci z polynómov rovný nule, pozri znova sekciu 3.2.1, resp. 3.2.2, v praktickom počítaní to takto nie je. Nasledujúca tabuľka 3.1 (i) ukazuje ako vyzerá posledný krok v poslednom behu v prípade použitia ETM a CST.

Vidíme, že v prípade CST je veľkosť koeficientov rádovo 10^{-13}, podstatne horšie je to v prípade ETM, kde sú koeficienty len rádu 10^{-4}.

To ale znamená, že ak chceme, aby program vrátil správny výsledok, musíme určiť, čo má "považovať za nulu". Preto sa v programoch zaviedol vstupný parameter *tol*,

(i)

	ETM	CST
x^8	$1,52e^{-4}$	$1,50e^{-13}$
x^7	$-3,82e^{-4}$	$-7,06e^{-13}$
x^6	$1,92e^{-4}$	$5,96e^{-13}$
x^5	$1,54e^{-4}$	$2,01e^{-13}$
x^4	$-1,22e^{-4}$	$-3,44e^{-13}$
x^3	$-1,34e^{-5}$	$-2,26e^{-15}$
x^2	$-2,71e^{-5}$	$6,95e^{-14}$
x^1	$2,74e^{-7}$	$-1,81e^{-15}$
x^0	$-1,42e^{-6}$	$-4,47e^{-15}$

(ii)

	ETM	CST
x^8	$-3,05e^{-11}$	$2,60e^{-11}$
x^7	$1,62e^{-10}$	$-1,21e^{-10}$
x^6	$-1,56e^{-10}$	$1,14e^{-10}$
x^5	$-4,28e^{-11}$	$4,37e^{-11}$
x^4	$9,27e^{-11}$	$-5,93e^{-11}$
x^3	$2,45e^{-13}$	$-1,06e^{-12}$
x^2	$-1,76e^{-11}$	$1,31e^{-11}$
x^1	$5,32e^{-13}$	$-2,10e^{-13}$
x^0	$1,13e^{-12}$	$-6,77e^{-13}$

Tabuľka 3.1: Vplyv zaokrúhľovacích chýb na správnosť algoritmov výpočtu GCD dvoch súdeliteľných polynómov. Tabuľka obsahuje koeficienty polynómu, ktorý podľa teórie má byť nulovým polynómom, pri použití metód ETM a CST. Prípad (i) obsahuje výsledok pre nenormované polynómy, (ii) polynómy normované geometrickým priemerom.

medz, ktorou sa "ohraničuje nula". Pri CST je potrebné voliť maximálnu toleranciu $tol = 10^{-12}$ [4], pre $tol = 10^{-13}$ je výsledok chybný, ale v prípade ETM musí byť tolerancia $tol = 10^{-3}$. Vidíme teda, že transformácia CST je v tomto ako i ostatných, menej triviálnych, testovaných príkladoch stabilnejšia než ETM. Poznamenajme ešte, že prípad transformácie Sylvestrovej matice pomocou ETM je v podstate prepis Euklidovho algoritmu, ktorý, ako vieme, je nestabilný.

Samozrejme, môže nás napadnúť polynómy istým spôsobom normovať. Uvažme totiž príklad, v ktorom koeficienty polynómu f sú rádovo veľkosti 10^{10} a koeficienty polynómu g sa pohybujú v intervale $[-10, 10]$. V takom prípade číslo podmienenosti príslušnej Sylvestrovej matice je vysoké, avšak vhodným prenormovaním koeficientov polynómu f sa číslo podmienenosti a tým i náchylnosť na chyby môže podstatne znížiť. V našich programoch sme volili normovanie polynómov geometrickým priemerom koeficientov. Bližšie sa o jeho vhodnosti zmieňujeme v poslednej kapitole. Normovaním polynómov dostávame podobné výsledky u oboch transformačných metód. Zaujímavosťou je, že v prípade CST nastáva v našom príklade dokonca zhoršenie. Maximálna tolerancia je pri oboch metódach $tol = 10^{-9}$. Pozri tabuľku 3.1 (ii). Rozdiel riešenia získaného ETM od skutočného je v norme $1.25e^{-10}$, u CST $7.44e^{-11}$. □

Avšak normovaním polynómov a zavedením tolerancií nie je otázka voľby vhodného ukončovacieho kritéria uzavrená. Je to otázka, ktorú sa nám nepodarilo s úspechom vyriešiť a ktorá je predmetom ďalšieho bádania.

[4] To znamená, že čísla menšie než 10^{-12} sú považované za nulu.

Kapitola 4

STLN

V predchádzajúcej časti sme sa zaoberali výpočtom GCD. Z teórie vieme, že ak mierne porušíme koeficienty súdeliteľných polynómov, dostaneme polynómy nesúdeliteľné, pozri [9]*. Bude nás preto zaujímať nájdenie minimálnych perturbácií koeficientov nepresných polynómov tak, aby tieto polynómy, ktorých koeficienty sa upravia pričítaním získaných perturbácií, mali znova nekonštantný GCD s najväčším možným stupňom. Tento problém vyriešime metódou STLN (Structured Total Least Norm), ktorej podrobný popis možno nájsť napríklad v* [9, 10].

4.1 Metóda STLN

Cieľom poslednej kapitoly je určiť pre nepresné polynómy $f(x)$ a $g(x)$ tvaru

$$f(x) = a_0x^m + a_1x^{m-1} + \cdots + a_{m-1}x + a_m = \sum_{i=0}^{m} a_i x^{m-i},$$

$$g(x) = b_0x^n + b_1x^{n-1} + \cdots + b_{n-1}x + b_n = \sum_{i=0}^{n} b_i x^{n-i}$$

perturbácie koeficientov $\delta a = \{\delta a_i\}_{i=0}^m$, $\delta b = \{\delta b_j\}_{j=0}^n$ jednotlivých koeficientov $a = \{a_i\}_{i=0}^m$, $b = \{b_j\}_{j=0}^n$ tak, aby polynómy

$$\tilde{f}(x) = \sum_{i=0}^{m}(a_i + \delta a_i)x^{m-i}, \qquad \tilde{g}(x) = \sum_{i=0}^{n}(b_i + \delta b_i)x^{n-i}$$

mali netriviálny GCD (pozri [9, 10]). Avšak pretože polynómy f a g sú s pravdepodobnosťou takmer jeden nesúdeliteľné, zavádzame pojem aproximavaný GCD polynómov f a g (značíme AGCD = AGCD(f, g)). Platí AGCD(f, g) = GCD$(\tilde{f}, \tilde{g})$.

Od perturbácií budeme požadovať, aby spĺňali podmienku

$$\|\delta a\|_2 \leq e_f \ll \|a\|_2, \qquad \|\delta b\|_2 \leq e_g \ll \|b\|_2 \tag{4.1}$$

pre vhodne zvolené konštanty e_f a e_g. Je to prirodzená požiadavka, pretože nekontrolovaním daných perturbácií polynómy $\tilde{f}$ a $\tilde{g}$ síce budú mať netriviálny GCD, avšak

samotné polynómy $\tilde{f}$, $\tilde{g}$ budú natoľko odlišné od teoreticky presných polynómov $\hat{f}$ a $\hat{g}$, že daný výsledok bude neadekvátny (nepoužiteľný). Očakávame totiž, že $\tilde{f}$, $\tilde{g}$ budú do istej miery "legitímne" s $\hat{f}$, resp. $\hat{g}$. Pretože presné polynómy $\hat{f}$, $\hat{g}$ majú netriviálny GCD, tak perturbácie δa a δb existujú, avšak nie sú všeobecne určené jednoznačne a teda AGCD budeme považovať za platný, ak perturbácie budú splňovať vyššie uvedenú podmienku (4.1) [1].

Dokonca nemusí platiť rovnosť $\deg \mathrm{AGCD}(f,g) = \deg \mathrm{GCD}(\hat{f},\hat{g})$. Preto budeme hľadať také perturbácie, pre ktoré má AGCD najväčší možný stupeň.

Výpočet AGCD dvoch nepresných polynómov f, g pomocou mnohých metód prebieha v dvoch krokoch:

- Určenie stupňa AGCD a perturbácií δa, δb.
- Vlastný výpočet AGCD.

Výpočtom GCD dvoch polynómov sme sa zaoberali v predchádzajúcej kapitole, preto sa ďalej sústredíme na prvý bod. Dokonca z pomedzi všetkých perturbácií splňujúcich (4.1) budeme hľadať takú perturbáciu, ktorá je v euklidovskej norme minimálna.

To znamená, že posledný problém, ktorým sa budeme zaoberať je porušiť nesúdeliteľné polynómy v minimálnom množstve tak, aby GCD porušených polynómov bol netriviálny s najväčším možným stupňom. Stupeň AGCD a perturbácie δa, δb umožňuje spočítať metóda STLN, ktorá spočíva v určení aproximácie $S(\tilde{f},\tilde{g})$ matice $S(f,g)$ s rovnakou štruktúrou, pričom matica $S(\tilde{f},\tilde{g})$ je singulárna (v angličtine sa užíva termín "Structured low rank approximation of the Sylvester matrix"). Stupeň $\mathrm{AGCD}(f,g)$ sa podľa vety 3.2.1 rovná strate hodnosti matice $S(\tilde{f},\tilde{g})$ [2].

Metóda STLN vychádza z vety, ktorá dáva do súvislosti existenciu spoločného deliteľa dvoch polynómov a riešenia istej sústavy rovníc. Dôkaz je možné nájsť v [9].

Veta 4.1.1. *Nech $f(x)$ a $g(x)$ sú polynómy definované v (3.1), resp. (3.2) a nech $k \leq \min\{m,n\}$, kde m je stupeň $f(x)$ a n stupeň $g(x)$, $m \geq n$. Potom nutnou a postačujúcou podmienkou k tomu, aby polynómy $f(x)$ a $g(x)$ mali spoločného deliteľa stupňa k je, že sústava*

$$A_k y = c_k \tag{4.2}$$

má práve jedno netriviálne riešenie, kde podľa (3.3) je

$$S_k(f,g) = [c_k \mid A_k]. \tag{4.3}$$

k-tý Sylvestrov subrezultant, $c_k \in \mathbb{R}^{m+n-k+1}$ a $A_k \in \mathbb{R}^{(m+n-k+1)\times(m+n-2k+1)}$.

[1] Ešte raz zdôrazníme značenie polynómov. Teoreticky presné polynómy s netriviálnym GCD značíme $\hat{f}$, resp. $\hat{g}$. Nepresné realizácie presných polynómov, ktoré sú nesúdeliteľné f, resp. g a polynómy, ktoré získame metódou STLN perturbovaním nepresných polynómov a ktoré znova majú netriviálny GCD $\tilde{f}$ a $\tilde{g}$.

[2] Problém určovania hodnosti matice je zložitý. Zo súčasných prác odkážeme na [11]. Bežný spôsob je určovanie numerickej hodnosti na základe poznania singulárnych čísel matice, avšak toto kritériu nie je spoľahlivé. Preto sa na overenie hodnosti matice uspokojíme s grafickým zobrazením singulárnych čísel.

Zaveďme vektor perturbácií $z = \{z_i\}_{i=0}^{m+n+1}$ koeficientov $f(x)$ a $g(x)$, presne buď $z_i = \delta a_i$ pre $i = 0, \ldots, m$ a $z_{m+1+i} = \delta b_i$ pre $i = 0, \ldots, n$. Ďalej definujme maticu B_k s rovnakou štruktúrou ako S_k, pričom v prvých $m - k + 1$ stĺpcoch nech sú umiestnené perturbácie koeficientov polynómu f a vo zvyšných $n - k + 1$ perturbácie koeficientov polynómu g. Maticu B_k rozdeľme rovnakým spôsobom ako S_k na vektor h_k a maticu E_k. Buď teda $B_k \in \mathbb{C}^{(m+n-k+1)\times(m+n-2k+2)}$,

$$B_k = [h_k \mid E_k] = \begin{bmatrix} z_0 & & & & z_{m+1} & & & \\ z_1 & z_0 & & & z_{m+2} & z_{m+1} & & \\ \vdots & z_1 & \ddots & & \vdots & z_{m+2} & \ddots & \\ z_{m-1} & \vdots & \ddots & z_0 & z_{m+n} & \vdots & \ddots & z_{m+1} \\ z_m & z_{m-1} & \ddots & z_1 & z_{m+n+1} & z_{m+n} & \ddots & z_{m+2} \\ & z_m & \ddots & \vdots & & z_{m+n+1} & \ddots & \vdots \\ & & \ddots & z_{m-1} & & & \ddots & z_{m+n} \\ & & & z_m & & & & z_{m+n+1} \end{bmatrix},$$

kde $h_k = h_k(z) \in \mathbb{R}^{m+n-k+1}$ je prvý stĺpec B_k a $E_k = E_k(z) \in \mathbb{R}^{(m+n-k+1)\times(m+n-2k+1)}$ pozostáva z ostatných stĺpcov B_k. Všimnime si teraz, že z tvaru koeficientov polynómov $\tilde{f}$, $\tilde{g}$ môžeme maticu $S_k(\tilde{f}, \tilde{g})$ písať v tvare

$$S_k(\tilde{f}, \tilde{g}) = S_k(f, g) + B_k.$$

Z toho plynie, že k určeniu hľadanej aproximácie $S(\tilde{f}, \tilde{g})$ potrebujeme spočítať maticu B_k, v ktorej sú uložené hľadané perturbácie δa a δb.

A teda problém nájdenia takého vektora $z = [\delta a, \delta b]$, ktorý je v norme minimálny sa spolu s podmienkou existencie $\mathrm{GCD}(\tilde{f}, \tilde{g})$ vďaka vete 4.1.1 transformuje do podoby minimalizácie $\|z\|_2^2 := \|\delta a\|_2^2 + \|\delta b\|_2^2$ s obmedzením na existenciu riešenia príslušného systému

$$(A_k + E_k(z))y = c_k + h_k(z). \tag{4.4}$$

Nás zaujíma stupeň GCD dvoch polynómov, pričom veta 4.1.1 hovorí o stupni spoločného deliteľa. Preto sa v praktickom počítaní postupuje nasledovne: Položíme $k = n$ a pokúsime sa vyriešiť (4.4) [3]. V prípade, že existuje práve jedno riešenie, je stupeň spoločného deliteľa rovný k, avšak ten už je stupňom hľadaného GCD, pretože k je najväčšie možné. Ak riešenie neexistuje, tak to podľa vety 4.1.1 znamená, že dané polynómy nemajú spoločného deliteľa so stupňom k, a tak položíme $k := k - 1$ a celú úvahu opakujeme a to dovtedy, kým (4.4) nebude mať riešnie a $k \neq 0$. V prípade, že pre niektoré k a spočítaný vektor z existuje riešenie (4.4), prevedieme test na vlastnosť perturbácií (4.1) a až vtedy ak (4.1) platí, môžeme algoritmus ukončiť. V opačnom prípade opäť položíme $k = k - 1$ [4].

[3] Pretože predpokladáme $m \geq n$, všeobecne bude $k = min\{m, n\}$, čo je najväčší možný stupeň GCD.

[4] Z týchto úvah plynie, že skutočne môže nastať $\deg \mathrm{AGCD}(f, g) \neq \deg \mathrm{GCD}(\hat{f}, \hat{g})$. Samozrejme môžeme trvať na požiadavke rovnosti stupňov. V takom prípade je potrebné metódu STLN modifikovať zavedením ďalšieho parametra α. Viac informácií podáme v závere, poprípade odkážeme na [10].

Položme

$$r(z, y) = c_k + h_k(z) - (A_k + E_k(z))y. \tag{4.5}$$

Metóda STLN teda vedie na riešenie problému najmenších štvorcov s obmedzením na existenciu riešenia nelineárnych algebraických rovníc (4.4), tj.

$$\min_{r(z,y)=0} \|z\|,$$

avšak nelineárny LSE problém s obmedzujúcou podmienkou $r(z, y) = 0$ nevieme riešiť, nakoľko v 2. kapitole sme sa venovali riešeniu lineárnych LSE problémov [5]. To znamená, že systém $m + n - k + 1$ nelineárnych rovníc $r(z, y) = 0$ potrebujeme určitým spôsobom linearizovať, tj. iteračný algoritmus, ktorým riešime $r(z, y) = 0$ vyžaduje, aby bol linearizovaný. Preto pokračujme v úvahách.

Definujme maticu $P_k \in \mathbb{R}^{(m+n-k+1)\times(m+n+2)}$ tak, že

$$h_k = P_k z = \begin{bmatrix} I_{m+1} & 0_{m+1,n+1} \\ 0_{n-k,m+1} & 0_{n-k,n+1} \end{bmatrix}$$

a maticu $Y_k = Y_k(y) \in \mathbb{R}^{(m+n-k+1)\times(m+n+2)}$, pre ktorú platí

$$Y_k z = E_k y,$$

a teda aj pre prírastok δz platí

$$Y_k(\delta z) = (\delta E_k)y.$$

Príklad 6. Majme polynómy

$$f(x) = a_3x^3 + a_2x^2 + a_1x + a_0,$$

$$g(x) = b_3x^3 + b_2x^2 + b_1x + b_0,$$

tj. buď $m = n = 3$, $k = 2$. Potom

$$A_2 = \begin{bmatrix} 0 & b_0 & 0 \\ a_0 & b_1 & b_0 \\ a_1 & b_2 & b_1 \\ a_2 & b_3 & b_2 \\ a_3 & 0 & b_3 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ 0 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$z = [z_0, z_1, z_2, z_3, z_4, z_5, z_6, z_7]^T, \qquad \delta z = [\delta z_0, \delta z_1, \delta z_2, \delta z_3, \delta z_4, \delta z_5, \delta z_6, \delta z_7]^T$$

$$E_2(z) = \begin{bmatrix} 0 & z_4 & 0 \\ z_0 & z_5 & z_4 \\ z_1 & z_6 & z_5 \\ z_2 & z_7 & z_6 \\ z_3 & 0 & z_7 \end{bmatrix}, \qquad h_2(z) = \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \\ 0 \end{bmatrix}, \qquad \delta E_2(z) = \begin{bmatrix} 0 & \delta z_4 & 0 \\ \delta z_0 & \delta z_5 & \delta z_4 \\ \delta z_1 & \delta z_6 & \delta z_5 \\ \delta z_2 & \delta z_7 & \delta z_6 \\ \delta z_3 & 0 & \delta z_7 \end{bmatrix}$$

[5] Nelinearita systému (4.4) vyplýva z požiadavky na existenciu jediného riešenia $y \in \mathbb{R}^{m+n-2k+1}$ tohto systému rovníc a z toho, že ak dvojice (z_1, y), (z_2, y) sú riešením $r(z, y) = 0$, tak sa ľahko presvedčíme, že $(z_1 + z_2, y)$ už riešením byť nemusí.

a konečne

$$Y_2(y) = \begin{bmatrix} 0 & 0 & 0 & 0 & y_2 & 0 & 0 & 0 \\ y_1 & 0 & 0 & 0 & y_3 & y_2 & 0 & 0 \\ 0 & y_1 & 0 & 0 & 0 & y_3 & y_2 & 0 \\ 0 & 0 & y_1 & 01 & 0 & 0 & y_3 & y_2 \\ 0 & 0 & 0 & y_1 & 0 & 0 & 0 & y_3 \end{bmatrix}.$$

□

Ľahko sa teraz nahliadne, že (4.5) možno písať v tvare

$$r(z, y) = c_k + h_k(z) - A_k y - Y_k(y)z.$$

V každom kroku iteračného algoritmu, ktorý rieši $r(z, y) = 0$, získavame vektory δz, δy a teda aj $z + \delta z$ a $y + \delta y$ [6]. Pretože ale požadujeme, aby polynómy $f(x)$ a $g(x)$ boli porušené v minimálnom množstve, tak normy $\|\delta z\|_2$ a $\|\delta y\|_2$ musia byť malé. Preto pre linearizáciu systému zanedbáme členy druhého rádu. Potom v ľubovoľnom kroku iteračného algoritmu je

$$r(z + \delta z, y + \delta y) = c_k + h_k(z + \delta z) - A_k \cdot (y + \delta y) - Y_k(y + \delta y) \cdot (z + \delta z) =$$

$$\underbrace{c_k + h_k(z) - A_k \cdot y - Y_k(y) \cdot z}_{= r(z,y)} + \underbrace{h_k(\delta z)}_{= P_k \cdot \delta z} - A_k \cdot \delta y - Y_k(y) \cdot \delta z - \underbrace{Y_k(\delta y) \cdot z}_{= E_k(z) \cdot \delta y} - \underbrace{Y_k(\delta y) \cdot \delta z}_{= (\delta E_k) \cdot \delta y \approx 0} \approx$$

$$r(z, y) - (Y_k - P_k) \cdot \delta z - (A_k + E_k) \cdot \delta y \quad [7].$$

Teda, skúmaný problém sa nám vďaka požiadavke na lineárny LSE problém znova transformuje do podoby

$$\min_{z+\delta z} \|D(z + \delta z)\| \qquad \text{s obmedzením} \qquad r(z + \delta z, y + \delta y) = 0,$$

kde $D \in \mathbb{R}^{(m+n+2)\times(m+n+2)}$ je diagonálna matica tvaru

$$D = \begin{bmatrix} (n - k + 1)I_{m+1} & 0 \\ 0 & (m - k + 1)I_{n+1} \end{bmatrix}.$$

Tvar matice D odvôvodňujeme snahou vybalansovať výskyt elementov vektoru z v matici B_k, totiž perturbácie z_i, $i = 0, \ldots, m$ odpovedajúce koeficientom polynómu f sa v B_k vyskytujú v $n - k + 1$ stĺpcoch, podobne je to s perturbáciami odpovedajucími polynómu g. Avšak pokračujme v úprave

$$\min_{r(z+\delta z, y+\delta y)=0} \|D(z + \delta z)\| = \min_{r(z+\delta z, y+\delta y)=0} \|D\delta z + Dz\| =$$

[6] Znova pripomenieme, že vektor z obsahuje perturbácie δa a δb. Pri riešení $r(z, y) = 0$ teda získavame aj samotné perturbácie, pričom počiatočné nastavenie v iteračnom algoritme je $z = 0$. To má samozrejme najmenšiu normu spomedzi všetkých perturbácií.

[7] Len na tomto mieste pre sprehľadnenie vzťahov zavádzame operátor $\cdot$ pre násobenie, pretože napríklad zo zápisu $A_k(y + \delta y)$ nie je hneď jasné, či sa jedná o funkciu A_k premennej $y + \delta y$ alebo o súčin A_k s $y + \delta y$. Výraz $A_k \cdot (y + \delta y)$ bude značiť súčin.

$$\min_{r(z+\delta z, y+\delta y)=0} \| [D,\, 0] \begin{bmatrix} \delta z \\ \delta y \end{bmatrix} - (-Dz) \|.$$

A z predchádzajúcich úvah máme

$$0 = r(z+\delta z, y+\delta y) = r(z,y) - (Y_k - P_k)\delta z - (A_k + E_k)\delta y,$$

ekvivalentne

$$r(z,y) = (Y_k - P_k)\delta z - (A_k + E_k)\delta y = [Y_k - P_k,\, A_k + E_k] \begin{bmatrix} \delta z \\ \delta y \end{bmatrix}.$$

Ak si konečne označíme

$$\begin{aligned}
&t = r(z,y) \in \mathbb{R}^{m+n-k+1}, \\
&C = [Y_k - P_k,\, A_k + E_k] \in \mathbb{R}^{(m+n-k+1)\times(2m+2n-2k+3)}, \\
&v = \begin{bmatrix} \delta z \\ \delta y \end{bmatrix} \in \mathbb{R}^{2m+2n-2k+3}, \delta z \in \mathbb{R}^{m+n+2}, \delta y \in \mathbb{R}^{m+n-2k+1}, \\
&s = -Dz \in \mathbb{R}^{m+n+2}, \\
&E = [D,\, 0] \in \mathbb{R}^{(m+n+2)\times(2m+2n-2k+3)},
\end{aligned}$$

dostávame LSE problém

$$\min_{Cv=t} \|Ev - s\|_2,$$

ktorým sme sa zaoberali v 2. kapitole.

Už sme spomenuli, že počiatočné nastavenie z je $z = 0$. Počiatočná hodnota y v tomto algoritme je pre hodnoty $z = 0$, tj. $h_k = 0$ a $E_k = 0$ daná ako riešenie systému

$$r(0,y) = 0 \Longleftrightarrow A_k y = c_k,$$

tzn. že pre y platí

$$y = \arg\min_u \|A_k u - c_k\|_2, \tag{4.6}$$

čo je klasický LS problém rozoberaný v časti 2.1 [8].

Ukončenie iteračného algoritmu, ktorým získavame δz a δy, nastáva v prípade, ak normalizované prírastky $\frac{\|\delta z\|}{\|z\|}$ a $\frac{\|\delta y\|}{\|y\|}$ súčasne klesnú pod dané malé kladné čísla e_z, resp. e_y. Vtedy spočítané δz, resp. δy prispievajú len veľmi málo k celkovému riešeniu a teda algoritmus môže byť ukončený.

4.2 Programová realizácia metódy STLN

V tejto časti práce zhrnieme predchádzajúce výsledky a ukážeme niekoľko príkladov. Skôr než pokročíme ďalej zmienime sa o dôležitom predspracovaní daných presných polynómov. Jednak pôjde o porušenie koeficientov polynómu, ktoré do určitej miery odpovedá prirodzeným poruchám vznikajúcim napr. pri fyzikálnych meraniach. Potom sa zmienime o normovaní koeficientov spracovávaných polynómov [9].

[8] Matica A_k ma plnú stĺpcovú hodnosť, pretože podľa vety 3.2.1 pre nepresné polynómy je Sylvestrova matica regulárna.

[9] Aj v praktickom programovaní metódy STLN sme postupovali tak, že sme si presné polynómy $\hat{f}$, $\hat{g}$ najprv porušili, následne normovali.

4.2.1 Porušovanie polynómov

V snahe čo najvernejšie napodobniť získanie nepresných dat, použijeme spôsob uvedený v [9, 10]. Predne, presné polynómy $\hat{f}$ a $\hat{g}$ budeme uvažovať v tvare

$$\hat{f}(x) = \hat{a}_0 x^m + \hat{a}_1 x^{m-1} + \cdots + \hat{a}_{m-1} x + \hat{a}_m, \tag{4.7}$$

a

$$\hat{g}(x) = \hat{b}_0 x^n + \hat{b}_1 x^{n-1} + \cdots + \hat{b}_{n-1} x + \hat{b}_n. \tag{4.8}$$

Výrazom $\|\hat{f}(x)\|$ budeme značiť euklidovskú normu koeficientov polynómu $\hat{f}$, tj.

$$\|\hat{f}(x)\| = \sqrt{\hat{a}_0^2 + \hat{a}_1^2 + \cdots + \hat{a}_m^2}.$$

Buďte ešte $c_f \in \mathbb{R}^{m+1}$ a $c_g \in \mathbb{R}^{n+1}$ vektory náhodných čísel rovnomerne rozmiestnených v intervale $[-1,\ 1]$ a ϵ malé kladné číslo signalizujúce "miesto poruchy" koeficientov. Je to konštanta určujúca desatinné miesto, na ktorom budú koeficienty porušené [10]. Potom poruchu i-tého koeficientu polynómu $\hat{f}$ vyjadríme vzťahom

$$\delta\hat{a}_i = \epsilon \frac{\|\hat{f}(x)\|}{\|c_f\|} c_{f,i}\,,$$

kde $c_{f,i}$ je i-tá zložka vektora c_f, $i = 0, 1, \ldots, m$. Podobne definujeme i perturbácie koeficientov polynómu g. Celkom tak môžeme písať

$$f(x) = \hat{f}(x) + \epsilon \frac{\|\hat{f}(x)\|}{\|c_f\|} c_f = \sum_{i=0}^{m} (\hat{a}_i + \delta\hat{a}_i) x^{m-i} =: \sum_{i=0}^{m} a_i x^{m-i}$$

v súlade s (3.1) a

$$g(x) = \hat{g}(x) + \epsilon \frac{\|\hat{g}(x)\|}{\|c_g\|} c_g = \sum_{i=0}^{n} (\hat{b}_i + \delta\hat{b}_i) x^{n-i} =: \sum_{i=0}^{n} a_i x^{n-i}$$

v súlade s (3.2). V ďalšom polynómy f a g vždy budeme uvažovať v takomto tvare.

4.2.2 Normovanie polynómov geometrickým priemerom

Predtým, než na dané polynómy bude aplikovaná metóda STLN, tieto polynómy si upravíme normovaním koeficientov geometrickým priemerom, ktorý poskytuje "lepší priemerovací koeficient". Uvažme jednoduchý príklad polynómu s koeficientami 1, 10^3 a 10^6. Normovaním pomocou $\|\cdot\|_1 \approx 10^6$, $\|\cdot\|_2 \approx 10^6$, $\|\cdot\|_\infty = 10^6$ zostanú koeficienty na podobnej úrovni, ale použitím geometrického priemeru 10^3 dostaneme čísla 10^{-3}, 1, 10^3. Normovaním koeficientov sa snažíme predísť nepríjemným numerickým problémom, ktoré sa môžu objaviť, ak napríklad koeficienty polynómu f sú podstatne

[10] V angličtine sa pre $\mu = 1/\epsilon$ používa termín *signal-to-noise ratio* - "koeficient signalizujúci poruchu".

väčšie než koeficienty polynómu g. Kvôli tomu môže narásť číslo podmienenosti matice $S(f,g)$ a tým aj vplyv zaokrúhľovacích chýb.

Polynómy $f(x)$ a $g(x)$ definované v (3.1), resp. (3.2) predefinujeme na tvar

$$f(x) = \sum_{i=0}^{m} \tilde{a}_i x^{m-i}, \qquad \tilde{a}_i = \frac{a_i}{(\prod_{k=0}^{m} |a_k|)^{\frac{1}{m+1}}}, \tag{4.9}$$

resp.

$$g(x) = \sum_{i=0}^{n} \tilde{b}_i x^{n-i}, \qquad \tilde{b}_i = \frac{b_i}{(\prod_{k=0}^{m} |b_k|)^{\frac{1}{n+1}}}, \tag{4.10}$$

4.2.3 Numerické výsledky

Na záver práce ukážeme niekoľko príkladov, na ktorých si predvedieme aplikáciu metódy STLN.

Pripomeňme, že v úvode kapitoly sme zaviedli dôležité kritérium (4.1), na základe ktorého buď akceptujeme alebo zamietame spočítané perturbácie. To znamená, že $\|z\|_2$ úzko súvisí so zavedeným koeficientom $\mu = 1/\epsilon$, ktorým si definujeme miesto poruchy koeficientov daných presných polynómov. Platí, že čím menšia je konštanta poruchy μ, tj. koeficienty sa porušujú na miestach bližšie k desatinej čiarke, tým väčšia môže byť prípustná hodnota $\|z\|$. V [10] sa pre presný polynóm $\hat{f}(x)$ definuje tzv. priestor legitímnych riešení, ktorý obsahuje všetky eventuálne poruchy získané spôsobom popísaným v časti 4.2.1 pre jedno dané μ, pre ktoré platí, že v norme sú menšie než ρ, kde

$$\rho = \frac{\|\hat{f}\|}{\mu}.$$

Podobne sa zavádza priestor legitímnych riešení pre polynóm $\hat{g}(x)$.

Preto sa zdá byť prirodzená požiadavka, aby aj získané perturbácie δa a δb metódou STLN spĺňali

$$\|\delta a\| \leq \frac{\|\hat{f}\|}{\mu}, \quad \|\delta b\| \leq \frac{\|\hat{g}\|}{\mu}.$$

V prípade, že nastávajú opačné nerovnosti nebudeme $\mathrm{GCD}(\tilde{f}, \tilde{g})$ považovať za legitímny s $\mathrm{GCD}(\hat{f}, \hat{g})$. Inak povedané, chceme, aby aj $\tilde{f}(x)$ a $\tilde{g}(x)$ patrili do priestoru legitímnych riešení príslušných $\hat{f}(x)$, resp. $\hat{g}(x)$.

Ale presné polynómy $\hat{f}(x)$, $\hat{g}(x)$ v praxi nepoznáme. Avšak ak budeme naviac predpokladať

$$\|\hat{f}\| \approx \|f\|, \quad \|\hat{g}\| \approx \|g\|$$

môžeme voliť

$$e_f = \frac{\|f\|}{\mu}, \qquad e_g = \frac{\|g\|}{\mu},$$

kde e_f, e_g sú konštanty vystupujúce v kritériu (4.1) [11].

[11] $\|\hat{f}\| \approx \|f\|$, $\|\hat{g}\| \approx \|g\|$ je možné docieliť napríklad voľbou väčšej konštanty poruchy μ.

Ďalej sa v algoritme na vstupe objavuje konštanta μ signalizujúca miesto poruchy a konštanty e_x a e_z, ktorými sa riadi iteračný algoritmus riešiaci (4.4).

Už sme sa zmienili o tom, že metódou STLN získavame perturbácie, pre ktoré všeobecne neplatí $\deg \mathrm{GCD}(\tilde{f}, \tilde{g}) = \deg \mathrm{GCD}(\hat{f}, \hat{g})$. Podľa teórie [10] však rovnosť môžeme dosiahnúť. Vychádza sa z pozorovania, že GCD polynómov $f(x)$ a $g(x)$ sa až na násobok nejakou konštantou rovná GCD polynómov $f(x)$ a $\alpha g(x)$, kde je $\alpha \neq 0$ konštanta. Ale platí $S(f, \alpha g) \neq \alpha S(f, g)$, a tak zavedením parametra α vieme metódou STLN spočítať celú množinu perturbácií δa, δb, s ktorými polynómy $\tilde{f}(x)$, $\tilde{g}(x)$ patria do legitímnych priestorov presných polynómov $\hat{f}(x)$, resp. $\hat{g}(x)$. Potom na základe dodatočného kritéria na hodnosť príslušnej Sylvestrovej matice vyberieme to α, pre ktoré je $\deg \mathrm{AGCD}(f, g) = \deg \mathrm{GCD}(\tilde{f}, \tilde{g}) = \deg \mathrm{GCD}(\hat{f}, \hat{g})$ [12]. Touto miernou modifikáciou sa viac zaoberať nebudeme, pretože ako sme už spomínali, problém určovania hodnosti matice nie je triviálny a moderná teória [11, 12] zaoberajúca sa výpočtom hodnosti Sylvestrových matíc presahuje rámec práce.

Príklad 7. Prvým z príkladov je demonštrácia metódy STLN na polynómoch stupňov 13 a 9 s konštantami $e_y = e_z = 10^{-6}$, $\mu = 10^6$. Pre presné polynómy

$$\hat{f}(x) = (x - 1,2)^4(x + 2)^5(x - 0,5)^4,$$
$$\hat{g}(x) = (x - 1,4)^2(x + 2)^3(x - 0,5)^4.$$

Je hneď vidieť, že

$$\begin{aligned}\mathrm{GCD}(\hat{f}, \hat{g}) &= (x + 2)^3(x - 0,5)^4 \\ &= x^7 + 4x^6 + 1,5x^5 + 7,5x^4 - 0,9375x^3 + 6,375x^2 - 3,25x + 0,5\end{aligned}$$

Porušením týchto polynómov spôsobom uvedeným vyššie s $\epsilon = 1/\mu$ získame polynómy $f(x)$, $g(x)$, ktoré sú nesúdeliteľné (tabuľka 4.1 (i)).

Použitím metódy STLN dostaneme polynómy $\tilde{f}(x)$ a $\tilde{g}(x)$ uvedené v tabuľke 4.1 (ii). Pre výpočet ich GCD sme použili c-s transformáciu Sylvestrovej matice [13]. V tabuľke 4.2 je pre porovnanie uvedený $\mathrm{GCD}(\hat{f}, \hat{g})$ a $\mathrm{GCD}(\tilde{f}, \tilde{g})$.

Z vety 3.2.1 vieme, že ak $\deg \mathrm{GCD}(\hat{f}, \hat{g}) = k$, tak potom $\mathrm{rank}\, S(\hat{f}, \hat{g}) = m + n - k$. V našom prípade je $m = 14$, $n = 9$ a $k = 7$. Na obrázku (4.1) sú symbolom $\Diamond$zobrazené singulárne čísla matice $S(\hat{f}, \hat{g})$ a vidíme, že je výrazný rozdiel medzi dominantnými singulárnymi číslami a singulárnymi číslami, ktoré neprispievajú do hodnosti matice a ktorých je 7. Máme na mysli rozdiel medzi singulárnymi číslami σ_{15} a σ_{16}.

V prípade, že polynómy mierne porušíme konštantou $\mu = 10^6$, tak všetky singulárne čísla, označené symbolom $\triangle$, sú považované za rovnako dôležité a Sylvestrova matica $S(f, g)$ má plnú hodnosť. Podľa vety 3.2.1 majú polynómy $f(x)$ a $g(x)$ triviálny GCD, čo ale presne odpovedá situácií, o ktorej sme sa zmieňovali a teda, že porušené polynómy sú s pravdepodobnosťou jeden nesúdeliteľné.

Aplikovaním metódy STLN na polynómy $f(x)$ a $g(x)$ získavame opravené polynómy $\tilde{f}(x)$ a $\tilde{g}(x)$. Singulárne čísla im odpovedajúcej Sylvestrovej matice sú vyznačené symbolom $\square$. Vidíme, že opäť vieme presne rozlíšiť dominantné singulárne čísla a teda aj určiť stupeň $\mathrm{GCD}(\tilde{f}, \tilde{g})$, ktorý je taktiež 7.

[12] Kritériom sa rozumie použitie vety 3.2.1.

[13] Paramter, ktorým sme "odhadli nulu" je $tol = 10^{-8}$.

Nakoniec si uveďme, že pre $\mu = 10^6$ je $e_f = \frac{\|f\|_2}{\mu} = 1,22e-5$, pričom $\|\hat{f} - \tilde{f}\|_2 = 6,60e-6$ a $e_g = \frac{\|g\|_2}{\mu} = 7,78e-6 > \|\hat{g} - \tilde{g}\|_2 = 7,58e-6$. □

(i)

	$f(x)$	$g(x)$
x^{13}	1	
x^{12}	3,20025	
x^{11}	−8,26093	
x^{10}	−26,49540	
x^9	38,00476	1
x^8	85,59627	1,199981
x^7	−121,21627	−7,739988
x^6	−109,89824	−3,859967
x^5	223,97294	23,002372
x^4	−17,51887	−5,699975
x^3	−156,15339	−22,937378
x^2	120,28351	22,094884
x^1	−36,63814	−7,769948
x^0	4,14757	0,979989

(ii)

	$\tilde{f}(x)$	$\tilde{g}(x)$
x^{13}	1	
x^{12}	3,19998	
x^{11}	−8,26007	
x^{10}	−26,49212	
x^9	38,00016	1
x^8	85,58606	1,199982
x^7	−121,20177	−7,739994
x^6	−109,88508	−3,859969
x^5	223,94605	23,002392
x^4	−17,51684	−5,699980
x^3	−156,13476	−22,937396
x^2	120,26907	22,094900
x^1	−36,63364	−7,769959
x^0	4,14719	0,979990

Tabuľka 4.1: (i) Porušené polynómy $f(x)$ a $g(x)$ odvodené od teoreticky presných polynómov $\hat{f}(x)$, resp. $\hat{g}(x)$. (ii) Výsledné polynómy $\tilde{f}(x)$ a $\tilde{g}(x)$ získané metódou STLN.

Príklad 8. Druhý z príkladov pracuje s polynómami vyšších stupňov. Buďte

$$\hat{f}(x) = (x-0.5)^5(x+0.4)^6(x-2)^8(x+2)^3,$$
$$\hat{g}(x) = (x-0.5)^3(x+0.4)^3(x-2)^3(x+3)^3(x-3)$$

polynómy stupňov 22 a 13, $\deg \mathrm{GCD}(\hat{f}, \hat{g}) = 9$

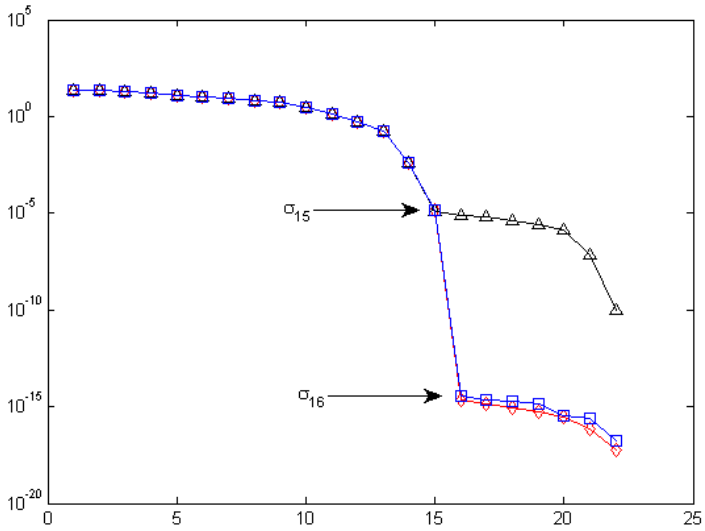
Ďalej zvoľme konštantu porušenia $\mu = 10^4$ a konštatny $e_y = 10^{-6}$, $e_z = 10^{-8}$. V tomto príklade nebudeme uvádzať koeficienty polynómov, s ktorým metóda STLN pracuje [14]. Poznamenajme ale, že výsledné polynómy $\tilde{f}(x)$ a $\tilde{g}(x)$, ktoré sme získali aplikáciou metódy STLN na pevne porušené polynómy $f(x)$, $g(x)$, majú netriviálny GCD stupňa 7.

Obrázok 4.2 znázorňuje singulárne čísla Sylvestrových matíc presných polynómov ($\Diamond$), porušených polynómov ($\triangle$) a výsledných polynómov získaných metódou STLN ($\square$). Všimnime si, že skutočne počet nedominantných singulárnych čísel u Sylvestrovej matice polynómov $\tilde{f}(x)$, $\tilde{g}(x)$ je 7. □

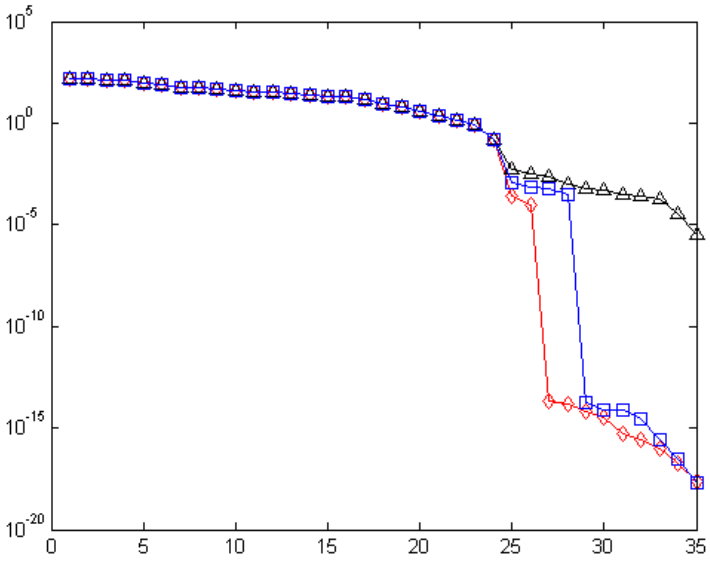
[14] Tie sú uložené v priloženom datovom súbore *STLN_pr.8.mat*. Data z predchádzajúceho príkladu 7 sú uložené v súbore *STLN_pr.7.mat*.

	$\mathrm{GCD}(\hat{f},\hat{g})$	$\mathrm{GCD}(\tilde{f},\tilde{g})$
x^7	1	1
x^6	4	$3,999978$
x^5	1.5	$1,499947$
x^4	$-7,5$	$-7,500006$
x^3	$-0,9375$	$-0,937463$
x^2	$6,375$	$6,375001$
x^1	$-3,25$	$-3,250011$
x^0	$0,5$	$0,499999$

Tabuľka 4.2: Porovnanie GCD teoreticky presných polynómov $\hat{f}(x)$ a $\hat{g}(x)$ a GCD polynómov $\tilde{f}(x)$ a $\tilde{g}(x)$ získaných metódou STLN.



Obrázok 4.1: Singulárne čísla Sylvestrových matíc sú pre pôvodné teoreticky presné polynómy $\hat{f}(x)$ a $\hat{g}(x)$ označené $\Diamond$, nepresné polynómy $f(x)$ a $g(x)$ označené $\triangle$ a "opravené" polynómy metódou STLN $\tilde{f}(x)$ a $\tilde{g}(x)$ označené $\square$. Všetky polynómy sú normované geometrickým priemerom.



Obrázok 4.2: Singulárne čísla Sylvestrových matíc sú pre pôvodné teoreticky presné polynómy $\hat{f}(x)$ a $\hat{g}(x)$ označené $\Diamond$, nepresné polynómy $f(x)$ a $g(x)$ označené $\triangle$ a "opravené" polynómy metódou STLN $\tilde{f}(x)$ a $\tilde{g}(x)$ označené $\square$. Všetky polynómy sú normované geometrickým priemerom.

Záver

Problematika, ktorou sme sa zaoberali, je veľmi obsiahla a schováva v sebe množstvo nevyriešených problémov (výpočet numerickej hodnosti matíc, hľadanie skoku v postupnosti singulárnych čísel, eliminovanie vplyvu zaokrúhľovacích chýb a i.). Aj napriek tomu práca podáva súhrn základných súvislostí, o ktoré sa možno oprieť pri ďalšom bádaní. Problémy, ktoré sa dajú ďalej študovať, sú napríklad:

1. otázka stability riešenia LSE problému, ktorá je popísaná v [1] a ktorá vyžaduje hlbšiu analýzu,

2. numerická realizácia algoritmov na výpočet GCD

3. obsah výsledkov v násobnej aritmetike a porovnanie so získanými výsledkami,

4. použitie vety 4.1.1 pre riešenie $(A_k + E_k(z))y = c_k + h_k(z)$, vhodná numerická realizácia, problém nelinearity tejto sústavy.

Tieto problémy však presahujú rámec bakalárskej práce.

Súčasne sa nám podarilo zostrojiť software, vďaka ktorému sme mohli teoretické poznatky testovať, ale programy je potrebné ďalej rozvýjať a dopĺňať o ďalšie moderné programovacie prvky a techniky.

Literatúra

[1] J. L. Barlow a S. L. Handy. *The direct solution of weighted and equality constrained least-squares problems*, SIAM J. ScI. Stat. Comput. 9(4), 1988.

[2] Åke Bjőrk. *Comment on the iterative refinement of least squares solutions*, J. Amer.Statist.Assoc., 73, s. 161– 166, 1978.

[3] Åke Bjőrk. *Numerical methods for least squares problems*, SIAM Society for industrial and Applied Mathematics, Philadelphia, 1996.

[4] A. J. Cox a N. J. Higham. *Row-wise backward stable elimination methods for the equality constrained least squares problem*, SIAM J. Matrix Anal. Appl., Vol. 21, No. 1, s. 313 – 326, 1999.

[5] G. H. Golub a Ch. Van Loan. *Matrix computations, third edition*, The Johns Hopkins University Press, Baltimore, 1996.

[6] K. Najzar a J. Zítko. *Numerické metody funkcionální analýzy*, SPN, Praha, 1984.

[7] J. Ortega a W. R. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, London, 1970.

[8] C. Van Loan. *On the method of weighting for equality-constrained least-squares problems*, SIAM Journal on Numerical Analysis, 851– 864, 1985.

[9] J. R. Winkler. *Polynomial roots and approximate greatest common divisors*, Lectured notes for a Summer School at The University of Oxford, England, 2007.

[10] J. R. Winkler a J. D. Allan. *Structured total least norm and approximate GCDs of inexact polynomials*, Journal of Computational and Applied Mathematics, 215:1– 13, 2008.

[11] J. R. Winkler a M. Hasan. *A non-linear structure preserving matrix method for the low rank approximation of the Sylvester resultant matrix*, preprint, 2009.

[12] J. R. Winkler a X. Y. Lao. *The calculation of the degree of an approximate greatest common divisor of two polynomials*, preprint, 2009.

[13] J. R. Winkler a J. Zítko. *The transformation of the Sylvester matrix and the calculation of the GCD of two polynomials*, preprint, 2009.