Polar Decompositions and Procrustes Problems in Finite Dimensional Indefinite Scalar Product Spaces

von

Dipl.-Inform. Ulric Kintzel

Von der Fakultät II
– Mathematik und Naturwissenschaften –
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuß

Vorsitzender:	Prof. Dr. Michael Scheutzow
Berichter:	Prof. Dr. Peter Benner
Berichter:	Prof. Dr. Volker Mehrmann
Gutachter:	Prof. Dr. Leiba Rodman

Tag der wissenschaftlichen Aussprache
11. April 2005

Berlin 2005
D 83

betreut von

Prof. Dr. Peter Benner

begutachtet durch

Prof. Dr. Peter Benner
Fakultät für Mathematik
Technische Universität Chemnitz
D-09107 Chemnitz
Germany
benner@mathematik.tu-chemnitz.de

und

Prof. Dr. Volker Mehrmann
Institut für Mathematik
Fakultät II: Mathematik und Naturwissenschaften
Technische Universität Berlin
D-10623 Berlin
Germany
mehrmann@math.tu-berlin.de

und

Prof. Dr. Leiba Rodman
Department of Mathematics
The College of William and Mary
Williamsburg, VA 23187-8795
USA
lxrodm@math.wm.edu

Bibliographische Beschreibung

Ulric Kintzel,
Polar Decompositions and Procrustes Problems in
Finite Dimensional Indefinite Scalar Product Spaces,
Dissertation,
Fakultät II: Mathematik und Naturwissenschaften, TU Berlin, 2005,
146 Seiten, 43 Literaturverweise.

Kontaktadresse

ukintzel@aol.com

Inhalt

In dieser Arbeit werden Prokrustesprobleme in endlichdimensionalen Vektorräumen mit indefiniten Skalarprodukten formuliert und untersucht. Es handelt sich dabei um Optimierungsaufgaben zur Bestimmung von Isometrien, mit deren Hilfe zwei gegebene Tupel von Vektoren im Sinne einer optimalen Kongruenz transformiert werden können. Als Kriterium für dieses Optimum werden Summen von Abstandsquadraten optimiert (Methode der kleinsten Quadrate).

Zur analytischen Untersuchung dieser Probleme werden H-Polarzerlegungen und die in dieser Arbeit eingeführten (G,H)-Polarzerlegungen verwendet. Dabei werden einerseits Kriterien für die Existenz dieser Zerlegungen angegeben und andererseits Verfahren zu ihrer numerische Berechnung entwickelt.

Nicht alle formulierten Prokrustesprobleme können analytisch gelöst werden. Daher wird auch ein Newton-Verfahren bereitgestellt, mit dessen Hilfe die numerische Lösung aller Optimierungsaufgaben zur Bestimmung von Isometrien möglich ist, bei denen die Optimierungsfunktion durch eine quadratische Form der vektorisierten Isometrie dargestellt werden kann. Diese Darstellung existiert insbesondere im Fall der Prokrustesprobleme, aber auch H-Polarzerlegungen können mit dem Verfahren berechnet werden.

Letztlich wird auch noch ein numerisches Verfahren entwickelt, mit dem die kanonische Form eines Paares $(\mathbf{A}, \mathbf{H})$ bestehend aus einer H-hermiteschen Matrix $\mathbf{A}$ und einer regulären hermiteschen Matrix $\mathbf{H}$ berechnet werden kann. Dieses Verfahren beruht auf der Berechnung der Jordanschen Normalform der Matrix $\mathbf{A}$ und einer Normalisierungsprozedur, die eine Verallgemeinerung des Cholesky-Verfahrens darstellt.

Schlagworte

Indefinite Skalarprodukte, Kanonische Formen, Polarzerlegungen, Prokrustesprobleme, Matrixgleichungen, Newton-Verfahren.

AMS Klassifikation

15A63, 15A21, 15A23, 15A24, 49M15.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Let $\mathbb{F}$ be the field of real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$ and let $\mathbb{F}^n$ be an n-dimensional vector space over $\mathbb{F}$. Furthermore, let $\mathbf{H} \in \mathbb{F}^{n\times n}$ be a fixed chosen nonsingular symmetric ($\mathbb{F} = \mathbb{R}$) or Hermitian ($\mathbb{F} = \mathbb{C}$) matrix and let $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{F}^n$ be column vectors. Then the bilinear or sesquilinear functional

$$[\mathbf{x}, \mathbf{y}] = (\mathbf{H}\mathbf{x}, \mathbf{y}) \text{ where } (\mathbf{x}, \mathbf{y}) = \sum_{\alpha=1}^{n} x_\alpha \overline{y}_\alpha \quad (\overline{y}_\alpha = y_\alpha \text{ if } \mathbb{F} = \mathbb{R})$$

defines an indefinite scalar product in $\mathbb{F}^n$. Indefinite scalar products have almost all the properties of ordinary scalar products, except for the fact that the value of $[\mathbf{x}, \mathbf{x}]$ for a vector $\mathbf{x} \neq \mathbf{0}$ can be positive, negative or zero. A corresponding vector is called positive (space-like), negative (time-like) or neutral (isotropic, light-like), respectively. The H-adjoint $\mathbf{A}^{[*]}$ of an arbitrary matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ is characterised by the property that

$$[\mathbf{A}\mathbf{x}, \mathbf{y}] = [\mathbf{x}, \mathbf{A}^{[*]}\mathbf{y}] \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{F}^n.$$

This is equivalent to the fact that between the H-adjoint $\mathbf{A}^{[*]}$ and the ordinary adjoint $\mathbf{A}^* = \overline{\mathbf{A}}^T$ there exists the relationship

$$\mathbf{A}^{[*]} = \mathbf{H}^{-1}\mathbf{A}^*\mathbf{H}.$$

If in particular $\mathbf{A}^{[*]} = \mathbf{A}$ or $\mathbf{A}^*\mathbf{H} = \mathbf{H}\mathbf{A}$, one speaks of an H-selfadjoint or H-symmetric or H-Hermitian matrix, and an invertible matrix $\mathbf{U}$ with $\mathbf{U}^{[*]} = \mathbf{U}^{-1}$ or $\mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H}$ is called an H-isometry or an H-orthogonal or H-unitary matrix [GLR]. If $\mathbb{F}^n$ provides several indefinite scalar products, we also write $[.,.]_H = (\mathbf{H}.,.)$ or $\mathbf{A}^H = \mathbf{A}^{[*]_H}$ to indicate the matrix $\mathbf{H}$ on which a particular scalar product is based.

Indefinite scalar products have been a central subject of research during recent years as can be seen by a large number of related publications, for example [BR], [BMRRR1–3], [GLR], [HO], [LMMR], [MMX]. Frequently, these publications generalise well-known results from an environment of ordinary (positive definite) scalar products to an environment of indefinite scalar products. This is

also the strategy of this thesis which is primarily concerned with the investigation of several variants of least-squares or Procrustes problems[1]. These problems occur in a branch of mathematics, known in psychology as factor analysis or multidimensional scaling (MDS) (for example see [BG], [D], [H]).

In a typical application of MDS test persons are first requested to estimate the dissimilarity (or similarity) of specified objects which are selected terms describing the subject of the analysis. In this way the comparison of N objects in pairs produces similarity measures, called proximities, p_{kl}, $1 \leq k, l \leq N$, from which the distances $d_{kl} = f(p_{kl})$ are then determined using a function f, for example $f(x) = ax + b$, which is called the MDS model. Based on these distances, the coordinates of points $\mathbf{x}_k$ in an n-dimensional Euclidean space are constructed such that $\|\mathbf{x}_k - \mathbf{x}_l\| = d_{kl}$ where $\|.\|$ denotes the Euclidean norm. Now each object is represented by a point in a coordinate system and the data can be analysed with regard to their geometric properties.

The results of interrogating the test persons are often categorised in groups, producing several descriptive constellations of points which must be mutually compared in the analysis. To make such a comparison of two constellations $\mathbf{x}_k$ and $\mathbf{y}_k$ possible, it is first of all necessary to compensate for irrelevant differences resulting from possibly different locations in space. This is done with an orthogonal transformation $\mathbf{U}$ selected such that $\sum_k \|\mathbf{U}\mathbf{x}_k - \mathbf{y}_k\|^2$ is minimised. Thereafter the constellations $\mathbf{x}'_k = \mathbf{U}\mathbf{x}_k$ and $\mathbf{y}_k$ are analysed.

The MDS model f is chosen in particular by adding a constant b (and by making further assumptions such as $d_{kk} = 0$), so that the triangle inequality is fulfilled and therefore the points can be embedded in a Euclidean space [BG, Chapter 18]. But this means that the transformed data d_{kl} describe completely different geometric properties than the original data p_{kl} do. It would thus be more reasonable to avoid the transformation and to interpret the proximities itself as distances which is possible if a pseudo-Euclidean geometry is admitted.

Following this approach we will show how to construct vectors $\mathbf{x}_k$ and an indefinite scalar product $[.,.] = (\mathbf{H}.,.)$ such that $[\mathbf{x}_k - \mathbf{x}_l, \mathbf{x}_k - \mathbf{x}_l] = q_{kl}$ where q_{kl} are given real numbers. If these numbers are defined by $q_{kl} = p_{kl}^2$, then the vectors $\mathbf{x}_k$ represent the objects of the analysis in an indefinite scalar product space and the proximities are their pseudo-Euclidean "distances".

Now assume that the vectors $\mathbf{y}_k$ represent a second set of proximities. Then, before comparing the constellations, a pseudo-Euclidean "rotation" $\mathbf{U}$ must be determined such that $\mathbf{x}'_k = \mathbf{U}\mathbf{x}_k$ and $\mathbf{y}_k$ are optimally congruent. This is the central subject of this thesis in which the following problems are considered:

Let $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ be two N-tuples ($N \geq 1$) of vectors in $\mathbb{F}^n$ and let $[.,.]_H$ be an indefinite scalar product. Then the basic task is to determine a solution of the constrained optimisation problem

$$f(\mathbf{U}) = \sum_{k=1}^{N} [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]_H \rightarrow \text{opt} \quad \text{with} \tag{1.1}$$

$$\mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H} = \mathbf{0}$$

[1] Procrustes, a robber in Greek mythology, who lived near Eleusis in Attica. Originally he was called Damastes or Polypemon. He was given the name Procrustes ("the stretcher") because he tortured his victims to fit them into a bed. If they were too tall, he chopped off their limbs or formed them with a hammer. If they were too small, he stretched them. He was overcome by Theseus who served him the same fate by chopping off his head to fit him into the bed.

which will be called the H-orthogonal or H-unitary Procrustes problem in agreement with the Euclidean case investigated in [S]. The wanted optimum depends on the matrix $\mathbf{H}$. For example, if $\mathbf{H}$ is positive or negative definite, then the minimum or maximum, respectively, of the function f has to be determined.

Whereas it turns out that (1.1) can always be solved in the case of a definite matrix $\mathbf{H}$, in the indefinite case it is possible that no solution exists. This leads to the optimisation problem

$$f(\mathbf{U}) = \sum_{k=1}^{N} [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]_H \to \text{opt} \quad \text{with} \tag{1.2}$$
$$\mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H} = \mathbf{0} \ \text{ and } \ \mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G} = \mathbf{0},$$

which will be called the (G,H)-orthogonal or (G,H)-unitary Procrustes problem. Here the geometry within the tuples is measured with the scalar product $[.,.]_G$ but the geometry between the tuples is measured with the scalar product $[.,.]_H$. The wanted matrix $\mathbf{U}$ has to be both an H-isometry and a G-isometry.

In addition to this, the problem

$$f(\mathbf{U}) = \sum_{k=1}^{N} [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]_H \to \text{opt} \quad \text{with} \tag{1.3}$$
$$\mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G} = \mathbf{0}$$

will also be investigated. Again the distances inside the tuples are measured with the internal metric $\mathbf{G}$ and the distances between the tuples are measured with the external metric $\mathbf{H}$. However, $\mathbf{U}$ is only required to be a G-isometry.

Here and elsewhere the matrix defining a (not necessarily indefinite) scalar product is called a metric in accordance with its meaning in tensor algebra where it, or more precisely its transpose, is called the metric tensor (for example see [WEY, §5]).

1.2 The orthogonal Procrustes problem

Several matrix factorisations play an important role for the analysis of Procrustes problems. In particular, the well-known singular value decomposition (SVD) is very useful.

Proposition 1.1 (Singular value decomposition). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{A} \in \mathbb{F}^{m\times n}$. Then there exist orthogonal or unitary matrices $\mathbf{P} \in \mathbb{F}^{m\times m}$ and $\mathbf{Q} \in \mathbb{F}^{n\times n}$ such that*

$$\mathbf{P}^*\mathbf{A}\mathbf{Q} = \mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \ \ p = \min(m, n),$$

where $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$.

Proof. See in textbooks on linear algebra, for example [GVL, Section 2.5.3]. □

Closely related to the SVD is the following factorisation which can be interpreted as a generalisation of the complex polar coordinates $z = e^{i \arg(z)} |z|$.

Definition 1.2 (Polar decomposition). Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{A} \in \mathbb{F}^{n\times n}$. A factorisation of the form

$$\mathbf{A} = \mathbf{UM} \text{ with } \mathbf{U}^*\mathbf{U} = \mathbf{I} \text{ and } \mathbf{M}^* = \mathbf{M},$$

where $\mathbf{U} \in \mathbb{F}^{n\times n}$ is an isometry and $\mathbf{M} \in \mathbb{F}^{n\times n}$ is selfadjoint, is called a polar decomposition of $\mathbf{A}$. ◊

Since the SVD of a matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ can be rewritten as

$$\mathbf{A} = (\mathbf{PQ}^*)(\mathbf{Q\Sigma Q}^*) = \mathbf{UM},$$

it is clear that every square matrix admits a polar decomposition.

With this background we are able to look at a first Procrustes problem. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N\times n}$. Then the least squares problem of determining an orthogonal matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that the squared Frobenius norm of the residual matrix $\mathbf{AT} - \mathbf{B}$ is a minimum can be formulated as

$$\operatorname{tr}[(\mathbf{AT} - \mathbf{B})^T(\mathbf{AT} - \mathbf{B})] \to \min \text{ with } \mathbf{TT}^T = \mathbf{I}.$$

In this form the so-called orthogonal Procrustes problem was investigated by Schönemann [S]. He showed that the solution is obtained from the singular value decomposition (which he called "Eckart-Young decomposition")

$$\mathbf{A}^T\mathbf{B} = \mathbf{Q\Sigma P}^T$$

by forming

$$\mathbf{T} = \mathbf{QP}^T.$$

Now, defining $\mathbf{A}^T = \mathbf{X}$, $\mathbf{B}^T = \mathbf{Y}$, $\mathbf{T}^T = \mathbf{U}$ and denoting the columns of $\mathbf{X}$ and $\mathbf{Y} \in \mathbb{R}^{n\times N}$ by $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{y}_1, \ldots, \mathbf{y}_N$, respectively, it immediately follows that the transposed problem

$$\operatorname{tr}[(\mathbf{UX} - \mathbf{Y})^T(\mathbf{UX} - \mathbf{Y})] \to \min \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{I}$$

or

$$\sum_{k=1}^{N}(\mathbf{Ux}_k - \mathbf{y}_k, \mathbf{Ux}_k - \mathbf{y}_k) \to \min \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{I}$$

is solved by the isometric factor of the polar decomposition

$$\mathbf{YX}^T = \mathbf{P\Sigma Q}^T = \mathbf{UM}, \ \mathbf{U} = \mathbf{PQ}^T, \ \mathbf{M} = \mathbf{Q\Sigma Q}^T.$$

Thus we have already found the solution of the problem (1.1) in the case $\mathbb{F} = \mathbb{R}$ and $\mathbf{H} = \mathbf{I}$.

It does not surprise that for $\mathbb{F} = \mathbb{C}$ and $\mathbf{H} = \mathbf{I}$ an analogous statement holds. Here the factor $\mathbf{U}$ of the complex polar decomposition $\mathbf{YX}^* = \mathbf{UM}$ is the wanted isometry. However, when $\mathbf{H}$ is a selfadjoint matrix having positive and negative eigenvalues, the things are getting more complicated. In this case the addends in the objectives of (1.1) – (1.3) can be positive as well as negative, so that even the criterion for the optimum of the function f has to be considered. Nevertheless there are a lot of analogies to the definite Procrustes problems, and it turns out that the following generalisation of the polar decomposition helps to solve the indefinite problems.

Definition 1.3 (H-polar decomposition). Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{H} \in \mathbb{F}^{n\times n}$ be nonsingular and selfadjoint. Furthermore, let $\mathbf{A} \in \mathbb{F}^{n\times n}$. A factorisation of the form

$$\mathbf{A} = \mathbf{UM} \text{ with } \mathbf{U}^*\mathbf{HU} = \mathbf{H} \text{ and } \mathbf{M}^*\mathbf{H} = \mathbf{HM},$$

where $\mathbf{U} \in \mathbb{F}^{n\times n}$ is an H-isometry and $\mathbf{M} \in \mathbb{F}^{n\times n}$ is H-selfadjoint, is called an H-polar decomposition of $\mathbf{A}$. ◊

In contrast to the ordinary polar decomposition, not every square matrix admits an H-polar decomposition, so that the theory of the latter is considerably more complicated. We will therefore have to study H-polar decompositions before the Procrustes problems can be investigated.

1.3 Contents and notation

In this thesis we will derive theoretical results and we will also develop numerical methods with which these results can be applied. The presentation is divided into seven chapters which are organised as follows:

In **Chapter 2** some essential properties of indefinite scalar product spaces are described. Several subspace decompositions are given and H-orthogonal bases are discussed. The chapter ends with the introduction of the HQR decomposition which is a generalisation of the QR decomposition in the presence of an indefinite scalar product.

Chapter 3 is concerned with H-polar decompositions. Some important results of the related theory are summarised and a new criterion for the existence of H-polar decompositions is given. Furthermore, the chapter presents algorithms for the numerical computation of H-polar decompositions of a complex matrix $\mathbf{A}$ for which either $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable or $\mathbf{A}^{[*]}\mathbf{A}$ has no non-positive eigenvalues. In this context the H-singular value decomposition is introduced.

Chapter 4 starts with an introduction to doubly structured indefinite polar decompositions, called (G,H)-polar decompositions, followed by a method for constructing points from given values of a quadratic form. Afterwards the Procrustes problems (1.1) and (1.2) are solved with the help of H- or (G,H)-polar decompositions, respectively. At the end of this chapter some more general results on (G,H)-polar decompositions are derived.

In **Chapter 5** a Newton method is developed with which all stated Procrustes problems and further related problems can be solved numerically. This method will in particular be adopted to solve the problem (1.3) for which no analytic solution has been found.

Chapter 6 presents an algorithm for computing the canonical form of the complex matrix pair $(\mathbf{A}, \mathbf{H})$ where $\mathbf{A}$ is H-Hermitian and $\mathbf{H}$ is nonsingular and Hermitian. The algorithm is based on the numerical computation of the Jordan normal form of $\mathbf{A}$ and a subsequent H-orthogonalisation of the bases of the generalised eigenspaces. In connection with results from Chapter 3 this algorithm allows to compute all H-polar decompositions of a matrix $\mathbf{A}$ for which $\mathbf{A}^{[*]}\mathbf{A}$ has no non-negative eigenvalues.

In the final **Chapter 7** the most important results on the Procrustes problems are summarised and explained with an illustrative example.

The chapters can more or less be read independently. For the presentation the following notation is used:

The kernel (null space), the image (range) and the rank of a matrix $\mathbf{A}$ are denoted by $\ker \mathbf{A}$, $\operatorname{im} \mathbf{A}$ and $\operatorname{rank} \mathbf{A}$, respectively. If the matrix $\mathbf{A}$ is square, then $\operatorname{tr} \mathbf{A}$, $\det \mathbf{A}$ and $\sigma(\mathbf{A})$ are its trace, determinant and spectrum, respectively. Furthermore, the abbreviation $\mathbf{A}^{-*} = (\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})^*$ is used.

The symbol $\mathbf{0}$ denotes zero vectors as well as zero matrices. In some places it is additionally provided with size attributes $\mathbf{0}_{p,q} \in \mathbb{F}^{p\times q}$ or $\mathbf{0}_p \in \mathbb{F}^{p\times p}$, but lower indices may also be intended as enumeration indices. This is evident from the respective context.

$\mathbf{I}_p$, $\mathbf{N}_p$ and $\mathbf{Z}_p$ specify the $p \times p$ identity matrix, the $p \times p$ matrix with ones on the superdiagonal and otherwise zeros, and the $p \times p$ matrix with ones on the antidiagonal and otherwise zeros. In particular $\mathbf{J}_p(\lambda) = \lambda\mathbf{I}_p + \mathbf{N}_p$ is an upper Jordan block with eigenvalue λ and $\mathbf{Z}_p$ is called a sip (standard involutary permutation) block. In explicit formulas we have

$$\mathbf{J}_p(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \text{ and } \mathbf{Z}_p = \begin{bmatrix} & & & 1 \\ & & \iddots & \\ & \iddots & & \\ 1 & & & \end{bmatrix}.$$

The notation $\mathbf{A}_1 \oplus \ldots \oplus \mathbf{A}_k$ represents a block diagonal matrix consisting of the specified blocks, and $\operatorname{diag}(\alpha_1, \ldots, \alpha_k)$ stands for a possibly rectangular diagonal matrix with the specified diagonal elements. Moreover, $X \oplus Y$ also denotes the direct sum of two subspaces $X, Y \subset \mathbb{F}^n$.

Whereas only the Euclidean vector norm $\|\mathbf{x}\| = \sqrt{\mathbf{x}^*\mathbf{x}}$ is required, different matrix norms are used. If $\mathbf{A} = [a_{ij}]$ is an $m \times n$ matrix, then

$$\|\mathbf{A}\|_F = \sqrt{\operatorname{tr} \mathbf{A}^*\mathbf{A}} \text{ and } \|\mathbf{A}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$$

are its Frobenius (Euclidean) norm and 1-norm, respectively.

Even when no further specifications are made, a nonsingular (real) symmetric or (complex) Hermitian matrix is always meant by $\mathbf{H}$.

1.4 Acknowledgements

The present work was developed in recent years while I was also working as a software engineer for a small company in Germany. After having accomplished the first presentable results (parts of Chapter 4 and 5) I started to look for an academic authority and found in Martin Buhmann (Justus-Liebig-Universität Gießen) an interested discussion partner. I would like to thank him for the given hints and support.

Since my investigations did not match Professor Buhmann's subject of research exactly, we tried to find out who might also be interested in my work, where the contact to Peter Benner (Technische Universität Chemnitz) resulted from. From that time on, Professor Benner accompanied my external research activities. He had open ears for all questions that arouse and supported me in

any possible way. I benefited very much from his help and would like to express my special thanks to him.

Next I thank Volker Mehrmann (Technische Universität Berlin) and Leiba Rodman (The College of William and Mary) for having refereed this work. Moreover, I would like to thank Martin L. Michaelis for his assistance in translating the first results into the English language and my employer, Klaus Nonne, for his generous working time regulation. Last not least, I also thank my girl friend Andrea for her patience and for always providing a place of recovery together with our lovely cats Max and Lucy.

Chapter 2

Indefinite scalar product spaces

2.1 Introduction

Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbb{F}^n$ be a vector space in which the scalar product $[.,.] = (\mathbf{H}.,.)$ is defined. Then the assumption that $\mathbf{H}$ has positive and negative eigenvalues has far reaching consequences. In particular, the subspaces $M \subset \mathbb{F}^n$ have properties which are not known from the subspaces of Euclidean or unitary spaces.

In this chapter we will therefore prepare our further investigations by studying some important aspects of indefinite scalar product spaces. In Section 2.2 several subspace decompositions will be derived. In Section 2.3 the construction of H-orthogonal bases of subspaces and their extension to H-orthogonal bases of $\mathbb{F}^n$ will be discussed. The results obtained in these studies will then be used to generalise the well-known QR factorisation so that it allows to compute an indefinite HQR factorisation. This takes place in the final Section 2.4.

2.2 Subspace decompositions

The properties of subspaces of real or complex indefinite scalar product spaces are discussed in detail in [GLR, Chapter I.1]. We find the following basic definitions and statements there:

Definition 2.1.

(i) A subspace $M \subset \mathbb{F}^n$ is called positive (non-negative, neutral, non-positive, negative) if

$$[\mathbf{x},\mathbf{x}] > 0 \;\; ([\mathbf{x},\mathbf{x}] \geq 0, \;\; [\mathbf{x},\mathbf{x}] = 0, \;\; [\mathbf{x},\mathbf{x}] \leq 0, \;\; [\mathbf{x},\mathbf{x}] < 0)$$

is satisfied for all $\mathbf{0} \neq \mathbf{x} \in M$.

(ii) A subspace M is called non-degenerate if $\mathbf{x} \in M$ and $[\mathbf{x},\mathbf{y}] = 0$ for all $\mathbf{y} \in M$ imply that $\mathbf{x} = \mathbf{0}$, otherwise M is called degenerate.

Proposition 2.2.

(*i*) *Let $M \subset \mathbb{F}^n$. The set defined by*

$$M^{[\perp]} = \{\mathbf{x} \in \mathbb{F}^n : [\mathbf{x}, \mathbf{y}] = 0 \text{ for all } \mathbf{y} \in M\}$$

is also a subspace of $\mathbb{F}^n$ and is termed the H-orthogonal companion of M.

(*ii*) *It is true that*

$$(M^{[\perp]})^{[\perp]} = M \quad \text{and} \quad \dim M + \dim M^{[\perp]} = n.$$

(*iii*) *It is true that*

$$M \cap M^{[\perp]} = \{\mathbf{0}\} \quad \text{and} \quad M \oplus M^{[\perp]} = \mathbb{F}^n$$

if and only if M is non-degenerate[2]*.*

It is furthermore shown in [GLR, Theorem I.1.4] that every non-negative (non-positive) subspace is a direct sum of a positive (negative) and a neutral subspace. In addition to this, the following more general theorem holds, whose proof is based on statements made in [GR, Sections 9.6, 9.7].

Theorem 2.3 (Decomposition of subspaces).

1. *Every non-degenerate subspace $M \subset \mathbb{F}^n$ can be expressed as a direct sum $M = M_+ \oplus M_-$ where M_+ is positive, M_- is negative and the spaces are H-orthogonal to each other.*

2. *Every subspace $M \subset \mathbb{F}^n$ can be expressed as a direct sum $M = M_0 \oplus M_1$ where M_0 is neutral, M_1 is non-degenerate and the spaces are H-orthogonal to each other.*

Proof. 1. Let M_+ be a positive subspace of M with maximum dimension. Then M_+ is non-degenerate and $M_+ \oplus M_+^{[\perp]} = \mathbb{F}^n$. Thus a representation

$$M_+ \oplus (M \cap M_+^{[\perp]}) = M, \quad M_- = M \cap M_+^{[\perp]},$$

exists with two H-orthogonal addends, and it remains to show that M_- is negative. Suppose that a vector $\mathbf{x} \in M_-$ exists with $[\mathbf{x}, \mathbf{x}] > 0$. Then it would follow that $[\mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y}] = [\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] > 0$ for all $\mathbf{y} \in M_+$. But this would mean that the subspace $M_+ \oplus \text{span}\{\mathbf{x}\}$ is also positive, in contradiction to the maximality of M_+. Thus M_- is non-positive and the Schwarz inequality [GLR, Chapter I.1.3][3]

$$|[\mathbf{x}, \mathbf{y}]|^2 \leq [\mathbf{x}, \mathbf{x}][\mathbf{y}, \mathbf{y}] \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in M_-$$

can be applied. Now assume that $\mathbf{x}_0 \in M_-$ with $[\mathbf{x}_0, \mathbf{x}_0] = 0$. Then the Schwarz inequality shows that $[\mathbf{x}_0, \mathbf{x}] = 0$ must hold for all $\mathbf{x} \in M_-$. Since it is also true that $[\mathbf{x}_0, \mathbf{y}] = 0$ for all $\mathbf{y} \in M_+$, it follows that $[\mathbf{x}_0, \mathbf{z}] = 0$ for all $\mathbf{z} \in M$. Consequently $\mathbf{x}_0 = \mathbf{0}$, because M is non-degenerate.

[2] This is an essential difference compared with the ordinary scalar product, for which these equations are always fulfilled for the ordinary orthogonal complement $M^{\perp}$.

[3] There is a typing error contained in equation (1.8): It must be read $|(Hy, z)|^2 \leq (Hy, y)(Hz, z)$.

2. Let $M_0 = M \cap M^{[\perp]}$. Then M_0 is neutral, because if a vector $\mathbf{x} \in M_0 \subset M$ were to exist with $[\mathbf{x}, \mathbf{x}] \neq 0$, it would follow that $\mathbf{x} \notin M^{[\perp]} \supset M_0$. Now let M_1 be a complementary subspace, so that

$$(M \cap M^{[\perp]}) \oplus M_1 = M, \quad M_0 = M \cap M^{[\perp]},$$

with two H-orthogonal addends is satisfied. To show that M_1 is non-degenerate, let $\mathbf{x}_0 \in M_1$ with $[\mathbf{x}_0, \mathbf{x}] = 0$ for all $\mathbf{x} \in M_1$. Furthermore, $[\mathbf{x}_0, \mathbf{y}] = 0$ for all $\mathbf{y} \in M_0$, so that $[\mathbf{x}_0, \mathbf{z}] = 0$ for all $\mathbf{z} \in M$. Thus it follows that $\mathbf{x}_0 \in M_1$ and $\mathbf{x}_0 \in M_0$, so that $\mathbf{x}_0 = \mathbf{0}$. □

On combining the two statements of the theorem, it is clear that every subspace $M \subset \mathbb{F}^n$ can be expressed in the form

$$M = M_+ \oplus M_- \oplus M_0$$

with a positive, a negative and a neutral — mutually H-orthogonal — subspace. In order to deduce the dimensions of these subspaces, we refer to the following classical result [GR, Sections 9.8, 9.9].

Remark 2.4 (Projection onto subspaces). Let $M = \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subspace of $\mathbb{F}^n$. Then every vector $\mathbf{y} \in M$,

$$\mathbf{y} = \sum_{\mu=1}^{m} \eta_\mu \mathbf{x}_\mu,$$

can be represented uniquely by its coordinates $\tilde{\mathbf{y}} = (\eta_1, \ldots, \eta_m)^T \in \mathbb{F}^m$ with respect to the given basis of M. If now $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_m] \in \mathbb{F}^{n \times m}$ is a matrix whose columns are the basis vectors, then $\mathbf{y} = \mathbf{X}\tilde{\mathbf{y}}$ and for $\tilde{\mathbf{H}} = \mathbf{X}^* \mathbf{H} \mathbf{X} \in \mathbb{F}^{m \times m}$ we obtain

$$(\mathbf{H}\mathbf{y}, \mathbf{z})_n = (\mathbf{H}\mathbf{X}\tilde{\mathbf{y}}, \mathbf{X}\tilde{\mathbf{z}})_n = (\mathbf{X}^*\mathbf{H}\mathbf{X}\tilde{\mathbf{y}}, \tilde{\mathbf{z}})_m = (\tilde{\mathbf{H}}\tilde{\mathbf{y}}, \tilde{\mathbf{z}})_m$$

where

$$(\mathbf{x}, \mathbf{y})_k = \sum_{\alpha=1}^{k} x_\alpha \overline{y}_\alpha.$$

Consequently the properties of the non-degenerate scalar product $\mathbf{H} : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}$ in the subspace M can be studied with the help of the possibly degenerate scalar product $\tilde{\mathbf{H}} : \mathbb{F}^m \times \mathbb{F}^m \to \mathbb{F}$. In particular, if $\sigma(\tilde{\mathbf{H}})$ contains p positive and q negative eigenvalues, and if $r = m - p - q$ is the multiplicity of the eigenvalue 0, then for the decomposition of M described above it holds that

$$\dim M_+ = p, \quad \dim M_- = q, \quad \dim M_0 = r.$$

The dimensions of these subspaces are uniquely determined. This is a consequence of Sylvester's law of inertia, according to which the numbers of positive, negative and vanishing elements are invariant for all diagonal representations of $\tilde{\mathbf{H}}$. Furthermore, the subspace M_0 is uniquely determined by the nullspace $\ker \tilde{\mathbf{H}}$ of the degenerate scalar product. In particular, if M is a non-degenerate subspace, then $\det \tilde{\mathbf{H}} \neq 0$, i.e. $r = 0$. In this case the maximum dimension of a neutral subspace of M is given by $\min(p, q)$ which is proved in [GLR, Theorem I.1.5]. ◊

According to Proposition 2.2 (iii) a non-degenerate subspace $M \subset \mathbb{F}^n$ induces a decomposition of $\mathbb{F}^n$ into the two complementary subspaces M and $M^{[\perp]}$. A generalisation of this statement is given by the following theorem. It describes the interesting fact that an arbitrary (degenerate) subspace induces a decomposition of $\mathbb{F}^n$ into four complementary subspaces.

Theorem 2.5 (Decomposition of the space). *Let $M \subset \mathbb{F}^n$. Then four subspaces $M_1, M_2, M_0', M_0'' \subset \mathbb{F}^n$ exist with the following properties:*

1. $\mathbb{F}^n = M_0 \oplus M_1 \oplus M_2$ *with* $M_0 = M_0' \oplus M_0''$.
2. $M_0' = M \cap M^{[\perp]}$ *and* $M = M_1 \oplus M_0'$ *as well as* $M^{[\perp]} = M_2 \oplus M_0'$.
3. M_0, M_1, M_2 *are non-degenerate and mutually H-orthogonal.*
4. M_0', M_0'' *are neutral and* $\dim M_0' = \dim M_0''$.

Proof. Let M_1 and M_2 be the complements of M_0' which exist according to Theorem 2.3 and for which the assertion 2. is fulfilled. Then $M_1 \subset M$ and $M_2 \subset M^{[\perp]}$ are H-orthogonal and non-degenerate, so that $M_1 \oplus M_2$ is also non-degenerate. Consequently

$$\mathbb{F}^n = (M_1 \oplus M_2) \oplus (M_1 \oplus M_2)^{[\perp]}$$

and, moreover, $M_0' \subset (M_1 \oplus M_2)^{[\perp]}$. If we now choose $M_0 = (M_1 \oplus M_2)^{[\perp]} = M_0' \oplus M_0''$, then assertions 1. and 3. are fulfilled, too. From

$$\mathbb{F}^n = (M_1 \oplus M_2 \oplus M_0') \oplus M_0'' = (M + M^{[\perp]}) \oplus M_0''$$

it furthermore follows that

$$\begin{aligned} \dim M_0'' &= n - \dim(M + M^{[\perp]}) \\ &= n - (\dim M + \dim M^{[\perp]} - \dim(M \cap M^{[\perp]})) \\ &= n - (n - \dim M_0') \\ &= \dim M_0', \end{aligned}$$

where the well-known dimension theorem [GR, Section 1.21]

$$\dim M + \dim N = \dim(M + N) + \dim(M \cap N) \quad \text{for} \quad M, N \subset \mathbb{F}^n$$

and Proposition 2.2 (ii) have been applied.

It remains to show that M_0'' is neutral. Let $r = \dim M_0' = \dim M_0''$. Then M_0 is a $2r$-dimensional non-degenerate subspace of $\mathbb{F}^n$, which can be split according to Theorem 2.1 into a positive and a negative subspace $M_0 = M_0^+ \oplus M_0^-$. Let $p = \dim M_0^+$ and $q = \dim M_0^-$. Since M_0 must contain the r-dimensional neutral subspace M_0' it follows that $r \leq \min(p, q)$ and thus $p \geq r$ and $q \geq r$ [GLR, Theorem I.1.5]. On the other hand $p + q = 2r$, so that $p = q = r$. Therefore, the subspace M_0 admits the decompositions

$$\begin{gathered} M_0 = M_0^+ \oplus M_0^- = M_0' \oplus M_0'' \text{ with} \\ \dim M_0^+ = \dim M_0^- = \dim M_0' = \dim M_0'' \end{gathered}$$

and H-orthogonal spaces M_0^+, M_0^-, so that the three bases

$$M_0^+ = \text{span}\{\mathbf{x}_1^+, \ldots, \mathbf{x}_r^+\}, \quad M_0^- = \text{span}\{\mathbf{x}_1^-, \ldots, \mathbf{x}_r^-\}, \quad M_0' = \text{span}\{\mathbf{x}_1', \ldots, \mathbf{x}_r'\}$$
$$\text{with } [\mathbf{x}_k^+, \mathbf{x}_l^+] > 0, \ [\mathbf{x}_k^-, \mathbf{x}_l^-] < 0, \ [\mathbf{x}_k^+, \mathbf{x}_l^-] = 0 \text{ and } [\mathbf{x}_k', \mathbf{x}_l'] = 0$$

for $1 \leq k, l \leq r$ can now be chosen. Since M_0^+ is positive, M_0^- is negative and M_0' is neutral, it must also be true that $M_0^+ \cap M_0' = M_0^- \cap M_0' = \{\mathbf{0}\}$, so that each basis vector of M_0' can be expressed in the form

$$\mathbf{x}_k' = \sum_{i=1}^{r} \alpha_{ki}\mathbf{x}_i^+ + \sum_{i=1}^{r} \beta_{ki}\mathbf{x}_i^- \text{ with } (\alpha_{k1}, \ldots, \alpha_{kr})^T, (\beta_{k1}, \ldots, \beta_{kr})^T \neq \mathbf{0}.$$

Furthermore, the vectors defined by

$$\tilde{\mathbf{x}}_k^+ = \sum_{i=1}^{r} \alpha_{ki}\mathbf{x}_i^+ \text{ and } \tilde{\mathbf{x}}_k^- = \sum_{i=1}^{r} \beta_{ki}\mathbf{x}_i^-$$

can be used as a new basis of M_0^+, M_0^-. Indeed, if it is assumed that the constants $(\lambda_1, \ldots, \lambda_r) \neq \mathbf{0}$ with $\lambda_1\tilde{\mathbf{x}}_1^+ + \ldots, \lambda_r\tilde{\mathbf{x}}_r^+ = \mathbf{0}$ exist, then $\mathbf{0} \neq \lambda_1\mathbf{x}_1' + \ldots + \lambda_r\mathbf{x}_r' = \lambda_1(\tilde{\mathbf{x}}_1^+ + \tilde{\mathbf{x}}_1^-) + \ldots + \lambda_r(\tilde{\mathbf{x}}_r^+ + \tilde{\mathbf{x}}_r^-) = \lambda_1\tilde{\mathbf{x}}_1^- + \ldots + \lambda_r\tilde{\mathbf{x}}_r^- \in M_0^-$ and thus $M_0^- \cap M_0' \neq \{\mathbf{0}\}$. The linear independence of the vectors $\tilde{\mathbf{x}}_1^-, \ldots, \tilde{\mathbf{x}}_r^-$ can be shown analogously. Finally, defining

$$\mathbf{x}_k'' = \tilde{\mathbf{x}}_k^+ - \tilde{\mathbf{x}}_k^- \text{ for } 1 \leq k \leq r \text{ and } M_0'' = \text{span}\{\mathbf{x}_1'', \ldots, \mathbf{x}_r''\},$$

then M_0'' is on the one hand a neutral subspace because

$$\begin{aligned}[\mathbf{x}_k'', \mathbf{x}_l''] &= [\tilde{\mathbf{x}}_k^+ - \tilde{\mathbf{x}}_k^-, \tilde{\mathbf{x}}_l^+ - \tilde{\mathbf{x}}_l^-] = [\tilde{\mathbf{x}}_k^+, \tilde{\mathbf{x}}_l^+] + [\tilde{\mathbf{x}}_k^-, \tilde{\mathbf{x}}_l^-] \\ &= [\tilde{\mathbf{x}}_k^+ + \tilde{\mathbf{x}}_k^-, \tilde{\mathbf{x}}_l^+ + \tilde{\mathbf{x}}_l^-] = [\mathbf{x}_k', \mathbf{x}_l'] = 0\end{aligned}$$

and on the other hand

$$\begin{aligned}M_0 &= M_0^+ \oplus M_0^- \\ &= \text{span}\{\tilde{\mathbf{x}}_1^+, \ldots, \tilde{\mathbf{x}}_r^+\} \oplus \text{span}\{\tilde{\mathbf{x}}_1^-, \ldots, \tilde{\mathbf{x}}_r^-\} \\ &= \text{span}\{\tilde{\mathbf{x}}_1^+ + \tilde{\mathbf{x}}_1^-, \ldots, \tilde{\mathbf{x}}_r^+ + \tilde{\mathbf{x}}_r^-\} \oplus \text{span}\{\tilde{\mathbf{x}}_1^+ - \tilde{\mathbf{x}}_1^-, \ldots, \tilde{\mathbf{x}}_r^+ - \tilde{\mathbf{x}}_r^-\} \\ &= M_0' \oplus M_0'',\end{aligned}$$

so that the assertion 4. of the theorem is fulfilled, too. □

2.3 H-orthogonal bases

Whereas the statements have been proved so far without reference to particular bases, we will also have to use H-orthogonal bases. The following two theorems contain generalisations of the Gram-Schmidt orthonormalisation method, with the help of which such bases can be constructed. Both theorems will in particular be applied for the H-orthogonalisation of eigenspaces of H-selfadjoint matrices (Theorem 2.6 for eigenspaces belonging to real and Theorem 2.7 for eigenspaces belonging to non-real eigenvalues).

Theorem 2.6 (H-Orthonormalisation of bases). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let X be a subspace of $\mathbb{F}^n$ with $\dim X = m$. Then there exists a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of X such that*

$$[\mathbf{u}_k, \mathbf{u}_l] = \varepsilon_k \delta_{kl}, \quad \varepsilon_k = \begin{cases} +1, & \text{for } 1 \leq k \leq p \\ -1, & \text{for } p+1 \leq k \leq p+q \\ 0, & \text{for } p+q+1 \leq k \leq p+q+r \end{cases}$$

where $p + q + r = m$. In particular, if X is non-degenerate, then $r = 0$.

Proof. (Induction) First assume that X is non-degenerate and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a basis of X. Also let k, l be two indices in $\{1, \ldots, m\}$ such that $|[\mathbf{x}_k, \mathbf{x}_l]|$ is maximised. Then it necessarily follows that $[\mathbf{x}_k, \mathbf{x}_l] \neq 0$, because otherwise X would be degenerate. For the case $k = l$ let the basis which is obtained by interchanging $\mathbf{x}_1$ and $\mathbf{x}_k$ still be denoted as $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Otherwise, let

$$\tilde{\mathbf{x}}_k = \frac{1}{\sqrt{2}}(\mathbf{x}_k + \varphi \mathbf{x}_l) \text{ and } \tilde{\mathbf{x}}_l = \frac{1}{\sqrt{2}}(\mathbf{x}_k - \varphi \mathbf{x}_l)$$

where $\varphi = [\mathbf{x}_k, \mathbf{x}_l]/|[\mathbf{x}_k, \mathbf{x}_l]|$. Then $\operatorname{span}\{\tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_l\} = \operatorname{span}\{\mathbf{x}_k, \mathbf{x}_l\}$ and

$$[\tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_k] = \alpha + |[\mathbf{x}_k, \mathbf{x}_l]| \text{ and } [\tilde{\mathbf{x}}_l, \tilde{\mathbf{x}}_l] = \alpha - |[\mathbf{x}_k, \mathbf{x}_l]|$$

where $\alpha = ([\mathbf{x}_k, \mathbf{x}_k] + [\mathbf{x}_l, \mathbf{x}_l])/2$. Hence, for

$$\tilde{\mathbf{y}} = \begin{cases} \tilde{\mathbf{x}}_k, & \text{if } \alpha \geq 0 \\ \tilde{\mathbf{x}}_l, & \text{if } \alpha < 0 \end{cases}$$

it is always true that $[\tilde{\mathbf{y}}, \tilde{\mathbf{y}}] \neq 0$. Let the particular basis obtained by replacing $\mathbf{x}_k, \mathbf{x}_l$ with $\tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_l$ and then exchanging $\mathbf{x}_1$ and $\tilde{\mathbf{y}}$ still be denoted as $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. If we now set

$$\mathbf{u}_1 = \mathbf{x}_1/\sqrt{|[\mathbf{x}_1, \mathbf{x}_1]|} \text{ and } \varepsilon_1 = \operatorname{sign}[\mathbf{x}_1, \mathbf{x}_1] \in \{+1, -1\},$$

then $[\mathbf{u}_1, \mathbf{u}_1] = [\mathbf{x}_1, \mathbf{x}_1]/|[\mathbf{x}_1, \mathbf{x}_1]| = \varepsilon_1$ and for the vectors defined by

$$\mathbf{x}'_i = \mathbf{x}_i - \varepsilon_1 [\mathbf{x}_i, \mathbf{u}_1] \mathbf{u}_1 \text{ for } 2 \leq i \leq m$$

we obtain

$$[\mathbf{x}'_i, \mathbf{u}_1] = [\mathbf{x}_i, \mathbf{u}_1] - \varepsilon_1 [\mathbf{x}_i, \mathbf{u}_1][\mathbf{u}_1, \mathbf{u}_1] = 0.$$

Thus X can be expressed as direct sum of its H-orthogonal subspaces $\operatorname{span}\{\mathbf{u}_1\}$ and $X' = \operatorname{span}\{\mathbf{x}'_2, \ldots, \mathbf{x}'_m\}$, so that X', too, is non-degenerate. Now according to the induction hypothesis there exists a basis $\{\mathbf{u}_2, \ldots, \mathbf{u}_m\}$ of X' with the demanded properties, so that finally $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is the wanted basis of X, if a suitable sorting is also made in the case of $\varepsilon_1 = -1$.

If X is a degenerate subspace, the same construction can be applied, but it then terminates after a certain number of steps, namely when no more non-zero scalar products can be found. The remaining r vectors $\mathbf{x}'_i$ then satisfy $[\mathbf{x}'_i, \mathbf{x}'_j] = 0$ for $m - r + 1 \leq i, j \leq m$. □

Theorem 2.7 (H-Orthonormalisation of pairs of bases). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let X, Y be two neutral subspaces of $\mathbb{F}^n$ with $\dim X = \dim Y = m$ and $X \cap Y = \{\mathbf{0}\}$. Then there exists a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of X and a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ of Y such that*

$$[\mathbf{u}_k, \mathbf{v}_l] = \varepsilon_k \delta_{kl}, \;\; \varepsilon_k = \begin{cases} 1, & \text{for } 1 \le k \le p \\ 0, & \text{for } p+1 \le k \le p+r \end{cases}$$

where $p + r = m$. In particular, if $X \oplus Y$ is non-degenerate, then $r = 0$.

Proof. (Induction) First assume that $X \oplus Y$ is non-degenerate and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a basis of X and $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ be a basis of Y. Also let k, l be two indices in $\{1, \ldots, m\}$ so that $|[\mathbf{x}_k, \mathbf{y}_l]|$ is maximised. Then it necessarily follows that $[\mathbf{x}_k, \mathbf{y}_l] \neq 0$, because otherwise $X \oplus Y$ would be degenerate. Let the particular bases obtained by exchanging $\mathbf{x}_1$ and $\mathbf{x}_k$ as well as $\mathbf{y}_1$ and $\mathbf{y}_l$ still be denoted as $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ respectively. In the case $\mathbb{F} = \mathbb{R}$ now let $\varepsilon_1 \in \{+1, -1\}$ such that $\lambda_1 = [\mathbf{x}_1, \varepsilon_1 \mathbf{y}_1] > 0$ and let

$$\mathbf{u}_1 = \mathbf{x}_1 / \sqrt{\lambda_1} \;\text{ and }\; \mathbf{v}_1 = \varepsilon_1 \mathbf{y}_1 / \sqrt{\lambda_1},$$

so that $[\mathbf{u}_1, \mathbf{v}_1] = [\mathbf{x}_1, \varepsilon_1 \mathbf{y}_1] / [\mathbf{x}_1, \varepsilon_1 \mathbf{y}_1] = 1$; in the case $\mathbb{F} = \mathbb{C}$ let $\omega_1^2 = \lambda_1 = [\mathbf{x}_1, \mathbf{y}_1]$ and let

$$\mathbf{u}_1 = \mathbf{x}_1 / \omega_1 \;\text{ and }\; \mathbf{v}_1 = \mathbf{y}_1 / \overline{\omega}_1,$$

so that $[\mathbf{u}_1, \mathbf{v}_1] = [\mathbf{x}_1, \mathbf{y}_1] / [\mathbf{x}_1, \mathbf{y}_1] = 1$. For the vectors defined by

$$\mathbf{x}_i' = \mathbf{x}_i - [\mathbf{x}_i, \mathbf{v}_1]\mathbf{u}_1 \;\text{ and }\; \mathbf{y}_i' = \mathbf{y}_i - [\mathbf{y}_i, \mathbf{u}_1]\mathbf{v}_1 \;\text{ for }\; 2 \le i \le m$$

we then obtain

$$\begin{aligned} [\mathbf{x}_i', \mathbf{v}_1] &= [\mathbf{x}_i, \mathbf{v}_1] - [\mathbf{x}_i, \mathbf{v}_1][\mathbf{u}_1, \mathbf{v}_1] = 0 \;\text{ and} \\ [\mathbf{y}_i', \mathbf{u}_1] &= [\mathbf{y}_i, \mathbf{u}_1] - [\mathbf{y}_i, \mathbf{u}_1][\mathbf{v}_1, \mathbf{u}_1] = 0. \end{aligned}$$

Thus $X \oplus Y$ can be expressed as direct sum of its H-orthogonal subspaces $\operatorname{span}\{\mathbf{u}_1, \mathbf{v}_1\}$ and $X' \oplus Y' = \operatorname{span}\{\mathbf{x}_2', \ldots, \mathbf{x}_m'\} \oplus \operatorname{span}\{\mathbf{y}_2', \ldots, \mathbf{y}_m'\}$, so that $X' \oplus Y'$, too, is non-degenerate. Now according to the induction hypothesis, two bases $\{\mathbf{u}_2, \ldots, \mathbf{u}_m\}$ and $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$ of X' and Y' exist with the demanded properties, so that finally $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ are the wanted bases of X and Y.

If $X \oplus Y$ is a degenerate subspace, the same construction can be applied, but it then terminates after a certain number of steps, namely when no more non-zero scalar products can be found. The remaining $2r$ vectors $\mathbf{x}_i', \mathbf{y}_i'$ then satisfy $[\mathbf{x}_i', \mathbf{y}_j'] = 0$ for $m - r + 1 \le i, j \le m$. □

These H-orthogonalisation methods are not only interesting for theoretical purposes. With some slight modifications they are also useful numerical methods. But before explaining this, we will first continue to develop the theory required for the investigation of H-polar decompositions in the next chapter.

By comparing Theorem 2.6 with Theorem 2.3 it is easily seen that the positive, negative and neutral vectors of the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ form bases of the subspaces M_+, M_-, and M_0, respectively. A corresponding basis representation of Theorem 2.5 is provided in the following statement, whose proof corrects an error made in [BR, Theorem 4.1] and [BMRRR2, Theorem 2.1].

Theorem 2.8 (Extension of bases). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let X be a subspace of $\mathbb{F}^n$ with $\dim X = m$. Then there exists a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ of $\mathbb{F}^n$ which has the following properties:*

1. *If $\mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_n]$ is a matrix whose columns are the basis vectors, then*

$$\mathbf{U}^*\mathbf{H}\mathbf{U} = \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_q \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0} & \mathbf{I}_r \\ \mathbf{I}_r & \mathbf{0} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_t \end{bmatrix}$$

   where $p + q + r = m$ and $p + q + 2r + s + t = n$.

2. *If the subspaces X_1, X_0', X_0'', X_2 are defined by*

$$\begin{aligned} X_1 &= \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{p+q}\}, \\ X_0' &= \operatorname{span}\{\mathbf{u}_{p+q+1}, \ldots, \mathbf{u}_{p+q+r}\}, \\ X_0'' &= \operatorname{span}\{\mathbf{u}_{p+q+r+1}, \ldots, \mathbf{u}_{p+q+2r}\}, \\ X_2 &= \operatorname{span}\{\mathbf{u}_{p+q+2r+1}, \ldots, \mathbf{u}_{p+q+2r+s+t}\}, \end{aligned}$$

   then X_0', X_0'' are neutral subspaces with equal dimensions, X_1, X_2 and $X_0 = X_0' \oplus X_0''$ are non-degenerate and mutually H-orthogonal, and $\mathbb{F}^n = X_0 \oplus X_1 \oplus X_2$ as well as

$$X = X_1 \oplus X_0', \quad X^{[\perp]} = X_2 \oplus X_0', \quad X \cap X^{[\perp]} = X_0'.$$

Proof. Let $p + q + r = m$ and let $E = \{\mathbf{e}_i\}_{i=1}^m$ be a basis of X which exists according to Theorem 2.6 such that

$$[\mathbf{e}_i, \mathbf{e}_j] = \begin{cases} +1, & \text{for } r+1 \le i = j \le r+p \\ -1, & \text{for } r+p+1 \le i = j \le r+p+q \\ 0, & \text{otherwise} \end{cases}.$$

Furthermore, let $\tilde{E} = \{\widetilde{\mathbf{e}}_i\}_{i=1}^m$ be a dual basis with respect to E , i.e.

$$[\mathbf{e}_i, \widetilde{\mathbf{e}}_j] = \delta_{ij} \text{ for } 1 \le i, j \le m.$$

Then the vectors defined by[4]

$$\widetilde{\widetilde{\mathbf{e}}}_k = \widetilde{\mathbf{e}}_k - \frac{1}{2}\sum_{\mu=1}^{r} [\widetilde{\mathbf{e}}_k, \widetilde{\mathbf{e}}_\mu]\, \mathbf{e}_\mu \text{ for } 1 \le k \le r$$

satisfy

$$\begin{aligned} [\widetilde{\widetilde{\mathbf{e}}}_k, \widetilde{\widetilde{\mathbf{e}}}_l] &= 0 \text{ for } 1 \le k, l \le r \text{ and} \\ [\mathbf{e}_i, \widetilde{\widetilde{\mathbf{e}}}_k] &= \delta_{ik} \text{ for } 1 \le i \le m, \ 1 \le k \le r. \end{aligned}$$

If we now set

$$\mathbf{e}_k' = \frac{1}{\sqrt{2}}(\mathbf{e}_k + \widetilde{\widetilde{\mathbf{e}}}_k) \text{ and } \mathbf{e}_k'' = \frac{1}{\sqrt{2}}(\mathbf{e}_k - \widetilde{\widetilde{\mathbf{e}}}_k) \text{ for } 1 \le k \le r,$$

[4]The same construction is also specified in [BR] and in [BMRRR2] within the scope of the proof for Witt's theorem. However, the necessary orthonormalisation of the vectors $\tilde{\mathbf{e}}_k$ is there not carried out completely, so that the basis $\{\mathbf{e}_i, \mathbf{e}_k', \mathbf{e}_k''\}$ also constructed there is **not** orthonormalised in general.

then it follows that

$$[\mathbf{e}'_k, \mathbf{e}'_l] = \delta_{kl}, \;\; [\mathbf{e}''_k, \mathbf{e}''_l] = -\delta_{kl}, \;\; [\mathbf{e}'_k, \mathbf{e}''_l] = 0 \;\text{ for }\; 1 \le k, l \le r \;\text{ and}$$
$$[\mathbf{e}_i, \mathbf{e}'_k] = 0, \;\; [\mathbf{e}_i, \mathbf{e}''_k] = 0 \;\text{ for }\; r+1 \le i \le m, \;\; 1 \le k \le r.$$

Thus the set of the vectors

$$\{\mathbf{u}_1, \ldots, \mathbf{u}_{m+r}\} = \{\mathbf{e}_i\}_{i=r+1}^{m} \cup \{\mathbf{e}'_k\}_{k=1}^{r} \cup \{\mathbf{e}''_k\}_{k=1}^{r}$$

forms an orthonormalised basis of a non-degenerate subspace $Y \subset \mathbb{F}^n$ which can be extended with $n-m-r$ further vectors $\mathbf{u}_{m+r+1}, \ldots, \mathbf{u}_n$ to an orthonormalised basis of $\mathbb{F}^n$

$$[\mathbf{u}_i, \mathbf{u}_j] = \varepsilon_i \delta_{ij}, \;\; \varepsilon_i \in \{+1, -1\} \;\text{ for }\; 1 \le i, j \le n.$$

For the matrix $\mathbf{U}$ consisting of these basis vectors we have

$$\mathbf{U}^* \mathbf{H} \mathbf{U} = (\mathbf{I}_p \oplus -\mathbf{I}_q) \oplus (\mathbf{I}_r \oplus -\mathbf{I}_r) \oplus (\mathbf{I}_s \oplus -\mathbf{I}_t),$$

where s specifies the number of positive and t specifies the number of negative extending vectors, and a suitable sorting is assumed. Instead of the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ it is also possible to use the basis

$$\{\mathbf{u}_1, \ldots, \mathbf{u}_{p+q}, \tilde{\mathbf{u}}_{p+q+1}, \ldots, \tilde{\mathbf{u}}_{p+q+2r}, \mathbf{u}_{p+q+2r+1}, \ldots, \mathbf{u}_n\}$$
$$\text{with } \{\tilde{\mathbf{u}}_{p+q+1}, \ldots, \tilde{\mathbf{u}}_{p+q+2r}\} = \{\mathbf{e}_k\}_{k=1}^{r} \cup \{\widetilde{\widetilde{\mathbf{e}}}_k\}_{k=1}^{r}.$$

For the matrix $\tilde{\mathbf{U}}$ consisting of these basis vectors we have

$$\tilde{\mathbf{U}}^* \mathbf{H} \tilde{\mathbf{U}} = \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_q \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0} & \mathbf{I}_r \\ \mathbf{I}_r & \mathbf{0} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_t \end{bmatrix},$$

and evidently the second part of the assertion is fulfilled by this basis, too. □

An important application of this result is the following Theorem of Witt concerning the extension of isometries, whose proof has been taken over from [BR, Theorem 4.1] and [BMRRR2, Theorem 2.1]. Here $\pi(\mathbf{H})$ denotes the number of positive eigenvalues of the selfadjoint matrix $\mathbf{H}$.

Theorem 2.9 (Witt, extension of isometries). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $[.,.]_1$, $[.,.]_2$ be two indefinite scalar products in $\mathbb{F}^n$ with the underlying nonsingular selfadjoint matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{F}^{n\times n}$ for which $\pi(\mathbf{H}_1) = \pi(\mathbf{H}_2)$. If X_1 and X_2 are subspaces of $\mathbb{F}^n$ and $\mathbf{U}_0 : X_1 \to X_2$ is a nonsingular transformation such that*

$$[\mathbf{U}_0\mathbf{x}, \mathbf{U}_0\mathbf{y}]_2 = [\mathbf{x}, \mathbf{y}]_1 \;\; \textit{for all} \;\; \mathbf{x}, \mathbf{y} \in X_1,$$

then there exists a nonsingular transformation $\mathbf{U} : \mathbb{F}^n \to \mathbb{F}^n$ such that

$$[\mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{y}]_2 = [\mathbf{x}, \mathbf{y}]_1 \;\; \textit{for all} \;\; \mathbf{x}, \mathbf{y} \in \mathbb{F}^n \;\; \textit{and} \;\; \mathbf{U}\mathbf{x} = \mathbf{U}_0\mathbf{x} \;\; \textit{for all} \;\; \mathbf{x} \in X_1.$$

Proof. Let $\dim X_1 = m$ and let $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be an orthonormalised (according to Theorem 2.6) basis of X_1 with

$$[\mathbf{e}_k, \mathbf{e}_l] = \varepsilon_k \delta_{kl}, \;\; \varepsilon_k = \begin{cases} +1, & \text{for } 1 \le k \le p \\ -1, & \text{for } p+1 \le k \le p+q \\ 0, & \text{for } p+q+1 \le k \le p+q+r \end{cases}$$

and $p+q+r=m$. Then $\{\mathbf{f}_1,\ldots,\mathbf{f}_m\}$ with $\mathbf{f}_k=\mathbf{U}_0\,\mathbf{e}_k$ for $1\le k\le m$ is an orthonormalised basis of X_2, and both bases can be extended to bases of $\mathbb{F}^n$ according to Theorem 2.8. For the matrices $\mathbf{R}_1=[\mathbf{e}_1\ldots\mathbf{e}_n]$ and $\mathbf{R}_2=[\mathbf{f}_1\ldots\mathbf{f}_n]$ consisting of the extended basis vectors we have

$$\mathbf{R}_1^*\mathbf{H}_1\mathbf{R}_1=\mathbf{R}_2^*\mathbf{H}_2\mathbf{R}_2=\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}_q\end{bmatrix}\oplus\begin{bmatrix}\mathbf{0} & \mathbf{I}_r\\ \mathbf{I}_r & \mathbf{0}\end{bmatrix}\oplus\begin{bmatrix}\mathbf{I}_s & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}_t\end{bmatrix}$$

with $r+s+t=n-m$. This results from the fact that the number of positive and the number of negative vectors, s and t respectively, must be identical for both bases, which is implied by the assumption of the equal signatures of the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. Thus the transformation defined by

$$\mathbf{U}\mathbf{R}_1=\mathbf{R}_2\ \text{ or }\ \mathbf{U}=\mathbf{R}_2\mathbf{R}_1^{-1}$$

fulfills the assertion of the theorem. □

2.4 The HQR decomposition

The H-orthogonalisation methods described in Theorems 2.6 and 2.7 allow to derive two matrix factorisations which may be seen as generalisations of the QR factorisation with column pivoting [GVL, Section 5.4]. For this purpose it is useful to make the following observations:

1st observation: The proof of Theorem 2.6 suggests to determine the pivot vector $\mathbf{y}$ for the H-orthonormalisation step by selecting k and l such that $|[\mathbf{x}_k,\mathbf{x}_l]|$ is maximised and then to use

$$\mathbf{y}=\begin{cases}\mathbf{x}_k, & \text{if } k=l\\ \tilde{\mathbf{x}}_k, & \text{if } k\ne l \text{ and } |[\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_k]|\ge|[\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_l]|\\ \tilde{\mathbf{x}}_l, & \text{if } k\ne l \text{ and } |[\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_k]|<|[\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_l]|\end{cases}$$

where

$$[\tilde{\mathbf{x}}_k\ \ \tilde{\mathbf{x}}_l]=\frac{1}{\sqrt{2}}[\mathbf{x}_k\ \ \mathbf{x}_l]\begin{bmatrix}1 & 1\\ \varphi & -\varphi\end{bmatrix}\ \text{ with }\ \varphi=\frac{[\mathbf{x}_k,\mathbf{x}_l]}{|[\mathbf{x}_k,\mathbf{x}_l]|}.$$

With this strategy it is ensured that $[\mathbf{y},\mathbf{y}]\ne 0$, but it is not ensured that

$$|[\mathbf{y},\mathbf{y}]|=\max_{1\le i<j\le m}\big\{|[\mathbf{x}_i,\mathbf{x}_i]|,|[\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_i]|,|[\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_j]|\big\}\tag{2.1}$$

which can be seen on the following example.

Example 2.10. Let $\mathbf{H}=\operatorname{diag}(1,-1)$ and $\mathbf{x}_1=(2,1)^T$, $\mathbf{x}_2=(1,0)^T$. Then

$$[\mathbf{x}_1,\mathbf{x}_1]=3,\quad [\mathbf{x}_1,\mathbf{x}_2]=2,\quad [\mathbf{x}_2,\mathbf{x}_2]=1.$$

On the other hand $\tilde{\mathbf{x}}_1=(\mathbf{x}_1+\mathbf{x}_2)/\sqrt{2}$ and $\tilde{\mathbf{x}}_2=(\mathbf{x}_1-\mathbf{x}_2)/\sqrt{2}$ fulfil

$$[\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_1]=4,\quad [\tilde{\mathbf{x}}_2,\tilde{\mathbf{x}}_2]=0,$$

so that $|[\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_1]|>|[\mathbf{x}_1,\mathbf{x}_1]|$ although $|[\mathbf{x}_1,\mathbf{x}_1]|$ maximises $|[\mathbf{x}_k,\mathbf{x}_l]|$. ◊

The method can thus be stabilised by determining $\mathbf{y}$ such that (2.1) holds which is numerically uncritical because the transformed vectors satisfy

$$\|\tilde{\mathbf{x}}_k\|^2 + \|\tilde{\mathbf{x}}_l\|^2 = \|\mathbf{x}_k\|^2 + \|\mathbf{x}_l\|^2.$$

2nd observation: The normalisation step in the proof of Theorem 2.7 can be modified with a factor α by setting

$$\begin{aligned} \mathbf{u}_1 &= \frac{\alpha}{\sqrt{\lambda_1}}\mathbf{x}_1, & \mathbf{v}_1 &= \frac{\varepsilon_1}{\alpha\sqrt{\lambda_1}}\mathbf{y}_1 & &\text{if } \mathbb{F} = \mathbb{R} \text{ or} \\ \mathbf{u}_1 &= \frac{\alpha}{\omega_1}\mathbf{x}_1, & \mathbf{v}_1 &= \frac{1}{\overline{\alpha\omega}_1}\mathbf{y}_1 & &\text{if } \mathbb{F} = \mathbb{C}. \end{aligned}$$

If the particular choice $\alpha = \sqrt{\|\mathbf{y}_1\|/\|\mathbf{x}_1\|}$ is made, then $\|\mathbf{u}_1\| = \|\mathbf{v}_1\|$ is ensured which has a stabilising effect. This is demonstrated with the following example in which $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}^*\mathbf{A})}$ denotes the Frobenius norm and $\mathrm{cond}_F(\mathbf{A}) = \|\mathbf{A}\|_F\,\|\mathbf{A}^{-1}\|_F$ denotes the condition number of a matrix $\mathbf{A}$.

Example 2.11. Let $\mathbf{H} = \mathrm{diag}(1,-1)$ and $\mathbf{x} = (x,\, x)^T$, $\mathbf{y} = (y,\, -y)^T$ with $x, y \in \mathbb{C}\backslash\{0\}$. Then $X = \mathrm{span}\{\mathbf{x}\}$ and $Y = \mathrm{span}\{\mathbf{y}\}$ are neutral subspaces of equal dimension and $X \cap Y = \{\mathbf{0}\}$, so that Theorem 2.7 can be applied. Let $\lambda = [\mathbf{x}, \mathbf{y}] = 2x\overline{y}$ and let ω be one of the two square roots of λ. Then the columns $[\mathbf{x}'\ \mathbf{y}']$ of the matrix

$$\mathbf{X}_1 = \begin{bmatrix} \dfrac{x}{\omega} & \dfrac{y}{\overline{\omega}} \\[2mm] \dfrac{x}{\omega} & -\dfrac{y}{\overline{\omega}} \end{bmatrix} \quad \text{with} \quad \mathbf{X}_1^{-1} = \frac{|\omega|^2}{2xy}\begin{bmatrix} \dfrac{y}{\overline{\omega}} & \dfrac{y}{\overline{\omega}} \\[2mm] \dfrac{x}{\omega} & -\dfrac{x}{\omega} \end{bmatrix}$$

are the vectors obtained by orthonormalisation **without** modification and

$$\mathrm{cond}_F(\mathbf{X}_1) = \frac{|x|^2 + |y|^2}{|x||y|} \quad \text{because}$$
$$\|\mathbf{X}_1\|_F^2 = \frac{2(|x|^2+|y|^2)}{|\omega|^2}, \quad \|\mathbf{X}_1^{-1}\|_F^2 = \frac{|\omega|^2(|x|^2+|y|^2)}{2|x|^2|y|^2}.$$

Now, let $\alpha = \sqrt{\|\mathbf{y}\|/\|\mathbf{x}\|} = \sqrt{|y|/|x|}$. Then the columns $[\mathbf{x}''\ \mathbf{y}'']$ of the matrix

$$\mathbf{X}_2 = \begin{bmatrix} \dfrac{\alpha x}{\omega} & \dfrac{y}{\alpha\overline{\omega}} \\[2mm] \dfrac{\alpha x}{\omega} & -\dfrac{y}{\alpha\overline{\omega}} \end{bmatrix} \quad \text{with} \quad \mathbf{X}_2^{-1} = \frac{|\omega|^2}{2xy}\begin{bmatrix} \dfrac{y}{\alpha\overline{\omega}} & \dfrac{y}{\alpha\overline{\omega}} \\[2mm] \dfrac{\alpha x}{\omega} & -\dfrac{\alpha x}{\omega} \end{bmatrix}$$

are the vectors obtained by orthonormalisation **with** modification and

$$\mathrm{cond}_F(\mathbf{X}_2) = \frac{\alpha^4|x|^2 + |y|^2}{\alpha^2|x||y|} = 2 \quad \text{because}$$
$$\|\mathbf{X}_2\|_F^2 = \frac{2(\alpha^4|x|^2+|y|^2)}{\alpha^2|\omega|^2}, \quad \|\mathbf{X}_1^{-1}\|_F^2 = \frac{|\omega|^2(\alpha^4|x|^2+|y|^2)}{2\alpha^2|x|^2|y|^2}.$$

But for arbitrary real numbers a, b with $ab > 0$ it is true that $0 \le (a-b)^2 = a^2 - 2ab + b^2$ or $2 \le (a^2+b^2)/ab$, so that in particular $\mathrm{cond}_F(\mathbf{X}_1) \ge 2$ and thus

$$\mathrm{cond}_F(\mathbf{X}_1) \ge \mathrm{cond}_F(\mathbf{X}_2).$$

Therefore, the matrix $\mathbf{X}_2$ obtained by modification is at least as well conditioned as $\mathbf{X}_1$; but in the case $|x| \ne |y|$ it is always better conditioned. $\Diamond$

Now all preparations are complete and the matrix factorisations can be described. We begin with the factorisation corresponding to Theorem 2.6 which will be called the HQR decomposition. Let $\mathbf{H} \in \mathbb{F}^{n\times n}$ be nonsingular and self-adjoint and let $\mathbf{A} \in \mathbb{F}^{n\times m}$. Then the HQR decomposition of $\mathbf{A}$ is given by

$$\mathbf{AP} = \mathbf{QR} \tag{2.2a}$$

where $\mathbf{P} \in \mathbb{F}^{m\times m}$ is unitary or orthogonal, $\mathbf{Q} \in \mathbb{F}^{n\times m}$ satisfies

$$\mathbf{Q}^*\mathbf{HQ} = \mathbf{D} = \mathbf{D}_1 \oplus \mathbf{0} \text{ with } \mathbf{D}_1 = \operatorname{diag}_p(\pm 1) \tag{2.2b}$$

and $\mathbf{R} \in \mathbb{F}^{m\times m}$ has the form

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{2.2c}$$

where $\mathbf{R}_{11} \in \mathbb{F}^{p\times p}$ is an upper triangular block such that $\mathbf{D}_1\mathbf{R}_{11}$ has positive real diagonal elements. Note that the columns of $\mathbf{A}$ and $\mathbf{Q}$ correspond to the vectors $\mathbf{x}_i$ and $\mathbf{u}_i$ used in the proof of Theorem 2.6.

The decomposition is obtained by stepwise transformation of $\mathbf{Q}_0 = \mathbf{A}$ into $\mathbf{Q}_p = \mathbf{Q}$. Suppose that after k steps

$$\mathbf{Q}_k = \mathbf{AP}_k\mathbf{R}_k^{-1}, \quad \mathbf{Q}_k^*\mathbf{HQ}_k = \mathbf{D}_k \oplus \mathbf{C}_k, \quad \mathbf{R}_k = \begin{bmatrix} \mathbf{R}_{11}^{(k)} & \mathbf{R}_{12}^{(k)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where $\mathbf{D}_k = \operatorname{diag}_k(\pm 1)$ and $\mathbf{R}_{11}^{(k)} \in \mathbb{F}^{k\times k}$. Then the symmetric or Hermitian matrix $\mathbf{C}_k = [c_{ij}]$ contains the scalar products of the not yet H-orthonormalised columns of $\mathbf{Q}_k$. To determine the pivot vector let

$$\chi_{ij} = \begin{cases} c_{ii}, & \text{if } i = j \\ (c_{ii} + c_{jj})/2 + |c_{ij}|, & \text{if } i < j \\ (c_{ii} + c_{jj})/2 - |c_{ji}|, & \text{if } i > j \end{cases}$$

and let μ, ν be indices such that $|\chi_{\mu\nu}| = \max|\chi_{ij}|$. If $\chi_{\mu\nu} = 0$, the transformation is complete. Otherwise, let

$$\mathbf{U}_k = \mathbf{I} \oplus \mathbf{U}_{22}, \quad \mathbf{U}_{22} = \begin{cases} \mathbf{\Pi}_{1\mu}, & \text{if } \mu = \nu \\ \mathbf{\Omega}_{\mu\nu}(c_{\mu\nu}/|c_{\mu\nu}|)\,\mathbf{\Pi}_{1\mu}, & \text{if } \mu < \nu \\ \mathbf{\Omega}_{\nu\mu}(c_{\nu\mu}/|c_{\nu\mu}|)\,\mathbf{\Pi}_{1\mu}, & \text{if } \mu > \nu \end{cases}$$

where $\mathbf{\Omega}_{\mu\nu}(\varphi) = [\omega_{ij}]$ and $\mathbf{\Pi}_{\mu\nu} = [\pi_{ij}]$ are defined by

$$\begin{Bmatrix} \omega_{\mu\mu} = \varphi/\sqrt{2}, & \omega_{\mu\nu} = \varphi/\sqrt{2}, \\ \omega_{\nu\mu} = 1/\sqrt{2}, & \omega_{\nu\nu} = -1/\sqrt{2}, \\ \omega_{ij} = \delta_{ij} & \text{otherwise} \end{Bmatrix} \text{ and } \begin{Bmatrix} \pi_{\mu\mu} = 0, & \pi_{\mu\nu} = 1, \\ \pi_{\nu\mu} = 1, & \pi_{\nu\nu} = 0, \\ \pi_{ij} = \delta_{ij} & \text{otherwise} \end{Bmatrix}.$$

Then $\mathbf{U}_k$ is orthogonal or unitary and the transformation

$$\mathbf{Q}'_k = \mathbf{AP}_k\mathbf{R}_k^{-1}\mathbf{U}_k = \mathbf{A}(\mathbf{P}_k\mathbf{U}_k)(\mathbf{U}_k^*\mathbf{R}_k\mathbf{U}_k)^{-1} = \mathbf{AP}_{k+1}(\mathbf{R}'_k)^{-1}$$

can be made. Now we have

$$(\mathbf{Q}'_k)^*\mathbf{H}(\mathbf{Q}'_k) = \mathbf{D}_k \oplus \mathbf{C}'_k, \quad \mathbf{C}'_k = \begin{bmatrix} c'_{11} & \mathbf{c}'_{12} \\ \mathbf{c}'_{21} & \mathbf{C}'_{22} \end{bmatrix}, \quad \mathbf{R}'_k = \begin{bmatrix} \mathbf{R}_{11}^{(k)} & \mathbf{R}_{12}^{(k)}\mathbf{U}_{22} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where $c'_{11} = \chi_{\mu\nu}$, so that the $k+1$-th column of $\mathbf{Q}'_k$ actually contains the selected pivot vector.

It remains to H-orthonormalise the columns of $\mathbf{Q}'_k$. For this purpose let $\lambda_1 = \sqrt{|c'_{11}|}$ and $\varepsilon_1 = \mathrm{sign}(c'_{11})$. Then

$$\tilde{\mathbf{C}}_k = \mathbf{\Lambda}^{-*}\mathbf{C}'_k\mathbf{\Lambda}^{-1} = \begin{bmatrix} \varepsilon_1 & \tilde{\mathbf{c}}_{12} \\ \tilde{\mathbf{c}}_{21} & \mathbf{C}'_{22} \end{bmatrix}, \; \tilde{\mathbf{c}}_{12} = \frac{\mathbf{c}'_{12}}{\lambda_1} \;\text{ for }\; \mathbf{\Lambda} = \lambda_1 \oplus \mathbf{I}$$

and

$$\mathbf{C}''_k = \mathbf{\Gamma}^{-*}\tilde{\mathbf{C}}_k\mathbf{\Gamma}^{-1} = \begin{bmatrix} \varepsilon_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}''_{22} \end{bmatrix}, \; \mathbf{C}''_{22} = \mathbf{C}'_{22} - \frac{\mathbf{c}'_{21}\mathbf{c}'_{12}}{c'_{11}} \;\text{ for }\; \mathbf{\Gamma} = \begin{bmatrix} 1 & \varepsilon_1\tilde{\mathbf{c}}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Hence, if we define

$$\mathbf{W}_k = \mathbf{I} \oplus \mathbf{W}_{22}, \quad \mathbf{W}_{22} = \mathbf{\Gamma\Lambda} = \begin{bmatrix} \lambda_1 & \varepsilon_1\tilde{\mathbf{c}}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

and use the transformation

$$\mathbf{Q}_{k+1} = \mathbf{A}\mathbf{P}_{k+1}(\mathbf{R}'_k)^{-1}\mathbf{W}_k^{-1} = \mathbf{A}\mathbf{P}_{k+1}(\mathbf{W}_k\mathbf{R}'_k)^{-1} = \mathbf{A}\mathbf{P}_{k+1}\mathbf{R}_{k+1}^{-1},$$

we obtain

$$\mathbf{Q}_{k+1}^*\mathbf{H}\mathbf{Q}_{k+1} = \mathbf{D}_k \oplus \mathbf{C}''_k = \mathbf{D}_{k+1} \oplus \mathbf{C}_{k+1},$$

$$\mathbf{R}_{k+1} = \begin{bmatrix} \mathbf{R}_{11}^{(k)} & \mathbf{R}_{12}^{(k)}\mathbf{U}_{22} \\ \mathbf{0} & \mathbf{W}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{11}^{(k+1)} & \mathbf{R}_{12}^{(k+1)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Thus, the transformation itself is described. However, we must also consider that the computation of one scalar product

$$[\mathbf{x}, \mathbf{y}] = \sum_{\alpha,\beta=1}^{n} h_{\alpha\beta}\, x_\beta\, \overline{y}_\alpha$$

already requires $n^2 - 1$ additions and $2n^2$ multiplications. It would therefore be too expensive to recompute the matrix $\mathbf{C}_k$ in each transformation step. This can be avoided in the following way:

Suppose that the matrix $\tilde{\mathbf{R}}$ is initialised with $\mathbf{C}_0 = \mathbf{A}^*\mathbf{H}\mathbf{A}$ and that after k transformation steps

$$\tilde{\mathbf{R}}_k = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^{(k)} & \tilde{\mathbf{R}}_{12}^{(k)} \\ \mathbf{0} & \mathbf{C}_k \end{bmatrix}.$$

Then using

$$\tilde{\mathbf{R}}'_k = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^{(k)} & \tilde{\mathbf{R}}_{12}^{(k)}\mathbf{U}_{22} \\ \mathbf{0} & \mathbf{U}_{22}^*\mathbf{C}_k\mathbf{U}_{22} \end{bmatrix}, \quad \mathbf{U}_{22}^*\mathbf{C}_k\mathbf{U}_{22} = \mathbf{C}'_k$$

and

$$\tilde{\mathbf{R}}_{k+1} = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^{(k)} & \tilde{\mathbf{R}}_{12}^{(k)}\mathbf{U}_{22} \\ \mathbf{0} & \mathbf{W}_{22}^{-*}\mathbf{C}'_k \end{bmatrix}, \quad \mathbf{W}_{22}^{-*}\mathbf{C}'_k = \begin{bmatrix} \varepsilon_1\lambda_1 & \tilde{\mathbf{c}}_{12} \\ \mathbf{0} & \mathbf{C}''_{22} \end{bmatrix}$$

we find that

$$\tilde{\mathbf{R}}_{k+1} = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^{(k+1)} & \tilde{\mathbf{R}}_{12}^{(k+1)} \\ \mathbf{0} & \mathbf{C}_{k+1} \end{bmatrix}, \quad \mathbf{C}_{k+1} = \mathbf{C}''_{22}$$

on the one hand contains the scalar products required for the next step. On the other hand the comparison with the matrix $\mathbf{R}_{k+1}$ shows that

$$\tilde{\mathbf{R}}_{11}^{(k+1)} = \mathbf{D}_{k+1}\mathbf{R}_{11}^{(k+1)} \text{ and } \tilde{\mathbf{R}}_{12}^{(k+1)} = \mathbf{D}_{k+1}\mathbf{R}_{12}^{(k+1)}.$$

In other words, when the transformation terminates in the p-th step and the block $\mathbf{C}_p = \mathbf{0}$ of $\tilde{\mathbf{R}}_p$ is replaced by $\mathbf{I}$ we have computed

$$\tilde{\mathbf{R}}_p = (\mathbf{D}_p \oplus \mathbf{I})\,\mathbf{R}_p.$$

This has the further advantage that the signs of the first p diagonal elements of $\tilde{\mathbf{R}}_p$ are just the diagonal elements of $\mathbf{D}_p$.

Finally, it must also be considered that the theoretical termination criterion $\chi_{\mu\nu} = 0$ is too hard and must be replaced by

$$\chi_{\mu\nu} \le \varepsilon$$

where ε is some user supplied constant. An appropriate choice might be $\varepsilon = \varepsilon_{mach}\|\mathbf{H}\|_F\|\mathbf{A}\|_F^2 \ge \varepsilon_{mach}\|\mathbf{A}^*\mathbf{H}\mathbf{A}\|_F$ where ε_{mach} is the machine accuracy.

Combining all of the above we obtain the following algorithm which is described using the "colon" notation

$$\begin{gathered}\mathbf{A} = [a_{ij}] \in \mathbb{F}^{m\times n},\\ \mathbf{A}(p:q, r:s) = \begin{bmatrix} a_{pr} & \cdots & a_{ps} \\ \vdots & & \vdots \\ a_{qr} & \cdots & a_{qs} \end{bmatrix} \in \mathbb{F}^{(q-p+1)\times(s-r+1)}, \qquad (2.3)\\ \mathbf{A}(p, r:s) = \mathbf{A}(p:p, r:s), \quad \mathbf{A}(p,:) = \mathbf{A}(p, 1:n),\\ \mathbf{A}(p:q, r) = \mathbf{A}(p:q, r:r), \quad \mathbf{A}(:, r) = \mathbf{A}(1:m, r)\end{gathered}$$

introduced in [GVL, Sections 1.1.8, 1.2.5].

Algorithm 2.12 (HQR decomposition). Given the matrices $\mathbf{A} \in \mathbb{F}^{n\times m}$, $\mathbf{H} \in \mathbb{F}^{n\times n}$ and a tolerance parameter $\varepsilon > 0$, the following algorithm computes the HQR decomposition (2.2). The matrix $\mathbf{A}$ is overwritten with $\mathbf{Q}$. The matrices $\mathbf{P}$ and $\mathbf{R}$ are stored in separate arrays. To keep the algorithm short the pivot transformation is given in matrix notation.

$\mathbf{R} = \mathbf{A}^*\mathbf{H}\mathbf{A}$, $\mathbf{P} = \mathbf{I}_m$
for $k = 1, \dots, m$ do
  $\alpha = 0$
  for $i = k, \dots, m$ do
    if $|\mathbf{R}(i,i)| > |\alpha|$ then
      $\alpha = \mathbf{R}(i,i)$, $\mu = i$, $\nu = i$
    end if
    for $j = i+1, \dots, m$ do
      $\beta = \big(\mathbf{R}(i,i) + \mathbf{R}(j,j)\big)/2 + \operatorname{sign}\big(\mathbf{R}(i,i) + \mathbf{R}(j,j)\big)\,|\mathbf{R}(i,j)|$
      if $|\beta| > |\alpha|$ then
        $\alpha = \beta$, $\mu = i$, $\nu = j$
      end if
    end for

```
    end for
    if |α| ≤ ε then
        p = k − 1
        R(k : m, k : m) = I_{m−p}
        return
    end if
    if μ ≠ ν then
        ρ = R(μ, ν)/|R(μ, ν)|
        A = AΩ_{μν}(ρ), P = PΩ_{μν}(ρ), R = Ω_{μν}(ρ)*RΩ_{μν}(ρ)
        if α < 0 then
            μ = ν
        end if
    end if
    if μ ≠ k then
        A = AΠ_{μk}, P = PΠ_{μk}, R = Π*_{μk}RΠ_{μk}
    end if
    σ = sign(α)
    α = √(σα)
    R(k, k) = σα
    ρ = 1/α
    A(:, k) = ρ A(:, k)
    R(k, k + 1 : m) = ρ R(k, k + 1 : m)
    for j = k + 1, . . . , m do
        ρ = σR(k, j)
        A(:, j) = A(:, j) − ρ A(:, k)
        R(j, k + 1 : m) = R(j, k + 1 : m) − ρ̄ R(k, k + 1 : m)
        R(j, k) = 0
    end for
end for
p = m
```

In order to understand how the matrix $\mathbf{Q}$ has to be interpreted let

$$\operatorname{im}(\mathbf{A}) = A_1 \oplus A_0 \;\text{ with }\; \dim A_1 = p \;\text{ and }\; \dim A_0 = q$$

be a subspace decomposition (according to Theorem 2.3) such that A_1 is non-degenerate and A_0 is neutral. Moreover, let

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_0 \end{bmatrix}$$

be a partitioning where $\mathbf{Q}_1 \in \mathbb{F}^{n\times p}$ and $\mathbf{Q}_0 \in \mathbb{F}^{n\times(m-p)}$. Then

$$\operatorname{im}(\mathbf{Q}_1) = A_1 \;\text{ and }\; \operatorname{im}(\mathbf{Q}_0) = A_0,$$

so that $\mathbf{Q}_1$ always has full rank. However, the rank of $\mathbf{Q}_0$ depends on $\operatorname{rank}(\mathbf{A}) = p + q$. Here the following cases can occur:

(1) $q = 0$, $p + q = m$: $\mathbf{Q}_0$ does not exist,

(2) $q = 0$, $p + q < m$: $\mathbf{Q}_0 = \mathbf{0}$,

(3) $q > 0$, $p + q = m$: $\mathbf{Q}_0$ has full rank,

(4) $q > 0$, $p + q < m$: $\mathbf{Q}_0$ is rank defective.

Whereas the columns of $\mathbf{Q}_0$ in case (3) form a basis of the neutral space A_0, in case (4) they only form a spanning set. If actually a basis of A_0 is required in case (4), it can be obtained via a QR decomposition with column pivoting. A QR decomposition (without column pivoting) may also be used to orthonormalise the columns of $\mathbf{Q}_0$ in case (3). Clearly, the same results can also be obtained via a further HQR decomposition by setting $\mathbf{H} = \mathbf{I}$.

We come to the matrix factorisation corresponding to Theorem 2.7 which will be called the HQR-2 decomposition. Let $\mathbf{H} \in \mathbb{F}^{n\times n}$ be nonsingular and selfadjoint and let $\mathbf{A}, \mathbf{B} \in \mathbb{F}^{n\times m}$. Then the HQR-2 decomposition of $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\mathbf{A}\mathbf{P}_A = \mathbf{Q}_A\mathbf{R}_A, \quad \mathbf{B}\mathbf{P}_B = \mathbf{Q}_B\mathbf{R}_B \tag{2.4a}$$

where $\mathbf{P}_A, \mathbf{P}_B \in \mathbb{F}^{m\times m}$ are permutations (one of them possibly signed if $\mathbb{F} = \mathbb{R}$), $\mathbf{Q}_A, \mathbf{Q}_B \in \mathbb{F}^{n\times m}$ satisfy

$$\mathbf{Q}_A^*\mathbf{H}\mathbf{Q}_B = \mathbf{D} = \mathbf{I}_p \oplus \mathbf{0} \tag{2.4b}$$

and $\mathbf{R}_A, \mathbf{R}_B \in \mathbb{F}^{m\times m}$ have the form

$$\mathbf{R}_A = \begin{bmatrix} \mathbf{R}_{A,11} & \mathbf{R}_{A,12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{R}_B = \begin{bmatrix} \mathbf{R}_{B,11} & \mathbf{R}_{B,12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{2.4c}$$

where $\mathbf{R}_{A,11}, \mathbf{R}_{B,11} \in \mathbb{F}^{p\times p}$ are upper triangular blocks with real or complex diagonal elements. Note that the columns of $\mathbf{A}, \mathbf{B}$ and $\mathbf{Q}_A, \mathbf{Q}_B$ correspond to the vectors $\mathbf{x}_i, \mathbf{y}_j$ and $\mathbf{u}_i, \mathbf{v}_j$ used in the proof of Theorem 2.7.

In the intended application of the HQR-2 decomposition the matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy

$$\mathbf{A}^*\mathbf{H}\mathbf{A} = \mathbf{B}^*\mathbf{H}\mathbf{B} = \mathbf{0} \text{ and } \operatorname{rank}(\mathbf{A}^*\mathbf{H}\mathbf{B}) = m.$$

This means that the columns of $\mathbf{A}$ and $\mathbf{B}$ form bases of the neutral subspaces $\operatorname{im}(\mathbf{A})$ and $\operatorname{im}(\mathbf{B})$ for which $C = \operatorname{im}(\mathbf{A}) \oplus \operatorname{im}(\mathbf{B})$ is non-degenerate. In this case the columns of $\mathbf{Q}_A$ and $\mathbf{Q}_B$ form a bi-H-orthogonal basis of C, but we do not discuss the further situations which can occur.

The decomposition is computed analogously to the HQR decomposition, so that not all the details need to be described again. For a better understanding of the following algorithm it should only be mentioned that the matrix $\mathbf{R}$, initialised with $\mathbf{C}_0 = \mathbf{A}^*\mathbf{H}\mathbf{B}$, is stepwise transformed such that

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{R}_{11}^{(k)} & \mathbf{R}_{B,12}^{(k)} \\ \mathbf{R}_{A,12}^{(k)*} & \mathbf{C}_k \end{bmatrix} \to \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{B,12} \\ \mathbf{R}_{A,12}^* & \mathbf{I} \end{bmatrix}.$$

Moreover, the scaling factors τ_k explained with Example 2.11 are computed in each step. Let $\mathbf{T} = \operatorname{diag}(\tau_1, \ldots, \tau_p) \oplus \mathbf{I}$ be a diagonal matrix containing this factors and let $\mathbf{R}_L, \mathbf{R}_D, \mathbf{R}_U$ contain the lower triangle, the diagonal and the upper triangle of the final matrix $\mathbf{R}$. Then at termination

$$\mathbf{R}_A = \mathbf{T}^{-1}\mathbf{R}_D^* + \mathbf{R}_L^* \text{ and } \mathbf{R}_B = \mathbf{T}\mathbf{R}_D + \mathbf{R}_U.$$

In the case $\mathbb{F} = \mathbb{R}$ it must additionally be ensured that the scalar product of the pivot vectors is positive for which a reflection with the sign matrix $\mathbf{\Sigma}_\mu = \operatorname{diag}(\sigma_i)$ where

$$\sigma_i = \begin{cases} -1, & \text{if } i = \mu \\ +1, & \text{otherwise} \end{cases}$$

is applied, if necessary.

Algorithm 2.13 (HQR-2 decomposition). Given the matrices $\mathbf{A}, \mathbf{B} \in \mathbb{F}^{n\times m}$, $\mathbf{H} \in \mathbb{F}^{n\times n}$ and a tolerance parameter $\varepsilon > 0$, the following algorithm computes the HQR-2 decomposition (2.4). The matrices $\mathbf{A}$ and $\mathbf{B}$ are overwritten with $\mathbf{Q}_A$ and $\mathbf{Q}_B$. The matrix $\mathbf{R}$ and the scaling factors τ are stored in separate arrays. The permutations $\mathbf{P}_A$ and $\mathbf{P}_B$ are stored in separate integer arrays[5]. To keep the algorithm short the pivot transformation is given in matrix notation.

$\mathbf{R} = \mathbf{A}^*\mathbf{H}\mathbf{B}\,,\ \mathbf{P}_A = \mathbf{I}_m\,,\ \mathbf{P}_B = \mathbf{I}_m$
`for` $k = 1, \ldots, m$ `do`
  $\alpha = 0$
  `for` $i = k, \ldots, m$ `do`
    `for` $j = i, \ldots, m$ `do`
      `if` $|\mathbf{R}(i,j)| > |\alpha|$ `then`
        $\alpha = \mathbf{R}(i,j)\,,\ \mu = i\,,\ \nu = j$
      `end if`
    `end for`
  `end for`
  `if` $|\alpha| \le \varepsilon$ `then`
    $p = k - 1$
    $\mathbf{R}(k : m, k : m) = \mathbf{I}_{m-p}$
    `return`
  `end if`
  `if` $\mu \neq k$ `then`
    $\mathbf{A} = \mathbf{A}\mathbf{\Pi}_{\mu k}\,,\ \mathbf{P}_A = \mathbf{P}_A\mathbf{\Pi}_{\mu k}\,,\ \mathbf{R} = \mathbf{\Pi}_{\mu k}^*\mathbf{R}$
  `end if`
  `if` $\nu \neq k$ `then`
    $\mathbf{B} = \mathbf{B}\mathbf{\Pi}_{\nu k}\,,\ \mathbf{P}_B = \mathbf{P}_B\mathbf{\Pi}_{\nu k}\,,\ \mathbf{R} = \mathbf{R}\mathbf{\Pi}_{\nu k}$
  `end if`
  `if` $\mathbb{F} = \mathbb{R}$ `and` $\alpha < 0$ `then`
    $\mathbf{B} = \mathbf{B}\mathbf{\Sigma}_k\,,\ \mathbf{P}_B = \mathbf{P}_B\mathbf{\Sigma}_k\,,\ \mathbf{R} = \mathbf{R}\mathbf{\Sigma}_k\,,\ \alpha = -\alpha$
  `end if`
  $\tau(k) = \sqrt{\|\mathbf{B}(:,k)\|/\|\mathbf{A}(:,k)\|}$
  $\alpha = \sqrt{\alpha}$
  $\mathbf{R}(k,k) = \alpha$
  $\rho = \tau(k)/\alpha$
  $\mathbf{A}(:,k) = \overline{\rho}\,\mathbf{A}(:,k)$
  $\mathbf{R}(k, k+1 : m) = \rho\,\mathbf{R}(k, k+1 : m)$
  $\rho = 1/\alpha/\tau(k)$
  $\mathbf{B}(:,k) = \rho\,\mathbf{B}(:,k)$
  $\mathbf{R}(k+1 : m, k) = \rho\,\mathbf{R}(k+1 : m, k)$
  `for` $j = k+1, \ldots, m$ `do`
    $\rho = \mathbf{R}(j,k)$
    $\mathbf{A}(:,j) = \mathbf{A}(:,j) - \overline{\rho}\,\mathbf{A}(:,k)$
    $\mathbf{R}(j, k+1 : m) = \mathbf{R}(j, k+1 : m) - \rho\,\mathbf{R}(k, k+1 : m)$
    $\rho = \mathbf{R}(k,j)$

[5]The reflections with $\mathbf{\Sigma}_k$ can also be stored in the array $(\pi_1, \ldots, \pi_m)$ representing $\mathbf{P}_B$ by negating the k-th element. The corresponding matrix contains the columns $\operatorname{sign}(\pi_\mu)\,\mathbf{e}_{|\pi_\mu|}$, $1 \le \mu \le m$, where $\{\mathbf{e}_\mu\}$ is the canonical basis of $\mathbb{F}^m$.

$\mathbf{B}(:,j) = \mathbf{B}(:,j) - \rho\,\mathbf{B}(:,k)$
end for
end for
$p = m$

There are several further decompositions which generalise the QR factorisation in the presence of an indefinite scalar product. In particular, the HR decomposition (for example see [BU])

$$\mathbf{A} = \mathbf{HR}, \quad \mathbf{H}^*\mathbf{D}_1\mathbf{H} = \mathbf{D}_2$$

where $\mathbf{A}, \mathbf{H}, \mathbf{R} \in \mathbb{F}^{n\times n}$, $\mathbf{R}$ is upper triangular and $\mathbf{D}_1, \mathbf{D}_2 = \mathrm{diag}_n(\pm 1)$, is closely related to the HQR decomposition. In order to avoid confusion let us rename the $\mathbf{H}$ in this decomposition to $\mathbf{Q}$. Then we obtain

$$\mathbf{A} = \mathbf{QR}, \quad \mathbf{Q}^*\mathbf{D}_1\mathbf{Q} = \mathbf{D}_2$$

which shows that the HR decomposition is just a HQR decomposition for square matrices in the particular case $\mathbf{H} = \mathbf{D}_1$ and $\mathbf{P} = \mathbf{I}$. Of course the HR decomposition breaks down when $\mathrm{im}(\mathbf{A})$ is degenerate. However, the particular choice of the metric allows to use hyperbolic Householder or Givens rotations to compute the HR decomposition. In contrast to this our approach is based on a generalised modified Gram-Schmidt method and might therefore not always be perfectly accurate. Nevertheless, we will see in the next chapter that the HQR decomposition mostly produces very well results.

We end this section with a first application of the HQR decomposition. The following method shows how an H-orthogonal basis of a subspace can be extended to a complete H-orthogonal basis of $\mathbb{F}^n$. This application of Theorem 2.8 will help to compute an indefinite generalisation of the singular value decomposition numerically.

Method 2.14 (Extension of bases). Let $m < n$ and let $\mathbf{X} \in \mathbb{F}^{n\times m}$ be a matrix with full column rank such that

$$\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2), \quad \mathbf{X}^*\mathbf{H}\mathbf{X} = \mathbf{D}_1 \oplus \mathbf{0}_q, \quad \mathbf{D}_1 = \mathrm{diag}_p(\pm 1)$$

where $\mathbf{X}_1 \in \mathbb{F}^{n\times p}$, $\mathbf{X}_2 \in \mathbb{F}^{n\times q}$ and $p + q = m$. Moreover, let

$$\mathbf{H}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \text{ with } \mathbf{U} \in \mathbb{F}^{n\times m} \text{ and } \mathbf{\Sigma}, \mathbf{V} \in \mathbb{F}^{m\times m}$$

be a thin singular value decomposition, and let

$$\mathbf{Y} = (\mathbf{Y}_1|\mathbf{Y}_2) = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^*$$

where $\mathbf{Y}$ is partitioned according to $\mathbf{X}$. Then $\mathbf{Y}^*\mathbf{H}\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{I}_m$, so that the matrix defined by

$$\mathbf{X}^{(1)} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{Y}_2)$$

satisfies

$$\mathbf{X}^{(1)*}\mathbf{H}\mathbf{X}^{(1)} = \mathbf{D}_1 \oplus \begin{bmatrix} \mathbf{0}_q & \mathbf{I}_q \\ \mathbf{I}_q & \mathbf{Y}_2^*\mathbf{H}\mathbf{Y}_2 \end{bmatrix}.$$

Now the block $\mathbf{Y}_2^*\mathbf{H}\mathbf{Y}_2$ must be eliminated and

$$\begin{bmatrix}\mathbf{I} & \mathbf{A}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^* \begin{bmatrix}\mathbf{0} & \mathbf{I}\\ \mathbf{I} & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{A}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{I}\\ \mathbf{I} & \mathbf{A}+\mathbf{A}^*+\mathbf{G}\end{bmatrix}$$

with $\mathbf{A} = -\frac{1}{2}\mathbf{G}$ shows, that this can be done by applying the transformation

$$\mathbf{I}_p \oplus \begin{bmatrix}\mathbf{I}_q & -\frac{1}{2}\mathbf{Y}_2^*\mathbf{H}\mathbf{Y}_2\\ \mathbf{0}_q & \mathbf{I}_q\end{bmatrix}$$

to $\mathbf{X}^{(1)}$. Hence, for

$$\mathbf{X}^{(2)} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3) \text{ with } \mathbf{X}_3 = \mathbf{Y}_2 - \tfrac{1}{2}\mathbf{X}_2(\mathbf{Y}_2^*\mathbf{H}\mathbf{Y}_2)$$

it follows that

$$\mathbf{X}^{(2)*}\mathbf{H}\mathbf{X}^{(2)} = \mathbf{D}_1 \oplus \mathbf{D}_{23}, \quad \mathbf{D}_{23} = \begin{bmatrix}\mathbf{0}_q & \mathbf{I}_q\\ \mathbf{I}_q & \mathbf{0}_q\end{bmatrix}.$$

In order to extend $\mathbf{X}^{(2)}$ to a nonsingular $n \times n$ matrix let

$$\mathbf{X}^{(2)} = \mathbf{Q}\mathbf{R} \text{ with } \mathbf{Q} = (\mathbf{Q}_1|\mathbf{Q}_2|\mathbf{Q}_3|\mathbf{Q}_4) \in \mathbb{F}^{n\times n}$$

be a QR decomposition where $\mathbf{Q}$ is partitioned according to $\mathbf{X}$. Then

$$\mathbf{X}^{(3)} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3|\mathbf{Q}_4) = (\mathbf{X}^{(2)}|\mathbf{Q}_4)$$

satisfies

$$\mathbf{X}^{(3)*}\mathbf{H}\mathbf{X}^{(3)} = \begin{bmatrix}\mathbf{D}_1 \oplus \mathbf{D}_{23} & \mathbf{X}^{(2)*}\mathbf{H}\mathbf{Q}_4\\ \mathbf{Q}_4^*\mathbf{H}\mathbf{X}^{(2)} & \mathbf{Q}_4^*\mathbf{H}\mathbf{Q}_4\end{bmatrix}.$$

Now the blocks $\mathbf{X}^{(2)*}\mathbf{H}\mathbf{Q}_4$ and $\mathbf{Q}_4^*\mathbf{H}\mathbf{X}^{(2)}$ must be eliminated and

$$\begin{bmatrix}\mathbf{I} & \mathbf{A}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^* \begin{bmatrix}\mathbf{E} & \mathbf{F}\\ \mathbf{F}^* & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{A}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} = \begin{bmatrix}\mathbf{E} & \mathbf{E}\mathbf{A}+\mathbf{F}\\ \mathbf{A}^*\mathbf{E}+\mathbf{F}^* & \mathbf{A}^*\mathbf{E}\mathbf{A}+\mathbf{F}^*\mathbf{A}+\mathbf{A}^*\mathbf{F}+\mathbf{G}\end{bmatrix}$$

with $\mathbf{A} = -\mathbf{E}^{-1}\mathbf{F}$ shows, that this can be done by applying the transformation

$$\begin{bmatrix}\mathbf{I}_{p+2q} & -(\mathbf{D}_1 \oplus \mathbf{D}_{23})^{-1}\mathbf{X}^{(2)*}\mathbf{H}\mathbf{Q}_4\\ \mathbf{0} & \mathbf{I}_{n-p-2q}\end{bmatrix}$$

to $\mathbf{X}^{(3)}$. Hence, for

$$\mathbf{X}^{(4)} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3|\mathbf{Y}_4) \text{ with } \mathbf{Y}_4 = \mathbf{Q}_4 - \mathbf{X}^{(2)}(\mathbf{D}_1 \oplus \mathbf{D}_{23})\mathbf{X}^{(2)*}\mathbf{H}\mathbf{Q}_4$$

it follows that

$$\mathbf{X}^{(4)*}\mathbf{H}\mathbf{X}^{(4)} = \mathbf{D}_1 \oplus \mathbf{D}_{23} \oplus \mathbf{Y}_4^*\mathbf{H}\mathbf{Y}_4.$$

Finally, using a HQR decomposition

$$\mathbf{Y}_4\mathbf{P}_4 = \mathbf{X}_4\mathbf{R}_4 \text{ with } \mathbf{X}_4^*\mathbf{H}\mathbf{X}_4 = \mathbf{D}_4 = \operatorname{diag}_{n-p-2q}(\pm 1)$$

and setting

$$\mathbf{X}^{(5)} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3|\mathbf{X}_4)$$

we obtain

$$\mathbf{X}^{(5)*}\mathbf{H}\mathbf{X}^{(5)} = \mathbf{D}_1 \oplus \mathbf{D}_{23} \oplus \mathbf{D}_4.$$

Thus the wanted extension of $\mathbf{X}$ has been determined. $\Diamond$

Chapter 3

H-polar decompositions

3.1 Introduction

In Definition 1.3 we have already introduced the H-polar decomposition of a matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ as a factorisation of the form

$$\mathbf{A} = \mathbf{UM} \tag{3.1}$$

where $\mathbf{U}$ is an H-isometry and $\mathbf{M}$ is H-selfadjoint. These decompositions are investigated in detail in [BMRRR1–3] and [MRR] as well as in the further references specified there. More specialised results concerning H-polar decompositions of H-normal matrices (i.e. matrices which commute with their H-adjoint) are derived in [LMMR]. An essential result of these studies is the fact that not every square matrix admits an H-polar decomposition unless $\mathbf{H}$ is definite.

H-polar decompositions are also the central subject of this chapter, in which theoretical as well as practical questions are discussed. In Section 3.2 some results of the investigations cited above are reviewed and in Section 3.3 a new criterion for the existence of H-polar decompositions is derived. The remaining Sections 3.4 – 3.6 are concerned with the numerical computation of H-polar decompositions.

3.2 Canonical forms and H-polar decompositions

The summary of well-known results begins with a theorem on the canonical form of a complex matrix pair $(\mathbf{A}, \mathbf{H})$ where $\mathbf{A}$ is H-Hermitian. This form is obtained under a transformation of the kind $(\mathbf{A}, \mathbf{H}) \to (\mathbf{S}^{-1}\mathbf{AS}, \mathbf{S}^*\mathbf{HS})$. It goes back to results of Kronecker and Weierstrass and is fundamental for exploring H-Hermitian matrices.

Theorem 3.1 (Canonical form). *Let $\mathbf{H} \in \mathbb{C}^{n\times n}$ be nonsingular and Hermitian and let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian. Then there exists a nonsingular matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ such that*

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{A}_1 \oplus \ldots \oplus \mathbf{A}_k \quad \text{and} \quad \mathbf{S}^*\mathbf{HS} = \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_k \tag{3.2a}$$

where the blocks $\mathbf{A}_j$ and $\mathbf{H}_j$ are of equal size and each pair $(\mathbf{A}_j, \mathbf{H}_j)$ has one and only one of the following forms:

1. *Pairs belonging to real eigenvalues*

$$\mathbf{A}_j = \mathbf{J}_p(\lambda) \text{ and } \mathbf{H}_j = \varepsilon \mathbf{Z}_p \tag{3.2b}$$

with $\lambda \in \mathbb{R}$, $p \in \mathbb{N}$ *and* $\varepsilon \in \{+1, -1\}$.

2. *Pairs belonging to non-real eigenvalues*

$$\mathbf{A}_j = \begin{bmatrix} \mathbf{J}_p(\lambda) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_p(\overline{\lambda}) \end{bmatrix} \text{ and } \mathbf{H}_j = \begin{bmatrix} \mathbf{0} & \mathbf{Z}_p \\ \mathbf{Z}_p & \mathbf{0} \end{bmatrix} \tag{3.2c}$$

with $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $\operatorname{Im}(\lambda) > 0$ *and* $p \in \mathbb{N}$.

Moreover, the canonical form $(\mathbf{S}^{-1}\mathbf{AS}, \mathbf{S}^*\mathbf{HS})$ *of* $(\mathbf{A}, \mathbf{H})$ *is uniquely determined up to the permutation of blocks.*

Proof. See [GLR, Theorem I.3.3]. □

The ordered set of the signs ε appearing in the blocks (3.2b) is an invariant of the canonical form and is called its sign characteristic. It allows to classify H-Hermitian matrices by H-unitary similarity [GLR, Section I.3.5]. Furthermore, an analogous form also exists for real matrices [GLR, Theorem I.5.3], but this is not required for our investigations.

The next statements summarise the most important results on the existence of H-polar decompositions.

Theorem 3.2 (H-polar decomposition of real matrices). *Let* $\mathbf{H} \in \mathbb{R}^{n\times n}$ *be nonsingular and symmetric, and let* $\mathbf{A} \in \mathbb{R}^{n\times n}$. *Then* $\mathbf{A}$ *admits a real H-polar decomposition* $\mathbf{A} = \mathbf{U}_r\mathbf{M}_r$ ($\mathbf{U}_r \in \mathbb{R}^{n\times n}$ *is H-orthogonal,* $\mathbf{M}_r \in \mathbb{R}^{n\times n}$ *is H-symmetric*) *if and only if it admits a complex H-polar decomposition* $\mathbf{A} = \mathbf{U}_c\mathbf{M}_c$ ($\mathbf{U}_c \in \mathbb{C}^{n\times n}$ *is H-unitary,* $\mathbf{M}_c \in \mathbb{C}^{n\times n}$*is H-Hermitian*).

Proof. See [BMRRR1, Lemma 4.2]. □

Theorem 3.3 (Existence of H-polar decompositions). *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$ *and let* $\mathbf{A} \in \mathbb{F}^{n\times n}$. *Then* $\mathbf{A}$ *admits an H-polar decomposition if and only if there exists an H-selfadjoint matrix* $\mathbf{M} \in \mathbb{F}^{n\times n}$ *such that* $\mathbf{M}^2 = \mathbf{A}^{[*]}\mathbf{A}$ *and* $\ker \mathbf{M} = \ker \mathbf{A}$.

Proof. See [BMRRR1, Theorem 4.1] and [BR, Lemma 4.1]. □

Theorem 3.4 (Existence of H-selfadjoint square roots). *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$ *and let* $\mathbf{A} \in \mathbb{F}^{n\times n}$. *Then there exists an H-selfadjoint matrix* $\mathbf{M}$ *such that* $\mathbf{M}^2 = \mathbf{A}^{[*]}\mathbf{A}$ *and* $\ker \mathbf{M} = \ker \mathbf{A}$ *if and only if the canonical form of the pair* $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H})$ *satisfies the following conditions:*

1. *Blocks belonging to a negative real eigenvalue* $\lambda < 0$ *can be represented in the form*

$$\left(\bigoplus_{i=1}^{r} [\mathbf{J}_{p_i}(\lambda) \oplus \mathbf{J}_{p_i}(\lambda)], \bigoplus_{i=1}^{r} [\mathbf{Z}_{p_i} \oplus -\mathbf{Z}_{p_i}]\right).$$

2. *Blocks belonging to the eigenvalue* 0 *can be represented in the form* $(\mathbf{J}^{(1)} \oplus \mathbf{J}^{(2)} \oplus \mathbf{J}^{(3)}, \mathbf{Z}^{(1)} \oplus \mathbf{Z}^{(2)} \oplus \mathbf{Z}^{(3)})$ *where*

$$(\mathbf{J}^{(1)}, \mathbf{Z}^{(1)}) = \left(\bigoplus_{i=1}^{r} [\mathbf{N}_{p_i} \oplus \mathbf{N}_{p_i}], \bigoplus_{i=1}^{r} [\mathbf{Z}_{p_i} \oplus -\mathbf{Z}_{p_i}] \right) \text{ with } p_i \geq 1,$$

$$(\mathbf{J}^{(2)}, \mathbf{Z}^{(2)}) = \left(\bigoplus_{j=1}^{s} [\mathbf{N}_{p_j} \oplus \mathbf{N}_{p_j-1}], \bigoplus_{j=1}^{s} [\varepsilon_j \mathbf{Z}_{p_j} \oplus \varepsilon_j \mathbf{Z}_{p_j-1}] \right) \text{ with } p_j > 1,$$

$$(\mathbf{J}^{(3)}, \mathbf{Z}^{(3)}) = \left(\bigoplus_{k=1}^{t} 0, \bigoplus_{k=1}^{t} \varepsilon_k \right).$$

3. *If a basis in which the blocks from 2. exist is denoted with* $E_1 \cup E_2 \cup E_3$,

$$E_1 = \{\mathbf{e}_{i,k}^{(1)}\}_{i=1}^{r}{}_{k=1}^{2p_i}, \quad E_2 = \{\mathbf{e}_{i,k}^{(2)}\}_{i=1}^{s}{}_{k=1}^{2p_i-1}, \quad E_3 = \{\mathbf{e}_{i,1}^{(3)}\}_{i=1}^{t},$$

then such a basis must exist in which

$$\ker \mathbf{A} = \operatorname{span}\{\mathbf{e}_{i,1}^{(1)} + \mathbf{e}_{i,p_i+1}^{(1)}\}_{i=1}^{r} \oplus \operatorname{span}\{\mathbf{e}_{i,1}^{(2)}\}_{i=1}^{s} \oplus \operatorname{span}\{\mathbf{e}_{i,1}^{(3)}\}_{i=1}^{t}.$$

(*Remark: From this condition it follows that* $\ker \mathbf{M} = \ker \mathbf{A}$.)

Proof. See [BMRRR1, Theorem 4.4] and [BMRRR3, Errata]. □

Whereas Theorem 3.2 makes it possible to transfer results concerning complex H-polar decompositions to real decompositions, Theorem 3.3 — whose proof is based on Witt's theorem — and Theorem 3.4 constitute the essential criterion for the existence of H-polar decompositions. Note that there is an error in the original Theorem 4.4 in [BMRRR1] which is pointed out by the following example.

Example 3.5. With the notation used in [BMRRR1], let

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{X} = \frac{1}{\sqrt{1-\xi^2}} \begin{bmatrix} 1+\xi & -1-\xi \\ 1+\xi & -1-\xi \end{bmatrix}, \quad \mathbf{X}^{[*]}\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

with $-1 < \xi < 1$. Then according to the statement (ii) of Theorem 4.4 in [BMRRR1] the equation $(\mathbf{X}^{[*]}\mathbf{X}, \mathbf{H}) = (\mathbf{B}_0, \mathbf{H}_0)$ is satisfied and $\ker \mathbf{B}_0 = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$. However, $\ker \mathbf{X} = \operatorname{span}\{\mathbf{e}_1 + \mathbf{e}_2\} \neq \ker \mathbf{B}_0$, so that according to the statement (iii) of the theorem the H-polar decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{A}, \quad \mathbf{U} = \frac{1}{\sqrt{1-\xi^2}} \begin{bmatrix} 1 & \xi \\ \xi & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

should not exist. ◊

This error is corrected by making a change in the second condition of Theorem 3.4. For the size of the blocks from $(\mathbf{J}^{(1)}, \mathbf{Z}^{(1)})$ now the condition $p_i \geq 1$ is imposed instead of the original condition $p_i > 1$. This correction is made in [BMRRR3, Errata] and the Theorem 3.21 contained in Section 3.4 also shows the need for modifying the condition.

If a matrix admits an H-polar decomposition, then it mostly admits several H-polar decompositions. The various decompositions are described in the following result.

Theorem 3.6 (Canonical forms of H-selfadjoint square roots). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{A} \in \mathbb{F}^{n\times n}$. If there exists an H-selfadjoint matrix $\mathbf{M}$ such that $\mathbf{M}^2 = \mathbf{A}^{[*]}\mathbf{A}$ and $\ker \mathbf{M} = \ker \mathbf{A}$, then the following relationships exist between the (complex) canonical form of the pairs $(\mathbf{M}^2, \mathbf{H})$ and $(\mathbf{M}, \mathbf{H})$:*

a. *Non-real eigenvalues. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a block of the form*

$$(\mathbf{J}_p(\alpha + i\beta) \oplus \mathbf{J}_p(\alpha - i\beta), \mathbf{Z}_{2p}) \text{ with } \alpha, \beta \in \mathbb{R} \text{ and } \beta > 0,$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a block of the form

$$(\mathbf{J}_p(\lambda) \oplus \mathbf{J}_p(\overline{\lambda}), \mathbf{Z}_{2p}) \text{ or } (\mathbf{J}_p(-\lambda) \oplus \mathbf{J}_p(-\overline{\lambda}), \mathbf{Z}_{2p}) \text{ with } \lambda^2 = \alpha + i\beta.$$

b. *Positive real eigenvalues. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a block of the form*

$$(\mathbf{J}_p(\alpha^2), \varepsilon\mathbf{Z}_p) \text{ with } \alpha > 0,$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a block of the form

$$(\mathbf{J}_p(\alpha), \varepsilon\mathbf{Z}_p) \text{ or } (\mathbf{J}_p(-\alpha), (-1)^{p+1}\varepsilon\mathbf{Z}_p).$$

c. *Negative real eigenvalues. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a block of the form*

$$(\mathbf{J}_p(-\beta^2) \oplus \mathbf{J}_p(-\beta^2), \mathbf{Z}_p \oplus -\mathbf{Z}_p) \text{ with } \beta > 0,$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a block of the form

$$(\mathbf{J}_p(i\beta) \oplus \mathbf{J}_p(-i\beta), \mathbf{Z}_{2p}).$$

d. *First case with eigenvalue 0. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a block of the form*

$$(\mathbf{N}_p \oplus \mathbf{N}_p, \mathbf{Z}_p \oplus -\mathbf{Z}_p) \in (\mathbf{J}^{(1)}, \mathbf{Z}^{(1)}),$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a block of the form

$$(\mathbf{N}_{2p}, \mathbf{Z}_{2p}) \text{ or } (\mathbf{N}_{2p}, -\mathbf{Z}_{2p}).$$

Moreover, a canonical basis can be chosen in such a way that the eigenvector of $\mathbf{M}$ coincides with the sum of the eigenvectors of the two Jordan blocks of $\mathbf{M}^2$.

e. *Second case with eigenvalue 0. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a block of the form*

$$(\mathbf{N}_p \oplus \mathbf{N}_{p-1}, \varepsilon\mathbf{Z}_p \oplus \varepsilon\mathbf{Z}_{p-1}) \in (\mathbf{J}^{(2)}, \mathbf{Z}^{(2)}),$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a block of the form

$$(\mathbf{N}_{2p-1}, \varepsilon\mathbf{Z}_{2p-1}).$$

Moreover, a canonical basis can be chosen in such a way that the eigenvector of $\mathbf{M}$ coincides with the eigenvector of the $p \times p$ Jordan block of $\mathbf{M}^2$.

f. Third case with eigenvalue 0. If the canonical form of $(\mathbf{M}^2, \mathbf{H})$ *contains a block of the form*

$$(0, \varepsilon) \in (\mathbf{J}^{(3)}, \mathbf{Z}^{(3)}),$$

then the canonical form of $(\mathbf{M}, \mathbf{H})$ *contains a block of the form*

$$(0, \varepsilon).$$

Proof. See [BMRRR1, Lemma 7.8]. □

We end this summary with a useful corollary which is obtained by combining the Theorems 3.3, 3.4 and 3.6 (a, b).

Corollary 3.7. *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$ *and let* $\mathbf{A} \in \mathbb{F}^{n\times n}$ *such that* $\sigma(\mathbf{A}^{[*]}\mathbf{A}) \subset \mathbb{C}\backslash(-\infty, 0]$. *Then* $\mathbf{A}$ *admits an H-polar decomposition* $\mathbf{A} = \mathbf{UM}$ *such that* $\sigma(\mathbf{M}) \subset \{z \in \mathbb{C} | \operatorname{Re}(z) > 0\}$.

3.3 A new criterion for the existence of H-polar decompositions

We will now derive a new criterion for the existence of H-polar decompositions. Thereby H-Hermitian as well as Z-Hermitian matrices occur, so that the notation $\mathbf{A}^H = \mathbf{A}^{[*]_H}$ will be used. The criterion is investigated only for complex matrices, with regard to Theorem 3.2, and is based on the following observation.

Lemma 3.8. *If* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *admits an H-polar decomposition, then the canonical forms of the pairs* $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ *and* $(\mathbf{AA}^H, \mathbf{H})$ *are identical.*

Proof. Let $\mathbf{A} = \mathbf{UM}$ be an H-polar decomposition of $\mathbf{A}$. Then $\mathbf{U}^H = \mathbf{U}^{-1}$ and $\mathbf{M}^H = \mathbf{M}$ imply

$$\mathbf{U}^{-1}\mathbf{AA}^H\mathbf{U} = \mathbf{U}^{-1}(\mathbf{UM})(\mathbf{M}^H\mathbf{U}^H)\mathbf{U} = \mathbf{M}^2 = (\mathbf{M}^H\mathbf{U}^H)(\mathbf{UM}) = \mathbf{A}^H\mathbf{A},$$

so that $\mathbf{A}^H\mathbf{A}$ and $\mathbf{AA}^H$ are H-unitary similar. If now $(\mathbf{R}^{-1}\mathbf{A}^H\mathbf{AR}, \mathbf{R}^*\mathbf{HR}) = (\mathbf{J}, \mathbf{Z})$ is the canonical form of the pair $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ and if $\mathbf{S} = \mathbf{UR}$, then $(\mathbf{S}^{-1}\mathbf{AA}^H\mathbf{S}, \mathbf{S}^*\mathbf{HS}) = (\mathbf{R}^{-1}\mathbf{U}^{-1}\mathbf{AA}^H\mathbf{UR}, \mathbf{R}^*\mathbf{U}^*\mathbf{HUR}) = (\mathbf{J}, \mathbf{Z})$ is the canonical form of the pair $(\mathbf{AA}^H, \mathbf{H})$, too. □

Clearly, the question arises whether the converse of this statement is also true. For nonsingular matrices it can be answered with the help of the following Lemma.

Lemma 3.9. *If* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *is nonsingular, then* $\mathbf{A}$ *has a square root which can be expressed as an invertible polynomial in* $\mathbf{A}$.

Proof. Let $\mathbf{R}^{-1}\mathbf{AR} = \mathbf{J}_{p_1}(\lambda_1) \oplus \ldots \oplus \mathbf{J}_{p_k}(\lambda_k)$ be the Jordan normal form of $\mathbf{A}$. Then every matrix defined by

$$\sqrt{\mathbf{A}} = \mathbf{R}\left(\sqrt{\mathbf{J}_{p_1}(\lambda_1)} \oplus \ldots \oplus \sqrt{\mathbf{J}_{p_k}(\lambda_k)}\right)\mathbf{R}^{-1}$$

where

$$\sqrt{\mathbf{J}_p(\lambda)} = \begin{bmatrix} f(\lambda) & \frac{f'(\lambda)}{1!} & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(p-1)}(\lambda)}{(p-1)!} \\ & f(\lambda) & \frac{f'(\lambda)}{1!} & \ddots & \vdots \\ & & f(\lambda) & \ddots & \frac{f''(\lambda)}{2!} \\ & & & \ddots & \frac{f'(\lambda)}{1!} \\ & & & & f(\lambda) \end{bmatrix} \quad \text{with } f(\lambda) = \sqrt{\lambda}$$

is a square root of $\mathbf{A}$. Moreover, if for all multiple eigenvalues $\lambda_0 \in \sigma(\mathbf{A})$ the same (!) branch of the multi-valued function $\sqrt{\lambda_0}$ is used in the blocks $\sqrt{\mathbf{J}_p(\lambda_0)}$, then $\sqrt{\mathbf{A}}$ according to [WED, Chapter VII, Theorem 2][6] can also be expressed as an invertible polynomial in $\mathbf{A}$. □

Theorem 3.10. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be nonsingular and let the canonical forms of the pairs $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ and $(\mathbf{A}\mathbf{A}^H, \mathbf{H})$ be identical. Then $\mathbf{A}$ admits an H-polar decomposition.*

Proof. Let $\mathbf{R}$ and $\mathbf{S}$ be nonsingular matrices in $\mathbb{C}^{n\times n}$ such that

$$(\mathbf{R}^{-1}\mathbf{A}^H\mathbf{A}\mathbf{R}, \mathbf{R}^*\mathbf{H}\mathbf{R}) = (\mathbf{J}, \mathbf{Z}) = (\mathbf{S}^{-1}\mathbf{A}\mathbf{A}^H\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S})$$

is the canonical form of the pairs $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ and $(\mathbf{A}\mathbf{A}^H, \mathbf{H})$. Then the nonsingular matrix defined by

$$\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{R}$$

is Z-normal, as can be seen with $\mathbf{Z}^{-1} = \mathbf{Z}^* = \mathbf{Z}$ from

$$\begin{aligned} (\mathbf{Z}\mathbf{B}^*\mathbf{Z})\mathbf{B} &= \mathbf{Z}(\mathbf{R}^*\mathbf{A}^*\mathbf{S}^{-*})\mathbf{Z}(\mathbf{S}^{-1}\mathbf{A}\mathbf{R}) = \mathbf{Z}\mathbf{R}^*\mathbf{A}^*\mathbf{H}\mathbf{A}\mathbf{R} \\ &= \mathbf{R}^{-1}\mathbf{H}^{-1}\mathbf{A}^*\mathbf{H}\mathbf{A}\mathbf{R} = \mathbf{J}, \\ \mathbf{B}(\mathbf{Z}\mathbf{B}^*\mathbf{Z}) &= (\mathbf{S}^{-1}\mathbf{A}\mathbf{R})\mathbf{Z}(\mathbf{R}^*\mathbf{A}^*\mathbf{S}^{-*})\mathbf{Z} = \mathbf{S}^{-1}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^*\mathbf{S}^{-*}\mathbf{Z} \\ &= \mathbf{S}^{-1}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^*\mathbf{H}\mathbf{S} = \mathbf{J}. \end{aligned}$$

Let $f(\mathbf{B})$ denote an arbitrary polynomial in $\mathbf{B}$. Then the commutability of $\mathbf{B}$ and $\mathbf{Z}\mathbf{B}^*\mathbf{Z}$ implies $(\mathbf{Z}\mathbf{B}^*\mathbf{Z})f(\mathbf{B}) = f(\mathbf{B})(\mathbf{Z}\mathbf{B}^*\mathbf{Z})$ or $(\mathbf{Z}f(\mathbf{B})^*\mathbf{Z})\mathbf{B} = \mathbf{B}(\mathbf{Z}f(\mathbf{B})^*\mathbf{Z})$ from which it follows that

$$(\mathbf{Z}f(\mathbf{B})^*\mathbf{Z})f(\mathbf{B}) = f(\mathbf{B})(\mathbf{Z}f(\mathbf{B})^*\mathbf{Z}).$$

Moreover, if $f(\mathbf{B})$ is invertible, then $f(\mathbf{B}) = f(\mathbf{B})(\mathbf{Z}f(\mathbf{B})^*\mathbf{Z})(\mathbf{Z}f(\mathbf{B})^{-*}\mathbf{Z}) = (\mathbf{Z}f(\mathbf{B})^*\mathbf{Z})f(\mathbf{B})(\mathbf{Z}f(\mathbf{B})^{-*}\mathbf{Z})$, so that

$$(\mathbf{Z}f(\mathbf{B})^{-*}\mathbf{Z})f(\mathbf{B}) = f(\mathbf{B})(\mathbf{Z}f(\mathbf{B})^{-*}\mathbf{Z}).$$

Consequently, if $\sqrt{\mathbf{B}}$ is a square root of $\mathbf{B}$ such that $\sqrt{\mathbf{B}} = f(\mathbf{B})$, then the matrices defined by

$$\begin{aligned} \mathbf{K} &= [\mathbf{Z}(\sqrt{\mathbf{B}})^*\mathbf{Z}](\sqrt{\mathbf{B}}) = (\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^*\mathbf{Z}], \\ \mathbf{T} &= (\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}] = [\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}](\sqrt{\mathbf{B}}) \end{aligned}$$

[6] For a deeper understanding of this statement, the corresponding fundamentals can be studied in [WED, Chapters VII, VIII] and in [G, Chapters V, VIII].

on the one hand satisfy

$$\begin{aligned}\mathbf{TK} &= (\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}][\mathbf{Z}(\sqrt{\mathbf{B}})^{*}\mathbf{Z}](\sqrt{\mathbf{B}}) = (\sqrt{\mathbf{B}})^2 = \mathbf{B},\\ \mathbf{KT} &= (\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^{*}\mathbf{Z}][\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}](\sqrt{\mathbf{B}}) = (\sqrt{\mathbf{B}})^2 = \mathbf{B}.\end{aligned}$$

On the other hand it is also true that

$$\begin{aligned}\mathbf{K}^*\mathbf{Z} &= (\sqrt{\mathbf{B}})^*\mathbf{Z}(\sqrt{\mathbf{B}}) = \mathbf{ZK},\\ \mathbf{T}^*\mathbf{ZT} &= [\mathbf{Z}(\sqrt{\mathbf{B}})^{-1}\mathbf{Z}](\sqrt{\mathbf{B}})^*\mathbf{Z}(\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}]\\ &= \mathbf{Z}(\sqrt{\mathbf{B}})^{-1}(\sqrt{\mathbf{B}})[\mathbf{Z}(\sqrt{\mathbf{B}})^{*}\mathbf{Z}][\mathbf{Z}(\sqrt{\mathbf{B}})^{-*}\mathbf{Z}] = \mathbf{Z},\end{aligned}$$

so that $\mathbf{B} = \mathbf{TK} = \mathbf{KT}$ is a Z-polar decomposition of $\mathbf{B}$ with (in this case) commuting factors. Finally, let

$$\mathbf{M} = \mathbf{RKR}^{-1} \text{ and } \mathbf{U} = \mathbf{STR}^{-1}.$$

Then

$$\begin{aligned}\mathbf{UM} &= (\mathbf{STR}^{-1})(\mathbf{RKR}^{-1}) = \mathbf{SBR}^{-1} = \mathbf{A},\\ \mathbf{M}^*\mathbf{H} &= (\mathbf{R}^{-*}\mathbf{K}^*\mathbf{R}^*)(\mathbf{R}^{-*}\mathbf{ZR}^{-1}) = (\mathbf{R}^{-*}\mathbf{ZR}^{-1})(\mathbf{RKR}^{-1}) = \mathbf{HM},\\ \mathbf{U}^*\mathbf{HU} &= (\mathbf{R}^{-*}\mathbf{T}^*\mathbf{S}^*)\mathbf{H}(\mathbf{STR}^{-1}) = \mathbf{R}^{-*}(\mathbf{T}^*\mathbf{ZT})\mathbf{R}^{-1} = \mathbf{R}^{-*}\mathbf{ZR}^{-1} = \mathbf{H}\end{aligned}$$

is the wanted H-polar decomposition of $\mathbf{A}$. □

A matrix $\mathbf{A}$ satisfying $\mathbf{A}^H\mathbf{A} = \mathbf{AA}^H$ is called an H-normal matrix. It is a trivial fact that for those matrices the canonical forms of $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ and $(\mathbf{AA}^H, \mathbf{H})$ are identical, and that $\mathbf{R} = \mathbf{S}$ in the proof of Theorem 3.10. This immediately implies the next result which has also been proved in a different way in [LMMR, Theorem 29].

Corollary 3.11. *Every nonsingular H-normal matrix in $\mathbb{C}^{n\times n}$ admits an H-polar decomposition with commuting factors.*

In the case of singular matrices the square root cannot be built according to Lemma 3.9 and the question of the validity of a statement corresponding to Theorem 3.10 has not yet been clarified completely. But it can be answered for the case in which all blocks of the canonical form $(\mathbf{J}, \mathbf{Z})$ belonging to the eigenvalue 0 are of size 1 (nilpotency of index 1).

Lemma 3.12. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ such that $\mathbf{A}^H\mathbf{A} = \mathbf{0}$. Then $\mathbf{A}$ admits an H-polar decomposition if and only if also $\mathbf{AA}^H = \mathbf{0}$.*

Proof. [⇒]: Let $\mathbf{A} = \mathbf{UM}$ be an H-polar decomposition of $\mathbf{A}$. Then $\mathbf{M}^2 = \mathbf{A}^H\mathbf{A} = \mathbf{0}$ implies $\mathbf{AA}^H = \mathbf{UM}^2\mathbf{U}^H = \mathbf{0}$.

[⇐]: For all matrices $\mathbf{A} \in \mathbb{C}^{n\times n}$ the easily proved equations

$$\operatorname{im}\mathbf{A}^H = (\ker\mathbf{A})^{[\perp]} \text{ and } \ker\mathbf{A}^H = (\operatorname{im}\mathbf{A})^{[\perp]}$$

are true [GLR, Proposition I.2.1], so that $\mathbf{AA}^H = \mathbf{0}$ on the one hand implies

$$(\ker\mathbf{A})^{[\perp]} = \operatorname{im}\mathbf{A}^H \subset \ker\mathbf{A}, \text{ i.e. } (\ker\mathbf{A})^{[\perp]} = \ker\mathbf{A} \cap (\ker\mathbf{A})^{[\perp]},$$

and $\mathbf{A}^H\mathbf{A} = \mathbf{0}$ on the other hand implies

$$\operatorname{im}\mathbf{A} \subset \ker\mathbf{A}^H = (\operatorname{im}\mathbf{A})^{[\perp]}, \text{ i.e. } \operatorname{im}\mathbf{A} = (\operatorname{im}\mathbf{A})^{[\perp]} \cap \operatorname{im}\mathbf{A}.$$

Thus, if

$$r = \operatorname{rank}\mathbf{A} = \dim(\operatorname{im}\mathbf{A}) = \dim(\operatorname{im}\mathbf{H}^{-1}\mathbf{A}^*\mathbf{H}) = \dim(\operatorname{im}\mathbf{A}^H) = \operatorname{rank}\mathbf{A}^H,$$

then $\mathbb{C}^n$ can be expressed according to Theorem 2.5 in the form

$$\mathbb{C}^n = X_1 \oplus X_0' \oplus X_0'' \text{ with } \ker\mathbf{A} = X_1 \oplus X_0' \text{ and } X_0' = \ker\mathbf{A} \cap (\ker\mathbf{A})^{[\perp]}$$

as well as in the form

$$\mathbb{C}^n = Y_2 \oplus Y_0' \oplus Y_0'' \text{ with } (\operatorname{im}\mathbf{A})^{[\perp]} = Y_2 \oplus Y_0' \text{ and } Y_0' = (\operatorname{im}\mathbf{A})^{[\perp]} \cap \operatorname{im}\mathbf{A}$$

where X_0', X_0'', Y_0', Y_0'' are neutral subspaces of dimension r and X_1, Y_2 are non-degenerate subspaces of dimension $p+q = n-2r$.

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_{p+q}, \mathbf{x}_1', \dots, \mathbf{x}_r'\}$ be an H-orthonormal basis of $\ker\mathbf{A}$. Then this basis can be extended to a complete basis of $\mathbb{C}^n$ by r further vectors $\mathbf{x}_1'', \dots, \mathbf{x}_r''$ according to Theorem 2.8, so that the matrix

$$\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_{p+q}\, \mathbf{x}_1' \dots \mathbf{x}_r'\, \mathbf{x}_1'' \dots \mathbf{x}_r'']$$

consisting of these basis vectors satisfies

$$\mathbf{X}^*\mathbf{H}\mathbf{X} = \mathbf{Z} \text{ with}$$
$$\mathbf{Z} = \begin{bmatrix} \mathbf{I}_p & \\ & -\mathbf{I}_q \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_r & \mathbf{I}_r \\ \mathbf{I}_r & \mathbf{0}_r \end{bmatrix} \text{ and } \mathbf{A}\mathbf{X} = [\mathbf{0}_1 \dots \mathbf{0}_{p+q+r}\, \mathbf{y}_1' \dots \mathbf{y}_r'].$$

Here $\{\mathbf{y}_1', \dots, \mathbf{y}_r'\}$ is a basis of the neutral space $\operatorname{im}\mathbf{A}$, which can also be extended to a complete basis of $\mathbb{C}^n$ by $p+q+r$ further vectors $\mathbf{y}_1, \dots, \mathbf{y}_{p+q}$, $\mathbf{y}_1'', \dots, \mathbf{y}_r''$ according to Theorem 2.8, so that the matrix

$$\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_{p+q}\, \mathbf{y}_1' \dots \mathbf{y}_r'\, \mathbf{y}_1'' \dots \mathbf{y}_r'']$$

consisting of these basis vectors satisfies

$$\mathbf{Y}^*\mathbf{H}\mathbf{Y} = \mathbf{Z} \text{ and}$$
$$\mathbf{Y}\mathbf{K} = [\mathbf{0}_1 \dots \mathbf{0}_{p+q+r}\, \mathbf{y}_1' \dots \mathbf{y}_r'] \text{ with } \mathbf{K} = \begin{bmatrix} \mathbf{0}_p & \\ & \mathbf{0}_q \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_r & \mathbf{I}_r \\ \mathbf{0}_r & \mathbf{0}_r \end{bmatrix}.$$

Moreover, the matrices $\mathbf{Z}$ and $\mathbf{K}$ fulfil $\mathbf{A}\mathbf{X} = \mathbf{Y}\mathbf{K}$, $\mathbf{Z}^{-1} = \mathbf{Z}^* = \mathbf{Z}$ and $\mathbf{K}^*\mathbf{Z} = \mathbf{0}_{p+q} \oplus \mathbf{I}_r \oplus \mathbf{0}_r = \mathbf{Z}\mathbf{K}$. Finally, let

$$\mathbf{M} = \mathbf{X}\mathbf{K}\mathbf{X}^{-1} \text{ and } \mathbf{U} = \mathbf{Y}\mathbf{X}^{-1}.$$

Then

$$\begin{aligned} \mathbf{U}\mathbf{M} &= \mathbf{Y}(\mathbf{X}^{-1}\mathbf{X})\mathbf{K}\mathbf{X}^{-1} = \mathbf{Y}\mathbf{K}\mathbf{X}^{-1} = \mathbf{A}, \\ \mathbf{M}^*\mathbf{H} &= \mathbf{X}^{-*}\mathbf{K}^*(\mathbf{X}^*\mathbf{H}) = \mathbf{X}^{-*}(\mathbf{K}^*\mathbf{Z})\mathbf{X}^{-1} \\ &= (\mathbf{X}^{-*}\mathbf{Z})\mathbf{K}\mathbf{X}^{-1} = \mathbf{H}\mathbf{X}\mathbf{K}\mathbf{X}^{-1} = \mathbf{H}\mathbf{M}, \\ \mathbf{U}^*\mathbf{H}\mathbf{U} &= \mathbf{X}^{-*}(\mathbf{Y}^*\mathbf{H}\mathbf{Y})\mathbf{X}^{-1} = \mathbf{X}^{-*}\mathbf{Z}\mathbf{X}^{-1} = \mathbf{H} \end{aligned}$$

is the wanted H-polar decomposition of $\mathbf{A}$. □

Remark 3.13. In addition to the H-polar decomposition of $\mathbf{A}$ given in the proof,

$$\mathbf{A}^H = \tilde{\mathbf{U}}\tilde{\mathbf{M}} \text{ with } \tilde{\mathbf{M}} = \mathbf{U}\mathbf{M}\mathbf{U}^{-1} = \mathbf{Y}\mathbf{K}\mathbf{Y}^{-1} \text{ and } \tilde{\mathbf{U}} = \mathbf{U}^{-1} = \mathbf{X}\mathbf{Y}^{-1}$$

is an H-polar decomposition of $\mathbf{A}^H$ and, furthermore, $\mathbf{A}^H\mathbf{Y} = \mathbf{X}\mathbf{K}$ as can be verified using $\mathbf{A}^H = \mathbf{H}^{-1}\mathbf{A}^*\mathbf{H}$. Thus the basis vectors can be assigned to the subspaces as follows

$$\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{p+q}, \overbrace{\mathbf{x}'_1, \ldots, \mathbf{x}'_r}^{\operatorname{im}\mathbf{A}^H}}_{\ker\mathbf{A}}, \mathbf{x}''_1, \ldots, \mathbf{x}''_r, \quad \underbrace{\mathbf{y}_1, \ldots, \mathbf{y}_{p+q}, \overbrace{\mathbf{y}'_1, \ldots, \mathbf{y}'_r}^{\operatorname{im}\mathbf{A}}}_{\ker\mathbf{A}^H}, \mathbf{y}''_1, \ldots, \mathbf{y}''_r.$$

This is also evident from the equations

$$\operatorname{im}\mathbf{A}^H \subset \ker\mathbf{A} = (\operatorname{im}\mathbf{A}^H)^{[\perp]} \text{ and } (\ker\mathbf{A}^H)^{[\perp]} = \operatorname{im}\mathbf{A} \subset \ker\mathbf{A}^H$$

which were not used in the proof but are valid, too. ◊

The following examples explain Lemma 3.12 and allow to introduce the concept of H-indecomposability.

Example 3.14.

1. If $\lambda \in \mathbb{C}\backslash\{0\}$ and

$$\mathbf{H} = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{0}_p & \mathbf{0}_p \\ \mathbf{J}_p(\lambda) & \mathbf{J}_p(\lambda) \end{bmatrix}, \quad \mathbf{A}^H = \begin{bmatrix} \mathbf{J}_p(\lambda)^* & \mathbf{0}_p \\ \mathbf{J}_p(\lambda)^* & \mathbf{0}_p \end{bmatrix},$$

   then $\mathbf{A}^H\mathbf{A} = \mathbf{0}$ but $\mathbf{A}\mathbf{A}^H \neq \mathbf{0}$. Therefore $\mathbf{A}$ has no H-polar decompositions.

2. If $\lambda \in \mathbb{C}\backslash\{0\}$ and

$$\mathbf{H} = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{0}_p & \\ & \mathbf{J}_p(\lambda) \end{bmatrix}, \quad \mathbf{A}^H = \begin{bmatrix} \mathbf{J}_p(\lambda)^* & \\ & \mathbf{0}_p \end{bmatrix},$$

   then $\mathbf{A}^H\mathbf{A} = \mathbf{0}$ and $\mathbf{A}\mathbf{A}^H = \mathbf{0}$. Therefore $\mathbf{A}$ has H-polar decompositions, for example

$$\mathbf{A} = \mathbf{U}\mathbf{M} \text{ with } \mathbf{U} = \begin{bmatrix} & \mathbf{J}_p(\lambda)^{-*} \\ \mathbf{J}_p(\lambda) & \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{0}_p & \end{bmatrix}. \qquad \lozenge$$

A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is called H-decomposable if there exists a non-degenerate proper subspace $M \subset \mathbb{C}^n$ such that both M and $M^{[\perp]}$ are invariant under $\mathbf{A}$, otherwise $\mathbf{A}$ is called H-indecomposable. In particular, it is shown in the proof of [HO, Theorem 1] that the H-normal matrix $\mathbf{A}$ from the second example is H-indecomposable. This is important in connection with normal forms of H-normal matrices [LMMR, Theorem 10] and will help to give some explanations at the end of this section.

The following sufficient condition for the existence of H-polar decompositions can now be proved with the help of Lemma 3.12.

Theorem 3.15. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and let the canonical forms of the pairs* $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ *and* $(\mathbf{A}\mathbf{A}^H, \mathbf{H})$ *be identical. Furthermore, let all blocks of the canonical form belonging to the eigenvalue* 0 *be of size* 1. *Then* $\mathbf{A}$ *admits an H-polar decomposition.*

Proof. Let $\mathbf{R}, \mathbf{S}, \mathbf{J}, \mathbf{Z}$ and $\mathbf{B}$ be as in the proof of Theorem 3.10, so that $\mathbf{B}^Z\mathbf{B} = \mathbf{B}\mathbf{B}^Z = \mathbf{J}$ holds. Furthermore, let

$$\mathbf{J} = \mathbf{J}_1 \oplus \mathbf{J}_0, \;\; \mathbf{J}_0 \in \mathbb{C}^{m\times m} \;\text{ and }\; \mathbf{Z} = \mathbf{Z}_1 \oplus \mathbf{Z}_0, \;\; \mathbf{Z}_0 \in \mathbb{C}^{m\times m}$$

where $\mathbf{J}_0, \mathbf{Z}_0$ denotes the part of the canonical form belonging to the eigenvalue 0. Then the spectra of the blocks $\mathbf{J}_1$ and $\mathbf{J}_0$ are disjoint and, moreover, the matrices $\mathbf{J}$ and $\mathbf{B}$ commute, so that $\mathbf{B}$ must also have the form

$$\mathbf{B} = \mathbf{B}_1 \oplus \mathbf{B}_0, \;\; \mathbf{B}_0 \in \mathbb{C}^{m\times m}.$$

Now $\mathbf{B}_1^{Z_1}\mathbf{B}_1 = \mathbf{B}_1\mathbf{B}_1^{Z_1} = \mathbf{J}_1$ implies that $\mathbf{B}_1$ is nonsingular and it consequently admits a $\mathbf{Z}_1$-polar decomposition

$$\mathbf{B}_1 = \mathbf{T}_1\mathbf{K}_1 \; (= \mathbf{K}_1\mathbf{T}_1) \;\text{ with }\; \mathbf{T}_1^*\mathbf{Z}_1\mathbf{T}_1 = \mathbf{Z}_1 \;\text{ and }\; \mathbf{K}_1^*\mathbf{Z}_1 = \mathbf{Z}_1\mathbf{K}_1$$

constructed according to Theorem 3.10. Moreover, the assumption of the theorem yields $\mathbf{B}_0^{Z_0}\mathbf{B}_0 = \mathbf{B}_0\mathbf{B}_0^{Z_0} = \mathbf{J}_0 = \mathbf{0}_m$ and therefore $\mathbf{B}_0$ admits a $\mathbf{Z}_0$-polar decomposition

$$\mathbf{B}_0 = \mathbf{T}_0\mathbf{K}_0 \;\text{ with }\; \mathbf{T}_0^*\mathbf{Z}_0\mathbf{T}_0 = \mathbf{Z}_0 \;\text{ and }\; \mathbf{K}_0^*\mathbf{Z}_0 = \mathbf{Z}_0\mathbf{K}_0$$

constructed according to Lemma 3.12. Finally, let

$$\mathbf{T} = \mathbf{T}_1 \oplus \mathbf{T}_0, \;\; \mathbf{U} = \mathbf{S}\mathbf{T}\mathbf{R}^{-1} \;\text{ and }\; \mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_0, \;\; \mathbf{M} = \mathbf{R}\mathbf{K}\mathbf{R}^{-1}.$$

Then $\mathbf{T}\mathbf{K}$ is a Z-polar decomposition of $\mathbf{B}$ and $\mathbf{U}\mathbf{M}$ is an H-polar decomposition of $\mathbf{A}$. □

Corollary 3.16. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *be H-normal and let all blocks of the canonical form of the pair* $(\mathbf{A}^H\mathbf{A}, \mathbf{H})$ *belonging to the eigenvalue 0 be of size* 1. *Then* $\mathbf{A}$ *admits an H-polar decomposition.*

Remark 3.17. If the assumption regarding the part of the canonical form belonging to the eigenvalue 0 is not made in Theorem 3.15, the matrix $\mathbf{J}_0$ from the proof is a block diagonal matrix consisting of nilpotent Jordan blocks. Thus, if it could be proved that every H-normal matrix $\mathbf{A}$ with $(\mathbf{A}^H\mathbf{A})^k = \mathbf{0}$ for some k in $\mathbb{N}$ admits an H-polar decomposition, then the present restrictions regarding the block sizes for the eigenvalue 0 would no longer be needed. Moreover, the corresponding corollary would state that every H-normal matrix admits an H-polar decomposition. (*Note*: In the meantime it turned out that the conjecture expressed with this remark actually holds. A corresponding proof has been found by Mehl, Ran and Rodman [MERR, Theorem 4]. Corollary 5 and Corollary 6 of this paper show that Theorem 3.15 and Corollary 3.16 are valid even if the blocks of the canonical form belonging to the eigenvalue 0 have arbitrary structure.) ♢

A further criterion for the existence of H-polar decompositions of H-normal matrices is given in Theorem 34 of [LMMR]. The theorem states that an H-normal matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ admits an H-polar decomposition if each of its singular H-indecomposable blocks over $\mathbb{C}$ (if any) either

(i) has two distinct complex eigenvalues (one of them must be zero), or

(ii) is similar to one (necessary nilpotent) Jordan block.

In Theorem 35 of the same paper H-polar decompositions of singular H-normal Matrices $\mathbf{X} = \mathbf{UA} \in \mathbb{C}^{n\times n}$ with $\mathbf{U}^H = \mathbf{U}^{-1}$, $\mathbf{A}^H = \mathbf{A}$ are presented for all possible nontrivial cases in which $\mathbf{H}$ has exactly two negative eigenvalues and whose existence is not guaranteed by Theorem 34. In the listed cases, the index of nilpotency k of the matrices $\mathbf{X}^H\mathbf{X} = \mathbf{X}\mathbf{X}^H$ is

$$k = \begin{Bmatrix} 1, & \text{in case(I),(VI)-(VII)} \\ 2, & \text{in case(II)-(III),(VIII)-(XII)} \\ 3, & \text{in case(IV)-(V)} \end{Bmatrix}.$$

Thus, the existence of the given H-polar decompositions in the cases (I), (VI)-(VII) is ensured by Corollary 3.16 and, moreover, the hypothesis expressed in Remark 3.17 is supported, too. On the other hand for $\alpha \in \mathbb{R}$ and

$$\mathbf{H} = \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 0 & 1 & i\alpha \\ & 0 & 1 \\ & & 0 \end{bmatrix} \text{ with } \mathbf{X}^H\mathbf{X} = \mathbf{X}\mathbf{X}^H = \begin{bmatrix} 0 & 0 & 1 \\ & 0 & 0 \\ & & 0 \end{bmatrix}$$

the existence of the H-polar decomposition

$$\mathbf{X} = \mathbf{UA} \text{ with } \mathbf{U} = \begin{bmatrix} 1 & i\alpha & -\frac{1}{2}\alpha^2 + i\beta \\ & 1 & i\alpha \\ & & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ & 0 & 1 \\ & & 0 \end{bmatrix} \quad (\beta \in \mathbb{R})$$

is guaranteed by Theorem 34(ii) but not by Corollary 3.16, so that the two criteria are mutually supplementary.

3.4 Canonical forms and H-polar decompositions in the case of diagonalisable matrices

In addition to the theoretical results presented in the previous sections we will now consider the numerical computation of H-polar decompositions. We start with the following observation.

Theorem 3.18. *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{H} \in \mathbb{F}^{n\times n}$ be selfadjoint and positive definite. Then every matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ admits an H-polar decomposition.*

Proof. According to the assumption there always exists a nonsingular matrix $\mathbf{S} \in \mathbb{F}^{n\times n}$ such that $\mathbf{H} = \mathbf{S}^*\mathbf{S}$. For example, if $\mathbf{H} = \mathbf{L}\mathbf{L}^*$ is a Cholesky decomposition or $\mathbf{H} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^*$ is an eigenvalue decomposition, then $\mathbf{S} = \mathbf{L}^*$ or $\mathbf{S} = \sqrt{\mathbf{\Lambda}}\mathbf{R}^*$ can be chosen. Let $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}\mathbf{S}^{-1}$ and let $\tilde{\mathbf{A}} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^*$ be a singular value decomposition. Moreover, let

$$\tilde{\mathbf{U}} = \mathbf{P}\mathbf{Q}^*, \quad \tilde{\mathbf{M}} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^* \text{ and } \mathbf{U} = \mathbf{S}^{-1}\tilde{\mathbf{U}}\mathbf{S}, \quad \mathbf{M} = \mathbf{S}^{-1}\tilde{\mathbf{M}}\mathbf{S}.$$

Then $\tilde{\mathbf{U}}\tilde{\mathbf{M}}$ obviously is an ordinary polar decomposition of $\tilde{\mathbf{A}}$ and $\mathbf{UM}$ is an H-polar decomposition of $\mathbf{A}$ as a simple verification shows. □

The proof directly provides a numerical method. This suggests that in the case of an indefinite matrix $\mathbf{H}$ a similar approach can be adopted to compute an H-polar decomposition. However, there are currently no numerical methods for the computation of H-singular value decompositions available, although a related theory is already contained in [BR]. In Section 8 thereof as in Section 3.2 the statements regarding H-polar decomposition are derived from the canonical form of the pair $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H})$. This suggests the following steps for computing an H-polar decomposition:

1. Compute the canonical form of the pair $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H})$,
2. Compute an H-selfadjoint matrix $\mathbf{M}$ such that $\mathbf{M}^2 = \mathbf{A}^{[*]}\mathbf{A}$ and $\ker \mathbf{M} = \ker \mathbf{A}$,
3. Compute an H-isometry $\mathbf{U}$ such that $\mathbf{A} = \mathbf{UM}$.

In the following section we will specify a corresponding numerical method for the case of a complex matrix $\mathbf{A}$ for which $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable. The necessary preparations, contained in this section, begin with the description of a simplified canonical form of the pair $(\mathbf{A}, \mathbf{H})$ where $\mathbf{A}$ is H-Hermitian and diagonalisable. This form is based on the following facts taken from [GLR, Chapter I.2.2]:

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian. Then for every non-real eigenvalue $\lambda \in \sigma(\mathbf{A})$ also $\overline{\lambda} \in \sigma(\mathbf{A})$ and the Jordan structures of both eigenvalues are equal. Let

$$E_A(\lambda) = \{\mathbf{x} \in \mathbb{C}^n : (\mathbf{A} - \lambda\mathbf{I})^k\mathbf{x} = \mathbf{0} \text{ for a } k \in \mathbb{N}\}$$

be the generalised eigenspace for the eigenvalue λ. Moreover, let $\lambda_1, \ldots, \lambda_r$ be the real and $\lambda_{r+1}, \ldots, \lambda_s$ be the non-real eigenvalues with positive imaginary parts, and let

$$\begin{aligned} &X_i = E_A(\lambda_i) \text{ for } 1 \le i \le r \text{ and} \\ &X_i = X_{i,1} \oplus X_{i,2} = E_A(\lambda_i) \oplus E_A(\overline{\lambda}_i) \text{ for } r+1 \le i \le s. \end{aligned}$$

Then $\mathbb{C}^n$ can be decomposed as the direct sum of the non-degenerate eigenspaces

$$\mathbb{C}^n = X_1 \oplus \ldots \oplus X_r \oplus X_{r+1} \oplus \ldots \oplus X_s$$

and the following equations hold

$$\begin{aligned} &[\mathbf{x}_k, \mathbf{x}_l] = 0 \text{ for } \mathbf{x}_k \in X_k, \ \mathbf{x}_l \in X_l \text{ and } 1 \le k \ne l \le s, \\ &[\mathbf{x}_k, \mathbf{y}_k] = 0 \text{ for } \mathbf{x}_k, \mathbf{y}_k \in X_{k,1} \text{ or } \mathbf{x}_k, \mathbf{y}_k \in X_{k,2} \text{ and } r+1 \le k \le s. \end{aligned}$$

Therefore, if $\mathbf{R} \in \mathbb{C}^{n\times n}$ is a matrix whose columns are bases of the subspaces X_i, then $\mathbf{R}$ is nonsingular and

$$\begin{aligned} \mathbf{R}^{-1}\mathbf{AR} &= \mathbf{A}_1 \oplus \ldots \oplus \mathbf{A}_r \oplus \begin{bmatrix} \mathbf{A}_{r+1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{r+1,2} \end{bmatrix} \oplus \ldots \oplus \begin{bmatrix} \mathbf{A}_{s1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{s2} \end{bmatrix}, \\ \mathbf{R}^*\mathbf{HR} &= \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_r \oplus \begin{bmatrix} \mathbf{0} & \mathbf{H}_{r+1} \\ \mathbf{H}_{r+1}^* & \mathbf{0} \end{bmatrix} \oplus \ldots \oplus \begin{bmatrix} \mathbf{0} & \mathbf{H}_s \\ \mathbf{H}_s^* & \mathbf{0} \end{bmatrix}. \end{aligned} \tag{3.3}$$

Here the sizes of the blocks are given by $p_i = \dim X_i$ for $1 \le i \le r$ and $p_i = \dim X_{i,1} = \dim X_{i,2}$ for $r+1 \le i \le s$. In the special case of a diagonalisable

matrix $\mathbf{A}$ the generalised eigenspaces contain exclusively eigenvectors and we obtain the simplified form

$$\mathbf{R}^{-1}\mathbf{A}\mathbf{R} = \lambda_1\mathbf{I}_{p_1} \oplus \ldots \oplus \lambda_r\mathbf{I}_{p_r} \oplus \begin{bmatrix} \lambda_{r+1}\mathbf{I}_{p_{r+1}} & \mathbf{0} \\ \mathbf{0} & \overline{\lambda}_{r+1}\mathbf{I}_{p_{r+1}} \end{bmatrix} \oplus \ldots \oplus \begin{bmatrix} \lambda_s\mathbf{I}_{p_s} & \mathbf{0} \\ \mathbf{0} & \overline{\lambda}_s\mathbf{I}_{p_s} \end{bmatrix},$$
$$\mathbf{R}^*\mathbf{H}\mathbf{R} = \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_r \oplus \begin{bmatrix} \mathbf{0} & \mathbf{H}_{r+1} \\ \mathbf{H}_{r+1}^* & \mathbf{0} \end{bmatrix} \oplus \ldots \oplus \begin{bmatrix} \mathbf{0} & \mathbf{H}_s \\ \mathbf{H}_s^* & \mathbf{0} \end{bmatrix}. \tag{3.4}$$

The following theorem is easily derived from this representation. Its proof will be given in two ways in order to, on the one hand, show the connection to Theorem 3.1 and, on the other hand, to provide the foundation for a corresponding numerical method.

Theorem 3.19 (Simplified canonical form). *Let $\mathbf{H} \in \mathbb{C}^{n\times n}$ be nonsingular and Hermitian and let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian and diagonalisable. Then there exists a nonsingular matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ such that*

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{A}_1 \oplus \ldots \oplus \mathbf{A}_k \quad \text{and} \quad \mathbf{S}^*\mathbf{H}\mathbf{S} = \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_k, \tag{3.5a}$$

where the blocks $\mathbf{A}_j$ and $\mathbf{H}_j$ are of equal size and the pairs $(\mathbf{A}_j, \mathbf{H}_j)$ have one and one only of the following forms:

1. *Pairs belonging to real eigenvalues.*

$$\mathbf{A}_j = \lambda\mathbf{I}_p \quad \text{and} \quad \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_{p-q} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_q \end{bmatrix} \tag{3.5b}$$

   with $\lambda \in \mathbb{R}$ and $p, q \in \mathbb{N}$, $q \leq p$.

2. *Pairs belonging to non-real eigenvalues*

$$\mathbf{A}_j = \begin{bmatrix} \lambda\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \overline{\lambda}\mathbf{I}_p \end{bmatrix} \quad \text{and} \quad \mathbf{H}_j = \begin{bmatrix} \mathbf{0} & \mathbf{I}_p \\ \mathbf{I}_p & \mathbf{0} \end{bmatrix} \tag{3.5c}$$

   with $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $\mathrm{Im}(\lambda) > 0$ and $p \in \mathbb{N}$.

Moreover, the simplified canonical form $(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S})$ of $(\mathbf{A}, \mathbf{H})$ is uniquely determined up to the permutation of blocks.

First proof. According to Theorem 3.1 a nonsingular matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ exists such that the pair $(\mathbf{A}, \mathbf{H})$ is in the canonical form (3.2). Because of the assumed diagonalisability, the size of the blocks appearing therein is always 1. Now combining all p blocks of the form (3.2b) or (3.2c) which belong to the same eigenvalue $\lambda \in \mathbb{R}$ or $\lambda \in \mathbb{C}\backslash\mathbb{R}$, respectively, then after a suitable permutation it is always possible to build one block of the form (3.5b) or (3.5c). From $p - q$ blocks of (3.2b) with $\varepsilon = +1$ and q blocks of (3.2b) with $\varepsilon = -1$ this gives one block of the form (3.5b).

Second proof. For all real eigenvalues $\lambda_\rho \in \sigma(\mathbf{A})$, $1 \leq \rho \leq r$, let $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}_\rho$ be an H-orthonormalised (according to Theorem 2.6) basis of eigenvectors of $E_A(\lambda_\rho)$, ordered such that $(\mathbf{H}\mathbf{u}_j, \mathbf{u}_j) = 1$ for $1 \leq j \leq p - q$ and $(\mathbf{H}\mathbf{u}_j, \mathbf{u}_j) = -1$ for $p - q + 1 \leq j \leq p$. For all non-real eigenvalues $\lambda_\sigma, \overline{\lambda}_\sigma \in \sigma(\mathbf{A})$, $r + 1 \leq \sigma \leq s$, let $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}_\sigma$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}_\sigma$ be two H-orthonormalised (according to Theorem 2.7) bases of eigenvectors of $E_A(\lambda_\sigma)$ and $E_A(\overline{\lambda}_\sigma)$. Now by combining these bases as columns of the matrix $\mathbf{S}$, the pair $(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S})$ takes on the assumed form. □

A useful application of this theorem is the following corollary which is also stated in [MMX, Corollary 2.4].

Corollary 3.20 (Non-defective matrix pencils). *Let $\rho\mathbf{H} - \mathbf{G} \in \mathbb{C}^{n\times n}$ be a non-defective Hermitian matrix pencil where both $\mathbf{H}$ and $\mathbf{G}$ are nonsingular. Then there exists a nonsingular matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ such that*

$$\begin{aligned}
\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{G}\mathbf{S} &= \Bigg(\bigoplus_{j=1}^{r} \lambda_j \mathbf{I}_{p_j}\Bigg) \oplus \Bigg(\bigoplus_{j=r+1}^{s} \begin{bmatrix} \lambda_j \mathbf{I}_{p_j} & \\ & \overline{\lambda}_j \mathbf{I}_{p_j} \end{bmatrix}\Bigg),\\
\mathbf{S}^{*}\mathbf{H}\mathbf{S} &= \Bigg(\bigoplus_{j=1}^{r} \begin{bmatrix} \mathbf{I}_{p_j - q_j} & \\ & -\mathbf{I}_{q_j} \end{bmatrix}\Bigg) \oplus \Bigg(\bigoplus_{j=r+1}^{s} \begin{bmatrix} & \mathbf{I}_{p_j} \\ \mathbf{I}_{p_j} & \end{bmatrix}\Bigg),\\
\mathbf{S}^{*}\mathbf{G}\mathbf{S} &= \Bigg(\bigoplus_{j=1}^{r} \lambda_j \begin{bmatrix} \mathbf{I}_{p_j - q_j} & \\ & -\mathbf{I}_{q_j} \end{bmatrix}\Bigg) \oplus \Bigg(\bigoplus_{j=r+1}^{s} \begin{bmatrix} & \overline{\lambda}_j \mathbf{I}_{p_j} \\ \lambda_j \mathbf{I}_{p_j} & \end{bmatrix}\Bigg)
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_r \in \mathbb{R}\backslash\{0\}$ and $\lambda_{r+1}, \ldots, \lambda_s \in \mathbb{C}\backslash\mathbb{R}$.

Proof. Since the pencil $\rho\mathbf{H} - \mathbf{G}$ is non-defective by definition there exist nonsingular matrices $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n\times n}$ such that both

$$\mathbf{\Lambda}_H = \mathbf{P}^{-1}\mathbf{H}\mathbf{Q} \text{ and } \mathbf{\Lambda}_G = \mathbf{P}^{-1}\mathbf{G}\mathbf{Q}$$

are diagonal [MMX, Definition 1.3]. Thus the matrix $\mathbf{H}^{-1}\mathbf{G} = \mathbf{Q}\mathbf{\Lambda}_H^{-1}\mathbf{\Lambda}_G\mathbf{Q}^{-1}$ is diagonalisable and because $(\mathbf{H}^{-1}\mathbf{G})^{*}\mathbf{H} = \mathbf{H}(\mathbf{H}^{-1}\mathbf{G})$ it is H-Hermitian. The assumption follows by application of Theorem 3.19. □

The criteria for the existence of H-Hermitian square roots of a diagonalisable matrix $\mathbf{A}^{[*]}\mathbf{A}$ can also be simplified. The following theorem contains a corresponding specialisation of Theorem 3.4. Its somewhat longer proof is presented completely in order to derive a numerical method therefrom. Furthermore, it corrects the error made in the proof of [BMRRR1, Theorem 4.4].

Theorem 3.21 (Existence of H-Hermitian square roots). *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ and let $\mathbf{B} = \mathbf{A}^{[*]}\mathbf{A}$ be diagonalisable. Then there exists an H-Hermitian matrix $\mathbf{M}$ such that $\mathbf{M}^2 = \mathbf{B}$ and $\ker\mathbf{M} = \ker\mathbf{A}$ if and only if the following conditions are satisfied:*

1. *The part of the (simplified) canonical form of the pair $(\mathbf{B}, \mathbf{H})$ belonging to negative real eigenvalues $\lambda = -\alpha^2$ consists of blocks of the form*

$$\mathbf{B}_j = -\alpha^2\mathbf{I}_{2p} \text{ and } \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_p \end{bmatrix}$$

   with $\alpha > 0$ and $p \in \mathbb{N}$.

2. *The part of the (simplified) canonical form of the pair $(\mathbf{B}, \mathbf{H})$ belonging to the eigenvalue 0 consists of the blocks*

$$\mathbf{B}_j = \mathbf{0}_p \text{ and } \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_{r+s} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{r+t} \end{bmatrix}$$

   with $p, r, s, t \in \mathbb{N}$, $2r + s + t = p$. Moreover, there exists a basis

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_{r+s}, \mathbf{f}_1, \ldots, \mathbf{f}_{r+t}\}$$

in which these blocks appear and for which

$$\begin{aligned}\ker \mathbf{A} = \operatorname{span}\{\mathbf{e}_1 + \mathbf{f}_1, \ldots, \mathbf{e}_r + \mathbf{f}_r\} \\ \oplus \operatorname{span}\{\mathbf{e}_{r+1}, \ldots, \mathbf{e}_{r+s}\} \oplus \operatorname{span}\{\mathbf{f}_{r+1}, \ldots, \mathbf{f}_{r+t}\}.\end{aligned}$$

Proof. [⇒]: Let $\mathbf{B} \in \mathbb{C}^{n\times n}$ be a diagonalisable matrix and let

$$\mathbf{R}^{-1}\mathbf{B}\mathbf{R} = \lambda_1\mathbf{I}_{p_1} \oplus \ldots \oplus \lambda_k\mathbf{I}_{p_k}$$

be its Jordan normal form, so that the columns of $\mathbf{R}$ form a basis of $\mathbb{C}^n$ consisting of eigenvectors of $\mathbf{B}$. Then every matrix $\mathbf{M}$ such that $\mathbf{M}^2 = \mathbf{B}$ can be expressed in the form

$$\mathbf{M} = \mathbf{R}(\sqrt{\lambda_1\mathbf{I}_{p_1}} \oplus \ldots \oplus \sqrt{\lambda_k\mathbf{I}_{p_k}})\mathbf{R}^{-1}$$

where

$$\begin{aligned}\sqrt{\lambda\mathbf{I}_p} &= \mathbf{X}_p(\sqrt{\lambda}\mathbf{I}_{p-q} \oplus -\sqrt{\lambda}\mathbf{I}_q)\mathbf{X}_p^{-1} \text{ if } \lambda \neq 0, \\ \sqrt{\mathbf{0}_p} &= \mathbf{X}_p\left(\begin{bmatrix} & \mathbf{I}_r \\ \mathbf{0}_r & \end{bmatrix} \oplus \mathbf{0}_{p-2r}\right)\mathbf{X}_p^{-1} \text{ if } \lambda = 0,\end{aligned}$$

and $\mathbf{X}_p \in \mathbb{C}^{p\times p}$ denotes an arbitrary nonsingular matrix. The simplified canonical form of an H-Hermitian matrix $\mathbf{M}$ whose square is diagonalisable

$$\mathbf{R}^{-1}\mathbf{M}\mathbf{R} = \mathbf{M}_1 \oplus \ldots \oplus \mathbf{M}_k, \quad \mathbf{R}^*\mathbf{H}\mathbf{R} = \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_k$$

therefore consists of blocks $(\mathbf{M}_j, \mathbf{H}_j)$ of the form

$$\begin{aligned}&\left(\begin{bmatrix}\lambda\mathbf{I}_p & \\ & \overline{\lambda}\mathbf{I}_p\end{bmatrix}, \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix}\right) && \text{if } \lambda \in \mathbb{C}\backslash\mathbb{R}, \\ &\left(\lambda\mathbf{I}_{p+q}\ ,\ \mathbf{I}_p \oplus -\mathbf{I}_q\right) && \text{if } \lambda \in \mathbb{R}\backslash\{0\}, \\ &\left(\begin{bmatrix} & \mathbf{I}_{p+q} \\ \mathbf{0}_{p+q} & \end{bmatrix} \oplus \mathbf{0}_{s+t}, \begin{bmatrix} & \mathbf{I}_p \oplus -\mathbf{I}_q \\ \mathbf{I}_p \oplus -\mathbf{I}_q & \end{bmatrix} \oplus \mathbf{I}_s \oplus -\mathbf{I}_t\right) && \text{if } \lambda = 0,\end{aligned}$$

where the blocks belonging to the eigenvalue 0 have been combined in evident manner to the ordinary canonical form

$$\begin{aligned}\mathbf{M}_0 &= \left(\bigoplus_{i=1}^{p+q}\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\right) \oplus \left(\bigoplus_{i=1}^{s+t}[0]\right), \\ \mathbf{H}_0 &= \left(\bigoplus_{i=1}^{p}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \oplus \bigoplus_{i=1}^{q}\begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix}\right) \oplus \left(\bigoplus_{i=1}^{s}[1] \oplus \bigoplus_{i=1}^{t}[-1]\right).\end{aligned}$$

Now, let $(\mathbf{M}, \mathbf{H})$ be such a simplified canonical form and let $\{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$ be the basis in which it exists. Then the simplified canonical form of the pair $(\mathbf{M}^2, \mathbf{H})$ can be constructed as follows:

(i) If $\lambda \in \mathbb{C}\backslash\mathbb{R} \cup i\mathbb{R}$ and if the canonical form of $(\mathbf{M}, \mathbf{H})$ contains the blocks

$$\begin{aligned}\mathbf{M}_{j_1} \oplus \mathbf{M}_{j_2} &= \begin{bmatrix}\lambda\mathbf{I}_{p_1} & \\ & \overline{\lambda}\mathbf{I}_{p_1}\end{bmatrix} \oplus \begin{bmatrix}-\lambda\mathbf{I}_{p_2} & \\ & -\overline{\lambda}\mathbf{I}_{p_2}\end{bmatrix}, \\ \mathbf{H}_{j_1} \oplus \mathbf{H}_{j_2} &= \begin{bmatrix} & \mathbf{I}_{p_1} \\ \mathbf{I}_{p_1} & \end{bmatrix} \oplus \begin{bmatrix} & \mathbf{I}_{p_2} \\ \mathbf{I}_{p_2} & \end{bmatrix},\end{aligned}$$

then the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a pair of blocks of the form

$$\mathbf{M}_j^2 = \begin{bmatrix} \lambda^2 \mathbf{I}_{p_1+p_2} & \\ & \overline{\lambda}^2 \mathbf{I}_{p_1+p_2} \end{bmatrix}, \ \mathbf{H}_j = \begin{bmatrix} & \mathbf{I}_{p_1+p_2} \\ \mathbf{I}_{p_1+p_2} & \end{bmatrix}.$$

(ii) If $\lambda \in \mathbb{R}\backslash\{0\}$ and if the canonical form of $(\mathbf{M}, \mathbf{H})$ contains the blocks

$$\begin{aligned} \mathbf{M}_{j_1} \oplus \mathbf{M}_{j_2} &= \lambda \mathbf{I}_{p_1+q_1} \oplus -\lambda \mathbf{I}_{p_2+q_2}, \\ \mathbf{H}_{j_1} \oplus \mathbf{H}_{j_2} &= (\mathbf{I}_{p_1} \oplus -\mathbf{I}_{q_1}) \oplus (\mathbf{I}_{p_2} \oplus -\mathbf{I}_{q_2}), \end{aligned}$$

then the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a pair of blocks of the form

$$\mathbf{M}_j^2 = \lambda^2 \mathbf{I}_{(p_1+p_2)+(q_1+q_2)}, \ \mathbf{H}_j = \mathbf{I}_{p_1+p_2} \oplus -\mathbf{I}_{q_1+q_2}.$$

(iii) If $\lambda = i\alpha \in i\mathbb{R}\backslash\{0\}$ and if the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a pair of blocks

$$\mathbf{M}_j = \begin{bmatrix} i\alpha \mathbf{I}_p & \\ & -i\alpha \mathbf{I}_p \end{bmatrix}, \ \mathbf{H}_j = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix},$$

then the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a pair of blocks

$$\mathbf{M}_j^2 = \begin{bmatrix} -\alpha^2 \mathbf{I}_p & \\ & -\alpha^2 \mathbf{I}_p \end{bmatrix} \text{ and } \mathbf{H}_j.$$

If a new basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_p, \mathbf{f}_1, \ldots, \mathbf{f}_p\}$ is now chosen with

$$\mathbf{e}_k = \frac{1}{\sqrt{2}}(\mathbf{g}_k + \mathbf{g}_{k+p}), \ \ \mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{g}_k - \mathbf{g}_{k+p}) \ \text{ for } \ 1 \le k \le p,$$

then

$$(\mathbf{H}_j \mathbf{e}_k, \mathbf{e}_l) = \delta_{kl}, \ \ (\mathbf{H}_j \mathbf{e}_k, \mathbf{f}_l) = 0, \ \ (\mathbf{H}_j \mathbf{f}_k, \mathbf{f}_l) = -\delta_{kl} \ \text{ for } \ 1 \le k, l \le p$$

and the following blocks appear

$$\tilde{\mathbf{M}}_j^2 = \mathbf{M}_j^2 \ \text{ and } \ \tilde{\mathbf{H}}_j = \begin{bmatrix} \mathbf{I}_p & \\ & -\mathbf{I}_p \end{bmatrix}.$$

(iv) If $\lambda = 0$ and if the canonical form of $(\mathbf{M}, \mathbf{H})$ contains a pair of blocks

$$\mathbf{M}_j = \begin{bmatrix} & \mathbf{I}_{p+q} \\ \mathbf{0}_{p+q} & \end{bmatrix} \oplus \mathbf{0}_{s+t}, \ \ \mathbf{H}_j = \begin{bmatrix} & \mathbf{I}_p \oplus -\mathbf{I}_q \\ \mathbf{I}_p \oplus -\mathbf{I}_q & \end{bmatrix} \oplus \mathbf{I}_s \oplus -\mathbf{I}_t,$$

then the canonical form of $(\mathbf{M}^2, \mathbf{H})$ contains a pair of blocks

$$\mathbf{M}_j^2 = \mathbf{0}_{2p+2q+s+t} \ \text{ and } \ \mathbf{H}_j.$$

If a new basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_{p+q+s}, \mathbf{f}_1, \ldots, \mathbf{f}_{p+q+t}\}$ is now chosen with

$$\begin{aligned} \mathbf{e}_k &= \frac{1}{\sqrt{2}}(\mathbf{g}_k + \mathbf{g}_{k+p+q}), \quad \mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{g}_k - \mathbf{g}_{k+p+q}) \quad \text{for } \ 1 \le k \le p, \\ \mathbf{e}_k &= \frac{1}{\sqrt{2}}(\mathbf{g}_k - \mathbf{g}_{k+p+q}), \quad \mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{g}_k + \mathbf{g}_{k+p+q}) \quad \text{for } \ p+1 \le k \le p+q, \\ \mathbf{e}_{k+p+q} &= \mathbf{g}_{k+2p+2q} \qquad \text{for } \ 1 \le k \le s, \\ \mathbf{f}_{k+p+q} &= \mathbf{g}_{k+2p+2q+s} \quad \text{for } \ 1 \le k \le t, \end{aligned}$$

then

$$(\mathbf{H}_j\mathbf{e}_k, \mathbf{e}_l) = \delta_{kl}, \;\; (\mathbf{H}_j\mathbf{e}_k, \mathbf{f}_\nu) = 0, \;\; (\mathbf{H}_j\mathbf{f}_\mu, \mathbf{f}_\nu) = -\delta_{\mu v}$$
$$\text{for } \; 1 \leq k, l \leq p+q+s \;\text{ and }\; 1 \leq \mu, \nu \leq p+q+t$$

and the following blocks appear

$$\tilde{\mathbf{M}}_j^2 = \mathbf{M}_j^2 \;\text{ and }\; \tilde{\mathbf{H}}_j = \begin{bmatrix} \mathbf{I}_{p+q+s} & \\ & -\mathbf{I}_{p+q+t} \end{bmatrix}.$$

Moreover, it is true that

$$\begin{aligned} \ker \mathbf{M} &= \operatorname{span}\{\mathbf{g}_1, \ldots, \mathbf{g}_{p+q}, \mathbf{g}_{2p+2q+1}, \ldots, \mathbf{g}_{2p+2q+s+t}\} \\ &= \operatorname{span}\{\mathbf{e}_1 + \mathbf{f}_1, \ldots, \mathbf{e}_{p+q} + \mathbf{f}_{p+q}\} \\ &\oplus \operatorname{span}\{\mathbf{e}_{p+q+1}, \ldots, \mathbf{e}_{p+q+s}\} \oplus \operatorname{span}\{\mathbf{f}_{p+q+1}, \ldots, \mathbf{f}_{p+q+t}\}. \end{aligned}$$

[$\Leftarrow$]: Let $\mathbf{B}$ be diagonalisable and let $\mathbf{R}^{-1}\mathbf{B}\mathbf{R} = \mathbf{B}_1 \oplus \ldots \oplus \mathbf{B}_k$, $\mathbf{R}^*\mathbf{H}\mathbf{R} = \mathbf{H}_1 \oplus \ldots \oplus \mathbf{H}_k$ be the (simplified) canonical form of the pair $(\mathbf{B}, \mathbf{H}) = (\mathbf{M}^2, \mathbf{H})$. Furthermore, let the conditions 1. and 2. be satisfied, and let $\mathbf{\Sigma}_k$ be $k \times k$ diagonal matrices with diagonal elements in $\{+1, -1\}$. Then the matrix $\mathbf{M}$ can be constructed as follows:

(i) If $\lambda = \omega^2 \in \mathbb{C}\backslash\mathbb{R}$ and if the canonical form contains a pair of blocks of the form

$$\mathbf{B}_j = \begin{bmatrix} \lambda\mathbf{I}_p & \\ & \overline{\lambda}\mathbf{I}_p \end{bmatrix} \;\text{ and }\; \mathbf{H}_j = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix},$$

then for

$$\mathbf{M}_j = \begin{bmatrix} \omega\mathbf{\Sigma}_p & \\ & \overline{\omega}\mathbf{\Sigma}_p \end{bmatrix}$$

the equations $\mathbf{M}_j^2 = \mathbf{B}_j$ and $\mathbf{M}_j^*\mathbf{H}_j = \mathbf{H}_j\mathbf{M}_j$ are satisfied.

(ii) If $\lambda \in \mathbb{R} \cap (0, \infty)$ and if the canonical form contains a pair of blocks of the form

$$\mathbf{B}_j = \lambda\mathbf{I}_p \;\text{ and }\; \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_{p-q} & \\ & -\mathbf{I}_q \end{bmatrix},$$

then for

$$\mathbf{M}_j = \sqrt{\lambda}\mathbf{\Sigma}_p$$

the equations $\mathbf{M}_j^2 = \mathbf{B}_j$ and $\mathbf{M}_j^*\mathbf{H}_j = \mathbf{H}_j\mathbf{M}_j$ are satisfied.

(iii) If $\lambda = -\alpha^2 \in \mathbb{R} \cap (-\infty, 0)$ and if the canonical form contains a pair of blocks of the form

$$\mathbf{B}_j = \begin{bmatrix} -\alpha^2\mathbf{I}_p & \\ & -\alpha^2\mathbf{I}_p \end{bmatrix} \;\text{ and }\; \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_p & \\ & -\mathbf{I}_p \end{bmatrix},$$

then for

$$\tilde{\mathbf{M}}_j = \begin{bmatrix} i\alpha\mathbf{\Sigma}_p & \\ & -i\alpha\mathbf{\Sigma}_p \end{bmatrix}, \;\; \tilde{\mathbf{H}}_j = \begin{bmatrix} & \mathbf{I}_p \\ \mathbf{I}_p & \end{bmatrix} \;\text{ and }\; \mathbf{S}_j = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_p & \mathbf{I}_p \\ \mathbf{I}_p & -\mathbf{I}_p \end{bmatrix}$$

the equations $\mathbf{S}_j^{-1}\mathbf{B}_j\mathbf{S}_j = \mathbf{B}_j = \tilde{\mathbf{M}}_j^2$ and $\mathbf{S}_j^*\mathbf{H}_j\mathbf{S}_j = \tilde{\mathbf{H}}_j$ are satisfied ($\mathbf{S}_j^{-1} = \mathbf{S}_j^* = \mathbf{S}_j$). Therefore by setting

$$\mathbf{M}_j = \mathbf{S}_j\tilde{\mathbf{M}}_j\mathbf{S}_j^{-1} = \begin{bmatrix} & i\alpha\mathbf{\Sigma}_p \\ i\alpha\mathbf{\Sigma}_p & \end{bmatrix},$$

we obtain $\mathbf{M}_j^2 = \mathbf{B}_j$ and $\mathbf{M}_j^*\mathbf{H}_j = \mathbf{H}_j\mathbf{M}_j$.

(iv) If $\lambda = 0$, then by arranging the basis vectors in the order

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_r, \mathbf{e}_{r+1}, \ldots, \mathbf{e}_{r+s}, \mathbf{f}_{r+1}, \ldots, \mathbf{f}_{r+t}\}$$

it is always possible to achieve that the blocks $\mathbf{B}_j$ and $\mathbf{H}_j$ exist in the form

$$\mathbf{B}_j = \mathbf{0}_p \text{ and } \mathbf{H}_j = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_r \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_t \end{bmatrix}$$

and for

$$\tilde{\mathbf{M}}_j = \begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_r \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \oplus \mathbf{0}_{s+t}, \quad \tilde{\mathbf{H}}_j = \begin{bmatrix} \mathbf{0} & \mathbf{I}_r \\ \mathbf{I}_r & \mathbf{0} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_t \end{bmatrix}$$
$$\text{and } \mathbf{S}_j = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_r & \mathbf{I}_r \\ \mathbf{I}_r & -\mathbf{I}_r \end{bmatrix} \oplus \mathbf{I}_{s+t}$$

the equations $\mathbf{S}_j^{-1}\mathbf{B}_j\mathbf{S}_j = \mathbf{B}_j = \tilde{\mathbf{M}}_j^2$ and $\mathbf{S}_j^*\mathbf{H}_j\mathbf{S}_j = \tilde{\mathbf{H}}_j$ are satisfied ($\mathbf{S}_j = \mathbf{S}_j^* = \mathbf{S}_j^{-1}$). Therefore by setting

$$\mathbf{M}_j = \mathbf{S}_j\tilde{\mathbf{M}}_j\mathbf{S}_j^{-1} = \frac{1}{2} \begin{bmatrix} \mathbf{\Sigma}_r & -\mathbf{\Sigma}_r \\ \mathbf{\Sigma}_r & -\mathbf{\Sigma}_r \end{bmatrix} \oplus \mathbf{0}_{s+t}$$

we obtain $\mathbf{M}_j^2 = \mathbf{B}_j$ and $\mathbf{M}_j^*\mathbf{H}_j = \mathbf{H}_j\mathbf{M}_j$ and, moreover,

$$\ker \mathbf{M} = \operatorname{span}\{\mathbf{e}_1 + \mathbf{f}_1, \ldots, \mathbf{e}_r + \mathbf{f}_r, \mathbf{e}_{r+1}, \ldots, \mathbf{e}_{r+s}, \mathbf{f}_{r+1}, \ldots, \mathbf{f}_{r+t}\} = \ker \mathbf{A}. \qquad \square$$

Using the notation of Theorem 3.4, the part of the canonical form of the pair $(\mathbf{B}, \mathbf{H}) = (\mathbf{M}^2, \mathbf{H})$ belonging to the eigenvalue 0 in the basis

$$\{\mathbf{e}_1, \mathbf{f}_1, \ldots, \mathbf{e}_r, \mathbf{f}_r, \mathbf{e}_{r+1}, \ldots, \mathbf{e}_{r+s}, \mathbf{f}_{r+1}, \ldots, \mathbf{f}_{r+t}\}$$

can be expressed as

$$\mathbf{J}^{(1)} \oplus \mathbf{J}^{(3)} = \left(\bigoplus_{i=1}^{r} [\mathbf{N}_1 \oplus \mathbf{N}_1]\right) \oplus \left(\bigoplus_{j=1}^{s} 0 \oplus \bigoplus_{k=1}^{t} 0\right),$$
$$\mathbf{Z}^{(1)} \oplus \mathbf{Z}^{(3)} = \left(\bigoplus_{i=1}^{r} [\mathbf{Z}_1 \oplus -\mathbf{Z}_1]\right) \oplus \left(\bigoplus_{j=1}^{s} 1 \oplus \bigoplus_{k=1}^{t} -1\right).$$

This confirms again the correction of Theorem 3.4 explained with Example 3.5. Furthermore, the diagonal matrices $\mathbf{\Sigma}_k$ used in the proof comply with the relationships between the canonical forms of the pairs $(\mathbf{M}^2, \mathbf{H})$ and $(\mathbf{M}, \mathbf{H})$ listed in Theorem 3.6. Using a particular choice of these sign matrices we can easily derive a statement concerning the class of H-polar decompositions described in the following definition [BMRRR2, Section 5].

Definition 3.22 (Semidefinite H-polar decompositions). A matrix $\mathbf{M}$ is said to be H-nonnegative, if $\mathbf{HM}$ is positive semidefinite. The particular H-polar decompositions in which the matrix $\mathbf{M}$ is H-nonnegative are called *semidefinite* H-polar decompositions. $\lozenge$

Corollary 3.23 (Existence of H-nonnegative square roots). *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and let* $\mathbf{B} = \mathbf{A}^{[*]}\mathbf{A}$. *Then there exists an H-nonnegative matrix* $\mathbf{M}$ *such that* $\mathbf{M}^2 = \mathbf{B}$ *and* $\ker \mathbf{M} = \ker \mathbf{A}$ *if and only if* $\mathbf{B}$ *is diagonalisable with* $\sigma(\mathbf{B}) \subset [0,\infty)$ *and if condition 2. of Theorem 3.21 is satisfied.*

Proof. If the matrix $\mathbf{B}$ from Theorem 3.21 has only non-negative eigenvalues, $\lambda_1,\dots,\lambda_k > 0$, $\lambda_{k+1} = 0$, and if it is assumed that condition 2. holds, then the blocks $\mathbf{M}_j = \sqrt{\lambda_j}\mathbf{\Sigma}_p$ $(1 \le j \le k)$ in case (ii) of the "if" part can be chosen such that $\mathbf{\Sigma}_p = \mathbf{I}_{p-q} \oplus -\mathbf{I}_q = \mathbf{H}_j$ and the block $\tilde{\mathbf{M}}_j$ $(j = k+1)$ in case (iv) can be chosen such that $\mathbf{\Sigma}_r = \mathbf{I}_r$. Then the canonical form of the pair $(\mathbf{M},\mathbf{H})$ has the form

$$\begin{aligned} \mathbf{S}^{-1}\mathbf{M}\mathbf{S} &= \bigoplus_{j=1}^{k}(\omega_j\mathbf{I}_{p_j-q_j} \oplus -\omega_j\mathbf{I}_{q_j}) \oplus \begin{bmatrix} & \mathbf{I}_r \\ \mathbf{0}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_s & \\ & \mathbf{0}_t \end{bmatrix}, \\ \mathbf{S}^{*}\mathbf{H}\mathbf{S} &= \bigoplus_{j=1}^{k}(\mathbf{I}_{p_j-q_j} \oplus -\mathbf{I}_{q_j}) \oplus \begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \\ & -\mathbf{I}_t, \end{bmatrix}, \end{aligned} \tag{3.6}$$

where $\omega_j = \sqrt{\lambda_j} > 0$. Thus

$$\mathbf{S}^{*}\mathbf{H}\mathbf{M}\mathbf{S} = \bigoplus_{j=1}^{k}\omega_j\mathbf{I}_{p_j} \oplus \mathbf{0}_r \oplus \mathbf{I}_r \oplus \mathbf{0}_{s+t},$$

so that the matrix $\mathbf{H}\mathbf{M}$ is positive semidefinite Hermitian.

Conversely, if $\mathbf{M}$ is H-nonnegative, then the canonical form of the pair $(\mathbf{M},\mathbf{H})$ must have the form (3.6) in which some of the blocks may not exist. This implies

$$\mathbf{S}^{-1}\mathbf{M}^2\mathbf{S} = \bigoplus_{j=1}^{k}\omega_j^2\mathbf{I}_{p_j} \oplus \mathbf{0}_{2r+s+t},$$

so that $\mathbf{B} = \mathbf{M}^2$ is diagonalisable and $\sigma(\mathbf{B}) \subset [0,\infty)$. □

A similar statement is also given in [BMRRR2, Theorem 5.3] where the kernel condition is formulated slightly different.

3.5 Numerical computation of H-polar decompositions of a matrix A for which $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable

Now we describe the method for computing H-polar decompositions of a complex matrix $\mathbf{A}$ for which $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable. This method, suggested at the beginning of the previous section, begins with the computation of a simplified canonical form, so that a corresponding algorithm is derived first. Note that this algorithm is a very simple version of the algorithm for computing a general canonical form presented in Chapter 6.

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian and diagonalisable. Then, according to the second proof of Theorem 3.19, the simplified canonical form of the pair $(\mathbf{A},\mathbf{H})$ can be determined with the following steps:

Step 1. Computing the eigenvalues and eigenvectors. First of all the Jordan normal form of $\mathbf{A}$ must be computed

$$\mathbf{R}_1^{-1}\mathbf{A}\mathbf{R}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Here $\lambda_1, \ldots, \lambda_n$ are the eigenvalues and $\mathbf{R}_1 = [\mathbf{r}_1 \ \ldots \ \mathbf{r}_n]$ is a nonsingular matrix consisting of corresponding eigenvectors. For this purpose usually the QR algorithm should be used [F], [GVL, Section 7.5].

Step 2. Grouping the eigenvalues. Now the eigenvalues must be combined in groups of numerical multiple eigenvalues, and the eigenvalues and eigenvectors must be permuted such that eigenvalues belonging to the same group are adjacent

$$\mathbf{R}_2^{-1}\mathbf{A}\mathbf{R}_2 = \lambda_1^*\mathbf{I}_{p_1} \oplus \ldots \oplus \lambda_k^*\mathbf{I}_{p_k}.$$

Here $\lambda_1^*, \ldots, \lambda_k^*$ are the numerical multiple eigenvalues, which are the average eigenvalues of the specified groups, and $\mathbf{R}_2$ is the permuted matrix of eigenvectors.

For this purpose there are several algorithms available: Baveley and Steward use norm estimates [BS], Kågström and Ruhe use Gershgorin circles [KR1], we will use a modified cluster analysis algorithm in Chapter 6. All these algorithms require a user-supplied tolerance parameter, all work well when the eigenvalues are well-separated, and all fail in particular cases. In other words, the grouping of the eigenvalues is an extremely difficult problem for which no absolute reliant method exists. In our implementation, with which the numerical results presented at the end of this section were obtained, we have chosen to apply the following algorithm:

```
k = 0
i = 1
while i < n do
    * Determine the pivot eigenvalue *
    p = i
    for j = i + 1, ..., n do
        if |λ_j| > |λ_p| then
            p = j
        end if
    end for
    if p ≠ i then
        λ_i ↔ λ_p , r_i ↔ r_p
        p = i
    end if
    * Determine the adjacent eigenvalues *
    α = λ_i
    β = α
    δ_rel = max(|α|δ, δ)
    for j = i + 1, ..., n do
        if |λ_j − β| ≤ δ_rel then
            p = p + 1
            λ_j ↔ λ_p , r_j ↔ r_p
            α = α + λ_p
```

```
            β = α/(p − i + 1)
        end if
    end for
    * Update the eigenvalues *
    for j = i, ..., p do
        λ_j = β
    end for
    * Store the block boundary *
    k = k + 1
    b_k = p
    i = p + 1
end while
```

Here $\delta > 0$ is a tolerance parameter with which the relative tolerances δ_{rel} are computed. When the algorithm terminates, k contains the number of groups, the integer array b contains the block boundaries, and the eigenvalues are replaced by the average eigenvalues of the groups.

Step 3. Pairing the eigenvalues. Now the real eigenvalues and pairs of non-real eigenvalues must be determined. A simple approach for this is to use a further (or the same) tolerance parameter δ' and to determine the pairs of non-real eigenvalues by demanding that $|\lambda_i^* - \overline{\lambda_j^*}| \leq \delta'$. However, this is not the most reliant method. More stable results are obtained by considering that the matrix

$$\mathbf{R}_2^* \mathbf{H} \mathbf{R}_2 = \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1k} \\ \vdots & & \vdots \\ \mathbf{C}_{k1} & \cdots & \mathbf{C}_{kk} \end{bmatrix}$$

should theoretically have the form (3.4). Let $\pi(i)$ be indices such that $\mathbf{C}_{i,\pi(i)}$ is the block with the maximum Frobenius norm in row i for $1 \leq i \leq k$. Then, if $\pi(i) = i$, λ_i^* is a real eigenvalue, and if $\pi(i) = j$ and $\pi(j) = i$ for $i \neq j$, $(\lambda_i^*, \lambda_j^*)$ is a pair of non-real eigenvalues. If this classification fails or the block sizes do not satisfy $p_i = p_j$, the algorithm terminates with an error. Possibly it may succeed with a better grouping parameter δ.

Step 4. H-orthonormalising the eigenvectors. Finally, the eigenvectors contained in the blocks $(\mathbf{Y}_1 | \ldots | \mathbf{Y}_k)$ of $\mathbf{R}_2$ must be H-orthonormalised. For this purpose the HQR and HQR-2 decomposition were developed in Section 2.4. Let

$$\mathbf{Y}_i \mathbf{P}_i = \mathbf{Q}_i \mathbf{R}_i \text{ if } \lambda_i^* \in \mathbb{R} \text{ and } \left\{ \begin{matrix} \mathbf{Y}_i \mathbf{P}_i = \mathbf{Q}_i \mathbf{R}_i \\ \mathbf{Y}_j \mathbf{P}_j = \mathbf{Q}_j \mathbf{R}_j \end{matrix} \right\} \text{ if } \lambda_i^* = \overline{\lambda_j^*} \in \mathbb{C} \backslash \mathbb{R}$$

be HQR and HQR-2 decompositions, respectively. Then the matrix $\mathbf{R}_3$ consisting of the blocks $(\mathbf{Q}_1 | \ldots | \mathbf{Q}_k)$ transforms the pair $(\mathbf{A}, \mathbf{H})$ into its simplified canonical form (3.5). Note that the blocks $\mathbf{C}_{ii}$ and $\mathbf{C}_{ij}$ computed in the previous step should be used to initialise the HQR and HQR-2 decompositions, so that they are not computed twice.

If the HQR and HQR-2 decompositions are not available, this step can also

be performed in the following way: Let

$$\mathbf{Y}_i^*\mathbf{H}\mathbf{Y}_i = \mathbf{C}_{ii} = \mathbf{U}_i\mathbf{\Phi}_{ii}\mathbf{U}_i^* \text{ if } \lambda_i^* \in \mathbb{R} \text{ and}$$
$$\mathbf{Y}_i^*\mathbf{H}\mathbf{Y}_j = \mathbf{C}_{ij} = \mathbf{U}_i\mathbf{\Psi}_{ij}\mathbf{V}_j^* \text{ if } \lambda_i^* = \overline{\lambda_j^*} \in \mathbb{C}\backslash\mathbb{R}$$

be eigenvalue and singular value decompositions, respectively. Then $\mathbf{U}_i$ and $\mathbf{V}_j$ are unitary, $\mathbf{\Phi}_{ii}$ is diagonal with non-zero real diagonal elements, and $\mathbf{\Psi}_{ij}$ is diagonal with positive real diagonal elements. Therefore, the matrix $\mathbf{R}_3$ consisting of the blocks

$$\mathbf{Q}_i = \mathbf{Y}_i\mathbf{U}_i|\mathbf{\Phi}_{ii}|^{-1/2} \text{ and } \mathbf{Q}_i = \mathbf{Y}_i\mathbf{U}_i\mathbf{\Psi}_{ij}^{-1/2}, \ \mathbf{Q}_j = \mathbf{Y}_j\mathbf{V}_j\mathbf{\Psi}_{ij}^{-1/2}$$

also transforms the pair $(\mathbf{A}, \mathbf{H})$ into its simplified canonical form. ◊

To simplify the quotation of this algorithm we summarise as follows:

Method 3.24 (Simplified canonical form). *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian and diagonalisable. Then the steps 1 – 4 described above compute the simplified canonical form of the pair $(\mathbf{A}, \mathbf{H})$.*

Now, let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be such that $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable. Then, according to the proof of Theorem 3.21, an H-polar decomposition of $\mathbf{A}$ can be determined with the following steps:

Step 1. Computing the canonical form. First of all the matrix $\mathbf{B} = \mathbf{A}^{[*]}\mathbf{A}$ must be computed by solving the linear system

$$\mathbf{H}\mathbf{B} = \mathbf{A}^*\mathbf{H}\mathbf{A}$$

for $\mathbf{B}$. Then the simplified canonical form of the pair $(\mathbf{B}, \mathbf{H})$ must be determined with Method 3.24. Assume that this form is given by

$$\mathbf{R}^{-1}\mathbf{A}^{[*]}\mathbf{A}\mathbf{R} = \mathbf{J} = \mathbf{J}_3 \oplus \mathbf{J}_2 \oplus \mathbf{J}_1 \oplus \mathbf{J}_0,$$
$$\mathbf{J} = \begin{bmatrix} \omega^2\mathbf{I}_{p_3} & \\ & \overline{\omega}^2\mathbf{I}_{p_3} \end{bmatrix} \oplus \begin{bmatrix} \alpha^2\mathbf{I}_{p_2} & \\ & \alpha^2\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} -\beta^2\mathbf{I}_{p_1} & \\ & -\beta^2\mathbf{I}_{p_1} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_{r+s} & \\ & \mathbf{0}_{r+t} \end{bmatrix},$$
$$\mathbf{R}^*\mathbf{H}\mathbf{R} = \mathbf{Z}_J = \mathbf{Z}_{J,3} \oplus \mathbf{Z}_{J,2} \oplus \mathbf{Z}_{J,1} \oplus \mathbf{Z}_{J,0},$$
$$\mathbf{Z}_J = \begin{bmatrix} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_2} & \\ & -\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_1} & \\ & -\mathbf{I}_{p_1} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{r+s} & \\ & -\mathbf{I}_{r+t} \end{bmatrix},$$
$$\mathbf{R} = (\mathbf{R}_3|\mathbf{R}_2|\mathbf{R}_1|\mathbf{R}_0)$$

where $\omega \in \mathbb{C}\backslash\mathbb{R} \cup i\mathbb{R}$, $0 < \alpha, \beta \in \mathbb{R}$, and the rectangular blocks $\mathbf{R}_j$, $0 \le j \le 3$, correspond to the blocks of $\mathbf{J}$ and $\mathbf{Z}_J$.

Step 2. Computing the H-Hermitian square root. Now an H-Hermitian square root $\mathbf{M}$ of the matrix $\mathbf{A}^{[*]}\mathbf{A}$ must be determined using

$$\mathbf{S}^{-1}\mathbf{M}\mathbf{S} = \mathbf{K} = \mathbf{K}_3 \oplus \mathbf{K}_2 \oplus \mathbf{K}_1 \oplus \mathbf{K}_0,$$
$$\mathbf{K} = \begin{bmatrix} \omega\mathbf{I}_{p_3} & \\ & \overline{\omega}\mathbf{I}_{p_3} \end{bmatrix} \oplus \begin{bmatrix} \alpha\mathbf{I}_{p_2} & \\ & \alpha\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} i\beta\mathbf{I}_{p_1} & \\ & i\beta\mathbf{I}_{p_1} \end{bmatrix} \oplus \left(\begin{bmatrix} & \mathbf{I}_r \\ \mathbf{0}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_s & \\ & \mathbf{0}_t \end{bmatrix} \right),$$

$$\mathbf{S}^*\mathbf{H}\mathbf{S} = \mathbf{Z}_K = \mathbf{Z}_{K,3} \oplus \mathbf{Z}_{K,2} \oplus \mathbf{Z}_{K,1} \oplus \mathbf{Z}_{K,0},$$

$$\mathbf{Z}_K = \begin{bmatrix} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_2} & \\ & -\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_1} & \\ & -\mathbf{I}_{p_1} \end{bmatrix} \oplus \left(\begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \\ & -\mathbf{I}_t \end{bmatrix} \right),$$

$$\mathbf{S} = (\mathbf{R}_3|\mathbf{R}_2|\mathbf{R}_1|\mathbf{R}_0'')$$

where $\mathbf{R}_0''$ is defined by the kernel transformation described below. For the negative eigenvalue $-\beta^2$ this construction is possible only if condition 1. of Theorem 3.21 holds, i.e. only if $\mathbf{R}_1$ consists of an equal number of positive and negative eigenvectors. It would then also be possible to use the blocks

$$\mathbf{K}_1' = \begin{bmatrix} i\beta\mathbf{I}_{p_1} & \\ & -i\beta\mathbf{I}_{p_1} \end{bmatrix}, \quad \mathbf{Z}_{K,1}' = \begin{bmatrix} & \mathbf{I}_{p_1} \\ \mathbf{I}_{p_1} & \end{bmatrix}, \quad \mathbf{R}_1' = \frac{1}{\sqrt{2}}\,\mathbf{R}_1 \begin{bmatrix} \mathbf{I}_{p_1} & \mathbf{I}_{p_1} \\ \mathbf{I}_{p_1} & -\mathbf{I}_{p_1} \end{bmatrix},$$

but this is found to be less convenient when constructing the H-isometry in the third step. Furthermore, in the blocks of $\mathbf{K}$ sign matrices $\mathbf{\Sigma}_{p_3}$, $\mathbf{\Sigma}_{p_2}$, $\mathbf{\Sigma}_{q_2}$, $\mathbf{\Sigma}_{p_1}$, $\mathbf{\Sigma}_r = \mathrm{diag}(\pm 1)$ could also be used instead of the identity matrices. This would then produce another H-polar decomposition of $\mathbf{A}$. For example, if $\mathbf{A}^{[*]}\mathbf{A}$ has only non-negative real eigenvalues, then by Corollary 3.23 a semidefinite H-polar decomposition can be computed this way. Finally, the required treatment of the eigenvalue 0 consists of the following transformation.

Kernel transformation. Let $\mathbf{R}_0 = (\mathbf{R}_+|\mathbf{R}_-)$ be a partitioning where $\mathbf{R}_+$ contains the $r+s$ positive and $\mathbf{R}_-$ contains the $r+t$ negative eigenvectors in $\ker \mathbf{A}^{[*]}\mathbf{A}$, so that

$$\mathbf{R}_+^*\mathbf{H}\mathbf{R}_+ = \mathbf{I}_{r+s} \text{ and } \mathbf{R}_-^*\mathbf{H}\mathbf{R}_- = -\mathbf{I}_{r+t}.$$

Then, if condition 2. of Theorem 3.21 holds, there must exist unitary transformations $\mathbf{T}_+$ and $\mathbf{T}_-$ such that

$$\mathbf{A}\mathbf{R}_+' = [-\mathbf{a}_1 \ldots -\mathbf{a}_r\, \mathbf{0}_1 \ldots \mathbf{0}_s] \text{ and } \mathbf{A}\mathbf{R}_-' = [\mathbf{a}_1 \ldots \mathbf{a}_r\, \mathbf{0}_1 \ldots \mathbf{0}_t]$$

is satisfied for

$$\mathbf{R}_+' = \mathbf{R}_+\mathbf{T}_+ = [\mathbf{e}_1 \ldots \mathbf{e}_{r+s}] \text{ and } \mathbf{R}_-' = \mathbf{R}_-\mathbf{T}_- = [\mathbf{f}_1 \ldots \mathbf{f}_{r+t}].$$

To determine this transformations, let

$$\mathbf{A}\mathbf{R}_+ = \mathbf{U}_+\mathbf{\Sigma}_+\mathbf{V}_+^* \text{ and } \mathbf{A}\mathbf{R}_- = \mathbf{U}_-\mathbf{\Sigma}_-\mathbf{V}_-^*$$

be singular value decompositions. Then the ranks of $\mathbf{\Sigma}_+$ and $\mathbf{\Sigma}_-$ must be equal

$$\begin{aligned} \mathbf{A}\mathbf{R}_+\mathbf{V}_+ = \mathbf{U}_+\mathbf{\Sigma}_+ &= [\mathbf{b}_1 \ldots \mathbf{b}_r\, \mathbf{0}_1 \ldots \mathbf{0}_s], \quad r = \mathrm{rank}(\mathbf{\Sigma}_+), \\ \mathbf{A}\mathbf{R}_-\mathbf{V}_- = \mathbf{U}_-\mathbf{\Sigma}_- &= [\mathbf{a}_1 \ldots \mathbf{a}_r\, \mathbf{0}_1 \ldots \mathbf{0}_t], \quad r = \mathrm{rank}(\mathbf{\Sigma}_-), \end{aligned}$$

and the matrix $\mathbf{W} \in \mathbb{C}^{r\times r}$ defined by $[\mathbf{b}_1 \ldots \mathbf{b}_r]\,\mathbf{W} = -[\mathbf{a}_1 \ldots \mathbf{a}_r]$, i.e.

$$\mathbf{W} = -(\mathbf{\Sigma}_+)_r^{-1}(\mathbf{U}_+^*\mathbf{U}_-)_r(\mathbf{\Sigma}_-)_r \quad (r \times r \text{ submatrices}),$$

must be unitary, to ensure that

$$\mathbf{T}_+ = \mathbf{V}_+(\mathbf{W} \oplus \mathbf{I}_s) \text{ and } \mathbf{T}_- = \mathbf{V}_-$$

exist. Now permuting the columns of $\mathbf{R}_0' = (\mathbf{R}_+'|\mathbf{R}_-')$ into the order

$$\mathbf{R}_0' = [\mathbf{e}_1 \dots \mathbf{e}_r\, \mathbf{f}_1 \dots \mathbf{f}_r\, \mathbf{e}_{r+1} \dots \mathbf{e}_{r+s}\, \mathbf{f}_{r+1} \dots \mathbf{f}_{r+t}]$$

we obtain

$$\begin{aligned} \mathbf{A}\mathbf{R}_0' &= [-\mathbf{a}_1 \dots -\mathbf{a}_r\, \mathbf{a}_1 \dots \mathbf{a}_r\, \mathbf{0}_1 \dots \mathbf{0}_{s+t}] \text{ and} \\ (\mathbf{R}_0')^*\mathbf{H}(\mathbf{R}_0') &= (\mathbf{I}_r \oplus -\mathbf{I}_r) \oplus (\mathbf{I}_s \oplus -\mathbf{I}_t). \end{aligned}$$

Hence, for

$$\mathbf{R}_0'' = \mathbf{R}_0' \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_r & \mathbf{I}_r \\ \mathbf{I}_r & -\mathbf{I}_r \end{bmatrix} \oplus \mathbf{I}_{s+t} \right)$$

it follows that

$$\begin{aligned} \mathbf{A}\mathbf{R}_0'' &= -\sqrt{2}\, [\mathbf{0}_1 \dots \mathbf{0}_r\, \mathbf{a}_1 \dots \mathbf{a}_r\, \mathbf{0}_1 \dots \mathbf{0}_{s+t}] \text{ and} \\ (\mathbf{R}_0'')^*\mathbf{H}(\mathbf{R}_0'') &= \begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \\ & -\mathbf{I}_t \end{bmatrix}, \end{aligned}$$

so that the wanted basis of $\ker \mathbf{A}^{[*]}\mathbf{A}$ has been determined. If the ranks of $\mathbf{\Sigma}_+$ and $\mathbf{\Sigma}_-$ differ or the transformation $\mathbf{W}$ is not unitary, the matrix $\mathbf{A}$ does not have H-polar decompositions.

In an implementation of this transformation the ranks must be computed with the help of a tolerance parameter $\tau > 0$. Suppose that the singular values in $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_m)$ are sorted in descending order. Then $\operatorname{rank}(\mathbf{\Sigma})$ is the smallest index such that $\sigma_r \le \sigma_1\tau$ or 0, if $\sigma_1 \le \tau$. Moreover, $\mathbf{W}$ may be regarded as unitary when $\|\mathbf{W}^*\mathbf{W} - \mathbf{I}\|_F \le \tau$.

Step 3. Computing the H-isometry. After the second step $\mathbf{M} = \mathbf{S}\mathbf{K}\mathbf{S}^{-1}$ is the H-Hermitian factor, and in the nonsingular case, in which the blocks $\mathbf{J}_0$, $\mathbf{Z}_{J,0}$, $\mathbf{K}_0$, $\mathbf{Z}_{K,0}$ do not exist, $\mathbf{U} = \mathbf{A}\mathbf{M}^{-1} = \mathbf{A}\mathbf{S}\mathbf{K}^{-1}\mathbf{S}^{-1}$ is the H-unitary factor of an H-polar decomposition of $\mathbf{A}$. Here the inverse of $\mathbf{S}$ can be computed using $\mathbf{S}^{-1} = \mathbf{Z}_K\mathbf{S}^*\mathbf{H}$ which follows from $\mathbf{S}^*\mathbf{H}\mathbf{S} = \mathbf{Z}_K$. In the singular case let

$$\tilde{\mathbf{K}} = \mathbf{K}_3 \oplus \mathbf{K}_2 \oplus \mathbf{K}_1 \oplus \tilde{\mathbf{K}}_0 \text{ with } \tilde{\mathbf{K}}_0 = \tilde{\mathbf{K}}_0^{-1} = \begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \mathbf{I}_{s+t},$$

and also let

$$\mathbf{S}_0 = (\mathbf{S}_0'|\mathbf{S}_0''|\tilde{\mathbf{S}}_0) \text{ with } \mathbf{S}_0', \mathbf{S}_0'' \in \mathbb{C}^{n\times r} \text{ and } \tilde{\mathbf{S}}_0 \in \mathbb{C}^{n\times(s+t)}$$

be a partitioning of the matrix $\mathbf{S}_0 = \mathbf{R}_0''$. Then we have

$$\mathbf{A}\mathbf{S}_0 = (\mathbf{0}_{n,r}|\mathbf{T}_0'|\mathbf{0}_{n,s+t}) \text{ and } \mathbf{S}_0\mathbf{K}_0 = (\mathbf{0}_{n,r}|\mathbf{S}_0'|\mathbf{0}_{n,s+t})$$

where $\mathbf{T}_0' = -\sqrt{2}\, [\mathbf{a}_1 \dots \mathbf{a}_r]$ by the kernel transformation. Therefore, the matrices $\mathbf{A}\mathbf{S}\tilde{\mathbf{K}}^{-1}$ and $\mathbf{M}\mathbf{S}\tilde{\mathbf{K}}^{-1}$ have the form

$$\begin{aligned} \mathbf{A}\mathbf{S}\tilde{\mathbf{K}}^{-1} &= (\mathbf{A}\mathbf{S}_3\mathbf{K}_3^{-1}|\mathbf{A}\mathbf{S}_2\mathbf{K}_2^{-1}|\mathbf{A}\mathbf{S}_1\mathbf{K}_1^{-1}|\mathbf{A}\mathbf{S}_0\tilde{\mathbf{K}}_0^{-1}) \\ &= (\mathbf{T}_3|\mathbf{T}_2|\mathbf{T}_1|\mathbf{T}_0'|\mathbf{0}_{n,r+s+t}), \\ \mathbf{M}\mathbf{S}\tilde{\mathbf{K}}^{-1} &= (\mathbf{S}_3\mathbf{K}_3\mathbf{K}_3^{-1}|\mathbf{S}_2\mathbf{K}_2\mathbf{K}_2^{-1}|\mathbf{S}_1\mathbf{K}_1\mathbf{K}_1^{-1}|\mathbf{S}_0\tilde{\mathbf{K}}_0\tilde{\mathbf{K}}_0^{-1}) \\ &= (\mathbf{S}_3|\mathbf{S}_2|\mathbf{S}_1|\mathbf{S}_0'|\mathbf{0}_{n,r+s+t}) \end{aligned}$$

where $\mathbf{T}_j = \mathbf{A}\mathbf{S}_j\mathbf{K}_j^{-1}$ for $1 \leq j \leq 3$. Their respective first $m = n - r - s - t$ columns

$$(\mathbf{T})_m = (\mathbf{A}\mathbf{S}\tilde{\mathbf{K}}^{-1})_m \text{ and } (\mathbf{S})_m = (\mathbf{M}\mathbf{S}\tilde{\mathbf{K}}^{-1})_m$$

are bases of $\mathrm{im}(\mathbf{A})$ and $\mathrm{im}(\mathbf{M})$ which satisfy

$$(\mathbf{T})_m^*\mathbf{H}(\mathbf{T})_m = (\mathbf{S})_m^*\mathbf{H}(\mathbf{S})_m = (\mathbf{Z}_K)_m = \begin{bmatrix} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_2} & \\ & -\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_1} & \\ & -\mathbf{I}_{p_1} \end{bmatrix} \oplus \mathbf{0}_r.$$

Now both bases can be extended according to Theorem 2.8 to bases of $\mathbb{C}^n$,

$$(\mathbf{T})_n = (\mathbf{T}_3|\mathbf{T}_2|\mathbf{T}_1|\mathbf{T}_0'|\mathbf{T}_0''|\tilde{\mathbf{T}}_0) \text{ and } (\mathbf{S})_n = (\mathbf{S}_3|\mathbf{S}_2|\mathbf{S}_1|\mathbf{S}_0'|\mathbf{S}_0''|\tilde{\mathbf{S}}_0),$$

for which $(\mathbf{T})_n^*\mathbf{H}(\mathbf{T})_n = (\mathbf{S})_n^*\mathbf{H}(\mathbf{S})_n = \mathbf{Z}_K$. This is obviously trivial in the case of $\mathrm{im}(\mathbf{M})$, because here $(\mathbf{S})_n = \mathbf{S}$ can be chosen. In the case of $\mathrm{im}(\mathbf{A})$ it is convenient to start the necessary application of Method 2.14 with the matrix

$$(\hat{\mathbf{T}})_m = (\mathbf{T})_m \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_{p_3} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & -\mathbf{I}_{p_3} \end{bmatrix} \oplus \mathbf{I}_{p_2+2p_1+r} \right),$$

$$(\hat{\mathbf{T}})_m^*\mathbf{H}(\hat{\mathbf{T}})_m = \begin{bmatrix} \mathbf{I}_{p_3} & \\ & -\mathbf{I}_{p_3} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_2} & \\ & -\mathbf{I}_{q_2} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{p_1} & \\ & -\mathbf{I}_{p_1} \end{bmatrix} \oplus \mathbf{0}_r,$$

whose columns already contain an H-orthonormal basis of $\mathrm{im}(\mathbf{A})$. Finally,

$$\mathbf{U} = (\mathbf{T})_n(\mathbf{S})_n^{-1} = \mathbf{T}\mathbf{S}^{-1} = \mathbf{T}\mathbf{Z}_K\mathbf{S}^*\mathbf{H}$$

is an H-isometry such that $\mathbf{U}\mathbf{M}$ is an H-polar decomposition of $\mathbf{A}$. ◊

Altogether this gives:

Method 3.25 (H-polar decomposition). *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be such that $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable. Then the steps 1 – 3 described above compute an H-polar decomposition of $\mathbf{A}$.*

In some cases an H-polar decomposition can also be obtained without computing the canonical form. This is described in the following remark.

Remark 3.26. If the matrix $\mathbf{A}^{[*]}\mathbf{A}$ has only non-real $\lambda_i, \overline{\lambda}_i$ and positive eigenvalues $\mu_j > 0$,

$$\mathbf{R}^{-1}\mathbf{A}^{[*]}\mathbf{A}\mathbf{R} = \bigoplus_{i=1}^{k} (\lambda_i\mathbf{I}_{p_i} \oplus \overline{\lambda}_i\mathbf{I}_{p_i}) \oplus \bigoplus_{j=1}^{m} (\mu_i\mathbf{I}_{q_i}),$$

the canonical form does not necessarily have to be computed. If in this case

$$\mathbf{K} = \bigoplus_{i=1}^{k} (\omega_i\mathbf{I}_{p_i} \oplus \overline{\omega}_i\mathbf{I}_{p_i}) \oplus \bigoplus_{j=1}^{m} (\varepsilon_j\sqrt{\mu_j}\mathbf{I}_{q_j})$$

is a diagonal matrix with $\omega_i^2 = \lambda_i$ and $\varepsilon_j \in \{+1, -1\}$, then $\mathbf{M} = \mathbf{R}\mathbf{K}\mathbf{R}^{-1}$ and $\mathbf{U} = \mathbf{A}\mathbf{M}^{-1}$ are already the factors of an H-polar decomposition of $\mathbf{A}$. In the more general case considered above, the canonical form is required to decide whether an H-Hermitian square root exists and to determine a suitable kernel transformation. ◊

With some minor modifications Method 3.25 also computes the factorisation

$$\mathbf{A} = \mathbf{TKS}^{-1} \text{ with } \mathbf{T}^*\mathbf{HT} = \mathbf{S}^*\mathbf{HS} = \mathbf{Z} \text{ and } \mathbf{K}^*\mathbf{Z} = \mathbf{ZK} \tag{3.7}$$

where $(\mathbf{K}, \mathbf{Z})$ is a simplified canonical form and the index at $\mathbf{Z} = \mathbf{Z}_K$ is omitted. It is only necessary to transform the blocks $\mathbf{K}_1$, $\mathbf{Z}_1$ into its canonical form $\mathbf{K}_1'$, $\mathbf{Z}_1'$ in step 2, and to return the transformation $\mathbf{T}$ instead of $\mathbf{U}$. In this form the method allows to derive all H-polar decompositions

$$\mathbf{A} = \mathbf{UM} \text{ with } \mathbf{U} = \mathbf{T\Sigma S}^{-1} \text{ and } \mathbf{M} = \mathbf{S\Sigma KS}^{-1} \tag{3.8}$$

where $\mathbf{\Sigma} = \mathbf{\Sigma}^* = \mathbf{\Sigma}^{-1}$ must be a sign matrix commuting with $\mathbf{K}$ and $\mathbf{Z}$.

In the particular case $\mathbf{H} = \mathbf{I}$, the matrix $\mathbf{K}$ is a diagonal matrix with non-negative diagonal elements and $\mathbf{Z} = \mathbf{I}$, too. In other words, in this case (3.7) is just a singular value decomposition which leads to the following definition.

Definition 3.27 (H-singular value decomposition). Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{A} \in \mathbb{F}^{n\times n}$. A factorisation of the form (3.7) is called an H-singular value decomposition of $\mathbf{A}$. ◊

This factorisation must in no way be mixed up with the trivial factorisation described in the following remark.

Remark 3.28. Let $\mathbf{H} = \mathrm{diag}_n(\pm 1)$ and let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be nonsingular. Furthermore, let $\mathbf{R\Lambda R}^*$ be an eigenvalue decomposition of $\mathbf{AHA}^*$. Then $\mathbf{R}$ is orthogonal or unitary, respectively, and the diagonal matrix of the real eigenvalues can be decomposed as $\mathbf{\Lambda} = \mathbf{\Omega H \Omega}^*$ where $\mathbf{\Omega} = \mathbf{\Omega}^*$ has positive diagonal elements. Now $\mathbf{AHA}^* = \mathbf{R\Omega H\Omega}^*\mathbf{R}^*$ implies that the matrix $\mathbf{U}$ defined by $\mathbf{AU} = \mathbf{R\Omega}$ is an H-isometry, and therefore the decomposition $\mathbf{A} = \mathbf{R\Omega U}^{-1}$ is called a hyperbolic singular value decomposition by some authors [BOS]. ◊

In the rest of this section we present some statistical experiments which were made to assess the numerical properties of the Methods 3.24 and 3.25. The philosophy in implementing the test programme was to use highly reliant standard linear algebra software in combination with our own building blocks. Thus, we have chosen to implement the programme in Fortran 77 using the `DOUBLE COMPLEX` versions of `LAPACK` and the `BLAS` [LUG]. Our extensions for computing the HQR decompositions (Algorithms 2.12 and 2.13), the extension of isometries (Method 2.14), and the further steps of methods were implemented as careful as the `LAPACK` codes. All results were obtained on a `PENTIUM 4` processor, for which the machine accuracy is

$$\varepsilon_{mach} \approx 2.22 \cdot 10^{-16}.$$

To test the Method 3.24 we specified the canonical form

$$\mathbf{J} = (\lambda \mathbf{I}_{p_3} \oplus \overline{\lambda}\mathbf{I}_{p_3}) \oplus \alpha \mathbf{I}_{p_2} \oplus \beta \mathbf{I}_{p_1} \oplus \mathbf{0}_r,$$

$$\mathbf{Z} = \begin{bmatrix} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & \end{bmatrix} \oplus \mathrm{diag}_{p_2}(\pm 1) \oplus \mathrm{diag}_{p_1}(\pm 1) \oplus \mathrm{diag}_r(\pm 1),$$

$$p_3 = 5, \;\; p_2 = 10, \;\; p_1 = 10, \;\; r = 10, \;\; n = 40$$

and used random eigenvalues $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $\alpha, \beta \in \mathbb{R}\backslash\{0\}$ and random transformations $\mathbf{R} \in \mathbb{C}^{n\times n}$ to construct the test matrices

$$\mathbf{A} = \mathbf{R}^{-1}\mathbf{JR}, \quad \mathbf{H} = \mathbf{R}^*\mathbf{ZR}.$$

Here the transformation $\mathbf{R}$ and the non-zero eigenvalues of $\mathbf{J}$ were initialised with normally distributed random numbers from the interval $[-2, 2]$, but it was controlled that the magnitudes of α, β and λ were at least 0.2.

Then the canonical form of the test matrix pair was computed

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}, \quad \mathbf{S}^*\mathbf{H}\mathbf{S} = \mathbf{Z},$$

whose numerical accuracy is estimated via the residuals

$$r_{AJ} = \|\mathbf{A}\mathbf{S} - \mathbf{S}\mathbf{J}\|_F, \quad r_{SZ} = \|\mathbf{S}^*\mathbf{H}\mathbf{S} - \mathbf{Z}\|_F$$

and the reciprocal condition number

$$c_S^{-1} = \mathrm{cond}_1(\mathbf{S})^{-1}, \quad \mathrm{cond}_1(\mathbf{S}) = \|\mathbf{S}\|_1 \, \|\mathbf{S}^{-1}\|_1.$$

The results of two experiments with $K = 30$ repetitions are presented in Table 3.1 where the columns "min" and "max" contain the observed minimum and maximum values. The column "avg" lists the respective average values which were computed as

$$\mathrm{avg}(x) = 10^{\bar{x}} \ \text{ with } \ \bar{x} = \frac{1}{K}\sum_{k=1}^{K} \log(x_k)$$

to avoid the domination of the large quantities. The tolerance parameter for the grouping of the eigenvalues, required in step 2 of Method 3.24, was always given by

$$\delta = 10^{-8}.$$

In the first experiment we used HQR and HQR-2 decompositions to H-orthogonalise the eigenvectors in step 3 of Method 3.24. In the second experiment this process was carried out with eigenvalue and singular value decompositions. To permit the comparability of the data, both experiments were made with the same test matrices.

Before assessing the results in Table 3.1 it is first of all necessary to note that the computed residuals are *absolute* errors of complex 40×40 matrices. In view of this fact even the maximum values for r_{SZ} appear to be acceptable, so that the canonical forms were in any case computed with an acceptable accuracy.

The comparison of the two methods for H-orthogonalising the eigenvectors shows that the eigenvalue and singular value decompositions were minimally more accurate, whereas the HQR and HQR-2 decompositions produced minimally better conditioned transformations. We may therefore conclude that both variants produce qualitatively similar results.

Finally, we implemented and installed some benchmark routines and counted the numbers of floating point operations (flops) which were required for computing the canonical forms. The average values

$$\mathrm{flops}_{HQR/HQR-2} = 2.034 \cdot 10^7 \ \text{ and } \ \mathrm{flops}_{EVD/SVD} = 2.043 \cdot 10^7$$

reveal a minor advantage for the HQR/HQR-2 variant, but it is without significance in the overall process.

Table 3.1: Results of two experiments with Method 3.24

	with HQR/HQR-2			with EVD/SVD		
	min	avg	max	min	avg	max
r_{AJ}	3.44e-14	2.76e-13	5.17e-11	3.23e-14	2.69e-13	5.14e-11
r_{SZ}	1.35e-13	1.14e-12	9.31e-11	1.12e-13	9.35e-13	7.55e-11
c_S^{-1}	1.17e-04	7.80e-04	2.40e-03	1.29e-04	8.40e-04	3.50e-03

Table 3.2: Results of two experiments with Method 3.25

	original transformations			modified transformations		
	min	avg	max	min	avg	max
r_{AK}	2.30e-12	1.97e-11	7.17e-10	6.75e-12	9.27e-11	2.04e-09
r_{TZ}	1.24e-11	7.94e-11	5.85e-10	5.14e-13	2.49e-12	1.05e-11
r_{SZ}	2.40e-12	2.45e-11	3.08e-10	4.57e-14	1.04e-13	2.46e-13
c_T^{-1}	1.03e-05	2.97e-05	1.20e-04	1.03e-05	2.97e-05	1.20e-04
c_S^{-1}	2.96e-04	8.59e-04	2.70e-03	2.96e-04	8.59e-04	2.70e-03
r_{UM}	9.82e-11	1.39e-09	3.73e-08	4.19e-11	6.33e-10	1.57e-08
r_{MH}	4.27e-13	9.21e-13	2.18e-12	4.17e-13	9.14e-13	2.28e-12
r_{UH}	7.99e-10	5.92e-09	5.19e-08	3.46e-11	1.86e-10	2.78e-09
c_U^{-1}	3.35e-07	3.15e-06	2.53e-05	3.35e-07	3.15e-06	2.53e-05

To test the Method 3.25 which was implemented such that is computes an H-singular value decomposition (H-SVD), we used a similar scenario. Here we specified the canonical form

$$\mathbf{K} = (\lambda\mathbf{I}_{p_3} \oplus \overline{\lambda}\mathbf{I}_{p_3}) \oplus \alpha\mathbf{I}_{p_2} \oplus (i\beta\mathbf{I}_{p_1} \oplus -i\beta\mathbf{I}_{p_1}) \oplus (\mathbf{0}_{2r} \oplus \mathbf{0}_s),$$

$$\mathbf{Z} = \begin{bmatrix} & \mathbf{I}_{p_3} \\ \mathbf{I}_{p_3} & \end{bmatrix} \oplus \mathrm{diag}_{p_2}(\pm 1) \oplus \begin{bmatrix} & \mathbf{I}_{p_1} \\ \mathbf{I}_{p_1} & \end{bmatrix} \oplus \left(\begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \mathrm{diag}_s(\pm 1) \right),$$

$$p_3 = 5, \;\; p_2 = 10, \;\; p_1 = 5, \;\; r = 3, \;\; s = 4, \;\; n = 40$$

and used random eigenvalues $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $\alpha, \beta \in \mathbb{R}\backslash\{0\}$, random transformations $\mathbf{R} \in \mathbb{C}^{n\times n}$ and random H-isometries $\mathbf{V} \in \mathbb{C}^{n\times n}$ to construct the test matrices

$$\mathbf{A} = \mathbf{V}\mathbf{R}^{-1}\mathbf{K}\mathbf{R}, \quad \mathbf{H} = \mathbf{R}^*\mathbf{Z}\mathbf{R}. \tag{3.9}$$

Whereas the transformation $\mathbf{R}$ and the non-zero eigenvalues of $\mathbf{K}$ were initialised as in the tests of Method 3.24, the H-isometries $\mathbf{V}$ were built according to the following remark.

Remark 3.29. Let $\mathbf{w} \in \mathbb{F}^n$ be a non-neutral vector and let

$$\mathbf{W} = \mathbf{I} - 2\frac{\mathbf{w}\mathbf{w}^*\mathbf{H}}{\mathbf{w}^*\mathbf{H}\mathbf{w}} \in \mathbb{F}^{n\times n}$$

be a generalised Householder reflection. Then

$$\mathbf{W}^*\mathbf{H} = \mathbf{H}\mathbf{W} \;\text{ and }\; \mathbf{W}^*\mathbf{H}\mathbf{W} = \mathbf{H},$$

so that $\mathbf{W}$ is an H-selfadjoint H-isometry. However, if $\mathbf{V}$ is a product of at least two generalised Householder reflections $\mathbf{W}_k$ $(k = 1, 2, \ldots)$, then it clearly is an

H-isometry, but it is not H-selfadjoint in general. Therefore, if the reflections are defined by some non-neutral random vectors $\mathbf{w}_k$, the matrix $\mathbf{V}$ is a random H-isometry which has no further structure. ♢

For computing the test matrix $\mathbf{A}$ we applied two generalised Householder reflections to the matrix $\mathbf{R}^{-1}\mathbf{K}\mathbf{R}$. The corresponding random vectors $\mathbf{w}_1, \mathbf{w}_2$ were initialised with normally distributed random numbers from the interval $[-1, 1]$. To bound the norms of the reflections, the vectors $\mathbf{w}_k$ were only accepted when they satisfied $|\mathbf{w}_k^*\mathbf{H}\mathbf{w}_k| \geq 1$.

Then an H-SVD of the test matrix

$$\mathbf{A} = \mathbf{T}\mathbf{K}\mathbf{S}^{-1}, \quad \mathbf{T}^*\mathbf{H}\mathbf{T} = \mathbf{Z}, \quad \mathbf{S}^*\mathbf{H}\mathbf{S} = \mathbf{Z}$$

and also an H-polar decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{M}, \quad \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H}, \quad \mathbf{M}^*\mathbf{H} = \mathbf{H}\mathbf{M}$$

was computed. Here the tolerance parameters for the grouping of the eigenvalues in step 2 of Method 3.24 and the rank determination in the kernel transformation of Method 3.25 were given by

$$\delta = 10^{-8} \text{ and } \tau = 10^{-8}.$$

Again two experiments with 30 repetitions we made. The results are presented in Table 3.2 which contains the residuals

$$\begin{aligned} r_{AK} &= \|\mathbf{A}\mathbf{S} - \mathbf{T}\mathbf{K}\|_F, & r_{UM} &= \|\mathbf{A} - \mathbf{U}\mathbf{M}\|_F, \\ r_{TZ} &= \|\mathbf{T}^*\mathbf{H}\mathbf{T} - \mathbf{Z}\|_F, & r_{MH} &= \|\mathbf{M}^*\mathbf{H} - \mathbf{H}\mathbf{M}\|_F, \\ r_{SZ} &= \|\mathbf{S}^*\mathbf{H}\mathbf{S} - \mathbf{Z}\|_F, & r_{UH} &= \|\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H}\|_F \end{aligned} \tag{3.10}$$

as well as the reciprocal 1-condition numbers of $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{U}$.

In the first experiment the H-polar decompositions were computed using

$$\mathbf{U} = \mathbf{T}\mathbf{S}_{\text{inv}} \text{ and } \mathbf{M} = \mathbf{S}\mathbf{K}\mathbf{S}_{\text{inv}} \text{ with } \mathbf{S}_{\text{inv}} = \mathbf{Z}\mathbf{S}^*\mathbf{H}.$$

Although the residuals for the H-SVDs do not indicate serious errors, the residuals r_{UM} and r_{UH} show a significant loss of accuracy for the H-unitary factor of the H-polar decompositions. We therefore repeated the experiment and tried to improve the transformations $\mathbf{T}$ and $\mathbf{S}$. For this purpose the linear systems

$$(\mathbf{Z}\mathbf{T}^*\mathbf{H})\mathbf{Y} = \mathbf{I} \text{ and } (\mathbf{Z}\mathbf{S}^*\mathbf{H})\mathbf{X} = \mathbf{I} \tag{3.11a}$$

were solved and the modified transformations

$$\tilde{\mathbf{T}} = \tfrac{1}{2}(\mathbf{T} + \mathbf{Y}) \text{ and } \tilde{\mathbf{S}} = \tfrac{1}{2}(\mathbf{S} + \mathbf{X}) \tag{3.11b}$$

were used instead of $\mathbf{T}$ and $\mathbf{S}$. Indeed this approach helped to reduce most of the residuals. Only r_{AK} was incremented a little.

Whereas these experiments indicate that Method 3.24 is stable for computing simplified canonical forms, we must admit that there is still some doubt concerning the stability of Method 3.25. The major problems of this method are:

(*a*) If the matrix $\mathbf{A}$ is singular, then several matrix factorisations are required for the kernel transformation and for the extension of the matrix $\mathbf{T}$.

(*b*) If the matrix $\mathbf{A}$ is ill-conditioned, then the formation of $\mathbf{A}^{[*]}\mathbf{A}$ can lead to a severe loss of accuracy.

The first problem is obvious and the second is a well-known fact which for the case $\mathbf{H} = \mathbf{I}$ is discussed in many text books on numerical linear algebra (for example see [GVL, Section 8.6.2] or [ST2, Kapitel 6.7]).

To solve these problems we have tried to modify the standard algorithm for computing ordinary SVDs [GVL, Section 8.6] such that it computes H-SVDs. This approach failed, because we were not able to determine the canonical form of the pair $(\mathbf{K}, \mathbf{Z})$ by directly transforming $\mathbf{A}$ (and $\mathbf{H}$). In the case $\mathbf{H} = \mathbf{I}$, this form is simply $(\mathrm{diag}(\kappa_i), \mathbf{I})$ and the standard algorithm is inherently based on this fact. A further approach succeeded and resulted in the algorithms presented in the following section.

3.6 Numerical computation of H-polar decompositions of a matrix A for which $\mathbf{A}^{[*]}\mathbf{A}$ has no non-positive real eigenvalues

The numerical computation of ordinary polar decompositions is not only possible via the singular value decomposition. There also exists an iteration method which is closely related to the Newton iteration for computing the matrix sign function [HI1], [HMMT]. In this section we will adopt this method for computing H-polar decompositions and we will also describe an extension for computing H-singular value decompositions.

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be nonsingular. Then the iteration method

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{X}_k^{-*}), \quad \mathbf{X}_0 = \mathbf{A}, \quad k = 0, 1, \ldots \tag{3.12a}$$

quadratically converges to the isometric factor of an ordinary polar decomposition $\mathbf{A} = \mathbf{UM}$ where $\mathbf{U}^* = \mathbf{U}^{-1}$ and $\mathbf{M}^* = \mathbf{M}$. The optimum selfadjoint factor corresponding to the approximate isometry

$$\tilde{\mathbf{U}} = \mathbf{X}_k \ \text{ with } \ \|\mathbf{X}_k^*\mathbf{X}_k - \mathbf{I}\|_F \le \varepsilon \tag{3.12b}$$

is given by [HP, Lemma 2.1]

$$\tilde{\mathbf{M}} = \frac{1}{2}(\tilde{\mathbf{U}}^*\mathbf{A} + \mathbf{A}^*\tilde{\mathbf{U}}), \tag{3.12c}$$

and, furthermore, the particular polar decomposition is computed for which $\mathbf{M}$ is positive definite. By substituting the adjoints in (3.12) with H-adjoints we obtain the iteration method

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{X}_k^{-H}), \quad \mathbf{X}_0 = \mathbf{A}, \quad k = 0, 1, \ldots, \tag{3.13}$$

$$\tilde{\mathbf{U}} = \mathbf{X}_k \ \text{ with } \ \|\mathbf{X}_k^H\mathbf{X}_k - \mathbf{I}\|_F \le \varepsilon \ \text{ and } \ \tilde{\mathbf{M}} = \frac{1}{2}(\tilde{\mathbf{U}}^H\mathbf{A} + \mathbf{A}^H\tilde{\mathbf{U}}).$$

The following theorem shows that this method actually computes an H-polar decomposition of $\mathbf{A}$. Note that a similar result has been derived independently by Higham [HI2, Theorem 5.2].

Theorem 3.30. *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a matrix such that $\sigma(\mathbf{A}^H\mathbf{A}) \subset \mathbb{C}\backslash(-\infty, 0]$. Then the method (3.13) applied to $\mathbf{A}$ computes the particular H-polar decomposition $\mathbf{A} = \mathbf{UM}$, for which $\sigma(\mathbf{M})$ lies in the open right complex half-plane.*

Proof. Let $\mathbf{A}$ be a matrix such that $\sigma(\mathbf{A}^H\mathbf{A})$ has the assumed property. Then Corollary 3.7 implies that $\mathbf{A}$ admits an H-polar decomposition $\mathbf{A} = \mathbf{U}_0\mathbf{M}_0$ such that $\sigma(\mathbf{M}_0) \subset \{z \in \mathbb{C} | \operatorname{Re}(z) > 0\}$. Now applying (3.13) to $\mathbf{A}$ we obtain

$$\begin{aligned} 2\mathbf{X}_1 &= \mathbf{U}_0\mathbf{M}_0 + (\mathbf{U}_0\mathbf{M}_0)^{-H} = \mathbf{U}_0\mathbf{M}_0 + (\mathbf{M}_0^H\mathbf{U}_0^H)^{-1} \\ &= \mathbf{U}_0\mathbf{M}_0 + (\mathbf{M}_0\mathbf{U}_0^{-1})^{-1} = \mathbf{U}_0\mathbf{M}_0 + \mathbf{U}_0\mathbf{M}_0^{-1} = \mathbf{U}_0(\mathbf{M}_0 + \mathbf{M}_0^{-1}) \end{aligned}$$

or

$$\mathbf{X}_1 = \mathbf{U}_0\mathbf{M}_1 \text{ with } \mathbf{M}_1 = \frac{1}{2}(\mathbf{M}_0 + \mathbf{M}_0^{-1}) = \mathbf{M}_1^H,$$

from which it follows that

$$\mathbf{X}_{k+1} = \mathbf{U}_0\mathbf{M}_{k+1} \text{ with } \mathbf{M}_{k+1} = \frac{1}{2}(\mathbf{M}_k + \mathbf{M}_k^{-1}) = \mathbf{M}_{k+1}^H.$$

Moreover, $\operatorname{Re}\lambda > 0$ for all $\lambda \in \sigma(\mathbf{M}_0)$ according to [R] implies

$$\lim_{k\to\infty} \mathbf{M}_k = \mathbf{I},$$

so that we finally have

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{U}_0. \qquad \square$$

In an implementation of method (3.13) the matrices $\mathbf{X}_k^{-H} = \mathbf{H}^{-1}\mathbf{X}_k^{-*}\mathbf{H}$ must be computed by solving the linear system $(\mathbf{X}_k^*\mathbf{H})\mathbf{Y}_k = \mathbf{H}$ for $\mathbf{Y}_k = \mathbf{X}_k^{-H}$. Analogously, the matrix $\tilde{\mathbf{M}}$ must be computed by solving $\mathbf{HY} = (\tilde{\mathbf{U}}^*\mathbf{HA} + \mathbf{A}^*\mathbf{H}\tilde{\mathbf{U}})/2$ for $\mathbf{Y} = \tilde{\mathbf{M}}$.

The termination criterion can be applied in the form $\|\mathbf{X}_k^*\mathbf{H}\mathbf{X}_k - \mathbf{H}\|_F \le \varepsilon$. It is based on an absolute error, but it is also possible to use a criterion based on a relative error. To obtain the latter we have adopted a similar consideration made in [HMMT, Section 4] and [HI2, Section 5]:

Let $\|.\|$ be a submultiplicative matrix norm, and let $\mathbf{U}$ and $\Delta\mathbf{U}$ satisfy

$$\mathbf{U}^*\mathbf{HU} = \mathbf{H} \text{ and } \|\Delta\mathbf{U}\| \le \varepsilon\|\mathbf{U}\|.$$

Then the inequality

$$\begin{aligned} &\|(\mathbf{U} + \Delta\mathbf{U})^*\mathbf{H}(\mathbf{U} + \Delta\mathbf{U}) - \mathbf{H}\| \\ &\quad = \|\mathbf{U}^*\mathbf{H}(\Delta\mathbf{U}) + (\Delta\mathbf{U})^*\mathbf{HU} + (\Delta\mathbf{U})^*\mathbf{H}(\Delta\mathbf{U})\| \\ &\quad \le \|\mathbf{U}^*\mathbf{H}(\Delta\mathbf{U})\| + \|(\Delta\mathbf{U})^*\mathbf{HU}\| + \|(\Delta\mathbf{U})^*\mathbf{H}(\Delta\mathbf{U})\| \\ &\quad \le (2\varepsilon + \varepsilon^2)\,\|\mathbf{H}\|\,\|\mathbf{U}\|^2 \end{aligned}$$

shows that $\mathbf{U}$ is an H-isometry to working precision when it fulfills

$$\rho_{UH} = \frac{\|\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H}\|}{\|\mathbf{H}\| \, \|\mathbf{U}\|^2} \approx \varepsilon_{mach}.$$

Hence, the iteration should be terminated when the iterate satisfies

$$\rho = \frac{\|\mathbf{X}_k^*\mathbf{H}\mathbf{X}_k - \mathbf{H}\|_F}{\|\mathbf{H}\|_F \, \|\mathbf{X}_k\|_F^2} \leq \varepsilon \text{ for some } \varepsilon \approx \varepsilon_{mach}. \tag{3.14}$$

Here the submultiplicative property of the Frobenius norm has been used.

Algorithm 3.31. Given the matrices $\mathbf{A}, \mathbf{H} \in \mathbb{F}^{n\times n}$, a tolerance parameter ε and a maximum number of iteration steps $maxits$, the following algorithm performs method (3.13). The matrices $\mathbf{U}$ and $\mathbf{M}$ are stored in separate arrays and $\mathbf{W}$ is a working array. The notation "$\mathbf{WY} = \mathbf{M} \to \mathbf{M} = \mathbf{Y}$" is to be read as "solve the system $\mathbf{WY} = \mathbf{M}$ and overwrite $\mathbf{M}$ with $\mathbf{Y}$".

$\mathbf{U} = \mathbf{A}$
`for` $its = 1, \ldots, maxits$ `do`
    $\mathbf{W} = \mathbf{U}^*\mathbf{H}$
    $\mathbf{M} = \mathbf{H}$
    $\mathbf{M} = \mathbf{W}\mathbf{U} - \mathbf{M}$ `(=` $\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H}$`)`
    $\rho = \|\mathbf{M}\|_F$ `(or` $\rho = \|\mathbf{M}\|_F/\|\mathbf{H}\|_F/\|\mathbf{U}\|_F/\|\mathbf{U}\|_F$`)`
    `if` $\rho \leq \varepsilon$ `then`
        `break`
    `end if`
    $\mathbf{M} = \mathbf{H}$
    $\mathbf{W}\mathbf{Y} = \mathbf{M} \to \mathbf{M} = \mathbf{Y}$
    $\mathbf{U} = (\mathbf{U} + \mathbf{M})/2$
`end for`
`if` $its > maxits$ `then`
    `return` *"divergent"*
`end if`
$\mathbf{M} = \mathbf{W}\mathbf{A}$ `(=` $\mathbf{U}^*\mathbf{H}\mathbf{A}$`)`
$\mathbf{M} = (\mathbf{M} + \mathbf{M}^*)/2$
$\mathbf{W} = \mathbf{H}$
$\mathbf{W}\mathbf{Y} = \mathbf{M} \to \mathbf{M} = \mathbf{Y}$
`return` *"convergent"*

This algorithm represents the basic form of the Newton iteration. It may be improved with factors for convergence acceleration analogously to [HP, Section 2], but this is not considered here. We will rather turn our interest to the fact that the algorithm only allows to compute two particular H-polar decompositions, namely, the decompositions (a) $\mathbf{A} = \mathbf{U}\mathbf{M}$ and (b) $\mathbf{A} = (-\mathbf{U})(-\mathbf{M})$ where $\mathbf{M}$ is the principal square root of $\mathbf{A}^H\mathbf{A}$. However, there are applications in which another decomposition is required. In particular, it turns out that the solution of the Procrustes problems requires semidefinite H-polar decompositions which have been introduced in Definition 3.22. Now, assume that $\mathbf{A} \in \mathbb{F}^{n\times n}$ is nonsingular. Then, according to Corollary 3.23, an H-polar decomposition of $\mathbf{A}$ is

definite if and only if the canonical form of the pair $(\mathbf{M}, \mathbf{H})$ is

$$\mathbf{S}^{-1}\mathbf{M}\mathbf{S} = \bigoplus_{j=1}^{k}(\omega_j \mathbf{I}_{p_j - q_j} \oplus -\omega_j \mathbf{I}_{q_j}), \quad \mathbf{S}^*\mathbf{H}\mathbf{S} = \bigoplus_{j=1}^{k}(\mathbf{I}_{p_j - q_j} \oplus -\mathbf{I}_{q_j})$$

where $\omega_j > 0$. This means that the decomposition (a) is definite when $\mathbf{H}$ is positive definite, and the decomposition (b) is definite when $\mathbf{H}$ is negative definite. But if $\mathbf{H}$ is indefinite, it is impossible to compute a definite H-polar decomposition with Algorithm 3.31.

The solution of this problem is very simple. We only need to extend the algorithm such that it computes an H-SVD and can then compute arbitrary H-polar decompositions by applying Equation (3.8). The extension consists of computing the canonical form of the pair $(\mathbf{M}, \mathbf{H})$ which for non-diagonalisable H-Hermitian matrices is described in Chapter 6.

Algorithm 3.32. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a matrix such that $\sigma(\mathbf{A}^H\mathbf{A}) \subset \mathbb{C}\backslash(-\infty, 0]$. Then an H-SVD of $\mathbf{A}$ can be computed with the following steps:*

1. *Compute an H-polar decomposition $\mathbf{A} = \mathbf{U}\mathbf{M}$ with Algorithm 3.31.*
2. *Compute the canonical form $(\mathbf{S}^{-1}\mathbf{M}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mathbf{K}, \mathbf{Z})$ of the pair $(\mathbf{M}, \mathbf{H})$ with Method 3.24 or, if $\mathbf{M}$ is not diagonalisable, with the algorithm described in Chapter 6.*
3. *Compute $\mathbf{T} = \mathbf{U}\mathbf{S}$, so that $\mathbf{A} = \mathbf{T}\mathbf{K}\mathbf{S}^{-1}$ is an H-SVD.*

Now, let $\mathbf{A}$ be a (nonsingular) matrix such that $\mathbf{A}^H\mathbf{A}$ is diagonalisable with positive eigenvalues, and let $\mathbf{A} = \mathbf{T}\mathbf{K}\mathbf{S}^{-1}$ be an an H-SVD computed with Algorithm 3.32 or with Method 3.25. Then the application of (3.8) with $\mathbf{\Sigma} = \mathbf{Z}$ results in

$$\mathbf{A} = \mathbf{U}\mathbf{M} \text{ with } \mathbf{U} = \mathbf{T}\mathbf{S}^*\mathbf{H} \text{ and } \mathbf{M} = \mathbf{S}\mathbf{K}\mathbf{S}^*\mathbf{H} \tag{3.15}$$

which is a definite H-polar decomposition of $\mathbf{A}$. If $\mathbf{A}$ is singular and $\mathbf{A}^H\mathbf{A}$ is diagonalisable with non-negative eigenvalues, the H-SVD can only be computed with Method 3.25. Then (3.15) is a semidefinite H-polar decomposition of $\mathbf{A}$.

To be able to test the algorithms we extended our software package with a Fortran 77 implementation of Algorithm 3.31. Again the codes were written using the `DOUBLE COMPLEX` versions of `LAPACK` and the `BLAS`. Then we made some experiments in which H-polar decompositions and definite H-polar decompositions were computed. For this purpose the canonical forms

$$\mathbf{K} = \mathbf{J}_p(\lambda) \oplus \mathbf{J}_p(\overline{\lambda}) \oplus \mathbf{J}_q(\alpha) \oplus \mathbf{J}_r(\beta), \quad \mathbf{Z} = \mathbf{Z}_{2p} \oplus \pm\mathbf{Z}_q \oplus \pm\mathbf{Z}_r,$$
$$p = q = r = 10, \quad n = 40$$

with $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $\alpha, \beta \in \mathbb{R}\backslash\{0\}$, were specified to test Algorithm 3.31, and the canonical forms

$$\mathbf{K} = \bigoplus_{j=1}^{4} \lambda_j \mathbf{I}_{p_j}, \quad \mathbf{Z} = \mathrm{diag}_n(\pm 1), \quad p_j = 10, \quad n = 40$$

with $\lambda_j \in \mathbb{R}\backslash\{0\}$, were specified to test Algorithm 3.32. Using randomly chosen eigenvalues, transformations $\mathbf{R} \in \mathbb{C}^{n\times n}$ and H-isometries $\mathbf{V} \in \mathbb{C}^{n\times n}$, the test

Table 3.3: Results of two experiments with Algorithm 3.31

	moderately conditioned			badly conditioned		
	min	avg	max	min	avg	max
c_A^{-1}	1.15e-07	4.14e-06	2.65e-05	8.41e-12	5.60e-09	9.69e-08
r_{UM}	2.31e-12	9.10e-11	4.75e-08	2.71e-09	3.65e-07	8.21e-04
r_{MH}	6.79e-13	1.42e-12	5.63e-12	7.81e-13	1.68e-12	1.83e-11
r_{UH}	1.68e-12	1.50e-11	1.69e-10	2.37e-12	1.93e-11	1.66e-09
c_U^{-1}	7.91e-07	1.47e-05	1.47e-04	5.90e-08	9.91e-06	1.86e-04
its	6	7.33	8	7	8.20	10

Table 3.4: Results of two experiments with Algorithm 3.32 and Method 3.25

	Algorithm 3.32			Method 3.25		
	min	avg	max	min	avg	max
r_{AK}	3.26e-13	8.02e-12	4.36e-10	9.94e-16	3.10e-15	1.64e-14
r_{TZ}	5.13e-13	2.28e-12	1.14e-11	1.50e-12	2.13e-11	3.04e-10
r_{SZ}	2.85e-13	1.40e-12	1.13e-11	1.84e-13	8.63e-13	1.75e-11
c_T^{-1}	1.44e-05	9.21e-05	9.82e-04	6.78e-06	5.12e-05	3.07e-04
c_S^{-1}	3.68e-04	9.94e-04	2.52e-03	4.12e-04	8.14e-04	2.85e-03
its	5	6.10	7			
r_{UM}	1.37e-11	1.67e-10	1.90e-09	1.19e-11	1.15e-10	2.36e-09
r_{MH}	1.17e-12	2.76e-12	1.15e-11	8.26e-13	1.57e-12	4.92e-12
r_{UH}	6.63e-11	3.85e-10	2.76e-09	1.39e-10	2.68e-09	6.53e-08
c_U^{-1}	2.12e-07	3.31e-06	5.41e-05	1.45e-07	1.48e-06	2.64e-05

matrices $\mathbf{A} = \mathbf{V}\mathbf{R}^{-1}\mathbf{K}\mathbf{R}$ and $\mathbf{H} = \mathbf{R}^*\mathbf{Z}\mathbf{R}$ were constructed as in (3.9). The magnitudes of the eigenvalues of $\mathbf{K}$ were at least 0.4. The machine accuracy was $\varepsilon_{mach} \approx 2.22 \cdot 10^{-16}$.

To test Algorithm 3.31 we made two experiments with 30 repetitions in which moderately conditioned and badly conditioned test matrices were used. The iteration was terminated when the relative error (3.14) was at most $\varepsilon = 10^{-16}$. The results are listed in Table 3.3 where c_A^{-1} is the reciprocal 1-condition number of the test matrix, *its* is the number of iteration steps and the further meanings are as in (3.10).

The table shows that the iteration required only between 6 and 8 steps in the first experiment and also the 7 to 10 steps used in the second experiment are not too much. Taking into account that the absolute errors r_{UH} correspond to relative errors near machine accuracy, the precision of the computations is very satisfactory. The only weak point is the residual r_{UM} obtained for the badly conditioned test matrices. In further experiments with ill-conditioned matrices where $c_A^{-1} < 10^{-12}$, this residual often indicated that the computed H-polar decomposition was utterly erroneous.

To test Algorithm 3.32 we made an experiment with 30 repetitions in which H-SVDs and then definite H-polar decompositions were computed. The parameter for Algorithm 3.31 was $\varepsilon = 10^{-16}$ and for Method 3.24 it was $\delta = 10^{-8}$. The same experiment was also made with Method 3.25 which was configured with

$\delta = \tau = 10^{-8}$. To permit comparison of the algorithms the transformations $\mathbf{T}$ and $\mathbf{S}$ were not modified according to (3.11). The results are listed in Table 3.4 which contains the residuals and reciprocal condition numbers for the H-SVDs and for the definite H-polar decompositions.

The residuals obtained with Algorithm 3.32 are more balanced than those obtained with Method 3.25. In particular the values of r_{UH} suggest to consider Algorithm 3.32 as the preferable method. However, its advantage is not tremendous and we may conclude that both algorithms are appropriate for computing H-SVDs and H-polar decompositions.

Chapter 4

Procrustes problems and (G,H)-polar decompositions

4.1 Introduction

In Chapter 1 we have seen that classical multidimensional scaling (MDS) is a technique for analysing empirical data which essentially consists of

(1) a method for constructing vectors $\mathbf{x}_k$ such that $\|\mathbf{x}_k - \mathbf{x}_l\| = d_{kl}$ for given distances d_{kl} $(1 \leq k, l \leq N)$ which satisfy the triangle inequality and

(2) a method for determining an isometry $\mathbf{U}$ such that $\sum_k \|\mathbf{U}\mathbf{x}_k - \mathbf{y}_k\|$ is minimal for given vectors $\mathbf{x}_k$ and $\mathbf{y}_k$ $(1 \leq k \leq N)$.

Moreover, we have formulated the goal to generalise these methods for real and complex indefinite scalar product spaces which requires

(1′) a method for constructing vectors $\mathbf{x}_k$ and an indefinite scalar product such that $[\mathbf{x}_k - \mathbf{x}_l, \mathbf{x}_k - \mathbf{x}_l] = q_{kl}$ for given real numbers q_{kl} $(1 \leq k, l \leq N)$ and

(2′) methods for solving the indefinite Procrustes problems (1.1) – (1.3).

In this chapter most of these generalisations are presented: In Section 4.3 the method (1') is given and in the Sections 4.4 and 4.5 the Procrustes problems (1.1) and (1.2) are solved.

Whereas the discussion of problem (1.1) is possible with the results of the previous chapter, the investigation of problem (1.2) requires some further statements on doubly structured indefinite polar decompositions which are provided in Section 4.2 and generalised in Section 4.6. In the final Section 4.7 the numerical computation of these doubly structured decompositions is considered.

4.2 Introduction to (G,H)-polar decompositions

In the previous chapter H-polar decompositions of real and complex matrices have been discussed extensively. In particular, the semidefinite H-polar decompositions introduced in Definition 3.22 and characterised in Corollary 3.23 are an important tool for solving the Procrustes problems.

This chapter will also make use of indefinite polar decompositions where the factors $\mathbf{U}$ and $\mathbf{M}$ are doubly structured with respect to two selfadjoint matrices $\mathbf{G}$ and $\mathbf{H}$.

Definition 4.1. ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$) Let $\mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$ be nonsingular and selfadjoint and let $\mathbf{A} \in \mathbb{F}^{n\times n}$. A factorisation of the form

$$\mathbf{A} = \mathbf{UM} \text{ with } \mathbf{U}^H = \mathbf{U}^G = \mathbf{U}^{-1} \text{ and } \mathbf{M}^H = \mathbf{M}^G = \mathbf{M}$$

is called a *(G,H)-polar decomposition* of $\mathbf{A}$. A matrix having the properties of $\mathbf{U}$ is said to be (G,H)-isometric (-orthogonal or -unitary), and a matrix having the properties of $\mathbf{M}$ is said to be (G,H)-selfadjoint (-symmetric or -Hermitian). If the factor $\mathbf{M}$ in particular is H-nonnegative ($\mathbf{HM} \geq 0$), the factorisation is called an *H-semidefinite* (G,H)-polar decomposition. ◊

These factorisations will be of interest in the special case in which the matrices $\mathbf{G}$ and $\mathbf{H}$ satisfy

$$\mathbf{H}^{-1}\mathbf{G} = \mu^2\mathbf{G}^{-1}\mathbf{H} \text{ for some } \mu \in \mathbb{R}\backslash\{0\}. \tag{4.1}$$

A pair of matrices which has this property can be characterised as follows[7].

Lemma 4.2. ($\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$) *Let* $\mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$ *be nonsingular and selfadjoint. Then* (4.1) *is satisfied if and only if there exists a nonsingular matrix* $\mathbf{S} \in \mathbb{F}^{n\times n}$ *such that*

$$\mathbf{S}^*\mathbf{HS} = \mathbf{I}_p \oplus -\mathbf{I}_q \oplus \mathbf{I}_r \oplus -\mathbf{I}_s \text{ and } \mathbf{S}^*\mathbf{GS} = \mu(\mathbf{I}_p \oplus -\mathbf{I}_q \oplus -\mathbf{I}_r \oplus \mathbf{I}_s)$$

for suitable constants $p, q, r, s \in \mathbb{N}$ *with* $p + q + r + s = n$.

Proof. [⇒]: Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a nonsingular matrix such that $\mathbf{A} = \mu^2\mathbf{A}^{-1}$ for some $\mu \in \mathbb{R}\backslash\{0\}$. Then $\mathbf{A}^2 = \mu^2\mathbf{I}$ so that the Jordan normal form of $\mathbf{A}$ must have the form

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{J} = \operatorname{diag}(\pm\mu).$$

In particular, if $\mathbf{A} = \mathbf{H}^{-1}\mathbf{G}$, it follows that

$$\begin{aligned}(\mathbf{P}^*\mathbf{HP})^{-1}(\mathbf{P}^*\mathbf{GP}) &= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{GP} = \mathbf{J} \\ &= \mathbf{J}^* = \mathbf{P}^*\mathbf{GH}^{-1}\mathbf{P}^{-*} = (\mathbf{P}^*\mathbf{GP})(\mathbf{P}^*\mathbf{HP})^{-1}.\end{aligned}$$

Thus the selfadjoint matrices $\mathbf{P}^*\mathbf{HP}$ and $\mathbf{P}^*\mathbf{GP}$ commute and can therefore be diagonalised simultaneously, so that an orthogonal or unitary matrix $\mathbf{Q}$ consisting of eigenvectors of $\mathbf{P}^*\mathbf{HP}$ (or $\mathbf{P}^*\mathbf{GP}$) can now be chosen for which

$$\mathbf{P}^*\mathbf{HP} = \mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^* \text{ and } \mathbf{P}^*\mathbf{GP} = \mathbf{Q}\mathbf{\Lambda}_G\mathbf{Q}^*$$

where $\mathbf{\Lambda}_H$, $\mathbf{\Lambda}_G$ are diagonal matrices containing the real eigenvalues. This means that

$$(\mathbf{\Lambda}_H^{-1}\mathbf{\Lambda}_G)^2 = \left(\mathbf{Q}^*(\mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^*)^{-1}(\mathbf{Q}\mathbf{\Lambda}_G\mathbf{Q}^*)\mathbf{Q}\right)^2 = (\mathbf{Q}^*\mathbf{JQ})^2 = \mu^2\mathbf{I}$$

[7] In the case $\mathbb{F} = \mathbb{C}$ the statement of the lemma also follows from Corollary 3.20. Here we have chosen to present a further proof which holds in the case $\mathbb{F} = \mathbb{R}$ as well.

and consequently $\mathbf{\Lambda}_H^{-1}\mathbf{\Lambda}_G$ can also be written in the form[8]

$$\mathbf{\Lambda}_H^{-1}\mathbf{\Lambda}_G = \mu\mathbf{\Sigma} \text{ with } \mathbf{\Sigma} = \operatorname{diag}(\pm 1).$$

Setting $\mathbf{\Lambda}_H = |\mathbf{\Lambda}_H|\mathbf{\Sigma}_H$, $\mathbf{\Lambda}_G = |\mathbf{\Lambda}_G|\mathbf{\Sigma}_G$, $\mu = \varepsilon|\mu|$, where $\mathbf{\Sigma}_H = \operatorname{sign}(\mathbf{\Lambda}_H)$, $\mathbf{\Sigma}_G = \operatorname{sign}(\mathbf{\Lambda}_G)$, $\varepsilon = \operatorname{sign}(\mu)$, we obtain

$$|\mathbf{\Lambda}_H|^{-1}|\mathbf{\Lambda}_G| = |\mu|\mathbf{I} \text{ and } \mathbf{\Sigma}_H\mathbf{\Sigma}_G = \varepsilon\mathbf{\Sigma}.$$

Hence, for $\mathbf{S} = \mathbf{PQ}|\mathbf{\Lambda}_H|^{-1/2}$ we finally have

$$\mathbf{S}^*\mathbf{HS} = (|\mathbf{\Lambda}_H|^{-1/2})^*\mathbf{Q}^*\mathbf{P}^*\mathbf{HPQ}|\mathbf{\Lambda}_H|^{-1/2} = |\mathbf{\Lambda}_H|^{-1/2}\mathbf{\Lambda}_H|\mathbf{\Lambda}_H|^{-1/2} = \mathbf{\Sigma}_H,$$
$$\mathbf{S}^*\mathbf{GS} = (|\mathbf{\Lambda}_H|^{-1/2})^*\mathbf{Q}^*\mathbf{P}^*\mathbf{GPQ}|\mathbf{\Lambda}_H|^{-1/2} = |\mathbf{\Lambda}_H|^{-1/2}\mathbf{\Lambda}_G|\mathbf{\Lambda}_H|^{-1/2} = \mu(\varepsilon\mathbf{\Sigma}_G).$$

The asserted form can always be obtained by suitable permutation. (The operations on $\mathbf{\Lambda}$ are to be applied to its diagonal elements.)

[$\Leftarrow$]: The assertion follows directly from $\mathbf{H}^{-1}\mathbf{G} = \mu\mathbf{S}(\mathbf{I}_{p+q} \oplus -\mathbf{I}_{r+s})\mathbf{S}^{-1}$ and $\mathbf{G}^{-1}\mathbf{H} = \mu^{-1}\mathbf{S}(\mathbf{I}_{p+q} \oplus -\mathbf{I}_{r+s})\mathbf{S}^{-1}$. □

Obviously, a (G,H)-polar decomposition of a matrix $\mathbf{A}$ can exist only if

$$\mathbf{H}^{-1}\mathbf{A}^*\mathbf{H} = \mathbf{H}^{-1}\mathbf{M}^*\mathbf{H}\mathbf{H}^{-1}\mathbf{U}^*\mathbf{H} = \mathbf{G}^{-1}\mathbf{M}^*\mathbf{G}\mathbf{G}^{-1}\mathbf{U}^*\mathbf{G} = \mathbf{G}^{-1}\mathbf{A}^*\mathbf{G}$$

or $\mathbf{A}^H = \mathbf{A}^G$. These matrices allow the following representation.

Lemma 4.3. ($\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$) *Let* $\mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$ *be nonsingular and selfadjoint such that* (4.1) *is satisfied and let* $\mathbf{A} \in \mathbb{F}^{n\times n}$ *such that* $\mathbf{A}^H = \mathbf{A}^G$. *Then there exists a nonsingular matrix* $\mathbf{S} \in \mathbb{F}^{n\times n}$ *such that*

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{A}_1 \oplus \mathbf{A}_2, \quad \mathbf{S}^*\mathbf{HS} = \mathbf{J}_1 \oplus \mathbf{J}_2, \quad \mathbf{S}^*\mathbf{GS} = \mu\mathbf{J}_1 \oplus -\mu\mathbf{J}_2,$$

where $\mathbf{A}_1 \in \mathbb{F}^{(p+q)\times(p+q)}$, $\mathbf{A}_2 \in \mathbb{F}^{(r+s)\times(r+s)}$ *and* $\mathbf{J}_1 = \mathbf{I}_p \oplus -\mathbf{I}_q$, $\mathbf{J}_2 = \mathbf{I}_r \oplus -\mathbf{I}_s$.

Proof. For the nonsingular matrix $\mathbf{S} \in \mathbb{F}^{n\times n}$ from Lemma 4.2, the matrices $\mathbf{S}^*\mathbf{HS}$ and $\mathbf{S}^*\mathbf{GS}$ take on the asserted form and $\mathbf{H}^{-1}\mathbf{G} = \mathbf{SFS}^{-1}$ where $\mathbf{F} = \mu\mathbf{I}_{p+q} \oplus -\mu\mathbf{I}_{r+s}$. According to the assumption $\mathbf{HAH}^{-1} = \mathbf{GAG}^{-1}$ we also have $\mathbf{F}(\mathbf{S}^{-1}\mathbf{AS}) = \mathbf{S}^{-1}(\mathbf{H}^{-1}\mathbf{GA})\mathbf{S} = \mathbf{S}^{-1}(\mathbf{AH}^{-1}\mathbf{G})\mathbf{S} = (\mathbf{S}^{-1}\mathbf{AS})\mathbf{F}$, which is possible only if $\mathbf{S}^{-1}\mathbf{AS}$ has the asserted form. □

If the matrix $\mathbf{A}$ satisfies $\mathbf{A}^H = \mathbf{A}^G$ and, furthermore, admits an H-polar decomposition, then although

$$\mathbf{G}^{-1}\mathbf{M}^*\mathbf{U}^*\mathbf{G} = \mathbf{H}^{-1}\mathbf{M}^*\mathbf{U}^*\mathbf{H} = \mathbf{H}^{-1}\mathbf{M}^*\mathbf{H}\mathbf{H}^{-1}\mathbf{U}^*\mathbf{H} = \mathbf{MU}^{-1}$$

or $\mathbf{M}^*\mathbf{U}^*\mathbf{GU} = \mathbf{GM}$, it cannot be concluded that the matrix also admits a G- or a (G,H)-polar decomposition. However, the following statement holds.

Lemma 4.4. ($\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$) *Let* $\mathbf{G}, \mathbf{H}, \mathbf{A}, \mathbf{S} \in \mathbb{F}^{n\times n}$ *be as in Lemma 4.3. Then* $\mathbf{A}$ *admits a (G,H)-polar decomposition if and only if* $\mathbf{A}_1$ *admits a* $\mathbf{J}_1$*-polar decomposition and* $\mathbf{A}_2$ *admits a* $\mathbf{J}_2$*-polar decomposition. Moreover, such a decomposition is H-semidefinite if and only if both* $\mathbf{J}_k$*-polar decompositions are semidefinite.*

Proof. Let $\mathbf{A} = \mathbf{UM}$ be a (G,H)-polar decomposition. Then $\mathbf{U}^H = \mathbf{U}^G$ and $\mathbf{M}^H = \mathbf{M}^G$ imply $\mathbf{S}^{-1}\mathbf{US} = \mathbf{U}_1 \oplus \mathbf{U}_2$ and $\mathbf{S}^{-1}\mathbf{MS} = \mathbf{M}_1 \oplus \mathbf{M}_2$, where the blocks $\mathbf{A}_k, \mathbf{J}_k, \mathbf{U}_k, \mathbf{M}_k$ have the same size ($k = 1, 2$). A simple calculation shows that $\mathbf{U}_k\mathbf{M}_k$ is a $\mathbf{J}_k$-polar decomposition of $\mathbf{A}_k$.

[8] The matrices $\mu\mathbf{\Sigma}$ and $\mathbf{J}$ have the same diagonal elements, but their ordering may be different.

If conversely $\mathbf{A}_1 = \mathbf{U}_1\mathbf{M}_1$ and $\mathbf{A}_2 = \mathbf{U}_2\mathbf{M}_2$ are given $\mathbf{J}_1$- and $\mathbf{J}_2$-polar decompositions, then these are also $(\mu\mathbf{J}_1)$- and $(-\mu\mathbf{J}_2)$-polar decompositions and therefore $\mathbf{A} = \mathbf{UM}$ with $\mathbf{U} = \mathbf{S}(\mathbf{U}_1 \oplus \mathbf{U}_2)\mathbf{S}^{-1}$ and $\mathbf{M} = \mathbf{S}(\mathbf{M}_1 \oplus \mathbf{M}_2)\mathbf{S}^{-1}$ is a (G,H)-polar decomposition.

The second part of the assertion follows from the fact that $\mathbf{HM} \geq 0$ if and only if $\mathbf{J}_k\mathbf{M}_k \geq 0$ for $k = 1, 2$. □

A useful application of this lemma is the next result which ensures the existence of a (G,H)-polar decomposition in an important particular case.

Lemma 4.5. ($\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$) *Let* $\mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$ *be nonsingular and self-adjoint such that* (4.1) *is satisfied and let* $\mathbf{A} \in \mathbb{F}^{n\times n}$ *such that* $\mathbf{A}^H = \mathbf{A}^G$. *If* $\mathbf{A} = \mathbf{UM}$ *is an H-polar decomposition with* $\sigma(\mathbf{M}) \subset \mathbb{C}^+ = \{z \in \mathbb{C} | \operatorname{Re}(z) > 0\}$, *then this is also a G-polar decomposition.*

Proof. Let $\mathbf{S}$ be as in Lemma 4.3. Then from $\sigma(\mathbf{S}^{-1}\mathbf{A}^H\mathbf{AS}) = \sigma(\mathbf{A}^H\mathbf{A}) = \sigma(\mathbf{M}^2) \subset \mathbb{C}\backslash(-\infty, 0]$ and

$$(\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-*})(\mathbf{S}^*\mathbf{A}^*\mathbf{S}^{-*})(\mathbf{S}^*\mathbf{HS})(\mathbf{S}^{-1}\mathbf{AS}) = \bigoplus_{k=1}^{2} \mathbf{J}_k^{-1}\mathbf{A}_k^*\mathbf{J}_k\mathbf{A}_k$$

it follows that $\sigma(\mathbf{A}_k^{J_k}\mathbf{A}_k) \subset \mathbb{C}\backslash(-\infty, 0]$ for $k = 1, 2$. Thus, according to Corollary 3.7, both blocks $\mathbf{A}_k$ admit a $\mathbf{J}_k$-polar decomposition $\mathbf{U}_k\mathbf{M}_k$ with $\sigma(\mathbf{M}_k) \subset \mathbb{C}^+$. Moreover, Lemma 4.4 implies that

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{M}} \text{ with } \tilde{\mathbf{U}} = \mathbf{S}(\mathbf{U}_1 \oplus \mathbf{U}_2)\mathbf{S}^{-1} \text{ and } \tilde{\mathbf{M}} = \mathbf{S}(\mathbf{M}_1 \oplus \mathbf{M}_2)\mathbf{S}^{-1}$$

is a (G,H)-polar decomposition with $\sigma(\tilde{\mathbf{M}}) \subset \mathbb{C}^+$. On the other hand, according to [PJ, Section 4], there exists one and only one matrix $\mathbf{M}$ for which $\mathbf{A}^H\mathbf{A} = \mathbf{M}^2$ and $\sigma(\mathbf{M}) \subset \mathbb{C}^+$, so that $\mathbf{M} = \tilde{\mathbf{M}}$ and thus also $\mathbf{U} = \tilde{\mathbf{U}}$ must be true. □

In conclusion of this preparatory section, the statements of the lemmas will be explained with the help of three examples. More general results on (G,H)-polar decompositions will be discussed in Section 4.6.

Example 4.6. Let $\mathbf{H} = \mathbf{I}_p \oplus \mathbf{I}_r$ and $\mathbf{G} = \mathbf{I}_p \oplus -\mathbf{I}_r$. Then a matrix $\mathbf{A} \in \mathbb{F}^{(p+r)\times(p+r)}$ for which $\mathbf{A}^H = \mathbf{A}^G$, according to Lemma 4.3, takes on the form $\mathbf{A} = \mathbf{A}_1 \oplus \mathbf{A}_2$, where $\mathbf{A}_1 \in \mathbb{F}^{p\times p}$ and $\mathbf{A}_2 \in \mathbb{F}^{r\times r}$. Let

$$\mathbf{A}_1 = \mathbf{P}_1\mathbf{\Sigma}_1\mathbf{Q}_1^* \text{ and } \mathbf{A}_2 = \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{Q}_2^*$$

be singular value decompositions and let

$$\mathbf{U} = \mathbf{P}_1\mathbf{Q}_1^* \oplus \mathbf{P}_2\mathbf{Q}_2^* \text{ and } \mathbf{M} = \mathbf{Q}_1\mathbf{\Sigma}_1\mathbf{Q}_1^* \oplus \mathbf{Q}_2\mathbf{\Sigma}_2\mathbf{Q}_2^*.$$

Then $\mathbf{A} = \mathbf{UM}$ is an H-semidefinite (G,H)-polar decomposition. ◊

Example 4.7. Let $\alpha, \beta, \mu \in \mathbb{R}$ with $\mu \neq 0$ and let $\mathbf{H} = \operatorname{diag}(1, -1, 1, -1)$ and $\mathbf{G} = \mu \operatorname{diag}(1, -1, -1, 1)$. The matrix

$$\mathbf{A}_1 = \begin{bmatrix} 0 & \beta \\ \alpha & 0 \end{bmatrix} \oplus \begin{bmatrix} 0 & \alpha \\ \beta & 0 \end{bmatrix}$$

satisfies $\mathbf{A}_1^H\mathbf{A}_1 = \mathbf{A}_1^G\mathbf{A}_1 = \mathrm{diag}(-\alpha^2, -\beta^2, -\beta^2, -\alpha^2)$ and admits the H-polar decomposition

$$\mathbf{A}_1 = \mathbf{U}_1\mathbf{M}_1 \text{ with } \mathbf{U}_1 = \begin{bmatrix} & & -i & 0 \\ & & 0 & -i \\ -i & 0 & & \\ 0 & -i & & \end{bmatrix}, \ \mathbf{M}_1 = \begin{bmatrix} & & 0 & i\alpha \\ & & i\beta & 0 \\ 0 & i\beta & & \\ i\alpha & 0 & & \end{bmatrix}.$$

But it is not a G-polar decomposition because $\mathbf{U}_1^*\mathbf{G}\mathbf{U}_1 = -\mathbf{G}$ and $\mathbf{M}_1^*\mathbf{G} = -\mathbf{G}\mathbf{M}_1$. In fact when $\alpha \neq \beta$, the matrix pair $(\mathbf{A}_1^G\mathbf{A}_1, \mathbf{G})$, which is already in canonical form, does not satisfy the condition 1. of Theorem 3.4. So $\mathbf{A}_1$ does not have any G-polar decompositions in this case. The matrix

$$\mathbf{A}_2 = \begin{bmatrix} 0 & \beta \\ \alpha & 0 \end{bmatrix} \oplus \begin{bmatrix} 0 & \beta \\ \alpha & 0 \end{bmatrix}$$

satisfies $\mathbf{A}_2^H\mathbf{A}_2 = \mathbf{A}_2^G\mathbf{A}_2 = \mathrm{diag}(-\alpha^2, -\beta^2, -\alpha^2, -\beta^2)$ and admits the G-polar decomposition

$$\mathbf{A}_2 = \mathbf{U}_2\mathbf{M}_2 \text{ with } \mathbf{U}_2 = \begin{bmatrix} & & 0 & -i \\ & & -i & 0 \\ 0 & -i & & \\ -i & 0 & & \end{bmatrix}, \ \mathbf{M}_2 = \begin{bmatrix} & & i\alpha & 0 \\ & & 0 & i\beta \\ i\alpha & 0 & & \\ 0 & i\beta & & \end{bmatrix}.$$

But it is not an H-polar decomposition because $\mathbf{U}_2^*\mathbf{H}\mathbf{U}_2 = -\mathbf{H}$ and $\mathbf{M}_2^*\mathbf{H} = -\mathbf{H}\mathbf{M}_2$. Again when $\alpha \neq \beta$, the matrix pair $(\mathbf{A}_2^H\mathbf{A}_2, \mathbf{H})$, which is already in canonical form, does not satisfy the condition 1. of Theorem 3.4. So $\mathbf{A}_2$ does not have any H-polar decompositions in this case. But if $\alpha = \beta$, then $\mathbf{A} = \mathbf{A}_1 = \mathbf{A}_2$ admits the (G,H)-polar decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{M} \text{ with } \mathbf{U} = \begin{bmatrix} -i & \\ & -i \end{bmatrix} \oplus \begin{bmatrix} -i & \\ & -i \end{bmatrix}, \ \mathbf{M} = \begin{bmatrix} & i\alpha \\ i\alpha & \end{bmatrix} \oplus \begin{bmatrix} & i\alpha \\ i\alpha & \end{bmatrix}$$

which evidently satisfies Lemma 4.4. ◊

Example 4.8. Let $\mathbf{G}, \mathbf{H}$ be matrices with (4.1) and let $\mathbf{A}$ be a matrix with $\mathbf{A}^H = \mathbf{A}^G$. If $\mathbf{H}$ is positive definite and $\mathbf{A}$ nonsingular, then there exists a definite H-polar decomposition which, according to Lemma 4.5, is a G-polar decomposition too. However, if $\mathbf{A}$ is singular or $\mathbf{H}$ is indefinite, this may not be always true. Consider the (semi)definite H-polar decompositions

$$\mathbf{H}_1 = \mathrm{diag}(1,1,1), \ \ \mathbf{G}_1 = \mathrm{diag}(1,1,-1), \ \ x \in \mathbb{R},$$
$$\mathbf{A}_1 = \begin{bmatrix} \cos(x) & 0 & 0 \\ \sin(x) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \cos(x) & 0 & -\sin(x) \\ \sin(x) & 0 & \cos(x) \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{U}_1\mathbf{M}_1$$

where $\sigma(\mathbf{H}_1\mathbf{M}_1) = \{0, 1\}$ and

$$\mathbf{H}_2 = \mathrm{diag}(1,-1), \ \ \mathbf{G}_2 = \mathrm{diag}(1,1), \ \ a > |b| > 0, \ \ u = b/a,$$
$$\mathbf{A}_2 = \begin{bmatrix} -\sqrt{a^2-b^2} & 0 \\ 0 & \sqrt{a^2-b^2} \end{bmatrix} = \frac{-1}{\sqrt{1-u^2}} \begin{bmatrix} 1 & u \\ u & 1 \end{bmatrix} \begin{bmatrix} a & b \\ -b & -a \end{bmatrix} = \mathbf{U}_2\mathbf{M}_2.$$

where $\sigma(\mathbf{H}_2\mathbf{M}_2) = \{a \pm b\}$. Here $\mathbf{U}_1^*\mathbf{G}_1\mathbf{U}_1 = \mathrm{diag}(1,-1,1) \neq \mathbf{G}_1$ and neither $\mathbf{U}_2$ is orthogonal nor $\mathbf{M}_2$ symmetric, so that both factorisations are not G-polar decompositions. In contrast to this, the "blockwise" (semi)definite H-polar decompositions

$$\mathbf{A}_1 = \left(\begin{bmatrix} \cos(x) & -\sin(x) \\ \sin(x) & \cos(x) \end{bmatrix} \oplus 1 \right) \mathbf{M}_1 \quad \text{and} \quad \mathbf{A}_2 = (-\mathbf{I}_2)(-\mathbf{A}_2),$$

according to Lemma 4.4, are also G-polar decompositions. ◊

With this background on (G,H)-polar decompositions we are now able to investigate the problems stated in the introduction, starting with the determination of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and an indefinite scalar product $[.,.]$ such that $[\mathbf{x}_k - \mathbf{x}_l, \mathbf{x}_k - \mathbf{x}_l] = q_{kl}$ for given real numbers q_{kl} $(1 \leq k, l \leq N)$.

4.3 Construction of vectors from values of a quadratic form

The construction of vectors from given values of a quadratic form presented in this section is a generalisation of the work [YH] for complex vector spaces and indefinite scalar products.

Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $[.,.]$ be an indefinite scalar product in $\mathbb{F}^n$ with the underlying nonsingular symmetric or Hermitian matrix $\mathbf{H} \in \mathbb{F}^{n\times n}$. Then for arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ in the case $\mathbb{F} = \mathbb{R}$ it is true that

$$[\mathbf{x}, \mathbf{y}] = \frac{1}{2}([\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] - [\mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}]) \tag{4.2a}$$

and in the case $\mathbb{F} = \mathbb{C}$ we have

$$\begin{aligned} \mathrm{Re}[\mathbf{x}, \mathbf{y}] &= \frac{1}{2}([\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] - [\mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}]) \\ &= \frac{1}{2}([\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] - [i\mathbf{y} - i\mathbf{x}, i\mathbf{y} - i\mathbf{x}]), \\ \mathrm{Im}[\mathbf{x}, \mathbf{y}] &= \frac{1}{2}([\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] - [\mathbf{x} - i\mathbf{y}, \mathbf{x} - i\mathbf{y}]) \\ &= -\frac{1}{2}([\mathbf{x}, \mathbf{x}] + [\mathbf{y}, \mathbf{y}] - [\mathbf{y} - i\mathbf{x}, \mathbf{y} - i\mathbf{x}]), \end{aligned} \tag{4.2b}$$

so that the scalar products of the vectors can be expressed in terms of the quadratic form $\Phi(\mathbf{x}) = [\mathbf{x}, \mathbf{x}]$.

Now let $N \geq n$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{F}^n$ be given and let $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_N] \in \mathbb{F}^{n\times N}$ be a matrix whose columns are these vectors. Then

$$\mathbf{W} = \mathbf{X}^*\mathbf{H}\mathbf{X}$$

is the Gramian matrix of the $\mathbf{x}_k$. Therefore, if $\mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} = \mathbb{F}^n$, then the number of positive and negative eigenvalues of $\mathbf{H}$ and $\mathbf{W}$ are equal, and furthermore the eigenvalue 0 appears in $\sigma(\mathbf{W})$ with the multiplicity $N - n$ (Sylvester's law of inertia, [GR, Chapter IX, §2]). Moreover, the elements $w_{kl} =$

$[\mathbf{x}_l, \mathbf{x}_k]$ of the matrix $\mathbf{W}$ according to (4.2) can be expressed in the form

$$w_{kl} = \frac{1}{2}(\rho_k + \rho_l - \sigma_{kl}) \text{ if } \mathbb{F} = \mathbb{R} \text{ or} \tag{4.3a}$$

$$w_{kl} = \frac{1}{2}(\rho_k + \rho_l - \sigma_{kl}) + \frac{i}{2}(\rho_k + \rho_l - \tau_{kl}) \text{ if } \mathbb{F} = \mathbb{C}. \tag{4.3b}$$

where

$$\rho_k = [\mathbf{x}_k, \mathbf{x}_k], \quad \sigma_{kl} = [\mathbf{x}_l - \mathbf{x}_k, \mathbf{x}_l - \mathbf{x}_k], \quad \tau_{kl} = [\mathbf{x}_l - i\mathbf{x}_k, \mathbf{x}_l - i\mathbf{x}_k] \tag{4.4}$$

$$\text{with } \rho_k, \sigma_{kl}, \tau_{kl} \in \mathbb{R}, \quad \sigma_{kl} = \sigma_{lk}, \quad \sigma_{kk} = 0, \quad \tau_{kl} + \tau_{lk} = 2(\rho_k + \rho_l) \tag{4.5}$$

for $1 \leq k, l \leq N$.

Conversely, let the real numbers $\rho_k, \sigma_{kl}, \tau_{kl}$ with (4.5) be given, and let the elements of a matrix $\mathbf{W}$ be defined by (4.3). Then this matrix is symmetric or Hermitian, respectively, and can therefore be written in the form

$$\mathbf{W} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^*.$$

Here $\mathbf{\Lambda}$ is a diagonal matrix of the real eigenvalues $\lambda_1, \ldots, \lambda_N$ of $\mathbf{W}$ and $\mathbf{R} = [\mathbf{r}_1 \ldots \mathbf{r}_N]$ is a matrix whose columns form a basis of $\mathbb{F}^N$ consisting of orthonormalised eigenvectors. Now if p is the number of positive and $n - p$ is the number of negative eigenvalues and if it is assumed that

$$\lambda_1, \ldots, \lambda_p > 0, \quad \lambda_{p+1}, \ldots, \lambda_n < 0 \text{ and } \lambda_{n+1} = \ldots = \lambda_N = 0,$$

then the matrices defined by

$$\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_p, \lambda_{p+1}, \ldots, \lambda_n) \text{ and } \mathbf{R}_1 = [\mathbf{r}_1 \ldots \mathbf{r}_n]$$

satisfy $\mathbf{W} = \mathbf{R}_1\mathbf{\Lambda}_1\mathbf{R}_1^*$ too. Consequently, if we set

$$\mathbf{\Sigma} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}, \sqrt{-\lambda_{p+1}}, \ldots, \sqrt{-\lambda_n}) \text{ and } \mathbf{H}_w = \mathbf{I}_p \oplus -\mathbf{I}_{n-p},$$

then the matrix

$$\mathbf{X} = \mathbf{\Sigma}^*\mathbf{R}_1^* \in \mathbb{F}^{n \times N}$$

fulfills on the one hand $\operatorname{rank}\mathbf{X} = n$ and on the other hand $\mathbf{X}^*\mathbf{H}_w\mathbf{X} = \mathbf{R}_1\mathbf{\Sigma}\mathbf{H}_w\mathbf{\Sigma}^*\mathbf{R}_1^* = \mathbf{R}_1\mathbf{\Lambda}_1\mathbf{R}_1^* = \mathbf{W}$. Therefore the columns $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{F}^n$ of $\mathbf{X}$ constitute a spanning set (or system of generators) for $\mathbb{F}^n$, and for the indefinite scalar product defined by $[\mathbf{x}, \mathbf{y}]_w = (\mathbf{H}_w\mathbf{x}, \mathbf{y})$ it is true that $w_{kl} = [\mathbf{x}_l, \mathbf{x}_k]_w$. This means that also

$$\begin{aligned}
[\mathbf{x}_k, \mathbf{x}_k]_w &= w_{kk} = \rho_k, \\
[\mathbf{x}_l - \mathbf{x}_k, \mathbf{x}_l - \mathbf{x}_k]_w &= w_{kk} + w_{ll} - w_{kl} - w_{lk} = \sigma_{kl}, \\
[\mathbf{x}_l - i\mathbf{x}_k, \mathbf{x}_l - i\mathbf{x}_k]_w &= w_{kk} + w_{ll} + iw_{kl} - iw_{lk} = \tau_{kl} \quad (\text{if } \mathbb{F} = \mathbb{C}),
\end{aligned}$$

so that the given numbers are values of the quadratic form $\Phi_w(\mathbf{x}) = [\mathbf{x}, \mathbf{x}]_w$ for particular combinations of the constructed vectors. We thus have proved the following theorem.

Theorem 4.9 (Construction of vectors). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let ρ_k, σ_{kl} be real numbers such that $\sigma_{kl} = \sigma_{lk}$ and $\sigma_{kk} = 0$ for all k, l in $\{1, \ldots, N\}$. Furthermore, for the case $\mathbb{F} = \mathbb{C}$ let τ_{kl} be real numbers such that $\tau_{kl} + \tau_{lk} = 2(\rho_k + \rho_l)$ for all k, l in $\{1, \ldots, N\}$. Then the following statements are equivalent:*

(*i*) *There exist vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{F}^n$ *constituting a spanning set for* $\mathbb{F}^n$, *for which* $[\mathbf{x}_k, \mathbf{x}_k] = \rho_k$ *as well as* $[\mathbf{x}_l - \mathbf{x}_k, \mathbf{x}_l - \mathbf{x}_k] = \sigma_{kl}$, *and in the case* $\mathbb{F} = \mathbb{C}$ *also* $[\mathbf{x}_l - i\mathbf{x}_k, \mathbf{x}_l - i\mathbf{x}_k] = \tau_{kl}$ *is satisfied. Thereby* $[.,.]$ *is an indefinite scalar product in* $\mathbb{F}^n$ *with underlying nonsingular symmetric or Hermitian matrix* $\mathbf{H} \in \mathbb{F}^{n\times n}$ *which has* p *positive eigenvalues.*

(*ii*) *The symmetric or Hermitian matrix* $\mathbf{W} \in \mathbb{F}^{N\times N}$ *whose elements* w_{kl} *are defined by* (4.3) *has* p *positive and* $n - p$ *negative eigenvalues, and the eigenvalue* 0 *appears with multiplicity* $N - n$.

For the case of a Euclidean or unitary space we immediately obtain the following corollary in which $\|.\|$ denotes the Euclidean norm.

Corollary 4.10. *Let* $\rho_k, \sigma_{kl}, \tau_{kl} \geq 0$ *be as in Theorem 4.9. Then there exist vectors* $\mathbf{x}_k$ *such that* $\|\mathbf{x}_k\| = \sqrt{\rho_k}$, $\|\mathbf{x}_l - \mathbf{x}_k\| = \sqrt{\sigma_{kl}}$, *and in the case* $\mathbb{F} = \mathbb{C}$ *also* $\|\mathbf{x}_l - i\mathbf{x}_k\| = \sqrt{\tau_{kl}}$ *if and only if the matrix* $\mathbf{W}$ *is positive semidefinite.*

Let $\mathbb{F} = \mathbb{R}$, $N = 2$ and $\rho_1, \rho_2, \sigma_{12} \geq 0$. Then

$$\begin{aligned}\det \mathbf{W} &= \frac{1}{2}(\rho_1\rho_2 + \rho_1\sigma_{12} + \rho_2\sigma_{12}) - \frac{1}{4}(\rho_1^2 + \rho_2^2 + \sigma_{12}^2)\\ &= \frac{1}{4}\big(\sigma_{12} - (\sqrt{\rho_1} - \sqrt{\rho_2})^2\big)\big((\sqrt{\rho_1} + \sqrt{\rho_2})^2 - \sigma_{12}\big)\end{aligned}$$

and this determinant is non-negative if and only if

$$|\sqrt{\rho_1} - \sqrt{\rho_2}| \leq \sqrt{\sigma_{12}} \text{ and } \sqrt{\sigma_{12}} \leq \sqrt{\rho_1} + \sqrt{\rho_2}.$$

But this is just the triangle inequality, so that Corollary 4.10 gives a generalisation of this essential property of Euclidean geometry.

In addition to these investigations concerning the geometrical properties of the vectors $\mathbf{x}_k$, the consideration of their physical properties provides some useful information for the application of Theorem 4.9 in MDS.

Remark 4.11 (Tensor of inertia). On interpreting the vectors $\mathbf{x}_k = (x_k^\alpha)$ constructed in Theorem 4.9 as the locations of point objects of mass 1, the matrix

$$\mathbf{T} = \mathbf{X}\mathbf{X}^*, \;\; \mathbf{T} = \big[T^{\alpha\beta}\big] \;\text{ with }\; T^{\alpha\beta} = \sum_{k=1}^{N} x_k^\alpha \overline{x}_k^\beta \;\text{ for }\; 1 \leq \alpha, \beta \leq n$$

gives their (contravariant) tensor of inertia in the sense of Hermann Weyl [WEY, §6][9]. Here $\mathbf{T} = \mathbf{\Sigma}^*\mathbf{R}_1^*\mathbf{R}_1\mathbf{\Sigma} = \mathbf{\Sigma}^2 = \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_n|)$ is a diagonal matrix, so that the axes of the coordinate system are also the inertial axes (principal axes) of the constellation. Moreover, the absolute values of the eigenvalues are the associated (contravariant) moments of inertia. From the viewpoint of MDS this means that the coordinates of the vectors can be interpreted, as usual, as the ratings of uncorrelated factors [BG, Section 7.10]. In addition to this, the space-like, time-like or light-like property of the vectors $\mathbf{x}_k$ and the canonical basis vectors $\mathbf{e}_\alpha$ may also provide some useful information. ♢

[9] Weyl's definition slightly differs from the definition given in the textbooks of classical physics. Nevertheless, it is more reasonable when considering the rotational motion in n-dimensional spaces.

Remark 4.12 (Centroid). Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^n$ be real vectors whose centroid lies at the coordinates' origin, i.e. $\sum_k \mathbf{x}_k = \mathbf{0}$, and let $\Phi(\mathbf{x}) = [\mathbf{x}, \mathbf{x}]$. Then the scalar products satisfy

$$[\mathbf{x}_l, \mathbf{x}_k] = \frac{1}{2}\Bigg(\frac{1}{N}\sum_{j=1}^{N} \Phi(\mathbf{x}_k - \mathbf{x}_j) + \frac{1}{N}\sum_{i=1}^{N} \Phi(\mathbf{x}_i - \mathbf{x}_l) \\ - \Phi(\mathbf{x}_k - \mathbf{x}_l) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \Phi(\mathbf{x}_i - \mathbf{x}_j) \Bigg),$$

as can easily be verified [T]. Conversely, let the real numbers $\sigma_{kl} = \sigma_{lk}$, $\sigma_{kk} = 0$, $1 \leq k, l \leq N$ be given. Then

$$w_{kl} = \frac{1}{2}\Bigg(\frac{1}{N}\sum_j \sigma_{kj} + \frac{1}{N}\sum_i \sigma_{il} - \sigma_{kl} - \frac{1}{N^2}\sum_i\sum_j \sigma_{ij} \Bigg)$$

defines the elements of a symmetric matrix $\mathbf{W}$ whose row and column sums vanish. Using the method of Theorem 4.9 again vectors $\mathbf{x}_k$ and an indefinite scalar product can be constructed such that $w_{kl} = [\mathbf{x}_l, \mathbf{x}_k]$. But now the centroid of these vectors lies at the origin. An analogous construction also applies in the complex case, but the conditions that must be assumed for the values τ_{kl} are rather complicated there. ♢

Remark 4.13 (Approximation). Assume that the p positive and $q = n - p$ negative eigenvalues of $\mathbf{W}$ are sorted such that

$$\lambda_1 \geq \ldots \geq \lambda_p > 0 > \lambda_{p+1} \geq \ldots \lambda_{p+q}$$

when defining $\mathbf{H} = \mathbf{I}_p \oplus -\mathbf{I}_q$ and $\mathbf{X}$. Moreover, let the columns of $\mathbf{X}^*$ be denoted by $\mathbf{u}_1, \ldots, \mathbf{u}_p; \mathbf{v}_q, \ldots, \mathbf{v}_1$, so that $\mathbf{u}_1$ belongs to a maximal positive and $\mathbf{v}_1$ belongs to a minimal negative eigenvalue. Then for $\mathbf{H}' = \mathbf{I}_r \oplus -\mathbf{I}_s$ and

$$(\mathbf{X}')^* = \begin{cases} [\mathbf{u}_1 \ldots \mathbf{u}_r; \mathbf{v}_s \ldots \mathbf{v}_1], & \text{if } r \leq p \text{ and } s \leq q \\ [\mathbf{u}_1 \ldots \mathbf{u}_r; \mathbf{0}_s \ldots \mathbf{0}_{q+1}\, \mathbf{v}_q \ldots \mathbf{v}_1], & \text{if } r \leq p \text{ and } s > q \\ [\mathbf{u}_1 \ldots \mathbf{u}_p\, \mathbf{0}_{p+1} \ldots \mathbf{0}_r; \mathbf{v}_s \ldots \mathbf{v}_1], & \text{if } r > p \text{ and } s \leq q \\ [\mathbf{u}_1 \ldots \mathbf{u}_p\, \mathbf{0}_{p+1} \ldots \mathbf{0}_r; \mathbf{0}_s \ldots \mathbf{0}_{q+1}\, \mathbf{v}_q \ldots \mathbf{v}_1], & \text{if } r > p \text{ and } s > q \end{cases}$$

it holds that

$$\mathbf{X}^*\mathbf{H}\mathbf{X} = (\mathbf{X}')^*(\mathbf{H}')(\mathbf{X}') + \mathbf{E}$$

where the residual matrix $\mathbf{E}^* = \mathbf{E} \in \mathbb{F}^{N \times N}$ satisfies

$$\|\mathbf{E}\|_F^2 = \sum_{\alpha=r+1}^{n-s} \lambda_\alpha^2.$$

In other words, if $r < p$ and $s < q$ but the magnitudes of the eigenvalues $\lambda_{r+1}, \ldots, \lambda_{n-s}$ are small, then it is still true that

$$\mathbf{X}^*\mathbf{H}\mathbf{X} \approx (\mathbf{X}')^*(\mathbf{H}')(\mathbf{X}'),$$

so that the matrices $\mathbf{X}'$ and $\mathbf{H}'$ form an $r+s$-dimensional approximation of the matrices $\mathbf{X}$ and $\mathbf{H}$. On the other hand, if $r \geq p$ and $s \geq q$, then it is always true that

$$\mathbf{X}^*\mathbf{H}\mathbf{X} = (\mathbf{X}')^*(\mathbf{H}')(\mathbf{X}').$$

Consequently, if $\mathbf{X}_1 \in \mathbb{F}^{(p_1+q_1)\times N}$ and $\mathbf{X}_2 \in \mathbb{F}^{(p_2+q_2)\times N}$ fulfil

$$\begin{aligned}\mathbf{W}_1 &= \mathbf{X}_1^*\mathbf{H}_1\mathbf{X}_1, \quad \mathbf{H}_1 = \mathbf{I}_{p_1} \oplus -\mathbf{I}_{q_1},\\ \mathbf{W}_2 &= \mathbf{X}_2^*\mathbf{H}_2\mathbf{X}_2, \quad \mathbf{H}_2 = \mathbf{I}_{p_2} \oplus -\mathbf{I}_{q_2},\end{aligned}$$

then it is always possible to choose matrices such that

$$\mathbf{W}_1 = (\mathbf{X}_1')^*\mathbf{H}(\mathbf{X}_1'), \;\; \mathbf{W}_2 = (\mathbf{X}_2')^*\mathbf{H}(\mathbf{X}_2'), \;\; \mathbf{H} = \mathbf{I}_{\max(p_1,p_2)} \oplus -\mathbf{I}_{\max(q_1,q_2)}.$$

Without loss of generality it can therefore be assumed that two constellations of vectors constructed from values of a quadratic form are embedded in a common indefinite scalar product space. ◊

4.4 Solution of the H-isometric Procrustes problem

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{F}^n$ be the vectors and let $[.,.] = (\mathbf{H}.,.)$ be the indefinite scalar product constructed from given scalars $\rho_k, \sigma_{kl}, \tau_{kl}$ according to Theorem 4.9, so that (4.4) holds. For every H-isometry $\mathbf{U} \in \mathbb{F}^{n\times n}$ it then follows that

$$[\mathbf{U}\mathbf{x}_l, \mathbf{U}\mathbf{x}_k] = [\mathbf{x}_l, \mathbf{x}_k] = w_{kl}$$

which can also be expressed in matrix equation form

$$\mathbf{X}^*\mathbf{U}^*\mathbf{H}\mathbf{U}\mathbf{X} = \mathbf{X}^*\mathbf{H}\mathbf{X} = \mathbf{W}.$$

Thus the columns $\mathbf{x}_k' = \mathbf{U}\mathbf{x}_k$ contained in the matrix $\mathbf{X}' = \mathbf{U}\mathbf{X}$ satisfy (4.4), too. Now assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are the vectors constructed from two measurements of a quadratic form. Then on comparing the constellations the question arises, what part of the observed differences is due to different positions in space, and what part is due to actual differences in the inner structure of the constellations. Expressed mathematically, the task is to determine an H-isometry $\mathbf{U} \in \mathbb{F}^{n\times n}$ which solves the optimisation problem

$$f(\mathbf{U}) = \sum_{k=1}^{N} [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k] \to \begin{cases} \min, & \text{if } \mathbf{H} > 0 \\ \max, & \text{if } \mathbf{H} < 0 \\ \min\,/\,\max, & \text{otherwise} \end{cases}, \tag{4.6a}$$

$$\mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H} = \mathbf{0}.$$

The sum of scalar products arising therein can be expressed in the form of a trace, so that an alternative expression with

$$f(\mathbf{U}) = \mathrm{tr}[(\mathbf{U}\mathbf{X} - \mathbf{Y})^*\mathbf{H}(\mathbf{U}\mathbf{X} - \mathbf{Y})] \tag{4.6b}$$

is given, where as above $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_N]$. Moreover, $\mathbf{H} < 0$ ($\mathbf{H} > 0$) stands for a positive (negative) definite matrix $\mathbf{H}$ and the symbol "min / max" stands for a particular saddle point, which will be explained more precisely below. Within the scope of Euclidean vector spaces a solution of this problem was found in [S] where it was called the orthogonal Procrustes problem ($\mathbb{F} = \mathbb{R}$, $\mathbf{H} = \mathbf{I}$). In the present context of indefinite scalar products it is furthermore called the H-orthogonal or H-unitary Procrustes problem.

The fact, that the addends in (4.6) can be positive as well as negative, whenever $\mathbf{H}$ is indefinite, causes severe difficulties. On first sight one may thus get the idea to avoid these difficulties by minimising one of the non-negative functions

$$f_1(\mathbf{U}) = f(\mathbf{U})^2 \geq 0 \;\text{ or }\; f_2(\mathbf{U}) = \textstyle\sum_k [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]^2 \geq 0.$$

But the example $\mathbf{H} = \mathrm{diag}(1,-1)$, $\mathbf{U}\mathbf{x}_k = (\xi,\xi)^T$, $\mathbf{y}_k = (\eta,\eta)^T$, i.e.

$$[\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k] = |\xi-\eta|^2 - |\xi-\eta|^2 = 0,$$

shows an addend which neither in f_1 nor in f_2 makes a contribution to the result although $|\xi - \eta|$ may be arbitrarily large. However, the intention of the optimisation is to converge the constellations in the sense of an optimum congruence which means, that the coordinate differences should become small. A first possibility to reach this goal is to measure the differences with a definite scalar product, e.g. $\|\mathbf{U}\mathbf{x}_k - \mathbf{y}_k\|^2$. This approach will be discussed in the next section. A further possibility is not to look for a minimum or maximum of the function f, but to determine a particular saddle point "min /max" where the coordinate differences are small. This is the subject of the following investigations.

Considering the case $\mathbb{F} = \mathbb{R}$ first and introducing a matrix of the (unknown) Lagrange multipliers $\mathbf{L} \in \mathbb{R}^{n\times n}$, the constraints can be stated in the form

$$h_L(\mathbf{U}) = \mathrm{tr}[\mathbf{L}(\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H})]$$

and the necessary first order condition for solving the problem is

$$\frac{\partial}{\partial \mathbf{U}}(f + h_L) = \mathbf{0}.$$

Differentiation of the trace [DP] gives

$$\frac{\partial f}{\partial \mathbf{U}} = 2\mathbf{H}\mathbf{U}\mathbf{X}\mathbf{X}^* - 2\mathbf{H}\mathbf{Y}\mathbf{X}^* \;\text{ and }\; \frac{\partial h_L}{\partial \mathbf{U}} = \mathbf{H}\mathbf{U}(\mathbf{L} + \mathbf{L}^*),$$

so that $\mathbf{U}$ must satisfy the equation

$$\mathbf{U}\mathbf{X}\mathbf{X}^*\mathbf{H} + \mathbf{U}\mathbf{\Lambda}\mathbf{H} = \mathbf{Y}\mathbf{X}^*\mathbf{H} \;\text{ with }\; \mathbf{\Lambda} = \frac{\mathbf{L}+\mathbf{L}^*}{2} = \mathbf{\Lambda}^*. \tag{4.7}$$

Now defining $\mathbf{M} = (\mathbf{X}\mathbf{X}^* + \mathbf{\Lambda})\mathbf{H}$, the necessary condition becomes

$$\mathbf{A} = \mathbf{U}\mathbf{M} \;\text{ with }\; \mathbf{A} = \mathbf{Y}\mathbf{X}^*\mathbf{H} \;\text{ and }\; \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H},\;\; \mathbf{M}^*\mathbf{H} = \mathbf{H}\mathbf{M}. \tag{4.8}$$

Thus, if a solution of the problem exists, it can be determined by an H-polar decomposition of the matrix $\mathbf{A}$. (The question which of the H-isometries contained in such an H-polar decomposition actually are solutions of the problem will be discussed after the complex case is complete.)

In the case $\mathbb{F} = \mathbb{C}$ the complex derivatives of f and h_L do not exist. However, the necessity for (4.8) can be shown by determining the real derivatives. For this, let the real and imaginary part of the matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ be denoted by $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Then the well-known linear map $T : \mathbb{C}^{m\times n} \to \mathbb{R}^{2m\times 2n}$,

$$T(\mathbf{A}) = \mathbf{Q}_{2m}^* \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{A}} \end{bmatrix} \mathbf{Q}_{2n} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ -\mathbf{A}_2 & \mathbf{A}_1 \end{bmatrix} \;\text{ where }\; \mathbf{Q}_{2n} = \frac{\sqrt{2}}{2}\begin{bmatrix} \mathbf{I}_n & -i\mathbf{I}_n \\ i\mathbf{I}_n & -\mathbf{I}_n \end{bmatrix},$$

allows the real representation $\mathbf{A}^\wedge = T(\mathbf{A})$ of $\mathbf{A}$. Moreover, for every Hermitian matrix $\mathbf{A}$ it is true that $2\,\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^\wedge)$ which follows from the unitarity of $\mathbf{Q}_{2n}$. Therefore, the objective function can be represented as

$$2f(\mathbf{U}) = f(\mathbf{U}^\wedge) = \mathrm{tr}[(\mathbf{U}^\wedge\mathbf{X}^\wedge - \mathbf{Y}^\wedge)^T\mathbf{H}^\wedge(\mathbf{U}^\wedge\mathbf{X}^\wedge - \mathbf{Y}^\wedge)]$$

having the real derivatives

$$\frac{\partial f(\mathbf{U}^\wedge)}{\partial \mathbf{U}^\wedge} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{U}_1} & \frac{\partial f}{\partial \mathbf{U}_2} \\ -\frac{\partial f}{\partial \mathbf{U}_2} & \frac{\partial f}{\partial \mathbf{U}_1} \end{bmatrix} = 2\mathbf{H}^\wedge\mathbf{U}^\wedge\mathbf{X}^\wedge(\mathbf{X}^\wedge)^T - 2\mathbf{H}^\wedge\mathbf{Y}^\wedge(\mathbf{X}^\wedge)^T.$$

The transformation of the constraints

$$\mathbf{h}_1(\mathbf{U}) = \mathrm{Re}(\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H}) = \mathbf{0} \;\text{ and }\; \mathbf{h}_2(\mathbf{U}) = \mathrm{Im}(\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H}) = \mathbf{0}$$

is more complicated. Introducing Lagrange multipliers $\mathbf{L}_1, \mathbf{L}_2 \in \mathbb{R}^{n\times n}$ and using $\mathbf{H} = \mathbf{H}_1 + i\mathbf{H}_2$, $\mathbf{U} = \mathbf{U}_1 + i\mathbf{U}_2$ we obtain

$$\begin{aligned} h_{L,1}(\mathbf{U}) &= \mathrm{tr}[\mathbf{L}_1(\mathbf{U}_1^T\mathbf{H}_1\mathbf{U}_1 - \mathbf{U}_1^T\mathbf{H}_2\mathbf{U}_2 + \mathbf{U}_2^T\mathbf{H}_1\mathbf{U}_2 + \mathbf{U}_2^T\mathbf{H}_2\mathbf{U}_1 - \mathbf{H}_1)], \\ h_{L,2}(\mathbf{U}) &= \mathrm{tr}[\mathbf{L}_2(\mathbf{U}_1^T\mathbf{H}_1\mathbf{U}_2 + \mathbf{U}_1^T\mathbf{H}_2\mathbf{U}_1 - \mathbf{U}_2^T\mathbf{H}_1\mathbf{U}_1 + \mathbf{U}_2^T\mathbf{H}_2\mathbf{U}_2 - \mathbf{H}_2)], \end{aligned}$$

from which it follows that

$$\begin{aligned} \frac{\partial h_{L,1}}{\partial \mathbf{U}_1} &= (\mathbf{H}_1\mathbf{U}_1 - \mathbf{H}_2\mathbf{U}_2)(\mathbf{L}_1 + \mathbf{L}_1^T), \qquad \frac{\partial h_{L,1}}{\partial \mathbf{U}_2} = (\mathbf{H}_1\mathbf{U}_2 + \mathbf{H}_2\mathbf{U}_1)(\mathbf{L}_1 + \mathbf{L}_1^T), \\ \frac{\partial h_{L,2}}{\partial \mathbf{U}_1} &= (\mathbf{H}_2\mathbf{U}_1 + \mathbf{H}_1\mathbf{U}_2)(\mathbf{L}_2 - \mathbf{L}_2^T), \qquad \frac{\partial h_{L,2}}{\partial \mathbf{U}_2} = (\mathbf{H}_2\mathbf{U}_2 - \mathbf{H}_1\mathbf{U}_1)(\mathbf{L}_2 - \mathbf{L}_2^T), \end{aligned}$$

where $\mathbf{H}_1 = \mathbf{H}_1^T$ and $\mathbf{H}_2 = -\mathbf{H}_2^T$ must be taken into account. Now setting

$$\mathbf{\Lambda}_1 = \frac{\mathbf{L}_1 + \mathbf{L}_1^T}{2} = \mathbf{\Lambda}_1^T, \;\; \mathbf{\Lambda}_2 = \frac{\mathbf{L}_2 - \mathbf{L}_2^T}{2} = -\mathbf{\Lambda}_2^T \;\text{ and }\; \mathbf{\Lambda} = \mathbf{\Lambda}_1 + i\mathbf{\Lambda}_2 = \mathbf{\Lambda}^*,$$

it can be verified that

$$\frac{\partial h_L(\mathbf{U}^\wedge)}{\partial \mathbf{U}^\wedge} = \begin{bmatrix} \frac{\partial (h_{L,1}+h_{L,2})}{\partial \mathbf{U}_1} & \frac{\partial (h_{L,1}+h_{L,2})}{\partial \mathbf{U}_2} \\ -\frac{\partial (h_{L,1}+h_{L,2})}{\partial \mathbf{U}_2} & \frac{\partial (h_{L,1}+h_{L,2})}{\partial \mathbf{U}_1} \end{bmatrix} = 2\mathbf{H}^\wedge\mathbf{U}^\wedge(\overline{\mathbf{\Lambda}})^\wedge.$$

Consequently, the necessary first order conditions for an optimum

$$\frac{\partial}{\partial \mathbf{U}_1}(f + h_{L,1} + h_{L,2}) = \mathbf{0} \;\text{ and }\; \frac{\partial}{\partial \mathbf{U}_2}(f + h_{L,1} + h_{L,2}) = \mathbf{0}$$

can be stated as

$$\frac{\partial f(\mathbf{U}^\wedge)}{\partial \mathbf{U}^\wedge} + \frac{\partial h_L(\mathbf{U}^\wedge)}{\partial \mathbf{U}^\wedge} = 2(\mathbf{H}\mathbf{U}\mathbf{X}\mathbf{X}^* - \mathbf{H}\mathbf{Y}\mathbf{X}^* + \mathbf{H}\mathbf{U}\overline{\mathbf{\Lambda}})^\wedge = \mathbf{0},$$

showing that (4.7) and (4.8) must be satisfied in the complex case, too. (The conjugation is irrelevant since $\overline{\mathbf{\Lambda}}$ may simply be renamed to $\mathbf{\Lambda}'$.)

It remains to determine the particular H-polar decomposition (if existent) which leads to the optimum congruence. For this, let $\mathbf{U}\mathbf{M}$ be an H-polar decomposition of the matrix $\mathbf{A} = \mathbf{Y}\mathbf{X}^*\mathbf{H}$, and let

$$(\mathbf{R}^{-1}\mathbf{A}^{[*]}\mathbf{A}\mathbf{R}, \mathbf{R}^*\mathbf{H}\mathbf{R}) = (\mathbf{J}, \mathbf{Z}_J) \;\text{ and }\; (\mathbf{S}^{-1}\mathbf{M}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mathbf{K}, \mathbf{Z}_K)$$

be the canonical forms (see Theorem 3.1) of the pairs $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H}) = (\mathbf{M}^2, \mathbf{H})$ and $(\mathbf{M}, \mathbf{H})$, respectively. Returning to the initial equation (4.6b), we find that

$$\begin{aligned} f(\mathbf{U}) &= \operatorname{tr}[(\mathbf{UX}-\mathbf{Y})^*\mathbf{H}(\mathbf{UX}-\mathbf{Y})] \\ &= \operatorname{tr}(\mathbf{X}^*\mathbf{U}^*\mathbf{HUX} - \mathbf{X}^*\mathbf{U}^*\mathbf{HY} - \mathbf{Y}^*\mathbf{HUX} + \mathbf{Y}^*\mathbf{HY}) \\ &= \operatorname{tr}(\mathbf{X}^*\mathbf{HX}) + \operatorname{tr}(\mathbf{Y}^*\mathbf{HY}) - 2\operatorname{Re}\operatorname{tr}[(\mathbf{YX}^*\mathbf{H})(\mathbf{H}^{-1}\mathbf{U}^*\mathbf{H})] \\ &= \tau - 2\operatorname{Re}\operatorname{tr}(\mathbf{AU}^{-1}) = \tau - 2\operatorname{Re}\operatorname{tr}(\mathbf{UMU}^{-1}) \\ &= \tau - 2\operatorname{Re}\operatorname{tr}(\mathbf{SKS}^{-1}) = \tau - 2\operatorname{Re}\operatorname{tr}(\mathbf{K}) \end{aligned}$$

where $\tau = \operatorname{tr}(\mathbf{X}^*\mathbf{HX}) + \operatorname{tr}(\mathbf{Y}^*\mathbf{HY})$. The optimum can be found from this equation by considering three cases:

Case (a): If $\mathbf{H}$ is definite, then the canonical forms are of the kind

$$\begin{aligned} (\mathbf{J}, \mathbf{Z}_J) &= \left(\bigoplus_{j=1}^{k} \lambda_j \mathbf{I}_{p_j} \oplus \mathbf{0}_r, \bigoplus_{j=1}^{k} \varepsilon \mathbf{I}_{p_j} \oplus \varepsilon \mathbf{I}_r \right) \\ (\mathbf{K}, \mathbf{Z}_K) &= \left(\bigoplus_{j=1}^{k} \sqrt{\lambda_j} \boldsymbol{\Sigma}_{p_j} \oplus \mathbf{0}_r, \bigoplus_{j=1}^{k} \varepsilon \mathbf{I}_{p_j} \oplus \varepsilon \mathbf{I}_r \right). \end{aligned}$$

Here $\lambda_j > 0$, $\boldsymbol{\Sigma}_{p_j} = \operatorname{diag}(\pm 1)$ for $1 \le j \le k$ and $\varepsilon = +1$ if $\mathbf{H} > 0$, $\varepsilon = -1$ if $\mathbf{H} < 0$. In the case $\mathbf{H} > 0$ the value $f(\mathbf{U})$ takes its minimum, when $\boldsymbol{\Sigma}_{p_j} = +\mathbf{I}_{p_j}$ is chosen and in the case $\mathbf{H} < 0$ the value $f(\mathbf{U})$ takes its maximum, when $\boldsymbol{\Sigma}_{p_j} = -\mathbf{I}_{p_j}$ is chosen. This means that in both cases

$$\mathbf{Z}_K\mathbf{K} = \bigoplus_{j=1}^{k} \sqrt{\lambda_j} \mathbf{I}_{p_j} \oplus \mathbf{0}_r \ge 0$$

and thus $\mathbf{HM} \ge 0$, so that the wanted result is obtained via a *semidefinite* H-polar decomposition of $\mathbf{A}$. In particular, if $\mathbf{H} = \mathbf{I}$, then the solution is determined by an ordinary polar decomposition where $\mathbf{U}^* = \mathbf{U}^{-1}$ and $\mathbf{M}^* = \mathbf{M}$ is positive semidefinite.

Case (b): If $\mathbf{H}$ is indefinite and if $\mathbf{A}$ admits a semidefinite H-polar decomposition, then the following relationships exists between the canonical forms

$$\begin{aligned} \mathbf{J} &= \bigoplus_{j=1}^{k} \begin{bmatrix} \lambda_j \mathbf{I}_{p_j} & \\ & \lambda_j \mathbf{I}_{q_j} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_{r+s} & \\ & \mathbf{0}_{r+t} \end{bmatrix}, \\ \mathbf{Z}_J &= \bigoplus_{j=1}^{k} \begin{bmatrix} \mathbf{I}_{p_j} & \\ & -\mathbf{I}_{q_j} \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_{r+s} & \\ & -\mathbf{I}_{r+t} \end{bmatrix}, \\ \mathbf{K} &= \bigoplus_{j=1}^{k} \begin{bmatrix} \sqrt{\lambda_j}\boldsymbol{\Sigma}_{p_j} & \\ & \sqrt{\lambda_j}\boldsymbol{\Sigma}_{q_j} \end{bmatrix} \oplus \begin{bmatrix} & \boldsymbol{\Sigma}_r \\ \mathbf{0}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_s & \\ & \mathbf{0}_t \end{bmatrix}, \\ \mathbf{Z}_K &= \bigoplus_{j=1}^{k} \begin{bmatrix} \mathbf{I}_{p_j} & \\ & -\mathbf{I}_{q_j} \end{bmatrix} \oplus \begin{bmatrix} & \mathbf{I}_r \\ \mathbf{I}_r & \end{bmatrix} \oplus \begin{bmatrix} \mathbf{I}_s & \\ & -\mathbf{I}_t \end{bmatrix}, \end{aligned} \tag{4.9a}$$

where $\lambda_j > 0$ for $1 \le j \le k$ (see Corollary 3.23). If in this case $\boldsymbol{\Sigma}_{p_j} = \mathbf{I}_{p_j}$,

$\mathbf{\Sigma}_{q_j} = -\mathbf{I}_{q_j}$ and $\mathbf{\Sigma}_r = \mathbf{I}_r$ is chosen, then again

$$\mathbf{Z}_K\mathbf{K} = \bigoplus_{j=1}^{k} \sqrt{\lambda_j}\mathbf{I}_{p_j+q_j} \oplus \mathbf{0}_r \oplus \mathbf{I}_r \oplus \mathbf{0}_{s+t} \geq 0. \tag{4.9b}$$

By this choice the contributions to $f(\mathbf{U})$ take on their minimum along the positive space dimensions and their maximum along the negative space dimensions. This is what is meant by "min / max" in (4.6a). Moreover, the resulting coordinate differences are "small" which can be seen in the following way: Let $\mathbf{X}' = \mathbf{U}\mathbf{X}$. Then

$$\begin{aligned}\mathbf{Y}(\mathbf{X}')^* &= \mathbf{Y}\mathbf{X}^*\mathbf{U}^* = \mathbf{U}\mathbf{M}\mathbf{H}^{-1}\mathbf{U}^* \\ &= \mathbf{U}(\mathbf{S}\mathbf{K}\mathbf{S}^{-1})(\mathbf{S}\mathbf{Z}_K\mathbf{S}^*)\mathbf{U}^* \\ &= (\mathbf{U}\mathbf{S})\mathbf{K}\mathbf{Z}_K(\mathbf{U}\mathbf{S})^*\end{aligned}$$

is positive semidefinite since $\mathbf{Z}_K(\mathbf{Z}_K\mathbf{K})\mathbf{Z}_K = \mathbf{K}\mathbf{Z}_K$ is. Hence, the orthogonal or unitary Procrustes problem

$$\varphi(\mathbf{T}) = \mathrm{tr}[(\mathbf{T}\mathbf{X}' - \mathbf{Y})^*(\mathbf{T}\mathbf{X}' - \mathbf{Y})] \to \min \quad \text{with} \quad \mathbf{T}^*\mathbf{T} = \mathbf{I}, \tag{4.10}$$

whose solution, according to case (a), is determined by an ordinary polar decomposition

$$\mathbf{T}\mathbf{M}' = \mathbf{Y}(\mathbf{X}')^* \quad \text{with} \quad \mathbf{M}' = (\mathbf{M}')^* \geq 0,$$

is solved for $\mathbf{T} = \mathbf{I}$. In other words, the coordinate differences $\sum_k \|\mathbf{x}'_k - \mathbf{y}_k\|^2$ obtained with the "min / max" solution $\mathbf{x}'_k = \mathbf{U}\mathbf{x}_k$ are at minimum with respect to an orthogonal or unitary transformation in the sense of problem (4.10). This is exactly what one would expect of a transformation to an optimum congruence.

Case (c): If $\mathbf{H}$ is indefinite and if $\mathbf{A}$ admits an H-polar decomposition but not a semidefinite H-polar decomposition, then by definition $\mathbf{Z}_K\mathbf{K}$ and thus also $\mathbf{K}\mathbf{Z}_K$ cannot be positive semidefinite. Therefore, there always exists a solution $\mathbf{T}_0$ of the problem (4.10) for which $\varphi(\mathbf{T}_0) < \varphi(\mathbf{I})$. Hence, the wanted result of an optimum congruence of the constellations $\mathbf{X}'$ and $\mathbf{Y}$ cannot be achieved in this case.

This investigation shows that an H-isometry for which $\mathbf{X}' = \mathbf{U}\mathbf{X}$ and $\mathbf{Y}$ are at optimum congruence can only exist if $\mathbf{A}$ admits a semidefinite H-polar decomposition. Conversely, let $\mathbf{X}'$ and $\mathbf{Y}$ be matrices which are at optimum congruence, i.e. for which $\mathbf{Y}(\mathbf{X}')^*$ is positive semidefinite and selfadjoint. Moreover, let $\mathbf{U}$ be an H-isometry and let $\mathbf{X} = \mathbf{U}^{-1}\mathbf{X}'$. Then $\mathbf{A} = \mathbf{Y}\mathbf{X}^*\mathbf{H} = \mathbf{Y}(\mathbf{X}')^*\mathbf{H}\mathbf{U}$ admits the semidefinite H-polar decomposition $\mathbf{A} = \mathbf{U}\mathbf{M}$ where $\mathbf{M} = \mathbf{U}^{-1}\mathbf{Y}(\mathbf{X}')^*\mathbf{H}\mathbf{U}$ is H-nonnegative. All in all, we thus have found the following result.

Theorem 4.14 (Solution of the H-isometric Procrustes problem). *A solution of the H-orthogonal or H-unitary Procrustes problem* (4.6) *exists if and only if the matrix* $\mathbf{A} = \mathbf{Y}\mathbf{X}^*\mathbf{H}$ *admits a semidefinite H-polar decomposition. In this case the H-isometry* $\mathbf{U}$ *contained in such a decomposition* $\mathbf{A} = \mathbf{U}\mathbf{M}$ *optimises the function* f*. Moreover,* $\mathbf{X}' = \mathbf{U}\mathbf{X}$ *and* $\mathbf{Y}$ *are at optimum congruence in the sense, that the orthogonal or unitary Procrustes problem* (4.10) *is solved for* $\mathbf{T} = \mathbf{I}$.

4.5 Solution of the (G,H)-isometric Procrustes problem

Whereas the H-isometric Procrustes problem can always be solved in the case of a definite matrix $\mathbf{H}$, in the case of an indefinite matrix $\mathbf{H}$ it is possible that no solution exists. But now let $\mathbf{G}$ and $\mathbf{H}$ be nonsingular selfadjoint matrices in $\mathbb{F}^{n\times n}$, and let the geometry within the tuples $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ and $(\mathbf{y}_1,\ldots,\mathbf{y}_N)$ be measured with the scalar product $[.,.]_G = (\mathbf{G}.,.)$, but the geometry between the tuples be measured with the scalar product $[.,.]_H = (\mathbf{H}.,.)$. Then the problem can be expressed, instead of (4.6), as

$$f(\mathbf{U}) = \sum_{k=1}^{N} [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]_H \rightarrow \begin{cases} \min, & \text{if } \mathbf{H} > 0 \\ \max, & \text{if } \mathbf{H} < 0 \\ \min \,/\, \max, & \text{otherwise} \end{cases} \tag{4.11a}$$

$$\text{with } \mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G} = \mathbf{0} \text{ and } \mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H} = \mathbf{0}$$

or in matrix notation

$$f(\mathbf{U}) = \operatorname{tr}[(\mathbf{U}\mathbf{X} - \mathbf{Y})^*\mathbf{H}(\mathbf{U}\mathbf{X} - \mathbf{Y})] \tag{4.11b}$$

which will be called the (G,H)-orthogonal or (G,H)-unitary Procrustes problem. If the vectors $\mathbf{x}_k$ and $\mathbf{y}_k$ result from a construction according to Theorem 4.9, the internal metric $\mathbf{G}$ is fixed, but the external metric $\mathbf{H}$ may be chosen within the scope of the "compatibility condition"

$$\mathbf{H}^{-1}\mathbf{G} = \mu^2\mathbf{G}^{-1}\mathbf{H} \text{ for some } \mu \in \mathbb{R}\backslash\{0\} \tag{4.12}$$

which is characterised in Lemma 4.2. If this choice is made such that $\mathbf{H}$ is positive definite, then a sum of non-negative distance squares is minimised. In this case a solution of (4.11) under the assumption (4.12) always exists which will be shown in the sequel. (An analogous statement holds for a negative definite matrix $\mathbf{H}$.)

If again $\mathbf{L}_G, \mathbf{L}_H \in \mathbb{R}^{n\times n}$ are matrices of the (unknown) Lagrange multipliers and if the constraints in the case $\mathbb{F} = \mathbb{R}$ are stated in the form

$$g_L(\mathbf{U}) = \operatorname{tr}[\mathbf{L}_G(\mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G})] \text{ and } h_L(\mathbf{U}) = \operatorname{tr}[\mathbf{L}_H(\mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H})]$$

then the necessary first order condition

$$\frac{\partial}{\partial \mathbf{U}}(f + g_L + h_L) = \mathbf{0}$$

leads in the same way as above to the equation

$$\begin{gathered} \mathbf{G}\mathbf{U}\mathbf{A} + \mathbf{H}\mathbf{U}\mathbf{B} = \tilde{\mathbf{C}} \text{ with } \tilde{\mathbf{C}} = \mathbf{H}\mathbf{Y}\mathbf{X}^* \text{ and} \\ \mathbf{A} = \frac{\mathbf{L}_G + \mathbf{L}_G^*}{2} = \mathbf{A}^*, \quad \mathbf{B} = \mathbf{X}\mathbf{X}^* + \frac{\mathbf{L}_H + \mathbf{L}_H^*}{2} = \mathbf{B}^* \end{gathered} \tag{4.13}$$

which is also valid in the case $\mathbb{F} = \mathbb{C}$. Furthermore

$$\mathbf{G}\mathbf{U}\mathbf{G}^{-1} = \mathbf{U}^{-*} = \mathbf{H}\mathbf{U}\mathbf{H}^{-1},$$

so that the transformations

$$\tilde{\mathbf{C}} = \mathbf{GUG}^{-1}\mathbf{GA} + \mathbf{HUB} = \mathbf{HUH}^{-1}\mathbf{GA} + \mathbf{HUB} = \mathbf{HU}(\mathbf{H}^{-1}\mathbf{GA} + \mathbf{B}),$$
$$\tilde{\mathbf{C}} = \mathbf{GUA} + \mathbf{HUH}^{-1}\mathbf{HB} = \mathbf{GUA} + \mathbf{GUG}^{-1}\mathbf{HB} = \mathbf{GU}(\mathbf{A} + \mathbf{G}^{-1}\mathbf{HB})$$

can be made, yielding

$$\begin{aligned} \mathbf{UM} = \mathbf{H}^{-1}\tilde{\mathbf{C}}\mathbf{H} + \mathbf{G}^{-1}\tilde{\mathbf{C}}\mathbf{G} = \mathbf{C} \text{ with} \\ \mathbf{M} = \mathbf{H}^{-1}\mathbf{GAH} + \mathbf{BH} + \mathbf{AG} + \mathbf{G}^{-1}\mathbf{HBG}. \end{aligned} \tag{4.14}$$

If now (4.12) is taken into account, then on the one hand

$$\mathbf{M}^*\mathbf{H} - \mathbf{HM} = \mathbf{GBHG}^{-1}\mathbf{H} - \mathbf{HG}^{-1}\mathbf{HBG} = \mu^{-2}(\mathbf{GBG} - \mathbf{GBG}) = \mathbf{0},$$
$$\mathbf{M}^*\mathbf{G} - \mathbf{GM} = \mathbf{HAGH}^{-1}\mathbf{G} - \mathbf{GH}^{-1}\mathbf{GAH} = \mu^2(\mathbf{HAH} - \mathbf{HAH}) = \mathbf{0}$$

and on the other hand (4.14) implies

$$\mathbf{HCH}^{-1} = \tilde{\mathbf{C}} + \mathbf{HG}^{-1}\tilde{\mathbf{C}}\mathbf{GH}^{-1} = (\mu^2/\mu^2)\mathbf{GH}^{-1}\tilde{\mathbf{C}}\mathbf{HG}^{-1} + \tilde{\mathbf{C}} = \mathbf{GCG}^{-1}$$
$$\text{or } \mathbf{H}^{-1}\mathbf{C}^*\mathbf{H} = \mathbf{G}^{-1}\mathbf{C}^*\mathbf{G}.$$

Therefore, if (4.12) holds and if $\mathbf{U}$ is a (G,H)-isometry and if there exist self-adjoint matrices $\mathbf{A}, \mathbf{B}$ which solve (4.13), then there exists a (G,H)-selfadjoint matrix $\mathbf{M}$ such that $\mathbf{UM}$ is a (G,H)-polar decomposition of $\mathbf{C}$. In particular, it is true that $\mathbf{C}^H = \mathbf{C}^G$.

In order to prove that the existence of a (G,H)-polar decomposition $\mathbf{UM}=\mathbf{C}$ conversely implies the existence of the matrices $\mathbf{A}$ and $\mathbf{B}$, assume that (4.12) holds. Then, according to Lemma 4.4, there exists a nonsingular matrix $\mathbf{S}$ such that

$$\mathbf{S}^*\mathbf{HS} = \mathbf{J}_1 \oplus \mathbf{J}_2, \qquad \mathbf{S}^*\mathbf{GS} = \mu(\mathbf{J}_1 \oplus -\mathbf{J}_2), \tag{4.15a}$$
$$\mathbf{S}^{-1}\mathbf{US} = \mathbf{U}_1 \oplus \mathbf{U}_2, \qquad \mathbf{S}^{-1}\mathbf{MS} = \mathbf{M}_1 \oplus \mathbf{M}_2, \qquad \mathbf{S}^{-1}\mathbf{CS} = \mathbf{C}_1 \oplus \mathbf{C}_2,$$

where $\mathbf{J}_k$ has the form $\mathrm{diag}(\pm 1)$ and $\mathbf{U}_k\mathbf{M}_k = \mathbf{C}_k$ is a $\mathbf{J}_k$-polar decomposition $(k = 1, 2)$. Let

$$\mathbf{S}^*\tilde{\mathbf{C}}\mathbf{S}^{-*} = \begin{bmatrix} \tilde{\mathbf{C}}_{11} & \tilde{\mathbf{C}}_{12} \\ \tilde{\mathbf{C}}_{21} & \tilde{\mathbf{C}}_{22} \end{bmatrix} \tag{4.15b}$$

and

$$\mathbf{S}^{-1}\mathbf{AS}^{-*} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^* & \mathbf{A}_{22} \end{bmatrix}, \ \mathbf{S}^{-1}\mathbf{BS}^{-*} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{12}^* & \mathbf{B}_{22} \end{bmatrix} \tag{4.15c}$$

be compatible partitionings. Then from (4.14) it follows that

$$\begin{aligned} \mathbf{U}_1\mathbf{M}_1 \oplus \mathbf{U}_2\mathbf{M}_2 = \mathbf{C}_1 \oplus \mathbf{C}_2 = 2(\mathbf{J}_1\tilde{\mathbf{C}}_{11}\mathbf{J}_1 \oplus \mathbf{J}_2\tilde{\mathbf{C}}_{22}\mathbf{J}_2) \text{ or} \\ \tilde{\mathbf{C}}_{11} = \tfrac{1}{2}\,\mathbf{J}_1\mathbf{U}_1\mathbf{M}_1\mathbf{J}_1 \text{ and } \tilde{\mathbf{C}}_{22} = \tfrac{1}{2}\,\mathbf{J}_2\mathbf{U}_2\mathbf{M}_2\mathbf{J}_2. \end{aligned} \tag{4.15d}$$

On the other hand (4.13) requires $\mathbf{GA} + \mathbf{HB} = \mathbf{U}^*\tilde{\mathbf{C}}$ or

$$\begin{bmatrix} \mathbf{J}_1(\mu\mathbf{A}_{11} + \mathbf{B}_{11}) & \mathbf{J}_1(\mu\mathbf{A}_{12} + \mathbf{B}_{12}) \\ \mathbf{J}_2(-\mu\mathbf{A}_{12}^* + \mathbf{B}_{12}^*) & \mathbf{J}_2(-\mu\mathbf{A}_{22} + \mathbf{B}_{22}) \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1^*\tilde{\mathbf{C}}_{11} & \mathbf{U}_1^*\tilde{\mathbf{C}}_{12} \\ \mathbf{U}_2^*\tilde{\mathbf{C}}_{21} & \mathbf{U}_2^*\tilde{\mathbf{C}}_{22} \end{bmatrix},$$

yielding the system of equations

$$\begin{aligned} \mu\mathbf{A}_{11}+\mathbf{B}_{11}=\mathbf{J}_1\mathbf{U}_1^*\tilde{\mathbf{C}}_{11}=\tfrac{1}{2}\,\mathbf{M}_1\mathbf{J}_1, \qquad & \mu\mathbf{A}_{12}+\mathbf{B}_{12}=\mathbf{J}_1\mathbf{U}_1^*\tilde{\mathbf{C}}_{12}, \\ -\mu\mathbf{A}_{22}+\mathbf{B}_{22}=\mathbf{J}_2\mathbf{U}_2^*\tilde{\mathbf{C}}_{22}=\tfrac{1}{2}\,\mathbf{M}_2\mathbf{J}_2, \qquad & -\mu\mathbf{A}_{12}+\mathbf{B}_{12}=\tilde{\mathbf{C}}_{21}^*\mathbf{U}_2\mathbf{J}_2. \end{aligned}$$

Therefore, by selecting arbitrary selfadjoint blocks $\mathbf{B}_{11},\mathbf{B}_{22}$ and setting

$$\begin{aligned} \mathbf{A}_{11}=\tfrac{1}{\mu}(\tfrac{1}{2}\,\mathbf{M}_1\mathbf{J}_1-\mathbf{B}_{11})=\mathbf{A}_{11}^*, \qquad & \mathbf{A}_{12}=\tfrac{1}{2\mu}(\mathbf{J}_1\mathbf{U}_1^*\tilde{\mathbf{C}}_{12}-\tilde{\mathbf{C}}_{21}^*\mathbf{U}_2\mathbf{J}_2), \\ \mathbf{A}_{22}=\tfrac{1}{\mu}(\mathbf{B}_{22}-\tfrac{1}{2}\,\mathbf{M}_2\mathbf{J}_2)=\mathbf{A}_{22}^*, \qquad & \mathbf{B}_{12}=\tfrac{1}{2}(\mathbf{J}_1\mathbf{U}_1^*\tilde{\mathbf{C}}_{12}+\tilde{\mathbf{C}}_{21}^*\mathbf{U}_2\mathbf{J}_2), \end{aligned}$$

the two selfadjoint matrices $\mathbf{A}$ and $\mathbf{B}$ which solve (4.13) are determined. If the particular choice

$$\mathbf{B}_{11}=\tfrac{1}{4}\,\mathbf{M}_1\mathbf{J}_1, \quad \mathbf{B}_{22}=\tfrac{1}{4}\,\mathbf{M}_2\mathbf{J}_2$$

is made, then

$$\mathbf{A}_{11}=\tfrac{1}{4\mu}\,\mathbf{M}_1\mathbf{J}_1, \quad \mathbf{A}_{22}=\tfrac{-1}{4\mu}\,\mathbf{M}_2\mathbf{J}_2$$

and thus

$$\begin{aligned} \mathbf{A}&=\tfrac{1}{2}(\mathbf{G}^{-1}\mathbf{U}^*\tilde{\mathbf{C}}+\tilde{\mathbf{C}}^*\mathbf{U}\mathbf{G}^{-1})-\tfrac{1}{4}\mathbf{M}\mathbf{G}^{-1}, \\ \mathbf{B}&=\tfrac{1}{2}(\mathbf{H}^{-1}\mathbf{U}^*\tilde{\mathbf{C}}+\tilde{\mathbf{C}}^*\mathbf{U}\mathbf{H}^{-1})-\tfrac{1}{4}\mathbf{M}\mathbf{H}^{-1} \end{aligned}$$

which follows from (4.15). Summarising, the following result is proved.

Lemma 4.15. ($\mathbb{F}=\mathbb{R}$ *or* $\mathbb{F}=\mathbb{C}$) *Let* $\mathbf{G},\mathbf{H}\in\mathbb{F}^{n\times n}$ *be nonsingular selfadjoint matrices which satisfy* (4.12)*. Moreover, let* $\mathbf{U}\in\mathbb{F}^{n\times n}$ *be a (G,H)-isometry and let* $\tilde{\mathbf{C}}\in\mathbb{F}^{n\times n}$*. Then the following statements are equivalent:*

(*i*) *There exist selfadjoint matrices* $\mathbf{A},\mathbf{B}\in\mathbb{F}^{n\times n}$ *such that*

$$\mathbf{G}\mathbf{U}\mathbf{A}+\mathbf{H}\mathbf{U}\mathbf{B}=\tilde{\mathbf{C}}.$$

(*ii*) *There exists a (G,H)-selfadjoint matrix* $\mathbf{M}\in\mathbb{F}^{n\times n}$ *such that*

$$\mathbf{U}\mathbf{M}=\mathbf{G}^{-1}\tilde{\mathbf{C}}\mathbf{G}+\mathbf{H}^{-1}\tilde{\mathbf{C}}\mathbf{H}.$$

Using this lemma, the necessary condition (4.13) for solving the Procrustes problem (4.11) under the assumption (4.12) finally becomes

$$\begin{aligned} &\mathbf{C}=\mathbf{U}\mathbf{M} \text{ with } \mathbf{C}=\mathbf{Y}\mathbf{X}^*\mathbf{H}+\mathbf{G}^{-1}\mathbf{H}\mathbf{Y}\mathbf{X}^*\mathbf{G} \text{ and} \\ &\mathbf{C}^H=\mathbf{C}^G, \ \mathbf{U}^H=\mathbf{U}^G=\mathbf{U}^{-1}, \ \mathbf{M}^H=\mathbf{M}^G=\mathbf{M}. \end{aligned} \tag{4.16}$$

Thus the solution of the problem can be determined by a (G,H)-polar decomposition of the matrix $\mathbf{C}$.

Again, it remains to determine the particular (G,H)-polar decomposition (if existent) which leads to the optimum congruence. For this, let $\mathbf{U}\mathbf{M}$ be a (G,H)-polar decomposition of the matrix $\mathbf{C}$. Moreover, let $\mathbf{S}$ be a nonsingular matrix such that (4.15) holds and let

$$\mathbf{S}^{-1}\mathbf{X}=\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} \text{ and } \mathbf{S}^{-1}\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix}$$

be compatible partitionings. Then on the one hand from (4.13), (4.15a), (4.15b) it follows that

$$\begin{bmatrix}\tilde{\mathbf{C}}_{11} & \tilde{\mathbf{C}}_{12}\\ \tilde{\mathbf{C}}_{21} & \tilde{\mathbf{C}}_{22}\end{bmatrix} = \mathbf{S}^*\tilde{\mathbf{C}}\mathbf{S}^{-*} = (\mathbf{S}^*\mathbf{H}\mathbf{S})(\mathbf{S}^{-1}\mathbf{Y})(\mathbf{S}^{-1}\mathbf{X})^* = \begin{bmatrix}\mathbf{J}_1\mathbf{Y}_1\mathbf{X}_1^* & \mathbf{J}_1\mathbf{Y}_1\mathbf{X}_2^*\\ \mathbf{J}_2\mathbf{Y}_2\mathbf{X}_1^* & \mathbf{J}_2\mathbf{Y}_2\mathbf{X}_2^*\end{bmatrix},$$

so that according to (4.15d)

$$\mathbf{U}_k\mathbf{M}_k = \mathbf{C}_k = 2\mathbf{J}_k\tilde{\mathbf{C}}_{kk}\mathbf{J}_k = 2\mathbf{Y}_k\mathbf{X}_k^*\mathbf{J}_k \text{ for } k = 1, 2.$$

On the other hand we find from the initial equation (4.11b)

$$\begin{aligned} f(\mathbf{U}) &= \operatorname{tr}\Big(\big[(\mathbf{S}^{-1}\mathbf{U}\mathbf{S})(\mathbf{S}^{-1}\mathbf{X})-(\mathbf{S}^{-1}\mathbf{Y})\big]^*(\mathbf{S}^*\mathbf{H}\mathbf{S})\big[(\mathbf{S}^{-1}\mathbf{U}\mathbf{S})(\mathbf{S}^{-1}\mathbf{X})-(\mathbf{S}^{-1}\mathbf{Y})\big]\Big)\\ &= \operatorname{tr}\left(\begin{bmatrix}\mathbf{U}_1\mathbf{X}_1-\mathbf{Y}_1\\ \mathbf{U}_2\mathbf{X}_2-\mathbf{Y}_2\end{bmatrix}^*(\mathbf{J}_1\oplus\mathbf{J}_2)\begin{bmatrix}\mathbf{U}_1\mathbf{X}_1-\mathbf{Y}_1\\ \mathbf{U}_2\mathbf{X}_2-\mathbf{Y}_2\end{bmatrix}\right)\\ &= \textstyle\sum_k \operatorname{tr}[(\mathbf{U}_k\mathbf{X}_k-\mathbf{Y}_k)^*\mathbf{J}_k(\mathbf{U}_k\mathbf{X}_k-\mathbf{Y}_k)]. \end{aligned}$$

Now, using the canonical forms of the pairs $(\mathbf{C}_k^{J_k}\mathbf{C}_k, \mathbf{J}_k) = (\mathbf{M}_k^2, \mathbf{J}_k)$ and $(\mathbf{M}_k, \mathbf{J}_k)$ the argumentation from Section 4.4 can be applied twice, showing that the optimum congruence is achieved when $\mathbf{U}_k\mathbf{M}_k$ are semidefinite $\mathbf{J}_k$-polar decompositions of the matrices $\mathbf{C}_k$. If in this case we set

$$\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix} = \begin{bmatrix}\mathbf{U}_1\mathbf{X}_1\\ \mathbf{U}_2\mathbf{X}_2\end{bmatrix} = \mathbf{S}^{-1}\mathbf{U}\mathbf{X} \text{ and } \begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \mathbf{S}^{-1}\mathbf{Y}, \tag{4.17a}$$

then the orthogonal or unitary Procrustes problems

$$\varphi_k(\mathbf{T}_k) = \operatorname{tr}[(\mathbf{T}_k\mathbf{X}_k'-\mathbf{Y}_k)^*(\mathbf{T}_k\mathbf{X}_k'-\mathbf{Y}_k)] \to \min \text{ with } \mathbf{T}_k^*\mathbf{T}_k = \mathbf{I}, \tag{4.17b}$$

are solved for $\mathbf{T}_k = \mathbf{I}$. Moreover, if $\mathbf{H} > 0$ ($\mathbf{H} < 0$), then $\mathbf{J}_k = \mathbf{I}$ ($\mathbf{J}_k = -\mathbf{I}$), so that a solution then always exists (see Example 4.6). Summarising, the result can be expressed by the following theorem.

Theorem 4.16 (Solution of the (G,H)-isometric Procrustes problem). *A solution of the (G,H)-orthogonal or (G,H)-unitary Procrustes problem* (4.11) *under the assumption* (4.12) *exists if and only if the matrix* $\mathbf{C} = \mathbf{Y}\mathbf{X}^*\mathbf{H} + \mathbf{G}^{-1}\mathbf{H}\mathbf{Y}\mathbf{X}^*\mathbf{G}$ *admits an H-semidefinite (G,H)-polar decomposition. In this case the (G,H)-isometry* $\mathbf{U}$ *contained in such a decomposition* $\mathbf{C} = \mathbf{U}\mathbf{M}$ *optimises the function* f*. Moreover,* $\mathbf{X}' = \mathbf{U}\mathbf{X}$ *and* $\mathbf{Y}$ *are at optimum congruence in the sense, that the orthogonal or unitary Procrustes problems* (4.17) *are solved for* $\mathbf{T}_k = \mathbf{I}$.

4.6 More general results on (G,H)-polar decompositions

In the previous section it was found that the (G,H)-polar decompositions introduced in Section 4.2 have useful applications. For this reason, we will now generalise the Lemmas 4.2 – 4.4 for the case in which $\rho\mathbf{H} - \mathbf{G}$ is a non-defective matrix pencil. In view of Theorem 3.2 only the case $\mathbb{F} = \mathbb{C}$ is investigated.

A pencil $\rho\mathbf{H}-\mathbf{G}$ is said to be non-defective, if there exist nonsingular matrices $\mathbf{P},\mathbf{Q}$ such that both $\mathbf{PGQ}$ and $\mathbf{PHQ}$ are diagonal [MMX, Definition 1.3]. If $\mathbf{G}$ and $\mathbf{H}$ are in addition nonsingular and Hermitian, then these pencils have a particularly simple canonical form which has already been described in Corollary 3.20. Using this result, the following theorem is proved easily.

Theorem 4.17. *Let $\rho\mathbf{H}-\mathbf{G}\in\mathbb{C}^{n\times n}$ be a non-defective Hermitian matrix pencil where both $\mathbf{H}$ and $\mathbf{G}$ are nonsingular, and let $\mathbf{A}\in\mathbb{C}^{n\times n}$ be a matrix with $\mathbf{A}^H=\mathbf{A}^G$. Then there exists a nonsingular matrix $\mathbf{S}\in\mathbb{C}^{n\times n}$ such that*

$$\begin{aligned}\mathbf{S}^{-1}\mathbf{AS}&=\mathbf{A}_1\oplus\ldots\oplus\mathbf{A}_k,\\ \mathbf{S}^{*}\mathbf{HS}&=\mathbf{H}_1\oplus\ldots\oplus\mathbf{H}_k,\\ \mathbf{S}^{*}\mathbf{GS}&=\mathbf{G}_1\oplus\ldots\oplus\mathbf{G}_k,\end{aligned}\tag{4.18a}$$

where the blocks $\mathbf{A}_j$, $\mathbf{H}_j$ and $\mathbf{G}_j$ are of equal size and each triple $(\mathbf{A}_j,\mathbf{H}_j,\mathbf{G}_j)$ has one and only one of the following forms:

1. *Triples belonging to real eigenvalues of the pencil*

$$\mathbf{A}_j\in\mathbb{C}^{p\times p},\ \ \mathbf{H}_j=\mathbf{I}_{p-q}\oplus-\mathbf{I}_q,\ \ \mathbf{G}_j=\mu(\mathbf{I}_{p-q}\oplus-\mathbf{I}_q)\tag{4.18b}$$

   with $\mu\in\mathbb{R}\backslash\{0\}$ and $p,q\in\mathbb{N}$, $q\le p$.

2. *Triples belonging to non-real eigenvalues of the pencil*

$$\mathbf{A}_j=\begin{bmatrix}\mathbf{A}_{j,1}&\\&\mathbf{A}_{j,2}\end{bmatrix}\in\mathbb{C}^{2p\times 2p},\ \ \mathbf{H}_j=\begin{bmatrix}&\mathbf{I}_p\\\mathbf{I}_p&\end{bmatrix},\ \ \mathbf{G}_j=\begin{bmatrix}&\overline{\mu}\mathbf{I}_p\\\mu\mathbf{I}_p&\end{bmatrix}\tag{4.18c}$$

   with $\mu\in\mathbb{C}\backslash\mathbb{R}$, $\mathrm{Im}(\mu)>0$ and $p\in\mathbb{N}$.

Moreover, the matrix $\mathbf{A}$ admits a (G,H)-polar decomposition if and only if each block $\mathbf{A}_j$ of the form (4.18b) *admits a $\mathbf{H}_j$-polar decomposition and each block of the form* (4.18c) *admits a decomposition such that*

$$\mathbf{A}_j=\begin{bmatrix}\mathbf{A}_{j,1}&\\&\mathbf{A}_{j,2}\end{bmatrix}=\begin{bmatrix}\mathbf{U}_{j,1}&\\&\mathbf{U}_{j,1}^{-*}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{j,1}&\\&\mathbf{M}_{j,1}^{*}\end{bmatrix}=\mathbf{U}_j\mathbf{M}_j,\tag{4.19}$$

where $\mathbf{U}_{j,1},\mathbf{M}_{j,1}\in\mathbb{C}^{p\times p}$.

Proof. According to Corollary 3.20 there exists a nonsingular matrix $\mathbf{S}\in\mathbb{C}^{n\times n}$ such that $\mathbf{H}$ and $\mathbf{G}$ take on the asserted form and it holds that

$$\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{GS}=\bigoplus_{j=1}^{r}(\mu_j\mathbf{I}_{p_j})\oplus\bigoplus_{j=r+1}^{s}(\mu_j\mathbf{I}_{p_j}\oplus\overline{\mu}_j\mathbf{I}_{p_j})$$

where $\mu_1,\ldots,\mu_r$ are the real and $\mu_{r+1},\ldots,\mu_s$ are the non-real eigenvalues with positive imaginary part of $\rho\mathbf{H}-\mathbf{G}$ ($\mu_i\neq\mu_j$ for $i\neq j$). Furthermore, $\mathbf{HAH}^{-1}=\mathbf{GAG}^{-1}$ implies that the matrices $\mathbf{S}^{-1}\mathbf{AS}$ and $\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{GS}$ commute, so that $\mathbf{S}^{-1}\mathbf{AS}$ must also have the asserted form.

Now, let $\mathbf{UM}$ be a (G,H)-polar decomposition of $\mathbf{A}$. Then $\mathbf{U}^H=\mathbf{U}^G=\mathbf{U}^{-1}$ and $\mathbf{M}^H=\mathbf{M}^G=\mathbf{M}$, so that $\mathbf{U}$ and $\mathbf{M}$ must have the same block structure as $\mathbf{A}$. Furthermore, $\mathbf{U}_j^{*}\mathbf{H}_j\mathbf{U}_j=\mathbf{H}_j$ and $\mathbf{M}_j^{*}\mathbf{H}_j=\mathbf{H}_j\mathbf{M}_j$, from which in the case of the non-real eigenvalues it also follows that $\mathbf{U}_{j,2}=\mathbf{U}_{j,1}^{-*}$ and $\mathbf{M}_{j,2}=\mathbf{M}_{j,1}^{*}$.

Conversely, if each block $\mathbf{A}_j$ admits the asserted decomposition, then it is easy to verify that $\mathbf{U}=\mathbf{S}(\mathbf{U}_1\oplus\ldots\oplus\mathbf{U}_k)\mathbf{S}^{-1}$ and $\mathbf{M}=\mathbf{S}(\mathbf{M}_1\oplus\ldots\oplus\mathbf{M}_k)\mathbf{S}^{-1}$ are the factors of a (G,H)-polar decomposition of $\mathbf{A}$. □

Whereas the existence of the $\mathbf{H}_j$-polar decomposition of the blocks (4.18b) can be deduced from Theorem 3.3 and Theorem 3.4, there is not yet a criterion for the existence of the decomposition (4.19) of the blocks (4.18c). However, from (4.19) it follows that $\mathbf{M}_{j,1} = \mathbf{U}_{j,1}^{-1}\mathbf{A}_{j,1} = \mathbf{A}_{j,2}^{*}\mathbf{U}_{j,1}$, so that this decomposition exists if and only if the equation $\mathbf{A}_{j,1} = \mathbf{U}_{j,1}\mathbf{A}_{j,2}^{*}\mathbf{U}_{j,1}$ can be solved for $\mathbf{U}_{j,1}$. The investigation of this special non-Hermitian algebraic Riccati equation requires some properties of the singular value decomposition which for clarity are reviewed next.

Proposition 4.18. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ $(m \geq n)$ and let*

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\mathbf{\Sigma}\\ \mathbf{0}\end{bmatrix}\mathbf{V}^{*}, \quad \mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$$

be a singular value decomposition where $\mathbf{U} \in \mathbb{C}^{m\times m}$, $\mathbf{V} \in \mathbb{C}^{n\times n}$ are unitary and $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. Then the following statements hold:

(*i*) *If* $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{\Sigma} = r$, *and if the corresponding partitioning*

$$\mathbf{U} = [\mathbf{U}_r \;\; \mathbf{U}_{m-r}], \quad \mathbf{\Sigma} = \mathbf{\Sigma}_r \oplus \mathbf{0}_{n-r}, \quad \mathbf{V} = [\mathbf{V}_r \;\; \mathbf{V}_{n-r}]$$

is made, then the columns of $\mathbf{U}_r$, $\mathbf{U}_{m-r}$, $\mathbf{V}_r$ and $\mathbf{V}_{n-r}$ form orthonormal bases of the subspaces $\operatorname{im}\mathbf{A}$, $(\operatorname{im}\mathbf{A})^{\perp}$, $(\ker\mathbf{A})^{\perp}$ *and* $\ker\mathbf{A}$, *respectively.*

(*ii*) *If* $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{\Sigma} = n$, *then the columns of*

$$\mathbf{B} = \mathbf{U}\begin{bmatrix}\mathbf{\Sigma}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}^{*}$$

form a basis which is dual with respect to the basis formed by the columns of $\mathbf{A}$.

Proof. (i) is obvious and (ii) follows from $\mathbf{B}^{*}\mathbf{A} = \mathbf{I}_n$. □

With the help of these properties the following theorem can now be proved.

Theorem 4.19. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$. Then there exists a nonsingular matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{X}\mathbf{B}^{*}\mathbf{X} \tag{4.20}$$

if and only if there exists a matrix $\mathbf{M} \in \mathbb{C}^{n\times n}$ such that

$$\mathbf{B}^{*}\mathbf{A} = \mathbf{M}^2 \;\; \textit{and} \;\; \ker\mathbf{A} = \ker\mathbf{M}, \;\; \ker\mathbf{B} = \ker\mathbf{M}^{*}. \tag{4.21}$$

Proof. [⇒]: Let $\mathbf{X}$ be a nonsingular solution of (4.20). For $\mathbf{M} = \mathbf{X}^{-1}\mathbf{A} = \mathbf{B}^{*}\mathbf{X}$ it then follows that $\mathbf{B}^{*}\mathbf{A} = \mathbf{M}^2$ as well as $\ker\mathbf{A} = \ker\mathbf{M}$. Since $\mathbf{X}^{*}\mathbf{B} = \mathbf{M}^{*}$, we also have $\ker\mathbf{B} = \ker\mathbf{M}^{*}$.

[⇐]: Let $\mathbf{M}$ be a matrix which satisfies (4.21) and let $p = \operatorname{rank}\mathbf{M}$. If $p = n$, then $\mathbf{A}, \mathbf{B}$ and $\mathbf{M}$ are nonsingular and therefore $\mathbf{X} = \mathbf{A}\mathbf{M}^{-1}$ is a solution of (4.20). Now assume that $p < n$. Then there exists a singular value decomposition

$$\mathbf{M} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^{*} = \begin{bmatrix}\mathbf{P}_1 & \mathbf{P}_2\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_1 & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Q}_1^{*}\\ \mathbf{Q}_2^{*}\end{bmatrix}$$

where $\mathbf{P}_1, \mathbf{Q}_1 \in \mathbb{C}^{n\times p}$, $\mathbf{P}_2, \mathbf{Q}_2 \in \mathbb{C}^{n\times(n-p)}$ and $\mathbf{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, $\sigma_i > 0$ for $1 \leq i \leq p$. Thus, on the one hand

$$\begin{aligned}\mathbf{MQ} &= [\mathbf{M}_1 \;\; \mathbf{0}] \text{ with } \mathbf{M}_1 = \mathbf{MQ}_1 = \mathbf{P}_1\mathbf{\Sigma}_1, \;\; \mathbf{M}_1^{\perp} = \mathbf{P}_2 \text{ and}\\ \mathbf{M}^*\mathbf{P} &= [\mathbf{M}_2 \;\; \mathbf{0}] \text{ with } \mathbf{M}_2 = \mathbf{M}^*\mathbf{P}_1 = \mathbf{Q}_1\mathbf{\Sigma}_1, \;\; \mathbf{M}_2^{\perp} = \mathbf{Q}_2,\end{aligned}$$

and since $\ker \mathbf{A} = \ker \mathbf{M}$ and $\ker \mathbf{B} = \ker \mathbf{M}^*$ on the other hand

$$\begin{aligned}\mathbf{AQ} &= [\mathbf{A}_1 \;\; \mathbf{0}] \text{ with } \mathbf{A}_1 = \mathbf{AQ}_1 \text{ and}\\ \mathbf{BP} &= [\mathbf{B}_2 \;\; \mathbf{0}] \text{ with } \mathbf{B}_2 = \mathbf{BP}_1.\end{aligned}$$

Therefore, the matrix $\mathbf{Y} = \mathbf{P}^*\mathbf{B}^*\mathbf{AQ} = \mathbf{P}^*\mathbf{M}^2\mathbf{Q}$ has the form

$$\mathbf{Y} = \mathbf{Y}_1 \oplus \mathbf{0}_{n-p} \text{ with } \mathbf{Y}_1 = \mathbf{B}_2^*\mathbf{A}_1 = \mathbf{M}_2^*\mathbf{M}_1 \,(= \mathbf{\Sigma}_1\mathbf{Q}_1^*\mathbf{P}_1\mathbf{\Sigma}_1) \in \mathbb{C}^{p\times p}.$$

For $q = \operatorname{rank} \mathbf{Y}_1$ now two cases must be considered:

Case 1): If $q = p$, let $\mathbf{A}_1^{\perp}$ and $\mathbf{B}_2^{\perp}$ be matrices whose columns form bases of $(\operatorname{im} \mathbf{A}_1)^{\perp}$ or $(\operatorname{im} \mathbf{B}_2)^{\perp}$, respectively, and let

$$\mathbf{A}_1' = (\mathbf{A}_1|\mathbf{B}_2^{\perp}), \;\; \mathbf{B}_2' = (\mathbf{B}_2|\mathbf{A}_1^{\perp}) \text{ and } \mathbf{M}_1' = (\mathbf{M}_1|\mathbf{M}_2^{\perp}), \;\; \mathbf{M}_2' = (\mathbf{M}_2|\mathbf{M}_1^{\perp}).$$

Then

$$(\mathbf{B}_2')^*\mathbf{A}_1' = \mathbf{Y}_1 \oplus \mathbf{Y}_A \text{ and } (\mathbf{M}_2')^*\mathbf{M}_1' = \mathbf{Y}_1 \oplus \mathbf{Y}_M$$

where $\mathbf{Y}_A, \mathbf{Y}_M \in \mathbb{C}^{(n-p)\times(n-p)}$ are nonsingular. Hence, for

$$\mathbf{A}_1'' = \mathbf{A}_1'(\mathbf{I}_p \oplus \mathbf{Y}_A^{-1}) \text{ and } \mathbf{M}_1'' = \mathbf{M}_1'(\mathbf{I}_p \oplus \mathbf{Y}_M^{-1}),$$

we obtain

$$(\mathbf{B}_2')^*\mathbf{A}_1'' = \mathbf{Y}_1 \oplus \mathbf{I}_{n-p} = (\mathbf{M}_2')^*\mathbf{M}_1''.$$

Thus, the matrix

$$\mathbf{X} = \mathbf{A}_1''(\mathbf{M}_1'')^{-1}$$

is a solution of (4.20).

Case 2): If $q < p$, there exists a singular value decomposition

$$\mathbf{Y}_1 = \mathbf{R}\mathbf{\Omega}\mathbf{S}^* \text{ with } \mathbf{R}, \mathbf{S} \in \mathbb{C}^{p\times p}, \;\; \mathbf{\Omega} = \mathbf{\Omega}_1 \oplus \mathbf{0}_{p-q},$$

where $\mathbf{\Omega}_1 = \operatorname{diag}(\omega_1, \ldots, \omega_q)$ is nonsingular. By setting

$$\begin{aligned}\mathbf{A}_1' &= \mathbf{A}_1\mathbf{S}(\mathbf{\Omega}_1^{-1} \oplus \mathbf{I}_{p-q}), \;\; \mathbf{B}_2' = \mathbf{B}_2\mathbf{R} \text{ and}\\ \mathbf{M}_1' &= \mathbf{M}_1\mathbf{S}(\mathbf{\Omega}_1^{-1} \oplus \mathbf{I}_{p-q}), \;\; \mathbf{M}_2' = \mathbf{M}_2\mathbf{R},\end{aligned}$$

we obtain

$$(\mathbf{B}_2')^*\mathbf{A}_1' = (\mathbf{M}_2')^*\mathbf{M}_1' = \mathbf{I}_q \oplus \mathbf{0}_{p-q}.$$

Now, let $\mathbf{D}_2 = [\mathbf{d}_1 \ldots \mathbf{d}_p]$ and $\mathbf{C}_1 = [\mathbf{c}_1 \ldots \mathbf{c}_p]$ be matrices whose columns form dual bases with respect to the bases of $\operatorname{im} \mathbf{A}_1'$ or $\operatorname{im} \mathbf{B}_2'$ formed by the columns of $\mathbf{A}_1' = [\mathbf{a}_1 \ldots \mathbf{a}_p]$ or $\mathbf{B}_2' = [\mathbf{b}_1 \ldots \mathbf{b}_p]$, respectively. Then

$$(\mathbf{D}_2)^*\mathbf{A}_1' = (\mathbf{B}_2')^*\mathbf{C}_1 = \mathbf{I}_p,$$

and thus the matrices defined by

$$\mathbf{A}_1'' = [\mathbf{a}_1 \dots \mathbf{a}_q\ \mathbf{a}_{q+1} \dots \mathbf{a}_p\ \mathbf{c}_{q+1} \dots \mathbf{c}_p] \text{ and}$$
$$\mathbf{B}_2'' = [\mathbf{b}_1 \dots \mathbf{b}_q\ \mathbf{b}_{q+1} \dots \mathbf{b}_p\ \mathbf{d}_{q+1} \dots \mathbf{d}_p]$$

satisfy

$$(\mathbf{B}_2'')^* \mathbf{A}_1'' = \mathbf{I}_q \oplus \begin{bmatrix} \mathbf{0}_{p-q} & \mathbf{I}_{p-q} \\ \mathbf{I}_{p-q} & \mathbf{W} \end{bmatrix}$$

with $\mathbf{W} = [w_{\mu\nu}]$, $w_{\mu\nu} = (\mathbf{c}_{q+\nu}, \mathbf{d}_{q+\mu})$ for $1 \leq \mu, \nu \leq p-q$. Consequently, the matrix

$$\mathbf{A}_1''' = [\mathbf{a}_1 \dots \mathbf{a}_q\ \mathbf{a}_{q+1} \dots \mathbf{a}_p\ \mathbf{c}'_{q+1} \dots \mathbf{c}'_p] \text{ with } \mathbf{c}'_k = \mathbf{c}_k - \sum_{j=q+1}^{p} (\mathbf{c}_k, \mathbf{d}_j)\, \mathbf{a}_j$$

for $q+1 \leq k \leq p$, fulfills

$$(\mathbf{B}_2'')^* \mathbf{A}_1''' = \mathbf{I}_q \oplus \begin{bmatrix} \mathbf{0}_{p-q} & \mathbf{I}_{p-q} \\ \mathbf{I}_{p-q} & \mathbf{0}_{p-q} \end{bmatrix}.$$

Again, let $(\mathbf{A}_1''')^\perp$ and $(\mathbf{B}_2'')^\perp$ be matrices whose columns form bases of $(\operatorname{im} \mathbf{A}_1''')^\perp$ or $(\operatorname{im} \mathbf{B}_2'')^\perp$, respectively, and let

$$\widetilde{\mathbf{A}}_1 = \left(\mathbf{A}_1'''|(\mathbf{B}_2'')^\perp\right) \text{ and } \widetilde{\mathbf{B}}_2 = \left(\mathbf{B}_2''|(\mathbf{A}_1''')^\perp\right).$$

Then

$$(\widetilde{\mathbf{B}}_2)^* \widetilde{\mathbf{A}}_1 = \mathbf{I}_q \oplus \begin{bmatrix} \mathbf{0}_{p-q} & \mathbf{I}_{p-q} \\ \mathbf{I}_{p-q} & \mathbf{0}_{p-q} \end{bmatrix} \oplus \mathbf{Y}_A$$

where $\mathbf{Y}_A \in \mathbb{C}^{(n-2p+q)\times(n-2p+q)}$ is nonsingular. Hence, for

$$\widetilde{\widetilde{\mathbf{A}}}_1 = \widetilde{\mathbf{A}}_1 (\mathbf{I}_{2p-q} \oplus \mathbf{Y}_A^{-1})$$

we finally obtain

$$(\widetilde{\mathbf{B}}_2)^* \widetilde{\widetilde{\mathbf{A}}}_1 = \mathbf{I}_q \oplus \begin{bmatrix} \mathbf{0}_{p-q} & \mathbf{I}_{p-q} \\ \mathbf{I}_{p-q} & \mathbf{0}_{p-q} \end{bmatrix} \oplus \mathbf{I}_{n-2p+q} = \mathbf{Z} \text{ with } \mathbf{Z}^2 = \mathbf{I}.$$

Starting with $\mathbf{M}_1'$ and $\mathbf{M}_2'$ in the same way the matrices $\widetilde{\widetilde{\mathbf{M}}}_1$ and $\widetilde{\mathbf{M}}_2$ can be constructed which also satisfy

$$(\widetilde{\mathbf{M}}_2)^* \widetilde{\widetilde{\mathbf{M}}}_1 = \mathbf{Z}.$$

Thus, the matrix

$$\mathbf{X} = \widetilde{\widetilde{\mathbf{A}}}_1 (\widetilde{\widetilde{\mathbf{M}}}_1)^{-1} = \widetilde{\widetilde{\mathbf{A}}}_1 \mathbf{Z} (\widetilde{\mathbf{M}}_2)^*$$

is a solution of (4.20). □

The square roots of $\mathbf{B}^*\mathbf{A}$ can be calculated based on the well-known results derived in [CL], [G, Chapter VIII, §7], [WED, Section 8.06]. A corresponding numerical algorithm is given in [BF]. Thus, Theorem 4.17 presents a quite general necessary and sufficient condition for the existence of (G,H)-polar decompositions which can be applied using the solution of the Riccati equation $\mathbf{A} = \mathbf{X}\mathbf{B}^*\mathbf{X}$ provided in Theorem 4.19.

4.7 Numerical computation of (G,H)-polar decompositions

To be able to solve (G,H)-unitary Procrustes problems numerically this final section explains how (G,H)-polar decompositions can be computed. For this purpose not really new algorithms are required. It is merely necessary to apply the methods from Section 3.5 and Section 3.6 correctly. For simplification of the presentation it is assumed that $\mathbf{G}$ and $\mathbf{H}$ satisfy $\mathbf{H}^{-1}\mathbf{G} = \mu^2\mathbf{G}^{-1}\mathbf{H}$ for some real $\mu \neq 0$.

Let $\mathbf{A}$ be a matrix such that $\mathbf{A}^H\mathbf{A}$ has no non-positive eigenvalues. Then Algorithm 3.31 can be applied and according to Theorem 3.30 it computes the particular H-polar decomposition $\mathbf{A} = \mathbf{U}_0\mathbf{M}_0$ for which $\sigma(\mathbf{M}_0)$ lies in the open right complex half-plane. If $\mathbf{A}^H\mathbf{A}$ additionally is diagonalisable, this decomposition can also be computed with the method given in Remark 3.26 where it must be chosen such that $\mathrm{Re}(\omega_j) > 0$ and $\varepsilon_j = +1$.

Now assume that $\mathbf{A}$ furthermore satisfies $\mathbf{A}^H = \mathbf{A}^G$. Then the computed H-polar decomposition, according to Lemma 4.5, is a G-polar decomposition, too[10]. If $\mathbf{H}$ additionally is positive or negative definite, then $\mathbf{A} = \mathbf{U}_0\mathbf{M}_0$ or $\mathbf{A} = (-\mathbf{U}_0)(-\mathbf{M}_0)$, respectively, is a definite H-polar decomposition (see the explanations following Algorithm 3.31) and consequently it is also an H-definite (G,H)-polar decomposition. Thus, in these cases an H-definite (G,H)-polar decomposition is simply obtained by computing a definite H-polar decomposition with Algorithm 3.31 or with the method described in Remark 3.26.

In general it is more difficult to compute such a decomposition. In fact, Example 4.8 shows that if $\mathbf{H}$ is indefinite or $\mathbf{A}$ is singular, then a definite or semidefinite H-polar decomposition of $\mathbf{A}$ need not be a G-polar decomposition, too. Hence, in general it is necessary to transform $\mathbf{A}$ and $\mathbf{H}$ into the form of Lemma 4.3 and then to compute (semi)definite $\mathbf{J}_k$-polar decompositions of the blocks $\mathbf{A}_k$ according to Lemma 4.4. The required transformation $\mathbf{S}$ can be determined by computing the simplified canonical form of the pair $(\mathbf{H}^{-1}\mathbf{G}, \mathbf{H})$ with Method 3.24. Thus we obtain the following algorithm.

Algorithm 4.20. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a matrix which has H-semidefinite (G,H)-polar decompositions. Then such a decomposition $\mathbf{A} = \mathbf{U}\mathbf{M}$ can be computed with the following steps:*

1. *Compute the simplified canonical form*

$$(\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{G}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mu\mathbf{I}_{p+q} \oplus -\mu\mathbf{I}_{r+s}, \mathbf{J}_1 \oplus \mathbf{J}_2)$$
$$\mathbf{J}_1 = \mathbf{I}_p \oplus -\mathbf{I}_q, \quad \mathbf{J}_2 = \mathbf{I}_r \oplus -\mathbf{I}_s$$

   of the pair $(\mathbf{H}^{-1}\mathbf{G}, \mathbf{H})$ with Method 3.24.

2. *Compute $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{A}_1 \oplus \mathbf{A}_2$.*

3. *Compute semidefinite $\mathbf{J}_k$-polar decompositions $\mathbf{A}_k = \mathbf{U}_k\mathbf{M}_k$ for $k = 1, 2$.*

4. *Compute $\mathbf{U} = \mathbf{S}(\mathbf{U}_1 \oplus \mathbf{U}_2)\mathbf{S}^{-1}$ and $\mathbf{M} = \mathbf{S}(\mathbf{M}_1 \oplus \mathbf{M}_2)\mathbf{S}^{-1}$.*

[10] This statement also holds when $\rho\mathbf{H} - \mathbf{G}$ is a non-defective Hermitian pencil which can be shown using a corresponding generalisation of Lemma 4.5 obtained from Theorem 4.17 and Theorem 4.19.

This algorithm can easily be adopted for the case in which $\rho\mathbf{H} - \mathbf{G}$ is a non-defective Hermitian pencil. For its application the following rules hold:

(a) The method for computing the decompositions in step (3) depends on the matrices $\mathbf{A}_k$ and $\mathbf{J}_k$. If $\mathbf{J}_k = \pm\mathbf{I}$, then $\mathbf{A}_k = (\pm\mathbf{U}_k)(\pm\mathbf{M}_k)$ is an ordinary polar decomposition. Otherwise, if $\mathbf{A}_k$ is nonsingular, then Method 3.25 or Algorithm 3.32 can be applied. If $\mathbf{A}$ is singular, then Method 3.25 must be used.

(b) If the decompositions in step (3) are not semidefinite, the algorithm still computes a (G,H)-polar decomposition, but it is not H-semidefinite.

In the most important application for solving the (G,H)-isometric Procrustes problem Algorithm 4.20 becomes particularly simple. Indeed, if it is assumed that the matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{F}^{n\times N}$ appearing in (4.11) are constructed by Theorem 4.9, then the internal metric has the form $\mathbf{G} = \mathbf{I}_p \oplus -\mathbf{I}_{n-p}$ and the external metric will usually be defined by $\mathbf{H} = \mathbf{I}_n$. Consequently, the matrix $\mathbf{C}$ from (4.16) is given by

$$\mathbf{C} = \mathbf{Y}\mathbf{X}^*\mathbf{H} + \mathbf{G}^{-1}\mathbf{H}\mathbf{Y}\mathbf{X}^*\mathbf{G} = 2\begin{bmatrix}\mathbf{Y}_1\mathbf{X}_1^* & \mathbf{0}\\ \mathbf{0} & \mathbf{Y}_2\mathbf{X}_2^*\end{bmatrix} \text{ for } \mathbf{X} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix},\ \mathbf{Y} = \begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix}$$

where $\mathbf{X}_1, \mathbf{Y}_1 \in \mathbb{F}^{p\times N}$ and $\mathbf{X}_2, \mathbf{Y}_2 \in \mathbb{F}^{(n-p)\times N}$. Since $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{C}$ are already in the form of Lemma 4.3, only step (3) of Algorithm 4.20 is required to compute an H-semidefinite (G,H)-polar decomposition of $\mathbf{C}$. Moreover, $\mathbf{J}_1 = \mathbf{I}_p$ and $\mathbf{J}_2 = \mathbf{I}_{n-p}$, so that in step (3) merely the ordinary polar decompositions

$$\mathbf{Y}_1\mathbf{X}_1 = \mathbf{U}_1\mathbf{M}_1 \text{ and } \mathbf{Y}_2\mathbf{X}_2 = \mathbf{U}_2\mathbf{M}_2$$

have to be computed where the factor 2 is taken into $\mathbf{M}_1$ and $\mathbf{M}_2$. Finally, $\mathbf{U} = \mathbf{U}_1 \oplus \mathbf{U}_2$ is the wanted (G,H)-isometry and

$$\mathbf{X}' = \mathbf{U}\mathbf{X} = \begin{bmatrix}\mathbf{U}_1\mathbf{X}_1\\ \mathbf{U}_2\mathbf{X}_2\end{bmatrix}$$

are the transformed coordinates for which $\mathbf{X}'$ and $\mathbf{Y}$ are optimally congruent.

Chapter 5

A Newton method for the numerical solution of Procrustes problems

5.1 Introduction

In the previous chapter we discussed the Procrustes problems

$$f(\mathbf{U}) = \mathrm{tr}[(\mathbf{UX}-\mathbf{Y})^*\mathbf{H}(\mathbf{UX}-\mathbf{Y})] \to \begin{cases} \min, & \text{if } \mathbf{H} > 0 \\ \max, & \text{if } \mathbf{H} < 0 \\ \min \,/\, \max, & \text{otherwise} \end{cases} \tag{5.1}$$

subject to the constraints

$$\mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{HU} - \mathbf{H} = \mathbf{0} \tag{5.2}$$

or

$$\mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{GU} - \mathbf{G} = \mathbf{0} \;\text{ and }\; \mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{HU} - \mathbf{H} = \mathbf{0}. \tag{5.3}$$

The necessary conditions for solving these problems were obtained as

$$\mathbf{HU}(\mathbf{XX}^* + \mathbf{\Lambda}_H) = \mathbf{HYX}^* \tag{5.4}$$

or

$$\mathbf{GU}\mathbf{\Lambda}_G + \mathbf{HU}(\mathbf{XX}^* + \mathbf{\Lambda}_H) = \mathbf{HYX}^*, \tag{5.5}$$

where $\mathbf{\Lambda}_G$ and $\mathbf{\Lambda}_H$ are unknown selfadjoint matrices of the Lagrange multipliers. These equations were then transformed into the form

$$\mathbf{UM} = \mathbf{A}$$

where either

$$\mathbf{U}^H = \mathbf{U}^{-1}, \;\; \mathbf{M}^H = \mathbf{M}$$

or

$$\mathbf{U}^H = \mathbf{U}^G = \mathbf{U}^{-1}, \;\; \mathbf{M}^G = \mathbf{M}^H = \mathbf{M}$$

which is trivial for (5.4) and is possible for (5.5) by assuming that

$$\mathbf{H}^{-1}\mathbf{G} = \mu^2\mathbf{G}^{-1}\mathbf{H} \text{ for some } \mu \in \mathbb{R}\backslash\{0\}.$$

In this way the wanted H- or (G,H)-isometry $\mathbf{U}$ can be expressed as the isometric factor of an H- or a (G,H)-polar decomposition of a known matrix $\mathbf{A}$ (see Theorem 4.14 and Theorem 4.16).

In this chapter we are now interested in the optimisation of (5.1) subject to the constraints

$$\mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G} = \mathbf{0}, \tag{5.6}$$

where the associated necessary condition becomes

$$\mathbf{G}\mathbf{U}\mathbf{\Lambda}_G + \mathbf{H}\mathbf{U}\mathbf{X}\mathbf{X}^* = \mathbf{H}\mathbf{Y}\mathbf{X}^*. \tag{5.7}$$

To avoid that (5.7) can be reduced to (5.4) it is furthermore assumed that

$$\mathbf{G} \neq \mu\mathbf{H} \text{ for all } \mu \in \mathbb{R}.$$

Clearly, the most interesting of these Procrustes problems are those in which the internal metric $\mathbf{G}$ is indefinite but the external metric $\mathbf{H}$ is definite. For example, if $\mathbf{X}$ and $\mathbf{Y}$ are constructed according to Theorem 4.9, the internal metric has the form $\mathbf{G} = \mathbf{I}_p \oplus -\mathbf{I}_{n-p}$. If now $\mathbf{U}$ is determined such that $f(\mathbf{U})$ is a minimum for $\mathbf{H} = \mathbf{I}_n$, then the matrix $\mathbf{X}' = \mathbf{U}\mathbf{X}$ minimises $\|\mathbf{X}' - \mathbf{Y}\|_F$ under the constraints $(\mathbf{X}')^*\mathbf{G}(\mathbf{X}') = \mathbf{X}^*\mathbf{G}\mathbf{X}$. This is exactly what is wanted.

Unfortunately, we were not able to transform (5.7) similar to (5.4) or (5.5) and to derive a corresponding expression for $\mathbf{U}$. Therefore, this chapter presents a Newton method with which $\mathbf{U}$ can be determined numerically.

The method will be designed to solve various constrained optimisation problems where the constraints are given by (5.2), (5.3), or (5.6), respectively. Although it even applies for optimising (5.1) in the case of an indefinite matrix $\mathbf{H}$, we will mostly consider the case in which $\mathbf{H}$ is positive definite. Then the minimum of f has to be determined and since $f(\mathbf{U})$ is bounded from below it is ensured that it always exists.

The method is described in Section 5.2 where not only the algorithm is derived but also its applicability is discussed. In Section 5.3 some numerical results are presented.

5.2 Description of the method

For the description of the Newton method the notation

$$\mathbf{x} \cdot \mathbf{y} = \sum_{\alpha=1}^{n} x_\alpha y_\alpha \text{ where } \mathbf{x} = (x_\alpha),\ \mathbf{y} = (y_\alpha) \in \mathbb{F}^n$$

is used. Furthermore, the conjugation is written explicitly where it is to be applied, so that the ordinary scalar product in the case $\mathbb{F} = \mathbb{C}$ is given by

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \overline{\mathbf{y}}.$$

The method is based on the following well-known results on constrained optimisation problems which, for example, are proved in [ER, Kapitel V.6].

Let $U \subset \mathbb{R}^n$ be an open set and let $C^2(U)$ be the class of twice continuously differentiable functions in U. Furthermore, let $f : U \to \mathbb{R}$ be an objective function in $C^2(U)$, $\mathbf{g} : U \to \mathbb{R}^m$ a constraint function in $C^2(U)$, and let solution $\mathbf{u} \in U$ of the optimisation problem

$$\left\{\begin{matrix} f(\mathbf{u}) \to \min \\ \mathbf{g}(\mathbf{u}) = \mathbf{0} \end{matrix}\right\} \quad \text{or} \quad \left\{\begin{matrix} f \to \min \\ g_1 = 0 \\ \vdots \\ g_m = 0 \end{matrix}\right\} \tag{5.8}$$

to be determined. Then the associated Lagrange function is defined by

$$l(\mathbf{u}, \lambda) = f(\mathbf{u}) + \lambda \cdot \mathbf{g}(\mathbf{u}) = f(\mathbf{u}) + \sum_{j=1}^{m} \lambda_j g_j(\mathbf{u}) \tag{5.9}$$

and the necessary first order condition for solving the problem is

$$\mathbf{F}(\mathbf{u}, \lambda) = \begin{pmatrix} \nabla \mathbf{f}(\mathbf{u}) + \lambda \cdot \nabla \mathbf{g}(\mathbf{u}) \\ \mathbf{g}(\mathbf{u}) \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial u_1} + \sum_{j=1}^{m} \lambda_j \frac{\partial g_j}{\partial u_1} \\ \vdots \\ \frac{\partial f}{\partial u_n} + \sum_{j=1}^{m} \lambda_j \frac{\partial g_j}{\partial u_n} \\ g_1 \\ \vdots \\ g_m \end{pmatrix} = \mathbf{0}. \tag{5.10}$$

Here $\lambda \in \mathbb{R}^m$ represents the vector of the unknown Lagrange multipliers and $\mathbf{F}(\mathbf{u}, \lambda)$ the gradient of the Lagrange function. If $(\mathbf{u}_0, \lambda_0)$ is a solution of (5.10), and

$$\mathbf{L}(\mathbf{u}, \lambda) = \nabla^2 \mathbf{f}(\mathbf{u}) + \lambda \cdot \nabla^2 \mathbf{g}(\mathbf{u}) = \left[\frac{\partial^2 f}{\partial u_\alpha \partial u_\beta} + \sum_{j=1}^{m} \lambda_j \frac{\partial^2 g_j}{\partial u_\alpha \partial u_\beta} \right] \tag{5.11}$$

denotes the associated Lagrange matrix ($1 \leq \alpha, \beta \leq n$), and also

$$\begin{aligned} M &= \operatorname{span}\{\nabla \mathbf{g}_1(\mathbf{u}_0), \ldots, \nabla \mathbf{g}_m(\mathbf{u}_0)\}, \\ M^\perp &= \{\mathbf{y} \in \mathbb{R}^m : \mathbf{x} \cdot \mathbf{y} = 0 \text{ for all } \mathbf{x} \in M\}, \end{aligned} \tag{5.12}$$

then the necessary second order condition for a minimum is

$$\mathbf{L}(\mathbf{u}_0, \lambda_0)\mathbf{y} \cdot \mathbf{y} \geq 0 \ \text{ for all } \ \mathbf{y} \in M^\perp. \tag{5.13}$$

If $\mathbf{L}(\mathbf{u}_0, \lambda_0)$ is not only positive semidefinite, but positive definite on $M^\perp$

$$\mathbf{L}(\mathbf{u}_0, \lambda_0)\mathbf{y} \cdot \mathbf{y} > 0 \ \text{ for all } \ \mathbf{y} \in M^\perp, \tag{5.14}$$

the sufficient second order condition for a strict local minimum of f in $\mathbf{u}_0$ holds.

In order to apply these equations to a complex optimisation problem $f : U \subset \mathbb{C}^n \to \mathbb{R}$, $\mathbf{g} : U \subset \mathbb{C}^n \to \mathbb{C}^m$ it suffices to split f and $\mathbf{g}$ into their real and imaginary parts and to represent them as real functions $f^\wedge : U^\wedge \subset \mathbb{R}^{2n} \to \mathbb{R}$,

$\mathbf{g}^\wedge : U^\wedge \subset \mathbb{R}^{2n} \to \mathbb{R}^{2m}$. Furthermore, the system of equations $\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}$ can be solved iteratively using the Newton method

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{DF}^{-1}(\mathbf{z}_i)\,\mathbf{F}(\mathbf{z}_i) \;\text{ with }\; \mathbf{z}_i = (\mathbf{u}_i \;\; \lambda_i)^T \in \mathbb{R}^{n+m}, \tag{5.15}$$

for which the components of the function $\mathbf{F}$ and the components of its Jacobi matrix

$$\mathbf{DF}(\mathbf{u}, \lambda) = \begin{bmatrix} \nabla^2\mathbf{f}(\mathbf{u}) + \lambda \cdot \nabla^2\mathbf{g}(\mathbf{u}) & \nabla\mathbf{g}(\mathbf{u}) \\ \nabla\mathbf{g}(\mathbf{u})^T & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{L}(\mathbf{u}, \lambda) & \nabla\mathbf{g}(\mathbf{u}) \\ \nabla\mathbf{g}(\mathbf{u})^T & \mathbf{0} \end{bmatrix} \tag{5.16}$$

are required. If the Newton method converges to a solution $(\mathbf{u}_0, \lambda_0)$ of (5.10), i.e. to a stationary point of (5.9), then $\mathbf{DF}(\mathbf{u}_0, \lambda_0)$ contains the Lagrange matrix $\mathbf{L}(\mathbf{u}_0, \lambda_0)$ as well as a basis of M, with the help of which (5.14) can also be verified.

On the basis of these principles, the next two subsections will be concerned with bringing the objective functions (5.1) and the constraints (5.6) into a form from which the components of $\mathbf{F}$ and $\mathbf{DF}$ can be calculated. Thereafter the Newton method will be specified and its applicability and starting values will be discussed.

5.2.1 Transformation of the objective function

To transform the objective function

$$f(\mathbf{U}) = \operatorname{tr}[(\mathbf{UX} - \mathbf{Y})^*\mathbf{H}(\mathbf{UX} - \mathbf{Y})] \tag{5.17}$$

into an appropriate form, the Kronecker product and vectorisation operator are required.

Definition 5.1 (Kronecker product and vectorisation operator).

(*i*) Let $\mathbf{A} = [a_{\mu\nu}] \in \mathbb{F}^{m\times n}$ and let $\mathbf{B} \in \mathbb{F}^{p\times q}$. The matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \in \mathbb{F}^{mp\times nq}$$

is called the Kronecker product of the matrices $\mathbf{A}$ and $\mathbf{B}$.

(*ii*) Let $\mathbf{a}_1, \ldots, \mathbf{a}_m$ be the row vectors of the matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$. The vector

$$\operatorname{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix} \in \mathbb{F}^{mn}$$

is called the vector of the matrix $\mathbf{A}$.

The properties of these operators are discussed in detail in [GB, Chapter 8, 9], where in particular the bilinearity of the Kronecker product and the following calculation rules are proved.

Lemma 5.2. *Let* $\mathbf{A} \in \mathbb{F}^{m\times m}$, $\mathbf{B} \in \mathbb{F}^{n\times n}$ *and* $\mathbf{U}, \mathbf{V} \in \mathbb{F}^{m\times n}$. *Then*

$$(\mathbf{A} \otimes \mathbf{B}^T)\operatorname{vec}(\mathbf{U}) = \operatorname{vec}(\mathbf{AUB}) \;\; \textit{and} \;\; \operatorname{tr}(\mathbf{V}^*\mathbf{U}) = \operatorname{vec}(\mathbf{U}) \cdot \overline{\operatorname{vec}(\mathbf{V})}.$$

Using this lemma, the objective function (5.17) can be transformed into the form

$$\begin{aligned} f(\mathbf{U}) &= \operatorname{tr}(\mathbf{U}^*\mathbf{H}\mathbf{U}\mathbf{X}\mathbf{X}^*) - 2\operatorname{Re}\operatorname{tr}(\mathbf{U}^*\mathbf{H}\mathbf{Y}\mathbf{X}^*) + \operatorname{tr}(\mathbf{H}\mathbf{Y}\mathbf{Y}^*) \\ &= \operatorname{vec}(\mathbf{H}\mathbf{U}\mathbf{X}\mathbf{X}^*) \cdot \overline{\operatorname{vec}(\mathbf{U})} - 2\operatorname{Re}\operatorname{vec}(\mathbf{H}\mathbf{Y}\mathbf{X}^*) \cdot \overline{\operatorname{vec}(\mathbf{U})} + \operatorname{tr}(\mathbf{H}\mathbf{Y}\mathbf{Y}^*) \\ &= [\mathbf{H} \otimes (\mathbf{X}\mathbf{X}^*)^T] \operatorname{vec}(\mathbf{U}) \cdot \overline{\operatorname{vec}(\mathbf{U})} - 2\operatorname{Re}\operatorname{vec}(\mathbf{H}\mathbf{Y}\mathbf{X}^*) \cdot \overline{\operatorname{vec}(\mathbf{U})} + \operatorname{tr}(\mathbf{H}\mathbf{Y}\mathbf{Y}^*). \end{aligned}$$

This corresponds to the general function

$$\begin{aligned} &f(\mathbf{u}) = \mathbf{A}\mathbf{u} \cdot \overline{\mathbf{u}} - 2\operatorname{Re}\mathbf{b} \cdot \overline{\mathbf{u}} + \gamma \ \text{ with} \\ &\mathbf{A}^* = \mathbf{A} \in \mathbb{F}^{n^2 \times n^2}, \ \ \mathbf{b} \in \mathbb{F}^{n^2}, \ \ \gamma \in \mathbb{R} \end{aligned} \tag{5.18}$$

of the vector $\mathbf{u} = \operatorname{vec}(\mathbf{U}) \in \mathbb{F}^{n^2}$ by setting

$$\mathbf{A} = \mathbf{H} \otimes (\mathbf{X}\mathbf{X}^*)^T, \ \ \mathbf{b} = \operatorname{vec}(\mathbf{H}\mathbf{Y}\mathbf{X}^*), \ \ \gamma = \operatorname{tr}(\mathbf{H}\mathbf{Y}\mathbf{Y}^*). \tag{5.19}$$

In the case $\mathbb{F} = \mathbb{R}$ the gradient of this quadratic form needed in (5.10) is directly given as

$$\nabla \mathbf{f}(\mathbf{u}) = 2(\mathbf{A}\mathbf{u} - \mathbf{b}). \tag{5.20}$$

In the case $\mathbb{F} = \mathbb{C}$ it can also be obtained by taking the real derivatives, for which a real representation of (5.18) is required:

Let the real and imaginary part of a complex matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ be denoted by $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, and let the same apply to complex vectors $\mathbf{u} = \mathbf{u}_1 + i\mathbf{u}_2 \in \mathbb{C}^n$ and complex scalars $\lambda = \lambda_1 + i\lambda_2 \in \mathbb{C}$. Furthermore, let the real representations of $\mathbf{A}$ and $\mathbf{u}$ of the first and second kind be defined by

$$\begin{aligned} \mathbf{A}^\wedge &= \begin{bmatrix} \mathbf{A}_1 & -\mathbf{A}_2 \\ \mathbf{A}_2 & \mathbf{A}_1 \end{bmatrix} \in \mathbb{R}^{2m\times 2n}, & \mathbf{u}^\wedge &= \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \in \mathbb{R}^{2n}, \\ \mathbf{A}^\vee &= \begin{bmatrix} \mathbf{A}_2 & \mathbf{A}_1 \\ -\mathbf{A}_1 & \mathbf{A}_2 \end{bmatrix} \in \mathbb{R}^{2m\times 2n}, & \mathbf{u}^\vee &= \begin{pmatrix} \mathbf{u}_2 \\ -\mathbf{u}_1 \end{pmatrix} \in \mathbb{R}^{2n}. \end{aligned}$$

Then the following calculation rules hold, whose proof is obtained by simple verification.

Lemma 5.3 (Real representation of complex matrix equations). *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$, $\mathbf{C} \in \mathbb{C}^{n\times k}$, $\mathbf{D} \in \mathbb{C}^{p\times q}$, $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ *and* $\lambda \in \mathbb{C}$. *Then*

1. $(\lambda\mathbf{A})^\wedge = \lambda_1\mathbf{A}^\wedge - \lambda_2\mathbf{A}^\vee$, $(\lambda\mathbf{A})^\vee = \lambda_1\mathbf{A}^\vee + \lambda_2\mathbf{A}^\wedge$,
2. $(\lambda\mathbf{u})^\wedge = \lambda_1\mathbf{u}^\wedge - \lambda_2\mathbf{u}^\vee$, $(\lambda\mathbf{u})^\vee = \lambda_1\mathbf{u}^\vee + \lambda_2\mathbf{u}^\wedge$
3. $(\mathbf{A}+\mathbf{B})^\wedge = \mathbf{A}^\wedge + \mathbf{B}^\wedge$, $(\mathbf{A}+\mathbf{B})^\vee = \mathbf{A}^\vee + \mathbf{B}^\vee$,
4. $(\mathbf{u}+\mathbf{v})^\wedge = \mathbf{u}^\wedge + \mathbf{v}^\wedge$, $(\mathbf{u}+\mathbf{v})^\vee = \mathbf{u}^\vee + \mathbf{v}^\vee$
5. $(\mathbf{A}\mathbf{C})^\wedge = \mathbf{A}^\wedge\mathbf{C}^\wedge = -\mathbf{A}^\vee\mathbf{C}^\vee$, $(\mathbf{A}\mathbf{C})^\vee = \mathbf{A}^\wedge\mathbf{C}^\vee = \mathbf{A}^\vee\mathbf{C}^\wedge$,
6. $(\mathbf{A}\mathbf{u})^\wedge = \mathbf{A}^\wedge\mathbf{u}^\wedge = -\mathbf{A}^\vee\mathbf{u}^\vee$, $(\mathbf{A}\mathbf{u})^\vee = \mathbf{A}^\wedge\mathbf{u}^\vee = \mathbf{A}^\vee\mathbf{u}^\wedge$,
7. $\operatorname{Re}(\mathbf{u}\cdot\overline{\mathbf{v}}) = \mathbf{u}^\wedge\mathbf{v}^\wedge = \mathbf{u}^\vee\mathbf{v}^\vee$, $\operatorname{Im}(\mathbf{u}\cdot\overline{\mathbf{v}}) = -\mathbf{u}^\wedge\mathbf{v}^\vee = \mathbf{u}^\vee\mathbf{v}^\wedge$,
8. $(\mathbf{A}\otimes\mathbf{D})^\wedge = (\mathbf{A}^\wedge \otimes \mathbf{D}_1 - \mathbf{A}^\vee \otimes \mathbf{D}_2)$,
   $(\mathbf{A}\otimes\mathbf{D})^\vee = (\mathbf{A}^\vee \otimes \mathbf{D}_1 + \mathbf{A}^\wedge \otimes \mathbf{D}_2)$,

9. $(\mathbf{A}^*)^\wedge = (\overline{\mathbf{A}}^T)^\wedge = (\mathbf{A}^\wedge)^T$, $(\mathbf{A}^*)^\vee = (\overline{\mathbf{A}}^T)^\vee = -(\mathbf{A}^\vee)^T$,

10. $\mathbf{A}^\wedge = (\mathbf{A}^\wedge)^T$, $\mathbf{A}^\vee = -(\mathbf{A}^\vee)^T$, *if* $m = n$ *and* $\mathbf{A}^* = \mathbf{A}$,

11. $\mathbf{B}^\wedge = -(\mathbf{B}^\wedge)^T$, $\mathbf{B}^\vee = (\mathbf{B}^\vee)^T$, *if* $m = n$ *and* $\mathbf{B}^* = -\mathbf{B}$.

With the help of this Lemma it follows from (5.18) that

$$\mathbf{A}\mathbf{u} \cdot \overline{\mathbf{u}} = \mathrm{Re}(\mathbf{A}\mathbf{u} \cdot \overline{\mathbf{u}}) = (\mathbf{A}\mathbf{u})^\wedge \cdot \mathbf{u}^\wedge = \mathbf{A}^\wedge \mathbf{u}^\wedge \cdot \mathbf{u}^\wedge \quad \text{and} \quad \mathrm{Re}(\mathbf{b} \cdot \overline{\mathbf{u}}) = \mathbf{b}^\wedge \cdot \mathbf{u}^\wedge.$$

Hence, the equivalent real representation in the case $\mathbb{F} = \mathbb{C}$ is given by the quadratic form

$$\begin{aligned} &f(\mathbf{u}^\wedge) = \mathbf{A}^\wedge \mathbf{u}^\wedge \cdot \mathbf{u}^\wedge - 2\mathbf{b}^\wedge \cdot \mathbf{u}^\wedge + \gamma = f(\mathbf{u}) \ \text{ with} \\ &\mathbf{A}^\wedge = (\mathbf{A}^\wedge)^T \in \mathbb{R}^{2n^2 \times 2n^2}, \ \ \mathbf{b}^\wedge \in \mathbb{R}^{2n^2}, \ \ \gamma \in \mathbb{R}, \end{aligned} \tag{5.21}$$

whose gradient is obtained as

$$\nabla \mathbf{f}(\mathbf{u}^\wedge) = 2(\mathbf{A}^\wedge \mathbf{u}^\wedge - \mathbf{b}^\wedge) = 2(\mathbf{A}\mathbf{u} - \mathbf{b})^\wedge. \tag{5.22}$$

Summarising, we have:

Lemma 5.4 (Representation of the objective function). *The objective function* (5.17) *can be expressed according to* (5.18), (5.20) – (5.22) *in the form*

$$\begin{aligned} f(\mathbf{u}) &= \mathbf{A}\mathbf{u} \cdot \mathbf{u} - 2\mathbf{b} \cdot \mathbf{u} + \gamma \ \ \textit{with} \\ \nabla \mathbf{f}(\mathbf{u}) &= 2(\mathbf{A}\mathbf{u} - \mathbf{b}) \ \ \textit{if} \ \ \mathbb{F} = \mathbb{R}, \\ f(\mathbf{u}^\wedge) &= \mathbf{A}^\wedge \mathbf{u}^\wedge \cdot \mathbf{u}^\wedge - 2\mathbf{b}^\wedge \cdot \mathbf{u}^\wedge + \gamma \ \ \textit{with} \\ \nabla \mathbf{f}(\mathbf{u}^\wedge) &= 2(\mathbf{A}^\wedge \mathbf{u}^\wedge - \mathbf{b}^\wedge) \ \ \textit{if} \ \ \mathbb{F} = \mathbb{C}, \end{aligned}$$

where $\mathbf{u} = \mathrm{vec}(\mathbf{U})$, $\mathbf{A}^* = \mathbf{A}$, $\overline{\gamma} = \gamma$ *and* $\mathbf{A}$, $\mathbf{b}$, γ *are defined by* (5.19).

5.2.2 Transformation of the constraints

The constraints

$$\mathbf{g}(\mathbf{U}) = \mathbf{U}^* \mathbf{G} \mathbf{U} - \mathbf{G} = \mathbf{0} \tag{5.23}$$

consist of n^2 equations $\mathbf{g}(\mathbf{U}) = [\gamma_{\mu\nu}(\mathbf{U})]$, $\gamma_{\mu\nu} = \overline{\gamma}_{\nu\mu}$, which with $\mathbf{G} = [g_{\mu\nu}]$, $g_{\mu\nu} = \overline{g}_{\nu\mu}$ and $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_n]$ can be transformed into the form

$$\gamma_{\mu\nu}(\mathbf{U}) = \mathbf{G}\mathbf{u}_\nu \cdot \overline{\mathbf{u}}_\mu - g_{\mu\nu} = 0 \ \text{ for } \ 1 \le \mu, \nu \le n.$$

Considering the case $\mathbb{F} = \mathbb{C}$ first, a real representation according to Lemma 5.3 must be used again

$$\begin{aligned} \mathrm{Re}(\gamma_{\mu\nu}) &= \mathbf{G}^\wedge \mathbf{u}_\nu^\wedge \cdot \mathbf{u}_\mu^\wedge - \hat{g}_{\mu\nu} = \mathrm{Re}(\gamma_{\nu\mu}) \ \text{ with } \ \hat{g}_{\mu\nu} = \mathrm{Re}(g_{\mu\nu}), \\ \mathrm{Im}(\gamma_{\mu\nu}) &= \mathbf{G}^\vee \mathbf{u}_\nu^\wedge \cdot \mathbf{u}_\mu^\wedge - \check{g}_{\mu\nu} = -\,\mathrm{Im}(\gamma_{\nu\mu}) \ \text{ with } \ \check{g}_{\mu\nu} = \mathrm{Im}(g_{\mu\nu}). \end{aligned}$$

Because of the symmetry of the real parts and the antisymmetry of the imaginary parts, now all requirements imposed on $\mathbf{U}$ can be expressed as n^2 real

constraints

$$\begin{aligned} r_{\mu\mu} &= \tfrac{1}{2}\,\mathrm{Re}(\gamma_{\mu\nu}) \\ &= \tfrac{1}{2}(\mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}\cdot\mathbf{u}_{\mu}^{\wedge} - \hat{g}_{\mu\mu}) \;\text{ for }\; 1 \le \mu \le n, \end{aligned} \tag{5.24a}$$

$$\begin{aligned} r_{\mu\nu} &= \mathrm{Re}(\gamma_{\mu\nu}) = \mathbf{G}^{\wedge}\mathbf{u}_{\nu}^{\wedge}\cdot\mathbf{u}_{\mu}^{\wedge} - \hat{g}_{\mu\nu} \\ &= \mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}\cdot\mathbf{u}_{\nu}^{\wedge} - \hat{g}_{\nu\mu} \;\text{ for }\; 1 \le \mu < \nu \le n, \end{aligned} \tag{5.24b}$$

$$\begin{aligned} r_{\mu\nu} &= \mathrm{Im}(\gamma_{\mu\nu}) = \mathbf{G}^{\vee}\mathbf{u}_{\nu}^{\wedge}\cdot\mathbf{u}_{\mu}^{\wedge} - \check{g}_{\mu\nu} \\ &= -\mathbf{G}^{\vee}\mathbf{u}_{\mu}^{\wedge}\cdot\mathbf{u}_{\nu}^{\wedge} + \check{g}_{\nu\mu} \;\text{ for }\; 1 \le \nu < \mu \le n, \end{aligned} \tag{5.24c}$$

in which the factor 1/2 has been introduced for simplification of the further presentation. From these equations the gradients according to $\mathbf{u}_{\alpha}^{\wedge} \in \mathbb{R}^{2n}$ can be read immediately using the Kronecker symbol $\delta_{\alpha\mu}$

$$\begin{aligned} \nabla_{\alpha}\mathbf{r}_{\mu\mu} &= \mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}\delta_{\alpha\mu} \;\text{ for }\; 1 \le \mu \le n, \\ \nabla_{\alpha}\mathbf{r}_{\mu\nu} &= \mathbf{G}^{\wedge}\mathbf{u}_{\nu}^{\wedge}\delta_{\alpha\mu} + \mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}\delta_{\alpha\nu} \;\text{ for }\; 1 \le \mu < \nu \le n, \\ \nabla_{\alpha}\mathbf{r}_{\mu\nu} &= \mathbf{G}^{\vee}\mathbf{u}_{\nu}^{\wedge}\delta_{\alpha\mu} - \mathbf{G}^{\vee}\mathbf{u}_{\mu}^{\wedge}\delta_{\alpha\nu} \;\text{ for }\; 1 \le \nu < \mu \le n, \end{aligned}$$

and only need to be brought into the correct order for determining the gradients according to $\mathbf{u}^{\wedge} \in \mathbb{R}^{2n^2}$.

Considering for this purpose the matrices

$$\begin{aligned} \mathbf{R}_{\mu\mu} &= \begin{bmatrix} \mathbf{0} & \dots & \overset{\mu\text{-th column}}{\overset{\downarrow}{\mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}}} & \dots & \mathbf{0} \end{bmatrix}, \\ \mathbf{R}_{\mu\nu} &= \begin{bmatrix} \mathbf{0} & \dots & \overset{\mu\text{-th column}}{\overset{\downarrow}{\mathbf{G}^{\wedge}\mathbf{u}_{\nu}^{\wedge}}} & \dots & \overset{\nu\text{-th column}}{\overset{\downarrow}{\mathbf{G}^{\wedge}\mathbf{u}_{\mu}^{\wedge}}} & \dots & \mathbf{0} \end{bmatrix} \;\text{ for }\; \mu < \nu, \\ \mathbf{R}_{\mu\nu} &= \begin{bmatrix} \mathbf{0} & \dots & \overset{\nu\text{-th column}}{\overset{\downarrow}{-\mathbf{G}^{\vee}\mathbf{u}_{\mu}^{\wedge}}} & \dots & \overset{\mu\text{-th column}}{\overset{\downarrow}{\mathbf{G}^{\vee}\mathbf{u}_{\nu}^{\wedge}}} & \dots & \mathbf{0} \end{bmatrix} \;\text{ for }\; \nu < \mu, \end{aligned}$$

and setting $\mathbf{U} = [u_{\mu\nu}]$, $\hat{u}_{\mu\nu} = \mathrm{Re}(u_{\mu\nu})$, $\check{u}_{\mu\nu} = \mathrm{Im}(u_{\mu\nu})$, it is found that their elements are the partial derivatives of $r_{\mu\nu}$ according to $\hat{u}_{11}, \dots, \hat{u}_{nn}, \check{u}_{11}, \dots, \check{u}_{nn}$

$$\mathbf{R}_{\mu\nu} = \begin{bmatrix} \dfrac{\partial r_{\mu\nu}}{\partial \hat{u}_{11}} & \dots & \dfrac{\partial r_{\mu\nu}}{\partial \hat{u}_{1n}} \\ \vdots & & \vdots \\ \dfrac{\partial r_{\mu\nu}}{\partial \hat{u}_{n1}} & \dots & \dfrac{\partial r_{\mu\nu}}{\partial \hat{u}_{nn}} \\ \dfrac{\partial r_{\mu\nu}}{\partial \check{u}_{11}} & \dots & \dfrac{\partial r_{\mu\nu}}{\partial \check{u}_{1n}} \\ \vdots & & \vdots \\ \dfrac{\partial r_{\mu\nu}}{\partial \check{u}_{n1}} & \dots & \dfrac{\partial r_{\mu\nu}}{\partial \check{u}_{nn}} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_{\mu\nu,1} \\ \vdots \\ \mathbf{r}_{\mu\nu,n} \\ \mathbf{r}_{\mu\nu,n+1} \\ \vdots \\ \mathbf{r}_{\mu\nu,2n} \end{bmatrix} \in \mathbb{R}^{2n\times n}, \tag{5.25a}$$

and that the gradient of $r_{\mu\nu}$ according to $\mathbf{u}^{\wedge}$ is given by a corresponding arrangement of the row vectors

$$\nabla\mathbf{r}_{\mu\nu} = \mathrm{vec}(\mathbf{R}_{\mu\nu}) = \begin{pmatrix} \mathbf{r}_{\mu\nu,1}^{T} \\ \vdots \\ \mathbf{r}_{\mu\nu,2n}^{T} \end{pmatrix} \in \mathbb{R}^{2n^2}. \tag{5.25b}$$

On the other hand, we can write

$$\mathbf{R}_{\mu\mu} = \mathbf{G}^\wedge \begin{bmatrix} \mathbf{0} & \dots & \overset{(\mu)}{\mathbf{u}_\mu^\wedge} & \dots & \mathbf{0} \end{bmatrix} = \mathbf{G}^\wedge \hat{\mathbf{U}} \mathbf{J}_{\mu\mu}, \tag{5.26a}$$

$$\mathbf{R}_{\mu\nu} = \mathbf{G}^\wedge \begin{bmatrix} \mathbf{0} & \dots & \overset{(\mu)}{\mathbf{u}_\nu^\wedge} & \dots & \overset{(\nu)}{\mathbf{u}_\mu^\wedge} & \dots & \mathbf{0} \end{bmatrix} = \mathbf{G}^\wedge \hat{\mathbf{U}} \mathbf{J}_{\mu\nu} \text{ for } \mu < \nu, \tag{5.26b}$$

$$\mathbf{R}_{\mu\nu} = \mathbf{G}^\vee \begin{bmatrix} \mathbf{0} & \dots & \overset{(\nu)}{-\mathbf{u}_\mu^\wedge} & \dots & \overset{(\mu)}{\mathbf{u}_\nu^\wedge} & \dots & \mathbf{0} \end{bmatrix} = \mathbf{G}^\vee \hat{\mathbf{U}} \mathbf{K}_{\mu\nu} \text{ for } \nu < \mu \tag{5.26c}$$

by defining

$$\mathbf{J}_{\mu\nu} = \begin{bmatrix} 0 & & & 0 \\ & & 1_{\mu\nu} & \\ & 1_{\nu\mu} & & \\ 0 & & & 0 \end{bmatrix}, \quad \mathbf{K}_{\mu\nu} = \begin{bmatrix} 0 & & & 0 \\ & & 1_{\nu\mu} & \\ & -1_{\mu\nu} & & \\ 0 & & & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

and $\hat{\mathbf{U}} = \begin{bmatrix} \mathbf{u}_1^\wedge & \dots & \mathbf{u}_n^\wedge \end{bmatrix} \in \mathbb{R}^{2n\times n}$. Here it is important to note that

$$\hat{\mathbf{U}} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \neq \begin{bmatrix} \mathbf{U}_1 & -\mathbf{U}_2 \\ \mathbf{U}_2 & \mathbf{U}_1 \end{bmatrix} = \mathbf{U}^\wedge, \tag{5.27}$$

so that the position of the symbol $\wedge$ is of decisive importance. The constraints and their gradients are thus represented by the vector and the matrix

$$\begin{aligned} \mathbf{r}(\mathbf{u}^\wedge) = \mathbf{r} &= \begin{pmatrix} r_{11} \dots r_{nn} \end{pmatrix}^T \in \mathbb{R}^{n^2} \text{ and} \\ \nabla\mathbf{r}(\mathbf{u}^\wedge) = \mathbf{R} &= \begin{bmatrix} \nabla\mathbf{r}_{11} \dots \nabla\mathbf{r}_{nn} \end{bmatrix} \in \mathbb{R}^{2n^2\times n^2}, \end{aligned} \tag{5.28}$$

whose components are arranged in the order $\overbrace{(11),\dots,(1n)}^{n};\dots;\overbrace{(n1),\dots,(nn)}^{n}$.

Now introducing the vector of the Lagrange multipliers

$$\lambda = \begin{pmatrix} \lambda_{11} \dots \lambda_{nn} \end{pmatrix}^T \in \mathbb{R}^{n^2}, \tag{5.29}$$

it remains to find an easily differentiable, according to $\mathbf{u}^\wedge$, representation of $\lambda \cdot \nabla\mathbf{r}(\mathbf{u}^\wedge) = \mathbf{R}\lambda$. For this purpose we obtain from (5.26) – (5.29)

$$\begin{aligned} \sum_{\mu,\nu} \mathbf{R}_{\mu\nu}\lambda_{\mu\nu} &= \sum_{\mu\le\nu} \mathbf{R}_{\mu\nu}\lambda_{\mu\nu} + \sum_{\nu<\mu} \mathbf{R}_{\mu\nu}\lambda_{\mu\nu} = \sum_{\mu\le\nu} \mathbf{G}^\wedge\hat{\mathbf{U}}\mathbf{J}_{\mu\nu}\lambda_{\mu\nu} + \sum_{\nu<\mu} \mathbf{G}^\vee\hat{\mathbf{U}}\mathbf{K}_{\mu\nu}\lambda_{\mu\nu} \\ &= \mathbf{G}^\wedge\hat{\mathbf{U}}\big(\sum_{\mu\le\nu} \lambda_{\mu\nu}\mathbf{J}_{\mu\nu}\big) + \mathbf{G}^\vee\hat{\mathbf{U}}\big(\sum_{\nu<\mu} \lambda_{\mu\nu}\mathbf{K}_{\mu\nu}\big) = \mathbf{G}^\wedge\hat{\mathbf{U}}\mathbf{\Lambda}_1^T + \mathbf{G}^\vee\hat{\mathbf{U}}\mathbf{\Lambda}_2^T \end{aligned}$$

where

$$\mathbf{\Lambda}_1 = \mathbf{\Lambda}_1^T = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{1n} & \cdots & \lambda_{nn} \end{bmatrix}, \quad \mathbf{\Lambda}_2 = -\mathbf{\Lambda}_2^T = \begin{bmatrix} 0 & \cdots & -\lambda_{n1} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Consequently, on account of $\mathrm{vec}(\hat{\mathbf{U}}) = \mathrm{vec}(\mathbf{U})^\wedge = \mathbf{u}^\wedge$, Lemma 5.2 and Lemma 5.4 it follows that

$$\begin{aligned} \mathbf{R}\lambda = \mathrm{vec}\big(\sum_{\mu,\nu} \mathbf{R}_{\mu\nu}\lambda_{\mu\nu}\big) &= (\mathbf{G}^\wedge \otimes \mathbf{\Lambda}_1)\mathbf{u}^\wedge + (\mathbf{G}^\vee \otimes \mathbf{\Lambda}_2)\mathbf{u}^\wedge \\ &= (\mathbf{G}^\wedge \otimes \mathbf{\Lambda}_1^T - \mathbf{G}^\vee \otimes \mathbf{\Lambda}_2^T)\mathbf{u}^\wedge = (\mathbf{G} \otimes \mathbf{\Lambda}^T)^\wedge\mathbf{u}^\wedge = (\mathbf{B}\mathbf{u})^\wedge \end{aligned} \tag{5.30}$$

by defining

$$\mathbf{B} = \mathbf{G} \otimes \mathbf{\Lambda}^T \text{ and } \mathbf{\Lambda} = \mathbf{\Lambda}_1 + i\mathbf{\Lambda}_2 = \mathbf{\Lambda}^*. \tag{5.31}$$

Thus, in the case $\mathbb{F} = \mathbb{C}$ all preparations are complete for expressing the Newton method. In the case $\mathbb{F} = \mathbb{R}$ it is only necessary to note that the imaginary constraints (5.24c) are not required and that no splitting into real and imaginary parts is necessary. Hence, we obtain the following result.

Lemma 5.5 (Representation of the constraints). *The constraints* (5.23) *can be expressed according to* (5.24), (5.28) *in the form*

$$\mathbb{R}: \ \mathbf{r}(\mathbf{u}) = (r_{\mu\nu}) \in \mathbb{R}^{\frac{n(n+1)}{2}} \quad \text{with } r_{\mu\nu} = \begin{cases} \frac{1}{2}(\mathbf{G}\mathbf{u}_\mu \cdot \mathbf{u}_\mu - g_{\mu\mu}), & 1 \le \mu \le n \\ (\mathbf{G}\mathbf{u}_\nu \cdot \mathbf{u}_\mu - g_{\mu\nu}), & 1 \le \mu < \nu \le n \end{cases},$$

$$\mathbb{C}: \ \mathbf{r}(\mathbf{u}^\wedge) = (r_{\mu\nu}) \in \mathbb{R}^{n^2} \quad \text{with } r_{\mu\nu} = \begin{cases} \frac{1}{2}(\mathbf{G}^\wedge \mathbf{u}_\mu^\wedge \cdot \mathbf{u}_\mu^\wedge - \hat{g}_{\mu\mu}), & 1 \le \mu \le n \\ \mathbf{G}^\wedge \mathbf{u}_\nu^\wedge \cdot \mathbf{u}_\mu^\wedge - \hat{g}_{\mu\nu}, & 1 \le \mu < \nu \le n \\ \mathbf{G}^\vee \mathbf{u}_\nu^\wedge \cdot \mathbf{u}_\mu^\wedge - \check{g}_{\mu\nu}, & 1 \le \nu < \mu \le n \end{cases}.$$

Their gradients can be expressed according to (5.25) – (5.28) *in the form*

$$\mathbb{R}: \ \mathbf{R}(\mathbf{u}) = [\mathrm{vec}(\mathbf{R}_{\mu\nu})] \in \mathbb{R}^{n^2 \times \frac{n(n+1)}{2}} \quad \text{with } \mathbf{R}_{\mu\nu} = \begin{cases} \mathbf{G}\mathbf{U}\mathbf{J}_{\mu\nu}, & 1 \le \mu \le \nu \le n \end{cases},$$

$$\mathbb{C}: \ \mathbf{R}(\mathbf{u}^\wedge) = [\mathrm{vec}(\mathbf{R}_{\mu\nu})] \in \mathbb{R}^{2n^2 \times n^2} \quad \text{with } \mathbf{R}_{\mu\nu} = \begin{cases} \mathbf{G}^\wedge \hat{\mathbf{U}} \mathbf{J}_{\mu\nu}, & 1 \le \mu \le \nu \le n \\ \mathbf{G}^\vee \hat{\mathbf{U}} \mathbf{K}_{\mu\nu}, & 1 \le \nu < \mu \le n \end{cases}.$$

Moreover, the product of the gradients with the Lagrange multipliers, according to (5.29) – (5.31), *satisfies*

$$\mathbb{R}: \ \mathbf{R}\lambda = \mathbf{B}\mathbf{u} \qquad \text{with } \mathbf{B} = \mathbf{G} \otimes \mathbf{\Lambda}^T,\ \lambda = (\lambda_{\mu\nu}) \in \mathbb{R}^{\frac{n(n+1)}{2}},\ \mathbf{\Lambda} = \mathbf{\Lambda}_1,$$

$$\mathbb{C}: \ \mathbf{R}\lambda = (\mathbf{B}\mathbf{u})^\wedge \qquad \text{with } \mathbf{B} = \mathbf{G} \otimes \mathbf{\Lambda}^T,\ \lambda = (\lambda_{\mu\nu}) \in \mathbb{R}^{n^2},\ \mathbf{\Lambda} = \mathbf{\Lambda}_1 + i\mathbf{\Lambda}_2.$$

Here the elements of $\mathbf{r}, \lambda$ *and the columns of* $\mathbf{R}$ *are arranged in the order*

$$\mathbb{R}: \ \overbrace{(11), \dots, (1n)}^{n}; \overbrace{(22), \dots, (2n)}^{n-1}; \dots; \overbrace{(nn)}^{1} \ (= \tfrac{n(n+1)}{2} \text{ components}),$$

$$\mathbb{C}: \ \overbrace{(11), \dots, (1n)}^{n}; \dots; \overbrace{(n1), \dots, (nn)}^{n} \ (= n^2 \text{ components}).$$

5.2.3 Specification of the method

To specify the Newton method for solving the optimisation problem

$$f(\mathbf{u}) = \mathbf{A}\mathbf{u} \cdot \overline{\mathbf{u}} - 2\,\mathrm{Re}\,\mathbf{b} \cdot \overline{\mathbf{u}} + \gamma \to \min,$$

$$\mathbf{g}(\mathbf{U}) = \mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G} = \mathbf{0},$$

it remains to insert the results of Lemma 5.4 and Lemma 5.5 into (5.15).

Considering the case $\mathbb{F} = \mathbb{C}$ first, the gradient of the Lagrange function is obtained as

$$\mathbf{F}(\mathbf{u}^\wedge, \lambda) = \begin{pmatrix} \nabla \mathbf{f}(\mathbf{u}^\wedge) + \lambda \cdot \nabla \mathbf{r}(\mathbf{u}^\wedge) \\ \mathbf{r}(\mathbf{u}^\wedge) \end{pmatrix} = \begin{pmatrix} 2(\mathbf{A}^\wedge \mathbf{u}^\wedge - \mathbf{b}^\wedge) + \mathbf{R}\lambda \\ \mathbf{r} \end{pmatrix} \in \mathbb{R}^{3n^2}$$

which, after the permissible multiplication of the constraints by the factor 2 (taken into the Lagrange multipliers) becomes

$$\mathbf{F}(\mathbf{u}^\wedge, \lambda) = 2\begin{pmatrix} \mathbf{A}^\wedge\mathbf{u}^\wedge - \mathbf{b}^\wedge + \mathbf{R}\lambda \\ \mathbf{r} \end{pmatrix} = 2\begin{pmatrix} \mathbf{A}^\wedge\mathbf{u}^\wedge - \mathbf{b}^\wedge + \mathbf{B}^\wedge\mathbf{u}^\wedge \\ \mathbf{r} \end{pmatrix}.$$

The Jacobi matrix of $\mathbf{F}$ according to

$$\frac{\partial \mathbf{F}}{\partial \mathbf{u}^\wedge} = 2\begin{bmatrix} \mathbf{A}^\wedge + \mathbf{B}^\wedge \\ \mathbf{R}^T \end{bmatrix} \in \mathbb{R}^{3n^2 \times 2n^2} \quad \text{and} \quad \frac{\partial \mathbf{F}}{\partial \lambda} = 2\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3n^2 \times n^2}$$

is given by

$$\mathbf{DF}(\mathbf{u}^\wedge, \lambda) = 2\begin{bmatrix} \mathbf{A}^\wedge + \mathbf{B}^\wedge & \mathbf{R} \\ \mathbf{R}^T & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3n^2 \times 3n^2},$$

so that the corresponding method (after canceling the factor $\frac{1}{2} \cdot 2$) is

$$\begin{pmatrix} \mathbf{u}_{i+1}^\wedge \\ \lambda_{i+1} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_i^\wedge \\ \lambda_i \end{pmatrix} - \begin{bmatrix} \mathbf{A}^\wedge + \mathbf{B}_i^\wedge & \mathbf{R}_i \\ \mathbf{R}_i^T & \mathbf{0} \end{bmatrix}^{-1} \begin{pmatrix} (\mathbf{A}^\wedge + \mathbf{B}_i^\wedge)\mathbf{u}_i^\wedge - \mathbf{b}^\wedge \\ \mathbf{r}_i \end{pmatrix}. \tag{5.32a}$$

If the constraint $\mathbf{h}(\mathbf{U}) = \mathbf{U}^*\mathbf{H}\mathbf{U} - \mathbf{H} = \mathbf{0}$ is to be fulfilled instead of $\mathbf{g}(\mathbf{U}) = \mathbf{0}$, it is only necessary, when calculating the restrictions $\mathbf{r}$ and their gradients $\mathbf{R}$, to use the matrix $\mathbf{H}$ instead of the matrix $\mathbf{G}$. But if both constraints are to be fulfilled, and if

$$\omega, \;\; \mathbf{\Omega}, \;\; \mathbf{s}, \;\; \mathbf{S}, \;\; \mathbf{C} = \mathbf{H} \otimes \mathbf{\Omega}^T$$

are the Lagrange multipliers, restrictions, restriction gradients and matrices belonging to $\mathbf{h}(\mathbf{U})$ and

$$\lambda, \;\; \mathbf{\Lambda}, \;\; \mathbf{r}, \;\; \mathbf{R}, \;\; \mathbf{B} = \mathbf{G} \otimes \mathbf{\Lambda}^T$$

are those belonging to $\mathbf{g}(\mathbf{U})$, then the method

$$\begin{pmatrix} \mathbf{u}_{i+1}^\wedge \\ \lambda_{i+1} \\ \omega_{i+1} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_i^\wedge \\ \lambda_i \\ \omega_i \end{pmatrix} - \begin{bmatrix} \mathbf{A}^\wedge + \mathbf{B}_i^\wedge + \mathbf{C}_i^\wedge & \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{R}_i^T & \mathbf{0} & \mathbf{0} \\ \mathbf{S}_i^T & \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1} \begin{pmatrix} (\mathbf{A}^\wedge + \mathbf{B}_i^\wedge + \mathbf{C}_i^\wedge)\mathbf{u}_i^\wedge - \mathbf{b}^\wedge \\ \mathbf{r}_i \\ \mathbf{s}_i \end{pmatrix} \tag{5.32b}$$

must be used, whose vectors now contain $4n^2$ components. Here cases can arise in which the Jacobi matrix $\mathbf{DF}$ becomes singular which is possible when the gradients of the constraints $\mathbf{R}$ and $\mathbf{S}$ are linearly dependent. Instead of the Newton iteration

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{DF}^{-1}(\mathbf{z}_i)\,\mathbf{F}(\mathbf{z}_i) \tag{5.33a}$$

it may then be possible to use the Gauß-Newton iteration

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{DF}^{+}(\mathbf{z}_i)\,\mathbf{F}(\mathbf{z}_i), \tag{5.33b}$$

where $\mathbf{DF}^+$ denotes the pseudo inverse (Moore-Penrose inverse) of the Jacobi matrix [ST1, Kapitel 4.8.5]. We will discuss this point in the next subsection.

Assuming suitable starting values $\mathbf{U}_0$, $\mathbf{\Lambda}_0$, $\mathbf{\Omega}_0$, the iteration converges to a vector $\mathbf{z}_m$,

$$\|\mathbf{F}(\mathbf{z}_m)\| < \varepsilon \;\text{ with }\; \|\mathbf{z}\| = \sqrt{\mathbf{z}^T\mathbf{z}}, \tag{5.34}$$

from which the components of the matrices $\mathbf{U} = \mathbf{U}_m$, $\mathbf{\Lambda} = \mathbf{\Lambda}_m$, $\mathbf{\Omega} = \mathbf{\Omega}_m$ can be read according to

$$\mathbf{z} = (\mathbf{u}^\wedge\ \lambda)^T = (\hat{u}_{11} \dots \hat{u}_{nn}\ \check{u}_{11} \dots \check{u}_{nn}\ \lambda_{11} \dots \lambda_{nn})^T \text{ or} \tag{5.35a}$$

$$\mathbf{z} = (\mathbf{u}^\wedge\ \lambda\ \ \omega)^T = (\hat{u}_{11} \dots \hat{u}_{nn}\ \check{u}_{11} \dots \check{u}_{nn}\ \lambda_{11} \dots \lambda_{nn}\ \omega_{11}\ \dots\ \omega_{nn})^T. \tag{5.35b}$$

It is now still necessary to ensure that the sufficient second order condition for a minimum holds which is achieved as follows: Let

$$\mathbf{DF}(\mathbf{z}_m) = \begin{bmatrix} \mathbf{L} & \mathbf{M} \\ \mathbf{M} & \mathbf{0} \end{bmatrix}$$

be a partitioning of the Jacobi matrix obtained in the last iteration step, where $\mathbf{L} \in \mathbb{R}^{p\times p}$, $\mathbf{M} \in \mathbb{R}^{p\times q}$ and $p = 2n^2$, $q = n^2$ or $p = q = 2n^2$, respectively. Then $\mathbf{L}$ is the symmetric Lagrange matrix denoted by $\mathbf{L}(\mathbf{u}_0, \lambda_0)$ in (5.14), and the columns of $\mathbf{M}$ form a basis or possibly spanning set of the subspace M defined in (5.12). Now a basis of $M^\perp$ is required which can be derived from a QR factorisation with column pivoting

$$\mathbf{MP} = \mathbf{QR},$$

where $\mathbf{Q} \in \mathbb{R}^{p\times p}$ is orthogonal, $\mathbf{P} \in \mathbb{R}^{q\times q}$ is a permutation and $\mathbf{R} \in \mathbb{R}^{p\times q}$ is trapezoidal [GVL, Chapter 5.4]. If $r = \operatorname{rank}(\mathbf{M})$, then $\mathbf{Q}$ and $\mathbf{R}$ can be partitioned as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{R}_{11} \in \mathbb{R}^{r\times r}$ is upper triangular and nonsingular and the columns of $\mathbf{Q}_2 \in \mathbb{R}^{p\times(p-r)}$ form an orthogonal basis of $M^\perp = \operatorname{Im}(\mathbf{M})^\perp$. Hence, the sufficient condition (5.14) holds, if and only if the symmetric matrix

$$\mathbf{K} = \mathbf{Q}_2^T \mathbf{L} \mathbf{Q}_2 \in \mathbb{R}^{(p-r)\times(p-r)} \tag{5.36}$$

is positive definite. This can be verified with an eigenvalue computation.

In the case $\mathbb{F} = \mathbb{R}$ the equations stated above can be taken over by eliminating the operator $\wedge$ from them. The arrays in the real version of (5.32a) obtained this way are of order

$$n^2 + \frac{n(n+1)}{2} = \frac{3n^2 + n}{2} \tag{5.37a}$$

and those in the real version of (5.32b) are of order

$$n^2 + 2\frac{n(n+1)}{2} = 2n^2 + n. \tag{5.37b}$$

These dimensions are fairly smaller than the corresponding values of $3n^2$ or $4n^2$ in the case $\mathbb{F} = \mathbb{C}$, so that a real problem should not be solved with the complex method. The results obtained so far are summarised in the following statement.

Method 5.6 (Newton method). *Let $f(\mathbf{U})$ be an objective function which can be expressed in the form*

$$f(\mathbf{u}) = \mathbf{A}\mathbf{u} \cdot \overline{\mathbf{u}} - 2\operatorname{Re}\mathbf{b} \cdot \overline{\mathbf{u}} + \gamma,$$

where $\mathbf{u} = \mathrm{vec}(\mathbf{U})$, $\mathbf{A}^* = \mathbf{A}$ *and* $\overline{\gamma} = \gamma$. *Then the constrained optimisation problem*

$$f(\mathbf{u}) \to \min \quad \textit{with} \quad \mathbf{U}^*\mathbf{G}\mathbf{U} = \mathbf{G} \quad \textit{and/or} \quad \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H}$$

can be solved numerically using the Newton method (5.33a) *or possibly the Gauß-Newton method* (5.33b), *where* $\mathbf{z}$, $\mathbf{F}$ *and* $\mathbf{DF}$ *are determined by* (5.32) *(in the case* $\mathbb{F} = \mathbb{R}$ *without the* $\wedge$ *operator). If the iteration converges and the matrix* $\mathbf{K}$ *defined by* (5.36) *is positive definite, then the matrix* $\mathbf{U}$ *determined by* (5.35) *solves the considered problem.*

If $\mathbf{A}$, $\mathbf{b}$ and γ are chosen according to (5.19), the method is suitable for minimising (5.17). But if we are interested in minimising

$$f(\mathbf{U}) = \mathrm{tr}[(\mathbf{U}\mathbf{X} - \mathbf{Y})^*\mathbf{H}_f(\mathbf{U}\mathbf{X} - \mathbf{Y})] \tag{5.38a}$$

subject to

$$\mathbf{U}^*\mathbf{G}\mathbf{U} = \mathbf{G} \ \text{ and/or } \ \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H}, \tag{5.38b}$$

where the matrix $\mathbf{H}_f$ contained in the objective may not be equal to the matrix $\mathbf{H}$ contained in the constraints, we must also use $\mathbf{H}_f$ in (5.19). Then the upper left block of $\mathbf{DF}$ and the upper part of $\mathbf{F}$ are given by

$$\begin{aligned}\mathbf{A} + \mathbf{B} &= \mathbf{H}_f \otimes (\mathbf{X}\mathbf{X}^*)^T + \mathbf{G} \otimes \mathbf{\Lambda}^T,\\ (\mathbf{A} + \mathbf{B})\mathbf{u} - \mathbf{b} &= \mathrm{vec}(\mathbf{H}_f\mathbf{U}\mathbf{X}\mathbf{X}^* + \mathbf{G}\mathbf{U}\mathbf{\Lambda} - \mathbf{H}_f\mathbf{Y}\mathbf{X}^*),\end{aligned}$$

$$\begin{aligned}\mathbf{A} + \mathbf{B} + \mathbf{C} &= \mathbf{H}_f \otimes (\mathbf{X}\mathbf{X}^*)^T + \mathbf{G} \otimes \mathbf{\Lambda}^T + \mathbf{H} \otimes \mathbf{\Omega}^T,\\ (\mathbf{A} + \mathbf{B} + \mathbf{C})\mathbf{u} - \mathbf{b} &= \mathrm{vec}(\mathbf{H}_f\mathbf{U}\mathbf{X}\mathbf{X}^* + \mathbf{G}\mathbf{U}\mathbf{\Lambda} + \mathbf{H}\mathbf{U}\mathbf{\Omega} - \mathbf{H}_f\mathbf{Y}\mathbf{X}^*).\end{aligned}$$

Hence, $\mathbf{F}(\mathbf{z}) = \mathbf{0}$ implies

$$\mathbf{H}_f\mathbf{U}\mathbf{X}\mathbf{X}^* + \mathbf{G}\mathbf{U}\mathbf{\Lambda} = \mathbf{H}_f\mathbf{Y}\mathbf{X}^*, \tag{5.39a}$$

$$\mathbf{H}_f\mathbf{U}\mathbf{X}\mathbf{X}^* + \mathbf{G}\mathbf{U}\mathbf{\Lambda} + \mathbf{H}\mathbf{U}\mathbf{\Omega} = \mathbf{H}_f\mathbf{Y}\mathbf{X}^* \tag{5.39b}$$

which with $\mathbf{\Lambda} = \mathbf{\Lambda}_G$, $\mathbf{\Omega} = \mathbf{\Lambda}_H$ and $\mathbf{H}_f = \mathbf{H}$ is just (5.7) or (5.5), respectively. This confirms that the isometry $\mathbf{U}$ computed with the iteration satisfies the necessary conditions for solving the problem (5.1) subject to the constraints (5.6) or (5.3), respectively. Moreover, if $\mathbf{\Upsilon} \in \mathbb{F}^{n\times n}$ is an arbitrary matrix and we are defining

$$\mathbf{A} = \mathbf{H} \otimes \mathbf{I} \ \text{ and } \ \mathbf{b} = \mathrm{vec}(\mathbf{H}\mathbf{\Upsilon}\mathbf{H}^{-1}),$$

then (5.39a) with $\mathbf{H}_f = \mathbf{G} = \mathbf{H}$ becomes $\mathbf{H}\mathbf{U}(\mathbf{I} + \mathbf{\Lambda}) = \mathbf{H}\mathbf{\Upsilon}\mathbf{H}^{-1}$ or

$$\mathbf{U}\mathbf{M} = \mathbf{\Upsilon}. \tag{5.40}$$

Here $\mathbf{M} = (\mathbf{I} + \mathbf{\Lambda})\mathbf{H}$ satisfies $\mathbf{M}^*\mathbf{H} = \mathbf{H}\mathbf{M}$, so that Method 5.6 also applies for computing H-polar decompositions.

5.2.4 Application of the method

The iteration rules (5.32) need some refinements before they fit for the application. We start with determining the vector $\mathbf{dz} = \mathbf{DF}^{-1}\mathbf{F}$, i.e. with solving

the linear system $\mathbf{DF}\,\mathbf{dz} = \mathbf{F}$, which is necessary in each iteration step. This system has the general form

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix}, \tag{5.41}$$

where $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{u}, \mathbf{c} \in \mathbb{R}^p$ and $\mathbf{v}, \mathbf{d} \in \mathbb{R}^q$. It may be rewritten as

$$\left(\begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^T \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right) \left(\begin{bmatrix} \mathbf{I} & \mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \right) = \begin{bmatrix} \mathbf{I} & -\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^T \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix},$$

from which with $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B} \in \mathbb{R}^{p\times q}$ it follows that

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{bmatrix} \begin{pmatrix} \mathbf{u} + \mathbf{A}^{-1}\mathbf{B}\mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{c} \end{pmatrix}.$$

This suggests the following algorithm for solving (5.41).

Algorithm 5.7.

(1) *Solve* $\mathbf{AX} = \mathbf{B}$ *for* $\mathbf{X}$.

(2) *Solve* $(\mathbf{X}^T\mathbf{B})\mathbf{v} = \mathbf{X}^T\mathbf{c} - \mathbf{d}$ *for* $\mathbf{v}$.

(3) *Solve* $\mathbf{Au} = \mathbf{c} - \mathbf{Bv}$ *for* $\mathbf{u}$.

In the case $\mathbb{F} = \mathbb{C}$ the matrix $\mathbf{A}$ and the vectors $\mathbf{u}$, $\mathbf{c}$ are real representations of complex arrays and we shall exploit the further structure appearing in the system

$$\begin{bmatrix} \mathbf{A}^\wedge & \hat{\mathbf{B}} \\ \hat{\mathbf{B}}^T & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{u}^\wedge \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{c}^\wedge \\ \mathbf{d} \end{pmatrix}, \tag{5.42}$$

where $\mathbf{A} = \mathbf{A}^* \in \mathbb{C}^{p\times p}$, $\mathbf{B} \in \mathbb{C}^{p\times q}$, $\mathbf{u}, \mathbf{c} \in \mathbb{C}^p$ and $\mathbf{v}, \mathbf{d} \in \mathbb{R}^q$. Here the hat over the $\mathbf{B}$ has the same meaning as in (5.27). Defining $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$ and using $\hat{\mathbf{X}}$ in the same sense, it can be verified that Algorithm 5.7 also applies for solving (5.42). The only modification required is that the real system

$$\operatorname{Re}(\mathbf{X}^*\mathbf{B})\mathbf{v} = \operatorname{Re}(\mathbf{X}^*\mathbf{c}) - \mathbf{d} \tag{5.43a}$$

has to be used in step (2).

A further modification is necessary in the case that the gradients in iteration (5.32b) are linearly dependent. Then the system (5.43a) is overdetermined, so that $\mathbf{v}$ must be computed as a least squares solution of the problem

$$\| \operatorname{Re}(\mathbf{X}^*\mathbf{B})\mathbf{v} - \operatorname{Re}(\mathbf{X}^*\mathbf{c}) + \mathbf{d}\| \to \min. \tag{5.43b}$$

However, it is not ensured that the iteration then always converges to the wanted result.

Now considering the case $\mathbf{A} = \mathbf{G} \otimes \mathbf{C} + \mathbf{H} \otimes \mathbf{D}$ for some selfadjoint matrices $\mathbf{C}, \mathbf{D}, \mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$, we find that $\mathbf{A}$ is a general selfadjoint matrix, so that no further simplification of Algorithm 5.7 is possible. But if $\mathbf{G}$ and $\mathbf{H}$ are diagonal matrices

$$\mathbf{G} = \operatorname{diag}(g_1, \ldots, g_n), \quad \mathbf{H} = \operatorname{diag}(h_1, \ldots, h_n),$$

Table 5.1: Flop counts for some operations

Operation		Flop Count
$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{y}$,	$\mathbf{A} \in \mathbb{R}^{m\times n}$	$2mn$
$\mathbf{C} = \mathbf{A}\mathbf{B} + \mathbf{C}$,	$\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times n}$, $\mathbf{C} \in \mathbb{R}^{m\times n}$	$2mnp$
$\mathbf{P}\mathbf{A}\mathbf{P}^T = \mathbf{L}\mathbf{D}\mathbf{L}^T$,	$\mathbf{A} \in \mathbb{R}^{n\times n}$	$n^3/3$
$\mathbf{L}\mathbf{D}\mathbf{L}^T\mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{b}$,	$\mathbf{b} \in \mathbb{R}^n$	$2n^2$

then $\mathbf{A}$ is a block diagonal matrix

$$\mathbf{A} = \mathbf{A}_1 \oplus \ldots \oplus \mathbf{A}_n \quad \text{with} \quad \mathbf{A}_\nu = g_\nu \mathbf{C} + h_\nu \mathbf{D} \in \mathbb{F}^{n\times n},$$

so that with

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_n \end{bmatrix}, \quad \mathbf{B}_\nu \in \mathbb{F}^{n\times q} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_n \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix}, \quad \mathbf{c}_\nu, \mathbf{u}_\nu \in \mathbb{F}^n$$

the following algorithm for solving (5.41) and (5.42) can be formulated.

Algorithm 5.8.

(1) *For* $\nu = 1, \ldots, n$*: Solve* $\mathbf{A}_\nu \mathbf{X}_\nu = \mathbf{B}_\nu$ *for* $\mathbf{X}_\nu$.

(2) *Solve* $\big(\sum_\nu \mathrm{Re}(\mathbf{X}_\nu^* \mathbf{B}_\nu)\big)\mathbf{v} = \big(\sum_\nu \mathrm{Re}(\mathbf{X}_\nu^* \mathbf{c}_\nu)\big) - \mathbf{d}$ *for* $\mathbf{v}$.

(3) *For* $\nu = 1, \ldots, n$*: Solve* $\mathbf{A}_\nu \mathbf{u}_\nu = \mathbf{c}_\nu - \mathbf{B}_\nu \mathbf{v}$ *for* $\mathbf{u}_\nu$.

In order to estimate the amount of work required for the algorithms it is assumed that all the symmetric (or Hermitian) linear systems are solved with the LDL^* decomposition [GVL, Chapter 4.4] and that the decompositions for $\mathbf{A}$ or $\mathbf{A}_\nu$, respectively, are computed only once. Counting flops in the case $\mathbb{F} = \mathbb{R}$ based on Table 5.1 (see [GVL]) yields

$$\begin{aligned} f_{R,1} &= p^3/3 + 2p^2 q + 2p^2 + 4pq + 2q^2 + pq^2 + q^3/3 \qquad \text{for Algorithm 5.7,} \\ f_{R,2} &= 7n^3/3 + 2n^3 q + n^2 q^2 + 4n^2 q + 2q^2 + q^3/3 \qquad \text{for Algorithm 5.8.} \end{aligned}$$

Here the counts for computing $\mathbf{X}^T\mathbf{B}$ or $\sum_\nu \mathbf{X}_\nu^T \mathbf{B}_\nu$ are calculated by $2pq^2/2$ or $2n^2q^2/2$, respectively, because the products are known to be symmetric. The direct solution of (5.41) with the LDL^* decomposition requires

$$f_{R,0} = (p+q)^3/3 + 2(p+q)$$

flops, so that with $p = n^2$ and $q = n(n+1)/2$, according to (5.37a), it follows that

$$f_{R,0} \approx 9n^6/8, \quad f_{R,1} \approx 13n^6/8, \quad f_{R,2} \approx 7n^6/24. \tag{5.44a}$$

Setting $p = n^2$ and $q = n(n+1)$, according to (5.37b), results in

$$f_{R,0} \approx 8n^6/3, \quad f_{R,1} \approx 11n^6/3, \quad f_{R,2} \approx 4n^6/3. \tag{5.44b}$$

In other words, the amount of work for solving (5.41) with Algorithm 5.7 is by a factor $13/9 \approx 1.44$ or $11/8 \approx 1.38$, respectively, larger than directly computing

the solution. But if $\mathbf{A}$ is a block diagonal matrix, then the application of Algorithm 5.8 reduces the amount of work by a factor $7/27 \approx 0.26$ or $1/2$, respectively.

In the case $\mathbb{F} = \mathbb{C}$ an addition requires 2 and a multiplication requires 6 flops, so that an average factor of 4 must be taken into account for the decomposition of $\mathbf{A}$ and the solution of the systems in step (1) and (3) of Algorithm 5.7. The number of flops for computing $\mathrm{Re}(\mathbf{X}^*\mathbf{B})$ is estimated by $\frac{4}{2\cdot 2} \cdot 2pq^2$, because the product is Hermitian and only its real part is required. The estimation for $\mathrm{Re}(\mathbf{X}^*\mathbf{c})$ or $\mathbf{c} - \mathbf{B}\mathbf{v}$ is $\frac{4}{2} \cdot 2pq$ in each case, because only the real part is required or $\mathbf{v}$ is real, respectively. Using analogously considerations for Algorithm 5.8 the flop counts in the case $\mathbb{F} = \mathbb{C}$ are obtained as

$$\begin{aligned} f_{C,1} &= 4p^3/3 + 8p^2q + 8p^2 + 8pq + 2q^2 + 2pq^2 + q^3/3 \quad \text{for Algorithm 5.7,} \\ f_{C,2} &= 28n^3/3 + 8n^3q + 2n^2q^2 + 8n^2q + 2q^2 + q^3/3 \qquad \text{for Algorithm 5.8.} \end{aligned}$$

The direct solution of (5.42) with the LDL^* decomposition requires

$$f_{C,0} = (2p+q)^3/3 + 2(2p+q)$$

flops, so that with $p = q = n^2$, according to (5.32a), it follows that

$$f_{C,0} \approx 9n^6, \quad f_{C,1} \approx 35n^6/3, \quad f_{C,2} \approx 7n^6/3. \tag{5.45a}$$

Setting $p = n^2$ and $q = 2n^2$, according to (5.32b), results in

$$f_{C,0} \approx 64n^6/3, \quad f_{C,1} \approx 28n^6, \quad f_{C,2} \approx 32n^6/3. \tag{5.45b}$$

Thus in the complex case we obtain the factors $35/27 \approx 1.30$ or $21/16 \approx 1.31$ for Algorithm 5.7 and $7/29 \approx 0.26$ or $1/2$ for Algorithm 5.8. This corresponds to the real case discussed above. Moreover, the effort for solving (5.42) is approximately eight times larger than for solving (5.41) with equal dimensions p and q.

These flop counts suggest to use Algorithm 5.8 instead of a direct solution when $\mathbf{A}$ is block diagonal, but not to use Algorithm 5.7 instead of a direct solution when $\mathbf{A}$ is a general selfadjoint matrix. However, flop counts do not involve the costs for pivoting and subscripting which are smaller when using the algorithms. Moreover, if the matrix $\mathbf{B}$ does not have full rank and (5.43b) is to be applied, then the direct solution requires a singular value decomposition of the whole matrix contained in (5.41) or (5.42), whereas an application of the algorithms allows to compute only a singular value decomposition of $\mathbf{B}$. Thus it may be worth to consider Algorithm 5.7 for solving a corresponding linear system, too.

We are now able to perform an individual iteration step using

$$\mathbf{DF}_i\, \mathbf{dz}_i = \mathbf{F}_i \;\text{ and }\; \mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{dz}_i.$$

A more satisfactory convergence behaviour can be achieved with the line search method

$$\begin{aligned} \mathbf{z}_{i+1} &= \mathbf{z}_i - \omega_i\, \mathbf{dz}_i, \;\; \omega_i = 2^{-r} \;\text{ with} \\ r &= \min\{s \geq 0 : \|\mathbf{F}(\mathbf{z}_i - 2^{-s}\mathbf{dz}_i)\| < \|\mathbf{F}(\mathbf{z}_i)\|\}, \end{aligned} \tag{5.46}$$

whose properties are discussed in [ST1, Kapitel 5.4.2]. This iteration rule does not only improve convergence. It furthermore allows to terminate the iteration

in divergent cases when after a specified maximum number of bisection steps $s_{\max}$ no reduction of $\|\mathbf{F}(\mathbf{z}_i)\|$ is obtained. The iteration also fails when after a specified maximum number of iteration steps $i_{\max}$ the convergence criterion (5.34) does not hold.

5.2.5 Specification of the starting values

It remains to specify the starting values for the iteration. Since the objective (5.38a) with $\mathbf{H}_f = \mathbf{I}$ also appears in the orthogonal or unitary Procrustes problem

$$\operatorname{tr}[(\mathbf{UX}-\mathbf{Y})^*(\mathbf{UX}-\mathbf{Y})] \to \min \quad \text{with} \quad \mathbf{U}^*\mathbf{U} = \mathbf{I} \tag{5.47}$$

and in the unconstrained least squares problem

$$\operatorname{tr}[(\mathbf{UX}-\mathbf{Y})^*(\mathbf{UX}-\mathbf{Y})] \to \min, \tag{5.48}$$

we will make an attempt to derive $\mathbf{U}_0$, $\mathbf{\Lambda}_0$, $\mathbf{\Omega}_0$ from the solutions of these problems. Moreover, we will focus on the case that $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ are diagonal matrices which is required for an implementation of the Newton method based on Algorithm 5.8.

The orthogonal or unitary Procrustes problem (5.47) is solved by the isometry $\mathbf{U}_0$ contained in the ordinary polar decomposition

$$\mathbf{YX}^* = \mathbf{U}_0\mathbf{M}_0,$$

where $\mathbf{M}_0$ is positive semidefinite and selfadjoint (see Section 1.2 and Theorem 4.14). Inserting this result into the necessary condition (5.39) implies

$$\mathbf{GU}_0\mathbf{\Lambda}_0 = \mathbf{H}_f\mathbf{U}_0(\mathbf{M}_0 - \mathbf{XX}^*) \text{ or} \tag{5.49a}$$

$$\mathbf{GU}_0\mathbf{\Lambda}_0 + \mathbf{HU}_0\mathbf{\Omega}_0 = \mathbf{H}_f\mathbf{U}_0(\mathbf{M}_0 - \mathbf{XX}^*). \tag{5.49b}$$

Hence, if $\mathbf{G} = \varepsilon_G\mathbf{I}$, $\mathbf{H} = \varepsilon_H\mathbf{I}$, and $\mathbf{H}_f = \varepsilon_f\mathbf{I}$ with ε_G, ε_H, $\varepsilon_f \in \{+1,-1\}$, then $\mathbf{U}_0$ on the one hand fulfills (5.38b), and on the other hand

$$\mathbf{\Lambda}_0 = \varepsilon_G\varepsilon_f(\mathbf{M}_0 - \mathbf{XX}^*) \text{ or}$$

$$\mathbf{\Lambda}_0 = \frac{\varepsilon_G\varepsilon_f}{2}(\mathbf{M}_0 - \mathbf{XX}^*) \text{ and } \mathbf{\Omega}_0 = \frac{\varepsilon_H\varepsilon_f}{2}(\mathbf{M}_0 - \mathbf{XX}^*)$$

are selfadjoint Lagrange multipliers such that (5.49) holds. Therefore, $\mathbf{U}_0$ is an optimiser of (5.38a).

This observation suggests to determine the starting values in the case that $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ are diagonal matrices having diagonal elements in $\{+1,-1\}$ as follows: Let $\mathbf{P}$ be a permutation matrix such that

$$\mathbf{P}^*\mathbf{GP} = \bigoplus_{i=1}^{k} \varepsilon_G^{(i)}\mathbf{I}_{p_i}, \quad \mathbf{P}^*\mathbf{HP} = \bigoplus_{i=1}^{k} \varepsilon_H^{(i)}\mathbf{I}_{p_i}, \quad \mathbf{P}^*\mathbf{H}_f\mathbf{P} = \bigoplus_{i=1}^{k} \varepsilon_f^{(i)}\mathbf{I}_{p_i}, \tag{5.50}$$

where either $(\varepsilon_G^{(i)}, \varepsilon_f^{(i)})$ runs over all $k = 4$ or $(\varepsilon_G^{(i)}, \varepsilon_H^{(i)}, \varepsilon_f^{(i)})$ runs over all $k = 8$ combinations of signs. Now partitioning the matrices $\mathbf{P}^*\mathbf{YX}^*\mathbf{P}$ and $\mathbf{P}^*\mathbf{XX}^*\mathbf{P}$ accordingly and computing from their diagonal blocks

$$(\mathbf{P}^*\mathbf{YX}^*\mathbf{P})_i = \mathbf{U}_0^{(i)}\mathbf{M}_0^{(i)} \text{ and } \mathbf{\Lambda}_0^{(i)} = \varepsilon_G^{(i)}\varepsilon_f^{(i)}\big(\mathbf{M}_0^{(i)} - (\mathbf{P}^*\mathbf{XX}^*\mathbf{P})_i\big)$$

(analogously in the case of two constraints)

we obtain

$$\mathbf{U}_{OP} = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{U}_0^{(i)}\Big)\mathbf{P}^*, \;\; \mathbf{\Lambda}_{OP} = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{\Lambda}_0^{(i)}\Big)\mathbf{P}^*, \;\; \mathbf{\Omega}_{OP} = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{\Omega}_0^{(i)}\Big)\mathbf{P}^*. \tag{5.51}$$

If $\mathbf{P}^*\mathbf{Y}\mathbf{X}^*\mathbf{P}$ and $\mathbf{P}^*\mathbf{X}\mathbf{X}^*\mathbf{P}$ are actually block diagonal matrices consisting of the blocks used in the computation above, then $\mathbf{U}_{OP}$ is a solution of (5.38) and $\mathbf{\Lambda}_{OP}$, $\mathbf{\Omega}_{OP}$ satisfy (5.49). Although this situation almost never occurs in practice, these matrices will be used as starting values for the iteration. Moreover, if $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ are general nonsingular real diagonal matrices, the matrix $\mathbf{P}$ is computed by using sign($\mathbf{G}$), sign($\mathbf{H}$) and sign($\mathbf{H}_f$) in (5.50).

The second approach for specifying the starting values is derived from the solution of the unconstrained least squares problem (5.48) which is given by (for example see [GVL, Section 5.5.4], [ST1, Kapitel 4.8.5])

$$\mathbf{U}_0 = \mathbf{Y}\mathbf{X}^+.$$

If this choice is made, then

$$\mathbf{U}_0\mathbf{X}\mathbf{X}^* = \mathbf{Y}\mathbf{X}^+\mathbf{X}\mathbf{X}^* = \mathbf{Y}(\mathbf{X}^+\mathbf{X})^*\mathbf{X}^* = \mathbf{Y}(\mathbf{X}\mathbf{X}^+\mathbf{X})^* = \mathbf{Y}\mathbf{X}^*$$

which implies

$$\mathbf{H}_f\mathbf{Y}\mathbf{X}^* - \mathbf{H}_f\mathbf{U}_0\mathbf{X}\mathbf{X}^* = \mathbf{0}.$$

Consequently, the necessary condition (5.39) holds for every matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ by setting $\mathbf{\Lambda}_0 = \mathbf{0}$ and $\mathbf{\Omega}_0 = \mathbf{0}$. However, $\mathbf{U}_0$ is in general not a solution of (5.38) because it does not satisfy (5.38b). Nevertheless, we will try to use the starting values

$$\mathbf{U}_{LS} = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{U}_0^{(i)}\Big)\mathbf{P}^*, \;\; \mathbf{\Lambda}_{LS} = \mathbf{0}, \;\; \mathbf{\Omega}_{LS} = \mathbf{0}, \tag{5.52}$$

where

$$\mathbf{U}_0^{(i)} = (\mathbf{P}^*\mathbf{Y}\mathbf{X}^*\mathbf{P})_i(\mathbf{P}^*\mathbf{X}\mathbf{X}^*\mathbf{P})_i^+$$

and $\mathbf{P}$ is defined as above.

Finally, we will also make an attempt to use

$$\mathbf{U}_0 = \mathbf{I}.$$

If in this case $\mathbf{G} = \varepsilon_G\mathbf{I}$, $\mathbf{H} = \varepsilon_H\mathbf{I}$, and $\mathbf{H}_f = \varepsilon_f\mathbf{I}$, then (5.39) is solved for

$$\mathbf{\Lambda}_0 = \varepsilon_G\varepsilon_f(\mathbf{Y}\mathbf{X}^* - \mathbf{X}\mathbf{X}^*) \;\text{ or}$$

$$\mathbf{\Lambda}_0 = \frac{\varepsilon_G\varepsilon_f}{2}(\mathbf{Y}\mathbf{X}^* - \mathbf{X}\mathbf{X}^*) \;\text{ and }\; \mathbf{\Omega}_0 = \frac{\varepsilon_H\varepsilon_f}{2}(\mathbf{Y}\mathbf{X}^* - \mathbf{X}\mathbf{X}^*).$$

But these matrices are not selfadjoint in general, so that their symmetric or Hermitian parts may be used instead. Hence, we will try to use

$$\mathbf{U}_I = \mathbf{I}, \;\; \mathbf{\Lambda}_I = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{\Lambda}_0^{(i)}\Big)\mathbf{P}^*, \;\; \mathbf{\Omega}_I = \mathbf{P}\Big(\bigoplus_{i=1}^{k} \mathbf{\Omega}_0^{(i)}\Big)\mathbf{P}^*, \tag{5.53}$$

where

$$\mathbf{\Lambda}_0^{(i)} = \varepsilon_G^{(i)}\varepsilon_f^{(i)}\big(\frac{(\mathbf{P}^*\mathbf{Y}\mathbf{X}^*\mathbf{P})_i + (\mathbf{P}^*\mathbf{X}\mathbf{Y}^*\mathbf{P})_i}{2} - (\mathbf{P}^*\mathbf{X}\mathbf{X}^*\mathbf{P})_i\big)$$

(analogously in the case of two constraints)

and $\mathbf{P}$ is defined as above.

5.3 Numerical results

In order to test the performance and convergence behaviour we implemented Method 5.6 for solving (5.38) with diagonal matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ in the case $\mathbb{F} = \mathbb{C}$. The implementation performs the iteration rule (5.46) and uses Algorithm 5.8 for computing the vector $\mathbf{dz}$. It was made in Fortran 77 using the `DOUBLE COMPLEX` versions of `LAPACK` and the `BLAS` [LUG].

Although we have applied the algorithm for various combinations of matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$, we will present detailed results only for the most important case where

$$\mathbf{G} = \mathbf{I}_{n-p} \oplus -\mathbf{I}_p \text{ and } \mathbf{H}_f = \mathbf{I}_n$$

and a G-isometry $\mathbf{U}$ is to be determined. In the tests the matrices $\mathbf{X} = [x_{\alpha,k}]$, $\mathbf{Y} = [y_{\alpha,k}] \in \mathbb{C}^{n\times N}$ were initialised either with independent coordinates

$$x_{\alpha,k} = \theta_a + i\theta_b, \quad y_{\alpha,k} = \theta_c + i\theta_d$$

or with dependent coordinates

$$x_{\alpha,k} = \theta_a + i\theta_b, \quad y_{\alpha,k} = x_{\alpha,k} + \tfrac{1}{2}(\theta_c + i\theta_d),$$

where the θ's denote normally distributed random numbers from the interval $[-1, +1]$. Independent coordinates were additionally translated such that

$$\textstyle\sum_k x_{\alpha,k} = 0 \text{ and } \sum_k y_{\alpha,k} = 0.$$

The machine accuracy computed with the `LAPACK` routine `DLAMCH` was

$$\varepsilon_{mach} \approx 2.22 \cdot 10^{-16}.$$

In a first series of statistical experiments we tried to find out with which starting values $(\mathbf{U}_0, \mathbf{\Lambda}_0)$ the iteration converges. For this purpose the values

$$(\mathbf{U}_{OP},\, \mathbf{\Lambda}_{OP}) \text{ or } (\mathbf{U}_{LS},\, \mathbf{\Lambda}_{LS} = \mathbf{0}) \text{ or } (\mathbf{U}_I = \mathbf{I},\, \mathbf{\Lambda}_I)$$

specified in (5.51) – (5.53) were tested and an additional attempt was made to combine the three possibilities for $\mathbf{U}_0$ with $\mathbf{\Lambda}_0 = \mathbf{I}$ and with $\mathbf{\Lambda}_0 = \mathbf{0}$. Varying the values for N, n, p and using dependent as well as independent coordinates 10 experiments were made in each case. Here the maximum number of iteration steps, the maximum number of bisection steps and the convergence parameter for the iteration were given by

$$i_{\max} = 48, \quad s_{\max} = 16, \quad \varepsilon = 10^{-3}.$$

If an iteration converged, the matrix $\mathbf{K}$ defined in (5.36) was computed and the iteration was classified to be successful when $\mathbf{K}$ had only non-negative eigenvalues. Then actually the constrained minimum of the objective f was determined.

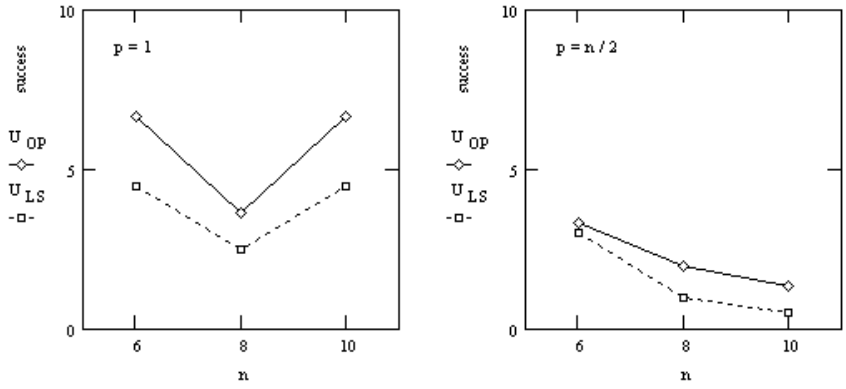
Table 5.2: Iteration results for dependent, unscaled coordinates

N, n, p		$\mathbf{U}_{OP}$	$\mathbf{U}_{LS}$	$\mathbf{I}$
9, 6, 1	$\mathbf{\Lambda}$	10/10 (4.2/ 1.0)	10/10 (5.1/ 1.3)	10/10 (4.9/ 1.0)
	$\mathbf{I}$	10/10 (4.3/ 1.0)	10/10 (5.9/ 1.7)	10/10 (4.8/ 1.0)
	$\mathbf{0}$	10/10 (4.1/ 1.0)		10/10 (4.7/ 1.0)
9, 6, 3	$\mathbf{\Lambda}$	10/10 (4.5/ 1.0)	10/10 (4.7/ 1.0)	10/10 (4.9/ 1.0)
	$\mathbf{I}$	10/10 (4.6/ 1.0)	10/10 (5.8/ 1.5)	10/10 (5.3/ 1.0)
	$\mathbf{0}$	10/10 (4.2/ 1.0)		10/10 (4.2/ 1.0)
12, 8, 1	$\mathbf{\Lambda}$	10/10 (4.2/ 1.0)	10/10 (6.2/ 1.7)	9/ 8 (6.8/ 2.3)
	$\mathbf{I}$	10/10 (4.3/ 1.0)	10/10 (6.2/ 1.6)	10/10 (5.4/ 1.3)
	$\mathbf{0}$	10/10 (4.1/ 1.0)		10/10 (5.3/ 1.2)
12, 8, 4	$\mathbf{\Lambda}$	10/10 (4.5/ 1.0)	10/10 (5.3/ 1.4)	10/10 (5.0/ 1.0)
	$\mathbf{I}$	10/10 (5.3/ 1.3)	10/10 (5.8/ 1.4)	10/10 (5.4/ 1.1)
	$\mathbf{0}$	10/10 (4.1/ 1.0)		10/10 (4.5/ 1.0)
15, 10, 1	$\mathbf{\Lambda}$	10/10 (4.4/ 1.2)	10/10 (5.5/ 1.9)	10/ 8 (7.5/ 2.1)
	$\mathbf{I}$	10/10 (4.3/ 1.0)	10/10 (6.3/ 2.0)	9/ 9 (5.0/ 1.0)
	$\mathbf{0}$	10/10 (4.1/ 1.0)		10/10 (5.2/ 1.1)
15, 10, 5	$\mathbf{\Lambda}$	10/10 (4.9/ 1.0)	10/10 (5.7/ 1.4)	10/10 (5.6/ 1.5)
	$\mathbf{I}$	10/10 (5.1/ 1.0)	10/10 (5.9/ 1.2)	10/10 (5.1/ 1.0)
	$\mathbf{0}$	10/10 (4.6/ 1.0)		10/10 (5.0/ 1.0)
Total	$\mathbf{\Lambda}$	60/60 (4.5/ 1.0)	60/60 (5.4/ 1.5)	59/56 (5.8/ 1.5)
	$\mathbf{I}$	60/60 (4.7/ 1.1)	60/60 (6.0/ 1.6)	59/59 (5.2/ 1.1)
	$\mathbf{0}$	60/60 (4.2/ 1.0)		60/60 (4.8/ 1.1)

Table 5.3: Iteration results for independent, unscaled coordinates

N, n, p		$\mathbf{U}_{OP}$	$\mathbf{U}_{LS}$	$\mathbf{I}$
9, 6, 1	$\mathbf{\Lambda}$	10/ 5 (9.7/ 3.3)	7/ 5 (10.9/ 6.7)	6/ 0 (19.0/ 7.3)
	$\mathbf{I}$	10/ 8 (8.0/ 3.2)	8/ 4 (15.4/ 8.1)	1/ 0 (28.0/ 8.0)
	$\mathbf{0}$	9/ 7 (7.6/ 2.7)		3/ 0 (23.7/ 7.3)
9, 6, 3	$\mathbf{\Lambda}$	5/ 2 (14.6/ 5.8)	5/ 2 (20.6/ 7.4)	5/ 0 (17.0/ 5.4)
	$\mathbf{I}$	3/ 3 (8.3/ 2.0)	7/ 4 (10.6/ 6.4)	1/ 0 (41.0/13.0)
	$\mathbf{0}$	8/ 5 (7.8/ 2.1)		3/ 0 (15.0/ 6.3)
12, 8, 1	$\mathbf{\Lambda}$	9/ 1 (15.9/ 5.6)	9/ 3 (22.6/ 8.8)	2/ 0 (36.0/ 6.0)
	$\mathbf{I}$	8/ 4 (10.1/ 3.1)	6/ 2 (11.0/ 7.0)	1/ 0 (24.0/ 7.0)
	$\mathbf{0}$	8/ 6 (8.4/ 2.4)		0/ 0
12, 8, 4	$\mathbf{\Lambda}$	1/ 0 (33.0/12.0)	4/ 1 (24.0/ 8.0)	0/ 0
	$\mathbf{I}$	3/ 2 (11.3/ 7.3)	3/ 1 (16.7/ 6.7)	1/ 0 (28.0/ 7.0)
	$\mathbf{0}$	5/ 4 (10.8/ 3.2)		2/ 0 (23.5/ 7.0)
15, 10, 1	$\mathbf{\Lambda}$	8/ 6 (15.0/ 5.6)	6/ 6 (12.7/ 7.8)	0/ 0
	$\mathbf{I}$	9/ 8 (8.3/ 2.8)	4/ 3 (15.0/ 7.3)	1/ 0 (43.0/ 6.0)
	$\mathbf{0}$	8/ 6 (8.4/ 2.3)		0/ 0
15, 10, 5	$\mathbf{\Lambda}$	0/ 0	3/ 1 (23.7/ 9.0)	1/ 0 (38.0/ 8.0)
	$\mathbf{I}$	3/ 1 (18.7/ 5.3)	1/ 0 (18.0/ 8.0)	0/ 0
	$\mathbf{0}$	7/ 3 (13.7/ 5.0)		0/ 0
Total	$\mathbf{\Lambda}$	33/14 (14.1/ 5.1)	34/18 (18.4/ 7.9)	14/ 0 (22.1/ 6.5)
	$\mathbf{I}$	36/26 (9.8/ 3.5)	29/14 (13.5/ 7.2)	5/ 0 (32.8/ 8.2)
	$\mathbf{0}$	45/31 (9.2/ 2.9)		8/ 0 (20.4/ 6.9)

Figure 5.1: Comparison of the iteration results for $\mathbf{U}_{OP}$ and $\mathbf{U}_{LS}$



The results of the experiments with dependent coordinates are listed in Table 5.2, the results with independent coordinates are listed in Table 5.3. For example, the entry

$$9/7\ (7.6/2.7)$$

in the third row and third column of Table 5.3 describes the 10 experiments with $N = 9$, $n = 6$, $p = 1$, the starting values $(\mathbf{U}_{OP}, \mathbf{0})$ and independent coordinates. It means that 9 of the 10 iterations converged in an average of 7.6 iteration steps and an average of at most 2.7 bisection steps. In 7 of the 9 convergent cases the minimum was found.

The entries for the starting values $(\mathbf{U}_{LS}, \mathbf{0})$ are empty because they are equal to those for $(\mathbf{U}_{LS}, \mathbf{\Lambda}_{LS})$. Moreover, the last three rows of the tables show the total results for the 60 experiments made with each combination of starting values.

Table 5.2 shows that when dependent coordinates were given the iteration in almost all experiments converged and determined the minimum. Thereby only between 4 and 6 iteration steps were required in which mostly no bisection steps were made. Only with the starting values $(\mathbf{I}, \mathbf{\Lambda}_I)$ and $(\mathbf{I}, \mathbf{I})$ the iteration failed in some cases. Thus, in these experiments none of the other combinations was found to be preferable.

The situation drastically changed when independent coordinates were given. Here the last column of Table 5.3 reveals that with $\mathbf{U}_0 = \mathbf{I}$ only a few iterations converged and none found the minimum. Thus $\mathbf{U}_0 = \mathbf{I}$ should not be used to initialise the iteration.

In order to compare the results for $\mathbf{U}_{OP}$ and $\mathbf{U}_{LS}$ we have computed the respective average numbers of successful iterations for each choice of N, n and p. For example, in the case $N = 9$, $n = 6$, $p = 1$ the average for $\mathbf{U}_{OP}$ is $(5+8+7)/3$ and for $\mathbf{U}_{LS}$ it is $(5+4)/2$. The results of these computations are shown in Figure 5.1 where the horizontal axis represents the spatial dimension n, the vertical axis the computed average and the cases $p = 1$ and $p = n/2$ are drawn separately.

In both graphics the solid line, representing the averages for $\mathbf{U}_{OP}$, lies above the dashed line, representing the averages for $\mathbf{U}_{LS}$. This indicates that $\mathbf{U}_{OP}$ is the better choice for $\mathbf{U}_0$ which can also be seen from the total results in

Table 5.3. In particular, with the starting values $(\mathbf{U}_{OP}, \mathbf{0})$ the best results were obtained.

To confirm these observations some further experiments were made with $N = 24$, $n = 16$ and $p = 1$ or $p = 8$. When dependent coordinates were given, the iteration found the minimum with all starting values, except for $(\mathbf{I}, \mathbf{\Lambda}_I)$. When independent coordinates were given and $p = 1$, the starting values $(\mathbf{U}_{OP}, \mathbf{0})$ again produced the best results: 7 iterations were convergent and in 5 iterations the minimum was determined. However, in the case $p = 8$ none of all the iterations converged.

We have then tried to force convergence based on the following heuristics: It is well-known that the Newton method converges only if the starting values $\mathbf{z}_0$ are "sufficiently" closed to the solution $\mathbf{z}$ [ST1, Kapitel 5.3]. Now using

$$\mathbf{z} = \begin{pmatrix} \mathbf{u}^\wedge \\ \lambda \end{pmatrix}, \quad \mathbf{u}^\wedge = \operatorname{vec}(\mathbf{U})^\wedge, \quad \lambda = \operatorname{vec}\left(\begin{bmatrix} \lambda_{11} & \cdots & \operatorname{Re}(\lambda_{1n}) \\ \vdots & & \vdots \\ \operatorname{Im}(\lambda_{n1}) & \cdots & \lambda_{nn} \end{bmatrix} \right)$$

and the notation $\operatorname{Diag}(\mathbf{\Lambda}) = \operatorname{diag}(\lambda_{11}, \ldots, \lambda_{nn})$ it can be verified that

$$\|\mathbf{z} - \mathbf{z}_0\|^2 = \|\mathbf{U} - \mathbf{U}_0\|_F^2 + \tfrac{1}{2}\big(\|\mathbf{\Lambda} - \mathbf{\Lambda}_0\|_F^2 + \|\operatorname{Diag}(\mathbf{\Lambda}) - \operatorname{Diag}(\mathbf{\Lambda}_0)\|_F^2\big).$$

Following the experiments described above the best choice to bound the right hand side is to set $\mathbf{U}_0 = \mathbf{U}_{OP}$ and $\mathbf{\Lambda}_0 = \mathbf{0}$ which implies

$$\|\mathbf{z} - \mathbf{z}_0\|^2 = \|\mathbf{U} - \mathbf{U}_{OP}\|_F^2 + \tfrac{1}{2}\big(\|\mathbf{\Lambda}\|_F^2 + \|\operatorname{Diag}(\mathbf{\Lambda})\|_F^2\big). \tag{5.54}$$

Moreover, if $\mathbf{U}$ solves (5.38), it also solves the scaled problems in which $\mathbf{X}$ and $\mathbf{Y}$ in the objective are replaced by $\alpha\mathbf{X}$ and $\alpha\mathbf{Y}$ for some $\alpha \in \mathbb{R}\backslash\{0\}$. The corresponding Lagrange multipliers, according to (5.39), then fulfil

$$\mathbf{G}\mathbf{U}\mathbf{\Lambda} = \mathbf{H}_f(\alpha\mathbf{Y})(\alpha\mathbf{X})^* - \mathbf{H}_f\mathbf{U}(\alpha\mathbf{X})(\alpha\mathbf{X})^*,$$

from which with $\mathbf{U}^*\mathbf{G}\mathbf{U} = \mathbf{G}$ it follows that

$$\mathbf{\Lambda} = \alpha^2\mathbf{G}^{-1}\mathbf{U}^*\mathbf{H}_f(\mathbf{Y} - \mathbf{U}\mathbf{X})\mathbf{X}^*.$$

Thus the terms containing $\mathbf{\Lambda}$ in (5.54) can be made arbitrarily small by choosing α small. On the other hand, the matrices $\alpha\mathbf{X}$ and $\alpha\mathbf{Y}$ must sufficiently differ from $\mathbf{0}$ to avoid rounding errors. We have therefore intuitively chosen α such that

$$\|\alpha\mathbf{X}\|_F \, \|\alpha\mathbf{Y}\|_F = 1 \;\text{ or }\; \alpha = (\|\mathbf{X}\|_F \, \|\mathbf{Y}\|_F)^{-1/2} \tag{5.55}$$

and repeated the experiments with the scaled independent coordinates.

Indeed it was found that the scaling approach succeeded for the starting values $(\mathbf{U}_{OP}, \mathbf{0})$. The results will be discussed in detail below. Also for the starting values $(\mathbf{U}_{OP}, \mathbf{\Lambda}_{OP})$ an improvement was observed. In the case $p = 1$ now 5 iterations converged and 3 found the minimum. However, when $\mathbf{U}_0 = \mathbf{U}_{LS}$ or $\mathbf{U}_0 = \mathbf{I}$ neither in the case $p = 1$ nor in the case $p = 8$ any of the iterations converged.

Table 5.4: Iteration results for independent, scaled coordinates

N, n, p	k	$\varphi(\mathbf{z}_0)$	i	s	c_A^{-1}	c_B^{-1}	$\varphi(\mathbf{z})$	κ_-	$f(\mathbf{I})$	$f(\mathbf{U})$	r_G	c_U^{-1}	$time(s)$
24, 16, 1	1	$1.78e-01$	9	3	$6.06e-04$	$1.26e-04$	$8.93e-15$	0	1.93	0.71	$1.30e-14$	$7.18e-02$	1.97
	2	$2.13e-01$	9	3	$1.33e-03$	$1.90e-05$	$3.33e-09$	0	2.03	0.73	$5.78e-09$	$6.81e-02$	1.97
	3	$1.92e-01$	11	5	$4.84e-06$	$1.08e-06$	$1.61e-15$	0	1.86	0.76	$2.43e-15$	$6.36e-02$	2.47
	4	$1.94e-01$	12	8	$8.89e-05$	$3.36e-05$	$8.52e-09$	0	2.05	0.75	$1.26e-08$	$9.02e-02$	2.73
	5	$2.06e-01$	17	4	$1.69e-04$	$5.96e-05$	$1.19e-14$	0	2.05	0.71	$1.72e-14$	$6.82e-02$	3.94
	6	$1.89e-01$	33	6	$1.60e-04$	$4.66e-05$	$6.00e-16$	0	2.05	0.72	$1.07e-15$	$5.37e-02$	7.91
	7	$2.00e-01$	14	4	$1.09e-03$	$1.38e-04$	$4.79e-16$	1	1.94	0.78	$8.06e-16$	$7.03e-02$	3.22
	8	$2.05e-01$	19	5	$2.23e-08$	$6.55e-09$	$3.43e-15$	1	1.90	0.72	$4.94e-15$	$6.97e-02$	4.48
	9	$1.84e-01$	30	17	$4.56e-04$	$4.19e-06$	$7.02e-02$	–	2.02	0.76	$1.04e-01$	$6.64e-02$	7.47
	10	$1.81e-01$	49	16	$4.20e-05$	$6.94e-07$	$1.18e-01$	–	2.13	0.69	$1.73e-01$	$6.15e-02$	11.94
24, 16, 8	11	$2.34e-01$	25	4	$2.12e-05$	$1.99e-06$	$3.86e-09$	0	2.05	0.80	$5.61e-09$	$2.01e-02$	5.95
	12	$2.54e-01$	26	4	$4.65e-04$	$1.95e-05$	$6.38e-10$	0	2.03	0.75	$9.34e-10$	$2.69e-02$	6.19
	13	$2.38e-01$	27	5	$3.13e-06$	$4.28e-07$	$1.09e-13$	0	2.05	0.85	$1.59e-13$	$2.76e-02$	6.47
	14	$2.56e-01$	29	5	$1.54e-05$	$1.19e-06$	$1.69e-13$	0	2.05	0.83	$2.46e-13$	$2.18e-02$	6.92
	15	$2.48e-01$	35	6	$1.34e-04$	$1.36e-05$	$3.87e-10$	0	1.90	0.83	$5.66e-10$	$3.81e-02$	8.44
	16	$2.27e-01$	39	7	$1.61e-06$	$4.11e-08$	$8.77e-15$	0	2.13	0.80	$1.29e-14$	$3.84e-02$	9.48
	17	$2.22e-01$	49	6	$1.45e-05$	$4.82e-07$	$7.77e-02$	–	1.93	0.78	$1.13e-01$	$2.04e-02$	11.95
	18	$2.32e-01$	49	9	$1.79e-05$	$4.50e-07$	$1.18e-01$	–	2.02	0.83	$1.71e-01$	$3.12e-02$	11.92
	19	$2.44e-01$	49	12	$5.52e-04$	$2.59e-05$	$9.77e-02$	–	1.94	0.88	$1.27e-01$	$4.48e-02$	11.95
	20	$2.31e-01$	49	14	$9.74e-06$	$1.37e-07$	$1.06e-01$	–	1.86	0.83	$1.50e-01$	$3.87e-02$	11.98

Table 5.4 contains the results of the experiments with the starting values $(\mathbf{U}_{OP}, \mathbf{0})$. Here the meanings of the columns are as follows:

- k is the number of the experiment,
- $\varphi(\mathbf{z}_0) = \|\mathbf{F}(\mathbf{z}_0)\|$ is the Euclidean norm of $\mathbf{F}$ at the beginning of the iteration,
- i is the number of iteration steps performed,
- s is the maximum number of bisection steps performed,
- $c_A^{-1} = \min\{\mathrm{cond}_1(\mathbf{A}_\nu)^{-1}\}$ is the worst reciprocal 1-condition number computed in step (1) of Algorithm 5.8 when solving $\mathbf{DF}\,\mathbf{dz} = \mathbf{F}$,
- $c_B^{-1} = \min\{\mathrm{cond}_1(\sum_\nu \mathrm{Re}(\mathbf{X}_\nu^* \mathbf{B}_\nu))^{-1}\}$ is the worst reciprocal 1-condition number computed in step (2) of Algorithm 5.8,
- $\varphi(\mathbf{z}) = \|\mathbf{F}(\mathbf{z})\|$ is the Euclidean norm of $\mathbf{F}$ at the end of the iteration,
- κ_- is the number of non-positive eigenvalues of the matrix $\mathbf{K}$ or "$-$" if the iteration did not converge,
- $f(\mathbf{I})$ is the value of the objective for the identity matrix,
- $f(\mathbf{U})$ is the value of the objective for the computed matrix $\mathbf{U}$,
- $r_G = \|\mathbf{U}^*\mathbf{G}\mathbf{U} - \mathbf{G}\|_F$ is the residual estimating the G-unitarity of $\mathbf{U}$,
- $c_U^{-1} = \mathrm{cond}_1(\mathbf{U})^{-1}$ is the reciprocal 1-condition number of $\mathbf{U}$,
- *time* is the time in seconds used for the iteration.

In these experiments the maximum number of iteration steps $i_{\max}$, the maximum number of bisection steps $s_{\max}$, and the tolerance parameter ε were specified by

$$i_{\max} = 48, \quad s_{\max} = 16, \quad \varepsilon = 10^{-8}.$$

Whereas the iteration in the experiments 1 – 8 and 11 – 16 converged with $\varphi(\mathbf{z}) < \varepsilon$, the experiments 10 and 17 – 20 were terminated after i_{max} iteration steps without satisfying the convergence criterion. Experiment 9 failed in the 30-th iteration step because $\varphi(\mathbf{z})$ was not reduced after s_{max} bisection steps.

In most experiments the reciprocal condition numbers c_A^{-1} and c_B^{-1} are sufficiently larger than the machine accuracy ε_{mach}, so that the iterations may be regarded as stable. Only in experiment 8 there is some doubt concerning the accuracy of the vectors $\mathbf{dz}$ computed during the iteration.

The comparison of $f(\mathbf{I})$ with $f(\mathbf{U})$ reveals a convincing reduction of the objective in all experiments and also the reciprocal condition numbers c_U^{-1} indicate well-conditioned iteration results $\mathbf{U}$. But, of course, only in the convergent cases the residuals r_G allow to interpret $\mathbf{U}$ as a G-isometry. Here the minimum and maximum residual obtained in experiment 7 and 4 show that approximately $\varepsilon_{mach} \le r_G \le \sqrt{\varepsilon_{mach}}$ which is a satisfactory result.

Table 5.5: Iteration results for various scaling factors

k	β_{opt}	i	s	κ_-	k	β_{opt}	i	s	κ_-
1	2.2 – 2.4	7	1	0	17	1.4 – 1.6	45	6	0
2	0.8 – 1.0	9	3	0	12	1.4 – 2.2	13	3	0
3	2.2 – 2.4	8	3	0	20	1.4 – 2.4	15	4	1
4	0.6	10	4	0	13	1.4 – 1.8	22	5	0
5	1.2 – 1.4	15	4	0	14	1.8	15	4	0
6	1.4	26	5	0	11	2.0 – 2.4	16	4	0
7	2.0 – 2.4	8	2	1	19	1.2	15	4	1
8	1.6 – 2.4	9	2	1	15	1.0	35	6	0
9	0.6	16	3	0	18	2.2 – 2.4	12	3	4
10	0.8	11	3	0	16	1.0	39	7	0

Although the iterations in experiment 7 and 8 converged, they did not determine the minimum since the real 256×256 matrix $\mathbf{K}$ had one non-positive eigenvalue λ_1 in each case. However, the ratio

$$|\lambda_1| / \sum_{i=1}^{256} |\lambda_i|, \quad \lambda_i \in \sigma(\mathbf{K})$$

was in both experiments very small, around 10^{-3} only. Thus the computed matrices $\mathbf{U}$ are "nearly" solutions of the problems which were still useful in real applications.

The numbers of iteration steps i show that the speed of convergence in the case $p = 8$ was considerably slower than in the case $p = 1$. In particular, the iteration in experiment 17 also converged when it was allowed to take 60 iteration steps. But these are too many steps, so that we have finally tried to accelerate convergence by choosing more appropriate scaling factors.

For this purpose some further experiments were made in which the matrices $\mathbf{X}$ and $\mathbf{Y}$ were scaled with various factors $\alpha\beta$ where α was chosen according to (5.55) and β such that $0.4 \leq \beta \leq 2.4$. The best results are listed in Table 5.5. Here β_{opt} denotes the observed optimum for β and the table is furthermore organised such that experiments made with equal coordinates are contained in one row.

The entries where $\beta_{\text{opt}} = 1.0$ show that α was the best scaling factor only in the experiments 2, 15 and 16. In the further experiments which converged with factor α it was possible to reduce the number of iteration steps. In the remaining experiments 9, 10 and 17 – 20 it was even possible to find the minimum or at least a nearby solution. Thus the iteration always succeeded when a suitable scaling factor was used.

Whereas these results are pleasant on the one hand, the different values of β_{opt} contained in one row also indicate that the optimum factor $\alpha_{\text{opt}} = \alpha\beta_{\text{opt}}$ does not only depend on the matrices $\mathbf{X}$ and $\mathbf{Y}$. Unfortunately, we are therefore not able to present an empirical formula with which it can be determined. Nevertheless, the results obtained with the starting values $(\mathbf{U}_{OP}, \mathbf{0})$ in combination with a scaling factor around α are very satisfactory.

The well behaviour of this combination has also been observed in tests with

randomly initialised matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{H}_f$ and computations for solving problems with one or two constraints. Moreover, the coordinates $\mathbf{X}$ and $\mathbf{Y}$ in real multidimensional scaling applications are usually not totally independent, so that the choice of appropriate starting values is not the major problem of the method. It is rather the fact that the amount of work is of order n^6. This also explains the times listed in Table 5.4 which were measured on a 2GHz processor. We may therefore conclude that our algorithm is suitable for solving (5.38) provided that the spatial dimension n is not too large.

Chapter 6

An algorithm for the numerical computation of canonical forms

6.1 Introduction

Almost all the results presented in Chapter 3 and Chapter 4 are based on Theorem 3.1: If $\mathbf{H} \in \mathbb{C}^{n\times n}$ is nonsingular and Hermitian and $\mathbf{A} \in \mathbb{C}^{n\times n}$ is H-Hermitian, then there exists a nonsingular matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ such that

$$\begin{aligned} \mathbf{S}^{-1}\mathbf{A}\mathbf{S} &= \bigoplus_{i=1}^{r} \mathbf{J}_{p_i}(\lambda_i) \oplus \bigoplus_{i=r+1}^{s} \begin{bmatrix} \mathbf{J}_{p_i}(\lambda_i) & \\ & \mathbf{J}_{p_i}(\overline{\lambda}_i) \end{bmatrix}, \\ \mathbf{S}^{*}\mathbf{H}\mathbf{S} &= \bigoplus_{i=1}^{r} \varepsilon_i \mathbf{Z}_{p_i} \oplus \bigoplus_{i=r+1}^{s} \begin{bmatrix} & \mathbf{Z}_{p_i} \\ \mathbf{Z}_{p_i} & \end{bmatrix} \end{aligned} \tag{6.1}$$

where $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$, $\varepsilon_1, \ldots, \varepsilon_r \in \{+1, -1\}$ and $\lambda_{r+1}, \ldots, \lambda_s \in \mathbb{C}\backslash\mathbb{R}$.

Whereas this canonical form of the pair $(\mathbf{A}, \mathbf{H})$ in the case of a diagonalisable matrix $\mathbf{A}$ can easily be computed numerically using HQR and HQR-2 or eigenvalue and singular value decompositions (see Method 3.24), its computation in the general case is a complicated problem. The major challenge is that (6.1) contains the Jordan normal form (JNF) of the matrix $\mathbf{A}$ whose numerical computation is known to be a very difficult task.

The most successful approach for computing the JNF was developed by Kågström and Ruhe [KR1]. Their procedure is available in the ACM algorithm collection as the CALGO algorithm 560 written in the programming language Fortran 66 [KR2]. Based on this work we present a method with which the canonical form $(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^{*}\mathbf{H}\mathbf{S})$ and an associated transformation matrix $\mathbf{S}$ can be computed numerically. The corresponding extension of the 560 algorithm essentially consists of a normalisation step which constitutes a generalisation of the Cholesky method.

The description of the algorithm is organised as follows: The required mathematical background is established in Section 6.2. In Section 6.3 the individual steps of the method are explained and in Section 6.4 some numerical results

are presented. The final Section 6.5 reconsiders the numerical computation of H-polar decompositions.

6.2 Mathematical background

Let $\mathbf{H} \in \mathbb{C}^{n\times n}$ be nonsingular and Hermitian and let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian. Then from Theorem 3.1 it follows that for every non-real eigenvalue $\lambda \in \sigma(\mathbf{A})$ also $\overline{\lambda} \in \sigma(\mathbf{A})$ and that the Jordan structures of the two eigenvalues are equal. Let $\lambda_1, \ldots, \lambda_r$ be the real and $\lambda_{r+1}, \ldots, \lambda_s$ be the non-real eigenvalues of $\mathbf{A}$ with positive imaginary part. Moreover, let $\alpha_i = \alpha(\lambda_i)$ and $\rho_i = \rho(\lambda_i)$ denote the algebraic and geometric multiplicity of the eigenvalue λ_i, so that $\alpha(\lambda_i) = \alpha(\overline{\lambda}_i)$ and $\rho(\lambda_i) = \rho(\overline{\lambda}_i)$ for $r+1 \le i \le s$. Then the canonical form

$$(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mathbf{J}, \mathbf{Z}) \tag{6.2}$$

of the pair $(\mathbf{A}, \mathbf{H})$ can be rewritten as

$$\mathbf{J} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{J}}(\lambda_i)\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{J}}(\lambda_i)\right], \quad \mathbf{Z} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{Z}}_i\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{Z}}_i\right], \tag{6.3a}$$

where the blocks which appear have the form

$$\begin{aligned} \hat{\mathbf{J}}(\lambda_i) &= \mathbf{J}_{p_1^{(i)}}(\lambda_i) \oplus \ldots \oplus \mathbf{J}_{p_{\rho_i}^{(i)}}(\lambda_i), \\ \hat{\mathbf{Z}}_i &= \varepsilon_1^{(i)}\mathbf{Z}_{p_1^{(i)}} \oplus \ldots \oplus \varepsilon_{\rho_i}^{(i)}\mathbf{Z}_{p_{\rho_i}^{(i)}} \qquad \text{for } 1 \le i \le r, \end{aligned} \tag{6.3b}$$

$$\begin{aligned} \check{\mathbf{J}}(\lambda_i) &= \begin{bmatrix} \mathbf{J}_{p_1^{(i)}}(\lambda_i) \oplus \ldots \oplus \mathbf{J}_{p_{\rho_i}^{(i)}}(\lambda_i) & \\ & \mathbf{J}_{p_1^{(i)}}(\overline{\lambda}_i) \oplus \ldots \oplus \mathbf{J}_{p_{\rho_i}^{(i)}}(\overline{\lambda}_i) \end{bmatrix}, \\ \check{\mathbf{Z}}_i &= \begin{bmatrix} & \mathbf{Z}_{p_1^{(i)}} \oplus \ldots \oplus \mathbf{Z}_{p_{\rho_i}^{(i)}} \\ \mathbf{Z}_{p_1^{(i)}} \oplus \ldots \oplus \mathbf{Z}_{p_{\rho_i}^{(i)}} & \end{bmatrix} \qquad \text{for } r+1 \le i \le s. \end{aligned} \tag{6.3c}$$

Furthermore, the eigenvalues can be sorted such that $\lambda_1 > \ldots > \lambda_r$ and

$$\mathrm{Re}(\lambda_{r+1}) \ge \ldots \ge \mathrm{Re}(\lambda_s) \;\text{ where }\; \mathrm{Im}(\lambda_i) > \mathrm{Im}(\lambda_{i+1}) \;\text{ if }\; \mathrm{Re}(\lambda_i) = \mathrm{Re}(\lambda_{i+1})$$

and $p_1^{(i)} \ge \ldots \ge p_{\rho_i}^{(i)}$ where $p_1^{(i)} + \ldots + p_{\rho_i}^{(i)} = \alpha_i$.

Now, let $\mathbf{X} \in \mathbb{C}^{n\times n}$ be a nonsingular matrix such that

$$\mathbf{X}^{-1}\mathbf{J}\mathbf{X} = \mathbf{J} \;\text{ or }\; \mathbf{J}\mathbf{X} = \mathbf{X}\mathbf{J}.$$

Then the commutability of $\mathbf{X}$ and $\mathbf{J}$ implies [G, Chapter VIII, §2], that $\mathbf{X}$ must

be a block diagonal matrix

$$\mathbf{X} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{X}}_i\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{X}}_i\right] \quad \text{with} \quad \hat{\mathbf{X}}_i = \begin{bmatrix} \mathbf{X}_{1,1}^{(i)} & \cdots & \mathbf{X}_{1,\rho_i}^{(i)} \\ \vdots & & \vdots \\ \mathbf{X}_{\rho_i,1}^{(i)} & \cdots & \mathbf{X}_{\rho_i,\rho_i}^{(i)} \end{bmatrix},$$

$$\check{\mathbf{X}}_i = \begin{bmatrix} \mathbf{X}_{1,1}^{(i)} & \cdots & \mathbf{X}_{1,\rho_i}^{(i)} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{X}_{\rho_i,1}^{(i)} & \cdots & \mathbf{X}_{\rho_i,\rho_i}^{(i)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Y}_{1,1}^{(i)} & \cdots & \mathbf{Y}_{1,\rho_i}^{(i)} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Y}_{\rho_i,1}^{(i)} & \cdots & \mathbf{Y}_{\rho_i,\rho_i}^{(i)} \end{bmatrix}, \tag{6.4a}$$

whose blocks $\mathbf{X}_{kl}^{(i)}$ and $\mathbf{Y}_{kl}^{(i)}$ are $p_k^{(i)} \times p_l^{(i)}$ matrices of the upper triangular Toeplitz form described by the examples

$$\underset{(p_k = p_l = 3)}{\mathbf{X}_{kl} = \begin{bmatrix} x & y & z \\ 0 & x & y \\ 0 & 0 & x \end{bmatrix}}, \quad \underset{(p_k = 3, p_l = 4)}{\mathbf{X}_{kl} = \begin{bmatrix} 0 & x & y & z \\ 0 & 0 & x & y \\ 0 & 0 & 0 & x \end{bmatrix}}, \quad \underset{(p_k = 4, p_l = 3)}{\mathbf{X}_{kl} = \begin{bmatrix} x & y & z \\ 0 & x & y \\ 0 & 0 & x \\ 0 & 0 & 0 \end{bmatrix}}. \tag{6.4b}$$

Therefore, the matrix

$$\mathbf{C} = \mathbf{X}^* \mathbf{Z} \mathbf{X}$$

must be an Hermitian block diagonal matrix

$$\mathbf{C} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{C}}_i\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{C}}_i\right] \quad \text{with} \quad \hat{\mathbf{C}}_i = \begin{bmatrix} \mathbf{C}_{1,1}^{(i)} & \cdots & \mathbf{C}_{1,\rho_i}^{(i)} \\ \vdots & & \vdots \\ \mathbf{C}_{1,\rho_i}^{(i)*} & \cdots & \mathbf{C}_{\rho_i,\rho_i}^{(i)} \end{bmatrix},$$

$$\check{\mathbf{C}}_i = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_{1,1}^{(i)*} & \cdots & \mathbf{C}_{\rho_i,1}^{(i)*} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_{1,\rho_i}^{(i)*} & \cdots & \mathbf{C}_{\rho_i,\rho_i}^{(i)*} \\ \mathbf{C}_{1,1}^{(i)} & \cdots & \mathbf{C}_{1,\rho_i}^{(i)} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{C}_{\rho_i,1}^{(i)} & \cdots & \mathbf{C}_{\rho_i,\rho_i}^{(i)} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \tag{6.5a}$$

whose blocks $\mathbf{C}_{kl}^{(i)}$ are also $p_k^{(i)} \times p_l^{(i)}$ matrices, but now of the lower anti-triangular Hankel form

$$\underset{(p_k = p_l = 3)}{\mathbf{C}_{kl} = \begin{bmatrix} 0 & 0 & a \\ 0 & a & b \\ a & b & c \end{bmatrix}}, \quad \underset{(p_k = 3, p_l = 4)}{\mathbf{C}_{kl} = \begin{bmatrix} 0 & 0 & 0 & a \\ 0 & 0 & a & b \\ 0 & a & b & c \end{bmatrix}}, \quad \underset{(p_k = 4, p_l = 3)}{\mathbf{C}_{kl} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a \\ 0 & a & b \\ a & b & c \end{bmatrix}}. \tag{6.5b}$$

Furthermore, the diagonal blocks of the matrices $\hat{\mathbf{C}}_i$ can contain only real elements. Summarising, we have

Theorem 6.1. *Let $\mathbf{H} \in \mathbb{C}^{n\times n}$ be nonsingular and Hermitian and let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian. Moreover, let $\mathbf{R} \in \mathbb{C}^{n\times n}$ be a nonsingular matrix such that $\mathbf{J} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R}$ is the Jordan normal form of $\mathbf{A}$. Then the matrix $\mathbf{C} = \mathbf{R}^*\mathbf{H}\mathbf{R}$ has the form* (6.5).

Proof. According to (6.2) it is true that $\mathbf{J} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R} = \mathbf{R}^{-1}(\mathbf{S}\mathbf{J}\mathbf{S}^{-1})\mathbf{R} = (\mathbf{S}^{-1}\mathbf{R})^{-1}\mathbf{J}(\mathbf{S}^{-1}\mathbf{R})$. Hence, for $\mathbf{X} = \mathbf{S}^{-1}\mathbf{R}$ we obtain $\mathbf{C} = (\mathbf{S}^{-1}\mathbf{R})^*\mathbf{Z}(\mathbf{S}^{-1}\mathbf{R}) = \mathbf{R}^*(\mathbf{S}^{-*}\mathbf{Z}\mathbf{S}^{-1})\mathbf{R} = \mathbf{R}^*\mathbf{H}\mathbf{R}$. □

Starting out from this theorem, the proposed algorithm for determining the canonical form is first to compute the (numerical) Jordan normal form $\mathbf{J} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R}$ of $\mathbf{A}$, and then to normalise the matrix $\mathbf{C} = \mathbf{R}^*\mathbf{H}\mathbf{R}$ with a transformation such that

$$\mathbf{X}^{-1}\mathbf{J}\mathbf{X} = \mathbf{J}, \quad \mathbf{X}^*\mathbf{C}\mathbf{X} = \mathbf{Z}. \tag{6.6}$$

Then $(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mathbf{J}, \mathbf{Z})$ with $\mathbf{S} = \mathbf{R}\mathbf{X}$ gives the wanted canonical form.

Whereas the computation of the Jordan normal form can essentially be carried out with the 560 algorithm by Kågström and Ruhe, the normalisation step (6.6) must be described next. For this purpose, we define

$$\mathbf{N}_p = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \qquad \mathbf{C}_p(\alpha_1 .. \alpha_p) = \begin{bmatrix} & & & \alpha_1 \\ & & \iddots & \alpha_2 \\ & \iddots & \iddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_p \end{bmatrix},$$

$$\mathbf{N}_{p,q}^k = \begin{bmatrix} \mathbf{N}_q^k \\ \mathbf{0}_{p-q,q} \end{bmatrix}, \qquad \mathbf{C}_{p,q}(\alpha_1 .. \alpha_q) = \begin{bmatrix} \mathbf{0}_{p-q,q} \\ \mathbf{C}_q(\alpha_1 .. \alpha_q) \end{bmatrix} \text{ for } p \geq q,$$

$$\mathbf{N}_{p,q}^k = \begin{bmatrix} \mathbf{0}_{p,q-p} & \mathbf{N}_p^k \end{bmatrix}, \qquad \mathbf{C}_{p,q}(\alpha_1 .. \alpha_p) = \begin{bmatrix} \mathbf{0}_{p,q-p} & \mathbf{C}_p(\alpha_1 .. \alpha_p) \end{bmatrix} \text{ for } p \leq q$$

where the powers of the nilpotent matrix $\mathbf{N}_p$ are given by $\mathbf{N}_p^0 = \mathbf{I}_p$, $\mathbf{N}_p^1 = \mathbf{N}_p$,

$$\mathbf{N}_p^2 = \begin{bmatrix} 0 & 0 & 1 & & \\ & 0 & 0 & \ddots & \\ & & 0 & \ddots & 1 \\ & & & \ddots & 0 \\ & & & & 0 \end{bmatrix}, \quad \ldots, \quad \mathbf{N}_p^{p-1} = \begin{bmatrix} 0 & 0 & 0 & & 1 \\ & 0 & 0 & \ddots & \\ & & 0 & \ddots & 0 \\ & & & \ddots & 0 \\ & & & & 0 \end{bmatrix},$$

and $\mathbf{N}_p^p = \mathbf{0}_{p,p}$. Now, let

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{p_1}(\alpha_1^{(1)} .. \alpha_1^{(p_1)}) & \mathbf{C}_{p_1,p_2}(\rho_{12}^{(1)} .. \rho_{12}^{(p_2)}) & \cdots & \mathbf{C}_{p_1,p_m}(\rho_{1m}^{(1)} .. \rho_{1m}^{(p_m)}) \\ \mathbf{C}_{p_2,p_1}(\rho_{21}^{(1)} .. \rho_{21}^{(p_2)}) & \mathbf{C}_{p_2}(\alpha_2^{(1)} .. \alpha_2^{(p_2)}) & \cdots & \mathbf{C}_{p_2,p_m}(\rho_{2m}^{(1)} .. \rho_{2m}^{(p_m)}) \\ \vdots & \vdots & & \vdots \\ \mathbf{C}_{p_m,p_1}(\rho_{m1}^{(1)} .. \rho_{m1}^{(p_m)}) & \mathbf{C}_{p_m,p_2}(\rho_{m2}^{(1)} .. \rho_{m2}^{(p_m)}) & \cdots & \mathbf{C}_{p_m}(\alpha_m^{(1)} .. \alpha_m^{(p_m)}) \end{bmatrix} \tag{6.7}$$

be a nonsingular matrix with $p_1 \geq \ldots \geq p_m$, and assume that $\alpha_1^{(1)} \neq 0$. Then the elements $\rho_{1j}^{(h)}$, $2 \leq j \leq m$, $1 \leq h \leq p_j$, can be eliminated successively by multiplication on the right with the matrices

$$\mathbf{U}_k = \begin{bmatrix} \mathbf{I}_{p_1} & \varphi_{12}^{(k)}\mathbf{N}_{p_1,p_2}^{k-1} & \varphi_{13}^{(k)}\mathbf{N}_{p_1,p_3}^{k-1} & \cdots & \varphi_{1m}^{(k)}\mathbf{N}_{p_1,p_m}^{k-1} \\ & \mathbf{I}_{p_2} & \mathbf{0}_{p_2,p_3} & \cdots & \mathbf{0}_{p_2,p_m} \\ & & \mathbf{I}_{p_3} & \cdots & \mathbf{0}_{p_3,p_m} \\ & & & \ddots & \vdots \\ & & & & \mathbf{I}_{p_m} \end{bmatrix}, \quad 1 \leq k \leq p_2, \tag{6.8a}$$

by setting

$$\varphi_{1j}^{(k)} = \left\{ \begin{array}{cc} -\dfrac{\rho_{1j}^{(k)}}{\alpha_1^{(1)}}, & \text{for } 1 \leq k \leq p_j \\ 0, & \text{otherwise} \end{array} \right\}, \quad 2 \leq j \leq m. \tag{6.8b}$$

In the same way the elements $\rho_{j1}^{(h)}$, $2 \leq j \leq m$, $1 \leq h \leq p_j$, can be eliminated by multiplication on the left with the matrices

$$\mathbf{V}_k^* = \begin{bmatrix} \mathbf{I}_{p_1} & \overline{\psi}_{12}^{(k)}\mathbf{N}_{p_1,p_2}^{k-1} & \overline{\psi}_{13}^{(k)}\mathbf{N}_{p_1,p_3}^{k-1} & \cdots & \overline{\psi}_{1m}^{(k)}\mathbf{N}_{p_1,p_m}^{k-1} \\ & \mathbf{I}_{p_2} & \mathbf{0}_{p_2,p_3} & \cdots & \mathbf{0}_{p_2,p_m} \\ & & \mathbf{I}_{p_3} & \cdots & \mathbf{0}_{p_3,p_m} \\ & & & \ddots & \vdots \\ & & & & \mathbf{I}_{p_m} \end{bmatrix}^*, \quad 1 \leq k \leq p_2, \tag{6.9a}$$

by setting

$$\psi_{j1}^{(k)} = \left\{ \begin{array}{cc} -\dfrac{\rho_{j1}^{(k)}}{\alpha_1^{(1)}}, & \text{for } 1 \leq k \leq p_j \\ 0, & \text{otherwise} \end{array} \right\}, \quad 2 \leq j \leq m. \tag{6.9b}$$

If $\mathbf{C}$ is even Hermitian, i.e. if $\alpha_j^{(h)} \in \mathbb{R}$ and $\rho_{ji}^{(h)} = \overline{\rho}_{ij}^{(h)} \in \mathbb{C}$ for $1 \leq i, j \leq m$, $1 \leq h \leq p_j$, then in particular $\mathbf{U}_k = \mathbf{V}_k$.

The manner of designation is to be understood here such that the elements modified by the multiplication with $\mathbf{U}_k$ and $\mathbf{V}_k^*$, namely $\tilde{\alpha}_j^{(h)}$, $\tilde{\rho}_{j1}^{(h)}$, $\tilde{\rho}_{j1}^{(k)} = 0$ and $\tilde{\alpha}_j^{(h)}$, $\tilde{\rho}_{1j}^{(h)}$, $\tilde{\rho}_{1j}^{(k)} = 0$, are renamed to $\alpha_j^{(h)}$ and $\rho_{j1}^{(h)}$ or $\rho_{1j}^{(h)}$ before carrying out the next transformation step ($2 \leq j \leq m$, $k \leq h \leq p_j$). This process is symbolised below as

$$\tilde{\mathbf{C}} = \mathbf{C}\mathbf{U}_k \to \mathbf{C} \;\text{ and }\; \tilde{\mathbf{C}} = \mathbf{V}_k^*\mathbf{C} \to \mathbf{C}.$$

Having performed all elimination steps we obtain the matrix

$$\begin{aligned} \mathbf{C}' &= \mathbf{V}_{p_2}^* \ldots \mathbf{V}_1^* \mathbf{C} \mathbf{U}_1 \ldots \mathbf{U}_{p_2} \\ &= \begin{bmatrix} \mathbf{C}_{p_1}(\alpha_1^{(1)} .. \alpha_1^{(p_1)}) & \mathbf{0}_{p_1,p_2} & \cdots & \mathbf{0}_{p_1,p_m} \\ \mathbf{0}_{p_2,p_1} & \mathbf{C}_{p_2}(\alpha_2^{(1)} .. \alpha_2^{(p_2)}) & \cdots & \mathbf{C}_{p_2,p_m}(\rho_{2m}^{(1)} .. \rho_{2m}^{(p_m)}) \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{p_m,p_1} & \mathbf{C}_{p_m,p_2}(\rho_{m2}^{(1)} .. \rho_{m2}^{(p_m)}) & \cdots & \mathbf{C}_{p_m}(\alpha_m^{(1)} .. \alpha_m^{(p_m)}) \end{bmatrix}, \end{aligned} \tag{6.10}$$

whose $(1,1)$-block remains unchanged. Using matrices of the form

$$\begin{aligned}\mathbf{W}_1 &= \begin{bmatrix} \chi_1^{(1)}\mathbf{I}_{p_1} & & & \\ & \mathbf{I}_{p_2} & & \\ & & \ddots & \\ & & & \mathbf{I}_{p_m} \end{bmatrix}, \\ \mathbf{W}_k &= \begin{bmatrix} \mathbf{I}_{p_1} + \chi_1^{(k)}\mathbf{N}_{p_1}^{k-1} & & & \\ & \mathbf{I}_{p_2} & & \\ & & \ddots & \\ & & & \mathbf{I}_{p_m} \end{bmatrix}, \quad 2 \le k \le p_1, \end{aligned} \tag{6.11a}$$

and transformations of the kind $\tilde{\mathbf{C}}' = \mathbf{W}_k^T\mathbf{C}'\mathbf{W}_k \to \mathbf{C}'$ it is furthermore possible to normalise $\alpha_1^{(1)}$ to $\varepsilon_1 = \pm 1$, and the remaining elements $\alpha_1^{(k)}$, $2 \le k \le p_1$ can be eliminated successively by setting

$$\chi_1^{(1)} = \left\{ \begin{array}{ll} \dfrac{1}{\sqrt{|\alpha_1^{(1)}|}}, & \text{if } \mathbf{C}^* = \mathbf{C} \\ \dfrac{1}{\sqrt{\alpha_1^{(1)}}}, & \text{otherwise} \end{array} \right\} \text{ and } \chi_1^{(k)} = -\frac{\alpha_1^{(k)}}{2\varepsilon_1}, \quad 2 \le k \le p_1. \tag{6.11b}$$

In the case $\mathbf{C}^* = \mathbf{C}$ the elements of $\mathbf{W}_k$ are real so that in particular $\mathbf{W}_k^T = \mathbf{W}_k^*$.

After all elimination steps we obtain the matrix

$$\begin{aligned}\mathbf{C}'' &= \mathbf{W}_{p_1}^T \dots \mathbf{W}_1^T\mathbf{C}'\mathbf{W}_1 \dots \mathbf{W}_{p_1} \\ &= \begin{bmatrix} \varepsilon_1\mathbf{Z}_{p_1} & \mathbf{0}_{p_1,p_2} & \cdots & \mathbf{0}_{p_1,p_m} \\ \mathbf{0}_{p_2,p_1} & \mathbf{C}_{p_2}(\alpha_2^{(1)} .. \alpha_2^{(p_2)}) & \cdots & \mathbf{C}_{p_2,p_m}(\rho_{2m}^{(1)} .. \rho_{2m}^{(p_m)}) \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{p_m,p_1} & \mathbf{C}_{p_m,p_2}(\rho_{m2}^{(1)} .. \rho_{m2}^{(p_m)}) & \cdots & \mathbf{C}_{p_m}(\alpha_m^{(1)} .. \alpha_m^{(p_m)}) \end{bmatrix} \\ &\text{where } \varepsilon_1 = \left\{ \begin{array}{ll} \pm 1, & \text{if } \mathbf{C}^* = \mathbf{C} \\ 1, & \text{otherwise} \end{array} \right\}. \end{aligned} \tag{6.12}$$

However, the same result is also achieved when the transformations are carried out in the order

$$\begin{aligned} &\mathbf{C}'' = \mathbf{Y}^*\mathbf{C}\mathbf{X} \text{ with} \\ &\mathbf{X} = \mathbf{W}_1\mathbf{U}_1\mathbf{W}_2\mathbf{U}_2 \dots \mathbf{W}_{p_2}\mathbf{U}_{p_2}\mathbf{W}_{p_2+1} \dots \mathbf{W}_{p_1}, \\ &\mathbf{Y} = \overline{\mathbf{W}}_1\mathbf{V}_1\overline{\mathbf{W}}_2\mathbf{V}_2 \dots \overline{\mathbf{W}}_{p_2}\mathbf{V}_{p_2}\overline{\mathbf{W}}_{p_2+1} \dots \overline{\mathbf{W}}_{p_1}, \\ &\mathbf{Y} = \mathbf{X} \text{ if } \mathbf{C}^* = \mathbf{C}. \end{aligned} \tag{6.13a}$$

With this sequence the element $\alpha_1^{(1)}$ takes on the constant value $\varepsilon_1 = \pm 1$ after the first transformation, and the parameters for the subsequent transformations are

$$\varphi_{1j}^{(k)} = \pm\rho_{1j}^{(k)}, \quad \psi_{j1}^{(k)} = \pm\rho_{j1}^{(k)} \text{ and } \chi_1^{(k)} = \pm\frac{\alpha_1^{(k)}}{2}. \tag{6.13b}$$

Thus, the divisions by $\alpha_1^{(1)}$ are not necessary, so that (6.13) presents the numerically favourable order.

To apply this transformation to a general nonsingular matrix $\mathbf{C}$ of the form (6.7), we must now determine matrices $\mathbf{P}$ and $\mathbf{Q}$ such that

$$\tilde{\mathbf{C}} = \begin{Bmatrix} \mathbf{P}^*\mathbf{C}\mathbf{P}, & \text{if } \mathbf{C}^* = \mathbf{C} \\ \mathbf{Q}^*\mathbf{C}\mathbf{P}, & \text{otherwise} \end{Bmatrix} \to \mathbf{C}$$

satisfies the condition $\alpha_1^{(1)} \neq 0$. For this purpose, let μ be the index with $p = p_1 = \ldots = p_\mu > p_{\mu+1} \geq \ldots \geq p_m$, and let

$$\mathbf{\Pi}(1) = \mathbf{I}_q, \quad \mathbf{\Pi}(\nu) = \begin{bmatrix} \mathbf{0} & & \mathbf{I}_p & \\ & \mathbf{I}_{(\nu-2)p} & & \\ \mathbf{I}_p & & \mathbf{0} & \\ & & & \mathbf{I}_{q-\nu p} \end{bmatrix}, \quad 2 \leq \nu \leq \mu, \quad q = \sum_{j=1}^{m} p_j,$$

be a permutation matrix. Furthermore, in the case $\mathbf{C}^* \neq \mathbf{C}$, let $\rho_{jj}^{(1)} = \alpha_j^{(1)}$, and let κ, λ be indices such that $|\rho_{\kappa\lambda}^{(1)}| = \max |\rho_{ij}^{(1)}|$ for $1 \leq i, j \leq \mu$. Then $\rho_{\kappa\lambda}^{(1)} \neq 0$, because otherwise $\mathbf{C}$ would have zero rows and columns and would therefore be singular. Thus, the wanted result is obtained with

$$\mathbf{P} = \mathbf{\Pi}(\lambda) \text{ and } \mathbf{Q} = \mathbf{\Pi}(\kappa). \tag{6.14a}$$

In the case $\mathbf{C}^* = \mathbf{C}$, let ν be an index such that $|\alpha_\nu^{(1)}| = \max |\alpha_j^{(1)}|$ for $1 \leq j \leq \mu$. If $\alpha_\nu^{(1)} \neq 0$, it is possible to use

$$\mathbf{P} = \mathbf{\Pi}(\nu). \tag{6.14b}$$

Otherwise $\alpha_1^{(1)} = \ldots = \alpha_\mu^{(1)} = 0$, and let κ, λ be indices such that $|\rho_{\kappa\lambda}^{(1)}| = \max |\rho_{ij}^{(1)}|$ for $1 \leq i < j \leq \mu$. Then, with the same argumentation as above, it must be true that $\rho_{\kappa\lambda}^{(1)} \neq 0$. If now the unitary transformation

$$\tilde{\mathbf{C}} = \mathbf{\Sigma}^*\mathbf{C}\mathbf{\Sigma} \text{ with } \mathbf{\Sigma} = \begin{bmatrix} \mathbf{I}_{(\kappa-1)p} & & & & \\ & \frac{e^{i\varphi}}{\sqrt{2}}\mathbf{I}_p & & \frac{e^{i\varphi}}{\sqrt{2}}\mathbf{I}_p & \\ & & \mathbf{I}_{(\lambda-\kappa-1)p} & & \\ & \frac{1}{\sqrt{2}}\mathbf{I}_p & & \frac{-1}{\sqrt{2}}\mathbf{I}_p & \\ & & & & \mathbf{I}_{q-\lambda p} \end{bmatrix}$$

for $\varphi = \arg(\rho_{\kappa\lambda}^{(1)})$ is applied, then $\tilde{\alpha}_\kappa^{(1)} = -\tilde{\alpha}_\lambda^{(1)} = |\rho_{\kappa\lambda}^{(1)}|$. With a subsequent permutation

$$\mathbf{P} = \mathbf{\Sigma}\mathbf{\Pi}(\kappa) \tag{6.14c}$$

or

$$\mathbf{P} = \mathbf{\Sigma}\mathbf{\Pi}(\lambda) \tag{6.14d}$$

we can therefore also achieve that $|\tilde{\alpha}_1^{(1)}| = |\rho_{\kappa\lambda}^{(1)}| \neq 0$.

If the elements $\alpha_\kappa^{(1)}$, $\alpha_\lambda^{(1)}$ are non-zero when applying $\mathbf{\Sigma}$, we obtain $\tilde{\alpha}_\kappa^{(1)} = \omega_{\kappa\lambda}^{(1)} + |\rho_{\kappa\lambda}^{(1)}|$ and $\tilde{\alpha}_\lambda^{(1)} = \omega_{\kappa\lambda}^{(1)} - |\rho_{\kappa\lambda}^{(1)}|$, where $\omega_{\kappa\lambda}^{(1)} = (\alpha_\kappa^{(1)} + \alpha_\lambda^{(1)})/2$. It thus makes sense to specify the *pivot element* $\alpha_1^{(1)}$ in the case $\mathbf{C}^* = \mathbf{C}$ as follows:

1. Determine ν such that $|\alpha_\nu^{(1)}|$ is at maximum.

2. Determine κ, λ such that $|\omega_{\kappa\lambda}^{(1)}| + |\rho_{\kappa\lambda}^{(1)}|$ is at maximum.

3. If $|\alpha_\nu^{(1)}| \geq |\omega_{\kappa\lambda}^{(1)}| + |\rho_{\kappa\lambda}^{(1)}|$ use (6.14b). Otherwise, if $\omega_{\kappa\lambda}^{(1)} \geq 0$ use (6.14c). Otherwise, use (6.14d).

In this way a high numeric robustness of the procedure is ensured.

Using the transformations described so far, any nonsingular matrix of the form (6.7) can be transformed into the form

$$\mathbf{Y}_1^*\mathbf{Q}_1^*\mathbf{C}\mathbf{P}_1\mathbf{X}_1 = \begin{bmatrix} \varepsilon_1\mathbf{Z}_{p_1} & & & \\ & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2m} \\ & \vdots & & \vdots \\ & \mathbf{C}_{m2} & \cdots & \mathbf{C}_{mm} \end{bmatrix}. \tag{6.15a}$$

Here the submatrix consisting of the blocks $\mathbf{C}_{ij}$, $2 \leq i, j \leq m$, again is a nonsingular (possibly Hermitian) matrix of the form (6.7), so that the inductive application of the procedure yields

$$\begin{aligned} &\mathbf{Y}^*\mathbf{C}\mathbf{X} = \mathbf{Z} = \varepsilon_1\mathbf{Z}_{p_1} \oplus \ldots \oplus \varepsilon_m\mathbf{Z}_{p_m} \text{ with} \\ &\mathbf{X} = \mathbf{P}_1\mathbf{X}_1 \ldots \mathbf{P}_m\mathbf{X}_m, \;\; \mathbf{Y} = \mathbf{Q}_1\mathbf{Y}_1 \ldots \mathbf{Q}_m\mathbf{Y}_m, \\ &\mathbf{Y} = \mathbf{X}, \;\; \varepsilon_j = \pm 1 \text{ if } \mathbf{C}^* = \mathbf{C} \text{ and } \varepsilon_j = 1 \text{ if } \mathbf{C}^* \neq \mathbf{C}. \end{aligned} \tag{6.15b}$$

Finally, considering that

$$\begin{bmatrix} \mathbf{X}^* & \\ & \mathbf{Y}^* \end{bmatrix} \begin{bmatrix} & \mathbf{C}^* \\ \mathbf{C} & \end{bmatrix} \begin{bmatrix} \mathbf{X} & \\ & \mathbf{Y} \end{bmatrix} = \begin{bmatrix} & \mathbf{X}^*\mathbf{C}^*\mathbf{Y} \\ \mathbf{Y}^*\mathbf{C}\mathbf{X} & \end{bmatrix} = \begin{bmatrix} & \mathbf{Z} \\ \mathbf{Z} & \end{bmatrix}, \tag{6.15c}$$

we have all in all

Theorem 6.2. *Let $\mathbf{H} \in \mathbb{C}^{n\times n}$ be nonsingular and Hermitian and let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be H-Hermitian. Moreover, let $\mathbf{R} \in \mathbb{C}^{n\times n}$ be a nonsingular matrix such that $\mathbf{J} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R}$ is the Jordan normal form of $\mathbf{A}$ and let $\mathbf{C} = \mathbf{R}^*\mathbf{H}\mathbf{R}$. Then there exists a nonsingular matrix $\mathbf{X}$ satisfying $\mathbf{J}\mathbf{X} = \mathbf{X}\mathbf{J}$ such that $\mathbf{Z} = \mathbf{X}^*\mathbf{C}\mathbf{X}$ has the form* (6.3).

Proof. Application of the procedure (6.7) – (6.15) to the blocks $\hat{\mathbf{C}}_i$ and $\check{\mathbf{C}}_i$ of the matrix $\mathbf{C}$ according to (6.5) produces the blocks $\hat{\mathbf{X}}_i$ and $\check{\mathbf{X}}_i$ of the matrix $\mathbf{X}$. These blocks commute with the blocks $\hat{\mathbf{J}}_i$ and $\check{\mathbf{J}}_i$ of a matrix $\mathbf{J}$ according to (6.3). Furthermore, the blocks $\hat{\mathbf{Z}}_i = \hat{\mathbf{X}}_i^*\hat{\mathbf{C}}_i\hat{\mathbf{X}}_i$ and $\check{\mathbf{Z}}_i = \check{\mathbf{X}}_i^*\check{\mathbf{C}}_i\check{\mathbf{X}}_i$ of the matrix $\mathbf{Z}$ take on the asserted form. □

This normalisation procedure seems to be difficult on first sight, but it is in fact based on a simple relationship. To explain this, let $n \geq m$ and let $\mathbf{B} \in \mathbb{C}^{n\times m}$ be a matrix with full column rank. Then $\mathbf{B}$ admits the well-known and uniquely determined QR decomposition

$$\mathbf{B} = \mathbf{Q}\mathbf{R}, \;\; \mathbf{Q} \in \mathbb{C}^{n\times m}, \;\; \mathbf{R} \in \mathbb{C}^{m\times m}, \;\; \mathbf{R} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1m} \\ & \ddots & \vdots \\ & & \rho_{mm} \end{bmatrix},$$

where $\mathbf{Q}$ is a matrix with orthogonal columns $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_m$ and $\mathbf{R}$ is an upper triangular matrix with positive diagonal elements [ST1, Kapitel 4.7]. On the

other hand, the matrix $\mathbf{A} = \mathbf{B}^*\mathbf{B} = \mathbf{A}^* \in \mathbb{C}^{m\times m}$ is positive definite, so that it admits the uniquely determined Cholesky decomposition

$$\mathbf{A} = \mathbf{L}\mathbf{L}^*, \;\; \mathbf{L} \in \mathbb{C}^{m\times m}, \;\; \mathbf{L} = \begin{bmatrix} \lambda_{11} & & \\ \vdots & \ddots & \\ \lambda_{m1} & \cdots & \lambda_{mm} \end{bmatrix},$$

where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements [ST1, Kapitel 4.3]. Since $\mathbf{L}\mathbf{L}^* = \mathbf{A} = \mathbf{B}^*\mathbf{B} = \mathbf{R}^*\mathbf{Q}^*\mathbf{Q}\mathbf{R} = \mathbf{R}^*\mathbf{R}$ and both decompositions are unique, it must be true that $\mathbf{L}^* = \mathbf{R}$. Thus, the orthonormalisation of the columns of $\mathbf{B}$ can also be obtained via a Cholesky decomposition of the matrix $\mathbf{A}$.

Now, let the matrix $\mathbf{R}$ from Theorem 6.2 be partitioned in

$$\mathbf{R} = \begin{bmatrix} \hat{\mathbf{R}}_1 & \cdots & \hat{\mathbf{R}}_r & \check{\mathbf{R}}_{r+1} & \cdots & \check{\mathbf{R}}_s \end{bmatrix}.$$

Then the matrix $\mathbf{C}$ consists of the blocks $\hat{\mathbf{C}}_i = \hat{\mathbf{R}}_i^*\mathbf{H}\hat{\mathbf{R}}_i$ and $\check{\mathbf{C}}_i = \check{\mathbf{R}}_i^*\mathbf{H}\check{\mathbf{R}}_i$. Therefore, the transformations $\hat{\mathbf{X}}_i^*\hat{\mathbf{C}}_i\hat{\mathbf{X}}_i = \hat{\mathbf{Z}}_i$ and $\check{\mathbf{X}}_i^*\check{\mathbf{C}}_i\check{\mathbf{X}}_i = \check{\mathbf{Z}}_i$ are merely the H-orthogonalisation of the blocks of $\mathbf{R}$ via generalised (inverse) Cholesky decompositions of the blocks of $\mathbf{C}$. Here the following correspondences exist

$$\mathbf{B} \leftrightarrow \hat{\mathbf{R}}_i, \check{\mathbf{R}}_i, \;\; \mathbf{I}_n \leftrightarrow \mathbf{H}, \;\; \mathbf{A} \leftrightarrow \hat{\mathbf{C}}_i, \check{\mathbf{C}}_i, \;\; \mathbf{L} \leftrightarrow \hat{\mathbf{X}}_i^{-*}, \check{\mathbf{X}}_i^{-*}, \;\; \mathbf{I}_m \leftrightarrow \hat{\mathbf{Z}}_i, \check{\mathbf{Z}}_i.$$

Moreover, the normalisation procedure is a generalisation of the HQR and HQR-2 decomposition (Algorithm 2.12 and Algorithm 2.13), in which the relationship to the Cholesky decomposition is more obvious.

6.3 Description of the algorithm

This section describes the overall algorithm (`CNF`) which has been implemented in ANSI Fortran 77 using the `DOUBLE COMPLEX` versions of `LAPACK` and the `BLAS` [LUG]. The description contains both, the mathematical theory of the steps performed as well as the most important details on the actual Fortran implementation. Throughout the section $(\mathbf{A}, \mathbf{H})$ is assumed to be a matrix pair, where $\mathbf{A}$ is H-Hermitian and $\mathbf{H}$ is nonsingular and Hermitian.

The two major tasks of the algorithm are the computation of the Jordan normal form of $\mathbf{A}$ (`JNF`) encapsulated in the subroutines `ZGEES`, `ZTRGRP`, `ZTRBLK`, `ZTRDFL`, `ZTRJNF`, and the subsequent normalisation of the eigenspaces encapsulated in the subroutine `ZSPCNF`[11]. Whereas the theory of the latter has essentially been developed in Section 6.2, the theory for `JNF` is mostly taken over from [KR1]. Nevertheless, we have modified some steps contained in the original implementation which will be referred to as the 560 algorithm [KR2]. The individual steps of `CNF` are:

[11] The naming scheme has been assimilated to `LAPACK`. The first letter `Z` indicates double complex data type, and the matrix types `GE`, `TR`, `SP` denote general, triangular and sip block matrices.

Step 1. Computing the Schur decomposition (ZGEES) The matrix $\mathbf{A}$ is transformed into upper triangular form using the Schur decomposition

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{T}_1 = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ & \lambda_2 & \cdots & * \\ & & \ddots & \vdots \\ & & & \lambda_n \end{bmatrix}$$

computed with the LAPACK routine ZGEES. Here $\mathbf{Q}$ is unitary and λ_i, $1 \leq 1 \leq n$, are the eigenvalues of $\mathbf{A}$. This well-known transformation is based on the QR algorithm which is, for example, described in [GVL, Algorithm 7.5.2]. In the 560 algorithm, the LR algorithm is used to obtain $\mathbf{T}_1$.

Step 2. Grouping the eigenvalues (ZTRGRP) The grouping of the eigenvalues is the most sensitive step in the overall algorithm. Kågström and Ruhe use for this purpose eigenvalue estimates based on Gershgorin circles. However, in practical application of the 560 algorithm it was found that these estimates strongly depend on the tolerance parameter EIN, whose satisfactory choice is not always easy. Our approach tries to avoid this difficulty by grouping the eigenvalues with the help of a hierarchical clustering algorithm [A, Chapter 6] which has been modified by introducing a maximum cluster distance.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be given points in $\mathbb{R}^m$ and let d be a real distance function defined for two sets of points in $\mathbb{R}^m$. Moreover, let $p_{\min} \in \{1, \ldots, n\}$ and $\delta_{\max} \in [0, \infty)$ be parameters. Then the points can be partitioned into clusters by applying the following algorithm.

$C_i = \{\mathbf{x}_i\},\ i = 1 \ldots n$
$p = n$
$\delta_0 = 0$
forever do
    $\delta_1 = \infty$
    for $1 \leq i < j \leq p$ **do**
        if $d(C_i, C_j) < \delta_1$ **then**
            $\delta_1 = d(C_i, C_j),\ r = i,\ s = j$
        end if
    end for
    if $p < p_{\min}$ **or** $\delta_1 > \delta_{\max}$ **then**
        break
    end if
    $C_r = C_r \cup C_s$
    $C_i = C_{i+1},\ i = s \ldots p-1$
    $p = p - 1$
    $\delta_0 = \delta_1$
end forever

The number of clusters finally generated p satisfies $p \geq p_{\min}$. Moreover, the maximum distance of the combined clusters δ_0 and the minimum distance of the remaining clusters δ_1 fulfil $\delta_0 \leq \delta_{\max}$ as well as $\delta_0 \leq \delta_1$. Thereby one of the

functions

$$d_{\min}(X,Y) = \min \|\mathbf{x}_i - \mathbf{y}_j\|_2, \qquad d_{\max}(X,Y) = \max \|\mathbf{x}_i - \mathbf{y}_j\|_2,$$
$$d_{\text{avg}}(X,Y) = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s} \|\mathbf{x}_i - \mathbf{y}_j\|_2, \quad d_{\text{mean}}(X,Y) = \left\| \frac{1}{r}\sum_{i=1}^{r}\mathbf{x}_i - \frac{1}{s}\sum_{j=1}^{s}\mathbf{y}_j \right\|_2,$$

where $\mathbf{x}_i \in X$, $1 \le i \le r$, and $\mathbf{y}_j \in Y$, $1 \le j \le s$, may be used for the distance determination.

A further distance function can be defined as follows: Let $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_r\}$ be a set of points $\mathbf{z}_j = (z_{i,j}) \in \mathbb{R}^m$ $(1 \le i \le m,\ 1 \le j \le r)$. Then the vectors $\mu = (\mu_i)$ and $\sigma^2 = (\sigma_i^2)$ having the components

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} z_{i,j} \ \text{ and } \ \sigma_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(z_{i,j} - \mu_i)^2 \ \ (\sigma_i^2 = 0 \text{ if } r = 1)$$

contain the (empirical) averages and variances of the points. Therefore, the non-negative scalar

$$\sigma^2(Z) = \sum_{i=1}^{m} \sigma_i^2$$

gives a total scatter measure for the set Z, and the function

$$d_{\text{var}}(X,Y) = \sigma^2(X \cup Y)$$

defines a distance of some sets X and Y. This measure does not seem to be known in the literature. Nevertheless, it has turned out to be particularly useful for grouping the eigenvalues $\lambda_1, \ldots, \lambda_n \in \sigma(\mathbf{A})$ which are considered as points in $\mathbb{R}^2$ for this purpose.

The implementation in the routine `ZTRGRP` allows to specify the distance measure to be used and to control the grouping by an expected number of eigenvalues $p_{\min}$ and/or a maximum grouping tolerance $\delta_{\max}$, respectively. It returns the number of groups (clusters) finally generated p, the assignments of the eigenvalues to the groups, and the distances δ_0 and δ_1.

Step 3. Sorting the eigenvalues (`ZTRBLK`, `JOB='S'`) The matrix $\mathbf{T}_1$ is unitarily transformed such that eigenvalues belonging to the same group are adjacent

$$\mathbf{U}^*\mathbf{T}_1\mathbf{U} = \mathbf{T}_2 = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \cdots & \mathbf{T}_{1p} \\ & \mathbf{T}_{22} & \cdots & \mathbf{T}_{2p} \\ & & \ddots & \vdots \\ & & & \mathbf{T}_{pp} \end{bmatrix}.$$

Now each upper triangular $\mathbf{T}_{kk}$ block contains the t_k eigenvalues which have been assigned to cluster C_k for $1 \le k \le p$.

This transformation is based on a sequence of eigenvalue interchanges which is (for the case $\mathbb{F} = \mathbb{R}$) described in [GVL, Algorithm 7.6.1]. The implementation in the routine `ZTRBLK` uses the `LAPACK` routine `ZTREXC` for this purpose. It returns the number of diagonal blocks p in the variable `LBLOCK` and the block boundaries $\sum_{i=1}^{k} t_i$ for $1 \le k \le p$ in the array `BLOCK`.

Step 4. Computing the block diagonal form (ZTRBLK, JOB='D') The matrix $\mathbf{T}_2$ is transformed into upper triangular block diagonal form

$$\mathbf{Y}^{-1}\mathbf{T}_2\mathbf{Y} = \mathbf{T}_3 = \mathbf{T}_{11} \oplus \ldots \oplus \mathbf{T}_{pp}$$

using the matrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{I} & \mathbf{Y}_{12} & \cdots & \mathbf{Y}_{1p} \\ & \mathbf{I} & \cdots & \mathbf{Y}_{2p} \\ & & \ddots & \vdots \\ & & & \mathbf{I} \end{bmatrix}$$

which is obtained by subsequent solutions of Sylvester equations as is described in [GVL, Algorithm 7.6.3]. The norms of the spectral projectors

$$\mathbf{P}_k = \begin{bmatrix} \mathbf{Y}_{1,k} \\ \vdots \\ \mathbf{Y}_{k-1,k} \\ \mathbf{I} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I} & (\mathbf{Y}^{-1})_{k,k+1} & \cdots & (\mathbf{Y}^{-1})_{k,p} \end{bmatrix}$$

satisfy the estimate

$$\|\mathbf{P}_k\|_2 \leq \left(1 + \sqrt{\sum_{i=1}^{k-1} \|\mathbf{Y}_{ik}\|_F^2}\right) \left(1 + \sqrt{\sum_{j=k+1}^{p} \|(\mathbf{Y}^{-1})_{kj}\|_F^2}\right), \tag{6.16}$$

where $(\mathbf{Y}^{-1})_{kj}$ stands for the corresponding block of $\mathbf{Y}^{-1}$. These norms are relevant condition numbers for the associated groups of eigenvalues.

The implementation in the routine ZTRBLK uses the LAPACK routine ZTRSYL to solve the Sylvester equations. It returns the reciprocal right hand sides of (6.16) in the array COND.

Step 5. Computing unitary bases of the invariant subspaces (ZTRDFL, JOB='O') Let $\mathbf{S} = \mathbf{QUY}$ be the accumulated transformation matrix as computed in the foregoing steps. Moreover, let $\mathbf{S}_k = \mathbf{Q}'_k\mathbf{R}_k$, $1 \leq k \leq p$, be QR decompositions of the $n \times t_k$ blocks corresponding to the $\mathbf{T}_{kk}$ blocks. Then $\mathbf{Q}' = \mathbf{S}(\mathbf{R}_1 \oplus \ldots \oplus \mathbf{R}_p)^{-1}$ consists of the unitary $\mathbf{Q}'_k$ blocks and

$$(\mathbf{Q}')^{-1}\mathbf{A}(\mathbf{Q}') = \mathbf{T}'_{11} \oplus \ldots \oplus \mathbf{T}'_{pp},$$

where $\mathbf{T}'_{kk} = \mathbf{R}_k\mathbf{T}_{kk}\mathbf{R}_k^{-1}$ is still upper triangular. It must be considered here, that the $\mathbf{S}_1$ block is already unitary, so that its QR decomposition consists of the factors $\mathbf{Q}'_1 = \mathbf{S}_1$ and $\mathbf{R}_1 = \mathbf{I}$.

After this orthogonalisation process the bases of the invariant subspaces formed by the columns of the $\mathbf{Q}'_k$ blocks are unitary. In the 560 algorithm, the modified Gram-Schmidt method is used to perform this transformation.

Step 6. Computing the block structure (ZTRDFL, JOB='D') Each $\mathbf{T}'_{kk}$ block is unitarily transformed such that

$$\mathbf{W}_k^*(\mathbf{T}'_{kk} - \lambda_k^*\mathbf{I})\mathbf{W}_k = \mathbf{E}_k + \mathbf{B}_k, \tag{6.17a}$$

where $\lambda_k^* = \operatorname{tr}(\mathbf{T}_{kk}) \,/\, t_k$ is the average eigenvalue and the blocks have the form

$$\mathbf{E}_k = \begin{bmatrix} \mathbf{E}_{11} & & & \\ \mathbf{E}_{21} & \mathbf{E}_{22} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{E}_{q1} & \mathbf{E}_{q2} & \cdots & \mathbf{E}_{qq} \end{bmatrix}, \quad \mathbf{B}_k = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1q} \\ & \mathbf{0} & \cdots & \mathbf{B}_{2q} \\ & & \ddots & \vdots \\ & & & \mathbf{0} \end{bmatrix}. \tag{6.17b}$$

The matrices $\mathbf{B}_k$ are nilpotent, the matrices $\mathbf{E}_k$ have a negligible Frobenius norm $\|\mathbf{E}_k\|_F \approx 0$, and the sizes of the square diagonal blocks furthermore satisfy $m_1 \geq \ldots \geq m_q$. This transformation is based on the following algorithm:

Let $\mathbf{A} = \mathbf{A}_{11} \in \mathbb{C}^{t\times t}$ and let

$$\mathbf{A}_{11} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$$

be a singular value decomposition, where the singular values are sorted in ascending order $\sigma_1^{(1)} \leq \ldots \leq \sigma_t^{(1)}$. Moreover, let $\tau > 0$ be a constant and let m_1 be an index such that $\sigma_{m_1}^{(1)} \leq \tau < \sigma_{m_1+1}^{(1)}$. Then the singular values are just the Euclidean column norms of

$$\mathbf{V}_1^* \mathbf{A}_{11} \mathbf{V}_1 = \mathbf{V}_1^* \mathbf{U}_1 \mathbf{\Sigma}_1,$$

so that the partitioning

$$\mathbf{V}_1^* \mathbf{A}_{11} \mathbf{V}_1 = \begin{bmatrix} \mathbf{E}_{11} & \mathbf{B}_{12}^{(1)} \\ \mathbf{E}_{21}^{(1)} & \mathbf{A}_{22} \end{bmatrix} \text{ with } \mathbf{E}_{11} \in \mathbb{C}^{m_1\times m_1},\ \eta_1^2 = \left\| \begin{matrix} \mathbf{E}_{11} \\ \mathbf{E}_{21}^{(1)} \end{matrix} \right\|_F^2 = \sum_{i=1}^{m_1} \left[\sigma_i^{(1)}\right]^2$$

can be made. The same procedure applied to $\mathbf{A}_{22} \in \mathbb{C}^{(t-m_1)\times(t-m_1)}$ yields

$$\mathbf{V}_2^* \mathbf{A}_{22} \mathbf{V}_2 = \begin{bmatrix} \mathbf{E}_{22} & \mathbf{B}_{23}^{(2)} \\ \mathbf{E}_{32}^{(2)} & \mathbf{A}_{33} \end{bmatrix} \text{ with } \mathbf{E}_{22} \in \mathbb{C}^{m_2\times m_2},\ \eta_2^2 = \left\| \begin{matrix} \mathbf{E}_{22} \\ \mathbf{E}_{32}^{(2)} \end{matrix} \right\|_F^2 = \sum_{i=1}^{m_2} \left[\sigma_i^{(2)}\right]^2.$$

Hence, for

$$\mathbf{W}_2 = \mathbf{V}_1(\mathbf{I}_{m_1} \oplus \mathbf{V}_2) \text{ and } \begin{bmatrix} \mathbf{E}_{21}^{(2)} \\ \mathbf{E}_{31}^{(2)} \end{bmatrix} = \mathbf{V}_2^* \mathbf{E}_{21}^{(1)},\ \begin{bmatrix} \mathbf{B}_{12}^{(2)} & \mathbf{B}_{13}^{(2)} \end{bmatrix} = \mathbf{B}_{12}^{(1)} \mathbf{V}_2$$

we obtain

$$\mathbf{W}_2^* \mathbf{A} \mathbf{W}_2 = \begin{bmatrix} \mathbf{E}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{E}_{21}^{(2)} & \mathbf{E}_{22} & \mathbf{0} \\ \mathbf{E}_{31}^{(2)} & \mathbf{E}_{32}^{(2)} & \mathbf{A}_{33} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12}^{(2)} & \mathbf{B}_{13}^{(2)} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{23}^{(2)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{E}_2 + \mathbf{B}_2,$$

where $\|\mathbf{E}_2\|_F^2 = \eta_1^2 + \eta_2^2 + \|\mathbf{A}_{33}\|_F^2$. The iteration can be continued until in the q-th step either all $m_q = t - m_1 - \ldots - m_{q-1}$ singular values are smaller or all are greater than the specified constant τ. Then

$$\mathbf{V}_q^* \mathbf{A}_{qq} \mathbf{V}_q = \begin{bmatrix} \mathbf{E}_{qq} \end{bmatrix} \text{ with } \mathbf{E}_{qq} \in \mathbb{C}^{m_q\times m_q},\ \eta_q^2 = \|\mathbf{E}_{qq}\|_F^2 = \sum_{i=1}^{m_q} \left[\sigma_i^{(q)}\right]^2.$$

Hence, for $\mathbf{W} = \mathbf{V}_1(\mathbf{I}_{m_1} \oplus \mathbf{V}_2) \ldots (\mathbf{I}_{m_1+\ldots+m_{q-1}} \oplus \mathbf{V}_q)$ we finally obtain

$$\mathbf{W}^* \mathbf{A} \mathbf{W} = \mathbf{E} + \mathbf{B},$$

where $\mathbf{E}$ and $\mathbf{B}$ have the forms as in (6.17). Moreover, if $\sigma_i^{(q)} \leq \tau$, then $\|\mathbf{E}\|_F^2 = \eta_1^2 + \ldots + \eta_q^2 \leq (m_1 + \ldots + m_q)\tau^2 = t\tau^2$.

In the application of the algorithm for computing (6.17) it is important that relations of the kind

$$\sigma_{m_h}^{(h)} \leq \tau_1 \ll \tau_2 < \sigma_{m_h+1}^{(h)}, \quad 1 \leq h \leq q,$$

hold at the splitting points. Namely, the coupling elements $\beta_{h,j}$ computed in the next step from the blocks $\mathbf{B}_{h,h+1}$ satisfy the estimate

$$|\beta_{h,j}| > \sigma_{m_h+1}^{(h)} > \tau_2, \quad 1 \leq j \leq m_{h+1},$$

and these elements must not become too small [KR1]. Therefore, the implementation in the routine `ZTRDFL` first of all determines the index m_h with $\sigma_{m_h}^{(h)} \leq \tau_1 = \texttt{TOL}$ and checks the validity of the relation $\sigma_{m_h+1}^{(h)} > \tau_2 = 10^3\tau_1$ thereafter. If this check fails, m_h is incremented by one. Furthermore, at least one deflation step is performed, so that in the case $\sigma_1^{(h)} > \tau_1$ the index $m_h = 1$ is chosen. The relation of the constants $\tau_2 = 10^3\tau_1$ is thereby not compelling, so that other factors were possible, too. The values of $\|\mathbf{E}_k\|_F$ are stored in the array `DELE` to provide information on the deflation process. They should not significantly exceed the optimum value $\|\mathbf{E}_k\|_F \leq \sqrt{t_k} \cdot \texttt{TOL}$.

After this step, the structure of the diagonal blocks can be represented by the block sizes

$$\underbrace{\begin{matrix}\tilde{m}_1 & & \tilde{m}_{q_1} \\ m_1^{(1)} & \ldots, & m_{q_1}^{(1)}\end{matrix}}_{t_1} \quad \ldots, \quad \underbrace{\begin{matrix}\tilde{m}_{q_1+\ldots+q_{p-1}+1} & & \tilde{m}_{q_1+\ldots+q_p} \\ m_1^{(p)} & \ldots, & m_{q_p}^{(p)}\end{matrix}}_{t_p},$$

where $\tilde{m}_j$ is merely a consecutive numbering scheme for $m_h^{(k)}$. The routine `ZTRDFL` stores $\sum_{i=1}^{k} q_i$ for $1 \leq k \leq p$ in the array `DBLK` and $\sum_{j=1}^{h} \tilde{m}_j$ for $1 \leq h \leq q_1 + \ldots + q_p$ in the array `DEFL`. Thus, the block structure is represented such that $\texttt{BLOCK}(k) = \texttt{DEFL}(\texttt{DBLK}(k))$ for $1 \leq k \leq p$ (see Step 3).

Whereas in the 560 algorithm the blocks $\mathbf{T}'_{kk}$ are overwritten with $\mathbf{B}_k + \lambda_k^*\mathbf{I} + (\mathbf{E}_{11} \oplus \ldots \oplus \mathbf{E}_{pp})$, the routine `ZTRDFL` overwrites these blocks with $\mathbf{B}_k + \lambda_k^*\mathbf{I}$. This has no effect on the following transformations, but allows a more efficient implementation.

Step 7. Computing the coupling elements (`ZTRJNF`, `JOB='C'`) Each $\mathbf{B}_k + \lambda_k^*\mathbf{I}$ block is transformed such that

$$\mathbf{M}_k^{-1}\mathbf{B}_k\mathbf{M}_k = \mathbf{B}'_k, \tag{6.18a}$$

where the blocks have the form

$$\mathbf{B}'_k = \begin{bmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{12} & & \\ & \mathbf{0} & \ddots & \\ & & \ddots & \boldsymbol{\Sigma}_{q-1,q} \\ & & & \mathbf{0} \end{bmatrix} \quad \text{with} \quad \boldsymbol{\Sigma}_{h,h+1} = \begin{bmatrix} \beta_{h,1} \oplus \ldots \oplus \beta_{h,m_{h+1}} \\ \mathbf{0} \end{bmatrix}. \tag{6.18b}$$

The non-zero real elements β_{hj} are called the coupling elements. Their determination is based on the following algorithm which is sufficiently explained on the example

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12} & \mathbf{B}_{13} & \mathbf{B}_{14} \\ & \mathbf{0} & \mathbf{B}_{23} & \mathbf{B}_{24} \\ & & \mathbf{0} & \mathbf{B}_{34} \\ & & & \mathbf{0} \end{bmatrix} = \mathbf{B}^{(0)}.$$

Let $\mathbf{B}_{34}^{(0)} = \mathbf{U}_3 \mathbf{\Sigma}_{34} \mathbf{V}_4^*$ be a singular value decomposition, and let $\mathbf{M}_{34} = \mathbf{I} \oplus \mathbf{I} \oplus \mathbf{U}_3 \oplus \mathbf{V}_4$. Then

$$\mathbf{M}_{34}^* \mathbf{B}^{(0)} \mathbf{M}_{34} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12}^{(0)} & \mathbf{B}_{13}^{(0)} \mathbf{U}_3 & \mathbf{B}_{14}^{(0)} \mathbf{V}_4 \\ & \mathbf{0} & \mathbf{B}_{23}^{(0)} \mathbf{U}_3 & \mathbf{B}_{24}^{(0)} \mathbf{V}_4 \\ & & \mathbf{0} & \mathbf{U}_3^* \mathbf{B}_{34}^{(0)} \mathbf{V}_4 \\ & & & \mathbf{0} \end{bmatrix} = \mathbf{B}^{(1)}.$$

Now using Gauss eliminations of the form

$$\mathbf{G}_{ij}^k = \mathbf{I} + \frac{b_{ij}}{b_{kj}} \, \mathbf{e}_i \mathbf{e}_k^T, \tag{6.19}$$

where $\mathbf{e}_i$ denotes the canonical basis of $\mathbb{C}^{t_k}$, the elements b_{ij} above the diagonal elements b_{kj} of the $\mathbf{B}_{34}^{(1)} = \mathbf{\Sigma}_{34}$ block can subsequently be eliminated which is all in all described by

$$\mathbf{G}_{34}^{-1} \mathbf{B}^{(1)} \mathbf{G}_{34} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12}^{(1)} & \mathbf{B}_{13}^{(1)} & \mathbf{0} \\ & \mathbf{0} & \mathbf{B}_{23}^{(1)} & \mathbf{0} \\ & & \mathbf{0} & \mathbf{\Sigma}_{34} \\ & & & \mathbf{0} \end{bmatrix} = \mathbf{B}^{(2)}.$$

Let $\mathbf{B}_{23}^{(2)} = \mathbf{Q}_2 \mathbf{R}_{23}$ be a QR decomposition, and let $\mathbf{M}_{23} = \mathbf{I} \oplus \mathbf{Q}_2 \oplus \mathbf{I} \oplus \mathbf{I}$. Then

$$\mathbf{M}_{23}^* \mathbf{B}^{(2)} \mathbf{M}_{23} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12}^{(2)} \mathbf{Q}_2 & \mathbf{B}_{13}^{(2)} & \mathbf{0} \\ & \mathbf{0} & \mathbf{Q}_2^* \mathbf{B}_{23}^{(2)} & \mathbf{0} \\ & & \mathbf{0} & \mathbf{\Sigma}_{34} \\ & & & \mathbf{0} \end{bmatrix} = \mathbf{B}^{(3)}.$$

Again using Gauss eliminations of the form (6.19), the elements above the diagonal of the $\mathbf{B}_{23}^{(3)} = \mathbf{R}_{23}$ block can subsequently be eliminated which is all in all described by

$$\mathbf{G}_{23}^{-1} \mathbf{B}^{(3)} \mathbf{G}_{23} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12}^{(3)} & \mathbf{0} & \mathbf{0} \\ & \mathbf{0} & \mathbf{\Sigma}_{23} & \mathbf{0} \\ & & \mathbf{0} & \mathbf{\Sigma}_{34} \\ & & & \mathbf{0} \end{bmatrix} = \mathbf{B}^{(4)}.$$

In the same manner the $\mathbf{B}_{12}^{(4)}$ block can be transformed, so that for $\mathbf{M} = \mathbf{M}_{34} \mathbf{G}_{34} \ldots \mathbf{M}_{12} \mathbf{G}_{12}$ finally (6.18) holds.

This algorithm differs from the one given by Kågström and Ruhe in that it uses a singular value decomposition for the first elimination step. It has the advantage that if the $\mathbf{B}_k$ block only contains a $\mathbf{B}_{12}$ block, the whole step can be performed with a unitary transformation. In contrast to the 560 algorithm, where the complete process is carried out with Gauss eliminations, the implementation in the routine `ZTRJNF` uses the algorithm described here.

Step 8. Permuting the coupling elements (`ZTRJNF`, `JOB='P'`) Each $\mathbf{B}'_k + \lambda_k^*\mathbf{I}$ block is permuted such that

$$\mathbf{P}_k^*(\mathbf{B}'_k + \lambda_k^*\mathbf{I})\mathbf{P}_k = \mathbf{J}'_k = \mathbf{J}'_{r_1}(\lambda_k^*) \oplus \ldots \oplus \mathbf{J}'_{r_m}(\lambda_k^*),$$

where $\mathbf{J}'_{r_s}(\lambda_k^*)$ is a not yet normalised Jordan block of the form

$$\mathbf{J}'_{r_s}(\lambda_k^*) = \begin{bmatrix} \lambda_k^* & \beta_{1,s} & & \\ & \lambda_k^* & \ddots & \\ & & \ddots & \beta_{r_s-1,s} \\ & & & \lambda_k^* \end{bmatrix}$$

with $r_s = \max\{h \mid m_h \geq s\}$ for $1 \leq s \leq m = m_1$. The permutation matrix $\mathbf{P}_k$ required for this step is defined as follows: Let the canonical basis of $\mathbb{C}^{t_k}$ be partitioned in

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_{\mu_1}\} \cup \{\mathbf{e}_{\mu_1+1}, \ldots, \mathbf{e}_{\mu_2}\} \cup \ldots \cup \{\mathbf{e}_{\mu_{q-1}+1}, \ldots, \mathbf{e}_{\mu_q}\},$$

where $\mu_h = \sum_{i=1}^{h} m_i$. Now, taking from each of the sets the first vector and inserting it as column of $\mathbf{P}_k$, then taking the second vectors and continuing the process until all vectors have been taken, we obtain

$$\mathbf{P}_k = [\, \mathbf{e}_1\, \mathbf{e}_{\mu_1+1}\, \ldots\, \mathbf{e}_{\mu_{q-1}+1}\, ;\, \ldots\, ;\, \mathbf{e}_m\, \mathbf{e}_{\mu_1+m}\, \ldots\, \mathbf{e}_{\mu_{r_m-1}+m}\,].$$

This matrix contains $m = m_1 = \mu_1$ groups of basis vectors corresponding to the Jordan blocks. Inspecting the s-th group reveals that $r_s = \max\{h \mid \mu_{h-1} + s \leq \mu_h\} = \max\{h \mid s \leq m_h\}$ which explains the block sizes already specified above.

The implementation in the routine `ZTRJNF` updates the arrays `DBLK` and `DEFL` such that they describe the Jordan structure analogously to Step 6. This and all subsequent steps are not contained in the 560 algorithm.

Step 9. Normalising the coupling elements (`ZTRJNF`, `JOB='D'`) Each $\mathbf{J}'_k$ block is transformed such that

$$\mathbf{D}_k\mathbf{J}'_k\mathbf{D}_k^{-1} = \mathbf{J}_k = \mathbf{J}_{r_1}(\lambda_k^*) \oplus \ldots \oplus \mathbf{J}_{r_m}(\lambda_k^*),$$

where $\mathbf{J}_{r_s}(\lambda_k^*)$ is a normalised Jordan block for $1 \leq s \leq m$. The diagonal matrix $\mathbf{D}_k$ required for this step is defined by

$$\begin{aligned} &\mathbf{D}_k = \boldsymbol{\Delta}_1 \oplus \ldots \oplus \boldsymbol{\Delta}_m \;\text{ with} \\ &\boldsymbol{\Delta}_s = 1 \oplus \beta_{1,s} \oplus \beta_{1,s}\beta_{2,s} \oplus \ldots \oplus \beta_{1,s}\beta_{2,s}\ldots\beta_{r_s-1,s}. \end{aligned}$$

Step 10. Pairing the eigenvalues (`ZSPCNF`, `JOB='P'`) Accumulating all transformations made in the foregoing steps we obtain the matrix

$$\mathbf{R} = \mathbf{QUY}\Big(\bigoplus_{k=1}^{p} \mathbf{R}_k^{-1}\mathbf{W}_k\mathbf{M}_k\mathbf{P}_k\mathbf{D}_k^{-1}\Big).$$

Now, according to Theorem 6.1, the matrices $\mathbf{J} = \mathbf{R}^{-1}\mathbf{AR}$ and $\mathbf{C} = \mathbf{R}^*\mathbf{HR}$ must theoretically have the form

$$\mathbf{J} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{J}}(\lambda_i^*)\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{J}}(\lambda_i^*)\right], \quad \mathbf{C} = \left[\bigoplus_{i=1}^{r} \hat{\mathbf{C}}_i\right] \oplus \left[\bigoplus_{i=r+1}^{s} \check{\mathbf{C}}_i\right]$$

up to a permutation of the blocks. Here the notation of Section 6.2 is used, so that $\lambda_1^*, \ldots, \lambda_r^*$ are the real and $\lambda_{r+1}^*, \ldots, \lambda_s^*$ are the non-real eigenvalues lying in the open upper complex half-plane. In order to classify the blocks of

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \ldots & \mathbf{C}_{1p} \\ \vdots & & \vdots \\ \mathbf{C}_{p1} & \ldots & \mathbf{C}_{pp} \end{bmatrix} \quad (p = 2s - r)$$

accordingly, let $\pi(k)$ be indices such that

$$\|\mathbf{C}_{k,\pi(k)}\|_F = \max_{1 \le j \le p} \|\mathbf{C}_{kj}\|_F$$

for $1 \le k \le p$. Now, if $\pi(k) = k$, the $\mathbf{C}_{kk}$ block is classified as a $\hat{\mathbf{C}}_i$ block. If $\pi(k) = l$ and $\pi(l) = k$ and if the Jordan structures of the corresponding eigenvalues are equal, the $\mathbf{C}_{kl}$ and $\mathbf{C}_{lk}$ block are classified to belong to a $\check{\mathbf{C}}_i$ block. Moreover, the Frobenius norms of the matrices

$$\mathbf{C}_k = [\mathbf{C}_{k1} \ldots \mathbf{C}_{k,\pi(k)-1}\, \mathbf{C}_{k,\pi(k)+1} \ldots \mathbf{C}_{kp}]$$

are used to estimate the H-orthogonality of the eigenspace $E_A(\lambda_k^*)$ spanned by the columns of $\mathbf{R}_k$ to its H-orthogonal companion $E_A(\lambda_k^*)^{[\perp]}$ spanned by the columns of $\mathbf{R}$ without the $\mathbf{R}_{\pi(k)}$ block. (For the definition and properties of the H-orthogonal companion see Section 2.2.)

The implementation in the routine `ZSPCNF` returns the indices $\pi(k)$ in the array `PAIR` and the norms $\|\mathbf{C}_k\|_F$ in the array `ORTH`. It returns an error, if the classification of the $\mathbf{C}_{kl}$ blocks fails.

Step 11. Normalising the eigenspaces (`ZSPCNF`, `JOB='N'`) The $\mathbf{C}_{kk}$ blocks and pairs of $\mathbf{C}_{kl}$ and $\mathbf{C}_{lk}$ blocks are normalised such that

$$\mathbf{X}_k^* \mathbf{C}_{kk} \mathbf{X}_k = \mathbf{Z}_{kk} = \varepsilon_1 \mathbf{Z}_{r_1} \oplus \ldots \oplus \varepsilon_m \mathbf{Z}_{r_m} \tag{6.20a}$$

and

$$\begin{bmatrix} \mathbf{X}_k & \\ & \mathbf{X}_l \end{bmatrix}^* \begin{bmatrix} & \mathbf{C}_{kl} \\ \mathbf{C}_{lk} & \end{bmatrix} \begin{bmatrix} \mathbf{X}_k & \\ & \mathbf{X}_l \end{bmatrix} = \begin{bmatrix} & \mathbf{Z}_{kl} \\ \mathbf{Z}_{lk} & \end{bmatrix} \tag{6.20b}$$
$$\text{with } \mathbf{Z}_{kl} = \mathbf{Z}_{lk} = \mathbf{Z}_{r_1} \oplus \ldots \oplus \mathbf{Z}_{r_m},$$

where $\mathbf{X}_k$ and $\mathbf{X}_k \oplus \mathbf{X}_l$ commute with $\mathbf{J}_k$ or $\mathbf{J}_k \oplus \mathbf{J}_l$, respectively. Thereafter the matrix $\mathbf{S} = \mathbf{R}(\mathbf{X}_1 \oplus \ldots \oplus \mathbf{X}_p)$ transforms the pair $(\mathbf{A}, \mathbf{H})$ to its canonical form.

The implementation in the routine `ZSPCNF` uses the subroutine `ZSPEVR` to compute (6.20a) and the subroutine `ZSPEVC` to compute (6.20b). These routines implement the algorithm described in Section 6.2 for the cases $\mathbf{C}^* = \mathbf{C}$ and $\mathbf{C}^* \ne \mathbf{C}$, respectively. `ZSPCNF` returns the signs ε_s in the array `SIGN`.

In addition to the routines performing the individual steps, the driver routine `ZHHCNF`[12] has been implemented for convenience. This routine expects a matrix pair $(\mathbf{A}, \mathbf{H})$ and computes its canonical form by subsequent calls to the routines described above. Thereby the parameter `STEP` allows to specify which of the steps are to be performed. If `STEP` is set to at most 9, the routine may also be used for computing the Jordan normal form of an arbitrary matrix $\mathbf{A}$.

[12] `HH` stands for H-Hermitian.

6.4 Numerical results

In this section some numerical results of the routine ZHHCNF are presented. In order to make the data more comprehensible we start with a brief discussion of the parameters and control variables of the algorithm (see also [KR1]).

6.4.1 Parameters and control variables

The grouping of the eigenvalues is controlled by the method of distance determination JOBGRP and the parameters LBLOCK $= p_{\min}$ and/or DEL $= \delta_{\max}$ (see Step 2). Although the routine ZHHCNF allows to select any of the distance measures described in the previous section, we recommend the utilisation of the function d_{var} which has produced appropriate groupings in all examples investigated.

If the number of (multiple) eigenvalues of the matrix $\mathbf{A}$ is known to be p_A, the parameter configuration $p_{min} = p_A$, $\delta_{\max} = \infty$ should be used. Otherwise, the configuration $p_{min} = 1$, $\delta_{\max} \approx \|\mathbf{A}\|\sqrt{\varepsilon_{mach}}$ is a good initial guess. (ε_{mach} denotes the machine accuracy.) If LBLOCK $= p$ is the computed number of groups and δ_0, δ_1 are the computed cluster distances, then any of the parameter configurations $p_{\min} = p$, $\delta_{\max} = \infty$ or $p_{\min} = 1$, $\delta_0 \leq \delta_{\max} < \delta_1$ produces the same groups. Moreover, $\delta_0 \approx \delta_1$ indicates, that the groups are not well separated. In this case, the choice of either $\delta_{\max} \geq \delta_1$ or $\delta_{\max} < \delta_0$ may be of advantage.

Further information on the groups of eigenvalues is provided by the reciprocal condition numbers COND$(k) \geq \|P_k\|_2^{-1}$ (see Step 4) and the deflation norms DELE$(k) = \|E_k\|_F$ (see Step 6). Too small values of $\|P_k\|_2^{-1}$ (say $\ll \sqrt{\varepsilon_{mach}}$) combined with too large values of $\|E_k\|_F$ (say $\gg \sqrt{\varepsilon_{mach}}$) indicate, that the groups of eigenvalues should be made larger by increasing $\delta_{\max}$ or decreasing $p_{\min}$.

Too large values of $\|E_k\|_F$ may also indicate, that the Jordan structure is not well defined. In this case not all coupling elements $\beta_{h,j}$ are sufficiently larger than $\|E_k\|_F$ (see Step 6 and 7). If the computed Jordan chains are too long, the deflation parameter TOL $= \tau$ must be increased, if too short, the parameter must be decreased. A good initial value is $\tau \approx \sqrt{\varepsilon_{mach}}$.

An easy way for inspecting the coupling elements is to perform the computations up to STEP $= 8$. Then the superdiagonal of the matrix A $= \mathbf{J}^{(8)}$ contains these elements. Moreover, the Euclidean norms of the corresponding principal vectors $\|\mathbf{s}_i^{(8)}\|_2$ contained in the matrix S $= \mathbf{S}^{(8)}$ should also be inspected. Too large norms may indicate inappropriate groups of eigenvalues.

If the computed Jordan normal form of $\mathbf{A}$ is incorrect, the classification of the blocks of the matrix $\mathbf{C}$ usually fails (see Step 10). Otherwise, the H-orthogonalities of the eigenspaces ORTH$(k) = \|\mathbf{C}_k\|_F$ estimate the deviation of the numerical eigenspaces to their theoretical counterparts. If these values are too large, the normalized transformation matrix S $= \mathbf{S}^{(11)}$ obtained after STEP $= 11$ cannot be expected to be perfectly accurate. Nevertheless, the computed canonical form may still be correct, when the magnitudes of the corresponding sip block elements of H $= \mathbf{Z}^{(11)}$ are near unity. (All variables in typewriter style denote parameters of the routine ZHHCNF.)

6.4.2 Results of the routine ZHHCNF

We come to the presentation of some numerical results which were obtained on an INTEL PENTIUM 4 processor (2GHz) in a Cygwin environment CYGWIN_NT-5.1. The program binary was created with the GNU compiler g77/gcc 2.95.3-4, and the machine accuracy computed with the LAPACK routine DLAMCH was

$$\varepsilon_{mach} \approx 2.22 \cdot 10^{-16}.$$

In addition to the parameters and control variables discussed above, the residual

$$r_{AH} = \|\mathbf{A}^*\mathbf{H} - \mathbf{H}\mathbf{A}\|_F$$

estimates the accuracy of the test matrices, and the residuals

$$r_{AJ} = \|\mathbf{A}\mathbf{S} - \mathbf{S}\mathbf{J}\|_F, \quad r_{HZ} = \|\mathbf{S}^*\mathbf{H}\mathbf{S} - \mathbf{Z}\|_F$$

as well as the condition number

$$c_S = \|\mathbf{S}\|_1 \, \|\mathbf{S}^{-1}\|_1$$

estimate the accuracy of the computed Jordan normal form $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$ or canonical form $(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \mathbf{S}^*\mathbf{H}\mathbf{S}) = (\mathbf{J}, \mathbf{Z})$, respectively. Moreover, g_i denotes the grade of the i-th principal vector.

For the first example the Jordan normal form of the test matrix

$$\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 1 & 1 & -2 & 1 & -1 & 2 & -2 & 4 & -3 \\ -1 & 2 & 3 & -4 & 2 & -2 & 4 & -4 & 8 & -6 \\ -1 & 0 & 5 & -5 & 3 & -3 & 6 & -6 & 12 & -9 \\ -1 & 0 & 3 & -4 & 4 & -4 & 8 & -6 & 16 & -12 \\ -1 & 0 & 3 & -6 & 5 & -4 & 10 & -10 & 20 & -15 \\ -1 & 0 & 3 & -6 & 2 & -2 & 12 & -12 & 24 & -18 \\ -1 & 0 & 3 & -6 & 2 & -5 & 15 & -13 & 28 & -21 \\ -1 & 0 & 3 & -6 & 2 & -5 & 12 & -11 & 32 & -24 \\ -1 & 0 & 3 & -6 & 2 & -5 & 12 & -14 & 37 & -26 \\ -1 & 0 & 3 & -6 & 2 & -5 & 12 & -14 & 36 & -25 \end{bmatrix}$$

has been computed. This example was taken from [KR1] to permit comparison with the 560 algorithm. All variables contained in Table 6.1 indicate stable computation of the Jordan Normal form

$$\mathbf{J}^{(1)} = [\mathbf{J}_2(3) \oplus \mathbf{J}_2(3)] \oplus [\mathbf{J}_3(2) \oplus \mathbf{J}_2(2)] \oplus \mathbf{J}_1(1)$$

which corresponds to the results presented by Kågström and Ruhe.

The further examples from [KR1] have also been recomputed. In case of the Frank matrix ($n = 12$) the parameters $p_{\min} = 12$, $\delta_0 = 0.0$, $\delta_1 = 1.7 \cdot 10^{-4}$ apply for computing the Jordan normal form $\mathbf{J} = \lambda_1 \oplus \ldots \oplus \lambda_{12}$ with $r_{AJ} = 4.76 \cdot 10^{-14}$ and $c_S^{-1} = 5.03 \cdot 10^{-9}$. In case of the "ill conditioned" matrix ($\alpha = 10^{-4}$) the parameters $p_{\min} = 8$, $\delta_0 = 0.0$, $\delta_1 = 5.0 \cdot 10^{-9}$ apply for computing the Jordan normal form $\mathbf{J} = \lambda_1 \oplus \ldots \oplus \lambda_8$ with $r_{AJ} = 3.47 \cdot 10^{-12}$ and $c_S^{-1} = 1.67 \cdot 10^{-6}$. Since both test matrices are diagonalisable, there are no deflation steps to be performed, so that the parameter τ may be chosen arbitrarily. We do not present all details on these examples, because the computation of a trivial Jordan normal form is rather a test of the LAPACK routine ZGEES than of ZHHCNF.

Table 6.1: Results for the test matrix $\mathbf{A}^{(1)}$

k	λ_k^*	$\|\mathbf{P}_k\|_2^{-1}$	$\|\mathbf{E}_k\|_F$	i	g_i	$\mathbf{J}_{i,i+1}^{(8)}$	$\|\mathbf{s}_i^{(8)}\|_2$	$\mathbf{J}_{i,i+1}^{(9)}$	$\|\mathbf{s}_i^{(9)}\|_2$
1	2.0000	$4.626 \cdot 10^{-2}$	$1.863 \cdot 10^{-15}$	1	1	-5.3748	1.0000	1.0	1.0000
				2	2	1.4412	9.9672	1.0	1.8544
				3	3	0.0000	1.0000	0.0	0.1291
				4	1	3.2249	4.1112	1.0	4.1112
				5	2	0.0000	1.0000	0.0	0.3101
2	1.0000	$2.622 \cdot 10^{-2}$	0.0000	6	1	0.0000	1.0000	0.0	1.0000
3	3.0000	$2.813 \cdot 10^{-2}$	$2.610 \cdot 10^{-15}$	7	1	48.1250	1.0000	1.0	1.0000
				8	2	0.0000	1.0000	0.0	0.0208
				9	1	1.9885	1.0000	1.0	1.0000
				10	2	–	1.0000	0.0	0.5029

Parameters: $p_{\min} = 3$, $\delta_0 = 1.807 \cdot 10^{-15}$, $\delta_1 = 0.1667$, $\tau = 10^{-12}$
Residuals: $r_{AJ} = 2.388 \cdot 10^{-14}$, $c_S^{-1} = 9.680 \cdot 10^{-4}$

Table 6.2: Results for the test matrix pair $(\mathbf{A}^{(2)}, \mathbf{H}^{(2)})$

k, $\pi(k)$	λ_k^*	$\|\mathbf{P}_k\|_2^{-1}$	$\|\mathbf{E}_k\|_F$	$\|\mathbf{C}_k\|_F$	i	j	g_i	$\mathbf{J}_{i,i+1}^{(9)}$	$\|\mathbf{s}_i^{(9)}\|_2$	$\mathbf{Z}_{i,j}^{(11)}$	$\|\mathbf{s}_i^{(11)}\|_2$
1	1.0000+	$1.429 \cdot 10^{-1}$	$2.580 \cdot 10^{-15}$	$1.585 \cdot 10^{-15}$	1	9	1	1.0	1.0000	1.0	4.1124
(3)	1.0000i				2	8	2	0.0	0.2807	1.0	1.5113
					3	10	1	0.0	1.0000	1.0	4.5954
2	2.0000	$4.920 \cdot 10^{-2}$	$6.935 \cdot 10^{-15}$	$2.132 \cdot 10^{-15}$	4	5	1	1.0	1.0000	-1.0	3.0462
(2)					5	4	2	0.0	0.0886	-1.0	1.3990
					6	7	1	1.0	1.0000	1.0	5.2230
					7	6	2	0.0	0.4725	1.0	1.2276
3	1.0000−	$1.228 \cdot 10^{-1}$	$1.619 \cdot 10^{-15}$	$2.797 \cdot 10^{-15}$	8	2	1	1.0	1.0000	1.0	4.1124
(1)	1.0000i				9	1	2	0.0	0.1372	1.0	1.1269
					10	3	1	–	1.0000	1.0	3.4041

Parameters: $p_{\min} = 3$, $\delta_0 = 7.675 \cdot 10^{-15}$, $\delta_1 = 0.5714$, $\tau = 10^{-12}$
Residuals: $r_{AH} = 0.0000$, $r_{AJ} = 3.564 \cdot 10^{-14}$, $r_{HZ} = 3.668 \cdot 10^{-14}$, $c_S^{-1} = 1.452 \cdot 10^{-2}$

Table 6.3: Results for the test matrix pair $(\mathbf{A}^{(3)}, \mathbf{H}^{(3)})$ with $\mu = 0$, $\lambda = 2 + i$

k, $\pi(k)$	λ_k^*	$\|\mathbf{P}_k\|_2^{-1}$	$\|\mathbf{E}_k\|_F$	$\|\mathbf{C}_k\|_F$	i	j	g_i	$\mathbf{J}_{i,i+1}^{(9)}$	$\|\mathbf{s}_i^{(9)}\|_2$	$\mathbf{Z}_{i,j}^{(11)}$	$\|\mathbf{s}_i^{(11)}\|_2$
1	0.0000	$2.905 \cdot 10^{-1}$	$1.284 \cdot 10^{-15}$	$1.718 \cdot 10^{-14}$	1	2	1	1.0	1.0000	-1.0	0.4612
(1)					2	1	2	0.0	0.9770	-1.0	0.4738
					3	4	1	1.0	1.0000	1.0	0.4640
					4	3	2	0.0	1.0111	1.0	0.4659
2	2.0000+	$1.589 \cdot 10^{-1}$	$4.223 \cdot 10^{-15}$	$2.750 \cdot 10^{-14}$	5	13	1	1.0	1.0000	1.0	0.4750
(3)	1.0000i				6	13	2	1.0	0.9541	1.0	0.4591
					7	12	3	0.0	0.9416	1.0	0.4682
					8	15	1	1.0	1.0003	1.0	0.4856
					9	14	2	0.0	0.9848	1.0	0.4879
					10	16	1	0.0	1.0000	1.0	0.5022
3	2.0000−	$2.687 \cdot 10^{-1}$	$3.161 \cdot 10^{-15}$	$2.704 \cdot 10^{-14}$	11	7	1	1.0	1.0000	1.0	0.4750
(2)	1.0000i				12	6	2	1.0	0.9908	1.0	0.4744
					13	5	3	0.0	0.9373	1.0	0.4673
					14	9	1	1.0	1.0003	1.0	0.4724
					15	8	2	0.0	0.9602	1.0	0.4765
					16	10	1	–	1.0000	1.0	0.4755

Parameters: $p_{\min} = 3$, $\delta_0 = 3.213 \cdot 10^{-11}$, $\delta_1 = 1.091$, $\tau = 10^{-12}$
Residuals: $r_{AH} = 5.641 \cdot 10^{-14}$, $r_{AJ} = 5.249 \cdot 10^{-15}$, $r_{HZ} = 1.301 \cdot 10^{-14}$, $c_S^{-1} = 7.026 \cdot 10^{-2}$

In order to present some results on the computation of canonical forms, the literature was searched for corresponding examples. Although H-Hermitian matrices frequently appear in theoretical books or papers, we have not found concrete numerical examples which were large enough to give appropriate test cases. Therefore, the matrix pair

$$\operatorname{Re}\mathbf{A}^{(2)} = \begin{bmatrix} 2 & 0 & 0 & -1 & 0 & -1 & -1 & 0 & 0 & -1 \\ -2 & 3 & -4 & 2 & -2 & 2 & 1 & 0 & -3 & 2 \\ -1 & 1 & 1 & 1 & -1 & 1 & 1 & 0 & 0 & 1 \\ 0 & -1 & -3 & 0 & -1 & -1 & -2 & -1 & -2 & -1 \\ -1 & 1 & -3 & 1 & 0 & 1 & 0 & 0 & -2 & 1 \\ -1 & 1 & 0 & 1 & -1 & 2 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 3 & -2 & 7 & -2 & 4 & -2 & 0 & 2 & 5 & -2 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 2 & -1 & 4 & -2 & 2 & -2 & -1 & 0 & 3 & 0 \end{bmatrix}$$

$$\operatorname{Im}\mathbf{A}^{(2)} = \begin{bmatrix} 1 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & -1 & -2 & -1 & 0 & -1 & -1 & 0 & -2 & -3 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 3 & -2 & 1 & -2 & 2 & -1 & 1 & 0 & 0 & -3 \\ 1 & -1 & -1 & -1 & 1 & -1 & -1 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 1 \\ -4 & 2 & 1 & 2 & -2 & 2 & 1 & -1 & 2 & 6 \\ -1 & 1 & -1 & 1 & -1 & 1 & 1 & 0 & 0 & 1 \\ -2 & 1 & 2 & 1 & -1 & 1 & 1 & -1 & 2 & 4 \end{bmatrix}$$

$$\mathbf{H}^{(2)} = \begin{bmatrix} 0 & 1 & -1 & 1 & 0 & 2 & i & 1 & 0 & 0 \\ 1 & 0 & 2+i & 0 & 0 & 0 & i & 0 & -1 & 0 \\ -1 & 2-i & 0 & 1 & -1-i & 1 & i & 1 & 1+i & -i \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1+i & 0 & 0 & -i & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & i & 0 & 0 & 0 & -1 & 0 \\ -i & -i & -i & 0 & 0 & 0 & 0 & -i & 2 & i \\ 1 & 0 & 1 & 0 & 0 & 0 & i & 0 & 0 & 0 \\ 0 & -1 & 1-i & 0 & 0 & -1 & 2 & 0 & 0 & -1 \\ 0 & 0 & i & 0 & 0 & 0 & -i & 0 & -1 & 0 \end{bmatrix}$$

has been constructed from the canonical form

$$\mathbf{J}^{(2)} = \mathbf{J}_2(2) \oplus \mathbf{J}_2(2) \oplus \begin{bmatrix} \mathbf{J}_2(1+i) \oplus \mathbf{J}_1(1+i) & \\ & \mathbf{J}_2(1-i) \oplus \mathbf{J}_1(1-i) \end{bmatrix},$$

$$\mathbf{Z}^{(2)} = -\mathbf{Z}_2 \oplus \mathbf{Z}_2 \oplus \begin{bmatrix} & \mathbf{Z}_2 \oplus \mathbf{Z}_1 \\ \mathbf{Z}_2 \oplus \mathbf{Z}_1 & \end{bmatrix}$$

to serve as the next example. The results presented in Table 6.2 report on perfectly accurate computations resulting in very small residuals. It cannot be seen in the table that the eigenvalues are correct up to an error of 10^{-15}. They are much closer to the theoretical values than those computed with `ZGEES` in step 1 some of which had errors of order 10^{-8}. A similar effect was observed in all the computations made.

For the last examples the canonical form

$$\mathbf{J}^{(3)} = \mathbf{J}_2(\mu) \oplus \mathbf{J}_2(\mu) \oplus \begin{bmatrix} \mathbf{J}_3(\lambda) \oplus \mathbf{J}_2(\lambda) \oplus \mathbf{J}_1(\lambda) & \\ & \mathbf{J}_3(\overline{\lambda}) \oplus \mathbf{J}_2(\overline{\lambda}) \oplus \mathbf{J}_1(\overline{\lambda}) \end{bmatrix},$$

$$\mathbf{Z}^{(3)} = -\mathbf{Z}_2 \oplus \mathbf{Z}_2 \oplus \begin{bmatrix} & \mathbf{Z}_3 \oplus \mathbf{Z}_2 \oplus \mathbf{Z}_1 \\ \mathbf{Z}_3 \oplus \mathbf{Z}_2 \oplus \mathbf{Z}_1 & \end{bmatrix}$$

was specified and the test matrices were defined by

$$\mathbf{A}^{(3)} = \mathbf{R}^{-1}\mathbf{J}^{(3)}\mathbf{R} \text{ and } \mathbf{H}^{(3)} = \mathbf{R}^{*}\mathbf{Z}^{(3)}\mathbf{R},$$

where

$$\mathbf{R} = 2\mathbf{I}_n + \sum_{i=1}^{\frac{n}{2}-1} \big(\mathbf{J}_n(0)\big)^{2i} - \sum_{i=1}^{\frac{n}{2}} \big(\mathbf{J}_n(0)^{*}\big)^{2i-1} + i\mathbf{Z}_n$$

for $n = 16$. For all examples we used $\mu = 0$ but varied λ from $2+i$ to $2+10^{-3}i$.

Table 6.3 contains the complete results for the case $\lambda = 2 + i$ which are as satisfactory as in the last example. In particular, the eigenspaces obtained after step 9 were H-orthogonal up to machine precision (see $\|\mathbf{C}_k\|_F$) and allowed to compute the canonical form with small residuals r_{AJ}, r_{HZ} and an excellent reciprocal condition number c_S^{-1}.

However, when moving λ closer to the real axis, the results were getting worse. Table 6.4 shows an significant increase in $\|\mathbf{C}_k\|_F$ and thus also in r_{HZ} for the cases $\lambda = 2 + 10^{-m}i$, $m = 1, 2, 3$. Nevertheless, all canonical forms $(\mathbf{J}, \mathbf{Z})$ were computed correctly and the residuals r_{AJ} and reciprocal condition numbers c_S even pretend to be excellent.

In order to determine the reason for this behaviour the algorithm was in any case performed until step 4. Hence, the matrices $\mathbf{QUY}$ were computed such that

$$(\mathbf{QUY})^{-1}\mathbf{A}(\mathbf{QUY}) = \mathbf{T}_{11}(\mu) \oplus \mathbf{T}_{22}(\lambda) \oplus \mathbf{T}_{33}(\overline{\lambda}),$$

where $\mathbf{T}_{kk}(\omega)$ denotes an upper triangular block having diagonal elements close to ω. Theoretically, the matrices

$$\mathbf{F}_1 = \begin{bmatrix} \mathbf{F}_{12} & \mathbf{F}_{13} \end{bmatrix}, \quad \mathbf{F}_2 = \begin{bmatrix} \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix}, \quad \mathbf{F}_3 = \begin{bmatrix} \mathbf{F}_{31} & \mathbf{F}_{33} \end{bmatrix}$$

defined by the blocks of

$$(\mathbf{QUY})^{*}\mathbf{H}(\mathbf{QUY}) = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} & \mathbf{F}_{13} \\ \mathbf{F}_{21} & \mathbf{F}_{22} & \mathbf{F}_{23} \\ \mathbf{F}_{31} & \mathbf{F}_{32} & \mathbf{F}_{33} \end{bmatrix}$$

should then have been equal to or at least near to zero. However, the Frobenius norms $\|\mathbf{F}_k\|_F$ listed in Table 6.4 present another result. In other words, the invariant subspaces spanned by the columns of $\mathbf{QUY}$ were already not as H-orthogonal as they should have been.

Since the transformations $\mathbf{Y}$ according to the values of $\|\mathbf{P}_k\|_2^{-1}$ must be regarded as stable and the unitary transformations $\mathbf{U}$ are uncritical, the major error was already made during the computation of the Schur decomposition $\mathbf{Q}^{*}\mathbf{AQ} = \mathbf{T}$. But the `LAPACK` routine `ZGEES` used for this purpose is well-known for its stability, so that our last examples are at the bounds of the mathematical possibilities. Nevertheless, these examples also report on the well behaviour of the grouping algorithm and the stability of the overall transformations which did not increase the errors $\|\mathbf{F}_k\|_F$ until the final steps as is indicated by $\|\mathbf{C}_k\|_F$ and r_{HZ}.

Table 6.4: Further results for the test matrix pair $(\mathbf{A}^{(3)}, \mathbf{H}^{(3)})$

k, $\pi(k)$	λ_k^*	$\|\mathbf{P}_k\|_2^{-1}$	$\|\mathbf{E}_k\|_F$	$\|\mathbf{C}_k\|_F$	$\|\mathbf{F}_k\|_F$
1 (1)	0.0	$2.91 \cdot 10^{-1}$	$1.11 \cdot 10^{-15}$	$1.56 \cdot 10^{-14}$	$5.70 \cdot 10^{-14}$
2 (3)	$2.0 + 0.1i$	$1.78 \cdot 10^{-1}$	$5.20 \cdot 10^{-15}$	$2.04 \cdot 10^{-11}$	$2.50 \cdot 10^{-11}$
3 (2)	$2.0 - 0.1i$	$2.61 \cdot 10^{-1}$	$4.69 \cdot 10^{-15}$	$6.20 \cdot 10^{-11}$	$1.48 \cdot 10^{-10}$
r_{AH}, r_{AJ}, r_{HZ}, c_S^{-1}:		$4.78 \cdot 10^{-14}$	$5.39 \cdot 10^{-15}$	$1.47 \cdot 10^{-11}$	$7.03 \cdot 10^{-2}$
1 (1)	0.0	$2.91 \cdot 10^{-1}$	$1.15 \cdot 10^{-15}$	$1.31 \cdot 10^{-14}$	$6.27 \cdot 10^{-14}$
2 (3)	$2.0 + 0.01i$	$1.59 \cdot 10^{-1}$	$3.48 \cdot 10^{-15}$	$1.16 \cdot 10^{-6}$	$1.48 \cdot 10^{-6}$
3 (2)	$2.0 - 0.01i$	$2.69 \cdot 10^{-1}$	$2.30 \cdot 10^{-15}$	$4.23 \cdot 10^{-6}$	$1.20 \cdot 10^{-5}$
r_{AH}, r_{AJ}, r_{HZ}, c_S^{-1}:		$4.88 \cdot 10^{-14}$	$3.69 \cdot 10^{-15}$	$9.90 \cdot 10^{-7}$	$7.03 \cdot 10^{-2}$
1 (1)	0.0	$2.91 \cdot 10^{-1}$	$6.18 \cdot 10^{-16}$	$1.33 \cdot 10^{-14}$	$6.06 \cdot 10^{-14}$
2 (3)	$2.0 + 0.001i$	$1.77 \cdot 10^{-1}$	$3.22 \cdot 10^{-15}$	$3.99 \cdot 10^{-1}$	$4.85 \cdot 10^{-1}$
3 (2)	$2.0 - 0.001i$	$2.62 \cdot 10^{-1}$	$1.77 \cdot 10^{-15}$	$4.22 \cdot 10^{-3}$	$1.11 \cdot 10^{-2}$
r_{AH}, r_{AJ}, r_{HZ}, c_S^{-1}:		$5.55 \cdot 10^{-14}$	$3.46 \cdot 10^{-15}$	$9.01 \cdot 10^{-2}$	$6.99 \cdot 10^{-2}$

6.4.3 Further numerical considerations

In addition to the examples presented we have performed many tests which confirmed that the algorithm works well. But unfortunately we are not able to present a detailed error analysis.

Clearly, the numerical computation of the Jordan normal form of $\mathbf{A}$ or the canonical form of $(\mathbf{A}, \mathbf{H})$, respectively, is a discontinuous problem. A tiny perturbation may destroy the results in exact sense. Therefore, the algorithm tries to find the nearest matrix $\hat{\mathbf{A}}$ or the nearest pair of matrices $(\hat{\mathbf{A}}, \hat{\mathbf{H}})$, respectively, such that

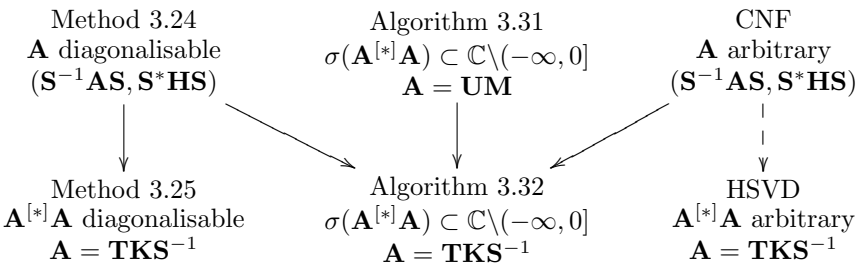
$$\|\hat{\mathbf{A}} - \mathbf{A}\| < \varepsilon_A \|\mathbf{A}\|, \qquad \hat{\mathbf{S}}^{-1}\hat{\mathbf{A}}\hat{\mathbf{S}} = \hat{\mathbf{J}},$$
$$\|\hat{\mathbf{H}} - \mathbf{H}\| < \varepsilon_H \|\mathbf{H}\|, \qquad \hat{\mathbf{S}}^{*}\hat{\mathbf{H}}\hat{\mathbf{S}} = \hat{\mathbf{Z}},$$

where $\hat{\mathbf{J}}$, $\hat{\mathbf{Z}}$, $\hat{\mathbf{S}}$ are the computed matrices and ε_A, ε_H should be small constants. Exact perturbation bounds for these constants are not available due to the complexity of the overall process. For this reason, the analysis of the stability information provided by the algorithm and the computation of the residuals, condition number, and vector norms are the only possibilities for assessing the quality of the computed normal forms.

6.5 Numerical computation of H-polar decompositions of arbitrary matrices

We conclude this chapter with some remarks on an extension of the `CNF` algorithm for computing H-singular value decompositions or H-polar decomposi-

tions, respectively. Consider the following diagram:



The diagram shows the dependencies of the methods and algorithms presented in Chapter 3. Here `CNF` is applied in step 2 of Algorithm 3.32 when the matrix $\mathbf{M}$, computed in step 1 with Algorithm 3.31, is not diagonalisable and Method 3.24 cannot be used for determining the canonical form of the pair $(\mathbf{M}, \mathbf{H})$. On the other hand, Method 3.24 is the foundation for Method 3.25 which computes an H-SVD (or H-polar decomposition) of a matrix $\mathbf{A}$ from the canonical form of the pair $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H})$ provided that $\mathbf{A}^{[*]}\mathbf{A}$ is diagonalisable. Now, using `CNF` instead of Method 3.24, it is possible in principle to generalise Method 3.25 such that H-SVDs of arbitrary matrices can be computed.

Assume that in step 1 of such an algorithm, called `HSVD`, the canonical form $(\mathbf{R}^{-1}\mathbf{A}^{[*]}\mathbf{AR}, \mathbf{R}^*\mathbf{HR}) = (\mathbf{J}, \mathbf{Z})$ has been computed with `CNF`. Then in step 2 square roots of the Jordan blocks contained in $\mathbf{J}$ must be determined. For non-zero eigenvalues this is possible using the formula given in Lemma 3.1 which can be written in the form

$$\mathbf{K}_p(\mu) = \begin{bmatrix} \mu & f_1(\mu) & f_2(\mu) & \cdots & f_{p-1}(\mu) \\ & \mu & f_1(\mu) & \ddots & \vdots \\ & & \mu & \ddots & f_2(\mu) \\ & & & \ddots & f_1(\mu) \\ & & & & \mu \end{bmatrix}$$

where

$$f_i(\mu) = \begin{cases} \dfrac{1}{2\mu}, & \text{if } i = 1 \\ \dfrac{(-1)^{i+1}(2i-3)!!}{(2i)!!\ \mu^{2i-1}}, & \text{if } 2 \le i \le p-1 \end{cases}$$

with

$$i!! = \begin{cases} 1 \cdot 3 \cdot \ldots \cdot i, & \text{if } i \text{ is odd} \\ 2 \cdot 4 \cdot \ldots \cdot i, & \text{if } i \text{ is even} \end{cases}.$$

These Toeplitz matrices satisfy $\mathbf{K}_p(\mu)^2 = \mathbf{J}_p(\mu^2)$ for all $\mu \in \mathbb{C}\backslash\{0\}$. Now, if $(\mathbf{J}, \mathbf{Z})$ contains the blocks

$$\begin{aligned} (\mathbf{J}^{(3)}, \mathbf{Z}^{(3)}) &= (\mathbf{J}_p(\omega^2) \oplus \mathbf{J}_p(\overline{\omega}^2), \mathbf{Z}_{2p}) \ \text{ with } \ \omega \in \mathbb{C}\backslash\mathbb{R} \cup i\mathbb{R}, \\ (\mathbf{J}^{(2)}, \mathbf{Z}^{(2)}) &= (\mathbf{J}_p(\alpha^2), \varepsilon\mathbf{Z}_p) \ \text{ with } \ \alpha \in \mathbb{R}, \ \ \varepsilon = \pm 1, \\ (\mathbf{J}^{(1)}, \mathbf{Z}^{(1)}) &= (\mathbf{J}_p(-\beta^2) \oplus \mathbf{J}_p(-\beta^2), \mathbf{Z}_p \oplus -\mathbf{Z}_p) \ \text{ with } \ \beta \in \mathbb{R}, \end{aligned}$$

then the matrices defined by

$$\mathbf{K}^{(3)} = \mathbf{K}_p(\omega) \oplus \mathbf{K}_p(\overline{\omega}), \quad \mathbf{K}^{(2)} = \mathbf{K}_p(\alpha) \text{ and} \tag{6.21}$$
$$\mathbf{K}^{(1)} = \mathbf{W}_{2p}^{-1}\big(\mathbf{K}_p(i\beta) \oplus \mathbf{K}_p(-i\beta)\big)\mathbf{W}_{2p} \text{ with } \mathbf{W}_{2p} = \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{I}_p & \mathbf{I}_p \\ \mathbf{I}_p & -\mathbf{I}_p\end{bmatrix}$$

fulfil

$$\big(\mathbf{K}^{(j)}\big)^2 = \mathbf{J}^{(j)} \text{ and } \big(\mathbf{K}^{(j)}\big)^*\mathbf{Z}^{(j)} = \mathbf{Z}^{(j)}\mathbf{K}^{(j)}$$

for $j = 1, 2, 3$ (see [BMRRR1, Theorem 4.4]). Here the blocks belonging to the negative eigenvalue $-\beta^2$ must satisfy condition 1 of Theorem 3.4 because otherwise the H-Hermitian square root cannot be formed.

Whereas these transformations are simple, the kernel transformation required in this step is extremely complicated. Let

$$\mathbf{J}^{(0)} = \bigoplus_{i=1}^{r} \mathbf{N}_{p_i}, \quad \mathbf{Z}^{(0)} = \bigoplus_{i=1}^{r} \varepsilon_i \mathbf{Z}_{p_i}$$

be the part of the canonical form $(\mathbf{J}, \mathbf{Z})$ belonging to the eigenvalue 0 and let $\mathbf{R}^{(0)}$ be the corresponding (rectangular) block of $\mathbf{R}$. Then the kernel transformation is possible only if condition 2 of Theorem 3.4 holds. It then requires to determine a matrix $\mathbf{W}^{(0)}$ such that

$$\mathbf{J}^{(0)} = \big(\mathbf{W}^{(0)}\big)^{-1}\mathbf{J}^{(0)}\mathbf{W}^{(0)}, \quad \mathbf{Z}^{(0)} = \big(\mathbf{W}^{(0)}\big)^*\mathbf{Z}^{(0)}\mathbf{W}^{(0)}$$

and such that the basis spanned by the columns of $\mathbf{R}^{(0)}\mathbf{W}^{(0)}$ expresses $\ker(\mathbf{A})$ as specified in condition 3 of Theorem 3.4. Here several possibilities for combining the blocks of $(\mathbf{J}^{(0)}, \mathbf{Z}^{(0)})$ exist. For example, if $\mathbf{J}^{(0)}$ consists of the blocks $\mathbf{N}_3$, $\mathbf{N}_3$, $\mathbf{N}_2$, $\mathbf{N}_2$, and $\mathbf{Z}^{(0)}$ consists of the block $\mathbf{Z}_3$, $-\mathbf{Z}_3$, $\mathbf{Z}_2$, $-\mathbf{Z}_2$, then

$$\mathbf{J}^{(0)} = (\mathbf{N}_3 \oplus \mathbf{N}_3) \oplus (\mathbf{N}_2 \oplus \mathbf{N}_2), \quad \mathbf{Z}^{(0)} = (\mathbf{Z}_3 \oplus -\mathbf{Z}_3) \oplus (\mathbf{Z}_2 \oplus -\mathbf{Z}_2)$$

as well as

$$\mathbf{J}^{(0)} = (\mathbf{N}_3 \oplus \mathbf{N}_2) \oplus (\mathbf{N}_3 \oplus \mathbf{N}_2), \quad \mathbf{Z}^{(0)} = (\mathbf{Z}_3 \oplus \mathbf{Z}_2) \oplus (-\mathbf{Z}_3 \oplus -\mathbf{Z}_2)$$

fulfil condition 2 of Theorem 3.4 (taken from [BMRRR1]). Moreover, even if it is clear which of the blocks have to be combined for building the square roots, it is still complicated to determine the transformations

$$\mathbf{N}_p \oplus \mathbf{N}_p = \mathbf{W}_{2p}^{-1}\big(\mathbf{N}_p \oplus \mathbf{N}_p\big)\mathbf{W}_{2p},$$
$$\mathbf{Z}_p \oplus -\mathbf{Z}_p = \mathbf{W}_{2p}^{*}\big(\mathbf{Z}_p \oplus -\mathbf{Z}_p\big)\mathbf{W}_{2p}$$

and

$$\mathbf{N}_p \oplus \mathbf{N}_{p-1} = \mathbf{W}_{2p-1}^{-1}\big(\mathbf{N}_p \oplus \mathbf{N}_{p-1}\big)\mathbf{W}_{2p-1},$$
$$\mathbf{Z}_p \oplus \mathbf{Z}_{p-1} = \mathbf{W}_{2p-1}^{*}\big(\mathbf{Z}_p \oplus \mathbf{Z}_{p-1}\big)\mathbf{W}_{2p-1}$$

which transform into an appropriate basis.

Confronted with these difficulties we have not found a way for the numerical computation of the matrix $\mathbf{W}^{(0)}$, so that this approach has not led to an algorithm for computing H-SVDs of arbitrary matrices. Nevertheless, it is still possible to combine (6.21) with the kernel transformation of Method 3.25 to

obtain an algorithm for computing H-SVDs of matrices $\mathbf{A}$, for which the part of the canonical form of the pair $(\mathbf{A}^{[*]}\mathbf{A}, \mathbf{H})$ belonging to the eigenvalue 0 has the form $(\mathbf{0}_{p+q,p+q}, \mathbf{I}_p \oplus -\mathbf{I}_q)$. On the other hand, equation (5.40) shows that the Newton method from Chapter 5 allows to compute H-polar decompositions of arbitrary matrices. Hence, by using this method in step 1 of Algorithm 3.32, we are in principle (and, in many cases, in practice) able to compute an H-SVD of an arbitrary matrix $\mathbf{A}$.

Chapter 7

Conclusions

This final chapter summarises the most important results on the Procrustes problems and gives an illustrative example. Moreover, some suggestions for the application in multidimensional scaling are made.

Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and let $\mathbf{G}, \mathbf{H} \in \mathbb{F}^{n\times n}$ be nonsingular and selfadjoint. Furthermore, let $\mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_N]$, $\mathbf{Y} = [\mathbf{y}_1 \ \dots \ \mathbf{y}_N] \in \mathbb{F}^{n\times N}$ be given coordinates of vectors and let $\mathbf{U} \in \mathbb{F}^{n\times n}$. Then the function

$$f(\mathbf{U}) = \sum_k [\mathbf{U}\mathbf{x}_k - \mathbf{y}_k, \mathbf{U}\mathbf{x}_k - \mathbf{y}_k]_H = \operatorname{tr}[(\mathbf{U}\mathbf{X} - \mathbf{Y})^*\mathbf{H}(\mathbf{U}\mathbf{X} - \mathbf{Y})]$$

measures the congruence of the constellations and the constrained optimisation problems

$$f(\mathbf{U}) \to \text{opt} \quad \text{with} \quad \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H}, \tag{7.1}$$

$$f(\mathbf{U}) \to \text{opt} \quad \text{with} \quad \mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{H} \ \text{ and } \ \mathbf{U}^*\mathbf{G}\mathbf{U} = \mathbf{G}, \tag{7.2}$$

$$f(\mathbf{U}) \to \text{opt} \quad \text{with} \quad \mathbf{U}^*\mathbf{G}\mathbf{U} = \mathbf{G} \tag{7.3}$$

determine isometries, with the help of which a transformation to an optimum congruence can be achieved. The wanted optimum depends on the matrix $\mathbf{H}$ and must be chosen according to

$$\text{opt} = \begin{cases} \min, & \text{if } \mathbf{H} > 0 \text{ (positive definite)} \\ \max, & \text{if } \mathbf{H} < 0 \text{ (negative definite)} \\ \min \,/\, \max, & \text{otherwise (indefinite)} \end{cases}$$

where $\min \,/\, \max$ describes a particular saddle point of the function f.

Our studies show that the H-isometric Procrustes problem (7.1) has a solution if and only if a semidefinite H-polar decomposition

$$\mathbf{A} = \mathbf{Y}\mathbf{X}^*\mathbf{H} = \mathbf{U}\mathbf{M} \ \text{ with}$$
$$\mathbf{U}^H = \mathbf{U}^{-1}, \ \ \mathbf{M}^H = \mathbf{M}, \ \ \mathbf{H}\mathbf{M} \geq 0$$

exists (see Section 4.4). Analogously, the (G,H)-isometric Procrustes problem (7.2) has a solution if and only if an H-semidefinite (G,H)-polar decomposition

$$\mathbf{C} = \mathbf{Y}\mathbf{X}^*\mathbf{H} + \mathbf{G}^{-1}\mathbf{H}\mathbf{Y}\mathbf{X}^*\mathbf{G} = \mathbf{U}\mathbf{M} \ \text{ with}$$
$$\mathbf{U}^H = \mathbf{U}^G = \mathbf{U}^{-1}, \ \ \mathbf{M}^H = \mathbf{M}^G = \mathbf{M}, \ \ \mathbf{H}\mathbf{M} \geq 0,$$

exists (see Section 4.5). Here it is additionally assumed that $\mathbf{G}$ and $\mathbf{H}$ satisfy

$$\mathbf{H}^{-1}\mathbf{G} = \mu^2\mathbf{G}^{-1}\mathbf{H} \text{ for some } \mu \in \mathbb{R}\backslash\{0\}.$$

In both cases the matrix $\mathbf{U}$ contained in the decomposition is the wanted isometry and the decomposition always exists when $\mathbf{H}$ is definite.

In contrast to these handy results, no analogous solution of the Procrustes problem (7.3) has been found. Here the necessary condition for determining the isometry is

$$\mathbf{GU\Lambda} + \mathbf{HUXX}^* = \mathbf{HYX}^*$$

where $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ beside $\mathbf{U}$ is unknown. Moreover, the additional assumption

$$\mathbf{G} \neq \mu\mathbf{H} \text{ for all } \mu \in \mathbb{R}$$

must be made to avoid that the problem can be reduced to (7.1). However, under these prerequisites we were not able to express $\mathbf{U}$ as a factor of some matrix decomposition (see Section 5.1).

For the numerical solution of the Procrustes problems the following methods are given:

(a) the Method 3.25 for computing H-singular value and H-polar decompositions of a matrix $\mathbf{A}$ for which $\mathbf{A}^H\mathbf{A}$ is diagonalisable,

(b) the Algorithm 3.32 for computing H-singular value and H-polar decompositions of a matrix $\mathbf{A}$ for which $\mathbf{A}^H\mathbf{A}$ has no non-positive eigenvalues,

(c) the Algorithm 4.20 for computing (G,H)-polar decompositions of a matrix $\mathbf{A}$ for which $\mathbf{A}^H\mathbf{A}$ is diagonalisable or has no non-positive eigenvalues,

(d) the Newton method from Chapter 4 for optimising functions of the form

$$f(\mathbf{u}) = (\mathbf{Au}, \mathbf{u}) - 2\,\mathrm{Re}(\mathbf{b}, \mathbf{u}) + \gamma \text{ with } \mathbf{u} = \mathrm{vec}(\mathbf{U})$$

under the constraints $\mathbf{U}^H = \mathbf{U}^{-1}$ and/or $\mathbf{U}^G = \mathbf{U}^{-1}$.

(a) and (b) apply for solving (7.1), (c) applies for solving (7.2), and (d) applies solving for (7.3).

Although (d) might also be used for computing solutions of (7.1) and (7.2), the other methods are preferable because they are more efficient and allow to decide whether the wanted decomposition exists. If the Newton method is applied to solve a problem in which $\mathbf{H}$ is indefinite and the iteration diverges, we can usually not decide whether this failure is caused by inappropriate starting values or by the fact that the considered problem is unsolvable. Fortunately, this difficulty does not arise in the case of a definite matrix $\mathbf{H}$ in which the existence of a solution is guaranteed, so that in this case only suitable starting values must be found.

In the following application the different Procrustes solutions are illustrated with a concrete numerical example.

Example 7.1. For $N = 5$, $n = 3$, $\mathbf{J}_{2,1} = \mathrm{diag}(1, 1, -1)$ and

$$\mathbf{X}^T = \begin{bmatrix} -0.2373 & 0.5122 & 1.7640 \\ 1.2910 & -0.8456 & -0.8393 \\ 1.6640 & 0.0378 & -0.8054 \\ -2.3035 & 0.8568 & 0.3262 \\ -0.4141 & -0.5612 & -0.4456 \end{bmatrix}, \quad \mathbf{Y}^T = \begin{bmatrix} -0.5979 & 0.5173 & -1.8355 \\ -0.5616 & -1.2324 & 0.4944 \\ -1.2682 & -0.5380 & 0.4147 \\ 1.7739 & 1.4836 & 0.2375 \\ 0.6539 & -0.2305 & 0.6889 \end{bmatrix}$$

the solution of (7.1) with $\mathbf{H} = \mathbf{J}_{2,1}$ computed with Algorithm 3.32 is

$$\mathbf{U}_1 = \begin{bmatrix} -0.9676 & -0.3915 & -0.2991 \\ -0.3681 & 0.9302 & 0.0268 \\ -0.2677 & -0.1360 & -1.0441 \end{bmatrix}, \ \mathbf{X}_1^T = \begin{bmatrix} -0.4986 & 0.6111 & -1.8480 \\ -0.6670 & -1.2842 & 0.6457 \\ -1.3839 & -0.5989 & 0.3903 \\ 1.7958 & 1.6536 & 0.1596 \\ 0.7537 & -0.3815 & 0.6524 \end{bmatrix}.$$

The solution of (7.2) with $\mathbf{G} = \mathbf{J}_{2,1}$ and $\mathbf{H} = \mathbf{I}_3$ computed with Algorithm 4.20 is

$$\mathbf{U}_2 = \begin{bmatrix} -0.9021 & -0.4315 & 0.0 \\ -0.4315 & 0.9021 & 0.0 \\ 0.0 & 0.0 & -1.0 \end{bmatrix}, \ \mathbf{X}_2^T = \begin{bmatrix} -0.0069 & 0.5645 & -1.7640 \\ -0.7998 & -1.3198 & 0.8393 \\ -1.5174 & -0.6839 & 0.8054 \\ 1.7084 & 1.7668 & -0.3262 \\ 0.6158 & -0.3276 & 0.4456 \end{bmatrix}$$

and the solution of (7.3) with $\mathbf{G} = \mathbf{J}_{2,1}$ and $\mathbf{H} = \mathbf{I}_3$ determined with the Newton method is

$$\mathbf{U}_3 = \begin{bmatrix} -0.9773 & -0.3858 & -0.3226 \\ -0.3744 & 0.9276 & -0.0248 \\ -0.3088 & -0.0965 & -1.0510 \end{bmatrix}, \ \mathbf{X}_3^T = \begin{bmatrix} -0.5347 & 0.5202 & -1.8302 \\ -0.6648 & -1.2468 & 0.5650 \\ -1.3811 & -0.5679 & 0.3290 \\ 1.8155 & 1.6490 & 0.2858 \\ 0.7650 & -0.3545 & 0.6503 \end{bmatrix}.$$

The residuals after the transformation $\mathbf{X}_k = \mathbf{U}_k\mathbf{X}$ are listed in the following table in which $\mathbf{X}_0 = \mathbf{X}$ denotes the original coordinates.

	congruence measure	$k=0$	$k=1$	$k=2$	$k=3$
(a)	$\mathrm{tr}[(\mathbf{X}_k - \mathbf{Y})^T(\mathbf{X}_k - \mathbf{Y})]$	48.4274	0.1427	1.2481	0.1015
(b)	$\mathrm{tr}[(\mathbf{X}_k - \mathbf{Y})^T\mathbf{J}_{2,1}(\mathbf{X}_k - \mathbf{Y})]$	13.3900	0.0807	-0.0590	0.0691

As one would expect, the solution of (7.3) results in the best congruence with respect to the Euclidean measure (a), but it is only a little better than the saddle point solution of problem (7.1). For solving (7.2) only a rotation in the xy-plane and a reflection along the z-axis is admitted ($\mathbf{H} = \mathbf{I}_2 \oplus \mathbf{I}_1$ and $\mathbf{G} = \mathbf{I}_2 \oplus -\mathbf{I}_1$), but this still produces an acceptable improvement in comparison with the starting situation.

These results can also be observed in Figure 7.1 in which the $\mathbf{Y}$ constellation is depicted with dashed lines and the $\mathbf{X}_k$ constellations are depicted with solid lines. The respective first point (first row of the transposed matrix) is surrounded by a circle and the figures show the projections onto the coordinate planes. ♢

With this example our studies are complete and it would now be interesting to apply the results in some real psychological multidimensional scaling investigations. In this context it could be analysed whether the mathematical property of vectors to be space-like, time-like or light-like can also be given a meaning in psychology (see Remark 4.12). Moreover, it were possible to search for the laws of cognition in indefinite scalar product spaces or even in Riemannian spaces. For this purpose, the following approach may be used:

Figure 7.1a: Projection onto the xy-plane

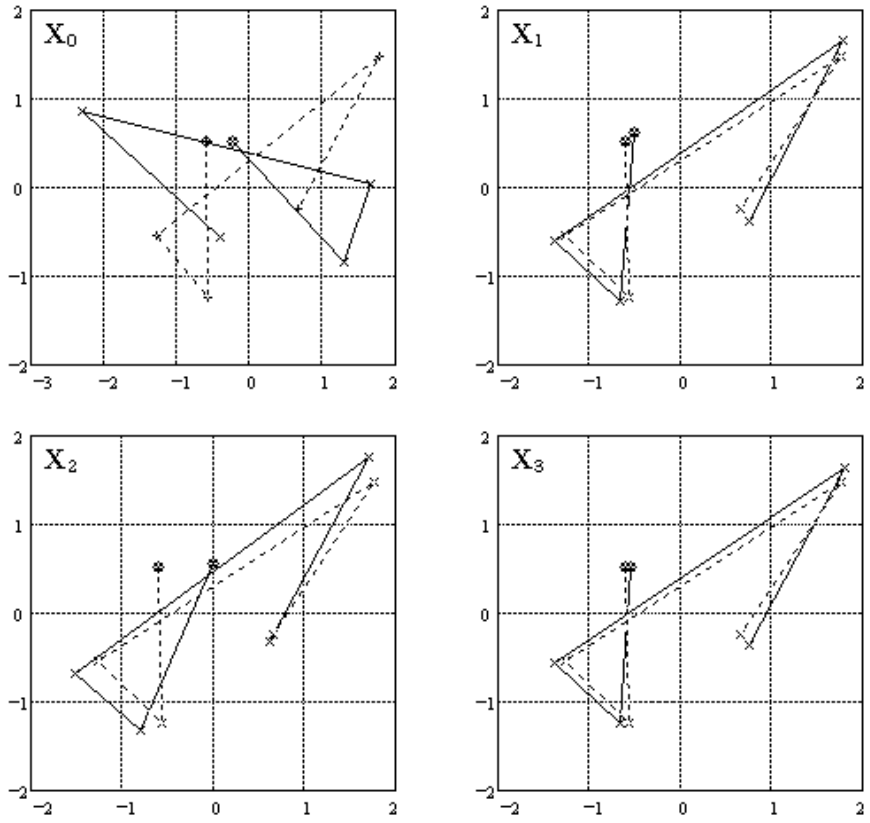
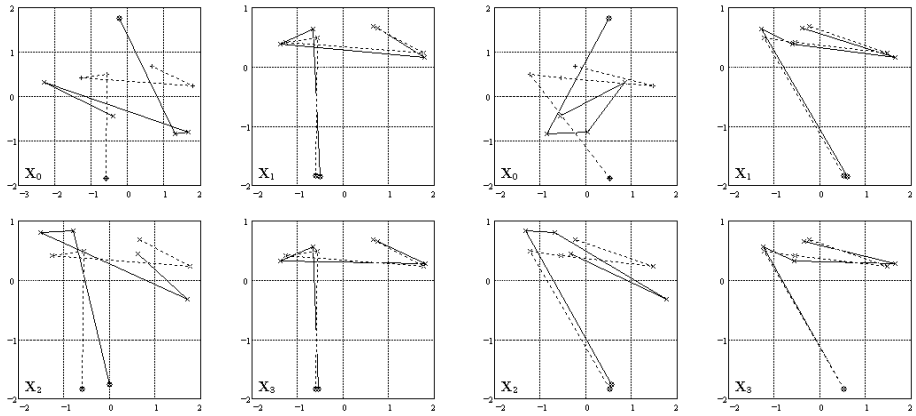


Figure 7.1b: Projections onto the xz- and yz-plane



Let ω_k be given objects and let $p_{kl}^{(0)}, \ldots, p_{kl}^{(T)}$ be sets of proximities measured at the times $t = 0, \ldots, T$ $(1 \leq k, l \leq N)$. Moreover, let $x_k^{(0)}, \ldots, x_k^{(T)}$ be constellations of vectors, determined according to Theorem 4.9, such that

$$\left[\mathbf{x}_k^{(t)} - \mathbf{x}_l^{(t)}, \mathbf{x}_k^{(t)} - \mathbf{x}_l^{(t)}\right] = \left(p_{kl}^{(t)}\right)^2$$

where $[.,.]$ is an indefinite scalar product with underlying matrix $\mathbf{G} = \mathbf{I}_p \oplus -\mathbf{I}_q$. Then by setting

$$\mathbf{y}_k^{(0)} = \mathbf{x}_k^{(0)}$$

and subsequently applying Procrustes solutions (for example of problem (7.2) with $\mathbf{H} = \mathbf{I}_{p+q}$)

$$\sum_k [\mathbf{U}^{(t)}\mathbf{x}_k^{(t)} - \mathbf{y}_k^{(t-1)}, \mathbf{U}^{(t)}\mathbf{x}_k^{(t)} - \mathbf{y}_k^{(t-1)}] \to \text{opt}, \quad \mathbf{y}_k^{(t)} = \mathbf{U}^{(t)}\mathbf{x}_k^{(t)}$$

for $t = 1, \ldots, T$, it can be achieved that adjacent constellations in $y_k^{(0)}, \ldots, y_k^{(T)}$ are optimally congruent. Now, for example using n-dimensional cubic splines, it is possible to determine curves $\mathbf{y}_k(\tau)$, $\tau \in [0, T]$, such that

$$\mathbf{y}_k(t) = \mathbf{y}_k^{(t)}.$$

Each curve $\mathbf{y}_k(t)$ describes the movement of an object ω_k in an indefinite scalar product space with metric $\mathbf{G}$, and if it is assumed that these curves are sort of "shortest curves" the situation appearing here has a well-known counterpart in physics: Einstein's theory of gravitation.

In general relativity the moving objects are cosmological objects such as galaxies or planets, and the curves are the geodesics of an Riemannian space with metric

$$\mathbf{G}(x, y, z, t) = \begin{bmatrix} -1 & & & \\ & -1 & & \\ & & -1 & \\ & & & 1 \end{bmatrix} + \mathbf{\Psi}(x, y, z, t)$$

where $\mathbf{\Psi}(x, y, z, t) = [\psi_{\alpha\beta}(x, y, z, t)]$ is symmetric [SU]. Isn't it possible that similar laws apply in cognition and that the curves $\mathbf{y}_k(t)$ are just geodesics in an Riemannian space? If this is true, are there more and deeper analogies between cognition and the universe?

Although this seem to be very interesting questions, their discussion is far beyond the scope of this work, so that they must be clarified at another place.

Bibliography

[A] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[BF] P. Benner, and H. Faßbender, *Computing Roots of Matrix Products*, Zeitschrift für angewandte Mathematik und Mechanik, Vol. 81, Suppl. 2, 717-718, 2001.

[BG] I. Borg, and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York, 1997.

[BMRRR1] Y. Bolshakov, C.V.M. van der Mee, A.C.M. Ran, B. Reichstein, and L. Rodman, *Polar decompositions in finite dimensional indefinite scalar product spaces: General Theory*, Linear Algebra Appl. 261, 91-141, 1997.

[BMRRR2] Y. Bolshakov, C.V.M. van der Mee, A.C.M. Ran, B. Reichstein, and L. Rodman, *Extension of isometries in finite-dimensional indefinite scalar product spaces and polar decompositions*, SIAM J. Matrix Anal. Appl. 18, 752-774, 1997.

[BMRRR3] Y. Bolshakov, C.V.M. van der Mee, A.C.M. Ran, B. Reichstein, and L. Rodman, *Polar decompositions in finite-dimensional indefinite scalar product spaces: special cases and applications*, in: *Recent Developments in Operator Theory and its Applications*, OT 87 (I. Gohberg, P. Lancaster, P.N. Shivakumar, Eds.), Birkhäuser, Basel, 1996, 61-94. *Errata*, Integral Equations and Operator Theory 17, 497-501, 1997.

[BOS] A.W. Bojanczyk, R. Onn, and A.O. Steinhardt, *Existence of the hyperbolic singular value decomposition*, Linear Algebra Appl. 185, 21-30, 1993.

[BR] Y. Bolshakov, and B. Reichstein, *Unitary equivalence in an indefinite scalar product: an analogue of singular-value decomposition*, Linear Algebra Appl. 222, 155-226, 1995.

[BS] C. Baveley, and G.W. Stewart, *An algorithm for computing reducing subspaces by block diagonalisation*, SIAM J. Num. Anal. 16, 359-367, 1979.

[BU] A. Bunse-Gerstner, *An analysis of the HR algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl. 35, 155-173, 1981.

[CL] G.W. Cross, and P. Lancaster, *Square Roots of Complex Matrices*, Linear and Multilinear Algebra 1, 289-293, 1974.

[D] M. Davidson, *Multidimensional Scaling*, Wiley, New York, 1983.

[DP] P.S.Dwyer, and M.S.McPhail, *Symbolic Matrix Derivatives*, Ann. Math. Statist. 19, 517-534, 1948.

[ER] F. Erwe, *Differential- und Integralrechnung I*, Bibl. Inst., Mannheim, 1962.

[F] J.F.G. Francis, *The QR transformation. An unitary analogue to the LR transformation*, Computer J. 4, 265-271, 332-345, 1961/62.

[G] F.R. Gantmacher, *The Theory of Matrices (Vol. I)*, Chelsea, New York, 1959.

[GB] F.A. Graybill, *Matrices with Applications in Statistics (2nd Ed.)*, Wadsworth, Belmont, 1983.

[GLR] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser, Basel, 1983.

[GR] W. Greub, *Linear Algebra (3rd Ed.)*, Springer, Berlin, 1967.

[GVL] G.H. Golub, and C.F. Van Loan, *Matrix Computations (3rd Ed.)*, Johns Hopkins University Press, Baltimore, 1996.

[H] H. Harman, *Modern Factor Analysis (3rd Ed.)*, Univ. of Chicago Press, Chicago, 1976.

[HI1] N.J. Higham, *The Matrix Sign Decomposition and its Relation to the Polar Decomposition*, Linear Algebra Appl. 212/213, 3-20, 1994.

[HI2] N.J. Higham, *J-Orthogonal Matrices: Properties and Generation*, SIAM Review Vol. 45, No. 3, 504-519, 2003.

[HMMT] N.J. Higham, D.S Mackey, N. Mackey, and F. Tisseur, *Computing the Polar Decomposition and the Matrix Sign Decomposition in Matrix Groups*, SIAM J. Matrix Anal. Appl. 25(4), 1178-1192, 2004.

[HP] N.J. Higham, and P. Papadimitrou, *A Parallel Algorithm for Computing the Polar Decomposition*, Parallel Computing 20(8), 1161-1173, 1994.

[HO] O.V. Holtz, *On indecomposable normal matrices in spaces with indefinite scalar product*, Linear Algebra Appl. 259, 155-168, 1997.

[KR1] B. Kågström, and A. Ruhe, *An Algorithm for Numerical Computation of the Jordan Normal Form of a Complex Matrix*, ACM Transactions on Mathematical Software (TOMS) Vol. 6, No. 3, 398-419, September 1980.

[KR2] B. Kågström, and A. Ruhe, *Algorithm 560, JNF, An Algorithm for Numerical Computation of the Jordan Normal Form of a Complex Matrix [F2]*, ACM Transactions on Mathematical Software (TOMS) Vol. 6, No. 3, 437-443, September 1980.

[LMMR] P. Lins, P. Meade, C. Mehl, and L. Rodman, *Normal Matrices and Polar Decompositions in Indefinite Inner Products*, Linear and Multilinear Algebra 49, 45-89, 2001.

[LUG] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide, 3rd Ed.*, SIAM, Philadelphia, PA, 1999.

[MERR] C. Mehl, A.C.M. Ran, and L. Rodman, *Polar decomposition of normal operators in indefinite inner product spaces*, submitted for publication, 2004.

[MMX] C. Mehl, V. Mehrmann, and H. Xu, *Canonical forms for doubly structured matrices and pencils*, Electron. J. Linear Algebra 7, 112-151, 2000.

[MRR] C.V.M. van der Mee, A.C.M. Ran, and L. Rodman, *Stability of self-adjoint square roots and polar decompositions in indefinite scalar product spaces*, Linear Algebra Appl. 302-303, 77-104, 1999.

[PJ] C.R. DePrima, and C.R. Johnson, *The Range of $A^{-1}A^*$ in $\mathbf{G}L(n, \mathbf{C})$*, Linear Algebra Appl. 9, 209-222, 1974.

[R] J.D. Roberts, *Linear Model Reduction and Solution of the Algebraic Riccati Equation by Use of the Sign Function*, Int. J. Control 32, 677-687, 1980 (Reprint of Technical Report No. TR-13, CUED/B-Control, Cambridge University, Engineering Department, 1971).

[S] P. Schönemann, *A generalized Solution of the Orthogonal Procrustes Problem*, Psychometrika, Vol. 31, No. 1, 1-10, 1966.

[ST1] J. Stoer, *Numerische Mathematik 1 (5. Aufl.)*, Springer, Berlin, 1989.

[ST2] J. Stoer und R. Bulirsch, *Numerische Mathematik II (3. Aufl.)*, Springer, Berlin, 1990.

[SU] R.U. Sexl und K.H. Urbantke, *Gravitation und Kosmologie: Eine Einführung in die allg. Relativitätstheorie (3. Aufl.)*, BI-Wissenschaftsverlag, Mannheim, 1987.

[T] W.S. Torgerson, *Theory and methods of scaling*, Wiley, New York, 1958.

[WEY] H. Weyl, *Raum, Zeit, Materie: Vorlesungen über allg. Relativitätstheorie (7. Aufl.)*, Springer, Berlin, 1988.

[WED] J.H.M. Wedderburn, *Lectures on Matrices*, AMS, Vol. 17, New York, 1934.

[YH] G. Young, and A.S. Householder, *Discussion of a set of points in terms of their mutual distances*, Psychometrika, Vol. 3, No. 1, 19-22, 1938.