Faculty of Mechatronics and
Interdisciplinary Engineering Studies
Technical University of Liberec, Liberec

Institute of Computer Science
Academy of Sciences of the Czech Republic, Prague

The Total Least Squares Problem and Reduction of Data in $AX \approx B$

Martin Plešinger

PhD Thesis, March 2008

PhD Thesis

The Total Least Squares Problem and Reduction of Data in $A\,X \approx B$

Author:	Martin Plešinger
	(martin.plesinger@tul.cz, mata@cs.cas.cz)
Supervisor:	Zdeněk Strakoš
	(strakos@cs.cas.cz)
Study programme:	P2612 Electrotechnics and informatics
Field of study:	3901V025 Science engineering
Departments:	Faculty of Mechatronics and Interdisciplinary Engineering,
	Institute of Novel Technologies and Applied Informatics,
	Technical University of Liberec,
	Studentská 2, 461 17 Liberec 1, Czech Republic
and	Institute of Computer Science,
	Department of Computational Methods,
	Academy of Sciences of the Czech Republic,
	Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic

Typeset by $\mathcal{AMS}$–LaTeX.

Prohlašuji,

že jsem tuto předkládanou disertační práci vypracoval samostatně a uvedl jsem veškeré prameny, kterých jsem použil.

Datum: _____ Podpis: _____

Homo nascitur imperfectus, sed perfectibilis est cum virtutibus et scienciis. Ad scienciam vero aquirendam duo principaliter sunt necessaria, scilicet karitas seu dileccio et sedulitas sive continuacio.

Magister Cristiannus de Prachaticz
Algorismus prosaycus, prefacio, ante 1439

Acknowledgements

First, I would like to express my deepest gratitude to all of those who gave me the possibility to complete this thesis, in particular to my supervisor professor Zdeněk Strakoš who introduced me to the area of numerical mathematics. His advice and support have been invaluable on both academic and personal level, for which I am extremely grateful.

I am indebted to my coworkers doctor Iveta Hnětynková from the Faculty of Mathematics and Physics, Charles University in Prague, doctor Diana Maria Sima and professor Sabine Van Huffel from the Department of Electrical Engineering, Katholieke Universiteit Leuven, for careful reading of parts of the manuscript and valuable suggestions which improved the presentation of the thesis. My thanks also go to all my colleagues and coworkers at the Department of Modelling of Processes, Technical University of Liberec, and at the Department of Computational Methods, Institute of Computer Science, Academy of Sciences of the Czech Republic, for a friendly working environment.

I wish to thank my parents, my family, and all my friends. Without their assistance I would never have been able to start my postgraduate studies and participate on the research work.

Financial support of the work presented in this thesis has been provided by the National Program of Research "Information Society" under project 1ET400300415.

This work I wish to dedicate to my grandmother doctor of medicine Hana Olga Zrůstová who is an excellent example for me through her professional as well as personal integrity.

Abstract

The presented thesis focuses on the solution of an orthogonally invariant linear approximation problem with multiple right-hand sides $AX \approx B$ through the *total least squares (TLS)* concept. With contribution of the early works of Golub and Reinsch (1970), Golub (1973), and van der Sluis (1975), the TLS theory for a problem with a single right-hand side was developed by Golub and Van Loan (1980). Then it was further extended by the so called *nongeneric solution* approach of Van Huffel and Vandewalle (1991), and finally revised by the *core problem theory* of Paige and Strakoš (2002, 2006). For a problem with multiple right-hand sides, a generalization of the TLS concept including a nongeneric solution was presented by Van Huffel and Vandewalle (1991).

Paige and Strakoš proved that for a problem with a single right-hand side, i.e., $Ax \approx b$, there is a reduction based on the singular value decomposition (SVD) of A which determines a *core problem* $A_{11}\,x_1 \approx b_1$, with all necessary and sufficient information for solving the original problem. The core problem always has the unique TLS solution, and, using the transformation to the original variables, it gives the solution of the original approximation problem identical to the minimum 2-norm solutions of all TLS formulations given by Van Huffel and Vandewalle. Moreover, the core problem can be efficiently computed using the (partial) upper bidiagonalization of the matrix $[\,b \,|\, A\,]$. Hnětynková, Plešinger and Strakoš (2006, 2007) derived, using the well known properties of Jacobi matrices, the core problem formulation from the relationship between the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization.

This thesis extends the classical analysis by Van Huffel and Vandewalle. It starts with an investigation of the necessary and sufficient conditions for the existence of the TLS solution. It is shown that the TLS solution is in some cases different from the output returned by the *TLS algorithm* by Van Huffel (1988), see also Van Huffel, Vandewalle (1991). The second goal of the presented thesis is an extension of the core problem theory concept to problems with multiple right-hand sides. Here the SVD-based reduction is related to the band generalization of the Golub-Kahan bidiagonalization algorithm, which was for this purpose for the first time considered by Björck (2005) and Sima (2006). We prove that the reduction results in a minimally dimensioned subproblem $A_{11}\,X_1 \approx B_1$, containing all necessary and sufficient information for solving the original problem. Unlike in the single right-hand side case, the core problem in the multiple right-hand side case may not have a TLS solution.

Keywords: linear approximation problem, multiple right-hand sides, total least squares, orthogonal transformation, data reduction, Golub-Kahan bidiagonalization algorithm, Jacobi matrices, core problem.

Abstrakt

Předkládaná disertační práce se zabývá řešením lineárních aproximačních úloh s vícenásobnou pravou stranou $A\,X \approx B$ metodou *úplných nejmenších čtverců (TLS z anglického total least squares)*. Analýza TLS problému pro úlohu s jednou pravou stranou byla, v návaznosti na dřívější práce Goluba a Reinsche (1970), Goluba (1973) a van der Sluise (1975), publikována v článku Goluba a Van Loana (1980). Tato analýza byla později rozšířena o koncept *negenerického řešení*, který zavádějí Van Huffel a Vandewalle (1991). Zcela nový vhled do teorie přináší myšlenka *core problému* Paige a Strakoše (2002, 2006). Zobecněním TLS problému na úlohy s více pravými stranami, včetně konceptu negenerického řešení, se jako první zabývali Van Huffel a Vandewalle (1991).

Paige a Strakoš dokázali za přirozeného předpokladu ortogonální invariance, tedy nezávislosti řešení na volbě souřadného systému, že pro libovolný problém s jednou pravou stranou $A\,x \approx b$ existuje transformace zkonstruovaná pomocí singulárního rozkladu matice A, která redukuje původní problém na tak zvaný *core problém* $A_{11}\,x_1 \approx b_1$, obsahující nutnou a postačující informaci k řešení původního problému. Dále ukázali, že core problém má vždy nezávisle na původních datech řešení ve smyslu TLS a toto řešení je jednoznačné. Navíc TLS řešení core problému transformované zpět do proměnných původního problému je identické s příslušným (klasickým nebo negenerickým) v normě minimálním řešením původního problému. Redukce na core problém může být provedena velmi jednoduše transformací matice $[\,b\,|A\,]$ na horní bidiagonální tvar. Hnětynková, Plešinger a Strakoš (2006, 2007) odvodili vlastnosti core problému alternativně pomocí vlastností Jakobiho matic a užitím vztahu mezi Golubovou-Kahanovou bidiagonalizací a Lanczosovou tridiagonalizací.

Předkládaná práce rozšiřuje klasické výsledky Van Huffelové a Vandewalleho pro úlohy s násobnou pravou stranou. Zabývá se analýzou nutných a postačujících podmínek existence TLS řešení. Práce ukazuje, že v některých zvláštních případech může mít TLS problém řešení, které je však různé od výsledku spočteného tak zvaným *TLS algoritmem*, viz Van Huffel (1988), případně Van Huffel, Vandewalle (1991). Dále se práce zabývá rozšířením myšlenky core problému na úlohy s vícenásobnou pravou stranou. Zobecňuje redukci dat založenou na singulárním rozkladu a zabývá se jejím vztahem k pásovému zobecnění Golubova-Kahanova bidiagonalizačního algoritmu, které bylo pro tento účel prvně doporučeno Björckem (2005) a Simou (2006). Ukážeme, že pro libovolné $A\,X \approx B$ existuje transformace, která původni problém redukuje na podproblém $A_{11}\,X_1 \approx B_1$ minimální dimenze, obsahující nutnou a postačující informaci k řešení původního problému. Ukážeme však, že na rozdíl od úloh s jednou pravou stranou core problém pro úlohy s více pravými stranami obecně nemusí mít TLS řešení.

Klíčová slova: lineární aproximační problém, vícenásobná pravá strana, úplný problém nejmenších čtverců, ortogonální transformace, redukce dat, Golubova-Kahanova bidiagonalizace, Jakobiho matice, core problém.

Contents

Acknowledgements **ix**

Abstract **xi**

Abstrakt (in Czech) **xiii**

Contents **xv**

Notation **xvii**

I Introduction 1

1 Introduction **3**
1.1 Linear approximation problems 3
1.2 Least squares and related techniques 4
1.3 TLS problem 7
1.4 Core problem theory of Paige and Strakoš 12
1.5 Goals of the thesis 26

II Theoretical Fundamentals of TLS 27

2 Problem formulation **29**
2.1 Introduction 29
2.2 Basic notation and selected theorems 30

3 Classification and the relationship to the work of Van Huffel and Vandewalle **33**
3.1 Introduction 33
3.2 Problems of the 1st class 35
3.3 Problems of the 2nd class 51
3.4 Summary and the TLS algorithm 53

III Data Reduction 57

4 SVD-based data reduction in $AX \approx B$ **59**
4.1 Introduction 59
4.2 Algorithm of the reduction 59
4.3 Summary 64
4.4 Properties of the subproblem $[\,B_1 \,|\, A_{11}\,]$ 65

5 Band generalization of the Golub-Kahan bidiagonalization 69
5.1 Introduction . . . 69
5.2 Description of the band reduction algorithm . . . 70
5.3 Basic properties of the subproblem $[\tilde{B}_1 \,|\, \tilde{A}_{11}]$ . . . 76
5.4 Generalization of Jacobi matrices . . . 77
5.5 Singular values in $[\tilde{B}_1 \,|\, \tilde{A}_{11}]$ problem . . . 84
5.6 Singular vector subspaces in $[\tilde{B}_1 \,|\, \tilde{A}_{11}]$ problem . . . 87
5.7 Summary . . . 90

6 Core problem 91
6.1 Core problem definition . . . 91
6.2 Basic properties of the core problem . . . 93
6.3 Solution of the core problem . . . 95
6.4 On the existence of independent subproblems within a core problem 100
6.5 Solution of the decomposable core problem . . . 105

IV Implementation and Computations 113

7 Bidiagonalization and core problem identification 115
7.1 Implementation remarks on bidiagonalization . . . 115
7.2 An example of the core problem identification . . . 116

8 Noise level revealing using the bidiagonalization, with application in hybrid methods 121
8.1 Introduction to ill-posed problems . . . 121
8.2 Noise level revealing example . . . 125
8.3 Approximation of the exact solution . . . 127
8.4 Summary . . . 131

V Conclusions 133

9 Conclusions and open questions 135
9.1 Conclusions . . . 135
9.2 Open questions . . . 136
9.3 List of related publications . . . 137

Bibliography 141

Notation

Scalars, vectors and matrices

$\mathbb{R}$	denotes the set of real numbers (scalars);
$\mathbb{R}^m$	linear vector space of real vectors of length m;
$\mathbb{R}^{m \times n}$	linear vector space of real m by n matrices.

Small Greek letters α, β, etc. usually denote scalars. Small Roman letters x, b, etc. denote real vectors, matrix components, and also integers. Capital Roman and sometimes Greek letters A, Q, Π, etc. denote real matrices. For example we denote:

e_j	jth Euclidean vector, e.g., $e_2 \equiv [\,0\,,\,1\,,\,0\,,\,\ldots\,,\,0\,]^T$;
m_{ij}	ith component of the jth column of the matrix M, i.e. $m_{ij} \equiv e_i^T\, M\, e_j$;
P, Q, R	orthogonal matrices;
$T_{2\rho+1}$	square symmetric band matrix with, in general, $2\rho+1$ nonzero diagonals (with the bandwidth equal to ρ);
I_n	n by n identity matrix;
Π	permutation matrix.

In particular we strictly use the following notation:

A, x, b	system matrix, vector of unknowns, and right-hand side vector, in the approximation problem $A\,x \approx b$, respectively;
A, X, B	system matrix, matrix of unknowns, and right-hand side matrix, in the approximation problem $A\,X \approx B$, respectively;
m, n, d	dimensions of the vectors and matrices as follows: $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, or $X \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{m \times d}$;

and we usually assume $m \geq n + d$ (otherwise add zero rows to the system matrix and the right-hand side).

Singular values and eigenvalues

$\sigma_j\,(M)$	denotes jth largest singular value of the matrix $M \in \mathbb{R}^{m \times n}$, for $j = 1\,,\,\ldots\,,\,\min\{\,m\,,\,n\,\}$;
$\lambda_j\,(K)$	jth largest eigenvalue value of the square symmetric (positive semidefinite) matrix $K \in \mathbb{R}^{n \times n}$, for $j = 1\,,\,\ldots\,,\,n$.
$\mathrm{sp}\,(K)$	denotes the spectrum (set of all eigenvalues) of the square symmetric (positive semidefinite) matrix $K \in \mathbb{R}^{n \times n}$, i.e. $\mathrm{sp}\,(K) \equiv \{\,\lambda_j\,(K)\,,\,j = 1\,,\,\ldots\,,\,n\,\}$.

Since $m \geq n + d$ we denote:

σ'_j	jth largest singular value of the system matrix $A \in \mathbb{R}^{m \times n}$, i.e. $\sigma'_j \equiv \sigma_j\,(A)$ and $\sigma'_1 \geq \ldots \geq \sigma'_n \geq 0$, and
ς'_j	jth largest of k *distinct and nonzero* singular values of $A \in \mathbb{R}^{m \times n}$, i.e. $\varsigma'_1 > \ldots > \varsigma'_k > 0$;

σ_j	jth largest singular value of the extended matrix $[\,B\,|\,A\,] \in \mathbb{R}^{m\times(n+d)}$ (possibly $[\,b\,|\,A\,] \in \mathbb{R}^{m\times(n+1)}$), i.e. $\sigma_j \equiv \sigma_j\,([\,B\,|\,A\,])$ and $\sigma_1 \geq \ldots \geq \sigma_{n+d} \geq 0$;
p, q, e	integers which characterize the multiplicity of the singular value $\sigma_{n+1} \equiv \sigma_{n+1}\,([\,B\,|\,A\,])$.

Vector and matrix norms

$\|v\|$	denotes the 2-norm (Euclidean norm) of the given vector x, $\|v\| \equiv (\sum_j v_j^2)^{1/2}$;
$\|M\|$	2-norm (spectral norm) of the given matrix $M \in \mathbb{R}^{m\times n}$, $\|M\| \equiv \max_{\|v\|=1}\{\,\|Mv\|\,\} = \sigma_1\,(M)$;
$\|M\|_F$	Frobenius norm of the given matrix $M \in \mathbb{R}^{m\times n}$, $\|M\|_F \equiv (\sum_{ij} m_{ij}^2)^{1/2} = (\sum_j \sigma_j^2\,(M))^{1/2}$.

Symbolic notation for matrices and matrix components

$\heartsuit$	denotes a generally nonzero matrix (or vector) component;
$\clubsuit$	denotes a matrix (or vector) component which is *different from zero* (often positive); occasionally it denotes a block (submatrix or subvector) or whole matrix (or vector) different from zero; (the zero components of matrices (or vectors) are not highlighted).

Standard matrix operations and properties

M^T	denotes the matrix transposed to $M \in \mathbb{R}^{m\times n}$;
$M^\dagger$	Moore-Penrose pseudoinverse of the matrix $M \in \mathbb{R}^{m\times n}$;
K^{-1}	denotes the inverse of the square nonsingular matrix $K \in \mathbb{R}^{n\times n}$;
$\mathrm{rank}\,(M)$	number of linearly independent rows and/or columns of $M \in \mathbb{R}^{m\times n}$.

Matrix associated subspaces, subspaces operations

$\mathcal{R}\,(M)$	denotes the range of the matrix $M \in \mathbb{R}^{m\times n}$, $\mathcal{R}\,(M) \equiv \{\,y : y = Mx\,,\, x \in \mathbb{R}^n\,\} \subset \mathbb{R}^m$;
$\mathcal{N}\,(M)$	the null space of the matrix $M \in \mathbb{R}^{m\times n}$, $\mathcal{N}\,(M) \equiv \{\,x : Mx = 0\,\} \subset \mathbb{R}^n$;
$[\,\mathcal{U}\,]^\perp$	orthogonal complement of $\mathcal{U}$;
$\mathcal{V} \oplus \mathcal{U}$	direct sum of mutually orthogonal subspaces $\mathcal{V}$ and $\mathcal{U}$;
$\mathrm{span}\,(\,v_j\,)$	span of vectors v_j, $\mathrm{span}\,(\,v_1\,,\,\ldots\,,\,v_n\,) \equiv \mathcal{R}\,([\,v_1\,,\,\ldots\,,\,v_n\,])$.

Approximation problems

$A\,x \approx b$	approximation problem with single right-hand side;
$A\,X \approx B$	approximation problem with multiple right-hand sides;
$A_{11}\,X_1 \approx B_1$	core problem within $A\,X \approx B$ (in particular in the SVD form);
$\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$	core problem within $A\,X \approx B$ in the banded form;
X_{TLS}	solution of TLS problem;
X_{NGN}	nongeneric solution of the TLS problem;
$X_{\mathrm{Const.}}$	solution of the constrained problem;
$X_{\mathrm{T\text{-}TLS}}$	solution of the truncated TLS problem;
$X_{\mathrm{Comp.}}$	composed solution of the composed problem;
$\mathscr{F}_1$, $\mathscr{F}_2$, $\mathscr{F}_3$	sets of problems of the 1st class;
$\mathscr{S}$	set of problems of the 2nd class.

List of abbreviations and acronyms

CS	denotes the CS (cosine sine) decomposition;
LQ	LQ decomposition, $M = L\,Q$, where L is lower triangular;
QR	QR decomposition, $M = Q\,R$, where R is upper triangular;
SVD	singular value decomposition, $M = U\,\Sigma\,V^T$ or $M = S\,\Theta\,W^T$;
LS, OLS	(ordinary) least squares;
DLS	data least squares;
ScTLS, STLS	scaled total least squares;
TLS	total least squares;
T-TLS	truncated total least squares.

Part I

Introduction

Chapter 1

Introduction

This chapter introduces the orthogonally invariant linear approximation problems. Such problems arise in many scientific and technical areas and various techniques are used to solve them. The least squares concept and the related techniques – data least squares, (scaled) total least squares – are briefly mentioned here, as well as the unification of all these concepts using the scaled total least squares (see also the series of papers [61, 62, 63]).

The rest of this chapter focuses namely on the total least squares concept applied on the problems with single right-hand sides. We briefly summarize the main idea of the so called basic and nongeneric solutions of total least squares, given by G. H. Golub and C. F. Van Loan in [31], and S. Van Huffel and J. Vandewalle in [84], respectively. Then we recall the core problem theory given by C. C. Paige and Z. Strakoš in [64]. At the end of this chapter we formulate the goals of the thesis.

1.1 Linear approximation problems

We are interested in the linear approximation problem

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

and its more general form

$$A\,X \approx B\,, \qquad A \in \mathbb{R}^{m\times n}\,, \quad X \in \mathbb{R}^{n\times d}\,, \quad B \in \mathbb{R}^{m\times d}\,. \tag{1.2}$$

Such linear approximation problems arise in a broad class of scientific and technical areas, for example in medical image deblurring (tomography), bioelectrical inversion problems, geophysics (seismology, radar or sonar imaging), astronomical observations. This thesis mainly focuses on the total least squares (TLS) formulation of (1.1), (1.2) that leads to a procedure that has been independently developed in various literature. It has been known by various names, for example, it is known as the *errors-in-variables modeling* in the statistical literature, see [84, 85, 86].

There exist a lot of approaches that are closely related to the TLS concept. For example, an additional difficulty appears when the system (1.1), (1.2) is *ill-posed*, here the matrix A is ill conditioned and typically a small perturbation of right-hand side causes large changes in the estimated solution. The matrix A is often numerically rank deficient and it has small singular values, but without a well defined numerical rank (singular values decay gradually without noticeable gap). In such cases the least squares (LS), the TLS or similar techniques might give a solution that is absolutely meaningless, because it is dominated by errors present in the data and possibly also by computational (rounding) errors. The *regularization*

techniques must be used in order to obtain a meaningful solution, see for example [36, 38].

Model reduction represents another important area of applications. Here the matrix A represents a model and the vector b or columns of the matrix B represent the observation vectors, e.g. measured data, that naturally contain errors. The idea is to approximate the high order system (1.1) or (1.2) by a lower order one while approximating well the behavior of the whole system. Truncation and projection techniques used to reduce the dimensions of the linear system may also be viewed as a type of regularization. Such methods are, for example the *truncated-least squares (T-LS)* also called the *truncated-singular value decomposition (T-SVD)*, the *truncated-total least squares (T-TLS)*, see [67], or Krylov subspace methods and Lanczos-type processes [20]. The system (1.2) can in such applications contain significantly more observations (columns of B) than is the dimension of range of A, or the number of columns of A, i.e. $d \gg n$; similar situation can occur in various statistical applications.

The systems (1.1), (1.2) can be compatible, i.e., $b \in \mathcal{R}(A)$, $\mathcal{R}(B) \subset \mathcal{R}(A)$, or incompatible, i.e., $b \notin \mathcal{R}(A)$, $\mathcal{R}(B) \not\subset \mathcal{R}(A)$. The compatible case is simpler because it reduces to finding a solution of the system of linear algebraic equations. Thus here the incompatible case is often considered. Another uninteresting case is excluded by the assumption $A^T b \neq 0$ or $A^T B \neq 0$. In this case it is meaningless to approximate b or columns of B by the columns of A and the systems (1.1), (1.2) have trivial solutions $x = 0$ or $X = 0$, respectively. In particular we assume a nonzero matrix A and a nonzero right-hand side vector b or matrix B. We assume for simplicity only the real case, an extension to the complex data being straightforward.

Since the incompatible problem does not have a solution in the classical meaning, the solution is obtained by solving a minimization (optimization) problem. It is senseful to assume orthogonally (unitarily) invariant minimization problems, i.e. problems such that their solutions do not depend on the particular choice of bases in $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^d$ in (1.1) or (1.2). In other words, when the original problem is transformed to another basis, this transformed problem is solved, and its solution is transformed back to the original basis, then this back-transformed solution is identical to the solution obtained directly from solving original problem.

1.2 Least squares and related techniques

Various orthogonally invariant minimization techniques can be used for solving the linear approximation problems. We introduce some of them on the problems with single right-hand sides (1.1), the extension of their definitions to problems with multiple right-hand sides (1.2) being straightforward.

The most common technique called *(ordinary) least squares (LS, OLS)* or *linear regression* is used to solve the system (1.1) when errors are confined to the right-hand side b but not to the matrix A. The LS method seeks a vector $g \in \mathbb{R}^n$ satisfying

$$\min_{x,g} \|g\| \qquad \text{subject to} \qquad A\,x = b + g\,,$$

i.e., $(b + g) \in \mathcal{R}(A)$. A minimal perturbation of the right-hand side b is searched such that the corrected system is compatible. In the multiple right-hand side case the Frobenius norm of the correction of B is minimized; it is easy to see that the LS problem with multiple right-hand side $B \in \mathbb{R}^{m\times d}$ represents nothing more than d independent LS problems with single right hand sides – the columns of B.

The opposite case to the LS is the *data least squares (DLS)*, see [33]. In DLS the correction is allowed only in A (errors are assumed to affect only the data matrix).

The matrix E is sought to minimize the Frobenius norm

$$\min_{x,E} \|E\|_F \qquad \text{subject to} \qquad (A+E)\,x \,=\, b\,,$$

i.e., $b \in \mathcal{R}\,(A+E)$. In words, we look for the minimal correction (in the Frobenius norm) of A such that the corrected system is compatible.

While in LS or DLS the correction is restricted only to the vector b or the matrix A, respectively (which corresponds to the assumption that all errors are confined to the vector of observations or the system matrix), in the *total least squares (TLS)* also called *orthogonal regression* the correction is allowed to compensate for errors in the system (data) matrix A as well as in the vector of observations b. Thus in TLS, E and g are sought to minimize the Frobenius norm in

$$\min_{x,E,g} \left\| \left[\; g \;\middle|\; E \;\right] \right\|_F \qquad \text{subject to} \qquad (A+E)\,x \,=\, b+g\,, \tag{1.3}$$

i.e., $(b+g) \in \mathcal{R}\,(A+E)$.

In all cases we look for a minimal correction such that the corrected system is compatible. The LS, DLS and TLS problems have statistical relevance for different situations, see Van Huffel and Vandewalle [84] for an excellent discussion and history. This book also carefully delineated the TLS theory and how it is related to LS. For comparison of the LS and the TLS solutions of (1.1) see also [88, 89].

From the definition of the LS solution it follows that the corrected right-hand side is the orthogonal projection of b onto the space generated by the columns of A. The LS solution always exists but it may be nonunique. Therefore the *minimum 2-norm LS solution* is defined, that is equal to

$$x_{\mathrm{LS}} \,\equiv\, A^{\dagger}\,b\,,$$

where $A^{\dagger}$ denotes the Moore-Penrose pseudoinverse of the matrix A, see, e.g., [32, p. 257]; such solution is naturally unique. Numerical methods for computing LS solution are direct (based on singular value decomposition or QR decomposition), or iterative (CGLS [40, 77] or LSQR [59, 60]); see also the classical books [52, 7]. The extension of the least squares concept to problems with multiple right-hand sides (1.2) is straightforward. As explained before, the matrix A does not change in the LS concept, thus the LS problem (1.2) with the matrix right-hand side B can be replaced by d independent LS problems (1.1) with the single right-hand sides – the columns of B. Putting these d independent solutions together gives the solution of the original problem. Consequently, the unique minimum norm LS solution is equal to

$$X_{\mathrm{LS}} \,\equiv\, A^{\dagger}\,B\,.$$

Unfortunately, the DLS and the TLS concepts applied on problems with multiple right-hand sides yield much more difficulties. In particular, the solution can not be, in general, rewritten as a set of solutions of independent problems with single right-hand sides – columns of B, see also [84].

In the following section it is shown how the concept of the scaled TLS unifies all the mentioned concepts LS, DLS, TLS. Consequently the LS and DLS concepts are interpreted and analyzed as a limit cases of the scaled TLS concept, see also [61, 62, 63].

The analysis of the TLS solution for the single right-hand side case is briefly discussed further in this chapter, see also [31, 84, 64]. Analysis of the TLS formulation in the multiple right-hand sides case (1.2) is discussed in [84], and it is one of the main goals of this thesis.

1.2.1 Scaled TLS and the importance of the TLS concept

All the mentioned concepts LS, DLS, TLS for solving (1.1) can be unified by considering the following very general *scaled TLS problem (ScTLS, STLS)*, [61, 62], see also the work of B. D. Rao [67], who called it *weighted TLS*. For a given $\vartheta > 0$, consider the problem

$$\min_{\tilde{x},\tilde{E},\tilde{g}} \left\| \left[\begin{array}{c|c} \tilde{g}\,\vartheta & \tilde{E} \end{array} \right] \right\|_F \qquad \text{subject to} \qquad (A+\tilde{E})\,\tilde{x} \,=\, b+\tilde{g}\,. \tag{1.4}$$

Here the relative sizes of the corrections in A and b are determined by a real parameter $\vartheta > 0$. As $\vartheta \longrightarrow 0$ in the ScTLS formulation, it leads to $\tilde{E} = 0$ and allows *arbitrary* $\tilde{g}$, and thus it approaches the LS formulation. When $\vartheta = 1$ the ScTLS problem obviously coincides with the TLS formulation. The case $\vartheta \longrightarrow \infty$ requires $\tilde{g} \longrightarrow 0$ leading to the DLS. See the papers [61, 62].

The formulation of the ScTLS problem can be used in a slightly different form. For any positive bounded ϑ, substitute $g \equiv \tilde{g}\,\vartheta$, $x \equiv \tilde{x}$ and $E \equiv \tilde{E}$ to obtain the following new formulation of the ScTLS problem. For a given $\vartheta > 0$,

$$\min_{x,E,g} \left\| \left[\begin{array}{c|c} g & E \end{array} \right] \right\|_F \qquad \text{subject to} \qquad (A+E)\,x\,\vartheta \,=\, b\,\vartheta + g\,. \tag{1.5}$$

In [61, 62] $x = x(\vartheta)$ that minimizes (1.5) is called the *ScTLS solution* of (1.5). And in analogy with the TLS problem, $x(\vartheta)\,\vartheta$ is called the *TLS solution* of (1.5). Consequently the ScTLS solution $x = x(\vartheta)$ of (1.5) is identical to the solution $\tilde{x}$ of (1.4). Therefore all results and discussions based on (1.5) apply fully to the scaled TLS problem (1.4). In particular, all the known TLS theory and algorithms can be applied directly to (1.5). The equivalence of (1.5) and (1.4) is extremely useful.

Using ϑ can have a statistical significance. Consider a model where the components of A are known to have independent zero-mean random error of equal standard deviation δ_A. Suppose also that the components of b have been observed with independent zero-mean random errors of equal standard deviation δ_b, and that the errors in b and A are independent. Taking $\vartheta = \delta_A/\delta_b$ in (1.5) will ensure that all the errors in that model have equal standard deviation (and thus variance), and (1.5) is an appropriate formulation for providing estimates. This agrees with the limiting behavior described above, for clearly if $\delta_A = 0$ and $\delta_b \neq 0$, then the LS is a suitable choice, while if $\delta_A \neq 0$ and $\delta_b = 0$, then the DLS is a suitable choice. However (1.5) can also be useful outside any statistical context, see [63].

Unifying the LS, DLS and TLS concepts using the ScTLS formulation (1.5) can be straightforwardly extended to the multiple right-hand side case (1.2). Because the TLS theory can be applied to the ScTLS problems, it is very important to understand the TLS formulation.

Remark 1.1. *In some applications it is suitable to preprocess the input data* $[\,b\,|\,A\,]$ *or* $[\,B\,|\,A\,]$ *before solving a given problem. Typically weighting of individual equations and/or scaling of unknowns is used. For example:*

(i) *In [31], a problem with single right-hand side is considered in the form*

$$(W\,A\,S^{-1})\,(S\,x) \,\approx\, (W\,b)\,,$$

where $W = \mathrm{diag}\,(\,w_1\,,\,\ldots\,,\,w_m\,)$ *and* $S = \mathrm{diag}\,(\,s_1\,,\,\ldots\,,\,s_n\,)$ *are diagonal weighting and scaling matrices, respectively.*

(ii) *The so called mixed LS-TLS problem, the case when some columns of the data matrix A are known exactly and the rest of columns of A contains errors, can also be viewed as a scaling, see [84, §3.6.3, p. 92].*

1.3 TLS problem

In this section the theory of solving the TLS problems with single right-hand sides is summarized. For better explanation and understanding of presented theory including detailed proofs we refer to [31, 84].

In the whole section we consider $A^T b \neq 0$, in particular $A \neq 0$, $b \neq 0$. First, it is worth to note that the TLS problem may not have a solution for a given data A, b, see the following example.

Example 1.1. *Consider the simple linear approximation problem*

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The TLS method seeks the smallest correction $[\,g \,|\, E\,]$ such that the resulting system is compatible. Since the original problem is incompatible, for any corrections E, g that makes it compatible, there exists a positive number $\varepsilon < \|[\,g \,|\, E\,]\|_F$ such that

$$\begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

is compatible too. Since the Frobenius norm of the new correction is equal to ε there is no minimal correction. Only a nonoptimal solution of the original problem can be obtained.

Moreover, the solution of the corrected system with the (nonoptimal) correction chosen as

$$\left[\begin{array}{c|cc} g & E \end{array}\right] \equiv \left[\begin{array}{c|cc} 0 & 0 & 0 \\ 0 & 0 & \varepsilon \end{array}\right], \qquad \left\| \left[\begin{array}{c|cc} g & E \end{array}\right] \right\|_F = \varepsilon\,,$$

is $x = [\,1\,, \varepsilon^{-1}\,]^T$. With $\varepsilon \longrightarrow 0$, the norm of this nonoptimal solution grows to infinity, $\|x\| \longrightarrow \infty$, and, furthermore, this solution depends on the arbitrarily chosen number ε.

Example 1.1 is taken over [31] and it illustrates that the TLS problem may not have a solution. Here the minimal correction does not exist and when we try to reach the greatest lower bound of the norm of the correction, the corresponding nonoptimal solution grows to infinity (in norm) and depends on an arbitrary value.

Further in the text, in Example 1.2, it is shown that for any problem which does not have a TLS solution the growing of the norm and the dependence on an arbitrary data (which is here given through the special choice of the correction matrix $[\,g \,|\, E\,]$) is a general behavior of nonoptimal solutions.

In [31] a sufficient condition for existence of a TLS solution is given. We follow the technique used in [31] to obtain a solution of (1.3). Analysis of several different cases that can occur yields different approaches to solving the TLS problem. We focus on each of them.

1.3.1 Basic solution by Golub and Van Loan

Consider an orthogonally invariant linear approximation problem (1.1)

$$A\,x \approx b$$

or, equivalently,

$$\left[\begin{array}{c|c} b & A \end{array}\right] \begin{bmatrix} -1 \\ \hline x \end{bmatrix} \approx 0\,. \tag{1.6}$$

In order to simplify the notation assume that $m > n$ (add zero rows if necessary). Denote $\sigma'_j \equiv \sigma_j(A)$ the jth largest singular value of A, and u'_j and v'_j the corresponding left and right singular vectors, respectively, $j = 1, \ldots, n$. Further denote $\sigma_j \equiv \sigma_j([\,b\,|\,A\,])$ the jth largest singular value of $[\,b\,|\,A\,]$, and u_j and v_j the corresponding left and right singular vectors, respectively, $j = 1, \ldots, n+1$.

Let A be of full column rank (i.e. $\sigma'_n > 0$ and, subsequently, $\sigma_{n+1} = 0$ iff the system (1.1) is compatible) and let σ_{n+1} be simple. Define the *correction matrix* $[\,g\,|\,E\,] \equiv -u_{n+1}\,\sigma_{n+1}\,v_{n+1}^T$, $\|[\,g\,|\,E\,]\|_F = \|[\,g\,|\,E\,]\| = \sigma_{n+1}$. The *corrected matrix* $[\,b+g\,|\,A+E\,]$ represents, by Eckart-Young-Mirsky theorem (see [84, Theorem 2.3, p. 31]), the unique best rank n approximation of $[\,b\,|\,A\,]$ in the Frobenius norm (and also in the 2-norm), see also Theorem 2.2. Since σ_{n+1} is simple, the correction as well as the corrected matrices are unique. The right singular vector v_{n+1} represents a basis of the null space of the corrected matrix, i.e. $[\,b+g\,|\,A+E\,]\,v_{n+1} = 0$.

If the first component of the vector v_{n+1} is nonzero, i.e. $\gamma \equiv e_1^T\,v_{n+1} \neq 0$, then scaling v_{n+1} such that the first component is equal to -1 gives

$$\left[\begin{array}{c} -1 \\ \hline x_{\mathrm{TLS}} \end{array}\right] \equiv -\,\frac{1}{\gamma}\,v_{n+1}\,, \qquad \text{and} \qquad \left[\;b+g\;\middle|\;A+E\;\right]\left[\begin{array}{c} -1 \\ \hline x_{\mathrm{TLS}} \end{array}\right] = 0\,;$$

compare with (1.6). Because σ_{n+1} is simple, the corrected and the correction matrices are unique, thus the vector x_{TLS} represents the *unique TLS solution* of the problem (1.3). This solution can be expressed in the closed form

$$x_{\mathrm{TLS}} \equiv (A^T\,A - \sigma_{n+1}^2\,I_n)^{-1}\,A^T\,b\,,$$

see [31, 84].

If the first component of the vector v_{n+1} is zero, i.e. $\gamma \equiv e_1^T\,v_{n+1} = 0$, then the TLS problem (1.3) does not have a solution, see also [31, 84].

Golub and Van Loan give in [31] a sufficient condition for the existence of the TLS solution, here formulated in the following theorem.

Theorem 1.1. *Let σ'_j be the jth largest singular value of $A \in \mathbb{R}^{m\times n}$ and σ_j the jth largest singular value of $[\,b\,|\,A\,]$ with v_j the corresponding right singular vector, $m > n$. If*

$$\sigma'_n > \sigma_{n+1}\,, \tag{1.7}$$

then $\sigma_n > \sigma_{n+1}$ and $e_1^T\,v_{n+1} \neq 0$.

The first part follows immediately from the interlacing theorem for singular values, see [79], or [66, p. 203], or see Theorem 2.1 with Remark 2.3, the inequalities (2.12), further in the text. For the proof of the second part see for example [84, proof of Lemma 3.1, pp. 64–65]. For a more general form of Theorem 1.1, see Theorem 3.1 together with Corollary 3.1 in the further text.

The Golub, Van Loan condition (1.7) ensures that the smallest singular value of the extended matrix $[\,b\,|\,A\,]$ is simple and the corresponding right singular vector has nonzero first component, and, subsequently, it ensures existence of the TLS solution. This condition is, however, intricate because it is only sufficient *but not necessary* for the existence of a TLS solution. (In fact, the condition (1.7) is necessary and sufficient for the existence of the *unique* TLS solution.) If $\sigma'_n = \sigma_{n+1}$, then it may happen either $\sigma_n > \sigma_{n+1}$ with $e_1^T\,v_{n+1} = 0$, which means that the TLS problem does not have a solution, or $\sigma_n = \sigma_{n+1}$. In this case a TLS solution still may exist or may not exist. Thus, now we focus on the case when the smallest singular value of $[\,b\,|\,A\,]$ is multiple, i.e. $\sigma_n = \sigma_{n+1}$.

1.3.2 Problems with nonunique TLS solution

We still assume $A^T b \neq 0$ and $m > n$. Let A be of full column rank and let σ_{n+1} be multiple. In particular there is an integer p such that

$$\sigma_p > \sigma_{p+1} = \ldots = \sigma_{n+1} .$$

The case $p = n$ reduces to the previous case. If $p = 0$, i.e. $\sigma_1 = \ldots = \sigma_{n+1}$, then $[\, b \,|\, A \,]^T [\, b \,|\, A \,] = \sigma_1^2 I_{n+1}$, and thus the columns of $[\, b \,|\, A \,]$ are mutually orthogonal (and σ_p is nonexistent). In this case the TLS problem has a nonunique solution, and from the construction below it will be clear that the minimum 2-norm TLS solution is trivial, $x_{\mathrm{TLS}} = 0$. Therefore for simplification of notation we consider $0 < p < n$ in the further text.

Since σ_{n+1} is multiple, a minimal correction matrix reducing the rank of $[\, b \,|\, A \,]$ to n is no longer unique. For an arbitrary given matrix $Q \in \mathbb{R}^{(n-p+1)\times(n-p+1)}$, $Q^{-1} = Q^T$, denote $\tilde{v} \equiv [\, v_{p+1}, \ldots, v_{n+1} \,] Q\, e_{n-p+1}$, a unit vector from the right singular vector subspace associated with σ_{n+1}, and $\tilde{u} \equiv [\, u_{p+1}, \ldots, u_{n+1} \,] Q\, e_{n-p+1}$, the corresponding unit vector from the left singular vector subspace. The matrix $[\, g \,|\, E \,] \equiv -\tilde{u}\, \sigma_{n+1}\, \tilde{v}^T$, $\|[\, g \,|\, E \,]\|_F = \|[\, g \,|\, E \,]\| = \sigma_{n+1}$, represents, by Eckart-Young-Mirsky theorem, a minimal norm correction such that $[\, b + g \,|\, A + E \,]$ is a rank n approximation of $[\, b \,|\, A \,]$. Because Q is arbitrary, the correction as well as the corrected matrices are not unique.

Similarly to the previous section, if $e_1^T \tilde{v} \neq 0$, then $\tilde{v}$ can be used for the construction of a solution of the TLS problem (1.3), by scaling $\tilde{v}$ such that the first component is equal to -1. Consequently, if there exists a vector with nonzero first component in the subspace $\mathcal{R}([\, v_{p+1}, \ldots, v_{n+1} \,])$, i.e. if $e_1^T [\, v_{p+1}, \ldots, v_{n+1} \,] \neq 0$, then the TLS problem (1.3) has a solution, but, clearly, this solution is not unique. The goal is to find the *minimum 2-norm TLS solution.*

Denote $\tilde{v} = (\, \tilde{\gamma}, w^T \,)^T$; the norm of the solution constructed from $\tilde{v}$ is equal to $\tilde{\gamma}^{-1} \|w\|$, where $\|w\|^2 = \|\tilde{v}\|^2 - \tilde{\gamma}^2 = 1 - \tilde{\gamma}^2$. Thus the goal is to minimize $\tilde{\gamma}^{-1} (1 - \tilde{\gamma}^2)^{1/2}$, i.e., to maximize $\tilde{\gamma}$. The minimum 2-norm TLS solution is obtained by choosing Q such that the first component of $\tilde{v}$ is maximal over all unit vectors in $\mathcal{R}([\, v_{p+1}, \ldots, v_{n+1} \,])$. Put $Q \equiv H$, the Householder reflection matrix such that

$$\left(e_1^T \left[\; v_{p+1}, \ldots, v_{n+1} \;\right] \right) H = \left[\; 0, \ldots, 0, \gamma \;\right], \qquad \text{where}$$
$$\gamma \equiv \left\| e_1^T \left[\; v_{p+1}, \ldots, v_{n+1} \;\right] \right\|,$$

and put $v \equiv [\, v_{p+1}, \ldots, v_{n+1} \,] H\, e_{n-p+1}$. Scaling v gives the minimum 2-norm TLS solution

$$\left[\frac{-1}{x_{\mathrm{TLS}}} \right] \equiv -\frac{1}{\gamma} \left[\; v_{p+1}, \ldots, v_{n+1} \;\right] H\, e_{n-p+1} = -\frac{1}{\gamma}\, v ,$$

with $\|x_{\mathrm{TLS}}\| = \gamma^{-1} (1 - \gamma^2)^{1/2}$. This minimum 2-norm TLS solution can be expressed in the closed form

$$x_{\mathrm{TLS}} \equiv (A^T A - \sigma_{n+1}^2 I_n)^{-1} A^T b ,$$

see [31, 84].

If all (unit) vectors in the subspace $\mathcal{R}([\, v_{p+1}, \ldots, v_{n+1} \,])$ have zero first components, i.e. if $e_1^T [\, v_{p+1}, \ldots, v_{n+1} \,] = 0$, then the TLS problem (1.3) does not have a solution, see also [31, 84].

Van Huffel and Vandewalle give in [84] the following equivalence which generalizes Theorem 1.1.

Theorem 1.2. *Let σ'_j be the jth largest singular value of $A \in \mathbb{R}^{m\times n}$ and σ_j the jth largest singular value of $[\,b\,|\,A\,]$ with v_j the corresponding right singular vector, $m > n$. Then the following two conditions are equivalent:*

(i) $\sigma'_p > \sigma_{p+1} = \ldots = \sigma_{n+1}\,,$

(ii) $\sigma_p > \sigma_{p+1} = \ldots = \sigma_{n+1}$ *and* $e_1^T\,[\,v_{p+1}\,,\,\ldots\,,\,v_{n+1}\,] \neq 0\,.$

For a more general form of this theorem and its proof see [84, Corollary 3.4, p. 65], and also Corollary 3.2 in the further text.

In fact, the condition (i) is necessary and sufficient for the existence of the TLS solution. If $p = n$, then it reduces to (1.7) and the statement of this theorem says that $\sigma'_n > \sigma_n$ iff $\sigma_n > \sigma_{n+1}$ with $e_1^T\,v_{n+1} \neq 0$ (i.e., it reduces to the necessary and sufficient condition for existence the *unique* TLS solution, as mentioned before). But the condition (i) is very complicated because it combines singular values of the matrix A and the extended matrix $[\,b\,|\,A\,]$ together with their multiplicities. The situation becomes transparent with the usage of the core problem concept [64], see also Section 1.4.

In both Section 1.3.1 and Section 1.3.2 we omitted to analyze the problems with rank deficient A. Thus let A be rank deficient, i.e. $\sigma'_n = 0$. If $b \notin \mathcal{R}\,(A)$ (otherwise there exists a solution in the classical meaning, and $x \equiv A^\dagger\,b$ represents the minimum 2-norm solution), then singular values $\sigma'_n \equiv \sigma_{n+1} = 0$ have the same multiplicities, all the right singular vectors corresponding to σ_{n+1} have zero first components, and thus the TLS problem does not have a solution. This assertion will be clarified through the core problem concept.

Now we focus on the case when the right singular vector subspace associated with the smallest singular value of $[\,b\,|\,A\,]$ does not contain any vector with nonzero first component. Here as well as in the previous section is curtly asserted that in this case the TLS problem (1.3) does not have a solution. The following section introduces the so called *nongeneric concept* used to solving such problems. For more detailed discussion about the nonexistence of a TLS solution, as well as about the meaning of the nongeneric solution we refer to the core problem theory [64] in the further text, see Section 1.4 and particularly Section 1.4.2, Case C.

1.3.3 Nongeneric solution by Van Huffel and Vandewalle

An unpleasant situation occurs in both previous cases (in Section 1.3.1 and Section 1.3.2) when the right singular vector subspace associated with the smallest singular value σ_{n+1} of $[\,b\,|\,A\,]$ does not contain a vector with nonzero first component. This situation is provided by the fact that the correlation between columns of the matrix A is stronger than the correlation between the column space of A and the right-hand side b. The extreme case when the columns of A are linearly dependent and the problem is incompatible is illustrated by Example 1.1 above. In such case there is no right singular vector that can be used for construction of a solution.

The idea of the so called nongeneric concept is the following, see [84]: because the solution can not be constructed from a vector corresponding to the smallest singular value, we try to use another, bigger, singular value and the corresponding left and right singular vectors for construction of a correction matrix and a solution. But, such a solution does not solve the original TLS problem (1.3).

Recall that we still assume $A^T\,b \neq 0$ and $m > n$. Let $\sigma_t > \sigma_{n+1}$ be the smallest singular value of $[\,b\,|\,A\,]$ such that $e_1^T\,v_t \neq 0$, i.e. $e_1^T\,[\,v_{t+1}\,,\,\ldots\,,\,v_{n+1}\,] = 0$ (this case includes all incompatible problems with rank deficient A, as mentioned).

Since $V \equiv [\, v_1 \,, \ldots \,, v_{n+1} \,]$ is an orthogonal matrix, such a singular value always exists. Put $[\, g \,|\, E \,] \equiv u_t \, \sigma_t \, v_t^T$, $\|[\, g \,|\, E \,]\|_F = \|[\, g \,|\, E \,]\| = \sigma_t$. Similarly to the previous cases $[\, b + g \,|\, A + E \,]\, v_t = 0$ and thus scaling the vector v_t such that the first component is equal to -1 gives the solution of the corrected system. This solution is in [84] called *nongeneric solution*.

Obviously, if $\sigma_{t-1} = \sigma_t$ (with $t > 1$) or $\sigma_t = \sigma_{t+1}$ (with $t < n$), then the correction as well as the solution are not unique. In the case of nonuniqueness the goal is to find the *minimum 2-norm nongeneric solution*. In order to handle a possible nonuniqueness define an integer $\tilde{p}$ such that

$$\sigma_{\tilde{p}} > \sigma_{\tilde{p}+1} = \ldots = \sigma_t \geq \ldots \geq \sigma_p > \sigma_{p+1} = \ldots = \sigma_{n+1} \,.$$

If $\tilde{p} = 0$, then it can be shown that the right-hand side b is orthogonal to the column space of A, and from the construction below it will be clear that the minimum 2-norm nongeneric solution becomes trivial, $x_{\mathrm{NGN}} = 0$ (and $\sigma_{\tilde{p}}$ is nonexistent).

Similarly to Section 1.3.2, define $H \in \mathbb{R}^{(n-\tilde{p}+1)\times(n-\tilde{p}+1)}$, the Householder reflection matrix such that

$$\begin{aligned} \left(e_1^T \left[\; v_{\tilde{p}+1} \,, \ldots \,, v_{n+1} \;\right] \right) H &= \left[\; 0 \,, \ldots \,, 0 \,, \gamma \;\right] \,, \qquad \text{where} \\ \gamma \equiv \left\| e_1^T \left[\; v_{\tilde{p}+1} \,, \ldots \,, v_{n+1} \;\right] \right\| &= \left\| e_1^T \left[\; v_{\tilde{p}+1} \,, \ldots \,, v_t \;\right] \right\| , \end{aligned}$$

and put $u \equiv [\, u_{\tilde{p}+1} \,, \ldots \,, u_{n+1} \,] \, H \, e_{n-\tilde{p}+1}$, and $v \equiv [\, v_{\tilde{p}+1} \,, \ldots \,, v_{n+1} \,] \, H \, e_{n-\tilde{p}+1}$. Because $e_1^T \, [\, v_{t+1} \,, \ldots \,, v_{n+1} \,] = 0$, the matrix H can be chosen such that it has a block skew diagonal structure with two orthogonal blocks on the skew diagonal, with the lower left block the identity matrix I_{n-t+1}. Then the upper right block $H_{12} \in \mathbb{R}^{(t-\tilde{p})\times(t-\tilde{p})}$ represents the Householder reflection matrix such that $[\, v_{\tilde{p}+1} \,, \ldots \,, v_t \,] \, H_{12} = [\, 0 \,, \ldots \,, 0 \,, \gamma \,]$. Obviously $v \in \mathcal{R}\,(\, v_{\tilde{p}+1} \,, \ldots \,, v_t \,)$ and $v \perp \mathcal{R}\,(\, v_{t+1} \,, \ldots \,, v_{n+1} \,)$. The matrix $[\, g \,|\, E \,] \equiv -\, u \, \sigma_t \, v^T$ has Frobenius norm (and also the 2-norm) equal to σ_t. Scaling v such that the first component is equal to -1 gives the minimum 2-norm nongeneric solution

$$\left[\begin{array}{c} -1 \\ \hline x_{\mathrm{NGN}} \end{array} \right] \equiv -\, \frac{1}{\gamma} \left[\; v_{\tilde{p}+1} \,, \ldots \,, v_{n+1} \;\right] H \, e_{n-\tilde{p}+1} = -\, \frac{1}{\gamma} \, v \,,$$

see [84].

The following remark specifies the minimization problem which is solved by the minimum 2-norm nongeneric solution.

Remark 1.2. *The nongeneric solution x_{NGN} defined above represents a solution of the constrained minimization problem*

$$\begin{aligned} &\min_{x,E,g} \left\| \left[\; g \;\middle|\; E \;\right] \right\|_F \qquad \textit{subject to} \qquad (A + E)\, x = b + g \\ &\textit{with the constraint} \quad \left[\; g \;\middle|\; E \;\right] \left[\; v_{t+1} \,, \ldots \,, v_{n+1} \;\right] = 0 \,. \end{aligned} \tag{1.8}$$

The minimum norm nongeneric solution represents the minimum norm solution of (1.8). See also [84, Definition 3.2, pp. 68–69].

The nongeneric solution is also related to the so called truncated TLS problem, see [84, note on p. 82], or [88, 17].

The additional constraint in Remark 1.2 ensures that the correction matrix is constructed from the right singular vector which is orthogonal to the *unwanted directions* $v_{t+1} \,, \ldots \,, v_{n+1}$. Without this constraint, (1.8) reduces to (1.3), and, e.g., the unit vector $\tilde{v} \equiv \sin(\alpha)\, v_t + \cos(\alpha)\, v_{n+1}$, with nonzero first component for any $0 < \alpha \leq \pi/2$, can be used for the construction of a solution. Obviously, for

$\alpha \longrightarrow 0$ the norm of the correction matrix constructed using $\tilde{v}$ goes to the greatest lower bound equal to σ_{n+1}, but, simultaneously the norm of the solution goes to the infinity. Thus the TLS problem (1.3) does not have a solution. Obviously, the minimum norm nongeneric solution does not solve (1.3).

From the algorithmic point of view, the minimum 2-norm nongeneric solution concept is a consistent extension of the minimum 2-norm TLS solution, see e.g. the TLS classical algorithm in [84, Algorithm 3.1, pp. 87–88], or, alternatively, Algorithm 3.1 in the further text. This algorithm computes, in general, the minimum 2-norm nongeneric solution, and, for $d = 1$, it automatically returns the minimum 2-norm TLS solution if it exits.

Now, we already described all the possibilities that can occur. It remains to justify the addition of zero rows in (1.1) in order to satisfy the condition $m > n$, and to show that the incompatible problem with rank deficient matrix A does not have a TLS solution. Both can be easily shown through the core problem concept.

1.4 Core problem theory of Paige and Strakoš

A new contribution to the theory and computation of linear approximation problems was published in a series of papers [61, 62, 64]. Here the authors define a core problem within the orthogonally (unitarily) invariant linear approximation problem (1.1). It is proposed to orthogonally transform the original approximation problems to a block form that allows to separate the necessary and sufficient information present in the data A, b, from the redundant information. It is shown that the so called *core reduction* represents a theoretical basis for several well known techniques as well as for new future developments.

Assuming $A^T b \neq 0$ and that the approximation problem (1.1) is orthogonally invariant, i.e. that the solution is independent on a particular choice of bases in $\mathbb{R}^m$ and $\mathbb{R}^n$, it easy to see that there exists an orthogonal transformation of the form

$$P^T \left[\begin{array}{c|c} b & A \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q \end{array}\right] = P^T \left[\begin{array}{c|c} b & A\,Q \end{array}\right] = \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline\hline 0 & 0 & A_{22} \end{array}\right], \qquad (1.9)$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, and where A_{22} might have row and/or column dimensions equal to zero. In the nontrivial case (when A_{22} has at least one row and one column, even if $A_{22} = 0$) both the singular value decompositions (SVD) of $[\,b\,|\,A\,]$ and A can be easily got as a direct sum of the SVDs of the blocks $[\,b_1\,|\,A_{11}\,]$ and A_{22}, and A_{11} and A_{22}, respectively. The original approximation problem $A\,x \approx b$ is in this way decomposed into two *independent* approximation subproblems,

$$A_{11}\,x_1 \approx b_1\,, \qquad A_{22}\,x_2 \approx 0\,, \qquad \text{where} \qquad x \equiv Q \left[\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right].$$

The second subproblem $A_{22}\,x_2 \approx 0$ has a trivial solution $x_2 = 0$, and thus only the first subproblem $A_{11}\,x_1 \approx b_1$ needs to be solved, see [64]. Paige and Strakoš formulate the following definition.

Definition 1.1 (Core problem)**.** *The subproblem $A_{11}\,x_1 \approx b_1$ is a core problem within the approximation problem $A\,x \approx b$ if $[\,b_1\,|\,A_{11}\,]$ is minimally dimensioned and A_{22} maximally dimensioned subject to (1.9).*

For any transformation (1.9) the subproblem $A_{11}\,x_1 \approx b_1$ contains all the sufficient information for solving the original problem. Since the core problem is the minimally

dimensioned subproblem, i.e. the subproblem can not be reduced more, it must contain all the sufficient and only the necessary information for solving $A\,x \approx b$.

Understanding of the minimal dimensionality of $A_{11}\,x_1 \approx b_1$ can be gained by the following construction, which shows how to concentrate the relevant information into A_{11} and b_1, while moving the irrelevant and redundant information into A_{22}, see [64]. Let A have rank r and consider the SVD

$$A \;=\; U'\,\Sigma'\,(V')^T\,, \qquad \Sigma' \;\equiv\; \left[\begin{array}{c|c} \Xi & 0 \\ \hline 0 & 0 \end{array}\right] \;\equiv\; \left[\begin{array}{c|c} \mathrm{diag}\,(\,\sigma'_1\,,\,\ldots\,,\,\sigma'_r\,) & 0 \\ \hline 0 & 0 \end{array}\right],$$

with singular values $\sigma'_1 \;\geq\; \ldots \;\geq\; \sigma'_r \;>\; 0$, $U' \in \mathbb{R}^{m\times m}$, $(U')^{-1} \;=\; (U')^T$, and $V' \in \mathbb{R}^{n\times n}$, $(V')^{-1} \;=\; (V')^T$. Moreover assume only k of the nonzero singular values of A to be distinct. Then

$$(U')^T\;\left[\;b\;\middle|\;A\,V'\;\right] \;=\; \left[\begin{array}{c||c|c} c & \Xi & 0 \\ \hline c_{k+1} & 0 & 0 \end{array}\right], \qquad \text{and} \qquad c \;\equiv\; \left[\;c_1^T\,,\,\ldots\,,\,c_k^T\;\right]^T\,,$$

where the partitioning of c respects the multiplicities of the singular values of A.

The singular values are unique in any SVD representation. But their ordering, and sometimes some singular vectors, are not unique. In order to obtain the core problem, the matrix $(U')^T\,[\,b\,|\,A\,V'\,]$ will be transformed further, while maintaining the SVD of A. For c_j, choose an orthogonal matrix H_j (e.g. the Householder reflection matrix) such that $H_j\,c_j \;=\; e_1\,\delta_j$, where $\delta_j \;\equiv\; \|c_j\|$, for $j \;=\; 1\,,\,\ldots\,,\,k\,,\,k+1$. Then put

$$\begin{aligned} G \;&\equiv\; \mathrm{diag}\,(\,H_1\,,\,\ldots\,,\,H_k\,,\,H_{k+1}\,)\,, \\ H \;&\equiv\; \mathrm{diag}\,(\,H_1\,,\,\ldots\,,\,H_k\,,\,I_{n-r}\,)\,, \end{aligned}$$

and replace the matrix U' by $U'\,G$ and V' by $V'\,H$. This transformation will leave Σ' unchanged and therefore preserves the SVD of A. In this way the vector c is transformed into a vector having at most one nonzero component corresponding to each block of equal singular values of A, and therefore the original right-hand side vector b is transformed into a vector having at most $k+1$ nonzero entries. Clearly, $\delta_j \;\neq\; 0$, $j \;=\; 1\,,\,\ldots\,,\,k$, if and only if the right-hand side b has nonzero projection onto the corresponding left singular vector subspace of A (i.e., $\delta_1 \;=\; \ldots \;=\; \delta_k \;=\; 0$ iff $b \;\perp\; \mathcal{R}$), and finally $\delta_{k+1} \;\neq\; 0$ iff $b \notin \mathcal{R}\,(A)$. Next permute the columns of $U'\,G$ and $V'\,H$ identically, in order to move the zero elements in the transformed c to the bottom of this vector, leaving d, the subvector of c with nonzero components only, at the top, while keeping Ξ diagonal. Finally if $\delta_{k+1} \;\neq\; 0$ move its row so that δ_{k+1} is immediately below d by a further permutation from the left to give, with obvious new notation and indexing,

$$\begin{aligned} (\,U'\,G\,\Pi_L\,)^T\;\left[\;b\;\middle|\;A\,(\,V'\,H\,\Pi_R\,)\;\right] \;&=\; \left[\begin{array}{c||c|c} d & \Xi_1 & 0 \\ \delta_{k+1} & 0 & 0 \\ \hline 0 & 0 & \Xi_2 \end{array}\right] \\ &\equiv\; \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right], \end{aligned} \tag{1.10}$$

the matrices Π_L, Π_R denote the permutations from the left and right, respectively, the vector d contains only the nonzero scalars $\delta_1\,,\,\ldots\,,\,\delta_k$, the matrix Ξ_1 is diagonal with simple and nonzero singular values; the row beginning with the scalar δ_{k+1} is nonexistent iff the problem (1.1) is compatible. The final partitioning in (1.10) corresponds to that in (1.9) with $P \;\equiv\; U'\,G\,\Pi_L$ and $Q \;\equiv\; V'\,H\,\Pi_R$. Denote $\bar{m}$, and $\bar{n} \;\equiv\; k$ the dimensions in (1.10) such that $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, $x_1 \in \mathbb{R}^{\bar{n}}$, and $b_1 \in \mathbb{R}^{\bar{m}}$; obviously $\bar{n} \;\leq\; \bar{m} \;\leq\; \bar{n}+1$.

It can be easily shown that the subproblem $A_{11}\,x_1 \approx b_1$ obtained by the transformation process (1.10) described above has indeed the desired minimality property, and thus it represents the core problem within $A\,x \approx b$, see also [64]. The core problem in the form given in (1.10) is called the *SVD form of the core problem.*

Remark 1.3. *Let $A\,x \approx b$ be an approximation problem with $m \le n$. From the construction of the SVD form of the core problem it immediately follows that the addition of zero rows in order to satisfy the condition $m > n$ required in Section 1.3.1, 1.3.2, or 1.3.3, does not change the core problem within $A\,x \approx b$. It extends only the matrix A_{22} which does not involve the solution $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T$.*

Remark 1.4. *Let $A\,x \approx b$ be an approximation problem and $A^T\,b = 0$. It is easy to see that core reduction results in a trivial matrix A_{11}, i.e., with no columns. Moreover, if $b = 0$, then both b_1 and A_{11} are trivial having no rows. In these cases the solution $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T = 0$ is equal to zero.*

A decomposition of the form (1.9) can also be computed directly by choosing orthogonal matrices P and Q in order to reduce $[\,b \,|\, A\,]$ to a real upper bidiagonal matrix, see [64]. It can be done using for example Householder reflection matrices, see [32, §5.4.3, pp. 251–252]. The first zero element on the main diagonal or on the first superdiagonal determines the desired partitioning. The matrix A_{22} needs not be bidiagonalized. Alternatively the *partial* Golub-Kahan iterative bidiagonalization algorithm [27, 57] can be used. Putting $w_0 \equiv 0$ and the starting vector $s_1 \equiv b/\beta_1$, where $\beta_1 \equiv \|b\|$, the algorithm computes for $j = 1\,,\,2\,,\,\ldots$

$$\begin{aligned} w_j\,\alpha_j &\equiv A^T\,s_j - w_{j-1}\,\beta_j\,, \\ s_{j+1}\,\beta_{j+1} &\equiv A\,w_j - s_j\,\alpha_j\,, \end{aligned} \tag{1.11}$$

where $\|w_j\| = 1$, $\alpha_j \ge 0$, and $\|s_{j+1}\| = 1$, $\beta_{j+1} \ge 0$, until $\alpha_j = 0$ or $\beta_{j+1} = 0$, or until the dimensions of A are exceeded, i.e. $j = \min\{\,m\,,\,n\,\}$.

We present, for completeness, the basic properties of the Golub-Kahan bidiagonalization as given in [57]. Consider $\alpha_j > 0$, $\beta_j > 0$, for $j = 1\,,\,\ldots\,,\,k$, and $\beta_{k+1} > 0$, and denote $S_j \equiv [\,s_1\,,\,\ldots\,,\,s_j\,]$, $W_j \equiv [\,w_1\,,\,\ldots\,,\,w_j\,]$,

$$L_j \equiv \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_j & \alpha_j \end{bmatrix} \in \mathbb{R}^{j\times j} \qquad \text{and} \qquad L_{j+} \equiv \begin{bmatrix} L_j \\ \beta_{j+1} e_j^T \end{bmatrix} \in \mathbb{R}^{(j+1)\times j}\,.$$

Consequently $A^T\,S_j = W_j\,L_j^T$, $A\,W_j = S_{j+1}\,L_{j+}$, giving the fundamental properties $S_{j+1}^T\,S_{j+1} = I_{j+1}$, $W_j^T\,W_j = I_j$, and $S_j^T\,A\,W_j = L_j$, $W_j^T\,A^T\,S_{j+1} = L_{j+}^T$, for $j = 1\,,\,\ldots\,,\,t$. Summarizing, the Golub-Kahan bidiagonalization (1.11) of the matrix A with $s_1 \equiv b/\|b\|$ yields one of the following two situations.

Case 1. If $\alpha_j > 0$, $\beta_j > 0$, $j = 1\,,\,\ldots\,,\,\tilde{n}$, and $\beta_{\tilde{n}+1} = 0$ or $\tilde{n} = m$, then $S_{\tilde{n}}^T\,A\,W_{\tilde{n}} = L_{\tilde{n}}$ and

$$\left[\;\tilde{b}_1 \;\middle|\; \tilde{A}_{11}\;\right] \equiv S_{\tilde{n}}^T\left[\;b \;\middle|\; A\,W_{\tilde{n}}\;\right] = \left[\begin{array}{c|ccc} \beta_1 & \alpha_1 & & \\ & \beta_2 & \alpha_2 & \\ & & \ddots & \ddots \\ & & & \beta_{\tilde{n}} \quad \alpha_{\tilde{n}} \end{array}\right]. \tag{1.12}$$

Case 2. If $\alpha_j > 0$, $\beta_j > 0$, $j = 1\,,\,\ldots\,,\,\tilde{n}$, $\beta_{\tilde{n}+1} > 0$, and $\alpha_{\tilde{n}+1} = 0$ or

$\tilde{n} = n$, then $S_{\tilde{n}+1}^T A W_{\tilde{n}} = L_{\tilde{n}+}$ and

$$\left[\begin{array}{c|c} \tilde{b}_1 & \tilde{A}_{11} \end{array}\right] \equiv S_{\tilde{n}+1}^T \left[\begin{array}{c|c} b & A W_{\tilde{n}} \end{array}\right] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \ddots & \ddots & \\ & & & \beta_{\tilde{n}} & \alpha_{\tilde{n}} \\ & & & & \beta_{\tilde{n}+1} \end{array}\right]. \tag{1.13}$$

In both cases, the matrices $S_{\tilde{n}}$ or $S_{\tilde{n}+1}$, and $W_{\tilde{n}}$ represent the first $\tilde{n}$ or $\tilde{n}+1$ columns of the matrix P, and the first $\tilde{n}$ columns of the matrix Q in (1.9), respectively.

The subproblems (1.12), (1.13) have many properties, see [64]; some important of them can be easily derived from the well known properties of tridiagonal matrices. The following definition introduces a commonly used terminology [14, 21], and some classical results are summarized in the subsequent theorem [90, 66] and lemma.

Definition 1.2 (Jacobi matrix)**.** *A symmetric tridiagonal matrix with positive off-diagonal components is called Jacobi matrix.*

Theorem 1.3. *Let $H \in \mathbb{R}^{\rho\times\rho}$ be a (symmetric tridiagonal) Jacobi matrix. Denote by $H_j \in \mathbb{R}^{j\times j}$ its leading principal submatrix, $j = 1, \ldots, \rho$. Then the following (characteristic) polynomials,*

$$\begin{aligned} p_0(\lambda) &\equiv 1, \\ p_1(\lambda) &\equiv H_1 - \lambda, \\ p_2(\lambda) &\equiv \det(H_2 - \lambda I_2), \\ &\vdots \\ p_{\rho-1}(\lambda) &\equiv \det(H_{\rho-1} - \lambda I_{\rho-1}), \\ p_\rho(\lambda) &\equiv \det(H - \lambda I_\rho), \end{aligned} \tag{1.14}$$

have the Sturm sequence property, i.e., two subsequent polynomials can not have the same root, and, moreover, zeros of $p_{j-1}(\lambda)$ and $p_j(\lambda)$ strictly interlace, for $j = 1, \ldots, \rho$.

Proof. Denote $h_{i,j} \equiv e_i^T H e_j$, recall that $h_{i,j} = h_{j,i}$, and, if $|i-j| = 1$, then $h_{i,j} \neq 0$. Obviously

$$\begin{aligned} p_1(\lambda) &= (h_{1,1} - \lambda)\ p_0(\lambda), \\ p_j(\lambda) &= (h_{j,j} - \lambda)\ p_{j-1}(\lambda) - h_{j,j-1}^2\ p_{j-2}(\lambda), \end{aligned}$$

for $j = 2, \ldots, \rho$. Suppose that there exists $\xi \in \mathbb{R}$ such that $p_k(\xi) = p_{k-1}(\xi) = 0$ for some $k \geq 2$. Since $h_{k,k-1} \neq 0$, we have $p_{k-2}(\xi) = 0$. Induction gives $p_0(\xi) = 0$ which contradicts the fact that $p_0(\lambda) = 1$. Thus two subsequent polynomials can not have the same root. See also [90, Chapter 5, §36., §37., pp. 299–302].

The following rest of the proof is taken over [46, Chapter 4, p. 168]. The separation property can be proved by mathematical induction. That is, a simple plot of $p_2(\lambda)$ shows that the simple zero of $p_1(\lambda)$ separates the two simple zeros of $p_2(\lambda)$. Assume that the $j-2$ simple zeros of $p_{j-2}(\lambda)$ separate the $j-1$ simple zeros of $p_{j-1}(\lambda)$. Now, from (1.14), at each zero of $p_{j-1}(\lambda)$, the sign of $p_j(\lambda)$ is opposite

to the sign of $p_{j-2}(\lambda)$. But, by the induction hypothesis, $p_{j-2}(\lambda)$ changes sign between each pair of neighboring zeros of $p_{j-1}(\lambda)$. Therefore, $p_j(\lambda)$ also changes sign and hence has a zero *between* each neighboring pair of zeros of $p_{j-1}(\lambda)$. Now

$$\lim_{\lambda \to +\infty} p_j(\lambda) = +\infty, \qquad \lim_{\lambda \to -\infty} p_j(\lambda) = (-1)^j \infty, \qquad j = 1, 2, \ldots .$$

Therefore $p_j(\lambda)$ has a zero to the right of the largest zero of $p_{j-1}(\lambda)$ and a zero to the left of the smallest zero of $p_{j-1}(\lambda)$. On the other hand, $p_j(\lambda)$ can have no more than j zeros. Therefore, we have shown that the $j-1$ simple zeros of $p_{j-1}(\lambda)$ separate the j simple zeros of $p_j(\lambda)$. This proves the *strict interlacing (or separation) property* in the theorem. □

The second part of the proof can be shown alternatively as follows. Because the roots of $p_j(\lambda)$ represent eigenvalues of T_j, the interlacing theorem for eigenvalues, see [66, p. 203] (see also Theorem 2.1 and Remark 2.3, the inequalities (2.12), in the further text) gives the strict interlacing property.

Thus Theorem 1.3 says that the eigenvalues of H_j are *strictly interlaced* by the eigenvalues of H_{j-1}, i.e.,

$$\lambda_1(H_j) > \lambda_1(H_{j-1}) > \lambda_2(H_j) > \ldots > \lambda_{j-1}(H_{j-1}) > \lambda_j(H_j), \tag{1.15}$$

and thus eigenvalues of H_j must be *simple*. Consequently, an arbitrary (symmetric tridiagonal) Jacobi matrix has distinct (and simple) eigenvalues.

Lemma 1.1. *Let $H \in \mathbb{R}^{\rho \times \rho}$ be a (symmetric tridiagonal) Jacobi matrix. Then:*

(i) *The matrix H has distinct eigenvalues.*

(ii) *The first (as well as the last) components of all eigenvectors of H are nonzero.*

Proof. Property (i) is the direct consequence of Theorem 1.3 as shown above. Further, (ii) can be shown by contradiction. Let x, $\|x\| = 1$, $Hx = x\lambda$, with $x_1 \equiv e_1^T x = 0$. Equating the corresponding components in $Hx = x\lambda$ gives $h_{1,1}x_1 + h_{1,2}x_2 = x_1\lambda$, and $h_{j-1,j}x_{j-1} + h_{j,j}x_j + h_{j,j+1}x_{j+1} = x_j\lambda$, for $j = 2, \ldots, \rho-1$. Since $h_{j,j+1} \neq 0$ we have $x_{j+1} = 0$, for $j = 1, \ldots, \rho-1$, which contradicts $\|x\| = 1$. Both (i) and (ii) are the basic and well known properties of Jacobi matrices, see e.g. [66, Lemma 7.7.1, Theorem 7.9.3]. □

Now the properties of (1.12), (1.13) can be analyzed. Because $\alpha_j > 0$ and $\beta_j > 0$ for $j = 1, \ldots, \tilde{n}$ in both cases (1.12) and (1.13), it is easy to see that the tridiagonal matrix

$$\begin{aligned} \tilde{A}_{11}^T \tilde{A}_{11} &= L_{\tilde{n}}^T L_{\tilde{n}} + \beta_{\tilde{n}+1}^2 e_{\tilde{n}} e_{\tilde{n}}^T \\ &= \begin{bmatrix} \alpha_1^2+\beta_2^2 & \alpha_2\beta_2 & & \\ \alpha_2\beta_2 & \alpha_2^2+\beta_3^2 & \ddots & \\ & \ddots & \ddots & \alpha_{\tilde{n}}\beta_{\tilde{n}} \\ & & \alpha_{\tilde{n}}\beta_{\tilde{n}} & \alpha_{\tilde{n}}^2+\beta_{\tilde{n}+1}^2 \end{bmatrix}, \end{aligned} \tag{1.16}$$

is a Jacobi matrix (in the compatible case $\beta_{\tilde{n}+1} = 0$). Thus in both cases the matrix $\tilde{A}_{11}$ has simple singular values (square roots of eigenvalues of $\tilde{A}_{11}^T \tilde{A}_{11}$), by (i) in Lemma 1.1. Obviously $\tilde{A}_{11}$ is of full column rank (and $[\,\tilde{b}_1 \,|\, \tilde{A}_{11}\,]$ of full row rank) in both cases, see e.g. [64, Remark 3.1]. Thus the singular values of $\tilde{A}_{11}$ are

nonzero. Further, the matrices

$$\tilde{A}_{11}\,\tilde{A}_{11}^T \;=\; L_{\tilde{n}}\,L_{\tilde{n}}^T \;=\; \begin{bmatrix} \alpha_1^2 & \alpha_1\,\beta_2 & & \\ \alpha_1\,\beta_2 & \alpha_2^2+\beta_2^2 & \ddots & \\ & \ddots & \ddots & \alpha_{\tilde{n}-1}\,\beta_{\tilde{n}} \\ & & \alpha_{\tilde{n}-1}\,\beta_{\tilde{n}} & \alpha_{\tilde{n}}^2+\beta_{\tilde{n}}^2 \end{bmatrix}, \tag{1.17}$$

in Case 1, and

$$\tilde{A}_{11}\,\tilde{A}_{11}^T \;=\; L_{\tilde{n}+}\,L_{\tilde{n}+}^T \;=\; \begin{bmatrix} \alpha_1^2 & \alpha_1\,\beta_2 & & \\ \alpha_1\,\beta_2 & \alpha_2^2+\beta_2^2 & \ddots & \\ & \ddots & \ddots & \alpha_{\tilde{n}}\,\beta_{\tilde{n}+1} \\ & & \alpha_{\tilde{n}}\,\beta_{\tilde{n}+1} & \beta_{\tilde{n}+1}^2 \end{bmatrix}, \tag{1.18}$$

in Case 2, also represent Jacobi matrices. Eigenvectors of (1.17), (1.18) represent left singular vectors of $\tilde{A}_{11} \equiv L_{\tilde{n}}$ and $\tilde{A}_{11} \equiv L_{\tilde{n}+}$, respectively. Because $\tilde{b}_1 \equiv e_1\,\beta_1$, it has nonzero projections onto all left (one dimensional) singular vector subspaces of $\tilde{A}_{11}$ in both cases, by (ii) in Lemma 1.1.

Consequently, the SVD-based reduction described by (1.10) applied on subproblems (1.12) or (1.13) can not reduce their dimensions more. Vice versa, since (1.10) represents the core problem transformation, i.e. the dimensions of $[\,b_1 \,|\, A_{11}\,]$ are minimal, the bidiagonalization of (1.10) can not terminate sooner than in the $(\tilde{n}+1)$st step; this happens either with $\beta_{\tilde{n}+1} = 0$, in the compatible case, or with $\alpha_{\tilde{n}+1} = 0$, in the incompatible case.

Thus the Golub-Kahan algorithm (1.11) yields the core problem, i.e., $\tilde{n} \equiv \bar{n}$. Particularly, the subproblems (1.12) and (1.13) represent the compatible and the incompatible case, respectively. The core problem in the form $\tilde{A}_{11}\,\tilde{x}_1 \approx \tilde{b}_1$ given in (1.12), (1.13) is called the *banded (bidiagonal) form of the core problem.*

As shown, a subproblem representing the core problem has several properties (that are independent on a particular form, e.g., SVD form, bidiagonal, or any other). We summarize the most important of them:

(G1) The matrix A_{11} is of full *column* rank equal to $\bar{n}$.

(G2) The right-hand side b_1 is of full *column* rank (i.e., b_1 is nonzero).

(G3) The matrices $(U_j')^T\,b_1$ are of full *row* rank for all j, where U_j' denotes an orthonormal basis of the left singular vector subspace corresponding to the jth distinct singular value of A_{11}.

(G4) The matrix $[\,b_1 \,|\, A_{11}\,]$ is of full *row* rank.

(G5) The matrix A_{11} has no zero or multiple singular values, so any zero singular values or repeats that A has, must appear in A_{22}.

First we explain the property (G3) which looks complicated. The right-hand b_1 is a column vector, thus for any matrix U_j', $(U_j')^T\,b_1$ is a column vector, which has full *row* rank iff it is a nonzero scalar. The number of rows of $(U_j')^T\,b_1$, i.e. the number of columns of U_j', is identical to the dimension of the jth left singular vector subspace, i.e. the multiplicity of the jth distinct thus singular value of A_{11}, singular values of A_{11} must be simple. Moreover, the right-hand side b_1 has nonzero component in all (one dimensional) left singular vector subspaces of A_{11}. It is easy to see that the property (G2) (which in the single right-hand side case reduces to the assertion $b_1 \neq 0$) is implied by the property (G3) under the assumption $A^T\,b \neq 0$ (and thus $A_{11}^T\,b_1 \neq 0$). Further, (G5) is also implied by (G3). Finally it can be shown, that

property (G4) is implied by (G1) and (G3). (Even if the property (G2) seems trivial it will be found useful in the multiple right-hand side case.)

Summarizing, it was shown that for any orthogonally invariant problem $A\,x \approx b$ there exists an orthogonal transformation of the form (1.9) yielding the core problem that has some notable properties. It contains all the necessary, and only the sufficient information for solving the original problem. A solution of the original problem can be given by $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T$ where x_1 is the solution of the core problem. Consequently, we focus on the TLS formulation applied on the core problem, in the next section.

1.4.1 Core problem and the TLS formulation

Let $A\,x \approx b$ be a linear approximation problem and $\tilde{A}_{11}\,\tilde{x}_1 \approx \tilde{b}_1$ the core problem within $A\,x \approx b$ in the bidiagonal form, either (1.12) or (1.13). Because $\alpha_j > 0$ and $\beta_j > 0$ for $j = 1\,,\,\ldots\,,\,\bar{n}$, the matrix (1.16), i.e. $\tilde{A}_{11}^T\,\tilde{A}_{11}$, is a Jacobi matrix, in both cases. Similarly, the matrix

$$\left[\;\tilde{b}_1 \;\middle|\; \tilde{A}_{11}\;\right]^T \left[\;\tilde{b}_1 \;\middle|\; \tilde{A}_{11}\;\right] = \left[\begin{array}{c|c} \tilde{b}_1^T\,\tilde{b}_1 & \tilde{b}_1^T\,\tilde{A}_{11} \\ \hline \tilde{A}_{11}^T\,\tilde{b}_1 & \tilde{A}_{11}^T\,\tilde{A}_{11} \end{array}\right] \tag{1.19}$$

is a Jacobi matrix – thus, all its eigenvalues are simple, all its eigenvectors have nonzero first and last components, by Lemma 1.1. The matrix (1.19) contains (1.16) as a trailing principal submatrix, which is crucial in the forthcoming analysis.

Note that Theorem 1.3 can be easily reformulated for the characteristic polynomials of the trailing principal submatrices, i.e., matrices $H_j \in \mathbb{R}^{j\times j}$ obtained from $H \in \mathbb{R}^{\rho\times\rho}$ by removing the first $\rho - j$ rows and columns, $j = 0\,,\,\ldots\,,\,\rho$ (in the notation of Theorem 1.3). Together with the interlacing theorem for eigenvalues, see [66, p. 203] (see also Theorem 2.1 and Remark 2.3, the inequalities (2.12), in the further text) this modification of Theorem 1.3 implies the strict interlacing property (1.15) for matrices (1.16), (1.19); see also discussion after Theorem 1.3.

Thus the eigenvalues of $[\,\tilde{b}_1 \,|\, \tilde{A}_{11}\,]^T\,[\,\tilde{b}_1 \,|\, \tilde{A}_{11}\,]$ are *strictly interlaced* by the eigenvalues of $\tilde{A}_{11}^T\,\tilde{A}_{11}$. Because the singular values are independent on the given form of the core problem, we omit tildes in the further text; we obtain:

Case 1. In the compatible case (1.12), distinct and nonzero singular values of A_{11} strictly interlace the singular values of $[\,b_1 \,|\, A_{11}\,]$ together with zero. But here we are not interested in the compatible case, because the compatible problems always have solution in the classical sense.

Case 2. In the incompatible case (1.13), both matrices A_{11} and $[\,b_1 \,|\, A_{11}\,]$ have distinct and nonzero singular values and the singular values of A_{11} strictly interlace the singular values of $[\,b_1 \,|\, A_{11}\,]$,

$$\sigma_{\bar{n}}\,(A_{11}) \;>\; \sigma_{\bar{n}+1}\,(\,[\,b_1 \,|\, A_{11}\,]\,)\,. \tag{1.20}$$

Appending the right-hand side vector b_1 to the core problem matrix A_{11} decreases the smallest singular value. The core problem always satisfies the Golub, Van Loan condition (1.7) and thus it always has the unique TLS solution (i.e., the smallest singular value $\sigma_{\bar{n}+1}\,([\,b_1 \,|\, A_{11}\,])$ is simple and the corresponding right singular vector has nonzero first component).

It remains to compare the solution $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T$ obtained using the core problem transformation (1.9) to all the TLS formulation in [31, 84], presented also in Sections 1.3.1, 1.3.2, and 1.3.3.

1.4.2 Relation to earlier work

Let $A\,x \approx b$ be a general linear approximation problem and $A_{11}\,x_1 \approx b_1$ a core problem within $A\,x \approx b$ obtained by a transformation to the form (1.9). Denote for simplicity x_1 the unique TLS solution of this core problem. Now, the question is, what this solution represents in the original variables.

As in Sections 1.3.1, 1.3.2, and 1.3.3, assume $m > n$ (add zero rows if necessary). Moreover we focus on the incompatible case, i.e. $b \notin \mathcal{R}(A)$. Consequently the matrix $[\,b_1 \,|\, A_{11}\,]$ is square and the matrix A_{22} is either square (iff $m = n+1$), or is has more rows than columns. Denote $\sigma_{\min}(M)$ the smallest singular value of M for simplicity (for all $[\,b_1 \,|\, A_{11}\,]$, A_{11} and A_{22}, the index of the smallest singular value is equal to the number of their columns). Recall that the SVD of $[\,b \,|\, A\,]$ can be obtained as a direct sum of SVDs of $[\,b_1 \,|\, A_{11}\,]$ and A_{22}, just by extending the singular vectors corresponding to the first block by zeros on the bottom and the singular vectors corresponding to the second block by zeros on the top.

There are three different possibilities:

Case A. If

$$\sigma_{\min}(A_{22}) > \sigma_{\min}(\,[\,b_1 \,|\, A_{11}\,]\,),$$

then, because $\sigma_{\min}(A_{11}) > \sigma_{\min}([\,b_1 \,|\, A_{11}\,])$ by (1.20), the smallest singular value of $[\,b \,|\, A\,]$ is simple and

$$\min\{\,\sigma_{\min}(A_{11})\,,\,\sigma_{\min}(A_{22})\,\} \equiv \sigma_n(A) > \sigma_{n+1}(\,[\,b \,|\, A\,]\,) \equiv \sigma_{\min}(\,[\,b_1 \,|\, A_{11}\,]\,).$$

Consequently the original problem $A\,x \approx b$ has by (1.7) the unique TLS solution.

Consider the SVD of $[\,b_1 \,|\, A_{11}\,] = U_1\,\Sigma_1\,V_1^T$ and the SVD of $A_{22} = U_2\,\Sigma_2\,V_2^T$, the matrix of right singular vectors of the original problem is

$$V = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q \end{array}\right] \bar{V}\,, \qquad \text{where} \qquad \bar{V} \equiv \left[\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array}\right] \Pi\,,$$

Π is a permutation matrix which sorts singular values of Σ_1 and Σ_2 in nonincreasing sequence, and Q is the orthogonal matrix given in (1.9). Because the smallest singular value of $[\,b \,|\, A\,]$ is simple and because the transformation (1.9) does not change the first components of the right singular vectors of $[\,b \,|\, A\,]$ (i.e., $v_j = \mathrm{diag}(\,1\,,\,Q\,)\,\bar{v}_j$, where v_j and $\bar{v}_j$ are columns of V and $\bar{V}$, respectively, $j = 1\,,\,\ldots\,,\,n+1$), the right singular vector corresponding to $\sigma_{n+1}([\,b \,|\, A\,])$ has nonzero first component. The TLS solution of the original problem is given by this right singular vector and obviously it is identical to the solution of the core problem transformed back to the original variables, i.e. $x_{\mathrm{TLS}} \equiv Q\,[\,x_1^T \,|\, 0\,]^T$.

Case B. If

$$\sigma_{\min}(A_{22}) = \sigma_{\min}(\,[\,b_1 \,|\, A_{11}\,]\,),$$

then the smallest singular value of $[\,b \,|\, A\,]$ is multiple and it is equal to $\sigma_n(A)$. The Golub, Van Loan condition (1.7) is no more satisfied. From (1.20) it follows that *the multiplicity of the smallest singular value of A increase by appending the right-hand side b.*

As in the previous case, because the transformation (1.9) does not change the first components of the right singular vectors of $[\,b \,|\, A\,]$, there exists a right singular vector $\bar{v}_\ell$ corresponding to $\sigma_{n+1}([\,b \,|\, A\,])$ (from the SVD of $[\,b_1 \,|\, A_{11}\,]$) which has nonzero first component, all other singular vectors corresponding to $\sigma_{n+1}([\,b \,|\, A\,])$ (from the SVD of A_{22}) have zero first components. Consequently the original problem $A\,x \approx b$ has a TLS solution but it is not unique. Obviously, the minimum norm TLS solution of the original problem is given by the right singular vector of $v_\ell = \mathrm{diag}(\,1\,,\,Q\,)\,\bar{v}_\ell$, i.e. it is identical to the solution of the core problem $[\,b_1 \,|\, A_{11}\,]$ transformed back to the original variables, i.e. $x_{\mathrm{TLS}} \equiv Q\,[\,x_1^T \,|\, 0\,]^T$.

Case C. If

$$\sigma_{\min}(A_{22}) < \sigma_{\min}([\,b_1 \,|\, A_{11}\,]),$$

then the singular values $\sigma_n(A) \equiv \sigma_{n+1}([\,b \,|\, A\,]) \equiv \sigma_{\min}(A_{22})$ have the same multiplicities. All the right singular vectors corresponding to $\sigma_{n+1}([\,b \,|\, A\,])$ have zero first components. The original problem $A\,x \approx b$ does not have a TLS solution.

For similar reasons as in the two previous cases, the smallest singular value of $[\,b \,|\, A\,]$ whose right singular vector has nonzero first component must be equal to $\sigma_{\min}([\,b_1 \,|\, A_{11}\,])$. The nongeneric solution of the original problem is unique iff $\sigma_{\min}([\,b_1 \,|\, A_{11}\,])$ is not present in the set of singular values of A_{22}, and, vice versa, the nongeneric solution is nonunique (i.e., σ_t is multiple in the notation used in Section 1.3.3) iff there exist $\sigma_j(A_{22})$ such that $\sigma_j(A_{22}) = \sigma_{\min}([\,b_1 \,|\, A_{11}\,])$. The (minimum 2-norm) nongeneric solution of $A\,x \approx b$ is given by the solution of the core problem transformed back to the original variables, i.e. $x_{\mathrm{NGN}} \equiv Q\,[\,x_1^T \,|\, 0\,]^T$.

Summarizing, for any approximation problem (1.1) the vector $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T$, where x_1 is the unique TLS solution of the core problem within $A\,x \approx b$, represents the corresponding minimum 2-norm solution given in [31, 84]. For the given $A\,x \approx b$ it is reasonable, and Paige and Strakoš in [64] also recommended, first to find a core problem $A_{11}\,x_1 \approx b_1$ using orthogonal transformations (or by Golub-Kahan iterative bidiagonalization), then solve the core problem $A_{11}\,x_1 \approx b_1$, put $x_2 = 0$, and define the solution of the original problem define as $x \equiv Q\,[\,x_1^T \,|\, 0\,]^T$. The assumption $x_2 = 0$ here does not follow from a theory, it is a postulate: do not mix the useful (necessary and sufficient) information with the useless data contained in A_{22} in the solution of $A\,x \approx b$. Consequently the core problem theory is consistent with earlier work and it explains and clarifies the concept of nongeneric solution. The nongeneric concept becomes justified although the minimum 2-norm nongeneric solution does not solve the TLS problem (1.3).

Clearly, from the core problem concept,

$$\sigma_{\min}(A_{22}) \geq \sigma_{\min}([\,b_1 \,|\, A_{11}\,]) \tag{1.21}$$

is the *necessary and sufficient condition for the existence of a TLS solution.* (If the matrix A_{22} is trivial, i.e. it has no columns, then the problem always has the unique TLS solution.)

The following example taken from [64] generalizes observations discussed in Example 1.1 and throughout the preceding text. It shows, for general A, b, that the TLS solution does not exist if and only if the minimal correction $[\,g \,|\, E\,]$ that makes $A\,x \approx b$ compatible does not exist.

Example 1.2. *Consider $A\,x \approx b$ and $A_{11}\,x_1 \approx b_1$ a core problem within $A\,x \approx b$ obtained by a transformation in the form (1.9). Assume that*

$$\sigma_{\min}(A_{22}) < \sigma_{\min}([\,b_1 \,|\, A_{11}\,]),$$

i.e., the original problem does not have a TLS solution, and let u, v be the left and right singular vectors such that $A_{22}\,v = u\,\sigma_{\min}(A_{22})$, $u^T\,A_{22} = \sigma_{\min}(A_{22})\,v^T$.

For any real vector z define $r_1 \equiv b_1 - A_{11}\,z$. Then for any arbitrary small real scalar $\varepsilon > 0$

$$\left[\begin{array}{c|c} A_{11} & r_1\,\varepsilon\,v^T \\ \hline 0 & A_{22} - u\,\sigma_{\min}(A_{22})\,v^T \end{array}\right] \left[\begin{array}{c} z \\ \hline v\,\varepsilon^{-1} \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline 0 \end{array}\right].$$

The square of the Frobenius norm of the corresponding correction to A is equal to $(\,\|r_1\|^2\,\varepsilon^2 + \sigma_{\min}^2(A_{22})\,) \longrightarrow \sigma_{\min}^2(A_{22})$ with $\varepsilon \longrightarrow 0$.

Thus by applying the TLS directly on the original problem we get a nonoptimal TLS distance less than $\sigma_{\min}([\,b_1 \,|\, A_{11}\,])$*, the TLS distance for the core problem* $A_{11}\,x_1 \approx b_1$*. The above nonoptimal solution has nothing to do with the TLS solution vector for* $A_{11}\,x_1 \approx b_1$*, because it is essentially determined by* v *from the non-core part of the problem* $A_{22}\,x_2 \approx 0$ *and by the arbitrary chosen vector* z*. This does not reflect any useful information contained in the data. Moreover in this case the norm of the nonoptimal solution grows to infinity* $\|[\,z^T \,|\, v^T\,\varepsilon^{-1}\,]^T\| \longrightarrow \infty$ *with* $\varepsilon \longrightarrow 0$*, and the optimal solution does not even exist.*

This example shows a general behavior of the nonoptimal solution, which was also illustrated by Example 1.1 for particular A, b. The attempt to reach the minimal correction causes the grow of the norm of the solution, and, what is more important, the dependency of the solution on an arbitrary data.

1.4.3 Alternative core problem definition

In the previous text it was shown that a subproblem representing the core problem, i.e. the minimal subproblem, by Definition 1.1, has several properties, namely (G1)–(G5). (Here, Definition 1.1 is identical to the Paige and Strakoš definition used in [64].) These properties are not independent, the properties (G1) and (G3) imply all the others. Here we ask whether (G1), (G3) can be used for an alternative core problem definition, i.e. whether (G1), (G3) automatically imply the minimality. This section is motivated by the work on the core problem theory for problems with multiple right-hand sides.

First, we show that the core reduction in the form (1.9) applied on a problem $A\,x \approx b$ satisfying (G1) and (G3) yields trivial A_{22} (having no columns as well as no rows). Equivalently, we show that any problem having properties (G1) and (G3) represents the core problem within itself. Then we introduce an alternative definition.

Let $A\,x \approx b$ be an approximation problem having properties (G1) and (G3). Since by (G3) all the left singular vector subspaces $\mathcal{R}(U_j')$ are one dimensional, and $(U_j')^T\,b$ is nonzero, the transformation (1.9) immediately reduces to the form

$$P^T \left[\begin{array}{c|c} b & A\,Q \end{array}\right] = \left[\begin{array}{c|c|c} b_1 & A_{11} & 0 \end{array}\right],$$

i.e. A_{22} has no rows. Now, since $A = P\,[\,A_{11}\,|\,0\,]\,Q^T$ is of full column rank, by (G1), the transformation must reduce further to the form

$$P^T \left[\begin{array}{c|c} b & A\,Q \end{array}\right] = \left[\begin{array}{c|c} b_1 & A_{11} \end{array}\right],$$

i.e. A_{22} has no columns. Thus the core problem transformation described by (1.9) applied to the problem $A\,x \approx b$ having properties (G1) and (G3) yields the subproblem $A_{11}\,x_1 \approx b_1$ with the same dimensions. However, it represents the minimally dimensioned subproblem. Consequently any problem satisfying (G1), (G3) represents a core problem within itself.

Now, it is straightforward to use properties (G1), (G3) for an alternative definition of the core problem.

Definition 1.3 (Core problem; alternative definition)**.** *Any approximation problem* $A\,x \approx b$ *having properties (G1), (G3) is called core problem.*

This alternative core problem definition does not contain the link to the original data, in the other words, that the core problem here is not defined as a subproblem within any other problem, as it is in the original Paige and Strakoš definition, [64], see Definition 1.1.

1.4.4 Lanczos tridiagonalization and core problems

Hnětynková, Plešinger and Strakoš [41, 42, 43] show alternative proofs of the fundamental properties of the core problem – especially the minimality of its dimensions, based on the relationship between the Golub-Kahan bidiagonalization, the Lanczos tridiagonalization and the properties of Jacobi matrices.

Consider the *partial* lower Golub-Kahan bidiagonalization of the matrix A in the form (1.11). Recall that it can be rewritten in the matrix form $A^T S_j = W_j L_j^T$, $A W_j = S_{j+1} L_{j+}$. The Golub-Kahan bidiagonalization (1.11) of A with $s_1 = b/\|b\|$ results in one of the two situations (1.12), (1.13), which will be distinguished throughout this section.

The bidiagonalization algorithm is closely connected with the Lanczos tridiagonalization, see [50]. Let $K \in \mathbb{R}^{\rho\times\rho}$ be a symmetric matrix. Given the initial vector $f_1 \in \mathbb{R}^\rho$ such that $\|f_1\| = 1$, $f_0 \equiv 0$, $\delta_1 \equiv 0$, the partial tridiagonalization algorithm computes for $j = 1, 2, \ldots$

$$\begin{aligned} y_j &\equiv K f_j - \delta_j f_{i-1}\,, \\ \gamma_j &\equiv y_j^T f_j\,, \\ \delta_{j+1} f_{j+1} &\equiv y_j - \gamma_j f_j\,, \end{aligned} \tag{1.22}$$

where $\|f_{j+1}\| = 1$, $\delta_{j+1} \geq 0$, until $\delta_{j+1} = 0$ or until $j = \rho$. Consider $\delta_j > 0$ for $j = 1, \ldots, k$. Denote $F_j \equiv [f_1, \ldots, f_j]$ and

$$H_j \equiv \begin{bmatrix} \gamma_1 & \delta_2 & & \\ \delta_2 & \gamma_2 & \ddots & \\ & \ddots & \ddots & \delta_j \\ & & \delta_j & \gamma_j \end{bmatrix} \in \mathbb{R}^{j\times j}\,,$$

for $j = 1, \ldots, k$. Then F_j has orthonormal columns and H_j represents a (symmetric tridiagonal) Jacobi matrix with positive elements on the first sub- and superdiagonal. The Lanczos algorithm (1.22) can be equivalently written in the matrix form $K F_j = F_j H_j + \delta_{j+1} f_{j+1} e_j^T$, $F_j^T f_{j+1} = 0$. For a given real symmetric K, the algorithm (1.22), called *Lanczos process* is fully determined by the starting vector f_1.

The properties of Jacobi matrices, see Theorem 1.3 and Lemma 1.1, yield the fundamental properties of the matrices H_j. The following lemma presents well known properties of the Lanczos process.

Lemma 1.2. *Let $K \in \mathbb{R}^{\rho\times\rho}$ be a symmetric matrix, $f_1 \in \mathbb{R}^\rho$ and $\|f_1\| = 1$. Assume that algorithm (1.22) does not stop before step k. Then for $j = 1, \ldots, k$:*

(i) *The matrix H_j has distinct eigenvalues.*

(ii) *The first (as well as the last) components of all eigenvectors of H_j are nonzero.*

(iii) *If K is real symmetric positive semidefinite and $f_1 \perp \mathcal{N}(K)$, then all the eigenvalues of H_j are positive (i.e. H_j is symmetric positive definite).*

Proof. The properties (i) and (ii) are the basic properties of Jacobi matrices, see Lemma 1.1, [66, Lemma 7.7.1, Theorem 7.9.3]. Further, (iii) follows recursively from the interlacing property (1.15), using the fact that the final Jacobi matrix H_l, for which $K F_l = F_l H_l$ (i.e. l is the index for which $\delta_l \neq 0$ and $\delta_{l+1} = 0$) must be under the assumption in (iii) nonsingular and thus symmetric positive definite, see [66, Theorem 10.1.1]. □

The relationship between the Lanczos tridiagonalization and the Golub-Kahan bidiagonalization can be described in several ways, see [5, pp. 662–663], [6, pp. 513–515], [27, pp. 212–214] and also [57, pp. 199–200], [59, pp. 44–48], [51, pp. 115–118]. Consider the coefficients of the Golub-Kahan bidiagonalization $\alpha_j > 0$, $\beta_j > 0$ for $j = 1, \ldots, k$. Then $A^T S_j = W_j L_j^T$ multiplied by A from the left, together with the substitution $A W_j = S_{j+1} L_{j+}$, gives,

$$A A^T S_j = S_j L_j L_j^T + \alpha_j \beta_{j+1} s_{j+1} e_j^T , \tag{1.23}$$

where $L_j L_j^T$ is the Jacobi matrix having the form (1.17), with j instead of $\tilde{n}$. The identity (1.23) represents j steps of the Lanczos tridiagonalization of the matrix $A A^T$ with starting vector $s_1 = b/\beta_1 = b/\|b\|$. Here, according to the notation used in (1.22), we have $K^{(1)} \equiv A A^T \in \mathbb{R}^{m\times m}$, $F_j^{(1)} \equiv S_j$, $H_j^{(1)} \equiv L_j L_j^T$ and $\delta_j^{(1)} \equiv \alpha_{j-1}\beta_j > 0$, for $j = 1, \ldots, k$.

Similarly $A W_j = S_{j+1} L_{j+}$ multiplied by A^T from the left, together with the substitution $A^T S_{j+1} = W_{j+1} L_{j+1}^T$, gives

$$A^T A W_j = W_j L_{j+}^T L_{j+} + \alpha_{j+1} \beta_{j+1} w_{j+1} e_j^T , \tag{1.24}$$

where $L_{j+}^T L_{j+}$ is the Jacobi matrix having the form (1.16), with j instead of $\tilde{n}$. The identity (1.24) represents j steps of the Lanczos tridiagonalization of the matrix $A^T A$ with starting vector $w_1 = A^T s_1/\alpha_1 = A^T b/\|A^T b\|$. Here we have $K^{(2)} \equiv A^T A \in \mathbb{R}^{n\times n}$, $F_j^{(2)} \equiv W_j$, $H_j^{(2)} \equiv L_{j+}^T L_{j+}$ and $\delta_j^{(2)} \equiv \alpha_j \beta_j > 0$, for $j = 1, \ldots, k$.

Remark 1.5. *The relationship between the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization algorithms can also be described using the following relation. The Lanczos tridiagonalization applied to the augmented matrix*

$$K^{(3)} \equiv \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$$

with the starting vector $f_1^{(3)} \equiv [\, s_1^T , 0\,]^T$ yields the Jacobi matrix with zero main diagonal and the sub- and superdiagonals equal to $(\, \alpha_1 , \beta_2 , \alpha_2 , \beta_2 , \alpha_3 , \ldots)$. The orthonormal vectors $f_{2j-1}^{(3)} \equiv [\, s_j^T , 0\,]^T$ and $f_{2j}^{(3)} \equiv [\, 0 , w_j^T\,]^T$ are generated in odd and even steps of algorithm, respectively.

Now the fundamental properties of a core problem can be related to the properties of the Lanczos tridiagonalization process and the Jacobi matrices given in Lemma 1.2. We include the proof in the same form as it is in [41, 43]. The compatible and the incompatible case are distinguished.

Case 1. In the compatible case, $\alpha_j > 0$, $\beta_j > 0$, for $j = 1, \ldots, \bar{n}$, $\beta_{\bar{n}+1} = 0$ or $\bar{n} = m$ (i.e. $m \le n$), see (1.12). The square matrix $\tilde{A}_{11} \equiv L_{\bar{n}}$ represents a Cholesky factor of the Jacobi matrix $H_{\bar{n}}^{(1)} \equiv L_{\bar{n}} L_{\bar{n}}^T \in \mathbb{R}^{\bar{n}\times\bar{n}}$, which results from the Lanczos tridiagonalization of $K^{(1)} \equiv A A^T$ with the starting vector $s_1 = b/\|b\|$ (1.23), and which stops exactly in $\bar{n}$ steps, i.e.,

$$(A A^T) S_{\bar{n}} = S_{\bar{n}} (L_{\bar{n}} L_{\bar{n}}^T) . \tag{1.25}$$

Consider the SVD of $L_{\bar{n}} = R \Sigma T^T$, where $\Sigma = \mathrm{diag}\,(\, \sigma_1 , \ldots , \sigma_{\bar{n}}\,)$, and $R \in \mathbb{R}^{\bar{n}\times\bar{n}}$, $T \in \mathbb{R}^{\bar{n}\times\bar{n}}$ are orthogonal matrices. Then $H_{\bar{n}}^{(1)} = L_{\bar{n}} L_{\bar{n}}^T = R \Sigma^2 R^T$ is the spectral decomposition of the matrix $H_{\bar{n}}^{(1)}$, σ_j^2 are its eigenvalues and $r_j = R e_j$ its eigenvectors, $j = 1, \ldots, \bar{n}$.

From (i) in Lemma 1.2 the eigenvalues σ_j^2 are distinct and thus the singular values of $\tilde{A}_{11} \equiv L_{\bar{n}}$ are distinct. The matrix $L_{\bar{n}}$ is square with positive elements on its diagonal. Therefore A_{11} is of full column rank (G1) and all its distinct singular values must be positive (G5). Property (G3) follows immediately from (ii) in Lemma 1.2, since $\tilde{b}_1^T r_j = \beta_1 e_1^T r_j \neq 0$ for $j = 1, \ldots, \bar{n}$.

Definition 1.1 formulates the core problem as the minimally dimensioned subproblem. As shown in Section 1.4.3 (G1) and (G3) ensure the minimality property. However, the minimality can be shown directly using the properties of the Lanczos algorithm and Jacobi matrices. Assume by contradiction that there exists $\hat{P}$, $\hat{P}^{-1} = \hat{P}^T$, and $\hat{Q}$, $\hat{Q}^{-1} = \hat{Q}^T$, such that the transformation (1.9) gives $\hat{A}_{11} \in \mathbb{R}^{q \times q}$ with $q < \bar{n}$. (Because the system is compatible by considering, for example, the QR decomposition of $\hat{A}_{11}$, we can with no loss of generality assume that $\hat{A}_{11}$ is square.) Substituting

$$A = \hat{P} \left[\begin{array}{c|c} \hat{A}_{11} & 0 \\ \hline 0 & \hat{A}_{22} \end{array}\right] \hat{Q}^T$$

into the Lanczos tridiagonalization (1.25) gives

$$\hat{P} \left[\begin{array}{c|c} \hat{A}_{11} & 0 \\ \hline 0 & \hat{A}_{22} \end{array}\right] \left[\begin{array}{c|c} \hat{A}_{11} & 0 \\ \hline 0 & \hat{A}_{22} \end{array}\right]^T \hat{P}^T S_{\bar{n}} = S_{\bar{n}} H_{\bar{n}}^{(1)},$$

i.e.

$$\left[\begin{array}{c|c} \hat{A}_{11}\,\hat{A}_{11}^T & 0 \\ \hline 0 & \hat{A}_{22}\,\hat{A}_{22}^T \end{array}\right] (\hat{P}^T S_{\bar{n}}) = (\hat{P}^T S_{\bar{n}})\, H_{\bar{n}}^{(1)}, \tag{1.26}$$

with $\hat{P}^T s_1 = \hat{P}^T b / \|b\| = [\,\tilde{b}_1^T \,|\, 0\,]^T / \|b\|$. Since $\hat{A}_{11}\,\hat{A}_{11}^T \in \mathbb{R}^{q \times q}$ and $\hat{b}_1 \in \mathbb{R}^q$, the Lanczos tridiagonalization represented by (1.26) must stop in at most q steps, and $H_{\bar{n}}^{(1)}$ must have $\delta_{q+1}^{(1)} = 0$, which contradicts the fact that $H_{\bar{n}}^{(1)}$ is a Jacobi matrix.

Case 2. In the incompatible case, $\alpha_j > 0$, $\beta_j > 0$, for $j = 1, \ldots, \bar{n}$, and $\beta_{\bar{n}+1} > 0$, $\alpha_{\bar{n}+1} = 0$ or $\bar{n} = n$ (i.e. $m \geq n+1$), see (1.13). The rectangular matrix $\tilde{A}_{11} \equiv L_{\bar{n}+}$ can be linked to the matrix $H_{\bar{n}}^{(2)} \equiv L_{\bar{n}+}^T L_{\bar{n}+} \in \mathbb{R}^{\bar{n} \times \bar{n}}$ (note that here $\tilde{A}_{11}$ does not represent the Cholesky factor). The matrix $H_{\bar{n}}^{(2)}$ results from the Lanczos tridiagonalization of $K^{(2)} \equiv A^T A$ with the starting vector $w_1 = A^T b / \|A^T b\|$ (see (1.24)), and which stops exactly in $\bar{n}$ steps, i.e.,

$$(A^T A)\, W_{\bar{n}} = W_{\bar{n}}\, (L_{\bar{n}+}^T L_{\bar{n}+}). \tag{1.27}$$

Consider the SVD of $L_{\bar{n}+} = R\,\Sigma\,T^T$, where $\Sigma = \mathrm{diag}\,(\,\sigma_1\,, \ldots, \sigma_{\bar{n}}\,)$, $R \in \mathbb{R}^{(\bar{n}+1) \times \bar{n}}$ is now a rectangular matrix with an orthonormal columns, $T \in \mathbb{R}^{\bar{n} \times \bar{n}}$ is orthogonal matrix. Then $H_{\bar{n}}^{(2)} = L_{\bar{n}+}^T L_{\bar{n}+} = T\,\Sigma^2\,T^T$ is the spectral decomposition of the matrix $H_{\bar{n}}^{(2)}$, σ_j^2 are its eigenvalues and $t_j = T\,e_j$ its eigenvectors, $j = 1, \ldots, \bar{n}$.

Similarly to the previous case, from (i) in Lemma 1.2 it follows that the singular values of $L_{\bar{n}+}$ are distinct. Since by construction v_1 does not have any nonzero component in the null space of $A^T A$, the property (iii) in Lemma 1.2 yields that all these distinct singular values of $L_{\bar{n}+}$ are positive and therefore we obtain properties (G1) and (G5). Moreover, $e_1^T t_j \neq 0$, $j = 1, \ldots, \bar{n}$, by the property (ii) in Lemma 1.2. Considering $L_{\bar{n}+}\,T = R\,\Sigma$ and the fact that $L_{\bar{n}+}$ is lower bidiagonal with nonzero bidiagonal elements, it follows that $e_1^T r_j \neq 0$, $j = 1, \ldots, \bar{n}$. Consequently $\tilde{b}_1^T r_j = \beta_1\, e_1^T r_j \neq 0$, $j = 1, \ldots, \bar{n}$, which gives the property (G3).

As mentioned, the minimality property used in Definition 1.1 is implied by (G1) and (G3), see also Section 1.4.3. However, it can be shown directly using the properties of Lanczos process and Jacobi matrices by contradiction, analogously

to the compatible case. Assume by contradiction that there exist $\hat{P}$, $\hat{P}^{-1} = \hat{P}^T$, and $\hat{Q}$, $\hat{Q}^{-1} = \hat{Q}^T$, such that the transformation (1.9) gives $\hat{A}_{11} \in \mathbb{R}^{(q+1)\times q}$ with $q < \bar{n}$. (Here the system is incompatible and therefore we can with no loss of generality assume that $\hat{A}_{11}$ is rectangular of the given dimensions.) Substituting $A = \hat{P}\,\mathrm{diag}\,(\,\hat{A}_{11}\,,\,\hat{A}_{22}\,)\,\hat{Q}^T$ into the Lanczos tridiagonalization (1.27) gives

$$\left[\begin{array}{c|c} \hat{A}_{11}^T\hat{A}_{11} & 0 \\ \hline 0 & \hat{A}_{22}^T\hat{A}_{22} \end{array}\right] (\hat{Q}^T\,W_{\bar{n}}) \;=\; (\hat{Q}^T\,W_{\bar{n}})\,H_{\bar{n}}^{(2)}\,, \tag{1.28}$$

with $\hat{Q}^T\,w_1 \;=\; \hat{Q}^T\,(\,A^T\,b\,)/\|A^T\,b\| \;=\; [\,(\,\hat{A}_{11}^T\,\hat{b}_1\,)^T\,|\,0\,]^T/\|A^Tb\|$, which leads to a contradiction exactly in the same way as in Case 1.

Summarizing, it was shown that the fundamental properties of the core problem can be proved in an elegant way without using the SVD of the whole matrix $[\,b\,|\,A\,]$. Here the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization are used as very strong mathematical tools for constructing proofs.

1.5 Goals of the thesis

The thesis focuses on solution of an orthogonally invariant linear approximation problem with multiple right-hand sides $AX \approx B$ through the TLS concept. The main goal of the thesis is to generalize the analysis of the TLS concept given for problems with single right-hand sides in [31, 84, 64] and to build up a consistent theory which would cover the multiple right-hand sides case.

For a problem with multiple right-hand sides, a partial generalization of the TLS concept was presented by S. Van Huffel and J. Vandewalle in [84]. They cover some particular cases for which they define a TLS solution. They also present an algorithm which for any data gives an output, which is, however, not identified with a theoretically justified TLS theory. Therefore we attempt in Chapter 3, as the first goal of the presented thesis, to revise and complete, within our abilities, their analysis.

C. C. Paige and Z. Strakoš proved in [64] that for a problem with a single right-hand side $Ax \approx b$ there is a reduction which determines a core problem $A_{11}\,x_1 \approx b_1$ within the original problem, with all necessary and sufficient information for solving the original problem. The core problem always has the unique TLS solution, and, using the transformation to the original variables, it gives the solution of the original approximation problem identical to the minimum 2-norm solutions of all TLS formulations given in [31, 84]. The core problem theory represents a new approach to understanding of the TLS concept. It makes the theory complete and transparent, and it also fundamentally changes a view to practical computations. The second goal of the presented thesis is therefore to extend the core problem theory, if possible, to problems with multiple right-hand sides. The reduction based on the SVD of A, motivated by the work of D. M. Sima and S. Van Huffel [73, 74], is given in Chapter 4. Another approach, based on a banded generalization of the Golub-Kahan bidiagonalization algorithm, is given in Chapter 5, motivated by the series of lectures [8, 9, 10] by Å. Björck, and also the work [64, 41, 42, 43] of C. C. Paige, Z. Strakoš, I. Hnětynková and partially of the author of this thesis.

Chapter 6 investigates the relationship between the SVD-based and the banded reduction approaches. An extension of the minimally dimensioned subproblem concept to the multiple right-hand side case has some difficulties. In particular, the minimally dimensioned reduced subproblem may not have a TLS solution.

Core problem computation in finite precision arithmetic must resolve a problem of relevant stopping criteria. Difficulties connected with revealing of core problem are illustrated on examples in Chapter 7. We do not address this question fully in the thesis, but present an example of the noise-revealing property of the Golub-Kahan bidiagonalization, which can be very useful in hybrid methods for solving ill-posed problems, see Chapter 8.

The thesis ends with conclusions, some open questions and directions for further research, in Chapter 9.

Part II

Theoretical Fundamentals of Total Least Squares Formulation in $AX \approx B$

Chapter 2

Problem formulation

This chapter introduces the total least squares formulation for the problems with the multiple right-hand sides, the basic notation used in the thesis, and, finally, the selected useful theorems repeatedly used throughout the text.

2.1 Introduction

Consider an orthogonally invariant linear approximation problem

$$A\,X \;\approx\; B\,, \qquad A \in \mathbb{R}^{m\times n}\,, \quad X \in \mathbb{R}^{n\times d}\,, \quad B \in \mathbb{R}^{m\times d}\,, \tag{2.1}$$

or, equivalently,

$$\left[\;\begin{array}{c|c} B & A \end{array}\;\right] \left[\begin{array}{c} -\,I_d \\ \hline X \end{array}\right] \;\approx\; 0\,. \tag{2.2}$$

We assume $A^T B \neq 0$, otherwise the columns of the *right-hand side* (*observation matrix*) B are not correlated with the columns of the *system matrix* A and it does not make sense to look for an approximation of B by the columns of A.

Definition 2.1 (Total least squares problem)**.** *The linear approximation problem (2.1) specified by*

$$\min_{X,E,G} \left\| \left[\;\begin{array}{c|c} G & E \end{array}\;\right] \right\|_F \qquad \textit{subject to} \qquad (A+E)\,X \;=\; B+G \tag{2.3}$$

is called the total least squares (TLS) problem with the TLS solution $X_{\mathrm{TLS}} \equiv X$ and the correction matrix $[\,G\,|\,E\,]$, $G \in \mathbb{R}^{m\times d}$, $E \in \mathbb{R}^{m\times n}$.

The TLS problem has been investigated for decades, see [29], [24, Section 6], [81, 31], [26, pp. 324–326], [84]. Even with $d = 1$ it may not have a solution, and when the solution exists, it may not be unique, see [84] for the classical description and [64] for the recent refinements. In our text we investigate existence and uniqueness of the TLS solution for the formulation (2.1)–(2.3) with $d \geq 1$.

Remark 2.1. *A more general form of (2.3) including weighting and scaling was considered in the literature, see, e.g., the first fully analytic paper on the subject [31]. It is worth to note that other norms than the Frobenius norm are also relevant in practice, see, e.g., [87]. In our text we restrict ourselves to the problem (2.1)–(2.3).*

Remark 2.2. *Equivalent approximation problems have been used in many applications. For a survey and relationship between the algebraic formulation and analysis and descriptions used in application areas we refer to [84, Chapters 1, 8 and 9].*

2.2 Basic notation and selected theorems

In order to simplify the notation we assume $m \geq n+d$ (add zero rows if necessary). Consider a singular value decomposition (SVD) of A, $r \equiv \operatorname{rank}(A)$,

$$A = U' \Sigma' (V')^T , \tag{2.4}$$

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

$$\sigma'_1 \geq \ldots \geq \sigma'_r > \sigma'_{r+1} = \ldots = \sigma'_n \equiv 0 . \tag{2.5}$$

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

$$\left[\begin{array}{c|c} B & A \end{array}\right] = U \Sigma V^T , \tag{2.6}$$

where $U^{-1} = U^T$, $V^{-1} = V^T$, $\Sigma = \operatorname{diag}(\sigma_1 , \ldots , \sigma_s , 0)$, and

$$\sigma_1 \geq \ldots \geq \sigma_s > \sigma_{s+1} = \ldots = \sigma_{n+d} \equiv 0 . \tag{2.7}$$

In the further text $\sigma_j(M)$ denotes the jth largest singular value, $\mathcal{R}(M)$ and $\mathcal{N}(M)$ the range and the null space, $\|M\|_F$ and $\|M\|$ the Frobenius norm and the 2-norm of the given matrix M, $\|v\|$ the 2-norm of the given vector v; $\operatorname{sp}(K)$ denotes the spectrum and $\lambda_j(K)$ the jth largest eigenvalue of the given symmetric positive semidefinite matrix K; $I_k \in \mathbb{R}^{k \times k}$ denotes the k by k identity matrix.

We will repeatedly use the following theorem. The statement follows from the classical form of the interlacing theorem, see [79], [66, p. 203 (in the original Prentice-Hall edition 1980, p. 186)], [90, Chapter 2, §47, pp. 103–4], [32, Theorem 8.1.7, p. 396, and Corollary 8.6.3, p. 449], [84, Theorem 2.4, p. 32].

Theorem 2.1 (Interlacing theorem for singular values)**.** *Consider $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m \times d}$, $m \geq n+d$. Let (2.4) be the SVD of A and (2.6) the SVD of $[\, B \,|\, A \,]$. Then*

$$\sigma_j \geq \sigma'_j \geq \sigma_{j+d} , \qquad j = 1 , \ldots , n . \tag{2.8}$$

For the proof see [66, p. 203]. Please note that the proof in the original edition from 1980, p. 186, is different. We include it here for its relevance to techniques used throughout the thesis.

Proof. Consider the symmetric matrices

$$M \equiv \left[\begin{array}{c|c} B & A \end{array}\right]^T \left[\begin{array}{c|c} B & A \end{array}\right] \in \mathbb{R}^{(n+d) \times (n+d)} , \qquad H \equiv A^T A \in \mathbb{R}^{n \times n} ,$$

then $\operatorname{sp}(H) = \{ \sigma'^2_j : j = 1 , \ldots , n \}$ represents the spectrum of the matrix H and $\operatorname{sp}(M) = \{ \sigma^2_j : j = 1 , \ldots , n+d \}$ represents the spectrum of the matrix M, ordered according to (2.5) and (2.7), respectively. For $\xi \in \mathbb{R}$, $\xi \notin \operatorname{sp}(H)$ define the matrices

$$\begin{aligned} K(\xi) &\equiv (H - I_n \xi)^{-1} A^T B \in \mathbb{R}^{n \times d} , \\ W(\xi) &\equiv (B^T B - I_d \xi) - B^T A (H - I_n \xi)^{-1} A^T B \in \mathbb{R}^{d \times d} . \end{aligned}$$

Then

$$(M - I_{n+d} \xi) = \left[\begin{array}{c|c} I_d & K(\xi)^T \\ \hline 0 & I_n \end{array}\right] \left[\begin{array}{c|c} W(\xi) & 0 \\ \hline 0 & H - I_n \xi \end{array}\right] \left[\begin{array}{c|c} I_d & 0 \\ \hline K(\xi) & I_n \end{array}\right]$$

is a congruence transformation. Consequently the matrices $(M - I_{n+d}\,\xi)$ and $\operatorname{diag}(W(\xi), H - I_n\,\xi)$ have the same inertia (π, ν, ζ), i.e., they have the same number of positive $\pi(\cdot)$, negative $\nu(\cdot)$ and zero $\zeta(\cdot)$ eigenvalues, see, e.g., [44, Theorem 4.5.8, p. 223]. Obviously

$$\begin{aligned}\pi(H - I_n\,\xi) &\le \pi(M - I_{n+d}\,\xi)\\ &= \pi(H - I_n\,\xi) + \pi(W(\xi)) \le \pi(H - I_n\,\xi) + d\,,\end{aligned} \tag{2.9}$$

$$\begin{aligned}\nu(H - I_n\,\xi) &\le \nu(M - I_{n+d}\,\xi)\\ &= \nu(H - I_n\,\xi) + \nu(W(\xi)) \le \nu(H - I_n\,\xi) + d\,.\end{aligned} \tag{2.10}$$

Assume by contradiction that there exists an index j such that $\sigma_j < \sigma'_j$. Choose $\xi_1 \notin \operatorname{sp}(H)$ such that $\sigma_j^2 < \xi_1 < \sigma'^2_j$. Then $\pi(H - I_n\,\xi_1) \ge j$, and $\pi(M - I_{n+d}\,\xi_1) < j$, which contradicts (2.9),

$$j \le \pi(H - I_n\,\xi_1) \le \pi(M - I_{n+d}\,\xi_1) < j\,,$$

and thus $\sigma_j \ge \sigma'_j$.

Similarly, assume by contradiction that there exists an index j such that $\sigma'_j < \sigma_{j+d}$. Choose $\xi_2 \notin \operatorname{sp}(H)$ such that $\sigma'^2_j < \xi_2 < \sigma^2_{j+d}$. Then $\nu(H - I_n\,\xi_2) \ge n - j + 1$, and $\nu(M - I_{n+d}\,\xi_2) < (n+d) - (j+d) + 1$, which contradicts (2.10),

$$n - j + 1 \le \nu(H - I_n\,\xi_2) \le \nu(M - I_{n+d}\,\xi_2) < n - j + 1\,,$$

and thus $\sigma'_j \ge \sigma_{j+d}$. □

Corollary 2.1. *Let σ'_j be a singular value of A with multiplicity r'_j. Then the extended matrix $[B\,|\,A]$ has a singular value σ_i, $\sigma_i \equiv \sigma'_j$, with multiplicity r_i, where*

$$\max\{0, r'_j - d\} \le r_i \le r'_j + d\,. \tag{2.11}$$

Proof. Consider σ'_j with the multiplicity $r'_j > d$. The lower bound follows from

$$\sigma'_j \ge \sigma_{j+d} \ge \ldots \ge \sigma_{j+d+(r'_j-d-1)} \equiv \sigma_{j+r'_j-1} \ge \sigma'_{j+r'_j-1} = \sigma'_j\,.$$

If $r'_j \le d$, the argument is analogous.

On the other hand, if $\sigma'_1 = \ldots = \sigma'_n$, then the upper bound is trivial. If $\sigma'_1 = \ldots = \sigma'_{r'_j} > \sigma_{r'_j+1}$ (with $r'_j < n$), then $\sigma'_{r'_j+1} \ge \sigma_{r'_j+1+d}$ gives the result. Analogously, if $\sigma'_{j-1} > \sigma'_j = \ldots = \sigma'_n$ (with $j > 1$), then $\sigma_{j-1} \ge \sigma'_{j-1}$ gives the result. In all other cases

$$\sigma_{j-1} \ge \sigma'_{j-1} > \sigma'_j = \ldots = \sigma'_{j+r'_j-1} > \sigma'_{j+r'_j} \ge \sigma_{j+r'_j+d}\,,$$

which gives the maximal possible multiplicity. □

Remark 2.3. *For $d = 1$, (2.8) reduces to*

$$\sigma_1 \ge \sigma'_1 \ge \sigma_2 \ge \sigma'_2 \ge \ldots \ge \sigma'_n \ge \sigma_{n+1} \ge 0\,. \tag{2.12}$$

As a consequence of this remark, appending one vector (column of B) to A causes interlacing of the singular values. Subsequently, the multiplicity of each singular value of A can increase by one, stagnate, or decrease by one. The assertion of Corollary 2.1 can then be obtained by induction.

Another theorem frequently used throughout the text describes a low-rank matrix approximations [84, Theorem 2.3, p. 31].

Theorem 2.2 (Eckart-Young-Mirsky matrix approximation theorem)**.** *Consider $C \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(C) = r$, $m \geq n$. Let $C = \sum_{j=1}^{r} u_j \sigma_j v_j^T$ be the SVD of C, and $\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n \equiv 0$ the singular values of C. Then for any $k \leq r$ the matrix $C_k \equiv \sum_{j=1}^{k} u_j \sigma_j v_j^T$ represents the rank k approximation to C with the following minimization property*

$$\min_{\operatorname{rank}(D)\leq k} \| C - D \| = \| C - C_k \| = \sigma_{k+1}\,, \tag{2.13}$$

and

$$\min_{\operatorname{rank}(D)\leq k} \| C - D \|_F = \| C - C_k \|_F = \left(\sum_{j=k+1}^{n} \sigma_j^2 \right)^{1/2}. \tag{2.14}$$

For the proof see [16, 55], for the 2-norm see also [32, Theorem 2.5.3, pp. 72–73]. Two alternative proofs for the Frobenius norm are in [4, Theorem 3, pp. 213–214, and Exercise 29, p. 216].

The history of Theorem 2.2 is described in [76, p. 210]. The *Eckart-Young-Mirsky* (or, alternatively, *Schmidt-Mirsky*) *theorem* is commonly attributed to Eckart and Young [16] (1936), who established it for the Frobenius norm. But Schmidt [71] (1907) proved it for integral operators and the Hilbert-Schmidt norm, the natural extension of the Frobenius norm. Mirsky [55] (1964) generalized it to unitarily invariant norms. For another generalization by Golub, Hoffman, and Stewart see [26] (1987), or [76, Theorem 4.18, p. 208].

It is worth to note some properties of the Moore-Penrose pseudoinverse which is used throughout the thesis.

Lemma 2.1. *Let K, L, M be arbitrary matrices, Z be an orthogonal matrix such that M and Z can be multiplied. Then*

$$(\operatorname{diag}(K, L))^\dagger = \operatorname{diag}(K^\dagger, L^\dagger)\,, \tag{2.15}$$

$$(M Z)^\dagger = Z^\dagger M^\dagger = Z^{-1} M^\dagger = Z^T M^\dagger\,, \tag{2.16}$$

$$\left[\begin{array}{c|c} M & 0 \end{array}\right]^\dagger = \left[\begin{array}{c} M^\dagger \\ \hline 0 \end{array}\right], \tag{2.17}$$

where the zero blocks in the third equality have transposed dimensions.

The proof is based on the SVD form of the Moore-Penrose pseudoinverse, or by using the Moore-Penrose conditions, see, e.g., [32, p. 257].

Chapter 3

Classification and the relationship to the work of Van Huffel and Vandewalle

This chapter recapitulates the classical analysis and results given by S. Van Huffel and J. Vandewalle in [84, Chapter 3] for the problems with multiple right-hand sides. Some extensions of the analysis and the completion of the classification of the total least squares problems are presented here.

The goal of this chapter is to formulate conditions for existence of a TLS solution for problems with multiple right-hand sides.

3.1 Introduction

In this chapter we concentrate on the *incompatible problem* (2.1), $\mathcal{R}(B) \not\subset \mathcal{R}(A)$. The compatible case is simpler because it reduces to finding a solution of a system of linear algebraic equations. With no loss of generality it is assumed $m \geq n + d$ (add zero rows if necessary; note that this assumption and the addition of zero rows is justified through the data reduction concept, see Chapter 4).

In order to handle a possible multiplicity of σ_{n+1}, we introduce the following notation

$$\sigma_{n-q} > \underbrace{\sigma_{n-q+1} = \ldots = \sigma_n}_{q} = \underbrace{\sigma_{n+1} = \ldots = \sigma_{n+e}}_{e} > \sigma_{n+e+1}\,, \tag{3.1}$$

where q singular values to the left and $e-1$ singular values to the right are equal to σ_{n+1}, and $q \geq 0$, $e \geq 1$. For convenience we denote $n - q \equiv p$. If $q = n$, then σ_p is nonexistent. Similarly, if $e = d$, then σ_{n+e+1} is nonexistent.

It will be useful to consider the following partitioning

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1^{(\Delta)} & \Sigma_2^{(\Delta)} \end{array}\right], \tag{3.2}$$

where $\Sigma_1^{(\Delta)} \in \mathbb{R}^{m\times(n-\Delta)}$, $\Sigma_2^{(\Delta)} \in \mathbb{R}^{m\times(d+\Delta)}$, and consistently with (3.2),

$$V = \left[\begin{array}{c|c} V_{11}^{(\Delta)} & V_{12}^{(\Delta)} \\ \hline V_{21}^{(\Delta)} & V_{22}^{(\Delta)} \end{array}\right], \tag{3.3}$$

where $V_{11}^{(\Delta)} \in \mathbb{R}^{d\times(n-\Delta)}$, $V_{12}^{(\Delta)} \in \mathbb{R}^{d\times(d+\Delta)}$, $V_{21}^{(\Delta)} \in \mathbb{R}^{n\times(n-\Delta)}$, $V_{22}^{(\Delta)} \in \mathbb{R}^{n\times(d+\Delta)}$, see Figure 3.1.

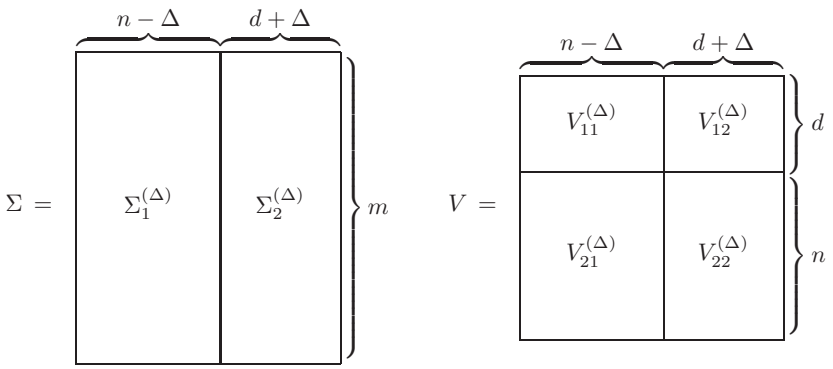


Figure 3.1: Dimensions of the individual matrix blocks in the partitioning (3.2), (3.3).

Depending on the data and convenience, the value of Δ, as described later, can be positive, zero, or negative. When $\Delta = 0$, the partitioning conforms in a straight way to the fact that $[\,B\,|\,A\,]$ is created by A appended by the matrix B with d columns. Then the upper index is omitted, $\Sigma_1 \equiv \Sigma_1^{(0)}$, etc.

The next Lemma follows from the general version of the CS decomposition of orthonormal matrices.

Lemma 3.1. *Let $V \in \mathbb{R}^{(n+d)\times(n+d)}$ be an orthogonal matrix with the partitioning given by (3.3). Then the following two assertions are equivalent:*

(i) $V_{12}^{(\Delta)}$ *is of full row (column) rank,*

(ii) $V_{21}^{(\Delta)}$ *is of full column (row) rank,*

respectively. Similarly for the matrices $V_{11}^{(\Delta)}$ and $V_{22}^{(\Delta)}$.

Proof. The CS decomposition [58, p. 402], see also [32, Theorem 2.6.3, p. 78], gives

$$\begin{aligned} &\operatorname{diag}(W_1, W_2)^T \left[\begin{array}{c|c} V_{11}^{(\Delta)} & V_{12}^{(\Delta)} \\ \hline V_{21}^{(\Delta)} & V_{22}^{(\Delta)} \end{array}\right] \operatorname{diag}(T_1, T_2) \\ &\qquad = \left[\begin{array}{ccc|ccc} I & 0 & 0 & 0 & 0 & 0 \\ 0 & C & 0 & 0 & S & 0 \\ 0 & 0 & 0 & 0 & 0 & I \\ \hline 0 & 0 & 0 & I & 0 & 0 \\ 0 & -S & 0 & 0 & C & 0 \\ 0 & 0 & I & 0 & 0 & 0 \end{array}\right], \end{aligned} \tag{3.4}$$

where $W_1 \in \mathbb{R}^{d\times d}$, $W_2 \in \mathbb{R}^{n\times n}$, $T_1 \in \mathbb{R}^{(n-\Delta)\times(n-\Delta)}$, $T_2 \in \mathbb{R}^{(d+\Delta)\times(d+\Delta)}$ are orthogonal matrices, I stands for the identity matrices with appropriate dimensions, C and S are square diagonal matrices with positive entries on the main diagonals. We do not need to specify the dimensions of the individual blocks; some of the zero blocks may be empty.

Clearly, the matrix $V_{12}^{(\Delta)}$ is of full row rank iff the first block row in (3.4) vanishes, i.e. the first block column in (3.4) is nonexistent and the matrix $V_{21}^{(\Delta)}$ is of full column rank. The rest of proof is fully analogous. □

The classical analysis of the total least square problem with a single right-hand side ($d = 1$) in [31, Theorem 4.1], and the whole theory developed in [84] were based on the relationship between the singular values of A and $[\,B \,|\, A\,]$. For $d = 1$, in particular, $\sigma'_n > \sigma_{n+1}$ represents a *sufficient (not necessary) condition* for the existence of the TLS solution. This view has changed radically by the works [61] and [64], which eliminate the difficulties caused by the fact that $\sigma'_n > \sigma_{n+1}$ is not a necessary condition by reduction of the data into the form of the so called core problem. In this chapter we stay with the classical theory based on a generalization of [31, Theorem 4.1] for $d > 1$. Here the following theorem is instrumental.

Theorem 3.1. *Let (2.4) be the SVD of A and (2.6) the SVD of $[\,B \,|\, A\,]$ with the partitioning given by (3.2)–(3.3), $m \geq n + d$, $\Delta \geq 0$. If*

$$\sigma'_{n-\Delta} > \sigma_{n-\Delta+1}\,, \tag{3.5}$$

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 Theorem 2.1. For the proof of the second part see [88, Lemma 2.1] or [84, Lemma 3.1, pp. 64–65]. Please note the different ordering of the partitioning of V in [88, 84].

Please note that here we do not specify the relationship between $\sigma_{n-\Delta+1}$ and σ_{n+1}, i.e. the partitioning (3.2)–(3.3) can be independent on (3.1).

3.2 Problems of the 1st class

Since the classification of situations which can occur when $d > 1$ is complicated, we can not follow the basic, generic and nongeneric terminology used in [84].

Definition 3.1 (Problem of the 1st class)**.** *Consider a TLS problem (2.3), $m \geq n + d$. Let (2.6) be the SVD of $[\,B \,|\, A\,]$ with the partitioning given by (3.2)–(3.3), $\Delta \equiv q$, where q is the integer related to the multiplicity of σ_{n+1}, given by (3.1). Let $V_{12}^{(q)}$ be of rank d. Then we call (2.3) a TLS problem of the 1st class.*

For $d = 1$ the TLS problem of the first class reduces to the case when the right singular vector subspace corresponding to the smallest singular value contains a singular vector with a nonzero first component. Consequently, the TLS problem has a (possibly nonunique) solution. As we will see, for $d > 1$ such a property is not preserved.

3.2.1 Problems of the 1st class with unique TLS solution

Consider a TLS problem of the 1st class. Let in (3.1) $\sigma_n > \sigma_{n+1}$, i.e. $q = 0$ ($p = n$). We set $\Delta \equiv q = 0$ in (3.2)–(3.3). Then $V_{12}^{(q)} \equiv V_{12}$ is a square (and nonsingular) matrix. Define the *correction matrix*

$$\left[\begin{array}{c|c} G & E \end{array}\right] \equiv -U \left[\begin{array}{c|c} 0 & \Sigma_2 \end{array}\right] V^T = -U\,\Sigma_2 \left[\begin{array}{c|c} V_{12}^T & V_{22}^T \end{array}\right]. \tag{3.6}$$

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 Theorem 2.2, 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,

$$\left[\;\; B+G \;\middle|\; A+E \;\;\right] \left[\begin{array}{c} -\,I_d \\ \hline -\,V_{22}\,V_{12}^{-1} \end{array}\right] \;=\; 0\,,$$

which gives the uniquely determined TLS solution

$$X_{\mathrm{TLS}} \;\equiv\; X^{(0)} \;\equiv\; -\,V_{22}\,V_{12}^{-1}\,. \tag{3.7}$$

For the Frobenius norm and the 2-norm of the solution (3.7) see (3.28) in Lemma 3.2, see also [84, Theorem 3.6, pp. 55–56]. We summarize the result as a theorem, see [84, Theorem 3.1, pp. 52–53].

Theorem 3.2. *Consider a TLS problem of the 1st class. If*

$$\sigma_n \;>\; \sigma_{n+1}\,, \tag{3.8}$$

then with the partitioning of the SVD of $[\,B \,|\, A\,]$ given by (3.2)–(3.3), $\Delta \equiv q = 0$, $V_{12} \in \mathbb{R}^{d\times d}$ is square and nonsingular, and (3.7) represents the unique TLS solution of the problem (2.3) with the corresponding correction matrix $[\,G \,|\, E\,]$ given by (3.6).

Theorem 3.1 with $\Delta = 0$ then has the following corollary.

Corollary 3.1. *Let (2.4) be the SVD of A and (2.6) the SVD of $[\,B \,|\, A\,]$ with the partitioning given by (3.2)–(3.3), $m \geq n+d$, $\Delta \equiv 0$. If*

$$\sigma'_n \;>\; \sigma_{n+1}\,, \tag{3.9}$$

then (2.3) is a problem of the first class, $\sigma_n > \sigma_{n+1}$ and (3.7) represents the unique TLS solution of the problem (2.3) with the corresponding correction matrix $[\,G \,|\, E\,]$ given by (3.6).

We see that (3.9) represents a sufficient condition for the existence of the TLS solution of the problem (2.3). If (3.9) is satisfied, then the TLS solution is unique. The condition (3.9) is, however, intricate. It may look as the key to the analysis of the TLS problem, in particular when one considers the following corollary of Theorems 2.1 and 3.1, see [84, Corollary 3.4, p. 65].

Corollary 3.2. *Let (2.4) be the SVD of A and (2.6) the SVD of $[\,B \,|\, A\,]$ with the partitioning given by (3.2)–(3.3), $m \geq n+d$, $\Delta \equiv q \geq 0$. Then the following conditions are equivalent:*

(i) $\sigma'_{n-q} > \sigma_{n-q+1} = \ldots = \sigma_{n+d}\,,$

(ii) $\sigma_{n-q} > \sigma_{n-q+1} = \ldots = \sigma_{n+d}$ *and* $V_{12}^{(q)}$ *is of rank* $d\,.$

Clearly, the condition (i) implies that the TLS problem is of the 1st class. If $d = 1$ and $q = 0$, then it reduces to (3.9) and the statement of Corollary 3.2 says that $\sigma'_n > \sigma_{n+1}$ if and only if $\sigma_n > \sigma_{n+1}$ and $e_1^T\,v_{n+1} \neq 0$.

In order to show the difficulty and motivate the further classification we now consider all remaining possibilities for the case $d = 1$. It should be 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 either

$$\sigma_n \;>\; \sigma_{n+1} \qquad \text{and} \qquad e_1^T\, v_{n+1} \;=\; 0\,,$$

which means that the TLS problem is not of the 1st class and it does not have a solution, or

$$\sigma_n \;=\; \sigma_{n+1}\,.$$

Depending on the relationship between σ'_{n-q} and $\sigma_{n-q+1} \;=\; \ldots \;=\; \sigma_{n+1}$ for some $q \;>\; 0$, see Corollary 3.2, the TLS problem may have in the last case 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 transparent with the use of the core problem concept from [64]. For any linear approximation problem $A\,x \;\approx\; b$ (we still consider $d \;=\; 1$) there are orthogonal matrices P, Q such that

$$P^T \;\left[\; b \;\middle|\; A \;\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q \end{array}\right] \;=\; \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right]\,, \tag{3.10}$$

where:

(i) A_{11} is of minimal dimensions and A_{22} is of maximal dimensions (it may also be nonexistent);

(ii) all singular values of A_{11} are simple and nonzero;

(iii) first components of all right singular vectors of $[\,b_1\,|\,A_{11}\,]$ are nonzero;

(iv) $\sigma_{\min}\,(A_{11}) \;>\; \sigma_{\min}\,([\,b_1\,|\,A_{11}\,])$,

see [64, Section 3] (and also Section 1.4). The minimally dimensioned subproblem $A_{11}\,x_1 \;\approx\; b_1$ is then called a *core problem* within $A\,x \;\approx\; b$. Please notice that the SVD of the block structured matrix on the right hand side can easily be got 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 the singular vectors corresponding to the second block by zeros on the top. Consequently, considering the special structure of the orthogonal transformation $\mathrm{diag}\,(\,1\,,\,Q\,)$ in (3.10), 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 the zero first components correspond to A_{22}. Now we are ready to finish the argument by reviewing all possible situations, see also Table 3.1.

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

2. Let $\sigma_{\min}\,(A) \;=\; \sigma_{\min}\,([\,b\,|\,A\,])$. Now we have to distinguish two cases.

   2a. Let $\sigma_{\min}\,(A) \;=\; \sigma_{\min}\,([\,b\,|\,A\,])$ be at the same time also the minimal singular value $\sigma_{\min}\,([\,b_1\,|\,A_{11}\,])$ which guarantees the existence of the (minimum norm) TLS solution. Since $\sigma_{\min}\,(A_{11}) \;>\; \sigma_{\min}\,([\,b_1\,|\,A_{11}\,])$ $=\; \sigma_{\min}\,(A)$, all singular values of A equal to $\sigma_{\min}\,(A)$ must be the singular values of the block A_{22}. Consequently, the multiplicity of $\sigma_{\min}\,([\,b\,|\,A\,])$ is larger by one than the multiplicity of $\sigma_{\min}\,(A)$.

2b. Let $\sigma_{\min}(A) = \sigma_{\min}([\,b\,|\,A\,])$ and let $\sigma_{\min}([\,b_1\,|\,A_{11}\,]) > \sigma_{\min}(A) \equiv \sigma_{\min}(A_{22})$. 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 the first components zero and the TLS solution does not exist.

Summarizing the argument, the TLS solution exists if and only if $\sigma_{\min}(A) > \sigma_{\min}([\,b\,|\,A\,])$ or $\sigma_{\min}(A) = \sigma_{\min}([\,b\,|\,A\,])$ with the multiplicities of $\sigma_{\min}(A)$ and $\sigma_{\min}([\,b\,|\,A\,])$ not equal. 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 [84].

$\sigma_{\min}(A_{22}) > \sigma_{\min}([\,b_1\,|\,A_{11}\,])$	$\Longleftrightarrow$	$\sigma_{\min}([\,b\,|\,A\,])$ is simple, the corresponding right singular vector has nonzero first component;
	$\Longleftrightarrow$	TLS solution exists, it is unique;
$\sigma_{\min}(A_{22}) = \sigma_{\min}([\,b_1\,|\,A_{11}\,])$	$\Longleftrightarrow$	$\sigma_{\min}([\,b\,|\,A\,])$ is multiple, there exists a corresponding right singular vector with nonzero first component;
	$\Longleftrightarrow$	TLS solution exists, it is nonunique;
$\sigma_{\min}(A_{22}) < \sigma_{\min}([\,b_1\,|\,A_{11}\,])$	$\Longleftrightarrow$	all the right singular vectors corresponding to $\sigma_{\min}([\,b\,|\,A\,])$ have zero first components;
	$\Longleftrightarrow$	TLS solution does not exist;

Table 3.1: Necessary and sufficient conditions for (non)existence of a TLS solution in the *single* right-hand side case, $d = 1$, see [64].

3.2.2 Problems of the 1st class with nonunique TLS solutions – a special case

Consider a TLS problem of the 1st class. Let in (3.1) $e \equiv d$, i.e. let all the singular values starting from $\sigma_{n-q+1} \equiv \sigma_{p+1}$ be equal,

$$\sigma_1 \geq \ldots \geq \sigma_p > \sigma_{p+1} = \ldots = \sigma_{n+1} = \ldots = \sigma_{n+d} \geq 0\,. \tag{3.11}$$

The case $q = 0$ ($p = n$) reduces to the problem with unique TLS solution discussed in Section 3.2.1. If $q = n$ ($p = 0$), i.e. $\sigma_1 = \ldots = \sigma_{n+d}$, then the columns of $[\,B\,|\,A\,]$ are mutually orthogonal, $[\,B\,|\,A\,]^T\,[\,B\,|\,A\,] = \sigma_1^2\,I_{n+d}$. Then it seems meaningless to approximate B by the columns of A, and we will get consistently with [84] the trivial solution $X_{\mathrm{TLS}} \equiv 0$. Therefore in the further text the nontrivial case is represented by $n > q > 0$ ($0 < p < n$).

As shown below, the correction matrix minimal in the Frobenius norm can be in this special case constructed from any d vectors (obtained as unitary linear combination of the last $q + d$ columns $v_{p+1}\,,\,\ldots\,,\,v_{n+d}$ 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 correction matrix is equal to* $\sigma_{n+1}\sqrt{d}$. Consequently, the TLS problem has infinitely many solutions. We concentrate on the construction of 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

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q \equiv \left[\begin{array}{c|c} v_{p+1}\,, \ldots\,, v_{n+d} \end{array}\right] Q = \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array}\right], \tag{3.12}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ is square and nonsingular. If $\sigma_{p+1} = \ldots = \sigma_{n+d} = 0$, then

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

and $X_{\mathrm{TLS}} = -Z\,\Gamma^{-1}$ is the minimum norm solution of the *compatible problem* $A\,X = B$. In the rest of the section we will consider $\sigma_{p+1} = \ldots = \sigma_{n+d} > 0$.

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 $V_{12}^{(q)T}$, and $Q_1 \in \mathbb{R}^{(q+d)\times q}$ is an orthonormal basis of its orthogonal complement, and

$$\left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] = \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2\,, \qquad V_{12}^{(q)} = \Gamma\, Q_2^T\,. \tag{3.13}$$

Such orthogonal matrix Q can be obtained, e.g., using the LQ decomposition

$$\begin{aligned} V_{12}^{(q)} &= \left[\begin{array}{c|c} L & 0 \end{array}\right] \tilde{Q}^T \\ &= \left[\begin{array}{c|c} L & 0 \end{array}\right] \left[\begin{array}{c|c} 0 & I_d \\ \hline I_q & 0 \end{array}\right] \left[\begin{array}{c|c} 0 & I_d \\ \hline I_q & 0 \end{array}\right]^T \tilde{Q}^T \\ &\equiv \left[\begin{array}{c|c} 0 & \Gamma \end{array}\right] Q^T\,, \end{aligned} \tag{3.14}$$

here $\Gamma \equiv L \in \mathbb{R}^{d\times d}$ is lower triangular. Alternatively, decomposition (3.12) can be constructed by application of d Householder transformation matrices on the matrix $V_{12}^{(q)}$ from the right, such that

$$V_{12}^{(q)}\, H_1 = \left[\begin{array}{ccc|ccccc} 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \nu_1 \\ \heartsuit & \cdots & \heartsuit & \heartsuit & \cdots & \heartsuit & \heartsuit & \heartsuit \\ \heartsuit & \cdots & \heartsuit & \heartsuit & \cdots & \heartsuit & \heartsuit & \heartsuit \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \heartsuit & \cdots & \heartsuit & \heartsuit & \cdots & \heartsuit & \heartsuit & \heartsuit \end{array}\right],$$

where ν_1 is the norm of the first row of $V_{12}^{(q)}$, and $\heartsuit$ denotes, in general, a nonzero component of a matrix. Further

$$V_{12}^{(q)}\, H_1\, H_2 = \left[\begin{array}{ccc|ccccc} 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \nu_1 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \nu_2 & \heartsuit \\ \heartsuit & \cdots & \heartsuit & \heartsuit & \cdots & \heartsuit & \heartsuit & \heartsuit \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \heartsuit & \cdots & \heartsuit & \heartsuit & \cdots & \heartsuit & \heartsuit & \heartsuit \end{array}\right],$$

where ν_2 is the norm of the left $(q+d-1)$-subrow of the second row of $V_{12}^{(q)}\, H_1$. Finally, after d steps, we obtain

$$V_{12}^{(q)} \underbrace{H_1 \ldots H_d}_{\equiv Q} = \left[\begin{array}{ccc|ccccc} 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \nu_1 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \nu_2 & \heartsuit \\ 0 & \cdots & 0 & 0 & & \nu_3 & \heartsuit & \heartsuit \\ \vdots & \ddots & \vdots & & \iddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \nu_d & \cdots & \heartsuit & \heartsuit & \heartsuit \end{array}\right] \equiv \left[\begin{array}{c|c} 0 & \Gamma \end{array}\right], \tag{3.15}$$

where ν_d is the norm of the left $(q+1)$-subrow of the last row of $V_{12}^{(q)}\, H_1 \,\ldots\, H_{d-1}$. The matrix $\Gamma \in \mathbb{R}^{d\times d}$ obtained form (3.15) is lower skew triangular (with zeros above the main skew diagonal, i.e., $e_i^T\,\Gamma\,e_j \;=\; 0$ if $i+j \;\leq\; d$). Obviously, the matrices Q obtained from (3.14) and (3.15) are different.

Define the correction matrix

$$\begin{aligned}
\left[\;\; G \;\middle|\; E \;\;\right] \;&\equiv\; -\;\left[\;\; B \;\middle|\; A \;\;\right] \left[\frac{\Gamma}{Z}\right] \left[\frac{\Gamma}{Z}\right]^T \\
&=\; -\; U\,\Sigma\,V^T \left[\frac{V_{12}^{(q)}}{V_{22}^{(q)}}\right] Q_2\,Q_2^T \left[\frac{V_{12}^{(q)}}{V_{22}^{(q)}}\right]^T \\
&=\; -\; \sigma_{n+1} \left[\;\; u_{p+1}\,,\,\ldots\,,\,u_{n+d} \;\;\right] \\
&\qquad\qquad\qquad\qquad Q_2\,Q_2^T \left[\;\; v_{p+1}\,,\,\ldots\,,\,v_{n+d} \;\;\right]^T\,,
\end{aligned} \tag{3.16}$$

where u_j and v_j are the corresponding left and right singular vectors of the matrix $[\,B\,|\,A\,]$, respectively. Please note that with the choice of any other matrix $Q' \;=\; [\,Q_1'\,|\,Q_2'\,]$ giving a decomposition of the form (3.12), Q_2' represents an orthonormal basis of the subspace spanned by the columns of $V_{12}^{(q)T}$, and therefore $Q_2 \;=\; Q_2'\,\Psi_2$ for some orthogonal matrix Ψ_2. Consequently, (3.16) is uniquely determined independently on the choice of Q in (3.12). Clearly, $\|\,[\,G\,|\,E\,]\,\|_F \;=\; \sigma_{n+1}\,\|\,Q_2\,Q_2^T\,\|_F \;=\; \sigma_{n+1}\,\sqrt{d}$, and the corrected matrix

$$\left[\;\; B+G \;\middle|\; A+E \;\;\right] \;\equiv\; \left[\;\; B \;\middle|\; A \;\;\right] \left(I_{n+d} \;-\; \left[\frac{\Gamma}{Z}\right] \left[\frac{\Gamma}{Z}\right]^T \right)$$

represents, by Theorem 2.2, the rank n approximation of $[\,B\,|\,A\,]$ such that the Frobenius norm of the correction matrix $[\,G\,|\,E\,]$ is minimal.

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,

$$\left[\;\; B+G \;\middle|\; A+E \;\;\right] \left[\frac{-\,I_d}{-\,Z\,\Gamma^{-1}}\right] \;=\; 0\,,$$

which gives the TLS solution

$$\begin{aligned}
X_{\mathrm{TLS}} \;\equiv\; X^{(q)} \;\equiv\; -\,Z\,\Gamma^{-1} \;&=\; -\;\left[\;\; Y \;\middle|\; Z \;\;\right] \left[\frac{0}{\Gamma^{-1}}\right] \\
&=\; -\;\left[\;\; Y \;\middle|\; Z \;\;\right] Q^T\,Q \left[\frac{0}{\Gamma^{-1}}\right] \\
&=\; -\; V_{22}^{(q)}\,V_{12}^{(q)\dagger}\,.
\end{aligned} \tag{3.17}$$

It can be expressed in the closed form

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

see [84, Theorem 3.10, pp. 62–64]. The solution (3.17) and the correction (3.16) do not depend on the choice of the matrix Q. Summarizing, the solution (3.17) represents the *unique TLS solution* of the problem (2.3), which is minimal in both the Frobenius norm and the 2-norm. For the Frobenius norm and the 2-norm of the solution (3.17) see (3.28) in Lemma 3.2, see also [84, Theorem 3.9, pp. 60–62]. We formulate the result as a theorem, see [84, Theorem 3.9, pp. 60–62].

Theorem 3.3. *Consider a TLS problem of the 1st class. If*

$$\sigma_p \;>\; \sigma_{p+1} \;=\; \ldots \;=\; \sigma_{n+d}\,, \tag{3.18}$$

then with the partitioning of the SVD of $[\,B\,|\,A\,]$ *given by (3.2)–(3.3),* $\Delta \equiv q < n$*, (3.17) represents the unique TLS solution of the problem (2.3), which is minimal in both the Frobenius norm and the 2-norm, with the corresponding unique correction matrix* $[\,G\,|\,E\,]$ *given by (3.16).*

Corollary 3.2 and Theorem 3.3 imply that the condition

$$\sigma'_p \;>\; \sigma_{p+1} \;=\; \ldots \;=\; \sigma_{n+d}\,, \tag{3.19}$$

represents a sufficient condition for the existence of the TLS solution of the problem (2.3). If $d = 1$, then (3.19) reduces to

$$\sigma'_p \;>\; \sigma_{p+1} \;=\; \ldots \;=\; \sigma_{n+1}\,, \tag{3.20}$$

i.e. $\sigma'_n > \sigma_{n+1}$ or the multiplicity of $\sigma_{\min}(A)$ is smaller than the multiplicity of $\sigma_{\min}([\,b\,|\,A\,])$.

For $d = 1$, it is clear from construction that any TLS problem of the first class must have a solution. For $d > 1$ this is, unfortunately, no longer true.

3.2.3 Problems of the 1st class – a general case

Consider a TLS problem of the 1st class with a general distribution of singular values, $e < d$

$$\begin{aligned}\sigma_1 \;\geq\; \ldots \;\geq\; \sigma_p \;>\; \sigma_{p+1} \;=\; \ldots \;=& \sigma_{n+1} \;=\; \ldots \\ \ldots \;=& \sigma_{n+e} \;>\; \sigma_{n+e+1} \;\geq\; \ldots \;\geq\; \sigma_{n+d} \;\geq\; 0\,.\end{aligned}$$

We will see that in this general case the problem (2.3) may not have a solution.

First, we apply the same approach as in Section 3.2.2. Since with the partitioning (3.2)–(3.3), $\Delta \equiv q$, $V_{12}^{(q)} \in \mathbb{R}^{d\times(q+d)}$ is of full row rank, there exists an orthogonal matrix $Q \in \mathbb{R}^{(q+d)\times(q+d)}$ such that

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q \;\equiv\; \left[\;\; v_{p+1}\,,\;\ldots\,,\;v_{n+d}\;\;\right] Q \;=\; \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array}\right], \tag{3.21}$$

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

$$\left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \;=\; \left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] Q_2\,, \qquad V_{12}^{(q)} \;=\; \Gamma\, Q_2^T\,. \tag{3.22}$$

Such orthogonal matrix Q can be obtained, similarly as in the previous section, e.g. by decompositions (3.14) or (3.15).

Define the correction matrix

$$\begin{aligned}\left[\;\; G \;\middle|\; E \;\;\right] \;\equiv\; -\left[\;\; B \;\middle|\; A \;\;\right] & \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T \\ =\; -\left[\;\; u_{p+1}\,,\;\ldots\,,\;u_{n+d}\;\;\right] & \operatorname{diag}\left(\,\sigma_{p+1}\,,\;\ldots\,,\;\sigma_{n+d}\,\right) \\ & Q_2\, Q_2^T \left[\;\; v_{p+1}\,,\;\ldots\,,\;v_{n+d}\;\;\right]^T .\end{aligned} \tag{3.23}$$

Similarly to the previous section, the matrix (3.23) is uniquely determined independently on the choice of Q in (3.21), see the argumentation below (3.16).

The columns of the matrix $[\,\Gamma^T \,|\, Z^T\,]^T$ represent a basis for the null space of the corrected matrix

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

Since Γ is square and nonsingular,

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

and we can construct

$$X^{(q)} \equiv -\,Z\,\Gamma^{-1} = -\,V_{22}^{(q)}\,V_{12}^{(q)\dagger}\,. \tag{3.24}$$

The matrices (3.24) and (3.23) do not depend on the choice of Q.

In the further text a minimality property of $X^{(q)}$ will be shown. First, the matrix $X^{(q)}$ given by (3.24) represents the solution of the compatible system $(A+E)\,X = B+G$, with E, G given by (3.23), minimal in the Frobenius norm. Both the compatibility of the given system and the fact that $X^{(q)}$ solves it follow from the construction, and can be verified simply by insertion. The range of the solution $X^{(q)}$ is orthogonal to the null space of $A+E$, i.e., it is a subset of the range of $(A+E)^T$, which proves the minimality of the solution (3.24) for the given fixed correction E, G. (Alternatively, the equality $X^{(q)} = (A+E)^\dagger\,(B+G)$ can be shown by insertion.)

Now we focus on the question whether there exists another correction $\tilde{E}$, $\tilde{G}$ obtained from the last $q+d$ columns of V, that makes the original system compatible, too, but which yields a solution smaller in norm. (Note that here we do not discuss about the norm of corrections, or, equivalently, whether such correction yields a solution of the TLS problem. This will be discussed later, below Theorem 3.4.)

Obviously, another correction can be defined similarly to (3.21) by considering an orthogonal matrix $\tilde{Q}$ such that

$$\left[\frac{V_{12}^{(q)}}{V_{22}^{(q)}}\right] \tilde{Q} = \left[\begin{array}{ccc} v_{p+1}\,, & \cdots\,, & v_{n+d} \end{array}\right] \tilde{Q} = \left[\begin{array}{c|c} \Omega & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right] , \tag{3.25}$$

where $\tilde{\Gamma} \in \mathbb{R}^{d\times d}$ is *nonsingular*. Define the correction matrix

$$\left[\begin{array}{c|c} \tilde{G} & \tilde{E} \end{array}\right] \equiv -\left[\begin{array}{c|c} B & A \end{array}\right] \left[\frac{\tilde{\Gamma}}{\tilde{Z}}\right] \left[\frac{\tilde{\Gamma}}{\tilde{Z}}\right]^T . \tag{3.26}$$

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

$$\tilde{X} \equiv -\,\tilde{Z}\,\tilde{\Gamma}^{-1} \tag{3.27}$$

solves this corrected system.

In order to compare the norms of the solutions $X^{(q)}$ and $\tilde{X}$ the following two lemmas will be useful, see also [84, Theorem 3.6, p. 55–56, and Theorem 3.9, p. 60–62].

Lemma 3.2. *Let* $[\,\tilde{\Gamma}^T \,|\, \tilde{Z}^T\,]^T \in \mathbb{R}^{(n+d)\times d}$ *have orthogonal 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*

$$\|\,\tilde{X}\,\|_F^2 = \|\,\tilde{\Gamma}^{-1}\,\|_F^2 - d\,, \qquad \|\,\tilde{X}\,\|^2 = \frac{1-\sigma_{\min}^2(\,\tilde{\Gamma}\,)}{\sigma_{\min}^2(\,\tilde{\Gamma}\,)}\,, \tag{3.28}$$

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

Proof. We follow the proof in [84]. From the CS decomposition [58], see also [32, pp. 77–79], it follows that there exist orthogonal matrices $W_1 \in \mathbb{R}^{d\times d}$, $W_2 \in \mathbb{R}^{n\times n}$, $T_2 \in \mathbb{R}^{d\times d}$ such that

$$\operatorname{diag}(W_1, W_2)^T \begin{bmatrix} \tilde{\Gamma} \\ \hline \tilde{Z} \end{bmatrix} T_2 = \begin{bmatrix} S \\ \hline C \end{bmatrix},$$

where

$$S = \operatorname{diag}(c_1, \ldots, c_k, I_{d-k}) \in \mathrm{R}^{d\times d},$$
$$C = \operatorname{diag}(s_1, \ldots, s_k, 0_{n-k,d-k}) \in \mathrm{R}^{n\times d},$$

for some k; and where $0 < s_1 \le \ldots \le s_k < 1$, $1 > c_1 \ge \ldots \ge c_k > 0$, and $c_j^2 + s_j^2 = 1$, for $j = 1, \ldots, k$. Then,

$$\|\tilde{X}\|_F^2 = \|C\,S^{-1}\|_F^2 = \sum_{j=1}^{k} \frac{1-s_j^2}{s_j^2} = \left(\sum_{j=1}^{k} \frac{1}{s_j^2}\right) - k = \|\tilde{\Gamma}^{-1}\|_F^2 - d,$$

and

$$\|\tilde{X}\|^2 = \|C\,S^{-1}\|^2 = \frac{1-s_1^2}{s_1^2} = \frac{1-\sigma_{\min}^2(\tilde{\Gamma})}{\sigma_{\min}^2(\tilde{\Gamma})},$$

because $\tilde{\Gamma}$ and S have the same singular values. □

Lemma 3.3. *Consider* $X^{(q)} = -Z\,\Gamma^{-1} = -V_{22}^{(q)}\,V_{12}^{(q)\dagger}$ *given by (3.21)–(3.24) and* $\tilde{X} = -\tilde{Z}\,\tilde{\Gamma}^{-1}$ *given by (3.25)–(3.27). Then*

$$\|\tilde{X}\|_F \ge \|X^{(q)}\|_F \qquad \textit{and} \qquad \|\tilde{X}\| \ge \|X^{(q)}\|, \tag{3.29}$$

and, moreover, equality of the Frobenius norms holds iff $\tilde{X} = X^{(q)}$.

Proof. Lemma 3.2 gives that the solution $\tilde{X}$ is going to be minimal in Frobenius or 2-norm when $\|\tilde{\Gamma}^{-1}\|_F$ is minimized, or when $\sigma_d(\tilde{\Gamma})$ is maximized, respectively. Note that $0 < \sigma_j(\tilde{\Gamma}) \le 1$. For any $\tilde{Q}$, the matrices $V_{12}^{(q)}$ and $V_{12}^{(q)}\,\tilde{Q} = [\,\Omega \,|\, \tilde{\Gamma}\,]$ have the same singular values. Theorem 2.1 applied to the matrices $[\,\Omega \,|\, \tilde{\Gamma}\,]$ and $\tilde{\Gamma}$ gives

$$\sigma_j(V_{12}^{(q)}) = \sigma_j\left(\left[\begin{array}{c|c} \Omega & \tilde{\Gamma} \end{array}\right]\right) \ge \sigma_j(\tilde{\Gamma}), \qquad j = 1, \ldots, d, \tag{3.30}$$

and all these inequalities become equalities iff $\Omega = 0$. (Moreover, Theorem 2.1 says that the singular values can not decrease while successively adding columns of Ω.)

The choice $\tilde{Q} \equiv Q$ ensures $\Omega = 0$ and thus it implies that all singular values are maximal. Consequently $X^{(q)}$ constructed using Q has minimal Frobenius as well as 2-norm among all $\tilde{X}$. (Note that the minimum for the Frobenius norm is reached iff all the inequalities in (3.30) become equalities. The minimum for the 2-norm is reached when only the dth (smallest) singular values in (3.30) are equal.) □

Summarizing, it was shown that $X^{(q)}$ is the solution of $(A+E)\,X = B+G$, where the correction is given by (3.23); $X^{(q)}$ is minimal in the Frobenius norm, and it is minimal among all solutions having the form (3.27) of all compatible systems $(A+\tilde{E})\,X = B+\tilde{G}$, with the correction having the form (3.26). The first part of this assertion can be shown elementarily, the second part was originally shown by Van Huffel, Vandewalle [84, Theorem 3.9].

The so called *classical TLS algorithm* [82, 83], see also [84], applied to a TLS problem of the 1st class returns as output the matrix $X^{(q)}$ given by (3.24) with the

matrices G, E given by (3.23). In the rest of this section we will show that $X^{(q)}$ does not generally represent a TLS solution of the given problem (2.3).

Now, we concentrate on an important question: whether (3.24) represents a TLS solution of the problem (2.3), or equivalently, whether the Frobenius norm of the correction matrix (3.23) is minimal. The following remark shows on an example that $X^{(q)}$ need not be a TLS solution.

Remark 3.1. *Let $q = n$, then the solution (3.24) is $X^{(q)} = 0$ and the correction matrix given by (3.23) is $[\,G\,|\,E\,] \equiv -[\,B\,|\,0\,]$. If $e < d$, then, for example, the ordinary least squares yields a correction $[\,\bar{G}\,|\,\bar{E}\,] \equiv [\,(A\,A^{\dagger} - I)\,B\,|\,0\,]$ having in general smaller Frobenius norm than $[\,G\,|\,E\,]$. Equivalently, $X^{(q)}$ does not represent the TLS solution in general.*

Further investigation is based on the following theorem.

Theorem 3.4. *Consider a TLS problem of the 1st class. Let (2.6) be the SVD of $[\,B\,|\,A\,]$ with the partitioning given by (3.2)–(3.3), $\Delta \equiv q < n$. Consider an orthogonal matrix $\tilde{Q}$ such that*

$$\left[\begin{array}{c|c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] \tilde{Q} = \left[\begin{array}{c|c} \Omega & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right], \qquad \tilde{Q} = \left[\begin{array}{c|c} \tilde{Q}_1 & \tilde{Q}_2 \end{array}\right] \tag{3.31}$$

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

$$\begin{aligned} \left[\begin{array}{c|c} \tilde{G} & \tilde{E} \end{array}\right] &\equiv - \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right] \left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right]^T \\ &= - \left[\begin{array}{c} u_{p+1}\,,\,\ldots\,,\,u_{n+d} \end{array}\right] \operatorname{diag}\left(\,\sigma_{p+1}\,,\,\ldots\,,\,\sigma_{n+d}\,\right) \\ &\qquad\qquad \tilde{Q}_2\,\tilde{Q}_2^T \left[\begin{array}{c} v_{p+1}\,,\,\ldots\,,\,v_{n+d} \end{array}\right]^T . \end{aligned} \tag{3.32}$$

Then the following two assertions are equivalent:

(i) *There exists an index k, $0 \le k \le e < d$, and an orthogonal matrix $\hat{Q}$ in the block diagonal form*

$$\hat{Q} = \left[\begin{array}{c|c} Q' & 0 \\ \hline 0 & I_{d-k} \end{array}\right] \in \mathbb{R}^{(q+d)\times(q+d)}, \qquad Q' \in \mathbb{R}^{(q+k)\times(q+k)}, \tag{3.33}$$

and using $\hat{Q}$ in (3.31), (3.32) instead of $\tilde{Q}$ gives the same $[\,\tilde{G}\,|\,\tilde{E}\,]$.

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

$$\left\| \left[\begin{array}{c|c} \tilde{G} & \tilde{E} \end{array}\right] \right\|_F = \left(\sum_{j=n+1}^{n+d} \sigma_j^2 \right)^{1/2} . \tag{3.34}$$

Remark 3.2. *If the condition* (i) *in Theorem 3.4 is satisfied for some index k, $0 \le k \le e$, then it is satisfied for any l, $k \le l \le e$, too, in particular for $l \equiv e$. The assertion* (i) *is usually used with e instead of k in the further text.*

Proof. First we prove the implication (i) $\Longrightarrow$ (ii). From (3.33) we get

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

where $\hat{Q}_2 \in \mathbb{R}^{(q+d)\times d}$, $Q_2' \in \mathbb{R}^{(q+k)\times k}$, which gives, using (3.32),

$$\begin{aligned} \| \left[\begin{array}{c|c} \tilde{G} & \tilde{E} \end{array} \right] \|_F^2 &= \| \operatorname{diag}(\sigma_{p+1}, \ldots, \sigma_{n+d}) \hat{Q}_2 \hat{Q}_2^T \|_F^2 \\ &= \sigma_{n+1}^2 \| Q_2' (Q_2')^T \|_F^2 + \sigma_{n+1}^2 (e-k) + \sum\nolimits_{j=n+e+1}^{n+d} \sigma_j^2 \\ &= \sigma_{n+1}^2 e + \sum\nolimits_{j=n+e+1}^{n+d} \sigma_j^2 , \end{aligned}$$

and (3.34) is proved.

Now we prove the implication (ii) $\Longrightarrow$ (i). Let $[\tilde{G} \,|\, \tilde{E}]$ be given by (3.31), (3.32) and let (3.34) holds. We prove that there exists $\hat{Q}$ of the form (3.33) giving $[\tilde{G} \,|\, \tilde{E}]$. Define the splitting

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

such that $\tilde{Q}_{11} \in \mathbb{R}^{(q+e)\times q}$, $\tilde{Q}_{21} \in \mathbb{R}^{(d-e)\times q}$, $\tilde{Q}_{12} \in \mathbb{R}^{(q+e)\times d}$, $\tilde{Q}_{22} \in \mathbb{R}^{(d-e)\times d}$, see Figure 3.2.

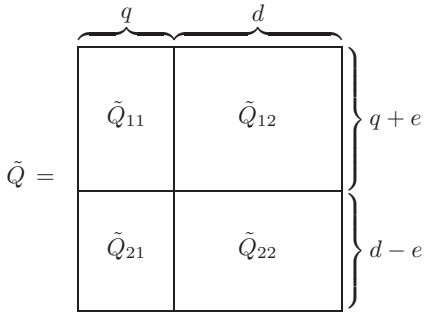


Figure 3.2: Dimensions of the individual matrix blocks in the partitioning of matrix $\tilde{Q}$.

The matrix $[\tilde{G} \,|\, \tilde{E}]$ given by (3.31) satisfies

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

where $D \equiv \operatorname{diag}(\sigma_{n+e+1}, \ldots, \sigma_{n+d})$. 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

$$\begin{aligned} \| \left[\begin{array}{c|c} \tilde{G} & \tilde{E} \end{array} \right] \|_F^2 &= \sigma_{n+1}^2 (d - \| \tilde{Q}_{22} \|_F^2) + \| D \tilde{Q}_{22} \|_F^2 \\ &= \sigma_{n+1}^2 d - \| (\sigma_{n+1}^2 I_{d-e} - D^2)^{1/2} \tilde{Q}_{22} \|_F^2 . \end{aligned}$$

The assumption (3.34) requires

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

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

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

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

$$R \equiv W \left[\begin{array}{c|c} 0 & S^T \\ \hline I_e & 0 \end{array}\right] \in \mathbb{R}^{d\times d}$$

and

$$\hat{Q} \equiv \tilde{Q} \left[\begin{array}{c|c} I_q & 0 \\ \hline 0 & R \end{array}\right] = \left[\begin{array}{c|c} \tilde{Q}_{11} & \tilde{Q}_{12}\,R \\ \hline 0 & [\,0\,|\,I_{d-e}\,] \end{array}\right] .$$

Because $\hat{Q}_2\,\hat{Q}_2^T = (\tilde{Q}_2\,R)\,(\tilde{Q}_2\,R)^T = \tilde{Q}_2\,\tilde{Q}_2^T$, the matrix $\hat{Q}$ yields the same correction (3.32) as $\tilde{Q}$. Since $\hat{Q}$ is orthogonal, the columns of $\tilde{Q}_{12}\,R$ corresponding to the block I_{d-e} must be zero, thus

$$\hat{Q} = \left[\begin{array}{c|c} Q' & 0 \\ \hline 0 & I_{d-e} \end{array}\right]$$

has the form (3.33) with $Q' \in \mathbb{R}^{(q+e)\times(q+e)}$. □

The following remark slightly extends the assertion of Theorem 3.4.

Remark 3.3. *Let $\tilde{Q}$ be the a matrix having the general block diagonal form $\tilde{Q} = \mathrm{diag}\,(\,Q'\,,\,Q''\,)$, where $Q' \in \mathbb{R}^{(q+k)\times(q+k)}$, $Q'' \in \mathbb{R}^{(d-k)\times(d-k)}$, $0 \le k \le e < d$. Then define $\hat{Q} \equiv \tilde{Q}\,\mathrm{diag}\,(\,I_{q+k}\,,\,(Q'')^T\,) = \mathrm{diag}\,(\,Q'\,,\,I_{d-k}\,)$, obviously $\hat{Q}$ yields the same correction (3.32) as $\tilde{Q}$, and has the form (3.33). The other implication is trivial.*

Consequently, the identity block I_{d-k} in $\hat{Q}$ in the condition (i) *of Theorem 3.4 can be replaced by any orthogonal matrix Q'' having the same dimensions.*

Application of Theorem 3.4 on the construction (3.21)–(3.24) immediately gives the necessary and sufficient condition for $X^{(q)}$ given by (3.24) to be a TLS solution, and thus a sufficient condition for existence of a TLS solution. We formulate it as the following corollary.

Corollary 3.3. *Consider a TLS problem (2.3) of the 1st class. The construction (3.21)–(3.24) 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 (3.33) such that substituting $\hat{Q}$ for Q in (3.21)–(3.23) gives the same correction $[\,E\,|\,G\,]$.*

Proof. Let $X^{(q)}$ given by (3.24) represent a TLS solution of the problem (2.3). Then the correction $[\,E\,|\,G\,]$ given by (3.23) is minimal in Frobenius norm, i.e. it satisfies (3.34). Consequently, there exists an orthogonal matrix $\hat{Q}$ in the form (3.33) by Theorem 3.4.

Let the matrix Q from (3.21) have the block diagonal form (3.33). Then $[\,E\,|\,G\,]$ given by (3.23) satisfies (3.34) by Theorem 3.4 and, further, it represents a correction (reducing rank of $[\,B\,|\,A\,]$ to n) which is minimal in the Frobenius norm, by Theorem 2.2. Consequently, $X_{\mathrm{TLS}} \equiv X^{(q)}$ defined by (3.24) represents a TLS solution of the problem (2.3) by Theorem 3.4. □

Now we describe three disjoint sets of problems of the 1st class. The first set $\mathscr{F}_1$ contains problems for which there always exist Q in the block diagonal form (3.33) satisfying (3.21) (i.e. (3.31) with $\Omega = 0$). For such problems the TLS solution

in the form (3.24) (i.e., the TLS solution having the minimality property (3.29)) always exists, by Corollary 3.3.

The second set $\mathscr{F}_2$ contains problems for which there always exists Q in the block diagonal form (3.33) satisfying (3.31) but only with $\Omega \neq 0$. Such problems always have a TLS solution but not in the form (3.24). Here $X^{(q)}$ does not represent a TLS solution.

The third set $\mathscr{F}_3$ contains problems for which there is no Q in the block diagonal form (3.33) yielding Γ nonsingular in (3.31). These problems do not have a TLS solution in the form (3.24), (3.27).

The solution $X^{(q)}$ given by (3.24) is commonly used for all problems of the 1st class (e.g., in the TLS algorithm, see [84]). However for any problem from the set $\mathscr{F}_2 \cup \mathscr{F}_3$ it does not solve the TLS problem; in these cases we call $X^{(q)}$ *nongeneric (nonoptimal) solution.*

In order to describe these three sets we introduce the following notation. Define the partitioning of the matrix $V_{12}^{(q)}$ for a given k, $0 \leq k \leq e < d$,

$$V_{12}^{(q)} = \left[\begin{array}{c|c} W^{(q,k)} & V_{12}^{(-k)} \end{array} \right], \tag{3.35}$$

where $W^{(q,k)} \in \mathbb{R}^{d\times(q+k)}$, $V_{12}^{(-k)} \in \mathbb{R}^{d\times(d-k)}$, see Figure 3.3.

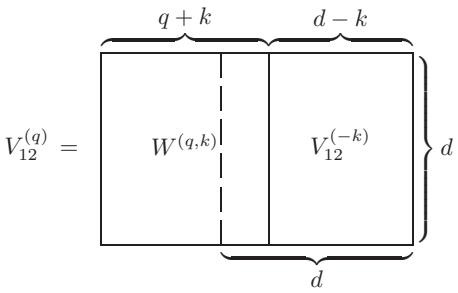


Figure 3.3: Dimensions of the individual matrix blocks in the partitioning (3.35).

Obviously, since $\operatorname{rank}(V_{12}^{(q)}) = d$ (TLS problem of the 1st class), $\operatorname{rank}(V_{12}^{(-k)}) \leq d-k$ implies $\operatorname{rank}(W^{(q,k)}) \geq k$.

3.2.4 Problems of the 1st class for which $\operatorname{rank}(V_{12}^{(-e)}) = d-e$ and $\operatorname{rank}(W^{(q,e)}) = e$, set $\mathscr{F}_1$

Consider a TLS problem of the 1st class. Let $\operatorname{rank}(W^{(q,e)}) = e$ in (3.35), then $V_{12}^{(-e)}$ must be of full column rank, i.e. $\operatorname{rank}(V_{12}^{(-e)}) = d-e$.

First we give a lemma which will allow to relate the partitioning (3.35) to the construction of a solution as in (3.21)–(3.24).

Lemma 3.4. *Let (2.6) be the SVD of $[\,B\,|\,A\,]$ with the partitioning given by (3.2)–(3.3), $\Delta \equiv q < n$. Consider the partitioning (3.35) of $V_{12}^{(q)}$ with $k \equiv e$. The following two assertions are equivalent:*

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

(ii) *There exists Q of the block diagonal form (3.33) with $k \equiv e$ satisfying (3.21).*

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 $\Gamma \equiv [\,M\,|\,V_{12}^{(-e)}\,] \in \mathbb{R}^{d\times d}$ has rank equal to d.

On the contrary, let Q have the form (3.33) with $k \equiv e$ and satisfy (3.21). 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 $\Gamma_1 = W^{(q,e)}\,Q'$, $\Gamma_2 = V_{12}^{(-e)}\,I_{d-e} = V_{12}^{(-e)}$. Since Γ is nonsingular, $\mathrm{rank}\,(\Gamma_1) = e$. Moreover, Q' is an orthogonal matrix and thus $\mathrm{rank}\,(W^{(q,e)}) = e$. □

Obviously, if the problem is of the 1st class and $\mathrm{rank}\,(W^{(q,k)}) = k$ for some $0 \le k \le e$, then for any l, $k \le l \le e$, $\mathrm{rank}\,(W^{(q,l)}) = l$.

The following theorem summarizes the case in which our construction (3.21)–(3.24) gives the TLS solution with the minimality property (3.29).

Theorem 3.5. *Let (2.6) be the SVD of $[\,B\,|\,A\,]$ with the partitioning given by (3.2)–(3.3), $m \ge n+d$, $\Delta \equiv q < n$ ($p = n-q$). Let the TLS problem (2.3) be of the 1st class, i.e. $V_{12}^{(q)}$ is of full row rank equal to d. Let $\sigma_p > \sigma_{p+1} = \ldots = \sigma_{n+1} = \ldots = \sigma_{n+e}$, $1 \le e \le d$. Consider the partitioning of $V_{12}^{(q)}$ given by (3.35) with $k \equiv e$. If*

$$\mathrm{rank}\,(W^{(q,e)}) = e\,, \tag{3.36}$$

then $X_{\mathrm{TLS}} \equiv -\,V_{22}^{(q)}\,V_{12}^{(q)\dagger}$, see also (3.24), represents the TLS solution of the problem (2.3) having the minimality property (3.29), with the corresponding correction $[\,G\,|\,E\,]$ given by (3.23).

The proof follows immediately from Lemma 3.4 and Corollary 3.3.

Remark 3.4. *Naturally, both problems of the 1st class discussed earlier in Sections 3.2.1 and 3.2.2 belong to the set $\mathscr{F}_1$, too. In the first case $q \equiv 0$, $V_{12}^{(q)} \equiv V_{12}$ is square nonsingular. Thus any partitioning of the form (3.35) yields $W^{(0,k)}$ with the full column rank equal to k. In the second case $e \equiv d$. Thus $W^{(q,d)} \equiv V_{12}^{(q)}$ is of full row rank equal to d.*

In both cases, one of the diagonal blocks of $Q = \mathrm{diag}\,(Q'\,,\,I_{d-e})$ from the condition (ii) *in Lemma 3.4 is nonexistent in general.*

Please note that unlike in Sections 3.2.1 and 3.2.2 here the information about singular values of A and $[\,B\,|\,A\,]$, except for the special cases $q \equiv 0$ or $e \equiv d$ does not guarantee that the construction (3.24) gives the TLS solution (see also Corollaries 3.1 and 3.2).

3.2.5 Problems of the 1st class for which $\mathrm{rank}\,(V_{12}^{(-e)}) = d-e$ and $\mathrm{rank}\,(W^{(q,e)}) > e$, set $\mathscr{F}_2$

Consider a TLS problem of the 1st class. Let $\mathrm{rank}\,(V_{12}^{(-e)}) = d-e$ and $\rho \equiv \mathrm{rank}\,(W^{(q,e)}) > e$ in (3.35).

Because $V_{12}^{(-e)}$ is of full column rank, there exists $\tilde{Q}$ having the block diagonal form (3.33) with $k \equiv e$, such that (3.31) holds and $\tilde{\Gamma}$ is nonsingular. Consequently there exists the correction $[\,\tilde{G}\,|\,\tilde{E}\,]$ defined by (3.32) minimal in Frobenius norm, see Theorem 3.4, and the matrix

$$\tilde{X} \equiv -\,\tilde{Z}\,\tilde{\Gamma}^{-1} \tag{3.37}$$

represents a TLS solution.

Consequently problems of the first class having full column rank $V_{12}^{(-e)}$ (i.e. problems from the set $\mathscr{F}_1 \cup \mathscr{F}_2$) always have a TLS solution. But if $W^{(q,e)}$ has rank bigger than e, then $X^{(q)} = -V_{22}^{(q)} V_{12}^{(q)\dagger}$ given by (3.24) does not represent a TLS solution. Recall that in such case $X^{(q)}$ is called *nongeneric (nonoptimal) solution*, i.e. $X_{\mathrm{NGN}} \equiv X^{(q)}$.

Note that here the information about singular values of A and $[\,B\,|\,A\,]$ can be used to guarantee that the TLS solution exists, i.e., that the given problem belongs to the set $\mathscr{F}_1 \cup \mathscr{F}_2$, but the condition is only sufficient and practically useless. We do not formulate this condition here explicitly, but it is based on the assertion of Theorem 3.1. The idea is analogous to Corollaries 3.1 and 3.2: suitable singular value inequalities guarantee that $V_{12}^{(q)}$ is of full row rank and $V_{12}^{(-e)}$ is of full column rank. Together with the interlacing property (Theorem 2.1) it gives a sufficient condition for existence of the TLS solution.

3.2.6 Problems of the 1st class for which $\mathrm{rank}\,(V_{12}^{(-e)}) < d - e$ and $\mathrm{rank}\,(W^{(q,e)}) > e$, set $\mathscr{F}_3$

Consider a TLS problem of the 1st class. Let $\mathrm{rank}\,(V_{12}^{(-e)}) < d - e$ in (3.35). Then $V_{12}^{(-e)}$ is rank deficient and $\mathrm{rank}\,(W^{(q,e)}) > e$.

Obviously in this case there does not exist $\tilde{Q}$ in the block diagonal form (3.33) yielding $\tilde{\Gamma}$ from (3.31) nonsingular. Consequently problems with rank deficient $V_{12}^{(-e)}$ do not have the TLS solution in the form (3.27). Similarly as in the $\mathscr{F}_2$ set, the matrix $X^{(q)} = -V_{22}^{(q)} V_{12}^{(q)\dagger}$ given by (3.24) is called the *nongeneric (nonoptimal) solution*, i.e. $X_{\mathrm{NGN}} \equiv X^{(q)}$.

For clarification of the structure of sets $\mathscr{F}_1$, $\mathscr{F}_2$, and $\mathscr{F}_3$ of problems see also Figure 3.5 (p. 55).

3.2.7 Correction corresponding to the solution $X^{(q)}$

In the further text we focus on the correction and solution given by (3.23), (3.24), respectively. In particular we focus on the cases with nongeneric solution. First we establish bounds for the corresponding correction. Obviously, the correction $[\,G\,|\,E\,]$ given by (3.23) is not optimal in the meaning of reducing rank in general; the norm of the correction can grow beyond $(\sum_{j=n+1}^{n+d} \sigma_j^2)^{1/2}$.

Lemma 3.5. *Let $[\,G\,|\,E\,]$ be the correction matrix given by (3.21), (3.23). Then the Frobenius norm of the correction satisfies*

$$\left(\sum_{j=p+1}^{p+d} \sigma_j^2 \right)^{1/2} \geq \left\| \left[\begin{array}{c|c} G & E \end{array} \right] \right\|_F \geq \left(\sum_{j=n+1}^{n+d} \sigma_j^2 \right)^{1/2} . \tag{3.38}$$

See also [89, Equation (5.4)]. Before the proof of Lemma 3.5, we quickly discuss the inequality (3.38) and prove an auxiliary lemma. Obviously:

(i) The norm of the correction is equal to the lower bound in (3.38) iff the condition (3.36) is satisfied, i.e. iff the problem belongs to the set $\mathscr{F}_1$, see also Theorem 3.5.

(ii) Both (left and right) inequalities in (3.38) become equalities iff

$$\sigma_{p+j} = \sigma_{n+j}\,, \qquad \forall\, j = 1\,, \ldots\,, d\,. \tag{3.39}$$

Note that $n = p + q$. The equalities in (3.39) imply that either

a) $q = 0$ (the simplest case discussed in Section 3.2.1), or

b) $\sigma_{p+1} = \ldots = \sigma_{n+d}$ (the special case discussed in Section 3.2.2).

Lemma 3.6. *Let $[\,G\,|\,E\,]$ be the correction matrix given by (3.21), (3.23). Then the rank of the correction matrix satisfies*

$$\min\{\,s\,,\,d\,\} \;\geq\; \mathrm{rank}\,\left(\left[\begin{array}{c|c} G & E \end{array}\right]\right) \;\geq\; \max\{\,0\,,\,s-n\,\}\,, \tag{3.40}$$

and the rank of the corrected matrix satisfies

$$\max\{\,0\,,\,s-d\,\} \;\leq\; \mathrm{rank}\,\left(\left[\begin{array}{c|c} B+G & A+E \end{array}\right]\right) \;\leq\; \min\{\,s\,,\,n\,\}\,, \tag{3.41}$$

where $s \equiv \mathrm{rank}\,([\,B\,|\,A\,])$.

Proof. The upper bound in (3.40) follows from the fact that the correction matrix is the rank s matrix $[\,B\,|\,A\,]$ projected onto a d dimensional subspace by the orthogonal projector $[\frac{\Gamma}{Z}][\frac{\Gamma}{Z}]^T$. The lower bound in (3.40) follows from the fact that the correction is constructed such that it makes the system compatible, it has to reduce the rank of $[\,B\,|\,A\,]$ at least to n (and, naturally, it must be positive).

Similarly, the upper bound in (3.41) is given through the fact that the corrected matrix is the rank s matrix $[\,B\,|\,A\,]$ projected onto a n dimensional subspace by the orthogonal projector $(I_{n+d} - [\frac{\Gamma}{Z}][\frac{\Gamma}{Z}]^T)$, or equivalently from the fact that the correction matrix is constructed such that the corrected matrix have rank at most n (and, naturally, it must be smaller than the rank of the original matrix). The lower bound in (3.41) can not be smaller than $s - d$ because the rank of $[\,G\,|\,E\,]$ is at most d (naturally it must be positive). □

Proof of Lemma 3.5. The lower bound in (3.38) is clear, it follows directly from Theorem 2.2. Now, we derive the upper bound. From (3.40), the matrix $[\,G\,|\,E\,]$ has rank at most $\rho \equiv \min\{\,s\,,\,d\,\}$, and it is obtained by projecting of $M \equiv U\,\Sigma_2^{(q)}\,[\,v_{p+1}\,,\,\ldots\,,\,v_{n+d}\,]$ onto a subspace with dimension at most ρ. When we approximate the matrix M by the nearest (in the Frobenius norm) matrix with rank equal to ρ, we obtain the correction with the maximal Frobenius norm equal to the upper bound in (3.39). □

Recall that the matrix $X^{(q)} = -\,V_{22}^{(q)}\,V_{12}^{(q)\dagger}$ given by (3.24) represents the minimum norm solution of the compatible system $(A+E)\,X = B+G$, where the correction $[\,G\,|\,E\,]$ is defined by (3.23). Two another important and useful interpretations of this (possibly nongeneric) solution of the original TLS problem (2.3) can be established.

Lemma 3.7. *The matrix $X^{(q)} = -\,V_{22}^{(q)}\,V_{12}^{(q)\dagger}$ given by (3.24) represents the unique solution of the constrained minimization problem*

$$\begin{array}{l} \displaystyle\min_{X,E,G} \left\|\left[\begin{array}{c|c} G & E \end{array}\right]\right\|_F \qquad \textit{subject to} \qquad (A+E)\,X \;=\; B+G \\[2ex] \textit{and} \quad \left[\begin{array}{c|c} G & E \end{array}\right] \left[\begin{array}{c} 0 \\ \hline w \end{array}\right] \;=\; 0\,, \quad \forall\, \left[\begin{array}{c} 0 \\ \hline w \end{array}\right] \in \mathcal{R}\left(\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right]\right)\,, \end{array} \tag{3.42}$$

with the corresponding correction $[\,G\,|\,E\,]$ defined by (3.23).

Since $\sigma_{n-q} > \sigma_{n-q+1}$, the additional constraint ensures that the correction matrix in (3.42) is given uniquely. Consequently, the constrained problem (3.42) always has unique solution $X_{\mathrm{Const.}} \equiv X^{(q)}$. See [84, Definition 3.3, p. 78 and Theorem 3.15, pp. 80–82], see also Lemma 3.11.

Lemma 3.8. *The matrix* $X^{(q)} = -V_{22}^{(q)} V_{12}^{(q)\dagger}$ *given by (3.24) represents the unique minimum norm TLS solution of the modified TLS problem*

$$\min_{X,\hat{E},\hat{G}} \left\| \left[\begin{array}{c|c} \hat{G} & \hat{E} \end{array} \right] \right\|_F \qquad \text{subject to} \qquad (\hat{A}+\hat{E})\, X = \hat{B}+\hat{G}\,, \tag{3.43}$$

where

$$\left[\begin{array}{c|c} \hat{B} & \hat{A} \end{array} \right] = \left(\sum\nolimits_{j=1}^{n-q} u_j\, \sigma_j\, v_j^T \right) + \sigma_{n+1} \left(\sum\nolimits_{j=n-q+1}^{n+d} u_j\, v_j^T \right),$$

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

Obviously the problem (3.43) is a TLS problem of the 1st class (from the set $\mathscr{F}_1$); moreover it is the special case problem described in Section 3.2.2. The problem is called *truncated total least squares (T-TLS) problem* for given A, B, with the solution $X_{\text{T-TLS}} \equiv X^{(q)}$. See [84, note on p. 82], see also Lemma 3.12.

3.3 Problems of the 2nd class

Problems which are not of the 1st class we call *problems of the 2nd class.*

Definition 3.2 (Problem of the 2nd class)**.** *Consider a TLS problem (2.3),* $m \geq n+d$*. Let (2.6) be the SVD of* $[\,B\,|\,A\,]$ *with the partitioning given by (3.2)–(3.3),* $\Delta \equiv q$*, where* q *is the integer related to the multiplicity of* σ_{n+1}*, given by (3.1). Let* $V_{12}^{(q)}$ *be rank deficient. Then we call (2.3) a TLS problem of the 2nd class.*

In this section we focus on problems (2.1) for which $V_{12}^{(q)}$ does not have full row rank. 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 any solution in the form (3.27). Thereby the problems of the 2nd class can not have a TLS solution having the form (3.27). In order to get at least some solution Van Huffel and Vandewalle follow the single right-hand side concept. The right singular vector subspace $[\,(V_{12}^{(q)})^T\,|\,(V_{22}^{(q)})^T\,]^T$ used for the construction (3.21)–(3.24) in all previous cases, is successively extended with additional right singular vectors until a full row rank block $V_{12}^{(t)} \in \mathbb{R}^{d\times(t+d)}$ is found in the upper right corner of V (i.e. $V_{12}^{(t-1)}$ is rank deficient), see Figure 3.4. Then the matrix $X^{(t)} = -V_{22}^{(t)} V_{12}^{(t)\dagger}$ with the corresponding correction can be constructed analogously to (3.21)–(3.24), with q replaced by t. Obviously, this matrix might not be uniquely defined 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 following notation

$$\sigma_{n-\tilde{q}} > \sigma_{n-\tilde{q}+1} = \ldots = \sigma_{n-t+1}\,, \tag{3.44}$$

where $\tilde{q} \geq t$; put for simplicity $n-\tilde{q} \equiv \tilde{p}$. (If $\sigma_{n-\tilde{q}} \equiv \sigma_{\tilde{p}}$ is nonexistent, then $\tilde{q} = n$.) 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 (3.21)–(3.24), with q replaced by $\tilde{q}$. The construction is completely analogous.

Since $V_{12}^{(\tilde{q})} \in \mathbb{R}^{d\times(\tilde{q}+d)}$ is of full row rank, there always exists an orthogonal matrix $Q \in \mathbb{R}^{(\tilde{q}+d)\times(\tilde{q}+d)}$ such that

$$\left[\begin{array}{c} V_{12}^{(\tilde{q})} \\ \hline V_{22}^{(\tilde{q})} \end{array} \right] Q \equiv \left[\begin{array}{ccc} v_{\tilde{p}+1}\,, & \ldots\,, & v_{n+d} \end{array} \right] Q = \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array} \right], \tag{3.45}$$

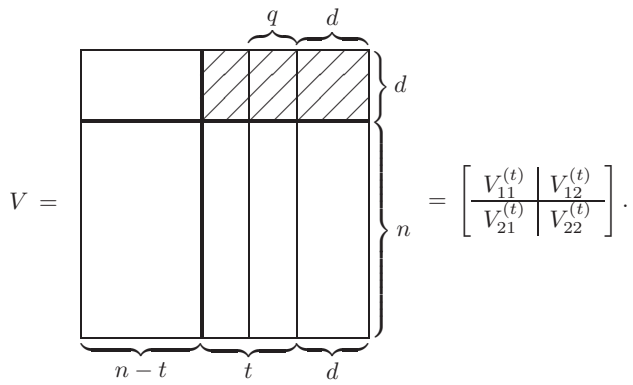


Figure 3.4: Dimensions and the partitioning of the matrix V for problems of the 2nd class.

where $\Gamma \in \mathbb{R}^{d\times d}$ is nonsingular. Define the correction matrix

$$\left[\begin{array}{c|c} G & E \end{array}\right] \equiv - \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T \tag{3.46}$$

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

$$X^{(\tilde{q})} \equiv - Z\,\Gamma^{-1} = - V_{22}^{(\tilde{q})}\, V_{12}^{(\tilde{q})\dagger}\,. \tag{3.47}$$

The matrices (3.47) and (3.46) do not depend on the choice of Q.

Similarly to the problems of the 1st class, the minimality property of $X^{(\tilde{q})}$ can be shown. It is the solution of the compatible corrected system $(A+E)\,X = B+G$ with E, G given by (3.46), minimal in the Frobenius norm. The Frobenius and the 2-norm of the solution $X^{(\tilde{q})}$ are given by Lemma 3.2 (the extension of Lemma 3.3 is straightforward, too). Further, $X^{(\tilde{q})}$ has minimal Frobenius and 2-norm over all solutions $\tilde{X}$ that can be obtained from the construction (3.25)–(3.27) with q replaced by $\tilde{q}$. (Note that the solution obtained from $V_{12}^{(t)}$, where t is the smallest index for which $V_{12}^{(t)}$ is of full column rank, is one of these $\tilde{X}$ solutions.) Thus, the substitution of $\tilde{q}$ for t ensures, except of the uniqueness of the construction, the solution with smallest norm, on the other hand it causes increasing the correction norm. Clearly, the Frobenius norm of the correction (3.46) is

$$\left\|\left[\begin{array}{c|c} G & E \end{array}\right]\right\|_F > \left(\sum_{j=n+1}^{n+d} \sigma_j^2\right)^{1/2},$$

always bigger that the smallest possible correction reducing rank of $[\,B \,|\, A\,]$ to n. Thus the matrix $X^{(\tilde{q})}$ given by (3.47) does not represent the TLS solution. The solution $X^{(\tilde{q})} = - V_{22}^{(\tilde{q})}\, V_{12}^{(\tilde{q})\dagger}$ is called *nongeneric (nonoptimal) solution* of the original TLS problem (2.3), see [84, Definition 3.3, p. 78].

Similarly to the problems of the 1st class two other important and useful interpretations of this nongeneric solution can be established.

Lemma 3.9. *The matrix $X^{(\tilde{q})} = - V_{22}^{(\tilde{q})}\, V_{12}^{(\tilde{q})\dagger}$ given by (3.47) represents the unique*

solution of the constrained minimization problem

$$\begin{aligned} &\min_{X,E,G} \left\| \left[\begin{array}{c|c} G & E \end{array}\right] \right\|_F \qquad \textit{subject to} \qquad (A+E)\,X \,=\, B+G \\ &\textit{and} \quad \left[\begin{array}{c|c} G & E \end{array}\right] \left[\begin{array}{c} 0 \\ \hline w \end{array}\right] \,=\, 0\,, \quad \forall\, \left[\begin{array}{c} 0 \\ \hline w \end{array}\right] \in \mathcal{R}\left(\left[\begin{array}{c} V_{12}^{(\tilde{q})} \\ \hline V_{22}^{(\tilde{q})} \end{array}\right]\right)\,, \end{aligned} \tag{3.48}$$

with the corresponding correction $[\,G\,|\,E\,]$ *defined by (3.23).*

Since $\sigma_{n-\tilde{q}} > \sigma_{n-\tilde{q}+1}$, the additional constraint ensures that the correction matrix in (3.48) is given uniquely. Consequently, the constrained problem (3.48) always has unique solution $X_{\text{Const.}} \equiv X^{(\tilde{q})}$. See [84, Definition 3.3, p. 78 and Theorem 3.15, pp. 80–82], the problem (3.48) from Lemma 3.9 is in [84] considered as the definition of the nongeneric solution; see also Lemma 3.11.

Lemma 3.10. *The matrix* $X^{(\tilde{q})} = -\,V_{22}^{(\tilde{q})}\,V_{12}^{(\tilde{q})\dagger}$ *given by (3.47) represents the unique minimum norm TLS solution of the modified TLS problem*

$$\min_{X,\hat{E},\hat{G}} \left\| \left[\begin{array}{c|c} \hat{G} & \hat{E} \end{array}\right] \right\|_F \qquad \textit{subject to} \qquad (\hat{A}+\hat{E})\,X \,=\, \hat{B}+\hat{G}\,, \tag{3.49}$$

where

$$\left[\begin{array}{c|c} \hat{B} & \hat{A} \end{array}\right] \,=\, \left(\sum\nolimits_{j=1}^{\tilde{p}} u_j\,\sigma_j\,v_j^T\right) \,+\, \sigma_{n-t+1}\,\left(\sum\nolimits_{j=\tilde{p}+1}^{n+d} u_j\,v_j^T\right)\,,$$

with the corresponding correction $[\,\hat{G}\,|\,\hat{E}\,]$, $\|[\,\hat{G}\,|\,\hat{E}\,]\|_F \,=\, \sigma_{n-t+1}\,\sqrt{d}$. *(Recall that* $\sigma_{n-\tilde{q}+1} \,=\, \sigma_{n-t+1}$, *see (3.44).)*

Obviously the problem (3.49) is a TLS problem of the 1st class (from the set $\mathscr{F}_1$); moreover it is the special case problem described in Section 3.2.2. The problem is called *truncated total least squares (T-TLS) problem* for given A, B, with the solution $X_{\text{T-TLS}} \equiv X^{(\tilde{q})}$. See [84, note on p. 82], see also Lemma 3.12.

3.4 Summary and the TLS algorithm

The following theorem unifies concepts of the (possibly nongeneric) solutions (3.7), (3.17), (3.24) and (3.47) of the TLS problem (2.3) established for different classes of problems independently on particular properties described in the previous sections.

Let (2.6) be the SVD of $[\,B\,|\,A\,]$ with the partitioning given by (3.2)–(3.3), $\Delta \equiv \kappa$, $0 \le \kappa \le n$, where κ is the smallest integer such that:

(i) the submatrix $V_{12}^{(\kappa)}$ is of full row rank, and

(ii) either $\sigma_{n-\kappa} > \sigma_{n-\kappa+1}$, or $\kappa = n$.

Since $V_{12}^{(\kappa)} \in \mathbb{R}^{d\times(\kappa+d)}$ is of full row rank, there always exists an orthogonal matrix $Q \in \mathbb{R}^{(\kappa+d)\times(\kappa+d)}$ such that

$$\left[\begin{array}{c} V_{12}^{(\kappa)} \\ \hline V_{22}^{(\kappa)} \end{array}\right] Q \,\equiv\, \left[\begin{array}{ccc} v_{n-\kappa+1}\,, & \cdots\,, & v_{n+d} \end{array}\right] Q \,=\, \left[\begin{array}{c|c} 0 & \Gamma \\ \hline Y & Z \end{array}\right]\,, \tag{3.50}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ is nonsingular. Define the correction matrix

$$\left[\begin{array}{c|c} G & E \end{array}\right] \,\equiv\, -\,\left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right] \left[\begin{array}{c} \Gamma \\ \hline Z \end{array}\right]^T\,. \tag{3.51}$$

And finally define the matrix

$$X^{(\kappa)} \equiv -Z\,\Gamma^{-1} = -V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger}\,. \tag{3.52}$$

Then $X^{(\kappa)}$, given by (3.52), is the solution of the compatible system $(A+E)\,X = B+G$, with E, G given by (3.51), minimal in the Frobenius norm. Matrices (3.51) and (3.52) are independent on the particular choice of Q in (3.50). Obviously the value of κ is either equal to q, if the problem is of the 1st class, or to $\tilde{q}$, if the problem is of the 2nd class.

The solution $X^{(\kappa)}$ given by (3.52) is identical to the solution computed by the *classical TLS algorithm* introduced by Van Huffel and Vandewalle, see [84, Algorithm 3.1, pp. 87–88]. A fully documented Fortran 77 program of this *classical TLS algorithm* is given in [83, 82]. (The code can be obtained through Netlib (cf. `http://www.netlib.org/vanhuffel/`).)

Algorithm 3.1 (classical TLS algorithm)**.**

00:	SET $j = 0$
01:	IF $\mathrm{rank}\,(V_{12}^{(j)}) = d$ AND $j = n$, THEN GOTO LINE 05
02:	IF $\mathrm{rank}\,(V_{12}^{(j)}) = d$ AND $\sigma_{n-j} > \sigma_{n-j+1}$, THEN GOTO LINE 05
03:	SET $j = j+1$
04:	GOTO LINE 01
05:	SET $\kappa = j$
06:	COMPUTE $X^{(\kappa)} = -V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger}$
07:	RETURN κ, $X^{(\kappa)}$

Summarizing, let (2.1) be an approximation problem, and $X^{(\kappa)}$ the solution given by (3.52), i.e., the solution returned by the classical TLS algorithm. If $\sigma_{n-\kappa+1} = \sigma_{n+1}$, then the problem is of the 1st class (in particular for $\kappa = 0$), and if $\sigma_{n-\kappa+1} > \sigma_{n+1}$, then the problem is of the 2nd class. In more details:

(i) If the problem is of 1st class and $\mathrm{rank}\,(W^{(q,e)}) = e$, then $X^{(\kappa)} \equiv X_{\mathrm{TLS}}$ represents the TLS solution (it solves the TLS problem (2.3)), $\kappa \equiv q$, the problem belongs to the set $\mathscr{F}_1$.

(ii) If the problem is of 1st class and $\mathrm{rank}\,(W^{(q,e)}) > e$, then $X^{(\kappa)} \equiv X_{\mathrm{NGN}}$ represents the nongeneric solution (solution of the constrained problem (3.42)), $\kappa \equiv q$, i.e., the problem belongs to the set $\mathscr{F}_2 \cup \mathscr{F}_3$.

(iii) If the problem is of 2nd class, then $X^{(\kappa)} \equiv X_{\mathrm{NGN}}$ represents the nongeneric solution (solution of the constrained problem (3.48)), $\kappa \equiv \tilde{q}$.

(Recall that the problems from the set $\mathscr{F}_1 \cup \mathscr{F}_2$ always have a TLS solution.) Figure 3.5 quickly recapitulates properties of problems and differences between problems in these individual classes.

As in the previous sections, the matrix $X^{(\kappa)} = -V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger}$ given by (3.52), i.e. the result of the classical TLS algorithm, represents the minimum norm solution of the compatible system $(A+E)\,X = B+G$, with E, G given by (3.51). This possibly nongeneric solution of the original TLS problem (2.3) has two further interpretations.

The following lemma summarizes the previous Lemmas 3.7 and 3.9 in a general form (independent on the particular assumptions).

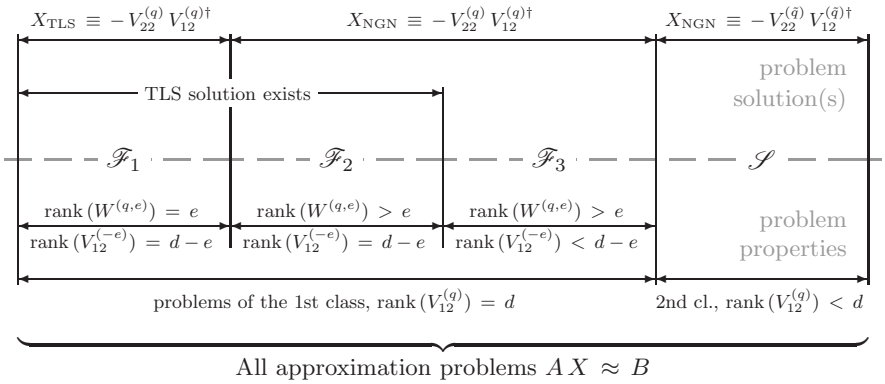


Figure 3.5: Properties of problems belonging to the sets $\mathscr{F}_1$, $\mathscr{F}_2$ and $\mathscr{F}_3$. The set $\mathscr{S}$ here denotes the set of problems of 2nd class.

Lemma 3.11. *The matrix* $X^{(\kappa)} = -V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger}$ *given by (3.52) represents the unique solution of the constrained minimization problem*

$$\begin{aligned} &\min_{X,E,G} \left\| \left[\; G \;\middle|\; E \;\right] \right\|_F \qquad \textit{subject to} \qquad (A+E)\,X = B+G \\ &\textit{and} \quad \left[\; G \;\middle|\; E \;\right] \left[\frac{0}{w}\right] = 0\,, \quad \forall\, \left[\frac{0}{w}\right] \in \mathcal{R}\left(\left[\frac{V_{12}^{(\kappa)}}{V_{22}^{(\kappa)}}\right]\right), \end{aligned} \tag{3.53}$$

with the corresponding correction $[\,G\,|\,E\,]$ *defined by (3.51).*

The additional constraint in (3.53) can be equivalently rewritten as $[\,G\,|\,E\,]\,[\frac{0}{Y}] = 0$ where Y is given in (3.50). Since $\sigma_{n-\kappa} > \sigma_{n-\kappa+1}$, the correction matrix in (3.53) is given uniquely. Consequently, the constrained problem (3.53) always has unique solution $X_{\mathrm{Const.}} \equiv X^{(\kappa)}$. See [84, Definition 3.3, p. 78 and Theorem 3.15, pp. 80–82], see also Lemmas 3.7 and 3.9.

Remark 3.5. *Since the matrix (3.50) has orthonormal columns,* $Z^T Y = 0$ *and consequently* $X^{(\kappa)T} Y = -(\Gamma^{-1})^T\,Z^T Y = 0$. *Because the constrained problem (3.53) has unique solution equal to* $X^{(\kappa)}$, *the additional constraint implies that*

$$X^T w = 0\,, \qquad \forall\, \left[\frac{0}{w}\right] \in \mathcal{R}\left(\left[\frac{V_{12}^{(\kappa)}}{V_{22}^{(\kappa)}}\right]\right).$$

See also [84, Eq. 3.101, p. 79]. Similarly in (3.42) and (3.48).

The following lemma summarizes the previous Lemmas 3.8 and 3.10 in a general form (independent on the particular assumptions).

Lemma 3.12. *The matrix* $X^{(\kappa)} = -V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger}$ *given by (3.52) represents the unique minimum norm TLS solution of the modified TLS problem*

$$\min_{X,\hat{E},\hat{G}} \left\| \left[\; \hat{G} \;\middle|\; \hat{E} \;\right] \right\|_F \qquad \textit{subject to} \qquad (\hat{A}+\hat{E})\,X = \hat{B}+\hat{G}\,, \tag{3.54}$$

where

$$\left[\; \hat{B} \;\middle|\; \hat{A} \;\right] = \left(\sum\nolimits_{j=1}^{n-\kappa} u_j\,\sigma_j\,v_j^T\right) + \sigma_{n-\kappa+1}\left(\sum\nolimits_{j=n-\kappa+1}^{n+d} u_j\,v_j^T\right),$$

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

Obviously the problem (3.54) is a TLS problem of the 1st class (from the set $\mathscr{F}_1$); moreover it is the special case problem described in Section 3.2.2. The problem is called *truncated total least squares (T-TLS) problem* for given A, B, with the solution $X_{\text{T-TLS}} \equiv X^{(\kappa)}$. See [84, note on p. 82], see also Lemmas 3.8 and 3.10.

It is worth to note that the T-TLS concept allows us to assume that the original problem $A\,X \approx B$ is a perturbation of the modified problem $\hat{A}\,X \approx \hat{B}$, or contrariwise. From the T-TLS point of view any problem may be interpreted as a perturbed problem of the 1st class with the special singular values distribution (3.11). This approach can be used as a relatively simple and useful regularization technique see, e.g., [88, 17] (for $d = 1$) and also [84, Algorithm and comments in §3.6.1, pp. 87–90].

On the other hand, it may be reasonable to expect that some information, which was originally contained in the problem $[\,B \,|\, A\,]$, is lost in the nongeneric solution, which may yield some troubles. See also Example 6.1 in Chapter 6 for illustration of such "loss-of-information".

Remark 3.6. *In whole this section κ is the smallest integer satisfying conditions (i) and (ii) in the beginning of this section, equivalently, the integer returned by Algorithm 3.1. However, in Lemmas 3.11 and 3.12 κ can be substituted by any integer $\tilde{\kappa}$ (not necessarily the smallest) satisfying both these conditions (i), (ii).*

Part III

Data Reduction

Chapter 4

SVD-based data reduction in $A\,X \approx B$

The SVD-based reduction given by C. C. Paige and Z. Strakoš in [64], described also in Chapter 1, yields for a problem with the single right-hand side the core problem.

In this chapter we extend such reduction to problems with multiple right-hand sides. We also investigate properties of the resulting reduced subproblem.

4.1 Introduction

Consider a general orthogonally invariant approximation problem (2.1), where, consistently with the right-hand side case, $A^T B \neq 0$. Since the problem is orthogonally invariant, it can be transformed into

$$(P^T A Q)(Q^T X R) \approx (P^T B R), \tag{4.1}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$. Equivalently (2.2) becomes

$$\left(P^T \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right]\right) \left(\left[\begin{array}{c|c} R^T & 0 \\ \hline 0 & Q^T \end{array}\right] \left[\begin{array}{c} -I_d \\ \hline X \end{array}\right] R\right) \approx 0\,. \tag{4.2}$$

In this chapter we construct a transformation of the form (4.1) which transforms the *original data* $[\,B\,|\,A\,]$ into the block form

$$\begin{aligned} P^T \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] &= \left[\begin{array}{c|c} P^T B R & P^T A Q \end{array}\right] \\ &\equiv \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \end{aligned} \tag{4.3}$$

where B_1 and A_{11} are of minimal dimensions and all irrelevant and redundant information in (4.3) is thus moved into the block A_{22}; the proof of minimality is given further in the text, in Chapter 6.

4.2 Algorithm of the reduction

The transformation is realized in four consequent steps:

(i) Preprocessing of the right-hand side. Here linearly dependent columns are removed from the right-hand side B, see Section 4.2.1.

(ii) Transformation of the system matrix. Here the matrix A is transformed to diagonal form, see Section 4.2.2.

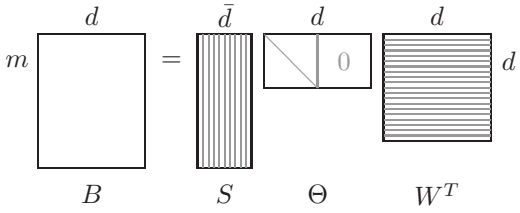
(iii) Transformation of the right-hand side. Here the right-hand side is transformed to introduce as many zeros as possible, see Section 4.2.3.

(iv) Final permutation. Here the rows and columns of the extended matrix are permuted to make the block structure visible, see Section 4.2.4.

4.2.1 Preprocessing of the right-hand side

Consider the SVD of B, $\bar{d} \equiv \operatorname{rank}(B) \leq \min\{m, d\}$,

$$B = S\,\Theta\,W^T, \qquad S \in \mathbb{R}^{m\times\bar{d}}, \quad \Theta \in \mathbb{R}^{\bar{d}\times d}, \quad W \in \mathbb{R}^{d\times d}, \tag{4.4}$$



where S has mutually orthonormal columns, Θ is of full row rank, and $W^{-1} = W^T$. In the case that $d > \bar{d} = \operatorname{rank}(B)$, the right-hand side B contains linearly dependent columns representing redundant information that can be removed from the problem (2.1).

Using the orthogonal matrix W the problem (2.1) transforms to

$$A\,\tilde{X} \approx \tilde{B}, \tag{4.5}$$

where $\tilde{X} \equiv X\,W$, $\tilde{B} \equiv B\,W = S\,\Theta$. Define

$$\begin{aligned} \tilde{B} &\equiv \left[\begin{array}{c|c} C & 0 \end{array}\right] \in \mathbb{R}^{m\times d}, \qquad C \in \mathbb{R}^{m\times\bar{d}}, \\ \tilde{X} &\equiv \left[\begin{array}{c|c} Y & Z \end{array}\right] \in \mathbb{R}^{n\times d}, \qquad Y \in \mathbb{R}^{n\times\bar{d}} \end{aligned} \tag{4.6}$$

(if $d = \bar{d}$, then $\tilde{B} \equiv C$, $\tilde{X} \equiv Y$, and Z is nonexistent). Equivalently (2.2) becomes

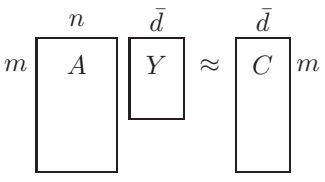
$$\begin{aligned} \left[\begin{array}{c|c} \tilde{B} & A \end{array}\right] \left[\begin{array}{c} -I_d \\ \hline \tilde{X} \end{array}\right] &= \left[\begin{array}{c|c} C & 0 \end{array}\right] \left\| \, A \right. \left[\begin{array}{c|c} -I_{\bar{d}} & 0 \\ \hline 0 & -I_{d-\bar{d}} \\ \hline\hline Y & Z \end{array}\right] \\ &= \left[\begin{array}{c|c} AY - C & AZ - 0 \end{array}\right] \approx 0\,. \end{aligned} \tag{4.7}$$

Thus the original problem is split into two subproblems

$$AY \approx C \qquad \text{and} \qquad AZ \approx 0\,.$$

As in the single right-hand side case, we consider $Z \equiv 0$. A nonzero Z would not increase the quality of approximation; it can only increase the norm of the solution. With $Z \equiv 0$, only the subproblem $AY \approx C$ has to be solved. For this reason the rest of the reduction algorithm considers the approximation problem

$$AY \approx C, \qquad A \in \mathbb{R}^{m\times n}, \quad Y \in \mathbb{R}^{n\times\bar{d}}, \quad C \in \mathbb{R}^{m\times\bar{d}}, \tag{4.8}$$



or, equivalently,

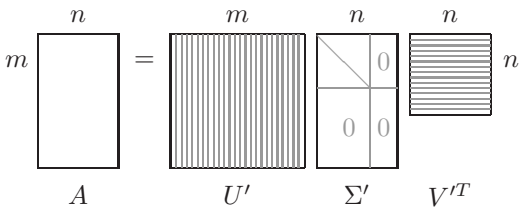
$$\left[\begin{array}{c|c} C & A \end{array}\right] \left[\begin{array}{c} -I_{\bar{d}} \\ \hline Y \end{array}\right] \approx 0\,, \tag{4.9}$$

where C is of full column rank. Moreover, as follows from (4.4), C has mutually orthogonal columns.

4.2.2 Transformation of the system matrix

In this section we use the SVD of A to transform the matrix to diagonal form. Consider the SVD (2.4) of A, $r \equiv \mathrm{rank}\,(A) \leq \min\{m, n\}$,

$$A = U' \Sigma' V'^T\,, \qquad U' \in \mathbb{R}^{m\times m}\,, \quad \Sigma' \in \mathbb{R}^{m\times n}\,, \quad V' \in \mathbb{R}^{n\times n}\,, \tag{4.10}$$



where $(U')^{-1} = (U')^T$, $(V')^{-1} = (V')^T$. Consider that A has k *distinct* nonzero singular values

$$\varsigma'_1 > \varsigma'_2 > \ldots > \varsigma'_k > 0\,, \tag{4.11}$$

with multiplicities m_j, $j = 1, \ldots, k$ (obviously $\sum_{j=1}^{k} m_j = r$), i.e.,

$$\Sigma' = \mathrm{diag}\,(\varsigma'_1 I_{m_1}, \varsigma'_2 I_{m_2}, \ldots, \varsigma'_k I_{m_k}, 0)\,.$$

Using orthogonal matrices U', V', the problem (4.8) transforms to

$$\Sigma' \tilde{Y} \approx \tilde{C}\,, \tag{4.12}$$

where $\tilde{Y} \equiv (V')^T Y$, $\tilde{C} \equiv (U')^T C$. Equivalently, (4.9) becomes

$$\begin{aligned} \left(U'^T \left[\begin{array}{c|c} C & A \end{array}\right] \left[\begin{array}{c|c} I_{\bar{d}} & 0 \\ \hline 0 & V' \end{array}\right] \right) \left(\left[\begin{array}{c|c} I_{\bar{d}} & 0 \\ \hline 0 & V'^T \end{array}\right] \left[\begin{array}{c} -I_{\bar{d}} \\ \hline Y \end{array}\right] \right) \\ = \left[\begin{array}{c|c} \tilde{C} & \Sigma' \end{array}\right] \left[\begin{array}{c} -I_{\bar{d}} \\ \hline \tilde{Y} \end{array}\right] \approx 0\,. \end{aligned} \tag{4.13}$$

The system matrix in (4.12)–(4.13) has diagonal form, possibly containing zero rows and columns.

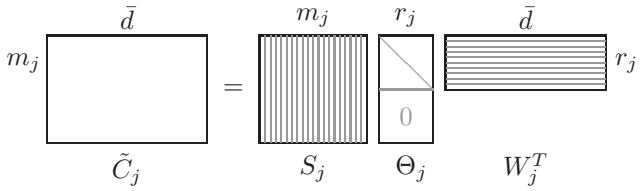
4.2.3 Transformation of the right-hand side

In order to obtain the required block structure (4.3) we seek to transform the right-hand side $\tilde{C}$, while maintaining the diagonal form of the matrix Σ'. Consider horizontal splitting of $\tilde{C}$ with respect to the multiplicities of the singular values of the system matrix A

$$\tilde{C} = \left[\begin{array}{c|c|c|c|c} \tilde{C}_1^T & \tilde{C}_2^T & \ldots & \tilde{C}_k^T & \tilde{C}_{k+1}^T \end{array}\right]^T\,,$$

where $\tilde{C}_j \in \mathbb{R}^{m_j \times \bar{d}}$, $j = 1, \ldots, k, k+1$, and $m_{k+1} \equiv m - r$ is the dimension of the null space $\mathcal{N}(A^T)$. For each $\tilde{C}_j$ denote $r_j \equiv \mathrm{rank}(\tilde{C}_j) \leq \min\{m_j, \bar{d}\}$ and consider the SVD

$$\tilde{C}_j = S_j \,\Theta_j \, W_j^T, \qquad S_j \in \mathbb{R}^{m_j \times m_j}, \quad \Theta_j \in \mathbb{R}^{m_j \times r_j}, \quad W_j \in \mathbb{R}^{\bar{d} \times r_j}, \tag{4.14}$$



where $S_j^{-1} = S_j^T$, Θ_j is of full column rank, and W_j has mutually orthonormal columns, $j = 1, \ldots, k, k+1$. Consider the orthogonal block diagonal matrices

$$\begin{aligned} G &\equiv \mathrm{diag}\,(S_1, S_2, \ldots, S_k, S_{k+1}) \in \mathbb{R}^{m \times m}, \\ H &\equiv \mathrm{diag}\,(S_1, S_2, \ldots, S_k, I_{n-r}) \in \mathbb{R}^{n \times n}. \end{aligned} \tag{4.15}$$

Using the orthogonal matrices G, H, the problem (4.12) transforms to

$$\Sigma'\,(H^T \tilde{Y}) \approx (G^T \tilde{C}), \tag{4.16}$$

because $\Sigma' = G^T \Sigma' H$. Equivalently, (4.13) becomes

$$\begin{aligned} \left(G^T \left[\, \tilde{C} \,\middle|\, \Sigma' \,\right] \left[\begin{array}{c|c} I_{\bar{d}} & 0 \\ \hline 0 & H \end{array}\right] \right) \left(\left[\begin{array}{c|c} I_{\bar{d}} & 0 \\ \hline 0 & H^T \end{array}\right] \left[\begin{array}{c} -I_{\bar{d}} \\ \hline \tilde{Y} \end{array}\right] \right) \\ = \left[\, G^T \tilde{C} \,\middle|\, \Sigma' \,\right] \left[\begin{array}{c} -I_{\bar{d}} \\ \hline H^T \tilde{Y} \end{array}\right] \approx 0\,. \end{aligned} \tag{4.17}$$

The extended system matrix from (4.17) has the form

$$\left[\, G^T \tilde{C} \,\middle|\, \Sigma' \,\right] = \left[\begin{array}{c|c|c|c|c||c} \Theta_1 W_1^T & \varsigma_1' I_{m_1} & & & & \\ \hline \Theta_2 W_2^T & & \varsigma_2' I_{m_2} & & & \\ \hline \vdots & & & \ddots & & 0 \\ \hline \Theta_k W_k^T & & & & \varsigma_k' I_{m_k} & \\ \hline\hline \Theta_{k+1} W_{k+1}^T & & & 0 & & 0 \end{array}\right]. \tag{4.18}$$

If $m_j > r_j$, for any $j = 1, \ldots, k, k+1$, then the block $S_j^T \tilde{C}_j = \Theta_j W_j^T$ contains zero rows, see (4.14). Therefore it is useful to denote

$$S_j^T \tilde{C}_j = \Theta_j W_j^T \equiv \left[\frac{D_j}{0}\right], \qquad D_j \in \mathbb{R}^{r_j \times \bar{d}}, \qquad j = 1, \ldots, k, k+1 \tag{4.19}$$

(if $r_j = m_j$, then $\Theta_j W_j^T \equiv D_j$; on the other hand, if $r_j = 0$, then the block D_j is nonexistent). In any case, D_j is of full row rank. Moreover, it has mutually orthogonal rows. Our aim is to remove the zero rows from (4.19) to the bottom part of the extended matrix (4.18), while maintaining the block diagonal form of the system matrix. This can be done by the following permutation.

4.2.4 Final permutation

Consider the permutation matrices

$\Pi_L \equiv$

$$\left[\begin{array}{c|c|c|c||c||c|c|c|c||c}
I_{r_1} & & & & & & & & & \\ \hline
 & & & & & I_{m_1-r_1} & & & & \\ \hline
 & I_{r_2} & & & & & & & & \\ \hline
 & & & & & & I_{m_2-r_2} & & & \\ \hline
 & & \ddots & & & & & & & \\ \hline
 & & & & & & & \ddots & & \\ \hline
 & & & I_{r_k} & & & & & & \\ \hline
 & & & & & & & & I_{m_k-r_k} & \\ \hline\hline
 & & & & I_{r_{k+1}} & & & & & \\ \hline
 & & & & & & & & & I_{m_{k+1}-r_{k+1}}
\end{array}\right] \tag{4.20}$$

and

$$\Pi_R \equiv \left[\begin{array}{c|c|c|c||c||c|c|c|c||c}
I_{r_1} & & & & & & & & & \\ \hline
 & & & & I_{m_1-r_1} & & & & & \\ \hline
 & I_{r_2} & & & & & & & & \\ \hline
 & & & & & I_{m_2-r_2} & & & & \\ \hline
 & & \ddots & & & & & & & \\ \hline
 & & & & & & \ddots & & & \\ \hline
 & & & I_{r_k} & & & & & & \\ \hline
 & & & & & & & I_{m_k-r_k} & & \\ \hline\hline
 & & & & & & & & I_{n-r} &
\end{array}\right] , \tag{4.21}$$

where $\Pi_L \in \mathbb{R}^{m\times m}$, $\Pi_R \in \mathbb{R}^{n\times n}$. If $r_j \equiv m_j$, then the block $I_{m_j-r_j}$ and the corresponding row and column are nonexistent; if $r_j = 0$, then the block I_{r_j} and the corresponding row and column are nonexistent, $j = 1, \ldots, k, k+1$.

Multiplication of the extended matrix (4.18) by the matrix Π_L^T from the left permutes the rows of the extended matrix such that the block rows beginning with D_j, see (4.18) and (4.19), are permuted up, while the block rows beginning with zero blocks are permuted down. Similarly, multiplication of the extended matrix (4.18) by the matrix $\mathrm{diag}\,(I_{\bar{d}}, \Pi_R)$ from the right permutes its block columns in such a way that the block columns corresponding to the block rows permuted down are now permuted to the right, while the other block columns are assembled to the

left. Finally, the following block structured extended matrix is obtained

$$\Pi_L^T \left[\begin{array}{c|c|c|c|c} \Theta_1 W_1^T & \varsigma_1' I_{m_1} & & & \\ \hline \vdots & & \ddots & & 0 \\ \hline \Theta_k W_k^T & & & \varsigma_k' I_{m_k} & \\ \hline\hline \Theta_{k+1} W_{k+1}^T & & 0 & & 0 \end{array}\right] \left[\begin{array}{c|c} I_{\bar{d}} & 0 \\ \hline 0 & \Pi_R \end{array}\right]$$

$$= \left[\begin{array}{c|c|c|c|c|c|c} D_1 & \varsigma_1' I_{r_1} & & & & & \\ \hline \vdots & & \ddots & & & 0 & 0 \\ \hline D_k & & & \varsigma_k' I_{r_k} & & & \\ \hline\hline D_{k+1} & & 0 & & & 0 & 0 \\ \hline\hline 0 & & & & \varsigma_1' I_{m_1-r_1} & & \\ \hline \vdots & & 0 & & & \ddots & 0 \\ \hline 0 & & & & & & \varsigma_k' I_{m_k-r_k} \\ \hline\hline 0 & & 0 & & & 0 & 0 \end{array}\right] \tag{4.22}$$

$$\equiv \left[\begin{array}{c|c|c} B_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right].$$

The described transformation can be summarized as follows.

4.3 Summary

Consider a general orthogonally invariant approximation problem (2.1). Further consider the SVD decompositions defined in (4.4), (4.10), (4.14), the orthogonal matrices G, H given by (4.15), and the permutation matrices Π_L, Π_R given by (4.20), (4.21). Denote

$$P \equiv U' G \Pi_L \qquad \text{and} \qquad Q \equiv V' H \Pi_R, \tag{4.23}$$

obviously $P^{-1} = P^T$, $Q^{-1} = Q^T$. Then the transformation

$$P^T \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c|c} W & 0 \\ \hline 0 & Q \end{array}\right] = \left[\begin{array}{c|c|c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right], \tag{4.24}$$

is obviously of the form (4.3). Matrices B_1, A_{11}, and A_{22} are given by (4.22), recall that

$$\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] = \left[\begin{array}{c|c|c|c} D_1 & \varsigma_1' I_{r_1} & & \\ \hline \vdots & & \ddots & \\ \hline D_k & & & \varsigma_k' I_{r_k} \\ \hline\hline D_{k+1} & & 0 & \end{array}\right] \in \mathbb{R}^{\bar{m} \times (\bar{n}+\bar{d})}, \tag{4.25}$$

where

$$\bar{m} \equiv \left(\sum\nolimits_{j=1}^{k+1} r_j\right), \qquad \bar{n} \equiv \left(\sum\nolimits_{j=1}^{k} r_j\right).$$

The corresponding transformation of the matrix of unknowns in (4.2) is given by

$$\left[\begin{array}{c|c} W & 0 \\ \hline 0 & Q \end{array}\right]^T \left[\begin{array}{c} -I_d \\ \hline X \end{array}\right] W = \left[\begin{array}{c|c} -I_{\bar{d}} & 0 \\ \hline 0 & -I_{d-\bar{d}} \\ \hline\hline X_1 & Z_1 \\ \hline X_2 & Z_2 \end{array}\right], \tag{4.26}$$

where the horizontal splitting $[\frac{X_1}{X_2}]$ and $[\frac{Z_1}{Z_2}]$ corresponds to the vertical splitting of the system matrix in (4.24). By insertion of (4.24) and (4.26) into the formula (4.2), the problem (2.2) is equivalently rewritten in the form

$$\left[\begin{array}{c|c} A_{11}\, X_1 \,-\, B_1 & A_{11}\, Z_1 \\ \hline A_{22}\, X_2 & A_{22}\, Z_2 \end{array}\right] \,\approx\, 0\,. \tag{4.27}$$

Analogously to the argumentation above, in the approximation problems

$$A_{11}\, Z_1 \,\approx\, 0\,, \qquad A_{22}\, X_2 \,\approx\, 0\,, \qquad A_{22}\, Z_2 \,\approx\, 0$$

it is natural to consider $X_2 \,\equiv\, 0$, $Z_1 \,\equiv\, 0$, $Z_2 \,\equiv\, 0$, and only the remaining subproblem, called *reduced problem*, has to be solved

$$A_{11}\, X_1 \,\approx\, B_1\,, \qquad A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}\,, \quad X_1 \in \mathbb{R}^{\bar{n}\times\bar{d}}\,, \quad B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}\,. \tag{4.28}$$

Equivalently,

$$\left[\;\, B_1 \;\middle|\; A_{11} \;\right] \left[\frac{-I_{\bar{d}}}{X_1}\right] \,\approx\, 0\,. \tag{4.29}$$

The solution of the original problem is given by

$$X \,\equiv\, Q \left[\begin{array}{c|c} X_1 & 0 \\ \hline 0 & 0 \end{array}\right] W^T\,, \tag{4.30}$$

and it is fully determined by the solution of the reduced problem. Consequently, the reduced problem contains all information sufficient for solving the original problem.

4.4 Properties of the subproblem $[\, B_1 \,|\, A_{11} \,]$

This section focuses on the properties of the problem (4.28), (4.29), with the extended matrix given by (4.25). Using (4.11), denote $U'_j \in \mathbb{R}^{\bar{m}\times r_j}$ the matrix of columns representing a basis of the left singular vectors subspace of A_{11} corresponding to ς'_j, $j \,=\, 1\,,\,\ldots\,,\,k$, and $U'_{k+1} \in \mathbb{R}^{\bar{m}\times r_{k+1}}$ the matrix of columns representing a basis of the null space of A_{11}^T (i.e. $\mathcal{R}\,(A_{11}) \,=\, \mathcal{R}\,(U'_1) \,\oplus\, \ldots \,\oplus\, \mathcal{R}\,(U'_k)$ and $\mathcal{N}\,(A_{11}^T) \,=\, \mathcal{R}\,(U'_{k+1})$). Obviously, the reduced problem has the following properties:

(G1) The matrix $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$ is of *full column rank* equal to $\bar{n} \,\leq\, \bar{m}$.

(G2) The matrix $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$ is of *full column rank* equal to $\bar{d} \,\leq\, \bar{m}$.

(G3) The matrices $(U'_j)^T\, B_1 \,\equiv\, D_j \in \mathbb{R}^{r_j\times\bar{d}}$ are of *full row rank* equal to $r_j \,\leq\, \bar{d}$, for $j \,=\, 1\,,\,\ldots\,,\,k+1$.

It can be easily proved that properties (G1)–(G3) imply:

(G4) The extended matrix $[\, B_1 \,|\, A_{11} \,] \in \mathbb{R}^{\bar{m}\times(\bar{n}+\bar{d})}$ is of *full row rank* equal to $\bar{m} \,\equiv\, \bar{n} + r_{k+1} \,\leq\, \bar{n} + \bar{d}$.

(G5) The matrix A_{11} does not have any zero singular value. Its singular values have multiplicities at most $\bar{d}$.

Because $\bar{m} - \bar{n} \,=\, r_{k+1} \,\leq\, \bar{d}$, the matrix A_{11} has at most $\bar{d}$ extra rows. Consequently the dimensions of the reduced problem satisfy the inequality

$$\max\{\,\bar{n}\,,\,\bar{d}\,\} \,\leq\, \bar{m} \,\leq\, \bar{n} + \bar{d}\,.$$

It can be simply verified that the properties (G1)–(G5) of the reduced problem are invariant with respect to an orthogonal transformation of the form

$$\hat{P}^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} \hat{R} & 0 \\ \hline 0 & \hat{Q} \end{array}\right] \equiv \left[\begin{array}{c|c} \hat{B}_1 & \hat{A}_{11} \end{array}\right], \tag{4.31}$$

where $\hat{P}^{-1} = \hat{P}^T$, $\hat{Q}^{-1} = \hat{Q}^T$, $\hat{R}^{-1} = \hat{R}^T$. Thus we call them *general properties*. Further, in Chapter 6, it will be shown that the dimensions of the reduced problem (4.25) are minimal (they can not be reduced more), and that (G1)–(G5) ensure this minimality of dimensions.

Remark 4.1. *Instead of the SVD preprocessing (4.4) of the right-hand side B (see Section 4.2.1) one can use the LQ decomposition (producing a matrix with m by $\bar{d}$ lower triangular block in the column echelon form and an orthogonal matrix). Similarly, in the transformation of the right-hand side (see Section 4.2.3) instead of the SVD decompositions (4.14) of $\tilde{C}_j$ one can use QR decompositions (producing a matrix with r_j by $\bar{d}$ upper triangular block in the row echelon form and an orthogonal matrix), $j = 1, \ldots, k+1$.*

It is not difficult to show that it this way the modified algorithm of reduction produces a subproblem having the same dimension as (4.25) which can be obtained from (4.25) by an orthogonal transformation of the form (4.31).

In contrast to the general properties, the subproblem $[\,B_1 \,|\, A_{11}\,]$ determined by the algorithm described above (in Sections 4.2, 4.3), see (4.25), has some properties which are not invariant with respect to the orthogonal transformation of the form (4.31). We call them *special properties*:

(S1) The matrix $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$ is diagonal with positive components ordered in nonincreasing sequence on the main diagonal.

(S2) The matrix $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$ has mutually orthogonal nonzero columns, ordered in a nonincreasing sequence with respect to their norms.

(S3) The matrices $(U_j')^T B_1 \equiv D_j \in \mathbb{R}^{r_j\times\bar{d}}$ have mutually orthogonal nonzero rows ordered in a nonincreasing sequence with respect to their norms, for $j = 1, \ldots, k+1$.

From (S1) and (S3) it follows that the blocks of rows $[\,D_j \,||\, 0 \ldots 0 \,|\, \varsigma_j' I_{r_j} \,|\, 0 \ldots 0\,]$ of $[\,B_1 \,|\, A_{11}\,]$, $j = 1, \ldots, k$, have mutually orthogonal rows.

Further, from the relationship between the solutions of reduced and original problems (4.30) it follows that the reduced problem contains all information sufficient for solving the original problem. (In Chapter 6, it will be shown that the problem contains only necessary information for solving the original problem. Note that for the problem with single right-hand side, i.e. $d = 1$, the reduction described above gives the core problem [64].)

Remark 4.2. *From (4.24) it follows that the SVD of A can be expressed by the SVDs of A_{11} and A_{22}, and, similarly, the SVD of $[\,B \,|\, A\,]$ can be expressed by the SVDs of $[\,B_1 \,|\, A_{11}\,]$ and A_{22}.*

Remark 4.3. *Note that $\bar{d}$ can be smaller, equal, or larger than $\bar{n}$. For example, $\bar{d} > \bar{n}$ in the reduced problem*

$$\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] = \left[\begin{array}{c|c|c|c} \clubsuit & \varsigma_1' & & \\ \hline & \clubsuit & \varsigma_2' & \\ \hline & \clubsuit & & \varsigma_3' \\ \hline & \clubsuit & & \\ & \clubsuit & & \\ & \clubsuit & & \end{array}\right], \tag{4.32}$$

where ♣ denotes nonzero components. An example with $\bar{d} = \bar{n}$ *can be obtained easily, e.g., from (4.32) by removing the last (4th) column of* B_1 *and the last (6th) row of* $[\, B_1 \,|\, A_{11} \,]$*. (The third possibility* $\bar{d} < \bar{n}$ *arises, e.g., in problems with single right-hand side b having nonzero projections onto at least two left singular vector spaces corresponding to distinct nonzero singular values of A, see [64].)*

Consequently, if $r_j < \bar{d}$*, then the matrix* $(U_j')^T B_1 \equiv D_j$ *has linearly dependent columns,* $j = 1, \ldots, k+1$*. All the matrices* D_j*,* $j = 1, \ldots, k+1$*, may have linearly dependent columns, see (4.32).*

Chapter 5

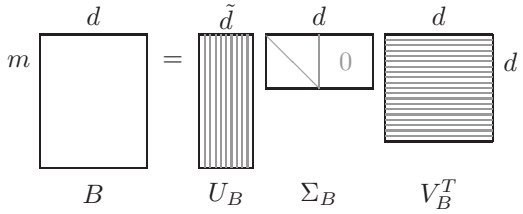
Band generalization of the Golub-Kahan bidiagonalization

The Golub-Kahan bidiagonalization algorithm and its relationship to the Lanczos process and Jacobi matrices plays an important role in the core problem theory for the single right-hand sides problems. Here we present and develop further a generalization of the Golub-Kahan algorithm, proposed for the purpose of the reduction to the core problem by Å. Björck, [8, 9, 10], D. M. Sima and S. Van Huffel, [73, 74].

A relationship between the generalized Golub-Kahan algorithm and the band formulation of the block Lanczos process allows us to analyze properties of the reduced problem. This chapter introduces so called generalized Jacobi matrices, which are useful for analysis of the reduction process. The whole chapter assumes exact arithmetic.

5.1 Introduction

Consider a general orthogonally invariant approximation problem (2.1), where, consistently with the single right-hand side case, $A^T B \neq 0$. With no loss of generality, see Section 4.2.1, we assume that B is of full column rank. If the right-hand side B does not have full column rank, then use the SVD $B = U_B \Sigma_B V_B^T$ of the form (4.4), to obtain the equivalent problem with the full column rank right-hand side $\tilde{B} \in \mathbb{R}^{m \times \tilde{d}}$, $[\,\tilde{B}\,|\,0\,] \equiv U_B \Sigma_B$, $\tilde{d} \equiv \operatorname{rank}(B)$, see also (4.4), (4.8) in Section 4.2.1.



The *SVD-preprocessed right-hand side* matrix $\tilde{B}$ has, moreover, mutually orthogonal columns sorted in such a way that their norm is nonincreasing. The preprocessing may be suitable also in the case when B is of full column rank. When started with such SVD-preprocessed right-hand side, the band algorithm, described in the further text in this chapter, yields a reduced problem with the right-hand side having

nonzero components only on the main diagonal. Alternatively, the LQ decomposition $B = [\,\hat{B}\,|\,0\,]\,\hat{Q}^T$, $\hat{B} \in \mathbb{R}^{m\times\tilde{d}}$ yields an equivalent problem with a full column rank right-hand side $\hat{B}$. The *LQ-preprocessed right-hand side* $\hat{B}$ is in the lower triangular (column echelon) form.

Remark 5.1. *In this section we frequently use the following symbolic notation for components of matrices:*

(i) ♣ *denotes a nonzero component,* ♣ $\neq 0$,

(ii) ♡ *denotes a component which may be zero as well as nonzero,*

(iii) *all other entries are equal to zero.*

5.2 Description of the band reduction algorithm

The band generalization of the partial Golub-Kahan iterative bidiagonalization algorithm, which is described in this section, and its application for the data decomposition in approximation problems $A\,X \approx B$ with multiple right-hand sides was first proposed and published in a series of presentations by Åke Björck [8, 9, 10], and by Diana M. Sima in [73], [74, Section 2.3.3, pp. 31–39]. Here we successively derive the algorithm, for the complete description of the algorithm see also the PhD thesis of Diana M. Sima [74, Algorithm 2.4, p. 38].

Assume the right-hand side $B \in \mathbb{R}^{m\times d}$ of full column rank (i.e. $m \geq d$). In order to simplify the notation, in the rest of this chapter d denotes the number of columns of B as well as the rank of B, according to the previous assumption. Consider the QR decomposition of B in the form

$$B = Q\,R = \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \left[\frac{R_1}{0}\right] = Q_1\,R_1\,, \tag{5.1}$$

where $Q \in \mathbb{R}^{m\times m}$ is an orthogonal matrix; if $m = d$ Q_2 is nonexistent. Here

$$R_1 = \begin{bmatrix} \gamma_1 & \beta_{1,2} & \beta_{1,3} & \cdots & \beta_{1,d} \\ & \gamma_2 & \beta_{2,3} & \cdots & \beta_{2,d} \\ & & \ddots & \ddots & \vdots \\ & & & \gamma_{d-1} & \beta_{d-1,d} \\ & & & & \gamma_d \end{bmatrix} \in \mathbb{R}^{d\times d}, \tag{5.2}$$

is a square upper triangular matrix with a positive diagonal, $\gamma_j > 0$, $j = 1\,,\,\ldots\,,\,d$. Clearly, for the SVD-preprocessed right-hand side $\tilde{B} \equiv U_B\,\Sigma_B$ the matrix R_1 is diagonal with nonzero singular values of B on the main diagonal, and $Q_1 \equiv U_B$ contains the corresponding left singular vectors.

We will orthogonally transform the extended matrix $[\,B\,|\,A\,]$ into the matrix with a banded left upper triangular block, such that

$$S_k^T \left[\begin{array}{c|c} B & A\,W_k \end{array}\right] \equiv \left[\begin{array}{c|c} R & L_k \end{array}\right] \tag{5.3}$$

where $S_k^{-1} = S_k^T$, $W_k^{-1} = W_k^T$,

$$L_k \equiv \left[\begin{array}{ccccc|ccc}
\alpha_1 & & & & & & & \\
\beta_{2,d+1} & \alpha_2 & & & & & & \\
\vdots & \beta_{3,d+2} & \alpha_3 & & & & & \\
\beta_{d,d+1} & \vdots & \beta_{4,d+3} & \ddots & & & & \\
\gamma_{d+1} & \beta_{d+1,d+2} & \vdots & \ddots & \alpha_k & & & \\
 & \gamma_{d+2} & \beta_{d+2,d+3} & & \beta_{k+1,d+k} & \heartsuit & \cdots & \heartsuit \\
 & & \gamma_{d+3} & \ddots & \vdots & \heartsuit & \cdots & \heartsuit \\
 & & & \ddots & \beta_{d+k-1,d+k} & \vdots & & \vdots \\
 & & & & \gamma_{d+k} & \heartsuit & \cdots & \heartsuit \\
\hline
 & & & & & \heartsuit & \cdots & \heartsuit \\
 & & & & & \vdots & \ddots & \vdots \\
 & & & & & \heartsuit & \cdots & \heartsuit
\end{array}\right], \quad (5.4)$$

and for the given k, $0 \le k \le \min\{ n , m-d \}$, the components α_j and γ_{d+j} are positive, i.e., $\alpha_j > 0$ and $\gamma_{d+j} > 0$, $j = 1 , \ldots , k$, if $k > 0$. Here we assume the existence of $k \ge 1$ with the given property, and derive the iterative representation of the partial decomposition above. The situation $k = 0$ will be discussed later. for some k, $0 \le k \le \min\{ n , m-d \}$. The orthogonal matrices $S_k \in \mathbb{R}^{m\times m}$ and $W_k \in \mathbb{R}^{n\times n}$ can be determined, e.g., as products of elementary Householder transformation matrices. Obviously, the first d columns of the matrices S_k, from (5.3)–(5.4), are given by Q_1 from the QR decomposition (5.1)–(5.2). Consequently, $S_k = Q \operatorname{diag}(I_d , \hat{S}_k) = [Q_1 \,|\, Q_2 \hat{S}_k]$, where $\hat{S}_k \in \mathbb{R}^{(m-d)\times(m-d)}$ is orthogonal.

Denote s_j and w_j the jth columns of the matrices S_k and W_k, respectively. Since $S_k^T A W_k = L_k$ where S_k, W_k are orthogonal matrices, equating the corresponding columns in $A W_k = S_k L_k$ and $A^T S_k = W_k L_k^T$ gives

$$A w_j = \left[\, s_j , s_{j+1} , \ldots , s_{j+d-1} , s_{j+d} \, \right] \left[\begin{array}{c} \alpha_j \\ \beta_{j+1,d+j} \\ \vdots \\ \beta_{j+d-1,d+j} \\ \gamma_{d+j} \end{array}\right],$$

and

$$A^T s_j = \left[\, w_{j-d} , w_{j-d+1} , \ldots , w_{j-1} , w_j \, \right] \left[\begin{array}{c} \gamma_j \\ \beta_{j,j+1} \\ \vdots \\ \beta_{j,j+d-1} \\ \alpha_j \end{array}\right],$$

for $j = 1 , 2 , \ldots , k$, where, for convenience, we set $w_j \equiv 0$ for $j \le 0$. Rearranging the previous equations gives, together with the orthogonality relations $S_k^T S_k = I_m$, $W_k^T W_k = I_n$, the recurrence formulas for computing the principal banded $(d+k)$ by k block of L_k, and the columns s_{d+j}, w_j, $j = 1 , 2 , \ldots , k$, of S_k and W_k, respectively. Given the vectors $s_1 , \ldots , s_d$ (columns of Q_1) and the matrix R, see

(5.1)–(5.2), $w_{-d+1} = \ldots = w_0 \equiv 0$, for $j = 1, 2, \ldots, k$

$$w_j \alpha_j \equiv A^T s_j - w_{j-d} \gamma_j - \left(\sum_{i=1}^{d-1} w_{j-d+i} \beta_{j,j+i} \right), \tag{5.5}$$

$$\text{for} \quad i = 1, \ldots, d-1, \qquad \beta_{j+i,d+j} \equiv s_{j+i}^T A w_j, \tag{5.6}$$

$$s_{d+j} \gamma_{d+j} \equiv A w_j - s_j \alpha_j - \left(\sum_{i=1}^{d-1} s_{j+i} \beta_{j+i,d+j} \right). \tag{5.7}$$

Coefficients α_j and γ_{j+d} are determined by the normalization conditions $\|w_j\| = 1$ and $\|s_{j+d}\| = 1$, respectively, and coefficients $\beta_{j+i,j+d}$ are determined by the orthogonality relations $s_{d+j}^T s_{j+i} = 0$, $i = 1, \ldots, d-1$, for $j = 1, 2, \ldots, k$.

Note that the described algorithm (5.5)–(5.7) produces the principal upper left block of L_k columnwisely. For $d = 1$ (in the single right-hand side case) the relations (5.6) vanish, and the algorithm (5.5)–(5.7) reduces to the Golub-Kahan iterative bidiagonalization.

Now we focus on the situation with $k = 0$. This happens in two different cases, either α_1 or γ_{d+1} (or both) in (5.4) are equal to zero; possibly, if $m = d$, then γ_{d+1} does not exist. Recall that the QR decomposition (5.1)–(5.2) of the right-hand side B transforms the original problem $[B \,|\, A]$ to the new problem $Q^T [B \,|\, A] = [R \,|\, Q^T A]$.

The first case is provided by the fact that the first row of the matrix $Q^T A$ is equal to zero, i.e. $e_1^T Q^T A = 0$. Then we try to construct a Householder transformation matrix producing positive α_1 component in the second row. As before, such transformation exists only if $e_2^T Q^T A \neq 0$, otherwise we try to construct a transformation producing positive α_1 in the third row, etc. The assumption $A^T B = A^T Q R = (Q^T A)^T R \neq 0$, where $R = [R_1^T \,|\, 0]^T$ and $R_1 \in \mathbb{R}^{d \times d}$, guarantees that at least one of $e_j^T Q^T A \neq 0$ for $j = 1, \ldots, d$ (at least one of the first d rows of $Q^T A$ is nonzero). Then the α_1 component corresponds to the first nonzero row in the matrix $Q^T A$.

The second case, $\gamma_{d+1} = 0$ in (5.4), is analogous. If $m > d$, then the second case is provided by the fact that the first column of $Q^T A$ ends with a zero subvector of length $m - d$. Then we try to construct a Householder transformation matrix producing positive γ_{d+1} in the second column of $Q^T A$. Contrary to the first case the matrix $Q^T A$ may not contain a column which ends with nonzero $m - d$ subvector and thus γ_{d+1} may not exists. If $m = d$, then the situation is obvious.

In all above described cases providing the situation with $k = 0$ can be obtained formulas analogous to (5.5)–(5.7) with shorter recurrences, by equating the corresponding columns. For a further description see the next section which discusses this phenomenon in more generally.

5.2.1 Deflation

Let occur the unfortunate situation when the matrix (5.4) does not exist, i.e. $k = 0$; or let l be the first index such that (5.5) or (5.7) becomes $w_l \alpha_l \equiv 0$ or $s_{d+l} \gamma_{d+l} \equiv 0$, respectively. Then the algorithm described by equations (5.5)–(5.7) applied on A breaks down in the 1st (if $k = 0$), or lth iteration. This occurs, e.g., for (5.7) when the number of rows of the matrix A is reached, i.e. $l = m - d + 1$ (generally it may occur sooner). (Note that the number of columns of A can not be reached because the algorithm (5.5)–(5.7) is designed columnwisely.)

The equation (5.5) reduces to $w_l \alpha_l = 0$ if all entries of the lth row of L_{l-1} to the right from $\beta_{l,d+l-1}$ are zero. Then obviously, there is a Householder transformation

matrix H_l such that

$$S_{l-1}^T \left[\begin{array}{c|c} B & A\,W_{l-1}\,H_l \end{array} \right] = \left[\begin{array}{ccc|ccc} \ddots & & & & & \\ \ddots & \alpha_{l-1} & & & & \\ \ddots & \beta_{l,d+l-1} & & 0 & \cdots & 0 \\ & \beta_{l+1,d+l-1} & & \alpha_l & \cdots & 0 \\ & \vdots & & \heartsuit & \cdots & \heartsuit \\ & & & \vdots & & \vdots \end{array} \right], \tag{5.8}$$

and we speak about *upper deflation*. For *lower deflation* the situation is analogous. Clearly, the upper and lower deflations can occur repeatedly. In the finite precision computations this phenomenon is called *exact deflation*, see, e.g., [1]. Each deflation reduces the *current width of the band* by one. After d deflations it is reduced to one, and the process terminates.

Consequently the orthogonal matrices $S \in \mathbb{R}^{m\times m}$ and $W \in \mathbb{R}^{n\times n}$ can be determined, e.g., as products of elementary Householder transformation matrices such that, $S^T\,[\,B\,|\,A\,W\,] \equiv [\,R\,|\,L\,]$, $S^T\,A\,W \equiv L$ have block diagonal form with a leading banded block. For $d = 3$ this is illustrated on the following scheme:

$$S^T \left[\begin{array}{c|c} B & A\,W \end{array} \right] \equiv \left[\begin{array}{c|c|c} \tilde{B}_1 & \tilde{A}_{11} & 0 \\ \hline 0 & 0 & \tilde{A}_{22} \end{array} \right] \tag{5.9}$$

$$= \left[\begin{array}{ccc|ccccc|ccc} \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & & & & & \\ & \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & & & & \\ & & \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & & & \\ & & & \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & & \\ & & & & 0 & \clubsuit & \heartsuit & \clubsuit & & & \\ & & & & & \clubsuit & \heartsuit & 0 & & & \\ & & & & & & \clubsuit & \clubsuit & & & \\ & & & & & & & \clubsuit & & & \\ \hline & & & & & & & 0 & \heartsuit & \cdots & \heartsuit \\ & & & & & & & & \vdots & \ddots & \vdots \\ & & & & & & & & \heartsuit & \cdots & \heartsuit \end{array} \right] .$$

Thus this transformation reveals a *banded subproblem* with the extended system matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$.

Now we describe how to include deflations in (5.5)–(5.7). Obviously, deflation reduces the lengths of the recurrences in (5.5), (5.7), and the number of computed coefficients in (5.6), see also [9, 10, 74].

The upper deflations can be handled simply. Let l be the first index for which the deflation occurs, and let it occur as upper deflation. Then the formula for computing α_l and w_l is given by equating the $(l+1)$st (instead of the lth) columns of $A^T\,S = W\,L^T$. Formulas (5.5)–(5.7) are modified for $j = l\,,\,\ldots\,,$ *until another deflation occurs* such that

$$w_j\,\alpha_j \equiv A^T\,s_{j+1} - w_{j-d+1}\,\gamma_j - \left(\sum_{i=2}^{d-1} w_{j-d+i}\,\beta_{j+1,j+i} \right), \tag{5.10}$$

$$\text{for} \quad i = 2\,,\ldots\,,\,d-1\,, \qquad \beta_{j+i,d+j} \equiv s_{j+i}^T\,A\,w_j\,, \tag{5.11}$$

$$s_{d+j}\,\gamma_{d+j} \equiv A\,w_j - s_{j+1}\,\alpha_j - \left(\sum_{i=2}^{d-1} s_{j+i}\,\beta_{j+i,d+j} \right). \tag{5.12}$$

Consequently, upper deflation in the lth step reduces by one the length of both recurrences and the number of computed coefficients in the algorithm for all $j \geq l$. Any other upper deflation requires analogous modification – decreasing of the length of recurrences.

Now, let l be the first index for which the deflation occurs, and let it occurs as lower deflation. Then the formula for computing γ_{d+l} and s_{d+l} is given by equating the $(l+1)$st (instead of the lth) columns of $AW = SL$. Unfortunately, here the modification of the recurrent formulas (5.5)–(5.7) is more complicated. The lower deflation in the lth step reduces by one the number of computed coefficients in (5.6) and the length of the recurrence in (5.7), for $j \geq l+1$. The length of the recurrence in (5.5) is reduced later – if another deflation does not occur in the meanwhile, then it is reduced after d steps, for $j \geq d+l$. This is a consequence of the fact that the algorithm is designed columnwisely (the length of the recurrence in (5.5) for computing α_j depends on the number of β components in the corresponding row).

Denote $S_{1:j} \equiv [\, s_1 \,,\, s_2 \,,\, \ldots \,,\, s_j \,]$, $W_{1:j} \equiv [\, w_1 \,,\, w_2 \,,\, \ldots \,,\, w_j \,]$, and $l_{ij} \equiv e_i^T \, L \, e_j$. The banded algorithm written in pseudocode follows; see also [74, Algorithm 2.4, p. 38].

Algorithm 5.1 (Band algorithm)**.**

00:	COMPUTE QR DECOMPOSITION $B = Q_1 \, R_1$
01:	SET $S_{1:d} = Q_1$, $W_\emptyset = []$, $L = []$
02:	SET $j = 1$, $k = d$, $c_U = 0$, $c_L = 0$
	/* COMPUTATION OF α , w , HANDLING THE UPPER DEFLATION */
03:	SET $p = A^T \, s_{j+c_U} - (\sum_{i=1}^{j-1} w_i \, l_{j+c_U,i})$
04:	IF $p = 0$, THEN SET $c_U = c_U + 1$, GOTO LINE 14
05:	SET $\alpha_j = \|p\|$, $w_j = p \,/\, \alpha_j$, $W_{1:j} = [\, W_{1:j-1} \,|\, w_j \,]$
	/* COMPUTATION OF β , INNER ORTHOGONALIZATION */
06:	FOR $i = j + 1 + c_U : j + d - 1 - c_L$, SET $\beta_{i,j+d} = s_i^T \, A \, w_j$
07:	SET $L = [\, L \,|\, [\, 0_{1,j-1+c_U} \,,\, \alpha_j \,,\, \beta_{j+1+c_U,j+d} \,,\, \ldots \,,\, \beta_{j+d-1-c_L,j+d} \,]^T \,]$
	/* COMPUTATION OF γ , s , HANDLING THE LOWER DEFLATION */
08:	SET $p = A \, w_j - (\sum_{i=1}^{k} s_i \, l_{i,j})$
09:	IF $p = 0$, THEN SET $c_L = c_L + 1$, GOTO LINE 13
10:	SET $k = k + 1$
11:	SET $\gamma_k = \|p\|$, $s_k = p \,/\, \gamma_k$, $S_{1:k} = [\, S_{1:k-1} \,|\, s_k \,]$
12:	SET $L = [\, L^T \,|\, [\, 0_{1,j-1} \,,\, \gamma_k \,]^T \,]^T$
13:	SET $j = j + 1$
14:	IF $c_U + c_L < d$, THEN GOTO LINE 03
15:	SET $\tilde{m} = k$, $\tilde{n} = j - 1$, $\tilde{B}_1 = [\, R_1^T \,|\, 0_{d,\tilde{m}-d} \,]^T$, $\tilde{A}_{11} = L$
16:	RETURN $S_{1:\tilde{m}}$, $W_{1:\tilde{n}}$, $\tilde{B}_1$, $\tilde{A}_{11}$

This algorithm stops if $c_U + c_L = d$ (line 14), where c_U and c_L are counters of upper and lower deflations, respectively. Thus it stops immediately after the dth

deflation. If j (the number of columns) or k (the number of rows) of the computed matrix L becomes equal to n or m, respectively, sooner than d deflations occur, then the algorithm stays in the cycle [03, 04, 14, 03, ...], *or* [03:09, 13, 14, 03:09, ...], *respectively, until $c_U + c_L < d$ (line 14).*

Remark 5.2. *Because it is difficult to describe the length of the recurrences in the Band algorithm 5.1, as in (5.5)–(5.7) or (5.10)–(5.12), in a simple way (see the discussion about lower deflation above) we use all the previously computed vectors in the description, with understanding that the matrix L has zero entries outside the band.*

The Algorithm 5.1 returns the following rectangular matrices: $S_{1:\tilde{m}} \in \mathbb{R}^{m\times\tilde{m}}$, $S_{1:\tilde{m}}^T S_{1:\tilde{m}} = I_{\tilde{m}}$, where $\tilde{m} > d$ and $S_{1:d} \equiv Q_1$ represents the first d columns of Q, see (5.1); $W_{1:\tilde{n}} \in \mathbb{R}^{n\times\tilde{n}}$, $W_{1:\tilde{n}}^T W_{1:\tilde{n}} = I_{\tilde{n}}$; $\tilde{B}_1$ containing the upper triangular block R_1, see (5.1)–(5.2), and the banded matrix $\tilde{A}_{11}$, such that

$$S_{1:\tilde{m}}^T \left[\; B \;\middle|\; A\,W_{1:\tilde{n}} \;\right] = \left[\; \tilde{B}_1 \;\middle|\; \tilde{A}_{11} \;\right] \in \mathbb{R}^{\tilde{m}\times(\tilde{n}+d)}\,, \tag{5.13}$$

with $\alpha_i > 0$, $i = 1\,,\ldots,\tilde{n}$, and $\gamma_j > 0$, $j = 1\,,\ldots,\tilde{m}$, see (5.2), (5.3), (5.5)–(5.7), etc. As an illustration see the following reduced problem:

$$\left[\; \tilde{B}_1 \;\middle|\; \tilde{A}_{11} \;\right] = \left[\begin{array}{ccc||ccccccc} \gamma_1 & \beta_{12} & \beta_{13} & \alpha_1 & & & & & & \\ & \gamma_2 & \beta_{23} & \beta_{24} & \alpha_2 & & & & & \\ & & \gamma_3 & \beta_{34} & \beta_{35} & \alpha_3 & & & & \\ & & & \gamma_4 & \beta_{45} & \beta_{46} & \alpha_4 & & & \\ & & & & & \gamma_5 & \beta_{57} & \alpha_5 & & \\ & & & & & & \gamma_6 & \beta_{68} & & \\ & & & & & & & \gamma_7 & \alpha_6 & \\ & & & & & & & & \gamma_8 & \alpha_7 \end{array}\right].$$

Remark 5.3 (Golub-Kahan bidiagonalization)**.** *For $d = 1$, Algorithm 5.1 reduces to the (partial) Golub-Kahan iterative bidiagonalization algorithm given in [27], see also [57, 59, 41, 42]. (The adjective partial means that it stops when the first bidiagonal component is equal to zero.)*

Remark 5.4 (Block Lanczos algorithm)**.** *Let the Band algorithm 5.1 applied on A started from the full column rank B return $S_{1:\tilde{m}}$, $W_{1:\tilde{n}}$, and $\tilde{B}_1$, $\tilde{A}_{11}$. Obviously $\tilde{B}_1$ is fully determined by the QR decomposition (5.2). A suitable banded formulation of the block Lanczos algorithm, see [69, 13, 80, 30], applied to $A\,A^T$ started with the orthonormal basis Q_1 of B (5.2) yields*

$$S_{1:\tilde{m}}^T\,(\,A\,A^T\,)\,S_{1:\tilde{m}} \;=\; \tilde{A}_{11}\,\tilde{A}_{11}^T\,.$$

Similarly the block Lanczos algorithm applied to $A^T\,A$ started with an orthonormal basis of $A^T\,B$ yields

$$W_{1:\tilde{n}}^T\,(\,A^T\,A\,)\,W_{1:\tilde{n}} \;=\; \tilde{A}_{11}^T\,\tilde{A}_{11}\,.$$

Thus Algorithm 5.1 is fully determined by the block Lanczos applied simultaneously to $A\,A^T$ and $A^T\,A$ with starting vectors the orthonormal bases of B, $A^T\,B$, respectively; or one applied to the larger symmetric matrix $\left[\begin{smallmatrix} 0 & A \\ A^T & 0 \end{smallmatrix}\right]$ with starting vectors the orthonormal basis of $\left[\begin{smallmatrix} B \\ 0 \end{smallmatrix}\right]$.

For the banded formulation of the block Lanczos algorithm called band Lanczos algorithm see [69], [2, Section 4.6 and 7.10 (R. W. Freund, Band Lanczos Method)]. For some numerical aspects and implementation details of band Lanczos type methods (inexact deflation etc.) see [1].

Remark 5.5 (Numerical aspects)**.** *In practical computations the newly computed vectors w_j and s_{j+d} must be twice reorthogonalized against the corresponding previously computed vectors. It ensures that the lost of orthogonality among computed vectors, i.e. $\| W_{1:\tilde{n}}^T W_{1:\tilde{n}} - I_{\tilde{n}} \|$, $\| S_{1:\tilde{m}}^T S_{1:\tilde{m}} - I_{\tilde{m}} \|$, is comparable with the machine precision. This follows from results for the Gram-Schmidt orthogonalization process, see [22, 23].*

Further, in an applicable implementation the decision when to deflate (lines 04, 09 *in Algorithm 5.1) can not be based on the exact condition $p = 0$, see [20, 74].*

As already mentioned, Algorithm 5.1 performs a problem transformation of the form (4.24). The algorithm yields the subproblem

$$\tilde{A}_{11}\,\tilde{X}_1 \;\approx\; \tilde{B}_1\,, \qquad \tilde{A}_{11} \in \mathbb{R}^{\tilde{m}\times\tilde{n}}\,, \quad \tilde{X}_1 \in \mathbb{R}^{\tilde{n}\times d}\,, \quad \tilde{B}_1 \in \mathbb{R}^{\tilde{m}\times d}\,, \tag{5.14}$$

or, equivalently,

$$\left[\;\tilde{B}_1 \;\middle|\; \tilde{A}_{11}\;\right] \left[\frac{-I_d}{\tilde{X}_1}\right] \;\approx\; 0\,, \tag{5.15}$$

which contains all information sufficient for solving the original problem, analogously to the argumentation in the previous chapter, see Section 4.3. The solution of the original problem is given by

$$X \;\equiv\; W_{1:\tilde{n}}\,\tilde{X}_1\,.$$

Properties of the subproblem (5.14) will be discussed in the several next sections.

Remark 5.6. *Note that the Krylov subspace methods and Lanczos-type algorithms are often used for model reduction problems. A very instructive example of a similar approach can be found in a model reduction application in modeling of electrical circuits, see, e.g., [20, Chapter 5].*

5.3 Basic properties of the subproblem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$

In the rest of this chapter we prove that the reduced problem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ obtained by the band algorithm has properties (G1)–(G5), see Section 4.4.

For a given B, A (assuming with no loss of generality the full column rank of B) Algorithm 5.1 produces the reduced problem $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$, see (5.9), (5.13), and (5.14), with the structure illustrated by the example

$$\left[\;\tilde{B}_1 \;\middle|\; \tilde{A}_{11}\;\right] = \left[\begin{array}{ccc||ccccccc} \gamma_1 & \beta_{12} & \beta_{13} & \alpha_1 & & & & & & \\ & \gamma_2 & \beta_{23} & \beta_{24} & \alpha_2 & & & & & \\ & & \gamma_3 & \beta_{34} & \beta_{35} & \alpha_3 & & & & \\ & & & \gamma_4 & \beta_{45} & \beta_{46} & \alpha_4 & & & \\ & & & & & \gamma_5 & \beta_{57} & \alpha_5 & & \\ & & & & & & \gamma_6 & \beta_{68} & & \\ & & & & & & & \gamma_7 & \alpha_6 & \\ & & & & & & & & \gamma_8 & \alpha_7 \end{array}\right], \tag{5.16}$$

where all $\alpha_j > 0$, $\gamma_j > 0$. Obviously, the reduced problem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ has the following properties:

(G1) The matrix $\tilde{A}_{11}$ is of full column rank, $\mathrm{rank}\,(\tilde{A}_{11}) = \tilde{n} \le \tilde{m}$.

(G2) The matrix $\tilde{B}_1$ is of full column rank (here by assumption).

(G4) The matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is of full row rank equal to $\tilde{m}$.

The full column rank of $\tilde{B}_1$ follows from the assumption that the right-hand side B is of full column rank (in the general case the right-hand side must be preprocessed, see Section 5.1 (or Section 4.2.1)). The other two assertions follow immediately from the fact that the matrix $\tilde{A}_{11}$ is in the lower triangular (column echelon) form with nonzero columns, and the matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is in the upper triangular (row echelon) form with nonzero rows.

Moreover, the vectors s_j and w_j computed by Algorithm 5.1 are the first columns of the matrices S and W from (5.9), respectively, it follows from the relationship to the block Lanczos algorithm see [28], see also [8, 9, 10]. By extending these sets of vectors to full bases, we get the orthogonal matrices $S \equiv [\,S_{1:\tilde{m}} \,|\, s_{\tilde{m}+1}\,,\,\ldots\,,\,s_m\,]$ and $W \equiv [\,W_{1:\tilde{n}} \,|\, w_{\tilde{n}+1}\,,\,\ldots\,,\,w_n\,]$ satisfying (5.9). We see that

$$(\,S_{1:\tilde{m}}^T\,A\,)\,W \;=\; \left[\;\tilde{A}_{11}\;\middle|\;0\;\right]\,,$$

so that the band algorithm gives an LQ decomposition of the matrix $(\,S_{1:\tilde{m}}^T\,A\,)$, ensuring that $\tilde{A}_{11}$ is of full column rank. Next we see that

$$S^T\,\left[\;B\;\middle|\;A\,W_{1:\tilde{n}}\;\right] \;=\; \left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \\ \hline 0 & 0 \end{array}\right]\,,$$

so that the band algorithm gives a QR decomposition of the matrix $[\,B \,|\, A\,W_{1:\tilde{n}}\,]$, ensuring that $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is of full row rank.

The original problem is decomposed into two independent subproblems, see (5.9), where the second one is homogeneous (with zero right-hand side). Since the problem is orthogonally invariant, the original problem $[\,B \,|\, A\,]$ is compatible, i.e. $\mathcal{R}(B) \subset \mathcal{R}(A)$, if and only if the subproblem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is compatible. Further, in the compatible case the matrix $\tilde{A}_{11} \in \mathbb{R}^{\tilde{m}\times\tilde{n}}$ must be square; it follows from the facts that $\tilde{A}_{11}$ is of full column rank, $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is of full row rank, and thus $\tilde{n} \,=\, \mathrm{rank}\,(\tilde{A}_{11}) \,=\, \mathrm{rank}\,([\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]) \,=\, \tilde{m}$.

On the other hand, the extended matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is of full row rank, thus the system matrix $\tilde{A}_{11}$ can have at most $\tilde{n}+d$ rows, i.e. $\tilde{m} \,\leq\, \tilde{n}+d$, which happens when $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ is square. We call such problem *fully incompatible* and it fulfill $\mathcal{R}\,(\tilde{B}_1) \,\cap\, \mathcal{R}(\tilde{A}_{11}) \,=\, \{0\}$ (this will be clarified in Chapter 6 through the relationship to the SVD-based reduction).

One can see that each row of the matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ contains at least one nonzero (positive) component γ_j. The assumption $B \in \mathbb{R}^{m\times d}$ being of full column rank guarantees that $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ contains at least γ_j, $j \,=\, 1\,,\,\ldots\,,\,d$, components. Similarly, each column of $\tilde{A}_{11}$ contains at least one nonzero (positive) component α_j. The assumption $A^T\,B \,\neq\, 0$ guarantees that $\tilde{A}_{11}$ contains at least α_1 component in the first d rows of $\tilde{A}_{11}$ (see Section 5.2). Consequently, the first ℓ rows of matrix $\tilde{A}_{11}$ can be zero, $\ell \,<\, d$. Finally, the matrix $\tilde{A}_{11}$ may not contain γ_{d+j} components, see the remark below.

Remark 5.7. *For the particular compatible system having the special property $\mathcal{R}(B) \,\equiv\, \mathcal{R}(A)$ the band algorithm yields $\tilde{A}_{11}$ which does not contain γ_j components. From (5.9), together with the special property and the fact that $\tilde{A}_{11}$ is square in the compatible case, it follows that $d \,=\, \tilde{n} \,=\, \tilde{m}$. Thus in this special case $\tilde{B}_1$ is square upper triangular and $\tilde{A}_{11}$ is square lower triangular and $\tilde{A}_{11}$ does not contain any γ_j.*

5.4 Generalization of Jacobi matrices

Properties (G3) and (G5), see Section 4.4, are related with the singular values and singular vector subspaces of $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$. Equivalently, these properties are related

to the spectral properties of the symmetric positive semidefinite matrix

$$\left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \end{array}\right]^T \left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \end{array}\right] = \left[\begin{array}{c|c} \tilde{B}_1^T \tilde{B}_1 & \tilde{B}_1^T \tilde{A}_{11} \\ \hline \tilde{A}_{11}^T \tilde{B}_1 & \tilde{A}_{11}^T \tilde{A}_{11} \end{array}\right]$$

and of its positive definite diagonal blocks, as well as the symmetric positive definite matrix

$$\left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \end{array}\right] \left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \end{array}\right]^T = \tilde{A}_{11}\,\tilde{A}_{11}^T + \tilde{B}_1\,\tilde{B}_1^T\,,$$

and both its positive semidefinite summands.

Now we introduce commonly used notation. For any given matrix $T \in \mathbb{R}^{n\times n}$ denote

$$\mathrm{bw}\,(T) \equiv \max_{i,j} \{\, |i-j| \,:\, t_{i,j} \equiv e_i^T\, T\, e_j \neq 0 \,\} \leq n-1$$

the *bandwidth*. It will be convenient to consider the following classes of matrices.

Definition 5.1 (Proper band matrix)**.** *Let $T \in \mathbb{R}^{n\times n}$ be a symmetric band matrix with the bandwidth $\rho \equiv \mathrm{bw}\,(T)$. When all components on its ρth sub- and superdiagonal are nonzero, i.e., $t_{ij} \neq 0$ when $|i-j| = \rho$, we call T a proper band matrix.*

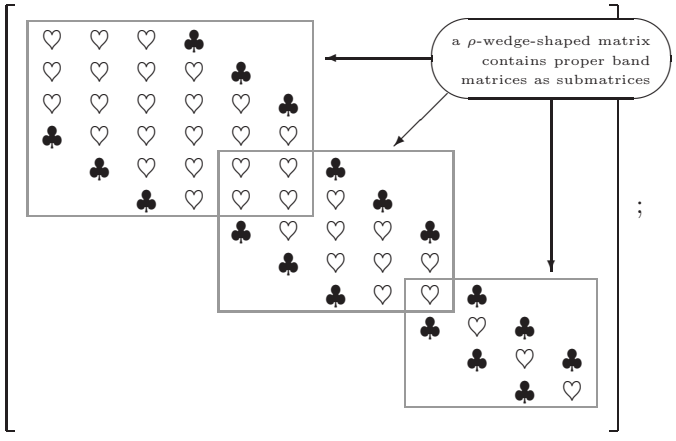
An illustration of a proper band matrix is given below for $\rho = 3$:

$$T = \left[\begin{array}{ccccccc} \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & \\ \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit \\ & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & & \clubsuit & \heartsuit & \heartsuit & \heartsuit \end{array}\right].$$

Recall that ♣ denotes a nonzero component, and ♡ denotes a component which may be zero as well as nonzero.

Definition 5.2 (ρ-wedge-shaped matrix)**.** *Let $T \in \mathbb{R}^{n\times n}$ be a symmetric band matrix. Denote $\nu(j)$ the column index of the first nonzero component in the jth row of T (using the symmetry of T gives that $\nu(j)$ denotes the row index of the first nonzero component in the jth column, too). When there exists a positive integer ρ, $1 \leq \rho < n$, such that the sequence $\{\, j - \nu(j) \,\}_{j=\rho+1}^{n}$ is positive and nonincreasing, we call T a ρ-wedge-shaped matrix.*

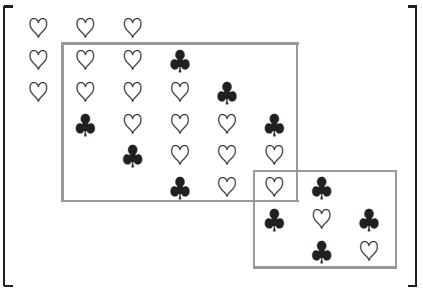
An illustration of a ρ-wedge-shaped matrix with $\rho = 3$ is given below:

$T =$ 

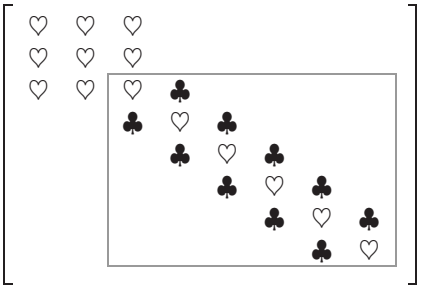
;

here $\nu(\rho+1) \equiv \nu(4) = 1$, $\nu(5) = 2$, $\nu(6) = 3$, $\nu(7) = 5$, etc., and the sequence $\{\, j - \nu(j) \,\}_{j=\rho+1}^{n}$ is $(\,3,3,3,2,2,2,1,1,1\,)$, and it is obviously positive and nonincreasing.

The definition of a ρ-wedge-shaped matrix does not require $\nu(\rho+1) = 1$. An illustration of a matrix with $\rho = 3$ and $\nu(\rho+1) = 2$ is given below:

$T =$ 

Further, an illustration of a matrix with $\rho = 3$ and $\nu(\rho+1) = 3$:

$T =$ 

Here it is worth to note that, if $T \in \mathbb{R}^{n\times n}$ is a ρ-wedge-shaped matrix with $\rho < n-1$, then it is $(\rho+1)$-wedge-shaped, too. Thus Definition 5.2 allows some nonuniqueness in classification of these matrices, which is necessary due to the fact that the $\heartsuit$ components may be nonzero as well as zero. We summarize some basic properties of these matrices:

- If $T \in \mathbb{R}^{n\times n}$ is a ρ-wedge-shaped matrix and $\nu(\rho+1) = 1$, then the bandwidth of T is equal to $\mathrm{bw}\,(T) \equiv \rho$.
- If $T \in \mathbb{R}^{n\times n}$ is a ρ-wedge-shaped matrix and $\nu(\rho+1) > 1$, then the bandwidth of T is smaller than ρ. Moreover, the first $\ell \equiv \nu(\rho+1) - 1$ columns (and rows) of a T can be zero; see the examples above recalling that $\heartsuit$ may be nonzero as well as zero.

The definition of ρ-wedge-shaped matrix requires positiveness of $\{\, j - \nu(j) \,\}_{j=\rho+1}^{n}$ sequence, and thus:

- The ρ-wedge-shaped matrix can not contain a 2 by 2 principal submatrix which is diagonal.

All these particularities follow from the fact that Definition 5.2 is chosen to fit the matrices returned by Algorithm 5.1. The relationship will be clarified in Lemma 5.1. Further:

- The proper band matrix T is a special case of ρ-wedge-shaped matrix, with $\rho \equiv \mathrm{bw}\,(T)$, $\nu(\rho+1) = 1$, and constant sequence $\{\, j - \nu(j) \,\}_{j=\rho+1}^{n}$.
- The k by k leading principal submatrix of a ρ-wedge-shaped matrix, $k > \rho$, is ρ-wedge-shaped, too. Similarly for the trailing principal submatrix (which means principal submatrix in the right bottom corner).

Finally note that, when we talk about ρ-wedge-shaped matrices and the value of ρ is not important, we often call them just *wedge-shaped matrices*. A ρ-wedge-shaped matrix we usually denote $T_{2\rho+1}$ (because it can have at most $2\rho+1$ nonzero diagonals).

In Sections 5.5 and 5.6 we generalize some properties of Jacobi matrices, i.e. symmetric tridiagonal matrices with positive subdiagonals, see [14, Chapter 2, pp. 13–35], [21, Chapter 1.3, pp. 10–20], [90, Section 5, §36–§48, pp. 299–316], [66, Chapter 7, pp. 119–150]; wedge-shaped matrices can be considered as generalization of the Jacobi matrices.

The following lemma formulates an important relationship between the subproblem obtained by the band algorithm and wedge-shaped matrices.

Lemma 5.1. *Let $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$ be the subproblem obtained by Algorithm 5.1, $\tilde{A}_{11} \in \mathbb{R}^{\tilde{m}\times\tilde{n}}$, $\tilde{B}_1 \in \mathbb{R}^{\tilde{m}\times d}$. Then the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,] \in \mathbb{R}^{(d+\tilde{n})\times(d+\tilde{n})}$ is d-wedge-shaped matrix. Furthermore, if $\tilde{m} > d$, then the matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T \in \mathbb{R}^{\tilde{m}\times\tilde{m}}$ is d-wedge-shaped matrix.*

The following proof is illustrated on a simple example in the further text, see Example 5.1 below the proof and subsequent notes.

Proof. In the proof we focus on the matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T$. A modification of the proof for the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ is straightforward (use $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T$ instead of $\tilde{A}_{11}$ and exchange the roles of the α and γ components).

Consider $\tilde{m} > d$. Denote $\tilde{a}_i^T \equiv e_i^T\,\tilde{A}_{11}$ the ith row of $\tilde{A}_{11}$, $i = 1\,,\,\ldots\,,\,\tilde{m}$. Then $e_k^T\,(\tilde{A}_{11}\,\tilde{A}_{11}^T)\,e_l = \tilde{a}_k^T\,\tilde{a}_l$ denotes the lth component of the kth row of the matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T$, i.e. components of $\tilde{A}_{11}\,\tilde{A}_{11}^T$ are dot products of rows of $\tilde{A}_{11}$.

Further we denote $\nu_U(i) \in \{\,1\,,\,\ldots\,,\,\tilde{m}\,\}$ the row index of the first nonzero component in the ith column of $\tilde{A}_{11}$, $i = 1\,,\,\ldots\,,\,\tilde{n}$ (i.e. the row index of α_i). Analogously, denote $\nu_L(j) \in \{\,1\,,\,\ldots\,,\,\tilde{n}\,\}$ the column index of the first nonzero component in the jth row of $\tilde{A}_{11}$, $j = d+1\,,\,\ldots\,,\,\tilde{m}$ (i.e. the column index of γ_j). Obviously, for $j = d+1\,,\,\ldots\,,\,\tilde{m}$,

$$\tilde{a}_j^T = [\underbrace{0\,,\,\ldots\,,\,0}_{\nu_L(j)-1}\,|\,\gamma_j\,|\,\heartsuit\,,\,\ldots\,,\,\heartsuit\,]\,. \tag{5.17}$$

where the components in the left or the right part of the row may be nonexistent. Similarly, for $i = 1\,,\,\ldots\,,\,\tilde{n}$,

$$\tilde{a}_{\nu_U(i)}^T = [\underbrace{\heartsuit\,,\,\ldots\,,\,\heartsuit}_{i-1}\,|\,\alpha_i\,|\,0\,,\,\ldots\,,\,0\,]\,, \tag{5.18}$$

where the components in the left or the right part of row may be nonexistent. Further, because $\nu_L(j) \in \{\,1\,,\,\ldots\,,\,\tilde{n}\,\}$ we choose from (5.18) only the following rows, for $j = d+1\,,\,\ldots\,,\,\tilde{m}$,

$$\tilde{a}_{\nu_U(\nu_L(j))}^T = [\underbrace{\heartsuit\,,\,\ldots\,,\,\heartsuit}_{\nu_L(j)-1}\,|\,\alpha_{\nu_L(j)}\,|\,0\,,\,\ldots\,,\,0\,]\,. \tag{5.19}$$

Compare with Example 5.1 below this proof and subsequent notes.

Recall that $\alpha_{\nu_L(j)}$ is the first nonzero component in the $(\nu_L(j))$th column of $\tilde{A}_{11}$, for $j = d+1\,,\,\ldots\,,\,\tilde{m}$. Consequently the components of the jth row of $\tilde{A}_{11}\,\tilde{A}_{11}^T$ are

$$\tilde{a}_j^T\,\tilde{a}_l = 0\,, \qquad \text{for} \quad l = 1\,,\,\ldots\,,\,\nu_U(\nu_L(j))-1\,,$$

and the first nonzero component is

$$\tilde{a}_j \, \tilde{a}^T_{\nu_U(\nu_L(j))} \, = \, \alpha_{\nu_L(j)} \, \gamma_j \, > \, 0 \, , \tag{5.20}$$

for $j \, = \, d+1 \, , \, \ldots \, , \, \tilde{m}$. This first nonzero component may be followed by several generally nonzero components, and, always, by the diagonal component $\tilde{a}^T_j \, \tilde{a}_j \, = \, \|\tilde{a}_j\|^2 \, \geq \, \gamma_j^2 \, > \, 0$. The other components in the jth row of $\tilde{A}_{11} \, \tilde{A}^T_{11}$ follow from the symmetry of $\tilde{A}_{11} \, \tilde{A}^T_{11}$.

It was shown that (5.20) represents the first nonzero component in the jth row, $j \, = \, d+1 \, , \, \ldots \, , \, \tilde{m}$, of $\tilde{A}_{11} \, \tilde{A}^T_{11} \in \mathbb{R}^{\tilde{m} \times \tilde{m}}$. Thus $\nu_U(\nu_L(j))$ is the row index of this first nonzero component, the index $\nu(j)$ from the definition of a ρ-wedge shaped matrix, i.e., $\nu_U(\nu_L(j)) \, \equiv \, \nu(j)$, see Definition 5.2.

It remains to show that the sequence $\{ \, j \, - \, \nu_U(\nu_L(j)) \, \}_{j=d+1}^{\tilde{m}}$ is positive and nonincreasing. The positiveness means that the first nonzero component (5.20) is not the diagonal component, which follows simply from the fact that $\alpha_{\nu_L(j)}$ is always over γ_j in $\tilde{A}_{11}$, see also the example (5.16). Further, because the integer sequences $\{ \, \nu_U(i) \, \}_{i=1}^{\tilde{n}}$ and $\{ \, \nu_L(j) \, \}_{j=d+1}^{\tilde{m}}$ are increasing, $\{ \, \nu_U(\nu_L(j)) \, \}_{j=d+1}^{\tilde{m}}$ is increasing too, and thus $\{ \, j \, - \, \nu_U(\nu_L(j)) \, \}_{j=d+1}^{\tilde{m}}$ must be nonincreasing. Thus $\tilde{A}_{11} \, \tilde{A}^T_{11} \in \mathbb{R}^{\tilde{m} \times \tilde{m}}$ is a d-wedge-shaped-matrix. □

It is worth to note that the condition $\tilde{m} \, > \, d$ in Lemma 5.1 is only formal. The matrix $\tilde{A}_{11}$ must have the α_1 component in the first d rows $\tilde{A}_{11}$ (it is a consequence of the assumption $A^T \, B \, \neq \, 0$), but it may not have the γ_{d+1} component, see also Sections 5.2 and 5.3. The condition $\tilde{m} \, > \, d$ ensures existence of γ_{d+1} and simplifies the proof. If $\tilde{m} \, = \, d$, then $\tilde{A}_{11} \, \tilde{A}^T_{11} \in \mathbb{R}^{d \times d}$ is general, and it does not make sense to call this matrix d-wedge-shaped. But we will see that all results achieved further in Sections 5.5 and 5.6 become trivial and are still valid for the matrix $\tilde{A}_{11}$ with $\tilde{m} \, = \, d$.

Further, note that if the α_1 component is not in the first row, i.e. $\nu_U(1) \, > \, 1$, or if the γ_{d+1} component is not in the first column, i.e. $\nu_L(d+1) \, > \, 1$, then the first nonzero component in the $(d+1)$st row of the d-wedge-shaped matrix $\tilde{A}_{11} \, \tilde{A}^T_{11}$ is not in the first column, i.e. $\nu(d+1) \, > \, 1$, see also Definition 5.2 and the subsequent notes. The consequence of this fact is that the bandwidth of the matrix $\tilde{A}_{11} \, \tilde{A}^T_{11}$ is smaller than d. In particular, if the α_1 component is not in the first row, i.e. $\nu_U(1) \, \equiv \, \ell + 1 \, > \, 1$, then the first ℓ rows (and columns) of $\tilde{A}_{11} \, \tilde{A}^T_{11}$ are zero. See also the following example.

Example 5.1. *Consider* $\tilde{A}_{11} \in \mathbb{R}^{\tilde{m} \times \tilde{n}}$, $\tilde{m} \, = \, 7$, $\tilde{n} \, = \, 5$ *and* $d \, = \, 5$, *such that*

$$\tilde{A}_{11} \, = \, \left[\begin{array}{ccccc} 0 & & & & \\ \hline \alpha_1 & & & & \\ \beta_{3,6} & & & & \\ \beta_{4,6} & \alpha_2 & & & \\ \beta_{5,6} & \beta_{5,7} & \alpha_3 & & \\ \hline & \gamma_6 & \beta_{6,8} & \alpha_4 & \\ & & & \gamma_7 & \alpha_5 \end{array} \right] \begin{array}{l} \} \, \ell \, \equiv \, 1 \\ \\ \\ \\ \\ \\ \\ \end{array} \left. \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \end{array} \right\} d \, \equiv \, 5 \quad ,$$

where α_1 *is in the second row and* $\gamma_{d+1} \, = \, \gamma_6$ *in the second column. Because* $7 \, = \, \tilde{m} \, > \, d \, = \, 5$, $\tilde{A}_{11} \, \tilde{A}^T_{11}$ *is a* d*-wedge shaped matrix; we will follow the proof above. The row indices of the first nonzero component in the* i*th column,* $i \, = \, 1 \, , \, \ldots \, , \, \tilde{n}$, *are:*

$$\nu_U(1) \, = \, 2 \, , \;\; \nu_U(2) \, = \, 4 \, , \;\; \nu_U(3) \, = \, 5 \, , \;\; \nu_U(4) \, = \, 6 \, , \;\; \nu_U(5) \, = \, 7 \, ;$$

we see that $\nu_U(i) \in \{ \, 1 \, , \, \ldots \, , \, \tilde{m} \, \}$. *Similarly the column indices of the first nonzero components in the* j *row,* $j \, = \, d+1 \, , \, \ldots \, , \, \tilde{m}$, *are:*

$$\nu_L(6) \, = \, 2 \, , \;\; \nu_L(7) \, = \, 4 \, ;$$

and we see that $\nu_L(j) \in \{\,1\,,\,\ldots\,,\,\tilde{n}\,\}$. *Then the values composite function* $\nu_U(\nu_L(j))$ *for* $j = d+1\,,\,\ldots\,,\,\tilde{m}$ *follows:*

$$\nu_U(\nu_L(6)) \,=\, \nu_U(2) \,=\, 4\,,\;\; \nu_U(\nu_L(7)) \,=\, \nu_U(4) \,=\, 6\,.$$

The dot products of the 6*th row with the* 1*st,* 2*nd, and* 3*rd rows of* $\tilde{A}_{11}$ *are zero; the product of the* 9*th row with* 4*th row,* $\nu_U(\nu_L(6)) = 4$*, is the first nonzero; the products of the* 6*th row with the* 5*th row may be nonzero as well as zero, and with* 6*th and* 7*th rows are nonzero. Similarly for the* 7*th row; the first nonzero dot product is with the* 6*th row,* $\nu_U(\nu_L(7)) = 6$*. These dot products represents components of the* 6*th and* 7*th row of the* $\tilde{A}_{11}\,\tilde{A}_{11}^T$ *matrix. The whole matrix* $\tilde{A}_{11}\,\tilde{A}_{11}^T$ *follows:*

$$\tilde{A}_{11}\,\tilde{A}_{11}^T \,=\, \left[\begin{array}{c|cccc|cc} 0 & & & & & & \\ \hline & \clubsuit & \heartsuit & \heartsuit & \heartsuit & & \\ & \heartsuit & \heartsuit & \heartsuit & \heartsuit & & \\ & \heartsuit & \heartsuit & \clubsuit & \heartsuit & \clubsuit & \\ & \heartsuit & \heartsuit & \heartsuit & \clubsuit & \heartsuit & \\ \hline & & & \clubsuit & \heartsuit & \clubsuit & \clubsuit \\ & & & & & \clubsuit & \clubsuit \end{array}\right] \begin{array}{l} \} \ell \equiv 1 \\ \\ \\ \\ \\ \\ \end{array} \left.\begin{array}{l} \\ \\ \\ \\ \\ \\ \end{array}\right\} d \equiv 5$$

$$= \left[\begin{array}{c|ccccc|cc} 0 & & & & & & & \\ \hline & \alpha_1^2 & \alpha_1\beta_{3,6} & \alpha_1\beta_{4,6} & \alpha_1\beta_{5,6} & & & \\ & \alpha_1\beta_{3,6} & \beta_{3,6}^2 & \beta_{3,6}\beta_{4,6} & \beta_{3,6}\beta_{5,6} & & & \\ & \alpha_1\beta_{4,6} & \beta_{3,6}\beta_{4,6} & \beta_{4,6}^2+\alpha_2^2 & \beta_{4,6}\beta_{5,6}+\alpha_2\beta_{5,7} & \alpha_2\gamma_6 & & \\ & \alpha_1\beta_{5,6} & \beta_{3,6}\beta_{5,6} & \beta_{4,6}\beta_{5,6}+\alpha_2\beta_{5,7} & \beta_{5,6}^2+\beta_{5,7}^2+\alpha_3^2 & \beta_{5,7}\gamma_6+\alpha_3\beta_{6,8} & & \\ \hline & & & \alpha_2\gamma_6 & \beta_{5,7}\gamma_6+\alpha_3\beta_{6,8} & \gamma_6^2+\beta_{6,8}^2+\alpha_4^2 & \alpha_4\gamma_7 & \\ & & & & & \alpha_4\gamma_7 & \gamma_7^2+\alpha_5^2 & \end{array}\right]$$

$$= \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & T_{2(d-1)+1} \end{array}\right].$$

The sequence $\{\,j-\nu(j)\,\}_{j=d+1}^{\tilde{m}} \equiv \{\,j-\nu_U(\nu_L(j))\,\}_{j=d+1}^{\tilde{m}}$ *is* $(\,2,1\,)$*, i.e., it is positive and nonincreasing, and thus the matrix* $T_{2d+1} \equiv \tilde{A}_{11}\,\tilde{A}_{11}^T$ *is a* d*-wedge-shaped matrix.*

Because $\nu_U(1) = 2$*, the first row and column of the matrix* $\tilde{A}_{11}\,\tilde{A}_{11}^T$ *is zero,* T_{2d+1} *contains the* $(d-1)$*-wedge-shaped matrix* $T_{2(d-1)+1}$ *with smaller dimension. Further, because* $\nu_L(d+1) = 2$*, the bandwidth of both matrices* T_{2d+1}*,* $T_{2(d-1)+1}$ *is, in general, equal to* 3*, i.e. smaller than* d *as well as* $d-1$*. Moreover, recall that all the coefficients* $\beta_{i,j}$ *can be zero. If, for example,* $\beta_{5,6} = 0$*, then the bandwidth of* $T_{2d+1} \equiv \tilde{A}_{11}\,\tilde{A}_{11}^T$ *as well as* $T_{2(d-1)+1}$ *is equal to* 2*. See also Definition 5.2 and the subsequent notes.*

For a bigger example of a ρ-wedge-shaped matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ with $\rho = 9$ see Figure 5.1.

Note that the situation with the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ from Lemma 5.1 is simpler. The component γ_1 is always in the first row of the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T$. Further $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T$ must contain the component α_1 (because $\tilde{A}_{11}$ must contain α_1 in the first d rows); the component α_1 can be in the $(\ell+1)$st column $\ell < d$.

Remark 5.8. *From Lemma 5.1 it follows that* $\tilde{A}_{11}\,\tilde{A}_{11}^T$ *and* $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ *are wedge-shaped matrices. Because* $\tilde{A}_{11}^T\,\tilde{A}_{11}$ *is the trailing principal submatrix of the second one, it is also a wedge-shaped matrix.*

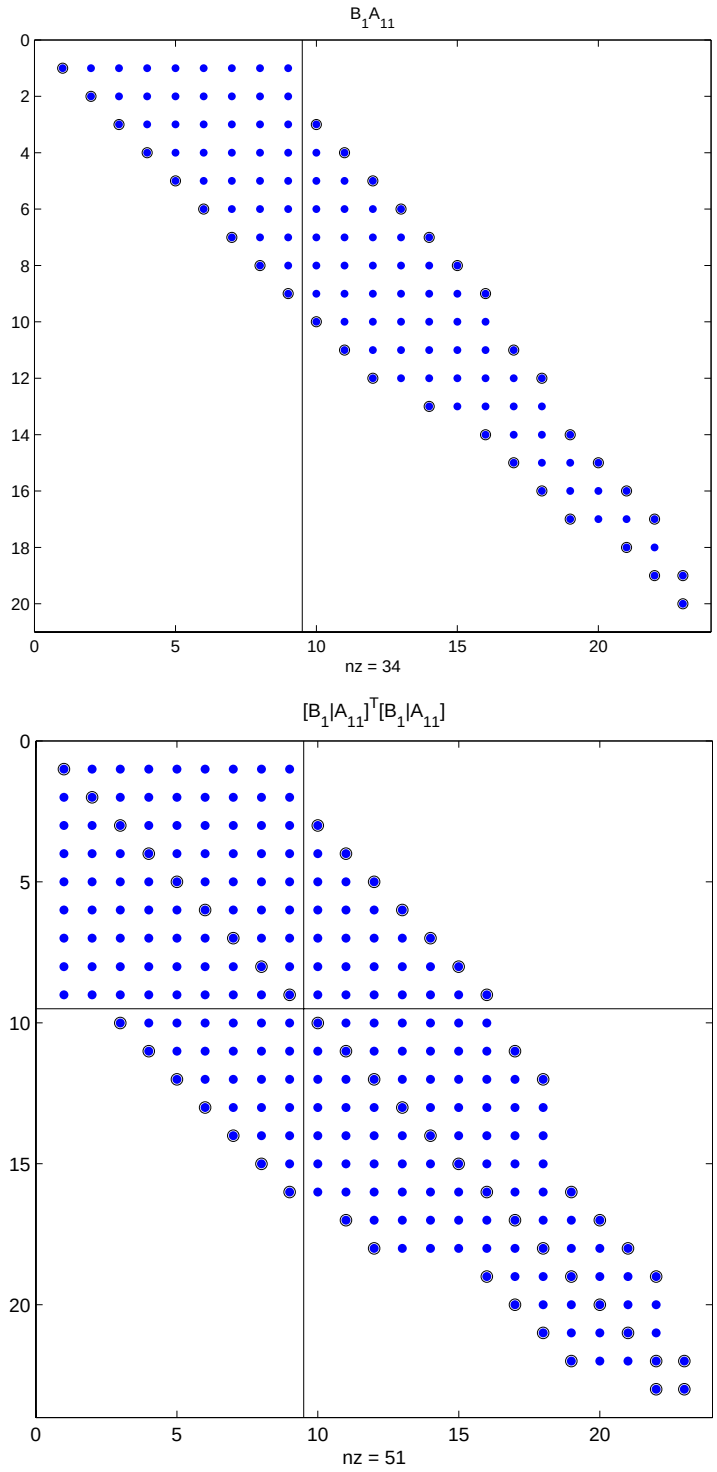


Figure 5.1: The top plot shows an example of a banded problem, the bottom plot shows the corresponding ρ-wedge-shaped matrix, $\rho = 9$. Bigger dots represent nonzero components (♣); smaller the components which may be nonzero as well as zero (♡).

5.5 Singular values in $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ problem

Here we study multiplicities of singular values of the matrices $\tilde{A}_{11}$ and $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$, i.e. multiplicities of eigenvalues of proper band and wedge-shaped matrices. It is well known that a Jacobi tridiagonal matrix has simple eigenvalues, see, e.g., [66, Section 7.7, Lemma 7.7.1, p. 134]. In the following text we give an analogous assertion for proper band and wedge-shaped matrices.

5.5.1 Eigenvalues of generalized Jacobi matrices

Lemma 5.2. *Let $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ be a symmetric proper band matrix with the bandwidth equal to $\mathrm{bw}\,(T_{2\rho+1}) \equiv \rho \geq 1$. Then the eigenvalues of the matrix $T_{2\rho+1}$ have multiplicities at most ρ.*

Proof. Let $\xi \in \mathbb{C}$ be an eigenvalue of $T_{2\rho+1}$. Denote $t_{i,j} \equiv e_i^T\, T_{2\rho+1}\, e_j$ the jth component of the ith row of $T_{2\rho+1}$. The matrix $(T_{2\rho+1} - \xi\, I_n)$ has a nonzero $(n-\rho)$ by $(n-\rho)$ minor $\prod_{j=\rho+1}^{n} t_{j,j-\rho}$. Thus $\mathrm{rank}\,(T_{2\rho+1} - \xi\, I_n) \geq (n-\rho)$, and $0 < \dim\,(\mathcal{N}\,(T_{2\rho+1} - \xi\, I_n)\,) \leq \rho$, cf. [66]. Finally, since $T_{2\rho+1}$ has a full set of orthogonal eigenvectors, the multiplicity of ξ is smaller than or equal to ρ. □

Example 5.2. *Consider $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ a symmetric proper band matrix with the bandwidth equal to $\mathrm{bw}\,(T_{2\rho+1}) \equiv \rho \geq 1$ such that, e.g.,*

$$T_{2\rho+1} \equiv \begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & \\ \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit \\ & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit \\ & & & \clubsuit & \heartsuit & \heartsuit & \heartsuit \end{bmatrix}, \qquad \rho = 3\,.$$

The nonzero $(n-\rho)$ by $(n-\rho)$ minor of $(T_{2\rho+1} - \xi\, I_n)$ is, e.g.,

$$\det\left(\begin{bmatrix} \clubsuit & \heartsuit & \heartsuit & \tilde{\heartsuit} \\ & \clubsuit & \heartsuit & \heartsuit \\ & & \clubsuit & \heartsuit \\ & & & \clubsuit \end{bmatrix}\right) = \prod_{j=\rho+1}^{n} \clubsuit \neq 0\,.$$

For $\rho = 1$, Lemma 5.2 reduces to the well known assertion that the Jacobi matrix – symmetric tridiagonal matrix with nonzero first sub- and superdiagonal has simple eigenvalues, see, e.g., [66, Section 7.7, Lemma 7.7.1, p. 134].

Lemma 5.3. *Let $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ be a symmetric ρ-wedge-shaped matrix with $\rho \geq 1$. Then the eigenvalues of the matrix $T_{2\rho+1}$ have multiplicities at most ρ.*

Proof. Let $\xi \in \mathbb{C}$ be an eigenvalue of $T_{2\rho+1}$. Denote $t_{i,j} \equiv e_i^T\, T_{2\rho+1}\, e_j$ the jth component of the ith row of $T_{2\rho+1}$. The sequence $\{\, j - \nu(j)\,\}_{j=\rho+1}^{n}$ is, from the definition of a ρ-wedge-shaped matrix, positive and nonincreasing. Recall that $\nu(j)$ denotes the first nonzero component in the jth row of the given matrix, see also the definition of a ρ-wedge-shaped matrix (Definition 5.2).

Because the sequence $\{\, j - \nu(j)\,\}_{j=\rho+1}^{n}$ is positive, all the components $t_{j,\nu(j)}$, $j = \rho+1\,,\,\ldots\,,\,n$ are in the strict lower triangular part (under the main diagonal) of $T_{2\rho+1}$ (and of $(T_{2\rho+1} - \xi\, I_n)$, too); and thus $t_{j,\nu(j)} \neq 0$ independently on the value of ξ. Because the sequence $\{\, j - \nu(j)\,\}_{j=\rho+1}^{n}$ is nonincreasing, the value of $\nu(j)$ must grow as fast as j, or faster; and thus all the $t_{j,\nu(j)}$, $j = \rho+1\,,\,\ldots\,,\,n$ components are in different rows and different columns of $T_{2\rho+1}$ (and of $(T_{2\rho+1} - \xi\, I_n)$, too).

Consequently, the matrix $(T_{2\rho+1} - \xi\, I_n)$ has a nonzero $(n-\rho)$ by $(n-\rho)$ minor $\prod_{j=\rho+1}^{n} t_{j,\nu(j)}$. Thus $\operatorname{rank}(T_{2\rho+1} - \xi\, I_n) \geq (n-\rho)$, and $0 < \dim(\mathcal{N}(T_{2\rho+1} - \xi\, I_n)) \leq \rho$, cf. [66]. Finally, since $T_{2\rho+1}$ has a full set of orthogonal eigenvectors, the multiplicity of ξ is smaller than or equal to ρ. □

Example 5.3. *Consider $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ a symmetric ρ-wedge-shaped matrix with $\rho \geq 1$ such that, e.g.,*

$$T_{2\rho+1} \equiv \begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & & \\ \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\ & & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & & \\ & & & & \clubsuit & \heartsuit & \heartsuit & \clubsuit & \\ & & & & & & \clubsuit & \heartsuit & \clubsuit \\ & & & & & & & \clubsuit & \heartsuit \end{bmatrix}, \qquad \rho = 3\,.$$

Using the column permutation $\Pi \equiv [\,e_1\,,\, e_2\,|\,e_4\,,\, e_5\,|\,e_7\,,\, e_8\,\|\,e_3\,|\,e_6\,|\,e_9\,]$ gives

$$(T_{2\rho+1} - \xi\, I_n)\,\Pi = \left[\begin{array}{cc|cc|cc||c|c|c} \tilde{\heartsuit} & \heartsuit & \clubsuit & & & & \heartsuit & & \\ \heartsuit & \tilde{\heartsuit} & \heartsuit & \clubsuit & & & \heartsuit & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & \tilde{\heartsuit} & & \\ \hline\hline \clubsuit & \heartsuit & \tilde{\heartsuit} & \heartsuit & & & \heartsuit & \clubsuit & \\ & \clubsuit & \heartsuit & \tilde{\heartsuit} & \clubsuit & & \heartsuit & \heartsuit & \\ & & \clubsuit & \heartsuit & \heartsuit & & & \tilde{\heartsuit} & \\ & & & \clubsuit & \tilde{\heartsuit} & \clubsuit & & \heartsuit & \\ & & & & \clubsuit & \tilde{\heartsuit} & & & \clubsuit \\ & & & & & \clubsuit & & & \tilde{\heartsuit} \end{array}\right],$$

and the nonzero $(n-\rho)$ by $(n-\rho)$ minor of $(T_{2\rho+1} - \xi\, I_n)$ is revealed.

Example 5.4. *For clarity we further consider $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ a symmetric ρ-wedge-shaped matrix with $\rho \geq 1$, and with $\nu(\rho+1) > 1$ (see Definition 5.2), such that, e.g.,*

$$T_{2\rho+1} \equiv \begin{bmatrix} \heartsuit & \heartsuit & \heartsuit & & & & & & \\ \heartsuit & \heartsuit & \heartsuit & & & & & & \\ \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & & & \\ & & \clubsuit & \heartsuit & \clubsuit & & & & \\ & & & \clubsuit & \heartsuit & \clubsuit & & & \\ & & & & \clubsuit & \heartsuit & \clubsuit & & \\ & & & & & \clubsuit & \heartsuit & \clubsuit & \\ & & & & & & \clubsuit & \heartsuit & \clubsuit \\ & & & & & & & \clubsuit & \heartsuit \end{bmatrix}, \qquad \rho = 3\,.$$

Using the column permutation $\Pi \equiv [\,e_3\,,\,\ldots\,,\,e_7\,\|\,e_1\,,\,e_2\,|\,e_8\,]$ gives

$$(T_{2\rho+1} - \xi\, I_n)\,\Pi = \left[\begin{array}{ccccc||cc|c} \heartsuit & & & & & \tilde{\heartsuit} & \heartsuit & \\ \heartsuit & & & & & \heartsuit & \tilde{\heartsuit} & \\ \tilde{\heartsuit} & \clubsuit & & & & \heartsuit & \heartsuit & \\ \hline\hline \clubsuit & \heartsuit & \clubsuit & & & & & \\ & \clubsuit & \tilde{\heartsuit} & \clubsuit & & & & \\ & & \clubsuit & \tilde{\heartsuit} & \clubsuit & & & \\ & & & \clubsuit & \tilde{\heartsuit} & & & \clubsuit \\ & & & & \clubsuit & & & \tilde{\heartsuit} \end{array}\right],$$

and the nonzero $(n-\rho)$ by $(n-\rho)$ minor of $(T_{2\rho+1} - \xi\, I_n)$ is revealed.

Obviously Lemma 5.3 generalizes the assertion of Lemma 5.2.

5.5.2 Singular values of matrices $\tilde{A}_{11}$ and $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$

Theorem 5.1. *Let $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$ be the subproblem obtained by the Algorithm 5.1. Then the singular values of the $(d+1)$-diagonal matrices $\tilde{A}_{11}$ and $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$, respectively, have multiplicities at most d.*

Proof. The singular values of $\tilde{A}_{11}$ are the square roots of the nonzero eigenvalues of the (symmetric positive semidefinite) d-wedge-shaped matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T$. Similarly, the singular values of $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ are the square roots of the eigenvalues of the (symmetric positive semidefinite) d-wedge-shaped matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]^T\,[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$. See Lemma 5.1. Consequently Lemma 5.3 proves the assertion. □

Remark 5.9. *Note that the assertion of Theorem 5.1 can be in some particular cases improved. If the first ℓ rows of the matrix $\tilde{A}_{11}$ are zero (such as in Example 5.1), then*

$$\tilde{A}_{11}\,\tilde{A}_{11}^T \;=\; \mathrm{diag}\,(\,0_{\ell,\ell}\,,\,T_{2(d-\ell)+1}\,)\,,$$

where $T_{2(d-\ell)+1}$ is a (symmetric positive semidefinite) $(d-\ell)$-wedge-shaped submatrix of $\tilde{A}_{11}\,\tilde{A}_{11}^T$. Consequently $\tilde{A}_{11}$ has singular values with multiplicities at most $d-\ell$. Further improving may be obtained using the (symmetric positive definite) wedge-shaped matrix $\tilde{A}_{11}^T\,\tilde{A}_{11}$. As an illustration see the following example: For

$$\tilde{A}_{11} \;=\; \begin{bmatrix} \clubsuit & & & & & \\ \heartsuit & \clubsuit & & & & \\ \heartsuit & \heartsuit & & & & \\ \clubsuit & \heartsuit & & & & \\ & \clubsuit & \clubsuit & & & \\ & & \clubsuit & \clubsuit & & \\ & & & \clubsuit & \clubsuit & \\ & & & & \clubsuit & \clubsuit \end{bmatrix}, \qquad d \;=\; 3\,,$$

we obtain

$$\tilde{A}_{11}\,\tilde{A}_{11}^T \;=\; \left[\begin{array}{ccc|ccccc} \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & & \\ \heartsuit & \clubsuit & \heartsuit & \heartsuit & \clubsuit & & & \\ \heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & \\ \hline \clubsuit & \heartsuit & \heartsuit & \clubsuit & \heartsuit & & & \\ & \clubsuit & \heartsuit & \heartsuit & \clubsuit & \clubsuit & & \\ & & & & \clubsuit & \clubsuit & \clubsuit & \\ & & & & & \clubsuit & \clubsuit & \end{array}\right], \qquad \tilde{A}_{11}^T\,\tilde{A}_{11} \;=\; \left[\begin{array}{cc|ccc} \clubsuit & \heartsuit & \clubsuit & & \\ \heartsuit & \clubsuit & \clubsuit & & \\ \hline & \clubsuit & \clubsuit & \clubsuit & \\ & & \clubsuit & \clubsuit & \clubsuit \\ & & & \clubsuit & \clubsuit \end{array}\right].$$

The first matrix is d-wedge-shaped with the bandwidth equal to $d = 3$, thus $\tilde{A}_{11}$ has singular values with multiplicities at most 3 (in agreement with Theorem 5.1).

The second matrix may be interpreted as ρ-wedge-shaped with $\rho = 2$, thus $\tilde{A}_{11}$ has singular values with multiplicities at most 2. But because $\nu(\rho+1) \equiv \nu(3) = 2 > 1$, the bandwidth of this matrix is smaller than ρ, i.e., equal to one.

One can see that if the $\heartsuit$ components in the second matrix are zeros, then $\tilde{A}_{11}$ can have at most one double singular value, the other must be simple; and if the $\heartsuit$ components are nonzero, $\tilde{A}_{11}$ has only simple singular values.

Summarizing, the subproblem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ obtained by the band algorithm has the following property:

(G5) The matrix $\tilde{A}_{11}$ has singular values with multiplicities at most d, and because $\tilde{A}_{11}$ is of full column rank, it does not have zero singular values.

5.6 Singular vector subspaces in $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ problem

We need to describe the singular vectors and the left singular vector subspaces of $\tilde{A}_{11}$, i.e. eigenvectors and eigenspaces of the matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T$, together with the corresponding components of the right-hand side $\tilde{B}_1$.

The right singular vectors and the right singular vector subspaces of the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ can be studied analogously as the eigenvectors and eigenspaces of the matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$. The results will be useful in Chapter 6.

It is well known that the eigenvector of a Jacobi tridiagonal matrix has nonzero first and last components, and it can not contain two subsequent zero components, see, e.g., [66, Section 7.9, Theorem 7.9.3, p. 140], see also [78, pp. 3–4]. In the following text we give an analogous assertion for proper band matrices and partially for wedge-shaped matrices.

5.6.1 Eigenvectors of generalized Jacobi matrices

Lemma 5.4. *Let $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ be a symmetric proper band matrix with the bandwidth equal to $\mathrm{bw}\,(T_{2\rho+1}) \equiv \rho \geq 1$, and let $\lambda \in \mathbb{R}$, $x = [\,\xi_1\,,\,\ldots\,,\,\xi_n\,]^T \in \mathbb{R}^n$, $\|x\| = 1$ be an eigenpair of $T_{2\rho+1}$,*

$$T_{2\rho+1}\,x \,=\, \lambda\,x\,.$$

Then:

(i) *The subvector $[\,\xi_1\,,\,\ldots\,,\,\xi_\rho\,]^T$ of x of length ρ is nonzero.*

(ii) *The subvector $[\,\xi_{n-\rho+1}\,,\,\ldots\,,\,\xi_n\,]^T$ of x of length ρ is nonzero.*

(iii) *All subvectors $[\,\xi_l\,,\,\ldots\,,\,\xi_{l+2\rho-1}\,]^T$ of x, $l \,=\, 1\,,\,\ldots\,,\,n-2\rho+1$, of length 2ρ are nonzero.*

Proof. Denote $t_{i,j} \equiv e_i^T\,T_{2\rho+1}\,e_j$. Assume by contradiction $\xi_1 \,=\, \ldots \,=\, \xi_\rho \,\equiv\, 0$. Using $T_{2\rho+1}\,x \,=\, \lambda\,x$ and comparing the kth components on the left and on the right, $k \,=\, 1\,,\,\ldots\,,\,n-\rho$, gives

$$\begin{aligned} e_k^T\,(\,T_{2\rho+1}\,x\,) \,&=\, \lambda\,\xi_k\,,\\ \left(\sum\nolimits_{j=1}^{k+\rho-1} t_{k,j}\,\xi_j\right) \,+\, t_{k,k+\rho}\,\xi_{k+\rho} \,&=\, \lambda\,\xi_k\,. \end{aligned}$$

Since $t_{k,k+\rho} \neq 0$, we have $\xi_{k+\rho} \,=\, 0$. Consequently $\|x\| \,=\, 0$ which contradicts the assumption $\|x\| \,=\, 1$. The proof of assertion (ii) is fully analogous.

Now we prove (iii). If $n \,\leq\, 2\rho$, then the assertion (iii) is trivially satisfied. Let $n \,>\, 2\rho$ and assume by contradiction $\xi_l \,=\, \ldots \,=\, \xi_{l+2\rho-1} \,\equiv\, 0$. Obviously, it is sufficient to prove it only for $2 \,\leq\, l \,\leq\, n-2\rho$ (the cases $l \,=\, 1$ and $l \,=\, n-2\rho+1$ are given by (i) and (ii)). Because $t_{l+\rho-1,l-1} \,\neq\, 0$, equating

$$\begin{aligned} e_{l+\rho-1}^T\,(\,T_{2\rho+1}\,x\,) \,&=\, \lambda\,\xi_{l+\rho-1}\,,\\ t_{l+\rho-1,l-1}\,\xi_{l-1} \,+\, \left(\sum\nolimits_{j=l}^{l+2\rho-1} t_{l+\rho-1,j}\,\xi_j\right) \,&=\, \lambda\,\xi_{l+\rho-1}\,, \end{aligned}$$

gives $x_{l-1} \,=\, 0$; similarly because $t_{l+\rho,l+2\rho} \,\neq\, 0$ equating

$$\begin{aligned} e_{l+\rho}^T\,(\,T_{2\rho+1}\,x\,) \,&=\, \lambda\,\xi_{l+\rho}\,,\\ \left(\sum\nolimits_{j=l}^{l+2\rho-1} t_{l+\rho,j}\,\xi_j\right) \,+\, t_{l+\rho,l+2\rho}\,\xi_{l+2\rho} \,&=\, \lambda\,\xi_{l+\rho}\,, \end{aligned}$$

gives $x_{l+2\rho} \,=\, 0$. Using induction, all components of x are zero and thus $\|x\| \,=\, 0$ which contradicts the assumption $\|x\| \,=\, 1$. □

For $\rho = 1$, Lemma 5.4 reduces to the well known assertion that eigenvectors of the Jacobi matrix – symmetric tridiagonal matrix with nonzero first sub- and super-diagonal, have nonzero first and last component, and can not have two subsequent zero components, see, e.g., [66, Section 7.9, Theorem 7.9.3, p. 140].

Lemma 5.5. *Let $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ be a symmetric ρ-wedge-shaped matrix with $\rho \geq 1$, and let $\lambda \in \mathbb{R}$, $x = [\,\xi_1\,,\,\ldots\,,\,\xi_n\,]^T \in \mathbb{R}^n$, $\|x\| = 1$ be an eigenpair of $T_{2\rho+1}$,*

$$T_{2\rho+1}\, x \;=\; \lambda\, x\,.$$

Then the subvector $[\,\xi_1\,,\,\ldots\,,\,\xi_\rho\,]^T$ of x of length ρ is nonzero.

Proof. Denote $t_{i,j} \equiv e_i^T\, T_{2\rho+1}\, e_j$. Recall that $\nu(j)$ denotes the column index of the first nonzero component in the jth row (or, using the symmetry, the row index of the first nonzero component in the jth column) of the matrix $T_{2\rho+1}$, see also the definition of a ρ-wedge-shaped matrix (Definition 5.2).

Assume by contradiction $\xi_1 = \ldots = \xi_\rho \equiv 0$. Using $T_{2\rho+1}\, x = \lambda\, x$ and comparing $(\nu(k))$th components on the left and on the right, $k = \rho+1\,,\,\ldots\,,\,n$, gives

$$\begin{aligned} e_{\nu(k)}^T\,(\,T_{2\rho+1}\, x\,) \;&=\; \lambda\, \xi_{\nu(k)}\,, \\ \left(\sum\nolimits_{j=1}^{k-1} t_{\nu(k),j}\, \xi_j\right) \;+\; t_{\nu(k),k}\, \xi_k \;&=\; \lambda\, \xi_{\nu(k)}\,. \end{aligned}$$

Because $\xi_1 = \ldots\, \xi_{k-1} = 0$, and because $\nu(k) < k$ (follows from the positiveness of the sequence $\{\, k - \nu(k)\,\}_{k=\rho+1}^{n}$) it is $\xi_{\nu(k)} = 0$, too. Since $t_{\nu(k),k} = t_{k,\nu(k)} \neq 0$, we have $\xi_k = 0$. Consequently $\|x\| = 0$ which contradicts the assumption $\|x\| = 1$. (Compare this proof with the proof of assertion (i) from Lemma 5.4 recalling that for a proper band matrix with the bandwidth ρ is $\nu(k) \equiv k-\rho$, $k = \rho+1\,,\,\ldots\,,\,n$.) □

The assertions (ii) and (iii) of Lemma 5.4 can not be generalized for wedge-shaped matrices. But the following example shows that there exists another nonzero subvector of the given eigenvector of a wedge-shaped matrix.

Example 5.5. *Let $\lambda \in \mathbb{R}$, $x \in \mathbb{R}^n$, $\|x\| = 1$ be an eigenpair of the ρ-wedge-shaped matrix with $\rho = 3$, $n = 9$,*

$$\begin{bmatrix}
\heartsuit & \heartsuit & \heartsuit & \clubsuit & & & & & \\
\heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & & \\
\heartsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & & & & \\
\clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & & \\
 & \clubsuit & \heartsuit & \heartsuit & \heartsuit & \heartsuit & \clubsuit & & \\
 & & & \clubsuit & \heartsuit & \heartsuit & \heartsuit & & \\
 & & & & \clubsuit & \heartsuit & \heartsuit & \clubsuit & \\
 & & & & & \clubsuit & \heartsuit & \clubsuit & \\
 & & & & & & \clubsuit & \heartsuit &
\end{bmatrix}.$$

First we illustrate the proof of Lemma 5.5. Here the sequence $\{\,\nu(j)\,\}_{j=\rho+1}^{n}$ is $(\,1,2,4,5,7,8\,)$. Assume by contradiction $\xi_1 = \xi_2 = \xi_3 = 0$. Comparing of the $(\nu(\rho+1))$th, i.e. 1st, components of $T_{2\rho+1}\, x = \lambda\, x$ gives $\xi_4 = 0$; further comparing the $(\nu(\rho+2))$th, i.e. 2nd, component gives $\xi_5 = 0$; comparing the $(\nu(\rho+3))$th, i.e. 4th, component gives $\xi_6 = 0$; etc., finally comparing the $(\nu(n))$th, i.e. 8th, component gives $\xi_9 = 0$ and thus $\|x\| = 0$ which contradicts the assumption $\|x\| = 1$.

Similarly, the subvector $[\,x_3\,,\,x_6\,,\,x_9\,]^T$ of x of length ρ must be nonzero. The proof is analogous to the proofs of Lemmas 5.4 and 5.5.

5.6.2 Eigenspaces of generalized Jacobi matrices

The goal of this section is to show that the subproblem obtained by the band algorithm has the property (G3), see Section 4.4. It was already proved that an eigenvector $x = [\,\xi_1\,,\,\ldots\,,\,\xi_n\,]^T$, $\|x\| = 1$ of a ρ-wedge-shaped matrix must contain a nonzero subvector $[\,\xi\,,\,\ldots\,,\,\xi_\rho\,]^T$ of length ρ. But such matrix can have multiple eigenvalues. It will be useful to show that the subvectors of length ρ, of the basis vectors corresponding to the given eigenspace, are linearly independent.

Lemma 5.6. *Let $T_{2\rho+1} \in \mathbb{R}^{n\times n}$ be a symmetric ρ-wedge-shaped matrix with $\rho \geq 1$, and let $\lambda \in \mathbb{R}$ be its eigenvalue with the multiplicity k, and*

$$\begin{aligned} x^{(1)} &= \left[\; \xi_1^{(1)}\,,\,\ldots\,,\,\xi_n^{(1)} \;\right]^T \in \mathbb{R}^n\,,\\ x^{(2)} &= \left[\; \xi_1^{(2)}\,,\,\ldots\,,\,\xi_n^{(2)} \;\right]^T \in \mathbb{R}^n\,,\\ &\vdots\\ x^{(k)} &= \left[\; \xi_1^{(k)}\,,\,\ldots\,,\,\xi_n^{(k)} \;\right]^T \in \mathbb{R}^n\,, \end{aligned}$$

be an orthonormal basis of the corresponding eigenspace,

$$T_{2\rho+1}\,\left[\; x^{(1)}\,,\,\ldots\,,\,x^{(k)} \;\right] = \lambda\,\left[\; x^{(1)}\,,\,\ldots\,,\,x^{(k)} \;\right]\,.$$

Then the leading submatrix

$$\Xi \equiv \begin{bmatrix} \xi_1^{(1)} & \xi_1^{(2)} & \cdots & \xi_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_\rho^{(1)} & \xi_\rho^{(2)} & \cdots & \xi_\rho^{(k)} \end{bmatrix} \in \mathbb{R}^{\rho\times k}$$

of the matrix $[\,x^{(1)}\,,\,\ldots\,,\,x^{(k)}\,]$ is of full column rank.

Proof. First, Lemma 5.3 ensures that $k \leq \rho$. Let $[\,\eta_1\,,\,\ldots\,,\eta_k\,]^T \neq 0$ be the vector of coefficients of an arbitrary nontrivial linear combination

$$y \equiv x^{(1)}\,\eta_1 + x^{(2)}\,\eta_2 + \ldots + x^{(k)}\,\eta_k\,.$$

Obviously $y = [\,\mu_1\,,\,\ldots\,,\,\mu_n\,]^T$ represents an eigenvector of $T_{2\rho+1}$ containing by Lemma 5.5 the nonzero subvector

$$\left[\; \mu_1\,,\,\ldots\,,\,\mu_\rho \;\right]^T = \Xi\,\left[\; \eta_1\,,\,\ldots\,,\,\eta_k \;\right]^T \neq 0\,,$$

of length ρ. Consequently, any nontrivial linear combination of columns of Ξ is nonzero, thus Ξ must be of full column rank. □

The assertion of Lemma 5.6 can be extended to any basis of the given eigenspace (not necessarily orthonormal).

5.6.3 Left singular vector subspaces of $\tilde{A}_{11}$

The following theorem uses results of Lemma 5.6 for the problem $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$.

Theorem 5.2. *Let $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$ be the subproblem obtained by Algorithm 5.1. Let ς' be a singular value of $\tilde{A}_{11}$ with multiplicity k, and $u'_1\,,\,\ldots\,,\,u'_k$ an orthonormal basis of the corresponding left singular vector subspace of $\tilde{A}_{11}$. Then the matrix*

$$\left[\; u'_1\,,\,\ldots\,,\,u'_k \;\right]^T\,\tilde{B}_1 \in \mathbb{R}^{k\times d}$$

is of full row rank.

Proof. The left singular vector subspaces of $\tilde{A}_{11}$ are identical to the eigenspaces of the matrix $\tilde{A}_{11}\,\tilde{A}_{11}^T$, which is, by Lemma 5.1, a d-wedge-shaped matrix.

The multiplicity of the singular value ς' of $\tilde{A}_{11}$ (or, equivalently, the multiplicity of the eigenvalue $(\varsigma')^2$ of $\tilde{A}_{11}\,\tilde{A}_{11}^T$) is, by Theorem 5.1, equal to at most d, i.e. $k \leq d$.

The leading d by k submatrix Ξ of the matrix $[\,u'_1\,,\,\ldots\,,\,u'_k\,]$ is, by Lemma 5.6, of full column rank. Recall that the matrix $\tilde{B}_1 = [\,R_1^T\,|\,0\,]^T$ contains a square nonsingular leading submatrix $R_1 \in \mathbb{R}^{d\times d}$, see (5.2). Consequently, the product $R_1^T\,\Xi = \tilde{B}_1^T\,[\,u'_1\,,\,\ldots\,,\,u'_k\,]$ must have full column rank, which finishes the proof. □

Thus finally, the subproblem $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ obtained by the band algorithm has the following property:

(G3) Let ς'_j be the singular value of $\tilde{A}_{11}$ with the multiplicity k_j, let U'_j be the matrix with the corresponding orthonormal singular vectors as its columns. Then $(U'_j)^T\,\tilde{B}_1$ are of full row rank for all j.

The following remark will be useful in Chapter 6.

Remark 5.10. *Let $\tilde{A}_{11}\,\tilde{X}_1 \approx \tilde{B}_1$ be the subproblem obtained by Algorithm 5.1. Let ς be a singular value of $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ with multiplicity k, and $v_1\,,\,\ldots\,,\,v_k$ an orthonormal basis of the corresponding right singular vector subspace. Application of Lemma 5.6 on the d-wedge-shaped matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]^T\,[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ gives that the d by k leading submatrix of $[\,v_1\,,\,\ldots\,,\,v_k\,]$ is of full column rank.*

5.7 Summary

Here we briefly summarize the properties proved through this chapter. The band generalization of the Golub-Kahan bidiagonalization described in Algorithm 5.1 produces a subproblem $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ having the properties:

(G1) The matrix $\tilde{A}_{11}$ is of full *column* rank equal to $\tilde{n} \leq \tilde{m}$ (see Section 5.3).

(G2) The matrix $\tilde{B}_1$ is of full *column* rank equal to $\tilde{d} \equiv d$ (by assumption).

(G3) Let ς'_j be the singular value of $\tilde{A}_{11}$ with the multiplicity k_j, let U'_j be the matrix with the corresponding orthonormal singular vectors as its columns. Then $(U'_j)^T\,\tilde{B}_1$ are of full row rank for all j (see Theorem 5.2).

(G4) The matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ is of full *row* rank equal to $\tilde{m}$ (see Section 5.3).

(G5) The matrix $\tilde{A}_{11}$ has singular values with multiplicities at most d, and it does not have zero singular values (see Theorem 5.1).

Recall that B is assumed to have column rank equal to d with no loss of generality, see Section 5.1.

Thus it was shown that both, the problem $[\,B_1\,|\,A_{11}\,]$ obtained from a general problem $[\,B\,|\,A\,]$ by the data reduction in Chapter 4 (see Section 4.4), as well as the problem $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ obtained from $[\,B\,|\,A\,]$ by the band algorithm in this chapter, have properties (G1)–(G5).

Moreover $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ has the following two properties:

- The matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ has singular values with multiplicities at most d, this matrix is of full row rank, shown in Theorem 5.1.
- The property which is shown in Remark 5.10.

Chapter 6

Core problem

Chapters 4 and 5 show how to transform the linear approximation problem $AX \approx B$ with multiple right-hand sides into the specific block forms. In both cases we obtain the reduced subproblem having properties (G1)–(G5).

In this chapter we show that properties (G1)–(G3) imply the minimality of the dimensions of the reduced problems, i.e. we show that the subproblems obtained by the SVD-based reduction, in Chapter 4, as well as by the generalized Golub-Kahan algorithm, in Chapter 5, can not be further reduced. We use properties (G1)–(G3) for definition of the core problem in the multiple right-hand side case.

We show that the core problem within the problem with multiple right-hand sides can contain more than one independent subproblems (of smaller dimensions) that can not be neglect. Finally we show by a counterexample that the core problem in the multiple right-hand sides case does not have a TLS solution in general.

6.1 Core problem definition

Consider a general approximation problem (2.1) with $A^T B \neq 0$. Let $[\,B_1 \,|\, A_{11}\,]$ be the subproblem obtained by the data reduction given in Chapter 4 applied to $[\,B \,|\, A\,]$. The subproblem $[\,B_1 \,|\, A_{11}\,]$ has the properties (G1)–(G5), $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$. Assume that there exists an orthogonal transformation such that

$$\hat{P}^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} \hat{R} & 0 \\ \hline 0 & \hat{Q} \end{array}\right] = \left[\begin{array}{c|c||c|c} \hat{B}_1 & 0 & \hat{A}_{11} & 0 \\ \hline 0 & 0 & 0 & \hat{A}_{22} \end{array}\right] \tag{6.1}$$

where $\hat{B}_1 \in \mathbb{R}^{\hat{m}\times\hat{d}}$, $\hat{A}_{11} \in \mathbb{R}^{\hat{m}\times\hat{n}}$, and $\hat{d} \leq \bar{d}$, $\hat{n} \leq \bar{n}$, $\hat{m} \leq \bar{m}$ and at least one of these inequalities is strict; $\hat{P}^{-1} = \hat{P}^T$, $\hat{Q}^{-1} = \hat{Q}^T$, $\hat{R}^{-1} = \hat{R}^T$. Then the problem $A_{11}\,X_1 \approx B_1$ can be further reduced, and, possibly, decomposed into two independent subproblems $\hat{A}_{11}\,\hat{X}_1 \approx \hat{B}_1$, $\hat{A}_{22}\,\hat{X}_2 \approx 0$.

From (G2), see Section 4.4, the right-hand side matrix B_1 is of full column rank, $\operatorname{rank}(B_1) = \bar{d}$ ($\bar{d} \leq \bar{m}$). Thus the matrix on the right in the following equality,

$$\hat{P}^T B_1 \hat{R} = \left[\begin{array}{c|c} \hat{B}_1 & 0 \\ \hline 0 & 0 \end{array}\right],$$

must be of full column rank, too. It gives that $\hat{d} \equiv \bar{d}$. Consequently, the transformation (6.1) reduces to

$$\hat{P}^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} \hat{R} & 0 \\ \hline 0 & \hat{Q} \end{array}\right] = \left[\begin{array}{c||c|c} \hat{B}_1 & \hat{A}_{11} & 0 \\ \hline 0 & 0 & \hat{A}_{22} \end{array}\right]. \tag{6.2}$$

From (G3), see Section 4.4, for all left singular vectors u_i of A_{11}, the product $u_i^T B_1$ are nonzero. The set of left singular vectors of A_{11} is the union of the sets of left singular vectors $[\,\hat{u}_{1,j}^T \,|\, 0\,]^T$, $j = 1, \ldots, \hat{m}$, and $[\,0 \,|\, \hat{u}_{2,k}^T\,]^T$, $k = 1, \ldots, \bar{m} - \hat{m}$, where $\hat{u}_{1,j}$ and $\hat{u}_{2,k}$ are left singular vectors of $\hat{A}_{11}$ and $\hat{A}_{22}$, respectively. Since the product $[\,0 \,|\, \hat{u}_{2,k}^T\,]\,[\,\frac{\hat{B}_1}{0}\,] = 0$ is zero for all k, $\hat{A}_{22}$ must have no left singular vectors and therefore it has no rows, i.e., $\hat{A}_{22}$ and the corresponding block row are nonexistent, which gives $\hat{m} \equiv \bar{m}$. Consequently, the transformation (6.2) reduces to

$$\hat{P}^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} \hat{R} & 0 \\ \hline 0 & \hat{Q} \end{array}\right] = \left[\begin{array}{c\|c|c} \hat{B}_1 & \hat{A}_{11} & 0 \end{array}\right]. \tag{6.3}$$

And finally, from (G1), see Section 4.4, the system matrix A_{11} is of full column rank, $\operatorname{rank}(A_{11}) = \bar{n} \leq \bar{m}$. Thus the matrix on the right in the following equality,

$$\hat{P}^T A_{11}\, \hat{Q} = \left[\begin{array}{c|c} \hat{A}_{11} & 0 \end{array}\right],$$

must be of full columns rank, too. It gives that $\hat{n} \equiv \bar{n}$, and thus the transformation (6.3) reduces to

$$\hat{P}^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} \hat{R} & 0 \\ \hline 0 & \hat{Q} \end{array}\right] = \left[\begin{array}{c|c} \hat{B}_1 & \hat{A}_{11} \end{array}\right]. \tag{6.4}$$

Summarizing, we showed that any transformation of the form (6.1) applied on a problem having properties (G1)–(G3) always reduces to the form (6.4).

The same idea can be applied on the subproblem $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ having properties (G1)–(G3) (see Section 5.7), returned by the banded generalization of the Golub-Kahan bidiagonalization algorithm (given in Chapter 5) applied on the original problem $[\,B \,|\, A\,]$. (Recall that the assumption B of full column rank in Chapter 5 is only technical, see the discussion in Section 5.1.) Moreover, the same idea can be applied on any reduced subproblem having properties (G1)–(G3), obtained by an orthogonal transformation from the original problem $[\,B \,|\, A\,]$.

Corollary 6.1. *Let $[\,B \,|\, A\,]$ be a general approximation problem. Both subproblems $[\,B_1 \,|\, A_{11}\,]$ obtained by the data reduction given in Chapter 4 applied on $[\,B \,|\, A\,]$, and $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ obtained from the banded algorithm given in Chapter 5 applied on $[\,B \,|\, A\,]$ must have the same (minimal) dimension and thus both represent the same subproblem of the original problem $[\,B \,|\, A\,]$ however in different coordinates.*

The proof follows from the considerations above.

Because the construction above shows that the properties (G1)–(G3) define some minimality property of the given problem, the following a core problem definition can be formulated.

Definition 6.1 (Core problem)**.** *The subproblem $A_{11}\, X_1 \approx B_1$ is a core problem within the approximation problem $A\, X \approx B$ if $[\,B_1 \,|\, A_{11}\,]$ is minimally dimensioned and A_{22} maximally dimensioned subject to (4.3), i.e. subject to the orthogonal transformation*

$$\begin{aligned} P^T \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] &= \left[\begin{array}{c|c} P^T B\, R & P^T A\, Q \end{array}\right] \\ &\equiv \left[\begin{array}{c|c\|c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right]. \end{aligned}$$

This definition has motivation in the Definition 1.1 of the core problem in the single right-hand side problems, firstly used by Paige and Strakoš in [64].

In accordance with the Definition 6.1, the subproblem obtained by the data reduction given in Chapter 4 or equivalently, by the banded algorithm given in Chapter 5 from the given $[\,B\,|\,A\,]$, is called the *core problem within* $[\,B\,|\,A\,]$. The form $[\,B_1\,|\,A_{11}\,]$ (4.25) of the core problem obtained by the data reduction given in Chapter 4 is called the *SVD form of the core problem*. Similarly, the form $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$, e.g., (5.16), of the core problem obtained by the banded algorithm given in Chapter 5 is called the *banded form of the core problem*.

Definition 6.1 assumes existence of an original problem and defines the core problem through the relationship between them. But for any problem having properties (G1)–(G3) can be artificially constructed an "original problem"; e.g., in according to Definition 6.1, one can the problem having properties (G1)–(G3) call *core problem within itself*. Consequently, we introduce an alternative of a core problem which uses only properties (G1)–(G3) (contrary to the fact, that the alternative definition defies the native meaning of this term: "to be the core within an original problem"), but it will be useful in some cases.

Definition 6.2 (Core problem; alternative definition)**.** *Any approximation problem* $A\,X \approx B$ *having properties (G1)–(G3) is called a core problem.*

Both definitions are equivalent, it follows from the considerations above.

Remark 6.1. *Note that because properties (G1)–(G3) uniquely define the core problem (up to an orthogonal transformation (6.4)) properties (G4)–(G5), see Section 5.7, are implied by (G1)–(G3).*

An application the banded reduction, see Chapter 5, on any problem having properties (G1)–(G3) yields a problem having properties (G1)–(G5) with the same dimensions. Because properties (G1)–(G5) are invariant with respect to the orthogonal transformation of the form (6.4), thus the original problem must have properties (G4)–(G5), too.

An application the SVD-based reduction, see Chapter 4, on any problem having properties (G1)–(G3) yields a problem having properties (G1)–(G5) and (S1)–(S3), see Section 4.4, with the same dimensions. Properties (S1)–(S3) are not invariant with respect to a general orthogonal transformation of the form (6.4), thus properties (S1)–(S3) are not general properties of a core problem.

6.2 Basic properties of the core problem

Here we summarize the basic properties of a core problem in the multiple right-hand side case. A core problem has properties (G1)–(G3) has by definition, and, as already discussed, properties (G4)–(G5) were proved from the SVD-based data reduction as well as from the generalized Golub-Kahan approach. Moreover, the core problem can always be transformed into the SVD form (4.25) having properties (S1)–(S3).

The generalized Golub-Kahan approach further shows that the matrix of the reduced problem $[\,B_1\,|\,A_{11}\,]$ has singular values with multiplicities at most equal to $\operatorname{rank}(B)$, see Theorem 5.1; and that $[\,B_1\,|\,A_{11}\,]$ has the property described in Remark 5.10, this additional property of core problems (which will be useful later in the analysis of solvability in the TLS sense) is discussed in this section.

First we recapitulate some notation used in Chapter 3. Let $A_{11}\,X_1 \approx B_1$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$, be a core problem. Consider the SVD of $[\,B_1\,|\,A_{11}\,]$,

$$\left[\;B_1\;\middle|\;A_{11}\;\right] \;=\; U\,\Sigma\,V^T\,, \tag{6.5}$$

where $U^{-1} = U^T$, $V^{-1} = V^T$, $\Sigma = [\,\mathrm{diag}\,(\sigma_1\,,\,\ldots\,,\,\sigma_{\bar{m}}\,)\,|\,0_{\bar{m},\bar{n}+\bar{d}-\bar{m}}\,]$, and

$$\sigma_1 \geq \ldots \geq \sigma_{\bar{m}} > 0\,, \tag{6.6}$$

recalling that $[\,B_1\,|\,A_{11}\,]$ is of full row rank, by (G2).

Here we concentrate on the *incompatible problem* $\mathcal{R}(B_1) \not\subset \mathcal{R}(A_{11})$ (the compatible case is simpler because it reduces to finding a solution of a system of linear algebraic equations). Thus $\mathrm{rank}\,([\,B_1\,|\,A_{11}\,]) > \mathrm{rank}\,(A_{11})$ which gives $\bar{m} > \bar{n}$.

In order to handle a possible multiplicity of $\sigma_{\bar{n}+1}$, we consider the following notation

$$\sigma_{\bar{n}-q} > \underbrace{\sigma_{\bar{n}-q+1} = \ldots = \sigma_{\bar{n}}}_{q} = \underbrace{\sigma_{\bar{n}+1} = \ldots = \sigma_{\bar{n}+e}}_{e} > \sigma_{\bar{n}+e+1}\,, \tag{6.7}$$

where q singular values to the left and $e-1$ singular values to the right are equal to $\sigma_{\bar{n}+1}$, and $q \geq 0$, $e \geq 1$. For convenience we denote $\bar{n} - q \equiv p$. If $q = \bar{n}$, then σ_p is nonexistent. Similarly, if $e = \bar{m} - \bar{n}$, then $\sigma_{\bar{n}+e+1}$ is nonexistent.

It will be useful to consider the following partitioning

$$V = \left[\begin{array}{c|c} V_{11}^{(q)} & V_{12}^{(q)} \\ \hline V_{21}^{(q)} & V_{22}^{(q)} \end{array}\right], \tag{6.8}$$

where $V_{11}^{(q)} \in \mathbb{R}^{\bar{d}\times(\bar{n}-q)}$, $V_{12}^{(q)} \in \mathbb{R}^{\bar{d}\times(\bar{d}+q)}$, $V_{21}^{(q)} \in \mathbb{R}^{\bar{n}\times(\bar{n}-q)}$, $V_{22}^{(q)} \in \mathbb{R}^{\bar{n}\times(\bar{d}+q)}$, see Figure 6.1 (compare with Figure 3.1). Further, define the following partitioning

$$V_{12}^{(q)} = \left[\begin{array}{c|c} W^{(q,e)} & V_{12}^{(-e)} \end{array}\right], \tag{6.9}$$

where $W^{(q,e)} \in \mathbb{R}^{\bar{d}\times(q+e)}$, $V_{12}^{(-e)} \in \mathbb{R}^{\bar{d}\times(\bar{d}-e)}$, $1 \leq e < \bar{d}$, see Figure 6.1 (compare with Figure 3.3).

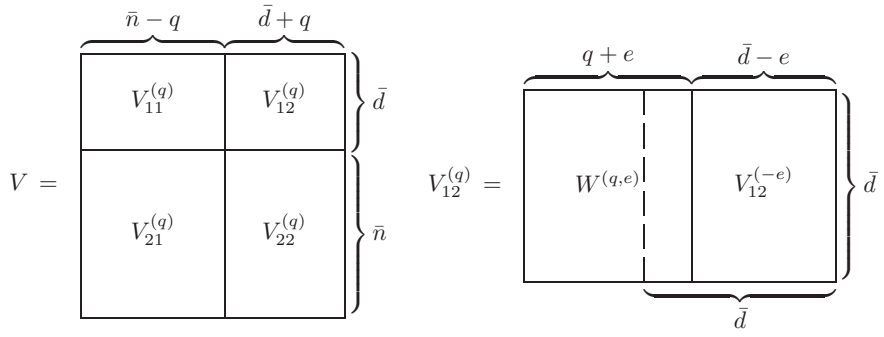


Figure 6.1: Dimensions of the individual matrix blocks in the partitioning (6.8) and (6.9). (Compare with Figure 3.1, p. 34, and Figure 3.3, p. 47.)

The matrix $W^{(q,e)}$ contains leading subvectors (of length $\bar{d}$) of the right singular vectors of $[\,B_1\,|\,A_{11}\,]$, corresponding to the singular value $\sigma_{\bar{n}+1}$ with the multiplicity $q+e$. The following theorem specifies the rank of this matrix.

Theorem 6.1. *Let $[\,B_1\,|\,A_{11}\,]$, $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$ $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$ be a core problem, and let $[\,B_1\,|\,A_{11}\,] = U\,\Sigma\,V^T$ be its SVD, with the partitioning given by (6.8), (6.9).*

Let the matrix $[\,B_1 \,|\, A_{11}\,]$ *have* l *distinct nonzero singular values* ς_j*, each with multiplicity* ρ_j*,* $\sum_{j=1}^{l} \rho_j = \bar{m}$*. Then each* $\bar{d}$ *by* ρ_j *block of* $[\,V_{11}^{(q)} \,|\, V_{12}^{(q)}\,]$ *corresponding to an individual singular value* ς_j *is of full column rank* ρ_j*,* $j = 1, \ldots, l$*.*

If $\bar{n}+\bar{d} > \bar{m}$*, then the trailing* $\bar{d}$ *by* $(\bar{n}+\bar{d}-\bar{m})$ *block of* $[\,V_{11}^{(q)} \,|\, V_{12}^{(q)}\,]$ *corresponding to the null space* $\mathcal{N}([\,B_1 \,|\, A_{11}\,])$ *is of full column rank* $\rho_{l+1} \equiv \bar{n} + \bar{d} - \bar{m}$*.*

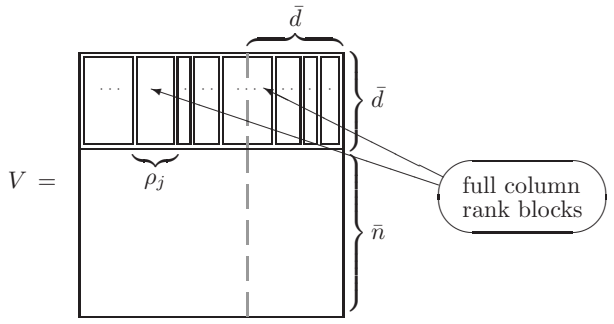


Figure 6.2: Full column rank blocks in the matrix of right singular vectors V obtained from the SVD of the core problem matrix $[\,B_1 \,|\, A_{11}\,]$. Each highlighted block corresponding to an individual singular value of $[\,B_1 \,|\, A_{11}\,]$ is of full column rank. If $[\,B_1 \,|\, A_{11}\,]$ is rectangular, then the last highlighted block (the rightmost) corresponding to the basis of $\mathcal{N}([\,B_1 \,|\, A_{11}\,])$ is of full column rank.

Proof. Any core problem (here represented by the matrix $[\,B_1 \,|\, A_{11}\,]$) can be orthogonally transformed into its banded form $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ described in Chapter 5, i.e.

$$P^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] = \left[\begin{array}{c|c} \tilde{B}_1 & \tilde{A}_{11} \end{array}\right],$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$. Denote $\tilde{V}$ the matrix of the right singular vectors of $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ with the splitting (6.8), then

$$R^T \left[\begin{array}{c|c} V_{11} & V_{12} \end{array}\right] = \left[\begin{array}{c|c} \tilde{V}_{11} & \tilde{V}_{12} \end{array}\right] \in \mathbb{R}^{\bar{d}\times(\bar{n}+\bar{d})}.$$

The ranks of the corresponding $\bar{d}$ by ρ_j blocks in V and $\tilde{V}$ must be the same, $j = 1, \ldots, l+1$, see also Figure 6.2. Lemma 5.6 applied on the wedge-shaped matrix $[\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]^T [\,\tilde{B}_1 \,|\, \tilde{A}_{11}\,]$ gives that these $\bar{d}$ by ρ_j blocks of $\tilde{V}$ must be of full column rank, see also Remark 5.10. □

Theorem 6.1 has an important corollary. The matrix $W^{(q,e)}$ (block of V), see (6.8), (6.9), and also Figure 6.1, corresponding to the singular value $\sigma_{\bar{n}+1}$ with the multiplicity $q+e$ is of full column rank. As already mentioned, this result will be useful later (in the following section) in the analysis of solvability of a core problem in the TLS sense.

6.3 Solution of the core problem

This section studies the solvability of a core problem. We focus on the relation between the results obtained by the TLS algorithm (Algorithm 3.1), applied to

the original problem and to its core problem. The TLS algorithm (Algorithm 3.1) returns a TLS solution if only if the problem belongs in the set $\mathscr{F}_1$, i.e. for the problem is $\operatorname{rank} W^{(q,e)} = e$, see the partitioning (6.8)–(6.9) and Figure 6.1 (see also Figure 3.5, p. 55). Thus this section further analyzes whether the core problem belongs to the set $\mathscr{F}_1$.

Let $[\, B_1 \,|\, A_{11} \,] = U \Sigma V^T$ be the SVD of the given core problem, $B_1 \in \mathbb{R}^{\bar{m} \times \bar{d}}$, $A_{11} \in \mathbb{R}^{\bar{m} \times \bar{n}}$; recall that $\max\{\, \bar{n}\,,\, \bar{d}\,\} \leq \bar{m} \leq \bar{n} + \bar{d}$. Because the whole analysis in Chapter 3 was done for problems having at least as many rows as columns, here at least $\bar{n} + \bar{d}$ rows, add $\rho_{l+1} \equiv \bar{m} - (\bar{n} + \bar{d})$ zero rows to the $[\, B_1 \,|\, A_{11} \,]$ matrix, if necessary, where

$$\left[\begin{array}{c|c} B_1 & A_{11} \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} U & 0 \\ \hline 0 & I_{\rho_{l+1}} \end{array}\right] \left[\begin{array}{c} \Sigma \\ \hline 0 \end{array}\right] V^T$$

represents the SVD of the modified system (the matrix of right singular vectors does not change). Consequently any solution which is constructed using columns of the matrix V, as well as any assertion based on properties of the matrix V, do not change. In this way we can apply to core problems all the tools developed in Chapter 3.

6.3.1 Solution computed by the TLS algorithm

This section shows that the TLS algorithm (see Algorithm 3.1, Section 3.4), applied on the given problem and on its core problem returns identical solutions (up to the orthogonal transformations which reveals the core problem).

Let $[\, B_1 \,|\, A_{11} \,]$ be a core problem within the given approximation problem $[\, B \,|\, A \,]$ and

$$P^T \left[\begin{array}{c|c} B & A \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] = \left[\begin{array}{c|c||c|c} B_1 & 0 & A_{11} & 0 \\ \hline 0 & 0 & 0 & A_{22} \end{array}\right] \tag{6.10}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$, see also (4.3). Let (6.5) be the SVD of $[\, B_1 \,|\, A_{11} \,]$ with partitioning given by (3.2)–(3.3), $\Delta \equiv \kappa$ (the partitioning (6.8) with κ instead of q, and the corresponding partitioning of the square matrix $[\, \frac{\Sigma}{0} \,]$), where κ is determined by Algorithm 3.1, applied on a core problem. Thus the TLS algorithm returns the matrix

$$X_1 \equiv - V_{22}^{(\kappa)} V_{12}^{(\kappa)\dagger} \tag{6.11}$$

as the solution of the core problem $[\, B_1 \,|\, A_{11} \,]$. Let $A_{22} = S \Theta W^T$ be the SVD of A_{22}. Then

$$\left[\begin{array}{c|c} B & A \end{array}\right] = \left(P \left[\begin{array}{c|c} U & 0 \\ \hline 0 & S \end{array}\right] \right) \left[\begin{array}{c||c|c|c} \Sigma_1^{(\kappa)} & \Sigma_2^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & \Theta & 0 \end{array}\right] \left(\left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] \left[\begin{array}{c||c|c|c} V_{11}^{(\kappa)} & V_{12}^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & 0 & I \\ \hline\hline V_{21}^{(\kappa)} & V_{22}^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & W & 0 \end{array}\right] \right)^T$$

represents (after permutation which sorts the singular values of Σ and Θ into one nonincreasing sequence) the SVD of the original problem. This permutation also changes the order of left and right singular vectors. Because the TLS algorithm chooses κ such that $V_{12}^{(\kappa)}$ is of full column rank, the upper right block

$$R \left[\begin{array}{c|c|c} V_{12}^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & I \end{array}\right] \tag{6.12}$$

is of full row rank, too. The permutation which sorts singular values of Σ and Θ rearranges columns of (6.12) in the following way:

The block column containing the identity matrix I corresponds to the zero singular values (originated in the zero block column in the (6.10)) and thus this block column can stay unchanged.

The singular values $\sigma_j(\Theta) > \sigma_{\max}(\Sigma_2^{(\kappa)})$ interlace with the singular values of $\Sigma_1^{(\kappa)}$. The corresponding columns of the zero block column in (6.12) are by this permutation moved left. Removing these corresponding zero columns from the full row rank matrix (6.12) does not change the rank of (6.12). Thus the TLS algorithm applied on $[\,B\,|\,A\,]$ chooses only the (permuted) submatrix of (6.12) without these zero columns.

The singular values $\sigma_j(\Theta) \leq \sigma_{\max}(\Sigma_2^{(\kappa)})$ are by this permutation interlaced with the singular values of $\Sigma_2^{(\kappa)}$. The corresponding columns of the zero block column in (6.12) are interlaced with columns of the left block column in (6.12). Denote Π the permutation matrix which represents this interlacing. Denote further $\tilde{W}$ the submatrix of the matrix W which contains right singular vectors of A_{22} corresponding to the singular values $\sigma_j(\Theta) \leq \sigma_{\max}(\Sigma_2^{(\kappa)})$.

Then

$$R\left[\begin{array}{c|c} [\,V_{12}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline 0 & I \end{array}\right], \qquad Q\left[\begin{array}{c|c} [\,V_{22}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline [\,0\,|\,W\,]\,\Pi & 0 \end{array}\right], \tag{6.13}$$

are the upper right, and lower right block of the matrix of right singular vectors of the original problem $[\,B\,|\,A\,]$ chosen by the TLS algorithm for computing the solution. Because the columns corresponding to the singular values $\sigma_j(\Theta) > \sigma_{\max}(\Sigma_2^{(\kappa)})$ are moved left, both matrices in (6.13) can have less columns than the matrix (6.12).

Consequently, the TLS algorithm returns the matrix

$$X \equiv -\left(Q\left[\begin{array}{c|c} [\,V_{22}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline [\,0\,|\,W\,]\,\Pi & 0 \end{array}\right]\right)\left(R\left[\begin{array}{c|c} [\,V_{12}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline 0 & I \end{array}\right]\right)^{\dagger} \tag{6.14}$$

as the solution of the original problem $[\,B\,|\,A\,]$. The successive using the properties (2.16), (2.15), (2.16), (2.17) of the Moore-Penrose pseudoinverse, see Lemma 2.1, gives

$$\begin{aligned}\left(R\left[\begin{array}{c|c} [\,V_{12}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline 0 & I \end{array}\right]\right)^{\dagger} &= \left(\left[\begin{array}{c|c} [\,V_{12}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline 0 & I \end{array}\right]\right)^{\dagger} R^T \\ = \left[\begin{array}{c|c} ([\,V_{12}^{(\kappa)}\,|\,0\,]\,\Pi)^{\dagger} & 0 \\ \hline 0 & I^{\dagger} \end{array}\right] R^T &= \left[\begin{array}{c|c} \Pi^T\,[\,V_{12}^{(\kappa)}\,|\,0\,]^{\dagger} & 0 \\ \hline 0 & I \end{array}\right] R^T \\ &= \left[\begin{array}{c|c} \Pi^T\begin{bmatrix} V_{12}^{(\kappa)\dagger} \\ \hline 0 \end{bmatrix} & 0 \\ \hline 0 & I \end{array}\right] R^T .\end{aligned}$$

Thus

$$\begin{aligned} X &= -Q\left[\begin{array}{c|c} [\,V_{22}^{(\kappa)}\,|\,0\,]\,\Pi & 0 \\ \hline [\,0\,|\,\tilde{W}\,]\,\Pi & 0 \end{array}\right]\left[\begin{array}{c|c} \Pi^T\begin{bmatrix} V_{12}^{(\kappa)\dagger} \\ \hline 0 \end{bmatrix} & 0 \\ \hline 0 & I \end{array}\right] R^T \\ &= -Q\left[\begin{array}{c|c} [\,V_{22}^{(\kappa)}\,|\,0\,]\,\Pi\,\Pi^T\begin{bmatrix} V_{12}^{(\kappa)\dagger} \\ \hline 0 \end{bmatrix} & 0 \\ \hline [\,0\,|\,\tilde{W}\,]\,\Pi\,\Pi^T\begin{bmatrix} V_{12}^{(\kappa)\dagger} \\ \hline 0 \end{bmatrix} & 0 \end{array}\right] R^T = -Q\left[\begin{array}{c|c} V_{22}^{(\kappa)}\,V_{12}^{(\kappa)\dagger} & 0 \\ \hline 0 & 0 \end{array}\right] R^T . \end{aligned} \tag{6.15}$$

Finally, comparing (6.11) and (6.15) gives

$$X \;=\; Q \left[\begin{array}{c|c} X_1 & 0 \\ \hline 0 & 0 \end{array}\right] R^T , \tag{6.16}$$

the relationship between the solutions the original problem and its core problem. We formulate this result as the following corollary.

Corollary 6.2. *Let* $[\,B \,|\, A\,]$ *be a general linear approximation problem and let* $[\,B_1 \,|\, A_{11}\,]$ *be a core problem within* $[\,B \,|\, A\,]$*. Then the core problem contains all necessary and sufficient information for computing the solution of the original problem by the TLS algorithm (Algorithm 3.1).*

Results obtained by Algorithm 3.1 applied on an original problem and a core problem within it are identical up to the corresponding orthogonal transformations (and adding some zero rows and columns).

Sufficiency of information follows from the construction (6.16), necessity from the basic property that the core problem has minimal dimensions.

It is worth to recall that this section does not discus whether the solution computed by Algorithm 3.1 is the TLS solution or not.

6.3.2 TLS solution of the core problem

This section focuses on the very important question whether a core problem has a TLS solution. First of all it is necessary to recall some results from Chapter 3.

Let $A_{11}\,X_1 \;\approx\; B_1$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$ be a core problem, and let (6.5) be the SVD of $[\,B_1 \,|\, A_{11}\,]$ with the partitioning (6.8), (6.9). If $V_{12}^{(q)} \in \mathbb{R}^{\bar{d}\times(\bar{d}+q)}$ is of full row rank equal to $\bar{d}$ (equivalently, the problem is of the 1st class), then:

(i) If $V_{12}^{(-e)} \in \mathbb{R}^{\bar{d}\times(\bar{d}-e)}$ is of full column rank equal to $\bar{d}-e$ ($A_{11}\,X_1 \;\approx\; B_1$ belongs in the set $\mathscr{F}_1 \cup \mathscr{F}_2$; see Figure 3.5, p. 55), then $A_{11}\,X_1 \;\approx\; B_1$ has a TLS solution.

(ii) Algorithm 3.1 applied on $A_{11}\,X_1 \;\approx\; B_1$ computes the TLS solution *if and only if* the rank of $W^{(q,e)} \in \mathbb{R}^{\bar{d}\times(q+e)}$ is equal to e ($A_{11}\,X_1 \;\approx\; B_1$ belongs in the set $\mathscr{F}_1$).

A TLS solution can be obtained using the following construction. Consider an orthogonal matrix $\tilde{Q} \;=\; \mathrm{diag}\,(\,Q'\,,\, I_{\bar{d}-e}\,) \in \mathbb{R}^{(\bar{d}+q)\times(\bar{d}+q)}$, $Q' \in \mathbb{R}^{(q+e)\times(q+e)}$ such that

$$\left[\begin{array}{c} V_{12}^{(q)} \\ \hline V_{22}^{(q)} \end{array}\right] \tilde{Q} \;=\; \left[\;\; v_{p+1}\,,\,\ldots\,,\,v_{\bar{n}+\bar{d}} \;\;\right] \tilde{Q} \;=\; \left[\begin{array}{c|c} \Omega & \tilde{\Gamma} \\ \hline \tilde{Y} & \tilde{Z} \end{array}\right] , \tag{6.17}$$

where $\tilde{\Gamma} \in \mathbb{R}^{\bar{d}\times\bar{d}}$ is *nonsingular*. Define the correction matrix

$$\left[\;\; \tilde{G} \;\middle|\; \tilde{E} \;\;\right] \;\equiv\; -\left[\;\; B_1 \;\middle|\; A_{11} \;\;\right] \left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right] \left[\begin{array}{c} \tilde{\Gamma} \\ \hline \tilde{Z} \end{array}\right]^T . \tag{6.18}$$

The corrected system $(A_{11} + \tilde{E})\,X_1 \;=\; B_1 + \tilde{G}$ is compatible and the matrix

$$\tilde{X}_1 \;\equiv\; -\,\tilde{Z}\,\tilde{\Gamma}^{-1} \tag{6.19}$$

solves the corrected system; see Theorem 3.4.

If $V_{12}^{(-e)}$ is of full column rank equal to $\bar{d} - e$ ($A_{11} X_1 \approx B_1$ belongs in the set $\mathscr{F}_1 \cup \mathscr{F}_2$), then the orthogonal matrix $\tilde{Q}$ exists.

If $W^{(q,e)}$ has rank equal to e ($A_{11} X_1 \approx B_1$ belongs in the set $\mathscr{F}_1$), then the orthogonal matrix $\tilde{Q}$ exists, and, moreover, $\Omega \equiv 0$ in (6.17). Consequently,

$$\begin{aligned}
\tilde{X}_1 = -\tilde{Z}\tilde{\Gamma}^{-1} &= -\left[\begin{array}{c|c} \tilde{Y} & \tilde{Z} \end{array}\right] \left[\begin{array}{c} 0 \\ \hline \tilde{\Gamma}^{-1} \end{array}\right] \\
&= -\left[\begin{array}{c|c} \tilde{Y} & \tilde{Z} \end{array}\right] \tilde{Q}^T \tilde{Q} \left[\begin{array}{c|c} 0 & \tilde{\Gamma} \end{array}\right]^\dagger \\
&= -\left[\begin{array}{c|c} \tilde{Y} & \tilde{Z} \end{array}\right] \tilde{Q}^T \left(\left[\begin{array}{c|c} 0 & \tilde{\Gamma} \end{array}\right] \tilde{Q}^T\right)^\dagger \\
&= -V_{22}^{(q)} V_{12}^{(q)\dagger},
\end{aligned}$$

thus the matrix (6.19) is identical with the matrix computed by Algorithm 3.1, see (6.11). The following theorem formulates the basic result for the TLS solution of a core problem.

Theorem 6.2. *Let $A_{11} X_1 \approx B_1$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, $B_1 \in \mathbb{R}^{\bar{m}\times\bar{n}}$ be a core problem, and let (6.5) be the SVD of $[\,B_1 \,|\, A_{11}\,]$ with the partitioning (6.8), (6.9). If $\operatorname{rank}(V_{12}^{(q)}) = \bar{d}$ (i.e. the core problem is a problem of the first class) then the following assertions are equivalent:*

(i) *The core problem belongs to the set $\mathscr{F}_1$ (see Figure 3.5 on p. 55);*

(ii) $\sigma_{\bar{n}}([\,B_1 \,|\, A_{11}\,]) > \sigma_{\bar{n}+1}([\,B_1 \,|\, A_{11}\,])$;

Proof. First we proof (i) $\Longrightarrow$ (ii). Any problem of the 1st class belongs in the set $\mathscr{F}_1$ if and only if the rank of the matrix $W^{(q,e)} \in \mathbb{R}^{\bar{d}\times(q+e)}$ is equal to e. For any core problem, the matrix $W^{(q,e)}$ is of full column rank, by Theorem 6.1. Thus if the core problem belongs in the set $\mathscr{F}_1$, then the matrix $W^{(q,e)}$ must have e columns, which gives $q = 0$ and thus the assertion (ii) of this lemma (see (6.7)).

The other implication (ii) $\Longrightarrow$ (i) follows immediately. The assertion (ii) gives $q = 0$. Thus $V_{12}^{(q)} \in \mathbb{R}^{\bar{d}\times(\bar{d}+q)}$ is square. Because $V_{12}^{(q)}$ is of full row rank, it has linearly independent columns and thus the rank of the matrix $W^{(q,e)} \in \mathbb{R}^{\bar{d}\times(q+e)}$ is equal to e (see (6.9) and Figure 6.1, see also Figure 6.3)). □

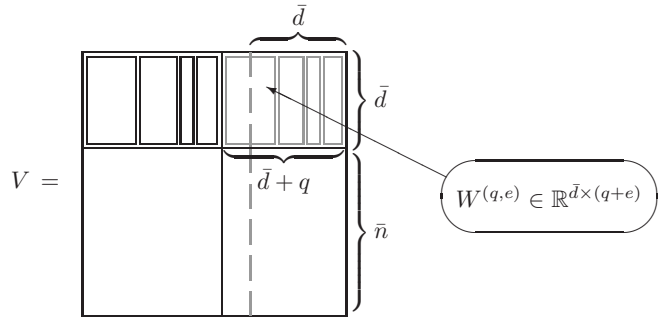


Figure 6.3: The matrix $W^{(q,e)}$ which corresponds to the singular value $\sigma_{\bar{n}}([\,B_1 \,|\, A_{11}\,])$ is of full column rank by Theorem 6.1.

Consequently, from Theorem 6.2 it follows that a core problem belongs to the set $\mathscr{F}_1$ if and only if it has the *unique TLS solution*, see Section 3.2.1.

In contrast to the single right-hand side case, we will show in Section 6.5 (see Examples 6.1, 6.2, and 6.3) below, that the value of q can be for some core problems nonzero. Thus the TLS algorithm (Algorithm 3.1) applied to the given core problem computes either the uniquely defined TLS solution $X_{\mathrm{TLS}} = -V_{22}^{(q)}\,(V_{12}^{(q)})^{-1}$, if $q = 0$; or the solution $X_{\mathrm{NGN}} = -V_{22}^{(\kappa)}\,(V_{12}^{(\kappa)})^{\dagger}$, where the value of κ is determined by the TLS algorithm, $\kappa \geq q > 0$. The X_{NGN} solution is called *nongeneric*, see also Figure 3.5, p. 55.

6.4 On the existence of independent subproblems within a core problem

In this section it will be shown that the core problem $[\,B_1 \,|\, A_{11}\,]$ can in some situations contain two or more independent subproblems which can be be solved independently (see Section 6.5).

Lemma 6.1. *Let* $[\,B_{1,\mathrm{I}} \,|\, A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}} \,|\, A_{11,\mathrm{II}}\,]$ *be matrices representing two independent approximation problems. Consider an approximation problem* $[\,B_1 \,|\, A_{11}\,]$ *such that*

$$\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \equiv P \left[\begin{array}{c|c||c|c} B_{1,\mathrm{I}} & 0 & A_{11,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} & 0 & A_{11,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right]^T, \tag{6.20}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$ *are chosen arbitrarily. Then the problems* $[\,B_{1,\mathrm{I}} \,|\, A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}} \,|\, A_{11,\mathrm{II}}\,]$ *have properties (G1)–(G3) if and only if* $[\,B_1 \,|\, A_{11}\,]$ *has properties (G1)–(G3).*

Proof. Property (G1): The matrices $A_{11,\mathrm{I}}$, $A_{11,\mathrm{II}}$ are of full column rank *iff* the matrix $\mathrm{diag}\,(\,A_{11,\mathrm{I}}\,,\,A_{11,\mathrm{II}}\,)$ is of full column rank. Because P, Q are square orthogonal matrices,

$$\mathrm{rank}\,(\,\mathrm{diag}\,(\,A_{11,\mathrm{I}}\,,\,A_{11,\mathrm{II}}\,)\,) \;=\; \mathrm{rank}\,(A_{11})\,,$$

where $A_{11} \equiv P\,\mathrm{diag}\,(\,A_{11,\mathrm{I}}\,,\,A_{11,\mathrm{II}}\,)\,Q^T$. Thus the matrices $A_{11,\mathrm{I}}$, $A_{11,\mathrm{II}}$ are of full column rank *iff* the matrix A_{11} is of full column rank.

Property (G2): The proof is analogous, the matrices $B_{1,\mathrm{I}}$, $B_{1,\mathrm{II}}$ are of full column rank *iff* the matrix $\mathrm{diag}\,(\,B_{1,\mathrm{I}}\,,\,B_{1,\mathrm{II}}\,)$ is of full column rank. Because P, R are square orthogonal matrices,

$$\mathrm{rank}\,(\,\mathrm{diag}\,(\,B_{1,\mathrm{I}}\,,\,B_{1,\mathrm{II}}\,)\,) \;=\; \mathrm{rank}\,(B_1)\,,$$

where $B_1 \equiv P\,\mathrm{diag}\,(\,B_{1,\mathrm{I}}\,,\,B_{1,\mathrm{II}}\,)\,R^T$. Thus the matrices $B_{1,\mathrm{I}}$, $B_{1,\mathrm{II}}$ are of full column rank *iff* the matrix B_1 is of full column rank.

Property (G3): Let $\varsigma_{i,\mathrm{I}}$ be the singular value of $A_{11,\mathrm{I}}$ with the multiplicity $k_{i,\mathrm{I}}$, and let $U'_{i,\mathrm{I}}$ be the matrix with the corresponding orthonormal singular vectors as its columns. Analogously let $\varsigma_{j,\mathrm{II}}$ be the singular value of $A_{11,\mathrm{II}}$ with the multiplicity $k_{j,\mathrm{II}}$, and let $U'_{j,\mathrm{II}}$ be the matrix with the corresponding orthonormal singular vectors as its columns. Then

$$\begin{aligned} (U'_{i,\mathrm{I}})^T\,B_{1,\mathrm{I}} \quad &\text{are of full row rank for all } i, \quad \text{and} \\ (U'_{j,\mathrm{II}})^T\,B_{1,\mathrm{II}} \quad &\text{are of full row rank for all } j \end{aligned}$$

iff

$$\left[\begin{array}{c|c} (U'_{i,\mathrm{I}})^T & 0 \end{array}\right] \left[\begin{array}{c|c} B_{1,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} \end{array}\right] \quad \text{and} \quad \left[\begin{array}{c|c} 0 & (U'_{j,\mathrm{II}})^T \end{array}\right] \left[\begin{array}{c|c} B_{1,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} \end{array}\right]$$

are of full row rank for all i and all j, respectively. The columns of $[\,(U'_{i,\mathrm{I}})^T \,|\, 0\,]^T$ and $[\,0 \,|\, (U'_{j,\mathrm{II}})^T\,]^T$ represent the left singular vectors of the matrix $\mathrm{diag}\,(\,A_{11,\mathrm{I}}\,,\,A_{11,\mathrm{II}}\,)$.

Using square orthogonal matrices P, R gives full row rank matrices

$$\left(\left[\begin{array}{c|c} (U'_{i,\mathrm{I}})^T & 0 \end{array}\right] P^T\right)\left(P\left[\begin{array}{c|c} B_{1,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} \end{array}\right] R^T\right) = \left[\begin{array}{c|c} (U'_{i,\mathrm{I}})^T & 0 \end{array}\right] P^T B_1\,,$$

$$\left(\left[\begin{array}{c|c} 0 & (U'_{j,\mathrm{II}})^T \end{array}\right] P^T\right)\left(P\left[\begin{array}{c|c} B_{1,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} \end{array}\right] R^T\right) = \left[\begin{array}{c|c} 0 & (U'_{j,\mathrm{II}})^T \end{array}\right] P^T B_1\,,$$

where columns of $P\,[\,(U'_{i,\mathrm{I}})^T \,|\, 0\,]^T$ and $P\,[\,0 \,|\, (U'_{j,\mathrm{II}})^T\,]^T$ represent the left singular vectors of the matrix A_{11}.

Finally, let ς_l be the singular value of A_{11} with the multiplicity k_l, and let U'_l be the matrix with the corresponding orthonormal singular vectors as its columns. Denote $\{\,\varsigma_{i,\mathrm{I}}\,\}$ and $\{\,\varsigma_{j,\mathrm{II}}\,\}$ the sets of distinct singular values of $A_{11,\mathrm{I}}$ and $A_{11,\mathrm{II}}$, respectively. There are three different situations:

- If $\varsigma_l \in \{\,\varsigma_{i,\mathrm{I}}\,\}$ and $\varsigma_l \notin \{\,\varsigma_{j,\mathrm{II}}\,\}$, then $(U'_l)^T B_1 \equiv [\,(U'_{i,\mathrm{I}})^T \,|\, 0\,]\, P^T B_1$ for some i, and thus $(U'_l)^T B_1$ is of full row rank ($k_l = k_{i,\mathrm{I}}$).
- If $\varsigma_l \notin \{\,\varsigma_{i,\mathrm{I}}\,\}$ and $\varsigma_l \in \{\,\varsigma_{j,\mathrm{II}}\,\}$, then $(U'_l)^T B_1 \equiv [\,0 \,|\, (U'_{j,\mathrm{II}})^T\,]\, P^T B_1$ for some j, and thus $(U'_l)^T B_1$ is of full row rank ($k_l = k_{j,\mathrm{II}}$).
- If $\varsigma_l \in \{\,\varsigma_{i,\mathrm{I}}\,\}$ and $\varsigma_l \in \{\,\varsigma_{j,\mathrm{II}}\,\}$, then

$$(U'_l)^T B_1 \equiv \left[\begin{array}{c|c} (U'_{i,\mathrm{I}})^T & 0 \\ \hline 0 & (U'_{j,\mathrm{II}})^T \end{array}\right] P^T B_1$$

  for some i and some j, and thus $(U'_l)^T B_1$ is of full row rank ($k_l = k_{i,\mathrm{I}} + k_{j,\mathrm{II}}$).

Thus the matrices $(U'_{i,\mathrm{I}})^T B_{1,\mathrm{I}}$, $(U'_{j,\mathrm{II}})^T B_{1,\mathrm{II}}$ are of full row rank for all i and all j *iff* the matrices $(U'_l)^T B_1$ are of full row rank for all l. □

Thus Lemma 6.1 says that the (de)composing (6.20) of independent problems having the core problem properties (G1)–(G3) preserves these properties (G1)–(G3). The following definitions introduce the terminology used further in the text.

Definition 6.3 (Composed problem)**.** *Let* $[\,B_{1,\mathrm{I}} \,|\, A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}} \,|\, A_{11,\mathrm{II}}\,]$ *be matrices representing two independent approximation problems. The problem* $[\,B \,|\, A\,]$ *obtained as in (6.20) is called composed problem.*

Definition 6.4 (Decomposable core problem)**.** *The core problem* $[\,B_1 \,|\, A_{11}\,]$ *for which there exist orthogonal matrices* P, Q, R *such that (6.20) holds and* $B_{1,\mathrm{I}}$, $B_{1,\mathrm{II}}$, $A_{11,\mathrm{I}}$, $A_{11,\mathrm{II}}$, *are nontrivial (each of them have at least one column) is called decomposable core problem.*

Subproblems $[\,B_{1,\mathrm{I}} \,|\, A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}} \,|\, A_{11,\mathrm{II}}\,]$ *of the given decomposable core problem are called independent subproblems within the core problem.*

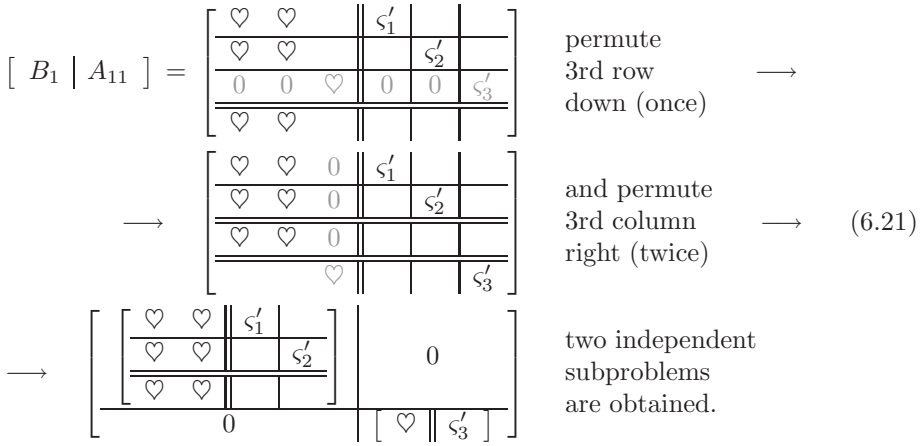
Remark 6.2. *The core problem with with the single right-hand side (i.e.* $\bar{d} = 1$*) is non decomposable. The decomposable core problem* $[\,B_1 \,|\, A_{11}\,]$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$ *can be decomposed into at most* $\bar{d}$ *non decomposable independent core subproblems, at most* $\bar{m} - \bar{n}$ *of them incompatible, at most* $\bar{d} - (\bar{m} - \bar{n})$ *of them compatible.*

The following sections give some examples of decomposable problems. It is worth to note here that the question how to identify whether the core problem is decomposable or not is, in general, not yet resolved.

6.4.1 Independent subproblems within the SVD form of a core problem

Let $[\, B_1 \,|\, A_{11} \,]$ be a core problem in the SVD form $B_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$, $A_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, see Chapter 4.

Because the matrix A_{11} is diagonal, for the existence of a decomposition of the given problem it is sufficient that the right-hand side matrix B_1 has a *chessboard structure* of zero and nonzero blocks, as in the following example:



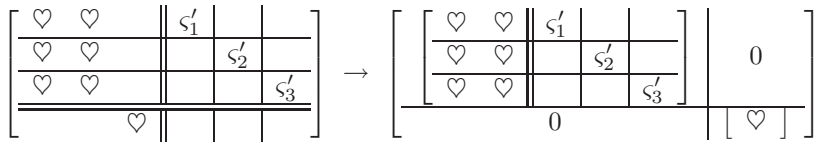
(6.21)

If the right-hand side has a chessboard structure, then the core problem is decomposed only using permutations, $B_1 = \Pi_1 \operatorname{diag}(\, B_{1,\mathrm{I}}\,,\, B_{1,\mathrm{II}}\,)\,\Pi_2^T$; the decomposition of A_{11} is trivial because it is diagonal.

Remark 6.3. *Recall that Definition 6.4 of the decomposable core problem does not cover the case*

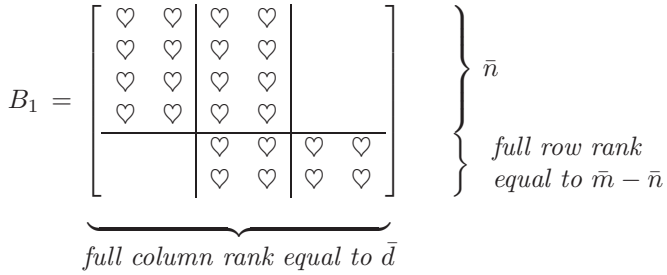
$$P^T \left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] = \left[\begin{array}{c|c||c} B_{1,\mathrm{I}} & 0 & A_{11,\mathrm{I}} \\ \hline 0 & B_{1,\mathrm{II}} & 0 \end{array}\right], \tag{6.22}$$

where $B_{1,\mathrm{II}}$ corresponds to the null space of A_{11}^T ($A_{11,\mathrm{II}}$ is trivial, it has no columns), e.g.,



(6.23)

where the third column of B_1 is orthogonal to $\mathcal{R}(A_{11})$. Motivated by such situation, we denote b_j, $j = 1, \ldots, \bar{d}$, the columns of B_1, then we distinguish between the following three cases: the column $b_j \in \mathcal{R}(A_{11})$ we call compatible column within the given core problem; the columns $b_j \perp \mathcal{R}(A_{11})$ we call $\mathcal{R}(A_{11})$-orthogonal column within the core problem; the column b_j, for which $b_j \notin \mathcal{R}(A_{11})$ and $b_j \not\perp \mathcal{R}(A_{11})$, we call mixed column within the core problem. In the SVD form of the core problem these three types of columns are given directly from the structure of nonzero components in B_1, e.g.,



(6.24)

The left block is formed of compatible columns, the middle block by mixed columns, and the right block by $\mathcal{R}(A_{11})$-orthogonal columns.

The situation (6.22) occurs iff B_1 contains only the compatible and the $\mathcal{R}(A_{11})$-orthogonal columns. It can be interpreted such that the problem $[\,B_1 \,|\, A_{11}\,]$ contains two independent subproblems: one is compatible (and thus it has a solution in the classical meaning) and the second with trivial system matrix (the second does not influence the solution, it influences only the correction).

6.4.2 Orthogonal transformation preserving the SVD form of the core problem

In both examples (6.21), (6.23), the decomposing of the given problem is guaranteed by the *chessboard structure* of zero and nonzero blocks in B_1. The right-hand side B_1 having chessboard structure always can be transformed into a block diagonal form $B_1 = \Pi_1 \operatorname{diag}(B_{1,\mathrm{I}}, B_{1,\mathrm{II}})\,\Pi_2^T$, using some permutation matrices Π_1, Π_2.

Any given decomposable core problem as well as all the independent subproblems within this decomposable core problem can be in the SVD form. Thus for a decomposable core problem in the SVD form there exists an orthogonal transformation preserving the SVD form, i.e. preserving properties (S1)–(S3), such that it reveals the chessboard structure of the right-hand side. This section presents a class of orthogonal transformations preserving the SVD form of a core problem.

Consider a core problem $A_{11}\,X_1 \approx B_1$ in the SVD form (4.25), i.e.

$$\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] = \left[\begin{array}{c|c|c|c} D_1 & \varsigma_1' I_{r_1} & & \\ \hline \vdots & & \ddots & \\ \hline D_k & & & \varsigma_k' I_{r_k} \\ \hline\hline D_{k+1} & & 0 & \end{array}\right] \in \mathbb{R}^{\bar m \times (\bar n + \bar d)}\,,$$

where the matrix $B_1 = [\,D_1^T, \ldots, D_k^T \,|\, D_{k+1}^T\,]^T \in \mathbb{R}^{\bar m \times \bar d}$ has orthogonal nonzero columns ordered in a nonincreasing sequence with respect to their norms, and the blocks $D_j \in \mathbb{R}^{r_j \times \bar d}$ have orthogonal nonzero rows ordered in a nonincreasing sequence with respect of their norms, for $j = 1, \ldots, k+1$. Further consider an orthogonal transformation of the form

$$P^T \left[\begin{array}{c|c|c|c} D_1 & \varsigma_1' I_{r_1} & & \\ \hline \vdots & & \ddots & \\ \hline D_k & & & \varsigma_k' I_{r_k} \\ \hline\hline D_{k+1} & & 0 & \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right], \tag{6.25}$$

where $P^{-1} = P^T$, $Q^{-1} = Q^T$, $R^{-1} = R^T$, which preserves the properties (S1)–(S3). First, because the transformation (6.25) preserves (S1), i.e. $P^T A_{11}\,Q = A_{11}$, matrices P, Q must be block diagonal such that

$$\begin{aligned} P^T A_{11}\,Q &\equiv (\operatorname{diag}(Q, Q_{k+1}))^T A_{11}\,Q \\ &= \left[\begin{array}{c|c|c|c} Q_1 & & & \\ \hline & \ddots & & \\ \hline & & Q_k & \\ \hline\hline & & & Q_{k+1} \end{array}\right]^T \left[\begin{array}{c|c|c} \varsigma_1' I_{r_1} & & \\ \hline & \ddots & \\ \hline & & \varsigma_k' I_{r_k} \\ \hline\hline & 0 & \end{array}\right] \left[\begin{array}{c|c|c} Q_1 & & \\ \hline & \ddots & \\ \hline & & Q_k \end{array}\right], \end{aligned}$$

where $Q_j^{-1} = Q_j^T$, $Q_j \in \mathbb{R}^{r_j \times r_j}$, $j = 1, \ldots, k+1$.

Here it is worth to recall the following assertion. Let M be a matrix with mutually orthogonal nonzero columns (i.e. M is of full column rank) ordered in the

nonincreasing sequence with respect to their norms. Then the norms of it columns represent singular values of M. The matrix of left singular vectors of M contains normalized columns of M, and the matrix of right singular vectors is the identity matrix. Analogously for the matrix with mutually orthogonal nonzero rows ordered in a nonincreasing sequence with respect to their norms (e.g. M^T).

Further we focus on preserving (S2) property. Let $B_1 = U_B \Sigma_B I_{\tilde{d}}$ be the SVD of B_1. The transformation (6.25) preserves (S2), i.e. the matrix $P^T B_1 R = (P^T U_B R)(R^T \Sigma_B R)$ has orthogonal nonzero columns ordered in a nonincreasing sequence with respect to their norms. The matrix P does not change norms of columns of B_1 as well as angles between columns of B_1, thus preserving (S2) does not restrict the matrix P. But for the matrix R it is allowed to form only (orthogonal) linear combinations of the columns of B_1 with the same norm, i.e. $R^T \Sigma_B R = \Sigma_B$. Consequently the matrix R must be block diagonal such that

$$P^T B_1 R = P^T B_1 \operatorname{diag}(R_1, \dots, R_\ell),$$

where $R_j^{-1} = R_j^T$, $R_j \in \mathbb{R}^{s_j \times s_j}$, s_j is the multiplicity of the jth greatest singular value of B_1, $j = 1, \dots, \ell$, and ℓ is the number of distinct singular values of B_1.

And, finally, we focus on preserving (S3) property; the considerations here are analogous to the previous case. For all D_j blocks, let $D_j = I_{r_j} \Sigma_j V_j^T$ be the SVD of D_j. The transformation (6.25) preserves (S3), i.e. the matrix $Q_j^T D_j R = (Q_j^T \Sigma_j Q_j)(Q_j^T V_j^T R)$ has orthogonal nonzero rows ordered in a nonincreasing sequence with respect to their norms. The matrix R does not change norms of rows of D_j as well as angles between rows of D_j, thus preserving (S3) does not restrict the matrix R. But for the matrix Q_j it is allowed to form only (orthogonal) linear combinations of the rows of D_j with the same norm, i.e. $Q_j^T \Sigma_j Q_j = \Sigma_j$. Consequently the matrix Q_j must be block diagonal such that

$$Q_j^T D_j R = (\operatorname{diag}(Q_{1,j}, \dots, Q_{\varkappa_j,j}))^T D_j R,$$

where $Q_{i,j}^{-1} = Q_{i,j}^T$, $Q_{i,j} \in \mathbb{R}^{t_{i,j} \times t_{i,j}}$, $t_{i,j}$ is the multiplicity of the ith greatest singular value of D_j, $i = 1, \dots, \varkappa_j$, and $\varkappa_j$ is the number of distinct singular values of D_j. All for $j = 1, \dots, k+1$.

The structure of the whole transformation preserving the SVD form of a core problem is shown on Figure 6.4.

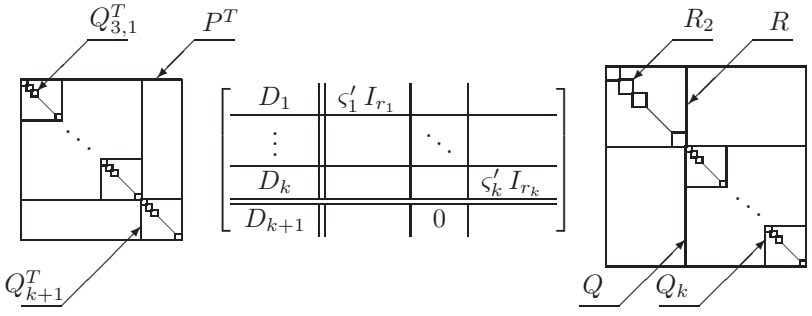


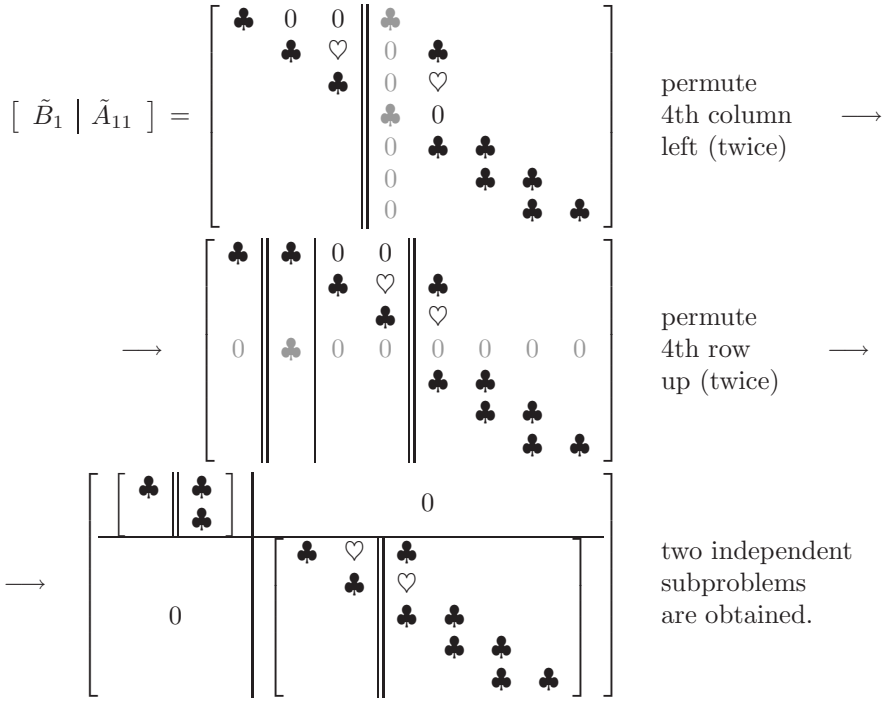
Figure 6.4: Structure of transformations preserving the SVD form of a core problem.

Summarized, if the core problem is decomposable, then there exists an orthogonal transformation with the described structure, see Figure 6.4, such that this transformation reveals the chessboard structure of the right-hand side. After a permutation is applied on the suitably transformed problem, all independent subproblems are revealed.

6.4.3 Independent subproblems within the banded form of a core problem

Let $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ be a decomposable core problem in the banded form $\tilde{B}_1 \in \mathbb{R}^{\bar{m}\times\bar{d}}$, $\tilde{A}_{11} \in \mathbb{R}^{\bar{m}\times\bar{n}}$, see Chapter 5.

When the whole matrix $[\,\tilde{B}_1\,|\,\tilde{A}_{11}\,]$ has the *chessboard structure*, it can be decomposed, see the following example:



This example uses the information that some ♡ components (β components, e.g. in (5.16)) are equal to zero.

Recall that in Sections 6.4.1 and 6.4.2 it is used the fact that the system matrix A_{11} is diagonal. Thus it is sufficient to reveal only a chessboard structure of the right-hand side matrix B_1 for decomposing the core problem.

Here it is worth to recall the SVD preprocessing of the right-hand side, see Section 5.1. Algorithm 5.1 applied on a problem with the SVD-preprocessed right-hand side gives $\tilde{B}_1$ diagonal (all β coefficients in the right-hand side are equal to zero). Consequently it is sufficient to reveal only a chessboard structure of the system matrix $\tilde{A}_{11}$ for decomposing the core problem. Thus the SVD preprocessing seems to be useful in practical computation.

6.5 Solution of the decomposable core problem

This section focuses on solving decomposable core problems. Thus it is closely connected with Section 6.3. We introduce this section by the following example.

Example 6.1. *Let* $[\,B_{1,\mathrm{I}}\,|\,A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}}\,|\,A_{11,\mathrm{II}}\,]$ *be matrices representing two independent approximation problems. Let* $U_{\mathrm{I}}\,\Sigma_{\mathrm{I}}\,V_{\mathrm{I}}^T$ *and* $U_{\mathrm{II}}\,\Sigma_{\mathrm{II}}\,V_{\mathrm{II}}^T$ *be the SVD of* $[\,B_{1,\mathrm{I}}\,|\,A_{11,\mathrm{I}}\,]$ *and* $[\,B_{1,\mathrm{II}}\,|\,A_{11,\mathrm{II}}\,]$*, respectively, both with partitioning given by (6.8), see also (3.2)–(3.3). Consider that the value of* q*, see (6.7), is equal to zero for both*

problems, thus matrices $V_{12,\mathrm{I}}$, $V_{12,\mathrm{II}}$ (we omit the upper index $^{(0)}$) are square. Thus

$$\begin{aligned}
\left[\begin{array}{c|c} B_{1,\mathrm{I}} & A_{11,\mathrm{I}} \end{array}\right] &= U_{\mathrm{I}} \left[\begin{array}{c|c} \Sigma_{1,\mathrm{I}} & \Sigma_{2,\mathrm{I}} \end{array}\right] \left[\begin{array}{c|c} V_{11,\mathrm{I}} & V_{12,\mathrm{I}} \\ \hline V_{21,\mathrm{I}} & V_{22,\mathrm{I}} \end{array}\right]^T , \\
\left[\begin{array}{c|c} B_{1,\mathrm{II}} & A_{11,\mathrm{II}} \end{array}\right] &= U_{\mathrm{II}} \left[\begin{array}{c|c} \Sigma_{1,\mathrm{II}} & \Sigma_{2,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} V_{11,\mathrm{II}} & V_{12,\mathrm{II}} \\ \hline V_{21,\mathrm{II}} & V_{22,\mathrm{II}} \end{array}\right]^T .
\end{aligned} \tag{6.26}$$

Consider further that both these matrices $V_{12,\mathrm{I}}$, $V_{12,\mathrm{II}}$ are nonsingular. In other words, we consider that both problems $[\,B_{1,\mathrm{I}}\,|\,A_{11,\mathrm{I}}\,]$ and $[\,B_{1,\mathrm{II}}\,|\,A_{11,\mathrm{II}}\,]$ have the unique TLS solutions

$$X_{\mathrm{I}} \equiv V_{22,\mathrm{I}}\, V_{12,\mathrm{I}}^{-1}, \qquad X_{\mathrm{II}} \equiv V_{22,\mathrm{II}}\, V_{12,\mathrm{II}}^{-1}, \tag{6.27}$$

see also Section 3.2.1; these solution are also computed by the TLS algorithm (Algorithm 3.1). Finally, consider the composed problem (6.20) with $P = I$, $Q = I$, $R = I$,

$$\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] \equiv \left[\begin{array}{c|c||c|c} B_{1,\mathrm{I}} & 0 & A_{11,\mathrm{I}} & 0 \\ \hline 0 & B_{1,\mathrm{II}} & 0 & A_{11,\mathrm{II}} \end{array}\right] . \tag{6.28}$$

Denote $[\,B_1\,|\,A_{11}\,] = U\,\Sigma\,V^T$ the SVD of the composed problem (6.28). Then using (6.26) gives the following decomposition

$$\begin{aligned}
\left[\begin{array}{c|c} B_1 & A_{11} \end{array}\right] &= \left[\begin{array}{c|c||c|c} B_{1,\mathrm{I}} & A_{11,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & B_{1,\mathrm{II}} & A_{11,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c|c|c} I & 0 & 0 & 0 \\ \hline 0 & 0 & I & 0 \\ \hline 0 & I & 0 & 0 \\ \hline 0 & 0 & 0 & I \end{array}\right] \\
&= \left[\begin{array}{c|c} U_{\mathrm{I}} & 0 \\ \hline 0 & U_{\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c||c|c} \Sigma_{1,\mathrm{I}} & \Sigma_{2,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & \Sigma_{1,\mathrm{II}} & \Sigma_{2,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c||c|c} V_{11,\mathrm{I}} & V_{12,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & V_{11,\mathrm{II}} & V_{12,\mathrm{II}} \\ \hline\hline V_{21,\mathrm{I}} & V_{22,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & V_{21,\mathrm{II}} & V_{22,\mathrm{II}} \end{array}\right]^T
\end{aligned}$$

which represents (after permutation which sorts singular values of Σ_{I} and Σ_{II} into one nonincreasing sequence) the SVD $[\,B_1\,|\,A_{11}\,] = U\,\Sigma\,V^T$. This permutation also changes the order of left and right singular vectors.

The question is whether a solution of the composed problem (6.28) is given by the solutions (6.27) of its independent subproblems.

In order to answer this question we focus on two special relationships between the singular values of Σ_{I} and Σ_{II} that may occur (recall that the problems are independent):

(a) *Assume* $\sigma_{\min}(\Sigma_{1,\mathrm{II}}) > \sigma_{\max}(\Sigma_{2,\mathrm{I}})$ *and* $\sigma_{\min}(\Sigma_{1,\mathrm{I}}) > \sigma_{\max}(\Sigma_{2,\mathrm{II}})$, *then*

$$V = \left[\begin{array}{c|c||c|c} V_{11,\mathrm{I}} & 0 & V_{12,\mathrm{I}} & 0 \\ \hline 0 & V_{11,\mathrm{II}} & 0 & V_{12,\mathrm{II}} \\ \hline\hline V_{21,\mathrm{I}} & 0 & V_{22,\mathrm{I}} & 0 \\ \hline 0 & V_{21,\mathrm{II}} & 0 & V_{22,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} \Pi_1 & 0 \\ \hline 0 & \Pi_2 \end{array}\right] \equiv \left[\begin{array}{c|c} V_{11}\,\Pi_1 & V_{12}\,\Pi_2 \\ \hline V_{21}\,\Pi_1 & V_{22}\,\Pi_2 \end{array}\right] ,$$

where Π_1, Π_2 represent column permutations; Π_1 sorts the singular values of $\Sigma_{1,\mathrm{I}}$ and $\Sigma_{1,\mathrm{II}}$ into one nonincreasing sequence, and Π_2 sorts the singular values of $\Sigma_{2,\mathrm{I}}$ and $\Sigma_{2,\mathrm{II}}$ into one nonincreasing sequence. The upper right block $V_{12}\,\Pi_2$ of the matrix V is square nonsingular. Because the assumed inequalities are strict, the composed problem has the unique TLS solution

$$X \equiv \left[\begin{array}{c|c} V_{22,\mathrm{I}} & 0 \\ \hline 0 & V_{22,\mathrm{II}} \end{array}\right] \Pi_2\,\Pi_2^T \left[\begin{array}{c|c} V_{12,\mathrm{I}} & 0 \\ \hline 0 & V_{12,\mathrm{II}} \end{array}\right]^{-1} = \left[\begin{array}{c|c} X_{\mathrm{I}} & 0 \\ \hline 0 & X_{\mathrm{II}} \end{array}\right] ,$$

which is also computed by the TLS algorithm (Algorithm 3.1).

(b) *Assume* $\sigma_{\max}(\Sigma_{2,\mathrm{I}}) \geq \sigma_{\max}(\Sigma_{1,\mathrm{II}})$, *then*

$$V = \left[\begin{array}{c|c||c|c} V_{11,\mathrm{I}} & V_{12,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & V_{11,\mathrm{II}} & V_{12,\mathrm{II}} \\ \hline\hline V_{21,\mathrm{I}} & V_{22,\mathrm{I}} & 0 & 0 \\ \hline 0 & 0 & V_{21,\mathrm{II}} & V_{22,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} I & 0 \\ \hline 0 & \Pi \end{array}\right] \equiv \left[\begin{array}{c|c} V_{11}^{(\kappa)} & V_{12}^{(\kappa)}\,\Pi \\ \hline V_{21}^{(\kappa)} & V_{22}^{(\kappa)}\,\Pi \end{array}\right],$$

where Π *represents column permutation, which sorts the singular values of* $\Sigma_{2,\mathrm{I}}$, $\Sigma_{1,\mathrm{II}}$, *and* $\Sigma_{2,\mathrm{II}}$ *into one nonincreasing sequence. The upper right block* $V_{12}^{(\kappa)}\,\Pi$ *is rectangular of full row rank (number of its columns is* κ + *number of its rows,* $\kappa > 0$*). The TLS algorithm (Algorithm 3.1) returns as solution the matrix*

$$\begin{aligned} X &\equiv \left[\begin{array}{c|c|c} V_{22,\mathrm{I}} & 0 & 0 \\ \hline 0 & V_{21,\mathrm{II}} & V_{22,\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c|c} V_{12,\mathrm{I}} & 0 & 0 \\ \hline 0 & V_{11,\mathrm{II}} & V_{12,\mathrm{II}} \end{array}\right]^{\dagger} \\ &= \left[\begin{array}{c|c} V_{22,\mathrm{I}}\,V_{12,\mathrm{I}}^{-1} & 0 \\ \hline 0 & [\,V_{21,\mathrm{II}}\,|\,V_{22,\mathrm{II}}\,]\,[\,V_{11,\mathrm{II}}\,|\,V_{12,\mathrm{II}}\,]^{\dagger} \end{array}\right] \\ &= \left[\begin{array}{c|c} X_{\mathrm{I}} & 0 \\ \hline 0 & 0 \end{array}\right]. \end{aligned}$$

This occurs, e.g., when one subproblem dominates the other subproblem (i.e. all the singular values of one subproblem are larger than all singular values of the second subproblem).

The example above has the following corollary.

Corollary 6.3. *The solution of a composed problem* $[\,B_1\,|\,A_{11}\,]$*, computed by Algorithm 3.1, differs from the solution composed from independent solutions, computed by Algorithm 3.1 too, of the independent subproblems within* $[\,B_1\,|\,A_{11}\,]$.

Recall that the transformation from the original problem $[\,B\,|\,A\,]$ to its core problem $[\,B_1\,|\,A_{11}\,]$ preserves the solution obtained by the TLS algorithm (Algorithm 3.1), see Corollary 6.2. Corollary 6.3 gives that the decomposing of $[\,B_1\,|\,A_{11}\,]$ (if it is decomposable) and treating subproblems independently does not preserve the solution obtained the by the TLS algorithm (Algorithm 3.1).

Assume for this moment that the composed problem $[\,B_1\,|\,A_{11}\,]$ from Example 6.1 represents a decomposable core problem within $[\,B\,|\,A\,]$, i.e. $[\,B_1\,|\,A_{11}\,]$ has properties (G1)–(G3). (Recall the composed problem $[\,B_1\,|\,A_{11}\,]$ has properties (G1)–(G3) if and only if both independent subproblems $[\,B_{1,\mathrm{I}}\,|\,A_{11,\mathrm{I}}\,]$ and $[\,B_{1,\mathrm{II}}\,|\,A_{11,\mathrm{II}}\,]$ have properties (G1)–(G3), by Lemma 6.1.)

The core problem $[\,B_1\,|\,A_{11}\,]$ belongs in the set $\mathcal{F}_1$ (see Figure 3.5, p. 55) if and only if Algorithm 3.1 applied on $[\,B_1\,|\,A_{11}\,]$ returns the unique TLS solution, by Theorem 6.2; it is proved in Section 6.3. In Section 6.3 there remains an unanswered question, whether there exists a core problem which does not belong in the set $\mathcal{F}_1$. Example 6.1 partially answers this question; further in the text, we show on simple examples that the (decomposable) core problem (of the form (6.28)) can be of the 2nd class (in the case (b), see Example 6.1). Recall that a problem is of the 2nd class if the matrix $V^{(q)}$, see (6.7)–(6.8), is rank deficient.

Example 6.2. *Using the same notation as in Example 6.1, let*

$$\left[\begin{array}{c|c} B_{1,\mathrm{I}} & A_{11,\mathrm{I}} \end{array}\right] = \left[\begin{array}{c|cc} \heartsuit & \sigma'_{1,\mathrm{I}} & 0 \\ \heartsuit & 0 & \sigma'_{2,\mathrm{I}} \\ \hline \heartsuit & 0 & 0 \end{array}\right],$$

$$\left[\begin{array}{c|c} B_{1,\mathrm{II}} & A_{11,\mathrm{II}} \end{array}\right] = \left[\begin{array}{c|cc} \heartsuit & \sigma'_{1,\mathrm{II}} & 0 \\ \heartsuit & 0 & \sigma'_{2,\mathrm{II}} \\ \hline \heartsuit & 0 & 0 \end{array}\right],$$

have properties (G1)–(G3). Then the singular values of $[\,B_{1,\mathrm{I}}\,|\,A_{11,\mathrm{I}}\,]$, $[\,B_{1,\mathrm{II}}\,|\,A_{11,\mathrm{II}}\,]$ *are simple. Let*

$$\left[\begin{array}{c|c} B_{1,\mathrm{I}} & A_{11,\mathrm{I}} \end{array}\right] = U_{\mathrm{I}}\ \mathrm{diag}\,(\,\sigma_{1,\mathrm{I}}\,,\,\sigma_{2,\mathrm{I}}\,,\,\sigma_{3,\mathrm{I}}\,) \left[\begin{array}{cc|c} v_{11,\mathrm{I}} & v_{12,\mathrm{I}} & v_{13,\mathrm{I}} \\ \hline v_{21,\mathrm{I}} & v_{22,\mathrm{I}} & v_{23,\mathrm{I}} \\ v_{31,\mathrm{I}} & v_{23,\mathrm{I}} & v_{33,\mathrm{I}} \end{array}\right],$$

$$\left[\begin{array}{c|c} B_{1,\mathrm{II}} & A_{11,\mathrm{II}} \end{array}\right] = U_{\mathrm{II}}\ \mathrm{diag}\,(\,\sigma_{1,\mathrm{II}}\,,\,\sigma_{2,\mathrm{II}}\,,\,\sigma_{3,\mathrm{II}}\,) \left[\begin{array}{cc|c} v_{11,\mathrm{II}} & v_{12,\mathrm{II}} & v_{13,\mathrm{II}} \\ \hline v_{21,\mathrm{II}} & v_{22,\mathrm{II}} & v_{23,\mathrm{II}} \\ v_{31,\mathrm{II}} & v_{23,\mathrm{II}} & v_{33,\mathrm{II}} \end{array}\right],$$

be the SVD decompositions and assume $\sigma_{3,\mathrm{I}} > \sigma_{1,\mathrm{II}}$, *i.e.*

$$\sigma_{1,\mathrm{I}} > \sigma_{2,\mathrm{I}} > \sigma_{3,\mathrm{I}} > \sigma_{1,\mathrm{II}} > \sigma_{2,\mathrm{II}} > \sigma_{3,\mathrm{II}}\,. \tag{6.29}$$

Because both these problems can be treated as core problems with single right-hand sides, both have the unique TLS solution, i.e. $v_{13,\mathrm{I}} \neq 0$, $v_{13,\mathrm{II}} \neq 0$. *The matrix of the right singular vectors of the composed problem (6.28) is*

$$V = \left[\begin{array}{cccc|cc} v_{11,\mathrm{I}} & v_{12,\mathrm{I}} & v_{13,\mathrm{I}} & 0 & 0 & 0 \\ 0 & 0 & 0 & v_{11,\mathrm{II}} & v_{12,\mathrm{II}} & v_{13,\mathrm{II}} \\ \hline v_{21,\mathrm{I}} & v_{22,\mathrm{I}} & v_{23,\mathrm{I}} & 0 & 0 & 0 \\ v_{31,\mathrm{I}} & v_{23,\mathrm{I}} & v_{33,\mathrm{I}} & 0 & 0 & 0 \\ 0 & 0 & 0 & v_{21,\mathrm{II}} & v_{22,\mathrm{II}} & v_{23,\mathrm{II}} \\ 0 & 0 & 0 & v_{31,\mathrm{II}} & v_{23,\mathrm{II}} & v_{33,\mathrm{II}} \end{array}\right] \equiv \left[\begin{array}{c|c} V_{11}^{(q)} & V_{12}^{(q)} \\ \hline V_{21}^{(q)} & V_{22}^{(q)} \end{array}\right],$$

with $q = 0$ *(see (6.7)); and the matrix* $V_{12}^{(q)}$ *is rank deficient. Problems with the rank deficient matrix* $V_{12}^{(q)}$ *are called problems of the 2nd class (the set of these problems is denoted* $\mathscr{S}$*), see Figure 3.5, p. 55.*

Example 6.3. *Assuming, e.g.,*

$$\sigma_{1,\mathrm{I}} > \sigma_{2,\mathrm{I}} > \sigma_{1,\mathrm{II}} > \sigma_{2,\mathrm{II}} > \sigma_{3,\mathrm{I}} > \sigma_{3,\mathrm{II}}$$

instead of (6.29), in Example 6.2, gives

$$V_{12}^{(q)} = \left[\begin{array}{c|c} v_{13,\mathrm{I}} & 0 \\ 0 & v_{13,\mathrm{II}} \end{array}\right] \equiv \left[\begin{array}{c|c} W^{(q,e)} & V_{12}^{(-e)} \end{array}\right],$$

the decomposable core problem of the 1st class belonging in the set $\mathscr{F}_1$ *(*$q = 0$, $e = 1$*). Assuming, e.g.,*

$$\sigma_{1,\mathrm{I}} > \sigma_{2,\mathrm{I}} > \sigma_{1,\mathrm{II}} > \sigma_{3,\mathrm{I}} = \sigma_{2,\mathrm{II}} > \sigma_{3,\mathrm{II}}$$

instead of (6.29) gives

$$V_{12}^{(q)} = \left[\begin{array}{cc|c} v_{13,\mathrm{I}} & 0 & 0 \\ 0 & v_{12,\mathrm{II}} & v_{13,\mathrm{II}} \end{array}\right] \equiv \left[\begin{array}{c|c} W^{(q,e)} & V_{12}^{(-e)} \end{array}\right],$$

the decomposable core problem of the 1st class belonging in the set $\mathscr{F}_2$ ($q = 1$, $e = 1$; note that $v_{12,\mathrm{II}} \neq 0$, by Theorem 6.1).

In the last case consider the decomposable core problem, composed from three problems having properties (G1)–(G3) with the single right-hand sides, all having the same dimensions as in Example 6.2; we use analogous notation. Assuming, e.g.,

$$\sigma_{1,\mathrm{I}} > \sigma_{2,\mathrm{I}} > \sigma_{1,\mathrm{II}} > \sigma_{2,\mathrm{II}} > \sigma_{1,\mathrm{III}} > \sigma_{3,\mathrm{I}} = \sigma_{3,\mathrm{II}} > \sigma_{2,\mathrm{III}} > \sigma_{3,\mathrm{III}}\,,$$

gives

$$V_{12}^{(q)} = \left[\begin{array}{ccc|cc} v_{13,\mathrm{I}} & 0 & & 0 & 0 \\ 0 & v_{13,\mathrm{II}} & & 0 & 0 \\ 0 & 0 & & v_{12,\mathrm{III}} & v_{13,\mathrm{III}} \end{array}\right] \equiv \left[\begin{array}{c|c} W^{(q,e)} & V_{12}^{(-e)} \end{array}\right],$$

where $v_{13,\mathrm{I}} \neq 0$, $v_{13,\mathrm{II}} \neq 0$, and $v_{13,\mathrm{III}} \neq 0$. Thus $V_{12}^{(q)}$ ($q = 1$) is of full row rank, and $V_{12}^{(-e)}$ ($e = 1$) is rank deficient. The this decomposable core problem is of 1st class belonging in the set $\mathscr{F}_3$. See also Figure 3.5, p. 55.

Examples 6.1, 6.2, and 6.3 give the following corollary.

Corollary 6.4. *The core problem $[\,B_1 \,|\, A_{11}\,]$ (in particular the decomposable core problem) may be of the 1st class (belonging to any of the sets $\mathscr{F}_1$, $\mathscr{F}_2$, or $\mathscr{F}_3$), as well as of the 2nd class.*

The solution of the core problem computed by Algorithm 3.1 may represent either the unique TLS solution (see Section 3.2.1), if $[\,B_1 \,|\, A_{11}\,]$ belongs in the set $\mathscr{F}_1$; or the nongeneric solution, if $[\,B_1 \,|\, A_{11}\,]$ belongs in the set $\mathscr{F}_2 \cup \mathscr{F}_3 \cup \mathscr{S}$. See also Figure 3.5, p. 55.

6.5.1 Composed solution of a composed problem

This section introduces the composed solution for composed problems, (e.g. for decomposable core problems).

Example 6.1 shows that the solution computed by the TLS algorithm (Algorithm 3.1) applied to a composed problem can be different from the solution obtained by the application of the TLS algorithm on the independent subproblems within the composed problem, in general (see Example 6.1, case (b)). The following lemma compares norms of these solutions.

Lemma 6.2. *Let $[\,B_{\mathrm{I}} \,|\, A_{\mathrm{I}}\,]$ and $[\,B_{\mathrm{II}} \,|\, A_{\mathrm{II}}\,]$ be two independent subproblems of the composed problem*

$$\left[\begin{array}{c|c} B & A \end{array}\right] \equiv P \left[\begin{array}{c|c||c|c} B_{\mathrm{I}} & 0 & A_{\mathrm{I}} & 0 \\ \hline 0 & B_{\mathrm{II}} & 0 & A_{\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right]^T,$$

not necessarily having properties (G1)–(G3). Denote X_{I}, $[\,G_{\mathrm{I}} \,|\, E_{\mathrm{I}}\,]$, X_{II}, $[\,G_{\mathrm{II}} \,|\, E_{\mathrm{II}}\,]$, and X, $[\,G \,|\, E\,]$ the solutions with the corresponding corrections of both subproblems and of the composed problem, respectively, all computed via the TLS algorithm (Algorithm 3.1). Further denote

$$X_{\textsc{Comp.}} \equiv Q \left[\begin{array}{c|c} X_{\mathrm{I}} & 0 \\ \hline 0 & X_{\mathrm{II}} \end{array}\right] R^T, \tag{6.30}$$

the composed solution, with the corresponding correction

$$\left[\begin{array}{c|c} G_{\textsc{Comp.}} & E_{\textsc{Comp.}} \end{array}\right] \equiv P \left[\begin{array}{c|c||c|c} G_{\mathrm{I}} & 0 & E_{\mathrm{I}} & 0 \\ \hline 0 & G_{\mathrm{II}} & 0 & E_{\mathrm{II}} \end{array}\right] \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right]^T. \tag{6.31}$$

Then

$$\begin{aligned} \|X\| &\leq \|X_{\text{Comp.}}\| \equiv \max\{\,\|X_{\text{I}}\|\,,\,\|X_{\text{II}}\|\,\}\,, \\ \|X\|_F &\leq \|X_{\text{Comp.}}\|_F \equiv \left(\,\|X_{\text{I}}\|_F^2 + \|X_{\text{II}}\|_F^2\,\right)^{1/2}. \end{aligned} \tag{6.32}$$

In words, the solution of a decomposable problem computed by the TLS algorithm directly can not have larger norm than the solution composed from solutions of its independent core subproblems.

Proof. Denote

$$\left[\begin{array}{c|c} B_{\text{I}} & A_{\text{I}} \end{array}\right] = U_{\text{I}} \left[\begin{array}{c|c} \Sigma_{1,\text{I}}^{(\kappa)} & \Sigma_{2,\text{I}}^{(\kappa)} \end{array}\right] \left[\begin{array}{c|c} V_{11,\text{I}}^{(\kappa)} & V_{12,\text{I}}^{(\kappa)} \\ \hline V_{21,\text{I}}^{(\kappa)} & V_{22,\text{I}}^{(\kappa)} \end{array}\right]^T,$$

$$\left[\begin{array}{c|c} B_{\text{II}} & A_{\text{II}} \end{array}\right] = U_{\text{II}} \left[\begin{array}{c|c} \Sigma_{1,\text{II}}^{(\nu)} & \Sigma_{2,\text{II}}^{(\nu)} \end{array}\right] \left[\begin{array}{c|c} V_{11,\text{II}}^{(\nu)} & V_{12,\text{II}}^{(\nu)} \\ \hline V_{21,\text{II}}^{(\nu)} & V_{22,\text{II}}^{(\nu)} \end{array}\right]^T,$$

the SVD decompositions of the independent subproblems. The values of κ and ν are chosen by the TLS algorithm such that, they are smallest for which:

(i) $V_{12,\text{I}}^{(\kappa)}$ and $V_{12,\text{II}}^{(\nu)}$ are rectangular of full row rank; and

(ii) $\sigma_{\min}(\Sigma_{1,\text{I}}^{(\kappa)}) > \sigma_{\max}(\Sigma_{2,\text{I}}^{(\kappa)})$ and $\sigma_{\min}(\Sigma_{1,\text{II}}^{(\nu)}) > \sigma_{\max}(\Sigma_{2,\text{I}}^{(\nu)})$.

The solutions of individual subproblems computed by the TLS algorithm and the composed solution are

$$X_{\text{I}} \equiv -V_{22,\text{I}}^{(\kappa)} V_{12,\text{I}}^{(\kappa)\dagger}, \qquad X_{\text{II}} \equiv -V_{22,\text{II}}^{(\nu)} V_{12,\text{II}}^{(\nu)\dagger},$$

$$X_{\text{Comp.}} \equiv -\operatorname{diag}\left(\,V_{22,\text{I}}^{(\kappa)} V_{12,\text{I}}^{(\kappa)\dagger}\,,\,V_{22,\text{II}}^{(\nu)} V_{12,\text{II}}^{(\nu)\dagger}\,\right),$$

for the Frobenius and 2-norm of these solutions see Lemma 3.2. Further, the matrix of right singular vector of the composed problem $[\,B\,|\,A\,]$ is

$$V \equiv \left[\begin{array}{c|c} R & 0 \\ \hline 0 & Q \end{array}\right] \left[\begin{array}{c|c||c|c} V_{11,\text{I}}^{(\kappa)} & V_{12,\text{I}}^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & V_{11,\text{II}}^{(\nu)} & V_{12,\text{II}}^{(\nu)} \\ \hline\hline V_{21,\text{I}}^{(\kappa)} & V_{22,\text{I}}^{(\kappa)} & 0 & 0 \\ \hline 0 & 0 & V_{21,\text{II}}^{(\nu)} & V_{22,\text{II}}^{(\nu)} \end{array}\right] \Pi,$$

where Π is a permutation which rearranges columns of the matrix while sorting the singular values of both subproblems in one nonincreasing sequence. The TLS algorithm applied on the composed problem chooses the integer μ analogously as for the subproblems; such that (i) the upper right block $V_{12}^{(\mu)}$ of V is of full column rank; and (ii) the singular value corresponding to the last column of $V_{11}^{(\mu)}$ is strictly bigger than the singular value corresponding to the first column of $V_{12}^{(\mu)}$.

Thus $V_{11}^{(\mu)}$ must contain at least both matrices $V_{12,\text{I}}^{(\kappa)}$, $V_{12,\text{II}}^{(\nu)}$ as submatrices, but it can contain some extra columns (depending on the relationship between singular values of individual subproblems), i.e. $\mu \geq \nu + \kappa$. Then the solution computed by the TLS algorithm applied on the composed problem is

$$X \equiv -V_{22}^{(\mu)} V_{12}^{(\mu)\dagger},$$

for the Frobenius and 2-norm of these solutions see Lemma 3.2. Adding the extra columns (discussed above) to the matrix

$$\operatorname{diag}\left(\,V_{12,\text{I}}^{(\kappa)}\,,\,V_{12,\text{II}}^{(\nu)}\,\right)$$

in order to obtain $V_{12}^{(\mu)}$ causes, by Theorem 2.1 (interlacing theorem for singular values), that the singular values can not decrease. The Frobenius and 2-norm of the solution computed by the pseudoinverse of this matrix can not increase which gives (6.32). See the proof of Lemma 3.3, see also the proof of Lemma 3.2. □

In Example 6.1, case (b), it is shown that the solution of the decomposable problem can not contain all the information which is contained in the solutions of all its subproblems, in general. Consequently, when we want to obtain the solutions of all subproblems within the composed problem, then these subproblems must be solved independently. The fact that some information can be removed when the independent subproblems are solved together may be a consequence of the regularizing properties of the nongeneric solution (recall the truncated TLS concept, see also Lemma 3.12).

The solution $X_{\text{COMP.}}$ defined as in (6.30) is important in the case when all its (non decomposable) core subproblems have unique TLS solution (as in Example 6.1; see also Section 3.2.1). Whether each non decomposable core problem has the unique TLS solution is not known yet. When this hypothesis is true and each non decomposable core problem has the unique TLS solution, then any given problem can be reduced to its core problem, which can be further decomposed into (non decomposable) subproblems within it. Solving these subproblems independently, composing, and transforming back to the coordinates of the original problem give a well defined and unique solution.

The core problem in the multiple right-hand side case can be of the 1st class (belonging to any of the sets $\mathscr{F}_1$, $\mathscr{F}_2$, or $\mathscr{F}_3$), as well as of the 2nd class, see Corollary 6.4. Thus the TLS algorithm (Algorithm 3.1) applied on a core problem can return a nongeneric solution. From a statistical point of view, this nongenericity of the composed core problem may be a consequence of the fact that the part $[\, B_{1,\mathrm{I}}^T \,|\, 0 \,]^T$ of the right-hand side is not correlated with the column space of $[\, 0 \,|\, A_{1,\mathrm{II}}^T \,]^T$, and, vice versa, the part $[\, 0 \,|\, B_{1,\mathrm{II}}^T \,]$ of the right-hand side is not correlated with the column space of $[\, A_{1,\mathrm{I}}^T \,|\, 0 \,]^T$. On the other hand, statistical methods can be useful in identification of the independent subproblems within the decomposable core problem.

If the TLS algorithm applied on a core problem returns a nongeneric solution, see Corollary 6.4, may it be interpreted such that there are two or more different and statistically independent subproblems mixed together in the core (and thus also in the original) problem; subproblems which must be solved independently? Is each core problem, for which the TLS algorithm returns the nongeneric solution, decomposable? Recall that core problem with single right-hand side is non decomposable and it has the unique TLS solutions (which is also computed by the TLS algorithm). On the other hand, recall that all problems, including core problems, belonging to the $\mathscr{F}_2$ set have the TLS solution (but the TLS algorithm does not compute it).

Part IV

Implementation, Computations, and Related Issues

Chapter 7

Bidiagonalization and core problem identification

This chapter, as well as whole Part II of this thesis, focuses on the numerical experiments related to the theory of problems with single right-hand sides, discussed in the introduction of the thesis.

First, in this chapter, several different and well known approaches to bidiagonalization are briefly mentioned. Then our aim is to identify a core problem within the given problem $A\,x \approx b$. We illustrate a difficulty of core problem revealing on an example, and shortly discuss the meaningfulness of using the TLS concept in various problems.

7.1 Implementation remarks on bidiagonalization

It is well known that the bidiagonalization of a given matrix can be computed by different algorithms, and can yield different bidiagonal forms, e.g. lower or upper bidiagonal matrices.

The first well known approach is called *Householder bidiagonalization algorithm*, see [32, Algorithm 5.4.2, p. 252]. When the Householder reflection matrices are suitably applied, see e.g. [32, § 5.1.2–4, pp. 209–211], this algorithm is sufficiently fast and numerically stable. This algorithm is usually used for bidiagonalization of small and dense matrices.

Another approach represents the *Golub-Kahan algorithm*. First, it is worth to note that although the bidiagonal matrix (lower or upper) given by the Householder algorithm is unique, the result of Golub-Kahan algorithm depends on a starting vector. But both approaches are closely related: For $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, the results of the upper bidiagonalization of $[\,b \,|\, A\,]$, obtained by the Householder algorithm, is identical to the results of the Golub-Kahan lower bidiagonalization of A started with the vector b, until the first zero bidiagonal element arises. In [64] and also in Section 1.4 it is shown that such bidiagonalization applied on the data A, b originated in an approximation problem $A\,x \approx b$ yields the core problem within $A\,x \approx b$ and the core problem is revealed exactly when the first zero bidiagonal component arises. In this chapter we aim at the numerical identification of a core problem, thus we are interested just in the described bidiagonalization.

Recall that the Golub-Kahan algorithm for computing the lower bidiagonal form of A started with the right-hand side b can be written as in (1.11). Putting $w_0 \equiv 0$ and the normalization of the starting vector $s_1 \equiv b/\beta_1$, where $\beta_1 \equiv \|b\|$, the

algorithm computes for $j = 1, 2, \ldots$

$$w_j \alpha_j \equiv A^T s_j - w_{j-1} \beta_j ,$$
$$s_{j+1} \beta_{j+1} \equiv A w_j - s_j \alpha_j ,$$

where $\|w_j\| = 1$, $\alpha_j \geq 0$, and $\|s_{j+1}\| = 1$, $\beta_{j+1} \geq 0$, until $\alpha_j = 0$ or $\beta_{j+1} = 0$, or until dimensions of A are exceeded, i.e. $j = \min\{m, n\}$. This algorithm is usually used for bidiagonalization of large and sparse matrices.

It is well known that the vectors w_j and s_j produced by the Golub-Kahan algorithm are mutually orthonormal in exact arithmetic [57]. But in finite precision arithmetic the orthogonality among them is lost in few iterations. In order to stabilize the computation, each newly computed vector w_j or s_j is reorthogonalized among $w_1, \ldots, w_{j-1}$ or $s_1, \ldots, s_{j-1}$, respectively. Really stable implementation is obtained only if the vectors are reorthogonalized against all the previously computed vectors two times, see e.g. [54]. More than two times reorthogonalization is useless. The Golub-Kahan algorithm with two times reorthogonalization is sufficiently fast and stable (comparably to the Householder algorithm). The lost of orthogonality among computed vector w_j and s_j is proportional to the machine precision ε_M. The disadvantage of such implementation of the Golub-Kahan algorithm is the growing of the computational time for one iteration during the computation. Our aim is to identify the core problem, or, because we use finite precision arithmetic, to approximate the core problem, in order to reduce dimensions of the original problem. Namely when we can expect a resulting subproblem of relatively small dimensions, the two times reorthogonalization does not prolong the computational time significantly.

Assuming that the matrix A is considerably rectangular, meaning that $m \gg n$ or $m \ll n$, the bidiagonalization algorithm can be further modified. For example when $m \gg n$ the QR decomposition of A is computed and then the triangular factor is bidiagonalized instead of A, see e.g. [32, §5.4.4, pp. 252–253]. The QR decomposition can be computed either using the Householder reflection matrices or better by the modified Gram-Schmidt algorithm with iterative refinement, see [32, Chapter 5.2, pp. 223–236]. Such QR-preprocessing reduces the computational time, but can influence the attainable accuracy of the computed result. The other case $m \ll n$ is similar, the LQ decomposition instead QR is used.

7.2 An example of the core problem identification

In all the following numerical experiments we try to identify the core problem of known dimensions. The original problem contains compatible core problem of given dimensions $\bar{m} \equiv \bar{n} \equiv k$. The bidiagonal form $L_k x_1 = \beta_1 e_1$ of the core problem is computed and we look at the computed bidiagonal components. One can expect that $\beta_{k+1} = 0$ or, at least, $\beta_{k+1} \approx 0$, but it is well known that the bidiagonal components are very sensitive (although the bidiagonalization is stable, i.e. the lost of orthogonality in both orthogonal factors is proportional to the relative machine precision ε_M, and the singular values of the bidiagonal factor well approximate the singular values of A). These experiments illustrate that the value of the computed β_{k+1} can be far from zero.

We use the SVD-form of core problem (1.10) for construction of the experiment. Consider the matrix and the right-hand side having the form

$$A = P \left[\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & \gamma A_{22} \end{array}\right] Q^T \equiv P \left[\begin{array}{c|c} \mathrm{diag}\,(\sigma_1, \ldots, \sigma_k) & 0 \\ \hline 0 & \gamma A_{22} \end{array}\right] Q^T \in \mathbb{R}^{n \times n} ,$$
$$b \equiv P \left[\begin{array}{c} \texttt{rand(k,1)} \\ \hline 0 \end{array}\right] \in \mathbb{R}^n ,$$

where P, Q are arbitrary orthogonal matrices (computed by `qr(rand(n))` command, see [53])[1]. In the experiment $n = 300$, $k = 20$ and

$$[\,\sigma_1\,,\,\sigma_2\,,\,\sigma_3\,,\,\ldots\,,\,\sigma_{20}\,] = [\,100\,,\,95\,,\,80\,,\,\ldots\,,\,5\,]\,,$$

i.e. singular values of A_{11} linearly decay, $\|A_{11}\| = 100$. Projections of the right-hand side onto the left singular subspaces of A_{11} are chosen randomly. For the non-core matrix we use Frank matrix $A_{22} \equiv$ `0.01*gallery('frank',n-k)`, see [53], with the norm $\|A_{22}\| \approx 198.1791$ comparable to A_{11}. The parameter γ represents a ratio between the core and non-core part, i.e. between the useful and irrelevant information, in the matrix A. We try to find the core problem for seven different γs chosen such that

$$\gamma = \texttt{10.\^{}[-2.5:0.7:2]} \approx 0.003162\,,\,0.01585\,,\\ 0.07943\,,\,0.3981\,,\,1.9953\,,\,10.0000\,,\,50.1187\,,$$

in MATLAB notation. For the singular values of A see Figure 7.1, where the singular values of A_{11} followed by singular values of $\gamma\,A_{22}$ are plotted for all used values of the parameter γ.

The values of γ are chosen such that for small values of γ the singular values of A_{11}, i.e. the useful information, dominates in the problem. On the other hand, for large values of γ the irrelevant information represented by A_{22} completely cover the useful data in the problem, e.g. for the largest γ, $100 = \|A_{11}\| \ll \|\gamma\,A_{22}\| \approx 9932.5$ and the information which is not correlated to the right-hand side dominates in the matrix A. It seems to be senseless to use the TLS formulation for solving the original problem for large γ, the problem must be incorrectly assembled.

The results of the core problem identification obtained by the Householder algorithm for selected values of γ are presented in Figures 7.2–7.6. In particular, Figure 7.2 shows the result for $\gamma \approx 0.003162$; results for the second and third γ are very similar. The value of β_{21} is approximately equal to 10^{-12} for all the first three values of γ. Figure 7.3 shows result for $\gamma \approx 0.3981$; here the norm of the non-core part of A is equal to 78.8965, the core problem is revealed too, but not as well as in the previous cases, the component β_{21} is approximately 10^{-9}, it is over the level of machine precision. For $\gamma \approx 1.995$ the core problem is not revealed correctly, the smallest of the bidiagonal components is $\beta_{23} \approx 10^{-5}$, see Figure 7.4. On Figures 7.5 and 7.6 the core problem is definitely not revealed, the useful information is completely lost.

On all graphs the solid line represents the computed alphas, the dashed line represents the computed betas. The horizontal line in the bottom of each graph represents the machine precision level related to the given problem, $\varepsilon \equiv n\,\|A\|_F\,\varepsilon_M$. One can see that for small values of γ the core problem reveals well, β_{21} is evidently smaller than the others bidiagonal components. With growing γ the results are not satisfactory yet. The wrong results obtained for large γ are probably the consequence of an inappropriately assembled problem, as discussed above. The experiment illustrates well the large sensitivity of bidiagonal components despite the fact that the stability of computation is held.

Similar results can be obtained for another choice of the A_{22} matrix, e.g., the random matrix $A_{22} =$ `rand(n-k,n-k)`, see [53].

Householder bidiagonalization algorithm was used (the implementation was written by the author of this thesis). The experiment was carried out on the computer Hewlett-Packard Compaq nx9110, Intel Pentium 4 CPU, 2.80 GHz, 448 MB RAM, in MATLAB 6.5.0.180913a Release 13 under Windows XP Home Edition operating system, using the IEEE 754 standard double precision floating point arithmetic.

[1] All phrases typeset by `TypeWriter font style` are MATLAB commands or subroutines, we refer to the MATLAB manual [53] for their description.

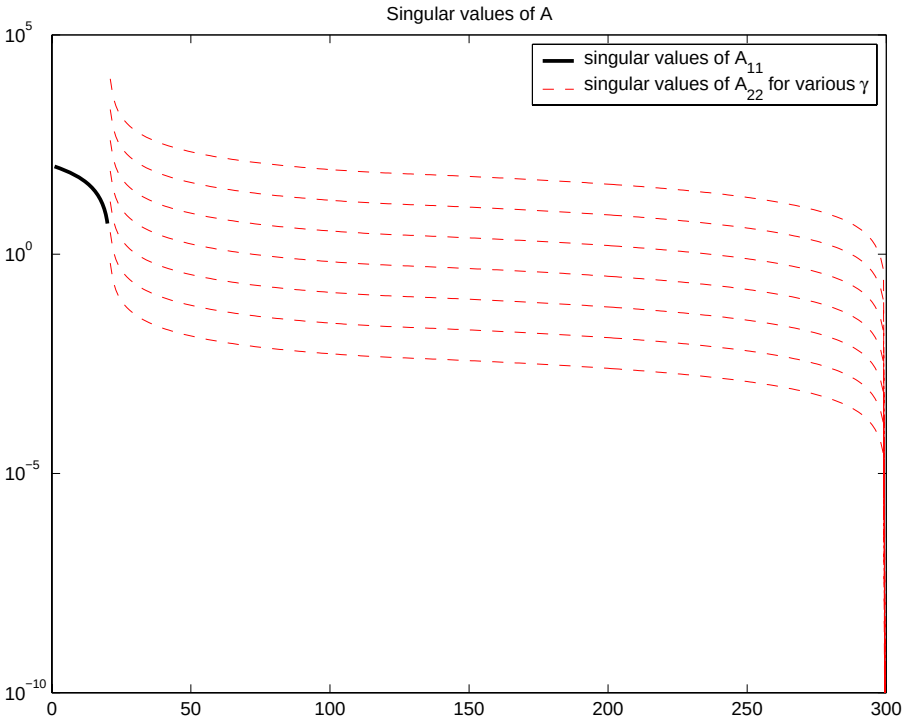


Figure 7.1: The singular values of A_{11} followed by the singular values $\gamma\, A_{22}$ for different values of γ.

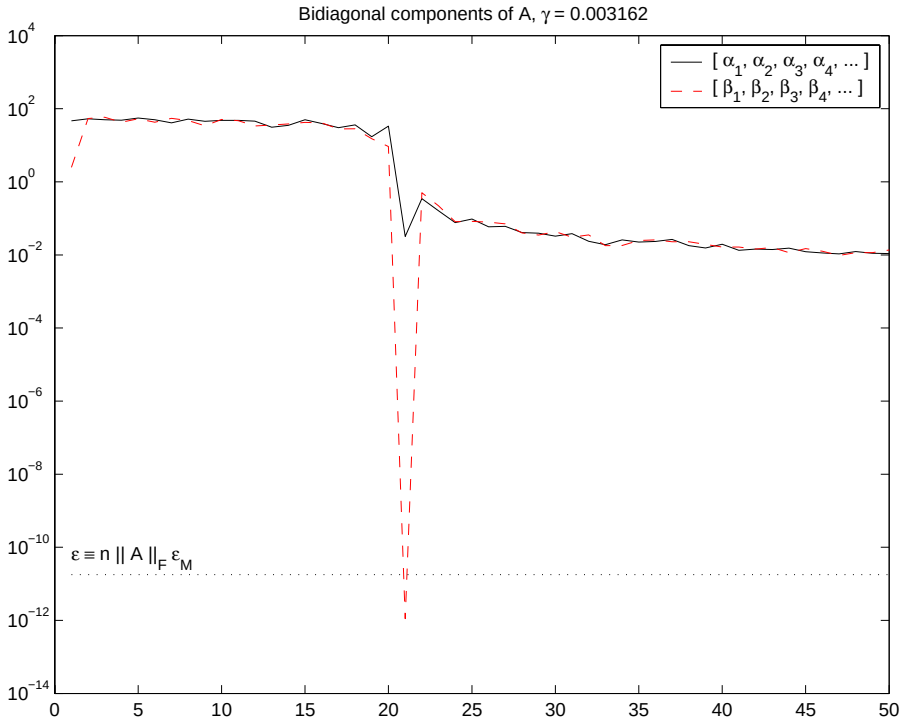


Figure 7.2: For $\gamma \approx 0.003162$ the core problem of dimension $k = 20$ is revealed, $\beta_{21} \approx 10^{-12}$, similarly for $\gamma \approx 0.01585\,, 0.07943$.

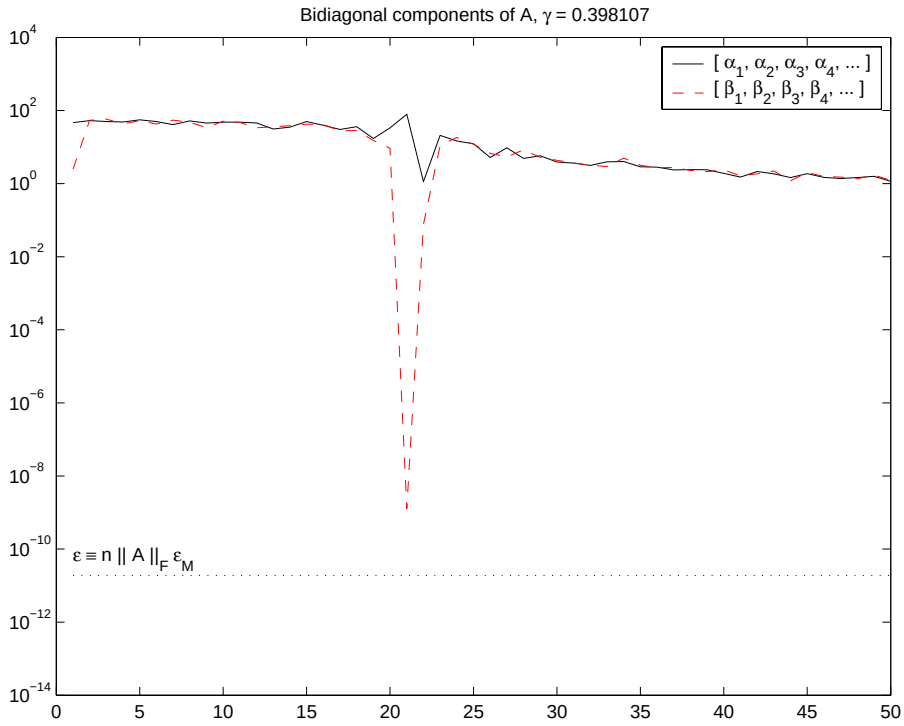


Figure 7.3: For $\gamma \approx 0.3981$ the core problem of dimension $k = 20$ is revealed, but $\beta_{21} \approx 10^{-9}$ is over the level of machine precision.

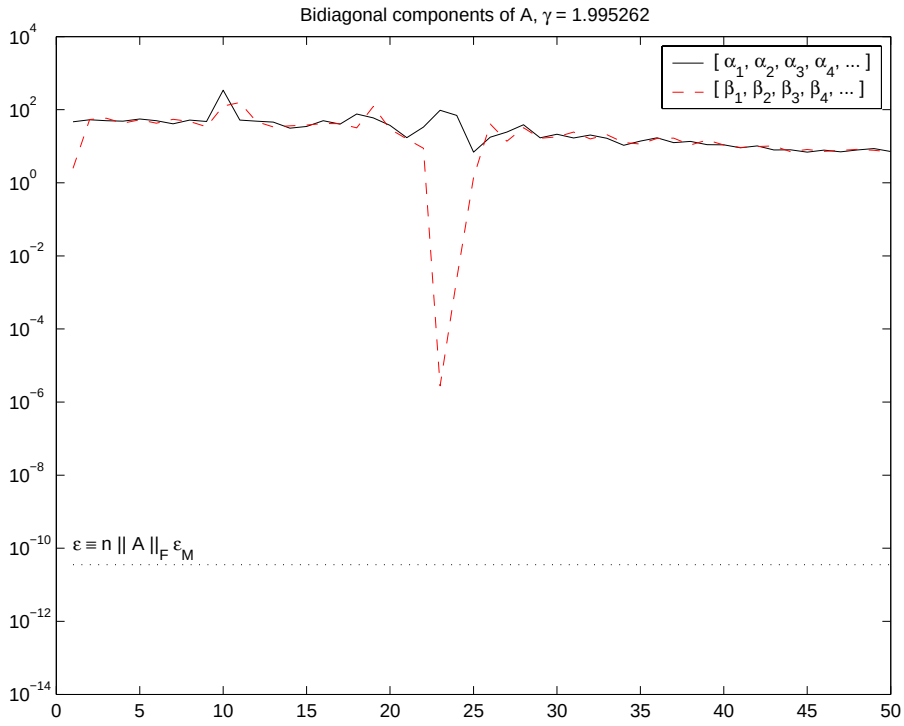


Figure 7.4: For $\gamma \approx 1.9953$ the core problem of dimension $k = 20$ is not revealed correctly. The smallest component is $\beta_{23} \approx 10^{-5}$.

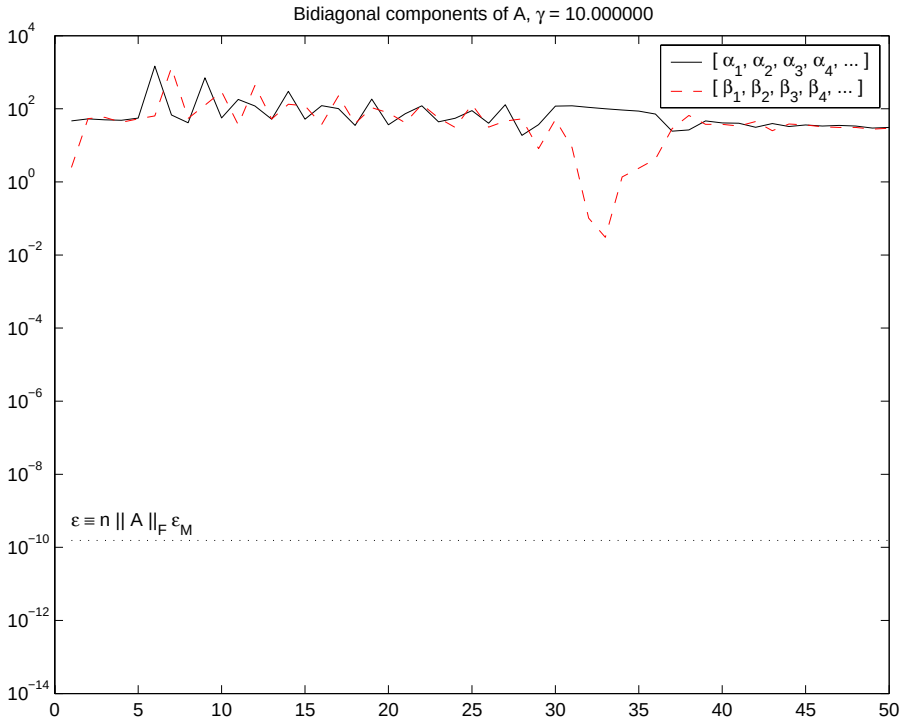


Figure 7.5: For $\gamma \approx 10.0000$ the core problem is not revealed, but there are still some small betas.

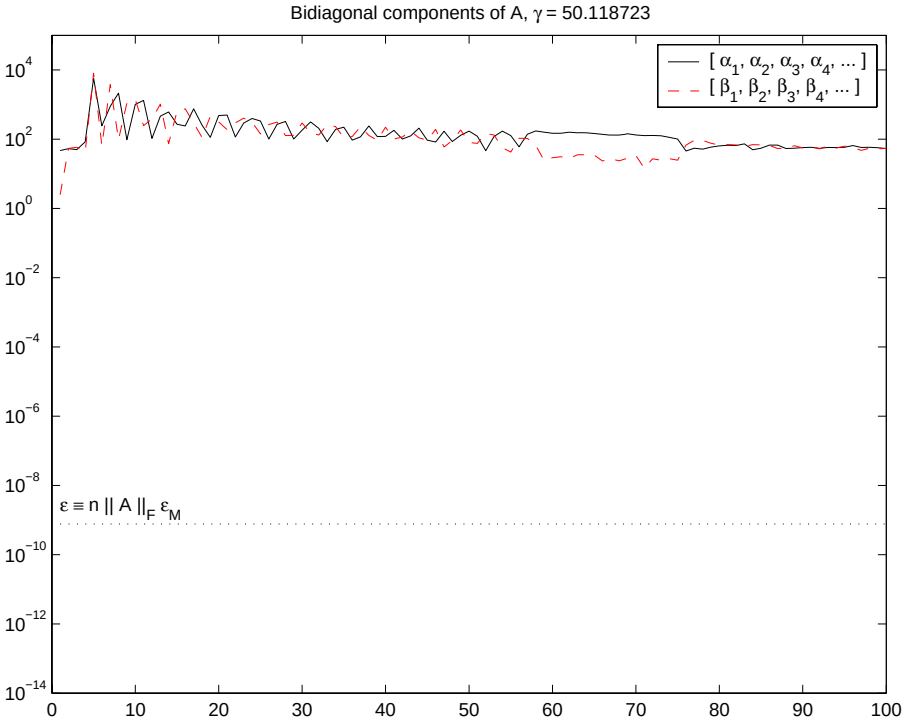


Figure 7.6: For $\gamma \approx 50.1187$ the core problem is not revealed, here the useful information is completely covered.

Chapter 8

Noise level revealing using the bidiagonalization, with application in hybrid methods

The Golub-Kahan bidiagonalization [27] leading to a fundamental decomposition of data, revealing the so called core problem [64], also has been used for iterative solving of large ill-posed and rank-deficient problems for years. It is the first step in hybrid methods which combine the outer bidiagonalization (or another Lanczos-type process that can also be viewed as a regularization) together with an inner regularization of the projected problem. Several hybrid methods are developed and considered in the literature, see for example [56, 19, 36, 48] or [49]. Ill-posed problems from the core problem point of view has been studied by D. Sima and S. Van Huffel [72, 74].

In solving practical problems it is necessary to decide when it is optimal to stop the (outer) bidiagonalization process, the regularization potential of the bidiagonalization has been noticed by Golub and Kahan [27], and later pointed out in relation with using the bidiagonalization for solving linear algebraic systems by Paige and Saunders, see [59, 60] or [70]. Recent examples can be found, e.g., in [34], where hybrid methods based on bidiagonalization are described as least squares projection methods, or [75], where bidiagonalization is used to compute low rank approximations of large sparse matrices. Numerical stability of the bidiagonalization algorithm has been studied, and new stable variants have been proposed, see, e.g., [3, 75].

An appropriate inner regularization method with suitably chosen parameters is based mostly on the estimation of the L-curve using the L-ribbon [11], the discrepancy principle and generalized cross validation [5, 6]. For an application of Tikhonov regularization method in TLS and LS concepts see [25]. These techniques have been widely studied and compared, e.g., in [36]. See also [38] for practical applications in image deblurring. The noise level revealing based on analysis of given vectors in the frequency domain is discussed in [39, 37].

8.1 Introduction to ill-posed problems

In this chapter we experimentally investigate on an example a possibility of the noise level detection from the bidiagonalization process. This noise level detection is further used in a simple hybrid method based on TSVD for reconstructing the solution.

Consider an ill-posed linear system, with square nonsingular matrix (i.e. compatible problem), with the right-hand side polluted by white noise

$$A\,x \;=\; b\,, \qquad A \in \mathbb{R}^{n\times n}\,, \qquad b \;=\; b^{\rm exact} + b^{\rm noise} \in \mathbb{R}^{n}\,, \tag{8.1}$$

where

$$\|b^{\rm exact}\| \;\gg\; \|b^{\rm noise}\|\,. \tag{8.2}$$

We wish to approximate

$$x^{\rm exact} \;=\; A^{-1}\,b^{\rm exact}\,. \tag{8.3}$$

In ill-posed problems the matrix A is usually a discretisation of a smoothing continuous operator and the singular values of A gradually decay to zero without a noticeable gap. The left as well as the right singular vectors of A corresponding to large singular values are dominated by low frequencies, and, vice versa, singular vectors corresponding to small singular values are dominated by high frequencies. The vector $b^{\rm exact}$ (as well as the noised right-hand side b) are typically dominated by low frequencies, see for example Figure 8.1. Ill-posed problem are often discussed in the literature, we refer to [36, 38].

The idea of revealing the noise will be explained on the following example. Consider the testing problem

$$A \;\equiv\; \texttt{shaw(400)} \;\in \mathbb{R}^{400\times 400}\,, \tag{8.4}$$

from Regularization Toolbox [35], with corresponding right-hand side. (Note that the `[A,b,x] = shaw(n)` command returns the matrix, with the noise-free right-hand side, as well as the exact solution. Thus we will be able to measure the relative errors of computed solutions further in the text.) White noise is added artificially; we approximate it by a random vector scaled such that

$$46.6225 \;\approx\; \|b^{\rm exact}\| \;\gg\; \|b^{\rm noise}\| \;=\; 10^{-12}\,, \tag{8.5}$$

using the following sequence of the MATLAB commands: `noise = rand(400,1)`, `noise = (1e-12)*noise/norm(noise)`, see [53].

It is well known that the direct solution $x^{\rm naive} \equiv A^{-1}\,b$ is inapplicable because it is completely dominated by noise. Using the SVD of A, $A \;=\; U'\,\Sigma'\,(V')^T$, this direct naive solution can be rewritten in the form

$$x^{\rm naive} \;=\; \left(\sum_{j=1}^{n} \frac{(u'_j)^T\,b^{\rm exact}}{\sigma'_j}\,v'_j\right) + \left(\sum_{j=1}^{n} \frac{(u'_j)^T\,b^{\rm noise}}{\sigma'_j}\,v'_j\right)\,. \tag{8.6}$$

The first sum in (8.6) must converge to $x^{\rm exact}$. It means that the numerators in the first sum in (8.6), i.e. the components of $b^{\rm exact}$ in the left singular vector subspaces of A, decay faster than (or, at least, as fast as) the singular values in denominators, with growing j. In other words, the exact right-hand side $b^{\rm exact}$ satisfies the so called *discrete Picard condition*, see [36].

On the other hand $b^{\rm noise}$ represents white noise, and because the left singular vectors of A constitute an orthonormal basis, i.e. the specific Fourier basis (representing frequencies), $b^{\rm noise}$ must have comparable components in all left singular vector subspaces. Thus the projections in the second sum in (8.6) do not decay with growing j, and thus they can not decay faster than the singular values in the denominators. The white noise can not satisfy the discrete Picard condition.

See also Figure 8.2: the left picture corresponds to the matrix (8.4), the right corresponds to a typical matrix that originates in image deblurring [38]. We see that first the right-hand side projections onto the left singular subspaces follow the

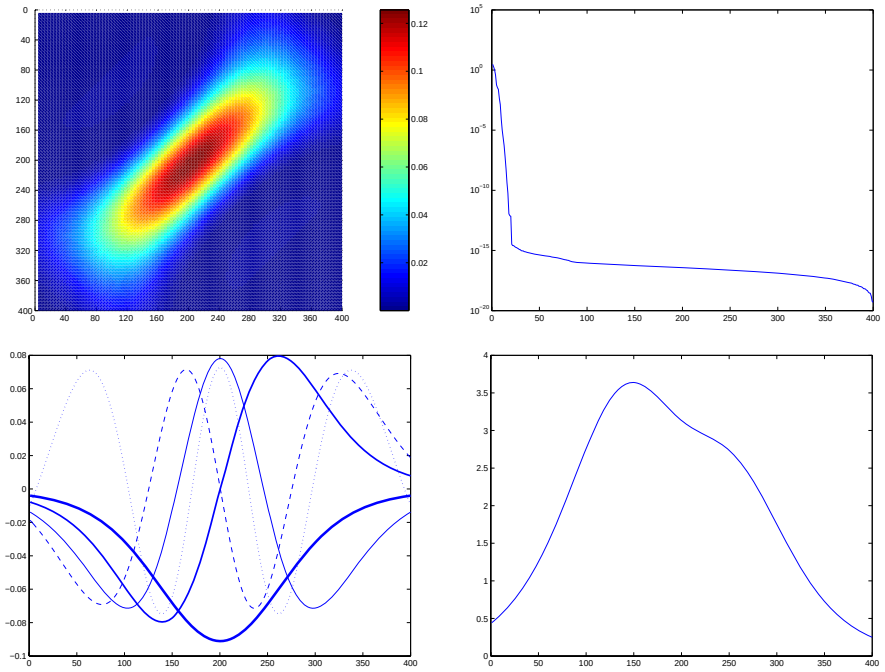


Figure 8.1: The top left picture shows the "surface" (see [53], command `surf`) of the matrix $A \equiv$ `shaw(400)`. The top right plot shows singular values of A in logarithmic scale. The bottom left plot shows the first five left singular vectors of A. The bottom right plot shows the noised right-hand side $b \equiv b^{\text{exact}} + b^{\text{noise}}$, see also (8.4)–(8.5).

singular values. Then, at some point, they begin to stagnate, the projections of the noise dominate the projections of the exact right-hand side. On both graphs the projections are plotted for three different noise levels, $\|b^{\text{noise}}\| = 10^{-10}$, 10^{-8} and 10^{-6}. (On the left graph is only a few projections that follows the singular values, therefore, for better illustration, we include the right graph. For clarity we do not plot the noise level 10^{-12} used in our experiment; the corresponding projections are very close to the singular values.) For the exact right-hand side, the rounding errors cause a similar effect.

Consequently, dividing each noise projection by the corresponding singular value which is smaller than this projection, magnifies the noise information in the second sum in (8.6). The magnified noise completely cover the useful information, $\|A^{-1}\, b^{\text{exact}}\| \ll \|A^{-1}\, b^{\text{noise}}\|$, and thus x^{naive} does not approximate (8.3).

The regularization methods can be interpreted as filtering methods where each summand in both sums in (8.6) is multiplied by a factor Φ_j, e.g.,

$$\Phi_j = \frac{(\sigma'_j)^2}{(\sigma'_j)^2 + \lambda^2}$$

for a given λ yields the well known *Tikhonov regularization method*. The aim of filtering is to eliminate the influence of the small singular values in the sum.

In the further text we will use the *truncated SVD (TSVD)* regularization method. The TSVD method, also called *truncated LS (T-LS)*, yields, for the given index r,

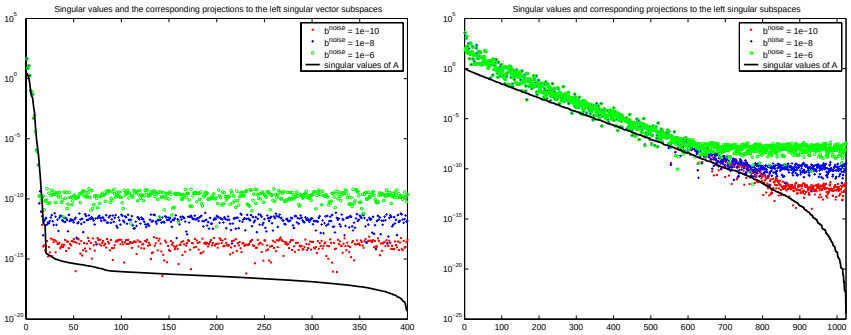


Figure 8.2: The left picture corresponds to the matrix $A \equiv$ `shaw(400)`, the right corresponds to a typical matrix arising in image deblurring, see [38]. The solid line represents singular values of the matrix, the dots represent projections of the noised right-hand side to the left singular vector subspaces of the matrix for three different noise levels: 10^{-10} (at the bottom), 10^{-8} and 10^{-6} (at the top).

the solution

$$x^{\mathrm{TSVD},r} \equiv A_r^{\dagger}\, b\,, \qquad \text{where} \qquad A_r \equiv \sum_{j=1}^{r} u'_j\, \sigma'_j\, {v'_j}^T$$

is the best rank r approximation of A in the Frobenius norm. This solution rewritten in the dyadic form (8.6), where each summand is multiplied by 1 or 0, i.e.,

$$\Phi_j = \begin{cases} 1 & \text{when } j \le r \\ 0 & \text{when } j > r \end{cases},$$

give us the filter function assigned to the TSVD method. The difficulty in using regularization methods is choosing the suitable regularization parameter, e.g. λ in the Tikhonov method, or r in the TSVD method.

It was shown in [64], see also Chapter 1, that the upper bidiagonalization of $[\,b\,|\,A\,]$, realized, e.g., by the Golub-Kahan bidiagonalization algorithm (1.11), yields a core problem within $A\,x = b$ which contains the necessary and sufficient information for solving the original problem. In particular the bidiagonalization algorithm yields the sequence of projected subproblems

$$L_j\, x = e_1\, \beta_1\,, \qquad L_j \in \mathbb{R}^{j\times j}, \qquad \text{for} \quad j = 1\,, 2\,, \ldots\,,$$

that approximate the (compatible) core problem. In other words, the bidiagonalization concentrates the useful information in the leading principal bidiagonal block while moving the irrelevant and redundant information into A_{22}, see (1.9).

Our goal is to approximate the core problem within $A\,x = b^{\mathrm{exact}}$ as well as possible, i.e., stop the bidiagonalization for some k by putting $\beta_{k+1} = 0$ such that $L_k\, x = e_1\, \beta_1$, in the best case, does not contain any noise pollution, or it is insignificantly polluted by noise (i.e., the subproblem $L_{k+1}\, x = e_1\, \beta_1$ is the first approximation of the core problem significantly polluted by noise).

In the following two sections, first, a method for revealing the noise level, i.e. for identifying such k, is introduced. Then we discus how to approximate the solution of (8.1)–(8.3). Both are illustrated on the given example.

8.2 Noise level revealing example

First we note, that similar ideas are used in [39, 37]. In these papers the authors use the information available in the residual vector and Arnoldi vectors, respectively, for choosing the regularization parameter in a discrete ill-posed problem. Authors show how to use statistical tools and Fourier analysis in the basis of singular vectors as well as in the standard trigonometric basis (using the fast Fourier transforms) to extract this information efficiently. In our experiment we use the left vectors produced by the Golub-Kahan bidiagonalization process (instead of residual vectors or Arnoldi vectors) to obtain an information about the noise level in the data.

In the following we use the Golub-Kahan bidiagonalization algorithm (1.11), see Chapter 1 (and also Chapter 7). Vectors s_j, and w_j, denote the vectors generated by the Golub-Kahan bidiagonalization of A started with the right-hand side b contaminated by noise, vectors u'_j, and v'_j are the left and right singular vectors of the matrix A.

The vector $s_1 = b/\|b\|$ is the normalized noised right-hand side and therefore it is contaminated by noise. The vector w_1 is obtained from s_1 by action of the smoothing operator A^T thus it has the information about the noise, but it is necessarily smoothed. The vector s_2 is obtained from w_1 by action of action of the smoothing operator A, or, equivalently from s_1 by the smoothing operator $A\,A^T$. The subsequent orthogonalization of $A\,A^T\,s_1$ against s_1 represents a linear combination of s_1 contaminated by noise and $A\,A^T\,s_1$ which is smooth. Therefore the contamination of s_1 by noise is transferred to s_2. Analogously for the other vectors $s_3\,,\,s_4\,,\,\ldots\,,\,s_j$, which are obtained from $(A\,A^T)^j\,s_1$ through the orthogonalization against the vectors $s_{j-1}\,,\,\ldots\,,\,s_1$ (here we deal with mathematical properties in exact arithmetic and do not consider specific implementation details). Equivalently, s_j can be considered as the normalized part of $A\,A^T\,s_{j-1}$ orthogonal to the subspace $\mathrm{span}\,(\,s_1\,,\,A\,A^T\,s_1\,,\,\ldots\,,\,(A\,A^T)^{j-2}\,s_1\,)$. Consequently, the noise contamination in s_1 is transferred to s_j, and it can be expected that subtracting the smooth components proportional to $A\,A^T\,s_1\,,\,\ldots\,,\,(A\,A^T)^{j-2}\,s_1$ will significantly increase the relative size of the high frequency noise. On the contrary the recurrence for the vectors w_j starts with $w_1 = A^T\,s_1/\|A^T\,s_1\|$ and all the vectors w_j are smoothed. orthogonalized against any noised vector. Consequently we are interested in the noised s_j vectors, in the further text.

In order to identify a procedure for noise revealing, the vectors s_j are analyzed in the frequency domain, i.e. we compute the Fourier coefficients with respect to a sequence of appropriate orthonormal vectors given by the left singular vectors u'_j of A which in $A\,x = b$ generate the right-hand side b. It can be expected that the vector s_1 has dominant components in directions of several first left singular vectors representing low frequencies, in particular u'_1. When projecting out the low frequency information, $s_2\,,\,s_3\,,\,\ldots$ will have dominating components in directions of the left singular vectors around $u'_2\,,\,u'_3\,,\,\ldots$ with the level of noise gradually increasing. Finally, for some k the vector s_k has comparable components in many singular vector subspaces, and the noise is revealed, see Figure 8.3.

A disadvantage of this process is the high computational cost of the SVD. The basis consisting of left singular vectors is very natural, but its computing is too expensive.

However, it is reasonable to expect that the noise revealing result can be observed in any other (suitable) Fourier basis. For illustration we use the standard trigonometric basis, i.e.

$$f_j(x) \equiv e^{(\frac{2\pi\mathrm{i}}{n})jx}\,, \qquad \text{for} \quad j = 0\,,\,\pm1\,,\,\pm2\,,\,\ldots\,, \tag{8.7}$$

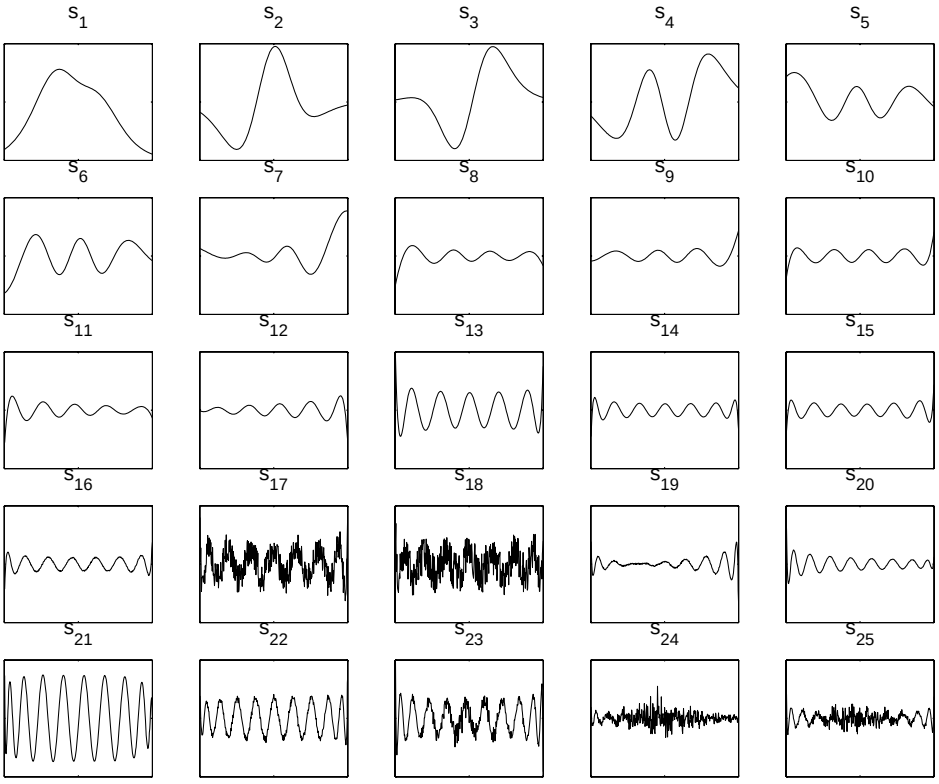


Figure 8.3: The first twenty-five left vectors computed by the double reorthogonalized Golub-Kahan bidiagonalization, applied on the ill-posed problem with noised right-hand side (8.4)–(8.5). Obviously, vectors s_{17}, s_{18} are strongly affected by noise.

for a vector of length n, here $\mathbf{i}$ is the imaginary unit. Computing the Fourier coefficients in trigonometric basis is very fast and efficient using e.g. *fast Fourier transforms (FFT) algorithm*, [12, 15], see also command `fft` in [53].

Thus we study the noisy vectors s_j produced by the Golub-Kahan algorithm using the Fourier analysis, i.e. in the frequency domain. Two orthogonal bases are used, the basis of the left singular vectors $U' = [\, u'_1 \,, \ldots \,, u'_n \,]$ of A (we are interested in vectors $(U')^T s_j$), and the standard trigonometric basis (8.7). Algorithmically, we use the Golub-Kahan bidiagonalization with double reorthogonalization, see Chapter 7. The Fourier coefficients in the singular vector basis are computed (using `[U,S,V] = svd(A)`) by the command `transpose(U)*s_j`. The coefficients in the (8.7) basis are computed by the command `fft(s_j)`, see [53]. Results of both Fourier analysis of s_j, $j = 1 \,, \ldots \,, 25$, for the testing problem (8.4)–(8.5) are plotted on Figures 8.4 and 8.5. (The data plotted in Figure 8.5 are obtained by `abs(fft(s_j))/n`, see [53].)

One can observe that the noise level in s_j grows until the vector s_{18} (the plot at the bottom left) is fully dominated by noise. The full dominance can be identified through the fact that the noise level in the subsequent vector s_{19} is significantly lower than in s_{18}. It is a consequence of the orthogonalization of s_{19} against s_{18}, the noise is partially projected out.

Similar behavior of vector s_j can be observed for another choice of of the noise level in the problem (8.1)–(8.3). For example when $\|b^{\rm noise}\| = 10^{-8}$ the first two

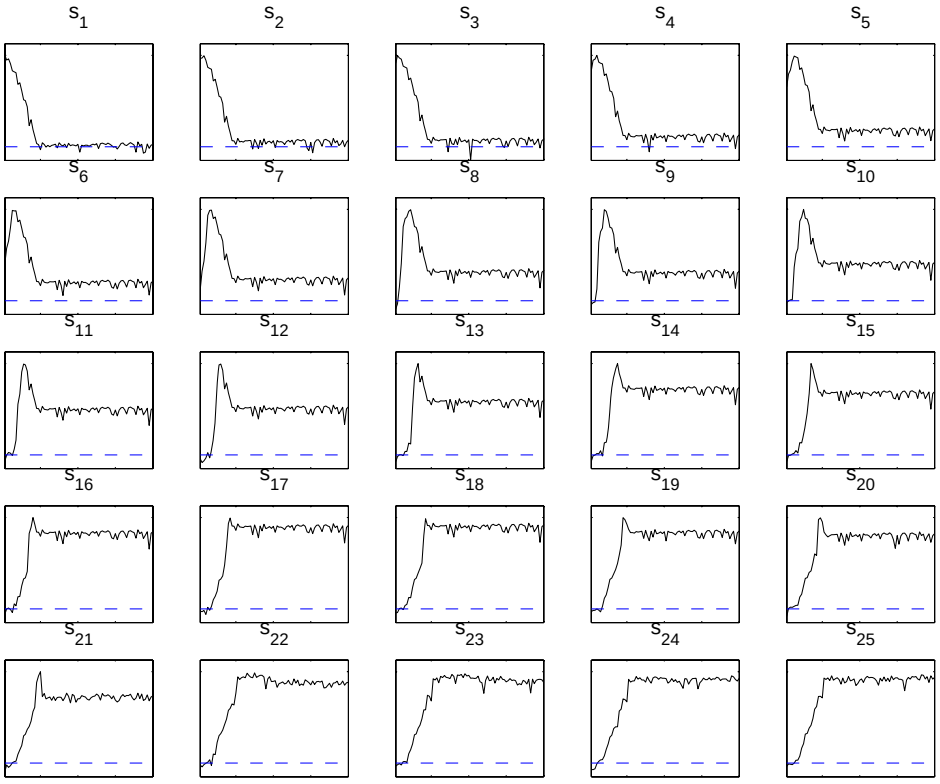


Figure 8.4: The first eighty Fourier coefficients of the vectors s_j in the basis of the left singular vectors of A, i.e., the first eighty components of the vectors $(U')^T s_j$, for $j = 1, \ldots, 25$. One can see the relative increase of the noise level. It can observed that for $j = 18$ the noise level reaches its relative maximum, while for $j = 19$ it relative decreases due to the orthogonality of s_{19} and s_{18} (compare with Figure 8.5). The dashed lines represent the machine precision level ε_M. All graphs are in logarithmic scale with range 10^{-18}–10^2.

noised vectors are s_{13}, s_{14}. In s_{14} the noise level reaches its relative maximum, while in s_{15} we can observe the relative decrease of the noise level. All the resulting graphs are very similar. Without the additional noise the vector s_{18} is the first vector strongly affected by noise. The results are very similar for our testing problem with $\|b^{\text{noise}}\| = 10^{-12}$. It is caused by rounding errors, recall that the matrix A is rank deficient with numerical rank approximately equal to 20.

In the further text we investigate a usage of this technique to the approximation of the solution (8.3).

8.3 Approximation of the exact solution

In the example presented in Section 8.2 the noise level has to reach the level of the useful information in the data at the step $j = 18$. An approximation of the solution x^{exact} of the original problem $A\,x = b$ computed through the bidiagonal problem

$$L_j\, y = e_1\, \beta_1\,, \qquad \text{for} \quad j > 18\,,$$

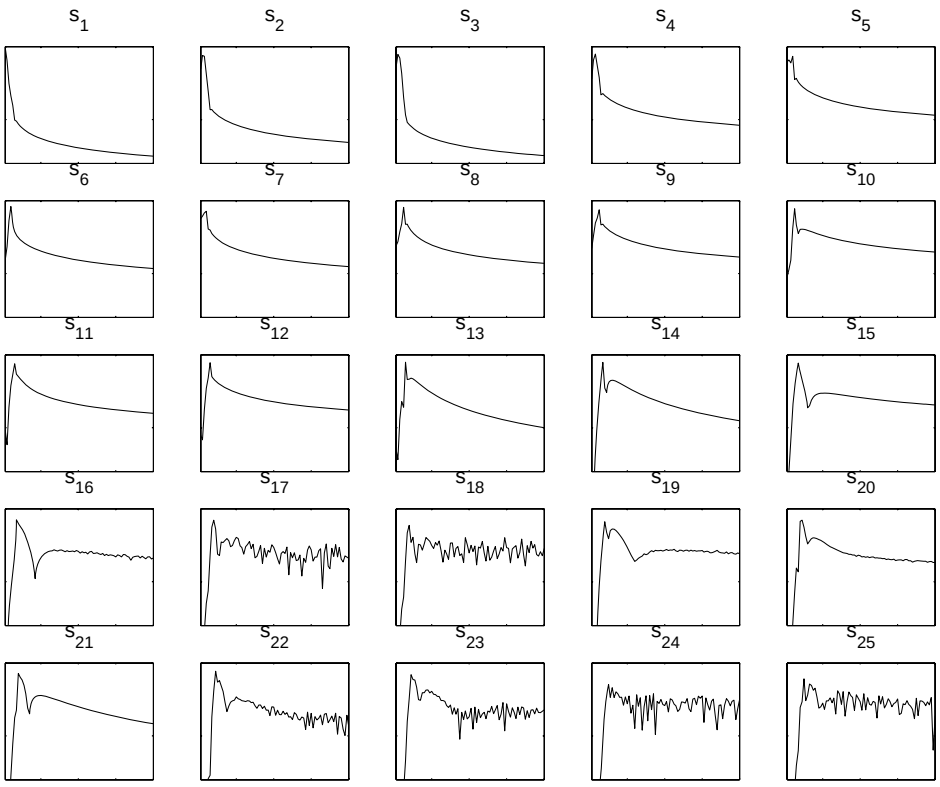


Figure 8.5: The first eighty Fourier coefficients of vectors s_j in the trigonometric basis; computed by `fft` MATLAB command, for $j = 1, \ldots, 25$. The noise level is maximal in the vector s_{18}, then it is partially projected out in s_{19}. All graphs are in logarithmic scale with range 10^{-8}–10^{0}.

can therefore be significantly polluted by the noise. The simplest way to approximation the solution is to compute it directly from the projected problem $L_{18}\, y = e_1\, \beta_1$. We will see that such solution does not approximate x^{exact} well and therefore we will use a simple inner regularization.

8.3.1 Direct solving

The simplest way to approximate the solution x^{exact} of the original problem is by stopping the Golub-Kahan algorithm directly in the 18th iteration and compute

$$x_k \;=\; (\, W_k\, L_k^{-1}\, S_k^T\,)\, b\,, \tag{8.8}$$

for $k = 18$, where $S_k = [\, s_1\,, \ldots, s_k\,]$ and $W_k = [\, w_1\,, \ldots, w_k\,]$. Because the exact solution is known (the command `shaw` returns the matrix A, noise-free right-hand side vector b^{exact}, as well as the exact solution x^{exact}, see [35]), then

$$\frac{\|x^{\text{exact}} - x_{18}\|}{\|x^{\text{exact}}\|} \;\approx\; 0.81902$$

is the relative error. Figure 8.6 presents both the exact solution and the solution computed by (8.8). The computed solution is clearly much more oscillating than the exact solution, the result is not satisfactory. The reason is that the original problem

is ill-posed, and the small projected problem $L_k\, x \;=\; e_1\, \beta_1$ may be ill-posed too. Moreover the termination occurs too early, and some useful information is lost with the eliminated noise.

In other words, the projected matrix L_k can have small singular values which magnify the "noise" in the computed intermediate results caused by rounding errors. In the following section we try to analyze the possible sources of troubles and improve the computed solution.

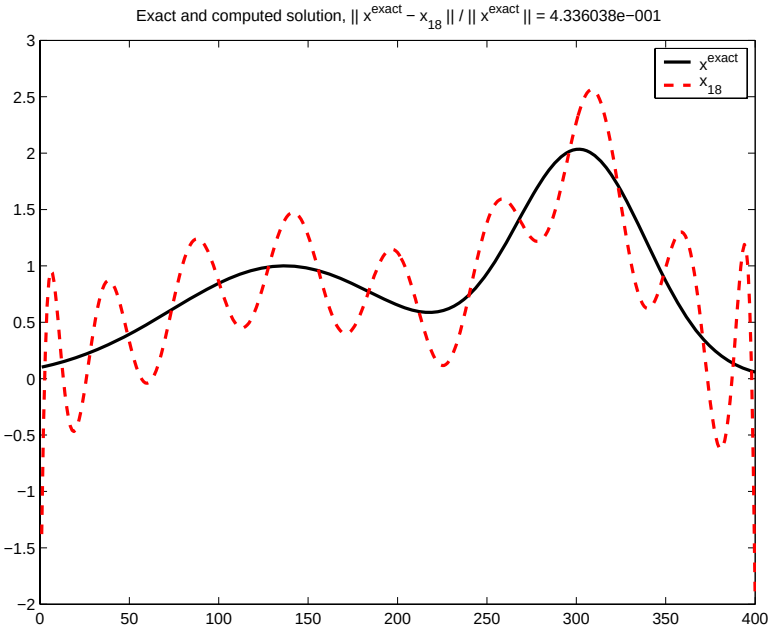


Figure 8.6: The exact solution of the ill-posed problem (8.4)–(8.5) and the approximate solution computed from $L_{18}\, x \;=\; e_1\, \beta_1$. The computed solution is obtained from an approximation of the core problem. The bidiagonalization is stopped when the noise level reaches the level of useful information in the data.

8.3.2 Inner regularization

First we study the singular values of the bidiagonal matrix L_k for different k in our example, see Figure 8.7. We see, that the smallest singular value of the matrix L_{18},

$$\sigma_{\min}\,(L_{18}) \;\approx\; 4.8029 \times 10^{-14}$$

is very small. The matrix L_{18} is nearly numerically rank deficient. Therefore x_{18} does not represent a good approximation of x^{exact}. We will apply the inner TSVD regularization for stabilizing the solution.

Consider the SVD of the bidiagonal matrix

$$L_k \;=\; P\,\Theta\,Q^T\,, \qquad \Theta \;=\; \operatorname{diag}\,(\,\theta_1\,,\,\ldots\,,\,\theta_k\,)\,,$$

where $L_k\,,\; P\,,\; \Theta\,,\; Q \in \mathbb{R}^{k\times k}$ and $P^{-1} \;=\; P^T$, $Q^{-1} \;=\; Q^T$. Denote

$$P_k \;\equiv\; S_k\, P \in \mathbb{R}^{n\times k}\,, \qquad Q_k \;\equiv\; W_k\, Q \in \mathbb{R}^{n\times k}\,,$$

then $P_k^T\,A\,Q_k\,=\,\Theta$. Further denote p_j, q_j, the jth columns of P_k and Q_k, $j\,=\,1\,,\,\ldots\,,\,k$, respectively. Consider the following approximate solution of $A\,x\,=\,b$ as

$$x_k^{\mathrm{TSVD},r}\,\equiv\,\sum_{j=1}^{r}\,\frac{p_j^T\,b}{\theta_j}\,q_j\,,\qquad\text{for}\quad r\,\leq\,k\,,$$

i.e. the TSVD concept is applied on the projected problem $L_k\,x\,=\,e_1\,\beta_1$ and the solution is then transformed back into the variables of the original problem. We study the relative error

$$\frac{\|x^{\mathrm{exact}}\,-\,x_k^{\mathrm{TSVD},r}\|}{\|x^{\mathrm{exact}}\|}\,,$$

for the fixed index $k\,=\,18$, see Figure 8.8. Recall, once again, that here the fixed index k represents the step of the bidiagonalization process revealed using the Fourier analysis above, and the index r is the truncation level in the inner TSVD regularization process when solving the problem with the bidiagonal matrix L_k.

On Figure 8.8 we see that the relative error of the TSVD regularized solution of $L_k\,x\,=\,e_1\,\beta_1$ is small when the singular values of L_k smaller than $\varepsilon\,=\,n\,\|A\|\,\varepsilon_M$ are removed from the solution process, see Figure 8.7 for the singular values of L_k. We can review the dependence of the error of the approximate solution:

(i) Stopping Golub-Kahan bidiagonalization too early (for $k\,\leq\,17$) does not allow the noise to affect the approximate the solution, but a significant useful information is lost and thus the approximation is not optimal.

(ii) When stopping Golub-Kahan bidiagonalization exactly when the noise is revealed (in our experiment $k\,=\,18$), the inner TSVD regularization with truncation level $r\,=\,k-1\;(=\,17)$ gives a reasonably good approximation of the exact solution.

(iii) Continuing the Golub-Kahan bidiagonalization (for $k\,\geq\,21$) and applying the inner TSVD regularization sometimes gives a further slight decrease of the error, see Table 8.1 and Figure 8.8.

The computed results for our experiment are summarized in Table 8.1. The best approximation of x^{exact} (measured by the relative error) we obtain when the Golub-Kahan algorithm was stopped in the 22th iteration with the TSVD truncation level $r\,=\,18$, see also Figure 8.8. Note that the improvement of the approximation while continuing the bidiagonalization, phase (iii), does not occur in general in our example, it depends on the particular noise generated by the `rand` command. We do not present neither the regularized solution $x_{18}^{\mathrm{TSVD},17}$ nor the slightly improved $x_{22}^{\mathrm{TSVD},18}$ (graphs analogous to Figure 8.6) because the differences between them and x^{exact} are not visible, both approximations look like the solid line in Figure 8.6.

GK algorithm stopped at	$k\,=$	16	17	...
minimal relative error is at	$r\,=$	16	17	...
relative error is approximately	$\approx\,10^{-4}\times$	7.5082	5.3364	...

...	18	19	20	21	22	25	32
...	17	17	17	18	18	18	18
...	4.5939	4.5946	4.0134	4.0420	**3.9834**	4.0480	40480

Table 8.1: The minimal relative errors of $x_k^{\mathrm{TSVD},r}$ for various k, see also Figure 8.8.

8.4 Summary

Here we have presented an example of the hybrid approach which uses information about the level of noise in the data revealed by the Golub-Kahan bidiagonalization. We believe that a similar idea can be used in practical problems, and in our further work we aim to focus on construction of an effective stopping criteria for hybrid methods based on the discrepancy principle.

All the numerical experiments in this chapter were carried out on the computer Hewlett-Packard Compaq nx9110, Intel Pentium 4 CPU, 2.80 GHz, 448 MB RAM, in MATLAB 6.5.0.180913a Release 13 under Windows XP Home Edition operating system, using the IEEE 754 standard double precision floating point arithmetic.

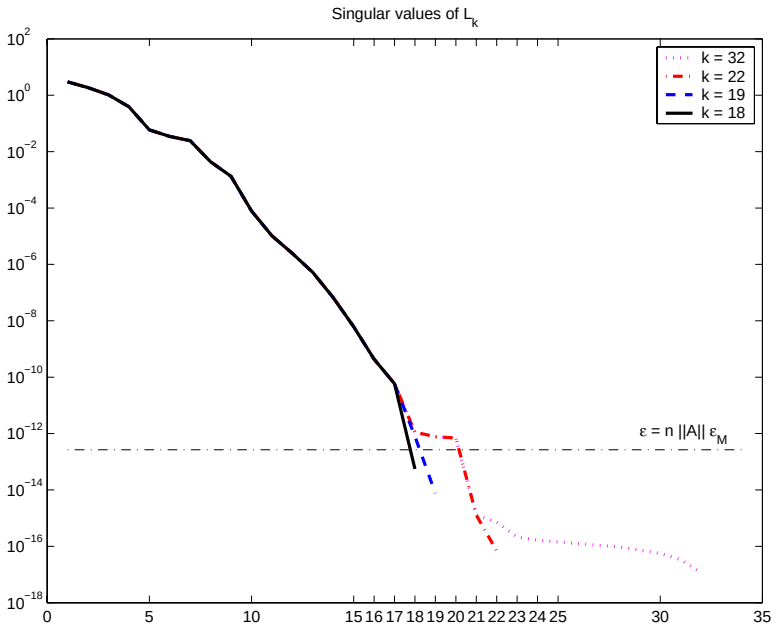


Figure 8.7: Singular values of bidiagonal matrices L_k, for $k = 18\,,\,19\,,\,22\,,\,32$. The horizontal dashed line is $\varepsilon = n\,\|A\|\,\varepsilon_M$. Obviously all L_k are numerically rank deficient.

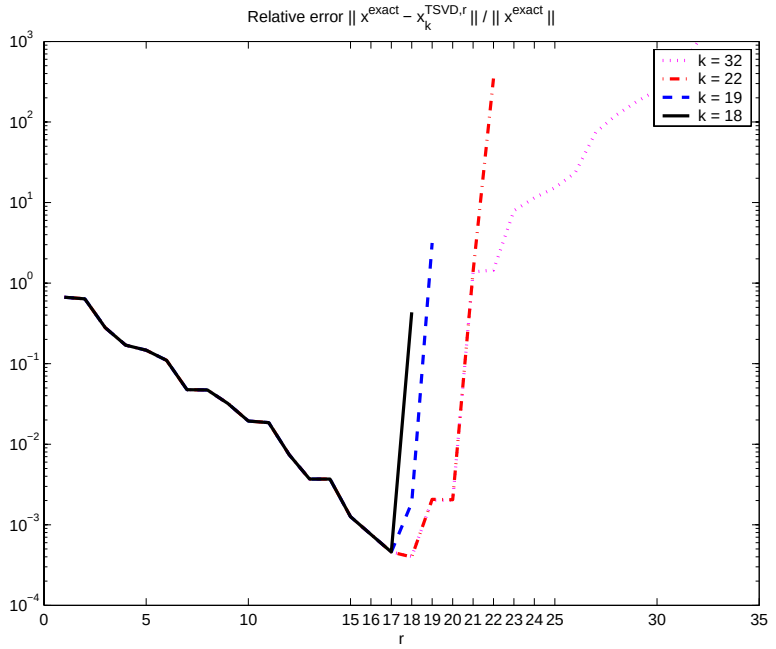


Figure 8.8: Relative errors of TSVD solutions $x_k^{\mathrm{TSVD},r}$ of $L_k\,x = e_1\,\beta_1$ for $k = 18\,,\,19\,,\,22\,,\,32$, for all the possible values of r, $1 \le r \le k$. See also Table 8.1.

Part V

Conclusions

Chapter 9

Conclusions and open questions

In this chapter we summarize results presented in the thesis. We formulate some open questions and mention some possible directions for further work.

9.1 Conclusions

Part I summarizes fundamentals of the total least squares theory in the single right-hand side case based on the work of Golub, Van Loan, Van Huffel, Vandewalle, Paige, Strakoš and others. Parts II and III of the presented thesis investigate an extension of the concept of the core reduction of Paige and Strakoš to a general unitary invariant linear algebraic approximation problem $AX \approx B$; we focus on the problems with *multiple right-hand sides*.

First, in Part II, starting from the results of Van Huffel and Vandewalle, we investigate the fundamental question of the existence of the TLS solution, and present a basic classification of the TLS problems. It is shown that the formulation of the TLS problem with multiple right-hand sides is significantly more complicated than the single right-hand side TLS problem and the results of Chapter 3 reflect the difficulties which have been revealed in our work on the subject.

The data reduction in Part III, which aims at the minimally dimensioned core problem containing the necessary and sufficient information for solving the problem with the original data, starts with the SVD-based transformation, which extends the work of Paige and Strakoš. Another reduction, in the single right-hand side case described by Paige, Strakoš, Hnětynková and the author of this thesis is based on the banded generalization of the Golub-Kahan iterative bidiagonalization, as suggested by Å. Björck and D. M. Sima. Using some properties of the class of generalized Jacobi matrices we investigate further properties of the suggested banded form of the reduced problem.

We have presented the proof of minimality of the SVD-based form as well as the banded form, and proved their equivalence. This allows to define the core problem for problems with multiple right-hand sides. In particular, we relate the solvability of the reduced problem obtained via the core problem approach to the result of the classical TLS algorithm by S. Van Huffel applied directly on the original problem. We showed that the solution computed by the classical TLS algorithm of Van Huffel is not necessarily the TLS solution of the given approximation problem.

Contrary to the single right-hand side case, the core problem may not have the TLS solution. We describe so called *decomposable core problems* and show that there exists a whole class of decomposable core problems which do not have the

TLS solution. Because the core problems in the problems with single right-hand sides are non decomposable, its TLS solution always exists. We formulate, with some ambiguity, the following conjecture:

Conjecture 9.1. *Any non decomposable core problem has the unique TLS solution.*

In decomposable core problems that we have presented the difficulty is caused by the fact that the problem links together data from different *independent subproblems*. If Conjecture 9.1 is correct, then the decomposing of decomposable core problem reveals the hidden structure of independent subproblems which should be treated separately. Then the obtained solution naturally differs form the solution obtained by the classical TLS algorithm by Van Huffel and Vandewalle which considers all data in one problem.

If Conjecture 9.1 is not correct, then the TLS formulation for the problems with multiple right-hand sides lacks in some cases a consistently defined solution. We still do not know how to identify and decompose *all* decomposable problems. Therefore we were unable to prove or disprove Conjecture 9.1.

Part IV of this thesis presents on an example a possible hybrid method for solving ill-posed problems, this method uses the Golub-Kahan bidiagonalization and it is based on core problem ideas, concerning fundamental data decomposition while accumulating necessary and sufficient data in a partially constructed A_{11} block. It is shown that the Golub-Kahan iterative bidiagonalization can be used for revealing the level of noise present in the data. In the example we combine the outer regularization accomplished by the bidiagonalization (the Lanczos-type process), which projects the original problem onto a Krylov subspace of small dimensions, with inner TSVD regularization.

Numerical results are presented. Unfortunately, they are not yet compared with results obtained by other hybrid methods. We believe that the presented idea can be used in practical computations as a contribution towards building efficient and reliable stopping criteria of the outer iterative process.

9.2 Open questions and possible directions for further research

Now we shortly summarize some questions which are interesting in the context of the material presented in this thesis but which are out of the scope of the presented text. In Part II, one can ask about the relationship between the TLS solution and the solution computed by the algorithm by Van Huffel, Vandewalle, and about an interpretation of such a relationship in application areas such as computational statistics. Similarly, it is desirable to give a possible statistical interpretation of the decomposability of the (core) problem. We believe that the statistical point of view and its combination with the matrix computation point of view can help in getting further understanding.

We are well aware of many important questions related to practical implementations and computations. For example, one can expect that a suitable preprocessing of the matrix right-hand side B can improve the behavior of the banded generalization of the Golub-Kahan algorithm. Numerical behavior can be studied in relationship with the block Lanczos algorithm.

Numerical analysis and solution of ill-posed problems illustrated in Part IV produces in the context of the core problem approach many very interesting problems. The presented noise-revealing idea is certainly worth of further effort. Hybrid methods for large ill-posed problems represent a very hot topic in scientific computing.

9.3 List of publications and conference talks of the author of this thesis related to the subject

Some parts of this work was presented on several international and local conferences, and papers for international journals are in preparation.

Papers in journals

[J1] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Lanczos tridiagonalization, Golub-Kahan bidiagonalization and core problem*, PAMM, Proceedings in Appl. Math. and Mechanics 6 (2006), pp. 717–718.

Papers in preparation

[J2] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, Sabine Van Huffel: *Classification of TLS problems in $AX \approx B$ and the relationship to the work of Van Huffel and Vandewalle*, in preparation.

[J3] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Noise revealing via Golub-Kahan bidiagonalization with application in hybrid methods*, in preparation.

Proceedings contributions

[P1] Martin Plešinger, Zdeněk Strakoš: *Core reduction and least squares problems* (in Czech), Proceedings of X. PhD. Conference '05 (F. Hakl, Ed.), Praha, ICS AS CR & Matfyzpress (2005), pp. 102–108. (`http://www.cs.cas.cz/hakl/doktorandsky-den/files/2005/dk05proc.pdf`)

[P2] Martin Plešinger, Zdeněk Strakoš: *Singular value decomposition – application in image deblurring* (in Czech), Seminar on Numerical Analysis '06, Praha, ICS AS CR (2006), pp. 78–81.

[P3] Martin Plešinger, Zdeněk Strakoš: *Some remarks on bidiagonalization and its implementation*, Proceedings of XI. PhD. Conference '06 (F. Hakl, Ed.), Praha, ICS AS CR & Matfyzpress (2006), pp. 104–114. (`http://www.cs.cas.cz/hakl/doktorandsky-den/files/2006/dk06proc.pdf`)

[P4] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Golub-Kahan Iterative Bidiagonalization and Stopping Criteria in Ill-Posed Problems*, In: Seminar on Numerical Analysis '07 (R. Blaheta, J. Starý, Eds.), Ostrava, Institute of Geonics AS CR (2007), pp. 43–45.

[P5] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš: *Total Least Squares Problem in Linear Algebraic Systems with Multiple Right-Hand Side*, In: Seminar on Numerical Analysis '07 (R. Blaheta, J. Starý, Eds.), Ostrava, Institute of Geonics AS CR (2007), pp. 81–84.

[P6] Martin Plešinger, Zdeněk Strakoš: *Total least squares formulation in problems with multiple right-hand sides* (in Czech), Proceedings of XII. PhD. Conference '07 (F. Hakl, Ed.), Praha, ICS AS CR & Matfyzpress (2007), pp. 70–74.

International conferences (talks and posters)

[I1] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Lanczos tridiagonalization and the core problem*, 77th Annual Meeting of the Gesell-

schaft für Angewandte Mathematik und Mechanik e.V., Technische Universität Berlin, Germany, March 27–31, 2006.

[I2] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Golub-Kahan bidiagonalization and stopping criteria in solving ill-posed problems*, Joint GAMM SIAM Conference on Applied Linear Algebra, Düsseldorf, Germany, July 24–27, 2006.

[I3] Poster: Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *On core problem formulation in linear approximation problems with multiple right-hand sides*, 4th International Workshop on Total Least Squares and Errors-in-Variables Modeling, Arenberg castle, Leuven, Belgium, August 21–23, 2006.

[I4] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Analysis of the TLS problem with multiple right-hand sides*, 22nd Biennial Conference on Numerical Analysis, University of Dundee, Scotland, UK, June 26–29, 2007.

[I5] Poster: Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, Sabine Van Huffel: *On total least squares formulation in linear approximation problems with multiple right-hand sides*, Computational Methods with Appl., Harrachov, Czech Republic, August 19–25, 2007.

[I6] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, Sabine Van Huffel: *On total least squares problem with multiple right-hand sides*, IMA Conference on Numerical Linear Algebra and Optimisation, University of Birmingham, UK, September 13–15, 2007.

[I7] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *On fundamentals of total least squares problems*, 13th Czech-French-German Conference on Optimization Heidelberg, Germany, September 17–21, 2007.

Local conferences (talks and posters)

[L1] Martin Plešinger: *Core reduction and least squares problems*, X. PhD. Conference, Hosty – Týn nad Vltavou, November 5–7, 2005.

[L2] Martin Plešinger, Zdeněk Strakoš: *Singular value decomposition – application in image deblurring*, SNA '06, Modelling and Simulation of Challenging Engineering Problems, Monínec – Sedlec-Prčice, January 16–20, 2006.

[L3] Martin Plešinger: *Two Topics from Theory of Linear Approximation Problems*[1], XI. PhD. Conference, Monínec – Sedlec-Prčice, September 18–20, 2006.

[L4] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Golub-Kahan Iterative Bidiagonalization and Stopping Criteria in Ill-Posed Problems*, Seminar on Numerical Analysis SNA '07, Ostrava, January 22–26, 2007.

[L5] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš: *Total Least Squares Problem in Linear Algebraic Systems with Multiple Right-Hand Side*, Seminar on Numerical Analysis SNA '07, Ostrava, January 22–26, 2007.

[L6] Poster: Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, Sabine Van Huffel: *On total least squares formulation in linear approximation problems with multiple right-hand sides*, Seminar on Numerical Analysis SNA '08, Liberec, January 28–February 1, 2008.

[1] This talk was awarded the price of Institute of Computer Science AS CR for the best lecture.

Seminar lectures

[S1] Martin Plešinger: *Singular Value Decomposition, Application in Image Deblurring*, Faculty of Mechatronics Seminar, TU Liberec, December 14, 2005.

[S2] Martin Plešinger: *Core reduction and least squares problems* $Ax \approx b$, Faculty of Mechatronics Seminar, TU Liberec, December 21, 2005.

[S3] Martin Plešinger: *Core problem, Golub-Kahan bidiagonalization, Lanczos tridiagonalization*, Depatrment of Modelling of Processes Seminar, FM, TU Liberec, April 13, 2006.

[S4] Martin Plešinger: *Reduction of data in* $AX \approx B$ Seminar at Institute of Computer Science, AS CR, November 14, 2006.

[S5] Martin Plešinger: *Solving total least squares problems with multiple right-hand sides* , Institute of Novel Technologies and Applied Informatics Seminar, FM, TU Liberec, February 27, 2007.

[S6] Martin Plešinger: *Solving total least squares problems with multiple right-hand sides*, Seminar at Inst. of Computer Science, AS CR, March 13, 2007.

Bibliography

[1] José Ignacio Aliaga, Daniel L. Boley, Roland W. Freund, Vincente Hernández: *A Lanczos-type method for multiple starting vectors*, Mathematics of Computation, Vol. 69 (2000), pp. 1577–1601.

[2] Zhaojun Bai, James Demmel, Jack J. Dongarra, Axel Ruhe, Henk A. van der Vorst (Eds.): *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide*, Philadelphia, SIAM Publications, 2000.

[3] Jesse Barlow, Nela Bosner, Zlatko Drmač: *A New Stable Bidiagonal Reduction Algorithm*, Linear Algebra Appl., Vol. 397 (2005), pp. 35–84.

[4] Adi Ben-Israel, Thomas N. E. Greville: *Generalized Inverses, Theory and Applications, Second Edition*, Springer-Verlag, 2003.

[5] Åke Björck: *A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations*, BIT, Vol. 28 (1988), pp. 659–670.

[6] Åke Björck, Eric Grimme, Paul Van Dooren: *An implicit shift bidiagonalization algorithm for ill-posed systems*, BIT, Vol. 34 (1994), pp. 510–534.

[7] Åke Björck: *Numerical Methods for Least Squares Problems*, Philadelphia, SIAM Publications, 1996.

[8] Åke Björck: *Bidiagonal Decomposition and Least Squares*, Presentation, Canberra (2005).
(`http://wwwmaths.anu.edu.au/events/sy2005/odatalks/canb05.pdf`)

[9] Åke Björck: *A Band-Lanczos Generalization of Bidiagonal Decomposition*, Presentation, Conference in Honor of G. Dahlquist, Stockholm (2006).
(`http://www.mai.liu.se/~akbjo/kth06.pdf`)

[10] Åke Björck: *A band-Lanczos algorithm for least squares and total least squares problems*, book of abstracts of 4th Total Least Squares and Errors-in-Variables Modeling Workshop, Leuven, Katholieke Universiteit Leuven (2006), pp. 22–23.

[11] Daniela Calvetti, Gene Howard Golub, Lothar Reichel: *Estimation of the L-curve via Lanczos bidiagonalization*, BIT, Vol. 39 (1999), pp. 603–619.

[12] James W. Cooley, John W. Tukey: *An Algorithm for the Machine Computation of the Complex Fourier Series*, Mathematics of Computation, Vol. 19 (1965), pp. 297–301.

[13] Jane K. Cullum, Wilm 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).

[14] Percy Deift: *Orthogonal Polynomials and Random Matrices, A Riemann-Hilbert Approach*, American Mathematical Society, Rhode Island, 2000.

[15] Pierre Duhamel, Martin Vetterli: *Fast Fourier Transforms: A Tutorial Review and a State of the Art*, Signal Processing, Vol. 19 (1990), pp. 259–299.

[16] Carl Eckart, Gale Young: *The approximation of one matrix by another of lower rank*, Psychometrica 1 (1936), pp. 211–218.

[17] Ricardo D. Fierro, J. R. Bunch: *Collinearity and total least squares*, SIAM J. Matrix Anal. Appl. 15 (1994), pp. 1167–1181.

[18] Ricardo D. Fierro, J. R. Bunch: *Orthogonal Projection and Total Least Squares*, Numerical Linear Algebra with Applications, Vol 2 (1995), pp. 135–153.

[19] Ricardo D. Fierro, Gene Howard Golub, Per Christian Hansen, Dianne P. O'Leary: *Regularization by Truncated Total Least Squares*, SIAM J. on Scientific Computing 18 (1997), pp. 1223–1241.

[20] Roland W. Freund: *Model reduction methods based on Krylov subspaces*, Acta Numerica, No. 12 (2003), pp. 267–319.

[21] Walter Gautschi: *Orthogonal Polynomials, Computation and Approximation* Oxford University Press, New York, 2004.

[22] Luc Giraud, Julien Langou, Miroslav Rozložník, Jasper van den Eshof: *Rounding error analysis of the classical Gram-Schmidt orthogonalization process*, Numerische Mathematik (2005) 101, pp. 87–100 (previous version without the result on CGS: L. Giraud, J. Langou, M. Rozložník: On the round-off error analysis of the Gram-Schmidt algorithm with reorthogonalization, Research Report TR/PA/02/33, CERFACS, Toulouse, France, April, 2002).

[23] Luc Giraud, Julien Langou, Miroslav Rozložník: *On the loss of orthogonality in the Gram-Schmidt orthogonalization process*, Computers & Mathematics with Applications 50 (2005), pp. 1069–1075.

[24] Gene Howard Golub: *Some modified matrix eigenvalue problems*, SIAM Review, 15 (1973), pp. 318–334.

[25] Gene Howard Golub, Per Christian Hansen, Dianne P. O'Leary: *Tikhonov Regularization and Total Least Squares*, SIAM J. Matrix Anal. Appl. 21 (1999), pp. 185–194.

[26] Gene Howard Golub, A. Hoffman, Gilbert W. Stewart: *A generalization of the Eckart-Young-Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.

[27] Gene Howard Golub, William Kahan: *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., Ser. B, 2 (1965), pp. 205–224.

[28] Gene Howard Golub, Franklin T. Luk, Michael L. Overton: *A Block Lanczos Method for Computing the Singular Values and Corresponding Singular Vectors of a Matrix* ACM Trans. Math. Software 7 (1981), pp. 149–169.

[29] Gene Howard Golub, C. Reinsch: *Singular value decomposition and least squares solutions*, Numerische Mathematik, 14 (1970), pp. 403–420.

[30] Gene Howard Golub, Richard R. Underwood: *The Block Lanczos Method for Computing Eigenvalues*, Mathematical Software, Vol. 3, (J. R. Rice, Ed.) Academic Press, New York (1977), pp. 364–377.

[31] Gene Howard Golub, Charles F. Van Loan: *An analysis of the total least squares problem*, Numer. Anal. 17 (1980), pp. 883–893.

[32] Gene Howard Golub, Charles F. Van Loan: *Matrix Computations, 3rd edition*, The Johns Hopkins University Press, Baltimore and London, 1996.

[33] R. D. D. Groat, E. M. Dowling: *The data least squares problem and channel equalization*, IEEE Transactions on Signal Processing 42:1 (1993), pp. 407–411.

[34] Martin Hanke: *On Lanczos based methods for the regularization of discrete ill-posed problems*, BIT, Vol. 41 (2001), pp. 1008–1018.

[35] Per Christian Hansen: *Regularization Tools – version 3.2 for MALTAB 6.0, a package for analysis and solution of discrete ill-posed problems.* (`http://www2.imm.dtu.dk/~pch/Regutools/index.html`).

[36] Per Christian Hansen: *Rank-Deficient and Discrete Ill-Posed Problems, Numerical Aspects of Linear Inversion*, Philadelphia, SIAM Publications, 1998.

[37] Per Christian Hansen, Toke Koldborg Jensen: *Noise propagation in regularizing iterations for image deblurring*, in preparation.

[38] Per Christian Hansen, James G. Nagy, Dianne P. O'Leary: *Deblurring Images: Matrices, Spectra, and Filtering*, Philadelphia, SIAM Publications, 2006.

[39] Per Christian Hansen, Misha Elena Kilmer, Rikke Kjeldsen: *Exploiting Residual Information in the Parameter Choice for Discrete Ill-Posed Problems*, BIT, Vol 46 (2006), pp. 41–59.

[40] Magnus R. Hestenes, Eduard Stiefel: *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, B49 (1952), pp. 409–436.

[41] Iveta Hnětynková, Zdeněk Strakoš: *Lanczos tridiagonalization and core problems*, Linear Algebra Appl., Vol. 421 (2007), pp. 243–251.

[42] Iveta Hnětynková, Martin Plešinger, Zdeněk Strakoš: *Lanczos tridiagonalization, Golub-Kahan bidiagonalization and core problem*, PAMM, Proc. Appl. Math. Mech. 6 (2006), pp. 717–718.

[43] Iveta Hnětynková: *Krylov subspace approximations in linear algebraic problems*, Ph.D. thesis, Dept. of Numerical Mathematics, Facutly of Mathematics and Physics, Charles University, Prague, 2006.

[44] Roger A. Horn, Charles R. Johnson: *Matrix Analysis*, Cambridge University Press, 1985 (reprint 1999).

[45] Roger A. Horn, Charles R. Johnson: *Topics in Matrix Analysis*, Cambridge University Press, 1991 (reprint 1999).

[46] Eugene Isaacson, Herbert Bishop Keller: *Analysis of Numerical Methods*, New York, Wiley, 1966 (reprint Dover, 1994).

[47] Er-Xiong Jiang: *Perturbation in eigenvalues of a symmetric tridiagonal matrix*, Lin Algebra Appl. 339 (2005), pp. 91–107.

[48] Misha Elena Kilmer, Dianne P. O'Leary: *Choosing Regularization Parameters in Iterative Methods for Ill-Posed Problems*, SIAM J. Matrix Anal. Appl. 22 (2001), pp. 1204–1221.

[49] Misha Elena Kilmer, Per Christian Hansen, Malena I. Español: *A Projection-based Approach to General Form Tikhonov Regularization*, SIAM J. on Scientific Computing 29 (2006), pp. 315–330.

[50] Cornelius Lanczos: *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards 45 (1950), pp. 255–282.

[51] Cornelius Lanczos: *Linear Differential Operators*, Van Nostrand, London, 1961.

[52] Charles L. Lawson, Richard J. Hanson: *Solving Least Squares Problems*, Philadelphia, SIAM Publications, 1995.

[53] The MathWorks, Inc.: *MATLAB: Documentation.* (`http://www.mathworks.com/access/helpdesk/help`)

[54] Gerard Meurant, Zdeněk Strakoš: *The Lanczos and conjugate gradients algorithms in finite precision arithmetic*, Acta Numerica, pp. 1–70, 2006.

[55] Leon Mirsky: *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math. Oxford 11 (1960), pp. 50–59.

[56] Dianne P. O'Leary, John A. Simmons: *A bidiagonalization regularization procedure for large scale discretizations of ill-posed problems*, SIAM J. on Scientific and Statistical Computing 2 (1981), pp. 474–489.

[57] Christopher C. Paige: *Bidiagonalization of matrices and solution of linear equations*, SIAM J. Numer. Anal. 11 (1974), pp. 197–209.

[58] Christopher C. Paige, Michael A. Saunders: *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal. 18 (1981), pp. 398–405.

[59] Christopher C. Paige, Michael A. Saunders: *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software 8 (1982), pp. 43–71.

[60] Christopher C. Paige, Michael A. Saunders: *ALGORITHM 583 LSQR: Spase Linear Equations and Least Squares Problems*, ACM Trans. Math. Software 8 (1982), pp. 195–209.

[61] Christopher C. Paige, Zdeněk Strakoš: *Scaled total least squares funamentals*, Numerische Mathematik, 91 (2002), pp. 117–146.

[62] Christopher C. Paige, Zdeněk 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.), Dordrecht, Kluwer Academic Publishers (2002), pp. 25–34.

[63] Christopher C. Paige, Zdeněk Strakoš: *Bounds for the least squares distance using scaled total least squares*, Numerische Mathematik, 91 (2002), pp. 93–115.

[64] Christopher C. Paige, Zdeněk Strakoš: *Core problem in linear algebraic systems*, SIAM J. Matrix Anal. Appl. 27 (2006), pp. 861–875.

[65] Christopher C. Paige, Musheng Wei: *Analysis of the generalized total least squares problem $AX \approx B$ when some columns of A are free of error*, Numerische Mathematik, 65 (1993), pp. 177–202.

[66] Beresford Neill Parlett: *The Symmetric Eigenvalue Problem*, Philadelphia, SIAM Publications, 1998.

[67] Bhaskar D. Rao: *Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework*, In: Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling (S. Van Huffel, Ed.), Philadelphia, SIAM Publications (1997), pp. 11–20.

[68] Radhakrishna C. Rao, Helge Toutenburg: *Linear Models: Least Squares and Alternatives, Second Edition*, Springer-Verlag, 1999.

[69] Axel Ruhe: *Implementation Aspects of Band Lanczos Algorithms for Computation of Eigenvalues of Large Sparse Matrices*, Math. of Comp., Vol. 33, No. 146 (1979), pp. 680–687.

[70] Michael A. Saunders: *Computing projections with LSQR*, BIT, Vol. 37 (1997), pp. 96–104.

[71] Erhard Schmidt: *Zur Theorie der linearen und nichtlinearen Integralgleichungen, I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener*, Mathematische Annalen, LXIII, Nr. 4 (1907), pp. 433–476.

[72] Diana Maria Sima, Sabine Van Huffel: *Using core formulations for ill-posed linear systems*, PAMM, Proc. Appl. Math. Mech. 5 (2005), pp. 795–796.

[73] Diana Maria Sima, Sabine Van Huffel: *Core problems in $AX \approx B$*, Technical Report, Dept. of Electrical Engineering, Katholieke Universiteit Leuven (2006).

[74] Diana Maria Sima: *Regularization techniques in model fitting and parameter estimation*, Ph.D. thesis, Dept. of Electrical Engineering, Katholieke Universiteit Leuven, 2006.

[75] Horst D. Simon, Hongyuan Zha: *Low-Rank Matrix Approximation Using the Lanczos Bidiagonalization Process with Applications*, SIAM J. on Scientific Computing 21 (2000), pp. 2257–2274.

[76] Gilbert W. Stewart, Ji-guang Sun: *Matrix Perturbation Theory*, Academic Press, Inc., 1990.

[77] Eduard Stiefel: *Ausgleichung ohne Aufstellung der Gausschen Normalgleichungen*, Wiss. Z. Tech. Hochsch. Dresden, 2 (1952/53), pp. 441–442.

[78] Zdeněk Strakoš, Anne Greenbaum: *Open questions in the convegence analysis of the Lanczos process for the real symmetric eigenvalue problem*, IMA Preprint 934, 1992.

[79] R. C. Thompson: *Principal submatrices IX: Interlacing inequalities for singular values of submatrices*, Linear Algebra and Applications 5 (1972), pp. 1–12.

[80] Richard Ray Underwood: *An Iterative Block Lanczos Method for the Solution of Large Sparse Symmetric Eigenproblems*, Ph.D. thesis, Stanford University, 1975.

[81] A. van der Sluis: *Stability of the solutions of linear least squares problems*, Numerische Mathematik, 23 (1975), pp. 241–254.

[82] Sabine 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, ESAT Lab., Dept. of Electrical Engrg., Katholieke Universiteit Leuven (1988).

[83] Sabine Van Huffel: *The extended classical total least squares algorithm*, J. Comput. Appl. Math., 25 (1989), pp. 111–119.

[84] Sabine Van Huffel, Joos Vandewalle: *The Total Least Squares Problem: Computational Aspects and Analysis*, Philadelphia, SIAM Publications, 1991.

[85] Sabine Van Huffel (Ed.): *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, Proceedings of the Second Int. Workshop on TLS and EIV, Philadelphia, SIAM Publications, 1997.

[86] Sabine Van Huffel, Philippe Lemmerling (Eds.): *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Dordrecht, Kluwer Academic Publishers, 2002.

[87] Alistair G. Watson: *Choice of norms for data fitting and function approximation*, Acta Numerica, No. 7 (1998), pp. 337–377.

[88] Musheng Wei: *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl. 13 (1992), pp. 746–763.

[89] Musheng Wei: *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numerische Mathematik, 62 (1992), pp. 123–148.

[90] James Hardy Wilkinson: *The Algebraic Eigenvalue Problem*, Oxford England, Clarendon Press, 1965 (reprint 2004).