Fakulta elektrotechniky a informatiky
Vysoká škola báňská — Technická univerzita Ostrava, Ostrava-Poruba

Fakulta přírodovědně-humanitní a pedagogická
Technická univerzita v Liberci, Liberec

Ústav informatiky AV ČR, v. v. i.
Akademie věd České republiky, Praha

Maticové a tenzorové výpočty. Analýza a aplikace

Martin Plešinger

Habilitační práce, Září 2016

Habilitační práce k získání titulu docent v oboru Aplikovaná matematika
Maticové a tenzorové výpočty. Analýza a aplikace

Autor:	Martin Plešinger
	(martin.plesinger@tul.cz, martin.plesinger@cs.cas.cz)
Obor:	Aplikovaná matematika
Habilitační řízení:	Fakulta elektrotechniky a informatiky,
	Katedra aplikované matematiky
	Vysoká škola báňská — Technická univerzita Ostrava,
	17. listopadu 15/2172, 708 33 Ostrava-Poruba
Pracoviště autora:	Fakulta přírodovědně-humanitní a pedagogická,
	Katedra matematiky a didaktiky matematiky,
	Technická univerzita v Liberci,
	Studentská 2, 461 17 Liberec 1
a	Ústav informatiky AV ČR, v. v. i.,
	Oddělení výpočetních metod,
	Akademie věd České republiky,
	Pod Vodárenskou věží 2, 182 07 Praha 8

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$—LaTeX.

Я знаю, как управлять Вселенной.
И скажите — зачем же мне бежать за миллионом?!

I know how to control the Universe.
So tell me — why should I run for a million?!

Já vím, jak ovládat Vesmír.
Tak mi řekněte — proč bych se měl hnát za milionem?!

Grigorij Perelman
rozhovor pro deník Komsomolskaja Pravda, 28. dubna 2011
http://www.kp.ru/daily/25677.3/836229

Poděkování

Rád bych na tomto místě poděkoval všem, díky kterým jsem mohl sepsat tento stručný textík. Určitě mezi ně patří ti, z jejichž bohatých zkušeností jsem mohl čerpat a učit se ať už matematice a oborům s ní blízce příbuzným, tak i odborné práci, a bez jejichž zápalu bych sám obtížněji hledal cestu ke svému současném povolání a zalíbení v matematice obecně. Rád bych na tomto místě jmenovitě poděkoval v první řadě Zdeňku Strakošovi, svému někdejšímu školiteli v doktorském studiu; Danielu Kressnerovi a Ivo Markovi, kteří postupně zastřešovali dvě postdoktorská místa, o která jsem se směl úspěšne ucházet, a to nejprve na Švýcarském federálním technologickém institutu (Eidgenösische technische Hochschule) v Curychu a později na Technické univerzitě v Liberci; Sabine Van Huffel a Per Christianu Hansenovi za možnost strávit část svého života na jejich pracovištích na Katolické univerzitě v Belgické Lovani a na Dánské technické univerzitě v Lyngby. Určitě mezi ně ale patří i služebně starší kolegové Láďa Lukšan, Mirek Tůma, Miro Rozložník, nedávno zesnulý Miroslav Fiedler, Jörg Liesen, Gérad Meurant, které jsem měl tu čest potkat díky svému působení na Ústavu informatiky Akademie věd České republiky, v. v. i., jakož i ti, které bych si dovolil zařadit mezi své vrstevníky – Péťu Tichého, Juru Tebbense, Ivetu Hnětynkovou, Marii Michenkovou, Dianu Simu, Kristýnu Tobler, Zbyňka Koldovského, Vaška Fiňka, Danu Černou, Jirku Hozmana, Jirku Kopala a řadu dalších, na které jsem neúmyslně zapomněl. Rád bych také poděkoval kolegům z pracovišť mě osobně blízkých, i když geograficky vzdálených, konkrétně Jiřímu Bouchalovi a Zdeňku Dostálovi z Vysoké školy báňské — Technické univerzity v Ostravě, kde jsem se rozhodl tuto práci obhájit, za jejich laskavý a velmi vstřícný přístup. Ve zdaleka neposlední řadě bych rád poděkoval svým studentům, převážně budoucím učitelům matematiky, zejména Janě Žákové a své kolegyni Martině Šimůnková, která mi s nimi pomáhá, za to, že jim smím předávat úlomky toho, co se mi honí hlavou, že mne zahrnují otázkami a nutí mne přemýšlet nad věcmi, co se mi zdají jasné, z nových úhlů a v nových perspektivách. Studentům pak zejména za to, že jsou ochotni se mnou spolupracovat i na tématech odbornějších než je jen středoškolská matematika, která snad ve většině případů bude jednou jejich denním chlebem. Samozřejmě bych také rád poděkoval všem svým bývalým i současným nadřízeným a chlebodárcům za shovívavost k mým vrtochům a za to, že jsem díky jejich péči nikdy nemusel jakkoliv strádat.

Rád bych také poděkoval své rodině a přátelům. V první řadě své milované téměř-manželce Romaně Fojtové za její lásku, trpělivost a za to, že to se mnou nevzdává. Rád bych poděkoval své zesnulé babičce Haně Olze Zrůstové a svým rodičům za to, že mne vychovali k zájmu o přírodu, lidi a věci kolem nás, zvídavosti, ale i pečlivosti při práci, protože to není samozřejmé; bratrovi a jeho ženě, jakož i rodině širší a vzdálenější, biologické i té opravdové. Rád bych poděkoval svým blízkým přátelům, kteří trpěli v posledních několika letech mou přítomností, Blahošovi Fajmonovi a jeho rodině, Honzovi Hálovi a jeho čajům a mnoha dalším lidem, na které jsem si nevzpomněl, za což se upřímně omlouvám. Bez těchto lidí a bez privilegia zkřížit svou cestu životem s cestami jejich bych nikdy nemohl tuto práci odevzdat.

Abstrakt

Tato habilitační práce se zabývá maticovými a tenzorovými výpočty aplikované a numerické lineární algebry. Sestává ze čtyř kapitol věnovaných čtyřem více-či-méně samostatným tématům. První kapitola je věnována lineárním aproximačním úlohám, konkrétně tzv. problému nejmenších čtverců (TLS). Zabývá se jednak (kap. 1.2) řešitelností tohoto problému v případě, že pracujeme s úlohou s více pravými stranami, jinak řečeno s maticovou pravou stranou; a jednak (kap. 1.3) tzv. teorií core problému v lineárních aproximačních úlohách a jeho TLS řešitelností. Kapitola také stručně zmiňuje možnost rozšíření TLS problémů na lineární aproximační úlohy zformulované ve světě tenzorů.

Druhá kapitola se zabývá tzv. ill-posed problémy a regularizačními metodami, které se používají k jejich řešení. Zabýváme se zejména odhadem velikosti odstupu signálu od šumu v datech pomocí Golubovy–Kahanovy bidiagonalizace pro šumy s různou charakteristikou (barvou) a velikostí, a pro úlohy různého původu. Dále se zabýváme odhadováním tzv. filtračních faktorů regularizačních metod postavených na tzv. truncated-TLS.

Ve třetí kapitole se zabýváme tenzorovými problémy, zejména (hierarchickými) rozklady tenzorů a jejich praktickému užití v tzv. low-rank aritmetice. Ukazujeme, jak lze metodu sdružených gradientů přeformulovat pro symetrikcý pozitivně definitní operátor žijící na prostoru tenzorů. Ukazujeme, že jsou-li navíc tenzory uloženy ve tvaru např. Tuckerova, nebo hierarchického Tuckerova rozkladu (formátu), můžeme v metodě sdružených gradientů snadno uplatnit právě low-rank aritmetiku. Na závěr také ukážeme, že existence tenzorových rozkladů obsahujících cyklické součiny pomáhá motivovat lineární aproximační úlohy zformulované jazykem tenzorů zmíněných v první kapitole.

Poslední a nejkratší kapitola se zaměřuje na analýzu krylovovských metod. Stručně zmiňujeme výsledky, z nichž většina byla obsažena již v předchozích třech kapitolách: použití krylovovských metod na řešení ill-posed úloh, numerické vlastnosti krylovovských metod s low-rank aritmetikou, vztahy mezi Golubovou–Kahanovou bidiagonalizací a Lanczosovou tridiagonalizací. Navíc se zde podrobněji zabýváme blokovým, resp. pásovým zobecněním dvou posledně jmenovaných algoritmů a jejich vztahu; přesněji řečeno věnujeme se tzv. wedge-shaped (klínovým) maticím, které jsou zobecněním tridiagonálních Jacobiho matic, a jejich spektrálním vlastnostem. Všechny části textu jsou na vhodných místech doplněny odkazy na reprinty původních publikací, které jsou přiloženy.

Klíčová slova: úplný problém nejmenších čtverců (TLS); TLS algoritmus; ortogonální regrese; modelování chyb v datech; násobná pravá strana; lineární aproximační problém; unitární invariance; core problém; (zobecněná) Golubova–Kahanova bidiagonalizace; Lanczosova tridiagonalizace; (zobecněná) Jacobiho matice; wedge-shaped matice; ill-posed úlohy; image deblurring; regularizace; identifikace hladiny šumu; truncated-TLS; filtrovací faktory; redukce modelu (MOR); ljapunovské rovnice, metoda sdružených gradientů (CG); předpodmiňování; Tuckerův rozklad tenzoru; low-rank aritmetika

Použité značení a zkratky

V textu se snažíme dodržovat následující značení. Značíme

vektory	pomocí malých písmen u_1, u_2, u_r, v_1, v_2, v_r, x, atd.;
matice	pomocí velkých písmen (latinských i řeckých) A, B, C, D, E, F, U, V, Σ, atd.;
tenzory řádu k, $k \geq 3$	pomocí velkých písmen psaných kaligraficky $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{T}$, $\mathcal{S}$, atd.;
množiny	pomocí velkých písmen psaných Scriptem $\mathscr{D}$, $\mathscr{S}$, $\mathscr{T}$, $\mathscr{X}_j$, atd.

Pomocí malých písmen (latinských i řeckých) také značíme prvky matic a tenzorů a také skaláry. Speciální význam pak mají písmena i, j, ℓ, jimiž zpravidla indexujeme prvky matic a tenzorů, a k, m, n, r, která používáme k označení řádu tenzoru, dimenze matice nebo tenzoru, resp. hodnosti (ranku) matice nebo tenzoru, atp.

Číselné množiny a prostory

Značení	**Význam**
$\mathbb{R}$, $\mathbb{C}$	pole reálných, resp. komplexních čísel;
$\mathbf{i}$	imaginární jednotka komplexních čísel, $\mathbf{i}^2 = -1$;
$\mathrm{re}(\gamma)$, $\mathrm{im}(\gamma)$	reálná a imaginární část komp. čísla, $\gamma = \mathrm{re}(\gamma) + \mathbf{i}\,\mathrm{im}(\gamma)$;
$\overline{\gamma}$	komplexně sdružené číslo, $\overline{\gamma} = \mathrm{re}(\gamma) - \mathbf{i}\,\mathrm{im}(\gamma)$;
$\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$	obecné, ale pevně zvolené pole reálných, nebo komp. čísel;
$\mathbb{F}^m$	prostor sloupcových vektorů (délky m) nad polem $\mathbb{F}$;
$\mathbb{F}^{m\times n}$	prostor matic (s m řádky a n) sloupci nad polem $\mathbb{F}$;
$\mathbb{F}^{m_1\times\cdots\times m_k}$	prostor tenzorů řádu k (s vlákny délky m_j v módu j) nad $\mathbb{F}$;
0, 0_m, $0_{m,n}$, ...	nula či nulový prvek prostoru, nulový prvek $\mathbb{F}^m$, $\mathbb{F}^{m\times n}$, atd.;
$\dim(\mathscr{V})$	dimenze prostoru $\mathscr{V}$;
$\mathrm{span}\{v_1, \ldots, v_\ell\}$	lineární obal vektorů $v_1, \ldots, v_\ell$;
$\mathscr{V} \oplus \mathscr{W}$	direktní součet podprostorů $\mathscr{V}$ a $\mathscr{W}$;
$\mathscr{V} \perp \mathscr{W}$ ($\mathscr{V} \not\perp \mathscr{W}$)	podprostory $\mathscr{V}$ a $\mathscr{W}$ (ne)jsou ortogonální;
$\mathscr{K}_\ell(A, v)$	ℓ-tý krylovovský prostor, $\mathscr{K}_\ell(A, v) = \mathrm{span}\{v, Av, \ldots, A^{\ell-1}v\}$.

Vektory

Značení	**Význam**
$v = [\nu_1, \ldots, \nu_m]^\mathsf{T}$	sloupcový vektor s prvky ν_i;
$\|v\|_2$	euklidovská norma vektoru, $\|v\|_2 = (\sum_i \|\nu_i\|^2)^{\frac{1}{2}}$;
$\langle v, w\rangle$	standardní skalární součin vektorů v a w, $\langle v, w\rangle = \sum_i \overline{\omega_i}\nu_i$;
$v \perp w$ ($v \not\perp w$)	vektory v a w (ne)jsou ortogonalní, $v \perp w \iff \langle v, w\rangle = 0$;
$v \perp \mathscr{W}$ ($v \not\perp \mathscr{W}$)	vektor v je (není) ortogonální na podprostor $\mathscr{W}$;
$v\|_{\mathscr{W}}$	projekce vektoru v na podprostor $\mathscr{W}$.

Matice

Značení	Význam
$A = (a_{i,j})$	matice s prvky $a_{i,j}$;
$A = [a_1, \dots, a_n]$	matice se sloupci a_j;
A^T	tranpozice matice A;
$A^\mathsf{H} = \overline{A}^\mathsf{T} = \overline{A^\mathsf{T}}$	Hermitovské sdružení matice A;
$\mathscr{R}(A)$	obor hodnot matice A;
$\mathscr{N}(A)$	jádro (nulový prostor) matice A;
$\|A\|_2$	spektrální norma matice, $\|A\|_2 = \max_{\|v\|_2=1} \|Av\|_2$;
$\|A\|_F$	Frobeniova norma matice, $\|A\|_F = (\sum_i \sum_j \|a_{i,j}\|^2)^{\frac{1}{2}}$;
$\mathrm{rank}(A)$	hodnost matice A
$\mathrm{trace}(A)$	stopa čtvercové matice, $\sum_i a_{i,i}$;
A^{-1}	inverze čtvercové regulární matice A;
$A^\dagger$	Mooreova–Penroseova pseudoinverze matice A;
I, I_n, e_i, $\delta_{i,j}$	jednotková matice řádu n, $I = [e_1, \dots, e_n] = (\delta_{i,j})$;
$\mathrm{diag}(A, B)$	blokově diagonální matice, direktní součet matic A a B;
$A \otimes B$	Kroneckerův součin dvou matic.

Tenzory

Značení	Význam
$\mathcal{A} = (a_{i_1,i_2,i_3})$	tenzor třetího řádu o rozměrech s prvky a_{i_1,i_2,i_3};
$\mathcal{A} = (a_{i_1,\dots,i_k})$	tenzor řádu k s prvky $a_{i_1,\dots,i_k}$, $\mathcal{A} \in \mathbb{F}^{m_1 \times \cdots \times m_k}$;
$a^{(\ell)}_{\mathscr{I}} \in \mathbb{F}^{m_\ell}$	$\mathscr{I}$-té vlákno ℓ-tého módu tenzoru $\mathcal{A}$;
$\mathrm{vec}(\mathcal{A}) \in \mathbb{R}^M$	vektorizace tenzoru, $M = \prod_j m_j$;
$\mathcal{A}^{(\ell)} \in \mathbb{F}^{m_\ell \times M_\ell}$	rozvoj tenzoru do matice v ℓ-tém módu, $M_\ell = M/m_\ell$;
$\mathrm{rank}_\ell(\mathcal{A})$	ℓ-rank tenzoru, $r_\ell = \mathrm{rank}_\ell(\mathcal{A}) = \mathrm{rank}(\mathcal{A}^{(\ell)})$;
$\mathrm{rank}(\mathcal{A})$	vektorová hodnost tenzoru, $\mathrm{rank}(\mathcal{A}) = (r_1, \dots, r_k)$;
$\mathcal{A} \times_\ell W$	násobení tenzoru maticí v ℓ-tém módu, $(\mathcal{A} \times_\ell W)^{(\ell)} = W\mathcal{A}^{(\ell)}$;
$[\![\mathcal{A} \mid W_1, \dots, W_k]\!]$	násobení tenzoru maticemi W_ℓ v módech ℓ, $\ell = 1, \dots, k$;
$\mathcal{A}_{\mathsf{T-core}}$	Tuckerovo jádro tenzoru $\mathcal{A}$;
$\mathcal{A} \times \mathcal{B}$	součin (úžení) tenzorů bez udání módů násobení;
$\mathcal{A} \times_{\ell,s} \mathcal{B}$	součin (úžení) tenzorů v módech ℓ a s;
$\mathcal{A} \times_{(\ell_1,\ell_2),(s_1,s_2)} \mathcal{B}$	součin (úžení) tenzorů ve dvojici módů;
$\mathrm{diag}_k(\mathcal{A}, \mathcal{B})$	blokově diagonální tenzor, direktní tenzorů $\mathcal{A}$ a $\mathcal{B}$ řádu k.

Nejpoužívanější zkratky a akronymy

Zkratka	Význam
SVD	singulární rozklad matice (singular value decomposition);
ESVD	ekonomický singulární rozklad matice (economy-size SVD);
HOSVD	Tuckerův rozklad tenzoru (higher-order SVD);
OLS, LS	obyčejný problém nejmenších čtverců ((ordinary) least squares);
TLS	úplný problém nejmenších čtverců (total least squares);
CP	core problém;
T-SVD, T-LS	truncated-SVD, truncated-LS;
T-TLS	truncated-TLS;
SNR	odstup signálu od šumu (signal-to-noise ratio);
CG	metoda sdružených gradientů (conjugate gradients);
TT	tenzorový vláček (tensor train decomposition);
TC	tenzorový řetízek (tensor chain decomposition).

Obsah

Kapitola 1

Problém nejmenších čtverců

1.1 Úvod

Mnoho reálných problémů původem ve fyzice, statistice, atp. vede na úlohy, které lze formulovat jazykem lineární algebry. Zahrnují-li tyto problémy data, která jsou zatížená chybami jakéhokoliv původu, např. veličiny získané pomocí fyzikálních měření, nemusí být obecně možné zapsat řešenou úlohu jako rovnici, viz např. [16], [80] a [53]. Měříme-li například proud procházející rezistorem (byť ideálním s lineární charakteristikou) při různých napětích, chyby v měření způsobí, že výsledná sada dat nebude lineární rovnici splňovat. Taková úloha pak nemá řešení v klasickém slova smyslu a nazýváme ji *aproximační úlohou*. Zpravidla se pak uchylujeme k formulování *optimalizačního problému* nad danou aproximační úlohou, který nám umožní nalézt alespoň řešení přibližné. Nejjednodušší lineární aproximační úlohu můžeme zapsat ve tvaru

$$Ax \approx b, \qquad \text{kde} \quad A \in \mathbb{F}^{m\times n}, \quad x \in \mathbb{F}^{n}, \quad b \in \mathbb{F}^{m}, \tag{1.1}$$

$$\text{resp.} \qquad AX \approx B, \qquad \text{kde} \quad A \in \mathbb{F}^{m\times n}, \quad X \in \mathbb{F}^{n\times d}, \quad B \in \mathbb{F}^{m\times d} \tag{1.2}$$

a nazýváme ji úlohou s jednou, resp. více pravými stranami. V dalším textu budeme navíc zpravidla předpokládat:

(i) $b \notin \mathscr{R}(A)$, resp. $\mathscr{R}(B) \nsubseteq \mathscr{R}(A)$ protože v opačném případě by existovalo řešení v klasickém slova smyslu,

(ii) $b \not\perp \mathscr{R}(A)$, resp. $\mathscr{R}(B) \not\perp \mathscr{R}(A)$ protože v opačném případě nejsou sloupce pravé strany korelovány se sloupci systémové matice A a snaha aproximovat je pomocí lineárních kombinací sloupců systémové matice postrádá smysl; přesněji řečeno jediná rozumná lineární kombinace bude triviální.

Případné výjimky z těchto dvou pravidel budou vždy explicitně zmíněny. V dalším textu se budeme zaměřovat převážně na úlohu s více pravými stranami; úlohu s jednou pravou stranou (1.1) získáme triviálním izomorfismem s úlohou (1.2) pro $d = 1$.

1.1.1 Obyčejný problém nejmenších čtverců

Optimalizačních problémů nad (1.2) lze uvažovat celou řadu. Klasickým postupem je použít *metodu nejmenších čtverců*, jinými slovy, zformulovat nad úlohou *obyčejný problém nejmenších čtveců*, anglicky *ordinary least squares problem* (OLS), případně jen *least squares* (LS); viz např. klasickou učebnici [16], případně články [94], [95].

Definice 1 (Obyčejný problém nejmenších čtverců)**.** *Nechť $AX \approx B$ je lineární aproximační úloha (1.2), pak optimalizační problém*

$$\min_{G \in \mathbb{F}^{m\times d}} \|G\|_F \quad \text{tak, že} \quad \mathscr{R}(B+G) \subseteq \mathscr{R}(A), \tag{1.3}$$

nazýváme obyčejným problémem nejmenších čtverců.

Libovolnou matici $X_{\mathsf{OLS}} \in \mathbb{F}^{n\times d}$, která splňuje $AX_{\mathsf{OLS}} = B + G$ pro (libovolnou) minimalizující matici G, nazýváme řešením ve smyslu nejmenších čtverců.

Poznamenejme, že výše zmíněný obyčejný problém nejmenších čtverců má řešení pro libovolné (A, B). Minimalizující matice G je navíc vždy dána jednoznačně. Označíme-li $B = [b_1, \ldots, b_d]$ a $G = [g_1, \ldots, g_d]$, pak $g_j \in \mathscr{N}(A^{\mathsf{H}})$ je záporně vzatá projekce b_j na jádro matice A^{H} a tudíž $b_j + g_j \in \mathscr{R}(A)$ je projekce b_j na obor hodnot matice A; tj. $b_j = b_j|_{\mathscr{R}(A)} + b_j|_{\mathscr{N}(A^{\mathsf{H}})} = (b_j + g_j) - g_j$, viz [26, kap. 6].

Z (jednoznačné) existence minimalizující matice G plyne existence řešení ve smyslu nejmenších čtverců, tj. matice X_{OLS}. Toto je jednoznačné tehdy a jen tehdy, má-li matice $A \in \mathbb{F}^{m\times n}$ lineárně nezávislé sloupce. V opačném případě, tj. pokud $r \equiv \operatorname{rank}(A) < n$, pak existuje nekonečně mnoho takových řešení. Zřejmě pokud $AX_{\mathsf{OLS}} = B + G$, pak také $A(X_{\mathsf{OLS}} + Y) = B + G$ pro libovolné $Y = [y_1, \ldots, y_d]$, $y_j \in \mathscr{N}(A)$. Za účelem zjednoznačnění řešení problém nejmenších čtverců *navíc* doplňujeme požadavkem na minimalitu $X_{\mathsf{OLS}} = [x_{1,\mathsf{OLS}}, \ldots, x_{d,\mathsf{OLS}}]$ ve Frobeniově normě, tj. ekvivalentně požadavkem $x_{j,\mathsf{OLS}} \perp \mathscr{N}(A)$; viz např. [26, kap. 6] Takové řešení získáme snadno známým vztahem

$$X_{\mathsf{OLS}} = A^{\dagger}B, \tag{1.4}$$

tj. užitím Mooreovy–Penroseovy pseudoinverze matice A. Ze vztahu (1.4) je zřejmé, že v případě OLS lze úlohu $AX \approx B$ s d pravými stranami ekvivalentě nahradit d zcela nezávislými úlohami s jednou pravou stranou $Ax_j \approx b_j$, $j = 1, \ldots, d$; viz [26, kap. 6].

V praktických úlohách je možné do problému (1.3) navíc zavést vážení nebo škálování jednotlivých řádků (rovnic), resp. sloupců (pozorování) pomocí diagonálních matic s kladnými diagonálními prvky; viz [16, kap. 4.4], [37]. Obecněji řečeno, je možné uvažovat jinou než Frobeniovu normu jak při hledání minimalizujícího G, tak při hledání minimálního řešení; viz [16, kap. 4.5], [129]. Dále je možné problém doplňovat dalšími omezujícími předpoklady; viz [16, kap. 5], případně [38, kap. 5 a 6], atd.

1.1.2 Úplný problém nejmenších čtverců

Dalším z blízce příbuzných optimalizačních problémů nad (1.2) je tzv. *úplný problém nejmenších čtverců*, anglicky *total least squares* (TLS), který je ústředním bodem první části této práce.

Definice 2 (Úplný problém nejmenších čtverců)**.** *Nechť $AX \approx B$ je lineární aproximační úloha (1.2), pak optimalizační problém*

$$\min_{\substack{E\in\mathbb{F}^{m\times n}\\ G\in\mathbb{F}^{m\times d}}} \|[G, E]\|_F \quad \text{tak, že} \quad \mathscr{R}(B+G) \subseteq \mathscr{R}(A+E), \tag{1.5}$$

nazýváme úplným problémem nejmenších čtverců.

Libovolnou matici $X_{\mathsf{TLS}} \in \mathbb{F}^{n\times d}$, která splňuje $(A+E)X_{\mathsf{TLS}} = B+G$ pro libovolný minimalizující pár (E, G), nazýváme řešením ve smyslu TLS.

Analýza existence a jednoznačnosti minimalizující korekce (E, G) a řešení X_{TLS} není zdaleka tak jednoduchá, jako v případě OLS. Nyní neupravujeme jen pravou

stranu $B = [b_1, \ldots, b_d]$, ale i matici zobrazení A. Ihned si všimneme, že nyní není možné úlohu s d pravými stranami ekvivaletně nahradit d nezávislými úlohami právě proto, že všechny musí sdílet stejnou korekci E matice zobrazení. Navíc TLS řešení obecně, tj. pro libovolné (A, B), nemusí existovat, narozdíl od OLS řešení; viz [37], [15], [99, Example 1.1, str. 7], případně [126, Example 2.2, str. 38].

1.2 Řešitelnost TLS problému

Pro další výklad bude vhodné zavést následující singulární rozklady (SVD). Uvažujme pro jednoduchost $m > n + d$ (v opačném případě rozšíříme úlohu $AX \approx B$ o vhodný počet nulových řádků, tj. přejdeme k $\left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right] X \approx \left[\begin{smallmatrix} B \\ 0 \end{smallmatrix}\right]$). Nechť

$$\begin{aligned} A = U'\Sigma' V'^{\mathsf{H}}, \quad \text{kde} \quad & U'^{-1} = U'^{\mathsf{H}}, \quad U' = [u'_1, \ldots, u'_m] = (u'_{i,j}), \\ & V'^{-1} = V'^{\mathsf{H}}, \quad V' = [v'_1, \ldots, v'_n] = (v'_{i,j}), \\ \Sigma' = \begin{bmatrix} \operatorname{diag}(\sigma'_1, \ldots, \sigma'_n) \\ 0_{m-n,n} \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad & \sigma'_1 \geq \sigma'_2 \geq \cdots \geq \sigma'_n \geq 0 \end{aligned} \tag{1.6}$$

je SVD systémové matice A a

$$\begin{aligned} [B, A] = U\Sigma V^{\mathsf{H}}, \quad \text{kde} \quad & U^{-1} = U^{\mathsf{H}}, \quad U = [u_1, \ldots, u_m] = (u_{i,j}), \\ & V^{-1} = V^{\mathsf{H}}, \quad V = [v_1, \ldots, v_{n+d}] = (v_{i,j}), \\ \Sigma = \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_{n+d}) \\ 0_{m-(n+d),n+d} \end{bmatrix} \in \mathbb{R}^{m \times (n+d)}, \quad & \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n+d} \geq 0 \end{aligned} \tag{1.7}$$

je SVD rozšířené matice $[B, A]$.

1.2.1 Úlohy s jednou pravou stranou

TLS problém s jednou pravou stranou ($d = 1$) byl analyzován nejprve v článku [37], který se zabývá zejména situací, kdy má matice A lineárně nezávislé sloupce. Je zde zformulována nutná (nikoliv však postačující) podmínka existence TLS řešení. Gene H. Golub a Charles F. Van Loan ukázali, že platí

$$\sigma'_n > \sigma_{n+1} \quad \Longrightarrow \quad X_{\mathsf{TLS}} = -\begin{bmatrix} v_{2,n+1} \\ \vdots \\ v_{n+1,n+1} \end{bmatrix} (v_{1,n+1})^{-1} \tag{1.8}$$

přičemž pro minimalizující korekci platí $[G, E] = -u_{n+1}\sigma_{n+1}{v_{n+1}}^{\mathsf{H}}$; TLS řešení i minimalizující korekce jsou dány jednoznačně.

Tento výsledek byl rozšířen v knize [126], kde je zformulována nutná a postačující podmínka pro existenci TLS řešení úlohy s jednou pravou stranou. Uvažujme p, $0 \leq p \leq n$ takové, že

$$\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1}, \tag{1.9}$$

kde formálně $\sigma_0 = \infty$; tj. $(n+1) - p$ je násobnost signulárního čísla σ_{n+1}. Uvažujme dále odpovídající dělení matice V, tj.

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{matrix} \}1 \\ \}n \end{matrix}; \tag{1.10}$$
$$\underbrace{\phantom{V_{11}}}_{p} \quad \underbrace{\phantom{V_{12}}}_{(n+1)-p}$$

zřejmě libovolná lineární kombinace sloupců $v_{p+1}, \ldots, v_{n+1}$ je opět pravým singulárním vektorem matice $[B, A]$ (analogicky pro levé singulární vekotry). Podmínka (1.8) lze nyní rozšířit. Platí

$$\sigma'_p > \sigma_{p+1} = \cdots = \sigma_{n+1} \quad \Longleftrightarrow \quad X_{\mathsf{TLS}} = -V_{22}V_{12}^{\dagger}, \tag{1.11}$$

kde X_{TLS} je *TLS řešení minimální ve Frobeniově normě*; viz [126, Corollary 3.4, str. 65]. Řešení je jediným řešením tehdy a jen tehdy když $p = n$. Poznamenejme, že zde $V_{12}^{\dagger} = V_{12}^{\mathsf{H}}\|V_{12}\|_F^{-2}$, neboť matice V_{12} má jen jeden řádek. Minimalizující korekce je opět ve tvaru $[G, E] = -u\sigma_{n+1}v^{\mathsf{H}}$, kde $u = [u_{p+1}, \ldots, u_{n+1}]V_{12}^{\mathsf{H}}\|V_{12}\|_F^{-1}$ a $v = [v_{p+1}, \ldots, v_{n+1}]V_{12}^{\mathsf{H}}\|V_{12}\|_F^{-1}$ jsou odpovídající levý a pravý singulární vektor matice $[B, A]$.

Poznamenejme také, že výše zmíněné TLS řešení minimální ve Frobeniově normě úlohy s jednou pravou stranou (1.11) lze (pokud existuje) vyjádřit v uzavřeném tvaru

$$X_{\mathsf{TLS}} = (A^{\mathsf{H}}A - \sigma_{n+1}I_n)^{-1}A^{\mathsf{H}}B; \tag{1.12}$$

viz [126, Corollary 3.1, p. 53].

Nutná a postačující podmínka (1.11) existence TLS řešení pro úlohy s jednou pravou stranou není příliš přehledná. Vyskytuje se v ní násobnost nejmenšího singulárního čísla rozšířené matice $[B, A]$ a řešení minimální ve Frobeniově normě. Navíc zůstává ne zcela vyjasněné, co to znamená, když TLS problém nemá pro dané (A, B) řešení. Sabine Van Huffel a Joos Vandewalle ve své knize (viz [126, kap. 3.4]) zavádějí v takovém případě tzv. *negenerické řešení*, které je však řešením jiného optimalizačního problému (s přidaným omezením; viz [63, Lemma 6.1, str. 767]), případně TLS řešením ovšem modifikované lineární aproximační úlohy; viz [63, Lemma 6.2, str. 768]. Jisté vysvětlení významu neexistence řešení v kontextu původních dat (A, B) dává tzv. *core problém*, kterým se budeme zabývat v kapitole 1.3. Nyní se podíváme, jak se situace změní v případě více pravých stran.

1.2.2 Úlohy s více pravými stranami

TLS problém s více pravými stranami ($d > 1$) byl analyzován zejména v knize [126], ale také v článcích [130], [131]. Analýza prezentovaná v těchto pracech podstatně sleduje postup použitý Golubem a Van Loanem v původním článku [37]. Postup je zobecněný a formulovaný jako tzv. *klasický TLS algoritmus*; viz [126, kap. 3.6, Algorithm 3.1, str. 87–88], viz také [63, Algorithm 1, str. 767]. Čistě algoritmicky motivované rozšíření (viz [122], [123]) je velmi praktické, snadno pochopitelné, snadno aplikovatelné a dává velmi dobré výsledky v řadě aplikací; viz sborníky [124], [125] konferencí zaměřených na terorii a aplikace TLS problémů. Je to dáno mimo jiné tím, že v praktických výpočtech je, např. díky zaokrouhlovacím chybám, velmi obtížné, ne-li zcena nemožné, určit kupříkladu násobnost daného singulárního čísla. Klasický TLS algoritmus navíc zcela přirozeně přechází k negenerickému řešení u úloh, které TLS řešení nemají. Snadno se tak stane, že dojde k záměně optimalizační úlohy a na místo původně zamýšleného TLS problému se řeší jeho *regularizovaná* varianta *truncated-TLS* (T-TLS), viz [63, kap. 6] a [69]. Naším cílem však bude striktně sledovat řešitelnost TLS problému, tj. optimalizačního problému zformulovaného v definici 2.

Sabine Van Huffel a Joos Vandewalle v knize [126] dokazují, že TLS problém pro více pravých stran má (za jistých předpokladů na hodnosti vybraných bloků matice V) řešení ve dvou speciálních případech charakterizovaných dvěma speciálními distribucemi singulárních čísel matice $[B, A]$. Jsou to

$$\sigma_n > \sigma_{n+1}, \tag{1.13}$$

kdy má problém jednoznačné řešení, a

$$\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1} = \cdots = \sigma_{n+d}, \tag{1.14}$$

kdy má problém nekonečně mnoho řešení, mezi kterými lze snadno nalézt jednoznačně dané řešení s minimální Frobeniovou a zároveň s minimální spektrální normou. Dá se ukázat (viz [63, kap. 3.1 a 3.2]), že oba případy jsou přímočarým rozšířením úlohy s jednou pravou stranou dvěma různými způsoby a že oba jsou jen speciální případy obecnější třídy problémů. Náznak toho lze vysledovat již v [126, kap. 3.3.2], kde je však obecnější distribuce singulárních čísel pro jednoduchost považována za perturbaci jednoho ze dvou právě zmíněných případů. První kroky analýzy obecného případu byly provedeny již v dizertaci [99] (z roku 2008) a v krátkém příspvěvku [66] (z roku 2008). Úplná analýza existence a jednoznačnosti TLS řešení obecného problému však byla publikována až o tři roky později v našem článku [63] (v roce 2011).

Vlastní přínos

V článku [63] zavádíme, analogicky k (1.9), obecnou distribuci singulárních čísel rozšířené matice $[B, A]$. Uvažujme p $(0 \le p \le n)$ a e $(1 \le e \le d)$ taková, že

$$\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1} = \cdots = \sigma_{n+e} > \sigma_{n+e+1}, \tag{1.15}$$

kde formálně $\sigma_0 = \infty$ a $\sigma_{n+d+1} = 0$; tj. $(n+e)-p$ je násobnost signulárního čísla σ_{n+1}. Opět, analogicky k (1.10), uvažujme odpovídající dělení matice V, tj. v clanku jsou matice pojmenovane jinak

$$V = \underbrace{\left[\begin{array}{ccc} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \end{array}\right]}_{p \quad (n+e)-p \quad d-e} \begin{matrix} \}d \\ \}n \end{matrix}. \tag{1.16}$$

Vektory v prvním, druhém a třetím blokovém sloupci tak odpovídají singulárním číslům ostře větším než, rovným a ostře menším než σ_{n+1}. (Zřejmě, pro $d = 1$, tj. v případě jediné pravé strany, je $d = e = 1$ a třetí blokový sloupec zmizí.) V závislosti na tomto dělení zavádíme následující klasifikaci.

Definice 3 (Klasifikace TLS problémů)**.** *Nechť $AX \approx B$ je lineární aproximační úloha (1.2), kde rozšířená matice $[B, A]$ má SVD ve tvaru (1.7), (1.15), (1.16). Řekneme, že TLS problém nad úlohou $AX \approx B$ patří do třídy:*

- $\mathscr{F}$ *(first class) pokud* $\operatorname{rank}([V_{12}, V_{13}]) = d$*, speciálně:*
  - $\mathscr{F}_1$ *pokud* $\operatorname{rank}(V_{12}) = d - e$ *a tudíž* $\operatorname{rank}(V_{13}) = e$*;*
  - $\mathscr{F}_2$ *pokud* $\operatorname{rank}(V_{12}) > d - e$ *a zároveň* $\operatorname{rank}(V_{13}) = e$*;*
  - $\mathscr{F}_3$ *pokud* $\operatorname{rank}(V_{13}) < e$ *a tudíž* $\operatorname{rank}(V_{12}) > d - e$*;*
- $\mathscr{S}$ *(second class) pokud* $\operatorname{rank}([V_{12}, V_{13}]) < d$.

Poznamenejme, že problémy patřící do první ($\mathscr{F} = \cup_{j=1}^{3} \mathscr{F}_j$), resp. druhé ($\mathscr{S}$) třídy odpovídají tzv. generickým, resp. negenerickým problémům v klasické analýze provedené v [126]. V tomto kontextu je nutné si uvědomit, že generické problémy, tj. problémy třídy $\mathscr{F}$ v naší klasifikaci jsou obecně (v mnoha knihách a článcích) považované za problémy mající TLS řešení; viz např. schéma v [126, str. 50]. Obecná rozšířenost této domněnky je dána do značné míry tím, že práve kniha [126] je první (a popravdě řečeno jediná) kniha, která se analýze TLS problémů věnuje v dostatečné obecnosti a šíři. V článku [63] ale ukazujeme, že toto tvrzení *není pravdivé*. Konkrétně dokazujeme následující tvrzení:

$\mathscr{F}_1$: Generické problémy charakterizované distribucí singulárních čísel (1.13) nebo (1.14), tj. oba případy analyzované v knize [126], jsou speciální případy problémů třídy $\mathscr{F}_1$.

- Všechny problémy třídy $\mathscr{F}_1$ mají TLS řešení. Toto řešení je dáno jednoznačně tehdy a jen tehdy, když $p = n$. Není-li řešení jednoznačné, existuje nekonečně mnoho řešení, mezi kterými lze vybrat jedno, které je minimální ve Frobeniově a zároveň ve spektrální normě. Toto jednoznačné minimální řešení lze zapsat ve tvaru
$$X_{\mathsf{TLS}} = -[V_{22}, V_{23}][V_{12}, V_{13}]^{\dagger}. \tag{1.17}$$
- Řešení minimální v obou normách z předchozího kroku lze pro každý problém třídy $\mathscr{F}_1$ nalézt pomocí klasického TLS algoritmu.

$\mathscr{F}_2$: Všechny problémy třídy $\mathscr{F}_2$ mají TLS řešení. Toto řešení nikdy není jednoznačné, vždy jich existuje nekonečně mnoho. Mezi všemi řešeními lze nalézt řešení minimální ve Frobeniově normě i řešení minimální ve spektrální normě. Tato dvě minimální řešení nemusí být identická.

- TLS řešení žádného problému třídy $\mathscr{F}_2$ nelze nalézt pomocí klasického TLS algoritmu (přesněji řečeno varianty prezentované v [63, Algorithm 1, str. 767], která uvažuje přesnou aritmetiku a pracuje s celými bloky matice V odpovídajícími násobným singulárním číslům).

$\mathscr{F}_3 \cup \mathscr{S}$: Žádný problém tříd $\mathscr{F}_3$ a $\mathscr{S}$ nemá TLS řešení.

- Klasický TLS algoritmus aplikovaný na problém třídy $\mathscr{F}_2$, $\mathscr{F}_3$, nebo $\mathscr{S}$ vrací řešení jiného optimalizačního problému (s dodatečným omezením, viz [63, Lemma 6.1, str. 767]), případně TLS řešení modifikované úlohy (viz [63, Lemma 6.2, str. 768]), tzv. T-TLS řešení. Pouze v případě třídy $\mathscr{S}$ je toto řešení v klasické literatuře nazýváno řešením negenerickým.

Klasifikace samozřejmě zahrnuje i problémy s jednou pravou stranou. Ukazuje, že všechny problémy s jednou pravou stranou patří buď do třídy $\mathscr{F}_1$ nebo do třídy $\mathscr{S}$, jiná možnost nastat nemůže.

Z definice 3 je patrné, že klasifikace jistým způsobem odráží, jak se množství informace měřené hodnostmi matice „přelévá“ z pravého horního rohu matice V směrem do jejího levého horního rohu. V tomto kontexu je dobré si uvědomit, že se jedná o matici pravých singulárních vektorů rozšířené matice $[B, A]$ a její $d \times d$ čtvercový blok v pravém horním rohu realizuje vazbu mezi systémovou maticí A a pravou stranou B (v případě kompatibilních dat, tj. $\mathscr{R}(B) \subseteq \mathscr{R}(A)$ je tento čtvercový blok regulární). Jednotlivá výše vypsaná tvrzení tento „přesun“ informace při předchodu od problémů třídy $\mathscr{F}_1$ směrem k problémům třídy $\mathscr{S}$ přirozeně reflektují postupným zhoršováním různých aspektů řešitelnosti ve smyslu TLS. Aspekty řešitelnosti lze také přehledně zobrazit na schématu prezentovananém v [63, str. 769].

V praktických úlohách je opět možné do problému (1.5) navíc zavést vážení nebo škálování jednotlivých řádků (rovnic), resp. sloupců (pozorování) pomocí diagonálních matic s kladnými diagonálními prkvy; viz [37]. Také je možné uvažovat jinou než Frobeniovu normu jak při hledání minimalizujícího páru (E, G), tak při hledání minimálního řešení; viz [129]. Poznamenejme, že právě práce [129] navazuje na naše výsledky a dovozuje zajímavé důsledky pro obecné unitárně invariantní normy.

Původní autorské články vztahující se k tomuto tématu publikované v impaktovaných časopisech jsou přiloženy, konkrétně:

[63] (2011) jako příloha A.1 na str. 51.

1.3 Teorie core problému

TLS problém (1.5) (narozdíl od OLS (1.3)) nad lineární aproximační úlohou (1.2) nemá řešení pro obecná vstupní data (A, B). Doposud jsme se zabývali pouze otázkou řešitelnosti, tedy kdy řešení neexistuje, kdy existuje a zda je jednoznačné. Zůstává však otevřená otázka, co to znamená, že řešení neexistuje, v kontextu vstupních dat. Na tuto otázku částečně odovídá tzv. *teorie core problému.*

1.3.1 Úlohy s jednou pravou stranou

Opět se nejprve podíváme na případ s jednou pravou stranou ($d = 1$). Teorii core problému zavedli a popsali Christopher C. Paige a Zdeněk Strakoš v sérii článků [94], [95] a zejména [96]. Základní idea stojí na pozorování, že norma, kterou jsme použili v TLS minimializaci (1.5), tj. Frobeniova norma, je ortogonálně (v reálném případě), resp. unitárně (v našem případě) invariantní. Místo úlohy (1.2) (připomeňme, že nyní uvažujeme $d = 1$) a optimalizačního problému (1.5), můžeme přejít k úloze

$$\widehat{A}\widehat{X} \equiv (P^{\mathsf{H}}AQ)(Q^{\mathsf{H}}X) \approx (P^{\mathsf{H}}B) \equiv \widehat{B}, \quad \text{kde} \quad P^{\mathsf{H}} = P^{-1},\ Q^{\mathsf{H}} = Q^{-1} \tag{1.18}$$

jsou unitární matice, a k optimalizačnímu problému

$$\min_{\substack{\widehat{E}\in\mathbb{F}^{m\times n}\\ \widehat{G}\in\mathbb{F}^{m\times d}}} \|[\widehat{G}, \widehat{E}]\|_F \quad \text{tak, že} \quad \mathscr{R}(\widehat{B} + \widehat{G}) \subseteq \mathscr{R}(\widehat{A} + \widehat{E}). \tag{1.19}$$

Nechť $(\widehat{E}, \widehat{G})$ je minimalizující pár problému (1.19). Označíme-li $E = P\widehat{E}Q^{\mathsf{H}}$, $G = P\widehat{G}$, pak z unitární invariance Frobeniovy normy, $\|[\widehat{E}, \widehat{G}]\|_F = \|P[\widehat{E}, \widehat{G}]\left[\begin{smallmatrix} Q^{\mathsf{H}} & 0 \\ 0 & 1 \end{smallmatrix}\right]\|_F = \|[E, G]\|_F$, plyne, že (E, G) je minimalizující pár problému (1.5) nad původní úlohou. Zacela analogicky transformujeme i TLS řešení, existuje-li, tj. $X_{\mathsf{TLS}} = Q\widehat{X}_{\mathsf{TLS}}$. Vidíme, že TLS řešitelnost je invariantní vzhledem k unitární transformaci (1.18) vstupních dat. Poznamenejme, že transformace představuje pouze změnu souřadnicového systému.

V článku [96] je pak ukázáno, že v množině všech transformovaných úloh $\widehat{A}\widehat{X} \approx \widehat{B}$ existuje podmnožina (v článku jsou ukázány dva konkrétní prvky této množiny), v níž úloha nabývá speciální tvar

$$[\widehat{B}, \widehat{A}] = P^{\mathsf{H}}[B, A]\begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \equiv \left[\begin{array}{c|cc} B_1 & A_{11} & 0 \\ 0 & 0 & A_{22} \end{array}\right], \tag{1.20}$$

přičemž matice $[B_1, A_{11}]$ má *minimální* a matice A_{22} *maximální* rozměry mezi všemi transformacemi vedoucími na stejnou blokovou strukturu.

Použijeme-li odpovídající blokové dělení vektoru $\widehat{X}$, dostaneme

$$\widehat{A}\widehat{X} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} A_{11}X_1 \\ A_{22}X_2 \end{bmatrix} \approx \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = \widehat{B} \tag{1.21}$$

a původní úloha se tak rozpadne na dvě zcela nezávislé úlohy

$$A_{11}X_1 \approx B_1 \qquad \text{a} \qquad A_{22}X_2 \approx 0, \tag{1.22}$$

přičemž druhá z nich má triviální řešení v klasickém slova smyslu, tj. $X_2 = 0$. Jediné, co zbývá vyřešit, je první úloha $A_{11}X_1 \approx B_1$, nejmenší netriviální podúloha v původních datech $AX \approx B$. S využitím minimality (a předpokladů $\mathscr{R}(B) \notin \mathscr{R}(A)$ a $\mathscr{R}(B) \not\perp \mathscr{R}(A)$) lze ukázat řadu důležitých vlastností této úlohy:

- matice $A_{11} \in \mathbb{F}^{\overline{m}\times\overline{n}}$ má lineárně nezávislé sloupce;

- matice A_{11} má jednoduchá singulární čísla;
- nechť u_j jsou levé singulární vektory A_{11}, pak $u_j^{\mathsf{H}} B_1 \neq 0$, $j = 1, \ldots, \overline{m}$;
- matice $[B_1, A_{11}] \in \mathbb{F}^{\overline{m} \times (\overline{n}+1)}$ má lineárně nezávislé řádky, tj. $\overline{m} = \overline{n} + 1$;
- matice $[B_1, A_{11}]$ má jednoduchá singulární čísla;

viz [96]. Poznamenejme, že tyto vlastnosti nejsou nezávislé, speciálně poslední dvě jsou důsledkem předchozích tří. Důležité však je, že tyto vlastnosti garantují, že nejmenší singulární číslo matice A_{11} je ostře větší než nejmenší singulární číslo matice $[B_1, A_{11}]$, tj.

$$\sigma_{\min}(A_{11}) \equiv \sigma'_{\overline{n}} > \sigma_{\overline{n}+1} \equiv \sigma_{\min}([B_1, A_{11}]). \tag{1.23}$$

Vynutí tedy splnění Golubovy–Van Loanovy podmínky (1.8) a tím i existenci jednoznačně daného TLS řešení $X_{1,\mathsf{TLS}}$.

Definice 4 (Core problém v úlohách s jednou pravou stranou)**.** *Nechť $AX \approx B$ je lineární aproximační úloha (1.2) a P, Q unitární matice realizující transformaci (1.18) takovou, že*

$$P^{\mathsf{H}}[B, A] \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \equiv \left[\begin{array}{c|cc} B_1 & A_{11} & 0 \\ 0 & 0 & A_{22} \end{array}\right],$$

kde $[B_1, A_{11}]$ má minimální rozměry mezi všemi transformacemi vedoucími na stejnou blokovou strukturu. Pak úlohu

$$A_{11} X_1 \approx B_1$$

nazýváme core problémem uvnitř původní úlohy $AX \approx B$.

Fakt, že core problém má vždy jednoznačně dané TLS řešení, tj. pro libovolná vstupní data (A, B), nám dovoluje revidovat analýzu řešitelnosti TLS problému nad původními vstupními daty. Zřejmě platí

$$\begin{aligned} \sigma'_n \equiv \sigma_{\min}(A) &= \min\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{22})\}, \\ \sigma_{n+1} \equiv \sigma_{\min}([B, A]) &= \min\{\sigma_{\min}([B_1, A_{11}]), \sigma_{\min}(A_{22})\}. \end{aligned} \tag{1.24}$$

Mohou tedy nastat následující situace (viz [63, kap. 3.1, str. 753]); připomeňme, že navíc platí (1.23):

- $\sigma_{\min}(A_{22}) > \sigma_{\min}([B_1, A_{11}])$, tj. nejmenší singulární číslo $[B, A]$ je zároveň singulárním číslem core problému $[B_{11}, A_{11}]$ a je jednoduché. Užitím nerovnosti (1.23) vidíme, že v tomto případě původní úloha splňuje Golubovu–Van Loanovu podmínku (1.8) a tudíž má původní úloha jednoznačně dané TLS řešení.
- $\sigma_{\min}(A_{22}) = \sigma_{\min}([B_1, A_{11}])$, tj. nejmenší singulární číslo $[B, A]$ je opět singulárním číslem core problému, ale zároveň i matice A_{22}, která neobsahuje žádnou smysluplnou informaci pro řešení problému. Nejmenší singulární číslo $\sigma_{\min}([B, A])$ původních dat je tedy v matici $[B, A]$ násobné. V takovém případě nemůže být splněna podmínka (1.8), singulární čísla původních dat ale splňují podmínku Van Huffelové a Wandewalleho (1.11) pro nějaké $p < n$. (Přítomnost nejmenšího singulárního čísla $[B, A]$ v bloku A_{22} způsobí jeho přítomnost také v matici A, ovšem s násobností právě o jedna menší.)

  TLS problém nad původní úlohou $AX \approx B$ má nekonečně mnoho řešení. Nalezneme-li mezi nimi řešení minimální ve Frobeniově normě, dostaneme

$$X_{\mathsf{TLS}} = Q \begin{bmatrix} X_{1,\mathsf{TLS}} \\ 0 \end{bmatrix}, \tag{1.25}$$

tj. jednoznačné TLS řešení core problému transformované do původních souřadnic.

- $\sigma_{\min}(A_{22}) < \sigma_{\min}([B_1, A_{11}])$, tj. nejmenší singulární číslo $[B, A]$ není singulárním číslem core problému, ale jen matice A_{22}, lhostejno zda jednoduché či násobné. Opět s využitím nerovnosti (1.23) vidíme, že nyní není možné splnit ani podmínku (1.11).

  TLS problém nad původní úlohou $AX \approx B$ tedy nemá řešení. Nalezneme-li však jednoznačné řešení core problému $A_{11}X_1 \approx B_1$ a provedeme transformaci (1.25), dostaneme právě v euklidovské normě minimální *negenerické řešení* původního problému; viz [126], [96].

Teorie core problému tak umožňuje jistým způsobem nahlédnout do útrob TLS problému a vysvětlit v kontextu původních dat, proč daná úloha má, či nemá TLS řešení. Je to dáno přítomností – vzhledem k řešení irelevantních a násobných kopií (matice core problému mají jednoduchá singulární čísla) dat – reprezentovaných maticí A_{22}, uvnitř matice $[B, A]$; a jejich strukturou (nejmenší singulární číslo A_{22} hraje klíčovou roli). Říkáme, že core problém obsahuje *informaci nutnou a postačující* k řešení původní úlohy.

Poznamenejme, že článek [96] odvozuje core problém dvěma způsoby, jednak pomocí SVD systémové matice A, a také pomocí tzv. Golubovy–Kahanovy bidiagonalizace; viz [35]. Alternativně lze také odvodit pomocí Lanczosovy tridiagonalizace s využitím vztahů mezi bidiagonalizaci a tridiagonalizací, jak jsme ukázali v krátkém příspěvku [65].

1.3.2 Úlohy s více pravými stranami

Překvapivě elegantní vysvětlení neexistence TLS řešení v případě úloh s jednou pravou stranou podnítilo snahu rozšířit tento koncept i na úlohy s více pravými stranami ($d > 1$). První pokusy byly provedeny v sérii přednášek Åke Björcka [11], [12], [13] a v jeho nepublikovaném rukopise [14]. V těchto pracech převažovala snaha zobecnit přístup založený na Golubově–Kahanově bidiagonalizaci. Další kroky byly provedeny v dizertačních pracech [108] (z roku 2006) a [99] (z roku 2008). Rigorozní definice core problému pro úlohy s více pravými stranami a důkaz jeho existence nicméně přichází o pět let později v článku [67] (v roce 2013).

Vlastní přínos

V článku [67] zavádíme core problém analogicky k postupu užitém v případě jedné pravé strany. Transformaci (1.18) nyní lze zobecnit na

$$\begin{gathered} \widehat{A}\widehat{X} \equiv (P^{\mathsf{H}}AQ)(Q^{\mathsf{H}}XR) \approx (P^{\mathsf{H}}BR) \equiv \widehat{B}, \\ \text{kde} \quad P^{\mathsf{H}} = P^{-1},\ Q^{\mathsf{H}} = Q^{-1},\ R^{\mathsf{H}} = R^{-1} \end{gathered} \tag{1.26}$$

jsou unitární matice. TLS problémy nad původní úlohou $AX \approx B$ a transformovanou úlohou $\widehat{A}\widehat{X} \approx \widehat{B}$ jsou opět ekvivalentní. Ukázali jsme, že pro každé (A, B) existují takové unitární matice P, Q, R, že

$$[\widehat{B}, \widehat{A}] = P^{\mathsf{H}}[B, A]\begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \equiv \left[\begin{array}{cc|cc} B_1 & 0 & A_{11} & 0 \\ 0 & 0 & 0 & A_{22} \end{array}\right], \tag{1.27}$$

kde $A_{11} \in \mathbb{F}^{\overline{m}\times\overline{n}}$ a $B_1 \in \mathbb{F}^{\overline{m}\times\overline{d}}$. Transformace je poměrně komplikovaná a využívá sekvenci několika singulárních rozkladů. V článku jsme navíc poměrně technicky komplikovaným způsobem dokázali, že dimenze $\overline{n}$, $\overline{m}$, $\overline{d}$ jsou minimální, a tedy,

že matice $[B_1, A_{11}]$ má *minimální* a matice A_{22} *maximální* rozměry mezi všemi transformacemi vedoucími na stejnou blokovou strukturu.

Použijeme-li odpovídající blokové dělení matice $\widehat{X}$, dostaneme

$$\begin{aligned}\widehat{A}\widehat{X} &= \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}X_{11} & A_{11}X_{12} \\ A_{22}X_{21} & A_{22}X_{21} \end{bmatrix} \approx \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix} = \widehat{B},\end{aligned} \tag{1.28}$$

podobně jako v (1.21) v úloze s jednou pravou stranu. Původní úloha se nyní ale rozpadne na čtyři zcela nezávislé úlohy

$$A_{11}X_{11} \approx B_1 \qquad \text{a} \qquad A_{11}X_{12} \approx 0,\ A_{22}X_{21} \approx 0,\ A_{22}X_{22} \approx 0, \tag{1.29}$$

přičemž poslední tři z nich mají triviální řešení v klasickém slova smyslu a jediné, co zbývá vyřešit, je první úloha $A_{11}X_{11} \approx B_1$. Minimalita sama o sobě je snad dostatečným ospravedlněním pro zformulování definice.

Definice 5 (Core problém v úlohách s více pravými stranami)**.** *Nechť $AX \approx B$ je lineární aproximační úloha (1.2) a P, Q, R unitární matice realizující transformaci (1.26) takovou, že*

$$P^{\mathsf{H}}[B, A]\begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \equiv \left[\begin{array}{cc|cc} B_1 & 0 & A_{11} & 0 \\ 0 & 0 & 0 & A_{22} \end{array}\right],$$

kde $[B_1, A_{11}]$ má minimální rozměry mezi všemi transformacemi vedoucími na stejnou blokovou strukturu. Pak úlohu

$$A_{11}X_{11} \approx B_1$$

nazýváme core problémem uvnitř původní úlohy $AX \approx B$.

V článku [67] pak ukazujeme, že takto zavedený core problém konzistentně rozšiřuje vlastnosti core problému s jednou pravou stranou. Konrétně:

(CP1) matice $A_{11} \in \mathbb{F}^{\overline{m}\times\overline{n}}$ má lineárně nezávislé sloupce;

(CP2) matice $B_{11} \in \mathbb{F}^{\overline{m}\times\overline{d}}$ má lineárně nezávislé sloupce;

(CP3) nechť A_{11} má k různých nenulových singulárních čísel $\sigma_j(A_{11})$ s násobnostmi r_j a $r_{k+1} \equiv \dim(\mathcal{N}(A))$;

nechť dále $U_j \in \mathbb{F}^{\overline{m}\times r_j}$ jsou matice jejichž sloupce tvoří ortonormální báze levých siguláních podprostorů matice A_{11} odpovídajících singulárním číslům $\sigma_j(A_{11})$, pro $j = 1, \ldots, k$, a jádra $\mathcal{N}(A_{11}^{\mathsf{H}})$, pro $j = k+1$;

matice $\Phi_j \equiv U_j^{\mathsf{H}}B_1 \in \mathbb{F}^{r_j\times\overline{d}}$ mají lineárně nezávislé řádky, pro $j = 1, \ldots, k+1$.

Tato trojice vlastností se ukazuje jako nejdůležitější, neboť (CP1)–(CP3) jsou nutnou a postačující podmínkou pro minimalitu rozměrů matice $[B_1, A_{11}]$. Tyto vlastnosti také dále implikují:

(CP4) matice $[B_1|A_{11}] \in \mathbb{F}^{\overline{m}\times(\overline{n}+\overline{d})}$ má lineárně nezávislé sloupce (důsledek vlastností (CP1) a (CP3); viz [68, kap. 2.1, str 420]);

(CP5) nechť $[B_1|A_{11}]$ má k různých nenulových singulárních čísel $\sigma_j([B_1|A_{11}])$ s násobnostmi ϱ_j a $\varrho_{k+1} \equiv \dim(\mathcal{N}([B_1|A_{11}]))$;

nechť dále $V_j \in \mathbb{F}^{(\overline{n}+\overline{d})\times\varrho_j}$ jsou matice jejichž sloupce tvoří ortonormální báze pravých singulárních podprostorů matice $[B_1|A_{11}]$ odpovídajících singulárním číslum $\sigma_j([B_1|A_{11}])$, pro $j = 1, \ldots, k$, a jádra $\mathcal{N}([B_1|A_{11}])$, pro $j = k+1$;

hlavní vedoucí $\overline{d}\times\varrho_j$ podmatice matic V_j mají lineárně nezávislé sloupce, pro $j = 1, \ldots, k+1$ (viz [68, Corollary 4.7 (b), str. 430–431]);

(CP6) nechť $\sigma_j(A_{11})$ jsou různá singulární čísla matice A_{11} s násobnostmi r_j, pak $r_j \leq \overline{d}$; navíc $\sum_j r_j = \overline{n}$ (důsledek vlastnosti (CP3); resp. (CP1));

(CP7) nechť $\sigma_j([B_1|A_{11}])$ jsou různá singulární čísla matice $[B_1|A_{11}]$ s násobnostmi ϱ_j, pak $\varrho_j \leq \overline{d}$; navíc $\sum_j \varrho_j = \overline{m}$ (důsledek vlastnosti (CP5); resp. (CP4)).

Porovnáním z předchozího výčtu na str. 7–8 vidíme, že první vlastnost je identická s (CP1), druhá prímo vyplývá z (CP6), třetí z (CP3), čtvrtá z (CP4), a pátá z (CP7). Naopak vlastnost (CP2) v předchozím výčtu není, ale triviálně vyplývá z faktu, že pravá strana je nenulová (tj. není ortogonální na obor hodnot A_{11}); vlastnost (CP5) v předchozím výčtu pouze chybí, neboť není až tak důležitá.

Důležité je však poznamenat, že vlastnosti (CP5) a tudíž i (CP7) se nám nepodařilo dokázat jen užitím singulárních rozkladů bloků vstupních dat a nejsou tedy dokázány v článku [67], ale až o dva roky později v článku [68].

Připomeňme, že původní článek [96] odvozuje core problém jednak pomocí singulárního rozkladu matice A, ale také pomocí Golubovy–Kahanovy bidiagonalizace, článek [65] pak pomocí Lanczosovy tridiagonalizace. Náš článek [68] pak rozšiřuje právě tento přístup, čímž navazuje na zejména na práce Åkeho Björcka [11]–[14]. Použitý algoritmus nazýváme *pásové zobecnění Golubovy–Kahanovy bidiagonalizace*[1]. Opět zde využíváme vztahu tentokrát mezi pásovým (případně blokovým) zobecněním bidiagonalizace obecné a pásovým (případně blokovým) zobecněním tridiagonalizace hermitovské matice. Z pohledu teorie core problému je článek [68] důležitý zejména proto, že explicitně dokazuje, že cesta naznačená Åke Björckem přes zobecnění bidiagonalizace ke core problému skutečně vede. Dále je teprve zde dokázáno, že obecný core problém splňuje již dříve zmíněné vlastnosti (CP5) a (CP7).

Otázka řešitelnosti samotného core problému ve smyslu TLS každopádně nebyla ani v jednom z článků [67], [68] otevřena. Vlastnosti (CP5) a (CP7) jsou však pro její otevření nezbytné. Řešitelností core problému se zabývá náš článek [62], který obsahuje tři důležité výsledky v tomto směru. První z nich zobecňuje fakt, že core problém s jednou pravou stranou má vždy jednoznačně dané TLS řešení. Víme, že každý core problém s jednou pravou stranou je problémem třídy $\mathscr{F}_1$ (problémy s jednou pravou stranou patří buď do třídy $\mathscr{F}_1$, pokud mají, nebo do třídy $\mathscr{S}$, pokud nemají TLS řešení). V článku dokazujeme tvrzení, že obecný core problém je problémem třídy $\mathscr{F}_1$ *tehdy a jen tehdy*, když má jen jediné TLS řešení; viz [62, Corolary 2.5, str. 866]. Jinými slovy, v případě core problému se už nemůže stát, že by byl třídy $\mathscr{F}_1$ a zároveň měl více než jedno řešení. Klíčovou částí důkazu je právě vlastnost (CP5).

Dále zavádíme pojem tzv. *složeného core problému*, konkrétně ukazujeme následující ekvivalenci:

$$\begin{array}{c} \text{data } \left(A_{11}^{(\alpha)}, B_1^{(\alpha)}\right) \text{ splňují (CP1)–(CP3), pro } \alpha = 1, 2, \ldots \\ \Updownarrow \\ \text{data } \Bigg(\underbrace{\begin{bmatrix} A_{11}^{(1)} & & \\ & A_{11}^{(2)} & \\ & & \ddots \end{bmatrix}}_{A_{11}}, \underbrace{\begin{bmatrix} B_1^{(1)} & & \\ & B_1^{(2)} & \\ & & \ddots \end{bmatrix}}_{B_1} \Bigg) \text{ splňují (CP1)–(CP3),} \end{array} \tag{1.30}$$

[1] Åke Björck používá název *pásový Lanczosův algorimus*, který vychází z faktu, že Golubově–Kahanově bidiagonalizaci se někdy říká Lanczosova, případně Golubova–Kahanova–Lanczosova bidiagonalizace. Zejména v prvním případě se pak snadno zamění s Lanczosovou tridiagonalizací, obzvlášť jedná-li se o pásové (případně blokové) zobecnění a místo specifikace počtu diagonál se používá jen slovo 'algorimtus'. Drobnou terminologickou poznámu k používání jmen Gena H. Goluba, Williama Kahana a Cornelia Lanczose v kontextu bidiagonalizace lze nalézt v [64, Appendix, str. 694].

přičemž je formálně přípustné, aby některé z matic $A_{11}^{(\alpha)}$ (nikoliv všechny) měly nulový počet sloupců a tedy hodnost nula. Tato ekvivalence umožňuje „uměle vytvářet“ větší core problémy z menších výše naznačeným diagonálním skládáním, komponováním. Takto složený core problém by bylo jednoduché opět rozložit na jednotlivé komponenty $(A_{11}^{(\alpha)}, B_1^{(\alpha)})$. V obecnějším případě, uvažujeme komponovaný problém navíc transformovaný tak, že

$$\begin{aligned} A_{11} &= P_1^{\mathsf{H}} \operatorname{diag}\Big(A_{11}^{(1)}, A_{11}^{(2)}, \dots\Big) Q_1 \in \mathbb{F}^{\overline{m}\times\overline{n}}, \\ B_1 &= P_1^{\mathsf{H}} \operatorname{diag}\Big(B_1^{(1)}, B_1^{(2)}, \dots\Big) R_1 \in \mathbb{F}^{\overline{m}\times\overline{d}}, \end{aligned} \tag{1.31}$$

kde $P_1^{\mathsf{H}} = P_1^{-1}$, $Q_1^{\mathsf{H}} = Q_1^{-1}$, $R_1^{\mathsf{H}} = R_1^{-1}$ jsou obecné unitární matice. Dle našeho názoru a dosavadního poznání nelze obecný komponovaný problém (1.31) algoritmicky efektivně zpětně rozložit na jednotlivé komponenty. Bylo by třeba identifikovat trojici unitárních matic, což by pravděpodobně vedlo na optimalizační problém na Liově grupě $\mathbb{SU}(\overline{m}) \times \mathbb{SU}(\overline{n}) \times \mathbb{SU}(\overline{d})$. Otázka rozložitelnosti případného core problému tak zůstává otevřená.

Lze se však vydat opačnou cestou. V článku ukazujeme, že ze dvou core problémů s jednou pravou stranou (tj. oba třídy $\mathscr{F}_1$) lze zkomponovat core problém třídy $\mathscr{F}_1$, $\mathscr{F}_2$ a $\mathscr{S}$ (viz [62, Example 4.1, 4.2, 4.3, str. 868–869]) a ze tří pak core problém třídy $\mathscr{F}_3$ (viz [62, Example 4.4, str. 869]); ukazujeme tedy, že komponované core problémy mohou náležet do jakékoliv třídy klasifikace z definice 3.

Vytváření obtížněji řešitelných core problémů pomocí skládání by mohlo motivovat domněnku, že core problém buď patří do třídy $\mathscr{F}_1$ nebo je rozložitlený. Třetím důležitým výsledkem týkajícím se řešitelnosti je vyvrácení této domněnky. V článku je prezentován protipříklad [62, kap. 4.2, Example 4.5, str. 870–871]. Ten prezentuje úlohu, která (i) patří do třídy $\mathscr{F}_2$, (ii) splňuje podmínky (CP1)–(CP3), je tedy core problémem, a zároveň (iii) se jedná o nerozložitelný problém. Nerozložitelnsot je v tomto (velmi jednoduchém) případě vyřešena vypsáním všech přípustných blokových diagonalizací (1.31) a následnou úplnou parametrizací Liovy grupy.

Na závěr v článku také ukazujeme několik důležitých pozorování. Jednak dokazujeme očekávatelný fakt, že klasický TLS algoritmus je invariantní vzhledem k redukci úlohy na core problém, tj. lhostejno zda pustím klasický TLS algorimus na původní data $[B, A]$ nebo na core problém $[B_1, A_{11}]$ uvnitř $[B, A]$. Výstup algoritmu (který ovšem nemusí mít s TLS řešením nic společného; viz [63]) je až na unitární transformaci stejný. Speciálně, pokud výstup získaný pro core problém transformujeme zpět do původních souřadnic, dostaneme výstup získaný pro původní data. Mnohem zajímavější je chování klasického TLS algoritmu vzhledem ke komponování core problémů. Zde je ukázáno, že klasický TLS algoritmus se chová k jednotlivým komponentám složeného problému různě, v závislosti na distribuci a vzájemné velikosti singulárních čísel jednotlivých komponent. Může nastat situace, kdy výstupem klasického TLS algoritmu aplikovaného na složený problém (1.30) je direktní součet výstupů získaných nezávisle pro jednotlivé komponenty. Může však nastat i situace, kdy výstup pro složený problém obsahuje pouze výstup získaný pro jedinou komponentu a výstupy pro ostatní komponenty jsou odstraněny. Tento jev je vysvětlen jako forma *vnitřní regularizace* TLS algoritmu.

Původní autorské články vztahující se k tomuto tématu publikované v impaktovaných časopisech jsou přiloženy, konkrétně:

[67] (2013) jako příloha A.2 na str. 77;

[68] (2015) jako příloha A.3 na str. 93;

[62] (2016) jako příloha A.4 na str. 111.

Výhled do budoucna

V současné době se pokoušíme celou teorii core problému dále zobecnit pro lineární aproximační úlohy pracující s tenzory. V nejobecnějším případě tedy uvažujeme místo matic A, X a B tenzory

$$\begin{aligned} \mathcal{A} &\in \mathbb{F}^{m_1 \times \cdots \times m_\mu \times n_1 \times \cdots \times n_\nu}, \\ \mathcal{X} &\in \mathbb{F}^{n_1 \times \cdots \times n_\nu \times d_1 \times \cdots \times d_\delta}, \\ \mathcal{B} &\in \mathbb{F}^{m_1 \times \cdots \times m_\mu \times d_1 \times \cdots \times d_\delta}, \end{aligned} \tag{1.32}$$

řádů $\mu + \nu$, $\nu + \delta$ a $\mu + \delta$. Místo maticové lineární aproximační úlohy $AX \approx B$ pak uvažujeme tenzorovou lineární aproximační úlohu

$$\mathcal{A} \times \mathcal{X} \equiv \Big(\sum_{j_1=1}^{n_1} \cdots \sum_{j_\nu=1}^{n_\nu} a_{i_1,\ldots,i_\mu,j_1,\ldots,j_\nu} \cdot x_{j_1,\ldots,j_\nu,k_1,\ldots,k_\delta} \Big) \approx \mathcal{B} \tag{1.33}$$

Jakkoliv se tenzorový problém zdá komplikovanější, některé jeho speciální varianty mohou pomoci vyjasnit některé otázky týkající se TLS řešitelnosti (nejen) core problémů s více pravými stranami. Toto téma je aktulálně rozpracované v rukopise [70]. Analýza je postavená na vícerozměrných analogiích maticových rozkladů, zejména na tzv. *Tuckerově rozkladu* tenzoru; jedná se o vícerozměrnou variantu singulárního rozkladu, proto je také nazýván *higher-order SVD* (HOSVD); viz např. [73, kap. 4.1]; podrobněji se budeme tenzorovým rozkladům věnovat také v kap. 3.

Kapitola 2

Regularizační metody

2.1 Úvod

Řada aplikací vede na tzv. *disrkétní ill-posed* (někdy překládáme jako *nekorektní* (viz [114]), nebo *špatně postavené*) úlohy. Frank Natterer ve své knize [88, kap. IV.1, str. 85] zavádí tento pojem s odkazem na Jacquese S. Hadamarda [44] a říká:

> Nechť $\mathscr{B}$, $\mathscr{X}$ jsou Hilbertovy prostory a nechť $\mathcal{A}$ je omezený lineární operátor z $\mathscr{X}$ do $\mathscr{B}$. Úlohu, kde je
>
> $$\text{dáno } b \in \mathscr{B} \text{ a hledáme } x \in \mathscr{X} \text{ tak, aby platilo } \mathcal{A}(x) = b, \qquad (2.1)$$
>
> nazýváme dle Hadamarda [44] *well-posed* (*korektní*, *dobře postavenou*), pokud má jednoznačné řešení pro každé $b \in \mathscr{B}$ a pokud je řešení závisí na b spojitě. V opačném případě nazýváme (2.1) ill-posed úlohou. To znamená, že pro ill-posed úlohy operátor $\mathcal{A}^{-1}$ buď neexistuje, nebo není definován na celém $\mathscr{B}$, nebo není spojitý.

Ve světě numerické a výpočetní lineární algebry se vyskytuje řada ill-posed úloh, jejichž původ je v nejrůznějších aplikacích. Typicky se jedná například o tomografické aplikace související s inverzí Radonovy transformace (viz [100], [101], [88], [54]); inverzí různých fyzikálních potenciálů a polí – tzv. *potential field inversion* (např. anomálnií magnetického a gravitačního pole země při identifikaci zrudnění v zemské kůře, viz [30]); distribuce tepla – tzv. *steady-state heat distribution* (viz [52]); rekonstrukce orientace mikrostruktury krystalů na základě Laueho difrakčních diagramů – tzv. *orientation distribution funcion* (viz [56]); v úlohách *image deblurring* (viz [57]); a mnoha dalších; viz např. [49], [51].

Ve všech výše zmíněných případech má zobrazení $\mathcal{A}$ v (2.1) „zhlazující“ povahu. Tím máme na mysli fakt, že pro obecný, třeba i nespojitý vstup x je výstup $\mathcal{A}(x)$ typicky velmi hladký – např. rozmazání ostrého (ostré hrany obsahujícího) obrazu v úloze image deblurring. Pravá strana úlohy (2.1) je tedy velmi hladká. Naším úkolem tak při řešení těchto úloh je z hladkého výstupu b zrekonstruovat nehladký, případně nespojitý vstup x. To je důvodem, proč je řešení podobných úloh přirozeně obtížné. Úlohy jsou o to obtížnější, když se na pravé straně, tj. ve výstupu – v datech získaných např. měřením, objeví navíc nějaký parazitní signál, který můžeme obecně považovat za šum.

V dalším textu tedy budem uvažovat úlohu (2.1) diskretizovanou. Budeme předpokládat, že pravá strana je poškozena a obsahuje nejen užitečná data, ale i neznámý šum

$$Ax \approx b \equiv b_{\mathsf{data}} + b_{\mathsf{noise}}, \quad \text{kde} \quad A \in \mathbb{F}^{m\times n}, \quad \text{a} \quad \|b_{\mathsf{data}}\|_2 \gg \|b_{\mathsf{noise}}\|_2. \qquad (2.2)$$

Našim úkolem je najít, resp. aproximovat vektor

$$x_{\mathsf{data}} \equiv A^{\dagger} b_{\mathsf{data}}. \tag{2.3}$$

Protože přesná pravá strana b_{data} je výstupem zhlazujícího zobrazení, platí $b_{\mathsf{data}} \in \mathscr{R}(A)$ a tedy také $Ax_{\mathsf{data}} = b_{\mathsf{data}}$.

Zřejmě $x_{\mathsf{data}} = A^{\dagger}b - A^{\dagger}b_{\mathsf{noise}}$, podmínka $\|b_{\mathsf{data}}\|_2 \gg \|b_{\mathsf{noise}}\|_2$ však neříká nic o vzorech těchto vektorů; ukazuje se, že často platí $\|x_{\mathsf{data}}\|_2 \ll \|A^{\dagger}b_{\mathsf{noise}}\|_2$, tedy užitečná data x_{data} jsou v $A^{\dagger}b$ zcela překryta invertovaným šumem, jak si ukážeme vzápětí. Toto řešení se často nazývá *naivním řešením* $x_{\mathsf{naive}} \equiv A^{\dagger}b$.

2.2 Vlastnosti ill-posed úloh

Výše popsané ill-posed úlohy, resp. matice těchto úloh mají řadu zajímavých vlastností. Uvažujme singulární rozklad matice A ve tvaru (1.6). Protože tyto matice jsou diskretizacemi zhlazujících integrálních operátorů, jejich singulární čísla i vektory vykazují specifické chování:

(i) Singulární čísla σ'_j prudce klesají s rostoucím indexem j, často bez snadno identifikovatelných skoků, které by nám mohly pomoci při určení numerické hodnosti. Navíc při zjemňování diskretizace původní úlohy se nejmenší singulární čísla rychle blíží nule; viz např. [64, Fig. 1, str. 675].

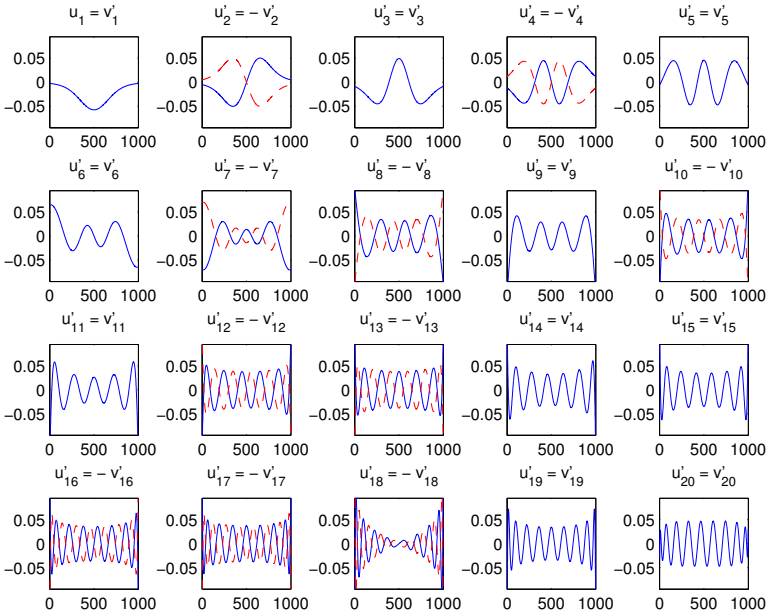
(ii) Singulární vektory u'_j a v'_j mají tendenci stále více oscilovat s rostoucím indexem j. Tím máme na mysli, že vektory odpovídající větším singulárním číslům jsou dominovány nižšími harmonickými frekvencemi, zatímco vektory odpovídající menším singulárním číslům vyššími frekvencemi. Tento jev lze u jednoduchých úloh pozorovat přímo nebo po Fourierově transformaci; viz [64, Fig. 2 a Fig. 4, str. 677–678]. U složitějších úloh je při vizualizaci potřeba zapojit geometrii původních dat; viz také obrázky 2.1 (jednorozměrná testovací úloha), 2.2 (dvourozměrná tomografická úloha) a 2.3 (dvourozměrná úloha image deblurring).

Víme, že platí $Ax_{\mathsf{data}} = b_{\mathsf{data}}$ a zároveň $x_{\mathsf{data}} = A^{\dagger}b_{\mathsf{data}}$. Rozepíšeme-li pseudoinverzi v druhém vztahu s využitím singulárního rozkladu, dostaneme

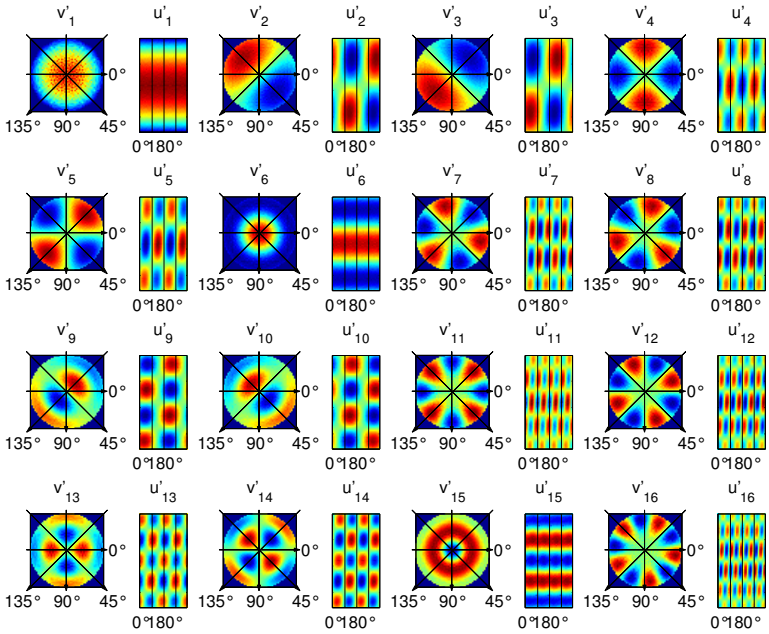
$$x_{\mathsf{data}} = A^{\dagger}b_{\mathsf{data}} = V'\Sigma'^{\dagger}U'^{\mathsf{H}}b_{\mathsf{data}} = \sum_{j=1}^{r} \frac{\langle b_{\mathsf{data}}, u'_j \rangle}{\sigma'_j} \cdot v'_j, \tag{2.4}$$

kde $r \equiv \operatorname{rank}(A)$. Vzhledem k tomu, že singulární čísla rapidně klesají k nule s rostoucím j a zároveň se zjemňováním diskretizace, a vzhledem k tomu, že vektor x_{data} představuje smysluplná data (vzor, objekt, jež byl zobrazován), nutně musí velikosti čitatelů $|\langle b_{\mathsf{data}}, u'_j \rangle|$ v zlomcích v (2.4), tj. projekce nepoškozené pravé strany do levých singulárních podprostorů matice A, klesat k nule v jistém smyslu rychleji, než singulární čísla σ'_j. V opačném případě by řada na pravé straně (2.4), která se zde objeví v limitě zjemňování diskretizace, nebyla konvergentní. Tento jev, jehož příčinou je skutečnost, že pracujeme s daty pocházejícími z reálného světa (je to tedy příčina „nematematická“), se nazývá *diskrétní Picardova podmínka*; viz např. [46], [57, kap. 5.6, str. 67–69]. Pro ilustraci splnění diskétní Picardovy podmínky viz [64, Fig. 1, str. 675].

Komponenta b_{data} má tedy v bázi $u'_1, u'_2, \ldots$ levých singulárních vektorů matice A dominantní komponenty zejména mezi prvními vektory. Tedy vektory, které jsou dominovány nízkými frekvencemi. Opět se tedy dostáváme k již dříve zmíněnému pozorování, že nepoškozená pravá strana bude typicky velmi hladká.

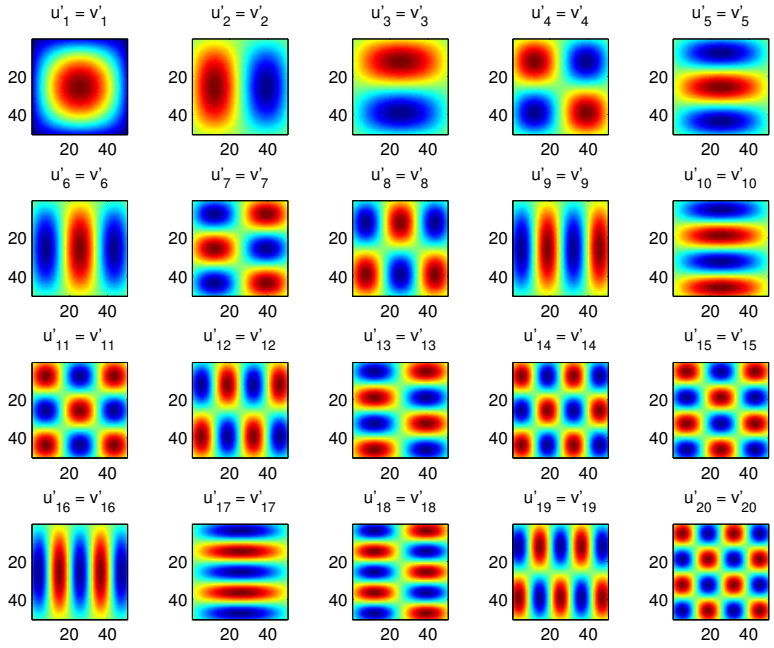


Obrázek 2.1: Prvních dvacet levých a pravých singulárních vektorů matice jednorozměrné testovací ill-posed úlohy *Shaw*; viz [107] a program `shaw.m` [48], [50].



Obrázek 2.2: Prvních šestnáct levých a pravých singulárních vektorů matice dvourozměrné tomografické úlohy; viz program `fanbeamtomo.m` [54], [55].

Oproti tomu druhá komponenta pravé strany b_{noise} obsahující šum více-či-méně neznámého původu má vztah k původní úloze, resp. k zobrazení realizovaném maticí A jen minimální. V řadě praktických úloh, kde je pravá strana $b = b_{\mathsf{data}} + b_{\mathsf{noise}}$



Obrázek 2.3: Prvních dvacet levých a pravých singulárních vektorů matice dvourozměrné úlohy image deblurring; viz např. [57].

skutečně získávána pozorováním případně měřením, může mít b_{noise} ryze náhodný charakter. Budeme-li předpokládat, že b_{noise} se chová např. podobně jako bílý šum, tj. že všechny harmonické přispívají ve Fourierově rozvoji b_{noise} zhruba stejnou měrou, můžeme očekávat, že projekce b_{noise} do levých singulární podprostorů A způsobí silné narušení diskrétní Picardovy podmínky. A skutečně se ukazuje, že hodnoty $|\langle b_{\text{noise}}, u'_j\rangle|$ jsou zhruba stejné, nezávisle na indexu j. Jejich velikost je dána velikostí tzv. *hladiny šumu*

$$\delta_{\text{noise}} = \frac{\|b_{\text{noise}}\|_2}{\|b_{\text{data}}\|_2}, \tag{2.5}$$

případně *odstupu signálu od šumu*, anglicky *signal-to-noise ratio* (SNR) $\delta_{\text{noise}}^{-1}$.

Protože $\|b_{\text{data}}\|_2 \gg \|b_{\text{noise}}\|_2$, tak pro *malé hodnoty* j jsou projekce pravé strany dominovány užitečnými daty $|\langle b, u'_j\rangle| \approx |\langle b_{\text{data}}, u'_j\rangle| \gg |\langle b_{\text{noise}}, u'_j\rangle|$. Pro *velké hodnoty* j však projekce užitečných dat klesají kvůli diskrétní Picardově podmínce, zatímco projekce šumu zůstavají; projekce pravé strany jako takové jsou pak dominovány šumem $|\langle b, u'_j\rangle| \approx |\langle b_{\text{noise}}, u'_j\rangle| \gg |\langle b_{\text{data}}, u'_j\rangle|$. Pro ilustraci tohoto jevu viz opět [64, Fig. 1, str. 675]. Díky tomu, že šum nesplňuje diskrétní Picardovu podmínku, dochází při dělení singulánírmi čísly σ'_j (viz (2.4)) k dramatickému zesílení šumu, který zcela překryje užitečná data, tj. $\|A^\dagger b_{\text{noise}}\|_2 \gg \|A^\dagger b_{\text{data}}\|_2$. Přímo spočtené naivní řešení tedy obsauje v podstatě jen zesílený šum.

2.3 Regularizace

Regularizační metody sloužící k řešení těchto úloh více-či-méně implicitně využívají právě faktu, že lze spektrálně oddělit oblasti, kde jsou v úloze dominantní data a kde šum. Mezi nejběžnější metody patří např. tzv. *truncated-SVD* (T-SVD), někdy též nazývané *truncated-LS* (T-LS), kdy je vliv nejmenších singulárních čísel omezen

de-facto tím, že matici A nahradíme její nejlepší aproximací hodnosti τ, $\tau \leq r = \operatorname{rank}(A)$ (viz [106], případně [27]), tj. odstraníme $r-\tau$ nejmenších singulárních čísel a pak teprve provedeme pseudoinverzi; viz [57, p. 11]. Druhou oblíbenou metodou je tzv. *tichonovská regularizace*, kdy úlohu (2.2)–(2.3) nahradíme optimalizačním problémem

$$x_{\mathsf{Tichonov}}(\lambda, I) \equiv \arg\min_{x\in\mathbb{F}^m}\{\|b-Ax\|_2^2 + \lambda^2\|Lx\|_2^2\}, \tag{2.6}$$

kde se první člen snaží minimalizovat reziduum a druhý penalizuje např. velikost řešení (když $L = I$), nebo velikost oscilací řešení (když je L bidiagonální s jedničkami, resp. mínus jedničkami na jedné z diagonál; tj. diferenční operátor), atp.; viz [57, p. 72], viz také [114].

Obě tyto regularizační metody, přesněji T-SVD a základní tichonovskou regularizaci s $L = I$ je možné snadno vyjádřit ve tvaru tzv. *filtrované pseudoinverze*

$$x_{\mathsf{filtered}} = \sum_{j=1}^{r} \varphi_j \cdot \frac{\langle b, u'_j\rangle}{\sigma'_j} \cdot v'_j, \tag{2.7}$$

kde φ_j jsou tzv. *filtrační faktory*; srovnej s (2.4). Platí

$$x_{\mathsf{T-SVD}}(\tau) \;=\; x_{\mathsf{filtered}} \qquad \text{pro} \qquad \varphi_j = \begin{cases} 1 & \text{když } j \leq \tau \\ 0 & \text{když } j > \tau \end{cases}, \tag{2.8}$$

$$x_{\mathsf{Tichonov}}(\lambda, I) \;=\; x_{\mathsf{filtered}} \qquad \text{pro} \qquad \varphi_j = \frac{{\sigma'_j}^2}{\lambda^2 + {\sigma'_j}^2}. \tag{2.9}$$

V obou dvou případech je zřejmě potřeba určit hodnotu nějakého regularizačního parametru. Buď τ v prvním, nebo λ v druhém případě. Existuje celá řada metod sloužících k určení těchto parametrů. Jmenujme např. tzv. *princip diskrepance* (*discrepancy principle*) V. A. Morozova (viz [86], [87]); metodu *generalized cross-validation* (GCV) G. H. Goluba (viz [34], [39], [89], [23]); kritéria postavená na *L-křivce* (*L-curve*) či *L-proužku* (*L-ribbon*) P. C. Hansena (viz [47], [20], [21], [22], viz také [80, pozn. ke kap. 25 a 26 na str. 292]); nebo metodu užívající tzv. *kumulativních periodogramů* B. W. Rusta (viz [103]); atd.

Vlastní přínos

V našem článku [64] provádíme poměrně detailní analýzu chování Golubovy–Kahanovy bidiagonalizace [35], je-li aplikovaná na matici A z ill-posed úlohy (2.2) a nastartovaná (normalizovanou) pravou stranou b. Připomeňme, že bidiagonalizace je jádrem řady Krylovovských metod, které lze též použít jako formu regularizace, zejména CGNR (CGLS) a LSQR (viz [58], [92], [93]; zkratky odpovídají matematicky ekvivalentním metodám, které se však liší v implementačních detailech), CGNE (CGME) a Craigova metoda (viz [24], [45], případně [105]), nebo nedávno publikované LSMR (viz [32]), a některé další. Provedli jsme přibližnou analýzu, jak se algoritmem bidiagonalizace šíří šum (vstupující algoritmu jako $b_{\mathsf{noise}}/\|b\|_2$) vzhledem k užitečné informaci (vstupující algoritmu jako $b_{\mathsf{data}}/\|b\|_2$); viz [64, kap. 3.2, str. 679–684].

Bidiagonalizace pracuje se dvěma sadami vektorů, ortonormálními bázemi krylovovských podprostorů

$$\mathscr{K}_\ell(AA^{\mathsf{H}}, b) \;=\; \operatorname{span}\{b, (AA^{\mathsf{H}})b, \ldots, (AA^{\mathsf{H}})^{\ell-1}b\}, \tag{2.10}$$

$$\mathscr{K}_\ell(A^{\mathsf{H}}A, A^{\mathsf{H}}b) \;=\; \operatorname{span}\{A^{\mathsf{H}}b, (A^{\mathsf{H}}A)A^{\mathsf{H}}b, \ldots, (A^{\mathsf{H}}A)^{\ell-1}A^{\mathsf{H}}b\}, \tag{2.11}$$

pro $\ell = 1, 2, \ldots$. Protože matice A zdědila zhlazující vlastnosti původního zobrazení, mají je i matice A^{H}, AA^{H} a $A^{\mathsf{H}}A$. Podíváme se nejprve na první krylovovský prostor. Pravá strana b obsauje šum, ale vektory $(AA^{\mathsf{H}})^j b$ jsou zhlazené. Při

stavbě ortonormální báze, tj. odečítáním projekcí $(AA^{\mathsf{H}})^j b$ do všech již dříve zkonstruovaných navzájem ortonormálních směrů, provádíme lineární kombinaci také s prvním bázovým vektorem $s_1 \equiv b/\|b\|_2$. Tímto mechanizmem se šum transportuje algoritmem a „kontaminuje“ postupně všechny konstruované bázové vektory s_j (které můžeme pro jednoduchost nazývat *Golubovy vektory*) tohoto prostoru. U druhého krylovovského prostoru (jehož bázi tvoří *Kahanovy vektory* w_j) tento jev nenastane, protože jej začínáme budovat s již zhlazeným vektorem $A^{\mathsf{H}}b$.

Tím, že při konstrukci bázových vektorů $s_1, \ldots, s_\ell$ ($\mathscr{K}_\ell(AA^{\mathsf{H}}, b) = \mathrm{span}\{s_1, \ldots, s_\ell\}$ pro $\ell = 1, 2, \ldots$; $\langle s_i, s_j \rangle = \delta_{i,j}$) odprojektováváme pryč část hladké komponenty, ve vektorech s_j postupně klesá SNR, tedy šum se zesiluje na úkor užitečných dat, až do okamžiku, kdy dosáhne maxima. Tento jev nazýváme *noise-revealing iterace* a příslušný vektor s_j je zcela dominován šumovou komponentou. Pro ilustraci tohoto jevu viz [64, Fig. 7, 8 a 9 (vektory s_j, jejich datové a šumové komponenty), Fig. 4 a 13 (Fourierovy koeficienty vektorů s_j v bázi levých singulárních vektorů matice A a v trigonometrické bázi)].

Navíc jsme ukázali, že jsme-li schopni identifikovat noise-revealing iteraci, pak lze z bidiagonalizačních (de-facto ortogonalizačních) koeficientů zpětně zrekonstruovat odstup signálu od šumu v původních datech, tj. určit doposud neznámou veličinu δ_{noise} (2.5). A-priorní znalost odstupu signálu od šumu je klíčová právě například při použití Morozovova principu diskrepance, ale je užitečná i při použití řady jiných kritérií používaných pro identifikaci parametrů regularizačních metod. Otázkou ale zůstává, zda lze identifikovat onu noise-revealing iteraci.

V článku [64] jsme popsali dva ukazatele, kterými se můžeme pokusit odhadovat velikost šumu ve vektorech s_j průběžně, za běhu bidiagonalizačního algoritmu. Jeden nazýváme *kumulativním poměrem* (*cumulative ratio*; viz [64, rov. (3.9) a Fig. 11]) a je postaven pouze na informaci obsažené v algoritmu bidiagonalizace, druhý je postaven na vztahu Golubovy–Kahanovy bidiagonalizace, Lanczosovy tridiagonalizace a aproximaci tzv. distribuční funkce (viz [112]) a vyžaduje navíc výpočet singulárního rozkladu vznikající bidiagonální matice; viz [64, rov. (4.6) a Fig. 14]. Navíc jsme v článku navrhli velmi primitivní a nedokonalý postup jak celý proces identifikace noise-revealing iterace zautomatizovat. Tento postup byl navržen jen pro ilustraci, pro použití numerických experimentů s jednoduchými testovacími ill-posed úlohami.

Původní autorské články vztahující se k tomuto tématu publikované v impaktovaných časopisech jsou přiloženy, konkrétně:

[64] (2009) jako příloha A.5 na str. 131.

Výhled do budoucna

Článek [64] vyvolal poměrně velký citační ohlas. My jsme na něj navázali prací na řadě složitějších experimentů. V rámci této činnosti jsme spolupracovali i s Per Christianem Hansenem na vývoji experimentálního prostředí AIR Tools pro Matlab (viz [54], [55]), které dokáže simulovat řadu jednoduchých tomografických úloh. Nedávno jsme formou konferenčního příspěvku publikovali případovou studii [59], která se zabývá úspěšností identifikace hladiny šumu v úlohách image deblurring, přičemž se soustředíme na různou velikost a barevnost šumu. Podobnou případovou studii zabývající se právě tomografickými úlohami připravujeme; zatím ale není ani ve formě rukopisu, máme jen napočítané experimenty.

Další prací, kterou jsme navázali, je analýza propagace šumu v reziduích $r_i = b - Ax_i$ iteračních metod LSQR (viz [92], [93]), Craigovy metody (viz [24]) a také právě nedávno publikované LSMR (viz [32]); analýza zatím není kompletně hotová a rukopis [60] je zatím stále rozpracovaný.

V úvodu této kapitoly jsme zmiňovali, že jednou z nejjednodušších regularizačních metod je tzv. truncated-SVD (T-SVD), které je aplikací metody nejmenších čtverců na úlohu (2.1) s modifikovanou maticí; A je nahrazeno nejlepší rank-τ aproximací matice A. Regularizace je tedy převedena na obyčejný problém nejmenších čtverců. Proto se této metodě také někdy říká truncated-LS (T-LS). Analogicky je možné provést regularizaci pomocí úplných nejmenších čtverců, konkrétně metody nazývané truncated-TLS (T-TLS). Podobně jako lze T-SVD interpretovat jako filtrovanou pseudoinverzi, viz (2.7)–(2.8), lze jako filtrovanou pseudoinverzi vyjádřit i T-TLS. Odpovídající filtrační faktory odvodili Fierro, Golub, Hansen a O'Leary v článku [31]. My jsme na regularizaci pomocí TLS narazili poprvé v našem článku [62, Example 5.4, str. 874–875]; konkrétně viz vysvětlující komentář za příkladem 5.4. Ukazujeme zde, že klasický TLS algoritmus, který ve skutečnosti místo TLS řešení hledá právě T-TLS řešení (viz [63, Lemma 6.2, str. 768]), provádí implicitně jistou formu regularizace na úlohách s více pravými stranami. Zejména v případě, obsahuje-li původní úloha složený core problém s komponentami s výrazně různými singulárními čísly; viz [62, rov. (5.21)].

Podstatný rozdíl mezi tímto výsledkem a prací Fierra et al. je v tom, že článek [31] uvažuje pouze úlohy s jednou pravou stranou. Pokusili jsme se tedy jejich přístup zobecnit i na úlohy s d pravými stranami (1.2). Toto zobecnění se povedlo provést a je prezentované v článku [69], který je odeslán k publikaci, ale zatím publikován nebyl. Ukazujeme zde, že v případě více pravých stran je situace podstatně složitější. Obecně k libovolnému sloupci $X_{\mathsf{T-TLS}}$ řešení přispívají vždy všechny pravé strany. Filtrační faktory (připomeňme, že v úloze s jednou pravou stranou jich je $r = \operatorname{rank}(A)$; viz (2.7)) tak tvoří tenzor třetího řádu rozměrů $d \times d \times r$. Jeho (i, j, k)-tý prvek popisuje k-tý filtrační faktor, kterým je filtrován vliv i-tého sloupce pravé strany na j-tý sloupec řešení.

Kapitola 3

Tenzorové výpočty

3.1 Úvod

Výpočetní nástroje lineární algebry, jejich popis a současně s tím i implementace přirozeným historickým vývojem přecházely od práce s jednotlivými souřadnicemi přes práci s celými vektory až po práci s jednotlivými bloky – de-facto (pod-)maticemi. Příkladem druhého jmenovaného přechodu je např. zobecňování dříve zmiňovaného úplného problému nejmenších čtverců z úloh s jednou pravou stranou na úlohy s více pravými stranami – tj. s maticovou pravou stranou.

Následujícím přirozeným krokem je přechod od blokových resp. maticových formulací k formulacím tenzorovým. Protože pojem tenzoru se v matematice vyskytuje v řadě různých disciplín již dlouho a jeho obsah je vnímán při různých příležitostech nepatrně odlišně, je dobré pojem jasně vymezit. My, v souladu s tím jak s pojmem pracuje současná numerická lineární algebra (viz přehledové články [73], [41] nebo knihy [28] a [17]), budeme pod pojmem *tenzor* (k-tého řádu) rozumět vícerozměrné (k-rozměrné) pole čísel

$$\mathcal{A} = (a_{i_1,i_2,\ldots,i_k}) \in \mathbb{F}^{m_1 \times m_2 \times \cdots \times m_k}. \tag{3.1}$$

Speciálně skaláry, vektory a matice, tak jak jsme s nimi pracovali doposud, budou tenzory nultého, prvého, respektive druhého řádu.

Přímá práce s tenzory vyšších řádů může být výpočetně velmi náročná ne-li nemožná; často je volně parafrázována mottem „*curse of dimensionality*“. Stačí si uvědomit, kolik složek bude mít např. tenzor řádu $k = 100$ s minimalistickými rozměry $n_1 = n_2 = \cdots = n_k = 2$; totiž $2^{100} \approx 1.27 \cdot 10^{30}$, přičemž pro uložení každé složky ve standardní dovjité přesnosti bychom potřebovali osm bytů. Ústředním nástrojem pro práci s takto velkými daty je tedy komprese, zpravidla postavená na rozkladu tenzoru (viz např. [73]) a využití tzv. *low-rank aritmetiky*; viz např. [77], [76], nebo [41].

3.2 Vybrané základní operace s tenzory

Jednorozměrné a dvourozměrné podtenzory tenzoru $\mathcal{A}$, které můžeme pomocí triviálního izomorfismu zobrazit na sloupcové vektory, resp. matice, zpravidla nazýváme vlákna a řezy v příslušných módech. Konkrétně

$$a^{(\ell)}_{\mathcal{J}} \equiv \begin{bmatrix} a_{i_1,\ldots,i_{\ell-1},1,i_{\ell+1},\ldots,i_k} \\ \vdots \\ a_{i_1,\ldots,i_{\ell-1},n_\ell,i_{\ell+1},\ldots,i_k} \end{bmatrix} \in \mathbb{F}^{m_\ell} \tag{3.2}$$

je $\mathscr{I}$-té vlákno ℓ-tého módu tenzoru $\mathcal{A}$, kde $\mathscr{I} \equiv (i_1, \ldots, i_{\ell-1}, i_{\ell+1}, \ldots, i_k)$ je multiindex upřesňující polohu vlákna v tenzoru. Řezy definujeme obdobně pomocí dvojice módů. Protože chceme při práci s tenzory využít obvyklý maticový aparát, zavedeme si dále pojem *rozvoje tenzoru do matice* v módu ℓ, anglicky *ℓ-mode matricization* nebo *unfolding*. Tímto rozvojem budeme rozumět matici

$$\mathcal{A}^{(\ell)} \equiv \left[a^{(\ell)}_{\mathscr{I}_1}, a^{(\ell)}_{\mathscr{I}_2}, \ldots, a^{(\ell)}_{\mathscr{I}_{M_\ell}} \right] \in \mathbb{F}^{m_\ell \times M_\ell}, \quad \text{kde} \quad M_\ell = \frac{M}{m_\ell}, \quad M = \prod_{j=1}^{k} m_j, \tag{3.3}$$

se sloupci tvořenými všemi vlákny ℓ-tého módu s multiindexy $\mathscr{I}_\mu$ seřazenými v lexikografickém pořadí. Rozvoj do matice lze snadno zobecnit pro dva a více módů; viz [134]. Obzvlášť důležitý je rozvoj podle všech módů, nazývaný vektorizace,

$$\operatorname{vec}(\mathcal{A}) \equiv \begin{bmatrix} a_{1,\ldots,1} \\ \vdots \\ a_{n_1,\ldots,n_k} \end{bmatrix} \in \mathbb{F}^M, \tag{3.4}$$

tj. všechny prvky tenzoru naskládáme do jednoho dlouhého vektoru; prvky opět řadíme tak, aby multiindexy $(i_1, \ldots, i_k)$ byly v lexikografickém pořadí.

Uvažujme dále matice $W_\ell = (w_{\nu_\ell, i_\ell}) \in \mathbb{F}^{n_\ell \times m_\ell}$, $\ell = 1, \ldots, k$. Součin tenzoru $\mathcal{A} \in \mathbb{F}^{m_1 \times m_2 \times \cdots \times m_k}$ (3.1) s maticí W_ℓ v ℓ-tém módu definujeme analogicky jako v případě součinu dvou matic vztahem

$$\mathcal{A} \times_\ell W_\ell = \Big(\sum_{i_\ell=1}^{m_\ell} a_{i_1,\ldots,i_\ell,\ldots,i_k} \, w_{\nu_\ell, i_\ell} \Big) \in \mathbb{F}^{m_1 \times \cdots \times m_{\ell-1} \times n_\ell \times m_{\ell+1} \times \cdots \times m_k}. \tag{3.5}$$

Dostaneme tedy tenzor, jehož vlákna ℓ-tého módu jsou vlákna ℓ-tého módu původního tenzoru $\mathcal{A}$ vynásobená maticí W_ℓ. S využitím rozvoje tenzoru do matice (3.3) snadno ověříme, že platí

$$(\mathcal{A} \times_\ell W_\ell)^{(\ell)} = W_\ell \mathcal{A}^{(\ell)} \in \mathbb{F}^{n_\ell \times M_\ell}. \tag{3.6}$$

Zřejmě také platí

$$\Big(\mathcal{A} \times_\ell W_\ell \Big) \times_s W_s = \Big(\mathcal{A} \times_s W_s \Big) \times_\ell W_\ell. \tag{3.7}$$

Součin tenzoru $\mathcal{A}$ se všemi maticemi W_ℓ v odpovídajících módech ℓ, $\ell = 1, \ldots, k$, můžeme tedy z úsporných důvodů psát např. ve tvaru

$$[\![\mathcal{A} \,|\, W_1, W_2, \ldots, W_k]\!] \in \mathbb{F}^{n_1 \times n_2 \times \cdots \times n_k}. \tag{3.8}$$

S využitím vektorizace (3.4) a analogicky k (3.6) platí

$$\operatorname{vec}\Big([\![\mathcal{A} \,|\, W_1, W_2, \ldots, W_k]\!] \Big) = \Big(W_1 \otimes W_2 \otimes \cdots \otimes W_k \Big) \operatorname{vec}(\mathcal{A}), \tag{3.9}$$

kde $\otimes$ značí Kroneckerův součin matic; viz [134].

Součin dvou tenzorů, někdy též nazývaný *úžení* tenzorů, opět zavádíme analogicky ke klasickému maticovému součinu. Mějme dva tenzory řádů k a t,

$$\mathcal{A} \in \mathbb{F}^{m_1 \times \cdots \times m_k}, \quad \mathcal{B} \in \mathbb{F}^{n_1 \times \cdots \times n_t}, \qquad \text{přičemž} \quad m_\ell = n_s \tag{3.10}$$

pro nějaké ℓ a s splňující $1 \le \ell \le k$, $1 \le s \le t$. Jejich součin v módech ℓ a s

$$\mathcal{A} \times_{\ell,s} \mathcal{B} = \Big(\sum_{\alpha=1}^{m_\ell} a_{i_1,\ldots,i_{\ell-1},\alpha,i_{\ell+1},\ldots,i_k} \cdot b_{j_1,\ldots,j_{s-1},\alpha,j_{s+1},\ldots,j_t} \Big) \tag{3.11}$$
$$\in \mathbb{F}^{m_1 \times \cdots \times m_{\ell-1} \times m_{\ell+1} \times \cdots \times m_k \times n_1 \times \cdots \times n_{s-1} \times n_{s+1} \times \cdots \times n_t}$$

je tenzor řádu $k+t-2$. Obdobně lze tenzory násobit ve více módech současně. Nechť platí $m_{\ell_1} = n_{s_1}$ a $m_{\ell_2} = n_{s_2}$ pro nějaké ℓ_1, ℓ_2, s_1 a s_2 splňující např. $1 \le \ell_1 < \ell_2 \le k$, $1 \le s_1 < s_2 \le t$. Pak součin tenzorů $\mathcal{A}$ a $\mathcal{B}$ ve dvojici módů (ℓ_1, ℓ_2) a (s_1, s_2)

$$\mathcal{A} \times_{(\ell_1,\ell_2),(s_1,s_2)} \mathcal{B} = \Big(\sum_{\alpha=1}^{m_{\ell_1}} \sum_{\beta=1}^{m_{\ell_2}} a_{i_1,\cdots,i_{\ell_1-1},\alpha,i_{\ell_1+1},\cdots,i_{\ell_2-1},\beta,i_{\ell_2+1},\cdots,i_k} \cdot b_{j_1,\cdots,j_{s_1-1},\alpha,j_{s_1+1},\cdots,j_{s_2-1},\beta,j_{s_2+1},\cdots,j_t} \Big) \tag{3.12}$$

je tenzor řádu $k+t-4$, atd; viz také (1.32)–(1.33) v kap. 1, kde používáme součin dvou tenzorů dokonce v obecné ν-tici módů.

3.3 Tenzorové rozklady

Pro kompresi maticových dat můžeme nejsnáze použít singulární rozklad. Pokud chceme danou matici hodnosti r aproximovat její nejlepší aproximací hodnosti τ, $\tau \le r$, použijeme právě singulární rozklad (viz (1.6)) a zanedbáme $r-\tau$ nejmenších singulárních čísel; viz [106] a [27], případně [85]; viz také [111]. Chceme-li provést kompresi, stačí místo původní matice ukládat pouze prvních τ singulárních tripletů (u'_i, σ'_i, v'_i), $i = 1, \ldots, \tau$; viz např. [26, kap. 5.7]. Takto uchovaný singulární rozklad se pro svou schopnost zmenšit místo na disku potřebné k uložení původních dat (tj. schopnost komprese dat) nazývá *ekonomický singulární rozklad* (ESVD). Analogií (ekonomického) singulárního rozkladu ve světě tenzorů je tzv. *Tuckerův rozklad*, někdy též nazývaný *higher-order SVD* (HOSVD); viz [117], [118], [119]; případně [73, kap. 4.1], [116, kap. 3.1.2], [134], či [71].

Definice 6 (Vektorová hodnost tenzoru, Tuckerovo jádro, Tuckerův rozklad)**.** *Nechť $\mathcal{A} \in \mathbb{F}^{m_1 \times m_2 \times \cdots \times m_k}$ je tenzor řádu k (3.1) a nechť $\mathcal{A}^{(\ell)} \in \mathbb{F}^{m_\ell \times M_\ell}$, kde $M_\ell = M/m_\ell$ a $M = \prod_{j=1}^{k} m_j$, jsou rozvoje tohoto tenzoru do matic pro všechna $\ell = 1, \ldots, k$ (3.3). Uvažujme dále*

$$r_\ell = \operatorname{rank}(\mathcal{A}^{(\ell)}) \quad a \quad \mathcal{A}^{(\ell)} = U^{(\ell)} \Sigma^{(\ell)} (V^{(\ell)})^{\mathsf{H}}, \tag{3.13}$$

hodnosti těchto rozvojů a jejich singulární rozklady, kde

$$U^{(\ell)} = [U_1^{(\ell)}, U_2^{(\ell)}] \in \mathbb{F}^{m_\ell \times m_\ell}, \quad U_1^{(\ell)} \in \mathbb{F}^{m_\ell \times r_\ell}. \tag{3.14}$$

Pak vektorovou hodností tenzoru nazýváme uspořádanou k-tici

$$\operatorname{rank}(\mathcal{A}) \equiv (r_1, r_2, \ldots, r_k) \tag{3.15}$$

a Tuckerovým jádrem nazýváme tenzor

$$\mathcal{A}_{\mathsf{T-core}} \equiv [\![\mathcal{A} \,|\, (U_1^{(1)})^{\mathsf{H}}, (U_1^{(2)})^{\mathsf{H}}, \ldots, (U_1^{(k)})^{\mathsf{H}}]\!] \in \mathbb{F}^{r_1 \times r_2 \times \cdots \times r_k}. \tag{3.16}$$

Rovnosti

$$\mathcal{A} = [\![\operatorname{diag}_k(\mathcal{A}_{\mathsf{T-core}}, 0_{m_1-r_1, m_2-r_2, \ldots, m_k-r_k}) \,|\, U^{(1)}, U^{(2)}, \ldots, U^{(k)}]\!] \tag{3.17}$$

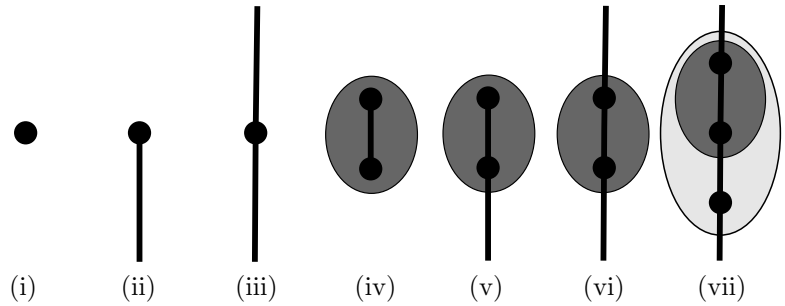
$$= [\![\mathcal{A}_{\mathsf{T-core}} \,|\, U_1^{(1)}, U_1^{(2)}, \ldots, U_1^{(k)}]\!], \tag{3.18}$$

kde $\operatorname{diag}_k(\cdot,\cdot)$ ze svých argumentů sestaví blokově diagonální tenzor k-tého řádu, nazýváme úplným, respektive ekonomickým Tuckerovým rozkladem.

Všimněme si, že ℓ-tý rozměr Tuckerova jádra je dán počtem lineárně nezávislých vláken ℓ-tého módu původního tenzoru. Obecně tedy nelze provést větší kompresi beze ztráty informace; viz [73]. Podobně jako u singulárního rozkladu můžeme ale

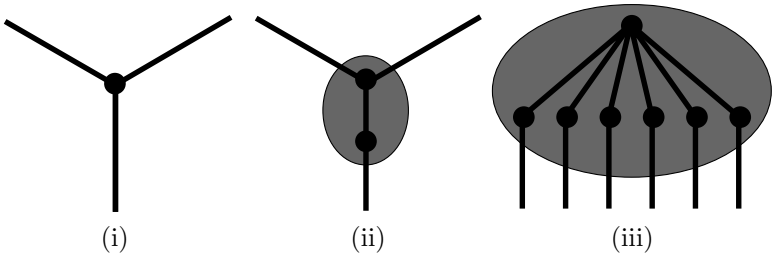
i zde zanedbávat nejmenší singulární čísla v některých, případně ve všech rozvojích do matic a tím zmenšovat paměťové nároky při uložení dat, viz např. [134]. Pro velké řády tenzorů (např. $k = 100$) však nebude takové snížení paměťových nároků dostatečné, stále budeme potřebovat řádově r^k paměti, kde $r = \max_\ell\{r_\ell\}$. V ideálním případě bychom chtěli, aby paměťové nároky závisely *lineárně* na řádu tenzoru. Za tím účelem se zavádí řada specializovanějších rozkladů či formátů, obecně nazývaných *tenzorové sítě*. Pro zjednodušení značení však bude užitečné nejprve zavést symbolický zápis tenzorových součinů.

Symbolický zápis je graf, jehož vrcholy jsou tenzory, přičemž počet hran incidentních s vrcholem odpovídá řádu tenzoru. Vrchol, z něhož vychází k hran, je tedy tenzorem řádu k; každá hrana koresponduje s právě jedním pevně daným módem ℓ. Spojuje-li nějaká hrana dva vrcholy grafu – tenzory – představuje vazbu mezi nimi, přesněji řečno sumu, která sčítá přes indexy příslušející daným módům (tyto módy tedy nutně musí mít stejné rozměry). Hranu, která inciduje jen s jedním vrcholem, nazýváme volnou hranou, případně fyzickým indexem a odpovídá některému z módů celého objektu. Na obrázku 3.1 je vidět sedm schémátek představujících několik základních objektů: nejprve vidíme (i) skalár, (ii) vektor a (iii) matici, tedy tenzory řádu nula, jedna a dva. Graf vždy obsahuje jen jediný vrchol a žádnou, jednu, respektive dvě volné hrany. Další čtyři schémátka zobrazují opět (iv) skalár, (v) vektor a (vi)–(vii) dvě matice. Skalár na schémátku (iv), tj. šedivý ovál bez volné hrany, je ovšem vyjádřen pomocí dvou tenzorů prvního řádu, vektorů, které sdílejí společnou hranu; tento skalár je tedy zapsán jako skalární součin dvou vektorů. Vektor na schémátku (v), opět šedý ovál, je vyjádřen jako součin matice a vektoru, atd. Snadno si dovedeme představit další možnosti, které tento zápis skýtá, např. stopu matice $\mathrm{trace}(A)$ je možné vyjádřit jako matici, tj. vrchol se dvěma hranami, které jsou spojené dohromady, atd.

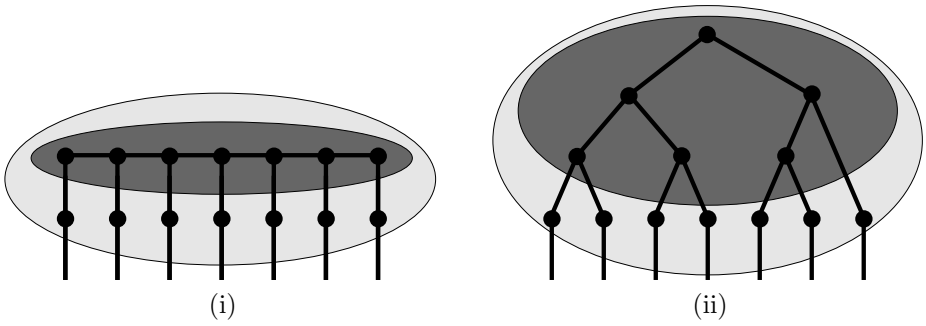


Obrázek 3.1: Symbolické vyjádření jednoduchých tenzorů a jejich interakcí. Zleva: (i) skalár, (ii) vektor, (iii) matice, (iv) skalár ve tvaru součinu (skalární součin) dvou vektorů, (v) vektor ve tvaru součinu matice s vektorem, (vi) matice ve tvaru součinu dvou matic, (vii) matice ve tvaru součinu tří matic

Pro úplnost obrázek 3.2 zobrazuje (i) tenzor třetího řádu, (ii) součin tenzoru s maticí a (iii) Tuckerův rozklad. Znovu vidíme, že tenzor řádu $k = 6$ zredukovaný Tuckerovým rozkladem a uložený v ekonomickém tvaru či formátu stále obsahuje tenzor jádra řádu k; dále pak k matic (obsahujících levé singulární vektory tvořící báze oborů hodnot jednotivých rozvojů) mapujících indexy jádra na fyzické indexy. Pro rozbití Tuckerova jádra velkého řádu na tenzory řádů nižších se používají nejčastěji dva postupy. První tzv. *tenzorový vláček*, anglicky *tenzor train decomposition* (TT; viz [90] a obrázek 3.3, schéma (i) vlevo) a *hierarchický Tuckerův rozklad*; viz [43], [40] a obrázek 3.3, schéma (ii) vpravo. Obě schémata představují Tuckerův rozklad tenzoru řádu sedm, kde Tuckerovo jádro (tmavě šedý ovál) vykazuje další strukturu; viz např. [120].



Obrázek 3.2: Symbolické vyjádření tenzorů vyšších řádů. Zleva: (i) tenzor třetího řádu, (ii) tenzor třetího řádu ve tvaru součinu tenzoru třetího řádu s maticí, (iii) tenzor šestého řádu ve tvaru Tuckerova rozkladu



Obrázek 3.3: Tenzor řádu sedm ve tvaru Tuckerova rozkladu. Vlevo (i) s Tuckerovým jádrem ve tvaru tenzorového vláčku, vpravo (ii) hierarchický Tuckerův rozklad.

Tuckerovo jádro obsahuje v případě vláčku právě $k-2$ tenzorů řádu tři a dva tenzory řádu dva. Pokud jsou rozměry všech zúčastněných tenzorů rozumně velké, např. omezené hodnotou r, budeme pro uložení jádra potřebovat řádově $2r^2 + (k-2)r^3$ paměti, narozdíl od původních r^k. Vidíme, že tenzorový vláček nám skutečně zajistí spotřebu paměti úměrnou řádu tenzoru. V případě hierarchického Tuckerova jádra je situace složitější. Vidíme, že zde je jeden tenzor řádu dva, a pak v ideálním případě $k-2$ tenzorů řádu tři tvořících vyvážený binární strom; spotřeba paměti na uložení jádra (pokud budou rozměry tenzorů rozumně omezeny) je $r^2 + (k-2)r^3$, tj. opět úměrná řádu tenzoru. Poznamenejme, že podobných formátů je více. Velké množství práce se věnovalo tomu, jaký rozklad, či formát je pro které aplikace vhodnější, který je programátorsky přívětivější atd.; viz např. [2], [3], [4], [116], [78], [79], případně [128]. Obecně lze říci, že hierarchický Tuckerův formát, pomineme-li požadavek na vyváženost stromu, je obecnější (ostatně „zavěsíme-li“ graf vláčku za jeden z tenzorů řádu dva, vznikne z něj také strom, byť extrémně nevyvážený), což dává větší možnosti a volnost ve volbě vhodného rozkladu dle dané aplikace a struktury dat; viz např. [75, kap. 4, zejm. Fig 4.1].

Vlastní přínos

Důležitým aspektem tenzorových výpočtů je nejen znát možnosti různých rozkladů a být schopen mnoharozměrná pole čísel do počítače uložit, ale také s nimi provádět operace, které je třeba. Tedy tenzory např. v hierarchickém Tuckerově formátu umět sčítat, případně na takový tenzor umět aplikovat operátor. Článek [77] se zabývá řešením soustavy rovnic jejíž matice a pravá strana (obecně tudíž i řešení) závisí na

sadě parametrů $\alpha^{(j)}$, tj.

$$A(\alpha)x(\alpha) = b(\alpha), \qquad \text{kde} \qquad \alpha = (\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(p)}). \tag{3.19}$$

Za předpokladu, že závislost $A(\alpha)$ a $b(\alpha)$ na parametrech $\alpha^{(j)}$ je lineární, nebo se dá v nějakém okolí předem daného zájmového bodu lineární závislostí rozumně aproximovat, můžeme úlohu po navzorkování, tj. diskretizaci prostoru parametrů přeformulovat tenzorově.

V našem článku [76] a v jeho rozšířené verzi [75] se zabýváme složitější úlohou, řešením zobecněné ljapunovské maticové rovnice

$$A(\alpha)X(\alpha)M(\alpha)^{\mathsf{T}} + M(\alpha)X(\alpha)A(\alpha)^{\mathsf{T}} = B(\alpha)B(\alpha)^{\mathsf{T}}, \tag{3.20}$$

kde $A(\alpha)$, $M(\alpha) \in \mathbb{R}^{n\times n}$ jsou čtvercové matice a $B(\alpha) \in \mathbb{R}^{n\times t}$ závislé na jednom (viz [76]), resp. p (viz [75]) parametrech. Opět předpokládáme, že parametr α nabývá hodnot z nějakého zájmového intervalu, kde je navzorkován a ve kterém lze závislost matic na α rozumně aproximovat lineární závislostí.

Snadno ověříme, že jsou-li matice $A(\alpha)$ a $M(\alpha)$ symetrické pozitivně definitní, lze celou rovnici (3.20) přepsat jako klasickou soustavu rovnic se symetrickou pozitivně definitní maticí $L(\alpha) \in \mathbb{R}^{N\times N}$, $N = n^2$, a k jejímu řešení tedy můžeme použít např. metodu sdružených gradientů (CG); viz [58]. V článku [76] jsme ukázali, jak metodu CG formulovat nejprve pro ljapunovskou rovnci $AXM^{\mathsf{T}} + MXA^{\mathsf{T}} = BB^{\mathsf{T}}$, která nezávisí na žádných parametrech a respektuje ljapunovskou strukturu úlohy. To se týká zejména akce pozitivně definitní matice, která není prováděna pomocí klasického součinu matice s vektorem

$$L : \mathbb{R}^N \longrightarrow \mathbb{R}^N, \qquad L : v \longmapsto w = Lv = (A \otimes M + M \otimes A)v \tag{3.21}$$

ale přímo pomocí ljapunovského operátoru

$$\mathcal{L} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}, \qquad \mathcal{L} : V \longmapsto W = \mathcal{L}(V) = AVM^{\mathsf{T}} + MVA^{\mathsf{T}}, \tag{3.22}$$

kde $v = \mathrm{vec}V$ a $w = \mathrm{vec}W$. S výhodou zde lze využít tzv. *low-rank* aritmetiky; viz např [98] a [81], [82]. Metodu CG startujeme *zpravidla* s nulovým počátečním odhadem řešení, tj. $X_0 = 0$, tedy maticí zjevně nízké hodnosti. Pokud je navíc pravá strana nízkého ranku, tj. $t \ll n$, pak singulární čísla přesného řešení X exponenciálně klesají k nule, tedy i přesné řešení lze snadno aproximovat maticí nízké hodnosti. Můžeme tedy s maticí X_0 pracovat např. ve tvaru ekonomického singulárního rozkladu (v článku [76] používáme analogický ekonomický spektrální rozklad, neboť všechny matice, které se ve výpočtu vyskytují jsou symetrické) a metodu CG přeformulovat tak, aby se celá matice nikdy nesestavovala; viz [76, Algorithm 1, str. 668]. To je výhodné zejména když A a M jsou řídké matice (v našem článku jsou to matice tuhosti a matice hmotnosti z FEM diskretizace) a n je tak velké, že by se hustá matice řádu n nevešla do paměti.

Důležitou součástí našeho algoritmu byl výběr předpodmiňovače pro CG, který zachovával ljapunovskou strukturu úlohy, tj. jehož aplikace šla zapsat anologicky jako akce operátoru (3.22). Používáme tzv. ADI iterace (*alternating directions implicite*; viz [72], viz také analogické použití na Sylvestrovu [8] případně Ricattiho [9] maticové rovnice) a iterace znaménkové funkce (*sign funcion iteration*; viz [10], [6]); viz [76, kap. 2.3.1, zejm. Algorithm 2 a 3, str. 671–674]. Explicitní inverzi řídké matice A pro potřebu iterace znaménkové funkce provádíme užitím tzv. *hierarchických* (viz [42] a [7]), případně *hierarchických semi-separabilních* (HSS) matic (viz [18], [19], [133] a [127]); pozn., že inverze se předpočítává jednorázově pro všechny CG-iterace.

Celý koncept je rozšířen pro úlohy závislé na jednom parametru v [76, kap. 3]. V této kapitole ukazujeme, že řešení $\mathcal{X} \in \mathbb{R}^{n\times n\times m}$ – tenzor třetího řádu obsahující

matice $X(\alpha) \in \mathbb{R}^{n\times n}$ jako své frontální řezy (obsahuje právě m takových matic pro m vzorků $\alpha_1, \ldots, \alpha_m$ parametru α) lze za předpokladu $t \ll n$ dobře aproximovat tenzorem v Tuckerově tvaru s malým Tuckerovým jádrem. Tedy lze aproximovat tenzorem nízkého ranku. Článek obsahuje větu s poměrně techickým důkazem (viz [76, Theorem 1, str. 676–678]) ukazující, jak klesají singulární čísla jednotlivých rozvojů $\mathcal{X}_{(1)} = \mathcal{X}_{(2)}$ a zejména $\mathcal{X}_{(3)}$. V článku studujeme strukturu tenzorů vznikajících při jednotlivých operacích v metodě CG (viz [76, rovnice (25) a (26), str. 679–680]), výpočetní náročnost těchto operací i náročnost celého algorimu.

Původní autorské články vztahující se k tomuto tématu publikované v impaktovaných časopisech jsou přiloženy, konkrétně:

[76] (2014) jako příloha A.6 na str. 161.

Poznamenejme, že v rozšířené verzi článku (viz [75, kap. 4]) je ukázáno, že lze stejný postup použít i pro více parametrů. Zde konkrétně pracujeme se čtyřmi parametry, které hrají v úloze analogické role; viz [75, Fig. 4.1 vlevo, str. 18]. Pro obecně p parametrů je tenzor řešení $\mathcal{X}$ právě řádu $p+2$. V případě čtyř parametrů tak již nebylo možné používat pouze obyčejný Tuckerův rozklad z důvodů paměťových nároků. Tuckerovo jádro bylo uloženo s užitím hierarchického Tuckerova formátu; struktura, resp. symetrie stromu hierarchického rozkladu koresponduje se symetrií rolí jednotlivých parametrů a úlohy jako takové (matice $X(\alpha)$ jsou symetrické), tj. se symetrií fyzických indexů tenzoru; viz [75, Fig. 4.1 vpravo, str. 18]. Výsledky získané pro hierarchický Tuckerův rozklad do článku nakonec, jak je patrné, zahrnuty nebyly. Důvodem pro to byl fakt, že ačkoliv jsme provedli řadu experimentů na reálných datech (převážně z Oberwolfach benchmark collection; viz [74]) s tenzory různých řádů, výsledky nejsou až tak slibné, jak jsme doufali. Přestože jsou výsledky oproti původnímu záměru méně ambiciózní, nejsou snad zanedbatelné; v nedávno zveřejněném preprintu [109] autoři píší:

> Although a lot of attention has been paid to Lyapunov equations, only very few publications dedicated to solving parametric algebraic Lyapunov equations (PALEs) can be found. To our knowledge, the method proposed in [76] is the only released printed work on this subject.

Výhled do budoucna

Tenzorové sítě dávají poměrně široké možnosti, jaké rozklady hledat. Speciálně nám dávají možnost vytvářet rozklady, které na rozdíl od všech předchozích obsahují cykly (uzavřené smyčky v grafu). První takovou smyčkou, kterou jsme již zmiňovali, je stopa stopa čtvercové matice $A = (a_{i,j}) \in \mathbb{F}^{m\times m}$, tedy číslo $\operatorname{trace}(A) = \sum_{i=1}^{m} a_{i,i}$. Stopu lze reprezentovat jako síť – graf obsahující jediný puntík se dvěma odchozími hranami (tenzor druhého řádu) spojenými dohromady. Podobně lze vytvořit např. „cyklický součin“ tří matic $A = (a_{i,j}) \in \mathbb{F}^{m\times n}$, $B = (b_{j,l}) \in \mathbb{F}^{n\times s}$, $C = (c_{l,i}) \in \mathbb{F}^{s\times m}$ ve tvaru $\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{s} a_{i,j}\cdot b_{j,l}\cdot c_{l,i}$, který bychom obtížně pomocí klasického maticového násobení zapsali. Odpovídající tenzorová síť bude graf obsahující tři puntíky spojené do trojúhelníku.

Tenzorový vláček tenzoru $\mathcal{A} = (a_{i_1,i_2,\ldots,i_k}) \in \mathbb{F}^{m_1\times m_2\times\cdots\times m_k}$ můžeme zapsat ve tvaru

$$\mathcal{A} = \Big(\sum_{\alpha_1}\sum_{\alpha_2}\cdots\sum_{\alpha_{k-1}} w^{(1)}_{i_1,\alpha_1}\cdot w^{(2)}_{\alpha_1,i_2,\alpha_2}\cdot\cdots\cdot w^{(k-1)}_{\alpha_{k-2},i_{k-1},\alpha_{k-1}}\cdot w^{(k)}_{\alpha_{k-1},i_k}\Big), \tag{3.23}$$

kde oba krajní vagónky jsou tenzory řádu dva (s jedním fyzickým indexem i_1, resp. i_k a jedním vnitřním sčítacím indexem α_1, resp. α_{k-1}) a všechny ostatní vagónky jsou tenzory řádu tři (opět s jedním fyzickým indexem i_ℓ a dvěma vnitřními sčítacími indexy $\alpha_{\ell-1}$ a α_ℓ, $\ell = 2, \ldots, k-1$). Jednoduchou modifikací tenzorové sítě můžeme z tenzorového vláčku vytvořit tzv. *tenzorový řetízek*, anglicky *tensor chain decomposition* (TC); viz např. [29]. Stačí do grafu přidat jedinou hranu propojující první a poslední vagónek. Pro původní tenzor tak dostaneme cyklický rozklad tvaru

$$\mathcal{A} = \Big(\sum_{\alpha_1} \sum_{\alpha_2} \cdots \sum_{\alpha_{k-1}} \sum_{\alpha_k} c^{(1)}_{\alpha_k, i_1, \alpha_1} \cdot c^{(2)}_{\alpha_1, i_2, \alpha_2} \cdot \cdots \cdot c^{(k-1)}_{\alpha_{k-2}, i_{k-1}, \alpha_{k-1}} \cdot c^{(k)}_{\alpha_{k-1}, i_k, \alpha_k} \Big), \tag{3.24}$$

kde se vyskytuje pouze k tenzorů třetího řádu, každý s jedním fyzickým a dvěma vnitřními sčítacími indexy.

Zatímco při výpočtu rozkladů do stromových struktur (jako je hierarchický Tucker nebo TT) vystačíme teoreticky jen se singulárním rozkladem různých (zobecněných) rozvojů tenzorů do matic, v případě cyklických struktur je situace principiálně složitější. Je potřeba použít nějaký iterační proces, který se snaží tenzor rozložit do předepsané struktury, zatímco postupně minimalizuje chybu aproximace. Velmi jednoduchý postup může být postaven například na klasické minimalizaci součtu čtverců (tedy euklidovské normy, resp. jejího kvadrátu) chyby, případně v kombinaci s nějakými dalšími předpoklady omezujícími např. velikosti jednotlivých tenzorů. Tím se dostáváme k tématu, o kterém již byla řeč na závěr kap. 1. Minimalizaci nemusíme formulovat jen jako klasický, obyčejný problém nejmenších čtverců, ale stejně dobře jako úplný problém nejmenších čtverců. Vhodnost té či oné metody vždy závisí na konkrétním tenzoru, konkrétní aplikaci a kontextu. To je motivací pro naši současnou práci o řešitelnosti TLS úloh s tenzory, která je aktuálně rozpracovaná v rukopise [70].

Vzhledem k aktuálnosti tématu tenzorových výpočtů a zejména tenzorových rozkladů i v technických aplikacích (např. při zpracování signálů [113], medicínských dat [110], [128], atd.) také připravujeme zatím útlé, česky psané skriptum k tomuto tématu [71].

Kapitola 4

Krylovovské metody

4.1 Úvod

Krylovovské metody jsou třídou metod pro řešení řady úloh lineární algebry, nejčastěji soustav lineárních rovnic $Ax = b$ se čtvercovou regulární maticí, případně pro řešení (částečného) problému vlastních čísel $Ax = x\lambda$, $x \neq 0$. Jádrem těchto metod je budování ortonormální báze krylovovského prostoru pro $\ell = 1, 2, \ldots$

$$\mathscr{K}_\ell(A, v) = \operatorname{span}\{v, Av, \ldots, A^{\ell-1}v\}, \quad \text{kde} \quad A \in \mathbb{F}^{m\times m},\ v \in \mathbb{F}^m,\ \|v\|_2 = 1. \tag{4.1}$$

K tomu slouží dva velmi úzce provázané algoritmy: Arnoldiho metoda (viz např. [26, kap. 7.1]) převádějící obecnou čtvercovou matici A na tzv. horní Hessenbergův tvar a Lanczosova tridiagonalizační metoda (viz např. [26, kap. 7.2]) použitelné pouze v případě, že matice A je čtvercová a hermitovská, tj. $A^{\mathsf{H}} = A$. Zjednodušeně bychom mohli říci, že Arnoldiho metoda aplikovaná na hermitovskou matici v podstatě splývá s metodou Lanczosovou.

U úloh s maticí, která není čtvercová, např. při řešení přeurčené aproximační úlohy $Ax \approx b$ ve smyslu nejmenších čtverců (ať už máme na mysli klasický LS problém, nebo TLS problém; viz kap. 1), můžeme využít krylovovské metody, které staví ortonormální báze prostorů pro $\ell = 1, 2, \ldots$

$$\mathscr{K}_\ell(AA^{\mathsf{H}}, s) = \operatorname{span}\{s, (AA^{\mathsf{H}})s, \ldots, (AA^{\mathsf{H}})^{\ell-1}s\}, \tag{4.2}$$

$$\mathscr{K}_\ell(A^{\mathsf{H}}A, w) = \operatorname{span}\{w, (A^{\mathsf{H}}A)w, \ldots, (A^{\mathsf{H}}A)^{\ell-1}w\}, \tag{4.3}$$

$$\text{kde} \quad A \in \mathbb{F}^{m\times n},\ s \in \mathbb{F}^m,\ w \in \mathbb{F}^n \quad \text{a} \quad \|s\|_2 = \|w\|_2 = 1.$$

K tomu slouží algoritmus Golubovy–Kahanovy bidiagonalizace. Položíme-li $s \equiv s_1 = b/\beta_1$, kde $\beta_1 = \|b\|_2$, a $w \equiv w_1 = A^{\mathsf{H}}s_1/\alpha_1$, kde $\alpha_1 = \|A^{\mathsf{H}}s_1\|_2$, pak bidiagonalizace počítá pro $\ell = 1, 2, \ldots$ Golubovy vektory $s_{\ell+1}$ a Kahanovy vektory $w_{\ell+1}$ pomocí rekurencí

$$s_{\ell+1}\beta_{\ell+1} = Aw_\ell - s_\ell\alpha_\ell, \tag{4.4}$$

$$w_{\ell+1}\alpha_{\ell+1} = A^{\mathsf{H}}s_{\ell+1} - w_\ell\beta_{\ell+1}, \tag{4.5}$$

kde $\beta_{\ell+1}$ a $\alpha_{\ell+1}$ jsou normalizační koeficienty volené tak, aby $\|s_{\ell+1}\|_2 = \|w_{\ell+1}\|_2 = 1$. Generuje tak ortonormální báze obou prostorů tak, že platí

$$\mathscr{K}_\ell(AA^{\mathsf{H}}, s) = \operatorname{span}\{s_1, s_2, \ldots, s_\ell\}, \tag{4.6}$$

$$\mathscr{K}_\ell(A^{\mathsf{H}}A, w) = \operatorname{span}\{w_1, w_2, \ldots, w_\ell\}, \quad \text{kde} \quad s_i^{\mathsf{H}}s_j = w_i^{\mathsf{H}}w_j = \delta_{i,j}. \tag{4.7}$$

Bidigaonalizace (4.4)–(4.5) úzce souvisí s Lanczosovsou tridiagonalizací díky tomu, že matice AA^{H} a $A^{\mathsf{H}}A$ jsou čtvercové a hermitovské. Fakticky bidiagonalizace provádí simultáně dvě Lanczosovy tridiagonalizace pro obě tyto matice se startovacími vektory s, resp. $A^{\mathsf{H}}s/\|A^{\mathsf{H}}s\|_2$; viz např. [26, kap. 7.3].

4.2 Vybrané aspekty chování některých krylovovských metod

Mezi nejznámější krylovovské metody, sloužící k řešení soustav lineárních rovnic se čtvercovou regulární maticí, patří metoda sdružených gradientů (CG; viz [58]), použitelná na soustavy se symetrickou pozitivně definitní maticí a minimalizující A-normu chyby; metoda minimálních reziduí (MinRes; viz [91]), použitelná na soustavy s obecnou symetrickou maticí a minimalizující reziduum; nebo zobecněná metoda minimálních reziduí (GMRes; viz [104]), použitelná na soustavy s obecnou čtvercovou maticí a opět minimalizující reziduum. Při řešení aproximačních úloh, tj. úloh nejmenších čtverců, nebo např. regularizačních úloh se používají např. metody LSQR (viz [92] a [93]), Craigova metoda (viz [24]), nebo LSMR (viz [32]), o kterých již byla řeč v kap. 2. Rodina krylovovských metod je nicméně mnohem bohatší a existuje celá řada dalších algoritmů, viz např. [26, kap. 8 a 9], případně [38, kap. 10 a 11].

V průběhu posledních desetiletí bylo investováno enormní úsilí do analýzy chování mnoha různých krylovovských metod v přesné aritmetice ale i jejich numerickému chování v počítačům vlastní aritmetice s plovoucí řádovou čárkou a omezenou přesností (FPA). Byly analyzovány paměťové nároky těchto metod, efektivita a stabilita implementací, atd; viz např. [84], [83], nebo [115]. Tento text si ani v nejmenší míře neklade za cíl provádět jakoukoliv rekapitulaci, výtah, či rešerši těchto výsledků. Nicméně v řadě již zmíněných prací jsme se různých krylovovských metod dotkli, používali je, případně dosadili drobné střípky vlastní analýzy do mnohem rozsáhlejší mozaiky chování těchto metod. Rádi bychom nyní velmi stručně tři již zmíněné příspěvky zrekapitulovali a navíc zmínili jeden, o kterém dosud nebyla řeč.

Vlastní přínos

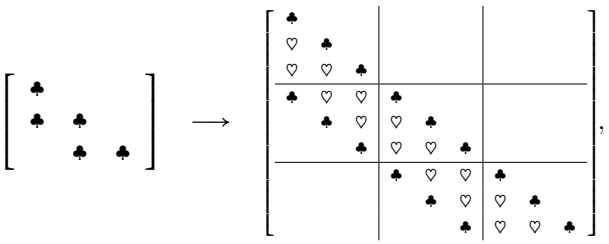
V krátkém příspěvku [65] jsme ukázali, jak spolu souvisí Golubova–Kahanova bidiagonalizace, Lanczosova tridiagonalizace a core problém. K vyjevení core problému v dané úloze dojde tehdy, pokud je pravá strana některé z rovnic (4.4) nebo (4.5) nulová, tj. buď $\beta_{\ell+1} = 0$, nebo $\alpha_{\ell+1} = 0$, což způsobí zastavení algoritmu bidiagonalizace. Dokud tento krok nenastane, máme k dispozici jen část celého core problému. V průběhu algoritmu tedy máme v rukou stále se postupně zlepšující aproximaci core problému. Právě toto je mechanizmus, který pomáhá pochopit, proč jsou metody jako např. LSQR tak úspěšné a proč tak dobře fungují. Tím, že stojí na bidiagonalizaci, tedy na stále lepší a lepší aproximaci core problému – tedy podproblému, který obsahuje veškeré nutné informace a zároveň informace právě postačující k řešení celé úlohy; viz kap. 1.3. Diskuze týkající se aproximačních vlastností LSQR ve vztahu ke core problému však v článku [65] přímo není a stejnou měrou bylo lze tuto souvislost tušit již v původním článku [96], kde byla idea core problému poprvé zformulována.

Druhým naším příspěvkem byla analýza chování Golubovy–Kahanovy bidiagonalizace v případě, že matice A je špatně podmíněná a pravá strana b obsahuje krom užitečných dat navíc nějaké chyby, resp. šum v článku [64]. Ukazujeme, jak se chyby propagují skrze algoritmus, jak se přenášejí mezi jednotlivými iteracemi a, díky špatné podmíněnosti matice, zesilují na úkor původních dat, o nichž přepdokládáme, že jsou hladká a dominována nízkými frekvencemi.

Třetím příspěvkem, který už byl také zmíněn, je analýza chování metody sdružených gradientů (CG), tedy algoritmu postaveném na Lancozosově tridiagonalizaci v prostředí s low-rank aritmetikou. V článku [76] studujeme, jak se mění hodnosti low-rank objektů v průběhu algoritmu CG, studujeme, jak agresivně lze nízkou

hodnost objektů vynucovat a jaký to má vliv na velikost rezidua. Věnujeme se zde také studiu vlivu *předpodmiňovačů* na hodnosti low-rank objektů, přípustnou míru agresivity při mechanickém snižování hodnosti, i velikost rezidua. Výsledky zde prezentované jsou obzvláště zajímavé a důležité proto, že objekty, se kterými pracujeme není možné uložit v paměti hustě a jejich low-rank aproximace je z principu nutná. Navíc možnost udržovat jejich hodnost relativně nízkou rapidním způsobem sníží nejen paměťové ale také časové nároky na výpočet.

Poslední příspěvek se týká tzv. *wedge-shaped* neboli *klínových matic*. V praktických numerických výpočtech se, v souladu s obecným trendem výpočetní lineání algebry několika posledních desetiletí, obecně přechází od formulací vektorových k formulacím blokovým, tj. maticovým. Řada metod, včetně algoritmu Lanczosovy tridiagonalizace a Golubovy–Kahanovy bidiagonalizace, tedy může být přeformulována blokově. Vezmeme-li např. rovnici (4.4), můžeme na pravé straně vektory w_ℓ a s_ℓ nahradit bloky (navzájem ortonormálních) vektorů a koeficient α_ℓ horní trojúhelníkovou maticí vhodných rozměrů. Objekt, který na pravé straně vznikne, bude maticí. Objekty na levé straně pak získáme QR rozkladem této matice, tj. $s_{\ell+1}$ bude blokem navzájem ortonormálních vektorů a koeficient $\beta_{\ell+1}$ bude horní trojúhelníkovou maticí. Analogicky můžeme naložit i s rovnicí (4.5). Místo bidiagonální matice tak dostáváme blokově bidiagonální matici s horními, resp. dolními (v rovnici (4.5) je systémová matice transponovaná) trojúhelníkovými bloky na diagonálách,

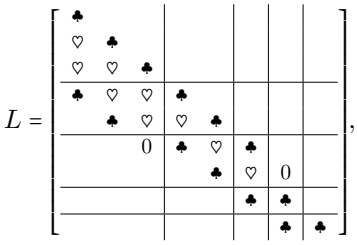


(4.8)

kde ♣ ≠ 0 a ♡ může být nulové i nenulové číslo.

Takto však bude zobecnění fungovat jen v ideálním případě, kdy bloky, které budou vycházet na pravých stranách rovnic (4.4) a (4.5), budou mít vždy více-či-méně lineárně nezávislé sloupce. Obecně v praktických výpočtech obsahujících zaokrouhlovací chyby můžeme předpokládat, že sloupce budou vždy lineárně nezávislé, navíc, bude-li úloha rozumně podmíněná, může být získaná aproximace použitelná. Zcela analogická je situace v případě blokové Lanczovy tridiagonalizace; viz např. [25], [121], [36], [102], nebo [5, kap. 4.6 a 7.10 (R. W. Freund, Band Lanczos Method)]. Pokud dochází k tomu, že vektory jsou v rámci bloku skoro lineárně závislé, provádí se tzv. *deflace* (přesněji *inexact deflation*, neboť se pohybujeme v aritmetice s plovoucí řádovou čárkou a omezenou přesností); viz např. [1].

V článku [68] jsme se však potřebovali vydat jiným směrem a provedli jsme tak detailní analýzu chování blokové Golubovy–Kahanovy bidiagonalizace (zapsané ovšem v nepatrně odlišné, tzv. pásové formě) v *přesné aritmetice* včetně anlýzy *přesných deflací*. Výstupem algoritmu je pricnipiálně analogický tvar jako má matice (4.8) vpravo, jen horní (resp. dolní) trojúhelníkové matice mají obecně horní (resp. dolní) schodovitý tvar, např.



(4.9)

kde schody (zde na dvou místech vyznačených nulami) indikují ony deflace. Tedy okamžiky – iterace, kdy se uzavírají jednotlivé podprostory invariantní vůči akcím operátorů AA^{H}, nebo $A^{\mathsf{H}}A$.

Ukázali jsme, že matice LL^{H}, $L^{\mathsf{H}}L$ mají speciální tvar, přesněji strukturu nenulových prvků, kterou jsme nazvali *wedge-shaped* (česky *klínová matice*); viz [68, kap. 4, str. 425–431]. (Nelze zde přitom použít klasický výsledek, že obálky pozitivně definitní matice a jejího Choleského faktoru jsou identické (viz [33, kap. 4.2]), neboť tento výsledek platí pouze pro regulární matice; zde jsou matice LL^{H}, $L^{\mathsf{H}}L$ obecně semidefinitní, přičemž mohou tedy být i singulární. Snadno lze najít protipříklady prve zmiňovaného výsledku o obálkách; viz [68, Fig. 1, str. 429].)

Ukázali jsme, že tyto matice jsou v jistém smyslu zobecněním Jacobiho tridiagonálních matic. Zobecňují totiž řadu jejich vlastností. Jacobiho tridiagonální matice mají například:

(i) jednoduchá vlastní čísla a

(ii) jejich vlastní vektory mají nenulové první

(iii) a poslední prvky,

(iv) navíc žádné dva po sobě jdoucí prvky vlastního vektoru nemohou být nulové;

viz např. [132], nebo [97]. První dvě vlastnosti byly zobecněny již v článku [68]. Ukázali jsme, že wedge-shaped matice s šířkou pásu ϱ (tj. s $2\varrho + 1$ nenulovými diagonálami) mají (i) vlastní čísla s násobností nevýše ϱ. Dále, vezmeme-li libovolnou bázi libovolného vlastního prostoru a sestavíme do matice, pak (ii) prvních ϱ řádků této matice tvoří blok s lineárně nezávislými sloupci. Podobně, ale poněkud složitěji lze zobecnit i vlastnost (iii) nenulového posledního a (iv) dvou po sobě jdoucích prvků. V článku [68] jsou první dvě důležité pro odvození vlastností core problémů. Plně se pak problematice věnujeme v článku [61]. Zde jsou zobecněny a dokázány všechny výše zmiňované vlastnosti (i)–(iv). Tyto matice a jejich vlastnosti tak snad mohou být užitečné a najít uplatnění při popisu dalším pochopení chování blokových krylovovských algoritmů.

Původní autorské články vztahující se k tomuto tématu publikované v impaktovaných časopisech jsou přiloženy, konkrétně:

[64] (2009) jako příloha A.5 na str. 131 (již zmíněno jako hlavní výsledek kap. 2);

[76] (2014) jako příloha A.6 na str. 161 (již zmíněno jako hlavní výsledek kap. 3);

[61] (2015) jako příloha A.7 na str. 183.

Závěr

Ve čtyřech předchozích kapitolách jsme se pokusili velmi stručně shrnout nejdůležitější z výsledků odborné práce habilitanta v oblasti numerické a aplikované lineární algebry, zejména pak maticových a tenzorových výpočtů. Výsledky jsou navzájem provázané mnoha různými souvislostmi, přesto jsme si však dovolili je rozčlenit do několika více-či-méně samostatných témat soustředících se kolem lineárních aproximačních úloh, regularizačních metod, tenzorových rozkladů a práce s nimi a krylovovských metod obecně.

Text je doprovozen přílohami, rozdělenými do pěti tématických sekcí z nichž první dvě odpovídají dvěma hlavním sekcím první kapitoly: řešitelnosti TLS problémů a core problémům, a každá ze zbývajících třech odpovídá jedné z dalších kapitol. Přílohy obsahují *sedm* článků publikovaných v impaktovaných časopisech vesměs zařazených do první (Q1) nebo druhé (Q2) čtvrtiny nejlépe hodnocených periodik. Většina článků (šest) je dohledatelných v databázi ISI Web-of-Knowledge (WoK), zbývající sedmý byl publikován v červenci 2016 a tudíž zatím není ve WoK dohledatelný. Těchto sedm článků má celkem *28 citací* (bez autocitací[2]) v publikacích, které buď již jsou (21 citací) evidované ve WoK, nebo byly přijaty a publikovány v impaktovaných časopisech, tj. zařazených do WoK. Celkový ohlas těchto článků zahrnuje *46 citací* (bez autotcitací), kam navíc zahrnujeme citace v monografiích nezarhnutých do WoK, preprinty uložené v arXivu, preprinty významných zahraničních univerzit, nebo dizertační práce obhájené na významných zahraničních univerzitách; viz detailní seznam v příloze A na str. 49, 75, 127, 159 a 181. Všichni spoluautoři všech sedmi článků přispěli k jejich vzniku rovným dílem.

[2]Autocitací, v souladu s WoK, rozumíme citaci provedenou autorem samým (tj. habilitantem), nikoliv některým ze spoluautorů; citující článek publikovaný některým ze spoluautorů bez autorské účasti habilitanta, není dle WoK považován za autocitaci. Ve výše uvedených 28 (resp. 46) citacích je 5 (resp. 8) těchto případů.

Literatura

[1] J. I. Aliaga, D. L. Boley, R. W. Freund, V. Hernández: *A Lanczos-type method for multiple starting vectors*, Mathematics of Computation, Vol. 69 (2000), pp. 1577–1601.
doi:10.1090/S0025-5718-99-01163-1

[2] B. W. Bader, T. G. Kolda: *Algorithm 862: MATLAB tensor classes for fast algorithm prototyping*, ACM Transactions on Mathematical Software, 32(4) (2006), pp. 635–653.
doi:10.1145/1186785.1186794

[3] B. W. Bader, T. G. Kolda: *Efficient MATLAB computations with sparse and factored tensors*, SIAM Journal on Scientific Computing, 30(1) (2007), pp. 205–231.
doi:10.1137/060676489

[4] B. W. Bader, T. G. Kolda: *MATLAB Tensor Toolbox, version 2.6.*, 2015. Available on http://www.sandia.gov/~tgkolda/TensorToolbox

[5] Z. Bai, J. Demmel, J. J. Dongarra, A. Ruhe, H. A. van der Vorst (eds.): *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM Publications, Philadelphia, PA, 2000.
doi:10.1137/1.9780898719581

[6] U. Baur, P. Benner: *Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic*, Computing, 78(3) (2006), pp. 211–234.
doi:10.1007/s00607-006-0178-y

[7] M. Bebendorf: *Hierarchical Matrices. A Means to Efficiently Solve Elliptic Boundary Value Problems*, Lecture Notes in Computational Science and Engineering 63, Springer Verlag, Berlin, Heidelberg, 2008.
doi:10.1007/978-3-540-77147-0

[8] P. Benner, J. R. Li, N. Truhar: *On the ADI method for Sylvester equations*, Journal of Computational and Applied Mathematics, 233(4) (2009), pp. 1035–1045.
doi:10.1016/j.cam.2009.08.108

[9] P. Benner, H. Mena, J. Saak: *On the parameter selection problem in the Newton-ADI iteration for large-scale Riccati equations*, Electronic Transactions on Numerical Analysis, 29 (2008), pp. 136–149.

[10] P. Benner, E. S. Quintana-Ortí: *Solving stable generalized Lyapunov equations with the matrix sign function*, Numerical Algorithms, 20(1) (1999), pp. 75–100.
doi:10.1023/A:1019191431273

[11] Å. BJÖRCK: *Bidiagonal Decomposition and Least Squares*, Presentation, Canberra, Australia, 2005.

[12] Å. BJÖRCK: *A Band-Lanczos Generalization of Bidiagonal Decomposition*, Presentation, Conference in Honor of G. Dahlquist, Stockholm, Sweden, 2006.

[13] Å. BJÖRCK: *A band-Lanczos algorithm for least squares and total least squares problems*, in Book of Abstracts of 4th Total Least Squares and Errors-in-Variables Modeling Workshop, Leuven, Katholieke Universiteit Leuven, Leuven, Belgium, 2006, pp. 22–23.

[14] Å. BJÖRCK: *Block bidiagonal decomposition and least squares problems with multiple right-hand sides*, unpublished manuscript.

[15] Å. BJÖRCK: *Least Squares Methods*, in Handbook of Numerical Analysis, Vol. I: Finite Difference Methods; Solution of Equations in R^n (P. G. CIARLET, J. L. LIONS eds.), North-Holland, Amsterdam, 1990.

[16] Å. BJÖRCK: *Numerical Methods for Least Squares Problem*, SIAM Publications, Philadelphia, PA, 1996.
doi:10.1137/1.9781611971484

[17] Å. BJÖRCK: *Numerical Methods in Matrix Computations*, Springer Verlag, Heidelberg, 2015.
doi:10.1007/978-3-319-05089-8

[18] S. BÖRM: *$\mathcal{H}_2$-Matrices – An Efficient Tool for the Treatment of Dense Matrices*, Habilitationsschrift, Christian-Albrechts-Universität zu Kiel, 2006.

[19] S. BÖRM: *Efficient Numerical Mehtods for Non-local Operators. $\mathscr{H}^2$-Matrix Compression, Algorithms and Analysis*, EMS Tracts in Mathematics Vol. 14, Zürich, 2010.
doi:10.4171/091

[20] D. CALVETTI, G. H. GOLUB, L. REICHEL: *Estimation of the L-curve via Lanczos bidiagonalization*, BIT Numerical Mathematics, 39(4) (1999), pp. 603–619.
doi:10.1023/A:1022383005969

[21] D. CALVETTI, S. MORIGI, L. REICHEL, F. SGALLARI: *Tikhonov regularization and the L-curve for large discrete ill-posed problems*, Journal of Computational and Applied Mathematics, 123(1–2) (2000), pp. 423–446.
doi:10.1016/S0377-0427(00)00414-3

[22] D. CALVETTI, L. REICHEL, A. SHUIBI: *L-curve and curvature bounds for Tikhonov regularization*, Numerical Algorithms, 35(2) (2004), pp. 301–314.
doi:10.1023/B:NUMA.0000021764.16526.47

[23] J. CHUNG, J. G. NAGY, D. P. O'LEARY: *A weighted GCV method for Lanczos hybrid regularization*, Electronic Transactions on Numerical Analysis, 28 (2008), pp. 149–167.

[24] E. J. CRAIG: *The N-step iteration procedures*, Studies in Applied Mathematics, 34(1–4) (1955), pp. 64–73.
doi:10.1002/sapm195534164

[25] J. K. CULLUM, W. E. DONATH: *A Block Generalization of the Symmetric S-Step Lanczos algorithm*, Rep. No. RC 4845, IBM, Thomas J. Watson Res. Center, Yorktown Heights, New York (1974).

[26] J. Duintjer Tebbens, I. Hnětynková, M. Plešinger, Z. Strakoš, P. Tichý: *Analýza metod pro maticové výpočty: základní metody*, Matfyzpress, Praha, 2012.

[27] C. Eckart, G. Young: *The approximation of the matrix by another of lower rank*, Psychometrica, 1(3) (1936), pp. 211–218.
doi:10.1007/BF02288367

[28] L. Eldén: *Matrix methods in data mining and pattern recognition*, SIAM Publications, Philadelphia, 2007.
doi:10.1137/1.9780898718867

[29] M. Espig, K. K. Naraparaju, J. Schneider: *A note on tensor chain approximation*, Computing and Visualization in Science, 15(6) (2012), pp. 331–344.
doi:10.1007/s00791-014-0218-7
Available on http://www.mis.mpg.de/preprints/2012/preprint2012_16.pdf

[30] M. Fedi, P. C. Hansen, V. Paoletti: *Analysis of depth resolution in potential field inversion*, Geophysics, 70(6) (2005), pp. A1–A11.
doi:10.1190/1.2122408

[31] R. D. Fierro, G. H. Golub, P. C. Hansen, D. P. O'Leary: *Regularization by truncated total least squares*, SIAM Journal on Scientific Computing, 18(4) (1997), 1223–1241.
doi:10.1137/S1064827594263837

[32] D. C. L. Fong, M. A. Saunders: *LSMR: An iterative algorithm for sparse least-squares problems*, SIAM Journal on Scientific Computing, 33(5) (2011), pp. 2950–2971.
doi:10.1137/10079687X

[33] A. George, J. W. H. Liu: *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981;

[34] G. H. Golub, M. Heat, G. Wahba: *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics 21(2) (1997), pp. 215–223.
doi:10.2307/1268518

[35] G. H. Golub, W. Kahan: *Calculating the singular values and pseudo-inverse of a matrix*, Journal of SIAM Series B Numerical Analysis, 2(2) (1965), pp. 205–224.
doi:10.1137/0702016

[36] G. H. Golub, R. R. Underwood: *The block Lanczos method for computing eigenvalues*, in Mathematical Software, Vol. 3 (J. R. Rice ed.), Academic Press, New York, 1977, pp. 364–377.

[37] G. H. Golub, C. F. Van Loan: *An analysis of the total least squares problem*, SIAM Journal on Numerical Analysis, 17(6) (1980), pp. 883–893.
doi:10.1137/0717073

[38] G. H. Golub, C. F. Van Loan: *Matrix Computations*, 4th Edition, Johns Hopkins University Press, Baltimore, MD, 2013.

[39] G. H. Golub, U. Von Matt: *Generalized cross-validation for large scale problems*, Journal of Computational and Graphical Statistics, 6(1) (1997), pp. 1–34.
doi:10.2307/1390722

[40] L. Grasedyck: *Hierarchical singular value decomposition of tensors*, SIAM Journal on Matrix Analysis and Applications, 31(4) (2010), pp. 2029–2054.
doi:10.1137/090764189

[41] L. Grasedyck, D. Kressner, C. Tobler: *A literature survey of low-rank tensor approximation techniques*, GAMM Mitteilungen, 36(1) (2013), pp. 53–78.
doi:10.1002/gamm.201310004

[42] W. Hackbusch: *Hierarchische Matrizen: Algorithmen und Analysis*, Springer Verlag, Berlin, Heidelberg, 2009.
doi:10.1007/978-3-642-00222-9

[43] W. Hackbusch, S. Kühn: *A new scheme for the tensor representation*, Journal of Fourier Analysis and Applications, 15(5) (2009), pp. 706–722.
doi:10.1007/s00041-009-9094-9

[44] J. S. Hadamard: *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*, Leçons professées à l'Université Yale, Herman et C^{ie} Éditeurs, Paris, 1932.

[45] M. Hanke: *On Lanczos based methods for regularization of discrete ill-posed problems*, BIT Numerical Mathematics, 41(5) (2001), pp. 1008–1018.
doi:10.1023/A:1021941328858

[46] P. C. Hansen: *The discrete Picard condition for discrete ill-posed problems*, BIT Numerical Mathematics, 30(4) (1990), pp. 658–672.
doi:10.1007/BF01933214

[47] P. C. Hansen: *Analysis of Discrete Ill-Posed Problems by Means of the L-Curve*, SIAM Review 34(4) (1992), pp. 561–580.
doi:10.1137/1034115

[48] P. C. Hansen: *Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems*, Numerical Algorithms, 6(1) (1994), pp. 1–35.
doi:10.1007/BF02149761

[49] P. C. Hansen: *Rank Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM Publications, Philadelphia, PA, 1996.
doi:10.1137/1.9780898719697

[50] P. C. Hansen: *Regularization Tools Version 4.0 for Matlab 7.3*, Numerical Algorithms, 46(2) (2007), pp. 189–194.
doi:10.1007/s11075-007-9136-9
Available on http://www.imm.dtu.dk/~pcha/Regutools

[51] P. C. Hansen: *Discrete Inverse Problems: Insight and Algorithms*, SIAM Publications, Philadelphia, PA, 2010.
doi:10.1137/1.9780898718836

[52] P. C. HANSEN, T. K. JENSEN: *Smoothing-norm preconditioning for regularizing minimum-residual methods*, SIAM Journal on Matrix Analysis and Applications, 29(1) (2006), pp. 1–14].
doi:10.1137/050628453

[53] P. C. HANSEN, V. PEREYRA, G. SCHERER: *Least Squares Data Fitting with Applications*, SIAM Publications, Philadelphia, PA, 2013.

[54] P. C. HANSEN, M. SAXILD-HANSEN: *AIR Tools — A MATLAB package of algebraic iterative reconstruction methods*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 2167–2178.
doi:10.1016/j.cam.s011.09.039.

[55] P. C. HANSEN, M. SAXILD-HANSEN, ET AL.: *AIR Tools — A MATLAB package of algebraic iterative reconstruction methods*, Version 1.3, revised version, 2015.
Available on http://www.imm.dtu.dk/~pcha/AIRtools

[56] P. C. HANSEN, H. O. SØRENSEN, Z. SÜKÖSD, H. F. POULSEN: *Reconstruction of single-grain orientation distribution functions for crystalline materials*, SIAM Journal on Imaging Sciences, 2(2) (2009), pp. 593–613.
doi:10.1137/080726021

[57] P. C. HANSEN, J. NAGY, D. P. O'LEARY: *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM Publications, Philadelphia, PA, 2006.
doi:10.1137/1.9780898718874

[58] M. R. HESTENES, E. STIEFEL: *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49(6) (1952), pp. 409–435.
doi:10.6028/jres.049.044

[59] I. HNĚTYNKOVÁ, M. KUBÍNOVÁ, M. PLEŠINGER: *Notes on performance of bidiagonalization-based noise level estimator in image deblurring*, in Proceedings of Algoritmy 2016 Conference (A. HANDLOVIČOVÁ ed.), Slovak University of Technology in Bratislava, Publishing House of STU, 2016, pp. 333–342.
Available on http://www.iam.fmph.uniba.sk/amuc/ojs [...]
/index.php/algoritmy/article/view/422

[60] I. HNĚTYNKOVÁ, M. KUBÍNOVÁ, M. PLEŠINGER: *On noise propagation in residuals of Krylov subspace iterative regularization methods*, in preparation, 21 pages.

[61] I. HNĚTYNKOVÁ, M. PLEŠINGER: *Complex wedge-shaped matrices: A generalization of Jacobi matrices*, Linear Algebra and its Applications, 487 (2015), pp. 203–219.
doi:10.1016/j.laa.2015.09.017

[62] I. HNĚTYNKOVÁ, M. PLEŠINGER, D. M. SIMA: *Solvability of the core problem with multiple right-hand sides in the TLS sense*, SIAM Journal on Matrix Analysis and Applications, 37(3) (2016), pp. 861–876.
doi:10.1137/15M1028339

[63] I. HNĚTYNKOVÁ, M. PLEŠINGER, D. M. SIMA, Z. STRAKOŠ, S. VAN HUFFEL: *The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works*, SIAM Journal on Matrix Analysis and Applications, 32(3) (2011), pp. 748–770.
doi:10.1137/100813348

[64] I. HNĚTYNKOVÁ, M. PLEŠINGER, Z. STRAKOŠ: *The regularizing effect of the Golub–Kahan iterative bidiagonalization and revealing the noise level in the data*, BIT Numerical Mathematics, 49(4) (2009), pp. 669–696.
doi:10.1007/s10543-009-0239-7

[65] I. HNĚTYNKOVÁ, M. PLEŠINGER, Z. STRAKOŠ: *Lanczos tridiagonalization, Golub–Kahan bidiagonalization and core problem*, Proceedings in Applied Mathematics and Mechanics, 6(1) (2006), pp. 717–718.
doi:10.1002/pamm.200610339

[66] I. HNĚTYNKOVÁ, M. PLEŠINGER, Z. STRAKOŠ: *On solution of total least squares problems with multiple right-hand sides*, Proceedings in Applied Mathematics and Mechanics, 8(1) (2008), pp. 10815–10816.
doi:10.1002/pamm.200810815

[67] I. HNĚTYNKOVÁ, M. PLEŠINGER, Z. STRAKOŠ: *The core problem within linear approximation problem $AX \approx B$ with multiple right-hand sides*, SIAM Journal on Matrix Analysis and Applications, 34(3) (2013), pp. 917–931.
doi:10.1137/120884237

[68] I. HNĚTYNKOVÁ, M. PLEŠINGER, Z. STRAKOŠ: *Band generalization of the Golub–Kahan bidiagonalization, generalized Jacobi matrices, and the core problem*, SIAM Journal on Matrix Analysis and Applications, 36(2) (2015), pp. 417–434.
doi:10.1137/140968914

[69] I. HNĚTYNKOVÁ, M. PLEŠINGER, J. ŽÁKOVÁ: *Filter factors of truncated TLS regularization with multiple observations*, submitted to Application of Mathematics in August 15, 2016, 15 pages.

[70] I. HNĚTYNKOVÁ, M. PLEŠINGER, J. ŽÁKOVÁ: *Tensor generalizations of the total least squares problem*, in preparation, 26 pages.

[71] I. HNĚTYNKOVÁ, M. PLEŠINGER, J. ŽÁKOVÁ: *Základní nástroje tenzorových výpočtů. Tenzorové rozklady*, rozpracované skriptum, 133 stran.

[72] M. HOCHBRUCK AND G. STARKE: *Preconditioned Krylov subspace methods for Lyapunov matrix equations*, SIAM Journal on Matrix Analysis and Applications, 16(1) (1995), pp. 156–171.
doi:10.1137/S0895479892239238

[73] T. G. KOLDA, B. W. BADER: *Tensor decompositions and applications*, SIAM Review, 51(3) (2009), pp. 455–500.
doi:10.1137/07070111X

[74] J. G. KORVINK, B. R. EVGENII: *Oberwolfach benchmark collection*, in Dimension Reduction of Large-Scale Systems (P. BENNER, V. MEHRMANN, D. C. SORENSEN eds.), Lecture Notes in Computational Science and Engineering 45, Springer Verlag, Heidelberg, 2005, pp. 311–316.
doi:10.1007/3-540-27909-1 (LNCSE45)
Available on https://portal.uni-freiburg.de/imteksimulation [...] /downloads/benchmark

[75] D. KRESSNER, M. PLEŠINGER, C. TOBLER: *A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations*, EFPL Mathicse technical report № 18.2012 (preprint), 2012, 20 pages.
Available on http://mathicse.epfl.ch/files/content/sites [...] /mathicse/files/Mathicse reports 2012/18.2012_DK-MP-CT.pdf

[76] D. KRESSNER, M. PLEŠINGER, C. TOBLER: *A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations*, Numerical Linear Algebra with Applications, 21(5) (2014), pp. 666–684.
doi:10.1002/nla.1919

[77] D. KRESSNER, C. TOBLER: *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM Journal on Matrix Analysis and Applications, 32(4) (2011), pp. 1288–1316.
doi:10.1137/100799010

[78] D. KRESSNER, C. TOBLER: *Algorithm 941: htucker—A Matlab toolbox for tensors in hierarchical Tucker format*, ACM Transactions on Mathematical Software, 40(3) (2014), article 22, 22 pages.
doi:10.1145/2538688
Available on http://anchp.epfl.ch/htucker

[79] D. KRESSNER, C. TOBLER: *htucker — A MATLAB toolbox for tensors in hierarchical Tucker format*, Technical Report 2012-02, Seminar for Applied Mathematics, ETH Zurich, 2012.
Available on http://anchp.epfl.ch/htucker and
http://sma.epfl.ch/~anchpcommon/publications/htucker_manual.pdf

[80] C. L. LAWSON, R. J. HANSON: *Solving Least Squares Problems* SIAM Publications, Philadelphia, PA, 1995.
doi:10.1137/1.9781611971217

[81] J.-R. LI, J. WHITE: *Low rank solution of Lyapunov equations*, SIAM Journal on Matrix Analysis and Applications, 24(1) (2002), pp. 260–280.
doi:10.1137/S0895479801384937

[82] J.-R. LI, J. WHITE: *Low-rank solution of Lyapunov equations*, SIAM Review, 46(4) (2004), pp. 693–713.
doi:10.1137/S0036144504443389

[83] J. LIESEN, Z. STRAKOŠ: *Krylov Subspace Methods. Principles and Analysis*, Numerical Mathematics and Scientific Computation, Oxford University Press, 2013.

[84] G. MEURANT: *The Lanczos and Conjugate Gradient Algorithms. From Theory to Finite Precision Computations*, SIAM Publications, Philadelphia, PA, 2006.
doi:10.1137/1.9780898718140

[85] L. MIRSKY: *Symmetric gauge functions and unitarily invariant norms*, The Quarterly Journal of Mathematics, 11(1) (1960), pp. 50–59.
doi:10.1093/qmath/11.1.50

[86] V. A. MOROZOV: *On the solution of functional equations by the method of regularization* (in Russian), Soviet Mathematics – Doklady, 7 (1966), pp. 414–417.

[87] V. A. MOROZOV: *Methods for Solving Incorrectly Posed Problems*, Springer, New York, 1984.
doi:10.1007/978-1-4612-5280-1

[88] F. NATTERER: *The Mathematics of Computerized Tomography*, B. G. Teubner, Stuttgart, 1986.

[89] N. NGUYEN, P. MILANFAR, G. H. GOLUB: *Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement*, IEEE Trans. Image Proces. 10(9) (2001), pp. 1299–1308.
doi:10.1109/83.941854

[90] I. V. OSELEDETS: *Tensor-train decomposition*, SIAM Journal on Scientific Computing, 33(5) (2011), Number 5, pp. 2295–2317.
doi:10.1137/090752286

[91] C. C. PAIGE, M. A. SAUNDERS: *Solutions of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12(4) (1975), pp. 617–629.
doi:10.1137/0712047

[92] C. C. PAIGE, M. A. SAUNDERS: *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Transactions on Mathematical Software, 8(1) (1982), pp. 43–71.
doi:10.1145/355984.355989

[93] C. C. PAIGE, M. A. SAUNDERS: *ALGORITHM 583 LSQR: Sparse Linear equations and least squares problems*, ACM Transactions on Mathematical Software, 8(2) (1982), pp. 195–209.
doi:10.1145/355993.356000

[94] C. C. PAIGE, Z. STRAKOŠ: *Scaled total least squares fundamentals*, Numerische Mathematik, 91(1) (2002), pp. 117–146.
doi:10.1007/s002110100314

[95] C. C. PAIGE, Z. STRAKOŠ: *Unifying least squares, total least squares and data least squares*, in Total Least Squares and Errors-in-Variables Modeling (S. VAN HUFFEL, P. LEMMERLING eds.), Kluwer Academic Publishers, Dordrecht, 2002, pp. 25–34.
doi:10.1007/978-94-017-3552-0_3

[96] C. C. PAIGE, Z. STRAKOŠ: *Core problem in linear algebraic systems*, SIAM Journal on Matrix Analysis and Applications, 27(3) (2006), pp. 861–875.
doi:10.1137/040616991

[97] B. N. PARLETT: *The Symmetric Eigenvalue Problem*, SIAM Publications, Philadelphia, PA, 1998.
doi:10.1137/1.9781611971163

[98] T. PENZL: *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM Journal on Scientific Computing, 21(4) (1999), pp. 1401–1418.
doi:10.1137/S1064827598347666

[99] M. PLEŠINGER: *The Total Least Squares Problem and Reduction of Data in* $AX \approx B$, Ph.D. thesis, Technická univerzita v Liberci, Liberec, 2008.

[100] J. K. G. RADON: *Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten*, Berichte über die Verhandlungen der Königlich-Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse 69 (1917), pp. 262–277.

[101] J. K. G. RADON, (P. C. PARKS translator): *On the determination of functions from their integral values along certain manifolds*, IEEE Transactions on Medical Imaging, 5(4) (1986), pp. 170–176.
doi:10.1109/TMI.1986.4307775

[102] A. Ruhe: *Implementation aspects of band Lanczos algorithms for cmputation of eigenvalues of large sparse matrices*, Mathematics of Computation, 33 (1979), pp. 680–687.
doi:10.1090/S0025-5718-1979-0521282-9

[103] B. W. Rust, D. P. O'Leary: *Residual periodograms for choosing regularization parameters for ill-posed problems*, Inverse Problems, 24(3) (2008), paper No. 034005.
doi:10.1088/0266-5611/24/3/034005

[104] Y. Saad, M. H. Schultz: *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7(3) (1986), pp. 856–869.
doi:10.1137/0907058

[105] M. A. Saunders: *Solution of sparse rectangular systems using LSQR and CRAIG*, BIT Numerical Mathematics, 35(4) (1995), pp. 588–604.
doi:10.1007/BF01739829

[106] E. Schmidt: *Zur Theorie der linearen un nichtlinearen Integralgleichungen. I. Tiel. Entwicklung willkürlichen Funkcionen nach System vorgechriebener*, Mathematische Annalen, 63 (1907), pp. 433–476.

[107] C. B. Shaw, Jr.: *Improvements of the resolution of an instrument by numerical solution of an integral equation*, Journal of Mathematical Analysis and Applications, 37(1) (1972), pp. 83–112.
doi:10.1016/0022-247X(72)90259-4

[108] D. M. Sima: *Regularization Techniques in Model Fitting and Parameter Estimation*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.

[109] N. T. Son, T. Stykel: *Solving parameter-dependent Lyapunov equations using reduced basis method with application to parametric model order reduction*, preprint of Augsburg university, Augsburg, 2015, 26 pages.
Availabe on https://opus.bibliothek.uni-augsburg.de/opus4 [...]
/files/3155/mpreprint_15_005.pdf

[110] L. Sorber, M. Van Barel, L. De Lathauwer: *Structured data fusion*, IEEE Journal of Selected Topics in Signal Processing, 9(4) (2015), pp. 586–600.
doi:10.1109/JSTSP.2015.2400415

[111] G. W. Stewart, J.-G. Sun: *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[112] Z. Strakoš: *Model reduction using the Vorobyev moment problem*, Numerical Algorithms, 51(3) (2009), pp. 363–379.
doi:10.1007/s11075-008-9237-0

[113] P. Tichavský, Z. Koldovský: *Weight adjusted tensor method for blind separation of underdetermined mixtures of nonstationary sources*, IEEE Transactions on Signal Processing, 59(3) (2011), pp. 1037–1047.
doi:10.1109/TSP.2010.2096221

[114] A. N. Tichonov, A. V. Gončarskij, V. V. Stepanov, A. G. Jagola, (J. Buša translator): *Numerické metódy riešenia nekorektných úloh*, Technická univerzita v Košiciach, Košice, 2000.

[115] P. TICHÝ: *Analysis of Krylov Subspace Methods*, habilitation thesis, Univerzita Karlova v Praze, Praha, 2015.

[116] C. TOBLER: *Low-rank tensor methods for linear systems and eigenvalue problems*, Ph.D. thesis, ETH Zürich, Zürich, 2012.
Available on http://sma.epfl.ch/~ctobler/diss.pdf

[117] L. R. TUCKER: *Implications of factor analysis of three-way matrices for measurement of change*, in Problems in Measuring Change (C. W. HARRIS ed.), University of Wisconsin Press, 1963, pp. 122–137.

[118] L. R. TUCKER: *The extension of factor analysis to three-dimensional matrices*, in Contributions to Mathematical Psychology (H. GULLIKSEN, N. FREDERIKSEN eds.), Holt, Rinehardt & Winston, New York, 1964, pp. 110–127.

[119] L. R. TUCKER: *Some mathematical notes on three-mode factor analysis*, Psychometrika 31(3) (1966), pp. 279–311.
doi:10.1007/BF02289464

[120] E. TYRTYSHNIKOV: *Numerical methods with tensor representations of data*, lecture of Summer Supercomputing Academy, 2012.
Available on http://academy2012.hpc-russia.ru/files [...]
/lectures/algebra/0703_1_tyrtyshnikov.pdf

[121] R. R. UNDERWOOD: *An Iterative Block Lanczos Method for the Solution of Large Sparse Symmetric Eigenproblems*, Ph.D. thesis, Stanford University, CA, 1975.

[122] S. VAN HUFFEL: *Documented Fortran 77 Programs of the Extended Classical Total Least Squares Algorithm, the Partial Singular Value Decomposition Algorithm and the Partial Total Least Squares Algorithm*, Internal report ESAT-KUL 88/1, Katholieke Universiteit Leuven, Leuven, Belgium 1988.

[123] S. VAN HUFFEL: *The extended classical total least squares algorithm*, Journal of Computational and Applied Mathematics, 25(1) (1989), pp. 111–119.
doi:10.1016/0377-0427(89)90080-0

[124] S. VAN HUFFEL (ed.): *Recent Advances in Total Least Squares Techniques and Error-in-Variables Modeling*, Proceedings of the Second Int. Workshop on TLS and EIV, SIAM Publications, Philadelphia, PA, 1997.

[125] S. VAN HUFFEL, P. LEMMERLING (eds.): *Total Least Squares and Error-in-Variables Modeling. Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[126] S. VAN HUFFEL, J. VANDEWALLE: *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM Publications, Philadelphia, PA, 1991.
doi:10.1137/1.9781611971002

[127] R. VANDEBRIL, M. VAN BAREL, N. MASTRONARDI: *Matrix Computations and Semiseparable Matrices*, Vol. 1., Johns Hopkins University Press, Baltimore, MD, 2008.

[128] N. VERVLIET, O. DEBALS, L. SORBER, M. VAN BAREL, L. DE LATHAUWER: *Tensorlab 3.0*, 2016.
Available on http://www.tensorlab.net

[129] X.-F. WANG: *Total least squares problem with the arbitrary unitarily invariant norms*, Linear and Multilinear Algebra, published online 29 May 2016.
doi:10.1080/03081087.2016.1189493

[130] M. WEI: *The analysis for the total least squares problem with more than one solution*, SIAM Journal on Matrix Analysis and Applications, 13(3) (1992), pp. 746–763.
doi:10.1137/0613047

[131] M. WEI: *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numerische Mathematik, 62(1) (1992), pp. 123–148.
doi:10.1007/BF01396223

[132] J. H. WILKINSON: *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, England, 1965 (reprint 2004).

[133] J. XIA, S. CHANDRASEKARAN, M. GU, X. S. LI: *Fast algorithms for hierarchically semiseparable matrices*, Numerical Linear Algebra with Applications, 17(6) (2010), pp 953–976.
doi:10.1002/nla.691

[134] J. ŽÁKOVÁ: *Tenzory a kanonické tenzorové rozklady: Tuckerův rozklad*, bakalářská práce, Technická univerzita v Liberci, Liberec, 2015.

Příloha A

Publikace, jejich citace a reprinty[3]

Řešitelnost TLS problému

A.1 **Článek:**[WoK] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, Sabine Van Huffel: *The total least squares problem in $AX \approx B$. A new classification with the relationship to the classical works*, SIAM Journal on Matrix Analysis and Applications (SIMAX) (ISSN 0895-4798, eISSN 1095-7162), Volume 32, Issue 3 (2011), pp. 748–770 (23 pages). (http://epubs.siam.org/doi/abs/10.1137/100813348)

Preprint: ETH SAM Research Report № 2010-38 (http://www.sam.math.ethz.ch/sam_reports/reports_final/reports2010 [...] /2010-38.pdf)

Doložitelné citace (9): Odborné články zařazené do databáze Web-of-Knowledge:

- [WoK] Z. Jia, B. Li: *On the condition number of the total least squares problem*, Numerische Mathematik, 125(1) (2013), pp. 61–78. (http://link.springer.com/article/10.1007/s00211-013-0533-9)
- [WoK] Z. Kong, S. Peng, Y. Zhang, L. Zhong: *EIV-based interference alignment scheme with CSI uncertainties*, Mathematical Problems in Engineering, 2015, Article ID 323214 (2015), 11 pages. (http://www.hindawi.com/journals/mpe/2015/323214)
- [WoK] N. Bagherpour, N. Mahdavi-Amiri: *A new error in variables model for solving positive definite linear system using orthogonal matrix decompositions*, Numerical Algorithm, 72(1) (2016), pp. 211–241. (http://link.springer.com/article/10.1007/s11075-015-0042-2)
- [WoK] S. Ito, K. Murota: *An algorithm for the generalized eigenvalue problem for nonsquare matrix pencils by minimal perturbation approach*, SIAM Journal on Matrix Analysis and Applications, 37(1) (2016), pp. 409–419. (http://epubs.siam.org/doi/abs/10.1137/14099231X)

[3] Publikace (článek, příspěvek, případně kniha) označená symbolem

WoK je zařazena v databázi ISI Web-of-Knowledge (https://www.webofknowledge.com);

(WoK) je publikována v časopise zařazeném, nebo sborníku obvykle zařazovaném do WoK, v databázi ale *zatím* (ke dni 14. listopadu 2016) není;

Scopus není ve WoK, je ale zařazena v databázi Scopus (https://www.scopus.com);

MathSciNet není ve WoK, je ale zařazena v databázi MathSciNet (http://www.ams.org/mathscinet).

- (WoK) X.-F. Wang: *Total least squares problem with the arbitrary unitarily invariant norms*, Linear and Multilinear Algebra, Available online, 2016, 20 pages. (http://www.tandfonline.com/doi/abs/10.1080/03081087.2016.1189493)

Další citace:

- PhD-thesis S. Park: *Matrix Reduction in Numerical Optimization*, PhD Thesis, Department of Computer Science, University of Maryland, College Park, MD, 2011. (http://drum.lib.umd.edu/handle/1903/11751)
- Book Scopus J. Liesen, Z. Strakoš[AutoCit]: *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, Oxford, 2013. (https://global.oup.com/academic/product [...] /krylov-subspace-methods-9780199655410)
- Report P. Novák, et al.: *Towards a better understanding of the Earth's interior and geological exploration research*, ESA Study Contract Report (final report), 2013. (http://www.lr.tudelft.nl/fileadmin/Faculteit/LR/Organisatie [...] /Afdelingen_en_Leerstoelen/Afdeling_SpE [...] /Astrodynamics_and_Space_Missions/Staff/doc [...] /final_report_blue__5_.pdf)
- Preprint P. Xie, Y. Wei, H. Xiang: *Perturbation analysis and randomized algorithms for large-scale total least squares problems*, arXiv:1401.6832, submitted 27 Jan 2014 (v1), 11 Nov 2014 (v2), 27 pages. (http://arxiv.org/abs/1401.6832)

SIAM J. MATRIX ANAL. APPL.
Vol. 32, No. 3, pp. 748–770

THE TOTAL LEAST SQUARES PROBLEM IN $AX \approx B$: A NEW CLASSIFICATION WITH THE RELATIONSHIP TO THE CLASSICAL WORKS*

IVETA HNĚTYNKOVÁ†, MARTIN PLEŠINGER‡, DIANA MARIA SIMA§, ZDENĚK STRAKOŠ†, AND SABINE VAN HUFFEL§

Abstract. This paper revisits the analysis of the total least squares (TLS) problem $AX \approx B$ with multiple right-hand sides given by Van Huffel and Vandewalle in the monograph, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991. The newly proposed classification is based on properties of the singular value decomposition of the extended matrix $[B|A]$. It aims at identifying the cases when a TLS solution does or does not exist and when the output computed by the classical TLS algorithm, given by Van Huffel and Vandewalle, is actually a TLS solution. The presented results on existence and uniqueness of the TLS solution reveal subtleties that were not captured in the known literature.

Key words. total least squares, multiple right-hand sides, linear approximation problems, orthogonally invariant problems, orthogonal regression, errors-in-variables modeling

AMS subject classifications. 15A24, 15A60, 65F20, 65F30

DOI. 10.1137/100813348

1. Introduction. This paper focuses on the *total least squares* (TLS) formulation of the linear approximation problem with *multiple right-hand sides*

$$AX \approx B, \quad A \in \mathbb{R}^{m\times n}, \quad X \in \mathbb{R}^{n\times d}, \quad B \in \mathbb{R}^{m\times d}, \quad A^T B \neq 0, \tag{1.1}$$

or, equivalently,

$$[B|A]\begin{bmatrix} -I_d \\ X \end{bmatrix} \approx 0. \tag{1.2}$$

We concentrate on the *incompatible problem* (1.1), i.e., $\mathcal{R}(B) \not\subset \mathcal{R}(A)$. The compatible case reduces to finding a solution of a system of linear algebraic equations. In TLS, contrary to the ordinary least squares, the correction is allowed to compensate for errors in the *system (data) matrix* A as well as in the *right-hand side (observation) matrix* B, and the matrices E and G are sought to minimize the Frobenius norm in

*Received by the editors October 28, 2010; accepted for publication (in revised form) by J. G. Nagy, April 19, 2011; published electronically July 28, 2011.
http://www.siam.org/journals/simax/32-3/81334.html

†Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, and Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic (hnetynkova@cs.cas.cz, strakos@cs.cas.cz). The research of these authors was supported by the research project MSM0021620839 financed by MŠMT ČR and by the GACR grant 201/09/0917.

‡Seminar for Applied Mathematics, Department of Mathematics, ETH Zurich, Zurich, Switzerland; Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic; and Faculty of Mechatronics, Technical University of Liberec, Liberec, Czech Republic (martin.plesinger@cs.cas.cz). The research of this author was supported by the GAAV grant IAA100300802 and by Institutional Research Plan AV0Z10300504.

§Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven (KUL), Leuven, Belgium, and IBBT-KUL Future Health Department, Leuven, Belgium (diana.sima@esat.kuleuven.be, sabine.vanhuffel@esat.kuleuven.be). The research of the third author, a postdoctoral fellow of the Fund for Scientific Research–Flanders, and the last author was supported by the Research Council KUL: GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), and the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization," 2007–2011).

$$\min_{X,E,G} \|[G|E]\|_F \quad \text{subject to } (A+E)X = B+G. \tag{1.3}$$

Throughout the paper, *any* matrix X which solves the corrected system in (1.3) is called a *TLS solution*. Similar to the ordinary least squares, we are often interested in TLS solutions *minimal in the* 2*-norm and/or in the Frobenius norm*.

Mathematically equivalent problems have been independently investigated in several areas as *orthogonal regression* and *errors-in-variables modeling*; see [18], [19]. It is worth noting that norms other than the Frobenius norm in (1.3) can also be relevant in practice; see, e.g., [20].

The TLS problem (1.1)–(1.3) has been investigated in its algebraic setting for decades; see the early works [6], [4, section 6], [14]. In [7] it is shown that even with $d = 1$ (which gives $Ax \approx b$, where b is an m-vector) the TLS problem may not have a solution and, when the solution exists, it may not be unique; see also [5, pp. 324–326]. The classical book [17] introduces the *generic–nongeneric* terminology representing the basic classification of TLS problems. If $d = 1$, then the *generic problems* simply represent problems that have a (possibly nonunique) solution, whereas *nongeneric problems* do not have a solution in the sense of (1.3). This is no longer true for multiple right-hand sides, where $d > 1$. The monograph [17] analyzes only two particular cases characterized by the special distribution of singular values of the extended matrix $[B|A]$. The so-called *classical TLS algorithm* presented in [17], however, for *any* A, B, computes some output X. The relationship of this output to the original problem is not always clear.

For $d = 1$, the TLS problem does not have a solution when the *collinearities among columns of* A *are stronger than the collinearities between* $\mathcal{R}(A)$ *and* b; see [9], [10], [11] for a recent description. An analogous situation may occur for $d > 1$, but here the difficulty can be caused for *different columns of* B *by different subsets of columns of* A. Therefore, it is no longer possible to stay with the generic–nongeneric classification of TLS problems. This is also the reason why the question remained open in [17]. In this paper we try to fill this gap and investigate existence and uniqueness of the TLS solution with $d > 1$ in full generality.

The organization of this paper is as follows. Section 2 recalls some basic results. Section 3 introduces *problems of what we call the* 1*st class*. After recalling known results for two special distributions of singular values in sections 3.1 and 3.2, we turn to the general case in section 3.3. The new classification is introduced in section 4. Section 5 introduces *problems of the* 2*nd class*. Section 6 links the new classification with the *classical TLS algorithm* from [17], and section 7 concludes the paper.

2. Preliminaries. As usual, $\sigma_j(M)$ denotes the jth largest singular value, $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote the range and the null space, $\|M\|_F$ and $\|M\|$ denote the Frobenius norm and the 2-norm of the given matrix M, respectively, and $M^\dagger$ denotes the Moore–Penrose pseudoinverse of M. Further, $\|v\|$ denotes the 2-norm of the given vector v; $I_k \in \mathbb{R}^{k\times k}$ denotes the k-by-k identity matrix.

In order to simplify the notation we assume, with no loss of generality, $m \geq n + d$ (otherwise, we can simply add zero rows). Consider the SVD of A, $r \equiv \operatorname{rank}(A)$,

$$A = U'\Sigma'(V')^T, \tag{2.1}$$

where $(U')^{-1} = (U')^T$, $(V')^{-1} = (V')^T$, $\Sigma' = \operatorname{diag}(\sigma'_1, \ldots, \sigma'_r, 0) \in \mathbb{R}^{m\times n}$, and

$$\sigma'_1 \geq \cdots \geq \sigma'_r > \sigma'_{r+1} = \cdots = \sigma'_n \equiv 0. \tag{2.2}$$

Similarly, consider the SVD of $[B|A]$, $s \equiv \mathrm{rank}([B|A])$,

$$[B|A] = U\Sigma V^T, \tag{2.3}$$

where $U^{-1} = U^T$, $V^{-1} = V^T$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_s, 0) \in \mathbb{R}^{m\times(n+d)}$, and

$$\sigma_1 \geq \cdots \geq \sigma_s > \sigma_{s+1} = \cdots = \sigma_{n+d} \equiv 0. \tag{2.4}$$

If $s = n + d$ (which implies $r = n$), then Σ' and Σ have no zero singular values. Among the singular values, a key role is played by σ_{n+1}, where n represents the number of columns of A. In order to handle possible higher multiplicity of σ_{n+1}, we introduce the notation

$$\sigma_p \equiv \sigma_{n-q} > \underbrace{\sigma_{n-q+1} = \cdots = \sigma_n}_{q} = \underbrace{\sigma_{n+1} = \cdots = \sigma_{n+e}}_{e} > \sigma_{n+e+1}, \tag{2.5}$$

where q singular values to the left and $e - 1$ singular values to the right are equal to σ_{n+1}, and hence $q \geq 0$, $e \geq 1$. For convenience we denote $n - q \equiv p$. (Clearly $\sigma_p \equiv \sigma_{n-q}$ is not defined if and only if $q = n$; similarly, σ_{n+e+1} is not defined if and only if $e = d$.)

For an integer Δ (not necessarily nonnegative) it will be useful to consider the partitioning

$$\Sigma = \begin{array}{|c|c|}\hline \overbrace{\Sigma_1^{(\Delta)}}^{n-\Delta} & \overbrace{\Sigma_2^{(\Delta)}}^{d+\Delta} \\ \hline \end{array}\Big\} m, \qquad V = \begin{array}{|c|c|}\hline \overbrace{V_{11}^{(\Delta)}}^{n-\Delta} & \overbrace{V_{12}^{(\Delta)}}^{d+\Delta} \\ \hline V_{21}^{(\Delta)} & V_{22}^{(\Delta)} \\ \hline \end{array}\begin{array}{l}\Big\} d \\ \Big\} n \end{array}, \tag{2.6}$$

where $\Sigma_1^{(\Delta)} \in \mathbb{R}^{m\times(n-\Delta)}$, $\Sigma_2^{(\Delta)} \in \mathbb{R}^{m\times(d+\Delta)}$, and $V_{11}^{(\Delta)} \in \mathbb{R}^{d\times(n-\Delta)}$, $V_{12}^{(\Delta)} \in \mathbb{R}^{d\times(d+\Delta)}$, $V_{21}^{(\Delta)} \in \mathbb{R}^{n\times(n-\Delta)}$, $V_{22}^{(\Delta)} \in \mathbb{R}^{n\times(d+\Delta)}$. When $\Delta = 0$, the partitioning conforms to the fact that $[B|A]$ is created by A appended by the matrix B with d columns, and in this case the upper index is omitted, $\Sigma_1 \equiv \Sigma_1^{(0)}$, etc.

The classical analysis of the TLS problem with a single right-hand side ($d = 1$) presented in [7] and the theory developed in [17] were based on relationships between the singular values of A and $[B|A]$. For $d = 1$, in particular, $\sigma'_n > \sigma_{n+1}$ represents a *sufficient (but not necessary) condition* for the existence and uniqueness of the solution. In order to extend this condition to the case $d > 1$, the following generalization of [7, Theorem 4.1] is useful.

THEOREM 2.1. *Let* (2.1) *be the SVD of* A *and* (2.3) *the SVD of* $[B|A]$ *with the partitioning given by* (2.6)*,* $m \geq n + d$*,* $\Delta \geq 0$*. If*

$$\sigma'_{n-\Delta} > \sigma_{n-\Delta+1}, \tag{2.7}$$

then $\sigma_{n-\Delta} > \sigma_{n-\Delta+1}$*. Moreover,* $V_{12}^{(\Delta)}$ *is of full row rank equal to* d*, and* $V_{21}^{(\Delta)}$ *is of full column rank equal to* $(n - \Delta)$*.*

The first part follows immediately from the interlacing theorem for singular values [17, Theorem 2.4, p. 32] (see also [13]). For the proof of the second part, see

[21, Lemma 2.1] or [17, Lemma 3.1, pp. 64–65]. (Please note the different ordering of the partitioning of V in [21], [17].)

We start our analysis with the following definition.

Definition 2.2 (problems of the 1st class and of the 2nd class). *Consider a TLS problem* (1.1)–(1.3), $m \geq n + d$. *Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning given by* (2.6). *Take* $\Delta \equiv q$, *where* q *is the "left multiplicity" of* σ_{n+1} *given by* (2.5).

- *If* $V_{12}^{(q)}$ *is of full row rank* d, *then we call* (1.1)–(1.3) *a* TLS problem of the 1st class.
- *If* $V_{12}^{(q)}$ *is rank deficient (i.e., has linearly dependent rows), then we call* (1.1)–(1.3) *a* TLS problem of the 2nd class.

The set of all problems of the 1*st class will be denoted by* $\mathcal{F}$. *The set of all problems of the* 2*nd class will be denoted by* $\mathcal{S}$.

3. Problems of the 1st class. For $d = 1$, the right singular vector subspace corresponding to the smallest singular value σ_{n+1} of $[b|A]$ contains for a TLS problem of the 1st class a singular vector with a nonzero first component. Consequently, the TLS problem has a (possibly nonunique) solution. As we will see, for $d > 1$ an analogous property does not hold. The TLS problem of the 1st class with $d > 1$ may not have a solution. First we recall known results for two special cases of problems of the 1st class.

3.1. Problems of the 1st class with unique TLS solution. Consider a TLS problem of the 1st class. Assume that $\sigma_n > \sigma_{n+1}$, i.e., $q = 0$ ($p = n$). Setting $\Delta \equiv q = 0$ in (2.6), $V_{12}^{(q)} \equiv V_{12}$ is a square (and nonsingular) matrix. Define the *correction matrix*

$$[G|E] \equiv -U[0|\Sigma_2]\,V^T = -U\Sigma_2[\,V_{12}^T|\,V_{22}^T]. \tag{3.1}$$

Clearly, $\|[G|E]\|_F = (\sum_{j=n+1}^{n+d} \sigma_j^2)^{1/2}$, and the *corrected matrix* $[B + G|A + E]$ represents, by the Eckart–Young–Mirsky theorem [1], [8], the unique rank n approximation of $[B|A]$ with minimal $[G|E]$ in the Frobenius norm.

The columns of the matrix $[\,V_{12}^T|\,V_{22}^T]^T$ represent a basis for the null space of the corrected matrix $[B + G|A + E] \equiv U\Sigma_1[\,V_{11}^T|\,V_{21}^T]$. Since V_{12} is square and nonsingular,

$$[B + G|A + E]\begin{bmatrix} -I_d \\ -V_{22}\,V_{12}^{-1} \end{bmatrix} = 0,$$

which gives the uniquely determined TLS solution

$$X_{\mathrm{TLS}} \equiv X^{(0)} \equiv -V_{22}\,V_{12}^{-1}. \tag{3.2}$$

We summarize these observations in the following theorem; see [17, Theorem 3.1, pp. 52–53].

Theorem 3.1. *Consider a TLS problem of the* 1*st class. If*

$$\sigma_n > \sigma_{n+1}, \tag{3.3}$$

then with the partitioning of the SVD of $[B|A]$ *given by* (2.6), $\Delta \equiv q = 0$, $V_{12} \in \mathbb{R}^{d\times d}$ *is square and nonsingular, and* (3.2) *represents the unique TLS solution of the problem* (1.1)–(1.3) *with the corresponding correction* $[G|E]$ *given by* (3.1).

Theorem 2.1 gives the following corollary.

Corollary 3.2. *Let* (2.1) *be the SVD of* A *and* (2.3) *the SVD of* $[B|A]$ *with the partitioning given by* (2.6), $m \geq n + d$, $\Delta \equiv 0$. *If*

$$\sigma'_n > \sigma_{n+1}, \tag{3.4}$$

then (1.1)–(1.3) *is a problem of the* 1*st class,* $\sigma_n > \sigma_{n+1}$*, and* (3.2) *represents the unique TLS solution of the problem* (1.1)–(1.3) *with the corresponding correction matrix* $[G|E]$ *given by* (3.1).

We see that (3.4) represents a sufficient condition for the existence and uniqueness of the TLS solution of the problem (1.1)–(1.3). This condition is, however, intricate. It may look like the key to the analysis of the TLS problem, in particular, when one considers the following corollary of the interlacing theorem for singular values and Theorem 2.1; see [17, Corollary 3.4, p. 65].

COROLLARY 3.3. *Let* (2.1) *be the SVD of* A *and* (2.3) *the SVD of* $[B|A]$ *with the partitioning given by* (2.6)*,* $m \geq n + d$*,* $\Delta \equiv q \geq 0$*. Then the following conditions are equivalent:*

(i) $\sigma'_{n-q} > \sigma_{n-q+1} = \cdots = \sigma_{n+d}$*,*
(ii) $\sigma_{n-q} > \sigma_{n-q+1} = \cdots = \sigma_{n+d}$ *and* $V_{12}^{(q)}$ *is of (full row) rank* d*.*

In the following discussion we restrict ourselves to the *single right-hand side* case. The condition (i) implies that the TLS problem is of the 1st class. If $d = 1$ and $q = 0$, then (i) reduces to (3.4) and the statement of Corollary 3.3 says that $\sigma'_n > \sigma_{n+1}$ if and only if $\sigma_n > \sigma_{n+1}$ and $[1, 0, \ldots, 0]^T v_{n+1} \neq 0$. In order to show the difficulty and motivate the classification in what follows, we now consider all remaining possibilities for the case $d = 1$. It should be, however, understood that they go beyond the problems of the 1st class and the unique TLS solution. If $\sigma'_n = \sigma_{n+1}$, then it may happen that *either* $\sigma_n > \sigma_{n+1}$ and $[1, 0, \ldots, 0]^T v_{n+1} = 0$, which means that the TLS problem is not of the 1st class and it does not have a solution, *or* $\sigma_n = \sigma_{n+1}$. In the latter case, depending on the relationship between σ'_{n-q} and $\sigma_{n-q+1} = \cdots = \sigma_{n+1}$ for some $q > 0$ (see Corollary 3.3), the TLS problem may have a nonunique solution if the TLS problem is of the 1st class (see the next section), or the solution may not exist. We see that an attempt to base the analysis on the relationship between σ'_n and σ_{n+1} becomes very involved.

The situation becomes more transparent with the use of the core problem concept from [11]. For any linear approximation problem $Ax \approx b$ (we still consider $d = 1$), there are orthogonal matrices P, R such that

$$P^T[b|A]\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & R \end{array}\right] = \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline\hline 0 & 0 & A_{22} \end{array}\right], \tag{3.5}$$

where the following hold:

(i) A_{11} is of minimal dimensions and A_{22} is of maximal dimensions (A_{22} may also have zero number of rows and/or columns) over all orthogonal transformations of $[b|A]$ yielding the structure (3.5) of zero and nonzero blocks. Suppose $b \not\perp \mathcal{R}(A)$ has nonzero projections on exactly ℓ left singular vector subspaces of A corresponding to distinct (nonzero) singular values. Then among all decompositions of the form (3.5) the minimally dimensioned A_{11} is $\ell \times \ell$ if $Ax \approx b$ is compatible and $(\ell + 1) \times \ell$ if $Ax \approx b$ is incompatible (see [11, Theorem 2.2]).
(ii) All singular values of A_{11} are simple and nonzero; all singular values of $[b_1|A_{11}]$ are simple and, since $b \notin \mathcal{R}(A)$, nonzero (recall that we consider only the incompatible problems).
(iii) The first components of all right singular vectors of $[b_1|A_{11}]$ are nonzero.
(iv) $\sigma_{\min}(A_{11}) > \sigma_{\min}([b_1|A_{11}])$. Moreover, singular values of A_{11} *strictly interlace* singular values of $[b_1|A_{11}]$.

(See [11, section 3].) The minimally dimensioned subproblem $A_{11}x_1 \approx b_1$ is then called the *core problem* within $Ax \approx b$. The SVD of the block structured matrix on the right-hand side in (3.5) can be obtained as a direct sum of the SVD decompositions of the blocks $[b_1|A_{11}]$ and A_{22}, just by extending the singular vectors corresponding to the first block by zeros on the bottom and by extending the singular vectors corresponding to the second block by zeros on the top. Consequently, considering the special structure of the orthogonal transformation $\mathrm{diag}(1, R)$ in (3.5), which does not change the first components of the right singular vectors, all right singular vectors of $[b|A]$ with nonzero first components correspond to the block $[b_1|A_{11}]$, and all right singular vectors of $[b|A]$ with zero first component correspond to A_{22}. Moreover,

$$\sigma'_n \equiv \sigma_{\min}(A) = \min\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{22})\},$$
$$\sigma_{n+1} \equiv \sigma_{\min}([b|A]) = \min\{\sigma_{\min}([b_1|A_{11}]), \sigma_{\min}(A_{22})\}.$$

We will review all possible situations.

Case 1. $\sigma'_n > \sigma_{n+1}$. This happens if and only if $\sigma_{\min}(A_{22}) > \sigma_{\min}([b_1|A_{11}]) = \sigma_{n+1}$, which is equivalent to the existence of the unique TLS solution.

Case 2. $\sigma_{\min}(A_{22}) \equiv \sigma'_n = \sigma_{n+1}$. Here we have to distinguish two cases:

Case 2a. $\sigma_{\min}(A) = \sigma_{\min}([b|A]) = \sigma_{\min}([b_1|A_{11}])$. This guarantees the existence of the (minimum norm) TLS solution. All singular values of A equal to $\sigma_{\min}(A)$ are the singular values of the block A_{22}. Consequently, the multiplicity of $\sigma_{\min}([b|A])$ is larger by one than the multiplicity of $\sigma_{\min}(A)$.

Case 2b. $\sigma_{\min}(A) = \sigma_{\min}([b|A]) < \sigma_{\min}([b_1|A_{11}])$. Then the multiplicities of $\sigma_{\min}(A)$ and $\sigma_{\min}([b|A])$ are equal, all right singular vectors of $[b|A]$ corresponding to $\sigma_{\min}([b|A])$ have zero first components, and the TLS solution does not exist.

Summarizing, the TLS solution exists if and only if *either* $\sigma_{\min}(A) > \sigma_{\min}([b|A])$, *or* $\sigma_{\min}(A) = \sigma_{\min}([b|A])$ with different multiplicities for $\sigma_{\min}(A)$ and $\sigma_{\min}([b|A])$. In terms of the singular values of subblocks in the core reduction (3.5),

$$\sigma_{\min}(A_{22}) > \sigma_{\min}([b_1|A_{11}]) \Leftrightarrow \text{TLS solution exists and is unique,}$$
$$\sigma_{\min}(A_{22}) = \sigma_{\min}([b_1|A_{11}]) \Leftrightarrow \text{TLS solution exists and is not unique,}$$
$$\sigma_{\min}(A_{22}) < \sigma_{\min}([b_1|A_{11}]) \Leftrightarrow \text{TLS solution does not exist.}$$

If the TLS solution exists, then the minimum norm TLS solution can always be computed, and it is automatically given by the core problem formulation. If the TLS solution does not exist, then the core problem formulation gives the solution equivalent to the minimum norm *nongeneric* solution constructed in [17].

We will see that in the multiple right-hand sides case the situation is much more complicated.

3.2. Problems of the 1st class with nonunique TLS solutions—a special case. Consider a TLS problem of the 1st class. Assume that $e \equiv d$ in (2.5); i.e., let all the singular values starting from $\sigma_{n-q+1} \equiv \sigma_{p+1}$ be equal:

$$\sigma_1 \geq \cdots \geq \sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1} = \cdots = \sigma_{n+d} \geq 0. \tag{3.6}$$

The case $q = 0$ ($p = n$) reduces to the problem with unique TLS solution discussed in section 3.1. If $q = n$ ($p = 0$), i.e., $\sigma_1 = \cdots = \sigma_{n+d}$, then the columns of $[B|A]$ are mutually orthogonal and $[B|A]^T[B|A] = \sigma_1^2 I_{n+d}$. Then it seems meaningless to approximate B by the columns of A, and we will get (consistently with [17]) the trivial solution $X_{\mathrm{TLS}} \equiv 0$ (this case does not satisfy the nontriviality assumption $A^T B \neq 0$ in (1.1)). Therefore, in this section the interesting case is represented by $n > q > 0$ ($0 < p < n$).

We first construct the solution minimal in norm. Since $V_{12}^{(q)} \in \mathbb{R}^{d\times(q+d)}$ is of full row rank, there exists an orthogonal matrix $Q \in \mathbb{R}^{(q+d)\times(q+d)}$ such that

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q \equiv [v_{p+1}, \ldots, v_{n+d}]\, Q = \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array}\right], \tag{3.7}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ is square and nonsingular. Such an orthogonal matrix Q can be obtained, e.g., using the LQ decomposition of $V_{12}^{(q)}$. Consider the partitioning $Q = [Q_1|Q_2]$, where $Q_2 \in \mathbb{R}^{(q+d)\times d}$ has d columns. Then the columns of Q_2 form an orthonormal basis of the subspace spanned by the columns of $V_{12}^{(q)T}$, $Q_1 \in \mathbb{R}^{(q+d)\times q}$ is an orthonormal basis of its orthogonal complement, and

$$\left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] = \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2, \qquad V_{12}^{(q)} = \Gamma Q_2^T. \tag{3.8}$$

Define the correction matrix

$$\begin{aligned} [G|E] &\equiv -[B|A] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T \\ &= -U\Sigma V^T \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2 Q_2^T \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]^T \\ &= -\sigma_{n+1}[u_{p+1}, \ldots, u_{n+d}]\, Q_2 Q_2^T [v_{p+1}, \ldots, v_{n+d}]^T, \end{aligned} \tag{3.9}$$

where u_j and v_j represent left and right singular vectors of the matrix $[B|A]$, respectively. If $\sigma_{p+1} = \cdots = \sigma_{n+d} = 0$, then the correction matrix is a zero matrix ($\sigma_{n+1} = 0$), and the problem is compatible; thus we consider $\sigma_{p+1} = \cdots = \sigma_{n+d} > 0$.

Note that with the choice of any other matrix $Q' = [Q_1'|Q_2']$ giving a decomposition of the form (3.7), Q_2' represents an orthonormal basis of the subspace spanned by the columns of $V_{12}^{(q)T}$, and therefore $Q_2' = Q_2\Psi$ for some orthogonal matrix $\Psi \in \mathbb{R}^{d\times d}$. Consequently, (3.9) is uniquely determined independently of the choice of Q in (3.7).

Clearly, $\|[G|E]\|_F = \sigma_{n+1}\|Q_2 Q_2^T\|_F = \sigma_{n+1}\sqrt{d}$, and the corrected matrix

$$[B + G|A + E] \equiv [B|A]\left(I_{n+d} - \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T\right)$$

represents the rank n approximation of $[B|A]$ such that the Frobenius norm of the correction matrix $[G|E]$ is minimal, by the Eckart–Young–Mirsky theorem.

The columns of the matrix $[\Gamma^T|Z^T]^T$ represent a basis for the null space of the corrected matrix $[B + G|A + E]$. Since Γ is square and nonsingular,

$$[B + G|A + E]\begin{bmatrix} -I_d \\ -Z\Gamma^{-1} \end{bmatrix} = 0,$$

which gives the TLS solution

$$X_{\mathrm{TLS}} \equiv -Z\Gamma^{-1} = -[Y|Z]Q^T Q\begin{bmatrix} 0 \\ \Gamma^{-1} \end{bmatrix} = -\,V_{22}^{(q)}\,V_{12}^{(q)}\; \equiv X^{(q)}. \tag{3.10}$$

This can be expressed as

$$X_{\mathrm{TLS}} = (A^T A - \sigma_{n+1}^2 I_n)^{\dagger} A^T B;$$

see [17, Theorem 3.10, pp. 62–64]. The solution (3.10) and the correction (3.9) do not depend on the choice of the matrix Q in (3.7). We summarize these observations in the following theorem (see [17, Theorem 3.9, pp. 60–62]).

Theorem 3.4. *Consider a TLS problem of the* 1*st class. Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning given by* (2.6)*,* $\Delta \equiv q < n$*,* $p \equiv n - q$*. If*

$$\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+d}, \tag{3.11}$$

then (3.10) *represents a TLS solution* X_{TLS} *of the problem* (1.1)–(1.3)*. This is the unique solution of the minimal Frobenius norm and* 2*-norm, with the corresponding unique correction matrix* $[G|E]$ *given by* (3.9)*.*

Using Corollary 3.3 we get

$$\sigma'_p > \sigma_{p+1} = \cdots = \sigma_{n+d}, \tag{3.12}$$

which represents a sufficient condition for the existence of the TLS solution of the TLS problem (1.1)–(1.3) minimal in the Frobenius norm and the 2-norm.

The correction matrix minimal in the Frobenius norm can be in this special case constructed from *any* d-vectors selected among $q + d$ columns $v_{p+1}, \ldots, v_{n+d}$ (or their orthogonal linear transformation) of the matrix V such that their top d-subvectors create a d-by-d square nonsingular matrix. The equality of the last $q + d$ singular values ensures that the Frobenius norm of the corresponding correction matrix is still equal to $\sigma_{n+1}\sqrt{d}$. It can be shown that, for any such choice, a norm of the corresponding solution $\tilde{X}$ is larger than or equal to the norm of $X^{(q)}$ given by (3.10), and any such $\tilde{X}$ represents a TLS solution. Consequently, the special TLS problem satisfying (3.6) has infinitely many solutions.

3.3. Problems of the 1st class—the general case. Here we consider a TLS problem of the 1st class with a general distribution of singular values. We will discuss only the remaining cases not covered in the previous two sections, i.e., $n \geq q > 0$ ($0 \leq p < n$; recall that $p = n - q$) and $e < d$, giving

$$\sigma_1 \geq \cdots \geq \sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1} = \cdots = \sigma_{n+e} > \sigma_{n+e+1} \geq \cdots \geq \sigma_{n+d} \geq 0$$

(note that σ_p does not exist for $q = n$ ($p = 0$)). We will see that in this general case the problem (1.1)–(1.3) *may not have a solution*.

We try to construct a TLS solution with the same approach used in section 3.2, and we will show that it may fail. Since, with the partitioning (2.6), $\Delta \equiv q$, the matrix $V_{12}^{(q)} \in \mathbb{R}^{d\times(q+d)}$ is of full row rank, there exists an orthogonal matrix $Q \in \mathbb{R}^{(q+d)\times(q+d)}$ such that

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q \equiv [v_{p+1}, \ldots, v_{n+d}] Q = \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array}\right], \tag{3.13}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ is square and nonsingular. With the partitioning $Q = [Q_1|Q_2]$, where $Q_1 \in \mathbb{R}^{(q+d)\times q}$, $Q_2 \in \mathbb{R}^{(q+d)\times d}$, the columns of Q_2 form an orthonormal basis of the subspace spanned by the columns of $V_{12}^{(q)T}$, and

$$\left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] = \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2, \qquad V_{12}^{(q)} = \Gamma Q_2^T. \tag{3.14}$$

Following [17], it is tempting to define the correction matrix

$$\begin{aligned} [G|E] &\equiv -[B|A] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T \\ &= -U\Sigma V^T \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2 Q_2^T \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]^T \\ &= -[u_{p+1}, \ldots, u_{n+d}] \mathrm{diag}(\sigma_{p+1}, \ldots, \sigma_{n+d}) Q_2 Q_2^T [v_{p+1}, \ldots, v_{n+d}]^T, \end{aligned} \tag{3.15}$$

which differs from (3.9) because the diagonal factor is no longer a scalar multiple of the identity matrix. Analogously to the previous section, the matrix (3.15) is uniquely determined independently of the choice of Q in (3.13).

The columns of the matrix $[\Gamma^T|Z^T]^T$ are in the null space of the corrected matrix

$$[B + G|A + E] \equiv [B|A] \left(I_{n+d} - \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T \right). \tag{3.16}$$

In general the columns of $[\Gamma^T|Z^T]^T$ do not represent *a basis* for the null space of the corrected matrix. If A is not of full column rank, the extended matrix $[B|A]$ has a zero singular value with the corresponding right singular vector having the first d entries equal to zero. Such a right singular vector is in the null space of the corrected matrix, but it cannot be obtained as a linear combination of the columns of $[\Gamma^T|Z^T]^T$. Since Γ is square and nonsingular,

$$[B + G|A + E] \left[\begin{array}{c} -I_d \\ \hline -Z\Gamma^{-1} \end{array}\right] = 0,$$

and we can construct

$$X^{(q)} \equiv -Z\Gamma^{-1} = -V_{22}^{(q)} V_{12}^{(q)\dagger}. \tag{3.17}$$

The matrices (3.17) and (3.15) do not depend on the choice of Q in (3.13). The matrix $X^{(q)}$ given by (3.17) is a natural generalization of $X^{(q)}$ given by (3.10). The *classical TLS algorithm* [15], [16] (see also [17]) applied to a TLS problem of the 1st class returns as output the matrix $X^{(q)}$ given by (3.17) with the matrices G, E given by (3.15). We will show, however, that $X^{(q)}$ is not necessarily a TLS solution.

We first focus on the question whether there exists another correction $\tilde{E}$, $\tilde{G}$ corresponding to the last $q + d$ columns of V that makes the corrected system compatible. Such a correction can be constructed analogously to (3.13) by considering an orthogonal matrix $\tilde{Q} = [\tilde{Q}_1|\tilde{Q}_2]$ such that

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] \tilde{Q} = [v_{p+1}, \ldots, v_{n+d}]\tilde{Q} = \left[\begin{array}{c|c} \boldsymbol{\Omega} & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right], \tag{3.18}$$

where $\tilde{\Gamma} \in \mathbb{R}^{d\times d}$ is *nonsingular* and $\boldsymbol{\Omega}$ is a matrix not necessarily equal to zero. Then define the correction matrix

$$[\tilde{G}|\tilde{E}] \equiv -[B|A]\left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right]\left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right]^T. \tag{3.19}$$

The corrected system $(A + \tilde{E})X = B + \tilde{G}$ is compatible and the matrix

$$\tilde{X} \equiv -\tilde{Z}\tilde{\Gamma}^{-1} = -V_{22}^{(q)}(V_{12}^{(q)}\tilde{Q}_2\tilde{Q}_2^T)^{\dagger} \tag{3.20}$$

solves this corrected system. The columns of $[\tilde{\Gamma}^T|\tilde{Z}^T]^T$ have to be in the null space of the corrected matrix $[B + \tilde{G}|A + \tilde{E}]$. As above, they do not necessarily represent a basis of this null space.

Now we show that $X^{(q)}$ does not necessarily represent a TLS solution; i.e., the Frobenius norm of the correction matrix (3.15) need not be minimal. This can be illustrated by a simple example. Let $q = n$ and $e < d$. Then in (3.13) we set $Q = [V_{22}^{(q)T}|V_{12}^{(q)T}]$. (Notice that $V_{11}^{(\Delta)}$ and $V_{21}^{(\Delta)}$ in the partitioning (2.6) vanish for $\Delta \equiv q = n$.) Therefore,

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right][V_{22}^{(q)T}|V_{12}^{(q)T}] = \left[\begin{array}{c|c} 0 & I_d \\ \hline I_n & 0 \end{array}\right]; \quad \text{i.e.,} \quad \Gamma = I_d, \quad Z = 0,$$

which gives from (3.13) $[G|E] = -[B|0]$, and, analogously, $X^{(q)} = 0$; see (3.17). If we solve the same problem in the ordinary least squares sense, then the corresponding correction matrix is $[\bar{G}|\bar{E}] \equiv [(AA^{\dagger} - I)B|0]$, having in general smaller Frobenius norm than $[G|E] = -[B|0]$, given by (3.15). Therefore, the constructed matrix $X^{(q)}$ given by (3.17) does not, in general, represent a TLS solution.

Summarizing, the classical TLS algorithm of Van Huffel computes for TLS problems of the 1st class the output (3.2), (3.10), or (3.17), which are formally analogous, but with different relationship to the TLS solution. While (3.2) and (in the particular case of a very special distribution of the singular values) (3.10) represent TLS solutions (having minimal Frobenius and 2-norm), the interpretation of (3.17) remains unclear. The partitioning of the set $\mathcal{F}$ of TLS problems of the 1st class according to the conditions valid in (3.2), (3.10), and (3.17) is unsatisfactory. In particular, apart from the simple case (3.2) and the very special case (3.10), we do not know whether a TLS solution exists.[1] We will therefore develop a different partitioning of the set $\mathcal{F}$ in section 4. First we briefly discuss some properties of matrices $X^{(q)}$ and $\tilde{X}$.

3.4. Note on the norms of matrices $X^{(q)}$ and $\tilde{X}$. It is obvious that $X^{(q)}$ given by (3.17) is a special case of $\tilde{X}$ given by (3.20). Lemma 3.5 gives simple formulas for the Frobenius norm and 2-norm of $\tilde{X}$. Lemma 3.6 shows that $X^{(q)}$ has the minimal norms among all $\tilde{X}$ of the form (3.20). The proofs are fully analogous to the proofs of [17, Theorems 3.6 and 3.9].

[1]The problems in the set $\mathcal{F}$ are called generic in [17]. Since a problem in this set may not have a TLS solution, we will no longer use the generic–nongeneric terminology.

LEMMA 3.5. *Let* $[\tilde{\Gamma}^T|\tilde{Z}^T]^T \in \mathbb{R}^{(n+d)\times d}$ *have orthonormal columns, and assume* $\tilde{\Gamma} \in \mathbb{R}^{d\times d}$ *is nonsingular. Then the matrix* $\tilde{X} = -\tilde{Z}\tilde{\Gamma}^{-1}$ *has the norms*

$$\|\tilde{X}\|_F^2 = \|\tilde{\Gamma}^{-1}\|_F^2 - d \quad \textit{and} \quad \|\tilde{X}\|^2 = \frac{1-\sigma_{\min}^2(\tilde{\Gamma})}{\sigma_{\min}^2(\tilde{\Gamma})}, \tag{3.21}$$

where $\sigma_{\min}(\tilde{\Gamma})$ *is the minimal singular value of* $\tilde{\Gamma}$.

LEMMA 3.6. *Consider* $X^{(q)} = -Z\Gamma^{-1} = -V_{22}^{(q)} V_{12}^{(q)\dagger}$ *given by* (3.13)–(3.17) *and* $\tilde{X} = -\tilde{Z}\tilde{\Gamma}^{-1}$ *given by* (3.18)–(3.20). *Then*

$$\|\tilde{X}\|_F \geq \|X^{(q)}\|_F, \quad \textit{and} \quad \|\tilde{X}\| \geq \|X^{(q)}\|. \tag{3.22}$$

Moreover, equality holds for the Frobenius norms if and only if $\tilde{X} = X^{(q)}$.

These lemmas can be easily seen as follows. A matrix $\tilde{X}$ of the form (3.20) is going to be minimal in the Frobenius or the 2-norm when $\|\tilde{\Gamma}^{-1}\|_F$ is minimized or $\sigma_{\min}(\tilde{\Gamma}) \equiv \sigma_d(\tilde{\Gamma})$ is maximized, respectively. The minimization/maximization are with respect to the orthogonal matrix $\tilde{Q}$ which is considered a free variable, with the constraint that $\tilde{\Gamma}$ has to be nonsingular. The interlacing theorem for singular values applied to the matrices $[\mathbf{\Omega}|\tilde{\Gamma}] = V_{12}^{(q)}\tilde{Q}$ and $\tilde{\Gamma}$ gives

$$\sigma_j(\Gamma) = \sigma_j(V_{12}^{(q)}) = \sigma_j([\mathbf{\Omega}|\tilde{\Gamma}]) \geq \sigma_j(\tilde{\Gamma}), \qquad j = 1, \ldots, d,$$

with all the inequalities becoming equalities if and only if $\mathbf{\Omega} = 0$. The minimum for the 2-norm is reached when the smallest singular values are equal, i.e., $\sigma_d(\Gamma) = \sigma_d(\tilde{\Gamma})$. Note that there can be more than one matrix of the form (3.20) reaching the minimum of the 2-norm.

If the corrected matrix $(A + \tilde{E})$ has linearly dependent columns, then the corrected system with the correction $[\tilde{G}|\tilde{E}]$ of the form (3.19) can have more than one solution. The following lemma shows that under some additional assumptions on the structure of $\tilde{Q}$, the matrix $(A + \tilde{E})$ is of full column rank, and therefore the matrix $\tilde{X}$ of the form (3.20) is the unique solution of the corrected system. (Note that the correction (3.15) is a special case of the correction (3.19).)

LEMMA 3.7. *Consider a TLS problem of the* 1*st class. Let* $[\tilde{G}|\tilde{E}]$ *be the correction matrix given by* (3.19), *and let* $\tilde{X}$ *be the matrix given by* (3.20). *If* $\tilde{Q}$ *in* (3.18) *has the block diagonal form* $\tilde{Q} = \operatorname{diag}(Q', I_{d-e})$, *where* $Q' \in \mathbb{R}^{(q+e)\times(q+e)}$ *is an orthogonal matrix, then* $(A + \tilde{E})$ *is of full column rank, and* $\tilde{X}$ *represents the unique solution of the corrected system* $(A + \tilde{E})\tilde{X} = B + \tilde{G}$.

Proof. Since $\tilde{Q} = \operatorname{diag}(Q', I_{d-e})$ has the block diagonal structure,

$$[B|A] = U\Sigma V^T = \left(U \left[\begin{array}{c|c|c} I_p & 0 & 0 \\ \hline 0 & \tilde{Q} & 0 \\ \hline 0 & 0 & I_{m-n-d} \end{array}\right] \right) \mathbf{\Sigma} \left(V \left[\begin{array}{c|c} I_p & 0 \\ \hline 0 & \tilde{Q} \end{array}\right] \right)^T \equiv \bar{U}\mathbf{\Sigma}\bar{V}^T;$$

i.e., $\bar{U}\mathbf{\Sigma}\bar{V}^T$ represents the SVD of $[B|A]$ with

$$\bar{U} = [\bar{u}_1, \ldots, \bar{u}_m], \qquad \bar{V} = [\bar{v}_1, \ldots, \bar{v}_{n+d}] = \left[\begin{array}{c|c|c} V_{11}^{(q)} & \mathbf{\Omega} & \tilde{\Gamma} \\ \hline V_{21}^{(q)} & \tilde{Y} & \tilde{Z} \end{array}\right].$$

Using this SVD, the corrected matrix can be written as

$$[B+\tilde{G}|A+\tilde{E}] = [\bar{u}_1, \ldots, \bar{u}_n]\,\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\left[\begin{array}{c|c} V_{11}^{(q)} & \boldsymbol{\Omega} \\ \hline V_{21}^{(q)} & \tilde{Y} \end{array}\right]^T.$$

If $\sigma_n = 0$, then $[\tilde{G}|\tilde{E}] = 0$ and the original system is compatible, i.e., $\mathcal{R}(B) \subseteq \mathcal{R}(A)$. Therefore, assume $\sigma_n > 0$. From the CS decomposition of $\tilde{V}$ it follows that since $\tilde{\Gamma}$ is square nonsingular, the matrix $[V_{21}^{(q)}|\tilde{Y}]$ is square nonsingular. Since $[\bar{u}_1, \ldots, \bar{u}_n]$ is of full column rank, the matrix

$$(A+\tilde{E}) = [\bar{u}_1, \ldots, \bar{u}_n]\mathrm{diag}(\sigma_1, \ldots, \sigma_n)[V_{21}^{(q)}|\ \tilde{Y}]^T$$

is of full column rank. The matrix $\tilde{X}$ is then the *unique* solution of the corrected system $(A+\tilde{E})\tilde{X} = B+\tilde{G}$. □

We will see in the next section that the form $\tilde{Q} = \mathrm{diag}(Q', I_{d-e})$ appears in a natural way.

4. Partitioning of the set of problems of the 1st class. We will base our partitioning and the subsequent classification of TLS problems with multiple right-hand sides on the following theorem.

THEOREM 4.1. *Consider a TLS problem of the* 1*st class. Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning given by* (2.6), $\Delta \equiv q \le n$, *where* q *is the "left multiplicity" of* σ_{n+1} *given by* (2.5), $p \equiv n-q$. *Consider an orthogonal matrix* $\tilde{Q}$ *such that*

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]\tilde{Q} = \left[\begin{array}{c|c} \boldsymbol{\Omega} & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right], \qquad \tilde{Q} = [\,\tilde{Q}_1\,|\,\tilde{Q}_2], \tag{4.1}$$

where $\tilde{Q}_1 \in \mathbb{R}^{(q+d)\times q}$, $\tilde{Q}_2 \in \mathbb{R}^{(q+d)\times d}$, *and define*

$$\begin{aligned}[\tilde{G}|\tilde{E}] &\equiv -[B|A]\left[\begin{array}{c}\tilde{\Gamma}\\ \hline \tilde{Z}\end{array}\right]\left[\begin{array}{c}\tilde{\Gamma}\\ \hline \tilde{Z}\end{array}\right]^T \\ &= -[u_{p+1}, \ldots, u_{n+d}]\,\mathrm{diag}(\sigma_{p+1}, \ldots, \sigma_{n+d})\,\tilde{Q}_2\tilde{Q}_2^T[v_{p+1}, \ldots, v_{n+d}]^T. \end{aligned} \tag{4.2}$$

Then the following two assertions are equivalent:

(i) *There exists an orthonormal matrix* $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d}$ *such that* $\hat{Q} \equiv \tilde{Q}\,\mathrm{diag}(I_q, \boldsymbol{\Psi})$ *has the block diagonal structure*

$$\hat{Q} = \left[\begin{array}{c|c} Q' & 0 \\ \hline 0 & I_{d-e} \end{array}\right] \in \mathbb{R}^{(q+d)\times(q+d)}, \qquad Q' \in \mathbb{R}^{(q+e)\times(q+e)}, \tag{4.3}$$

and using $\hat{Q}$ *in* (4.1)–(4.2) *instead of* $\tilde{Q}$ *yields the same* $[\tilde{G}|\tilde{E}]$.

(ii) *The matrix* $[\tilde{G}|\tilde{E}]$ *satisfies*

$$\|[\tilde{G}|\tilde{E}]\|_F = \left(\sum_{j=n+1}^{n+d}\sigma_j^2\right)^{1/2}. \tag{4.4}$$

Proof. First we prove the implication (i) ⇒ (ii). We partition $\hat{Q} = [\hat{Q}_1|\hat{Q}_2]$, where $\hat{Q}_1 \in \mathbb{R}^{(q+d)\times q}$, $\hat{Q}_2 \in \mathbb{R}^{(q+d)\times d}$, and $Q' = [Q'_1|Q'_2]$, where $Q'_1 \in \mathbb{R}^{(q+e)\times q}$, $Q'_2 \in \mathbb{R}^{(q+e)\times e}$. Then

$$\hat{Q}_2\hat{Q}_2^T = \left[\begin{array}{c|c} Q_2' & 0 \\ \hline 0 & I_{d-e} \end{array}\right]\left[\begin{array}{c|c} Q_2' & 0 \\ \hline 0 & I_{d-e} \end{array}\right]^T = \left[\begin{array}{c} Q_2' \\ \hline 0 \end{array}\right]\left[\begin{array}{c} Q_2' \\ \hline 0 \end{array}\right]^T + \left[\begin{array}{c} 0 \\ \hline I_{d-e} \end{array}\right]\left[\begin{array}{c} 0 \\ \hline I_{d-e} \end{array}\right]^T,$$

which gives, using (4.2) and (2.5),

$$\begin{aligned}\|[\tilde{G}|\tilde{E}]\|_F^2 &= \|\mathrm{diag}(\sigma_{p+1},\ldots,\sigma_{n+d})\,\hat{Q}_2\hat{Q}_2^T\|_F^2 \\ &= \sigma_{n+1}^2\|Q_2'(Q_2')^T\|_F^2 + \sum_{j=n+e+1}^{n+d}\sigma_j^2 = \sigma_{n+1}^2 e + \sum_{j=n+e+1}^{n+d}\sigma_j^2,\end{aligned}$$

i.e., (4.4). The implication (i) $\Rightarrow$ (ii) is proved.

Now we prove the implication (ii) $\Rightarrow$ (i). Let $[\tilde{G}|\tilde{E}]$ be given by (4.1), (4.2) and assume that (4.4) holds. We prove that there exists $\hat{Q}$ of the form (4.3) giving the same $[\tilde{G}|\tilde{E}]$. Define the splitting

$$\tilde{Q} = [\tilde{Q}_1|\tilde{Q}_2] = \left[\begin{array}{c|c} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \hline \tilde{Q}_{21} & \tilde{Q}_{22} \end{array}\right]$$

such that $\tilde{Q}_{11}\in\mathbb{R}^{(q+e)\times q}$, $\tilde{Q}_{21}\in\mathbb{R}^{(d-e)\times q}$, $\tilde{Q}_{12}\in\mathbb{R}^{(q+e)\times d}$, $\tilde{Q}_{22}\in\mathbb{R}^{(d-e)\times d}$. The matrix $[\tilde{G}|\tilde{E}]$ given by (4.2) satisfies

$$\begin{aligned}\|[\tilde{G}|\tilde{E}]\|_F^2 &= \|\mathrm{diag}(\sigma_{p+1},\ldots,\sigma_{n+d})\,\tilde{Q}_2\|_F^2 \\ &= \sigma_{n+1}^2\|\tilde{Q}_{12}\|_F^2 + \|D\tilde{Q}_{22}\|_F^2,\end{aligned}$$

where $D\equiv\mathrm{diag}(\sigma_{n+e+1},\ldots,\sigma_{n+d})$. Note that $\|\tilde{Q}_{12}\|_F^2 = d - \|\tilde{Q}_{22}\|_F^2$, since the matrix $\tilde{Q}_2$ consists of d orthonormal columns. Thus,

$$\|[\tilde{G}|\tilde{E}]\|_F^2 = \sigma_{n+1}^2(d-\|\tilde{Q}_{22}\|_F^2) + \|D\tilde{Q}_{22}\|_F^2 = \sigma_{n+1}^2 d - \|(\sigma_{n+1}^2 I_{d-e} - D^2)^{1/2}\tilde{Q}_{22}\|_F^2.$$

Using (4.4) this gives

$$\sigma_{n+1}^2(d-e) - \sum_{j=n+e+1}^{n+d}\sigma_j^2 = \|(\sigma_{n+1}^2 I_{d-e} - D^2)^{1/2}\tilde{Q}_{22}\|_F^2.$$

Since $\sigma_{n+1} > \sigma_{n+e+\ell}$ for all $\ell = 1,\ldots,d-e$, this implies that all rows of $\tilde{Q}_{22}$ have norm equal to one. Consequently, since $\tilde{Q}$ is an orthogonal matrix, $\tilde{Q}_{21} = 0$, i.e.,

$$\tilde{Q} = [\tilde{Q}_1|\tilde{Q}_2] = \left[\begin{array}{c|c} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \hline 0 & \tilde{Q}_{22} \end{array}\right],$$

and the matrix $\tilde{Q}_{22}$ has orthonormal rows. Consider the SVD $\tilde{Q}_{22} = S[I_{d-e}|0]P^T = [S|0]P^T$, where $S\in\mathbb{R}^{(d-e)\times(d-e)}$, $P\in\mathbb{R}^{d\times d}$ are square orthogonal matrices. Define orthogonal matrices

$$\boldsymbol{\Psi} \equiv P\left[\begin{array}{c|c} 0 & S^T \\ \hline I_e & 0 \end{array}\right]\in\mathbb{R}^{d\times d} \quad\text{and}\quad \hat{Q}\equiv\tilde{Q}\left[\begin{array}{c|c} I_q & 0 \\ \hline 0 & \boldsymbol{\Psi} \end{array}\right] = \left[\begin{array}{c|c} \tilde{Q}_{11} & \tilde{Q}_{12}\boldsymbol{\Psi} \\ \hline 0 & [0|I_{d-e}] \end{array}\right].$$

Because $\hat{Q}$ is orthogonal, the last $d-e$ columns of $\tilde{Q}_{12}\boldsymbol{\Psi}$ (i.e., corresponding to the block I_{d-e}) are zero, and

$$\hat{Q} = \mathrm{diag}(Q', I_{d-e})$$

is in the form (4.3) with $Q' = [\tilde{Q}_{11}|\tilde{Q}_{12}\mathbf{\Psi} I_{q+d}^{(e)}] \in \mathbb{R}^{(q+e)\times(q+e)}$, where $I_{q+d}^{(e)}$ represents the first e columns of I_{q+d}. Because $\hat{Q}_2\hat{Q}_2^T = (\tilde{Q}_2\mathbf{\Psi})(\tilde{Q}_2\mathbf{\Psi})^T = \tilde{Q}_2\tilde{Q}_2^T$, the matrix $\hat{Q}$ yields the same correction (4.2) as $\tilde{Q}$. ∎

The statement of this theorem says that *any* correction $[\tilde{G}|\tilde{E}]$ (reducing rank of $[B|A]$ to at most n) having the norm given by (4.4) can be obtained as in (4.1)–(4.2) with $\tilde{Q}$ in the *block diagonal* form (4.3).

Now we describe three disjoint subsets of problems of the 1st class representing the core of the proposed classification. Define the partitioning of the matrix $V_{12}^{(q)}$ with respect to e, the "right multiplicity" of σ_{n+1}, given by (2.5),

$$V_{12}^{(q)} = \begin{array}{|c|c|}\hline \overbrace{\phantom{W^{(q,e)}}}^{q+e} & \overbrace{\phantom{V_{12}^{(-e)}}}^{d-e} \\ \hline W^{(q,e)} & V_{12}^{(-e)} \\ \hline \underbrace{\phantom{W^{(q,e)}}}_{d} & \end{array} \Bigg\} d\,, \tag{4.5}$$

where $W^{(q,e)} \in \mathbb{R}^{d\times(q+e)}$, $V_{12}^{(-e)} \in \mathbb{R}^{d\times(d-e)}$. Note that since $\mathrm{rank}(V_{12}^{(q)}) = d$, i.e., the problem is of the 1st class, $\mathrm{rank}(V_{12}^{(-e)}) \le d-e$ implies that $\mathrm{rank}(W^{(q,e)}) \ge e$. On the other hand, $\mathrm{rank}(W^{(q,e)}) = e$ implies that $\mathrm{rank}(V_{12}^{(-e)}) = d-e$.

DEFINITION 4.2 (partitioning of the set of problems of the 1st class). *Consider a TLS problem* (1.1)–(1.3), $m \ge n+d$. *Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning given by* (2.6), $\Delta \equiv q$, *and the partitioning of* $V_{12}^{(q)}$ *given by* (4.5), *where* q *and* e *are the integers related to the multiplicity of* σ_{n+1}, *given by* (2.5). *Let the problem* (1.1)–(1.3) *be of the* 1*st class (i.e.,* $\mathrm{rank}(V_{12}^{(q)}) = d$*). The set of all problems for which*

- $\mathrm{rank}(W^{(q,e)}) = e$ *and* $\mathrm{rank}(V_{12}^{(-e)}) = d-e$ ($V_{12}^{(-e)}$ *has full column rank),*
- $\mathrm{rank}(W^{(q,e)}) > e$ *and* $\mathrm{rank}(V_{12}^{(-e)}) = d-e$ ($V_{12}^{(-e)}$ *has full column rank),*
- $\mathrm{rank}(W^{(q,e)}) > e$ *and* $\mathrm{rank}(V_{12}^{(-e)}) < d-e$ ($V_{12}^{(-e)}$ *is rank deficient)*

will be denoted by $\mathcal{F}_1$, $\mathcal{F}_2$, *and* $\mathcal{F}_3$, *respectively. Clearly,* $\mathcal{F}_1$, $\mathcal{F}_2$, *and* $\mathcal{F}_3$ *are mutually disjoint and* $\mathcal{F}_1 \cup \mathcal{F}_2 \cup \mathcal{F}_3 = \mathcal{F}$.

4.1. The set $\mathcal{F}_1$—problems of the 1st class having a TLS solution in the form $X^{(q)}$. Consider a TLS problem of the 1st class from the set $\mathcal{F}_1$, i.e., $\mathrm{rank}(W^{(q,e)}) = e$ in (4.5) which implies $V_{12}^{(-e)}$ is of full column rank, i.e., $\mathrm{rank}(V_{12}^{(-e)}) = d-e$. First we give a lemma which allows us to relate the partitioning (4.5) to the construction of a solution in (3.13)–(3.17).

LEMMA 4.3. *Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning* (2.6), $m \ge n+d$, $\Delta \equiv q \le n$. *Consider the partitioning* (4.5) *of* $V_{12}^{(q)}$. *The following two assertions are equivalent:*

(i) *The matrix* $W^{(q,e)}$ *has rank equal to* e.

(ii) *There exists* Q *in the block diagonal form* (4.3) *satisfying* (3.13).

Proof. Let $W^{(q,e)} \in \mathbb{R}^{d\times(q+e)}$ have rank equal to e. Then $\mathrm{rank}(V_{12}^{(-e)}) = d-e$. There exists an orthogonal matrix $H \in \mathbb{R}^{(q+e)\times(q+e)}$ (e.g., a product of Householder transformation matrices) such that $W^{(q,e)}H = [0|M]$, where $M \in \mathbb{R}^{d\times e}$ is of full column rank. Putting $Q \equiv \mathrm{diag}(H, I_{d-e})$ yields $V_{12}^{(q)}Q = [0|\Gamma]$, where the square matrix $\Gamma \equiv [M|\,V_{12}^{(-e)}] \in \mathbb{R}^{d\times d}$ is nonsingular.

Conversely, let $Q = \mathrm{diag}(Q', I_{d-e})$ and satisfy (3.13). Denote $\Gamma = [\Gamma_1|\Gamma_2]$, where $\Gamma_1 \in \mathbb{R}^{d\times e}$, $\Gamma_2 \in \mathbb{R}^{d\times(d-e)}$. Obviously $[0|\Gamma_1] = W^{(q,e)}Q'$, $\Gamma_2 = V_{12}^{(-e)} I_{d-e} = V_{12}^{(-e)}$. Since Γ is nonsingular, $\mathrm{rank}(\Gamma_1) = e$. Q' is an orthogonal matrix and thus $\mathrm{rank}(W^{(q,e)}) = e$. ∎

The following theorem formulates results for the set $\mathcal{F}_1$.

THEOREM 4.4. *Let* (2.3) *be the SVD of* $[B|A]$ *with the partitioning* (2.6), $m \geq n + d$, $\Delta \equiv q \leq n$ ($p \equiv n - q$). *Let the TLS problem* (1.1)–(1.3) *be of the* 1*st class; i.e.,* $V_{12}^{(q)}$ *is of full row rank equal to* d. *Let* $\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1} = \cdots = \sigma_{n+e}$, $1 \leq e \leq d$ (*if* $q = n$, *then* σ_p *is not defined). Consider the partitioning of* $V_{12}^{(q)}$ *given by* (4.5). *If*

$$\mathrm{rank}(W^{(q,e)}) = e \tag{4.6}$$

(the problem is from the set $\mathcal{F}_1$*), then* $X_{\mathrm{TLS}} \equiv X^{(q)} = -V_{22}^{(q)} V_{12}^{(q)\dagger}$ *given by* (3.17) *represents the TLS solution having the minimality property* (3.22). *The corresponding correction* $[G|E]$ *given by* (3.15) *has the norm* (4.4).

The proof follows immediately from Lemmas 4.3, 3.6, and 3.7.

The problems of the 1st class discussed earlier in sections 3.1 and 3.2 belong to the set $\mathcal{F}_1$. In the first case $q \equiv 0$ and $V_{12}^{(q)} \equiv V_{12}$ is square nonsingular. Thus, independently of the value of e (4.5) yields $W^{(0,e)}$ with the (full column) rank equal to e and the matrix Q' from $Q = \mathrm{diag}(Q', I_{d-e})$ in the assertion (ii) of Lemma 4.3 can always be chosen equal to the identity matrix I_e; i.e., $Q = I_d$. In the second case $e \equiv d$. Thus $W^{(q,d)} \equiv V_{12}^{(q)}$ is of (full row) rank equal to d. Here the identity block I_{d-e} in the assertion (ii) of Lemma 4.3 disappears; i.e., $Q = Q'$.

4.2. The set $\mathcal{F}_2$—problems of the 1st class having a TLS solution but not in the form $X^{(q)}$. Consider a TLS problem of the 1st class from the set $\mathcal{F}_2$, i.e., $\mathrm{rank}(V_{12}^{(-e)}) = d - e$ and $\mathrm{rank}(W^{(q,e)}) > e$ in (4.5). Because $V_{12}^{(-e)}$ is of full column rank, there exists $\tilde{Q} = \mathrm{diag}(Q', I_{d-e})$ having the block diagonal form (4.3) such that (4.1) holds, i.e.,

$$V_{12}^{(q)}\tilde{Q} = [\,W^{(q,e)}Q' | \,V_{12}^{(-e)}] = [\mathbf{\Omega}\|\tilde{\Gamma}_1|\,V_{12}^{(-e)}] \tag{4.7}$$

with $\tilde{\Gamma} = [\tilde{\Gamma}_1|\,V_{12}^{(-e)}]$ nonsingular. Consequently, the correction $[\tilde{G}|\tilde{E}]$ defined by (4.2) is minimal in the Frobenius norm (see Theorem 4.1), and the corresponding matrix $\tilde{X} \equiv -\tilde{Z}\tilde{\Gamma}^{-1}$ given by (3.20) represents a TLS solution (which is, by Lemma 3.7, the unique solution of the corrected system with the given fixed correction $[\tilde{G}|\tilde{E}]$). Because $\mathrm{rank}(W^{(q,e)}) > e$ and Q' is orthogonal, the product $W^{(q,e)}Q' = [\mathbf{\Omega}|\tilde{\Gamma}_1]$, where $\mathrm{rank}(\tilde{\Gamma}_1) = e$ ($\tilde{\Gamma}$ is nonsingular), leads always to a nonzero $\mathbf{\Omega}$. On the other hand, the construction (3.15)–(3.17) always leads to $\mathbf{\Omega} = 0$. Hence, the matrix $X^{(q)}$ given by (3.17) does not represent a TLS solution.

The following theorem completes the argument by showing that any problem from the set $\mathcal{F}_2$ always has a minimum norm TLS solution.

THEOREM 4.5. *Let* (1.1)–(1.3) *be the TLS problem of the* 1*st class belonging to the set* $\mathcal{F}_2$. *Then there exist TLS solutions given by* (3.18)–(3.20) *minimal in the* 2*-norm, and in the Frobenius norm, respectively.*

Proof. A TLS solution $\tilde{X} = -\tilde{Z}\tilde{\Gamma}^{-1}$ is obtained from the formula

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]\hat{Q} = \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]\left[\begin{array}{c|c|c} Q_1' & Q_2' & 0 \\ \hline 0 & 0 & I_{d-e} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{\Omega} & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right],$$

where the block diagonal matrix $\hat{Q}$ is the orthogonal matrix (4.3) from Theorem 4.1. The TLS solution is uniquely determined by the orthogonal matrix $Q' \equiv [Q_1'|Q_2'] \in \mathbb{R}^{(q+e)\times(q+e)}$.

In our construction, $Q' \in \mathbb{R}^{(q+e)\times(q+e)}$ is required to lead to a nonsingular $\tilde{\Gamma}$. Since the matrix inversion is a continuous function of entries of a nonsingular matrix, and matrix multiplication is a continuous function of entries of both factors, the matrix $\tilde{X} = -\tilde{Z}\tilde{\Gamma}^{-1}$ is a continuous matrix-valued function of Q'. Define two nonnegative functionals $\mathbf{N}_2(Q'):\mathbb{R}^{(q+e)\times(q+e)} \to [0,+\infty]$ and $\mathbf{N}_F(Q'):\mathbb{R}^{(q+e)\times(q+e)} \to [0,+\infty]$ on a set of all $(q+e)$-by-$(q+e)$ orthogonal matrices such that

$$\mathbf{N}_2(Q') \equiv \begin{cases} \|\tilde{X}(Q')\|_2 & \text{if } Q' \text{ gives } \tilde{\Gamma}(Q') \text{ nonsingular,} \\ +\infty & \text{if } Q' \text{ gives } \tilde{\Gamma}(Q') \text{ singular.} \end{cases}$$

The functional $\mathbf{N}_F(Q')$ is defined analogously. Note that both functionals are nonnegative and lower semicontinuous on the compact set of all $(q+e)$-by-$(q+e)$ orthogonal matrices, and thus both functionals have a minimum on this set. □

Theorem 4.5 does not address the uniqueness of the minimum norm solutions, and it also does not give any practical algorithm for computing them. Further note that the sets of solutions minimal in 2-norm and minimal in the Frobenius norm can be *different or even disjoint*. This fact can be illustrated with the following example. Consider the problem given by its SVD decomposition

$$[B|A] \equiv U \left[\begin{array}{c|ccc} 3 & 0 & 0 & 0 \\ \hline 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \left(\frac{1}{4} \left[\begin{array}{c|ccc} -1 & -3 & \sqrt{3} & \sqrt{3} \\ \hline 3 & -1 & \sqrt{3} & -\sqrt{3} \\ \sqrt{3} & \sqrt{3} & 1 & 3 \\ \sqrt{3} & -\sqrt{3} & -3 & 1 \end{array}\right] \right)^T, \tag{4.8}$$

where $A \in \mathbb{R}^{4\times 2}$, $B \in \mathbb{R}^{4\times 2}$ (it is easy to verify that $A^T B \neq 0$). Here $q = 1$, $e = 1$,

$$W^{(q,e)} = \frac{1}{4}\begin{bmatrix} -3 & \sqrt{3} \\ -1 & \sqrt{3} \end{bmatrix}, \qquad V_{12}^{(-e)} = \frac{1}{4}\begin{bmatrix} \sqrt{3} \\ -\sqrt{3} \end{bmatrix}$$

have rank two and one, respectively. This problem is of the 1st class and belongs to the set $\mathcal{F}_2$. The TLS solution is determined by the orthogonal matrix

$$\hat{Q} = \left[\begin{array}{c|c||c|c} Q_1' & Q_2' & 0 \\ \hline\hline 0 & 0 & I_{d-e} \end{array}\right] = \left[\begin{array}{c|c||c} \cos(\phi) & -\sin(\phi) & 0 \\ \hline \sin(\phi) & \cos(\phi) & 0 \\ \hline\hline 0 & 0 & 1 \end{array}\right],$$

which depends on only one real variable ϕ. Figure 4.1 shows how the 2-norm and the Frobenius norm of the TLS solution depend on the value of ϕ. From the behavior of the norms it is clear that the set of solutions minimal in the 2-norm has no intersection with the set of solutions minimal in the Frobenius norm. If we use in the previous example (4.8) the matrix of the right singular vectors

$$V = \frac{1}{2}\left[\begin{array}{c|ccc} 0 & 1 & 0 & \sqrt{3} \\ \hline -1 & 0 & \sqrt{3} & 0 \\ \sqrt{3} & 0 & 1 & 0 \\ 0 & -\sqrt{3} & 0 & 1 \end{array}\right],$$

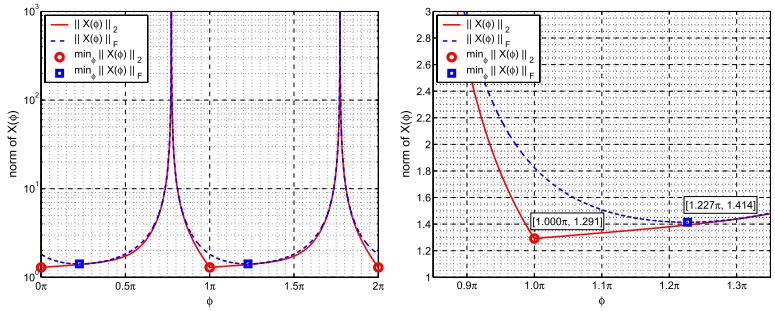


FIG. 4.1. *(Left plot) The* 2*-norm and the Frobenius norm of TLS solutions of the problem* (4.8) *belonging to the set* $\mathcal{F}_2$*. Solutions minimal in different norms are distinct. (Right plot) Detail of the solutions minimal in the* 2*-norm and in the Frobenius norm.*

then there exists a solution which is minimal in both the 2-norm and the Frobenius norm.

4.3. The set $\mathcal{F}_3$—problems of the 1st class which do not have a TLS solution. Consider a TLS problem of the 1st class from the set $\mathcal{F}_3$, i.e., the case with $\mathrm{rank}(V_{12}^{(-e)}) < d - e$. Since $V_{12}^{(-e)}$ in (4.5) is rank deficient, $\tilde{Q} \in \mathbb{R}^{(q+d)\times(q+d)}$ in the block diagonal form (4.3) leads to (4.7) with $\tilde{\Gamma} = [\tilde{\Gamma}_1 | V_{12}^{(-e)}]$ containing linearly dependent column(s). Thus $\tilde{\Gamma}$ in (4.1) is always singular. Consequently, in this case there does not exist $\tilde{Q}$ in the block diagonal form yielding $\tilde{\Gamma}$ nonsingular. Therefore, there is no correction $[\tilde{G}|\tilde{E}]$ having the norm (4.4) which makes the system (1.1) compatible; see Theorem 4.1.

Now we show that a TLS solution does not exist for the problems from the set $\mathcal{F}_3$. Using a general matrix $\tilde{Q}$ (see (3.18)), we construct a correction (3.19) which makes the system compatible, and the norm of this correction is arbitrarily close to the lower bound (4.4). Denote $\rho \equiv (d-e) - \mathrm{rank}(V_{12}^{(-e)})$ the rank defect of $V_{12}^{(-e)}$. Analogously to section 4.2, there exists an orthogonal matrix $Q' \in \mathbb{R}^{(q+e)\times(q+e)}$ such that

$$V_{12}^{(q)} \mathrm{diag}(Q', I_{d-e}) = [\, W^{(q,e)} Q' | \, V_{12}^{(-e)}] = [\mathbf{\Omega} \| \tilde{\Gamma}_1 | \, V_{12}^{(-e)}]$$

with $\mathrm{rank}([\tilde{\Gamma}_1 | \, V_{12}^{(-e)}]) = d - \rho$; compare with (4.7). Let $\mathcal{J} = \{j_1, \ldots, j_\rho\}$ denote indices of any ρ columns of $V_{12}^{(-e)}$ such that the remaining columns of $V_{12}^{(-e)}$ (with indices $\{1, \ldots, d-e\} \setminus \mathcal{J}$) are linearly independent. Because $\mathrm{rank}(V_{12}^{(q)}) = d$, the matrix $\mathbf{\Omega}$ has ρ linearly independent columns which are not in $\mathcal{R}([\tilde{\Gamma}_1 | \, V_{12}^{(-e)}])$; let $\mathcal{K} = \{k_1, \ldots, k_\rho\}$ denote their indices. Consider an angle θ, $0 < \theta < \pi$. A Givens rotation corresponding to θ applied subsequently on pairs of columns with indices j_ℓ and k_ℓ, for $\ell = 1, \ldots, \rho$, can be written as an orthogonal transformation

$$[\mathbf{\Omega} \| \tilde{\Gamma}_1 | \, V_{12}^{(-e)}] \begin{bmatrix} C_{11} & 0 & S_{12} \\ \hline 0 & I_e & 0 \\ \hline -S_{12}^T & 0 & C_{22} \end{bmatrix} = [\hat{\mathbf{\Omega}} \| \, \tilde{\Gamma}_1 \, | \, \hat{V}_{12}^{(-e)}],$$

where $C_{11} \in \mathbb{R}^{q\times q}$ and $C_{22} \in \mathbb{R}^{(d-e)\times(d-e)}$ are diagonal matrices having ρ diagonal entries (on the positions (k_ℓ, k_ℓ) and (j_ℓ, j_ℓ), $\ell = 1, \ldots, \rho$, respectively) equal to $\cos(\theta)$ (the other diagonal entries are equal to one), and $S_{12} \in \mathbb{R}^{q\times(d-e)}$ has entries on positions (k_ℓ, j_ℓ), $\ell = 1, \ldots, \rho$, equal to $\sin(\theta)$ (the other entries are zero). Since $0 < \theta < \pi$, the

matrix $\tilde{\Gamma} = [\tilde{\Gamma}_1 | \hat{V}_{12}^{(-e)}]$ is nonsingular, and thus the corresponding correction makes the system compatible. The transformation matrix

$$\tilde{Q} = \left[\begin{array}{c|c} Q' \mathrm{diag}(C_{11}, I_e) & Q' \left[\begin{smallmatrix} S_{12} \\ 0 \end{smallmatrix}\right] \\ \hline [-S_{12}^T | 0] & C_{22} \end{array}\right]$$

can be, with $\theta \to 0$, arbitrarily close to the block diagonal form (4.3), and moreover the Frobenius norm of the corresponding correction

$$\|[\tilde{G}|\tilde{E}]\|_F = \left(\sum_{j=n+1}^{n+d} \sigma_j^2 + \sin^2(\theta) \sum_{j \in \mathcal{J}} (\sigma_{n+1}^2 - \sigma_{n+e+j}^2) \right)^{1/2}$$

can be arbitrarily close to the lower bound given by (4.4).

Consequently, there is *no minimal correction* that makes the system (1.1) compatible. The TLS problem (1.1)–(1.3) with rank deficient $V_{12}^{(-e)}$ does not have a solution.

4.4. Correction corresponding to the matrix $X^{(q)}$. In the previous three sections we have shown that a TLS solution (if it exists) always has the correction matrix with the Frobenius norm (4.4). We can formulate the following corollary.

Corollary 4.6. *Consider a TLS problem* (1.1)–(1.3) *of the* 1*st class. The construction* (3.13)–(3.17) *yields the TLS solution* $X_{\mathrm{TLS}} \equiv X^{(q)}$ *if and only if there exists an orthogonal matrix* $\hat{Q}$ *in the block diagonal form* (4.3) *such that substituting* $\hat{Q}$ *for* Q *in* (3.13)–(3.15) *gives the same correction* $[G|E]$.

Now we focus on the properties of the correction $[G|E]$ given by (3.15) in general. First we prove an auxiliary lemma.

Lemma 4.7. *Let* $[G|E]$ *be the correction matrix given by* (3.15). *Denote* $s \equiv \mathrm{rank}([B|A])$. *Then the ranks of the correction and corrected matrix satisfy*

$$\min\{s, d\} \geq \mathrm{rank}([G|E]) \geq \max\{0, s - n\}, \tag{4.9}$$

$$\max\{0, s - d\} \leq \mathrm{rank}([B + G|A + E]) \leq \min\{s, n\}. \tag{4.10}$$

Proof. The upper bound in (4.9) follows immediately from (3.15). The lower bound in (4.9) follows from the fact that the correction matrix makes the system compatible, i.e., the resulting rank of $[B + G|A + E]$ is at most n, which also proves the upper bound in (4.10). Since the rank of $[G|E]$ is at most d, the lower bound in (4.10) follows trivially. □

The result of the following theorem can also be found in [22, eq. (5.4)].

Theorem 4.8. *Let* $[G|E]$ *be the correction matrix given by* (3.15). *Then its Frobenius norm satisfies*

$$\left(\sum_{j=p+1}^{p+d} \sigma_j^2 \right)^{1/2} \geq \|[G|E]\|_F \geq \left(\sum_{j=n+1}^{n+d} \sigma_j^2 \right)^{1/2}. \tag{4.11}$$

Proof. The lower bound in (4.11) is trivial. The matrix $[G|E]$ has from (4.9) the rank not greater than $\min\{s, d\}$, which immediately gives the upper bound. From the construction (3.15) a rank d matrix of the given form cannot have Frobenius norm larger than (4.11). □

Since the Frobenius norm of the correction $[G|E]$ given by (3.15) can be larger than $(\sum_{j=n+1}^{n+d}\sigma_j^2)^{1/2}$, the correction need not be minimal, and (3.17) need not represent (as described above) a TLS solution. Further note that the inequalities in (4.11) become equalities if and only if

$$\sigma_{p+j} = \sigma_{n+j}, \qquad j = 1, \ldots, d$$

(recall that $n = p + q$). This happens *either* if $q = 0$ (the case with the unique solution discussed in section 3.1), *or* if $\sigma_{p+1} = \cdots = \sigma_{n+d}$ (the special case discussed in section 3.2).

5. Problems of the 2nd class. In this section we briefly describe problems (1.1)–(1.3) of the 2nd class, i.e., the problems for which $V_{12}^{(q)}$ does not have full row rank; see Definition 2.2. Here the right singular vector subspace given by the last $(q + d)$ singular vectors $v_{p+1}, \ldots, v_{n+d}$ does not contain sufficient information for constructing a solution (3.20), and the problems of the 2nd class do not have a TLS solution (the argumentation is analogous to that in section 4.3).

The classical TLS algorithm, which gives an output also for problems of the 2nd class, is derived in [17] by a straightforward generalization of the single right-hand side concept. The right singular vector subspace $\mathcal{R}([\,V_{12}^{(q)\,T}\,|\,V_{22}^{(q)\,T}\,]^T)$ used for the construction (3.13)–(3.17) in previous cases is extended with additional right singular vectors until, for some t, a full row rank block $V_{12}^{(t)} \in \mathbb{R}^{d\times(t+d)}$ is found in the upper right corner of V (and $V_{12}^{(t-1)}$ is, at the same time, rank deficient),

$$V = \begin{array}{c} \begin{array}{cc} q & d \end{array} \\ \left[\begin{array}{c|c} & \\ & \\ \hline & \\ & \end{array}\right] \begin{array}{l} d \\ n \end{array} \\ \begin{array}{ccc} n-t & t & d \end{array} \end{array} = \left[\begin{array}{c|c} V_{11}^{(t)} & V_{12}^{(t)} \\ \hline V_{21}^{(t)} & V_{22}^{(t)} \end{array}\right].$$

Then the matrix $X^{(t)} = -V_{22}^{(t)} V_{12}^{(t)\dagger}$ with the corresponding correction can be constructed analogously to (3.13)–(3.17) with q replaced by t. Obviously, this matrix might not be uniquely defined when σ_{n-t+1} is not simple, in particular, when $\sigma_{n-t} = \sigma_{n-t+1}$. In order to handle a possible multiplicity of σ_{n-t+1}, it is convenient to consider the notation

$$\sigma_{n-\tilde{q}} > \sigma_{n-\tilde{q}+1} = \cdots = \sigma_{n-t} = \sigma_{n-t+1} \geq \sigma_{n-t+2},$$

where $\tilde{q} \geq t$; put for simplicity $n - \tilde{q} \equiv \tilde{p}$. (If such $\sigma_{n-\tilde{q}} \equiv \sigma_{\tilde{p}}$ does not exist, then put $\tilde{q} \equiv n$.) The condition that $V_{12}^{(\tilde{q})}$ is of full row rank equal to d is readily satisfied, since $V_{12}^{(\tilde{q})}$ extends $V_{12}^{(t)}$. Then $X^{(\tilde{q})}$ and $[G|E]$ can be constructed as in (3.13)–(3.17) with q replaced by $\tilde{q}$. Thus, the matrix $X^{(\tilde{q})} \equiv -V_{22}^{(\tilde{q})} V_{12}^{(\tilde{q})\dagger}$ represents a solution of the compatible corrected system $(A + E)X = B + G$. The Frobenius and the 2-norm of the matrix $X^{(\tilde{q})}$ are given by Lemma 3.5. Similarly to the problems of the 1st class, the minimality property (3.22) of $X^{(\tilde{q})}$ can be shown. Thus, $X^{(\tilde{q})}$ has minimal Frobenius and 2-norm over all matrices $\tilde{X}$ that can be obtained from the construction analogous to (3.18)–(3.20) with q replaced by $\tilde{q}$. The substitution of $\tilde{q}$ for t ensures the uniqueness of the construction and leads to the matrix with the smallest norm. On the other hand, it inevitably increases the norm of the correction, with

$$\|[G|E]\|_F > \left(\sum_{j=n+1}^{n+d} \sigma_j^2\right)^{1/2}.$$

The Frobenius norm of $[G|E]$ is strictly larger than the smallest possible correction reducing the rank of $[B|A]$ to n, and the matrix $X^{(\tilde{q})}$ does not represent a TLS solution.[2]

6. Summary of the relationship to the classical TLS algorithm. The classical TLS algorithm gives for any data the output $X^{(\kappa)}$ which is equal (in exact arithmetic) either to $X^{(q)}$ given by (3.2), or by (3.10), or by (3.17), or to $X^{(\tilde{q})}$ described in section 5.

Algorithm 1 (the classical TLS algorithm).
A fully documented Fortran 77 implementation is given in [15], [16]. The code can be obtained through Netlib.org, cf. http://www.netlib.org/vanhuffel.

Require: $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times d}$ {here the SVD of $[B|A]$ in the form (2.3)–(2.6)}
1: $\Delta \leftarrow 0$
2: **if** $\operatorname{rank}(V_{12}^{(\Delta)}) = d$ **and** $\Delta = n$, **then goto** 6
3: **if** $\operatorname{rank}(V_{12}^{(\Delta)}) = d$ **and** $\sigma_{n-\Delta} > \sigma_{n-\Delta+1}$, **then goto** 6
4: $\Delta \leftarrow \Delta + 1$
5: **goto** 2
6: $\kappa \leftarrow \Delta$
7: $X^{(\kappa)} \leftarrow -V_{22}^{(\kappa)} V_{12}^{(\kappa)\dagger}$
8: **return** κ, $X^{(\kappa)}$

The output $X^{(\kappa)}$ is called a generic (or TLS) solution in [17] for any problem of the 1st class, and it is called a nongeneric solution in [17] for any problem of the 2nd class. As our new partitioning and the included classification reveals,

(i) if the problem is of the 1st class and $\operatorname{rank}(W^{(q,e)}) = e$ (i.e., the problem belongs to the set $\mathcal{F}_1$), then $X^{(\kappa)} \equiv X_{\mathrm{TLS}}$ represents a TLS solution (it solves the TLS problem (1.1)–(1.3)), $\kappa \equiv q$;
(ii) if the problem is of the 1st class and $\operatorname{rank}(W^{(q,e)}) > e$ (i.e., the problem belongs to the set $\mathcal{F}_2 \cup \mathcal{F}_3$), then $X^{(\kappa)}$ does not represent a TLS solution, which exists for the problems in the set $\mathcal{F}_2$ but does not exist for the problems in the set $\mathcal{F}_3$, $\kappa \equiv q$;
(iii) if the problem is of the 2nd class, (i.e., the problem belongs to the set $\mathcal{S}$), then $X^{(\kappa)}$ does not represent a TLS solution (a TLS solution does not exist), $\kappa \equiv \tilde{q}$.

For $d = 1$ (single right-hand side case) the output $X^{(\kappa)}$ of Algorithm 1 represents the TLS solution of the core problem (3.5) transformed to the original coordinate system. The output $X^{(\kappa)}$ has two further important interpretations.

Lemma 6.1 (the constrained total least squares (C-TLS)). *The matrix* $X^{(\kappa)} = -V_{22}^{(\kappa)} V_{12}^{(\kappa)\dagger}$ *given by Algorithm* 1 *represents the unique solution of the constrained minimization problem*

$$\min_{X,E,G} \|[G|E]\|_F \qquad \textit{subject to} \qquad (A+E)X = B+G, \tag{6.1}$$

[2] The matrix $X^{(\tilde{q})} = -V_{22}^{(\tilde{q})} V_{12}^{(\tilde{q})\dagger}$ is called a *nongeneric solution* in [17, Definition 3.3, p. 78].

$$\text{and } [G|E]\begin{bmatrix}0\\ w\end{bmatrix}=0 \qquad \text{for all} \qquad \begin{bmatrix}0\\ w\end{bmatrix}\in\mathcal{R}\left(\begin{bmatrix}V_{12}^{(\kappa)}\\ V_{22}^{(\kappa)}\end{bmatrix}\right) \tag{6.2}$$

with the correction $[G|E]$ *given by* (3.15) *(with* q *possibly replaced by* $\tilde{q}$*).*

The additional constraint (6.2) can be equivalently rewritten as

$$[G|E]\begin{bmatrix}0\\ Y\end{bmatrix}=0,$$

where Y is defined analogously to (3.13). Since $\sigma_{n-\kappa}>\sigma_{n-\kappa+1}$, the correction matrix in (6.1)–(6.2) is unique. Consequently, the constrained problem (6.1)–(6.2) has the unique solution $X_{\text{C-TLS}}\equiv X^{(\kappa)}$. Furthermore, since the matrix in (3.13) (with q possibly replaced by $\tilde{q}$) has orthonormal columns, $X^{(\kappa)T}Y=-(\Gamma^{-1})^TZ^TY=0$, and the additional constraint implies that $X^{(\kappa)T}w=0$ for all w from (6.2); see [17, Eq. 3.101, p. 79], [21], [22]. Note that the problem (6.1)–(6.2) for $\kappa\equiv\tilde{q}$ is considered a *definition* of the non-generic solution in [17, Definition 3.3, p. 78 and Theorem 3.15, pp. 80–82].

LEMMA 6.2 (the truncated total least squares (T-TLS)). *The matrix* $X^{(\kappa)}=-V_{22}^{(\kappa)}V_{12}^{(\kappa)\dagger}$ *given by Algorithm* 1 *represents the unique minimum norm TLS solution of the modified TLS problem*

$$\min_{X,\hat{E},\hat{G}}\|[\hat{G}|\hat{E}]\|_F \qquad \text{subject to} \qquad (\hat{A}+\hat{E})X=\hat{B}+\hat{G}, \tag{6.3}$$

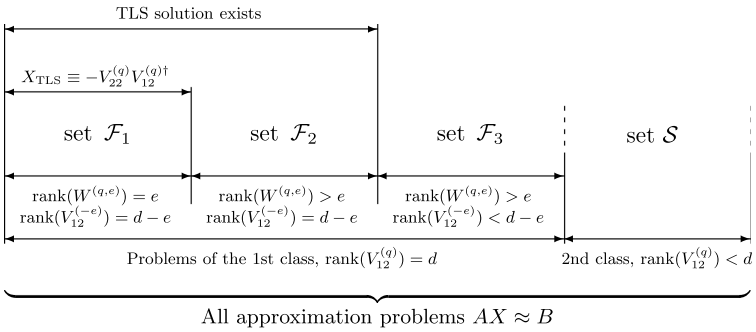
$$\text{where } [\hat{B}|\hat{A}]=\left(\sum_{j=1}^{n-\kappa}u_j\sigma_jv_j^T\right)+\sigma_{n-\kappa+1}\left(\sum_{j=n-\kappa+1}^{n+d}u_jv_j^T\right)$$

with the corresponding correction $[\hat{G}|\hat{E}]$, $\|[\hat{G}|\hat{E}]\|_F=\sigma_{n-\kappa+1}\sqrt{d}$.

The problem (6.3) is clearly a TLS problem of the 1st class (belonging to the set $\mathcal{F}_1$). Moreover, it is a special case described in section 3.2. This problem is called *truncated total least squares* (T-TLS) problem for the given A, B with the solution $X_{\text{T-TLS}}\equiv X^{(\kappa)}$; see [17, note on p. 82]. It is worth noting that the T-TLS concept allows us to assume that the original problem $AX\approx B$ is a perturbation of the modified problem $\hat{A}X\approx\hat{B}$. From the T-TLS point of view, any TLS problem may be interpreted as a perturbed problem of the 1st class with the special singular values distribution (3.6). Since $X_{\text{T-TLS}}=X^{(\kappa)}$, Algorithm 1 can be used as a relatively simple and useful regularization technique; see, e.g., [21], [2], [3] (for $d=1$) and also [17, algorithm and comments in section 3.6.1, pp. 87–90]. The distribution of the smallest singular values of $[B|A]$ plays no role in the algorithm output.

The *true* TLS solution (if it exists) does not have this regularization property. The TLS solution uses information about the smallest singular values of $[B|A]$.

7. Conclusions. We have presented a new classification of TLS problems with multiple right-hand sides. Each TLS problem falls into one of four distinct sets. The union of the first three sets $\mathcal{F}_j$, $j=1$, 2, 3, contains problems of the 1st class. It is complemented by the set $\mathcal{S}$ of problems of the 2nd class, as illustrated by the following schema.



It has been shown that the special cases analyzed in [17] belong to the set $\mathcal{F}_1$. We have proved that *any* problem from $\mathcal{F}_1 \cup \mathcal{F}_2$ has a TLS solution, whereas problems from $\mathcal{F}_3 \cup \mathcal{S}$ do *not* have a TLS solution. Moreover, for any problem from $\mathcal{F}_1 \cup \mathcal{F}_2$ there exist a TLS solution minimal in the 2-norm and a solution minimal in the Frobenius norm, but for the problems from the set $\mathcal{F}_2$ the minimum norm solutions can be distinct.

The classical TLS algorithm (Algorithm 1) computes a TLS solution only for problems belonging to the set $\mathcal{F}_1$. We have not provided an efficient algorithm for computing a TLS solution for the problems from $\mathcal{F}_2$ (where it exists). It can possibly be obtained using a nonlinear optimization over a parameterization of the set of corresponding orthogonal matrices. However, this optimization is hardly practically applicable.

The TLS problems with $d = 1$ have been clarified through the concept of the *core reduction*. An extension of this concept to a TLS problem with $d > 1$ could help to understand the discrepancy between the true TLS solution and the solution given by the classical TLS algorithm. An approach based on such a reduction, outlined in [12], will be discussed elsewhere.

Acknowledgments. We wish to thank Daniel Kressner and two anonymous referees for their comments which led to improvements of our manuscript.

REFERENCES

[1] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.
[2] R. D. Fierro and J. R. Bunch, *Collinearity and total least squares*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1167–1181.
[3] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Regularization by truncated total least squares*, SIAM J. Sci. Comput., 18 (1997), pp. 1223–1241.
[4] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[5] G. H. Golub, A. Hoffman, and G. W. Stewart, *A generalization of the Eckart–Young–Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.
[6] G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.
[7] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.
[8] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Q. J. Math, 11 (1960), pp. 50–59.
[9] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., 91 (2002), pp. 117–146.
[10] C. C. Paige and Z. Strakoš, *Unifying least squares, total least squares and data least squares*, in Total Least Squares and Errors-in-Variables Modeling, S. Van Huffel and P. Lemmerling, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 25–34.

[11] C. C. Paige and Z. Strakoš, *Core problems in linear algebraic systems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 861–875.
[12] M. Plešinger, *The Total Least Squares Problem and Reduction of Data in $AX \approx B$*, Ph.D. thesis, Academy of Sciences of the Czech Republic, Prague, Czech Republic, and Technical University of Liberec, Liberec, Czech Republic, 2008.
[13] R. C. Thompson, *Principal submatrices* IX*: Interlacing inequalities for singular values of submatrices*, Linear Algebra Appl., 5 (1972), pp. 1–12.
[14] A. van der Sluis, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.
[15] S. Van Huffel, *Documented Fortran* 77 *Programs of the Extended Classical Total Least Squares Algorithm, the Partial Singular Value Decomposition Algorithm and the Partial Total Least Squares Algorithm*, Internal report ESAT-KUL 88/1, Katholieke Universiteit Leuven, Leuven, Belgium, 1988.
[16] S. Van Huffel, *The extended classical total least squares algorithm*, J. Comput. Appl. Math., 25 (1989), pp. 111–119.
[17] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.
[18] S. Van Huffel, ed., *Recent advances in total least squares techniques and errors-in-variables modeling*, in Proceedings of the Second International Workshop on TLS and EIV, SIAM, Philadelphia, 1997.
[19] S. Van Huffel and P. Lemmerling, eds. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.
[20] A. G. Watson, *Choice of norms for data fitting and function approximation*, Acta Numer., 7 (1998), pp. 337–377.
[21] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 746–763.
[22] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math., 62 (1992), pp. 123–148.

Teorie core problému

A.2 **Článek:**[WoK] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *The core problem within a linear approximation problem $AX \approx B$ with multiple right-hand sides*, SIAM Journal on Matrix Analysis and Applications (SIMAX) (ISSN 0895-4798, eISSN 1095-7162), Volume 34, Issue 3 (2013), pp. 917–931 (15 pages). (http://epubs.siam.org/doi/abs/10.1137/120884237)

Doložitelné citace (1): Odborné články zařazené do databáze Web-of-Knowledge:

- [WoK] X.-F. Wang: *Total least squares problem with the arbitrary unitarily invariant norms*, Linear and Multilinear Algebra, Available online, 2016, 20 pages. (http://www.tandfonline.com/doi/abs/10.1080/03081087.2016.1189493)

A.3 **Článek:**[WoK] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Band generalization of the Golub–Kahan bidiagonalization, generalized Jacobi matrices, and the core problem*, SIAM Journal on Matrix Analysis and Applications (SIMAX) (ISSN 0895-4798, eISSN 1095-7162), Volume 36, Issue 2 (2015), pp. 417–434 (18 pages). (http://epubs.siam.org/doi/abs/10.1137/140968914)

Doložitelné citace (1): Odborné články zařazené do databáze Web-of-Knowledge:

- [WoK] X.-F. Wang: *Total least squares problem with the arbitrary unitarily invariant norms*, Linear and Multilinear Algebra, Available online, 2016, 20 pages. (http://www.tandfonline.com/doi/abs/10.1080/03081087.2016.1189493)

A.4 **Článek:**[WoK] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima: *Solvability of the core problem with multiple right-hand sides in the TLS sense*, SIAM Journal on Matrix Analysis and Applications (SIMAX) (ISSN 0895-4798, eISSN 1095-7162), Volume 37, Issue 3 (2016), pp. 861–876 (16 pages). (http://epubs.siam.org/doi/abs/10.1137/15M1028339)

SIAM J. MATRIX ANAL. APPL.
Vol. 34, No. 3, pp. 917–931

THE CORE PROBLEM WITHIN A LINEAR APPROXIMATION PROBLEM $AX \approx B$ WITH MULTIPLE RIGHT-HAND SIDES*

IVETA HNĚTYNKOVÁ†, MARTIN PLEŠINGER‡, AND ZDENĚK STRAKOŠ§

Abstract. This paper focuses on total least squares (TLS) problems $AX \approx B$ with multiple right-hand sides. Existence and uniqueness of a TLS solution for such problems was analyzed in the paper [I. Hnětynková et al., *SIAM J. Matrix Anal. Appl.*, 32, 2011, pp. 748–770]. For TLS problems with single right-hand sides the paper [C. C. Paige and Z. Strakoš, *SIAM J. Matrix Anal. Appl.*, 27, 2006, pp. 861–875] showed how necessary and sufficient information for solving $Ax \approx b$ can be revealed from the original data through the so-called core problem concept. In this paper we present a theoretical study extending this concept to problems with multiple right-hand sides. The data reduction we present here is based on the singular value decomposition of the system matrix A. We show minimality of the reduced problem; in this sense the situation is analogous to the single right-hand side case. Some other properties of the core problem, however, cannot be extended to the case of multiple right-hand sides.

Key words. total least squares problem, multiple right-hand sides, core problem, linear approximation problem, error-in-variables modeling, orthogonal regression, singular value decomposition

AMS subject classifications. 15A06, 15A18, 15A21, 15A24, 65F20, 65F25

DOI. 10.1137/120884237

1. Introduction.

Consider a linear approximation problem

$$AX \approx B, \qquad \text{or, equivalently,} \qquad [B|A] \begin{bmatrix} -I_d \\ X \end{bmatrix} \approx 0, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{m\times d}$, without any further assumption on the positive integers m, n, d, and $A^TB \neq 0$ (this eliminates the trivial case where it does not make sense to approximate B by a linear combination of the columns of A; see also [14]). The equivalent $[B|A]$ form in (1.1) has been chosen so that our transformations will take it to block form (as, for example, in (2.1) below for the $d = 1$ case) revealing the core problem most simply. We will focus on incompatible problems, i.e., $\mathcal{R}(B) \not\subset \mathcal{R}(A)$. If $\mathcal{R}(B) \subset \mathcal{R}(A)$, then the system $AX = B$ can be solved using standard methods. Consider changes of the coordinate systems in $\mathbb{R}^m$, $\mathbb{R}^n$, and $\mathbb{R}^d$ represented by orthogonal transformations

$$\widehat{A}\widehat{X} \equiv (P^TAQ)(Q^TXR) \approx (P^TBR) \equiv \widehat{B}, \tag{1.2}$$

*Received by the editors July 11, 2012; accepted for publication (in revised form) by J. L. Barlow April 24, 2013; published electronically July 9, 2013. This work has been supported by the GAČR grant P201/13–06684S.

http://www.siam.org/journals/simax/34-3/88423.html

†Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, and Institute of Computer Science, AS CR, Prague, Czech Republic (hnetynkova@cs.cas.cz).

‡Department of Mathematics, Technical University of Liberec, Liberec, Czech Republic, and Institute of Computer Science, AS CR, Prague, Czech Republic (martin.plesinger@tul.cz). The research of this author was partly supported by the ESF grants CZ.1.07/2.3.00/09.0155 and CZ.1.07/2.3.00/30.0065.

§Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic (strakos@karlin.mff.cuni.cz). The research of this author was partly supported by the GAČR grant 201/09/0917.

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$; or, equivalently,

$$[\widehat{B}|\widehat{A}] \begin{bmatrix} -I_d \\ \widehat{X} \end{bmatrix} \equiv \left(P^T [B|A] \begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} \right) \left(\begin{bmatrix} R^T & 0 \\ 0 & Q^T \end{bmatrix} \begin{bmatrix} -I_d \\ X \end{bmatrix} R \right) \approx 0. \tag{1.3}$$

We require that X solves (1.1) if and only if $\widehat{X} = Q^T X R$ solves (1.2) and call such problems *orthogonally invariant*. The *total least squares problem* (TLS)

$$\min_{X,E,G} \|[G|E]\|_F \qquad \text{subject to} \qquad (A+E)X = B+G \tag{1.4}$$

serves as an important example; see [12], [13], [7], [5, section 6], [18]. Mathematically equivalent problems have been independently investigated under the names *orthogonal regression* and *errors-in-variables modeling*; see [20], [21].

In [8] it is shown that even with $d = 1$ (which gives $Ax \approx b$, where b is an m-vector) the TLS problem may not have a solution and, when the solution exists, it may not be unique; see also [6, pp. 324–326]. In order to resolve this difficulty, the classical book [19] introduces the so-called *nongeneric solution*. This book also extends the TLS theory to problems with multiple right-hand sides, i.e., for $d > 1$. The existence and uniqueness of a TLS solution with $d > 1$ is then discussed in full generality in the recent paper [10], giving a new classification of all possible cases.

The sequence of papers [12], [13], and [14] by Paige and Strakoš investigates, using a unified framework, different least squares formulations for problems with $d = 1$. The last paper [14] introduced the so-called *core problem* that separates the necessary and sufficient information for solving the problem from the rest. It gives the necessary and sufficient condition for existence of the TLS solution, explains when the TLS solution exists and when it is unique, and clarifies the meaning of the nongeneric solution. For a brief summary, see also the recent paper [10, pp. 752–753], which also shows that there is a class of problems for which the classical TLS approach described in [19], [22], [23], [9, Chapter 6.3, pp. 320–327] (in particular, the so-called *classical TLS algorithm*; see [19, Chapter 3.6.1, pp. 87–90], [10, Algorithm 1, p. 767]) is unable to find the existing TLS solutions.

The first steps in generalizing the core problem theory for $d > 1$ were done by Björck in the series of talks [1], [2], [3], and also in the unpublished manuscript [4], and by Sima [16] and Sima and Van Huffel in [17]. In a theoretical study presented in this paper we further develop the data reduction suggested in [15] which gives the core problem, we investigate its properties, and prove its minimality. We do not advocate a computational technique for solving the problem. This is a matter of further work and results will be published elsewhere.

The organization of this paper is as follows. Section 2 recalls the core problem concept for a *single* right-hand side. Section 3 describes the data reduction for *multiple* right-hand sides, shows how to assemble the transformation matrices, and discusses basic properties of the reduced problem. Section 4 proves the minimality of the reduced problem and thus justifies the definition of the core problem. Section 5 concludes this paper.

Throughout this paper, $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote the range and null space of a matrix M, respectively; I_ℓ (or just I) denotes an $\ell \times \ell$ identity matrix; and $0_{\ell,\xi}$ (or just 0) denotes an $\ell \times \xi$ zero matrix. The matrices A, B, $[B|A]$, and X from (1.1) are called the *system matrix*, the *right-hand sides (or the observation) matrix*, the *extended (or data) matrix*, and the *matrix of unknowns*, respectively.

2. Data reduction in the single right-hand side case. Consider the linear approximation problem (1.1) with $d = 1$. In [14] it was shown that there exist orthogonal matrices P, Q that transform the original problem into the block form

$$P^T[b|A]\begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} -1 \\ x \end{bmatrix} = \left[\begin{array}{c|c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right]\begin{bmatrix} -1 \\ x_1 \\ x_2 \end{bmatrix} \approx 0, \tag{2.1}$$

where b_1 and A_{11} are of *minimal dimensions*. Such a transformation can be obtained using the singular value decomposition (SVD) of the system matrix A,

$$A = U\Sigma V^T, \qquad U \in \mathbb{R}^{m\times m}, \quad \Sigma \in \mathbb{R}^{m\times n}, \quad V \in \mathbb{R}^{n\times n}, \tag{2.2}$$

where $U^{-1} = U^T$, $V^{-1} = V^T$. Let A have k *distinct* nonzero singular values

$$\sigma_1 > \sigma_2 > \cdots > \sigma_k > 0, \tag{2.3}$$

and let their multiplicities be m_j, $j = 1, \ldots, k$; $\sum_{j=1}^{k} m_j = r \equiv \operatorname{rank}(A)$. Then

$$\Sigma = \operatorname{diag}(\sigma_1 I_{m_1}, \ldots, \sigma_k I_{m_k}, 0_{m-r,n-r}). \tag{2.4}$$

Consider the partitioning $U = [U_1, \ldots, U_k, U_{k+1}]$, $U_j \in \mathbb{R}^{m\times m_j}$, $j = 1, \ldots, k$, and $U_{k+1} \in \mathbb{R}^{m\times m_{k+1}}$, where $m_{k+1} \equiv m - r$ is the dimension of the null space $\mathcal{N}(A^T)$. Columns of U_j represent an orthonormal basis of the jth left singular vector subspace of A. Then

$$U^T[b|A]\begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} = U^T[b|AV] \equiv [f|\Sigma], \tag{2.5}$$

where

$$f \equiv U^T b = [f_1^T, \ldots, f_k^T, f_{k+1}^T]^T, \quad f_j \equiv U_j^T b, \quad j = 1, \ldots, k+1,$$

and

$$\varphi_j \equiv \|f_j\| \geq 0. \tag{2.6}$$

Note that $\varphi_j = 0$ if and only if b is orthogonal to the jth left singular vector subspace. In order to be conformal with the multiple right-hand sides case, we keep (unlike in [14]) the zero and nonzero components f_j together until the last permutation. Let $S_j \in \mathbb{R}^{m_j\times m_j}$, $S_j^{-1} = S_j^T$, be a Householder reflection matrix such that

$$S_j^T f_j = e_1\varphi_j, \quad e_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{m_j}, \quad j = 1, \ldots, k+1, \tag{2.7}$$

and let

$$S_\oplus = \operatorname{diag}(S_1, \ldots, S_k), \quad S_L = \operatorname{diag}(S_\oplus, S_{k+1}), \quad S_R = \operatorname{diag}(S_\oplus, I_{n-r}).$$

Note that $S_L^T\Sigma S_R = \Sigma$. The orthogonal transformation

$$(US_L)^T[b|A(VS_R)] = S_L^T[f|\Sigma S_R] = [S_L^T f|\Sigma]$$

maximizes the number of zero entries in the right-hand side vector

$$S_L^T f = [\varphi_1 e_1^T, \ldots, \varphi_k e_1^T, \varphi_{k+1} e_1^T]^T$$

(as mentioned above, we may have $\varphi_j = 0$ for some j). Let b have nonzero components f_{j_ℓ} in $\overline{n}$ left singular vector subspaces corresponding to nonzero singular values σ_j with indices $j_1, \ldots, j_{\overline{n}}$, $1 \le \overline{n} \le k$. The component f_{k+1} (the component of b in the null space of A^T) is nonzero due to the fact that the problem is incompatible. Consider the row permutation Π_L of the matrix $[S_L^T f|\Sigma]$ such that

$$\Pi_L^T S_L^T f = [b_1^T, 0]^T \equiv [\varphi_{j_1}, \ldots, \varphi_{j_{\overline{n}}}, \varphi_{k+1}, 0, \ldots, 0]^T,$$

i.e., all entries of

$$b_1 = [\varphi_{j_1}, \ldots, \varphi_{j_{\overline{n}}}, \varphi_{k+1}]^T = [\|f_{j_1}\|, \ldots, \|f_{j_{\overline{n}}}\|, \|f_{k+1}\|]^T \in \mathbb{R}^{\overline{n}+1} \tag{2.8}$$

are positive. Then there exists a column permutation Π_R of the matrix $\Pi_L^T \Sigma$ such that

$$\Pi_L^T \Sigma \Pi_R = \left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right],$$

where the block $A_{11} \in \mathbb{R}^{(\overline{n}+1)\times\overline{n}}$ is (with $b \notin \mathcal{R}(A)$) rectangular, containing *at most one* copy of each nonzero singular value σ_j on its diagonal and having the zero last row. All the other singular values are moved to the diagonal of the second block A_{22}, which can be of any shape, or nonexistent. Summarizing, we obtain

$$P^T[b|AQ] = \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline\hline 0 & 0 & A_{22} \end{array}\right], \quad P \equiv U S_L \Pi_L, \quad Q \equiv V S_R \Pi_R,$$

where $[b_1|A_{11}]$ and A_{11} are of minimal dimensions; see [14]. The corresponding transformation and conformal splitting of vector of unknowns is

$$Q^T x = (V S_R \Pi_R)^T x = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right].$$

The subproblem $[b_1|A_{11}]$, or $A_{11}x_1 \approx b_1$, contains the necessary and sufficient information for solving the original problem $Ax \approx b$, and it is called the *core problem*. The solution of the second subproblem $A_{22}x_2 \approx 0$ with the maximally dimensioned block $A_{22} \in \mathbb{R}^{(m-\overline{n}-1)\times(n-\overline{n})}$ can be considered to be $x_2 = 0$ (see the discussion in [14]) giving

$$x = Q \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = V S_R \Pi_R \left[\begin{array}{c} x_1 \\ 0 \end{array}\right]. \tag{2.9}$$

The core problem (in a different form) can also be revealed by the Golub–Kahan bidiagonalization; see [14, section 3] and [11].

For $d = 1$ the core problem has a unique TLS solution; see [14]. Using (2.9), the core problem defines the minimum norm TLS solution if it exists, or the minimum norm nongeneric solution of (1.1); see [14] and [10, section 3.1, pp. 752–753]. Computationally (assuming exact arithmetic) the solution (2.9) is for $d = 1$, therefore, identical to the output of the classical TLS algorithm given in [19, Chapter 3.6.1, pp. 87–90]; see also [10, Algorithm 1, p. 767]. Unlike in the classical TLS algorithm, the solution (2.9) is constructed in a straightforward way by separating the TLS-meaningful part $A_{11}x_1 \approx b_1$ of the problem $Ax \approx b$ from the redundant and irrelevant part represented by $A_{22}x_2 \approx 0$, $x_2 = 0$. In the rest of this paper it will be shown how to generalize the SVD-based data reduction and obtain the core problem for $d > 1$, i.e., for the problem $AX \approx B$.

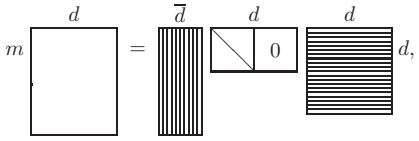
3. Data reduction in the multiple right-hand side case. Consider the problem (1.1) with $d > 1$. In this section, we construct orthogonal matrices P, Q, R that transform the original data matrix $[B|A]$ into the block form

$$P^T[B|A]\begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} = [P^TBR|P^TAQ] = \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right], \tag{3.1}$$

where B_1 and A_{11} are of minimal dimensions (the proof of minimality will be given in section 4). The orthogonal transformation (3.1) is done in four successive steps: preprocessing of the right-hand side B (section 3.1), transformation of the system matrix A (SVD of A) (section 3.2), transformation of the right-hand side B (section 3.3), and final permutation (section 3.4).

3.1. Preprocessing of the right-hand side. Let $\overline{d} \equiv \operatorname{rank}(B) \leq \min\{m, d\}$. Consider the SVD of B in the form

$$B = S\Theta R^T, \qquad S \in \mathbb{R}^{m\times\overline{d}}, \quad \Theta \in \mathbb{R}^{\overline{d}\times d}, \quad R \in \mathbb{R}^{d\times d}, \tag{3.2}$$



where S has mutually orthonormal columns, i.e., $S^TS = I_{\overline{d}}$, Θ is of *full row rank*, and R is square, i.e., $R^{-1} = R^T$. (Note that the schema illustrates the case $m > d > \overline{d}$; other cases can be illustrated analogously.) We will see later that this R plays the role of the transformation matrix R in (3.1). If $\overline{d} < d$, then B contains linearly dependent columns representing redundant information that can be removed from the original problem (1.1). Multiplication of (1.1) from the right by R gives

$$A(XR) \approx BR, \tag{3.3}$$

where

$$\begin{aligned} BR = S\Theta \equiv [C, 0] \in \mathbb{R}^{m\times d}, \qquad C \in \mathbb{R}^{m\times\overline{d}}, \qquad \text{and} \\ XR \equiv [Y, Y'] \in \mathbb{R}^{n\times d}, \qquad Y \in \mathbb{R}^{n\times\overline{d}}. \end{aligned} \tag{3.4}$$

If $d = \overline{d}$, then $BR = C$, $XR = Y$. With this notation

$$[BR|A]\begin{bmatrix} -I_d \\ XR \end{bmatrix} = [C, 0|A]\begin{bmatrix} -I_{\overline{d}} & 0 \\ 0 & -I_{d-\overline{d}} \\ Y & Y' \end{bmatrix} = [AY - C|AY'] \approx 0. \tag{3.5}$$

The original problem (1.1) is in this way split into two subproblems,

$$AY \approx C \qquad \text{and} \qquad AY' \approx 0, \tag{3.6}$$

where the second problem is homogeneous. Following the arguments in [14], we consider the meaningful solution $Y' \equiv 0$. In this way, the approximation problem (1.1) reduces to

$$AY \approx C, \qquad \text{or, equivalently,} \qquad [C|A]\begin{bmatrix} -I_{\overline{d}} \\ Y \end{bmatrix} \approx 0, \tag{3.7}$$

where $A \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times\overline{d}}$, and $C \in \mathbb{R}^{m\times\overline{d}}$ is of *full column rank*. From (3.2)–(3.4) it follows that the right-hand side matrix C has *mutually orthogonal columns*.

3.2. Transformation of the system matrix. Consider the SVD of A given by (2.2) with (2.3) and (2.4). The problem (3.7) can then be transformed analogously to (2.5),

$$\big(U^T A V\big)\big(V^T Y\big) = \Sigma Z \approx F, \tag{3.8}$$

where $Z \equiv V^T Y$, $F \equiv U^T C$. Equivalently,

$$[F|\Sigma] \begin{bmatrix} -I_{\overline{d}} \\ Z \end{bmatrix} \approx 0. \tag{3.9}$$

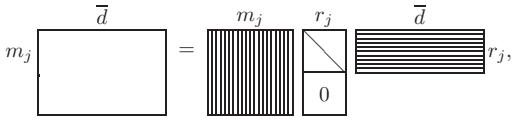
The approximation problem $\Sigma Z \approx F$ has the full column rank right-hand side matrix F with mutually orthogonal columns and the system matrix Σ in a diagonal form.

3.3. Transformation of the right-hand side. Similarly to the single right-hand side case, we now transform the right-hand side matrix of (3.8) in order to get as many zero rows as possible. Consider a partitioning of F into the block-rows with respect to the multiplicities of the singular values of the system matrix A, i.e.,

$$F = [F_1^T, \ldots, F_k^T, F_{k+1}^T]^T, \qquad \text{where} \qquad F_j \in \mathbb{R}^{m_j \times \overline{d}}, \qquad j = 1, \ldots, k, k+1.$$

Let $r_j \equiv \operatorname{rank}(F_j) \le \min\{m_j, \overline{d}\}$. Consider the SVD of F_j in the form

$$F_j = S_j \Theta_j W_j^T, \qquad S_j \in \mathbb{R}^{m_j \times m_j}, \quad \Theta_j \in \mathbb{R}^{m_j \times r_j}, \quad W_j \in \mathbb{R}^{\overline{d} \times r_j}, \tag{3.10}$$



where S_j is square, i.e., $S_j^{-1} = S_j^T$, Θ_j is of *full column rank*, and W_j has mutually orthonormal columns, i.e., $W_j^T W_j = I_{r_j}$, $j = 1, \ldots, k, k+1$. (Note that, as above, the schema illustrates the case $r_j < m_j < \overline{d}$.) The matrix S_j generalizes the role of the identically denoted matrix in section 2; see (2.7). Consider the block diagonal orthogonal matrices

$$S_\oplus = \operatorname{diag}(S_1, \ldots, S_k), \quad S_L = \operatorname{diag}(S_\oplus, S_{k+1}), \quad S_R = \operatorname{diag}(S_\oplus, I_{n-r}), \tag{3.11}$$

where as in (2.4) $r = \operatorname{rank}(A)$, and recall that

$$\Sigma = \operatorname{diag}(\sigma_1 I_{m_1}, \ldots, \sigma_k I_{m_k}, 0_{m-r,n-r}) \in \mathbb{R}^{m \times n}, \qquad \sigma_1 > \sigma_2 > \cdots > \sigma_k > 0;$$

see (2.3), (2.4). Then $\Sigma Z \approx F$ from (3.8) can be transformed to

$$\big(S_L^T \Sigma S_R\big)\big(S_R^T Z\big) \approx S_L^T F \qquad \text{and} \qquad S_L^T \Sigma S_R = \Sigma. \tag{3.12}$$

Equivalently, (3.9) becomes

$$\big[S_L^T F|\Sigma\big] \begin{bmatrix} -I_{\overline{d}} \\ S_R^T Z \end{bmatrix} \approx 0,$$

and the extended (data) matrix has the form

$$\big[S_L^T F|\Sigma\big] = \left[\begin{array}{c|ccc} \Theta_1 W_1^T & \sigma_1 I_{m_1} & & 0 \quad 0 \\ \vdots & & \ddots & \vdots \\ \Theta_k W_k^T & 0 & & \sigma_k I_{m_k} \quad 0 \\ \Theta_{k+1} W_{k+1}^T & 0 & \cdots & 0 \quad 0 \end{array}\right] \in \mathbb{R}^{m \times (n+\overline{d})}. \tag{3.13}$$

If $m_j > r_j$, then the block $S_j^T F_j = \Theta_j W_j^T$ contains zero rows at the bottom; see (3.10). Denote

$$\Theta_j W_j^T \equiv \begin{bmatrix} \Phi_j \\ 0 \end{bmatrix}, \qquad \Phi_j \in \mathbb{R}^{r_j \times \overline{d}}, \tag{3.14}$$

where Φ_j is the block of nonzero rows (if $r_j = 0$, then the block Φ_j has *no rows*). For $r_j = m_j$ we simply have $\Theta_j W_j^T \equiv \Phi_j$. It follows from (3.10) that Φ_j has *mutually orthogonal rows*. The matrix Φ_j generalizes the role of the number φ_j; see (2.6), (2.8).

3.4. Final permutation. Now the aim is to find a permutation of (3.13) that reveals the block diagonal structure (3.1). This can be done analogously to the single right-hand side case (see also [14, section 2]), by moving the rows of (3.13) with zero blocks in $\Theta_j W_j^T$ (see (3.14)) to the bottom submatrix of the whole matrix (see the matrix in the middle row of (3.15)) with subsequent moving of the corresponding columns with the diagonal blocks $\sigma_j I_{m_j - r_j}$ in the bottom to the right,

$$\Pi_L^T \left[\begin{array}{c|ccc|cc} \Theta_1 W_1^T & \sigma_1 I_{m_1} & & & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ \Theta_k W_k^T & & & \sigma_k I_{m_k} & 0 & 0 \\ \Theta_{k+1} W_{k+1}^T & 0 & \cdots & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{cc} I_{\overline{d}} & 0 \\ 0 & \Pi_R \end{array}\right]$$

$$= \left[\begin{array}{c|ccc|ccc} \Phi_1 & \sigma_1 I_{r_1} & & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots & & \vdots & \vdots \\ \Phi_k & 0 & & \sigma_k I_{r_k} & 0 & \cdots & 0 & 0 \\ \Phi_{k+1} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \hline 0 & 0 & \cdots & 0 & \sigma_1 I_{m_1 - r_1} & & 0 & 0 \\ \vdots & \vdots & & \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & & \sigma_k I_{m_k - r_k} & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \end{array}\right]$$

$$\equiv \left[\begin{array}{c|c|c} B_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right]. \tag{3.15}$$

Here $\Pi_L \in \mathbb{R}^{m \times m}$ is given by

$$\Pi_L \equiv \left[\begin{array}{cccc|cccc} \left[\begin{smallmatrix} I_{r_1} \\ 0 \end{smallmatrix}\right] & & 0 & 0 & \left[\begin{smallmatrix} 0 \\ I_{m_1 - r_1} \end{smallmatrix}\right] & & 0 & 0 \\ & \ddots & & \vdots & & \ddots & & \vdots \\ 0 & & \left[\begin{smallmatrix} I_{r_k} \\ 0 \end{smallmatrix}\right] & 0 & 0 & & \left[\begin{smallmatrix} 0 \\ I_{m_k - r_k} \end{smallmatrix}\right] & 0 \\ 0 & \cdots & 0 & \left[\begin{smallmatrix} I_{r_{k+1}} \\ 0 \end{smallmatrix}\right] & 0 & \cdots & 0 & \left[\begin{smallmatrix} 0 \\ I_{m_{k+1} - r_{k+1}} \end{smallmatrix}\right] \end{array}\right] \tag{3.16}$$

and it permutes the block-rows starting with Φ_j up, while moving the block-rows starting with zero blocks down. Analogously, $\Pi_R \in \mathbb{R}^{n \times n}$ given by

$$\Pi_R \equiv \left[\begin{array}{ccc|ccc} \left[\begin{smallmatrix} I_{r_1} \\ 0 \end{smallmatrix}\right] & & 0 & \left[\begin{smallmatrix} 0 \\ I_{m_1 - r_1} \end{smallmatrix}\right] & & 0 & 0 \\ & \ddots & & & \ddots & & \vdots \\ 0 & & \left[\begin{smallmatrix} I_{r_k} \\ 0 \end{smallmatrix}\right] & 0 & & \left[\begin{smallmatrix} 0 \\ I_{m_k - r_k} \end{smallmatrix}\right] & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & I_{n-r} \end{array}\right] \tag{3.17}$$

rearranges the block-columns of the system matrix. Note that if for some j we have $r_j = m_j$, then the block $I_{m_j - r_j}$ and the corresponding block-rows and block-columns vanish; if $r_j = 0$, then the block I_{r_j} and the corresponding block-rows and block-columns vanish. Let us briefly summarize the whole transformation.

3.5. Summary of the transformation. Using the SVD $B = S\Theta R^T$, defined in (3.2), the original approximation problem (1.1)

$$AX \approx B, \qquad A \in \mathbb{R}^{m\times n}, \quad X \in \mathbb{R}^{n\times d}, \quad B \in \mathbb{R}^{m\times d}$$

is transformed to

$$AY \approx C, \qquad A \in \mathbb{R}^{m\times n}, \quad Y \in \mathbb{R}^{n\times \overline{d}}, \quad C \in \mathbb{R}^{m\times \overline{d}},$$

with *the full column rank right-hand side*. Then using the SVD $A = U\Sigma V^T$ defined in (2.2)–(2.4), the problem is further transformed to

$$\Sigma Z \approx F, \qquad \Sigma \in \mathbb{R}^{m\times n}, \quad Z \in \mathbb{R}^{n\times \overline{d}}, \quad F \in \mathbb{R}^{m\times \overline{d}},$$

with *the diagonal system matrix* Σ. Using singular value decompositions $F_j = S_j \Theta_j W_j^T$ of block-rows of F (see (3.10)) and the orthogonal matrix $S_\oplus$ given by (3.11), the right-hand side matrix gets the structure with the full row rank block-rows Φ_j and the zero block-rows, while the diagonal system matrix Σ stays unchanged. Finally, the permutation matrices Π_L, Π_R given by (3.16) and (3.17) are used to collect the full row rank blocks, and to transform the system matrix to the block diagonal form with two diagonal (in general rectangular) blocks; see (3.15).

We give a quick summary of the mathematical transformations, each preceded by the relevant equation numbers:

(1.1), (3.2):
$$AX \approx B = S\Theta R^T = [C,\ 0]R^T,$$

(2.2), (3.4)–(3.8):
$$\begin{cases} (U^TAV)(V^TXR) \approx U^TBR, \\ \Sigma(V^TXR) \approx [U^TC,\ 0] = [F,\ 0], \end{cases}$$

(3.10)–(3.14):
$$\begin{cases} (S_L^TU^TAVS_R)(S_R^TV^TXR) \approx S_L^TU^TBR, \\ \Sigma(S_R^TV^TXR) \approx [S_L^TF,\ 0], \end{cases}$$

(3.15)–(3.17):
$$\begin{cases} (\Pi_L^TS_L^TU^TAVS_R\Pi_R)(\Pi_R^TS_R^TV^TXR) \approx \Pi_L^TS_L^TU^TBR, \\ (\Pi_L^T\Sigma\Pi_R)(\Pi_R^TS_R^TV^TXR) \approx [\Pi_L^TS_L^TF,\ 0], \\ \mathrm{diag}(A_{11}, A_{22})(\Pi_R^TS_R^TV^TXR) \approx \mathrm{diag}(B_1, 0), \end{cases}$$

so that the full transformations of A, X, and B can be summarized as

$$(P^TAQ)(Q^TXR) \approx P^TBR, \qquad P \equiv US_L\Pi_L, \qquad Q \equiv VS_R\Pi_R. \tag{3.18}$$

Clearly $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$, and

$$P^T[B|A]\begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} = \left[\begin{array}{c|c|c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \begin{matrix} \} \overline{m} \\ \} m - \overline{m} \end{matrix} \tag{3.19}$$
$$\underbrace{}_{\overline{d}}\ \underbrace{}_{d-\overline{d}}\ \underbrace{}_{\overline{n}}\ \underbrace{}_{n-\overline{n}}$$

is of the form (3.1), where from (3.15)

$$[B_1|A_{11}] \equiv \left[\begin{array}{c|ccc} \Phi_1 & \sigma_1 I_{r_1} & & 0 \\ \vdots & & \ddots & \\ \Phi_k & 0 & & \sigma_k I_{r_k} \\ \Phi_{k+1} & 0 & \cdots & 0 \end{array}\right] \in \mathbb{R}^{\overline{m}\times(\overline{n}+\overline{d})}, \tag{3.20}$$

and $\overline{m} \equiv \sum_{j=1}^{k+1} r_j$, $\overline{n} \equiv \sum_{j=1}^{k} r_j$, $\overline{d} \equiv \mathrm{rank}(B)$. The block A_{22} has the form

$$A_{22} \equiv \mathrm{diag}(\sigma_1 I_{m_1-r_1}, \ldots, \sigma_k I_{m_k-r_k}, 0_{m-r-r_{k+1},n-r}). \tag{3.21}$$

Thus the original problem $AX \approx B$ is transformed into the block form

$$\left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right] (Q^T X R) \approx \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \end{array}\right];$$

compare with (1.2). Using a conformal partitioning of the matrix of unknowns

$$Q^T X R = \underbrace{\left[\begin{array}{cc} X_1 & X_1' \\ X_2 & X_2' \end{array}\right]}_{\overline{d} \quad d-\overline{d}} \begin{array}{l} \} \, \overline{n} \\ \} \, n-\overline{n} \end{array}, \tag{3.22}$$

where $Q\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right] = Y$, $Q\left[\begin{smallmatrix} X_1' \\ X_2' \end{smallmatrix}\right] = Y'$, and $[Y, Y'] = XR$ (see (3.4)) the block diagonal structure of the system matrix and the right-hand sides matrix allows us to split the original problem $AX \approx B$ into (generally four) subproblems

$$A_{11}X_1 \approx B_1, \qquad \text{and} \qquad A_{22}X_2 \approx 0, \quad A_{11}X_1' \approx 0, \quad A_{22}X_2' \approx 0.$$

The last three subproblems are homogeneous and we consider, following the arguments in [14], $X_2 \equiv 0$, $X_1' \equiv 0$, $X_2' \equiv 0$. Only the subproblem

$$A_{11}X_1 \approx B_1, \qquad \text{or, equivalently,} \qquad [B_1|A_{11}] \left[\begin{array}{c} -I_{\overline{d}} \\ X_1 \end{array}\right] \approx 0, \tag{3.23}$$

where $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$, $X_1 \in \mathbb{R}^{\overline{n}\times\overline{d}}$, $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ has to be solved. If its solution is X_1, the solution X of the original problem $AX \approx B$ is then

$$X \equiv Q \left[\begin{array}{cc} X_1 & 0 \\ 0 & 0 \end{array}\right] R^T. \tag{3.24}$$

The dimensions of the reduced problem satisfy

$$\max\{\overline{n}, \overline{d}\} \le \overline{m} \equiv \overline{n} + r_{k+1} \le \overline{n} + \overline{d}.$$

Note that $\overline{d}$ can be smaller than, equal to, or even larger than $\overline{n}$. From the construction we immediately have the following properties:

(CP1) The matrix $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ is of *full column rank* equal to $\overline{n} \le \overline{m}$.

(CP2) The matrix $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ is of *full column rank* equal to $\overline{d} \le \overline{m}$.

(CP3) The matrices $\Phi_j \in \mathbb{R}^{r_j\times\overline{d}}$ are of *full row rank* equal to $r_j \le \overline{d}$, $j = 1, \ldots, k+1$.

Remark 3.1. Note that instead of the SVD preprocessing (3.2) of the right-hand side B (section 3.1), one can use an LQ decomposition (producing a matrix with $m \times \overline{d}$ lower triangular block in the column echelon form and an orthogonal matrix), or another decomposition giving a full column rank matrix multiplied by an orthogonal matrix from the right. Similarly, instead of the SVDs (3.10) of F_j (section 3.3) one can use QR decompositions (producing a matrix with an $r_j \times \overline{d}$ upper triangular block in the row echelon form and an orthogonal matrix), or another decompositions giving a full row rank matrix multiplied by an orthogonal matrix from the left. Such modifications lead, in general, to a different subproblem $[\widetilde{B}_1|\widetilde{A}_{11}]$ with the same dimension as (3.20) and satisfying (CP1)–(CP3). Moreover, there exist orthogonal matrices $\widetilde{P}^{-1} = \widetilde{P}^T$, $\widetilde{Q}^{-1} = \widetilde{Q}^T$, $\widetilde{R}^{-1} = \widetilde{R}^T$ such that

$$\widetilde{P}^T [B_1|A_{11}] \begin{bmatrix} \widetilde{R} & 0 \\ 0 & \widetilde{Q} \end{bmatrix} = [\widetilde{B}_1|\widetilde{A}_{11}]. \tag{3.25}$$

Properties (CP1)–(CP3) are *invariant* with respect to any orthogonal transformation of the form (3.25).

4. The core problem. Let $[B_1|A_{11}]$ be the subproblem of the given problem $[B|A]$ obtained by the transformation (3.18)–(3.19). The subproblem $[B_1|A_{11}]$, $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$, $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$, has the properties (CP1)–(CP3). Consider now *arbitrary* orthogonal transformations of the form in (1.2) of the original problem analogous to (3.18)–(3.19) such that

$$\widehat{P}^T [B|A] \begin{bmatrix} \widehat{R} & 0 \\ 0 & \widehat{Q} \end{bmatrix} = \left[\begin{array}{c|c||c|c} \widehat{B}_1 & 0 & \widehat{A}_{11} & 0 \\ \hline 0 & 0 & 0 & \widehat{A}_{22} \end{array}\right], \tag{4.1}$$

where $\widehat{P}^{-1} = \widehat{P}^T$, $\widehat{Q}^{-1} = \widehat{Q}^T$, $\widehat{R}^{-1} = \widehat{R}^T$, and $\widehat{B}_1 \in \mathbb{R}^{\widehat{m}\times\widehat{d}}$, $\widehat{A}_{11} \in \mathbb{R}^{\widehat{m}\times\widehat{n}}$. Substituting for $[B|A]$ from (3.19) gives

$$(P^T\widehat{P})^T \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \begin{bmatrix} R^T\widehat{R} & 0 \\ 0 & Q^T\widehat{Q} \end{bmatrix} = \left[\begin{array}{c|c||c|c} \widehat{B}_1 & 0 & \widehat{A}_{11} & 0 \\ \hline 0 & 0 & 0 & \widehat{A}_{22} \end{array}\right]. \tag{4.2}$$

We will show that the subproblem $[B_1|A_{11}]$ on the left-hand side of (4.2) has minimal dimensions over all possible subproblems $[\widehat{B}_1|\widehat{A}_{11}]$ on the right-hand side; i.e., $\overline{d} \le \widehat{d}$, $\overline{m} \le \widehat{m}$, and $\overline{n} \le \widehat{n}$. In analogy with the single right-hand side case, this justifies calling $[B_1|A_{11}]$ a *core problem* within $[B|A]$.

4.1. Proof of minimality. The proof consists of five successive steps presented, for an easy orientation, in separate subsections. The first gives $\overline{d} \le \widehat{d}$. The second step describes the structure of the orthogonal matrix $R^T\widehat{R}$. The proof of the inequality $\overline{m} \le \widehat{m}$ is then based on the projections of B to the left singular subspaces of A. The fourth step describes the structure of the orthogonal matrix $P^T\widehat{P}$ which is needed for finalizing the proof by showing $\overline{n} \le \widehat{n}$.

4.1.1. Number of columns of the reduced observation matrix. In (4.1) we have $\widehat{B}_1 \in \mathbb{R}^{\widehat{m}\times\widehat{d}}$, and

$$\widehat{P}^T B \widehat{R} = \left[\begin{array}{c|c} \widehat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right] \quad \text{has} \quad \operatorname{rank}(\widehat{B}_1) = \operatorname{rank}(B) = \overline{d}, \quad \text{so that} \quad \overline{d} \le \widehat{d}.$$

4.1.2. Structure of the $R^T\widehat{R}$ matrix. Consider the partitioning

$$R^T\widehat{R} = \begin{bmatrix} R'_{11} & R'_{12} \\ R'_{21} & R'_{22} \end{bmatrix}, \qquad R'_{11} \in \mathbb{R}^{\overline{d}\times\widehat{d}}, \quad R'_{22} \in \mathbb{R}^{(d-\overline{d})\times(d-\widehat{d})}.$$

Then (4.2) gives

$$(P^T\widehat{P})^T \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \end{array}\right] (R^T\widehat{R}) = (P^T\widehat{P})^T \underbrace{\left[\begin{array}{c|c} B_1R'_{11} & B_1R'_{12} \\ \hline 0 & 0 \end{array}\right]}_{\widehat{d} \qquad d-\widehat{d}} = \underbrace{\left[\begin{array}{c|c} \widehat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right]}_{\widehat{d} \quad d-\widehat{d}}.$$

Because B_1 is of full column rank, then $R'_{12} = 0$. Consequently

$$R^T\widehat{R} = \begin{bmatrix} R'_{11} & 0 \\ R'_{21} & R'_{22} \end{bmatrix}, \tag{4.3}$$

and therefore R'_{11} has $\overline{d}$ orthonormal rows.

4.1.3. Number of rows of the reduced data matrix. The number $\overline{m} = \sum_{j=1}^{k+1} r_j$ of rows of $[B_1|A_{11}]$ is the sum of dimensions of intersections of the range of B with the individual left singular subspaces of A, and these dimensions are invariant with respect to the orthogonal transformation of the form (3.18)–(3.19); see also (3.20). The desired inequality $\overline{m} \le \widehat{m}$ will be shown by comparing three different forms of the SVD of the system matrix A. Recall that $A = U\Sigma V^T$ (see (2.2)–(2.4)) is the standard SVD of A. Now consider the SVDs

$$\widehat{A}_{11} = \widehat{U}_1\widehat{\Sigma}_1\widehat{V}_1^T, \qquad \widehat{A}_{22} = \widehat{U}_2\widehat{\Sigma}_2\widehat{V}_2^T,$$

with square orthogonal matrices $\widehat{U}_1$, $\widehat{U}_2$, $\widehat{V}_1$, and $\widehat{V}_2$, and diagonal matrices $\widehat{\Sigma}_1$, $\widehat{\Sigma}_2$. Then, using (4.1),

$$A = \left(\widehat{P}\begin{bmatrix} \widehat{U}_1 & 0 \\ 0 & \widehat{U}_2 \end{bmatrix}\right)\begin{bmatrix} \widehat{\Sigma}_1 & 0 \\ 0 & \widehat{\Sigma}_2 \end{bmatrix}\left(\widehat{Q}\begin{bmatrix} \widehat{V}_1 & 0 \\ 0 & \widehat{V}_2 \end{bmatrix}\right)^T \tag{4.4}$$

represents another SVD of the system matrix A. Furthermore, A_{11}, A_{22} in (3.20), (3.21) are diagonal matrices. The transformation (3.18)–(3.19) gives

$$A = P\left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right] Q^T, \qquad \text{where} \qquad P = US_L\Pi_L, \tag{4.5}$$

which also represents the SVD of A. Because the singular values (and their multiplicities) of A are unique, for some permutation matrices $\widehat{\Pi}_L$, $\widehat{\Pi}_R$ (analogous to Π_L, Π_R in (3.16), (3.17))

$$\Sigma = \widehat{\Pi}_L\begin{bmatrix} \widehat{\Sigma}_1 & 0 \\ 0 & \widehat{\Sigma}_2 \end{bmatrix}\widehat{\Pi}_R^T = \Pi_L\left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right]\Pi_R^T.$$

The matrix U of the left singular vector subspaces of A can be expressed using (4.4), (4.5) and some orthogonal matrices

$$\widehat{S}_L = \operatorname{diag}(\widehat{S}_1, \ldots, \widehat{S}_k, \widehat{S}_{k+1}), \qquad \widehat{S}_j \in \mathbb{R}^{m_j\times m_j}, \quad j = 1, \ldots, k, k+1, \tag{4.6}$$

as

$$U = \widehat{P} \begin{bmatrix} \widehat{U}_1 & 0 \\ 0 & \widehat{U}_2 \end{bmatrix} \widehat{\Pi}_L^T \widehat{S}_L^T = P \Pi_L^T S_L^T . \tag{4.7}$$

Using (4.1) and (3.19), the projections of B in the left singular subspaces of A are

$$U^T B = \widehat{S}_L \widehat{\Pi}_L \left[\begin{array}{c|c} \widehat{U}_1^T \widehat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right] \widehat{R}^T = S_L \Pi_L \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \end{array}\right] R^T,$$

and, with (4.3),

$$\left[\begin{array}{c|c} \widehat{U}_1^T \widehat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right] = \widehat{\Pi}_L^T \widehat{S}_L^T S_L \Pi_L \left[\begin{array}{c|c} B_1 R'_{11} & 0 \\ \hline 0 & 0 \end{array}\right]. \tag{4.8}$$

The matrix $R'_{11} \in \mathbb{R}^{\overline{d} \times \widehat{d}}$, $\widehat{d} \geq \overline{d}$, has linearly independent rows; see (4.3) and section 4.1.1. Now every row of B_1 is nonzero (see (3.20) and (CP3)), so any row of B_1 multiplied by R'_{11} on the right must also be nonzero, and the matrix $B_1 R'_{11}$ has $\overline{m}$ nonzero rows. The matrix Π_L permutes rows (see (3.16)), and matrices S_L, $\widehat{S}_L$ (see (3.11), (4.6)) are orthogonal block diagonal matrices such that (see (3.15), (3.14)),

$$\widehat{S}_L^T S_L \Pi_L \left[\begin{array}{c|c} B_1 R'_{11} & 0 \\ \hline 0 & 0 \end{array}\right] = \begin{bmatrix} \widehat{S}_1^T S_1 \begin{bmatrix} \Phi_1 R'_{11} & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \Big\} m_1 \\ \vdots \\ \begin{bmatrix} \widehat{S}_k^T S_k \begin{bmatrix} \Phi_k R'_{11} & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \Big\} m_k \\ \begin{bmatrix} \widehat{S}_{k+1}^T S_{k+1} \begin{bmatrix} \Phi_{k+1} R'_{11} & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \Big\} m_{k+1} \end{bmatrix}, \tag{4.9}$$

where each block of m_j rows corresponds to one singular value σ_j (of the matrix Σ) with the multiplicity m_j. Because $\Phi_j R'_{11}$ has all r_j rows nonzero (since the rows of Φ_j are linearly independent, the rows of $\Phi_j R'_{11}$ are, in fact, also linearly independent and, therefore, each $\Phi_j R'_{11}$ is of full row rank) and $\widehat{S}_j^T S_j$ are orthogonal matrices, the block matrix on the right of (4.9) has at least $\overline{m}$ nonzero rows; see (3.20). The permutation matrix $\widehat{\Pi}_L^T$ then moves all the nonzero rows (and possibly also some zero rows) of (4.9) to the top block $[\widehat{U}_1^T \widehat{B}_1, 0] \in \mathbb{R}^{\widehat{m} \times d}$ of the leftmost matrix in (4.8). Thus $\widehat{U}_1^T \widehat{B}_1$ has at least $\overline{m}$ nonzero rows (and possibly some zero rows). Because $\widehat{U}_1$ is square, $\widehat{B}_1 \in \mathbb{R}^{\widehat{m} \times \widehat{d}}$ has at least $\overline{m}$ rows, so that $\overline{m} \leq \widehat{m}$.

4.1.4. Structure of the $P^T \widehat{P}$ matrix. Consider the partitioning of the permutation matrices Π_L in (3.16) with $\overline{m} = \sum_{j=1}^{k+1} r_j$ as in (3.20), and $\widehat{\Pi}_L$ in (4.8) where $\widehat{U}_1$ is $\widehat{m} \times \widehat{m}$,

$$\Pi_L = [\Pi_1, \Pi_2], \quad \Pi_1 \in \mathbb{R}^{m \times \overline{m}}, \qquad \widehat{\Pi}_L = [\widehat{\Pi}_1, \widehat{\Pi}_2], \quad \widehat{\Pi}_1 \in \mathbb{R}^{m \times \widehat{m}}.$$

Then (4.8) can be rewritten as

$$\left[\begin{array}{c|c} \widehat{U}_1^T \widehat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right] = \begin{bmatrix} \widehat{\Pi}_1^T \widehat{S}_L^T S_L \Pi_1 & \widehat{\Pi}_1^T \widehat{S}_L^T S_L \Pi_2 \\ \widehat{\Pi}_2^T \widehat{S}_L^T S_L \Pi_1 & \widehat{\Pi}_2^T \widehat{S}_L^T S_L \Pi_2 \end{bmatrix} \left[\begin{array}{c|c} B_1 R'_{11} & 0 \\ \hline 0 & 0 \end{array}\right],$$

yielding the condition

$$(\widehat{\Pi}_2^T \widehat{S}_L^T S_L \Pi_1)(B_1 R'_{11}) = 0. \tag{4.10}$$

Using the structure of the matrices Π_L, S (see (3.11), (3.16)) and analogously for $\widehat{\Pi}_L$ and $\widehat{S}$ (see section 4.1.3),

$$\widehat{\Pi}_2^T \widehat{S}_L^T S_L \Pi_1 = \begin{bmatrix} [0, I_{m_1-\widehat{r}_1}]\widehat{S}_1^T S_1 \left[\begin{smallmatrix} I_{r_1} \\ 0 \end{smallmatrix}\right] & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & [0, I_{m_k-\widehat{r}_k}]\widehat{S}_k^T S_k \left[\begin{smallmatrix} I_{r_k} \\ 0 \end{smallmatrix}\right] & 0 \\ 0 & \cdots & 0 & [0, I_{m_{k+1}-\widehat{r}_{k+1}}]\widehat{S}_{k+1}^T S_{k+1} \left[\begin{smallmatrix} I_{r_{k+1}} \\ 0 \end{smallmatrix}\right] \end{bmatrix},$$

and thus with (3.20) the condition (4.10) is split into $k+1$ block parts

$$\left([0, I_{m_j-\widehat{r}_j}]\widehat{S}_j^T S_j \begin{bmatrix} I_{r_j} \\ 0 \end{bmatrix} \right) (\Phi_j R'_{11}) = 0, \qquad j = 1, \ldots, k, k+1.$$

Because $\Phi_j R'_{11}$ are of full row rank, this gives for all indices j

$$[0, I_{m_j-\widehat{r}_j}]\widehat{S}_j^T S_j \begin{bmatrix} I_{r_j} \\ 0 \end{bmatrix} = 0, \qquad \text{and thus also} \qquad \widehat{\Pi}_2^T \widehat{S}_L^T S_L \Pi_1 = 0.$$

Using (4.7) we finally obtain

$$\begin{aligned} P^T \widehat{P} &= \Pi_L^T S_L^T \widehat{S}_L \widehat{\Pi}_L \begin{bmatrix} \widehat{U}_1 & 0 \\ 0 & \widehat{U}_2 \end{bmatrix}^T \\ &= \begin{bmatrix} \Pi_1^T S_L^T \widehat{S}_L \widehat{\Pi}_1 \widehat{U}_1^T & 0 \\ \Pi_2^T S_L^T \widehat{S}_L \widehat{\Pi}_1 \widehat{U}_1^T & \Pi_2^T S_L^T \widehat{S}_L \widehat{\Pi}_2 \widehat{U}_2^T \end{bmatrix} \equiv \underbrace{\begin{bmatrix} P'_{11} & 0 \\ P'_{21} & P'_{22} \end{bmatrix}}_{\widehat{m}} \begin{matrix} \} \overline{m} \\ \} m - \overline{m} \end{matrix}, \end{aligned} \tag{4.11}$$

and therefore P'_{11} has $\overline{m}$ orthonormal rows.

4.1.5. Number of columns of the reduced system matrix. The relation (4.2) gives, using the structure of $P^T\widehat{P}$ in (4.11),

$$\begin{bmatrix} P'_{11} & 0 \\ P'_{21} & P'_{22} \end{bmatrix}^T \left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right] (Q^T \widehat{Q}) = \left[\begin{array}{c|c} \widehat{A}_{11} & 0 \\ \hline 0 & \widehat{A}_{22} \end{array}\right].$$

Thus

$$[(P'_{11})^T A_{11}, (P'_{21})^T A_{22}](Q^T \widehat{Q}) = [\widehat{A}_{11}, 0].$$

Because $(P'_{11})^T$ has orthonormal columns and $Q^T\widehat{Q}$ is a square orthogonal matrix, $A_{11} \in \mathbb{R}^{\overline{m} \times \overline{n}}$ in (3.20) has rank $\overline{n}$, and $\widehat{A}_{11} \in \mathbb{R}^{\widehat{m} \times \widehat{n}}$ (see (4.1)) where

$$\overline{n} = \operatorname{rank}(A_{11}) = \operatorname{rank}((P'_{11})^T A_{11}) \le \operatorname{rank}([\widehat{A}_{11}, 0]) = \operatorname{rank}(\widehat{A}_{11}) \le \widehat{n}$$

completing our proof. □

5. Summary and concluding remarks. We formulate the minimality property as a theorem.

THEOREM 5.1 (minimality). *Consider the problem* $AX \approx B$ (1.1) *and its subproblem* $A_{11}X_1 \approx B_1$ (3.23)*, where* $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ *and* $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ *are obtained by the transformation* (3.19) *described in section* 3*. The subproblem* $A_{11}X_1 \approx B_1$ *has minimal dimensions over all subproblems* $\widehat{A}_{11}\widehat{X}_1 \approx \widehat{B}_1$*,* $\widehat{A}_{11} \in \mathbb{R}^{\widehat{m}\times\widehat{n}}$*,* $\widehat{B}_1 \in \mathbb{R}^{\widehat{m}\times\widehat{d}}$ *obtained by an orthogonal transformation of the form* (4.1)*, i.e.,* $\widehat{d} \geq \overline{d}$*,* $\widehat{m} \geq \overline{m}$*, and* $\widehat{n} \geq \overline{n}$*.*

This motivates the following definition of the core problem within (1.1).

DEFINITION 5.2 (core problem). *The subproblem* $A_{11}X_1 \approx B_1$ *is a core problem within the approximation problem* $AX \approx B$ *if* $[B_1|A_{11}]$ *is minimally dimensioned and* A_{22} *maximally dimensioned subject to* (3.1)*, i.e., subject to the orthogonal transformations of the form*

$$P^T[B|A]\left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] = [P^TBR|P^TAQ] \equiv \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right].$$

The core problem obtained by the data reduction based on the SVD described in section 3 *is called the core problem in the SVD form.*

This extends the core problem definition within the single right-hand side problems formulated by Paige and Strakoš in [14].

The data reduction presented here is based on the SVD of the system matrix A. The original paper [14] presents two ways of determining the core problem. The second is based on Golub–Kahan iterative bidiagonalization. This can also be generalized to problems with multiple right-hand sides. It leads to a band algorithm which was for this purpose proposed by Björck; see [1], [2], [3], and also the unpublished manuscript [4].

The core problem concept is a useful tool in understanding the TLS problems, which was the original motivation in [14]. The data reduction and core problem based on the generalization of the Golub–Kahan iterative bidiagonalization is under investigation. The results will be presented in the near future.

Acknowledgments. The authors thank Diana M. Sima for valuable discussions about the TLS problems. The numerous comments and suggestions of two anonymous referees led to significant improvements of the text.

REFERENCES

[1] Å. BJÖRCK, *Bidiagonal Decomposition and Least Squares*, Presentation, Canberra, Australia, 2005.
[2] Å. BJÖRCK, *A Band-Lanczos Generalization of Bidiagonal Decomposition*, Presentation, Conference in Honor of G. Dahlquist, Stockholm, Sweden, 2006.
[3] Å. BJÖRCK, *A Band-Lanczos algorithm for least squares and total least squares problems*, in Book of Abstracts of the 4th Total Least Squares and Errors-in-Variables Modeling Workshop, Leuven, Katholieke Universiteit Leuven, Leuven, Belgium, 2006, pp. 22–23.
[4] Å. BJÖRCK, *Block Bidiagonal Decomposition and Least Squares Problems with Multiple Right-Hand Sides*, unpublished manuscript.
[5] G. H. GOLUB, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[6] G. H. GOLUB, A. HOFFMAN, AND G. W. STEWART, *A generalization of the Eckart-Young-Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.
[7] G. H. GOLUB AND C. REINSCH, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.
[8] G. H. GOLUB AND C. F. VAN LOAN, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, London, 2012.
[10] I. Hnětynková, M. Plešinger, D. M. Sima, Z. Strakoš, and S. Van Huffel, *The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 748–770.
[11] I. Hnětynková and Z. Strakoš, *Lanczos tridiagonalization and core problems*, Linear Algebra Appl., 421 (2007), pp. 243–251.
[12] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., 91 (2002), pp. 117–146.
[13] C. C. Paige and Z. Strakoš, *Unifying least squares, total least squares and data least squares*, in Total Least Squares and Errors-in-Variables Modeling, S. Van Huffel and P. Lemmerling, eds., Kluwer Academic Publishers, Dordrecht, (2002), pp. 25–34.
[14] C. C. Paige and Z. Strakoš, *Core problem in linear algebraic systems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 861–875.
[15] M. Plešinger, *The Total Least Squares Problem and Reduction of Data in $AX \approx B$*, Ph.D. thesis, Technical University of Liberec, Liberec, Czech Rebuplic, 2008.
[16] D. M. Sima, *Regularization Techniques in Model Fitting and Parameter Estimation*, Ph.D. thesis, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.
[17] D. M. Sima and S. Van Huffel, *Core Problems in $AX \approx B$*, Technical Report, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.
[18] A. van der Sluis, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.
[19] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.
[20] S. Van Huffel, ed., *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, Proceedings of the Second International Workshop on TLS and EIV, Philadelphia, SIAM, Philadelphia, 1997.
[21] S. Van Huffel and P. Lemmerling, eds., *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications,* Kluwer Academic Publishers, Dordrecht, 2002.
[22] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 746–763.
[23] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math., 62 (1992), pp. 123–148.

SIAM J. MATRIX ANAL. APPL.
Vol. 36, No. 2, pp. 417–434

BAND GENERALIZATION OF THE GOLUB–KAHAN BIDIAGONALIZATION, GENERALIZED JACOBI MATRICES, AND THE CORE PROBLEM*

IVETA HNĚTYNKOVÁ†, MARTIN PLEŠINGER‡, AND ZDENĚK STRAKOŠ§

Abstract. The concept of the core problem in total least squares (TLS) problems with single right-hand side introduced in [C. C. Paige and Z. Strakoš, *SIAM J. Matrix Anal. Appl.*, 27 (2005), pp. 861–875] separates necessary and sufficient information for solving the problem from redundancies and irrelevant information contained in the data. It is based on orthogonal transformations such that the resulting problem decomposes into two independent parts. One of the parts has nonzero right-hand side and minimal dimensions and it always has the unique TLS solution. The other part has trivial (zero) right-hand side and maximal dimensions. Assuming exact arithmetic, the core problem can be obtained by the Golub–Kahan bidiagonalization. Extension of the core concept to the multiple right-hand sides case $AX \approx B$ in [I. Hnětynková, M. Plešinger, and Z. Strakoš, *SIAM J. Matrix Anal. Appl.*, 34 (2013), pp. 917–931], which is highly nontrivial, is based on application of the singular value decomposition. In this paper we prove that the band generalization of the Golub–Kahan bidiagonalization proposed in this context by Björck also yields the core problem. We introduce generalized Jacobi matrices and investigate their properties. They prove useful in further analysis of the core problem concept. This paper assumes exact arithmetic.

Key words. total least squares problem, multiple right-hand sides, core problem, Golub–Kahan bidiagonalization, generalized Jacobi matrices

AMS subject classifications. 15A06, 15A18, 15A21, 15A24, 65F20, 65F25

DOI. 10.1137/140968914

1. Introduction. This paper further elaborates on extending the core problem concept to total least squares (TLS) problems with multiple right-hand sides; see [13]. We will use the same notation as in [13] and very briefly recall some basic facts. Consider a linear approximation problem

$$AX \approx B \qquad \text{or, equivalently,} \qquad [B|A] \begin{bmatrix} -I_d \\ X \end{bmatrix} \approx 0, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{m\times d}$, and $A^TB \neq 0$, without any further assumption on the positive integers m, n, d. The matrices A, B, $[B|A]$, and X are called the *system matrix*, the *right-hand side (or the observation) matrix*, the *extended (or data) matrix*, and the *matrix of unknowns*, respectively. We will focus on incompatible problems, i.e., $\mathcal{R}(B) \not\subset \mathcal{R}(A)$, although, the compatible case is not

*Received by the editors May 12, 2014; accepted for publication (in revised form) by J. L. Barlow January 29, 2015; published electronically April 14, 2015. The research of the authors was supported by GAČR grant P201/13–06684S.
http://www.siam.org/journals/simax/36-2/96891.html
†Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, and Institute of Computer Science, AS CR, Prague, Czech Republic (hnetynkova@cs.cas.cz). This author is a member of the University center for mathematical modeling, applied analysis, and computational mathematics (Math MAC).
‡Department of Mathematics, Technical University of Liberec, Liberec, Czech Republic (martin.plesinger@tul.cz). The research of this author was partly supported by the ESF grant CZ.1.07/2.3.00/30.0065.
§Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic (strakos@karlin.mff.cuni.cz). The research of this author was partly supported by the ERC-CZ project LL1202.

strictly excluded. Consider the orthogonal transformations

$$\widehat{A}\widehat{X} \equiv (P^T A Q)(Q^T X R) \approx (P^T B R) \equiv \widehat{B}, \tag{1.2}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$, or, equivalently,

$$[\widehat{B}|\widehat{A}] \begin{bmatrix} -I_d \\ \widehat{X} \end{bmatrix} \equiv \left(P^T [B|A] \begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} \right) \left(\begin{bmatrix} R^T & 0 \\ 0 & Q^T \end{bmatrix} \begin{bmatrix} -I_d \\ X \end{bmatrix} R \right) \approx 0. \tag{1.3}$$

We call problems (1.1) and (1.2)–(1.3) *orthogonally invariant* and require that X solves (1.1) if and only if $\widehat{X} = Q^T X R$ solves (1.2)–(1.3). Within this paper we investigate the most common case of the TLS problem

$$\min_{X,E,G} \|[G|E]\|_F \qquad \text{subject to} \qquad (A+E)X = B+G, \tag{1.4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The TLS problem has been studied for a long time, including the works of Golub and Van Loan [10], Van Huffel and Vandewalle [21], Wei [22], [23], and many others. The paper [10] analyzes the single right-hand side case ($d = 1$) and uses a strict decrease of the smallest singular value of the extended matrix $[b|A]$ in comparison to the smallest singular value of A as the *sufficient condition* for existence of the TLS solution. The subsequent work [21] extends the concept of the TLS solution by projecting out the inappropriate right singular vectors of $[b|A]$ (called *unwanted directions*) associated with the smallest singular values. This allows construction of the *classical TLS algorithm* (see [21, section 3.6.1]) which always gives as an output *a computationally defined "solution"*, with a relatively straightforward computational extension to problems with multiple right-hand sides ($d > 1$). The analytic part of this computationally defined algorithmic output, i.e., explanation of its precise meaning in terms of the formulation of the TLS problem, remains very involved and is not fully explained in [21]. The work of Wei [22], [23] complements the previous results by focusing mainly on the rank deficient problems. All the theory in [10], [21], [22], [23] follows essentially the path outlined in [10], in particular, it is mostly based on the sufficient (not *necessary and sufficient*) condition for existence of the TLS solution. The solvability of multiple right-hand side problems has been analyzed in full generality only recently in [11], revealing the intriguing difficulties in the classical approach. In particular, it shows that the computationally defined "solution" of the TLS problem may be in some cases different from the true TLS solution, i.e., the *classical TLS algorithm may not reach the existing TLS solution.*

The *core problem* concept, introduced in [16] for $d = 1$, is based on a different reasoning. It asks what does it mean in terms of the original data A and b that the solution in the TLS sense does not exist. Van Huffel and Vandewalle indicate that this happens in the presence of the so-called *nonpredictive multicollinearities* (see [21, p. 71]), when the linear dependency between the columns of A is stronger than the linear dependency between the range of A and the right-hand side b. Projecting out some unwanted directions in construction of the computationally defined "solution" does not remove all redundancies and irrelevant information from $[b|A]$. For an orthogonally invariant linear approximation problem this is done by the core problem reduction allowing us to simply formulate the *necessary and sufficient condition* for the existence of the TLS solution for $d = 1$; see [16]. The core problem reduction can be done by the singular value decomposition (SVD) of A or by the Golub–Kahan iterative bidiagonalization [9]; see [16], [14]. Moreover, the core problem approach

reveals that *any partial result of the Golub–Kahan iterative bidiagonalization* contains a part of the necessary information for solving the original problem and it does not contain any redundancies or irrelevant information.

In the multiple right-hand side case the situation is more complicated. The first steps in generalizing the core problem concept for $d > 1$ were done by Björck in the series of talks [2], [3], [4], and in the unpublished manuscript [5], by Sima in [19], by Sima and Van Huffel in [20], and by Plešinger in [18]. Following these works and the paper [11] fully classifying situations which can occur when $d > 1$, the paper [13] provides a rigorous extension of the core problem concept to TLS with multiple right-hand sides (1.4). The orthogonal transformation (data reduction) used there is based on the SVD of the matrix A.

The results for single right-hand side problems give a motivation for using the *band generalization of the Golub–Kahan bidiagonalization* (or simply the *band algorithm*) also in the multiple right-hand side case, as proposed in [2], [3], [4], [5]. Here the deflation due to possible zero entries reducing the band shape of the transformed matrix plays a crucial role. We investigate the band algorithm and prove rigorously that it indeed provides a core problem in the sense of [13]. Furthermore, we derive additional properties of the core problem with multiple right-hand sides that might be useful in analysis of its solvability.

The paper is organized in the following way. Section 2 recalls the background results. Section 3 describes the band generalization of the Golub–Kahan bidiagonalization. Section 4 introduces generalized Jacobi matrices and analyzes properties of the band subproblem. Section 5 concludes the paper.

Throughout the text $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote the range and null space of a matrix M, respectively; I_ℓ (or just I) denotes an $\ell \times \ell$ identity matrix; e_k denotes the kth column of I; $0_{\ell,\xi}$ (or just 0) denotes an $\ell \times \xi$ zero matrix; and $\|v\|$ denotes the Euclidean norm of a vector v. The following convention concerning the entries of matrices will simplify the exposition:

- club ($\clubsuit$) stands for a nonzero entry, $\clubsuit \neq 0$;
- heart ($\heartsuit$) stands for a general entry which can also be zero;
- empty spaces in matrices always represent zero entries.

Throughout the paper we assume *exact arithmetic*.

2. The core problem and other background results. In order to make the text as self-consistent as possible, we briefly recall the known results used below.

2.1. Core problem. The *core problem* within the problem (1.1) is defined as follows (see [13, Definition 5.2]).

DEFINITION 2.1 (core problem). *The subproblem $A_{11}X_1 \approx B_1$ is a core problem within the approximation problem $AX \approx B$ if $[B_1|A_{11}]$ is minimally dimensioned and A_{22} maximally dimensioned subject to the orthogonal transformations of the form*

$$P^T[B|A]\begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} = P^T[BR|AQ] \equiv \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right], \tag{2.1}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$.

Let $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ have k distinct singular values σ_j with multiplicities r_j and the orthonormal bases of the corresponding left singular vector subspaces $U_j \in \mathbb{R}^{\overline{m}\times r_j}$, $j = 1, \ldots, k$. Let $r_{k+1} \equiv \dim(\mathcal{N}(A_{11}^T))$, with $U_{k+1} \in \mathbb{R}^{\overline{m}\times r_{k+1}}$ having the orthonormal basis vectors of $\mathcal{N}(A_{11}^T)$ as its columns. Then $U \equiv [U_1, \ldots, U_k, U_{k+1}] \in \mathbb{R}^{\overline{m}\times\overline{m}}$ and $U^T = U^{-1}$. The core problem $A_{11}X_1 \approx B_1$ has the following properties (see [13, p. 925]):

(CP1) The matrix $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ is of *full column rank* equal to $\overline{n} \le \overline{m}$.

(CP2) The matrix $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ is of *full column rank* equal to $\overline{d} \le \overline{m}$.

(CP3) The matrices $\Phi_j \equiv U_j^T B_1 \in \mathbb{R}^{r_j\times\overline{d}}$ are of *full row rank* equal to $r_j \le \overline{d}$, for $j = 1, \ldots, k+1$.

These properties guarantee minimality of the core problem; see [13, section 4]. Dimensions of any subproblem $A_{11}X_1 \approx B_1$ having the properties (CP1)–(CP3) cannot be reduced by any orthogonal transformation of the form (2.1). Moreover

$$U^T[B_1|A_{11}] = \underbrace{\left[\begin{array}{c|c} \Phi_1 & U_1^T A_{11} \\ \vdots & \vdots \\ \Phi_k & U_k^T A_{11} \\ \Phi_{k+1} & 0 \end{array}\right]}_{\overline{d}\qquad\overline{n}} \left.\begin{array}{c} \}r_1 \\ \vdots \\ \}r_k \\ \}r_{k+1} \end{array}\right\} \overline{m},$$

where $[U_1, \ldots, U_k]^T A_{11}$ is a square nonsingular matrix of the size $\overline{n} \times \overline{n}$, $\overline{n} = r_1 + \cdots + r_k$, and Φ_{k+1} is of full row rank $r_{k+1} = \overline{m} - \overline{n}$. Thus (CP1)–(CP3) imply that the extended matrix $[B_1|A_{11}]$ is of *full row rank* equal to $\overline{m}$, $\max\{\overline{n}, \overline{d}\} \le \overline{m} \le \overline{n} + \overline{d}$.

2.2. Golub–Kahan bidiagonalization. Consider first $d = 1$, i.e., the single right-hand side problem $Ax \approx b$. Here the core problem can be obtained by the Golub–Kahan iterative bidiagonalization.[1] Using the initial vectors $q_0 = 0$ and $p_1 = b/\gamma_1$, where $\gamma_1 = \|b\|$, it computes for $j = 1, 2, \ldots,$

$$q_j\alpha_j = A^T p_j - q_{j-1}\gamma_j, \tag{2.2}$$

$$p_{j+1}\gamma_{j+1} = Aq_j - p_j\alpha_j, \tag{2.3}$$

such that $\|q_j\| = \|p_{j+1}\| = 1$, and $\alpha_j > 0$, $\gamma_{j+1} > 0$. The matrices

$$P_j \equiv [p_1, \ldots, p_j] \in \mathbb{R}^{m\times j}, \quad Q_j \equiv [q_1, \ldots, q_j] \in \mathbb{R}^{n\times j},$$

have orthonormal columns, $P_j^T P_j = Q_j^T Q_j = I_j$; see [9]. The iterative process (2.2)–(2.3) terminates when the right-hand side of one of the equations becomes zero, i.e., either $q_j\alpha_j = 0$ (in the incompatible case) or $p_{j+1}\gamma_{j+1} = 0$ (in the compatible case) for some j. Consider that $Ax \approx b$ is incompatible, $b \notin \mathcal{R}(A)$, and let $q_{\overline{n}+1}\alpha_{\overline{n}+1} = 0$. Then, denoting $P_1^{\rm cp} \equiv P_{\overline{n}+1}$ and $Q_1^{\rm cp} \equiv Q_{\overline{n}}$,

$$(P_1^{\rm cp})^T[b|AQ_1^{\rm cp}] = \left[\begin{array}{c|ccc} \gamma_1 & \alpha_1 & & \\ & \gamma_2 & \ddots & \\ & & \ddots & \alpha_{\overline{n}} \\ & & & \gamma_{\overline{n}+1} \end{array}\right] = [b_1|A_{11}] \in \mathbb{R}^{(\overline{n}+1)\times(\overline{n}+1)} \tag{2.4}$$

represents the core problem within $[b|A]$, and

$$P^T[b|A]\begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} = \left[\begin{array}{c|c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right], \quad P \equiv [P_1^{\rm cp}, P_2^{\rm cp}], \quad Q \equiv [Q_1^{\rm cp}, Q_2^{\rm cp}],$$

where $P_2^{\rm cp}$, $Q_2^{\rm cp}$ are chosen such that $P^{-1} = P^T$, $Q^{-1} = Q^T$; see [16]. A generalization of the Golub–Kahan bidiagonalization for the problems with multiple right-hand sides is given in section 3 below.

[1] Due to the close connection to the Lanczos algorithm one can also find it under the name the Golub–Kahan–Lanczos bidiagonalization; see, e.g., [1], [3], [4].

2.3. Right-hand side preprocessing. In order to get an equivalent problem with the full column rank right-hand sides matrix, we preprocess B in an analogous way to [13, section 3.1]. Let $\overline{d} \equiv \operatorname{rank}(B) \leq \min\{m, d\}$, $B \in \mathbb{R}^{m\times d}$. Consider any decomposition of B in the form

$$B = [C, 0]R^T, \qquad C \in \mathbb{R}^{m\times\overline{d}}, \qquad R \in \mathbb{R}^{d\times d}, \tag{2.5}$$

where C is of *full column rank*, and R is square orthogonal, i.e., $R^{-1} = R^T$. Multiplication of (1.1) from the right by R gives

$$A(XR) \approx BR, \qquad \text{where} \qquad XR \equiv [Y, Y'] \in \mathbb{R}^{n\times d}, \qquad Y \in \mathbb{R}^{n\times\overline{d}} \tag{2.6}$$

(if $d = \overline{d}$, then it can be considered $R = I_d$, $B = C$, $X = Y$). The original problem (1.1) is in this way split into two subproblems,

$$AY \approx C \qquad \text{and} \qquad AY' \approx 0, \tag{2.7}$$

where the second problem is homogeneous. Following the arguments in [16], we consider the meaningful solution $Y' \equiv 0$. In this way, the approximation problem (1.1) reduces to $AY \approx C$ in (2.7). The full column rank matrix $C \in \mathbb{R}^{m\times\overline{d}}$ is called the *preprocessed right-hand side*.

Remark 2.2. A decomposition (2.5) can be obtained using the LQ decomposition of B (see [13, remark 3.1]) in the form

$$\Pi B = [\Lambda, 0]R^T, \qquad \Pi \in \mathbb{R}^{m\times m}, \quad \Lambda \in \mathbb{R}^{m\times\overline{d}}, \quad R \in \mathbb{R}^{d\times d},$$

where Π is a permutation matrix (representing possible row pivoting of B), and Λ is in a lower triangular column echelon form with nonzero columns. Then $C \equiv \Pi^T\Lambda$ is called the *LQ-preprocessed right-hand side*. Alternatively, one can use the SVD of B (see [13, section 3.1]) in the form

$$B = S[\Theta, 0]R^T, \qquad S \in \mathbb{R}^{m\times\overline{d}}, \quad \Theta \in \mathbb{R}^{\overline{d}\times\overline{d}}, \quad R \in \mathbb{R}^{d\times d}, \tag{2.8}$$

where S has mutually orthonormal columns, and the square nonsingular Θ contains the singular values of B on the diagonal. Then $C \equiv S\Theta$ has (nonzero) mutually orthogonal columns and it is called the *SVD-preprocessed right-hand side*.

3. Band generalization of the Golub–Kahan bidiagonalization. Now we describe in details the band algorithm. Consider the problem $AY \approx C$, where $C \in \mathbb{R}^{m\times\overline{d}}$ is of full column rank obtained above. As an extension of (2.4), we want to reduce $[C|A]$ to the upper triangular band matrix with (at most) $\overline{d}+1$ nonzero diagonals (all entries above the $\overline{d}$th superdiagonal are zero). We start with the QR decomposition of the right-hand side C. The basic band structure is then obtained using Householder reflections. The whole transformation can be reformulated as an iterative procedure that, employing deflations, reveals a subproblem representing the core problem analogously to (2.2)–(2.4).

3.1. Basic structure of the band algorithm. First, the right-hand side C is transformed to the upper triangular form. Consider the QR decomposition

$$C = P_{(0)}F, \qquad F = \begin{bmatrix} F_1 \\ 0 \end{bmatrix}, \qquad F_1 = \begin{bmatrix} \gamma_{1,1} & \beta_{1,2} & \cdots & \beta_{1,\overline{d}} \\ & \gamma_{2,2} & \ddots & \vdots \\ & & \ddots & \beta_{\overline{d}-1,\overline{d}} \\ & & & \gamma_{\overline{d},\overline{d}} \end{bmatrix}, \tag{3.1}$$

where $P_{(0)} \in \mathbb{R}^{m\times m}$, $P_{(0)}^{-1} = P_{(0)}^T$, and F_1 is the upper triangular square matrix with a positive diagonal, $\gamma_{j,j} > 0$, $j = 1, \ldots, \overline{d}$. If C is the *SVD-preprocessed right-hand side*, then $F_1 = \Theta$ is diagonal containing the singular values of the original right-hand side B, and the first $\overline{d}$ columns of $P_{(0)}$ are the columns of S; see (2.8). It should be noted that the matrix $P_{(0)}$ as well as the matrices $P_{(k)}$ below, which are all $\in \mathbb{R}^{m\times m}$, are distinct from the matrices $P_j \in \mathbb{R}^{m\times j}$ used above in the description of the Golub–Kahan iterative bidiagonalization. Denote $L_{(0)} \equiv P_{(0)}^T A$, then

$$P_{(0)}^T [C|A] = [F|L_{(0)}]. \tag{3.2}$$

It remains to transform $L_{(0)}$ to a lower triangular band matrix with (at most) $\overline{d}+1$ nonzero diagonals (all entries below the $\overline{d}$th subdiagonal are zero). This can be done, e.g., by multiplications of $L_{(0)}$ with suitable Householder matrices $H_{Q,j}$, $H_{P,j}$, $j = 1, 2, \ldots$, from the right and left, respectively. Let, for $k = 1, 2, \ldots$,

$$P_{(k)} = P_{(0)} H_{P,1} H_{P,2} \ldots H_{P,k} \in \mathbb{R}^{m\times m}, \quad Q_{(k)} = H_{Q,1} H_{Q,2} \ldots H_{Q,k} \in \mathbb{R}^{n\times n} \tag{3.3}$$

be orthogonal matrices yielding a transformation

$$P_{(k)}^T [C|AQ_{(k)}] = [F|P_{(k)}^T AQ_{(k)}] \tag{3.4}$$

$$= \left[\begin{array}{cccc|cccccc} \gamma_{1,1} & \beta_{1,2} & \cdots & \beta_{1,\overline{d}} & \alpha_{1,\overline{d}+1} & & & & & \\ & \gamma_{2,2} & \ddots & \vdots & \beta_{2,\overline{d}+1} & \ddots & & & & \\ & & \ddots & \beta_{\overline{d}-1,\overline{d}} & \vdots & \ddots & \alpha_{k,\overline{d}+k} & & & \\ & & & \gamma_{\overline{d},\overline{d}} & \beta_{\overline{d},\overline{d}+1} & & \beta_{k+1,\overline{d}+k} & \heartsuit & \cdots & \heartsuit \\ & & & & \gamma_{\overline{d}+1,\overline{d}+1} & \ddots & \vdots & \vdots & & \vdots \\ & & & & & \ddots & \beta_{k+\overline{d}-1,\overline{d}+k} & \heartsuit & \cdots & \heartsuit \\ & & & & & & \gamma_{\overline{d}+k,\overline{d}+k} & \heartsuit & \cdots & \heartsuit \\ & & & & & & & \heartsuit & \cdots & \heartsuit \\ & & & & & & & \vdots & & \vdots \\ & & & & & & & \heartsuit & \cdots & \heartsuit \end{array}\right], \tag{3.5}$$

$$\text{with} \quad \alpha_{j,\overline{d}+j} > 0, \quad \gamma_{\overline{d}+j,\overline{d}+j} > 0, \quad \text{for} \quad j = 1, \ldots, k. \tag{3.6}$$

Denote

$$L_{(j)} \equiv P_{(j)}^T AQ_{(j)} = H_{P,j}^T L_{(j-1)} H_{Q,j}\,, \tag{3.7}$$

$$\text{and} \qquad L_{(j-)} \equiv P_{(j-1)}^T AQ_{(j)} = L_{(j-1)} H_{Q,j}\,, \qquad j = 1, \ldots, k. \tag{3.8}$$

The entry $\alpha_{j,\overline{d}+j}$ represents the norm of the trailing subrow (of length $n-j+1$) of the jth row of $L_{(j-1)}$, and, analogously, the entry $\gamma_{\overline{d}+j,\overline{d}+j}$ represents the norm of the trailing subcolumn (of length $m-j-\overline{d}+1$) of the jth column of $L_{(j-)}$. If the first row of $L_{(0)}$ is zero (i.e., $\alpha_{1,\overline{d}+1} = 0$), or if the trailing subcolumn of $L_{(1-)}$ of length $m-\overline{d}$ is zero (i.e., $\gamma_{\overline{d}+1,\overline{d}+1} = 0$), or if the problem does not have enough rows (i.e., $\gamma_{\overline{d}+1,\overline{d}+1}$ does not exist), then the transformation (3.4) to the form (3.5) with the condition (3.6) does not exist. In such case we formally put $k = 0$ and $Q_{(0)} \equiv I_n$. This particular case is discussed later in section 3.2. In the rest of this paragraph for

simplicity we consider $\alpha_{j,\overline{d}+j} > 0$, $\gamma_{\overline{d}+j,\overline{d}+j} > 0$, $j = 1, \ldots, k$. Denote by $p_1, \ldots, p_{\overline{d}+k}$ the first $\overline{d}+k$ columns of $P_{(k)}$ and by $q_1, \ldots, q_k$ the first k columns of $Q_{(k)}$. Using (3.7) rewritten as

$$AQ_{(k)} = P_{(k)}L_{(k)} \qquad \text{and} \qquad A^T P_{(k)} = Q_{(k)}L_{(k)}^T, \tag{3.9}$$

we can write for Aq_j and $A^T p_j$

$$Aq_j = [p_j, p_{j+1}, \ldots, p_{j+\overline{d}-1}, p_{j+\overline{d}}][\alpha_{j,\overline{d}+j}, \beta_{j+1,\overline{d}+j}, \ldots, \beta_{j+\overline{d}-1,\overline{d}+j}, \gamma_{\overline{d}+j,\overline{d}+j}]^T, \tag{3.10}$$

$$A^T p_j = [q_{j-\overline{d}}, q_{j-\overline{d}+1}, \ldots, q_{j-1}, q_j][\gamma_{j,j}, \beta_{j,j+1}, \ldots, \beta_{j,j+\overline{d}-1}, \alpha_{j,\overline{d}+j}]^T. \tag{3.11}$$

Using the initial vectors $p_1, \ldots, p_{\overline{d}}$ given by (3.1) and $q_{1-\overline{d}} = \cdots = q_0 \equiv 0$, the columns of the $(\overline{d}+k) \times k$ leading principal block of $L_{(k)}$, and the columns $q_1, \ldots, q_k$, and $p_{\overline{d}+1}, \ldots, p_{\overline{d}+k}$ are iteratively generated by

$$q_j \alpha_{j,\overline{d}+j} \equiv A^T p_j - q_{j-\overline{d}}\gamma_{j,j} - \left(\sum_{i=1}^{\overline{d}-1} q_{j-\overline{d}+i}\beta_{j,j+i} \right), \tag{3.12}$$

$$\beta_{j+i,\overline{d}+j} \equiv p_{j+i}^T A q_j \qquad \text{for} \quad i = 1, \ldots, \overline{d}-1, \tag{3.13}$$

$$p_{\overline{d}+j}\gamma_{\overline{d}+j,\overline{d}+j} \equiv Aq_j - p_j\alpha_{j,\overline{d}+j} - \left(\sum_{i=1}^{\overline{d}-1} p_{j+i}\beta_{j+i,\overline{d}+j} \right), \tag{3.14}$$

where $\|q_j\| = \|p_{j+\overline{d}}\| = 1$, $\alpha_{j,\overline{d}+j} > 0$, $\gamma_{\overline{d}+j,\overline{d}+j} > 0$, for $j = 1, 2, \ldots, k$. The β entries represent orthogonalization coefficients.

3.2. Deflation in the band algorithm. Now we focus on the case when the right-hand side of (3.12) or (3.14) becomes zero (including the case $k = 0$). Let ℓ be the *first* index for which *either* $q_\ell \alpha_{\ell,\overline{d}+\ell} = 0$ (yielding formally $\alpha_{\ell,\overline{d}+\ell} = 0$) *or* $p_{\overline{d}+\ell}\gamma_{\overline{d}+\ell,\overline{d}+\ell} = 0$ (yielding formally $\gamma_{\overline{d}+\ell,\overline{d}+\ell} = 0$), $1 \le \ell \le \min\{n+1, m-\overline{d}+1\}$. The cases $\ell = n+1$ and $\ell = m-\overline{d}+1$ represent reaching the number of columns and rows of the system matrix, respectively.

3.2.1. Upper deflation. Let $q_\ell \alpha_{\ell,\overline{d}+\ell} = 0$ for some $\ell < n+1$. Recall that $\alpha_{\ell,\overline{d}+\ell}$ is the norm of a trailing subrow of the ℓth row of $L_{(\ell-1)} = P_{(\ell-1)}^T A Q_{(\ell-1)}$ to the right of $\beta_{\ell,\overline{d}+\ell-1}$; therefore,

$$P_{(\ell-1)}^T [C|AQ_{(\ell-1)}] = [F|L_{(\ell-1)}] = \begin{bmatrix} \ddots & & & & \\ \ddots & \alpha_{\ell-1,\overline{d}+\ell-1} & & & \\ \ddots & \beta_{\ell,\overline{d}+\ell-1} & 0 & \cdots & 0 \\ & \beta_{\ell+1,\overline{d}+\ell-1} & \heartsuit & \cdots & \heartsuit \\ & \vdots & \vdots & & \vdots \end{bmatrix}. \tag{3.15}$$

In this case the Householder matrix $H_{Q,\ell}$ is constructed to transform the first row below the ℓth row having a nonzero trailing subrow (say, the ξth row) while producing $\alpha_{\xi,\overline{d}+\ell} > 0$.

This *upper deflation* can be easily described using (3.12)–(3.14). Consider that the $(\ell+1)$th row of $[F|L_{(\ell-1)}]$ has the nonzero trailing subrow. The formula for computing $\alpha_{\ell+1,\overline{d}+\ell} > 0$ and q_ℓ is then given by equating the $(\ell+1)$th (instead of the ℓth) columns of $A^T P_{(\ell)} = Q_{(\ell)} L_{(\ell)}^T$; see also (3.9) and (3.11). Formulas (3.12)–(3.14) are then, for $j = \ell, \ell+1, \ldots$, modified to

$$q_j \alpha_{j+1,\overline{d}+j} \equiv A^T p_{j+1} - q_{j-\overline{d}+1}\gamma_{j+1,j+1} - \left(\sum_{i=2}^{\overline{d}-1} q_{j-\overline{d}+i}\beta_{j+1,j+i} \right), \tag{3.16}$$

$$\beta_{j+i,\overline{d}+j} \equiv p_{j+i}^T A q_j \qquad \text{for} \quad i = 2, \ldots, \overline{d}-1, \tag{3.17}$$

$$p_{\overline{d}+j}\gamma_{\overline{d}+j,\overline{d}+j} \equiv A q_j - p_{j+1}\alpha_{j+1,\overline{d}+j} - \left(\sum_{i=2}^{\overline{d}-1} p_{j+i}\beta_{j+i,\overline{d}+j} \right). \tag{3.18}$$

The *number of summands* and *computed coefficients* β is *reduced by one* for all $j \geq \ell$. Each upper deflation changes the pattern of nonzero entries in the band matrix by reducing the *effective bandwidth* from the top by one.

3.2.2. Lower deflation. Let $p_{\overline{d}+\ell}\gamma_{\overline{d}+\ell,\overline{d}+\ell} = 0$ for $\ell < m - \overline{d} + 1$. Recall that $\gamma_{\overline{d}+\ell,\overline{d}+\ell}$ is the norm of a trailing subcolumn of the ℓth column of $L_{(\ell-)} = P_{(\ell-1)}^T A Q_{(\ell)}$ below $\beta_{\ell+\overline{d}-1,\overline{d}+\ell}$, therefore

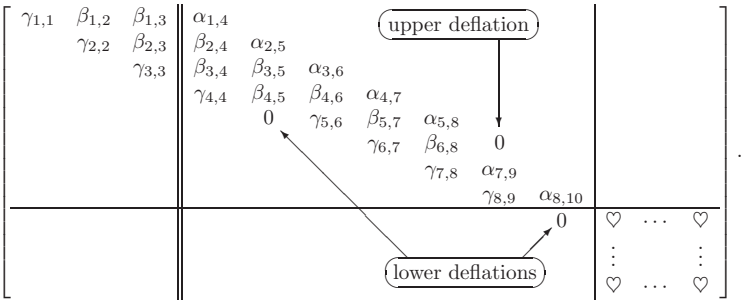
$$P_{(\ell-1)}^T [C|AQ_{(\ell)}] = [F|L_{(\ell-)}] = \begin{bmatrix} \ddots & \ddots & \vdots & \vdots & \\ & \gamma_{\overline{d}+\ell-1,\overline{d}+\ell-1} & \beta_{\ell+\overline{d}-1,\overline{d}+\ell} & \heartsuit & \cdots \\ & & 0 & \heartsuit & \cdots \\ & & \vdots & \vdots & \\ & & 0 & \heartsuit & \cdots \end{bmatrix}. \tag{3.19}$$

Then we take $H_{P,\ell} = I_m$. The matrix $L_{(\ell-)}$ in (3.19) is multiplied by $H_{Q,\ell+1}$ from the right, giving $\alpha_{\ell+1,\overline{d}+\ell+1}$, and the algorithm proceeds with transformation of the $(\ell+1)$th column (provided its trailing subcolumn is nonzero). For capturing this *lower deflation* analogously as above, it is convenient to consider a row-oriented formulation of (3.12)–(3.14). Each lower deflation modifies the pattern of nonzero entries in the band matrix by reducing the effective bandwidth from the bottom by one.

3.2.3. Band subproblem. Since the matrix $[F|L_{(k)}]$ has $(\overline{d}+1)$ nonzero diagonals (see (3.5)), after $\overline{d}$ deflations the effective bandwidth is reduced to one. Denote $P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{n\times n}$ as the products of the resulting Householder matrices (see (3.3)) and denote $L \equiv P^T A Q$. Then

$$P^T[C|AQ] = [F|L] \equiv \left[\begin{array}{c||c|c} B_1 & A_{11} & 0 \\ \hline\hline 0 & 0 & A_{22} \end{array}\right] \tag{3.20}$$

and the problem is decomposed into the desired subproblems; see, e.g., the following illustration:



Let $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$, $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$, and denote $P_1^{\mathrm{cp}} \equiv [p_1, \ldots, p_{\overline{m}}] \in \mathbb{R}^{m\times\overline{m}}$, $Q_1^{\mathrm{cp}} \equiv [q_1, \ldots, q_{\overline{n}}] \in \mathbb{R}^{n\times\overline{n}}$. In the rest of this paper we show that the *band subproblem*

$$(P_1^{\mathrm{cp}})^T [C|AQ_1^{\mathrm{cp}}] = [B_1|A_{11}] \in \mathbb{R}^{\overline{m}\times(\overline{n}+\overline{d})} \tag{3.21}$$

represents the core problem, by proving that it satisfies the properties (CP1)–(CP3); see section 2.1.

An implementation of the band algorithm with inputs A and B, and outputs A_{11}, B_1, P_1^{cp}, Q_1^{cp}, and R (see (2.5)) can be found in Appendix A. For alternative implementations see [19, Algorithm 2.4, p. 38] or [18, Algorithm 5.1, p. 74].

4. Core problem in the band form. Equations (3.20)–(3.21) immediately give

$$A^T P_1^{\mathrm{cp}} = Q \begin{bmatrix} A_{11}^T \\ \hline 0 \end{bmatrix} \qquad \text{and} \qquad [C|AQ_1^{\mathrm{cp}}] = P \left[\begin{array}{c|c} B_1 & A_{11} \\ \hline 0 & 0 \end{array}\right],$$

which represent QR decompositions of matrices $A^T P_1^{\mathrm{cp}}$ and $[C|AQ_1^{\mathrm{cp}}]$, respectively. The matrix A_{11} is in the lower triangular column echelon form with nonzero columns, thus it is of *full column rank* $\overline{n}$ giving the property (CP1). The right-hand side B_1 is in the upper triangular form with nonzero entries on the diagonal (see (3.1)), thus it is of *full column rank* $\overline{d}$ giving the property (CP2). Further $[B_1|A_{11}]$ is in the upper triangular row echelon form with nonzero rows, thus it is of *full row rank* $\overline{m}$ giving the inequality

$$\max\{\overline{n}, \overline{d}\} \le \overline{m} \le \overline{n} + \overline{d}. \tag{4.1}$$

Note that for $\overline{d} = 1$ the matrix B_1 becomes a vector b_1, the band algorithm becomes the standard Golub–Kahan bidiagonalization of A, the matrices $[b_1|A_{11}]$, A_{11} become bidiagonal with $[b_1|A_{11}]^T[b_1|A_{11}]$, $A_{11}A_{11}^T$, and $A_{11}^T A_{11}$ representing Jacobi matrices (symmetric tridiagonal matrices with positive subdiagonal entries). This relationship has been used in [14] and [12]. Jacobi matrices represent thoroughly studied objects with their origin in the first half of the 19th century; see the historical note 3.4.3 in [15, section 3.4, pp. 108–136]; see also [17, Chapter 7, pp. 119–150], [24, section 5, paragraphs 36–48, pp. 299–316], and [7, Chapter 1.3, pp. 10–20].

In the following we introduce generalized Jacobi matrices, discuss their spectral properties, and show their relationship to the band subproblem with $\overline{d} > 1$. In particular, we investigate bases of eigenspaces of generalized Jacobi matrices in section 4.1, and we show that $A_{11}A_{11}^T$ represents a generalized Jacobi matrix in section 4.2. As a

consequence, the bases of the left singular vector subspaces of A_{11} have the properties guaranteeing that the band subproblem $[B_1|A_{11}]$ satisfies also the property (CP3). Other generalizations of Jacobi matrices can be found, e.g., in [6, Chapter 3].

4.1. Generalized Jacobi matrices. Let $T \in \mathbb{R}^{n\times n}$ be a symmetric matrix with entries $t_{k,j}$. In analogy to the notation in, e.g., [8, section 4.1], we consider for $k = 1, \ldots, n$

$$f(k) = \min\{j \,:\, t_{k,j} \neq 0\} \qquad \text{and} \qquad h(k) = k - f(k). \tag{4.2}$$

The number $f(k)$ is the column index of the first nonzero entry in the kth row of T (provided it exists), and $h(k)$ is the distance between this and the diagonal entry. Consider the following matrices.

Definition 4.1 (ρ-wedge-shaped matrix). *Let $T \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and ρ, $1 \le \rho < n$, an integer. If $h(k)$ for $k = \rho+1, \ldots, n$ is positive and nonincreasing, then we call T a ρ-wedge-shaped matrix.*

For clarity we give some examples of 3-wedge-shaped matrices:

$$\begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & \\ \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit \\ & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & & \clubsuit & \heartsuit & \heartsuit & \heartsuit \end{bmatrix}, \quad \begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\ \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & \\ & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \\ & & & \clubsuit & \heartsuit & \heartsuit & \clubsuit \\ & & & & \clubsuit & \heartsuit & \heartsuit \end{bmatrix}, \quad \begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & & & & \\ \heartsuit & \heartsuit & \heartsuit & & & & \\ \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ & & \clubsuit & \heartsuit & \clubsuit & & \\ & & & \clubsuit & \heartsuit & \clubsuit & \\ & & & & \clubsuit & \heartsuit & \clubsuit \\ & & & & & \clubsuit & \heartsuit \end{bmatrix}.$$

Recall that clubs (♣) stand for nonzero entries, and hearts (♡) stand for general entries which can also be zero. Since 1-wedge-shaped matrices are symmetric tridiagonal with nonzero subdiagonal entries, the wedge-shaped matrices can be seen as a generalization of Jacobi matrices.

Jacobi matrices have simple eigenvalues; see, e.g., [17, Lemma 7.7.1]. In the text below it is shown that multiplicities of eigenvalues of a ρ-wedge-shaped matrix are bounded by ρ. The following example of a 2-wedge-shaped matrix

$$\begin{bmatrix} 0 & 0 & 1 & & & \\ 0 & 0 & 0 & \ddots & & \\ 1 & 0 & 0 & \ddots & 1 & \\ & \ddots & \ddots & \ddots & 0 & \\ & & 1 & 0 & 0 & \end{bmatrix} = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & 0 & 1 & \\ & & 1 & 0 & \end{bmatrix} \otimes I_2 \in \mathbb{R}^{8\times 8}$$

with eigenvalues

$$\lambda_{1,2} = -\frac{\sqrt{5}+1}{2}, \quad \lambda_{3,4} = -\frac{\sqrt{5}-1}{2}, \quad \lambda_{5,6} = \frac{\sqrt{5}-1}{2}, \quad \lambda_{7,8} = \frac{\sqrt{5}+1}{2}$$

illustrates that the bound is sharp, in the sense that the multiple eigenvalues with the multiplicity ρ can be present. This also shows that the *strict interlacing property* of eigenvalues of Jacobi matrices (see, e.g., [17, section 7.10]) does not hold for wedge-shaped matrices. Eigenvectors of Jacobi matrices have nonzero first and last entries; see, e.g., [17, Theorem 7.9.3 (7.9.5 in the original Prentice-Hall edition)]. The following theorem shows how to generalize the property of the nonzero first element to leading subvectors of eigenvectors of wedge-shaped matrices. This immediately gives the bound for the multiplicities of the individual eigenvalues. Subsequently, we show how

to generalize the property of the nonzero last element to eigenvectors of wedge-shaped matrices.

THEOREM 4.2. *Let $T \in \mathbb{R}^{n\times n}$ be a ρ-wedge-shaped matrix, $1 \le \rho < n$. Let $\lambda \in \mathbb{R}$, $v = [\nu_1, \dots, \nu_n]^T \in \mathbb{R}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Then the leading subvector $[\nu_1, \dots, \nu_\rho]^T \in \mathbb{R}^\rho$ of v is nonzero.*

Proof. Because $h(k)$, $k = \rho+1, \dots, n$, is nonincreasing, the first nonzero entry $t_{k,f(k)}$ in the kth row is also the last nonzero entry in the $(f(k))$th column of T. Using the symmetry of T, $t_{f(k),k}$ is the last nonzero entry in the $(f(k))$th row. Thus the $(f(k))$th row of $Tv = \lambda v$ can for $k = \rho+1, \dots, n$ be written as

$$\left(\sum_{\ell=1}^{k-1} t_{f(k),\ell}\,\nu_\ell\right) + t_{f(k),k}\,\nu_k = \lambda\,\nu_{f(k)}. \tag{4.3}$$

Let, by contradiction, $\nu_1 = \cdots = \nu_\rho = 0$. Then (4.3) is for $k = \rho+1$ reduced to

$$t_{f(\rho+1),\rho+1}\,\nu_{\rho+1} = \lambda\,\nu_{f(\rho+1)}.$$

Because $h(\rho+1)$ is positive, $f(\rho+1) < \rho+1$ and $\nu_{f(\rho+1)} = 0$. Since $t_{f(\rho+1),\rho+1} \neq 0$, then $\nu_{\rho+1} = 0$. Repeating the argument gives, for $k = \rho+2, \dots, n$, $\nu_{\rho+2} = \cdots = \nu_n = 0$, which contradicts $v \neq 0$. ∎

This theorem has the following corollary.

COROLLARY 4.3. *Let $T \in \mathbb{R}^{n\times n}$ be a ρ-wedge-shaped matrix, $1 \le \rho < n$. Let $\lambda \in \mathbb{R}$ be an eigenvalue of T with the multiplicity r. Let $v_\ell = [\nu_{1,\ell}, \dots, \nu_{n,\ell}]^T \in \mathbb{R}^n$, $\ell = 1, \dots, r$, be an arbitrary basis of the corresponding eigenspace, i.e., $TV = \lambda V$, where $V \equiv [v_1, \dots, v_r] \in \mathbb{R}^{n\times r}$. Then the leading $\rho \times r$ block of V,*

$$\Omega \equiv \begin{bmatrix} \nu_{1,1} & \cdots & \nu_{1,r} \\ \vdots & \ddots & \vdots \\ \nu_{\rho,1} & \cdots & \nu_{\rho,r} \end{bmatrix} \in \mathbb{R}^{\rho\times r}, \tag{4.4}$$

is of full column rank r.

Proof. Since $Vw = [\omega_1, \dots, \omega_n]^T$ represents for any $w \neq 0 \in \mathbb{R}^r$ an eigenvector of T, by Theorem 4.2, $\Omega w = [\omega_1, \dots, \omega_\rho]^T$ is nonzero, which gives the assertion. ∎

If $r > \rho$, then there exists a nontrivial linear combination of the columns of V which gives a vector with the first ρ entries zero, i.e., $\Omega \in \mathbb{R}^{\rho\times r}$ can obviously not have full column rank. This gives the bound for the multiplicities of individual eigenvalues.

COROLLARY 4.4. *An eigenvalue of a ρ-wedge-shaped matrix $T \in \mathbb{R}^{n\times n}$, $1 \le \rho < n$, has multiplicity at most ρ.*

The following theorem generalizes the property of the last nonzero element of eigenvectors of Jacobi matrices to eigenvectors of wedge-shaped matrices. The proof is analogous to the proof of Theorem 4.2.

THEOREM 4.5. *Let $T \in \mathbb{R}^{n\times n}$ be a ρ-wedge-shaped matrix, $1 \le \rho < n$. Let $\lambda \in \mathbb{R}$, $v = [\nu_1, \dots, \nu_n]^T \in \mathbb{R}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Denote*

$$\{s_1, \dots, s_\rho\} \equiv \{1, \dots, n\} \setminus \{f(k) : k = \rho+1, \dots, n\},$$
$$s_1 < s_2 < \cdots < s_\rho,$$

where $f(k)$ is given by (4.2)*. Then the subvector $[\nu_{s_1}, \dots, \nu_{s_\rho}]^T \in \mathbb{R}^\rho$ of v is nonzero.*

Proof. Since $t_{k,f(k)}$ is the first nonzero entry in the kth row of T, the kth row of $Tv = \lambda v$ can for $k = \rho+1, \ldots, n$ be written as

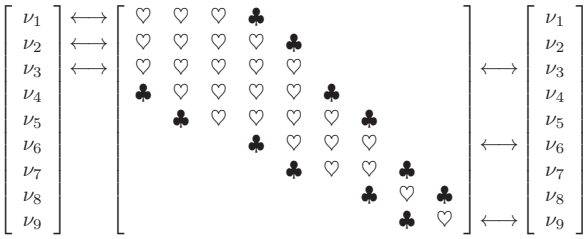
$$t_{k,f(k)}\,\nu_{f(k)} + \left(\sum_{\ell=f(k)+1}^{n} t_{k,\ell}\,\nu_\ell \right) = \lambda\,\nu_k. \tag{4.5}$$

Let, by contradiction, $\nu_{s_1} = \cdots = \nu_{s_\rho} = 0$. Because $h(n)$ is positive, $f(n) < n$, and $\nu_\ell = 0$ for all $\ell > f(n)$, in particular, $\nu_n \equiv \nu_{s_\rho} = 0$. Thus (4.5) is for $k = n$ reduced to

$$t_{n,f(n)}\,\nu_{f(n)} = 0,$$

and $t_{n,f(n)} \neq 0$ gives $\nu_{f(n)} = 0$. Repeating the argument for $k = n-1, n-2, \ldots$ up to $\rho+1$ gives $\nu_{f(n-1)} = \nu_{f(n-2)} = \cdots = \nu_{f(\rho+1)} = 0$, which contradicts $v \neq 0$. ∎

Note that $s_1, \ldots, s_\rho$ represent the row (and column) indices where the effective bandwidth of T is reduced by one, and $s_\rho = n$. Both nonzero subvectors of length ρ described by Theorems 4.2 and 4.5 can be observed from the pattern of a wedge-shaped matrix. As an illustration, eigenvectors of the following 3-wedge-shaped matrix of the size 9 have nonzero subvectors $[\nu_1, \nu_2, \nu_3]^T$ and $[\nu_3, \nu_6, \nu_9]^T$:



4.2. Singular values and vectors of the band subproblem. To prove (CP3), we first show the link of the band subproblem (3.20)–(3.21) to the wedge-shaped matrices. For a positive definite matrix $M = Z^T Z$ with its (upper triangular) Cholesky factor Z it is well known that

$$\mathrm{env}(M) = \mathrm{env}(Z^T + Z), \qquad \text{where} \qquad \mathrm{env}(T) = \{(k,j) \,:\, f(k) \le j < k\} \tag{4.6}$$

denotes the so-called *envelope* of a symmetric matrix T, and $f(k)$ is given by (4.2); see, e.g., [8, section 4.2]. Analogously, using the structure of the band subproblem (3.20)–(3.21), it is reasonable to expect that the symmetric positive semidefinite matrices $A_{11}A_{11}^T$ and $[B_1|A_{11}]^T[B_1|A_{11}]$ inherit the band structure and represent wedge-shaped matrices. However, their full row rank upper triangular factors A_{11}^T and $[B_1|A_{11}]$, respectively, do not represent the Cholesky factors, in general. Thus the abovementioned result cannot be used directly; see also an example in Figure 1. Therefore we state and prove the following lemma which shows when $A_{11}A_{11}^T$ and $[B_1|A_{11}]^T[B_1|A_{11}]$ are wedge-shaped matrices.

Lemma 4.6. *Let $A_{11}X_1 \approx B_1$, $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$, $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ be the band subproblem* (3.20)–(3.21). *If $\overline{m} > \overline{d}$, then the matrix*

$$A_{11}A_{11}^T \in \mathbb{R}^{\overline{m}\times\overline{m}}$$

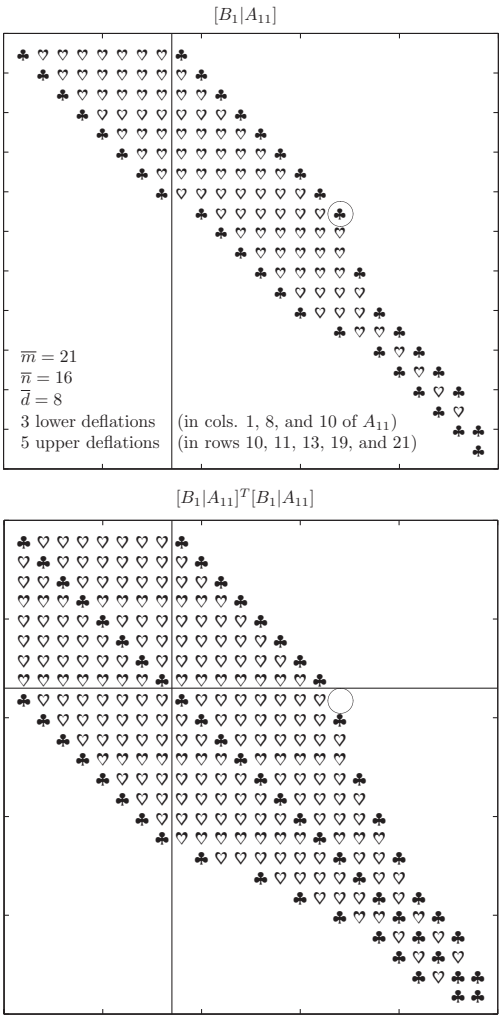


FIG. 1. *Top: A band problem* $Z \equiv [B_1|A_{11}]$ *with* $\overline{d} = 8$. *Bottom: One can see that the positive semidefinite matrix* $M \equiv [B_1|A_{11}]^T[B_1|A_{11}]$ *is an* 8*-wedge-shaped matrix. Since the upper triangular matrix* Z *in the top does not represent the Cholesky factor of the matrix* M *in the bottom, then* $\mathrm{env}(M) \neq \mathrm{env}(Z^T + Z)$; *see, e.g., the encircled nonzero entry in the top part and the encircled zero entry in the bottom part.*

is $\overline{d}$*-wedge-shaped. Since* $\overline{d} + \overline{n} > \overline{d}$, *then the matrix*

$$[B_1|A_{11}]^T[B_1|A_{11}] \in \mathbb{R}^{(\overline{d}+\overline{n})\times(\overline{d}+\overline{n})}$$

is also $\overline{d}$*-wedge-shaped.*

Proof. Denote by $a_k^T \equiv e_k^T A_{11}$ the kth *row* of A_{11}, thus $a_k^T a_j$ represents the (k,j)th entry of $A_{11}A_{11}^T$. We look for the first nonzero entry in the kth row of $A_{11}A_{11}^T$, $k = \overline{d}+1,\dots,\overline{m}$. Denote by $\varphi(j) \in \{1,\dots,\overline{m}\}$ the row index of the first nonzero entry in the jth column of A_{11}, i.e.,

$$a_{\varphi(j)}^T = [\underbrace{\heartsuit,\dots,\heartsuit}_{j-1}, \alpha_{\varphi(j),\overline{d}+j}, \underbrace{0,\dots,0}_{\overline{n}-j}], \qquad j = 1,\dots,\overline{n}. \tag{4.7}$$

Denote by $\psi(k) \in \{1,\dots,\overline{n}\}$ the column index of the first nonzero entry in the kth row of A_{11}, i.e.,

$$a_k^T = [\underbrace{0,\dots,0}_{\psi(k)-1}, \gamma_{k,\overline{d}+\psi(k)}, \underbrace{\heartsuit,\dots,\heartsuit}_{\overline{n}-\psi(k)}], \qquad k = \overline{d}+1,\dots,\overline{m}. \tag{4.8}$$

For $j = \psi(k)$ the entries $\alpha_{\varphi(j),\overline{d}+j}$ and $\gamma_{k,\overline{d}+\psi(k)}$ belong to the same column, i.e.,

$$a_k^T a_{\varphi(\psi(k))} = \gamma_{k,\overline{d}+\psi(k)}\, \alpha_{\varphi(\psi(k)),\overline{d}+\psi(k)} > 0. \tag{4.9}$$

Using the lower echelon form of A_{11}, all rows above $a_{\varphi(\psi(k))}^T$ ((4.7) with $j = \psi(k)$) are *structurally orthogonal* to a_k^T (4.8), thus all entries to the left of $a_k^T a_{\varphi(\psi(k))}$ (4.9) are zero. Consequently, $a_k^T a_{\varphi(\psi(k))}$ (4.9) is the *first nonzero entry* in the kth row of $A_{11}A_{11}^T$, $k = \overline{d}+1,\dots,\overline{m}$. Thus

$$f(k) = \varphi(\psi(k));$$

see (4.2). The matrix A_{11} has the band form with the α entries located on the top and the γ entries on the bottom of the band; see section 3.2.3 above. The row $a_{\varphi(\psi(k))}^T$ ((4.7) with $j = \psi(k)$) is always placed above the row (4.8). Thus (4.9) is on the left of the diagonal entry $a_k^T a_k$ in the kth row of $A_{11}A_{11}^T$, i.e., $\varphi(\psi(k)) < k$ and

$$h(k) = k - \varphi(\psi(k))$$

is positive for $k = \overline{d}+1,\dots,\overline{m}$. Because both $\varphi(j)$ and $\psi(k)$ are increasing, the composed function $\varphi(\psi(k))$ is also increasing, and $h(k)$ is nonincreasing for $k = \overline{d}+1,\dots,\overline{m}$. Consequently, $A_{11}A_{11}^T$ is a $\overline{d}$-wedge-shaped-matrix.

The proof for $[B_1|A_{11}]^T[B_1|A_{11}]$ is analogous: Replace A_{11} by $[B_1|A_{11}]^T$ and exchange the roles of the α and γ entries. □

Recall that $\overline{m} \geq \max\{\overline{n},\overline{d}\}$; see (4.1). Thus the only case not covered by the previous lemma is $\overline{m} = \overline{d}$, where $A_{11}A_{11}^T$ has no particular structure. Note that $\overline{d}+\overline{n} = \overline{d}$ (i.e., A_{11} has no columns) occurs in the excluded case $A^TB = 0$ (after the QR decomposition (3.1)–(3.2), the band algorithm starts with $\overline{d}$ upper deflations). The matrix $A_{11}^T A_{11} \in \mathbb{R}^{\overline{n}\times\overline{n}}$ is the trailing principal block of the $\overline{d}$-wedge-shaped matrix $[B_1|A_{11}]^T[B_1|A_{11}]$. If $\overline{n} > \overline{d}$, then $A_{11}^T A_{11}$ represents a $\overline{d}$-wedge-shaped matrix; see also Figure 1. If $\overline{n} \leq \overline{d}$, then $A_{11}^T A_{11}$ has no particular structure.

Now we are ready to apply the spectral properties of wedge-shaped matrices proved in section 4.1 to the band subproblem (3.20)–(3.21).

Corollary 4.7. *Let $A_{11}X_1 \approx B_1$, $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$, $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ be the band subproblem* (3.20)–(3.21)*, i.e., $[B_1|A_{11}]$ and A_{11} are the upper and lower triangular band matrices, respectively, with (at most) $\overline{d}+1$ nonzero diagonals. Then we have the following:*

(a) *The singular values of* $[B_1|A_{11}]$ *and* A_{11} *have multiplicities at most* $\overline{d}$.
(b) *Let* $v_1, \ldots, v_r$ *be an orthonormal basis of the right singular vector subspace of* $[B_1|A_{11}]$ *corresponding to a singular value with the multiplicity* r *(or of the null space* $\mathcal{N}([B_1|A_{11}])$ *with the dimension* r*). Then the leading* $\overline{d} \times r$ *block of* $[v_1, \ldots, v_r] \in \mathbb{R}^{(\overline{n}+\overline{d})\times r}$ *is of full column rank* r.
(c) *Let* $u_1, \ldots, u_r$ *be an orthonormal basis of the left singular vector subspace of* A_{11} *corresponding to a singular value with the multiplicity* r *(or of the null space* $\mathcal{N}(A_{11}^T)$ *with the dimension* r*). Then the matrix*

$$\Phi \equiv [u_1, \ldots, u_r]^T B_1 \in \mathbb{R}^{r\times\overline{d}}$$

is of full row rank r*, i.e., the band subproblem* $A_{11}X_1 \approx B_1$ *satisfies the condition* (CP3); *see section* 2.1.

Proof. Assertion (a) follows directly from Lemma 4.6 and Corollary 4.4, except for the case of A_{11} with $\overline{m} = \overline{d}$. Since $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ is of full column rank, $\overline{m} \geq \overline{n}$, the assertion becomes in this case trivial. Assertion (b) follows directly from Lemma 4.6 and Corollary 4.3. For assertion (c), the leading block $\Omega \in \mathbb{R}^{\overline{d}\times r}$ of $[u_1, \ldots, u_r] \in \mathbb{R}^{\overline{m}\times r}$ is of full column rank r by Corollary 4.3 (the case $\overline{m} = \overline{d}$ excluded in Lemma 4.6 becomes again trivial) and $B_1 = [F_1^T, 0]^T$, where $F_1 \in \mathbb{R}^{\overline{d}\times\overline{d}}$ is nonsingular; see (3.1). Thus Φ is of full row rank r. □

Consequently, we have proved that the band algorithm computes the problem $A_{11}X_1 \approx B_1$ that satisfies conditions (CP1)–(CP3) defining the core problem formulated in section 2.1. We state this result as the following theorem.

THEOREM 4.8. *The band subproblem* $A_{11}X_1 \approx B_1$ (3.20)–(3.21) *obtained as the output of the band algorithm described in section* 3 *applied on the problem* $AX \approx B$ *represents a core problem within* $AX \approx B$ *in the sense of Definition* 2.1. *It can therefore be called the core problem in the band form.*

5. Concluding remarks. We have shown that the band generalization of the Golub–Kahan iterative bidiagonalization algorithm always yields the minimally dimensioned subproblem within the original linear approximation problem $AX \approx B$. This consistently extends the results obtained in [16] to problems with multiple right-hand sides.

Assertions (a) and (b) of Corollary 4.7 give some additional properties of core problems. For $\overline{d} = 1$, these properties reduce to the well-known facts that singular values of $[b_1|A_{11}]$ are simple and the right singular vectors have nonzero first components, guaranteeing existence of the unique TLS solution of the core problem. The properties from Corollary 4.7 might be helpful in analysis of solvability of core problems for $\overline{d} > 1$. This issue is, however, still under investigation.

Appendix A. Implementation of the band algorithm. Algorithm 1 implements the band algorithm. Assuming exact arithmetic, it returns for general input data A, B the matrices A_{11}, B_1 of the core problem in the band form, and the corresponding transformation matrices P_1^{cp}, Q_1^{cp} (see (3.21)), and R (see (2.5)). For alternative implementations, see [19, Algorithm 2.4, p. 38] or [18, Algorithm 5.1, p. 74]. In the algorithm p_k, q_j denote the kth and jth column of P and Q, and $P_k \equiv [p_1, \ldots, p_k] \in \mathbb{R}^{m\times k}$, $Q_j \equiv [q_1, \ldots, q_j] \in \mathbb{R}^{n\times j}$ (Q_0 represents a matrix with no columns); $L_{k,j} \equiv P_k^T A Q_j$ denotes the $k \times j$ leading principal block of L, in particular $L_{\overline{m},\overline{n}} = A_{11}$ (see (3.21)), and $l_{k,j} \equiv e_k^T L e_j$ is the (k,j)th entry of L. The variables c_U and c_L are counters of the upper and lower deflations, respectively. The algorithm stops when $c_U + c_L = \overline{d} = \mathrm{rank}(B)$; see line 7. The indices j and k denote the

Algorithm 1. Band generalization of the Golub–Kahan bidiagonalization.

1: **input** $A \in \mathbb{R}^{m\times n}, \quad B \in \mathbb{R}^{m\times d}$ $\{A^TB \neq 0, \text{rank}(B) = \overline{d}\}$
2: **compute** $B = [C, 0]R^T$ {RHS preprocessing, $C \in \mathbb{R}^{m\times\overline{d}}$, $R^T = R^{-1}$}
3: **compute** $C = P_{\overline{d}} F_1$ {QR decomposition in the economic form}
4: **initialize** $Q_0 \leftarrow [\,], \quad L_{\overline{d},0} \leftarrow [\,]$ {data arrays/matrices}
5: **initialize** $c_U \leftarrow 0, \quad c_L \leftarrow 0$ {deflation counters}
6: **initialize** $j \leftarrow 1, \quad k \leftarrow \overline{d}$ {control variables/indices}
7: **while** $c_U + c_L < \overline{d}$ **do**
8: **if** $j = 1,$ **then** {compute an auxiliary vector}
9: $\mathsf{aux}_q \leftarrow A^T p_{j+c_U} = A^T p_1$
10: **else**
11: $\mathsf{aux}_q \leftarrow A^T p_{j+c_U} - \sum_{i=1}^{j-1} q_i l_{j+c_U,i}$
12: **end**
13: **if** $\mathsf{aux}_q \neq 0,$ **then**
14: $\alpha_{j+c_U,\overline{d}+j} \leftarrow \|\mathsf{aux}_q\|$ {compute α coefficient}
15: $q_j \leftarrow \mathsf{aux}_q/\alpha_{j+c_U,\overline{d}+j}$ {compute q vector}
16: $Q_j \leftarrow [Q_{j-1}, q_j]$ {update of Q matrix}
17: $\mathsf{beta} \leftarrow [\,]$
18: **if** $c_U + c_L < \overline{d} - 2,$ **then**
19: **for** $i = j + 1 + c_U, \ldots, j + \overline{d} - 1 - c_L$ **do**
20: $\beta_{i,\overline{d}+j} \leftarrow p_i^T A q_j$ {compute β coefficients}
21: $\mathsf{beta} \leftarrow [\mathsf{beta}, \beta_{i,\overline{d}+j}]$
22: **end**
23: **end**
24: $L_{k,j} \leftarrow [L_{k,j-1}, [0_{1,j-1+c_U}, \alpha_{j+c_U,\overline{d}+j}, \mathsf{beta}\,]^T]$ {update of L (add a col.)}
25: $\mathsf{aux}_p \leftarrow Aq_j - \sum_{i=1}^{k} p_i l_{i,j}$ {compute an auxiliary vector}
26: **if** $\mathsf{aux}_p \neq 0,$ **then**
27: $k \leftarrow k + 1$
28: $\gamma_{k,\overline{d}+j} \leftarrow \|\mathsf{aux}_p\|$ {compute γ coefficient}
29: $p_k \leftarrow \mathsf{aux}_p/\gamma_{k,\overline{d}+j}$ {compute p vector}
30: $P_k \leftarrow [P_{k-1}, p_k]$ {update of P matrix}
31: $L_{k,j} \leftarrow [L_{k-1,j}^T, [0_{1,j-1}, \gamma_{k,\overline{d}+j}]^T]^T$ {update of L matrix (add a row)}
32: **else**
33: $c_L \leftarrow c_L + 1$ {lower deflation}
34: **end**
35: $j \leftarrow j + 1$
36: **else**
37: $c_U \leftarrow c_U + 1$ {upper deflation}
38: **end**
39: **end**
40: $\overline{m} \leftarrow k, \overline{n} \leftarrow j - 1, B_1 \leftarrow [F_1^T, 0_{\overline{d},\overline{m}-\overline{d}}]^T, A_{11} \leftarrow L_{\overline{m},\overline{n}}, P_1^{\text{cp}} \leftarrow P_{\overline{m}}, Q_1^{\text{cp}} \leftarrow Q_{\overline{n}}$
41: **output** $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}, \quad B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}, \quad P_1^{\text{cp}} \in \mathbb{R}^{m\times\overline{m}}, \quad Q_1^{\text{cp}} \in \mathbb{R}^{n\times\overline{n}}, \quad R \in \mathbb{R}^{d\times d}$

number of columns and rows, respectively, of the currently computed part of the matrix L. If j or k becomes equal to n or m, respectively, then the algorithm stays in the loop of lines {7–13, 36–39, 7, etc.} or {7–26, 32–35, 38–39, 7, etc.}. The value of c_U, respectively, c_L increases until $c_U + c_L = \overline{d}$. The following schema illustrates

(on the example given below (3.20)) how the algorithm assembles the matrix A_{11}; the arrows represent updates in lines 24 or 31:

$$L_{\overline{d},0} = \left[\begin{array}{c}\end{array}\right] \xrightarrow{24} \left[\begin{array}{c}\alpha_{1,4}\\ \beta_{2,4}\\ \beta_{3,4}\end{array}\right] \xrightarrow{31} \left[\begin{array}{c}\alpha_{1,4}\\ \beta_{2,4}\\ \beta_{3,4}\\ \hline \gamma_{4,4}\end{array}\right] \xrightarrow{24} \left[\begin{array}{c|c}\alpha_{1,4} & 0\\ \beta_{2,4} & \alpha_{2,5}\\ \beta_{3,4} & \beta_{3,5}\\ \hline \gamma_{4,4} & \beta_{4,5}\end{array}\right] \xrightarrow{24} \left[\begin{array}{c|c|c}\alpha_{1,4} & 0 & 0\\ \beta_{2,4} & \alpha_{2,5} & 0\\ \beta_{3,4} & \beta_{3,5} & \alpha_{3,6}\\ \hline \gamma_{4,4} & \beta_{4,5} & \beta_{4,6}\end{array}\right]$$

$$\xrightarrow{31} \left[\begin{array}{c|c|c}\alpha_{1,4} & 0 & 0\\ \beta_{2,4} & \alpha_{2,5} & 0\\ \beta_{3,4} & \beta_{3,5} & \alpha_{3,6}\\ \hline \gamma_{4,4} & \beta_{4,5} & \beta_{4,6}\\ \hline 0 & 0 & \gamma_{5,6}\end{array}\right] \xrightarrow{24} \left[\begin{array}{c|c|c|c}\alpha_{1,4} & 0 & 0 & 0\\ \beta_{2,4} & \alpha_{2,5} & 0 & 0\\ \beta_{3,4} & \beta_{3,5} & \alpha_{3,6} & 0\\ \hline \gamma_{4,4} & \beta_{4,5} & \beta_{4,6} & \alpha_{4,7}\\ \hline 0 & 0 & \gamma_{5,6} & \beta_{5,7}\end{array}\right]$$

$$\xrightarrow{31} \cdots \xrightarrow{24} \left[\begin{array}{ccc|c|c|c|c}\alpha_{1,4} & 0 & 0 & 0 & 0 & 0 & 0\\ \beta_{2,4} & \alpha_{2,5} & 0 & 0 & 0 & 0 & 0\\ \beta_{3,4} & \beta_{3,5} & \alpha_{3,6} & 0 & 0 & 0 & 0\\ \hline \gamma_{4,4} & \beta_{4,5} & \beta_{4,6} & \alpha_{4,7} & 0 & 0 & 0\\ \hline 0 & 0 & \gamma_{5,6} & \beta_{5,7} & \alpha_{5,8} & 0 & 0\\ \hline 0 & 0 & 0 & \gamma_{6,7} & \beta_{6,8} & 0 & 0\\ \hline 0 & 0 & 0 & 0 & \gamma_{7,8} & \alpha_{7,9} & 0\\ \hline 0 & 0 & 0 & 0 & 0 & \gamma_{8,9} & \alpha_{8,10}\end{array}\right] = L_{\overline{m},\overline{n}} = A_{11}.$$

The sums in (3.12), (3.14), or in (3.16), (3.18), are implemented in lines 11 and 25, respectively. This implementation does not reflect, for simplicity, the structure of zero entries in the band matrix; i.e., as an example, the sum in line 11 computes the full matrix-vector product of the matrix $[q_1, \ldots, q_{j-1}]$ with the last column of $L_{k,j-1}^T$.

Acknowledgments. The authors wish to thank Diana M. Sima for valuable discussions about the TLS problems and Miroslav Tůma for his help and discussions about envelopes of band matrices. The authors are grateful to anonymous referees for their helpful comments.

REFERENCES

[1] J. L. Barlow, *Reorthogonalization for the Golub–Kahan–Lanczos bidiagonal reduction*, Numer. Math., 124 (2013), pp. 237–278.
[2] Å. Björck, *Bidiagonal Decomposition and Least Squares*, Presentation, Canberra, Australia, 2005.
[3] Å. Björck, *A Band-Lanczos Generalization of Bidiagonal Decomposition*, Presentation, Conference in Honor of G. Dahlquist, Stockholm, Sweden, 2006.
[4] Å. Björck, *A band-Lanczos algorithm for least squares and total least squares problems*, in Book of Abstracts of 4th Total Least Squares and Errors-in-Variables Modeling Workshop, Leuven, Katholieke Universiteit Leuven, Leuven, Belgium, 2006, pp, 22–23.
[5] Å. Björck, *Block Bidiagonal Decomposition and Least Squares Problems with Multiple Right-Hand Sides*, manuscript, 2008.
[6] B. Bohnhorst, *Beiträge zur Numerischen Behandlung des Unitären Eigenwertproblems*, Ph.D. thesis, Universität Bielefeld, Bielefeld, Germany, 1993.
[7] W. Gautschi, *Orthogonal Polynomials, Computation and Approximation*, Oxford University Press, New York, 2004.
[8] A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
[9] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, J. Soc. Ind. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 205–224.
[10] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[11] I. Hnětynková, M. Plešinger, D. M. Sima, Z. Strakoš, and S. Van Huffel, *The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works*, SIAM J. Matrix Anal. Appl., 32 (2011), pp, 748–770.
[12] I. Hnětynková, M. Plešinger, and Z. Strakoš, *Lanczos tridiagonalization, Golub–Kahan bidiagonalization and core problem*, Proc. Appl. Math. Mech., 6 (2006), pp, 717–718.
[13] I. Hnětynková, M. Plešinger, and Z. Strakoš, *The core problem within a linear approximation problem $AX \approx B$ with multiple right-hand sides*, SIAM J. Matrix Anal. Appl., 34 (2013), pp, 917–931.
[14] I. Hnětynková and Z. Strakoš, *Lanczos tridiagonalization and core problems*, Linear Algebra Appl., 421 (2007), pp, 243–251.
[15] J. Liesen and Z. Strakoš, *Krylov Subspace Methods, Principles and Analysis*, Oxford University Press, Oxford, 2013.
[16] C. C. Paige and Z. Strakoš, *Core problem in linear algebraic systems*, SIAM J. Matrix Anal. Appl., 27 (2005), pp, 861–875.
[17] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1998.
[18] M. Plešinger, *The Total Least Squares Problem and Reduction of Data in $AX \approx B$*, Ph.D. thesis, Technical University of Liberec, Liberec, Czech Republic, 2008.
[19] D. M. Sima, *Regularization Techniques in Model Fitting and Parameter Estimation*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.
[20] D. M. Sima and S. Van Huffel, *Core Problems in $AX \approx B$*, Technical report, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.
[21] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Front. Appl. Math. 9, SIAM, Philadelphia, 1991.
[22] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl., 13 (1992), pp, 746–763.
[23] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math., 62 (1992), pp, 123–148.
[24] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, England, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 37, No. 3, pp. 861–876

SOLVABILITY OF THE CORE PROBLEM WITH MULTIPLE RIGHT-HAND SIDES IN THE TLS SENSE*

IVETA HNĚTYNKOVÁ†, MARTIN PLEŠINGER‡, AND DIANA MARIA SIMA§

Abstract. Recently it was shown how necessary and sufficient information for solving an orthogonally invariant linear approximation problem $AX \approx B$ with multiple right-hand sides can be revealed through the so-called core problem reduction; see [I. Hnětynková, M. Plešinger, and Z. Strakoš, *SIAM J. Matrix Anal. Appl.*, 34 (2013), pp. 917–931]. The total least squares (TLS) serves as an important example of such approximation problem. Solvability of TLS was discussed in the full generality in [I. Hnětynková et al., *SIAM J. Matrix Anal. Appl.*, 32 (2011), pp. 748–770]. This theoretical study investigates solvability of core problems with multiple right-hand sides in the TLS sense. It is shown that, contrary to the single right-hand side case, a core problem with multiple right-hand sides may not have a TLS solution. Further possible internal structure of core problems is studied. Outputs of the classical TLS algorithm for the original problem $AX \approx B$ and for the core problem within $AX \approx B$ are compared.

Key words. total least squares (TLS) problem, multiple right-hand sides, core problem, linear approximation problem, error-in-variables modeling, orthogonal regression, classical TLS algorithm

AMS subject classifications. 15A06, 15A18, 15A21, 15A24, 65F20, 65F25

DOI. 10.1137/15M1028339

1. Introduction.

Consider a linear approximation problem

$$AX \approx B \qquad \text{or, equivalently,} \qquad [B|A]\begin{bmatrix} -I_d \\ X \end{bmatrix} \approx 0, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{m\times d}$, and $A^TB \neq 0$. (Otherwise the only meaningful solution is $X \equiv 0$.) Consider its orthogonal transformation

$$\widehat{A}\widehat{X} \equiv (P^TAQ)(Q^TXR) \approx (P^TBR) \equiv \widehat{B}, \tag{1.2}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$, or, equivalently,

$$[\widehat{B}|\widehat{A}]\begin{bmatrix} -I_d \\ \widehat{X} \end{bmatrix} \equiv \left(P^T[B|A]\begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix}\right)\left(\begin{bmatrix} R^T & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} -I_d \\ X \end{bmatrix}R\right) \approx 0. \tag{1.3}$$

We assume that the problem (1.1) is *orthogonally invariant,* meaning that X solves (1.1) *if and only if* $\widehat{X} = Q^TXR$ solves (1.2)–(1.3). In this paper, we focus on the *total least squares (TLS)* formulation of (1.1), i.e.,

$$\min_{X,E,G} \|[G|E]\|_F \qquad \text{subject to} \qquad (A+E)X = B+G, \tag{1.4}$$

*Received by the editors June 29, 2015; accepted for publication (in revised form) by J. L. Barlow February 29, 2016; published electronically July 19, 2016. This work was supported by GAČR grant P201/13–06684S.
http://www.siam.org/journals/simax/37-3/M102833.html

†Faculty of Mathematics and Physics, Charles University, Prague 8, Czech Republic, and Institute of Computer Science, AS CR, Prague, Czech Republic (hnetynkova@cs.cas.cz). This author is a member of the University Center for Mathematical Modeling, Applied Analysis, and Computational Mathematics (Math MAC).

‡Department of Mathematics, Technical University of Liberec, Liberec, Czech Republic (martin.plesinger@tul.cz). The research of this author was supported by internal grant FP-SGS/2016/21161 (carried together with J. Žáková) of the Technical University of Liberec.

§Department of Electrical Engineering, ESAT-STADIUS, Katholieke Universiteit Leuven, Leuven, Belgium (diana.sima@esat.kuleuven.be).

where $\|.\|_F$ denotes the Frobenius norm. The TLS problem (1.1)–(1.4) has been investigated for decades; see [3], [12], [13], [14]. The existence and uniqueness of a TLS solution with $d \geq 1$ was discussed in full generality in the paper [4], revealing some difficulties in previous concepts. For example, the output of the widely used computational approach—the *classical TLS algorithm* (see [10], [11], [12, Chapter 3.6.1, pp. 87–90])—in some cases differs from the existing TLS solution.

The so-called core problem reduction introduced in [7] for $d = 1$ brings a different point of view. It decomposes the original problem into two independent parts, one *always having the unique TLS solution* and minimal dimension (called the *core problem*), and the other having zero solution and maximal dimension. In this way, the core problem reduction allows one to separate the *necessary and sufficient information* for solving the problem from the rest. Moreover, the core problem reduction for $d = 1$ is *invariant with respect to the classical TLS algorithm* in the sense that the output of the algorithm for a problem $Ax \approx b$ is equal to the output for its core problem (after some orthogonal back-transformation).

The core concept was generalized to $d > 1$ recently in [5] and [6]; see also the first attempts in [1], [2], [9], and [8]. The construction of a core problem is based either on the singular value decomposition (SVD) in [5] or on the band generalization of the Golub–Kahan bidiagonalization in [6]. Each of the approaches allowed investigators to obtain several properties of core problems with $d > 1$. However, existence and possible uniqueness of a TLS solution of a core problem, as well as any influence of the core reduction on the outputs of the TLS algorithm, have not yet been analyzed. Here we fill in these gaps. In particular, we prove that a core problem with $d > 1$ may belong to any of the classification sets introduced in [4], and thus it may not have a TLS solution. We show this constructively by deriving examples of core problems belonging to all the sets. On the other hand, we prove that the invariance of the classical TLS algorithm under the core problem reduction holds also for $d > 1$.

Section 2 summarizes the classification of TLS problems and gives properties of core problems with $d \geq 1$. It describes the basic results on solvability of core problems in the TLS sense. Section 3 employs properties of core problems to explore their possible *internal structure*. Section 4 analyzes the solvability of core problems with $d \geq 1$ in the TLS sense. Section 5 discusses the core problem reduction in the context of the classical TLS algorithm. Section 6 concludes the paper.

Throughout the paper, $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote the range and null-space, respectively, of a matrix M; I_ℓ (or just I) denotes an $\ell \times \ell$ identity matrix; $0_{\ell,\xi}$ (or just 0) denotes an $\ell \times \xi$ zero matrix; and $M^\dagger$ denotes the Moore–Penrose pseudoinverse of M. The matrices A, B, $[B|A]$, and X from (1.1) are called the *system matrix*, the *right-hand sides (or the observation) matrix*, the *extended (or data) matrix*, and the *matrix of unknowns*, respectively.

2. Preliminaries. First we briefly recall some results from [4], [5], [6], and [7].

2.1. Classification of TLS problems. Consider the problem (1.1) and its SVD

$$[B|A] = U\Sigma V^T \quad \text{and} \quad \sigma_1 \geq \cdots \geq \sigma_n \geq \sigma_{n+1} \geq \cdots \geq \sigma_{n+d} \geq 0; \tag{2.1}$$

i.e., σ_j denote the singular values *including the multiplicities*. If $m < n + d$ (which is often the case for core problems), then

$$\sigma_1 \geq \cdots \geq \sigma_n \geq \sigma_{n+1} \geq \cdots \geq \sigma_m \geq 0 \quad \text{and} \quad \sigma_{m+1} = \cdots = \sigma_{n+d} \equiv 0.$$

Define integers q, $n \geq q \geq 0$, and e, $d \geq e \geq 1$, such that

$$\sigma_{n-q} > \underbrace{\sigma_{n-q+1} = \cdots = \sigma_n}_{q} = \underbrace{\sigma_{n+1} = \cdots = \sigma_{n+e}}_{e} > \sigma_{n+e+1}; \tag{2.2}$$

i.e., q is the "left" and e is the "right multiplicity" of σ_{n+1}. (Note that σ_{n-q} or σ_{n+e+1} may not exist if $q = n$ or $e = d$, respectively.) Further define a conformal partitioning[1] of the matrix $V \in \mathbb{R}^{(n+d)\times(n+d)}$,

$$V = \begin{bmatrix} \underbrace{V_{11}}_{n-q} & \underbrace{V_{12}}_{q+e} & \underbrace{V_{13}}_{d-e} \\ V_{21} & V_{22} & V_{23} \end{bmatrix} \begin{matrix} \}d \\ \}n \end{matrix}. \tag{2.3}$$

The block-columns correspond to singular values strictly larger than, equal to, and strictly smaller than σ_{n+1}, respectively. (Note that V_{11}, V_{21} or V_{13}, V_{23} may not exist if $n = q$ or $d = e$, respectively.)

The classification introduced in [4] distinguishes four disjoint sets $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$, and $\mathcal{S}$ of TLS problems based on ranks of individual blocks in (2.3):[2]

- $\mathcal{F}$: If $\operatorname{rank}([V_{12}, V_{13}]) = d$, then the problem is of the *first class* and the following hold:
  - $\mathcal{F}_1$: If $\operatorname{rank}(V_{12}) = e$, $\operatorname{rank}(V_{13}) = d-e$, then the problem has a TLS solution (that may be unique as well as nonunique), and there exists a TLS solution minimal in both the Frobenius and the 2-norm at the same time. Such a solution can be computed by the classical TLS algorithm.
  - $\mathcal{F}_2$: If $\operatorname{rank}(V_{12}) > e$, $\operatorname{rank}(V_{13}) = d - e$, then the problem has a nonunique TLS solution, and the TLS solutions minimal in the Frobenius and 2-norms may be different. The classical TLS algorithm cannot reach a TLS solution.
  - $\mathcal{F}_3$: If $\operatorname{rank}(V_{12}) > e$, $\operatorname{rank}(V_{13}) < d - e$, then the problem does not have a TLS solution.
- $\mathcal{S}$: If $\operatorname{rank}([V_{12}, V_{13}]) < d$, then the problem is of the *second class*, and it does not have a TLS solution.

Note that a problem $Ax \approx b$ with a single right-hand side ($d = 1$) belongs either to the set $\mathcal{F}_1$ or to the set $\mathcal{S}$. Further, the TLS solution minimal in both the Frobenius and the 2-norm is in $\mathcal{F}_1$ (and only here) given by

$$X_{\mathrm{TLS}} \equiv -[V_{22}, V_{23}][V_{12}, V_{13}]^{\dagger}, \tag{2.4}$$

and it represents the unique TLS solution if and only if $[V_{12}, V_{13}]$ is square invertible (or equivalently $\sigma_n > \sigma_{n+1}$, or V_{12} is of *full column rank*); see [12], [4].

2.2. Core problem properties. The core problem within (1.1) is defined as follows; see [5, Definition 5.2].

[1] Different partitionings of V are used in the literature. The paper [4] and the book [12] use partitioning into four blocks $V_{ij}{}^{(q)}$, $i = 1, 2$, $j = 1, 2$. The blocks V_{11}, $[V_{12}, V_{13}]$, V_{22}, $[V_{22}, V_{23}]$ of (2.3) correspond to the blocks $V_{11}{}^{(q)}$, $V_{12}{}^{(q)}$, $V_{21}{}^{(q)}$, $V_{22}{}^{(q)}$, respectively, in [4]. Since the analysis in [12] is based on the matrix $[A|B]$ instead of $[B|A]$, the roles of block-rows of V in [12] and [4] are exchanged.

[2] The classical literature (e.g., the book [12]) often uses the terminology *generic* and *nongeneric* problems, which in this classification correspond to the sets $\mathcal{F}$ and $\mathcal{S}$, respectively.

Definition 2.1 (core problem). *The subproblem $A_{11}X_1 \approx B_1$ is a core problem within the approximation problem $AX \approx B$ if $[B_1|A_{11}]$ is minimally dimensioned and A_{22} maximally dimensioned subject to the orthogonal transformations of the form*

$$P^T[B|A]\begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} = [P^TBR|P^TAQ] \equiv \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right], \tag{2.5}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$.

The following theorem describes the core problem by its properties.

Theorem 2.2. *Let $A_{11}X_1 \approx B_1$ be any subproblem obtained by an orthogonal transformation of the form (2.5). Then $A_{11}X_1 \approx B_1$ represents the core problem if and only if it satisfies the following properties:*

(CP1) *The matrix $A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}$ is of full column rank equal to $\overline{n} \le \overline{m}$.*

(CP2) *The matrix $B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}}$ is of full column rank equal to $\overline{d} \le \overline{m}$.*

(CP3) *Let A_{11} have k distinct nonzero singular values $\sigma_j(A_{11})$ with multiplicities r_j, and $r_{k+1} \equiv \dim(\mathcal{N}(A_{11}^T))$. Let $U_j \in \mathbb{R}^{\overline{m}\times r_j}$ be matrices having orthonormal bases of left singular vector subspaces of A_{11} corresponding to $\sigma_j(A_{11})$ for $j = 1,\dots,k$, and of $\mathcal{N}(A_{11}^T)$ for $j = k+1$, as their columns. The matrices $\Phi_j \equiv U_j^TB_1 \in \mathbb{R}^{r_j\times\overline{d}}$ are of full row rank equal to $r_j \le \overline{d}$ for $j = 1,\dots,k+1$.*

For the proof see [5, section 4]. Any core problem has also the following properties:

(CP4) The extended matrix $[B_1|A_{11}] \in \mathbb{R}^{\overline{m}\times(\overline{n}+\overline{d})}$ is of *full row rank* equal to $\overline{m}$, which follows from (CP1) and (CP3); see [6, section 2.1].

(CP5) Let $[B_1|A_{11}]$ have k distinct nonzero singular values $\sigma_j([B_1|A_{11}])$ with multiplicities ϱ_j, and $\varrho_{k+1} \equiv \dim(\mathcal{N}([B_1|A_{11}]))$. Let $V_j \in \mathbb{R}^{(\overline{n}+\overline{d})\times\varrho_j}$ be matrices having orthonormal bases of right singular vector subspaces of $[B_1|A_{11}]$ corresponding to $\sigma_j([B_1|A_{11}])$ for $j = 1,\dots,k$, and of $\mathcal{N}([B_1|A_{11}])$ for $j = k+1$, as their columns. The leading $\overline{d}\times\varrho_j$ submatrices of V_j are of *full column rank* equal to $\varrho_j \le \overline{d}$, for $j = 1,\dots,k+1$; see [6, Corollary 4.7(b)].

(CP6) Let $\sigma_j(A_{11})$ denote distinct singular values of A_{11} with multiplicities r_j, then $r_j \le \overline{d}$ and $\sum_j r_j = \overline{n}$, which follows from (CP3) and (CP1).

(CP7) Let $\sigma_j([B_1|A_{11}])$ denote distinct singular values of $[B_1|A_{11}]$ with multiplicities ϱ_j; then $\varrho_j \le \overline{d}$ and $\sum_j \varrho_j = \overline{m}$, which follows form (CP5) and (CP4).

It is worth noting that matrices A_{11} and B_1 of the core problem within $AX \approx B$ are not unique. Any problem $\widehat{A}_{11}\widehat{X}_1 \approx \widehat{B}_1$ obtained by an orthogonal transformation

$$[\widehat{B}_1|\widehat{A}_{11}] \equiv [\widehat{P}^TB_1\widehat{R}|\widehat{P}^TA_{11}\widehat{Q}] = \widehat{P}^T[B_1|A_{11}]\begin{bmatrix} \widehat{R} & 0 \\ 0 & \widehat{Q} \end{bmatrix}, \tag{2.6}$$

where $\widehat{P}^{-1} = \widehat{P}^T$, $\widehat{Q}^{-1} = \widehat{Q}^T$, $\widehat{R}^{-1} = \widehat{R}^T$, also represents a core problem within $AX \approx B$; see [5, Remark 3.1]. The properties (CP1)–(CP7) are *invariant* with respect to the orthogonal transformation of the form (2.6).

Remark 2.3. Revealing the core problem using the SVD of A with the SVD-preprocessing of B leads to the so-called *SVD form* of the core problem with B_1 having mutually orthogonal columns, and A_{11} diagonal (possibly with some additional zero rows); see [5]. The approach based on the band generalization of the Golub–Kahan bidiagonalization algorithm leads to the so-called *band form* with B_1 upper triangular, A_{11} lower triangular in a column echelon form, and $[B_1|A_{11}]$ upper triangular in a row echelon form; see [6].

2.3. Basic results on solvability of core problems. Using (2.5), the original problem $AX \approx B$ transforms into the block form

$$\left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right] (Q^T X R) \approx \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \end{array}\right].$$

A conformal partitioning of the matrix of unknowns

$$Q^T X R = \underbrace{\left[\begin{array}{c} X_1 \\ X_2 \end{array}\right.}_{\overline{d}} \underbrace{\left.\begin{array}{c} Z_1 \\ Z_2 \end{array}\right]}_{d-\overline{d}} \begin{array}{l} \} \, \overline{n} \\ \} \, n-\overline{n} \end{array} \tag{2.7}$$

allows us to split the original problem $AX \approx B$ into subproblems

$$A_{11}X_1 \approx B_1 \qquad \text{and} \qquad A_{11}Z_1 \approx 0,\ A_{22}X_2 \approx 0,\ A_{22}Z_2 \approx 0.$$

The last three subproblems are homogenous. Following the arguments in [7], $X_2 \equiv 0$, $Z_1 \equiv 0$, and $Z_2 \equiv 0$, and only the core problem

$$A_{11}X_1 \approx B_1 \quad \text{or, equivalently,} \quad [B_1|A_{11}] \left[\begin{array}{c} -I_{\overline{d}} \\ X_1 \end{array}\right] \approx 0, \tag{2.8}$$

has to be solved. If X_1 is some solution (e.g., the TLS solution) of (2.8), then

$$X \equiv Q \left[\begin{array}{cc} X_1 & 0 \\ 0 & 0 \end{array}\right] R^T \tag{2.9}$$

represents a reasonable solution of the original problem $AX \approx B$, *fully determined* by the solution of the core problem. This leads us to the fundamental question of existence (and possibly also uniqueness) of a TLS solution of (2.8).

For a problem with *single right-hand side*, i.e., $d = \overline{d} = 1$, the properties (CP5) and (CP7) guarantee solvability of the corresponding core problem in the TLS sense; see [3], [7], and [4].

COROLLARY 2.4. *Let $Ax \approx b$ be an approximation problem and $A_{11}x_1 \approx b_1$ the core problem within $Ax \approx b$. Then the following hold:*

- *The smallest singular value $\sigma_{\overline{n}+1}([b_1|A_{11}])$ of $[b_1|A_{11}]$ is simple.*
- *The first entry of the right singular vector of $[b_1|A_{11}]$ corresponding to the singular value $\sigma_{\overline{n}+1}([b_1|A_{11}])$ is nonzero.*

Consequently, the core problem $A_{11}x_1 \approx b_1$ belongs to the set $\mathcal{F}_1$ and has the unique TLS solution.

As a result, the core problem reduction simplifies the classification of single right-hand side problems and gives (through the back-transformation) a clear interpretation framework for a solution of the original problem.

Now consider a core problem with multiple right-hand sides,

$$A_{11}X_1 \approx B_1, \qquad A_{11} \in \mathbb{R}^{\overline{m}\times\overline{n}}, \qquad B_1 \in \mathbb{R}^{\overline{m}\times\overline{d}},$$

and its SVD analogous to (2.1)–(2.3), where $\overline{q} + \overline{e}$ is the multiplicity of the singular value $\sigma_{\overline{n}+1}$ of $[B_1|A_{11}]$, and the matrix V of right singular vectors of $[B_1|A_{11}]$ is partitioned with respect to integers $\overline{n}$, $\overline{d}$, $\overline{q}$, and $\overline{e}$. Since the columns of

$$\left[\begin{array}{c} V_{12} \\ V_{22} \end{array}\right] \in \mathbb{R}^{(\overline{n}+\overline{d})\times(\overline{q}+\overline{e})} \tag{2.10}$$

form an orthonormal basis of the right singular vectors subspace corresponding to $\sigma_{\overline{n}+1}$, the leading $\overline{d}\times(\overline{q}+\overline{e})$ submatrix of (2.10), i.e., the matrix V_{12}, is of *full column rank* by (CP5). This observation together with the results summarized in section 2.1 gives the following corollary.

Corollary 2.5. *Let $A_{11}X_1 \approx B_1$ be a core problem. Then the following assertions are equivalent:*

- *The problem $A_{11}X_1 \approx B_1$ has a unique TLS solution.*
- *The problem $A_{11}X_1 \approx B_1$ belongs to the set $\mathcal{F}_1$.*

For multiple right-hand sides problems with core problems in $\mathcal{F}_1$, this corollary states solvability results generalizing the results for single right-hand side problems. However, we show that a core problem with multiple right-hand sides may belong to any of the sets $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, or $\mathcal{S}$. We do it constructively, by building examples of core problems belonging to these sets. In order to do this, we first prove that core problems may have some further internal structure not present in problems with $d = 1$.

3. Internal structure of core problems. The following theorem shows that a core problem with multiple right-hand sides may be composed of two or more independent subproblems having the core problem properties.

Theorem 3.1. *The problems*

$$\begin{aligned} A_{11}^{(\alpha)}X_1^{(\alpha)} &\approx B_1^{(\alpha)}, \qquad A_{11}^{(\alpha)} \in \mathbb{R}^{\overline{m}_\alpha\times\overline{n}_\alpha}, \quad B_1^{(\alpha)} \in \mathbb{R}^{\overline{m}_\alpha\times\overline{d}_\alpha}, \\ A_{11}^{(\beta)}X_1^{(\beta)} &\approx B_1^{(\beta)}, \qquad A_{11}^{(\beta)} \in \mathbb{R}^{\overline{m}_\beta\times\overline{n}_\beta}, \quad B_1^{(\beta)} \in \mathbb{R}^{\overline{m}_\beta\times\overline{d}_\beta}, \end{aligned}$$

represent two core problems, i.e., both satisfy (CP1)–(CP3)*, if and only if the problem*

$$A_{11}X_1 \approx B_1, \quad \text{where } A_{11} = \begin{bmatrix} A_{11}^{(\alpha)} & 0 \\ 0 & A_{11}^{(\beta)} \end{bmatrix} \text{ and } B_1 = \begin{bmatrix} B_1^{(\alpha)} & 0 \\ 0 & B_1^{(\beta)} \end{bmatrix}, \tag{3.1}$$

represents a core problem, i.e., satisfies (CP1)–(CP3).

Proof. Obviously (CP1) are (CP2) hold; thus we concentrate on (CP3). Let

$$A_{11}^{(\ell)} = U^{(\ell)}\Sigma^{(\ell)}(V^{(\ell)})^T, \quad \ell = \alpha, \beta,$$

be SVDs with the standard ordering of singular values. Denote by $\sigma_i^{(\ell)}$ the *distinct nonzero* singular values of $A_{11}^{(\ell)}$, by $U_i^{(\ell)}$ the blocks of corresponding left singular vectors, and let $\Phi_i^{(\ell)} \equiv (U_i^{(\ell)})^T B_1^{(\ell)}$, $\ell = \alpha, \beta$. Let Ψ_L, Ψ_R be permutation matrices such that

$$A_{11} = \underbrace{\left(\begin{bmatrix} U^{(\alpha)} & 0 \\ 0 & U^{(\beta)} \end{bmatrix} \Psi_L \right)}_{U_{11}} \underbrace{\left(\Psi_L^T \begin{bmatrix} \Sigma^{(\alpha)} & 0 \\ 0 & \Sigma^{(\beta)} \end{bmatrix} \Psi_R \right)}_{\Sigma_{11}} \underbrace{\left(\Psi_R^T \begin{bmatrix} V^{(\alpha)} & 0 \\ 0 & V^{(\beta)} \end{bmatrix}^T \right)}_{V_{11}^T}$$

is an SVD of A_{11} with the standard ordering of singular values. Denote by σ_i the *distinct nonzero* singular values of A_{11}, by U_i the blocks of corresponding left singular vectors, and let $\Phi_i \equiv U_i^T B_1$. For each σ_i one of the following assertions is true:

- $\sigma_i = \sigma_j^{(\alpha)}$ for some j, and it is *not* a singular value of $A_{11}^{(\beta)}$;
- $\sigma_i = \sigma_k^{(\beta)}$ for some k, and it is *not* a singular value of $A_{11}^{(\alpha)}$;
- $\sigma_i = \sigma_j^{(\alpha)} = \sigma_k^{(\beta)}$ for some j and k, so it is a singular value of both $A_{11}^{(\alpha)}$, $A_{11}^{(\beta)}$.

Let the permutations Ψ_L, Ψ_R be chosen such that

$$U_i = \begin{bmatrix} U_j^{(\alpha)} \\ 0 \end{bmatrix} \quad \text{or} \quad U_i = \begin{bmatrix} 0 \\ U_k^{(\beta)} \end{bmatrix} \quad \text{or} \quad U_i = \begin{bmatrix} U_j^{(\alpha)} & 0 \\ 0 & U_k^{(\beta)} \end{bmatrix}$$

in these cases, respectively. This gives

$$\Phi_i = [\Phi_j^{(\alpha)}, 0] \quad \text{or} \quad \Phi_i = [0, \Phi_k^{(\beta)}] \quad \text{or} \quad \Phi_i = \begin{bmatrix} \Phi_j^{(\alpha)} & 0 \\ 0 & \Phi_k^{(\beta)} \end{bmatrix},$$

and Φ_i is of full row rank *if and only if* $\Phi_j^{(\alpha)}$ or $\Phi_k^{(\beta)}$ or both are of full row rank, respectively. The same result can be obtained analogously also for $\Phi_N \equiv U_N^T B_1$ and $\Phi_N^{(\ell)} \equiv (U_N^{(\ell)})^T B_1$, where U_N, $U_N^{(\ell)}$ are bases of $\mathcal{N}(A_{11}^T)$, $\mathcal{N}((A_{11}^{(\ell)})^T)$, respectively. □

It is worth introducing the following theorem, although it can be considered as a special case of the previous one.

THEOREM 3.2 (degenerate version of Theorem 3.1). *The problem*

$$A_{11}^{(\alpha)} X_1^{(\alpha)} \approx B_1^{(\alpha)}, \qquad A_{11}^{(\alpha)} \in \mathbb{R}^{\overline{m}_\alpha \times \overline{n}_\alpha}, \quad B_1^{(\alpha)} \in \mathbb{R}^{\overline{m}_\alpha \times \overline{d}_\alpha},$$

represents a core problem, i.e., satisfies (CP1)–(CP3)*, and* $B_1^{(\beta)} \in \mathbb{R}^{\overline{m}_\beta \times \overline{d}_\beta}$ *is square nonsingular, if and only if*

$$A_{11} X_1 \approx B_1, \qquad A_{11} = \begin{bmatrix} A_{11}^{(\alpha)} \\ 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} B_1^{(\alpha)} & 0 \\ 0 & B_1^{(\beta)} \end{bmatrix} \tag{3.2}$$

represents a core problem, i.e., satisfies (CP1)–(CP3).

Proof. The proof is straightforward. Technically, one can use Theorem 3.1 with $A_{11}^{(\beta)} \in \mathbb{R}^{\overline{m}_\beta \times 0}$, a system matrix with no columns. □

The following definition completes Theorems 3.1 and 3.2.

DEFINITION 3.3 (reducible core problem). *Any core problem* $[B_1|A_{11}]$ *having, after some transformation of the form* (2.6)*, the block structure* (3.1) *or* (3.2)*, i.e.,*

$$\widehat{P}^T [B_1|A_{11}] \begin{bmatrix} \widehat{R} & 0 \\ 0 & \widehat{Q} \end{bmatrix} = \left[\begin{array}{cc|cc} B_1^{(\alpha)} & 0 & A_{11}^{(\alpha)} & 0 \\ 0 & B_1^{(\beta)} & 0 & A_{11}^{(\beta)} \end{array}\right], \tag{3.3}$$

where $\widehat{P}^{-1} = \widehat{P}^T$, $\widehat{Q}^{-1} = \widehat{Q}^T$, $\widehat{R}^{-1} = \widehat{R}^T$*, including the case of* $A_{11}^{(\beta)}$ *with no columns, is called reducible (or composed or decomposable). A core problem which is not reducible is called irreducible.*

Individual subproblems $[B_1^{(\ell)}|A_{11}^{(\ell)}]$*,* $\ell = \alpha, \beta$*, are called components of* $[B_1|A_{11}]$*. If* $\mathcal{R}(B_1^{(\ell)}) \subset \mathcal{R}(A_{11}^{(\ell)})$*, then* $[B_1^{(\ell)}|A_{11}^{(\ell)}]$ *is called a compatible component. A component which is not compatible is called incompatible. If* $A_{11}^{(\ell)}$ *has no columns, then* $[B_1^{(\ell)}|A_{11}^{(\ell)}] = [B_1^{(\ell)}]$ *is called a degenerate component.*

Based on the previous results, we may conclude that when some core problem is reducible, it would be more reasonable to solve the individual subproblems independently. However, the question of constructing the block-structure-revealing transformation (3.3), or at least of detecting whether the given (core) problem is reducible or not, remains open.

Remark 3.4. An extreme example of an incompatible reducible core problem consists only of one compatible and one degenerate component. The compatible component $A_{11}^{(\alpha)}X_1^{(\alpha)} \approx B_1^{(\alpha)}$, $\mathcal{R}(B_1^{(\alpha)}) \subset \mathcal{R}(A_{11}^{(\alpha)})$ has a square nonsingular system matrix (see (CP1) and (CP4)), and thus (if it is treated separately) it has the unique solution $X_1^{(\alpha)} = (A_{11}^{(\alpha)})^{-1}B_1^{(\alpha)}$. The right-hand side of the degenerate component cannot be approximated by the columns of A_{11}; see (3.2). Thus, while solving a TLS problem (1.4), the degenerate component only increases the norm of the correction matrix G.

4. TLS solvability of core problems. Now we employ Theorem 3.1 to get examples of core problems with multiple right-hand sides in all four sets $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$, and $\mathcal{S}$.

4.1. Examples of reducible core problems. Consider three incompatible core problems with single right-hand sides $A_{11}^{(\ell)}x_1^{(\ell)} \approx b_1^{(\ell)}$, $A_{11}^{(\ell)} \in \mathbb{R}^{3\times 2}$, $b_1^{(\ell)} \in \mathbb{R}^3$, for $\ell = \alpha, \beta, \gamma$. Let

$$[b_1^{(\ell)}|A_{11}^{(\ell)}] = U^{(\ell)}\Sigma^{(\ell)}(V^{(\ell)})^T = U^{(\ell)} \left[\begin{array}{cc|c} \sigma_1^{(\ell)} & 0 & 0 \\ 0 & \sigma_2^{(\ell)} & 0 \\ 0 & 0 & \sigma_3^{(\ell)} \end{array}\right] \left[\begin{array}{cc|c} v_{11}^{(\ell)} & v_{12}^{(\ell)} & v_{13}^{(\ell)} \\ v_{21}^{(\ell)} & v_{22}^{(\ell)} & v_{23}^{(\ell)} \\ v_{31}^{(\ell)} & v_{32}^{(\ell)} & v_{33}^{(\ell)} \end{array}\right]^T \tag{4.1}$$

be the SVDs of their extended matrices. The partitioning (2.3) of $V^{(\ell)}$ is suggested by the lines. (The third block-column is not present since the "right multiplicities" of $\sigma_3^{(\ell)}$ are equal to one; see (2.2).)

Example 4.1. Consider a *reducible core problem* containing two subproblems (4.1) as components:

$$[B_1|A_{11}] \equiv \left[\begin{array}{cc|cc} b_1^{(\alpha)} & 0 & A_{11}^{(\alpha)} & 0 \\ 0 & b_1^{(\beta)} & 0 & A_{11}^{(\beta)} \end{array}\right] \in \mathbb{R}^{\overline{m}\times(\overline{n}+\overline{d})}, \qquad \overline{m} = 6,\ \overline{n} = 4,\ \overline{d} = 2.$$

There exist permutation matrices Ψ_L, Ψ_R such that

$$[B_1|A_{11}] = U\Sigma V^T, \qquad U = \left[\begin{array}{cc} U^{(\alpha)} & 0 \\ 0 & U^{(\beta)} \end{array}\right]\Psi_L, \qquad \Sigma = \Psi_L^T\left[\begin{array}{cc} \Sigma^{(\alpha)} & 0 \\ 0 & \Sigma^{(\beta)} \end{array}\right]\Psi_R,$$

$$V = \left[\begin{array}{ccc|ccc} v_{11}^{(\alpha)} & v_{12}^{(\alpha)} & v_{13}^{(\alpha)} & 0 & 0 & 0 \\ 0 & 0 & 0 & v_{11}^{(\beta)} & v_{12}^{(\beta)} & v_{13}^{(\beta)} \\ \hline v_{21}^{(\alpha)} & v_{22}^{(\alpha)} & v_{23}^{(\alpha)} & 0 & 0 & 0 \\ v_{31}^{(\alpha)} & v_{32}^{(\alpha)} & v_{33}^{(\alpha)} & 0 & 0 & 0 \\ 0 & 0 & 0 & v_{21}^{(\beta)} & v_{22}^{(\beta)} & v_{23}^{(\beta)} \\ 0 & 0 & 0 & v_{31}^{(\beta)} & v_{32}^{(\beta)} & v_{33}^{(\beta)} \end{array}\right]\Psi_R$$

represents an SVD with standard ordering of singular values. Denote by σ_j, $j = 1,\ldots,6$, the sorted singular values of $[B_1|A_{11}]$, and consider $V_{12} \in \mathbb{R}^{\overline{d}\times(\overline{q}+\overline{e})}$, $V_{13} \in \mathbb{R}^{\overline{d}\times(\overline{d}-\overline{e})}$, the blocks of V analogous to (2.3). If, e.g.,

$$\min\{\sigma_2^{(\alpha)}, \sigma_2^{(\beta)}\} > \max\{\sigma_3^{(\alpha)}, \sigma_3^{(\beta)}\},$$

then

$$\sigma_4 > \sigma_5 \geq \sigma_6, \quad \overline{q} = 0, \quad \overline{e} \in \{1, 2\}$$

and

$$[V_{12}, V_{13}] = \left[\begin{array}{c|c} v_{13}^{(\alpha)} & 0 \\ 0 & v_{13}^{(\beta)} \end{array}\right] \quad \text{or} \quad \left[\begin{array}{cc} v_{13}^{(\alpha)} & 0 \\ 0 & v_{13}^{(\beta)} \end{array}\right] \quad \text{or} \quad \left[\begin{array}{c|c} 0 & v_{13}^{(\alpha)} \\ v_{13}^{(\beta)} & 0 \end{array}\right],$$

depending on the relation between $\sigma_3^{(\alpha)}$ and $\sigma_3^{(\beta)}$ ($>$, $=$, $<$, respectively). The partitioning of $[V_{12}, V_{13}]$ is suggested by the vertical line. (If $\sigma_3^{(\alpha)} = \sigma_3^{(\beta)}$, then the matrix V_{13} has no columns.) The matrix $[V_{12}, V_{13}]$ is in any case square nonsingular, and this core problem belongs to the set $\mathcal{F}_1$ having the unique TLS solution. ∎

Example 4.2. Consider $[B_1|A_{11}]$ as in Example 4.1. If, e.g.,

$$\sigma_3^{(\alpha)} = \sigma_2^{(\beta)},$$

then

$$\sigma_3 > \sigma_4 = \sigma_5 > \sigma_6, \quad \overline{q} = 1, \quad \overline{e} = 1, \quad \text{and} \quad [V_{12}, V_{13}] = \left[\begin{array}{cc|c} v_{13}^{(\alpha)} & 0 & 0 \\ 0 & v_{12}^{(\beta)} & v_{13}^{(\beta)} \end{array}\right].$$

Since $\operatorname{rank}(V_{12}) = 2 > \overline{e}$ and $\operatorname{rank}(V_{13}) = 1 = \overline{d} - \overline{e}$, this core problem belongs to the set $\mathcal{F}_2$ having a nonunique TLS solution. ∎

Example 4.3. Consider $[B_1|A_{11}]$ as in Example 4.1. If, e.g.,

$$\sigma_3^{(\alpha)} > \sigma_2^{(\beta)},$$

then

$$\sigma_4 > \sigma_5 > \sigma_6 \quad \text{and} \quad [V_{12}, V_{13}] = \left[\begin{array}{c|c} 0 & 0 \\ v_{12}^{(\beta)} & v_{13}^{(\beta)} \end{array}\right].$$

The matrix $[V_{12}, V_{13}]$ is rank deficient, and this core problem belongs to the set $\mathcal{S}$ having no TLS solution. ∎

Now we present an example in the set $\mathcal{F}_3$.

Example 4.4. Consider a *reducible core problem* containing all three subproblems (4.1) as components, i.e.,

$$[B_1|A_{11}] \equiv \left[\begin{array}{ccc|ccc} b_1^{(\alpha)} & 0 & 0 & A_{11}^{(\alpha)} & 0 & 0 \\ 0 & b_1^{(\beta)} & 0 & 0 & A_{11}^{(\beta)} & 0 \\ 0 & 0 & b_1^{(\gamma)} & 0 & 0 & A_{11}^{(\gamma)} \end{array}\right] \in \mathbb{R}^{\overline{m} \times (\overline{n}+\overline{d})},$$

and $\overline{m} = 9$, $\overline{n} = 6$, $\overline{d} = 3$. An SVD of $[B_1|A_{11}]$ with standard ordering of singular values can be written in a way analogous to that in Example 4.1. Denote by σ_j, $j = 1, \ldots, 9$, the sorted singular values of $[B_1|A_{11}]$. If, e.g.,

$$\sigma_3^{(\alpha)} = \sigma_3^{(\beta)} = \sigma_1^{(\gamma)},$$

then

$$\sigma_4 > \sigma_5 = \sigma_6 = \sigma_7 > \sigma_8 > \sigma_9, \quad \overline{q} = 2, \quad \overline{e} = 1,$$

and

$$[V_{12}, V_{13}] = \left[\begin{array}{ccc|cc} v_{13}^{(\alpha)} & 0 & 0 & 0 & 0 \\ 0 & v_{13}^{(\beta)} & 0 & 0 & 0 \\ 0 & 0 & v_{11}^{(\gamma)} & v_{12}^{(\gamma)} & v_{13}^{(\gamma)} \end{array}\right].$$

Since $\operatorname{rank}(V_{12}) = 3 > \overline{e}$ and $\operatorname{rank}(V_{13}) = 1 < \overline{d} - \overline{e}$, this core problem belongs to $\mathcal{F}_3$ having no TLS solution. ∎

4.2. Example of irreducible core problem in $\mathcal{F}_2$. Examples 4.1–4.4 may motivate a tempting conjecture, that each core problem which does not belong to $\mathcal{F}_1$ is reducible. The following irreducible problem stands as a counterexample.

Example 4.5.[3] Consider a problem $A_{11}X_1 \approx B_1$, $A_{11} \in \mathbb{R}^{4\times 2}$, $B_1 \in \mathbb{R}^{4\times 2}$ given by its SVD, with the partitioning (2.3) of V revealing that it belongs to $\mathcal{F}_2$,

$$[B_1|A_{11}] \equiv I_4 \left[\begin{array}{c|c|c|c} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \underbrace{\left(\frac{1}{4} \left[\begin{array}{c|c|c|c} -1 & -3 & \sqrt{3} & \sqrt{3} \\ 3 & -1 & \sqrt{3} & -\sqrt{3} \\ \hline \sqrt{3} & \sqrt{3} & 1 & 3 \\ \sqrt{3} & -\sqrt{3} & -3 & 1 \end{array}\right] \right)^T}_{V}. \tag{4.2}$$

Because $[B_1|A_{11}]$ is of full column rank, then A_{11} and B_1 are of full column rank, and the problem satisfies (CP1) and (CP2). Employing the SVD of $A_{11} = U'\Sigma'(V')^T$ reveals that (CP3) is also satisfied; i.e., (4.2) represents a *core problem*. It will be useful to consider its SVD form

$$(U')^T[B_1|A_{11}V'] \equiv \left[\begin{array}{cc|cc} b_{11} & b_{12} & \sigma_1 & 0 \\ b_{21} & b_{22} & 0 & \sigma_2 \\ b_{31} & b_{32} & 0 & 0 \\ b_{41} & b_{42} & 0 & 0 \end{array}\right], \quad \text{where } \sigma_{1,2} = \sqrt{4 \pm \frac{3}{8}\sqrt{5}}. \tag{4.3}$$

Now we show by contradiction that this core problem is irreducible.

Let $[b_1^{(\ell)}|A_{11}^{(\ell)}]$, $\ell = \alpha, \beta$, be the components of (4.2)–(4.3) in their SVD forms. They are *either* both incompatible,

$$[b_1^{(\alpha)}|A_{11}^{(\alpha)}] \equiv \left[\begin{array}{c|c} b_{1\alpha} & \sigma_1 \\ b_{3\alpha} & 0 \end{array}\right], \qquad [b_1^{(\beta)}|A_{11}^{(\beta)}] \equiv \left[\begin{array}{c|c} b_{2\beta} & \sigma_2 \\ b_{4\beta} & 0 \end{array}\right], \tag{4.4}$$

or one is incompatible and one is degenerate (with, e.g., $A_{11}^{(\beta)}$ having no columns),

$$[b_1^{(\alpha)}|A_{11}^{(\alpha)}] \equiv \left[\begin{array}{c|cc} b_{1\alpha} & \sigma_1 & 0 \\ b_{2\alpha} & 0 & \sigma_2 \\ b_{3\alpha} & 0 & 0 \end{array}\right], \qquad [b_1^{(\beta)}|A_{11}^{(\beta)}] \equiv [\, b_{4\beta} \,]. \tag{4.5}$$

Since the SVD form (4.3) of a reducible core problem can be composed of SVD forms of its components (4.4) or (4.5), there exist orthogonal matrices G_L, $G_R \in \mathbb{R}^{2\times 2}$ (see (2.6)) such that

$$\left[\begin{array}{cc} I_2 & 0 \\ 0 & G_L^T \end{array}\right] \left[\begin{array}{cc} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{array}\right] G_R \quad \text{is equal to} \quad \left[\begin{array}{cc} b_{1\alpha} & 0 \\ 0 & b_{2\beta} \\ b_{3\alpha} & 0 \\ 0 & b_{4\beta} \end{array}\right] \text{ or } \left[\begin{array}{cc} b_{1\alpha} & 0 \\ b_{2\alpha} & 0 \\ b_{3\alpha} & 0 \\ 0 & b_{4\beta} \end{array}\right]. \tag{4.6}$$

In both cases,

$$\left[\begin{array}{cc} b_{31} & b_{32} \\ b_{41} & b_{42} \end{array}\right] = G_L \left[\begin{array}{cc} b_{3\alpha} & 0 \\ 0 & b_{4\beta} \end{array}\right] G_R^T \tag{4.7}$$

[3] We include MATLAB scripts `exa45sym.m` [local/web 5.66KB] and `exa45num.m` [local/web 5.56KB] in supplementary materials, so one can easily follow and verify the subsequent steps by symbolic (using the Symbolic Math Toolbox) and numeric calculations, respectively. The same problem was used in [4, eq. (4.8)], however in a different context.

represents an SVD (up to possibly negative signs of $b_{3\alpha}$ or $b_{4\beta}$). Since for the considered example $|b_{3\alpha}| \neq |b_{4\beta}|$, the decomposition in (4.7) is unique, fully determining the matrices G_L, G_R (up to signs of their columns). Computation of G_L and G_R for (4.2), followed by the transformation (4.6), however, does not yield either of the two patterns of nonzero entries in the right-hand-side matrix, nor a permutation of them, contradicting the assumption of reducibility. ∎

The question of whether there exist irreducible core problems also in sets $\mathcal{F}_3$ and $\mathcal{S}$, i.e., having no TLS solution, remains open.

5. Outputs of the classical TLS algorithm. The classical TLS algorithm [12, pp. 87–90] represents a powerful approach for solving TLS problems (1.1). In exact arithmetic, it returns for any input data $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times d}$ an output matrix

$$\mathtt{X} \leftarrow \mathtt{class\_TLS\_alg}(A, B), \qquad \mathtt{X} \in \mathbb{R}^{n\times d}, \tag{5.1}$$

constructed as follows. Consider the SVD (2.1). Find the *smallest* κ, $n \geq \kappa \geq 0$, such that

$$\Sigma = [\underbrace{\widetilde{\Sigma}_1}_{n-\kappa} \underbrace{\widetilde{\Sigma}_2}_{\kappa+d}], \qquad \sigma_{\min}(\widetilde{\Sigma}_1) \equiv \sigma_{n-\kappa} \ > \ \sigma_{n-\kappa+1} \equiv \sigma_{\max}(\widetilde{\Sigma}_2), \tag{5.2}$$

$$\text{and } V = \underbrace{\begin{bmatrix} \widetilde{V}_{11} & \widetilde{V}_{12} \\ \widetilde{V}_{21} & \widetilde{V}_{22} \end{bmatrix}}_{n-\kappa \quad \kappa+d} \begin{matrix} \}d \\ \}n \end{matrix}, \quad \operatorname{rank}(\widetilde{V}_{12}) = d. \tag{5.3}$$

(If $\kappa = n$, then $\sigma_{n-\kappa}$, $\widetilde{\Sigma}_1$, $\widetilde{V}_{11}$, and $\widetilde{V}_{21}$ do not exist.) The output (5.1) is then

$$\mathtt{X} = -\widetilde{V}_{22}\widetilde{V}_{12}^{\dagger}. \tag{5.4}$$

Recalling the classification in section 2.1, if $AX \approx B$ belongs to the sets $\mathcal{F}_1$, $\mathcal{F}_2$, or $\mathcal{F}_3$, then $\kappa = q$, and the partitioning (5.3) coincides with (2.3) such that $\widetilde{V}_{12} = [V_{12}, V_{13}]$; compare (2.2)–(2.3) with (5.2)–(5.3). If $AX \approx B$ belongs to the set $\mathcal{S}$, then $\kappa > q$; see also the discussion below Algorithm 1 in [4, p. 767]. Further, if $AX \approx B$ belongs to the set $\mathcal{F}_1$, then $\mathtt{X} = X_{\mathrm{TLS}}$ (compare (2.4) and (5.4)), which is not true for problems in the sets $\mathcal{F}_2$, $\mathcal{F}_3$, or $\mathcal{S}$. This allows us to extend Corollary 2.5 as follows.

COROLLARY 5.1. *Let $A_{11}X_1 \approx B_1$ be a core problem. Then the following assertions are equivalent:*

- *The problem $A_{11}X_1 \approx B_1$ has a unique TLS solution.*
- *The problem $A_{11}X_1 \approx B_1$ belongs to the set $\mathcal{F}_1$.*
- *The output of the classical TLS algorithm for the data A_{11}, B_1 represents a TLS solution.*

In order to study how core reduction influences the TLS algorithm, we now compare the outputs of the classical TLS algorithm for a problem $AX \approx B$ and its core problem, and for a reducible problem and its components.

5.1. Classical TLS algorithm and the core problem reduction. Consider a single right-hand side problem $Ax \approx b$ and its core problem

$$A_{11}x_1 \approx b_1, \qquad \text{where} \quad P^T[b|AQ] = \left[\begin{array}{c|c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right]. \tag{5.5}$$

Define $\mathtt{x} \leftarrow \mathtt{class\_TLS\_alg}\,(A, b)$ and $\mathtt{x}_1 \leftarrow \mathtt{class\_TLS\_alg}\,(A_{11}, b_1)$. In [7], it was shown that

$$\mathtt{x} = Q \begin{bmatrix} \mathtt{x}_1 \\ 0 \end{bmatrix}; \tag{5.6}$$

i.e., the algorithm gives the same output (up to the orthogonal transformation) for the original problem and for the core problem. The following theorem extends this result to $d > 1$.

THEOREM 5.2. *Let $AX \approx B$ be an approximation problem* (1.1) *and*

$$A_{11}X_1 \approx B_1, \qquad P^T[BR|AQ] = \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \tag{5.7}$$

be a core problem within $AX \approx B$. Let

$$\mathtt{X} \leftarrow \mathtt{class\_TLS\_alg}\,(A, B), \qquad \mathtt{X}_1 \leftarrow \mathtt{class\_TLS\_alg}\,(A_{11}, B_1) \tag{5.8}$$

be the outputs of the classical TLS algorithm. Then

$$\mathtt{X} = Q \begin{bmatrix} \mathtt{X}_1 & 0 \\ 0 & 0 \end{bmatrix} R^T; \tag{5.9}$$

i.e., the classical TLS algorithm gives the same output (up to the orthogonal transformation) for the original problem and for the core problem.

Notice that there is a clear resemblance between (5.9) and the basic solution equality (2.9). However, since the TLS algorithm returns the TLS solution only for problems from the set $\mathcal{F}_1$, neither of the matrices $\mathtt{X}$, $\mathtt{X}_1$ has to represent a TLS solution. The theorem states an important consistency result, that the original problem and the core problem are solved in the same sense by the TLS algorithm.

Proof. The SVD (2.1) of $[B|A] \in \mathbb{R}^{m\times(d+n)}$ with the partitioning (5.2)–(5.3) gives

$$P^T[BR|AQ] = (P^TU)[\widetilde{\Sigma}_1, \widetilde{\Sigma}_2] \begin{bmatrix} R^T\widetilde{V}_{11} & R^T\widetilde{V}_{12} \\ Q^T\widetilde{V}_{21} & Q^T\widetilde{V}_{22} \end{bmatrix}^T. \tag{5.10}$$

Consider further the SVD of $[B_1|A_{11}]$ with the partitioning analogous to (5.2)–(5.3),

$$[B_1|A_{11}] = U^{(1)}\Sigma^{(1)}(V^{(1)})^T = U^{(1)}[\widetilde{\Sigma}_1^{(1)}, \widetilde{\Sigma}_2^{(1)}] \begin{bmatrix} \widetilde{V}_{11}^{(1)} & \widetilde{V}_{12}^{(1)} \\ \widetilde{V}_{21}^{(1)} & \widetilde{V}_{22}^{(1)} \end{bmatrix}^T, \tag{5.11}$$

described by the smallest $\overline{\kappa}$, $\overline{n} \geq \overline{\kappa} \geq 0$, for which

$$\sigma_{\overline{n}-\overline{\kappa}}([B_1|A_{11}]) \equiv \sigma_{\min}(\widetilde{\Sigma}_1^{(1)}) \;>\; \sigma_{\max}(\widetilde{\Sigma}_2^{(1)}) \equiv \sigma_{\overline{n}-\overline{\kappa}+1}([B_1|A_{11}]) \tag{5.12}$$

and $\widetilde{V}_{12}^{(1)} \in \mathbb{R}^{\overline{d}\times(\overline{d}+\overline{\kappa})}$ is of full row rank. Then

$$\mathtt{X}_1 = -\,\widetilde{V}_{22}^{(1)}(\widetilde{V}_{12}^{(1)})^\dagger$$

is the output of the algorithm for the core problem. Consider also the SVD of A_{22},

$$A_{22} = U^{(2)}\Sigma^{(2)}(V^{(2)})^T = U^{(2)}[\widetilde{\Sigma}_1^{(2)}, \widetilde{\Sigma}_2^{(2)}]\,[\widetilde{V}_1^{(2)}, \widetilde{V}_2^{(2)}]^T,$$

where $U^{(2)}$, $V^{(2)}$ are square orthogonal; the partitioning of $\Sigma^{(2)}$ satisfies

$$\sigma_{\min}(\widetilde{\Sigma}_1^{(2)}) \;>\; \sigma_{\max}(\widetilde{\Sigma}_2^{(1)}) \;\geq\; \sigma_{\max}(\widetilde{\Sigma}_2^{(2)}); \tag{5.13}$$

and the partitioning of $V^{(2)}$ is conformal with the partitioning of $\Sigma^{(2)}$. Then

$$\left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \tag{5.14}$$

$$= \begin{bmatrix} U^{(1)} & 0 \\ 0 & U^{(2)} \end{bmatrix} \left[\begin{array}{cc|ccc} \widetilde{\Sigma}_1^{(1)} & 0 & \widetilde{\Sigma}_2^{(1)} & 0 & 0 \\ 0 & \widetilde{\Sigma}_1^{(2)} & 0 & \widetilde{\Sigma}_2^{(2)} & 0 \end{array}\right] \left[\begin{array}{cc|ccc} \widetilde{V}_{11}^{(1)} & 0 & \widetilde{V}_{12}^{(1)} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{d-\overline{d}} \\ \hline \widetilde{V}_{21}^{(1)} & 0 & \widetilde{V}_{22}^{(1)} & 0 & 0 \\ 0 & \widetilde{V}_1^{(2)} & 0 & \widetilde{V}_2^{(2)} & 0 \end{array}\right]^T.$$

From (5.12) and (5.13) it immediately follows that

$$\sigma_{\min}\left(\begin{bmatrix} \widetilde{\Sigma}_1^{(1)} & 0 \\ 0 & \widetilde{\Sigma}_1^{(2)} \end{bmatrix}\right) \;>\; \sigma_{\max}\left(\begin{bmatrix} \widetilde{\Sigma}_2^{(1)} & 0 & 0 \\ 0 & \widetilde{\Sigma}_2^{(2)} & 0 \end{bmatrix}\right) \equiv \sigma_{\max}(\widetilde{\Sigma}_2^{(1)}). \tag{5.15}$$

Since (5.10) and (5.14) represent two SVDs of the same matrix, and since (5.15) holds, there exist permutation matrices Ψ_L, Ψ_{R_1}, Ψ_{R_2} such that

$$[\widetilde{\Sigma}_1, \widetilde{\Sigma}_2] = \Psi_L^T \left[\begin{array}{c|c} \begin{bmatrix} \widetilde{\Sigma}_1^{(1)} & 0 \\ 0 & \widetilde{\Sigma}_1^{(2)} \end{bmatrix} \Psi_{R_1} & \begin{bmatrix} \widetilde{\Sigma}_2^{(1)} & 0 & 0 \\ 0 & \widetilde{\Sigma}_2^{(2)} & 0 \end{bmatrix} \Psi_{R_2} \end{array}\right]. \tag{5.16}$$

Consequently

$$\begin{bmatrix} R^T\widetilde{V}_{11} & R^T\widetilde{V}_{12} \\ Q^T\widetilde{V}_{21} & Q^T\widetilde{V}_{22} \end{bmatrix} = \left[\begin{array}{cc|ccc} \widetilde{V}_{11}^{(1)} & 0 & \widetilde{V}_{12}^{(1)} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{d-\overline{d}} \\ \hline \widetilde{V}_{21}^{(1)} & 0 & \widetilde{V}_{22}^{(1)} & 0 & 0 \\ 0 & \widetilde{V}_1^{(2)} & 0 & \widetilde{V}_2^{(2)} & 0 \end{array}\right] \begin{bmatrix} \Psi_{R_1} & 0 \\ 0 & \Psi_{R_2} \end{bmatrix}, \tag{5.17}$$

where $\widetilde{V}_{12}^{(1)} \in \mathbb{R}^{\overline{d}\times(\overline{d}+\overline{\kappa})}$ is of full row rank $\overline{d}$, and thus

$$\operatorname{rank}\left(\begin{bmatrix} \widetilde{V}_{12}^{(1)} & 0 & 0 \\ 0 & 0 & I_{d-\overline{d}} \end{bmatrix} \Psi_{R_2}\right) = d. \tag{5.18}$$

The partitioning of the matrices on the right-hand sides of (5.16)–(5.17) has properties (5.15) and (5.18) analogous to (5.2)–(5.3).

Recall that $\overline{\kappa}$ is the smallest integer in the partitioning (5.11) of $[B_1|A_{11}]$, and the largest singular value on the right-hand side of the inequality (5.15) is originated in $[B_1|A_{11}]$. Consequently, the blocks suggested by the lines in the partitioning (5.16)–(5.17) are of the same dimensions as the blocks in (5.2)–(5.3). Then (5.17) gives

$$\widetilde{V}_{12} = R\begin{bmatrix} \widetilde{V}_{12}^{(1)} & 0 & 0 \\ 0 & 0 & I_{d-\overline{d}} \end{bmatrix} \Psi_{R_2}, \qquad \widetilde{V}_{22} = Q\begin{bmatrix} \widetilde{V}_{22}^{(1)} & 0 & 0 \\ 0 & \widetilde{V}_2^{(2)} & 0 \end{bmatrix} \Psi_{R_2},$$

and the output $\mathtt{X}$ of the algorithm for the original data is given by

$$\begin{aligned} \mathtt{X} = -\widetilde{V}_{22}\widetilde{V}_{12}^{\dagger} &= -\left(Q\begin{bmatrix} \widetilde{V}_{22}^{(1)} & 0 & 0 \\ 0 & \widetilde{V}^{(2)} & 0 \end{bmatrix} \Psi_{R_2}\right) \left(\Psi_{R_2}^T \begin{bmatrix} (\widetilde{V}_{12}^{(1)})^{\dagger} & 0 \\ 0 & 0 \\ 0 & I_{d-\overline{d}} \end{bmatrix} R^T\right) \\ &= -Q\begin{bmatrix} \widetilde{V}_{22}^{(1)}(\widetilde{V}_{12}^{(1)})^{\dagger} & 0 \\ 0 & 0 \end{bmatrix} R^T = Q\begin{bmatrix} \mathtt{X}_1 & 0 \\ 0 & 0 \end{bmatrix} R^T. \end{aligned}$$

□

5.2. Classical TLS algorithm and reducible core problems. Now we study the outputs of the TLS algorithm for a reducible core problem and its components. Consider two core problems,

$$A_{11}^{(\ell)} X_1^{(\ell)} \approx B_1^{(\ell)}, \qquad A_{11}^{(\ell)} \in \mathbb{R}^{\overline{m}_\ell \times \overline{n}_\ell}, \quad B_1^{(\ell)} \in \mathbb{R}^{\overline{m}_\ell \times \overline{d}_\ell}, \qquad \ell = \alpha, \beta, \tag{5.19}$$

and their SVDs with the block partitionings (5.2)–(5.3),

$$[B_1^{(\ell)} | A_{11}^{(\ell)}] = U^{(\ell)} \Sigma^{(\ell)} (V^{(\ell)})^T = U^{(\ell)} \left[\widetilde{\Sigma}_1^{(\ell)}, \widetilde{\Sigma}_2^{(\ell)} \right] \begin{bmatrix} \widetilde{V}_{11}^{(\ell)} & \widetilde{V}_{12}^{(\ell)} \\ \widetilde{V}_{21}^{(\ell)} & \widetilde{V}_{22}^{(\ell)} \end{bmatrix}^T,$$

and denote

$$\mathtt{X}_1^{(\ell)} \leftarrow \mathtt{class\_TLS\_alg}\,(A_{11}^{(\ell)}, B_1^{(\ell)}), \quad \text{i.e.,} \quad \mathtt{X}_1^{(\ell)} = -\,\widetilde{V}_{22}^{(\ell)} (\widetilde{V}_{12}^{(\ell)})^\dagger.$$

Further consider a reducible core problem and its SVD,

$$[B_1 | A_{11}] \equiv \left[\begin{array}{cc|cc} B_1^{(\alpha)} & 0 & A_{11}^{(\alpha)} & 0 \\ 0 & B_1^{(\beta)} & 0 & A_{11}^{(\beta)} \end{array}\right] = U^{(1)} \Sigma^{(1)} (V^{(1)})^T, \tag{5.20}$$

and denote

$$\mathtt{X}_1 \leftarrow \mathtt{class\_TLS\_alg}\,(A_{11}, B_1).$$

We focus on two particular cases of relations among the singular values of the individual components in (5.20).

Example 5.3. If, e.g.,

$$\sigma_{\min}(\widetilde{\Sigma}_1^{(\alpha)}) > \sigma_{\max}(\widetilde{\Sigma}_1^{(\beta)}), \quad \sigma_{\min}(\widetilde{\Sigma}_1^{(\beta)}) > \sigma_{\max}(\widetilde{\Sigma}_2^{(\alpha)}), \quad \sigma_{\min}(\widetilde{\Sigma}_2^{(\alpha)}) > \sigma_{\max}(\widetilde{\Sigma}_2^{(\beta)}),$$

then, for some permutation matrix Ψ_L,

$$\Sigma^{(1)} = \Psi_L^T \left[\begin{array}{cc|cc} \widetilde{\Sigma}_1^{(\alpha)} & 0 & \widetilde{\Sigma}_2^{(\alpha)} & 0 \\ 0 & \widetilde{\Sigma}_1^{(\beta)} & 0 & \widetilde{\Sigma}_2^{(\beta)} \end{array}\right], \qquad V^{(1)} = \left[\begin{array}{cc|cc} \widetilde{V}_{11}^{(\alpha)} & 0 & \widetilde{V}_{12}^{(\alpha)} & 0 \\ 0 & \widetilde{V}_{11}^{(\beta)} & 0 & \widetilde{V}_{12}^{(\beta)} \\ \hline \widetilde{V}_{21}^{(\alpha)} & 0 & \widetilde{V}_{22}^{(\alpha)} & 0 \\ 0 & \widetilde{V}_{21}^{(\beta)} & 0 & \widetilde{V}_{22}^{(\beta)} \end{array}\right],$$

with the partitioning (5.2)–(5.3) suggested by the lines. Then

$$\mathtt{X}_1 = -\begin{bmatrix} \widetilde{V}_{22}^{(\alpha)} & 0 \\ 0 & \widetilde{V}_{22}^{(\beta)} \end{bmatrix} \begin{bmatrix} \widetilde{V}_{12}^{(\alpha)} & 0 \\ 0 & \widetilde{V}_{12}^{(\beta)} \end{bmatrix}^\dagger = \begin{bmatrix} \mathtt{X}_1^{(\alpha)} & 0 \\ 0 & \mathtt{X}_1^{(\beta)} \end{bmatrix}.$$

The output of the classical TLS algorithm for this reducible problem is a direct sum of outputs for the individual components. □

Example 5.4. If, e.g.,

$$\sigma_{\min}([B_1^{(\alpha)} | A_{11}^{(\alpha)}]) \equiv \sigma_{\min}(\widetilde{\Sigma}_2^{(\alpha)}) \;>\; \sigma_{\max}(\widetilde{\Sigma}_1^{(\beta)}) \equiv \left\| \left[B_1^{(\beta)} | A_{11}^{(\beta)} \right] \right\|_2, \tag{5.21}$$

then, for some permutation matrix Ψ_L,

$$\Sigma^{(1)} = \Psi_L^T \left[\begin{array}{c|ccc} \widetilde{\Sigma}_1^{(\alpha)} & \widetilde{\Sigma}_2^{(\alpha)} & 0 & 0 \\ 0 & 0 & \widetilde{\Sigma}_1^{(\beta)} & \widetilde{\Sigma}_2^{(\beta)} \end{array}\right], \qquad V^{(1)} = \left[\begin{array}{c|ccc} \widetilde{V}_{11}^{(\alpha)} & \widetilde{V}_{12}^{(\alpha)} & 0 & 0 \\ 0 & 0 & \widetilde{V}_{11}^{(\beta)} & \widetilde{V}_{12}^{(\beta)} \\ \hline \widetilde{V}_{21}^{(\alpha)} & \widetilde{V}_{22}^{(\alpha)} & 0 & 0 \\ 0 & 0 & \widetilde{V}_{21}^{(\beta)} & \widetilde{V}_{22}^{(\beta)} \end{array}\right],$$

with the partitioning (5.2)–(5.3) suggested by the lines. Then

$$\mathtt{X}_1 = -\begin{bmatrix} V_{22}^{(\alpha)} & 0 & 0 \\ 0 & V_{21}^{(\beta)} & V_{22}^{(\beta)} \end{bmatrix}\begin{bmatrix} V_{12}^{(\alpha)} & 0 & 0 \\ 0 & V_{11}^{(\beta)} & V_{12}^{(\beta)} \end{bmatrix}^{\dagger} = \begin{bmatrix} \mathtt{X}_1^{(\alpha)} & 0 \\ 0 & 0 \end{bmatrix}.$$

The output $\mathtt{X}_1^{(\beta)}$ corresponding to the second component is not present in the output of the algorithm for this reducible problem. □

Consequently, the output of the classical TLS algorithm for a reducible core problem depends not only on the individual components, but also on the *relations between* them. Neglecting the component with a smaller 2-norm in Example 5.4 can be seen as a sort of *regularization*.

Remark 5.5. Note that we have not used any of the particular properties (CPℓ), $\ell = 1, \ldots, 7$, of core problems in sections 5.1 and 5.2. Thus, Theorem 5.2 can be extended to *any subproblem* $A_{11}X_1 \approx B_1$ of the given problem $AX \approx B$ obtained by an orthogonal transformation of the form (5.7). The output of the classical TLS algorithm is for $AX \approx B$ fully determined by any $A_{11}X_1 \approx B_1$ obtained as in (5.7), not necessarily the core problem. Similarly, Examples 5.3 and 5.4 can be generalized to any approximation problems (5.19).

6. Conclusions. We have shown that core problems with multiple right-hand sides may have a specific internal structure, which has led us to introduce the so-called *reducible and irreducible core problems*. We have proved that, contrary to the case with a single right-hand side, a core problem with multiple right-hand sides may belong to any of the classification sets $\mathcal{F}_1$ (if and only if it has a unique TLS solution), $\mathcal{F}_2$ (having infinitely many TLS solutions), $\mathcal{F}_3$, or $\mathcal{S}$ (having no TLS solution). We have proved that the output of the classical TLS algorithm stays unchanged under the core problem reduction (up to an orthogonal transformation), which is an important consistency result fully generalizing the result obtained previously for problems with a single right-hand side. We have shown that the output of the TLS algorithm for a reducible core problem depends on its components as well as on the relations between them. The question of detecting possible reducibility of a given problem remains open.

Acknowledgments. The authors wish to thank Zdeněk Strakoš and Sabine Van Huffel for valuable discussions about TLS problems. They also thank the anonymous referees for their useful comments.

REFERENCES

[1] Å. Björck, *A band-Lanczos algorithm for least squares and total least squares problems*, in Book of Abstracts of the 4th Total Least Squares and Errors-in-Variables Modeling Workshop, Leuven, Katholieke Universiteit Leuven, Leuven, Belgium, 2006, pp. 22–23; http://homepages.vub.ac.be/~imarkovs/workshop/program.pdf.

[2] Å. Björck, *Block Bidiagonal Decomposition and Least Squares Problems with Multiple Right-Hand Sides*, unpublished manuscript, Department of Mathematics, University of Linköping, 2008.

[3] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893, doi:10.1137/0717073.

[4] I. Hnětynková, M. Plešinger, D. M. Sima, Z. Strakoš, and S. Van Huffel, *The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 748–770, doi:10.1137/100813348.

[5] I. Hnětynková, M. Plešinger, and Z. Strakoš, *The core problem within linear approximation problem $AX \approx B$ with multiple right-hand sides*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 917–931, doi:10.1137/120884237.

[6] I. Hnětynková, M. Plešinger, and Z. Strakoš, *Band generalization of the Golub–Kahan bidiagonalization, generalized Jacobi matrices, and the core problem*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 417–434, doi:10.1137/140968914.

[7] C. C. Paige and Z. Strakoš, *Core problem in linear algebraic systems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 861–875, doi:10.1137/040616991.

[8] M. Plešinger, *The Total Least Squares Problem and Reduction of Data in $AX \approx B$*, Ph.D. thesis, Faculty of Mechatronics, Technical University of Liberec, Liberec, Czech Republic, 2008.

[9] D. M. Sima, *Regularization Techniques in Model Fitting and Parameter Estimation*, Ph.D. thesis, ESAT, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.

[10] S. Van Huffel, *Documented Fortran* 77 *Programs of the Extended Classical Total Least Squares Algorithm, the Partial Singular Value Decomposition Algorithm and the Partial Total Least Squares Algorithm*, Internal report ESAT-KUL 88/1, Katholieke Universiteit Leuven, Leuven, Belgium, 1988.

[11] S. Van Huffel, *The extended classical total least squares algorithm*, J. Comput. Appl. Math., 25 (1989), pp. 111–119, doi:10.1016/0377-0427(89)90080-0.

[12] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers in Appl. Math. 9, SIAM, Philadelphia, 1991, doi:10.1137/1.9781611971002.

[13] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 746–763, doi:10.1137/0613047.

[14] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math., 62 (1992), pp. 123–148, doi:10.1007/BF01396223.

Regularizační metody

A.5 Článek:[WoK] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *The regularizing effect of the Golub–Kahan iterative bidiagonalization and revealing the noise level in the data*, BIT Numerical Mathematics (ISSN 0006-3835, eISSN 1572-9125), Volume 49, Number 4 (2009), pp. 669–696 (28 pages). (http://link.springer.com/article/10.1007/s10543-009-0239-7)

Doložitelné citace (27): Odborné články zařazené do databáze Web-of-Knowledge:

- [WoK] J. Chung, P. Sternberg, C. Yang: *High-performance three-dimensional image reconstruction for molecular structure determination*, International Journal of High Performance Computing Applications, 24 (2) (2010), pp. 117–135. (http://hpc.sagepub.com/content/24/2/117.short)
- [WoK] R. A. Renaut, I. Hnětynková[AutoCit], J. Mead: *Regularization parameter estimation for large-scale Tikhonov regularization using a priori information*, Computational Statistics & Data Analysis, 54(12) (2010), pp. 3430–3445. (http://www.sciencedirect.com/science/article/pii/S0167947309002278)
- [WoK] F. S. Viloche Bazán, L. S. Borges: *GKB-FP: an algorithm for large-scale discrete ill-posed problems*, BIT Numerical Mathematics, 50(3) (2010), pp. 481–507. (http://link.springer.com/article/10.1007/s10543-010-0275-3)
- [WoK] Z. Strakoš[AutoCit]: *Gene H. Golub and Gérard Meurant: Matrices, Moments and Quadrature with Applications*, Foundations of Computational Mathematics, 11(2) (2011), pp. 241–255. (http://link.springer.com/article/10.1007/s10208-010-9082-0)
- [WoK] L. Eldén, V. Simoncini: *Solving ill-posed linear systems with GMRES and a singular preconditioner*, SIAM Journal on Matrix Analysis and Applications, 33(4) (2012), pp. 1369–1394. (http://epubs.siam.org/doi/abs/10.1137/110832793)
- [WoK] L. Reichel, G. Rodriguez: *Old and new parameter choice rules for discrete ill-posed problems*, Numerical Algorithms, 63(1) (2013), pp. 65–87. (http://link.springer.com/article/10.1007/s11075-012-9612-8)
- [WoK] J. L. Barlow: *Reorthogonalization for the Golub–Kahan–Lanczos bidiagonal reduction*, Numerische Mathematik, 124(2) (2013), pp. 237–278. (http://link.springer.com/article/10.1007/s00211-013-0518-8)
- [WoK] C. Zhao, T.-Z. Huang, X.-L. Zhao, L.-J. Deng: *Two new efficient iterative regularization methods for image restoration problems*, Abstract and Applied Analysis, 2013 (2013), Article Number 129652, 9 pages. (http://www.hindawi.com/journals/aaa/2013/129652)
- [WoK] G. Lee, J. L. Barlow: *Two projection methods for Regularized Total Least Squares approximation*, Linear Algebra and its Application, 461 (2014), pp. 18–41. (http://www.sciencedirect.com/science/article/pii/S0024379514004960)
- [WoK] M. E. Hochstenbach, L. Reichel, G. Rodrigues: *Regularization parameter determination for discrete ill-posed problems*, Journal of Computational and Applied Mathematics, 273 (2015), pp. 132–149. (http://www.sciencedirect.com/science/article/pii/S037704271400274X)
- [WoK] L. Reichel, X. Yu: *Tikhonov regularization via flexible Arnoldi reduction*, BIT Numerical Mathematics, 55(4) (2015), pp. 1145–1168. (http://link.springer.com/article/10.1007/s10543-014-0542-9)

- [WoK] J. Chung, M. E. Kilmer, D. P. O'Leary: *A framework for regularization via operator approximation*, SIAM Journal on Scientific Computing, 37(2) (2015), pp. B332–B359. (http://epubs.siam.org/doi/abs/10.1137/130945363)
- [WoK] J. Chung, K. Palmer: *A hybrid LSMR algorithm for large-scale Tikhonov regularization*, SIAM Journal on Scientific Computing, 37(5) (2015), pp. S562–S580. (http://epubs.siam.org/doi/abs/10.1137/140975024)
- [WoK] G. Landi, E. Loli Piccolomini, I. Tomba: *A stopping criterion for iterative regularization methods*, Applied Numerical Mathematics, 106 (2016), pp. 53–68. (http://www.sciencedirect.com/science/article/pii/S0168927416300368)
- [WoK] S. Gazzola, P. Novati: *Inheritance of the discrete Picard condition in Krylov subspace methods*, BIT Numerical Mathematics, 56(3) (2016), pp. 893–918. (http://link.springer.com/article/10.1007/s10543-015-0578-5)
- [(WoK)] J. Chung, L. Ruthotto: *Computational methods for image reconstruction*, NMR (Nuclear Magnetic Resonance) in Biomedicine. Available online, 13 pages. (http://onlinelibrary.wiley.com/doi/10.1002/nbm.3545/abstract)

Dizertační práce:

- [PhD-thesis] J. Chung: *Numerical Approaches for Large-Scale Ill-Posed Inverse Problems*, Emory University, Atlanta, GA, 2009. (http://www.math.vt.edu/people/jmchung/resources [...] /Chung_PhDDissertation.pdf)
- [PhD-thesis] J. Lampe: *Solving Regularized Total Least Squares Problems Based on Eigenproblems*, Technical University Hamburg-Harburg, 2010. (http://doku.b.tu-harburg.de/volltexte/2010/959/pdf [...] /Lampe_Joerg.pdf, http://d-nb.info/1009720546)
- [PhD-thesis] X. Bonnefond: *Contributions a la tomographie thermoacoustique. Modélisation et inversion*, Université Toulouse 3 Paul Sabatier, Toulouse, 2010. (http://thesesups.ups-tlse.fr/1155)
- [PhD-thesis] S. Gazzola: *Regularization techniques based on Krylov subspace methods for ill-posed linear system*, Università degli Studi di Padova, 2013. (http://www.math.unipd.it/~gazzola/gazzola_silvia_tesi.pdf)

Knihy:

- [Book WoK] P. C. Hansen: *Discrete Inverse Problems, Insight and Algorithms*, SIAM, Philadelphia, PA, 2010. (http://epubs.siam.org/doi/book/10.1137/1.9780898718836)
- [Book Scopus] J. Liesen, Z. Strakoš[AutoCit]: *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, Oxford, 2013. (https://global.oup.com/academic/product [...] /krylov-subspace-methods-9780199655410)
- [Book] Å. Björck: *Numerical Methods in Matrix Computations* (Chapter 4, Iterative methods), Springer Verlag, 2015. (http://link.springer.com/book/10.1007/978-3-319-05089-8)

Preprinty z arXivu:

- Preprint X. Bonnefond, P. Maréchal: *Lagrange duality for the Morozov principle*, arXiv:1402.2999, submitted 12 Feb 2014 (v1), 7 pages. (http://arxiv.org/abs/1402.2999)
- Preprint Y. Huang, Z. Jia: *Some results on the regularization of LSQR for large-scale discrete ill-posed problems*, arXiv:1503.01864, submitted 6 Mar 2015 (v1), 15 Jul 2015 (v2), 19 Jan 2016 (v3), 19 pages. (http://arxiv.org/abs/1503.01864)
- Preprint J. Chung, A. K. Saibaba: *Generalized hybrid iterative methods for large-scale Bayesian inverse problems*, arXiv:1607.03943, submitted 13 Jul 2016 (v1), 27 pages. (http://arxiv.org/abs/1607.03943)
- Preprint Z. Jia: *The regularization theory of the Krylov iterative solvers LSQR, CGLS, LSMR and CGME for linear discrete ill-posed problems*, arXiv:1608.05907, submitted 21 Aug 2016 (v1), 77 pages. (http://arxiv.org/abs/1608.05907)

BIT Numer Math (2009) 49: 669–696
DOI 10.1007/s10543-009-0239-7

BIT

The regularizing effect of the Golub-Kahan iterative bidiagonalization and revealing the noise level in the data

Iveta Hnětynková · Martin Plešinger · Zdeněk Strakoš

Received: 31 October 2008 / Accepted: 27 August 2009 / Published online: 22 September 2009

Abstract Regularization techniques based on the Golub-Kahan iterative bidiagonalization belong among popular approaches for solving large ill-posed problems. First, the original problem is projected onto a lower dimensional subspace using the bidiagonalization algorithm, which by itself represents a form of regularization by projection. The projected problem, however, inherits a part of the ill-posedness of the original problem, and therefore some form of inner regularization must be applied. Stopping criteria for the whole process are then based on the regularization of the projected (small) problem.

In this paper we consider an ill-posed problem with a noisy right-hand side (observation vector), where the noise level is unknown. We show how the information

Communicated by Lars Eldén.

The work of the first author was supported by the research project MSM0021620839 financed by MŠMT. The work of the second and the third author was supported by the GAAS grant IAA100300802, and by the Institutional Research Plan AV0Z10300504.

I. Hnětynková (✉) · M. Plešinger · Z. Strakoš
Institute of Computer Science, Academy of Sciences, Pod Vodárenskou věží 2, Prague 8, Czech Republic
e-mail: hnetynkova@cs.cas.cz

M. Plešinger
e-mail: martin.plesinger@tul.cz

Z. Strakoš
e-mail: strakos@cs.cas.cz

I. Hnětynková · Z. Strakoš
Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, Prague 8, Czech Republic

M. Plešinger
Faculty of Mechatronics, Technical University of Liberec, Studentská 2/1402, Liberec, Czech Republic

from the Golub-Kahan iterative bidiagonalization can be used for estimating the noise level. Such information can be useful for constructing efficient stopping criteria in solving ill-posed problems.

Keywords Ill-posed problems · Golub-Kahan iterative bidiagonalization · Lanczos tridiagonalization · Noise revealing

Mathematics Subject Classification (2000) 15A06 · 15A18 · 15A23 · 65F10 · 65F22

1 Introduction

Consider an ill-posed linear algebraic system with the right-hand side b contaminated by *white noise*

$$Ax \approx b, \qquad A \in \mathbb{R}^{n\times n}, \qquad b = b^{\text{exact}} + b^{\text{noise}} \in \mathbb{R}^n, \tag{1.1}$$

with A nonsingular and the goal to numerically approximate the exact solution

$$x \equiv x^{\text{exact}} = A^{-1}b^{\text{exact}}. \tag{1.2}$$

Linear approximation problems of this form arise in a broad class of applications. In many cases the matrix A represents a discretized smoothing operator with the singular values of A decaying gradually without a noticeable gap. Since A is ill-conditioned, the presence of the noise makes the naive solution $x^{\text{naive}} = A^{-1}b$ meaningless. Therefore it is necessary to use regularization techniques for finding an acceptable numerical approximation to (1.2) which reflects a sufficient amount of information contained in the data, while suppressing the devastating influence of the noise. In image processing A typically represents a discretized blurring operator. In other applications A might be of different origin. Throughout the paper we assume A square and nonsingular. The presented methods can be extended to a general rectangular case, which is also confirmed by numerical experiments. However, the analysis contains some subtle points which would further extend the length of the paper.

For the unknown noise component b^{noise} we assume

$$\|b^{\text{noise}}\| \ll \|b^{\text{exact}}\|, \tag{1.3}$$

where $\|v\|$ denotes the standard Euclidean norm of the vector v. We will further assume that multiplication of a vector v by A and A^T results in smoothing which reduces the relative sizes of the high frequency components of v. In particular, in comparison to v, the vectors A^TAv and AA^Tv have significantly reduced high frequency components.

The Golub-Kahan iterative bidiagonalization algorithm [14] is widely used for solving large ill-posed problems. In hybrid methods, see, e.g., [18, Chap. 6.7, pp. 162–164] or [10, 27, 28, 36], the outer bidiagonalization (which itself represents

a regularization of the original large problem by projection) is combined with an inner regularization of the projected small problem. The bidiagonalization is stopped when the regularized solution of the projected problem matches some selected stopping criteria. They are typically based, amongst others (see [1, 2]) on estimation of the L-curve and it's curvature [4–6], estimation of the distance between the exact and regularized solution [35], the discrepancy principle [32, 33], and cross validation methods [7, 15, 34]. These techniques have been studied and compared in the context of regularization, e.g., in [18, Chap. 7, pp. 175–208] and in [27].

It is worthwhile recalling that the regularization idea was mentioned, although not fully developed, by Golub and Kahan [14]. It was described in detail in relation to using the Golub-Kahan iterative bidiagonalization, under the name damped least squares, by Paige and Saunders in [42]; see also [41] or [48]. Paige and Saunders alluded to some older interesting references in [42]. From the more recent literature we mention [16], where hybrid methods based on bidiagonalization are described as least-squares projection methods, and [51], where bidiagonalization is used to compute low-rank approximations of large sparse matrices. Numerical stability of the bidiagonalization algorithm was studied and new stable variants have been proposed, e.g., in [3, 51], with a simplified analysis of [3] presented in [40].

A new contribution to the theoretical background for hybrid methods has recently been presented in [43–45]. In exact arithmetic, the bidiagonalization provides a fundamental decomposition of the matrix of the data $[b, A]$. When the bidiagonalization stops, it reveals the so called *core problem* represented by the computed bidiagonal matrix. The core problem is minimally dimensioned and it contains the necessary and sufficient information for solving the original orthogonally invariant linear approximation problem $Ax \approx b$. Whenever the bidiagonalization is stopped before reaching the core problem, it gives (in exact arithmetic) its leading left upper part. Consequently, the approximate solution computed at the given step is based on *information which is necessary for solving the original problem* and it is not influenced by any part of the redundant or irrelevant information. From the properties of the core problem it then follows (see [23, 45]) that further possible steps can be considered a refinement of the current approximation. Ill-posed problems from the core problem point of view were studied in [49, 50]. It should be mentioned that the application of the core problem theory to results of finite precision computations needs further investigation.

In this paper we focus on the Golub-Kahan iterative bidiagonalization and investigate how the noise contained in the right-hand side b is propagated to the projected problem. We demonstrate under the given assumptions that *the unknown noise level* can be determined from the information available during the bidiagonalization process. The knowledge of the noise level can further be used in construction of stopping criteria for hybrid methods.

Similar ideas are used in [20, 46, 47] for selection of a value of the regularization parameter for which the residual vector changes from being dominated by the remaining signal to being white-noise like. This leads to a parameter-choice method based on Fourier analysis of residual vectors. In [25] regularization properties of iterative methods GMRES, MINRES, RRGMRES, MR-II and CGLS are studied. It is shown that MINRES, MR-II and CGLS filter out large singular value decomposition (SVD) components of the residual, but this is not always true for GMRES and

RRGMRES, where the SVD components are mixed [25, Sect. 3]. The noise propagation to reconstructed images computed by regularizing iterations is further studied in [19]. Our approach, which uses information from the iterative bidiagonalization algorithm for estimating the level of the noise in the data, offers another view on the noise propagation studied in [19, 20, 25, 46, 47].

The paper is organized as follows. Section 2 gives a brief recollection of the Golub-Kahan iterative bidiagonalization, its relationship to the Lanczos tridiagonalization and to approximation of the distribution function in the corresponding Riemann-Stieltjes integral. Section 3 describes propagation of the noise in the bidiagonalization vectors. Section 4 shows how to estimate at a negligible cost the unknown noise level in the original data. This can lead to construction of stopping criteria for the bidiagonalization process as well as to construction of the regularized solution using many different approaches which can be applied in subsequent steps, cf. [27, Sect. 3.2]. It is important to emphasize that the first four sections of our paper deal with *mathematical properties* of the problem and of the methods, i.e., they assume *exact arithmetic*. Therefore, unless specified otherwise, the presented experiments were performed with double reorthogonalization of the computed sequences of the bidiagonalization vectors, which ensures preserving orthogonality close to machine precision. Up to now, effects of rounding errors in solving ill-posed problems were not, to our knowledge, thoroughly investigated. In the references known to us, loss of orthogonality in solving ill-posed problems is not considered a fundamental issue although there are papers that acknowledge that it can be a potential problem. If the loss of orthogonality occurs, then the reader is referred to some form of reorthogonalization, with no further investigation of the differences between the non-reorthogonalized and reorthogonalized computations. In many cases the implementation details are missing, and it is unclear whether the experiments used reorthogonalization or not. We will show in Sect. 5 that effects of rounding errors in solving ill-posed problems can be substantial and the matter should be investigated further. Concluding remarks summarize the main ideas and formulate open questions.

2 Golub-Kahan iterative bidiagonalization

Given the initial vectors $w_0 \equiv 0, s_1 \equiv b/\beta_1$, where $\beta_1 \equiv \|b\| \neq 0$, the Golub-Kahan iterative bidiagonalization computes for $j = 1, 2, \ldots$

$$\begin{aligned} \alpha_j w_j &= A^T s_j - \beta_j w_{j-1}, \qquad \|w_j\| = 1, \\ \beta_{j+1} s_{j+1} &= A w_j - \alpha_j s_j, \qquad \|s_{j+1}\| = 1 \end{aligned} \tag{2.1}$$

until $\alpha_j = 0$ or $\beta_{j+1} = 0$, or until the process is stopped by reaching the dimensionality of the problem. Let $S_k = [s_1, \ldots, s_k]$ and $W_k = [w_1, \ldots, w_k]$ be the matrices with the left and right bidiagonalization vectors as their (orthonormal) columns, and

$$L_k \equiv \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{bmatrix}, \qquad L_{k+} \equiv \begin{bmatrix} L_k \\ \beta_{k+1} e_k^T \end{bmatrix}, \tag{2.2}$$

where e_k is the kth vector of the standard Euclidean basis. Then the first k steps of the Golub-Kahan iterative bidiagonalization can be written in the matrix form as

$$A^T S_k = W_k L_k^T, \qquad A W_k = [S_k, s_{k+1}] L_{k+}, \tag{2.3}$$

see [14, 38].

This algorithm is closely related to the Lanczos tridiagonalization of a symmetric matrix [29, 30]. The Lanczos tridiagonalization of the matrix AA^T with the starting vector $s_1 = b/\beta_1$, $\beta_1 = \|b\|$, yields in k steps the symmetric tridiagonal matrix T_k such that

$$AA^T S_k = S_k T_k + \alpha_k \beta_{k+1} s_{k+1} e_k^T, \tag{2.4}$$

and

$$T_k = L_k L_k^T = \begin{bmatrix} \alpha_1^2 & \alpha_1\beta_2 & & \\ \alpha_1\beta_2 & \alpha_2^2+\beta_2^2 & \ddots & \\ & \ddots & \ddots & \alpha_{k-1}\beta_k \\ & & \alpha_{k-1}\beta_k & \alpha_k^2+\beta_k^2 \end{bmatrix},$$

i.e. the matrix L_k from the Golub-Kahan iterative bidiagonalization represents a Cholesky factor of the matrix T_k. For a more detailed description of the outlined relationship we refer to [23] and to the literature given there.

Using the results from [22, Sect. 14], the Lanczos tridiagonalization of a given matrix B generates at each step k a non-decreasing piecewise constant distribution function $\omega^{(k)}$, with the nodes being the (distinct) eigenvalues of the Lanczos matrix T_k and the weights $\omega_j^{(k)}$ being the squared first components of the corresponding normalized eigenvectors. The distribution functions $\omega^{(k)}$, $k = 1, 2, \ldots$ represent approximations to the distribution function ω, a non-decreasing piecewise constant function with the nodes being the distinct eigenvalues $\lambda_1, \ldots, \lambda_t$ of the matrix B and the weights ω_j being the squared components of the normalized starting vector in the direction of the jth invariant subspace of B, $j = 1, \ldots, t$, for more details see, e.g., [31, Sect. 2.2], [11].

Consider the singular value decomposition (SVD) of the bidiagonal matrix of the projected problem

$$L_k = P_k \Theta_k Q_k^T, \tag{2.5}$$

where $P_k^{-1} = P_k^T$, $Q_k^{-1} = Q_k^T$, and Θ_k is a diagonal matrix with singular values $\theta_1^{(k)}, \ldots, \theta_k^{(k)}$ of L_k on its diagonal ordered in *increasing order*. Please note that here the increasing order of the singular values follows the standard ordering of the Ritz values in the Lanczos method. From (2.4) it follows that if $B = AA^T$ and the Lanczos tridiagonalization starts with the vector $s_1 = b/\beta_1$, $\beta_1 = \|b\|$, then

$$T_k \equiv L_k L_k^T = P_k \Theta_k^2 P_k^T$$

is the spectral decomposition of T_k, $(\theta_\ell^{(k)})^2$ are its eigenvalues and $p_\ell^{(k)} \equiv P_k e_\ell$ its eigenvectors, $\ell = 1, \ldots, k$. Furthermore, consider the SVD of A,

$$A = U \Sigma V^T = \sum_{j=1}^{n} \sigma_j u_j v_j^T, \tag{2.6}$$

where $U = [u_1, \ldots, u_n]$, $V = [v_1, \ldots, v_n]$, $U^{-1} = U^T$, $V^{-1} = V^T$, and Σ is a diagonal matrix with the singular values $\sigma_1, \ldots, \sigma_n$ of A on its diagonal ordered in nonincreasing order, $\sigma_n > 0$. Then

$$AA^T = U \Sigma^2 U^T$$

is the spectral decomposition of the matrix AA^T, σ_j^2 are its nonzero eigenvalues and u_j the corresponding eigenvectors, $j = 1, \ldots, n$. Summarizing, the Lanczos tridiagonalization (2.4) generates at each step k the distribution function $\omega^{(k)}$ with the nodes $(\theta_\ell^{(k)})^2$ and the weights $|(p_\ell^{(k)}, e_1)|^2$ that approximates the distribution function ω with the nodes $\sigma_n^2, \ldots, \sigma_1^2$ and the weights $|(b/\beta_1, u_n)|^2, \ldots, |(b/\beta_1, u_1)|^2$.

The fact that the weights of ω are determined by the projections of the normalized right-hand side vector b onto the left singular subspaces of A is especially important. As we will see in Sect. 3, until the level determined by the noise is reached, the absolute value of each of these projections is related to the size of the corresponding singular value. This is because of the discrete Picard condition (see Sect. 3).

Mathematically, the bidiagonalization algorithm yields two sequences of subproblems. Consider an approximation to the solution of (1.1) in the subspace generated by the vectors $w_1, \ldots, w_k$, i.e. $x_k = W_k y_k$, so that in seeking $AW_k y_k \approx b$,

$$r_k = b - AW_k y_k = S_{k+1}(e_1 \beta_1 - L_{k+} y_k) = S_k (e_1 \beta_1 - L_k y_k) - s_{k+1} \beta_{k+1} e_k^T y_k.$$

If the approximation is computed by ensuring that the residual r_k is orthogonal to S_k, then we obtain

$$L_k y_k^{\mathrm{CGME}} = e_1 \beta_1, \qquad L_k \in \mathbb{R}^{k \times k}, \tag{2.7}$$

which corresponds to the CGME method; see, e.g., [16]. If the approximation is computed by minimizing the norm of the residual r_k, then we get

$$\| L_{k+} y_k^{\mathrm{LSQR}} - e_1 \beta_1 \| = \min_y \| L_{k+} y - e_1 \beta_1 \|, \quad L_{k+} \in \mathbb{R}^{(k+1) \times k}, \tag{2.8}$$

which corresponds to the CGLS or LSQR method; see [22, 41, 42]. In both cases, $[\beta_1 e_1, L_k]$ respectively $[\beta_1 e_1, L_{k+}]$ approximate the core problem within $Ax \approx b$; see [23, 24, 43]. In short, the bidiagonalization *concentrates the useful information from the data* $[b, A]$ *to the leading principal bidiagonal block*. In noisy ill-posed problems (1.1), however, the subproblems (2.7), (2.8) can also be polluted by the noise.

3 Propagation of the noise in the Golub-Kahan iterative bidiagonalization

Noise propagation in the Golub-Kahan iterative bidiagonalization is illustrated for the problem `shaw` from the Regularization Toolbox [17]. The matrix A, the exact right-hand side b^{exact} and the exact solution x were determined by `[A,b_exact,x] = shaw(400)`. We add the noise b^{noise} as a random vector using the MATLAB function `randn(400,1)`, scaled in order to obtain different *noise levels* δ_{noise},

$$\delta_{\text{noise}} \equiv \frac{\|b^{\text{noise}}\|}{\|b^{\text{exact}}\|}. \tag{3.1}$$

Figure 1(a) shows the singular values σ_j of the matrix A (solid line) computed by the MATLAB function `svd`, and the absolute values of the projections $|(b, u_j)|$ of the noisy right-hand side b onto the left singular vector subspaces of the matrix A. We use the noise levels $\delta_{\text{noise}} = 10^{-14}$, 10^{-8}, and 10^{-4}. Until the noise level is reached, the absolute values of the right-hand side projections onto the left singular vector subspaces are close to the corresponding singular values. This is given by the fact that, for this problem, $|(b^{\text{exact}}, u_j)|$, $j = 1, 2, \ldots$ satisfy the discrete Picard condition: on average, they decay faster than the singular values of A; see [25], [20, Sect. 4] and Fig. 1(b). For the subspaces corresponding to small singular values, the projections of b are completely dominated by the noise, and the discrete Picard condition for b is thus drastically violated.

Consider the vectors s_k, w_k generated by the bidiagonalization algorithm (2.1) described in the previous section. The starting vector $s_1 = b/\|b\|$ is the normalized noisy right-hand side and therefore it is *contaminated by the noise*. The vector s_2 is obtained from s_1 as follows. First, application of the smoothing operator AA^T

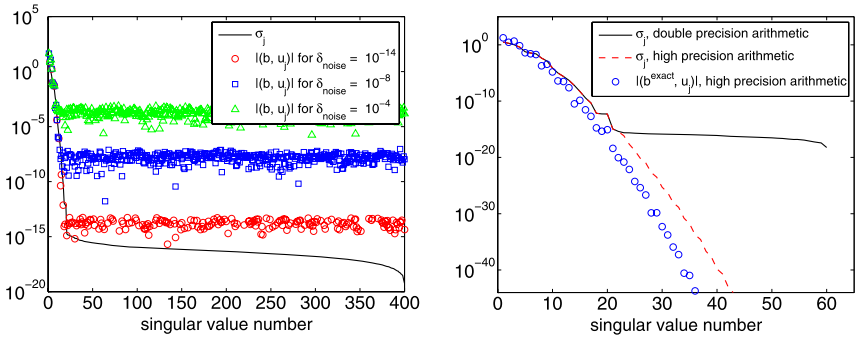


Fig. 1 The singular values σ_j and the absolute values of the projections of the noisy right-hand side b onto the left singular vectors of A for the problem `shaw(400)` and the noise levels $\delta_{\text{noise}} = 10^{-14}$, 10^{-8}, and 10^{-4} (**a**). The singular values of the matrix A computed by the statement `[A,b,x]=shaw(60)` using the standard routine from the Regularization Toolbox [17] (*solid line*) compared with the singular values of the matrix A computed by the modified routine `shaw_vpa`, cf. http://www.cs.cas.cz/krylov, section 'Software', (*dash-dotted line*) using high precision arithmetic guaranteeing 128 decimal digits (**b**). The discrete Picard condition is illustrated by plotting $|(b^{\text{exact}}, u_j)|$, where b^{exact} is computed using the routine `shaw_vpa`

(see (2.4)) smooths out the high frequency components in both b^{exact} and b^{noise}; see also [25, Sect. 3.1], [19, Sect. 4.3]. The subsequent orthogonalization of $AA^T s_1$ against s_1 represents a linear combination of s_1 contaminated by the noise and $AA^T s_1$ *which has been smoothed*. Therefore the contamination of s_1 by the *high frequency part* of the noise is transferred, with multiplication by a scalar coefficient, to s_2. Adding a multiple of $AA^T s_1$ eliminates *a portion of the smooth part* of s_1. Therefore the relative level of the high frequency part of the noise can be expected to be higher in s_2 (which is orthogonal to s_1) than in s_1. Analogous high frequency noise transfer takes place for any k with the vector s_{k+1} obtained from $AA^T s_k$ through the orthogonalization against the vectors s_{k-1} and s_k with a subsequent normalization.

The rest of Sect. 3 is structured into two complementary subsections. First we use the relationship between the Golub-Kahan iterative bidiagonalization and the Lanczos tridiagonalization described in Sect. 2. The noise amplification is described as an effect of damping (filtering out) the smooth (low frequency) components due to convergence of Ritz values to large eigenvalues. Then we show that smoothing and orthogonalization in the Golub-Kahan iterative bidiagonalization can be interpreted as a step-by-step elimination of the dominant smooth (low frequency) components, and it therefore leads to revealing of the high frequency noise.

3.1 Lanczos tridiagonalization and filtering out the low frequency components

The white noise amplification can be described in the frequency domain by computing the spectral coefficients of s_k with respect to the (noise-free) orthonormal left singular vectors of the matrix A. Using (2.6), (2.4) gives the matrix identity for the *spectral components*

$$\Sigma^2(U^T S_k) = (U^T S_k)(L_k L_k^T) + \alpha_k \beta_{k+1}(U^T s_{k+1}) e_k^T,$$

with the last column of the last term equal to

$$\alpha_k \beta_{k+1}(U^T s_{k+1}) = \Sigma^2(U^T s_k) - (\alpha_k^2 + \beta_k^2)(U^T s_k) - \alpha_{k-1}\beta_k(U^T s_{k-1}). \qquad (3.2)$$

Consider for example `shaw(400)` for which the vector s_1 has dominant components in the directions of several left singular vectors representing low frequencies, with a noticeable maximum in u_1. This is illustrated in Figs. 1 and 2 which shows the left singular vectors of A corresponding to $\sigma_1, \sigma_2, \ldots, \sigma_{10}$. An analogous assumption can be used for discrete ill-posed problems in general, see [20, Sect. 2]. At the first step

$$\alpha_1 \beta_2(U^T s_2) = \Sigma^2(U^T s_1) - \alpha_1^2(U^T s_1) = (\Sigma^2 - \alpha_1^2 I)U^T s_1,$$

and, apart from multiplication by $\alpha_1\beta_2$, the spectral components $U^T s_2$ are given by the spectral components of $U^T s_1$ scaled by $\Sigma^2 - \alpha_1^2 I$. The scaling acts as filtering which damps the dominant lowest frequency component in the direction of u_1. At the same time, the high frequency components are multiplied by the factors $\sigma_j^2/(\alpha_1\beta_2) - \alpha_1/\beta_2$, where the first term is negligible for large j, and the absolute value of the second term is likely to be significantly larger than one (see the argument below). As a consequence, the relative level of high frequency noise is likely to be much larger

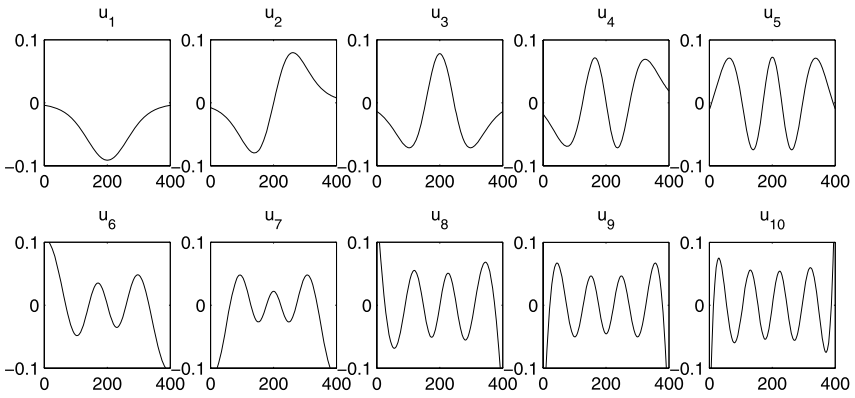


Fig. 2 The left singular vectors of A corresponding to $\sigma_1, \sigma_2, \ldots, \sigma_{10}$ for the problem `shaw(400)`

in s_2 than in s_1. At the general step, the Lanczos recurrence (3.2) can be rewritten in terms of the (Lanczos) polynomial in the diagonal matrix Σ^2,

$$U^T s_{k+1} = \varphi_k(\Sigma^2) U^T s_1, \tag{3.3}$$

where $\varphi_k(\lambda)$ is the kth orthonormal polynomial with respect to the Riemann-Stieltjes integral defined by the distribution function ω with the points of increase $\sigma_n^2, \ldots, \sigma_1^2$ (please recall that we assume, for simplicity of exposition, $\sigma_n > 0$) and the weights $|(s_1, u_n)|^2, \ldots, |(s_1, u_1)|^2$ respectively; see Sect. 2.[1] The roots of $\varphi_k(\lambda)$ are given by $(\theta_\ell^{(k)})^2$, $\ell = 1, \ldots, k$ (the Ritz values). Because of the dominance of the weights corresponding to the large nodes of ω, the large Ritz values $(\theta_k^{(k)})^2$, $(\theta_{k-1}^{(k)})^2, \ldots$ closely approximate $\sigma_1^2, \sigma_2^2, \ldots$; see Fig. 3. This creates the damping effect of $\varphi_k(\lambda)$ in the direction of the smooth components $(s_1, u_1)u_1$, $(s_1, u_2)u_2, \ldots$ of s_1. The constant term

$$\varphi_k(0) = \prod_{j=1}^{k} \frac{\alpha_j}{\beta_{j+1}} \tag{3.4}$$

then causes the relative amplification of the high frequency noise components present in s_1.

Summarizing, assume that the components of s_1 in the directions of the left singular vectors $u_1, u_2, u_3, \ldots$ decay faster than the associated singular values. Then, as a consequence of the orthogonalization process (3.2), $s_2, s_3, \ldots$ will have dominant components in the directions of the left singular vectors $u_2, u_3, \ldots$ respectively, with the relative levels of the high frequency noise components gradually increasing. Eventually, for some $k \equiv k_{\text{noise}}$, the vector $s_{k_{\text{noise}}+1}$ will have comparable com-

[1] It is worth to mention that the polynomials orthogonal with respect to a given Riemann-Stieltjes integral are called *Stieltjes polynomials*, and the associated three-term recurrence, given in its matrix form by Lanczos, was described by Stieltjes and others in the 19th century, cf. [13, Sect. 1.4, p. 80].

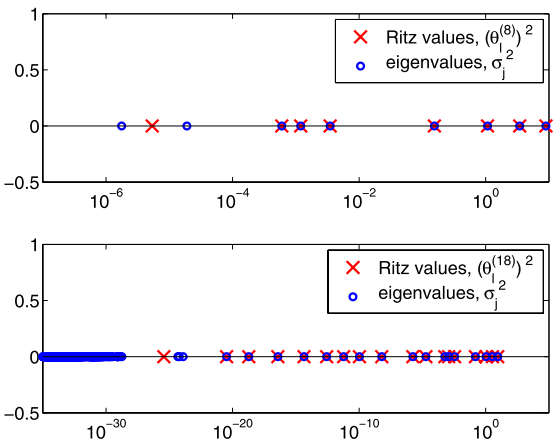


Fig. 3 The eigenvalues of AA^T (*circles*), and the Ritz values (*crosses*) corresponding to the 8th (*top plot*) and 18th (*bottom plot*) iterations respectively, for the problem `shaw(400)`. (All the *circles* in the *top plot* reappear as the nine rightmost circles in the bottom plot)

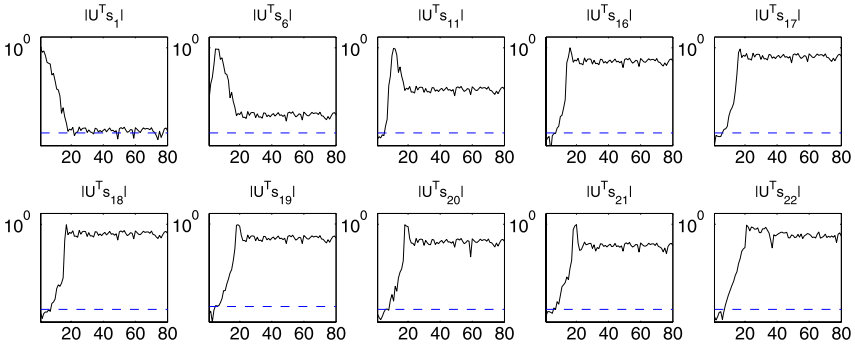


Fig. 4 The absolute values of the first 80 spectral components of the vectors s_k computed using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$. The *dashed line* represents the machine precision ε_M

ponents in practically all subspaces generated by singular vectors corresponding to $\sigma^2_{k_{\text{noise}}+1}, \sigma^2_{k_{\text{noise}}+2}, \ldots$, and this will reveal the noise. Figure 4 shows the absolute values of the components of several vectors $U^T s_k$ (the spectral components) for the problem `shaw(400)` with the noise level 10^{-14}. The absolute values of the components in the vector $U^T s_{18}$ corresponding to $\sigma_{18}, \sigma_{19}, \ldots$ are comparable, with no dominant maximum. Consequently, the noise is revealed in s_{18} for $k = k_{\text{noise}} = 17$. In the next step the high frequency components in s_{19} slightly decrease, and this is captured in Fig. 5. Components corresponding to low frequencies gradually decrease to the machine precision level ε_M (plotted by the dashed line). Figure 5 nicely illustrates why $k_{\text{noise}} = 17$ is the appropriate choice of the noise revealing iteration, i.e. why the noise is fully revealed in s_{18}.

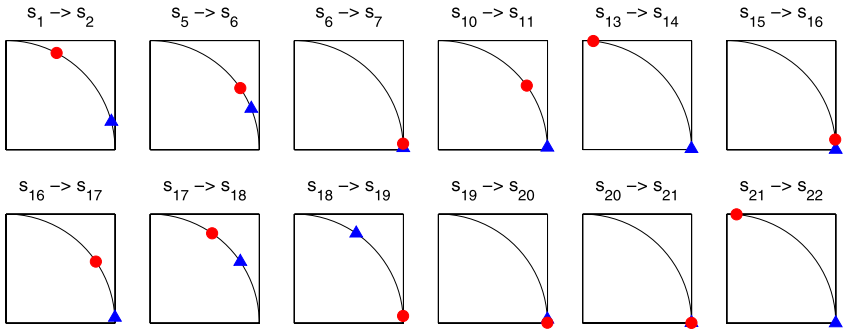


Fig. 5 Each graph shows for a given k the norms of the components of two consecutive vectors s_k (*the triangle*) and s_{k+1} (*the circle*) in the subspace span$\{u_1, \ldots, u_{k+1}\}$, also called the signal subspace (*the horizontal axis*), and the subspace span$\{u_{k+2}, \ldots, u_n\}$, also called the noise subspace (*the vertical axis*), for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$. Since the vectors s_j are normalized, the points are on the unit circle. One can see that until the noise level is reached, each step of the Golub-Kahan iterative bidiagonalization algorithm moves the subsequent vector counterclockwise, i.e. each s_{k+1} is closer to the noise subspace (*the vertical axis*) than s_k. The vector s_{19}, on the other hand, moves significantly clockwise due to the fact that the noise is partially projected out, and therefore the noise is revealed in s_{18}. This justifies the choice of the noise revealing iteration $k_{\text{noise}} = 17$. We are grateful to an anonymous referee for suggesting this visualization of the noise propagation process

3.2 Smoothing and orthogonalization in the Golub-Kahan iterative bidiagonalization

In order to further illustrate the (high frequency) white noise amplification, we consider the decomposition of s_1 into the exact component s_1^{exact} and the noise component s_1^{noise}, $s_1 = s_1^{\text{exact}} + s_1^{\text{noise}}$. Then the second equation in (2.1) gives

$$\beta_2 s_2 = Aw_1 - \alpha_1(s_1^{\text{exact}} + s_1^{\text{noise}}),$$

where Aw_1 is smooth with the low frequency components of the noise negligible relatively to the low frequency components of the exact data. This justifies the following *definition* of s_2^{exact} and s_2^{noise},

$$\beta_2 s_2^{\text{exact}} \equiv Aw_1 - \alpha_1 s_1^{\text{exact}},$$
$$\beta_2 s_2^{\text{noise}} \equiv Aw_1 - \alpha_1 s_1 - \beta_2 s_2^{\text{exact}} = -\alpha_1 s_1^{\text{noise}},$$

giving

$$\beta_2(s_2^{\text{exact}} + s_2^{\text{noise}}) = Aw_1 - \alpha_1(s_1^{\text{exact}} + s_1^{\text{noise}}).$$

Analogously, for $k = 2, 3, \ldots$

$$\beta_{k+1} s_{k+1}^{\text{exact}} \equiv Aw_k - \alpha_k s_k^{\text{exact}}, \tag{3.5}$$
$$\beta_{k+1} s_{k+1}^{\text{noise}} \equiv -\alpha_k s_k^{\text{noise}}, \tag{3.6}$$
$$s_{k+1} = s_{k+1}^{\text{exact}} + s_{k+1}^{\text{noise}}, \qquad \beta_{k+1} s_{k+1} = Aw_k - \alpha_k s_k. \tag{3.7}$$

It should be understood that s_k^{exact} and s_k^{noise} do not represent *true* exact data and noise components of s_k, respectively. If the multiplication by AA^T represents a significant smoothing of the high frequency components, then one can however expect that while $\|s_k^{\text{exact}}\| \gg \|s_k^{\text{noise}}\|$, $\|s_k^{\text{exact}}\|$ is close to the norm of the true data component and $\|s_k^{\text{noise}}\|$ is close to the norm of the true noise component of s_k. Using (3.6),

$$s_{k+1}^{\text{noise}} \equiv -\frac{\alpha_k}{\beta_{k+1}} s_k^{\text{noise}} \equiv (-1)^k \rho_k^{-1} s_1^{\text{noise}} \tag{3.8}$$

where the cumulative amplification ratio

$$\rho_k^{-1} \equiv \prod_{j=1}^{k} \frac{\alpha_j}{\beta_{j+1}} = \varphi_k(0), \tag{3.9}$$

see (3.4), on average (rapidly) grows as k increases.

In order to justify the expectation that ρ_k^{-1} on average (rapidly) grows with increasing k, we rewrite the Golub-Kahan bidiagonalization (2.1) for the spectral components $U^T s_j$ and $V^T w_j$,

$$\alpha_1(V^T w_1) = \Sigma(U^T s_1), \tag{3.10}$$

$$\beta_2(U^T s_2) = \Sigma(V^T w_1) - \alpha_1(U^T s_1), \tag{3.11}$$

and for $k = 2, 3, \ldots$

$$\alpha_k(V^T w_k) = \Sigma(U^T s_k) - \beta_k(V^T w_{k-1}), \tag{3.12}$$

$$\beta_{k+1}(U^T s_{k+1}) = \Sigma(V^T w_k) - \alpha_k(U^T s_k). \tag{3.13}$$

From (3.10) we see that $(V^T w_1)$ is dominated by the same components as $(U^T s_1)$, with the dominance even enlarged as a consequence of the scaling by Σ. In (3.11), however, $\Sigma(V^T w_1)$ is orthogonalized against $U^T s_1$ in order to get, after normalization, $U^T s_2$. This requires that the dominance in $\Sigma(V^T w_1)$ and $U^T s_1$ is cancelled out, otherwise the orthogonality $U^T s_2 \perp U^T s_1$ can not hold. If the dominance is significant, one can therefore expect $\beta_2 \ll \alpha_1$. An analogous argument can be applied to the general step (3.12) and (3.13). Since the dominance in $\Sigma(U^T s_k)$ and $(V^T w_{k-1})$ is shifted by one component, one can not expect a significant cancellation, and, as a consequence

$$\alpha_k \approx \beta_k$$

should roughly hold. On the other hand, the vectors $\Sigma(V^T w_k)$ and $U^T s_k$ do exhibit dominance in the direction of the same components. If this dominance is strong enough, then the required orthogonality of s_{k+1} and s_k can not be achieved without a significant cancellation, and

$$\beta_{k+1} \ll \alpha_k$$

can be expected, giving the large cumulative amplification ratio ρ_k^{-1} (small ρ_k) as k progresses. The process is illustrated in Fig. 6.

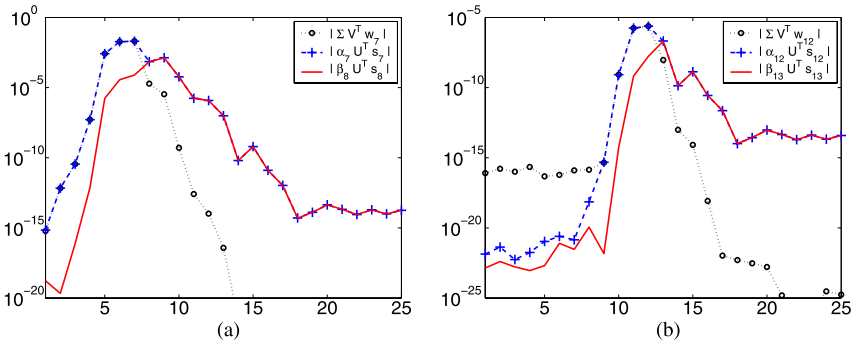


Fig. 6 The absolute values of the first 25 components of the vectors $\Sigma(V^T w_k)$, $\alpha_k(U^T s_k)$, and $\beta_{k+1}(U^T s_{k+1})$ for the problem `shaw(400)` for $k = 7$, $\beta_8/\alpha_7 = 0.0542$ (**a**), and $k = 12$, $\beta_{13}/\alpha_{12} = 0.0677$ (**b**)

Summarizing, starting with $s_1 = b/\beta_1$ contaminated by white noise, the Golub-Kahan iterative bidiagonalization algorithm applied to the discrete ill-posed problem $Ax \approx b$, where A represents *a smoothing operator*, amplifies the relative level of the noise in s_{k+1} as k increases with the cumulative amplification ratio ρ_k^{-1}. Note that the recurrence for the vectors w_k starts with the smoothed vector $w_1 = A^T s_1/\|A^T s_1\|$. Consequently, all vectors w_k are smoothed, and, in comparison to the vectors s_k, they do not contain significant information about the noise.

Remark Here we deal with mathematical properties (assuming exact arithmetic) and we do not need to consider specific implementation details. It should be noted, however, that the transfer of the high frequency noise from s_{k-1} and s_k to s_{k+1} is a local phenomenon which seemingly does not require preserving global orthogonality among the vectors $s_1, \ldots, s_k$. Since in practical implementations the local orthogonality among s_{k-1}, s_k and s_{k+1} is well preserved, cf. [39], one can intuitively expect that the preceding considerations are valid, with a small inaccuracy, also in practical finite precision computations. As we will demonstrate numerically, although such an intuitive argument is to some extent valid, the transfer of noise in finite precision computations is much more complicated; see Sect. 5.

The amplification of white noise in the Golub-Kahan iterative bidiagonalization is illustrated in Figs. 7–12. Figures 7–9 show individual components of several left bidiagonalization vectors s_k and their components s_k^{exact} and s_k^{noise}, computed using (3.5)–(3.7), for the problem `shaw(400)` with the noise level 10^{-14}. The Golub-Kahan iterative bidiagonalizations were performed with double reorthogonalization of the computed vector sequences, which ensures preserving orthogonality close to machine precision. Here we use the low noise level *on purpose* to illustrate the propagation of the noise through many steps of the iterative bidiagonalization. Results with larger noise levels will be presented below. As k increases, we observe the increasing oscillating pattern of s_k. For the vectors s_{17} and s_{18} the basic oscillating pattern (determined by s_k^{exact}) is strongly modulated by the high frequency noise; compare

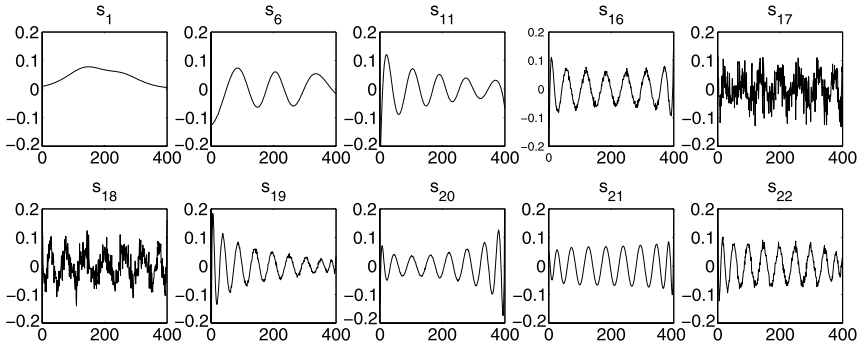


Fig. 7 Individual components of several left bidiagonalization vectors s_k computed using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$

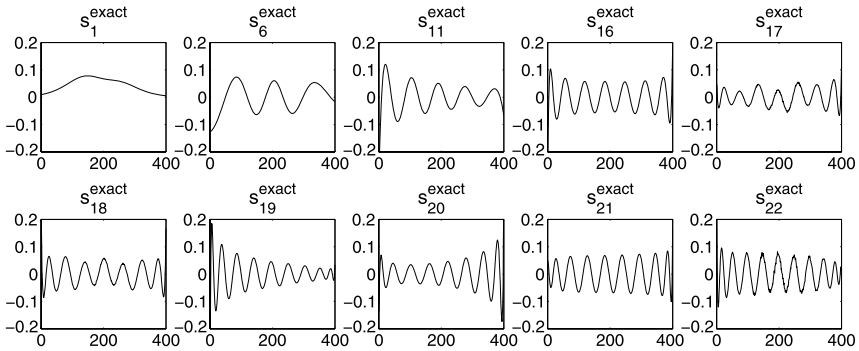


Fig. 8 Individual components of several vectors s_k^{exact} computed by (3.5) and (3.7) using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$

Figs. 7 and 8. It is interesting to observe that the level of the high frequency noise in the subsequent vector s_{19} is significantly lower than in s_{18}. As a consequence of smoothing and the orthogonality of s_{19} to s_{18} and s_{17}, the high frequency noise is partially projected out, cf. Fig. 5. Figure 10 presents the norms of the components s_k^{exact}, s_k^{noise} and the smallest singular value of the matrix S_k as k increases, which nicely complements the visual description in Figs. 7–9. Please notice that the decomposition of s_k into the components s_k^{exact} and s_k^{noise}, which should be close to the (unknown) exact data and noise components, is no longer relevant for $k \geq k_{\text{noise}}$. Finally, Fig. 11 gives the values of the normalization coefficients α_k, β_{k+1} and of their cumulative ratio ρ_k, and Fig. 12 depicts α_k, β_k and the cumulative ratio

$$\prod_{j=2}^{k} \frac{\beta_j}{\alpha_j}. \tag{3.14}$$

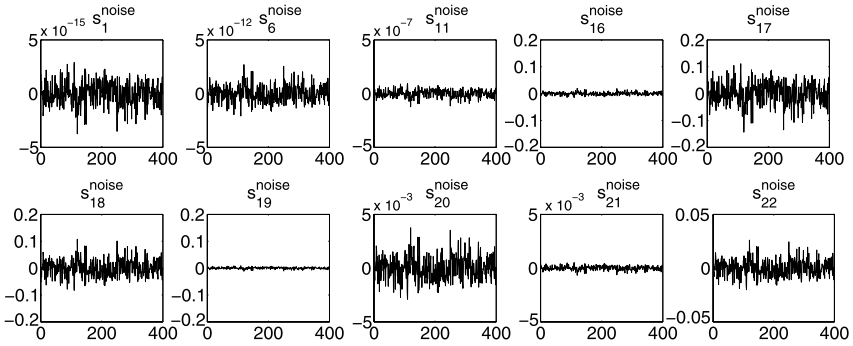


Fig. 9 Individual components of several vectors s_k^{noise} computed by (3.6) and (3.7) using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$. Note that the scale on the y axis is for different k different

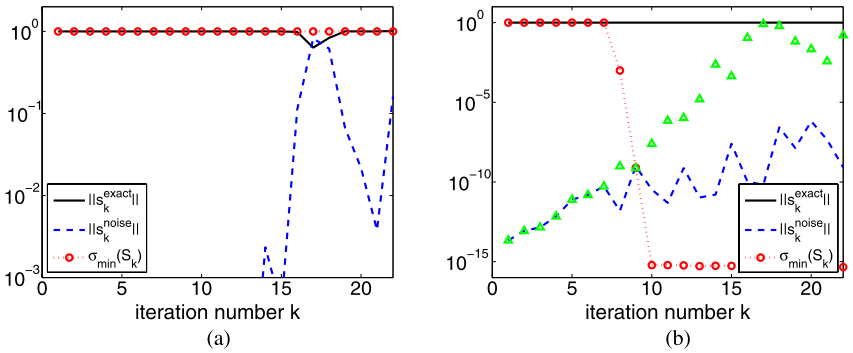


Fig. 10 The norms of $s_k^{\text{exact}}, s_k^{\text{noise}}$ and the smallest singular value of the matrix S_k for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$, computed by the Golub-Kahan iterative bidiagonalization with double reorthogonalization (**a**), and without reorthogonalization (**b**). For comparison, *the triangles* in (**b**) represent the norm of the components s_k^{noise} computed with double reorthogonalization. Please note the different vertical scales in (**a**) and (**b**)

The experimental results in Figs. 10(a), 11(a) and 12 correspond to the Golub-Kahan iterative bidiagonalization with double reorthogonalization, while Figs. 10(b) and 11(b) correspond to the results obtained without reorthogonalization. The differences will be discussed in Sect. 5 below.

The basis consisting of the left singular vectors of A is computationally expensive, and computing the corresponding components of the vectors s_k is practically infeasible. Analogous noise-revealing behavior can be observed in any other suitable orthonormal basis. Consider, e.g., the standard Fourier trigonometric basis

$$f_j(x) \equiv e^{(\frac{2\pi\imath}{n})jx}, \quad \text{for } j = 0, \pm 1, \ldots, \pm n, \tag{3.15}$$

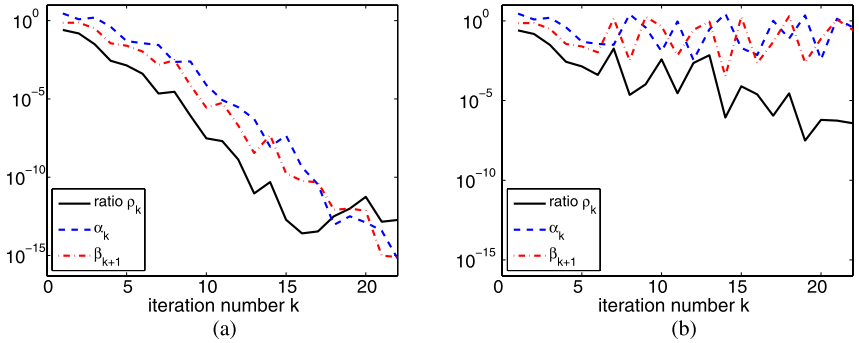
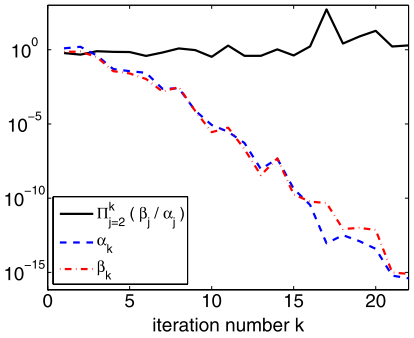


Fig. 11 Normalization coefficients α_k, β_{k+1} and their cumulated ratio ρ_k, see (3.9), for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$, computed by the Golub-Kahan iterative bidiagonalization with double reorthogonalization (**a**), and without reorthogonalization (**b**)

Fig. 12 Normalization coefficients α_k, β_k and their cumulated ratio (3.14) for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$, computed by the Golub-Kahan iterative bidiagonalization with double reorthogonalization



where ι is the imaginary unit. Fourier coefficients in this basis can be computed efficiently using the fast Fourier transform, [8, 9]. Figure 13 shows similar results as Fig. 4 computed for the basis (3.15) using the MATLAB function `fft`.

Similar behavior can be observed for other noise levels, as well as for noise-free problems. In this last case the "high frequency noise" is simply caused by local rounding errors from the finite precision arithmetic. Recall here that, because of the double reorthogonalization, the local rounding errors do not cause loss of orthogonality of the computed bidiagonalization vectors (the loss of orthogonality is kept close to the machine precision level). For more details see Sect. 5.

4 Determination of the noise level in the data and detection of the noise revealing iteration

As mentioned in Sect. 3, Fourier coefficients of the vectors s_k can be computed cheaply using, e.g., MATLAB function `fft`. Then some kind of statistical criteria can be used to identify the iteration where the Fourier coefficients $k+1, \ldots, n$ of the vector s_{k+1} are (on average) comparable, i.e. s_{k+1} resembles (except the first k

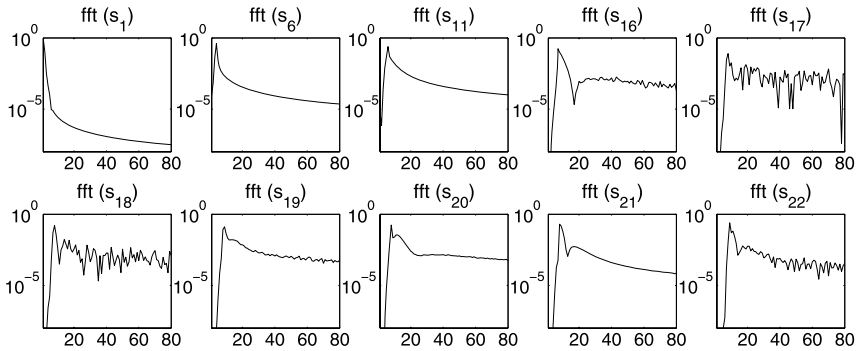


Fig. 13 The absolute values of the first 80 Fourier components of the vectors s_k, computed using the double reorthogonalized Golub-Kahan iterative bidiagonalization, in the trigonometric basis (3.15) for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$

components) white noise. Such statistical tools are used in a different context in [20, 46, 47] to decide whether a given residual vector is white-noise like. Here we propose a different approach which is more straightforward. The connection between the Golub-Kahan iterative bidiagonalization and the Lanczos tridiagonalization described in Sect. 2 suggests a new way of determining the noise level in the data without computing spectral or Fourier coefficients of s_k.

4.1 Automatic determination of the noise level based on approximation of the Riemann-Stieltjes distribution function

As recalled in Sect. 2, the Lanczos tridiagonalization generates a sequence of distribution functions $\omega^{(k)}$ with the nodes $(\theta_\ell^{(k)})^2$ and the weights $|(p_\ell^{(k)}, e_1)|^2$, $\ell = 1, \ldots, k$, that approximate the distribution function ω with the nodes σ_j^2 and the weights $|(b/\beta_1, u_j)|^2$, $j = n, n-1, \ldots, 1$; see [12, 31, 37, 52]. Depending on the noise level, for the smaller nodes of the distribution function ω the weights are completely dominated by noise, i.e., there exists an index J_{noise} such that for $j \geq J_{\text{noise}}$

$$|(b/\beta_1, u_j)|^2 \approx |(b^{\text{noise}}/\beta_1, u_j)|^2,$$

and the weight of the *set of the associated nodes* is given by

$$\delta^2 \equiv \sum_{j=J_{\text{noise}}}^{n} |(b^{\text{noise}}/\beta_1, u_j)|^2, \tag{4.1}$$

for illustration see Fig. 1. Since $\|b^{\text{noise}}\| \ll \|b^{\text{exact}}\|$, we can approximately write

$$\delta_{\text{noise}}^2 = \frac{\|b^{\text{noise}}\|^2}{\|b^{\text{exact}}\|^2} \approx \frac{\|b^{\text{noise}}\|^2}{\|b\|^2} = \sum_{j=1}^{n} |(b^{\text{noise}}/\beta_1, u_j)|^2. \tag{4.2}$$

With discrete ill-posed problems one can assume $J_{\text{noise}} \ll n$, and therefore, *assuming white noise*,

$$\delta^2 \approx \frac{n - J_{\text{noise}}}{n} \delta^2_{\text{noise}} \approx \delta^2_{\text{noise}}. \tag{4.3}$$

Please recall that the large nodes $\sigma_1^2, \sigma_2^2, \ldots$ are well-separated relative to the small ones and their weights on average decrease faster than $\sigma_1^2, \sigma_2^2, \ldots$. Therefore the distribution function $\omega^{(k)}$ approximates the distribution function ω in a special way—the large nodes essentially control the behavior of the early stages of the Lanczos tridiagonalization; see also Sect. 3, in particular Fig. 3. For $k = 1, 2, \ldots$ the large singular values of L_k become close to the largest singular values of A due to the dominance described above. At *any* iteration step the weight corresponding to $(\theta_1^{(k)})^2$ must be larger than the sum of weights of all σ_j^2 smaller than $(\theta_1^{(k)})^2$; see [12, 26], [11, Sect. 5.3]. As k increases, some $(\theta_1^{(k)})^2$ eventually approaches (or becomes smaller than) the node $\sigma^2_{J_{\text{noise}}}$, and its weight $|(p_1^{(k)}, e_1)|^2$ becomes

$$|(p_1^{(k)}, e_1)|^2 \approx \delta^2 \approx \delta^2_{\text{noise}}. \tag{4.4}$$

Using the notation of Sect. 3, this iteration step is equal to $k_{\text{noise}} + 1$. Indeed, $|(p_1^{(k)}, e_1)|$ is proportional to the noise level only *after* all smooth components of s_1 with the norms larger than the noise level are damped at the iteration step k_{noise}. The smallest nodes $(\theta_1^{(k_{\text{noise}}+1)})^2$, $(\theta_1^{(k_{\text{noise}}+2)})^2, \ldots$ strictly decrease due to the strict interlacing property of the Ritz values, but the corresponding weights (which also strictly decrease) remain in the subsequent steps approximately the same until the set of the smallest nodes $\{\sigma_n^2, \ldots, \sigma^2_{J_{\text{noise}}}\}$ is approximated by more than one Ritz value and the weight (4.1) is split, which happens when

$$(\theta_1^{(k)})^2 < \sigma_j^2 \quad \text{for } j \gg J_{\text{noise}}. \tag{4.5}$$

Summarizing, the weight $|(p_1^{(k)}, e_1)|^2$ corresponding to the smallest Ritz value $(\theta_1^{(k)})^2$ is strictly decreasing. At some iteration step $k = k_{\text{noise}} + 1$ it sharply starts to (almost) stagnate on the level close to the squared noise level δ^2_{noise}, see (4.3) and (4.4). In order to determine k_{noise} and to estimate δ_{noise}, it is therefore sufficient to monitor the first component of the left singular vector $|(p_1^{(k)}, e_1)|$ of the bidiagonal matrix L_k that is associated to its *smallest singular value* $\theta_1^{(k)}$; see (4.4). When it starts to stagnate,

$$\delta_{\text{noise}} \approx |(p_1^{(k_{\text{noise}}+1)}, e_1)|. \tag{4.6}$$

The stagnation can be detected by an automated procedure that does not rely on human interaction.

Figure 14 shows the positions of the leftmost points $[(\theta_1^{(k)})^2, |(p_1^{(k)}, e_1)|^2]$, $k = 1, 2, \ldots$ with respect to the distribution function $\omega(\lambda)$ (a) and $|(p_1^{(k)}, e_1)|$ as a function of k (b) for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$; (c) and (d) show the same for the noise level $\delta_{\text{noise}} = 10^{-4}$.

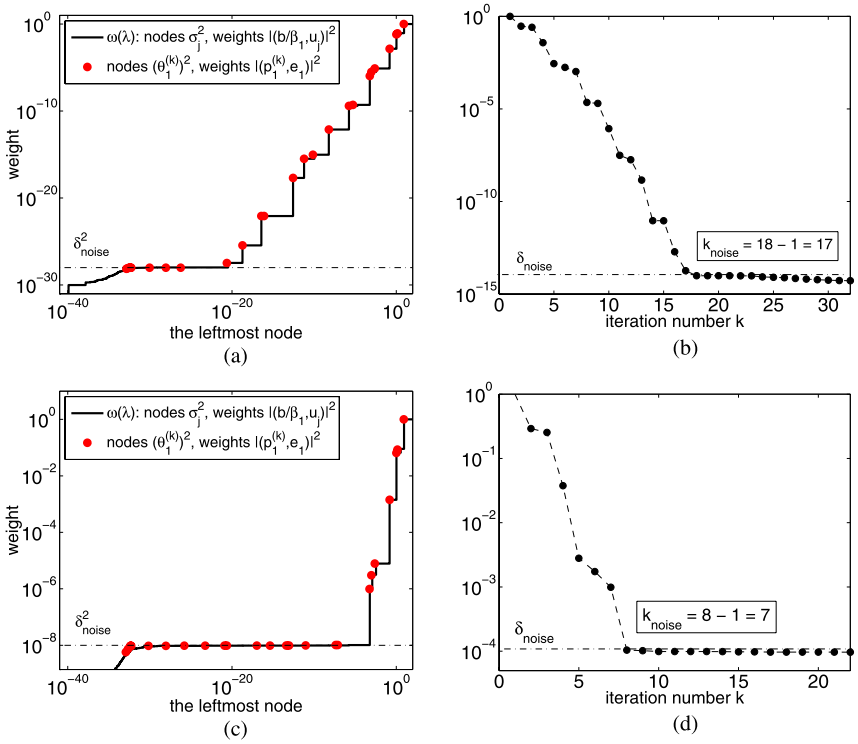


Fig. 14 The positions of the leftmost points $[(\theta_1^{(k)})^2, |(p_1^{(k)}, e_1)|^2]$, $k = 1, 2, \ldots$ with respect to the distribution function $\omega(\lambda)$ (**a**) and the absolute value of the first component $|(p_1^{(k)}, e_1)|$ of the left singular vector of L_k corresponding to its smallest singular value (**b**) for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-14}$; (**c**) and (**d**) show the same for the noise level $\delta_{\text{noise}} = 10^{-4}$. The *horizontal dashed-dotted lines* represent the squared noise level δ^2_{noise} and the noise level δ_{noise}, respectively

4.2 An additional way of estimating the noise level in the data

Knowing the iteration k_{noise} where the noise is revealed in the Golub-Kahan iterative bidiagonalization, one can also estimate the level of the noise in the original data from the bidiagonalization coefficients. Here k_{noise} can be determined as described in the previous section. This section therefore describes an additional way of checking the estimate of δ_{noise} obtained in Sect. 4.1. Using (3.8),

$$s_{k+1}^{\text{noise}} = (-1)^k \rho_k^{-1} \frac{b^{\text{noise}}}{\|b\|},$$

where ρ_k is defined by (3.9), which gives

$$\frac{\|b^{\text{noise}}\|}{\|b\|} = \rho_k \|s_{k+1}^{\text{noise}}\|. \tag{4.7}$$

At the iteration step k_{noise} the noise is revealed, and we can assume that *on average* the components s_k^{exact} and s_k^{noise} are of about the same norm, cf. s_{18}^{exact} and s_{18}^{noise} in Figs. 8 and 9. Therefore

$$1 = \|s_{k_{\text{noise}}+1}\| = \|s^{\text{exact}}_{k_{\text{noise}}+1} + s^{\text{noise}}_{k_{\text{noise}}+1}\| \le \|s^{\text{exact}}_{k_{\text{noise}}+1}\| + \|s^{\text{noise}}_{k_{\text{noise}}+1}\|$$

gives, considering that $s^{\text{exact}}_{k_{\text{noise}}+1}$ is much smoother than $s^{\text{noise}}_{k_{\text{noise}}+1}$ and thus there is not much cancellation between the individual components,

$$\|s^{\text{noise}}_{k_{\text{noise}}+1}\| \approx \frac{1}{2}.$$

Assuming $\|b^{\text{noise}}\| \ll \|b^{\text{exact}}\|$, the left part of (4.7) can be approximated, as in (4.2), by

$$\frac{\|b^{\text{noise}}\|}{\|b\|} = \frac{\|b^{\text{noise}}\|}{\|b^{\text{exact}} + b^{\text{noise}}\|} \approx \frac{\|b^{\text{noise}}\|}{\|b^{\text{exact}}\|},$$

which finally gives the estimate of the noise level in the original data

$$\delta_{\text{noise}} = \frac{\|b^{\text{noise}}\|}{\|b^{\text{exact}}\|} \approx \frac{1}{2}\rho_{k_{\text{noise}}}. \tag{4.8}$$

Table 1 shows the iterations k_{noise} (second row) and the corresponding estimates of the noise level $|(p_1^{(k_{\text{noise}}+1)}, e_1)|$ (third row) and $\frac{1}{2}\rho_{k_{\text{noise}}}$ (last row) for the problems `shaw(400)` and `ilaplace(100,1)` from the Regularization Toolbox [17] with different noise levels δ_{noise}. The estimates are average values computed using the set of 1000 randomly chosen sample vectors b^{noise}. In our experiments k_{noise} was

Table 1 Noise level in the data (first row), iteration k_{noise} where the noise is revealed (second row), the estimated noise level $|(p_1^{(k_{\text{noise}}+1)}, e_1)|$ (third row) and $\frac{1}{2}\rho_{k_{\text{noise}}}$ (last row), see (4.6) and (4.8) respectively, for problems `shaw(400)` and `ilaplace(100,1)`. The estimates represent average values computed using 1000 randomly chosen vectors b^{noise}

problem	`shaw(400)`				
noise level δ_{noise}	1×10^{-14}	1×10^{-10}	1×10^{-6}	1×10^{-4}	1×10^{-2}
k_{noise}	16	13	9	7	4
$\lvert(p_1^{(k_{\text{noise}}+1)}, e_1)\rvert$	1.80×10^{-14}	8.99×10^{-11}	1.31×10^{-6}	1.01×10^{-4}	1.03×10^{-2}
estimate $\frac{1}{2}\rho_{k_{\text{noise}}}$	8.93×10^{-15}	4.95×10^{-11}	6.55×10^{-7}	5.24×10^{-5}	5.55×10^{-3}
problem	`ilaplace(100,1)`				
noise level δ_{noise}	1×10^{-13}	1×10^{-10}	1×10^{-7}	1×10^{-2}	1×10^{-1}
k_{noise}	22	18.75	15.30	6.02	2
$\lvert(p_1^{(k_{\text{noise}}+1)}, e_1)\rvert$	9.12×10^{-14}	1.24×10^{-10}	1.34×10^{-7}	1.02×10^{-2}	1.11×10^{-1}
estimate $\frac{1}{2}\rho_{k_{\text{noise}}}$	4.77×10^{-14}	6.42×10^{-11}	7.11×10^{-8}	8.98×10^{-3}	5.57×10^{-2}

determined as the first iteration step k for which

$$\frac{|(p_1^{(k+1)}, e_1)|}{|(p_1^{(k+1+step)}, e_1)|} < \left(\frac{|(p_1^{(k)}, e_1)|}{|(p_1^{(k+1)}, e_1)|} \right)^{\zeta}, \tag{4.9}$$

where ζ was set to 0.5 and *step* was set to 3. We emphasize that (4.9) should be considered as *an example* of a possible automated stopping criteria. Its form has not resulted from an extensive research; such work is yet to be done and the results will most probably depend on particular application areas. For the problem `shaw` (4.9) works well except for the noise level $\delta_{\text{noise}} = 1 \times 10^{-14}$ where the automatically determined value $k_{\text{noise}} = 16$ is one less than the value determined in Sect. 3. The error is, however, negligible; see Fig. 14.

4.3 A comment on regularization and stopping criteria

When solving discrete ill-posed problems using hybrid methods based on the Golub-Kahan iterative bidiagonalization, for sufficiently large k, $k \ll n$, the absolute value of the first component of the left singular vector of L_k corresponding to its smallest singular value (almost) stagnates close to the *level of the noise present in the original data*. The beginning of the stagnation determines $|(p_1^{(k_{\text{noise}}+1)}, e_1)| \approx \delta_{\text{noise}}$ and the iteration k_{noise} when the noise level is revealed; see (4.6). Moreover, (4.8) gives the additional (secondary) noise level estimate.

Knowing the noise level, many different approaches can be applied in the subsequent steps, cf. [27, Sect. 3.2] including the straightforward application of the discrepancy principle [32, 33] (see also [18, Chap. 7.2, pp. 179–181]). It remains to determine which of these are effective in this context. Results in this direction will be reported elsewhere.

5 Noise propagation and the loss of orthogonality in the bidiagonalization vectors

To our knowledge, the effects of rounding errors have up to now not been thoroughly investigated in the literature on hybrid methods for solving discrete ill-posed problems, although they are sometimes acknowledged as a potential difficulty. As illustrated on Figs. 10 and 11, loss of orthogonality and subsequently loss of linear independence among the computed bidiagonalization vectors very significantly affect the propagation of the noise in the Golub-Kahan iterative bidiagonalization process. The corresponding Lanczos process without reorthogonalization computes multiple approximations to the well-separated squared large singular values. Consequently, the convergence of the Ritz values in the other parts of the spectrum can be significantly delayed, which affects convergence of the hybrid methods to the regularized solutions. The noise-revealing phenomenon is then complicated by the fact that computation of multiple approximations for the well-separated large singular values is connected with reappearance of the smooth components in the computed left bidiagonalization vectors s_k, which makes the propagation of the noise in the bidiagonalization process rather irregular. This is illustrated in Figs. 15–18, analogous to

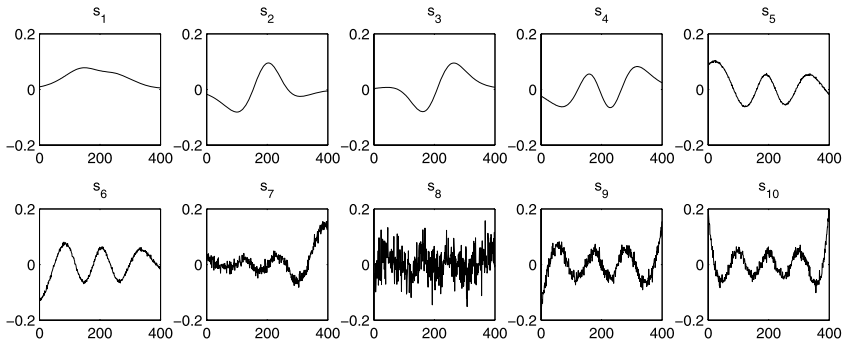


Fig. 15 Individual components of several left bidiagonalization vectors s_k computed by the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$

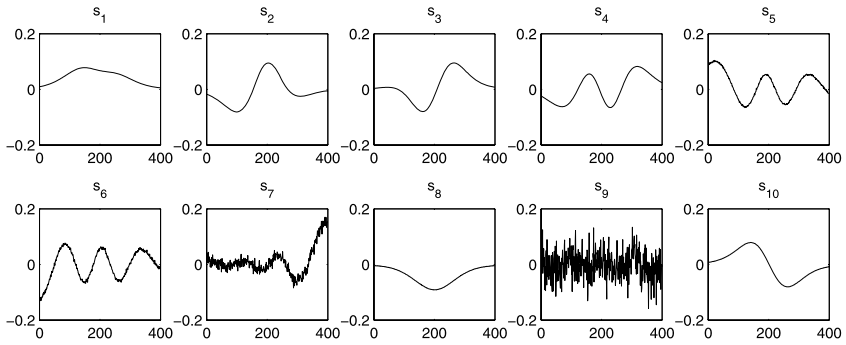


Fig. 16 Individual components of several left bidiagonalization vectors s_k computed by the Golub-Kahan iterative bidiagonalization without reorthogonalization for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$

Figs. 7, 10 and 11. Figures 15 and 16 show individual components of several left bidiagonalization vectors s_k computed by the Golub-Kahan iterative bidiagonalization *with double reorthogonalization and without reorthogonalization*, respectively, for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$. Figure 17 presents the norms of the components s_k^{exact}, s_k^{noise} and the smallest singular value of the matrix S_k as k increases. Finally, Fig. 18 gives the values of the normalization coefficients α_k, β_{k+1} and of their cumulated ratio ρ_k, see (3.9). We see that here the appearance of the smooth left bidiagonalization vector s_8 in the Golub-Kahan iterative bidiagonalization without reorthogonalization *delays* the revealing of the noise level by one iteration. In more difficult examples the loss of orthogonality among the bidiagonalization vectors can significantly delay the noise-revealing process. Nevertheless, as illustrated in Fig. 19 presenting results computed by the Golub-Kahan iterative bidiagonalization without reorthogonalization, the determination of the noise level based on monitoring the absolute value of the first component of the left singular vector

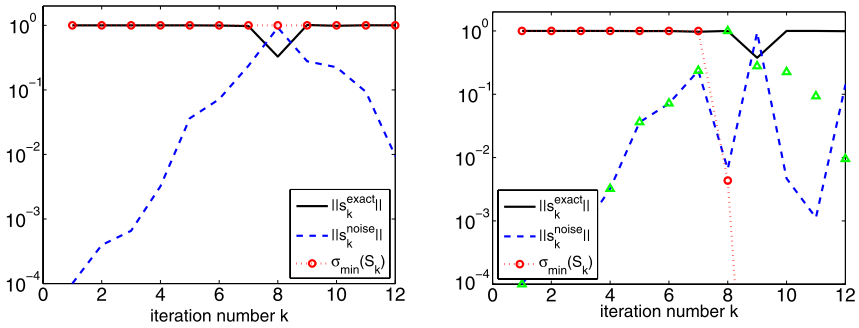


Fig. 17 The norms of s_k^{exact}, s_k^{noise} and the smallest singular value of the matrix S_k for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$, computed by the Golub-Kahan iterative bidiagonalization with double reorthogonalization (**a**), and without reorthogonalization (**b**). For comparison, *the triangles* in (**b**) represent the norm of the component s_k^{noise} computed with double reorthogonalization

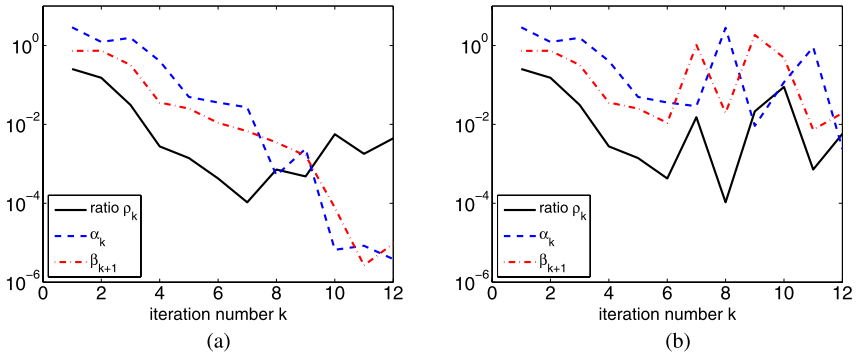


Fig. 18 Normalization coefficients α_k, β_{k+1} and their cumulated ratio ρ_k, see (3.9), for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$, computed by the Golub-Kahan iterative bidiagonalization with double reorthogonalization (**a**), and without reorthogonalization (**b**)

of L_k corresponding to its smallest singular value still works. The k_{noise} determined from Fig. 19(a) is, however, shifted by one step in comparison to Fig. 16, where $k_{\text{noise}} = 8$.

It is worth noting that the propagation of noise in the Golub-Kahan iterative bidiagonalization for discrete ill-posed problems is different from the propagation of the elementary finite precision rounding errors which are present in each iteration. An investigation of this interesting topic is, however, outside the scope of this paper.

6 Conclusion

This paper considers discrete ill-posed problems represented by the linear system (1.1) with the following properties:

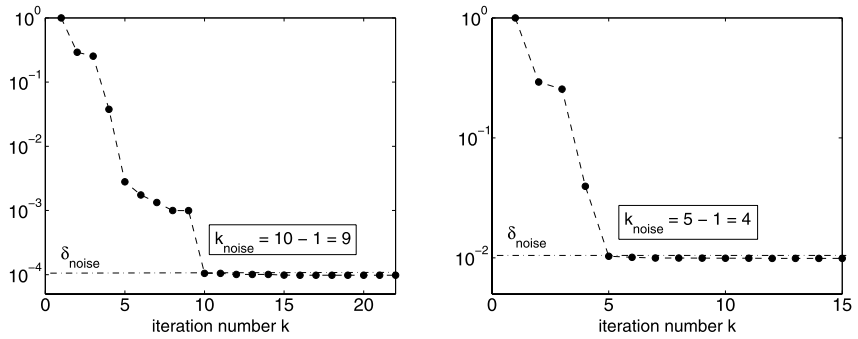


Fig. 19 The absolute value of the first component $|(p_1^{(k)}, e_1)|$ of the left singular vector of L_k corresponding to its smallest singular value for the problem `shaw(400)` with the noise level $\delta_{\text{noise}} = 10^{-4}$ (**a**) and 10^{-2} (**b**). The matrix L_k was computed using the Golub-Kahan iterative bidiagonalization without reorthogonalization. The *horizontal dashed-dotted line* represents the noise level δ_{noise}

- the matrices A, A^T, AA^T have a smoothing property;
- the left singular vectors u_j of A represent increasing frequencies as j increases;
- related to the last point, the system satisfies the discrete Picard condition; on average, the absolute value of the projections of the exact right-hand side b^{exact} to the left singular subspaces of A decays faster than the corresponding singular values;
- the noise component b^{noise} in the right-hand side represents white noise.

We showed that for this class of ill-posed problems it is possible to identify the iteration when the noise present in the data begins to propagate significantly to the projected problem computed by the Golub-Kahan iterative bidiagonalization. The unknown level of the noise in the original data can be estimated at a negligible cost from the absolute value of the first component of the left singular vector of the bidiagonal matrix of the projected system corresponding to its smallest singular value. This estimate is reliable and accurate. It can also be subsequently compared with the secondary estimate which uses the computed bidiagonalization coefficients. The estimated noise level can be used for construction of efficient stopping criteria based on many different approaches. We emphasize that throughout the paper the assumption on the white noise character of b^{noise} is substantial. Possible generalizations to cases with colored noise of varying degrees of dispersion, but dominated by high frequencies, need further investigation.

It is worth recognizing that if any of the assumptions made is not satisfied, then the presented noise revealing techniques may not succeed. Figures 20 and 21 show individual components of several left bidiagonalization vectors s_k computed by the double reorthogonalized Golub-Kahan iterative bidiagonalization, and their components in the basis of the left singular vectors of A, for the problem `phillips(256)` from the Regularization Toolbox [17] with the noise level 10^{-6}. Even though we observe an increase of the high frequency components in the vectors $s_1, s_2, \ldots$, it is not possible to identify the noise revealing iteration k_{noise} so clearly as proposed in Sect. 3. The reason is that in `phillips(256)` the discrete Picard condition is par-

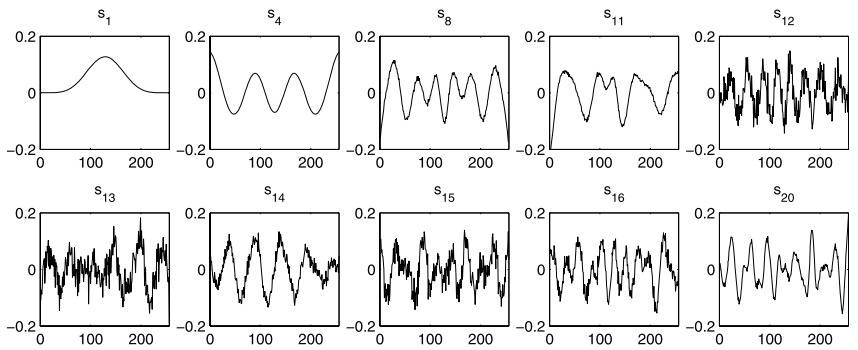


Fig. 20 Individual components of several left bidiagonalization vectors s_k computed using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `phillips(256)` with the noise level $\delta_{\text{noise}} = 10^{-6}$

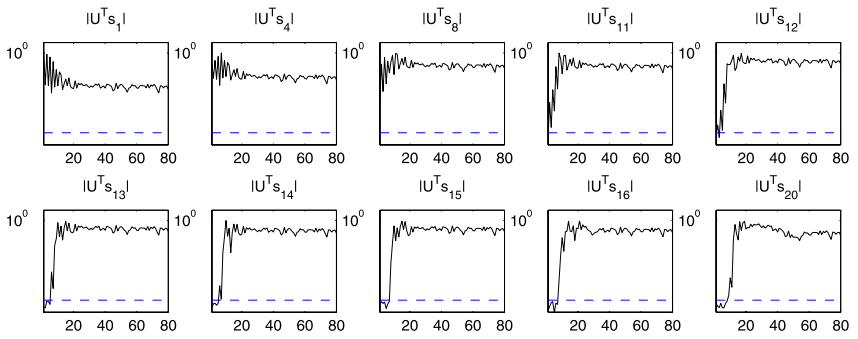


Fig. 21 The absolute values of the first 80 spectral components of the vectors s_k computed using the double reorthogonalized Golub-Kahan iterative bidiagonalization for the problem `phillips(256)` with the noise level $\delta_{\text{noise}} = 10^{-6}$. The *dashed line* represents machine precision ε_M

tially violated due to oscillations of the absolute value of the individual components. The technique presented in [20, 46] based on the Fourier analysis of the residual vectors works well also for this problem.

In this paper we have proposed noise revealing tools which may have potentially wide application for the solution of discrete ill-posed problems. The numerical experiments presented in this paper report only results obtained for the problem `shaw` (and to some extent also `ilaplace` and `phillips`) from the Regularization Toolbox [17]. Further experiments were performed with `ilaplace`, `phillips`, `deriv2`, and, without reorthogonalization, with ODF (here the matrix is rectangular of dimensions 5290×3375; see [21]) and with image deblurring examples Elephant (square problem of dimension 152280) and Barbara (two square problems of dimensions 65536 and 262144), cf. http://www.cs.cas.cz/krylov, section 'Software'. Based on the results, which can be found on the given www page, we believe that the conclusions derived from our observations and their mathematical justification offered

in Sects. 3 and 4 will find wide applications subject to the assumptions given above. We have also formulated several open questions which will be the subject of further investigation. The paper presents a step in understanding the noise revealing and regularizing properties of the Golub-Kahan iterative bidiagonalization. An application of the methods presented here to large scale problems needs further work. Results will be reported elsewhere.

Throughout the paper we have assumed A square and nonsingular. An extension of the results presented in this paper to problems with rectangular and even rank deficient matrices would require an additional assumption on the norm of the component of the right side b in the nullspace of A^T.

Acknowledgements We wish to thank Petr Tichý for his advice concerning numerical experiments and Gerard Meurant for useful comments. We are indebted to Per Christian Hansen for providing his codes and test examples used in additional tests, to Chris Paige for suggestion several corrections and clarifications, and to two anonymous referees for the very insightful remarks and suggestions which has led to significant improvements of our presentation.

Appendix

The terminology "*Golub-Kahan iterative bidiagonalization*" used throughout the paper is worth a short explanation. In their seminal paper [14], Golub and Kahan proposed two approaches for orthogonal bidiagonalization of a given matrix. In the first approach the bidiagonalization is computed by a sequence of Householder transformations from the left and right of A. The second approach, called here the Golub-Kahan iterative bidiagonalization, is introduced in the paper by the following words (here we use, for consistency, our notation for the lower bidiagonalization of A, while in [14] the authors consider the upper bidiagonalization):

> An alternative approach to bidiagonalization of A is to generate the columns of S and W sequentially as is done by the Lanczos algorithm for tridiagonalizing a symmetric matrix. The equation
>
> $$AW = SL \quad \text{and} \quad S^T A = LW^T$$
>
> can be expanded in terms of columns s_i of S and w_i of W to yield ...,

where W, S and L denote the results of the full bidiagonalization; see [14, p. 210]. For computation of the singular values of the bidiagonal matrix their paper refers to the idea of computation of the eigenvalues of the augmented matrix, which leads (through the Lanczos algorithm) to computation of the eigenvalues of the symmetric tridiagonal matrix, with reference to [29, Chap. 3].

Golub and Kahan gave by their reference to Lanczos an example of fairness which should be appreciated and followed. But their iterative bidiagonalization algorithm can not be attributed, according to our opinion, to Lanczos. The iterative bidiagonalization was proposed in [14]. Though the term "Lanczos bidiagonalization" is widespread in a part of literature, we concur with the remaining literature, in particular with [41, 42], that it is appropriate to use the name Golub-Kahan iterative bidiagonalization.

References

1. Björck, Å.: A bidiagonalization algorithm for solving large sparse ill-posed systems of linear equations. BIT **28**, 659–670 (1988)
2. Björck, Å., Grimme, E., Van Dooren, P.: An implicit shift bidiagonalization algorithm for ill-posed systems. BIT **34**, 510–534 (1994)
3. Barlow, J.L., Bosner, N., Drmač, Z.: A new stable bidiagonal reduction algorithm. Linear Algebra Appl. **397**, 35–84 (2005)
4. Calvetti, D., Golub, G.H., Reichel, L.: Estimation of the L-curve via Lanczos bidiagonalization. BIT **39**, 603–619 (1999)
5. Calvetti, D., Morigi, S., Reichel, L., Sgallari, F.: Tikhonov regularization and the L-curve for large discrete ill-posed problems. J. Comput. Appl. Math. **123**, 423–446 (2000)
6. Calvetti, D., Reichel, L., Shuibi, A.: L-curve and curvature bounds for Tikhonov regularization. Numer. Algorithms **35**, 301–314 (2004)
7. Chung, J., Nagy, J.G., O'Leary, D.P.: A weighted GCV method for Lanczos hybrid regularization. Electron. Trans. Numer. Anal. **28**, 149–167 (2008)
8. Cooley, J.W., Tukey, J.W.: An algorithm for the machine computation of the complex Fourier series. Math. Comput. **19**, 297–301 (1965)
9. Duhamel, P., Vetterli, M.: Fast Fourier transforms: A tutorial review and a state of the art. Signal Process. **19**, 259–299 (1990)
10. Fierro, R.D., Golub, G.H., Hansen, P.C., O'Leary, D.P.: Regularization by truncated total least squares. SIAM J. Sci. Statist. Comput. **18**, 1225–1241 (1997)
11. Fischer, B.: Polynomial Based Iteration Methods for Symmetric Linear Systems. Wiley-Teubner Series Advances in Numerical Mathematics. Wiley, New York (1996)
12. Fischer, B., Freund, R.W.: On adaptive weighted polynomial preconditioning for Hermitian positive definite matrices. SIAM J. Sci. Comput. **15**, 408–426 (1994)
13. Gautschi, W.: Gauss-Christoffel Quadrature Formulae. In: E.B. Christoffel, Aachen/Monschau, 1979, pp. 72–147. Birkhäuser, Basel (1981)
14. Golub, G.H., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. SIAM J. Numer. Anal. Ser. B **2**, 205–224 (1965)
15. Golub, G.H., Von Matt, U.: Generalized cross-validation for large scale problems. J. Comput. Graph. Stat. **6**, 1–34 (1997)
16. Hanke, M.: On Lanczos based methods for regularization of discrete ill-posed problems. BIT **41**, 1008–1018 (2001)
17. Hansen, P.C.: Regularization Tools—version 3.2 for MATLAB 6.0, a package for analysis and solution of discrete ill-posed problems, http://www2.imm.dtu.dk/~pch/Regutools/index.html
18. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems, Numerical Aspects of Linear Inversion. SIAM, Philadelphia (1998)
19. Hansen, P.C., Jensen, T.K.: Noise propagation in regularizing iterations for image deblurring. ETNA **31**, 204–220 (2008)
20. Hansen, P.C., Kilmer, M.E., Kjeldsen, R.: Exploiting residual information in the parameter choice for discrete ill-posed problems. BIT **46**, 41–59 (2006)
21. Hansen, P.C., Sørensen, H.O., Sükösd, Z., Poulsen, H.F.: Reconstruction of single-grain orientation distribution functions for crystalline materials. SIAM J. Image Sci. **2**(2), 593–613 (2009)
22. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Stand. **49**, 409–435 (1952)
23. Hnětynková, I., Strakoš, Z.: Lanczos tridiagonalization and core problems. Linear Algebra Appl. **421**, 243–251 (2007)
24. Hnětynková, I., Plešinger, M., Strakoš, Z.: Lanczos tridiagonalization, Golub-Kahan bidiagonalization and core problem. PAMM **6**, 717–718 (2006)
25. Jensen, T.K., Hansen, P.C.: Iterative regularization with minimum-residual methods. BIT **47**, 103–120 (2007)
26. Karlin, S., Shapley, L.S.: Geometry of Moment Spaces. American Mathematical Society, Providence (1953)
27. Kilmer, M.E., O'Leary, D.P.: Choosing regularization parameters in iterative methods for ill-posed problems. SIAM J. Matrix Anal. Appl. **22**, 1204–1221 (2001)
28. Kilmer, M.E., Hansen, P.C., Español, M.I.: A projection-based approach to general form Tikhonov regularization. SIAM J. Sci. Comput. **29**, 315–330 (2006)

29. Lanczos, C.: Linear Differential Operators. Van Nostrand, London (1961)
30. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Stand. **45**, 255–282 (1950)
31. Meurant, G., Strakoš, Z.: The Lanczos and conjugate gradient algorithms in finite precision arithmetic. Acta Numer. **15**, 471–542 (2006)
32. Morozov, V.A.: On the solution of functional equations by the method of regularization (in Russian). Sov. Math. Dokl. **7**, 414–417 (1966)
33. Morozov, V.A.: Methods for Solving Incorrectly Posed Problems. Springer, New York (1984)
34. Nguyen, N., Milanfar, P., Golub, G.H.: Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. IEEE Trans. Image Proces. **10**, 1299–1308 (2001)
35. O'Leary, D.P.: Near-optimal parameters for Tikhonov and other regularization methods. SIAM J. Sci. Comput. **23**, 1161–1171 (2001)
36. O'Leary, D.P., Simmons, J.A.: A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems. SIAM J. Sci. Stat. Comput. **2**, 474–489 (1981)
37. O'Leary, D.P., Strakoš, Z., Tichý, P.: On sensitivity of Gauss-Christoffel quadrature. Numer. Math. **107**, 147–174 (2007)
38. Paige, C.C.: Bidiagonalization of matrices and solution of linear equations. SIAM J. Numer. Anal. **11**, 197–209 (1974)
39. Paige, C.C.: Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. Linear Algebra Appl. **34**, 235–258 (1980)
40. Paige, C.C.: A useful form of unitary matrix obtained from any sequence of unit 2-norm n-vectors. SIAM J. Matrix Anal. Appl. **31**, 565–583 (2009)
41. Paige, C.C., Saunders, M.A.: LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Softw. **8**, 43–71 (1982)
42. Paige, C.C., Saunders, M.A.: ALGORITHM 583 LSQR: Sparse Linear equations and least squares problems. ACM Trans. Math. Softw. **8**, 195–209 (1982)
43. Paige, C.C., Strakoš, Z.: Scaled total least squares fundamentals. Numer. Math. **91**, 117–146 (2002)
44. Paige, C.C., Strakoš, Z.: Unifying least squares, total least squares and data least squares. In: Van Huffel, S., Lemmerling, P. (eds.) Total Least Squares and Errors-in-Variables Modeling, pp. 25–34. Kluwer Academic, Dordrecht (2002)
45. Paige, C.C., Strakoš, Z.: Core problem in linear algebraic systems. SIAM J. Matrix Anal. Appl. **27**, 861–875 (2006)
46. Rust, B.W.: Parameter selection for constrained solutions to ill-posed problems. Comput. Sci. Stat. **32**, 333–347 (2000)
47. Rust, B.W., O'Leary, D.P.: Residual periodograms for choosing regularization parameters for ill-posed problems. Inverse Probl. **24** (2008). doi:10.1088/0266-5611/24/3/034005
48. Saunders, M.A.: Computing projections with LSQR. BIT **37**, 96–104 (1997)
49. Sima, D.M., Van Huffel, S.: Using core formulations for ill-posed linear systems. PAMM **5**, 795–796 (2005)
50. Sima, D.M.: Regularization techniques in model fitting and parameter estimation. Ph.D. thesis, Dept. of Electrical Engineering, Katholieke Universiteit Leuven (2006)
51. Simon, H.D., Zha, H.: Low-rank matrix approximation using the Lanczos bidiagonalization process with applications. SIAM J. Sci. Stat. Comput. **21**, 2257–2274 (2000)
52. Strakoš, Z.: Model reduction using the Vorobyev moment problem. Numer. Algorithms **51**, 363–376 (2009)

Tenzorové výpočty

A.6 Článek:[WoK] Daniel Kressner, Martin Plešinger, Christine Tobler: *A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations*, Numerical Linear Algebra with Applications (NLAA) (ISSN 1070-5325, eISSN 1099-1506), Volume 21, Issue 5 (2014), pp. 666–684 (19 pages).
(http://onlinelibrary.wiley.com/doi/10.1002/nla.1919/abstract)

Preprint: EFPL Mathicse technical report № 18.2012 (rozšířená verze článku)
(http://mathicse.epfl.ch/files/content/sites/mathicse/files [...] /Mathicse reports 2012/18.2012_DK-MP-CT.pdf)

Doložitelné citace (8): Odborné články zařazené do databáze Web-of-Knowledge:

- [WoK] D. Kressner[AutoCit], P. Sirković: *Truncated low-rank methods for solving general linear matrix equations*, Numerical Linear Algebra with Applications, 22(3) (2015), pp. 564–583.
(http://onlinelibrary.wiley.com/doi/10.1002/nla.1973/abstract)
- [WoK] D. Niu, Y. Lu, X. Xu, B. Li: *Short-term power load point prediction based on the sharp degree and chaotic RBF neural network*, Mathematical Problems in Engineering, 2015, Article ID 231765 (2015), 8 pages.
(http://www.hindawi.com/journals/mpe/2015/231765)
- [WoK] D. Kressner[AutoCit], M. Steinlechner, B. Vandereycken: *Preconditioned low-rank Riemannian optimization for linear systems with tensor product structure*, SIAM Journal on Scientific Computing, 38(4) (2016), pp. A2018–A2044.
(http://epubs.siam.org/doi/abs/10.1137/15M1032909)
- [WoK] J. Ballani, D. Kressner[AutoCit]: *Reduced basis methods: From low rank matrices to low rank tensors*, SIAM Journal on Scientific Computing, 38(4) (2016), pp. A2045–A2067.
(http://epubs.siam.org/doi/abs/10.1137/15M1042784)

Další citace:

- [Preprint] M. Bollhöfer, A. K. Eppler: *Low-rank Cholesky factor Krylov subspace methods for generalized projected Lyapunov equations*, preprint of TU Braunschweig, Braunschweig, 2012, 32 pages.
(http://www.icm.tu-bs.de/~bolle/Publicat/tp6.pdf)
- [MathSciNet] L. Grasedyck, D. Kressner[AutoCit], C. Tobler[AutoCit]: *A literature survey of low-rank tensor approximation techniques*, GAMM Mitteilungen, 36(1) (2013), pp. 53–78.
(http://onlinelibrary.wiley.com/doi/10.1002/gamm.201310004/abstract)
- [Preprint] N. T. Son, T. Stykel: *Solving parameter-dependent Lyapunov equations using reduced basis method with application to parametric model order reduction*, preprint of Augsburg university, Augsburg, 2015, 26 pages. (https://opus.bibliothek.uni-augsburg.de/opus4/files [...] /3155/mpreprint_15_005.pdf)
- [PhD-thesis] M. M. Steinlechner: *Riemannian Optimization for Solving High-Dimensional Problems with Low-Rank Tensor Structure*, PhD Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 2016.
(http://infoscience.epfl.ch/record/217938/files/EPFL_TH6958.pdf)

NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2014; **21**:666–684
Published online 17 December 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nla.1919

A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations

Daniel Kressner [1], Martin Plešinger [2,*,†] and Christine Tobler [1]

[1] *ANCHP, MATHICSE, EPF Lausanne, Lausanne, Switzerland*
[2] *Department of Mathematics and Didactics of Mathematics, TU Liberec, Liberec, Czech Republic*

SUMMARY

This paper is concerned with the numerical solution of symmetric large-scale Lyapunov equations with low-rank right-hand sides and coefficient matrices depending on a parameter. Specifically, we consider the situation when the parameter dependence is sufficiently smooth, and the aim is to compute solutions for many different parameter samples. On the basis of existing results for Lyapunov equations and parameter-dependent linear systems, we prove that the tensor containing all solution samples typically allows for an excellent low multilinear rank approximation. Stacking all sampled equations into one huge linear system, this fact can be exploited by combining the preconditioned CG method with low-rank truncation. Our approach is flexible enough to allow for a variety of preconditioners based, for example, on the sign function iteration or the alternating direction implicit method. Copyright © 2013 John Wiley & Sons, Ltd.

Received 14 April 2012; Revised 7 May 2013; Accepted 3 November 2013

KEY WORDS: Lyapunov equations; CG method; preconditioning; ADI preconditioner; sign function preconditioner; tensors in Tucker format; model reduction

1. INTRODUCTION

Let us consider a Lyapunov matrix equation

$$A(\alpha)X(\alpha)M(\alpha)^T + M(\alpha)X(\alpha)A(\alpha)^T = B(\alpha)B(\alpha)^T, \tag{1}$$

where the coefficient matrices $A(\alpha), M(\alpha) \in \mathbb{R}^{n\times n}$, $B(\alpha) \in \mathbb{R}^{n\times t}$, and consequently also the solution matrix $X(\alpha) \in \mathbb{R}^{n\times n}$, depend on a parameter $\alpha \in \mathbb{R}$.

We are concerned with the problem of solving (1) for possibly many parameter samples. This is needed in interpolatory model reduction techniques for parameterized linear control systems, see [1, 2]. For the rest of this paper, we will assume that $A(\alpha)$ and $M(\alpha)$ are both symmetric positive definite for all parameter values α of interest. In particular, this implies that (1) has a unique symmetric positive definite solution. Our assumption is satisfied, for example, for Lyapunov equations (1) arising from the finite element discretization of an infinite-dimensional linear control system governed by a parameter-dependent parabolic PDE. In this case, $A(\alpha)$ and $M(\alpha)$ correspond to the stiffness and mass matrices, respectively. We will comment on extensions to the nonsymmetric case in Section 4.

For the rest of this paper, we suppose that the right-hand side of (1) has low rank, that is, $t \ll n$. Under our assumptions, this implies that $X(\alpha)$ admits an excellent low-rank approximation for

*Correspondence to: Martin Plešinger, Department of Mathematics and Didactics of Mathematics, TU Liberec, Liberec, Czech Republic.

†E-mail: martin.plesinger@tul.cz

fixed α, see, for example, [3–6]. Virtually all existing algorithms for large-scale Lyapunov equations exploit this observation. This includes the low-rank Smith iteration and alternating direction implicit (ADI) method [7, 8], subspace methods [9–13], and low-rank variants of the sign function iteration [14, 15]. All these methods deal efficiently with a single Lyapunov equation evaluated at an individual parameter sample, but none of them can be extended in an obvious way to deal efficiently with many parameter samples.

In this paper, we propose new Krylov subspace based techniques to solve (1) for many parameter samples simultaneously. For this purpose, we combine our recently developed low-rank techniques for solving parameter-dependent linear systems [4] with low-rank techniques for Lyapunov equations. For this purpose, we proceed as follows.

In Section 2, we consider (1) for a fixed parameter sample and treat it as a Kronecker-structured $n^2 \times n^2$ linear system. This view, which has already been promoted in [16], allows more flexibility in the choice of the solver and the preconditioner. More specifically, we combine the standard conjugate gradient (CG) method with preconditioners inspired by existing methods for solving Lyapunov equations. One obvious disadvantage of this approach is that each iterate in the CG method is a vector of length n^2, which is infeasible for large-scale applications. This can be avoided by applying low-rank truncations to the iterates, an idea that has been successfully used in [17–19].

In Section 3, the approach from Section 2 is extended to $m > 1$ parameter samples by considering all m Lyapunov equations simultaneously in one huge block diagonal linear system of size $mn^2 \times mn^2$. Again, a CG method is applied to solve this system. However, instead of low-rank matrix truncation, we now consider the iterates of length mn^2 as third-order tensors of size $n \times n \times m$ and apply multilinear low-rank approximation [20]. The success of this approach crucially depends on the approximability of the solution tensor. In the case of smooth dependence on a single parameter, we prove rapid decay of the approximation error for increasing multilinear ranks. This approach is also well suited for several parameters, provided that the number of samples does not grow too large. This can be achieved for several parameters by sparse collocation techniques, see for example [21].

In the technical report [22] accompanying this paper, we describe an alternative approach for $p > 1$ parameters. Assuming that the samples are arranged in a tensor grid, the solutions of (1) are collected into a tensor of order $2 + p$, where the first two modes correspond to the rows/columns of the solutions and each of the remaining p modes corresponds to a parameter. The associated linear system is then solved by combining CG with low-rank truncation in the so called hierarchical Tucker format [23, 24].

Remark 1
All numerical experiments in this paper have been performed in MATLAB version 7.8 (R2009a) on an Intel Core2 Duo (T8300) 2.40 GHz processor.

2. NO PARAMETER

We first consider (1) for a fixed parameter sample. For simplicity, we omit the dependence on the parameter:

$$AXM^T + MXA^T = BB^T, \tag{2}$$

where A and M are both symmetric positive definite. It is well known that (2) can be cast as a Kronecker product linear system

$$(M \otimes A + A \otimes M)x = b, \tag{3}$$

with $x = \mathrm{vec}(X)$ and $b = \mathrm{vec}(BB^T)$, where $\mathrm{vec}(\cdot)$ stacks the columns of an $n \times n$ matrix into a vector of length n^2.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

2.1. The basic form of preconditioned CG

As the matrix in the linear system (3) is symmetric positive definite, we can apply the preconditioned CG method to (3). We will base our preconditioner on existing methods for solving a standard Lyapunov equation of the form

$$\bar{A}\bar{X} + \bar{X}\bar{A}^T = \bar{B}\bar{B}^T. \tag{4}$$

Note that (2) and (4) are equivalent via the relations $\bar{A} = L_M^{-1} A L_M^{-T}$, $\bar{B} = L_M^{-1} B$, and $X = L_M^{-T} \bar{X} L_M^{-1}$, with the Cholesky factorization $M = L_M L_M^T$. Given a preconditioner $\mathcal{P}^{-1}$ for

$$\bar{\mathcal{A}} := I \otimes \bar{A} + \bar{A} \otimes I \tag{5}$$

the Kronecker product formulation of (4), a preconditioner for (3) is obtained as

$$\left(L_M^{-1} \otimes L_M^{-1}\right) \mathcal{P}^{-1} \left(L_M^{-T} \otimes L_M^{-T}\right). \tag{6}$$

Algorithm 1 is the standard CG method [25] applied to (3) with the preconditioner (6). The only difference to the standard formulation is that we recast all operations in terms of $n \times n$ matrices instead of vectors of length n^2. In particular, the inner product $\langle \cdot, \cdot \rangle$ should be understood as the matrix inner product and the preconditioner $\mathcal{P}^{-1}$ is considered as a linear operator on $\mathbb{R}^{n \times n}$. Such an approach is not new; it belongs to the class of so called global Krylov subspace methods, see, for example, [26].

Algorithm 1 Conjugate gradient method for solving $AXM^T + MXA^T = BB^T$.

Require: Symmetric positive definite matrices $A, M \in \mathbb{R}^{n \times n}$ and right-hand side matrix $B \in \mathbb{R}^{n \times t}$, tolerance $\mathsf{tol} > 0$.

Ensure: Approximation X_k to solution of Lyapunov equation (2).

1: $L_M \leftarrow \mathtt{chol}\,(M)$ {(sparse) Cholesky decomposition}
2: $k \leftarrow 0$,
3: $R_0 \leftarrow BB^T$ {initial residual corresponding to (2)}
4: Compute $\bar{R}_0 \leftarrow L_M^{-1} R_0 L_M^{-T}$.
5: Compute $\bar{Z}_0 = \mathcal{P}^{-1}\left(\bar{R}_0\right)$. {apply preconditioner for (4)}
6: Compute $Z_0 \leftarrow L_M^{-1} \bar{Z}_0 L_M^{-T}$.
7: $\rho_0^{\mathrm{CG}} \leftarrow \langle R_0, Z_0 \rangle$
8: $P_0 \leftarrow Z_0$
9: **repeat**
10: $k \leftarrow k + 1$
11: $W_k \leftarrow AP_{k-1}M^T + MP_{k-1}A^T$ {apply Lyapunov operator}
12: $\alpha_k^{\mathrm{CG}} \leftarrow \rho_{k-1}^{\mathrm{CG}} / \langle W_k, P_{k-1} \rangle$
13: $X_k \leftarrow X_{k-1} + \alpha_k^{\mathrm{CG}} P_{k-1}$
14: $R_k \leftarrow R_{k-1} - \alpha_k^{\mathrm{CG}} W_k$
15: Compute $\bar{R}_k \leftarrow L_M^{-1} R_k L_M^{-T}$.
16: Compute $\bar{Z}_k = \mathcal{P}^{-1}\left(\bar{R}_k\right)$. {apply preconditioner for (4)}
17: Compute $Z_k \leftarrow L_M^{-1} \bar{Z}_k L_M^{-T}$.
18: $\rho_k^{\mathrm{CG}} \leftarrow \langle R_k, Z_k \rangle$
19: $\beta_k^{\mathrm{CG}} \leftarrow \rho_k^{\mathrm{CG}} / \rho_{k-1}^{\mathrm{CG}}$
20: $P_k \leftarrow Z_k + \beta_k^{\mathrm{CG}} P_{k-1}$
21: **until** $\|BB^T - (AX_kM^T + MX_kA^T)\|_F < \mathsf{tol}$ {test true residual for convergence}

It is well known that the solution X of the Lyapunov equation (2) is symmetric. It turns out that Algorithm 1 automatically preserves this symmetry. More specifically, if the preconditioner $\mathcal{P}^{-1}$ applied to a symmetric matrix again results in a symmetric matrix, it can be easily seen that all iterates P_k, R_k, X_k, W_k, Z_k generated by Algorithm 1 are symmetric $n \times n$ matrices.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

2.2. Incorporating low-rank truncation into the preconditioned CG method

A serious drawback of Algorithm 1 is that the storage requirements are $O(n^2)$ as the iterates are generally dense $n \times n$ matrices. Motivated by the facts that the right-hand side BB^T has low rank (as we assumed $t \ll n$) and that the solution X can be well approximated by a low-rank matrix, we expect the iterates to be also approximable by low-rank matrices. The ranks of iterates can, however, grow in a transient phase of convergence, see Figure 1, right (compare also to Figure 2). This rank growth has to be avoided using a suitable preconditioner. We will explicitly enforce low rank by truncating the iterates repeatedly.

Any symmetric matrix $S \in \mathbb{R}^{n\times n}$ of rank $r \ll n$ can be stored in $O(nr)$ memory by means of a decomposition

$$S = U_S \Lambda_S U_S^T, \tag{7}$$

where $\Lambda_S \in \mathbb{R}^{r\times r}$ is symmetric and $U_S \in \mathbb{R}^{n\times r}$. For example, this can be achieved by the spectral decomposition of S. All iterates of Algorithm 1 will be represented in the factored form (7). This allows all operations needed in Algorithm 1 to be performed efficiently:

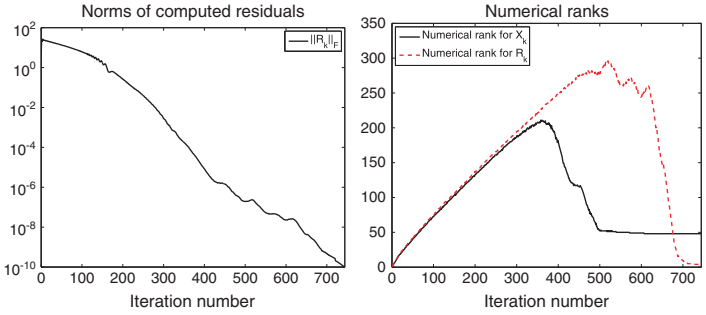


Figure 1. Algorithm 1 applied to the test problem, see Remark 3 or [27, Sec. 2.3.1]. The matrices A and M are the stiffness and mass matrices, respectively, $n = 11,036$. The matrix $B \in \mathbb{R}^{n\times 1}$ is chosen randomly. *Left:* Convergence of the residual norms. *Right:* Numerical ranks for X_k and R_k for $\varepsilon = 10^{-10}$. The CG method stops after 744 iterations (4 h, 1 min, 43 s) with residual norm $4.489 \cdot 10^{-11}$.

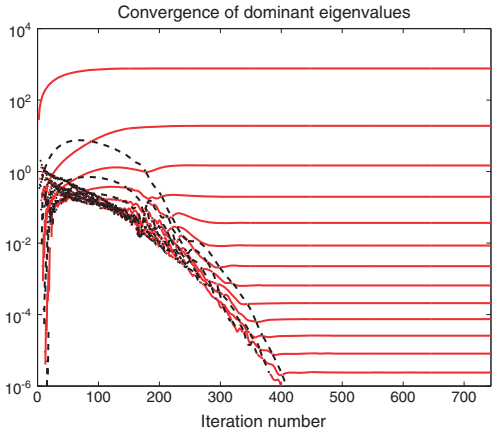


Figure 2. Evolution of 15 largest singular values of X_k as k increases. Solid lines correspond to positive eigenvalues, dashed lines to absolute values of negative eigenvalues.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Matrix multiplication. An operation of the form $\bar{S} = L_M^{-1} S L_M^{-T}$ is performed as

$$\bar{S} = L_M^{-1}\left(U_S \Lambda_S U_S^T\right) L_M^{-T} = \left(L_M^{-1} U_S\right) \Lambda_S \left(U_S^T L_M^{-T}\right) =: U_{\bar{S}} \Lambda_S U_{\bar{S}}^T,$$

not increasing the rank and requiring $O(\mathsf{nnz}(L_M)r)$ instead of $O(\mathsf{nnz}(L_M)n)$ operations, where nnz denotes the number of nonzero entries of a matrix.

Matrix addition. $\bar{S} = S + T$ is performed as

$$\begin{aligned} S + T &= U_S \Lambda_S U_S^T + U_T \Lambda_T U_T^T = [U_S, U_T] \begin{bmatrix} \Lambda_S & 0 \\ 0 & \Lambda_T \end{bmatrix} [U_S, U_T]^T \\ &=: U_{\bar{S}} \Lambda_{\bar{S}} U_{\bar{S}}^T. \end{aligned}$$

While this operation has zero cost, the rank generally increases to $r_S + r_T$, where r_S, r_T are the ranks of S, T, respectively.

Application of Lyapunov operator. $\bar{S} = ASM^T + MSA^T$ is performed in an analogous way:

$$\begin{aligned} ASM^T + MSA^T &= AU_S \Lambda_S U_S^T M^T + MU_S \Lambda_S U_S^T A^T \\ &= [AU_S, MU_S] \begin{bmatrix} 0 & \Lambda_S \\ \Lambda_S & 0 \end{bmatrix} [AU_S, MU_S]^T \\ &=: U_{\bar{S}} \Lambda_{\bar{S}} U_{\bar{S}}^T, \end{aligned}$$

doubling the rank and requiring $O(\mathsf{nnz}(M)r + \mathsf{nnz}(A)r)$ instead of $O(\mathsf{nnz}(M)n + \mathsf{nnz}(A)n)$ operations.

Matrix inner product. $\langle S, T\rangle$ is performed as

$$\langle S, T\rangle = \mathsf{tr}(ST) = \mathsf{tr}\left(U_T^T U_S \Lambda_S U_S^T U_T \Lambda_T\right),$$

where the last matrix product can be evaluated in $O(nr_S r_T)$ instead of $O(n^2)$ operations, provided that Λ_S, Λ_T are diagonal.

The only operation not covered in the list earlier is the application of the preconditioner $\mathcal{P}^{-1}$; this will be discussed in Section 2.3.

When applying Algorithm 1, the ranks of the iterates and consequently also the storage requirements will grow dramatically. This rank growth can be limited as follows. Given a factored matrix $S = U_S \Lambda_S U_S^T$ of rank r, we first perform a QR decomposition $U_S = QR$ with $Q \in \mathbb{R}^{n\times r}$ having orthonormal columns and $R \in \mathbb{R}^{r\times r}$ being upper triangular. Then,

$$S = U_S \Lambda_S U_S^T = Q\left(R\Lambda_S R^T\right) Q^T.$$

We then compute a spectral decomposition

$$R\Lambda_S R^T = [V_1, V_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} [V_1, V_2]^T,$$

where $\Lambda_1 \in \mathbb{R}^{\bar{r}\times\bar{r}}$ is a diagonal matrix containing the $\bar{r}$ eigenvalues of largest absolute magnitude. The truncated matrix $\bar{S}$ of rank $\bar{r}$ is obtained as

$$\bar{S} = U_{\bar{S}} \Lambda_{\bar{S}} U_{\bar{S}}^T, \quad \text{with} \quad U_{\bar{S}} = QV_1, \quad \Lambda_{\bar{S}} = \Lambda_1.$$

Note that $\|S - \bar{S}\|_2 = \|\Lambda_2\|_2 = \sigma_{\bar{r}+1}(S)$, where $\sigma_j(\cdot)$ denotes the jth largest singular value of a matrix. It remains to discuss the choice of the parameter $\bar{r}$, the numerical rank to which the different iterates of Algorithm 1 are truncated.

Numerical rank for X_k. For a desired user-specified accuracy $\varepsilon > 0$ and a safety factor $C_1 \leqslant 1$, we let the *numerical rank of* X_k denote the smallest j such that

$$\sigma_{j+1}(X_k) \leqslant C_1 \varepsilon \|X_k\|_2. \tag{8}$$

In our experiments, we have observed that $C_1 = 0.05$ gives good performance.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Numerical rank for R_k. As the residuals can be expected to become small with increasing k, a relative criterion of the form (8) would lead to unnecessarily high ranks in latter stages of the CG method. Instead, we let the *numerical rank of* R_k denote the smallest j such that

$$\sigma_{j+1}(R_k) \leqslant \max\{C_1\,\varepsilon\,\|R_k\|_2, C_2\,\varepsilon\,\|R_0\|_2\}, \tag{9}$$

for safety factors $C_1 \leqslant 1$ and $C_2 \leqslant 1$. Analogous criteria are used for the iterates P_k and W_k. In our experiments, we have observed that choosing $C_1 = 0$, $C_2 = 0.1$ for R_k, $C_1 = C_2 = 0.25$ for P_k, and $C_1 = 0$, $C_2 = 1$ for W_k gives good performance.

Summarizing the discussion, all iterates X_k, P_k, R_k, W_k, Z_k of Algorithm 1 are represented in the factored form (7). We apply truncation with the parameters described earlier after every operation that increases the rank of the iterates, that is, in lines 5, 11, 13, 14, 16, and 20.

Remark 2
In exact arithmetic, the matrix R_k corresponds to the residual of the approximation X_k produced in the kth step of CG. In finite-precision arithmetic, $\|R_k\|_F$ remains a faithful convergence indicator (up to machine precision) [28]. To a certain extent, this remains valid when introducing low-rank truncations. However, it is safer to use the explicitly recomputed residual for the stopping criterion in line 21 of Algorithm 1.

2.3. Preconditioners for Lyapunov equations

The convergence of Algorithm 1 is governed by classical results for the CG method. To attain fast convergence, the use of an effective preconditioner $\mathcal{P}^{-1}$ for $\bar{A}$ defined in (5) is therefore mandatory. In the context of low-rank truncation, there is another important reason for preconditioning. As illustrated in Figure 1, Algorithm 1 without any preconditioner not only suffers from slow convergence but also from a significant growth of the numerical ranks in the initial phase. As we will see later, this transient growth is diminished when using effective preconditioners.

Remark 3 (Test problem)
As an example in all the presented numerical experiments (except the last one), we consider a discretized heat equation

$$\begin{aligned} -\nabla(\sigma(x)\nabla u) &= f \qquad \text{in } \Omega = [-1,1]^2, \\ u &= 0 \qquad \text{on } \Gamma = \partial\Omega. \end{aligned}$$

The heat conductivity coefficient $\sigma(x)$ is assumed piecewise constant

$$\sigma(x) = \begin{cases} 1+\alpha & \text{for } x \in \mathcal{D}, \\ 1 & \text{for } x \notin \mathcal{D}, \end{cases}$$

where $\mathcal{D} \subset \Omega$ is a disc of radius 0.5 and $\alpha > 0$ is the parameter. This system is discretized by a finite element formulation with piecewise linear basis functions. This example is taken from [27, Sec. 2.3.1].

In the following, we will discuss various possibilities for the preconditioner. On the one hand, the preconditioner should preserve the symmetry and (approximately) low rank. The latter requirement is satisfied when $\mathcal{P}^{-1}$ can be written as a short sum of Kronecker products. This rules out, for example, the use of classical preconditioners such as Jacobi and SSOR [16]. On the other hand, the preconditioner should reduce the condition number of $\bar{A}$ significantly. Fully structure-preserving preconditioners, such as $\mathcal{P}^{-1} = (P \otimes P)^{-1}$, may not offer enough flexibility to achieve this goal, see [29] for a related discussion. We consider two preconditioners inspired by the ADI iteration and the sign function iteration.

2.3.1. Alternating direction implicit preconditioner. Preconditioning a Krylov subspace method with a few iterates of ADI was already proposed in [16] for Lyapunov equations, see also [9, 30].

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Algorithm 2 describes one cycle of ADI with ℓ real negative shifts $\varphi_1, \ldots, \varphi_\ell$. The optimal choice of shifts is discussed in [6, 31–33]. In each iteration of ADI, linear systems with the matrix $\bar{A} - \varphi_j I_n = L_M^{-1}(A - \varphi_j M)L_M^{-T}$ need to be solved. Typically, this is performed by a sparse direct solver, computing the sparse Cholesky factorization of $A - \varphi_j M$ only once in a preprocessing step.

Algorithm 2 ADI(ℓ) applied to $\bar{A}\bar{Z} + \bar{Z}\bar{A}^T = \bar{R}$.

1: $\bar{Z}^{(0)} \leftarrow 0$
2: **for** $j = 1, \ldots, \ell$
3: `solve:` $\left(\bar{A} - \varphi_j I_n\right) \bar{Z}^{(j-\frac{1}{2})} = \bar{R} - \bar{Z}^{(j-1)} \left(\bar{A} + \varphi_j I_n\right)^T$
4: `solve:` $\bar{Z}^{(j)} \left(\bar{A} - \varphi_j I_n\right)^T = \bar{R} - \left(\bar{A} + \varphi_j I_n\right) \bar{Z}^{(j-\frac{1}{2})}$
5: **end**

If the right-hand side $\bar{R}$ has low rank, then Algorithm 2 can be implemented efficiently in low-rank arithmetic. As each iteration of Algorithm 2 increases the rank, it is necessary to truncate after each iteration. While using $\ell > 1$ yields more effective preconditioners, our numerical experiments revealed that the computational time spent on low-rank repeated truncation offsets this benefit. We have therefore restricted ourselves to $\ell = 1$. In this case, Algorithm 2 reduces to

$$\bar{Z} \leftarrow -2\varphi_1 \left(\bar{A} - \varphi_1 I_n\right)^{-1} \bar{R} \left(\bar{A} - \varphi_1 I_n\right)^{-T}, \tag{10}$$

which preserves the rank of $\bar{R}$. The optimal value of the parameter φ_1 in ADI(1) is given by

$$\varphi_1 = -\sqrt{\lambda_{\max}\left(\bar{A}\right) \lambda_{\min}\left(\bar{A}\right)}, \tag{11}$$

where $\lambda_{\max}$, $\lambda_{\min}$ denote the largest/smallest eigenvalues of $\bar{A}$ and can be easily estimated by applying a few steps of the Lanczos method [33].

Figure 3 and Table I show the performance of Algorithm 1 with the preconditioner (10). Compared to Figure 1 (no preconditioner), the convergence is dramatically improved. The numerical ranks of the iterates do not grow larger than twice the numerical rank of the solution. Both improvements result into a dramatically reduced execution time: 16 s instead of 4 h. Figure 4 shows additional experiments for symmetric positive definite problems from the Oberwolfach Model Reduction Benchmark Collection [34].

2.3.2. Sign function preconditioner. Instead of ADI, one can also use a few iterations of the sign function method for solving Lyapunov equations as a preconditioner. A similar idea has been

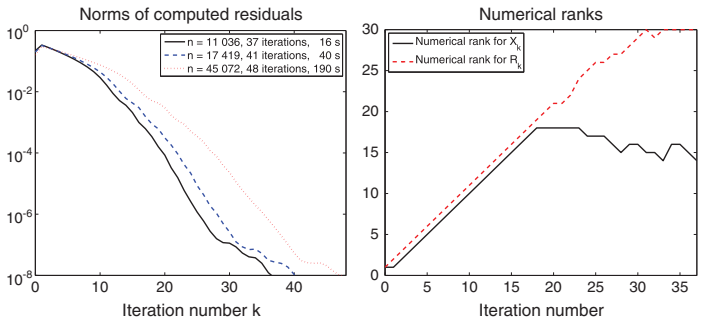


Figure 3. Algorithm 1 applied to the example from Figure 1 with the alternating direction implicit preconditioner (10) for three different mesh sizes n. *Left:* Convergence of the residual norms. *Right:* Numerical ranks for X_k and R_k for $\varepsilon = 10^{-8}$ and $n = 11,036$.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Table I. Breakdown of the total execution time. The preprocessing step for constructing the preconditioners is not included in the table and takes 0.248, 0.443, and 1.934 s, respectively.

Problem size	Plain CG (s)	Low-rank truncation (s)	Application of preconditioner (s)	Total time (s)
$n = 11,036$	1.888	10.951	3.186	16.025
$n = 17,419$	3.192	31.219	6.457	40.868
$n = 45,072$	9.835	154.122	26.533	190.491

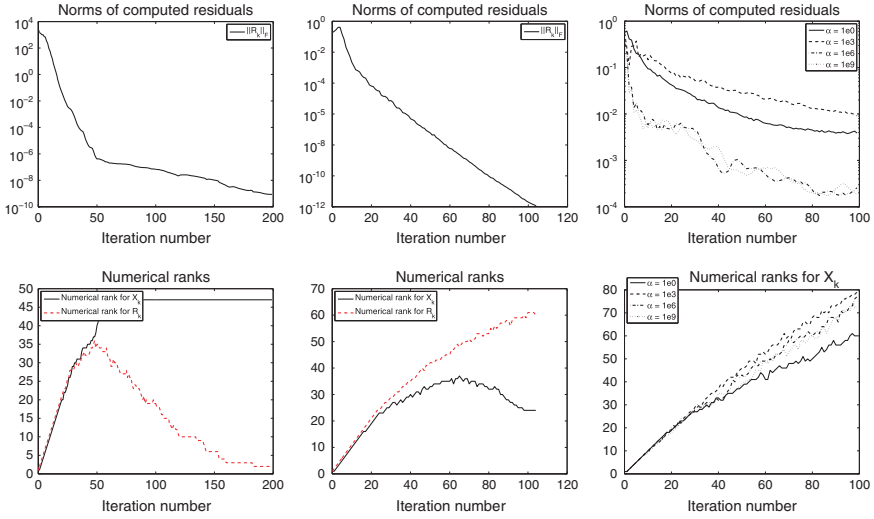


Figure 4. Algorithm 1 applied to three problems from [34]. *Left column:* Spiral Inductor PEEC Model, $n = 1434$, $\varepsilon = 10^{-12}$, the final norm of the true residual is $\|BB^T - \mathcal{A}(X_k)\|_F = 3.1 \cdot 10^{-11}$. *Middle column:* 2D Tunable Optical Filter, $n = 1668$, $\varepsilon = 10^{-12}$, the final norm of the true residual is $1.8 \cdot 10^{-18}$. *Right column:* Boundary Condition Independent Thermal Model, $n = 4257$, $\varepsilon = 10^{-6}$; the system matrix consists of four parts $A = -A_0 + \alpha(A_{\text{top}} + A_{\text{bottom}} + A_{\text{side}})$, the final norm of the true residual is $1.8 \cdot 10^{-6}$, $3.5 \cdot 10^{-6}$, $2.4 \cdot 10^{-7}$, and $2.3 \cdot 10^{-7}$, respectively.

successfully used for iterative refinement in the context of a hybrid CPU–GPU implementation [35].

Algorithm 3 performs the first ℓ iterations of the sign function method. A discussion on the choice of appropriate scaling parameters $\omega_j > 0$ can be found, for example, in [36]. These parameters can be estimated during the computation of the matrices $A^{(j)}$. In particular, for $\ell = 1$, the choice $\omega_1 = \sqrt{\lambda_{\max}(\bar{A})\,\lambda_{\min}(\bar{A})}$ is recommended, which coincides with (11).

Algorithm 3 Sign(ℓ) applied to $\bar{A}\bar{Z} + \bar{Z}\bar{A}^T = \bar{R}$.

1: $\bar{Z}^{(0)} \leftarrow \bar{R}, \qquad A^{(0)} \leftarrow \bar{A}$
2: **for** $j = 1, \ldots, \ell - 1$
3: $\quad \bar{Z}^{(j)} \leftarrow \frac{1}{2\omega_j}\left(\bar{Z}^{(j-1)} + \omega_j^2 (A^{(j-1)})^{-1} \bar{Z}^{(j-1)} (A^{(j-1)})^{-T}\right)$
4: $\quad A^{(j)} \leftarrow \frac{1}{2\omega_j}\left(A^{(j-1)} + \omega_j^2 (A^{(j-1)})^{-1}\right)$
5: **end**
6: $\bar{Z} \leftarrow \frac{1}{2\omega_\ell}\left(\bar{Z}^{(\ell-1)} + \omega_\ell^2 (A^{(\ell-1)})^{-1} \bar{Z}^{(\ell-1)} (A^{(\ell-1)})^{-T}\right)$

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

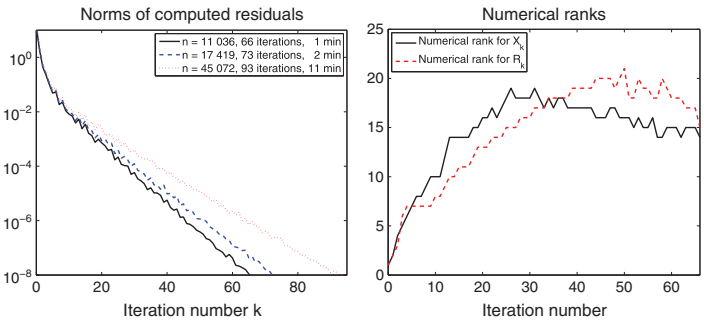


Figure 5. Algorithm 1 applied to the example from Figure 1 with the sign function preconditioner (12) for three different mesh sizes n. *Left:* Convergence of the residual norms. *Right:* Numerical ranks for X_k and R_k for $\varepsilon = 10^{-8}$ and $n = 11,036$.

Table II. Breakdown of the total execution time. The preprocessing step for constructing the preconditioners is not included in the table and takes 0.059, 0.101, and 0.332 s, respectively.

Problem size	Plain CG (s)	Low-rank truncation (s)	Application of preconditioner (s)	Total time (s)
$n = 11,036$	9.227	40.648	8.037	57.913
$n = 17,419$	17.446	101.295	18.448	137.189
$n = 45,072$	70.258	485.153	88.473	643.883

Because the iterates $A^{(j)}$, $j = 0, \ldots, \ell - 1$, in Algorithm 3 are independent of the right-hand side $\bar{R}$, they can be precomputed once in a preprocessing step. A major obstacle is that the matrices $A^{(j)}$ for $j \geq 1$ cannot be represented in terms of sparse matrices and must be stored as dense matrices. This can be avoided when using a data-sparse matrix format that allows for storage-efficient (approximate) inversion and addition. Examples for such formats include hierarchical matrices [37] and hierarchically semiseparable matrices [38–40]. Implementations of the sign function method in these formats are discussed in [14, 41, 42]. For $\ell = 1$, this is not needed. In this case, Algorithm 3 reduces to

$$\bar{Z} \leftarrow \frac{1}{2\omega_1}\left(\bar{R} + \omega_1^2 \bar{A}^{-1}\bar{R}\bar{A}^{-T}\right), \tag{12}$$

which can be performed efficiently via Cholesky factorizations of M and A.

Figure 5 and Table II summarize a numerical experiment with the sign function preconditioner for $\ell = 1$. For the example under consideration, the performance is worse than for the ADI-based preconditioner, see Figure 5, because of slower convergence of the preconditioned CG method. To test whether we could gain advantage from using $\ell > 1$, we have implemented Algorithm 3 in the hierarchical matrix format using the `HLib` library [43]. The iterates $\bar{Z}^{(j)}$ are stored in the low-rank format (7) and repeatedly truncated, see [14]. For the sake of simplicity, we have only tested $n = 11,036$ and set $M = I_n$. Algorithm 1 with this preconditioner for $\ell = 2$ takes 375 s and another 318 s are needed for setting up the matrix $A^{(1)}$. This compares poorly with $\ell = 1$, which leads to a total execution time of 58 s.

3. ONE PARAMETER

In this section, we extend the preconditioned CG method discussed earlier to a Lyapunov equation depending on a single parameter α:

$$A(\alpha)X(\alpha)M(\alpha)^T + M(\alpha)X(\alpha)A(\alpha)^T = B(\alpha)B(\alpha)^T, \qquad \alpha \in \mathbb{R}. \tag{13}$$

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

More specifically, we consider the solution of (13) for m parameter samples $\alpha_1, \ldots, \alpha_m$ with $\alpha_{\min} \equiv \alpha_1 < \ldots < \alpha_m \equiv \alpha_{\max}$.

For each sample α_l, we consider the corresponding linear system

$$(M(\alpha_l) \otimes A(\alpha_l) + A(\alpha_l) \otimes M(\alpha_l))\mathrm{vec}(X(\alpha_l)) = \mathrm{vec}\left(B(\alpha_l)B(\alpha_l)^T\right). \tag{14}$$

Similar to the approach in [27], we collect these m linear systems into one huge $mn^2 \times mn^2$ block diagonal system

$$\mathcal{A}x = b \tag{15}$$

with

$$\mathcal{A} = \mathrm{diag}((M(\alpha_1) \otimes A(\alpha_1) + A(\alpha_1) \otimes M(\alpha_1)), \ldots, (M(\alpha_m) \otimes A(\alpha_m) + A(\alpha_m) \otimes M(\alpha_m))).$$

and

$$x = \begin{bmatrix} \mathrm{vec}(X(\alpha_1)) \\ \vdots \\ \mathrm{vec}(X(\alpha_m)) \end{bmatrix}, \qquad b = \begin{bmatrix} \mathrm{vec}\left(B(\alpha_1)B(\alpha_1)^T\right) \\ \vdots \\ \mathrm{vec}\left(B(\alpha_m)B(\alpha_m)^T\right) \end{bmatrix}.$$

We now rewrite (15) in terms of tensors in the sense of multidimensional arrays. For this purpose, we collect the entries of the solution and the right-hand side into two tensors $\mathcal{X}, \mathcal{B} \in \mathbb{R}^{n \times n \times m}$ with the entries

$$\mathcal{X}_{i,j,l} = (X(\alpha_l))_{i,j}, \qquad \mathcal{B}_{i,j,l} = \left(B(\alpha_l)B(\alpha_l)^T\right)_{i,j}.$$

Then, the matrix $\mathcal{A}$ can be reinterpreted as a linear operator on $\mathbb{R}^{n \times n \times m}$ and (15) becomes

$$\mathcal{A}(\mathcal{X}) = \mathcal{B}. \tag{16}$$

3.1. The Tucker format

To develop an efficient algorithm for solving (16), the low-rank matrix format (7) needs to be replaced by a low-rank format for third-order tensors. For our purposes, a suitable low-rank format is given by the Tucker format [44], which we will briefly introduce.

The Tucker format of a third-order tensor $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ takes the form

$$\mathrm{vec}(\mathcal{S}) = (U_3 \otimes U_2 \otimes U_1)\mathrm{vec}(\mathcal{C}), \tag{17}$$

where $U_i \in \mathbb{R}^{n_i \times r_i}$ for $i = 1, 2, 3$ are the *basis matrices* and $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the *core tensor*. If all U_k have full column rank, then the triple (r_1, r_2, r_3) corresponds to the multilinear rank of $\mathcal{S}$.

The Tucker format (17) is closely linked to the three different matricizations of $\mathcal{S}$. The mode-1 matricization $S^{(1)} \in \mathbb{R}^{n_1 \times n_2 n_3}$ is obtained by arranging the 1-mode fibers $\mathcal{S}(:, i_2, i_3) \in \mathbb{R}^{n_1}$ for $i_2 = 1, \ldots, n_2$, $i_3 = 1, \ldots, n_3$ into the columns of $S^{(1)}$. Similarly, $S^{(2)} \in \mathbb{R}^{n_2 \times n_1 n_3}$ and $S^{(3)} \in \mathbb{R}^{n_3 \times n_1 n_2}$ are obtained from the 2-mode fibers $\mathcal{S}(i_1, :, i_3)$ and the 3-mode fibers $\mathcal{S}(i_1, i_2, :)$, respectively. Then, the multilinear rank satisfies

$$r = (r_1, r_2, r_3) = \left(\mathrm{rank}(S^{(1)}), \mathrm{rank}(S^{(2)}), \mathrm{rank}(S^{(3)})\right).$$

Moreover, the truncation of an explicitly given tensor to a given lower multilinear rank $\bar{r} = (\bar{r}_1, \bar{r}_2, \bar{r}_3)$ can be performed by means of singular value decompositions of $S^{(1)}, S^{(2)}, S^{(3)}$, the so called *higher order singular value decomposition* (HOSVD) [20]. The obtained quasibest approximation $\bar{\mathcal{S}}$ satisfies the error bound

$$\left\|\mathcal{S} - \bar{\mathcal{S}}\right\|^2 := \left\|\mathrm{vec}(\mathcal{S}) - \mathrm{vec}\left(\bar{\mathcal{S}}\right)\right\|_2^2 \leqslant \sum_{i=1}^{3} \sum_{j=\bar{r}_i+1}^{n_i} \sigma_j^2(S^{(i)}). \tag{18}$$

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

This shows that the neglected singular values determine the error, just as in the matrix case.

The right-hand side and solution tensors for the parameter-dependent Lyapunov equation (13) are symmetric in the first two indices. This property,

$$\mathcal{S}_{i_1,i_2,i_3} = \mathcal{S}_{i_2,i_1,i_3} \quad \text{for all} \quad i_1, i_2, i_3,$$

is equivalent to $S^{(1)} = S^{(2)}$. It can be easily seen that the HOSVD can be modified to preserve this symmetry in the sense that $U_1 = U_2$ holds for the basis matrices and $C^{(1)} = C^{(2)}$ holds for the core tensor. The corresponding symmetric variant of the Tucker format takes the form

$$\operatorname{vec}(\mathcal{S}) = (U_3 \otimes U_1 \otimes U_1)\operatorname{vec}(\mathcal{C}), \quad \text{with} \quad C^{(1)} = C^{(2)}. \tag{19}$$

3.2. Approximability of $\mathcal{X}$ in the Tucker format

In this section, we show that the solution tensor $\mathcal{X}$ of (13) can be well approximated in the Tucker format, provided that $t \ll n$ and the parameter dependence is sufficiently smooth. According to the error bound (18), this can be shown by bounding the singular values of the matricizations

$$X^{(1)} = X^{(2)} = [X(\alpha_1), \ldots, X(\alpha_m)], \quad X^{(3)} = [\operatorname{vec}(X(\alpha_1)), \ldots, \operatorname{vec}(X(\alpha_m))]^T.$$

For theoretical purposes, we may assume $M(\alpha) \equiv I_n$ without loss of generality, by a suitable transformation of the Lyapunov equation. Moreover, by a suitable parameter transformation, we may also assume that $\alpha \in [-1, 1]$.

The error bound for $X^{(3)}$ can be immediately obtained by applying an existing result [27, Thm 2.4] on parameter-dependent linear systems, as the columns of $X^{(3)}$ are solutions to the linear system (14) with the system matrix $\mathcal{A}(\alpha) = A(\alpha) \otimes I + I \otimes A(\alpha)$.

Lemma 1
Let $B(\alpha) : [-1, 1] \to \mathbb{R}^{n\times t}$ and $A(\alpha) : [-1, 1] \to \mathbb{R}^{n\times n}$ both have analytic extensions to the Bernstein ellipse $\mathcal{E}_{\rho_0}$ for some $\rho_0 > 1$, and assume that $\mathcal{A}(\alpha)$ is invertible for all $\alpha \in \mathcal{E}_{\rho_0}$. Then,

$$\sigma_k(X^{(3)}) \leq \frac{2\rho\sqrt{m}}{1-\rho^{-1}} \max_{\eta\in\partial\mathcal{E}_\rho} \|\mathcal{A}(\eta)^{-1}\|_2 \, C_B \, \rho^{-k}, \tag{20}$$

for any $1 < \rho < \rho_0$, where $C_B := \max_{\eta\in\partial\mathcal{E}_\rho} \|B(\eta)\|_F^2$.

In the case that $A(\alpha)$ is symmetric positive definite for all $\alpha \in [-1, 1]$, the bound (20) can be simplified. In particular, if $A(\alpha) = A_0 + \alpha A_1$, a perturbation argument can be used to show that

$$\max_{\eta\in\partial\mathcal{E}_\rho} \|\mathcal{A}(\eta)^{-1}\|_2 \lesssim \frac{1}{\mu_{\min} - (\rho-1)^2\|A_1\|_2}$$

for $\rho - 1$ sufficiently small, where $\mu_{\min}$ is the minimal value of $\lambda_{\min}(A_0 + \alpha A_1)$ for $\alpha \in [-1, 1]$.

The following theorem gives bounds for the singular values of $X^{(1)} = X^{(2)}$. The result is only shown for $t = 1$; bounds for general t can be obtained by superposition.

Theorem 1
Let $B(\alpha) : [-1, 1] \to \mathbb{R}^{n\times 1}$, $A(\alpha) : [-1, 1] \to \mathbb{R}^{n\times n}$ have analytic extensions to $\mathcal{E}_{\rho_0}$ for some $\rho_0 > 1$. Moreover, we assume that $A(\alpha)$ is symmetric positive definite on $[-1, 1]$ and remains positive definite[‡] on $\mathcal{E}_{\rho_0}$. Then, there exists a constant $C > 0$ not depending on k and B such that

$$\sigma_{k+1}(X^{(1)}) \leq C \, C_B \exp\left(-\pi \sqrt{\frac{\log(\rho)}{\log(8\kappa)}}\sqrt{k}\right),$$

[‡]A general matrix $B \in \mathbb{C}^{n\times n}$ is called positive definite if its Hermitian part $(B^* + B)/2$ is positive definite.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

for any $1 < \rho < \rho_0$, where $C_B := \max_{\eta\in\partial\mathcal{E}_\rho} \|B(\eta)\|_2^2$ and

$$\mu_{\min} = \frac{1}{2}\inf_{\alpha\in\mathcal{E}_\rho} \lambda_{\min}(A(\alpha)+A(\alpha)^*), \quad \mu_{\max} = \max_{\alpha\in[-1,1]} \lambda_{\max}(A(\alpha)), \quad \kappa = \frac{\mu_{\max}}{\mu_{\min}}.$$

Proof
We start by recalling existing results [4, 6, 45] on the singular value decay of $X(\alpha)$ for fixed $\alpha \in [-1, 1]$. On the basis of the approximation of $1/x$ on the interval $x \in [2\mu_{\min}, 2\mu_{\max}]$ by sums of exponentials, one obtains a low-rank approximation

$$X(\alpha) \approx \tilde{X}(\alpha) = \sum_{j=1}^{\tilde{k}} g_j(\alpha) g_j(\alpha)^T, \qquad g_j(\alpha) := \sqrt{\gamma_j}\, \exp(-\omega_j A(\alpha)) B(\alpha),$$

for certain parameters $\gamma_j > 0$ and $\omega_j > 0$ independent of α. The approximation error satisfies

$$\left\| X(\alpha) - \tilde{X}(\alpha) \right\|_F \leq \frac{8\,\|B(\alpha)\|_2^2}{\mu_{\min}} \exp\left(-\frac{\pi^2}{\log(8\kappa)}\,\tilde{k}\right).$$

For the 1-mode matricization of the corresponding tensor $\tilde{\mathcal{X}}$, this directly implies the error bound

$$\left\| X^{(1)} - \tilde{X}^{(1)} \right\|_F \leq \frac{8\sqrt{m}}{\mu_{\min}} C_B \exp\left(-\frac{\pi^2}{\log(8\kappa)}\,\tilde{k}\right). \tag{21}$$

By rearranging the terms contributing to each $\tilde{X}(\alpha_l)$, we can write

$$\tilde{X}^{(1)} = \tilde{U}_1 \tilde{V}_1^T + \cdots + \tilde{U}_{\tilde{k}} \tilde{V}_{\tilde{k}}^T$$

with

$$\tilde{U}_j = [\|g_j(\alpha_1)\|_2 \cdot g_j(\alpha_1),\ \ldots,\ \|g_j(\alpha_m)\|_2 \cdot g_j(\alpha_m)] \in \mathbb{R}^{n\times m}$$
$$\tilde{V}_j = \operatorname{diag}\left(\frac{1}{\|g_j(\alpha_1)\|_2} g_j(\alpha_1),\ \ldots,\ \frac{1}{\|g_j(\alpha_m)\|_2} g_j(\alpha_m)\right) \in \mathbb{R}^{nm\times m}$$

for $j = 1, \ldots, \tilde{k}$.

Note that the columns of $\tilde{U}_j$ are evaluations of the vector-valued function

$$\|g_j(\alpha)\|_2 \cdot g_j(\alpha),$$

which has an analytic extension to $\mathcal{E}_{\rho_0}$. By [27, Corollary 2.3], there is, for every $\hat{k} \leq \min\{m, n\}$, a matrix $\hat{U}_j \in \mathbb{R}^{n\times m}$ of rank $\hat{k}$ such that

$$\left\| \tilde{U}_j - \hat{U}_j \right\|_F \leq \frac{2\rho\sqrt{m}}{1-\rho^{-1}} \max_{\eta\in\partial\mathcal{E}_\rho} \|g_j(\eta)\|_2^2\, \rho^{-\hat{k}}.$$

A classical result by Dahlquist [46] implies

$$\|g_j(\eta)\|_2 \leq \sqrt{\gamma_j}\, \|\exp(-\omega_j A(\eta))\|_2 \|B(\eta)\|_2 \leq \sqrt{\gamma_j}\exp(-\omega_j \mu_{\min}) C_B$$

for all $\eta \in \mathcal{E}_\rho$. Then, the approximation $\hat{X}^{(1)} = \sum_{j=1}^{\hat{k}} \hat{U}_j \tilde{V}_j^T$ has rank $\tilde{k}\cdot\hat{k}$ and satisfies

$$\left\| \tilde{X}^{(1)} - \hat{X}^{(1)} \right\|_F \leq \sum_{j=1}^{\tilde{k}} \left\| \tilde{U}_j - \hat{U}_j \right\|_F \leq \frac{2\rho\sqrt{m}}{1-\rho^{-1}} C_B \left(\sum_{j=1}^{\tilde{k}} \gamma_j \exp(-2\omega_j \mu_{\min}) \right) \rho^{-\hat{k}}.$$

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

The remaining exponential sum turns out to be the approximation of $1/x$ at $x = 2\mu_{\min}$ from [47] and can be bounded by

$$\sum_{j=1}^{\tilde{k}} \gamma_j \exp(-2\omega_j \mu_{\min}) \leqslant \frac{1}{2\mu_{\min}} + \frac{8}{\mu_{\min}} \exp\left(-\frac{\pi^2}{\log(8\kappa)}\,\tilde{k}\right) \leqslant \frac{8.5}{\mu_{\min}}.$$

Hence, with $\hat{C} = \frac{17\rho\sqrt{m}}{\mu_{\min}(1-\rho^{-1})}$, we obtain

$$\left\| \tilde{X}^{(1)} - \hat{X}^{(1)} \right\|_F \leqslant \hat{C} C_B \rho^{-\hat{k}}. \tag{22}$$

Combining (21) with (22) yields

$$\begin{aligned} \left\| X^{(1)} - \hat{X}^{(1)} \right\|_F &\leqslant \left\| X^{(1)} - \tilde{X}^{(1)} \right\|_F + \left\| \tilde{X}^{(1)} - \hat{X}^{(1)} \right\|_F \\ &\leqslant C_1 C_B \left(\exp\left(-\frac{\pi^2 \tilde{k}}{\log(8\kappa)} \right) + \rho^{-\hat{k}} \right) \\ &= C_1 C_B \left(\exp\left(-\tau_1 \tilde{k}\right) + \exp\left(-\tau_2 \hat{k}\right) \right), \end{aligned} \tag{23}$$

where $C_1 = \max\left\{8\sqrt{m}/\mu_{\min}, \hat{C}\right\}$ and $\tau_1 = \pi^2/(\log(8\kappa))$, $\tau_2 = \log(\rho)$. Recall that $\hat{X}^{(1)}$ has rank $\bar{k} := \tilde{k} \cdot \hat{k}$.

For a given integer k, we now balance the influence of the two terms in (23) by choosing

$$\tilde{k} := \left\lfloor \sqrt{k \frac{\tau_2}{\tau_1}} \right\rfloor, \qquad \hat{k} := \left\lfloor \sqrt{k \frac{\tau_1}{\tau_2}} \right\rfloor.$$

Then,

$$\begin{aligned} \sigma_{k+1}(X^{(1)}) \leqslant \sigma_{\bar{k}+1}(X^{(1)}) &\leqslant \left\| X^{(1)} - \hat{X}^{(1)} \right\|_F \\ &\leqslant C_1 C_B \left(\exp\left(-\tau_1 \tilde{k}\right) + \exp\left(-\tau_2 \hat{k}\right) \right) \\ &\leqslant C_1 C_B \left(\exp\left(-\tau_1 \left(\sqrt{k\tau_2/\tau_1} - 1 \right)\right) + \exp\left(-\tau_2 \left(\sqrt{k\tau_1/\tau_2} - 1 \right)\right)\right) \\ &= C_1 C_B (\exp(\tau_1) + \exp(\tau_2)) \exp\left(-\sqrt{\tau_1 \tau_2 k}\right), \end{aligned}$$

which completes the proof by setting $C := C_1(\exp(\tau_1) + \exp(\tau_2))$. □

The bounds of Lemma 1 and Theorem 1 predict a pronounced difference between the singular value decays of $X^{(3)}$ and of $X^{(1)}$, $X^{(2)}$. Such a significantly slower decay for $X^{(1)}$, $X^{(2)}$ does not appear to be an artifact of the proof but it also shows up numerically, see Figure 6.

3.3. The CG method with low-rank truncation in Tucker format

The basic idea from Section 2 carries over in a relatively straightforward manner to the solution of the parameter-dependent Lyapunov equation (13). Formally, we apply the CG method to the $mn^2 \times mn^2$ linear system (15) and apply repeated low-rank truncation to keep the computational cost low. For this purpose, we view all iterates as tensors $\mathcal{X}_k, \mathcal{R}_k, \mathcal{P}_k, \mathcal{W}_k, \mathcal{Z}_k \in \mathbb{R}^{n\times n\times m}$ and store them in the Tucker format (19). Although the formal algorithmic description of this CG method is virtually identical with Algorithm 1, the efficient implementation of the required operations demands a more detailed discussion.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

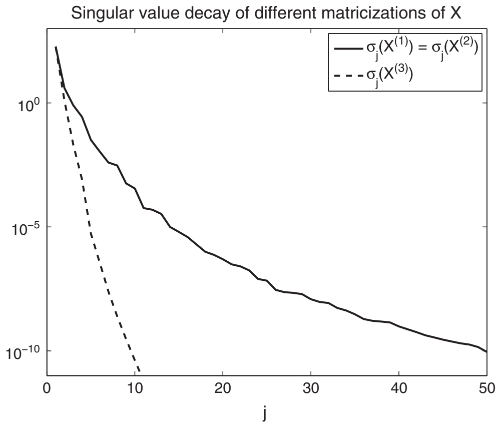


Figure 6. Solution of parameter-dependent Lyapunov equation $(A_0+\alpha A_1)X(\alpha)M^T+MX(A_0+\alpha A_1)^T = BB^T$ with $n = 371$. Singular values for matricizations of $\mathcal{X}$ when using samples $\alpha_j = j$, $j = 1, 2, \ldots, 100$.

3.3.1. Matrix multiplication. Lines 4, 6, 15, and 17 of the tensorized variant of Algorithm 1 require the multiplication of a tensor $\mathcal{S} \in \mathbb{R}^{n\times n\times m}$ with L_M^{-1} in modes 1 and 2. In the Tucker format (19), this can be easily performed; the resulting tensor $\bar{\mathcal{S}}$ takes the form

$$\operatorname{vec}\left(\bar{\mathcal{S}}\right) = \left(U_3 \otimes L_M^{-1}U_1 \otimes L_M^{-1}U_1\right)\operatorname{vec}(\mathcal{C}).$$

The ADI(1) preconditioner (10) from Section 2.3.1 can be applied in a similar fashion, by multiplying modes 1 and 2 with the matrix $(A(\bar{\alpha}) - \varphi_1 I_n)^{-1}$, and the core tensor with the scalar $-2\varphi_1$. Here, $\bar{\alpha}$ corresponds to an average of the parameter samples and $\varphi_1 = -\sqrt{\lambda_{\max}(A(\bar{\alpha}))\lambda_{\min}(A(\bar{\alpha}))}$.

3.3.2. Addition of tensors. Given two tensors in the Tucker format,

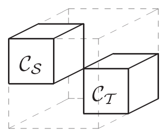
$$\operatorname{vec}(\mathcal{S}) = (U_3 \otimes U_1 \otimes U_1)\operatorname{vec}(\mathcal{C}_{\mathcal{S}}), \quad \operatorname{vec}(\mathcal{T}) = (V_3 \otimes V_1 \otimes V_1)\operatorname{vec}(\mathcal{C}_{\mathcal{T}}), \tag{24}$$

the sum $\bar{\mathcal{S}} = \mathcal{S} + \mathcal{T}$ can be represented by concatenating the factors:

$$\operatorname{vec}\left(\bar{\mathcal{S}}\right) = ([U_3, V_3] \otimes [U_1, V_1] \otimes [U_1, V_1])\operatorname{vec}\left(\mathcal{C}_{\bar{\mathcal{S}}}\right),$$

with

$$\mathcal{C}_{\bar{\mathcal{S}}} = \operatorname{diag}_{3D}(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{\mathcal{T}}) =$$



. (25)

3.3.3. Application of the linear operator $\mathcal{A}$. To discuss the efficient application of $\mathcal{A}$, we first assume that $A(\alpha) = A_0 + \alpha A_1$ is *affine linear* in α and $M(\alpha) \equiv M$ is constant. In this case, the matrix representation of $\mathcal{A}$ takes the form

$$\mathcal{A} = I \otimes M \otimes A_0 + I \otimes A_0 \otimes M + D \otimes M \otimes A_1 + D \otimes A_1 \otimes M,$$

where $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^{m\times m}$. Applied to a tensor $\mathcal{S}$, the operator $\mathcal{A}$ is a combination of matrix multiplication and addition. For $\mathcal{S}$ given in the Tucker format (19), the tensor $\bar{\mathcal{S}} = \mathcal{A}(\mathcal{S})$

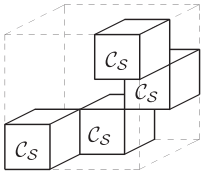
DOI: 10.1002/nla

takes the form

$$\begin{aligned}\operatorname{vec}\left(\bar{\mathcal{S}}\right) &= (I \otimes M \otimes A_0 + I \otimes A_0 \otimes M + D \otimes M \otimes A_1 + D \otimes A_1 \otimes M) \cdots \\ &\quad (U_3 \otimes U_1 \otimes U_1)\operatorname{vec}(\mathcal{C}_{\mathcal{S}}) \\ &= ([U_3, DU_3] \otimes [A_0U_1, A_1U_1, MU_1] \otimes [A_0U_1, A_1U_1, MU_1])\operatorname{vec}\left(\mathcal{C}_{\bar{\mathcal{S}}}\right),\end{aligned}$$

with

$$\mathcal{C}_{\bar{\mathcal{S}}} = \quad . \tag{26}$$



For general $A(\alpha), M(\alpha)$ depending nonlinearly on α, it is often also possible to apply $\mathcal{A}$ efficiently, in particular when $A(\alpha)$, $M(\alpha)$ are low-degree matrix polynomials in α or can be reparametrized to take this form. A more detailed discussion can be found in [27].

3.3.4. Inner product. The inner product between two tensors $\mathcal{S}$ and $\mathcal{T}$ in the Tucker format (24) can be written as

$$\begin{aligned}\langle \mathcal{S}, \mathcal{T} \rangle &:= \langle \operatorname{vec}(\mathcal{S}), \operatorname{vec}(\mathcal{T}) \rangle \\ &= \langle (U_3 \otimes U_1 \otimes U_1)\operatorname{vec}(\mathcal{C}_{\mathcal{S}}), (V_3 \otimes V_1 \otimes V_1)\operatorname{vec}(\mathcal{C}_{\mathcal{T}}) \rangle \\ &= \left\langle \operatorname{vec}(\mathcal{C}_{\mathcal{S}}), \left(U_3^T V_3 \otimes U_1^T V_1 \otimes U_1^T V_1\right) \operatorname{vec}(\mathcal{C}_{\mathcal{T}}) \right\rangle.\end{aligned}$$

In other words, we only need to form the two small matrices $U_1^T V_1$ and $U_3^T V_3$, apply them to one of the Tucker cores, and then compute the inner product between the Tucker cores.

The computation simplifies for $\mathcal{S} = \mathcal{T}$. In particular, when the columns of U_1 and U_3 are orthonormal, we have $\|\mathcal{S}\|^2 = \langle \mathcal{S}, \mathcal{S} \rangle = \langle \mathcal{C}_{\mathcal{S}}, \mathcal{C}_{\mathcal{S}} \rangle = \|\mathcal{C}_{\mathcal{S}}\|^2$.

3.3.5. Low-rank truncation. Repeated addition and application of $\mathcal{A}$ lets the multilinear ranks of the iterates of the CG method quickly grow. As in Section 2, this rank growth can be limited by repeatedly truncating the iterates to lower multilinear ranks.

Given $\operatorname{vec}(\mathcal{S}) = (U_3 \otimes U_1 \otimes U_1)\operatorname{vec}(\mathcal{C})$ with $\mathcal{C} \in \mathbb{R}^{r_1 \times r_1 \times r_3}$, the first step of low-rank truncation consists of computing QR decompositions $U_1 = Q_1 R_1$ and $U_3 = Q_3 R_3$, where $Q_j \in \mathbb{R}^{n_j \times r_j}$ has orthonormal columns and $R_j \in \mathbb{R}^{r_j \times r_j}$ is upper triangular. Then,

$$\operatorname{vec}(\mathcal{S}) = (Q_3 \otimes Q_1 \otimes Q_1)\operatorname{vec}(\mathcal{C}_Q), \quad \text{with} \quad \operatorname{vec}(\mathcal{C}_Q) := (R_3 \otimes R_1 \otimes R_1)\operatorname{vec}(\mathcal{C}).$$

Forming $\mathcal{C}_Q$ becomes expensive for larger (r_1, r_1, r_3). This cost can be reduced by exploiting the block structures (25) and (26) of the core tensor $\mathcal{C}$.

Upon completion of the orthogonalization step, we compress the Tucker core $\mathcal{C}_Q$ by computing SVDs of its matricizations $C_Q^{(1)}$ and $C_Q^{(3)}$. More specifically, for $\bar{r} = (\bar{r}_1, \bar{r}_1, \bar{r}_3)$ with $\bar{r}_1 \leqslant r_1$ and $\bar{r}_3 \leqslant r_3$, we compute the dominant left singular vectors $W_1 \in \mathbb{R}^{r_1 \times \bar{r}_1}$ and $W_3 \in \mathbb{R}^{r_3 \times \bar{r}_3}$ of $C_Q^{(1)}$ and $C_Q^{(3)}$, respectively. The truncated tensor $\bar{\mathcal{S}}$ is then obtained by projection

$$\operatorname{vec}\left(\bar{\mathcal{S}}\right) = \left(\bar{U}_3 \otimes \bar{U}_1 \otimes \bar{U}_1\right) \operatorname{vec}\left(\bar{\mathcal{C}}\right),$$

where

$$\bar{U}_1 := Q_1 W_1, \quad \bar{U}_3 := Q_3 W_3, \quad \operatorname{vec}\left(\bar{\mathcal{C}}\right) := \left(W_3^T \otimes W_1^T \otimes W_1^T\right) \operatorname{vec}(\mathcal{C}_Q).$$

Using (18), the truncation error can be bounded by

$$\left\|\mathcal{S} - \bar{\mathcal{S}}\right\|^2 \leqslant \sum_{i=1}^{3} \sum_{j=\bar{r}_i+1}^{r_i} \sigma_j^2(S^{(i)}).$$

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Table III. Computational cost of individual operations in low-rank tensor variant of the CG method applied to the $n \times n$ parameter-dependent Lyapunov equation (13) with m parameter samples. It is assumed that $n \gg r_1$ and $m \gg r_3$.

Operation	Computational cost
Addition	$O\left(nr_1^2 + mr_3^2 + r_1^3 r_3 + r_1^2 r_3^2\right)$
Application of $\mathcal{A}(\cdot)$	$O\left(\mathsf{nnz}(A_0 + A_1 + M)r_1 + nr_1^2 + mr_3^2 + r_1^3 r_3 + r_1^2 r_3^2\right)$
Truncation	$O\left(r_1^3 r_3 + r_1^2 r_3^2 + r_3^3\right)$
Inner product	$O\left(nr_1^2 + mr_3^2 + r_1^3 r_3 + r_1^2 r_3^2\right)$
ADI(1) preconditioner	$O(\mathsf{nnz}(L)r_1)$

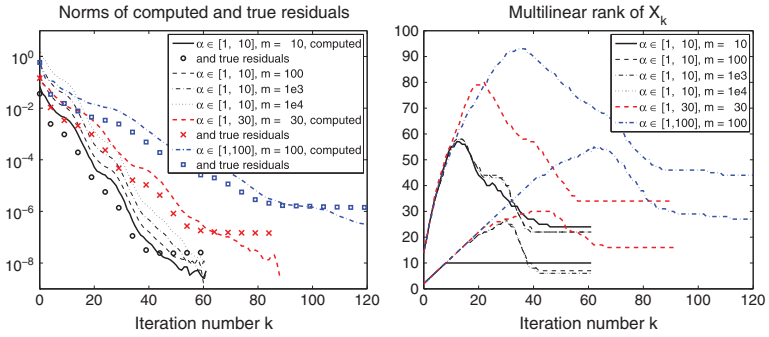


Figure 7. Solution of parameter-dependent Lyapunov equation $(A_0+\alpha A_1)X(\alpha)M^T + MX(A_0+\alpha A_1)^T = BB^T$ with $n = 371$. *Left:* Convergence of low-rank tensor CG measured by decay of computed (and true) residuals, and its dependence on number of samples of α, and on interval $[\alpha_{\min}, \alpha_{\max}]$. *Right:* Multilinear ranks (r_1, r_2, r_3) of $\mathcal{X}_k$. The upper graphs correspond to $r_1 = r_2$; the lower graphs correspond to r_3.

The choice of $\bar{r}_i$ is based on the singular values of $S^{(i)}$. For truncating the iterates $\mathcal{X}_k$ of the CG method, we use a relative criterion of the form (8). For truncating $\mathcal{R}_k$, $\mathcal{P}_k$, $\mathcal{W}_k$, we use a mixed relative/absolute criterion of the form (9).

3.3.6. Summary of computational cost. Table III summarizes the complexity of the operations discussed earlier. Note that the cost for the orthogonalization step needed for low-rank truncation (Section 3.3.5) is included in the cost for addition and application of $\mathcal{A}$. Moreover, it is assumed that $A(\alpha) = A_0 + \alpha A_1$ and $M(\alpha) \equiv M$. The cost for computing the Cholesky factorization $A(\bar{\alpha}) = LL^T$ needed for the ADI(1) preconditioner depends on the sparsity pattern of $A(\bar{\alpha})$ and is not included.

3.4. Numerical experiments

The described CG method with low-rank truncation in Tucker format has been applied to the example from [27, Sec. 2.3.1], see also Figure 1. Here, the parameter dependence of the stiffness matrix $A(\alpha) = A_0 + \alpha A_1$ arises from a parametrization of the material coefficient in parts of the domain. The right-hand side is a random rank-1 matrix independent of α.

We use the ADI(1) preconditioner with the optimal shift φ_1 with respect to $\bar{\alpha} = \sqrt{\alpha_{\min}\alpha_{\max}}$. The initial guess $\mathcal{X}_0$ is set to $(\mathcal{X}_0)_{i,j,l} = (X(\bar{\alpha}))_{i,j}$, where $X(\bar{\alpha})$ is the low-rank solution of (13) computed by Algorithm 1.

Figure 7 and Table IV display the obtained results. Compared to the corresponding results without parameter dependence (Figure 3), it becomes evident that the convergence is somewhat slower and the ranks grow significantly larger. These effects become more pronounced as the parameter variation increases. This is certainly because the average-based ADI(1) preconditioner becomes less effective. Finally, Figure 8 shows the obtained results for an example from the Oberwolfach

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Table IV. Detailed results for low-rank tensor CG applied to parameter-dependent Lyapunov equation. The individual columns contain the following: (1) parameter interval; (2) number of samples m; (3) number of CG iterations $k_{\max}$; (4) multilinear rank of approximate solution after $k_{\max}$ iterations; (5) residual norm $\mathsf{res} := \frac{1}{\sqrt{m}}\|\mathcal{B} - \mathcal{A}(\mathcal{X}_{k_{\max}})\|_F$; and (6) total computational time.

$[\alpha_{\min}, \alpha_{\max}]$	m	$k_{\max}$	rank $(\mathcal{X}_{k_{\max}})$	res	Computational time	
$[1, 10]$	10	60	$(24, 24, 10)$	$8.14 \cdot 10^{-9}$	21 s	(0.35 s/it)
$[1, 10]$	30	60	$(22, 22, 9)$	$6.27 \cdot 10^{-9}$	1 min, 07 s	(1.12 s/it)
$[1, 10]$	100	60	$(22, 22, 7)$	$5.52 \cdot 10^{-9}$	1 min, 58 s	(1.97 s/it)
$[1, 10]$	1 000	60	$(22, 22, 6)$	$5.11 \cdot 10^{-9}$	3 min, 16 s	(3.27 s/it)
$[1, 10]$	10 000	60	$(22, 22, 6)$	$5.05 \cdot 10^{-9}$	5 min, 32 s	(5.54 s/it)
$[1, 30]$	30	90	$(34, 34, 16)$	$2.60 \cdot 10^{-8}$	3 min, 57 s	(2.63 s/it)
$[1, 100]$	100	140	$(44, 44, 27)$	$1.39 \cdot 10^{-7}$	35 min, 35 s	(15.2 s/it)
$[1, 100]$	300	140	$(37, 37, 24)$	$1.18 \cdot 10^{-7}$	46 min, 26 s	(19.9 s/it)
$[1, 100]$	1 000	140	$(34, 34, 20)$	$7.16 \cdot 10^{-7}$	63 min, 31 s	(27.2 s/it)

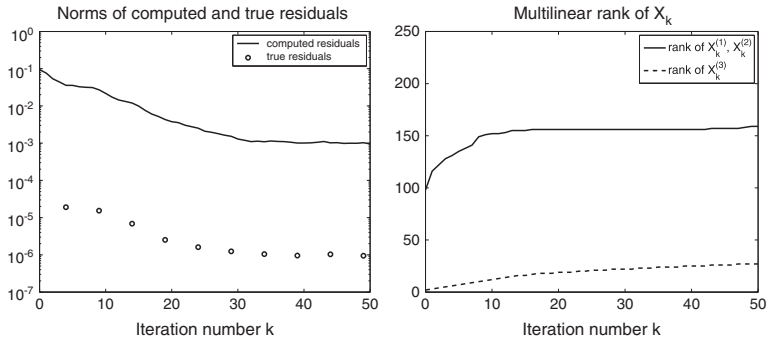


Figure 8. Boundary Condition Independent Thermal Model from [34], see also Figure 4. We used $\varepsilon = 10^{-6}$ and 30 samples of α in the interval $[10^7, 10^8]$. The final norm of the true residual is $9.30 \cdot 10^{-7}$. Note that the large gap between the true residual norms (for the generalized Lyapunov equation (13)) and the computed residual norms (for the transformed Lyapunov equation with $M = I$) is primarily due to $\|M\|_2 \approx 10^{-5}$.

Benchmark Collection [34]. This example is known to be quite a challenge, with system matrices having condition number 10^8 and larger.

4. CONCLUSIONS AND FUTURE WORK

We have presented methods for solving Lyapunov matrix equations, possibly depending on a parameter. Our methods consist of applying the preconditioned CG method combined with low-rank matrix truncation to a (huge) linear system formulation. While this point of view was mainly taken to have enough flexibility for dealing with the parameter-dependent case, the method appears to be quite competitive even for standard Lyapunov matrix equations. A more detailed comparison to existing methods based on an optimized implementation of our method remains to be performed. In the parameter-dependent case, the results are promising but also indicate that the rank growth in the transient phase of the method may pose a bottleneck for more complex problems, which can only be overcome by the development of more effective preconditioners, see [48] for an example.

While this paper has focused on Lyapunov equations with symmetric coefficients, the algorithmic extension to the nonsymmetric case is relatively straightforward, by replacing the CG method with BiCGstab or restarted GMRES, as proposed in [17, 19, 27, 49]. On the other hand, it is not

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

clear how to extend the theoretical results, in particular the bounds on the singular value decay from Theorem 1, to the nonsymmetric case.

As mentioned in Section 1, the solution of parameter-dependent Lyapunov equations plays an important role in interpolatory approaches to model reduction of parametrized linear control systems. The methods developed in this paper represent a first step towards efficient algorithms for such model reduction techniques.

ACKNOWLEDGEMENTS

We thank Zdeněk Strakoš, the home mentor of the second author in the SCIEX project that led to this paper, and two anonymous referees for their comments that led to improvements of our manuscript. The work of the second author has been supported by CRUS through the Scientific Exchange program SCIEX–NMS, project No. 09.071. The work of the third author has been supported by SNF Grant PDFMP2 124898.

REFERENCES

1. Baur U, Beattie C, Benner P, Gugercin S. Interpolatory projection methods for parameterized model reduction. *SIAM Journal on Scientific Computing* 2011; **33**(5):2489–2518.
2. Baur U, Benner P, Greiner A, Korvink JG, Lienemann J, Moosmann C. Parameter preserving model order reduction for MEMS applications. *Mathematical and Computer Modelling of Dynamical Systems* 2011; **17**(4):297–317.
3. Grasedyck L. Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation. *Numerical Linear Algebra with Applications* 2004; **11**(4):371–389.
4. Kressner D, Tobler C. Krylov subspace methods for linear systems with tensor product structure. *SIAM Journal on Matrix Analysis and Applications* 2010; **31**(4):1688–1714.
5. Penzl T. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems & Control Letters* 2000; **40**(2):139–144.
6. Sabino J. Solution of large-scale Lyapunov equations via the block modified smith methods. *Ph.D. Thesis*, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2006.
7. Li JR, White J. Low-rank solution of Lyapunov equations. *SIAM Review* 2004; **46**(4):693–713.
8. Penzl T. A cyclic low-rank Smith method for large sparse Lyapunov equations. *SIAM Journal on Scientific Computing* 1999; **21**(4):1401–1418.
9. Benner P, Li JR, Truhar N. On the ADI method for Sylvester equations. *Journal of Computational and Applied Mathematics* 2009; **233**:1035–1045.
10. Jaimoukha IM, Kasenally EM. Krylov subspace methods for solving large Lyapunov equations. *SIAM Journal on Numerical Analysis* 1994; **31**:227–251.
11. Saad Y. Numerical solution of large Lyapunov equations. In *Signal Processing, Scattering and Operator Theory, and Numerical Methods (Amsterdam, 1989)*, Vol. 5, Progress in Systems and Control Theory. Birkhäuser Boston: Boston, MA, 1990; 503–511.
12. Simoncini V. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM Journal on Scientific Computing* 2007; **29**(3):1268–1288.
13. Simoncini V, Druskin V. Convergence analysis of projection methods for the numerical solution of large Lyapunov equations. *SIAM Journal on Numerical Analysis* 2009; **47**:828–843.
14. Baur U, Benner P. Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic. *Computing* 2006; **78**(3):211–234.
15. Benner P, Quintana-Ortí ES. Solving stable generalized Lyapunov equations with the matrix sign function. *Numerical Algorithms* 1999; **20**(1):75–100.
16. Hochbruck M, Starke G. Preconditioned Krylov subspace methods for Lyapunov matrix equations. *SIAM Journal on Matrix Analysis and Applications* 1995; **16**(1):156–171.
17. Benner P, Breiten T. Low rank methods for a class of generalized Lyapunov equations and related issues. *MPI Magdeburg Preprints MPIMD/12-03* 2012. To appear in Numerische Mathematik.
18. Bollhöfer M, Eppler AK. A structure preserving FGMRES method for solving large Lyapunov equations. In *Progress in Industrial Mathematics at ECMI 2010*, Günther M, Bartel A, Brunk M, Schöps S, Striebel M (eds). Springer: Berlin Heidelberg, 2012; 131–136.
19. Eppler AK, Bollhöfer M. An alternative way of solving large Lyapunov equations. *PAMM* 2010; **10**(1):547–548.
20. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* 2000; **21**(4):1253–1278.
21. Babuška I, Nobile F, Tempone R. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis* 2007; **45**(3):1005–1034. DOI: 10.1137/050645142.
22. Kressner D, Plešinger M, Tobler C. A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations. *Technical Report*, MATHICSE, EPF Lausanne, Switzerland, 2012. Extended version of this paper. (Available from: http://anchp.epfl.ch) [Accessed date May 2013].

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

23. Grasedyck L. Hierarchical singular value decomposition of tensors. *SIAM Journal on Matrix Analysis and Applications* 2010; **31**(4):2029–2054.
24. Hackbusch W, Kühn S. A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications* 2009; **15**(5):706–722.
25. Barrett R, Berry M, Chan TF, Demmel JW, Donato J, Dongarra JJ, Eijkhout V, Pozo R, Romine C, van der Vorst H. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM: Philadelphia, PA, 1994.
26. Jbilou K, Messaoudi A, Sadok H. Global FOM and GMRES algorithms for matrix equations. *Applied Numerical Mathematics* 1999; **31**(1):49–63.
27. Kressner D, Tobler C. Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM Journal on Matrix Analysis and Applications* 2011; **32**(4):1288–1316.
28. Strakoš Z, Tichý P. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electronic Transactions on Numerical Analysis* 2002; **13**:56–80.
29. Ullmann E. A Kronecker product preconditioner for stochastic Galerkin finite element discretization. *SIAM Journal on Scientific Computing* 2010; **32**(2):923–946.
30. Benner P, Mena H, Saak J. On the parameter selection problem in the Newton-ADI iteration for large-scale Riccati equations. *Electronic Transactions on Numerical Analysis* 2008; **29**:136–149.
31. Wachspress EL. Optimum alternating-direction-implicit iteration parameters for a model problem. *Journal of the Society for Industrial and Applied Mathematics* 1962; **10**(2):339–350.
32. Wachspress EL. Extended application of alternating direction implicit iteration model problem theory. *Journal of the Society for Industrial and Applied Mathematics* 1963; **11**(4):994–1016.
33. Penzl T. Lyapack users guide. *Technical Report SFB393/00-33*, Sonderforschungsbereich 393 Numerische Simulation auf massiv parallelen Rechnern, TU Chemnitz, 09107, Chemnitz, FRG, 2000.
34. Korvink JG, Evgenii BR. Oberwolfach benchmark collection. In *Dimension Reduction of Large-Scale Systems*, Vol. 45, Benner P, Mehrmann V, Sorensen DC (eds), Lecture Notes in Computational Science and Engineering. Springer: Heidelberg, 2005; 311–316.
35. Benner P, Ezzatti P, Kressner D, Quintana-Ortí ES, Remón A. A mixed-precision algorithm for the solution of Lyapunov equations on hybrid CPU-GPU platforms. *Parallel Computing* 2011; **37**(8):439–450.
36. Sima V, Benner P. Experimental evaluation of the new SLICOT solvers for linear matrix equations based on the matrix sign function. *Proceedings of 2008 IEEE Multi-conference on Systems and Control, 9th IEEE International Symposium on Computer-Aided Control System Design (CACSD)*, San Antonio, Texas, 2008; 601–606.
37. Hackbusch W. *Hierarchische Matrizen: Algorithmen und Analysis*. Springer: Berlin Heidelberg, 2009.
38. Börm S. $\mathcal{H}_2$-matrices – an efficient tool for the treatment of dense matrices. *Habilitationsschrift*, Christian-Albrechts-Universität zu Kiel, 2006.
39. Xia J, Chandrasekaran S, Gu M, Li XS. Fast algorithms for hierarchically semiseparable matrices. *Numerical Linear Algebra with Applications* 2010; **17**(6):953–976.
40. Vandebril R, Van Barel M, Mastronardi N. *Matrix Computations and Semiseparable Matrices*, Vol. 1. Johns Hopkins University Press: Baltimore, MD, 2008.
41. Grasedyck L, Hackbusch W, Khoromskij BN. Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Computing* 2003; **70**(2):121–165.
42. Pauli S. A numerical solver for Lyapunov equations based on the matrix sign function iteration in HSS arithmetic. *Semester Thesis*, SAM, ETH Zurich, 2010.
43. Hackbusch W, Börm S, Grasedyk L. HLib library. (Available from: http://www.hlib.org) [Accessed date July 2011].
44. Kolda TG, Bader BW. Tensor decompositions and applications. *SIAM Review* 2009; **51**(3):455–500.
45. Hackbusch W. Approximation of $1/x$ by exponential sums. (Available from: http://www.mis.mpg.de/scicomp/EXP_SUM/1_x/tabelle) [Retrieved August 2008].
46. Dahlquist G. *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations*, Vol. 130. Kungliga Tekniska Högskolan (KTH): Stockholm, 1959. p. 87.
47. Braess D, Hackbusch W. Approximation of $1/x$ by exponential sums in $[1, \infty)$. *IMA Journal of Numerical Analysis* 2005; **25**(4):685–697.
48. Khoromskij BN, Schwab C. Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM Journal on Scientific Computing* 2011; **33**(1):364–385.
49. Ballani J, Grasedyck L. A projection method to solve linear systems in tensor format. *Numerical Linear Algebra with Applications* 2013; **20**(1):27–43.

Numer. Linear Algebra Appl. 2014; **21**:666–684
DOI: 10.1002/nla

Krylovovské metody a jejich chování

A.7 Článek:[WoK] Iveta Hnětynková, Martin Plešinger: *Complex wedge-shaped matrices: A generalization of Jacobi matrices*, Linear Algebra and its Applications (LAA) (ISSN 0024-3795), Volume 487 (2015), pp. 203–219 (17 pages).
(http://www.sciencedirect.com/science/article/pii/S0024379515005327)

Linear Algebra and its Applications 487 (2015) 203–219

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Complex wedge-shaped matrices: A generalization of Jacobi matrices ☆

Iveta Hnětynková [a,c,*,1], Martin Plešinger [b,2]

[a] *Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic*
[b] *Department of Mathematics, Technical University of Liberec, Liberec, Czech Republic*
[c] *Institute of Computer Science, AS CR, Prague, Czech Republic*

ARTICLE INFO

Article history:
Received 11 February 2015
Accepted 8 September 2015
Available online 19 September 2015
Submitted by V. Mehrmann

MSC:
15A18
15B57
65F15

Keywords:
Eigenvalues
Eigenvectors
Wedge-shaped matrices
Generalized Jacobi matrices
Band (or block) Krylov subspace methods

ABSTRACT

The paper by I. Hnětynková et al. (2015) [11] introduces real wedge-shaped matrices that can be seen as a generalization of Jacobi matrices, and investigates their basic properties. They are used in the analysis of the behavior of a Krylov subspace method: The band (or block) generalization of the Golub–Kahan bidiagonalization. Wedge-shaped matrices can be linked also to the band (or block) Lanczos method. In this paper, we introduce a complex generalization of wedge-shaped matrices and show some further spectral properties, complementing the already known ones. We focus in particular on nonzero components of eigenvectors.

☆ This work has been supported by the GACR grant No. P201/13–06684S.

* Corresponding author at: Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.
E-mail addresses: hnetynkova@cs.cas.cz (I. Hnětynková), martin.plesinger@tul.cz (M. Plešinger).

[1] This author is a member of the University center for mathematical modeling, applied analysis and computational mathematics (Math MAC).

[2] The research of this author has been supported by the ESF grant CZ.1.07/2.3.00/30.0065.

http://dx.doi.org/10.1016/j.laa.2015.09.017
0024-3795/

1. Introduction

Jacobi matrices

$$T = \begin{bmatrix} \delta_1 & \xi_1 & & & \\ \xi_1 & \delta_2 & \xi_2 & & \\ & \xi_2 & \ddots & \ddots & \\ & & \ddots & \delta_{n-1} & \xi_{n-1} \\ & & & \xi_{n-1} & \delta_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad \xi_\ell > 0, \quad \ell = 1,\dots,n-1, \tag{1}$$

i.e., symmetric tridiagonal matrices with positive sub-diagonal entries, represent thoroughly studied objects with the origin in the first half of the 19th century; see the historical note 3.4.3 in [14, Section 3.4]; see also [16, Chapter 7], [21, Section 5, §§ 36–48], and [7, Chapter 1.3]. Jacobi matrices have many interesting spectral properties such as: simple eigenvalues, strict interlacing of eigenvalues, eigenvectors with nonzero first and last entries, etc.; see, e.g., [16]. They are closely connected to the Lanczos tridiagonalization (see [13]), the Golub–Kahan bidiagonalization (see [9]), Gauss-type quadrature rules, moment problems, etc.; see [14]. Different generalizations of Jacobi matrices were proposed in different contexts; see [3,11,19].

In [12], the spectral properties of Jacobi matrices were used to prove fundamental properties of the so-called *core problem* (see [15]) within linear approximation problems

$$AX \approx B, \quad \text{where} \quad A \in \mathbb{R}^{n\times m},\ B \in \mathbb{R}^{n\times d},\ A^T B \neq 0, \tag{2}$$

with $d = 1$. (Note that the core problem is a useful tool in the analysis of the *total least squares* solution of (2); see [15].) In [11], it was shown that the core problem within (2) with $d > 1$ can be obtained by the *band (or block) generalization of the Golub–Kahan bidiagonalization* with *exact deflations*; see also [2]. (For further reading on finite precision computations we refer to [4,20,10,18,5,1], or [6, Chapter 5].) Starting with an orthonormal basis $s_1,\dots,s_\rho$ of the range of B, $\rho \equiv \operatorname{rank}(B)$, the generalized bidiagonalization yields a *block-bidiagonal matrix*, e.g.,

$$S_7^T A W_6 = \left[\begin{array}{ccc|cc|c} \alpha_1 & & & & & \\ \beta_{2,1} & \alpha_2 & & & & \\ \beta_{3,1} & \beta_{3,2} & \alpha_3 & & & \\ \hline \gamma_4 & \beta_{4,2} & \beta_{4,3} & \alpha_4 & & \\ & & \gamma_5 & \beta_{5,4} & \alpha_5 & \\ \hline & & & \gamma_6 & \beta_{6,5} & \\ & & & & \gamma_7 & \alpha_6 \end{array}\right] \equiv \begin{bmatrix} \Phi_1^T & & \\ \Psi_2 & \Phi_2^T & \\ & \Psi_3 & \Phi_3^T \end{bmatrix}, \qquad \rho = 3,$$

where $\alpha_k > 0$, $\gamma_{k+\rho} > 0$, $k = 1,2,\dots$; Φ_ℓ, $\Psi_{\ell+1}$ are full row rank blocks in upper triangular row echelon forms; $S_k \equiv [s_1,\dots,s_k]$, $W_k \equiv [w_1,\dots,w_k]$, $S_k^T S_k = W_k^T W_k = I_k$;

see [11]. The closely related *band (or block) Lanczos algorithm* (see [16, Chapter 13.10]) applied to AA^T with starting vectors $s_1, \ldots, s_\rho$ yields a *symmetric block-tridiagonal matrix*

$$S_7^T(AA^T)S_7 = \left[\begin{array}{ccc|cc|cc} \heartsuit & \heartsuit & \heartsuit & \alpha_1\gamma_4 & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \alpha_3\gamma_5 & & \\ \hline \alpha_1\gamma_4 & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \alpha_4\gamma_6 & \\ & & \alpha_3\gamma_5 & \heartsuit & \heartsuit & \heartsuit & \\ \hline & & & \alpha_4\gamma_6 & \heartsuit & \heartsuit & \alpha_5\gamma_7 \\ & & & & & \alpha_5\gamma_7 & \heartsuit \end{array}\right] \equiv \left[\begin{array}{cccc} \Delta_1 & \Xi_1^T & & \\ \Xi_1 & \Delta_2 & \Xi_2^T & \\ & \Xi_2 & \Delta_3 & \Xi_3^T \\ & & \Xi_3 & \Delta_4 \end{array}\right],$$

where $\heartsuit$ are in general nonzero dot-products of rows of $S_7^T A W_6$; blocks Δ_ℓ are symmetric, $\Delta_1 \in \mathbb{R}^{\rho\times\rho}$; and Ξ_ℓ are full row rank blocks in upper triangular row echelon forms. It was shown in [11] that real matrices of this block-tridiagonal form, called *wedge-shaped matrices*, can be understood as *a block generalization of Jacobi matrices*. The property of positive sub-diagonal entries of a Jacobi matrix is generalized to the property of full row rank sub-diagonal blocks.

Basic spectral properties of Jacobi matrices follow only from their nonzero structure and thus can be generalized to wedge-shaped matrices, while reflecting their block structure. It was proven in [11] that multiplicities of eigenvalues of a ρ-wedge-shaped matrix are bounded by ρ, and that the eigenvectors have nonzero leading and so-called quasi-trailing subvectors (i.e., subvectors with nonzero norms) of length ρ.

Here we extend the concept of wedge-shaped matrices to *the complex field* by preserving the block form required in the real case. Then we analyze properties of eigenvalues and eigenvectors following from the nonzero structure of wedge-shaped matrices. We give an illustrative schema for localization of nonzero subvectors of eigenvectors. We show how the results can be simplified when we consider proper band matrices.

Section 2 gives the definition of a ρ-wedge-shaped matrix and discusses some of its basic properties. Section 3 summarizes already known spectral properties of wedge-shaped matrices and reformulates them to the complex case. Section 4 discusses the structure of eigenvectors by describing a set of their nonzero subvectors, so-called running components. Section 5 comments on interlacing of eigenvalues, and Section 6 concludes the paper.

Throughout the text M^T, $M^H \equiv \overline{M}^T$, and $\mathcal{N}(M)$ denote the transposition, the conjugate transposition, and the null space of a matrix M, respectively; $M \otimes N$ denotes the Kronecker product of matrices (i.e., $m_{i,j}$, the (i,j)-th entry of M, is replaced by the block $Nm_{i,j}$ in the product); I denotes the square identity matrix, and e_k denotes the k-th column of I. The following convention concerning the entries of matrices will simplify the exposition:

- club (♣) stands for a nonzero entry, ♣ $\neq 0$;
- heart (♡) stands for a general entry which can also be zero;
- empty spaces in matrices always represent zero entries.

2. Definition and basic properties

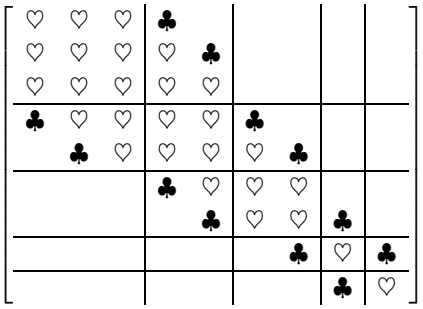
For a Hermitian (self-adjoint) matrix $T \in \mathbb{C}^{n\times n}$ with entries $t_{k,j}$, we consider the following notation

$$f(k,T) = \min\{j \,:\, t_{k,j} \neq 0\} \qquad \text{and} \qquad h(k,T) = k - f(k,T), \qquad k = 1,\dots,n, \tag{3}$$

in analogy to, e.g., [8, Chapter 4]. For simplicity, we often omit the second parameter and write $f(k)$ and $h(k)$. The number $f(k)$ is the column index of the first nonzero entry in the k-th row of T (provided it exists), and $h(k)$ is the distance between this and the diagonal entry. The number $h(k)$ is called the *k-th bandwidth of T*.[3] The following definition introduces complex wedge-shaped matrices; see also [11, Definition 4.1].

Definition 1 *(ρ-wedge-shaped matrix).* Let $T \in \mathbb{C}^{n\times n}$ be a Hermitian matrix, i.e., $T = T^H$, and ρ, $1 \le \rho < n$, an integer. If $h(k)$ is positive and non-increasing for $k = \rho+1,\dots,n$, then we call T a *ρ-wedge-shaped matrix*. We denote $\mathcal{WS}_\rho^{n\times n}$ the set of all ρ-wedge-shaped matrices of order n.

For clarity, we give an example of a 3-wedge-shaped matrix of order 9, and the corresponding values of $f(k)$ and $h(k)$:

$T =$ 

k	4	5	6	7	8	9
$f(k)$	1	2	4	5	7	8
$h(k)$	3	3	2	2	1	1

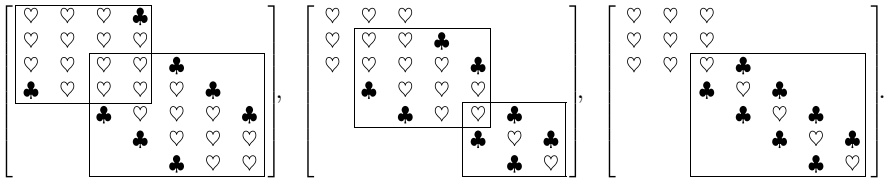
. (4)

The partitioning shows that a ρ-wedge-shaped matrix is a block-tridiagonal Hermitian matrix with full row rank sub-diagonal blocks (in upper triangular row echelon forms). Note that ρ-wedge shaped matrices satisfying

$$t_{k,k-\rho} \neq 0, \qquad \text{for} \qquad k = \rho+1,\dots,n, \tag{5}$$

[3] Note that $h \equiv \max_{k=1,\dots,n} h(k)$ is called the bandwidth of a Hermitian matrix; see [8]. Other authors call h the half-bandwidth and define the bandwidth as $2h+1$.

i.e., $h(k) = \rho$, for $k = \rho + 1, \ldots, n$, are called *matrices with a constant bandwidth* (or *proper band matrices*; see [17]). They can be considered as basic building-blocks of general wedge-shaped matrices. The following examples illustrate that any wedge-shaped matrix contains overlapping principal blocks with constant bandwidth (highlighted by frames); in all three cases $\rho = 3$, $n = 7$:



k	4	5	6	7
$f(k)$	1	3	4	5
$h(k)$	3	2	2	2

k	4	5	6	7
$f(k)$	2	3	5	6
$h(k)$	2	2	1	1

k	4	5	6	7
$f(k)$	3	4	5	6
$h(k)$	1	1	1	1

(6)

The following lemma gives some basic properties of wedge-shaped matrices.

Lemma 2. *Let* $T \in \mathcal{WS}_\rho^{n\times n}$.

(a) *If* $\rho < n-1$, *then* $T \in \mathcal{WS}_{\rho+1}^{n\times n}$. *Consequently,*

$$\mathcal{WS}_\rho^{n\times n} \subset \mathcal{WS}_{\rho+1}^{n\times n} \subset \cdots \subset \mathcal{WS}_{n-1}^{n\times n} \subset \mathbb{C}^{n\times n}.$$

(b) *If* $f(\rho+1) = 1$, *then* T *is a* $(2\rho+1)$*-diagonal matrix.*
(c) *If* $f(\rho+1) > 1$, *then* T *is at most a* $(2\rho-1)$*-diagonal matrix.*

Proof. Property (a) follows directly from Definition 1. If $f(\rho+1) = 1$, then $h(\rho+1) = \rho$, i.e., $h(k) \le \rho$, for $k = \rho+1, \ldots, n$. If $f(\rho+1) > 1$, then $h(\rho+1) < \rho$, i.e., $h(k) < \rho$, for $k = \rho+1, \ldots, n$. Since the bandwidth of the leading principal block of order ρ is trivially smaller than ρ, then T satisfying the assumption of (b) or (c) is a $(2\rho+1)$- or a $(2\rho-1)$-diagonal matrix, respectively. □

Note that we are not necessarily interested in the smallest value of ρ for which a given T is a wedge-shaped matrix. For example, in [11, Section 4.2], the aim is to verify that T is wedge-shaped for one particularly prescribed value of ρ. The following lemma on submatrices of wedge-shaped matrices will be useful later.

Lemma 3. *Let* $T \in \mathcal{WS}_\rho^{n\times n}$ *with the following* (n_1, n_2)*-partitioning,*

$$T = \underbrace{\begin{bmatrix} T_1 & L \\ L^H & T_2 \end{bmatrix}}_{n_1 \quad n_2} \begin{matrix} \} \, n_1 \\ \} \, n_2 \end{matrix}, \qquad n_1 + n_2 = n, \quad 0 \le n_1, n_2 \le n. \tag{7}$$

(a) *If $n_1 > \rho$, then T_1 is a ρ-wedge-shaped matrix, $T_1 \in \mathcal{WS}_{\rho}^{n_1 \times n_1}$.*
(b) *If $n_2 > h(n,T)$, then T_2 is a π-wedge-shaped matrix, $T_2 \in \mathcal{WS}_{\pi}^{n_2 \times n_2}$. The smallest possible value of π is defined as*

$$\pi \equiv \pi(n_1, n_2) \equiv \left(\arg \min_{\rho < k \le n} \{ f(k,T) > n_1 \} \right) - n_1 - 1, \quad \pi \le \rho. \tag{8}$$

Proof. Assertion (a) follows directly from Definition 1. Let us focus on assertion (b). If $n_2 > h(n,T)$, then

$$f(n,T) = n - h(n,T) > n_1,$$

i.e., the first nonzero entry in the last row of T is placed in the block T_2. Let $k_{\min}$, $\rho < k_{\min} \le n$, be the first row of T for which $f(k_{\min},T) > n_1$. Since $h(k,T) = k - f(k,T)$ is positive, then

$$k_{\min} = f(k_{\min},T) + h(k_{\min},T) > n_1 + 1,$$

i.e., the entry $t_{k_{\min}, f(k_{\min},T)}$ is placed in the block T_2. By employing

$$f(n_1 + \ell, T) = f(\ell, T_2) + n_1, \quad h(n_1 + \ell, T) = h(\ell, T_2), \quad \text{for} \quad \ell = k_{\min} - n_1, \dots, n_2,$$

we obtain $T_2 \in \mathcal{WS}_{\pi}^{n_2 \times n_2}$ (and $T_2 \notin \mathcal{WS}_{\pi-1}^{n_2 \times n_2}$), where $\pi \equiv k_{\min} - n_1 - 1$.

Clearly, $1 \le \pi < n_2$. If $n_2 \le \rho$, then $\pi < \rho$. If $n_2 > \rho$, then $h(n_1 + \rho + 1, T) \le \rho$ gives

$$f(n_1 + \rho + 1, T) \ge n_1 + 1, \quad \text{i.e.} \quad f(\rho + 1, T_2) \ge 1.$$

Consequently, $T_2 \in \mathcal{WS}_{\rho}^{n_2 \times n_2}$ and the minimality of π results in $\pi \le \rho$. □

The lemma directly implies that *any sufficiently large principal block* of a wedge-shaped matrix is again a wedge-shaped matrix. Let us illustrate how to determine the value of π from assertion (b). Consider, e.g., the $(5,4)$-partitioning of (4). One can see that

$$T_2 = \begin{bmatrix} \heartsuit & \heartsuit & & \\ \heartsuit & \heartsuit & \clubsuit & \\ & \clubsuit & \heartsuit & \clubsuit \\ & & \clubsuit & \heartsuit \end{bmatrix}, \qquad \begin{array}{|c|c|c|} \hline k & 3 = 8 - n_1 & 4 = 9 - n_1 \\ \hline f(k,T_2) & 2 = 7 - n_1 & 3 = 8 - n_1 \\ h(k,T_2) & 1 & 1 \\ \hline \end{array}, \tag{9}$$

is 2-wedge-shaped; i.e., $n_1 = 5$, $n_2 = 4$ yield $\pi = 2$.

Finally note that if $n_1 \le \rho$ or $n_2 \le h(n,T)$, then T_1 or T_2 have no particular structure, respectively. Moreover, for the given wedge-shaped matrix and the given

(n_1, n_2)-partitioning it may happen that none of the conditions in assertions (a) and (b) of Lemma 3 is satisfied; see, e.g., the following 2-wedge-shaped matrix:

$$T = \left[\begin{array}{cc|cc} \heartsuit & \heartsuit & \clubsuit & \\ \heartsuit & \heartsuit & \heartsuit & \clubsuit \\ \hline \clubsuit & \heartsuit & \heartsuit & \heartsuit \\ & \clubsuit & \heartsuit & \heartsuit \end{array}\right], \qquad \rho = 2, \quad h(n,T) = 2, \quad \text{and} \quad n_1 = n_2 = 2.$$

3. Spectral properties of wedge-shaped matrices

Since ρ-wedge-shaped matrices are Hermitian block-tridiagonal with full row rank sub-diagonal blocks, they can be seen as a block generalization of real symmetric tridiagonal matrices with nonzero sub-diagonal entries (including the special case of the Jacobi matrices). It is well known that a symmetric tridiagonal matrix with nonzero sub-diagonal entries has distinct eigenvalues and its eigenvectors have nonzero first and last entries; see, e.g., [16, Lemma 7.7.1 and Theorem 7.9.3 (7.9.5 in the original Prentice-Hall edition)]. These properties fully follow from the nonzero pattern of the matrix, allowing their extension to wedge-shaped matrices.

The property of nonzero first entry can be generalized in two ways. We can either stay with a single eigenvector and study its leading subvectors, or we can look at the whole eigenspace corresponding to the given λ. We start with the first approach. The proofs are straightforward generalizations of proofs in [11] for real wedge-shaped matrices. Thus we give only ideas.

Theorem 4. *Let $T \in \mathcal{WS}_\rho^{n\times n}$ and let $\lambda \in \mathbb{R}$, $v = [\nu_1, \ldots, \nu_n]^T \in \mathbb{C}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Then the subvector*

$$v^{\downarrow} \equiv [\nu_1, \ldots, \nu_\rho]^T \in \mathbb{C}^\rho, \tag{10}$$

called the leading component of the eigenvector v, is nonzero.

Assuming $[\nu_1, \ldots, \nu_\rho]^T = 0$, the comparison of the left and right-hand sides of the first row of $Tv = \lambda v$ gives $\nu_{\rho+1} = 0$. Repeating the argument gives $\nu_k = 0$ for $k = \rho+2, \ldots, n$ which contradicts $v \neq 0$; see [11, Theorem 4.2] for details.

The following result for the whole eigenspace clearly reflects the block structure of the wedge-shaped matrix.

Corollary 5. *Let $T \in \mathcal{WS}_\rho^{n\times n}$ and let $\lambda \in \mathbb{R}$ be an eigenvalue of T with multiplicity r. Let $v_\ell = [\nu_{1,\ell}, \ldots, \nu_{n,\ell}]^T \in \mathbb{C}^n$, $\ell = 1, \ldots, r$, be an arbitrary basis of the corresponding eigenspace, i.e., $TV = \lambda V$, where $V = [v_1, \ldots, v_r] \in \mathbb{C}^{n\times r}$. Then the leading $\rho \times r$ block of V,*

$$V^{\downarrow} \equiv \begin{bmatrix} \nu_{1,1} & \cdots & \nu_{1,r} \\ \vdots & \ddots & \vdots \\ \nu_{\rho,1} & \cdots & \nu_{\rho,r} \end{bmatrix} \in \mathbb{C}^{\rho \times r}, \tag{11}$$

is of full column rank r.

Proof. Since $Vw \equiv [\omega_1, \ldots, \omega_n]^T$ represents an eigenvector of T for any $w \neq 0 \in \mathbb{C}^r$, then $V^{\downarrow} w = [\omega_1, \ldots, \omega_\rho]^T$ is nonzero, by Theorem 4. Thus $V^{\downarrow}$ has linearly independent columns, which gives the assertion; see also [11, Corollary 4.3]. □

Another corollary bounds the dimension of eigenspaces of a wedge-shaped matrix.

Corollary 6. *An eigenvalue of $T \in \mathcal{WS}_{\rho}^{n \times n}$ has multiplicity at most ρ.*

The proof follows directly from Corollary 5; see also [11, Corollary 4.4]. It can also be derived *independently* of Theorem 4 as follows: Consider $S \in \mathbb{C}^{(n-\rho)\times(n-\rho)}$ a submatrix of $T - \lambda I$, $\lambda \in \mathbb{R}$, formed by rows $\rho+1, \ldots, n$ and columns $f(\rho+1), \ldots, f(n)$. Since $f(k) < k$, for $k = \rho+1, \ldots, n$, S is upper triangular with nonzero entries $t_{k,f(k)}$ on the diagonal. Thus it is nonsingular for any λ, giving $\dim(\mathcal{N}(T - \lambda I)) \leq \rho$.

Generalization of the property of nonzero last entry is more complicated. The particular structure of the band of the given wedge-shaped matrix has to be taken into account. The following theorem states the result on trailing subvectors of eigenvectors.

Theorem 7. *Let $T \in \mathcal{WS}_{\rho}^{n \times n}$ and let $\lambda \in \mathbb{R}$, $v = [\nu_1, \ldots, \nu_n]^T \in \mathbb{C}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Denote*

$$\mathcal{I}^{\uparrow}(T) \equiv \{s_1, \ldots, s_\rho\} \equiv \{1, \ldots, n\} \setminus \{f(k,T) \,:\, k = \rho+1, \ldots, n\}\,, \tag{12}$$
$$s_1 < s_2 < \cdots < s_\rho,$$

where $f(k,T)$ is given by (3). Then the subvector

$$v^{\uparrow} \equiv [\nu_{s_1}, \ldots, \nu_{s_\rho}]^T \in \mathbb{C}^\rho, \tag{13}$$

called the quasi-trailing component of the eigenvector v, is nonzero.

The proof is similar to the proof of Theorem 4. Assuming $[\nu_{s_1}, \ldots, \nu_{s_\rho}] = 0$, the comparison of the left and right-hand sides of the last row of $Tv = \lambda v$ gives $\nu_{f(n)} = 0$. Repeating this argument gives $\nu_{f(k)} = 0$ for $k = n-1, n-2, \ldots, \rho+1$ which contradicts $v \neq 0$; see [11, Theorem 4.5] for details.

Note that the vector (13) always contains the last entry of the eigenvector v, i.e., $s_\rho = n$, but in general it does not represent the trailing part of v. See the nonzero quasi-trailing components of the above given examples of wedge-shaped matrices:

- $[\nu_3, \nu_6, \nu_9]^T \in \mathbb{C}^3$ of an eigenvector $v \in \mathbb{C}^9$ of (4),
- $[\nu_2, \nu_6, \nu_7]^T \in \mathbb{C}^3$ of an eigenvector $v \in \mathbb{C}^7$ of the first matrix in (6),
- $[\nu_1, \nu_4, \nu_7]^T \in \mathbb{C}^3$ of an eigenvector $v \in \mathbb{C}^7$ of the second matrix in (6),
- $[\nu_1, \nu_2, \nu_7]^T \in \mathbb{C}^3$ of an eigenvector $v \in \mathbb{C}^7$ of the third matrix in (6).

A simplified assertion can be obtained for ρ-wedge-shaped matrices with a constant bandwidth. Here $f(k,T) = k - h(k,T) = k - \rho$ giving

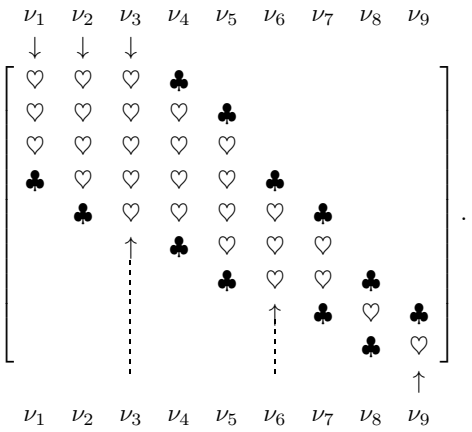
$$\mathcal{I}^{\uparrow}(T) = \{s_1, \ldots, s_\rho\} = \{n - \rho + 1, \ldots, n\}.$$

Thus $v^{\uparrow} = [\nu_{n-\rho+1}, \ldots, \nu_n]^T \in \mathbb{C}^\rho$ is the trailing part of v of length ρ.

Denote, similarly to (12),

$$\mathcal{I}^{\downarrow}(T) \equiv \{1, \ldots, \rho\}. \tag{14}$$

The sets $\mathcal{I}^{\downarrow}(T)$ and $\mathcal{I}^{\uparrow}(T)$ of indices describing components $v^{\downarrow}$ and $v^{\uparrow}$, respectively, can be observed from the pattern of T. See for example the matrix (4) for which $\mathcal{I}^{\downarrow}(T) = \{1,2,3\}$ and $\mathcal{I}^{\uparrow}(T) = \{3,6,9\}$:

$T =$ 

. (15)

The numbers $s_1, \ldots, s_{\rho-1} \in \mathcal{I}^{\uparrow}(T)$ in Theorem 7 represent row (and column) indices where the effective bandwidth of T is reduced.

Theorem 7 has a corollary analogous to Corollary 5 dealing with the whole eigenspace, which reflects the block structure of the wedge-shaped matrix.

Corollary 8. *Let $T \in \mathcal{WS}_\rho^{n\times n}$ and let $\lambda \in \mathbb{R}$ be an eigenvalue of T with multiplicity r. Let $v_\ell = [\nu_{1,\ell}, \ldots, \nu_{n,\ell}]^T \in \mathbb{C}^n$, $\ell = 1, \ldots, r$, be an arbitrary basis of the corresponding eigenspace, i.e., $TV = \lambda V$, where $V = [v_1, \ldots, v_r] \in \mathbb{C}^{n\times r}$. Let $\mathcal{I}^{\uparrow}(T) = \{s_1, \ldots, s_\rho\}$. Then the $\rho \times r$ submatrix of V,*

$$V^{\uparrow} \equiv \begin{bmatrix} \nu_{s_1,1} & \cdots & \nu_{s_1,r} \\ \vdots & \ddots & \vdots \\ \nu_{s_\rho,1} & \cdots & \nu_{s_\rho,r} \end{bmatrix} \in \mathbb{C}^{\rho \times r}, \tag{16}$$

is of full column rank r.

Proof. Since $Vw \equiv [\omega_1, \ldots, \omega_n]^T$ represents an eigenvector of T for any $w \neq 0 \in \mathbb{C}^r$, then $V^{\uparrow} w = [\omega_{s_1}, \ldots, \omega_{s_\rho}]^T$ is nonzero, by Theorem 7. Thus $V^{\uparrow}$ has linearly independent columns, which gives the assertion. □

4. Running nonzero components of eigenvectors

In this section we focus on the characterization of a set of nonzero subvectors of eigenvectors of wedge-shaped matrices. We start with another well-known property of symmetric tridiagonal matrices with nonzero sub-diagonal entries, that will be demonstrated on the Jacobi matrix T in (1). We include the derivation in order to motivate further steps. Let $\lambda \in \mathbb{R}$, $v = [\nu_1, \ldots, \nu_n]^T \in \mathbb{C}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Then

$$(\delta_1 - \lambda)\nu_1 + \xi_1 \nu_2 = 0, \tag{17}$$

$$\xi_{\ell-1}\nu_{\ell-1} + (\delta_\ell - \lambda)\nu_\ell + \xi_\ell \nu_{\ell+1} = 0, \qquad \ell = 2, \ldots, n-1, \tag{18}$$

$$\xi_{n-1}\nu_{n-1} + (\delta_n - \lambda)\nu_n = 0. \tag{19}$$

Assume that $\nu_\ell = \nu_{\ell+1} = 0$ for some $1 \leq \ell < n$. Then from (18) it successively follows that $\nu_1 = \cdots = \nu_n = 0$ contradicting $v \neq 0$. Thus *two subsequent entries of an eigenvector of a symmetric tridiagonal matrix with nonzero sub-diagonal entries* cannot be zero. In other words, violating this property would imply that *either* the leading principal submatrix of T of order ℓ has an eigenvector with zero last component, *or* the trailing principal submatrix of T of order $(n - \ell)$ has an eigenvector with zero first component. However, none of these situations can occur.

Now we use similar ideas to generalize this property to a wedge-shaped matrix T. Consider the *nontrivial* (n_1, n_2)-partitioning (7), i.e. satisfying $0 < n_1, n_2 < n$. Define two corresponding independent eigenvalue problems

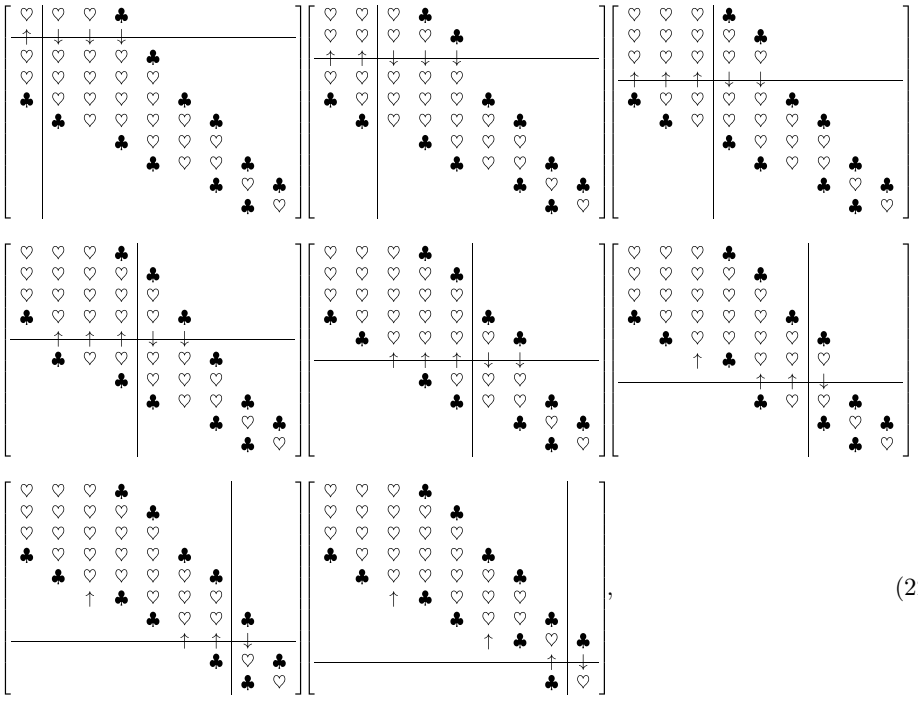
$$T_1 \widetilde{v}_1 = \lambda_1 \widetilde{v}_1, \quad \widetilde{v}_1 \neq 0, \qquad \text{and} \qquad T_2 \widetilde{v}_2 = \lambda_2 \widetilde{v}_2, \quad \widetilde{v}_2 \neq 0, \tag{20}$$

where, $T_1 \in \mathbb{C}^{n_1 \times n_1}$, $T_2 \in \mathbb{C}^{n_2 \times n_2}$. If $n_1 > \rho$, then T_1 represents a ρ-wedge-shaped matrix and indices $\mathcal{I}^{\uparrow}(T_1)$ form the nonzero quasi-trailing component $\widetilde{v}_1^{\uparrow} \in \mathbb{C}^\rho$ of $\widetilde{v}_1$; see Lemma 3 (a) and Theorem 7. Analogously, if $n_2 > h(n, T)$, then T_2 represents a π-wedge-shaped matrix for $\pi \equiv \pi(n_1, n_2) \leq \rho$ defined in (8) and indices $\mathcal{I}^{\downarrow}(T_2) = \{1, \ldots, \pi\}$ form the nonzero leading component $\widetilde{v}_2^{\downarrow} \in \mathbb{C}^\pi$ of $\widetilde{v}_2$; see Lemma 3 (b) and Theorem 4. Otherwise, if $n_1 \leq \rho$ or $n_2 \leq h(n, T)$, then T_1 or T_2 are general square

Hermitian matrices, respectively. For a *general square Hermitian matrix* $H \in \mathbb{C}^{k\times k}$ and its eigenvector $w \neq 0$ we formally define

$$\mathcal{I}^{\uparrow}(H) \equiv \mathcal{I}^{\downarrow}(H) \equiv \{1,\dots,k\} \qquad \text{and} \qquad w^{\uparrow} \equiv w^{\downarrow} \equiv w. \tag{21}$$

The following schema shows all possible nontrivial (n_1, n_2)-partitionings of (4); the up- and down-arrows denote positions of entries belonging to components $\widetilde{v}_1^{\uparrow}$ and $\widetilde{v}_2^{\downarrow}$, respectively, similarly to (15):



(22)

Note that $n_1 \leq \rho$ holds in the first three partitionings, and $n_2 \leq h(n,T)$ in the last one. Denote

$$\mathcal{I}^{(n_1,n_2)}(T) \equiv \mathcal{I}^{\uparrow}(T_1) \cup \Big(\mathcal{I}^{\downarrow}(T_2) + n_1\Big), \tag{23}$$

where the set $(\mathcal{I}^{\downarrow}(T_2) + n_1)$ contains indices from $\mathcal{I}^{\downarrow}(T_2)$ increased by n_1. Let $v = [\nu_1,\dots,\nu_n]^T \in \mathbb{C}^n$ be an eigenvector of T. Denote by $v^{(n_1,n_2)}$ a subvector (component) of v containing entries with the indices $\mathcal{I}^{(n_1,n_2)}$. The following set lists all $v^{(n_1,n_2)}$ components of v for the matrix (4):

$$\left\{ \begin{bmatrix} \nu_1 \\ \hline \nu_2 \\ \nu_3 \\ \nu_4 \end{bmatrix}, \begin{bmatrix} \nu_1 \\ \nu_2 \\ \hline \nu_3 \\ \nu_4 \\ \nu_5 \end{bmatrix}, \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \hline \nu_4 \\ \nu_5 \end{bmatrix}, \begin{bmatrix} \nu_2 \\ \nu_3 \\ \nu_4 \\ \hline \nu_5 \\ \nu_6 \end{bmatrix}, \begin{bmatrix} \nu_3 \\ \nu_4 \\ \nu_5 \\ \hline \nu_6 \\ \nu_7 \end{bmatrix}, \begin{bmatrix} \nu_3 \\ \nu_5 \\ \nu_6 \\ \hline \nu_7 \end{bmatrix}, \begin{bmatrix} \nu_3 \\ \nu_6 \\ \nu_7 \\ \hline \nu_8 \end{bmatrix}, \begin{bmatrix} \nu_3 \\ \nu_6 \\ \nu_8 \\ \hline \nu_9 \end{bmatrix} \right\}.$$

$$\begin{matrix} \| & \| & \| & \| & \| & \| & \| & \| \\ v^{(1,8)} & v^{(2,7)} & v^{(3,6)} & v^{(4,5)} & v^{(5,4)} & v^{(6,3)} & v^{(7,2)} & v^{(8,1)} \end{matrix}$$

The horizontal lines separate the entries determined by the indices of the quasi-trailing and leading components $\widetilde{v}_1^{\uparrow}$ and $\widetilde{v}_2^{\downarrow}$ of $\widetilde{v}_1$ and $\widetilde{v}_2$, respectively. We prove that $v^{(n_1,n_2)}$ are nonzero components of v, i.e. *indices* of nonzero subvectors of eigenvectors of submatrices T_1 and T_2 define indices of nonzero subvectors of eigenvectors of the matrix T.

Theorem 9. *Let $T \in \mathcal{WS}_{\rho}^{n\times n}$ and let $\lambda \in \mathbb{R}$, $v = [\nu_1, \dots, \nu_n]^T \in \mathbb{C}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. Consider the nontrivial (n_1, n_2)-partitioning (7) of T, $0 < n_1, n_2 < n$, and the conformal partitioning of v,*

$$T = \begin{bmatrix} T_1 & L \\ L^H & T_2 \end{bmatrix}, \; v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad T_\ell \in \mathbb{C}^{n_\ell \times n_\ell}, \; v_\ell \in \mathbb{C}^{n_\ell}, \; \ell = 1, 2, \quad n_1 + n_2 = n.$$

Denote

$$\mathcal{I}^{(n_1,n_2)}(T) \equiv \{t_1, \dots, t_\mu\} \equiv \mathcal{I}^{\uparrow}(T_1) \cup (\mathcal{I}^{\downarrow}(T_2) + n_1), \quad t_1 < t_2 < \cdots < t_\mu, \qquad (\textit{see (23)}),$$

where

$$\mathcal{I}^{\uparrow}(T_1) = \begin{cases} \{s_1, \dots, s_\rho\}, & \textit{when } n_1 > \rho \quad (\textit{see (12)}), \\ \{1, \dots, n_1\}, & \textit{when } n_1 \leq \rho \quad (\textit{see (21)}), \end{cases}$$

$$\mathcal{I}^{\downarrow}(T_2) = \begin{cases} \{1, \dots, \pi(n_1, n_2)\}, & \textit{when } n_2 > h(n, T) \quad (\textit{see (14)}), \\ \{1, \dots, n_2\}, & \textit{when } n_2 \leq h(n, T) \quad (\textit{see (21)}), \end{cases}$$

and where $\pi(n_1, n_2)$ is given in (8), and $h(n, T)$ in (3). Then subvectors

$$v^{(n_1,n_2)} \equiv [\nu_{t_1}, \dots, \nu_{t_\mu}]^T, \tag{24}$$

called running components of the eigenvector v, are nonzero.

Proof. The eigenvalue problem

$$Tv = \begin{bmatrix} T_1 & L \\ L^H & T_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda v, \quad v \neq 0 \tag{25}$$

yields

$$T_1 v_1 = \lambda v_1 - L v_2 \qquad \text{and} \qquad T_2 v_2 = \lambda v_2 - L^H v_1. \tag{26}$$

Assume that $v^{(n_1,n_2)} = 0$. The following situations may occur:

- If $n_1 > \rho$, then $\mathcal{I}^{\uparrow}(T_1) = \{1,\dots,n_1\} \setminus \{f(k,T_1) : k = \rho+1,\dots,n_1\}$, and, in particular, entries of v_1 with indices $f(n_1,T_1)+1,\dots,n_1$ are zero. Since $h(k,T)$ is positive and non-increasing for $k = \rho+1,\dots,n$, we get

$$\begin{aligned} h(n_1,T) &\geq h(n_1+1,T) \qquad \text{and} \\ f(n_1,T_1) = f(n_1,T) &< f(n_1+1,T), \end{aligned}$$

  i.e., the first $f(n_1,T_1)$ columns of L^H are zero. Consequently $L^H v_1 = 0$.
- If $n_1 \leq \rho$, then $\mathcal{I}^{\uparrow}(T_1) = \{1,\dots,n_1\}$ giving $v_1 = 0$, and $L^H v_1 = 0$ independently of the structure of L.
- If $n_2 > h(n,T)$, then $\mathcal{I}^{\downarrow}(T_2) = \{1,\dots,\pi(n_1,n_2)\}$ giving the first $\pi(n_1,n_2)$ entries of v_2 zero. Because T_2 is a $\pi(n_1,n_2)$-wedge-shaped matrix, we get

$$f(k,T_2) = f(k+n_1,T) - n_1, \qquad \text{for} \qquad k = \pi(n_1,n_2)+1,\dots,n_2,$$

  i.e., the first nonzero entry of the $(k+n_1)$-th row of T is localized in the block T_2 while the k-th row of L^H is zero. Consequently $L v_2 = 0$.
- If $n_2 \leq h(n,T)$, then $\mathcal{I}^{\downarrow}(T_2) = \{1,\dots,n_2\}$ giving $v_2 = 0$, and $L v_2 = 0$ independently of the structure of L.

Summarizing, in any case all nonzero entries of L are multiplied in (26) by the leading zeros in v_2, and all nonzero entries of L^H are multiplied by the trailing zeros in v_1 giving

$$L v_2 = 0 \qquad \text{and} \qquad L^H v_1 = 0; \tag{27}$$

see also (22) for an example. Thus (26) becomes

$$T_1 v_1 = \lambda v_1 \qquad \text{and} \qquad T_2 v_2 = \lambda v_2.$$

Since $v \neq 0$, then at least one of the vectors v_1, v_2 is nonzero and thus represents an eigenvector of T_1 giving $v_1 = \widetilde{v}_1$ or of T_2 giving $v_2 = \widetilde{v}_2$, respectively; see (20). Using (23) the assumption $v^{(n_1,n_2)} = 0$ contradicts the property that $\widetilde{v}_1^{\uparrow} \neq 0$ and $\widetilde{v}_2^{\downarrow} \neq 0$. □

Note that the theorem can be extended to the *trivial partitionings* with $n_1 = 0$, $n_2 = n$ (i.e., $T_1 \equiv [\,]$ is an empty matrix and $T_2 \equiv T$) and with $n_1 = n$, $n_2 = 0$ (i.e., $T_1 \equiv T$ and $T_2 \equiv [\,]$). Here the running components are identical to the leading and quasi-trailing components, respectively, i.e.,

$$v^{(0,n)} \equiv v^{\downarrow} \qquad \text{and} \qquad v^{(n,0)} \equiv v^{\uparrow}.$$

Although the statement of the theorem is complicated, schemata (15) and (22) give a simple and illustrative approach to localize nonzero components of eigenvectors based on the structure of the matrix. Finally note that Theorem 9 has a corollary analogous to Corollaries 5 and 8 that we do not formulate explicitly.

In the special case of a ρ-wedge shaped matrix $T \in \mathbb{C}^{n\times n}$ satisfying (5), i.e., with a constant bandwidth, and $n > 2\rho$, we get

$$v^{(\ell,n-\ell)} \in \begin{cases} \mathbb{C}^{\rho+\ell}, & \text{for} \quad \ell = 0,\dots,\rho-1, \\ \mathbb{C}^{2\rho}, & \text{for} \quad \ell = \rho,\dots,n-\rho, \\ \mathbb{C}^{\rho+n-\ell}, & \text{for} \quad \ell = n-\rho+1,\dots,n. \end{cases}$$

The running components have the constant length 2ρ (except for several leading and trailing components). Consequently, Theorems 4, 7, and 9 have the following corollary.

Corollary 10. *Let $T \in \mathcal{WS}_{\rho}^{n\times n}$ and let $\lambda \in \mathbb{R}$, $v = [\nu_1,\dots,\nu_n]^T \in \mathbb{C}^n$ be an eigenpair of T, i.e., $Tv = \lambda v$, $v \neq 0$. If*

$$t_{k,k-\rho} \neq, 0 \quad \textit{for} \quad k = \rho+1,\dots,n,$$

then:

(a) *The leading component $v^{(0,n)} = v^{\downarrow} = [\nu_1,\dots,\nu_\rho]^T \in \mathbb{C}^\rho$ of v is nonzero.*
(b) *The trailing component $v^{(n,0)} = v^{\uparrow} = [\nu_{n-\rho+1},\dots,\nu_n]^T \in \mathbb{C}^\rho$ of v is nonzero.*
(c) *Provided $n \geq 2\rho$, any running component $v^{(\ell,n-\ell)} = [\nu_{\ell-\rho+1},\dots,\nu_{\ell+\rho}]^T \in \mathbb{C}^{2\rho}$ for $\ell = \rho,\dots,n-\rho$ of v is nonzero.*

5. Note on the interlacing property

It is well-known that eigenvalues of Jacobi (and all 1-wedge-shaped) matrices have the so-called *strict interlacing* property. Let T_n be a 1-wedge-shaped matrix of order n, and T_j its leading principal submatrices of orders j, $j = 1,\dots,n-1$, i.e.,

$$T_1 = [\delta_1] \in \mathbb{C}^{1\times 1}, \qquad \delta_1 = \bar{\delta}_1,$$

$$T_j = \begin{bmatrix} T_{j-1} & e_{j-1}\bar{\xi}_{j-1} \\ \xi_{j-1}e_{j-1}^T & \delta_j \end{bmatrix} \in \mathbb{C}^{j\times j}, \qquad \delta_j = \bar{\delta}_j,\ \xi_{j-1} \neq 0,\ j = 2,\dots,n,$$

$$T_n = \begin{bmatrix} \delta_1 & \bar{\xi}_1 & & & \\ \xi_1 & \delta_2 & \bar{\xi}_2 & & \\ & \xi_2 & \ddots & \ddots & \\ & & \ddots & \delta_{n-1} & \bar{\xi}_{n-1} \\ & & & \xi_{n-1} & \delta_n \end{bmatrix} \in \mathbb{C}^{n\times n}. \tag{28}$$

Table 1
Interlacing of eigenvalues of *leading* principal submatrices of $\boldsymbol{T}_3 = T_3 \otimes I_\sigma$ of order k. Eigenvalues λ_ℓ of $T_3 \in \mathbb{C}^{3\times 3}$ are strictly interlaced by eigenvalues λ'_s of its leading principal submatrix $T_2 \in \mathbb{C}^{2\times 2}$, i.e., $\lambda_1 < \lambda'_1 < \lambda_2 < \lambda'_2 < \lambda_3$.

Submatrix of order k	Characteristic polynomial				
$\boldsymbol{T}_3 \in \mathbb{C}^{k\times k}$, $k = 3\sigma$	$(\lambda-\lambda_1)^{\sigma}$		$(\lambda-\lambda_2)^{\sigma}$		$(\lambda-\lambda_3)^{\sigma}$
$k = 3\sigma-1$	$(\lambda-\lambda_1)^{\sigma-1}$	$(\lambda-\lambda'_1)^{1}$	$(\lambda-\lambda_2)^{\sigma-1}$	$(\lambda-\lambda'_2)^{1}$	$(\lambda-\lambda_3)^{\sigma-1}$
$k = 3\sigma-2$	$(\lambda-\lambda_1)^{\sigma-2}$	$(\lambda-\lambda'_1)^{2}$	$(\lambda-\lambda_2)^{\sigma-2}$	$(\lambda-\lambda'_2)^{2}$	$(\lambda-\lambda_3)^{\sigma-2}$
$k = 3\sigma-3$	$(\lambda-\lambda_1)^{\sigma-3}$	$(\lambda-\lambda'_1)^{3}$	$(\lambda-\lambda_2)^{\sigma-3}$	$(\lambda-\lambda'_2)^{3}$	$(\lambda-\lambda_3)^{\sigma-3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k = 2\sigma+3$	$(\lambda-\lambda_1)^{3}$	$(\lambda-\lambda'_1)^{\sigma-3}$	$(\lambda-\lambda_2)^{3}$	$(\lambda-\lambda'_2)^{\sigma-3}$	$(\lambda-\lambda_3)^{3}$
$k = 2\sigma+2$	$(\lambda-\lambda_1)^{2}$	$(\lambda-\lambda'_1)^{\sigma-2}$	$(\lambda-\lambda_2)^{2}$	$(\lambda-\lambda'_2)^{\sigma-2}$	$(\lambda-\lambda_3)^{2}$
$k = 2\sigma+1$	$(\lambda-\lambda_1)^{1}$	$(\lambda-\lambda'_1)^{\sigma-1}$	$(\lambda-\lambda_2)^{1}$	$(\lambda-\lambda'_2)^{\sigma-1}$	$(\lambda-\lambda_3)^{1}$
$\boldsymbol{T}_2 \in \mathbb{C}^{k\times k}$, $k = 2\sigma$		$(\lambda-\lambda'_1)^{\sigma}$		$(\lambda-\lambda'_2)^{\sigma}$	

The eigenvalues λ_ℓ, $\ell = 1,\dots,j$, of T_j are strictly interlaced by the eigenvalues λ'_s, $s = 1,\dots,j-1$, of T_{j-1},

$$\lambda_1 < \lambda'_1 < \lambda_2 < \lambda'_2 < \cdots < \lambda_{j-1} < \lambda'_{j-1} < \lambda_j;$$

see, e.g., [16, section 7.10].

The ρ-wedge-shaped matrices have multiplicities of eigenvalues bounded by ρ, by Corollary 6. Employing the 1-wedge-shaped matrix (28) yields

$$\boldsymbol{T}_n \equiv T_n \otimes I_\sigma = \begin{bmatrix} I_\sigma\delta_1 & I_\sigma\bar{\xi}_1 & & & & \\ I_\sigma\xi_1 & I_\sigma\delta_2 & I_\sigma\bar{\xi}_2 & & & \\ & I_\sigma\xi_2 & \ddots & \ddots & & \\ & & \ddots & I_\sigma\delta_{n-1} & I_\sigma\bar{\xi}_{n-1} & \\ & & & I_\sigma\xi_{n-1} & I_\sigma\delta_n & \end{bmatrix} \in \mathbb{C}^{n\sigma\times n\sigma},$$

a σ-wedge shaped matrix satisfying (5), i.e., with a constant bandwidth. Matrices T_n and $\boldsymbol{T}_n$ have the same spectra; since all eigenvalues of (28) are simple, all eigenvalues of $\boldsymbol{T}_n$ have multiplicities σ, i.e., multiplicities of all eigenvalues reach the maximal bound given by Corollary 6. The *multiple* eigenvalues of $\boldsymbol{T}_j$ are strictly interlaced by *multiple* eigenvalues of its leading principal submatrix $\boldsymbol{T}_{j-1}$. Spectra of all $\sigma-1$ interjacent leading principal submatrices of $\boldsymbol{T}_j$, having $\boldsymbol{T}_{j-1}$ as the leading principal submatrix, are fully given by the standard (not the strict) interlacing, as illustrated on the example in Table 1. The *strong interlacing property therefore cannot hold* for general wedge-shaped matrices. Note that the wedge-shaped matrix $\boldsymbol{T}_n$ is for $\sigma > 1$ *reducible*, whereas Jacobi matrices are always irreducible.

6. Conclusion

We have extended some of the well-known spectral properties of symmetric tridiagonal matrices with nonzero sub-diagonal entries (including Jacobi matrices) to the class

of complex matrices called wedge-shaped. In particular, we have characterized a set of nonzero subvectors (running components) of eigenvectors of wedge-shaped matrices and described an illustrative schema for their localization based on the structure of the matrix. We have shown how the presented properties can be reformulated when we consider a wedge-shaped matrix with a constant bandwidth (a proper band matrix). The concept of (real) wedge-shaped matrices has been already used in [11] in the analysis of the band (or block) generalization of the Golub–Kahan bidiagonalization closely connected to the band (or block) Lanczos algorithm, and also in the analysis of core problems within linear approximation problems (2) with multiple right hand sides, i.e., with $d > 1$. Thus, we believe that the presented results will be useful, e.g., in further study of band and block Krylov subspace methods for complex data and related topics.

Acknowledgements

We wish to thank the anonymous referee for her or his comments which have led to improvements of our manuscript.

References

[1] J.I. Aliaga, D.L. Boley, R.W. Freund, V. Hernández, A Lanczos-type method for multiple starting vectors, Math. Comp. 69 (2000) 1577–1601.
[2] Å. Björck, Block bidiagonal decomposition and least squares problems with multiple right-hand sides, unpublished manuscript.
[3] B. Bohnhorst, Beiträge zur numerischen Behandlung des unitären Eigenwertproblems, Ph.D. thesis, Universität Bielefeld, Bielefeld, Germany, 1993.
[4] J.K. Cullum, W.E. Donath, A block generalization of the symmetric s-step Lanczos algorithm, Rep. No. RC 4845, IBM, Thomas J. Watson Res. Center, Yorktown Heights, New York, 1974.
[5] R.W. Freund, Band Lanczos method, in: Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst (Eds.), Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, SIAM, Philadelphia, 2000, pp. 80–87, 205–215 (sections 4.6 and 7.10).
[6] R.W. Freund, Model reduction methods based on Krylov subspaces, Acta Numer. 12 (2003) 267–319.
[7] W. Gautschi, Orthogonal Polynomials, Computation and Approximation, Oxford University Press, New York, 2004.
[8] A. George, J. Liu, E. Ng, Computer Solution of Sparse Linear Systems, Academic Press, Orlando, FL, 1994.
[9] G.H. Golub, W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, SIAM J. Numer. Anal. Ser. B 2 (2) (1965) 205–224.
[10] G.H. Golub, R.R. Underwood, The block Lanczos method for computing eigenvalues, in: J.R. Rice (Ed.), Mathematical Software, vol. 3, Academic Press, New York, 1977, pp. 364–377.
[11] I. Hnětynková, M. Plešinger, Z. Strakoš, Band generalization of the Golub–Kahan bidiagonalization, generalized Jacobi matrices, and the core problem, SIAM J. Matrix Anal. Appl. 36 (2) (2015) 417–434.
[12] I. Hnětynková, Z. Strakoš, Lanczos tridiagonalization and core problems, Linear Algebra Appl. 421 (2–3) (2007) 243–251.
[13] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Nat. Bur. Standards 45 (1950) 255–282.
[14] J. Liesen, Z. Strakoš, Krylov Subspace Methods, Principles and Analysis, Oxford University Press, New York, 2012.
[15] C.C. Paige, Z. Strakoš, Core problem in linear algebraic systems, SIAM J. Matrix Anal. Appl. 27 (3) (2006) 861–875.
[16] B.N. Parlett, The Symmetric Eigenvalue Problem, SIAM, Philadelphia, 1998.

[17] M. Plešinger, The total least squares problem and reduction of data in $AX \approx B$, Ph.D. thesis, Technical University of Liberec, Liberec, Czech Republic, 2008.
[18] A. Ruhe, Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse matrices, Math. Comp. 33 (146) (1979) 680–687.
[19] S. Pozza, M. Pranić, Z. Strakoš, Gauss quadrature for quasi-definite linear functionals, manuscript, 21 pp.
[20] R.R. Underwood, An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems, Ph.D. thesis, Stanford University, 1975.
[21] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, England, 1965 (reprint 2004).