An Extended QR-Solver for Large Profiled Matrices

M. Ruess[1,*], P.J. Pahl[2]

[1] Technische Universität München, Chair for Computation in Engineering, Arcisstraße 21, 80333 Munich, Germany

[2] Technische Universität Berlin, Fachgebiet Theoretische Methoden der Bau- und Verkehrstechnik, Gustav-Meyer-Allee 25, 13355 Berlin, Germany

Abstract

A new method for the solution of the standard eigenvalue problem with large symmetric profile matrices is presented. The method is based on the well-known QR-method for dense matrices. A new, flexible and reliable extension of the method is developed that is highly suited for the independent computation of any set of eigenvalues. In order to analyze the weak convergence of the method in the presence of clustered eigenvalues, the QR-method is studied. Two effective, stable and numerically cheap extensions are introduced to overcome the troublesome stagnation of the convergence. A repeated preconditioning process in combination with Jacobi rotations in the parts of the matrix with the strongest convergence is developed to significantly improve both local and global convergence. The extensions preserve the profile structure of the matrix. The efficiency of the new method is demonstrated with several examples.

Keywords: QR algorithm; Structured Matrix; Preconditioning; Jacobi; Plate Vibration

1 Introduction

Eigenvalue problems are common in engineering tasks. In particular the prediction of structural stability and dynamic behavior are important aspects of engineering that lead to eigenvalue problems for which a set of successive eigenvalues must be determined. The system matrices that arise from these tasks are typically large (dimension $N > 1000$) and possess some kind of sparsity and/or structure.

For algorithms based on similarity transformations, the structure of the large system matrices is one of the key aspects that determine the success or failure of the method. A significant number of articles on structure preserving QR algorithms emerged over the years that consider the consequences of different algebraic types and properties of matrices as well as the preservation of matrix shape [1, 2, 3, 4]. The publications on matrix shape focus mainly on Hessenberg and tridiagonal structure.

Symmetry can reduce storage and numerical effort by a factor of 2. The exploitation of structure can reduce the total effort by a much larger factor and is therefore of special importance. The classical QR-method for dense matrices [5, 6, 7] nowadays is still the method of choice for small problems [8, 3] since it combines stability and accuracy in an impressive manner. For

[*] Corresponding author:
Technische Universität München, Chair for Computation in Engineering, Arcisstraße 21, 80333 Munich, Germany, email: ruess@mytum.de

the solution of large problems, QR and its manifold and perfected derivatives are important methods for the solution of the reduced eigenvalue problem arising from reduction algorithms ([7, 9, 10] et.al.) as well as from projection methods ([11, 12, 13, 14] et.al). Particularly solvers for tridiagonal matrices can be found in all modern software packages of Linear Algebra [15, 16]. The developments of the past decades clearly favor QR as a suitable method for tridiagonal matrices since implicit shift and decoupling strategies significantly reduce the complexity of the calculations [5, 7, 12, 17, 13]. Compared to this area of application, the developments of QR-solvers for structured and banded matrices are less prominent. Already in 1971, ten years after Francis et.al. established the QR theory, Wilkinson and Reinsch [18] and later Parlett [5] documented the preservation of a band structure of the matrix, an insight that was not pursued in the following years. In 1995 Arbenz and Golub published a paper on matrix shapes that are invariant under the symmetric QR algorithm [2]. They pointed out that the QR method does not produce fill-in for decoupled diagonal blocks. For coupled diagonal blocks, they analysed the fill-in and proved the preservation of convex profile structures on a mathematical basis. Problems of numerical accuracy, particularly in the presence of multiple and clustered eigenvalues were not investigated. In numerical practice these aspects prove to be the significant difficulties for the successful application of the method.

In this paper we present the theory and implementation of a method for the eigenvalue computation of real symmetric matrices which utilizes the preservation of the natural convex profile of such matrices to develop a reliable and robust numerical procedure. In order to reduce the numerically expensive QR-decompositions, frequent preconditioning of the iterated matrix and a very effective local iteration scheme are introduced stepwise. These extensions of the QR-method may reduce the effort up to 35%, at the same time preserving and exploiting the stable and accurate nature of QR. These extensions also prove to be very advantageous for the local and global convergence in presence of multiple and clustered eigenvalues as illustrated with several examples.

This paper is organized as follows. In the remainder of this introduction, we briefly outline the concept of the QR iteration method in order to establish the necessary formulae. The preservation of a convex profile structure during decomposition of $\mathbf{A}$ into the product $\mathbf{QR}$ and its inverse recombination $\mathbf{RQ}$ is illustrated in section 2. Section 3 presents some aspects of local and global convergence of the iteration that explain the low rate of convergence in presence of clustered and poorly separated eigenvalues. With these insights the strategies for the extensions of the method are developed stepwise in section 4. The effects of the new developments on the accuracy and stability of the new method is demonstrated with several numerical examples in section 5. The paper closes with conclusions and a discussion of the results in section 6.

1.1 QR iteration

The concepts and developments of the presented extensions of the QR-method are introduced in the following for the real symmetric case. With matrix $\mathbf{A} \in \mathbb{R}^{(N \times N)}$ being real and symmetric the standard eigenvalue problem has N real eigenvalues $\lambda_i (i = 1, \dots, N)$. The corresponding eigenvectors $\mathbf{x}_i$ are linearly independent and form an orthonormal basis of the associated vectorspace $\mathcal{R}^N$ ([5, 7] et.al.) .

The standard eigenvalue problem

$$\mathbf{A}\,\mathbf{x}_i \;=\; \lambda_i\;\mathbf{x}_i \tag{1}$$

is transferred into Schur form

$$\mathbf{\Lambda}\,\mathbf{e}_i \;=\; \lambda_i\;\mathbf{e}_i \tag{2}$$

with

$$\begin{aligned} \mathbf{\Lambda} \;&:=\; \mathbf{X}^T\,\mathbf{A}\,\mathbf{X} && \text{Diagonalform with eigenvalues of } \mathbf{A} && (3)\\ \mathbf{x}_i \;&:=\; \mathbf{X}\,\mathbf{e}_i && \text{Eigenvector for eigenvalue } \lambda_i && (4)\end{aligned}$$

by a sequence of similarity transformations. The coefficients of $\mathbf{\Lambda}$ are the eigenvalues of eq. (1) in sorted order.

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{k-1}| \geq |\lambda_k| \geq |\lambda_{k+1}| \geq \ldots \geq |\lambda_n| \tag{5}$$

The eigenvectors of eq.(2) are unit vectors $\mathbf{e}_i$. The eigenvectors of $\mathbf{A}$ are the columns of $\mathbf{X}$ (eq.(4)) in sorted order corresponding to (5). The orthonormal matrix $\mathbf{X}$ is determined iteratively.

In each step s the spectral shifted matrix $(\mathbf{A}_s - \omega\mathbf{I})$ is decomposed into the product of $\mathbf{Q}_s$ and a right triangular matrix $\mathbf{R}_s$.

$$\mathbf{Q}_s\,\mathbf{R}_s \;:=\; \mathbf{A}_s - \omega\mathbf{I} \qquad\qquad \text{with } \mathbf{Q}_s\,\mathbf{Q}_s^T \;=\; \mathbf{Q}_s^T\,\mathbf{Q}_s \;=\; \mathbf{I} \tag{6}$$

The similarity transform is completed by recombination of the decomposition product (6) in inverse order.

$$\begin{aligned} \mathbf{A}_{s+1} \;&:=\; \mathbf{R}_s\,\mathbf{Q}_s + \omega\mathbf{I} && (7)\\ &=\; \mathbf{Q}_s^T\,\mathbf{A}_s\,\mathbf{Q}_s && (8)\end{aligned}$$

In the limit iterate $\mathbf{A}_{s+1}$ tends to diagonal form $\mathbf{\Lambda}$ (3) with coefficients (5) whereas the accumulated product of $\mathbf{Q}_s$ tends to the eigenmatrix $\mathbf{X}$. Spectral shifting is used to increase the rate of convergence from linear to cubic [5, 19, 20]. Without loss of generality the shift parameter ω is assumed zero in the following sections 2, 3 and 4.

2 Preservation of a convex profile structure

Instead of storing $\mathbf{A}$ in banded form it is advantageous to use its natural convex profile structure for the convergence acceleration as further explained in section 4. The profile structure is described with the following notation (Fig. 1):

1. $pl[i]$ and $pr[i]$ denote the left and right profile of row i storing the first and last non-zero entry of row i, respectively.

2. Due to symmetry of $\mathbf{A}$ the left profile in row i corresponds the upper profile in column i. Similarly the right profile in row i corresponds the lower profile in column i.

3. The profile of $\mathbf{A}$ is said to be convex, if the following holds for $m \geq i$:

$$pl[m] \geq pl[i] \quad \text{and} \quad pr[m] \geq pr[i] \tag{9}$$

4. The mean bandwidth of the convex profile matrix $\mathbf{A}$ is denoted with b.

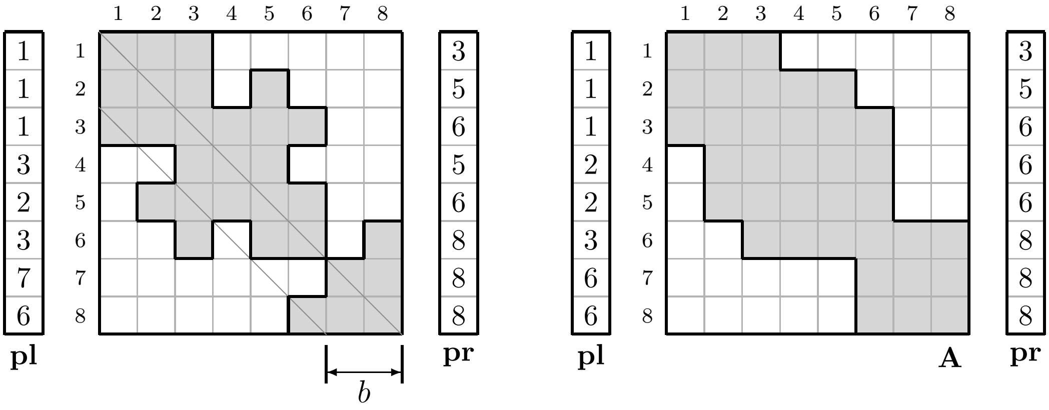


Figure 1: General (left) and convex (right) matrix profile

In general, the profile of $\mathbf{A}$ is not completely convex. The necessary extension to convex storage structure typically requires less than 1% additional storage space.

2.1 QR-decomposition

In step s of the iteration, matrix $\mathbf{A}_s$ is decomposed into the product $\mathbf{QR}$ by a stepwise reduction of $\mathbf{A}_s$ to a triangular form $\mathbf{R}_s$. The reduction process is carried out by columnwise premultiplication with a sequence of plane rotation matrices $\mathbf{P}^T_{ik}$, each eliminating a coefficient a_{ik} below the main diagonal (eq. (10)) as illustrated in Figure 2.

$$\mathbf{R}_s = \ldots \mathbf{P}^T_{pr[2],2} \cdots \mathbf{P}^T_{32} \mathbf{P}^T_{pr[1],1} \cdots \mathbf{P}^T_{21} \mathbf{A}_s = \mathbf{Q}^T_s \mathbf{A}_s \tag{10}$$

$$\mathbf{Q}_s = \mathbf{P}_{21} \ldots \mathbf{P}_{pr[1],1} \mathbf{P}_{32} \ldots \mathbf{P}_{pr[2],2} \cdots \tag{11}$$

Premultiplying $\mathbf{A}_s$ with rotation $\mathbf{P}^T_{ik}$ affects only the coefficients of rows i and k in the column range between $pl[i]$ and $pr[pr[k]]$. The profile of row k is temporarily extended from $pr[k]$ to $pr[pr[k]-1]$ with r $(r \leq b)$ matrix elements. The extension is cancelled after the coefficients in column k have been reduced to zero. The calculation of the last significant coefficient $\hat{a}_{i,pr[i]}$ in row i depends only on the coefficient $a_{k,pr[i]}$ if $pr[i-1] = pr[i]$. For $pr[i-1] < pr[i]$ coefficient $a_{k,pr[i]}$ equals zero. Hence, the number of additionally stored coefficients of the temporarily extended domain of row k depends on the right profile of row $i-1$, not on the right profile of

row i. Due to the columnwise approach and the convexity of the left and right profiles $\mathbf{pl}$ and $\mathbf{pr}$, the profile structure is retained during decomposition.
Figure 2 illustrates the elimination step for coefficient a_{52} during the decomposition of $\mathbf{A}_s$. With $k = 2$ and $i = 5$ the premultiplication of $\mathbf{P}_{52}^T$ only changes the shaded coefficients in rows 2 and 5, resulting in the modified matrix $\hat{\mathbf{A}}_s$ with $\hat{a}_{52} = 0$ [1]. Row 2 is used to eliminate all subdiagonal elements a_{i2} in column 2 in the range between row $k+1 = 3$ and row $pr[k] = 5$. Thus the right profile $\mathbf{pr}$ of row 2 is extended from $pr[k] = 5$ to $pr[pr[k]-1] = 6$.

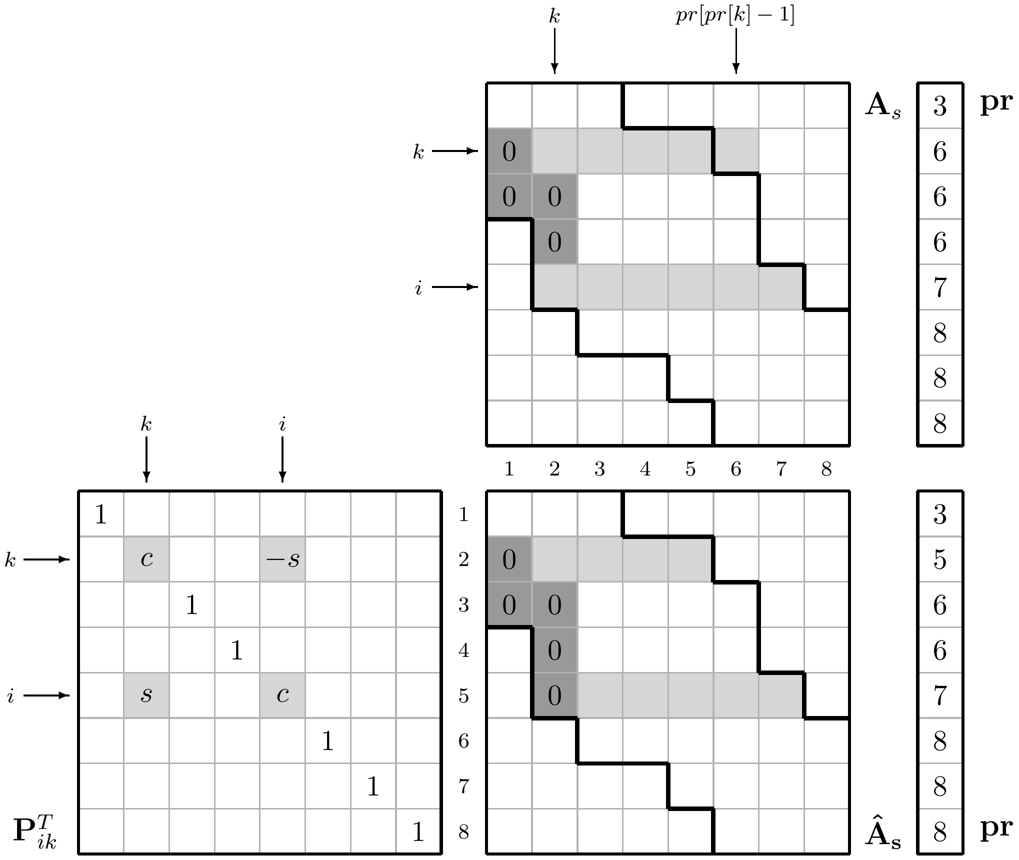


Figure 2: Decomposition of $\mathbf{A}_s$, Elimination of coefficient a_{52} : $\hat{\mathbf{A}}_s = \mathbf{P}_{52}^T \mathbf{A}_s$

2.2 RQ-recombination

Calculation of $\mathbf{A}_{s+1}$ starts with the destruction of zero-coefficient a_{21} by post-multiplying rotation $\mathbf{P}_{21}$ to $\mathbf{R}_s$. Continuing in the sequence of (11) rotation $\mathbf{P}_{ik}$ destroys the zero-coefficient a_{ik} and affects only the coefficients in columns i and k in the range between rows $(k+1)$ and $pr[k]$.

[1] symbol ˆ denotes a modified value

The coefficients $\hat{a}_{km}$ of row k with $m > k$ are implicitly determined by symmetry and therefore are not calculated (Fig. 3).
Figure 3 illustrates the destruction of the zero-coefficient a_{52} during the recombination of $\mathbf{A}_{s+1}$. With $k = 2$ and $i = 5$ the postmultiplication of $\mathbf{P}_{52}$ changes only the shaded coefficients in columns 2 and 5, resulting in the modified matrix $\hat{\mathbf{R}}_s$.

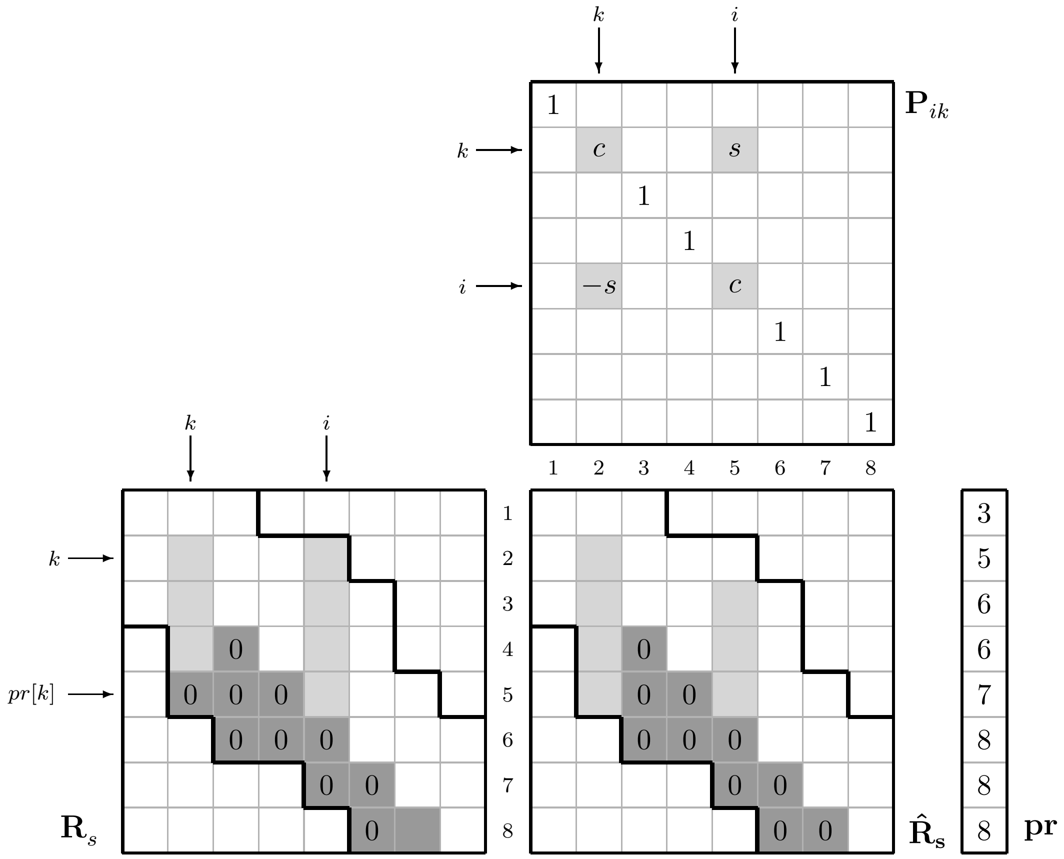


Figure 3: Recombination of $\mathbf{A}_{s+1}$, Destruction of zero-coefficient a_{52}: $\hat{\mathbf{R}}_s = \mathbf{R}_s \mathbf{P}_{52}$

3 Proof of convergence

In order to illustrate the convergence behavior of the QR iteration and to identify potential sources of stagnation and failure of the method, we show a complete formal proof for the real symmetric case.
The proof of the convergence of the basic QR algorithm can be found in numerous papers [20, 21, 22], many of them with focus on the various shift strategies or the different types of matrices like unitary Hessenberg or real symmetric tridiagonal [5, 23, 24]. Whereas Parlett [21] and

Wilkinson [20] reveal the global convergence behavior of the algorithm, a local convergence behavior is documented by Stewart [25] and Parlett in [5] et.al.
Our aim is to give a structured and formal proof in a form that reveals the potential for further developments to improve the convergence of separated, multiple and clustered eigenvalues. The proof shows that the numerical characteristics of the solution procedure depend essentially on the specific treatment that is chosen for the different types of block matrices which can occur in the course of the iteration. The proof is split into the following stages.

1. As a first step we prove convergence for well-separated eigenvalues of $\mathbf{A}$. The non-uniqueness of QR-decompositions of quadratic real matrices is shown and used in the following for the sake of generality.

2. In step two we extend our proof to the more general case by introducing multiple eigenvalues and eigenvalues of equal modulus but opposite sign.

3.1 Separated eigenvalues $|\lambda_i| \neq |\lambda_j| \neq 0$

Theorem 3.1 *Let the symmetric matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *have the QR-decompositions* $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$ *and* $\mathbf{A} = \mathbf{Q}_2\mathbf{R}_2$*. With* $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *being orthonormal and* $\mathbf{R}_1$ *and* $\mathbf{R}_2$ *having upper triangular form, for* $\mathbf{Q}_1 \neq \mathbf{Q}_2$ *and* $\mathbf{R}_1 \neq \mathbf{R}_2$ *the QR-decomposition is not unique.*

Proof Since $|\lambda_i| \neq |\lambda_j| \neq 0$ $\mathbf{R}_1$ and $\mathbf{R}_2$ are non-singular and therefore invertible. The Inverse $\mathbf{R}_1^{-1}$ is upper triangular, so is the product $(\mathbf{R}_2\,\mathbf{R}_1^{-1})$.

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2, \quad \text{with } \mathbf{Q}_i^T\mathbf{Q}_i = \mathbf{I} \tag{12}$$

$$\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{R}_2\mathbf{R}_1^{-1} := \mathbf{E} \tag{13}$$

Both products $(\mathbf{Q}_2^T\mathbf{Q}_1)$ and $(\mathbf{R}_2\mathbf{R}_1^{-1})$ are orthonormal. Due to the latter $\mathbf{E}$ is diagonal with arbitrary coefficients $e_{ii} = +1$ and $e_{ii} = -1$, respectively.

$$\mathbf{E}^T\mathbf{E} = (\mathbf{Q}_2^T\mathbf{Q}_1)^T(\mathbf{Q}_2^T\mathbf{Q}_1) = \mathbf{I}, \quad \text{with } e_{ii} = \pm 1 \tag{14}$$

With $(\mathbf{Q}_1 = \mathbf{Q}_2\mathbf{E})$ and $(\mathbf{R}_1 = \mathbf{E}\mathbf{R}_2)$ in (12) the QR-decomposition of $\mathbf{A}$ is not unique.

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = (\mathbf{Q}_2\mathbf{E})(\mathbf{E}\mathbf{R}_2) \tag{15}$$

□

With $\mathbf{E}$ from theorem 3.1 the columns of $\mathbf{Q}$ and the corresponding rows of $\mathbf{R}$ are of arbitrary algebraic sign. It is shown in the following, that the optional choice of the algebraic sign of the diagonal coefficients of $\mathbf{R}$ by a phase matrix $\mathbf{E}$ does not influence the overall convergence of the iteration.

Lemma 3.1 *In step s of the iteration the accumulated decomposition factors $\mathbf{P}_s$ and $\mathbf{U}_s$ correspond to the $s - th$ power of $\mathbf{A}$.*

The more generalized form of the QR-decomposition (15) is analyzed in cycle s of the iteration in order to find the relation between $\mathbf{A}$ and $\mathbf{A}_s$. In cycle s the product of the orthonormal matrices $(\mathbf{Q}_s\mathbf{E}_s)$ as well as the product of the right triangular matrices $(\mathbf{E}_s\mathbf{R}_s)$ is formed.

$$\begin{aligned} \mathbf{A}_s &= (\mathbf{Q}_s\mathbf{E}_s)(\mathbf{E}_s\mathbf{R}_s) \qquad (16)\\ \mathbf{P}_s &= \mathbf{Q}_1\mathbf{E}_1\ldots\mathbf{Q}_s\mathbf{E}_s \qquad (17)\\ \mathbf{U}_s &= \mathbf{E}_s\mathbf{R}_s\ldots\mathbf{E}_1\mathbf{R}_1 \qquad (18)\end{aligned}$$

Equation (16) may now be used to form the iterate $\mathbf{A}_{s+1}$:

$$\begin{aligned} \mathbf{A}_{s+1} &= \mathbf{E}_s\mathbf{R}_s\mathbf{Q}_s\mathbf{E}_s &&= \mathbf{E}_s\mathbf{Q}_s^T\mathbf{A}_s\mathbf{Q}_s\mathbf{E}_s \qquad (19)\\ \mathbf{A}_{s+1} &= \mathbf{E}_s^T\mathbf{Q}_s^T\ldots\mathbf{E}_1^T\mathbf{Q}_1^T\mathbf{A}\mathbf{Q}_1\mathbf{E}_1\ldots\mathbf{Q}_s\mathbf{E}_s &&= \mathbf{P}_s^T\mathbf{A}\mathbf{P}_s \qquad (20)\end{aligned}$$

Finally relation (20) is used to express the product $(\mathbf{P}_s\,\mathbf{U}_s)$ in terms of $\mathbf{A}$.

$$\begin{aligned} \mathbf{P}_s\mathbf{U}_s &= \mathbf{Q}_1\mathbf{E}_1\ldots\mathbf{Q}_s\mathbf{E}_s\mathbf{E}_s\mathbf{R}_s\ldots\mathbf{E}_1\mathbf{R}_1 \qquad (21)\\ &= \mathbf{P}_{s-1}\mathbf{A}_s\mathbf{U}_{s-1} \qquad (22)\\ &= \mathbf{A}\mathbf{P}_{s-1}\mathbf{U}_{s-1} \;=\; \mathbf{A}^2\mathbf{P}_{s-2}\mathbf{U}_{s-2} \;=\; \ldots \qquad (23)\\ \mathbf{P}_s\mathbf{U}_s &= \mathbf{A}^s \qquad (24)\end{aligned}$$

Since the real and symmetric matrix $\mathbf{A}$ has N real eigenvalues λ_i and N linearly independent eigenvectors $\mathbf{x}_i$, the $s - th$ power of $\mathbf{A}$ may be decomposed into the eigen-representation :

$$\mathbf{A}^s \;=\; (\mathbf{X}\,\mathbf{E})\,\mathbf{\Lambda}^s\,(\mathbf{X}\,\mathbf{E})^T, \quad \text{with } \mathbf{X}\,\mathbf{X}^T = \mathbf{X}^T\mathbf{X} \;=\; \mathbf{I} \qquad (25)$$

The eigenvalues λ_i are the non-zero coefficients of the diagonal matrix $\mathbf{\Lambda}$ in arbitrary order. The corresponding eigenvectors $\mathbf{x}_i$ are the columns of the eigenmatrix $\mathbf{X}$.
Lemma 3.1 and the eigendecomposition (25) of $\mathbf{A}^s$ are used to show by comparison the convergence of $\mathbf{A}_s$ to diagonal form.

Theorem 3.2 *If the transpose of the eigenmatrix $(\mathbf{X}\,\mathbf{E})^T$ has a unique decomposition into the product of a left triangular matrix $\mathbf{L}$(with $l_{ii} = 1$) and a right triangular matrix $\mathbf{U}$ then the convergence of $\mathbf{A}_s$ in iteration step s depends on the following assumptions :*

1. *The eigenvalues λ_i of $\mathbf{A}$ are separated and ordered in the sequence of descending order $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$.*
2. *$\mathbf{A}^s$ is definite and therefore has only eigenvalues $\lambda_i \neq 0$.*

Proof In order to show the independence of the convergence from the algebraic sign of the eigenvalues λ_i, the eigenvalue matrix $\mathbf{\Lambda}$ is decomposed into the product of the unit matrix $\mathbf{F}$ and a diagonal matrix $\mathbf{\Lambda}_+$ containing only positive diagonal values.

$$\mathbf{\Lambda} \;=\; \mathbf{F}\,\mathbf{\Lambda}_+, \quad \text{with } \lambda_{+_{ii}} \;>\; 0 \tag{26}$$

With $(\mathbf{X}\,\mathbf{E})^T \;=\; \mathbf{L}\,\mathbf{U}$ and (26), equation (25) leeds to :

$$\mathbf{A}^s \;=\; \mathbf{X}\,\mathbf{E}\,(\mathbf{F}\,\mathbf{\Lambda}_+)^s\,\mathbf{L}\,\mathbf{U} \tag{27}$$

$$\;=\; (\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\mathbf{C}_s)\,(\mathbf{\Lambda}_+^s\,\mathbf{U}) \tag{28}$$

$$\mathbf{C}_s \;=\; \mathbf{\Lambda}_+^s\,\mathbf{L}\,\mathbf{\Lambda}_+^{-s}, \qquad \lim_{s\to\infty}\mathbf{C}_s \;=\; \mathbf{I} \tag{29}$$

The product $(\mathbf{\Lambda}_+^s\,\mathbf{U})$ still has triangular form. In general the product $(\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\mathbf{C}_s)$ is not orthonormal since $\mathbf{C}_s$ has triangular form, thus not being orthonormal. But with assumption *(1)* of this theorem, $\mathbf{C}_s$ in the limit converges to the identity matrix.

$$\mathbf{C}_s \;=\; \begin{array}{|c|c|c|c|}\hline 1 & 0 & 0 & \cdots \\ \hline l_{21}\left[\frac{\lambda_2}{\lambda_1}\right]^s & 1 & 0 & \cdots \\ \hline l_{31}\left[\frac{\lambda_3}{\lambda_1}\right]^s & l_{32}\left[\frac{\lambda_3}{\lambda_2}\right]^s & 1 & \cdots \\ \hline \vdots & \vdots & \vdots & \ddots \\ \hline \end{array}$$

Figure 4: Leading convergence matrix $\mathbf{C}_s$

With (29), equation (28) is in the limit a QR-decomposition. The comparison of (28) with (24) leads to the decomposition factors of (16), thus showing the convergence of $\mathbf{A}_s$ to diagonal form $\mathbf{\Lambda}$:

$$\lim_{s\to\infty}\mathbf{A}_s \;=\; (\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s)(\mathbf{\Lambda}_+^s\mathbf{U}) \;=\; \mathbf{P}_s\mathbf{U}_s \tag{30}$$

The comparison of the decomposition factors is carried out stepwise :

1. The orthonormal factor converges in the limit to a unit matrix :

$$\mathbf{P}_s \;=\; \mathbf{P}_{s-1}\mathbf{Q}_s\,\mathbf{E}_s \tag{31}$$

$$\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s \;=\; \mathbf{X}\,\mathbf{E}\,\mathbf{F}^{s-1}\,\mathbf{Q}_s\,\mathbf{E}_s \tag{32}$$

$$\lim_{s\to\infty}\mathbf{Q}_s\,\mathbf{E}_s \;=\; \mathbf{F} \tag{33}$$

2. The right triangular factor converges in the limit to diagonal form :

$$\mathbf{U}_s = \mathbf{E}_s\mathbf{R}_s\,\mathbf{U}_{s-1} \tag{34}$$

$$\mathbf{\Lambda}_+^s\,\mathbf{U} = \mathbf{E}_s\mathbf{R}_s\,\mathbf{\Lambda}_+^{s-1}\,\mathbf{U} \tag{35}$$

$$\lim_{s\to\infty} \mathbf{E}_s\,\mathbf{R}_s = \mathbf{\Lambda}_+ \tag{36}$$

With (33) and (36) the limit of (16) is given by :

$$\lim_{s\to\infty} \mathbf{A}_s = \lim_{s\to\infty} (\mathbf{Q}_s\,\mathbf{E}_s)(\mathbf{E}_s\,\mathbf{R}_s) = \mathbf{F}\,\mathbf{\Lambda}_+ = \mathbf{\Lambda} \tag{37}$$

□

Matrix $\mathbf{C}_s$ (eq. 29) reveals the linear convergence rate of the iteration that clearly stagnates in presence of poorly separated eigenvalue clusters. Furthermore, with the descending order of the eigenvalues as assumed in theorem 3.2, convergence is fastest for λ_n in the last row and column n of $\mathbf{A}_s$.
For theorem 3.2 a descending order of the eigenvalues on the diagonal of $\mathbf{\Lambda}$ as well as a unique triangular decomposition $\mathbf{L}\,\mathbf{U}$ of the eigenmatrix $(\mathbf{X}\,\mathbf{E})^T$ were assumed. There are cases where the latter is not unique. The following theorem shows that this is accompanied by disorder of the eigenvalues on the diagonal of $\mathbf{\Lambda}$.

Theorem 3.3 *Let* $(\mathbf{X}\,\mathbf{E})^T = \hat{\mathbf{X}}^T$ *be the transpose of the eigenmatrix of equation* (25)*. The rows of* $\hat{\mathbf{X}}^T$ *are the scaled eigenvectors of* $\mathbf{A}$ *and therefore linearly independent, thus making* $\hat{\mathbf{X}}^T$ *non-singular. If the decomposition of* $\hat{\mathbf{X}}^T$ *into a left triangular matrix* $\mathbf{L}$(*with* $l_{ii} = 1$) *and a right triangular matrix* $\mathbf{U}$ *is not possible, there exists a triangular decomposition of a permutation of* $\hat{\mathbf{X}}^T$*. The permutation* $\mathbf{T}$ *directly influences the ordering of the converging eigenvalues.*

Proof The triangular decomposition of $\hat{\mathbf{X}}^T$ stops with the zero element $u_{ii} = 0$ on the diagonal of $\mathbf{U}$ since in general the equation for coefficient $x_{i+1,i}$ leads to a contradiction. In order to counteract the abortion of the decomposition, row i is exchanged with the first row $j > i$ that leads to $u_{ii} \neq 0$. The row interchange is carried out by a permutation matrix $\mathbf{T}$ that modifies only the rows of $\mathbf{L}$.

$$\mathbf{T}(\mathbf{X}\,\mathbf{E})^T = \mathbf{T}\,\mathbf{L}\,\mathbf{U}, \quad \text{with } \mathbf{T}\,\mathbf{T}^T = \mathbf{I} \tag{38}$$

With the decomposition of the permuted eigenmatrix $\mathbf{T}\hat{\mathbf{X}}^T$ the convergence of $\mathbf{C}_s$ changes to :

$$\mathbf{A}^s = \mathbf{X}\,\mathbf{E}\,\mathbf{\Lambda}^s\,\mathbf{T}\,\mathbf{L}\,\mathbf{U} \tag{39}$$

$$= \mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\mathbf{T}\,\hat{\mathbf{C}}_s\,\hat{\mathbf{\Lambda}}_+^s\,\mathbf{U} \tag{40}$$

$$\hat{\mathbf{C}}_s = \hat{\mathbf{\Lambda}}_+^s\,\mathbf{L}\,\hat{\mathbf{\Lambda}}_+^{-s}, \quad \text{with } \lim_{s\to\infty} \hat{\mathbf{C}}_s = \mathbf{I} \tag{41}$$

$$\hat{\mathbf{\Lambda}}_+^s = \mathbf{T}^T\,\mathbf{\Lambda}_+^s\,\mathbf{T} \tag{42}$$

$$\hat{\mathbf{C}}_s = \begin{array}{|c|c|c|c|c|c}
\multicolumn{1}{c}{} & \multicolumn{1}{c}{\overset{k}{\downarrow}} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \\
\hline
1 & 0 & 0 & 0 & \cdots & \\
\hline
\ddots & \ddots & 0 & 0 & \cdots & \\
\hline
\ddots & l_{jk}\left[\frac{\lambda_j}{\lambda_k}\right]^s & 1 & 0 & \cdots & \leftarrow j \\
\hline
\cdots & l_{ik}\left[\frac{\lambda_i}{\lambda_k}\right]^s & l_{ij}\left[\frac{\lambda_i}{\lambda_j}\right]^s & 1 & \ddots & \leftarrow i \\
\hline
\vdots & \vdots & \ddots & \ddots & \ddots & \\
\hline
\end{array}$$

Figure 5: Leading convergence matrix $\hat{\mathbf{C}}_s$

With assumption $j > i$ for the row interchange during the decomposition of $(\mathbf{X}\,\mathbf{E})^T$ and $k < i$ (Figure 5) follows $k < j$ and therefore assuring convergence of coefficients $\hat{c}_{ik}$ and $\hat{c}_{jk}$ to zero. For $j = k$ in the swapped row i it follows that $\hat{c}_{ij} = 0$ since $l_{ij} = 0$. In the limit $\hat{\mathbf{C}}_s$ again converges to an identity matrix. Permutation $\mathbf{T}$ changes the order of the eigenvalues on the diagonal of $\boldsymbol{\Lambda}$ but does not influence the convergence of the iteration. □

3.2 Multiple eigenvalues and multiple eigenvalues of opposite sign $|\lambda_i| = |\lambda_k|$

Theorem 3.4 *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ have a p-fold eigenvalue λ, a q-fold eigenvalue $-\lambda$ and $(N-p-q)$ separated eigenvalues. Then the iterated matrix $\mathbf{A}_s$ converges to a block diagonal matrix $\boldsymbol{\Lambda}_b$ with diagonal blocks of size $(p+q)\times(p+q)$.*

Proof With assumption (43) the diagonal matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_+$ from (26) have the following form :

$$\boldsymbol{\Lambda} = \begin{array}{|c|c|c|c|}
\hline
\boldsymbol{\Lambda}_1 & & & \\
\hline
 & \lambda\mathbf{I} & & \\
\hline
 & & -\lambda\mathbf{I} & \\
\hline
 & & & \boldsymbol{\Lambda}_2 \\
\hline
\end{array}
\qquad
\boldsymbol{\Lambda}_+ = \begin{array}{|c|c|c|c|}
\hline
\boldsymbol{\Lambda}_{1_+} & & & \\
\hline
 & \lambda\mathbf{I} & & \\
\hline
 & & \lambda\mathbf{I} & \\
\hline
 & & & \boldsymbol{\Lambda}_{2_+} \\
\hline
\end{array}$$

Figure 6: Eigenvalue distribution for $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_+$

$$\begin{aligned}|\lambda_1| \geq \ldots \geq |\lambda_{k-1}| \geq |\lambda_{k_1}| = \ldots = |\lambda_{k_p}| = |-\lambda_{k_1}| = \ldots \\ = |-\lambda_{k_q}| \geq |\lambda_{k+1}| \geq \ldots \geq |\lambda_n|\end{aligned} \tag{43}$$

The leading convergence matrix $\mathbf{C}_s$ from (29) will no longer converge to pure diagonal form. Instead $\mathbf{C}_s$ converges in the rows associated with the p-fold eigenvalue λ and the q-fold eigenvalue $-\lambda$ to lower triangular form of dimension $(p+q)$ with diagonal coefficients $l_{ii} = 1$ (Figure 7).

$$\lim_{s\to\infty} \mathbf{C}_s \;=\; \tilde{\mathbf{L}} \tag{44}$$

Thus convergence in the presence of multiple eigenvalues and multiple eigenvalues of opposite sign solely depends on the coefficients l_{ik} $(i \neq k)$ of the triangular submatrix of $\tilde{\mathbf{L}}$. For equation (30) follows :

$$\lim_{s\to\infty} \mathbf{A}^s \;=\; (\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\tilde{\mathbf{L}})(\mathbf{\Lambda}_+^s\mathbf{U}) \;=\; \mathbf{P}_s\mathbf{U}_s \tag{45}$$

Because of $\tilde{\mathbf{L}}$ the product $(\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\mathbf{C}_s)$ in (28) is no longer orthonormal. $\tilde{\mathbf{L}}$ is decomposed into the product of an orthonormal matrix $(\hat{\mathbf{Q}}\,\hat{\mathbf{E}})$ and an upper triangular matrix $(\hat{\mathbf{E}}\,\hat{\mathbf{R}})$.

$$\lim_{s\to\infty} \mathbf{A}^s \;=\; (\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\hat{\mathbf{Q}}\,\hat{\mathbf{E}})(\hat{\mathbf{E}}\,\hat{\mathbf{R}}\,\mathbf{\Lambda}_+^s\mathbf{U}) \;=\; \mathbf{P}_s\mathbf{U}_s \tag{46}$$

Convergence of $\mathbf{A}_s$ to diagonal form is shown in analogy to theorem 3.2 :

1. The orthonormal factor again converges in the limit to a unit matrix :

$$\mathbf{P}_s \;=\; \mathbf{P}_{s-1}\mathbf{Q}_s\,\mathbf{E}_s \tag{47}$$

$$\mathbf{X}\,\mathbf{E}\,\mathbf{F}^s\,\hat{\mathbf{Q}}\,\hat{\mathbf{E}} \;=\; \mathbf{X}\,\mathbf{E}\,\mathbf{F}^{s-1}\,\hat{\mathbf{Q}}\,\hat{\mathbf{E}}\,\mathbf{Q}_s\,\mathbf{E}_s \tag{48}$$

$$\lim_{s\to\infty} \mathbf{Q}_s\,\mathbf{E}_s \;=\; (\hat{\mathbf{Q}}\,\hat{\mathbf{E}})^T\,\mathbf{F}\,(\hat{\mathbf{Q}}\,\hat{\mathbf{E}}) \;=\; \mathbf{I} \tag{49}$$

2. The right triangular factor converges in the limit to diagonal form :

$$\mathbf{U}_s \;=\; \mathbf{E}_s\mathbf{R}_s\,\mathbf{U}_{s-1} \tag{50}$$

$$\hat{\mathbf{E}}\,\hat{\mathbf{R}}\,\mathbf{\Lambda}_+^s\,\mathbf{U} \;=\; \mathbf{E}_s\mathbf{R}_s\,\hat{\mathbf{E}}\,\hat{\mathbf{R}}\,\mathbf{\Lambda}_+^{s-1}\,\mathbf{U} \tag{51}$$

$$\lim_{s\to\infty} \mathbf{E}_s\,\mathbf{R}_s \;=\; \hat{\mathbf{E}}\,\hat{\mathbf{R}}\,\mathbf{\Lambda}_+(\hat{\mathbf{E}}\,\hat{\mathbf{R}})^{-1} \tag{52}$$

With (49), (52) and the relation $(\hat{\mathbf{E}}\,\hat{\mathbf{R}})^{-1} = \tilde{\mathbf{L}}^{-1}\hat{\mathbf{Q}}\,\hat{\mathbf{E}}$ it follows for the limit in equation (16) :

$$\begin{aligned}
\lim_{s\to\infty} \mathbf{A}_s &= \lim_{s\to\infty} (\mathbf{Q}_s\,\mathbf{E}_s)(\mathbf{E}_s\,\mathbf{R}_s) && (53)\\
\lim_{s\to\infty} \mathbf{A}_s &= (\hat{\mathbf{E}}\,\hat{\mathbf{Q}}\,\mathbf{F}\,\hat{\mathbf{Q}}\,\hat{\mathbf{E}})(\hat{\mathbf{E}}\,\hat{\mathbf{R}}\,\boldsymbol{\Lambda}_{+}\hat{\mathbf{R}}^{-1}\,\hat{\mathbf{E}}) && (54)\\
&= \hat{\mathbf{E}}\,\hat{\mathbf{Q}}\,\mathbf{F}\,\tilde{\mathbf{L}}\,\boldsymbol{\Lambda}_{+}\tilde{\mathbf{L}}^{-1}\,\hat{\mathbf{Q}}^{T}\,\hat{\mathbf{E}} && (55)\\
&= \hat{\mathbf{E}}\,\hat{\mathbf{Q}}\,\boldsymbol{\Lambda}\,\hat{\mathbf{Q}}^{T}\,\hat{\mathbf{E}} \;=\; \boldsymbol{\Lambda}_b && (56)
\end{aligned}$$

$$\begin{array}{|c|c|c|c|}\hline \mathbf{I} & & & \\ \hline & \tilde{\mathbf{L}}_{11} & & \\ \hline & \tilde{\mathbf{L}}_{21} & \tilde{\mathbf{L}}_{22} & \\ \hline & & & \mathbf{I} \\ \hline \end{array}\ \tilde{\mathbf{L}} \qquad \begin{array}{|c|c|c|c|}\hline \mathbf{I} & & & \\ \hline & \hat{\mathbf{Q}}_{11} & \hat{\mathbf{Q}}_{12} & \\ \hline & \hat{\mathbf{Q}}_{21} & \hat{\mathbf{Q}}_{22} & \\ \hline & & & \mathbf{I} \\ \hline \end{array}\ \hat{\mathbf{Q}} \qquad \begin{array}{|c|c|c|c|}\hline \hat{\mathbf{E}}_1 & & & \\ \hline & \hat{\mathbf{E}}_2 & & \\ \hline & & \hat{\mathbf{E}}_3 & \\ \hline & & & \hat{\mathbf{E}}_4 \\ \hline \end{array}\ \hat{\mathbf{E}}$$

Figure 7: Matrices $\tilde{\mathbf{L}}, \hat{\mathbf{Q}}, \hat{\mathbf{E}}$

With the matrices shown in Figures 6 and 7 the coefficients of the $\boldsymbol{\Lambda}_b$ are determined as follows :

$$\begin{aligned}
\mathbf{A}_{11} &= \lambda\,\hat{\mathbf{E}}_2\,(\hat{\mathbf{Q}}_{11}\,\hat{\mathbf{Q}}_{11}^{T} - \hat{\mathbf{Q}}_{12}\,\hat{\mathbf{Q}}_{12}^{T})\,\hat{\mathbf{E}}_2 && (57)\\
\mathbf{A}_{12} &= \lambda\,\hat{\mathbf{E}}_2\,(\hat{\mathbf{Q}}_{11}\,\hat{\mathbf{Q}}_{21}^{T} - \hat{\mathbf{Q}}_{12}\,\hat{\mathbf{Q}}_{22}^{T})\,\hat{\mathbf{E}}_3 \;=\; \mathbf{A}_{21}^{T} && (58)\\
\mathbf{A}_{22} &= \lambda\,\hat{\mathbf{E}}_3\,(\hat{\mathbf{Q}}_{21}\,\hat{\mathbf{Q}}_{21}^{T} - \hat{\mathbf{Q}}_{22}\,\hat{\mathbf{Q}}_{22}^{T})\,\hat{\mathbf{E}}_3 && (59)
\end{aligned}$$

$$\begin{array}{|c|c|c|c|}\hline \boldsymbol{\Lambda}_1 & & & \\ \hline & \mathbf{A}_{11} & \mathbf{A}_{12} & \\ \hline & \mathbf{A}_{12}^{T} & \mathbf{A}_{22} & \\ \hline & & & \boldsymbol{\Lambda}_2 \\ \hline \end{array}$$

Figure 8: Limit matrix $\boldsymbol{\Lambda}_b$

Depending on the assignment of the coefficients of $\boldsymbol{\Lambda}$ the following cases of convergence have to be distinguished :

1. *All eigenvalues λ_i are separated :* $\mathbf{\Lambda}_b = \mathbf{\Lambda}$, no diagonal blocks exist since $\hat{\mathbf{Q}} = \mathbf{I}$
2. $\mathbf{A}$ *has a p-fold eigenvalue λ :* $\mathbf{\Lambda}_b$ only contains the submatrices $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ and $\mathbf{A}_{11}$. With the orthonormality of the submatrices $\hat{\mathbf{Q}}_{11}$, $\mathbf{A}_{11}$ converges to diagonal form.

$$\mathbf{A}_{11} = \lambda\, \hat{\mathbf{E}}_2\, \hat{\mathbf{Q}}_{11}\, \hat{\mathbf{Q}}_{11}^T\, \hat{\mathbf{E}}_2 = \lambda\, \mathbf{I} \tag{60}$$

3. $\mathbf{A}$ *has a q-fold eigenvalue $-\lambda$:* $\mathbf{\Lambda}_b$ only contains the submatrices $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ and $\mathbf{A}_{22}$. In analogy to case 2, $\mathbf{A}_{11}$ converges to diagonal form.

$$\mathbf{A}_{22} = -\lambda\, \hat{\mathbf{E}}_3\, \hat{\mathbf{Q}}_{22}\, \hat{\mathbf{Q}}_{22}^T\, \hat{\mathbf{E}}_3 = -\lambda\, \mathbf{I} \tag{61}$$

4. $\mathbf{A}$ *has a p-fold eigenvalue λ and q-fold eigenvalue $-\lambda$:* $\mathbf{\Lambda}_b$ converges to block diagonal form with the submatrices $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ defined in equations (57) to (59).

□

4 Extensions

4.1 Preconditioning

The leading convergence matrix $\mathbf{C}_s$ clearly shows the critical domain of the iterated matrix in the presence of clusters of poorly separated eigenvalues. With convergence rates $|\lambda_i/\lambda_k|^s$ $(k < i)$ near 1.0, the convergence of the off-diagonal elements a_{ik} in rows and columns i and k towards zero stagnates despite the intensive use of sophisticated shift strategies. As a result of this local convergence lag the iterated matrix tends to diagonal dominance in parts other than the desired last rows and columns of the unreduced matrix. A monotonic convergence behavior is completely destroyed.

In order to overcome this troublesome convergence behavior and to remove local perturbance from the iteration the critical lower diagonal part of the matrix is frequently preconditioned with locally bounded similarity transformations. In step s the critical coefficients are reduced to zero within a single transformation. Though subsequent QR-iterations destroy the zero entries it can be observed that in most cases the modified coefficients are significantly smaller than the original values.

The preconditioning is based on a blockwise scheme of similarity transformations with orthonormal matrices $\mathbf{U}_k \in \mathbb{R}^{N\times N}$ that differ from the identity matrix $\mathbf{I}$ by an orthonormal block of size $(p \times p)$ on the diagonal. Thus the transformation

$$\hat{\mathbf{A}}_s = \mathbf{U}_k^T\, \mathbf{A}_s\, \mathbf{U}_k \qquad k = 1, \ldots, n,\ n \ll N \tag{62}$$

only affects coefficients a_{im} in rows and columns $i, i+1, \ldots, i+p$ and $m, m+1, \ldots, m+p$, respectively, of the iterated Matrix $\hat{\mathbf{A}}_s$.

In order to retain the matrix profile the block dimension p is chosen so that the transformation only affects rows and columns with constant profile $\mathbf{pl}$ and $\mathbf{pr}$, respectively. The number n of blocks and their location on the diagonal of $\mathbf{U}_k$ is therefore determined and limited by structural aspects rather than by the computational need. Even so, they influence the convergence positively as is shown in the numerical examples.

Figure 9 shows the partitioned and iterated matrix $\mathbf{A}_s$ with its diagonal blocks arising from the natural convex profile structure of the matrix. Each of these block zones provide a diagonal matrix $\mathbf{A}_k^{p \times p}$ that is readily diagonalized with little effort.

$$\mathbf{A}_k^{p \times p} \;=\; \mathbf{V}_k \, \mathbf{D}_k \, \mathbf{V}_k^T \tag{63}$$

The orthonormal matrix $\mathbf{V}_k^{p \times p}$ in (63) is used to diagonalize $\mathbf{A}_k^{p \times p}$in step s, thus removing the critical coefficients from the subsequent computation. The block size p is strongly influenced by the convex profile structure and therefore by the number of degrees of freedom of the discrete mesh points. A slight extension of the profile in order to increase the preconditioning effect often turns out to be inexpensive in regard to the convergence and the numerical effort. Block sizes typically vary between $p = 3$ for plane problems and $p = 12$ for spatial problems, respectively, up to $p = b/5$, with typical mean bandwidth b of $\sqrt{N}$ in 2D and $\sqrt[3]{N^2}$ in 3D.

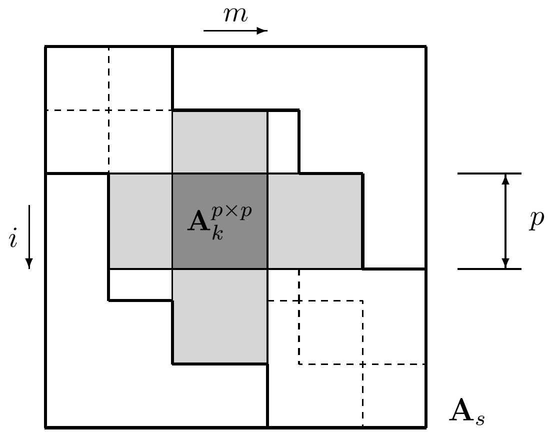


Figure 9: Partitioning of $\mathbf{A}_s$ in blocks with constant profile

In general matrix $\mathbf{A}_k^{p \times p}$ is symmetric, real and definite and therefore may contain positive and negative eigenvalues and even multiple eigenvalues with opposite sign, the latter being a very rare situation. The QR-algorithm for full matrices is suited for the diagonalization of $\mathbf{A}_k^{p \times p}$ but may also be replaced by the numerically stable cyclic Jacobi method. The latter may lead to the loss of the ordering of the vectors of $\mathbf{U}_k^{p \times p}$ corresponding to a descending ordering of the eigenvalues of $\mathbf{A}$.

4.2 Jacobi correction

Several parts of $\mathbf{A}$ are not covered by the preconditioning (Fig. 10). There are mainly three effects that still may cause slow convergence behavior :

1. The admissible block size on the diagonal of $\mathbf{A}$ is smaller than 2. Preconditioning would irretrievably destroy the profile structure of the adjacent rows and columns.
2. Adjacent diagonal blocks lock out zones due to their convexity.
3. Between the diagonal block and the left and upper margin of the profile there exist zones that may not profit from the eigeninformation of $\mathbf{A}_k^{p \times p}$.

In order to counteract stagnation of the convergence the critical parts of the matrix are iterated locally with Jacobi transformations.

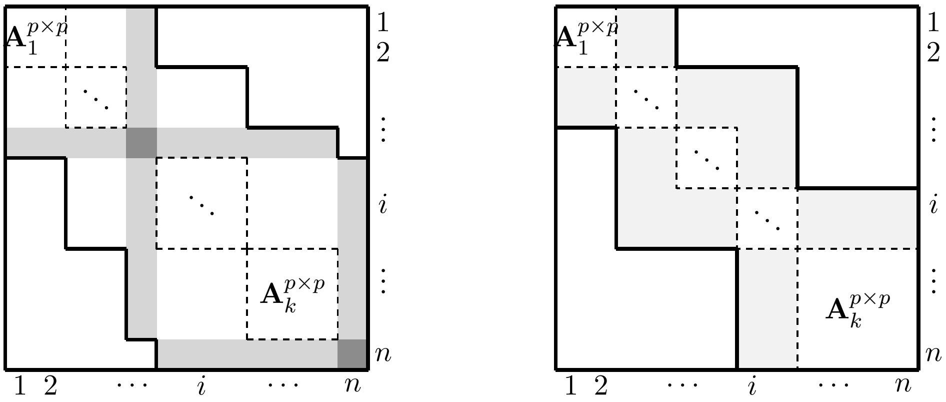

$\mathbf{A}_1^{p\times p}$
$\mathbf{A}_k^{p\times p}$
1
2
$\vdots$
i
$\vdots$
n
1 2 $\cdots$ i $\cdots$ n
$\mathbf{A}_1^{p\times p}$
$\mathbf{A}_k^{p\times p}$
1
2
$\vdots$
i
$\vdots$
n
1 2 $\cdots$ i $\cdots$ n

Figure 10: Insufficiently preconditioned zones of $\mathbf{A}$

In iteration cycle s the values of large magnitude in the last row and column n of $\mathbf{A}_s$ are successively eliminated by Jacobi rotations (64)

$$\mathbf{H}^T \mathbf{A}_s \mathbf{H} = \hat{\mathbf{A}}_s \tag{64}$$

$$\mathbf{H}^T \mathbf{H} = \mathbf{H} \mathbf{H}^T = \mathbf{I} \tag{65}$$

In general coefficients $\hat{a}_{nm} = \hat{a}_{mn}$ of the transformed matrix $\hat{\mathbf{A}}$ do not stay zero, since they are destroyed by subsequent transformations. At a sufficient level of convergence they are however significantly smaller than coefficients $a_{nm} = a_{mn}$ of $\mathbf{A}$. The orthonormal matrix $\mathbf{H}$ is implicitly constructed as the product of orthonormal matrices $\mathbf{H}_{nm}$(66) each representing a rotation θ in the (x, y)-plane.

$$\mathbf{H} = \mathbf{H}_{n,n-1} \mathbf{H}_{n,n-2} \mathbf{H}_{n,n-3} \cdots \tag{66}$$

The rotation angle θ_{mn} is chosen so that $(\mathbf{H}_{nm}^T \mathbf{A}_s \, \mathbf{H}_{nm})$ eliminates coefficients $a_{nm} = a_{mn}$ [20, 19]. The Jacobi rotations are numerically very cheap and may be applied to all coefficients of the last row and column in each cycle, though the effect of cleaning the row and column of strongest convergence from perturbing large coefficients is already noticeable for the elimination of a few coefficients of largest magnitude (Example 5.1, Figure 12). Jacobi rotations are used effectively in our implementation for coefficients with absolute value below an empirically chosen threshold δ. Variations in the factor δ in (67) by order of magnitude 10 do not influence the convergence of the iteration significantly.

$$\delta = \|\mathbf{A}[n]\|_\infty / 10 \tag{67}$$

In general the Jacobi rotations preserve the profile structure of the last k rows and columns of $\mathbf{A}$ if and only if the profile over these rows is constant. In many cases this situation is given only for a limited number of rows and columns. But, due to the fact that coefficients that are close to the profile border converge faster than coefficients that are close to the main diagonal of the matrix (cp. Figure 4), the profile can be readjusted continuously during the iteration within the allocated storage scheme, thus significantly improving the situation. The coefficients in the essential border regions are often below the threshold for off-diagonal elements after a few QR-decompositions.

4.3 Implementation aspects

We give a brief outline of the extended QR-Algorithm and discuss some implementation aspects concerning the extensions in more detail.

Algorithm
For $s = 1, \ldots, max\ s$:

1. Spectral shifting and deflation : $\mathbf{A}_s \leftarrow \mathbf{A}_s - c_s \mathbf{I}$
   (a) Several shift strategies are investigated in [5, 19] et.al. to ensure quadratic and cubic convergence rates. With increasing convergence the last diagonal element of the iterated matrix provides a simple yet effective choice as shift parameter.

   (b) The off-diagonal coefficients of the last row and column of $\mathbf{A}_s$ converge rapidly towards zero and are set to zero once they fall below a carefully chosen threshold value ε [7, 5].
2. QR-decomposition : $\mathbf{A}_s \leftarrow \mathbf{Q}_s^T \mathbf{A}_s$
3. RQ-recombination : $\mathbf{A}_s \leftarrow \mathbf{A}_s \, \mathbf{Q}_s$
4. Preconditioning : $\mathbf{A}_s \leftarrow \mathbf{U}_k^T \mathbf{A}_s \, \mathbf{U}_k$
   (a) The transformations of the preconditioning are applied directly to the corresponding rows and columns of the iterated matrix $\mathbf{A}_s$. The transformation matrix $\mathbf{U}_k$ is never formed explicitly. Instead $\mathbf{U}_k$ is implicitly applied to $\mathbf{A}_s$ by the set of plane rotation

matrices $\mathbf{P}_i$ that is necessary to diagonalize the diagonal block $\mathbf{A}_k^{p\times p}$ by shifted QR-decompositions.

(b) It turns out that for the diagonalization of $\mathbf{A}_k^{p\times p}$ a number of iterations in the range of the block size p is sufficient for effective preconditioning. Hence the progress of convergence is not observed explicitly.

(c) In order to avoid a repeated time consuming search process for domains that are suited for preconditioning the necessary information about the relevant diagonal blocks, size and location, are determined from the profile vectors $\mathbf{pl}$ and $\mathbf{pr}$ before entering the QR-iteration loop s. A readjustment of the block information due to profile changes resulting from an increasing convergence of the off-diagonal coefficients can be advantageous if a larger number of eigenvalues has to be computed.

(d) Preconditioning is most effective for the parts of $\mathbf{A}_s$ where convergence has already set in and is therefore limited to the lower third of the matrix.

5. Jacobi correction : $\mathbf{A}_s \leftarrow \mathbf{H}^T \mathbf{A}_s \, \mathbf{H}$

(a) The Jacobi rotations $\mathbf{H}_{ni}$ are directly applied to the last row and column n of $\mathbf{A}_s$.

(b) The decision to remove coefficients with Jacobi rotations is based solely on structural criteria. For a constant profile in rows $n, \ldots, n-k$ Jacobi is applied regardless of the size of the coefficients. For different profiles in rows n and i the transformation with matrix $\mathbf{H}_{ni}$ to remove coefficient a_{ni} is carried out only if the coefficients in row i do not affect the coefficients in row n outside the profile $pl[n]$ (see 4.2).

(c) The sequence of the coefficients in row/column n that are removed with Jacobi rotations is chosen according to the direction of increasing convergence of that row and column, starting at index $n-1$.

4.4 Complexity

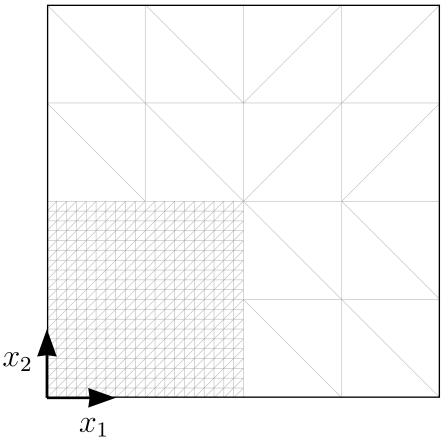
The construction of $\mathbf{R}$ during the decomposition of $\mathbf{A}$ requires $\mathcal{O}(b^2\, N)$ arithmetic floating point operations. During the decomposition of $\mathbf{A}^{N\times N}$ into $\mathbf{QR}$ symmetry cannot be exploited efficiently resulting in $\sim\; 6\, b^2\, N$ multiplications (b: mean bandwidth). The use of structure and symmetry in the recombination $\mathbf{RQ}$ may reduce the effort to $\sim\; 2\, b^2\, N$ multiplications. The factors 6 and 2 account for the stepwise reduction and recombination, respectively.
The preconditioning procedure on selected parts of the iterated matrix $\mathbf{A}$ is mainly influenced by the chosen block size p. The calculation of the transformation matrix $\mathbf{U}$ and the similarity transformation $(\mathbf{U}^T\mathbf{AU})$ requires $\sim\; (40\, p^3 + 4p^2 b + \mathcal{O}(p^2))$ multiplications for each preconditioned block of $\mathbf{A}$. Restricted to the lower and fastest converging part of the matrix, each preconditioning cycle results in $\mathcal{O}(N)$ multiplications.
With $\sim\; (k\, b)$ multiplications the numerical effort of the Jacobi correction is very low since it is limited to the last k rows of $\mathbf{A}$. Even for a repeated use of this method in each cycle the total effort usually stays below 1% of the decomposition effort.

5 Numerical examples

The numerical experiments which have been performed with the software implementation of our method show that the success in the practical application of the extended QR-method depends significantly on the specific treatment of the aspects that are presented in this paper, especially the convergence in presence of multiple and clustered eigenvalues. The examples shed light on the convergence and reliability of the iteration as well as the accuracy of the calculated results. All examples presented here are computed using FELiNA, an object-oriented FE-platform for linear and nonlinear analysis, developed at the Technische Universität Berlin [26].

All examples come from oscillation analyses of thin square plates with lumped mass distribution. Due to symmetry their vibration is characterized by a large number of multiple and clustered eigenvalues, thus being a severe test case for eigenvalue computation. Spectral shifting and deflation of converged eigenstates is intensively used for all computations. Both, preconditioning and Jacobi correction are a fixed part of the algorithm and applied in each iteration cycle (see 4.3).

5.1 Example 1 - Convergence behavior: Computation of the complete Eigenspectrum $\sigma(\mathbf{A})$ for a simply supported square plate, dimension $N = 47$



dimensions	:	$[0,1] \times [0,1]$
thickness	:	0.1
Youngs modulus	:	1.0
poisson ratio	:	0.0
green density	:	1.0

Figure 11: Sample mesh for the square plate

In the following we demonstrate how the developed extensions positively influence the global convergence behavior of the computation, thus significantly reducing the numerical effort.

The first example of rather small dimension is well suited to demonstrate the stepwise improvement of the convergence behavior by introducing preconditioning and Jacobi-correction. The results of this example may easily be transferred to large-scale problems as shown in Figure 14. Without the aforementioned extensions the complete spectrum of 47 eigenvalues of the simply supported plate completely converge after 97 iteration cycles ($i1$), with a relative accuracy of $\mathcal{O}(\epsilon)$, ($\epsilon \approx 2.22 \times 10^{-16}$ in IEEE double precision). The stagnations in progression ($i1$)

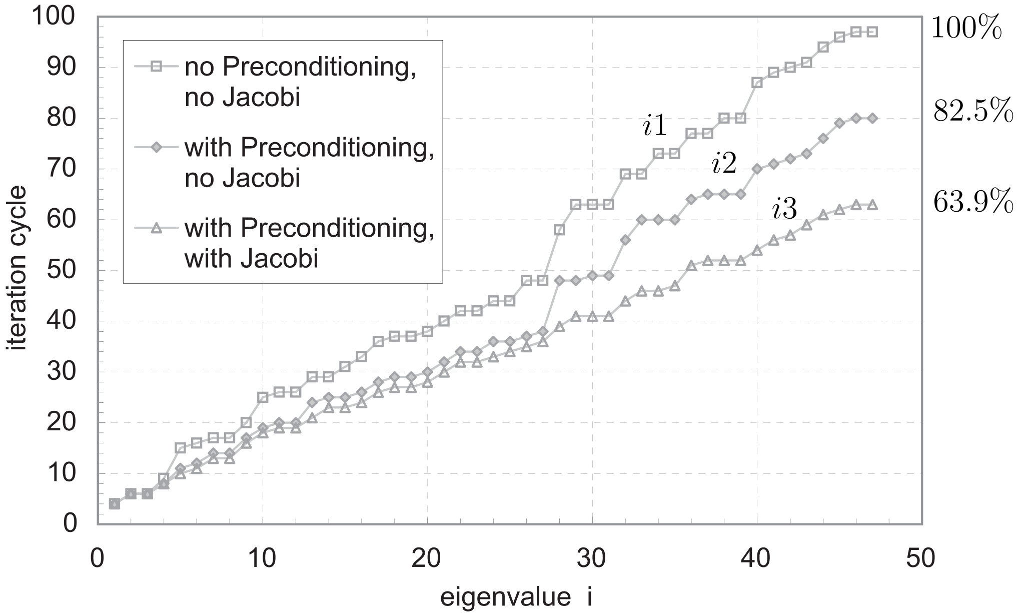


Figure 12: Improved convergence behavior for matrix $\mathbf{K}^{47\times 47}$

clearly identify multiple and clustered eigenvalues with relative separations that easily fall below a threshold of 10^{-4}. With preconditioning the convergence progression of the first half of the spectrum is much smoother. Figure 12 shows that the positive effect of preconditioning sets in already after a few iteration cycles and therefore counteracts the numerically expensive QR decompositions of this early stage. For larger examples some tendency to diagonal dominance is advantageous for the preconditioning to display its full effectiveness. This prerequisite is even more pronounced for larger matrices as can be seen in the essential lower diagonal parts of the matrix after a short starting phase of the iteration. In order to further reduce the numerical effort, the preconditioning is restricted to the part of fastest convergence and incorporates the whole matrix with the ongoing deflation of converged eigenstates. The local and global convergence behavior is essentially improved thus reducing the effort to approximately 85% even for large matrices. The large jumps that are still apparent in curve $i2$ (Figure 12) indicate parts in the matrix that may not profit from preconditioning due to structural deficiencies in the profile structure. They are completely removed with the repeated application of Jacobi-corrections. The numerical effort is further reduced to 64% ($i3$).

The eigenvalue distribution of the aforementioned examples and their relative gaps

$$relgap_i \quad := \quad \underset{i\neq j}{min} \frac{|\lambda_i - \lambda_j|}{\sqrt{(\lambda_i^2 + \lambda_j^2)}} \tag{68}$$

are shown in Figure 13. Eigenvalues with multiplicity 2 are marked with a filled square. The eigenvalue distribution of the second half $(\lambda_{28} - \lambda_{47})$ clearly shows the clustered eigenvalues with

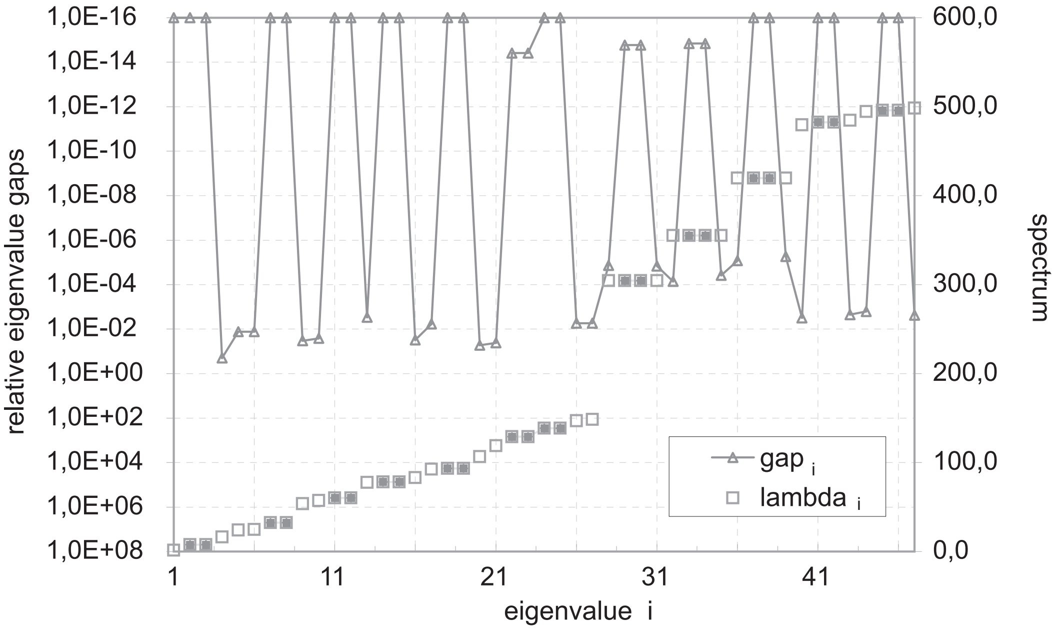


Figure 13: Gaps and eigenvalue distribution of $\mathbf{K}^{47\times 47}$

multiplicity 2. For these clusters the relative gaps between the eigenvalues have the magnitude of their absolute gaps thus explaining the jumps in the convergence progression of curves $i1$ and $i2$ in Figure 12.

5.2 Example 2 - Numerical effort: Computation of the complete eigenspectrum $\sigma(\mathbf{A}_i)$ of a simply supported square plate, various levels of mesh refinement, dimensions $N_i = 47, \ldots, 10687$

The computational results of this example (Figure 14) shows that the improved convergence behavior of the extended algorithm also holds for large matrices. Each square/triangle shows the numerical effort (*floating point operations*) as a percentage of the *pure*[2] QR-iteration for the determination of the complete eigenspectrum $\sigma(\mathbf{A})$ for different mesh refinements (Figure 11). The figure shows that even for small dimensions the numerical effort is reduced below 70%. Approximately a forth of the eigenvalues of all matrices had multiplicity ≥ 2. Convergence of the diagonal element a_{nn} to eigenvalue λ_n was detected by small off-diagonal elements $|a_{ni}| \leq (\|\mathbf{A}\|_\infty \cdot c \cdot \epsilon)$, $i = up[n], \ldots, n-1$ in row n (with c=cycle of iteration).

[2] QR-iteration with deflation and shifting but without introduced extensions of section 4

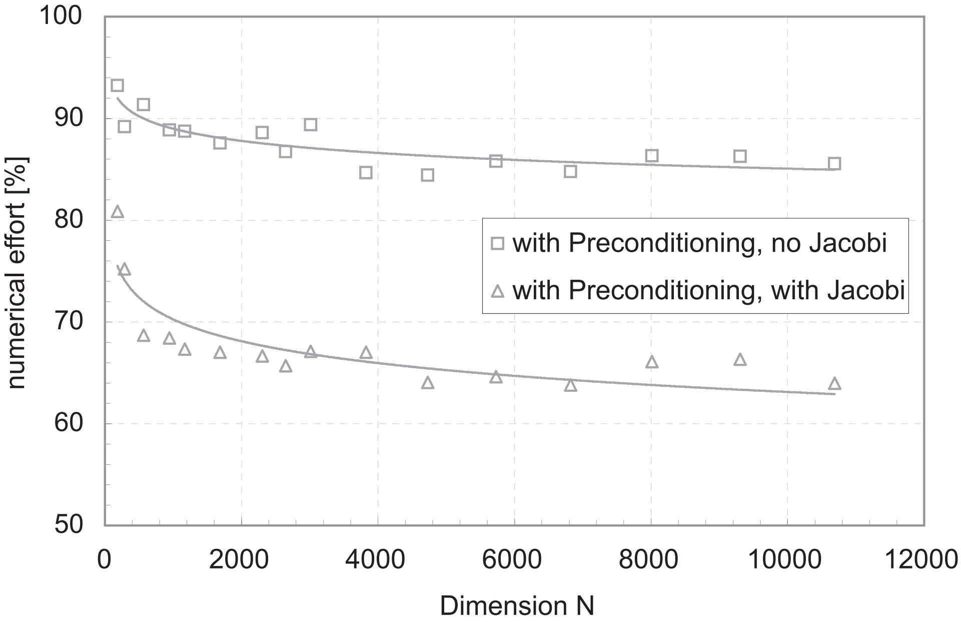


Figure 14: Reduction of the numerical effort for the computation of the complete eigenspectrum $\sigma(\mathbf{A})$

5.3 Example 3 - Accuracy: Computation of a subset of eigenvalues $\psi(\mathbf{A})$ of a simply supported square plate, dimension $N = 5727$

In order to prove the accuracy of the extended method, typical error bounds based on the residual $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \hat{\lambda}\hat{\mathbf{x}}$ are exploited. The indispensable knowledge of the corresponding eigenvector approximations $\hat{\mathbf{x}}$ to the determined spectrum $\sigma(\mathbf{A})$ is obtained from an explicit calculation step by accumulating the $(N \times N)$-transforms $\mathbf{Q}_i, (i = 1 \ldots conv =$ cycle of convergence) to the eigenmatrix $\hat{\mathbf{X}}$. For the practical approach of the method of solution a very reliable and accurate iteration scheme was developed that allows an independent and selective computation of eigenvectors in any range of a known eigenspectrum. The presentation of this method and its accuracy level is beyond the scope of this paper.

The influence of the introduced extensions (Section 4) on the accuracy of the eigenvalues is shown by comparison of the residual error norm $||\mathbf{r}||_2$ resulting from pure, shifted QR iteration and calculations with preconditioning and Jacobi correction. The analysed example has 5727 degrees of freedom and a high number of clustered and multiple eigenvalues. The determined spectrum includes the 1000 algebraically smallest eigenvalues and corresponding eigenvectors. For the sake of clarity the graphical representation of the following Figure 15 is limited to a small number of results in the lower range of the eigenspectrum. The figure clearly shows that the introduced extensions to the QR algorithm do not have negative effects on the accuracy of the approximated eigenpairs. The spread of the residual norm narrows and even averages slightly better than for calculations without the extensions. The figure shows the residual norms of the fifty eigenstates with smallest eigenvalue. The solid line represents the mean value for the extended QR method at $1.743e - 10$, whereas the dashed line at $3.136e - 10$ marks the mean value for pure QR-iteration (both smaller $\epsilon\|\mathbf{A}\|$).

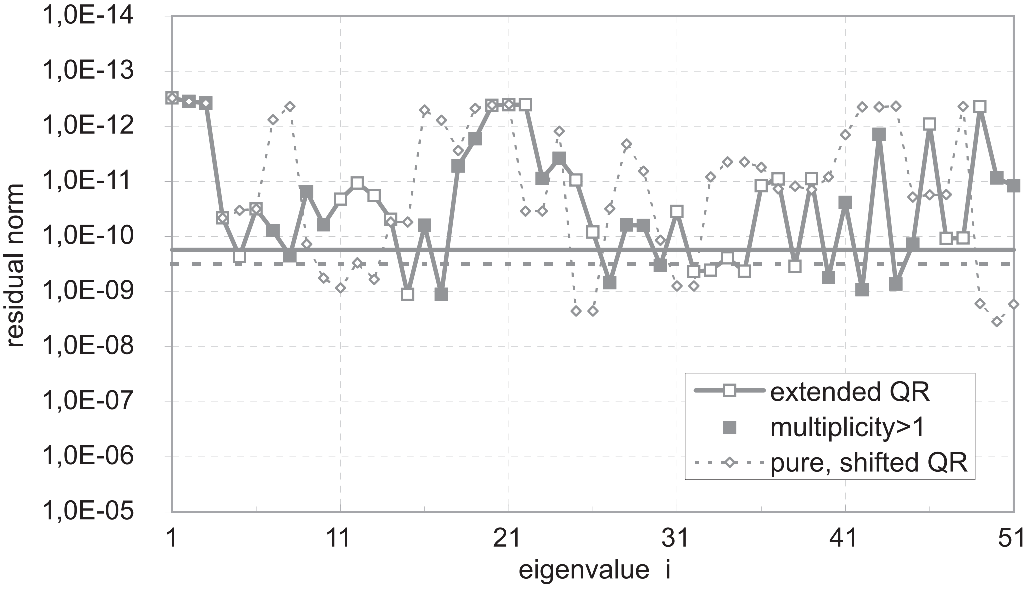


Figure 15: Residual Error Norm $\|\mathbf{r}\|_2$

5.4 Example 4 - Reliability: Computation of a subset of smallest/interior eigenvalues $\psi(\mathbf{A})$ of a structural slab, dimension $N = 6438$

In the following the convergence behavior of the developed method with regard to the partial eigenvalue problem is analysed. The system matrix $\mathbf{A}$ originates from an oscillation analysis of a structural slab. The slab is modelled as a thin plate with the Theory of Kirchhoff using the Finite Element Method and is partly supported by walls and columns (Figure 16). The slab has a length of $40m$, a maximum width of $13m$ and a thickness of $0.21m$. The material has a Young's modulus of $E = 3.0e + 07KN/m^2$ and a Poisson ratio of $\nu = 0.2$. The assumed mass of $m = 2960kg/m^3$ is concentrated at the nodes of the discretized model using a lumping scheme.

5.4.1 Eigenfrequencies $< 30Hz$:

In a first run we determine a subset of eigenvalues of smallest modulus $\psi_S(\mathbf{A})$. The subset contains the 65 eigenfrequencies $f_i(\mathbf{A})$ that are smaller than $30Hz$.
More than 20% of the determined eigenvalues have a relative separation (see eq. 68) of $\mathcal{O}(1.0e - 3)$ or even less, around 72% have a relative separation of $\mathcal{O}(1.0e - 2)$.
The complete subset was determined within 98 iteration cycles (∼1.5 cycles/eigenvalue), with an average restnorm error $\|\mathbf{r}\|_2$ of $\mathcal{O}(\epsilon\, 10\|\mathbf{A}\|_2)$. Compared to the shifted QR-algorithm without

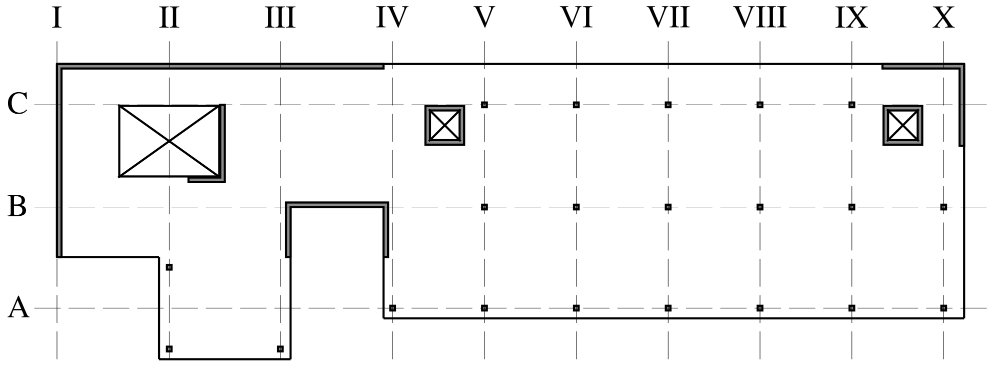


Figure 16: Structural Slab: Geometry and Support

extensions the total effort was reduced to $\sim 80\%$. The computation was carried out for various convergence criteria that lead to different levels of accuracy with a varying total effort but an almost constant effort reduction of $\sim 20\%$. Due to the early convergence of the eigenvalues as a result of the local iteration scheme, partial disorder of some converged values was observed. The completeness of $\psi_S(\mathbf{A})$ was proved using a Sturm sequence check after completion of the iteration. A monitoring of $\psi_S(\mathbf{A})$ during the iteration is not necessary since a disorder of the converging eigenvalues is locally confined and declines with the number of computed eigenvalues.

5.4.2 Eigenfrequency intervals $[400, 420]Hz$ and $[800, 820]Hz$:

In the second run we determine the complete subset of eigenvalues $\psi_I(\mathbf{A})$ in a predefined interval.

The effort for the computation of interior eigenvalues is larger than for the dominant eigenvalues of smallest modulus and essentially depends on the shift technique that requires a more conservative strategy. The interval $[400, 420]Hz$ contains 147 eigenfrequencies that are determined with 238 cycles of the extended QR-method (~1.6 cycles/eigenvalue). The total effort compared to the shifted QR-algorithm without extensions is reduced to $\sim 88\%$. The computation of the eigenfrequencies of a second interval in the range between $800Hz$ and $820Hz$ shows the same tendency. The complete subset of 193 eigenvalues is determined in 308 cycles of the extended QR-method (~1.6 cycles/eigenvalue) that corresponds to a reduction of 85% relative to the total effort required to the shifted QR-algorithm without extensions. The relative separation of the eigenvalues for both intervals lies completely below $\mathcal{O}(1.0e-3)$.

Examples 5.4.1 and 5.4.2 show that the improvement of the numerical effort decreases with the number of computed eigenvalues, a natural consequence of the much lower diagonal dominance in the starting phase of the iteration. Nevertheless the efficiency in all examples is significantly improved by the extensions.

The extended QR-algorithm computes a subset of k eigenvalues in $\mathcal{O}(k\,b^2\,N)$ arithmetic floating point operations using $\mathcal{O}(b\,N)$ storage locations. Despite the numerically expensive QR-decomposition the method is competitive with algorithms like divide and conquer or bisection that typically require a $\mathcal{O}(b\,N^2)$ reduction scheme to tridiagonal form [5, 7, 17, 18].
In [26] the extended QR-algorithm was intensively studied and compared to a powerful implementation of a restarted Lanczos method [27, 14]. The limiting factor for an efficient use of the Lanczos method for a larger number of eigenvalues $(> 1\%)$ is the number of Lanczos vectors needed to approximate an appropriate subspace that contains the desired set of eigenvalues. Even for restarted schemes the number of vectors that have to be kept orthogonal in fast storage often is significantly larger than the number of desired eigenvalues resulting in a dominant reorthogonalization effort. Particularly the computation of interior eigenvalues often requires additional effort to dampen the influence of the eigenvalues of smallest modulus. If only a few eigenvalues are desired $(< 0.5\%)$ the extended QR-algorithm is not competitive with the Lanczos method or its derivates since this is clearly the domain of subspace iteration schemes.

6 Concluding remarks

The classical QR-method for dense matrices is the method of choice for solving the standard eigenvalue problem for matrices of small dimension. In this paper we extended this stable and accurate iteration scheme with its well-known and well understood convergence properties to large-scale problems with real symmetric profile coefficient matrices. With two simple but effective extensions, a repeated preconditioning and a Jacobi correction step, the convergence behavior of the method is significantly improved. Stagnation of the convergence in presence of multiple and clustered eigenvalues is completely removed thus providing a stable and particular continuously converging method that is well suited for the calculation of an arbitrary number of eigenvalues. The extensions do neither destroy a convex profile structure of the coefficient matrix nor the basic property of convergence of the eigenvalues in sorted order and improve the accuracy of the calculated approximates. Symmetry and profile structure of the matrix are efficiently used in the algorithm in order to keep the numerical effort low. The introduced extensions may reduce the numerical effort to 65% of the effort of the shifted QR algorithm without extensions. For subsets of eigenvalues of smallest modulus or interior eigenvalues the developed method still reduces the total effort to around $80 - 90\%$. The stability and reliability of the method make the method attractive for many eigenvalue problems in engineering and science.

References

[1] Delvaux S, Van Barel M. Structures preserved by the QR-algorithm. *Journal of Computational and Applied Mathematics* 2005; **187**: 29-40.

[2] Arbenz P, Golub G. Matrix Shapes Invariant under Symmetric QR Algorithm. *Numerical Linear Algebra with Applications* 1995; **2(2)**: 87-93.

[3] Calvetti D, Kim SM, Reichel L. The restarted QR-algorithm for eigenvalue computation of structured matrices. *Journal of Computational and Applied Mathematics* 2002; **149**: 415-422.

[4] Kailath T, Sayed AH (Eds.). *Fast Reliable Algorithms for Matrices with Structure*. SIAM : Philadelphia, PA: 1999

[5] Parlett BN. *The Symmetric Eigenvalue Problem*. SIAM : Prentice-Hall, Englewood Cliffs, NJ, 1980.

[6] Watkins DS. The QR Algorithm Revisited. *SIAM Review archive* 2008; **50**: 133-145.

[7] Golub GH, Van Loan CF. *Matrix Computations*. Johns Hopkins University Press: Baltimore, 1996.

[8] Watkins DS. QR-like Algorithms for Eigenvalue Problems. *Journal of Computational and Applied Mathematics* 2000; **123**: 67-83.

[9] Lang B. Efficient Algorithms for Reducing Banded Matrices to Bidiagonal and Tridiagonal Form. In Arbenz P, Paprzycki M, Sameh A, and Sarin V (eds), Nova Science Publishers, Commack, NY, *High Performance Algorithms for Structured Matrix Problems, volume 2 of Advances in the Theory of Computation and Computational Mathematics* 1999; 75-89.

[10] Bischof CH, Sun X. *A Framework for Symmetric Band Reduction and Tridiagonalization*. Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, 1992.

[11] Calvetti D, Reichel L, Sorensen DC. An Implicitly Restarted Lanczos Method For Large Symmetric Eigenvalue Problems. *Electronic Transactions on Numerical Analysis* 1994; **2**: 1-21.

[12] Saad Y. Numerical Methods For Large Eigenvalue Problems. *Algorithms for Advanced Scientific Computing*. Manchester University Press Series: Manchester, 1991.

[13] van der Vorst HA. Computational Methods for Large Eigenvalue Problems. In: Ciarlet RG, Lions JL (eds.), *Handbook of Numerical Analysis*, North-Holland, Amsterdam, 2002; **8**: 3-179.

[14] Lehoucq RB, Sorensen DC, Yang C. *ARPACK Users' Guide : Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods* . SIAM : Philadelphia, PA, 1997.

[15] Golub GH, Van der Vorst HA. Eigenvalue Computation in the 20th Century. *Journal of Computational and Applied Mathematics* 2000; **123**: 35-65.

[16] Anderson E, Bai Z, Bischof C, Blackford S, Demmel JW, Dongarra JJ Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen DC. *LAPACK Users' Guide*. SIAM : Philadelphia, PA, 1999.

[17] Demmel J. *Applied Numerical Linear Algebra*. SIAM: Philadelphia, PA, 1997.

[18] Wilkinson JH, Reinsch C. *Handbook for Automatic Computation, Vol.2, Linear Algebra*. Springer Verlag: Heidelberg, Berlin, New York, 1971.

[19] Stewart GW. *Matrix Algorithms, Vol II: Eigensystems*. SIAM: Philadelphia, PA, 2001.

[20] Wilkinson JH. *The Algebraic Eigenvalue Problem*. The Clarendon Press, Oxford University Press : Oxford, 1965

[21] Parlett BN. Convergence of the QR Algorithm. *Numerische Mathematik* 1965; **7**: 187-193.

[22] Huang CP On the Convergence of the QR Algorithm with Origin Shifts for Normal Matrices. *Journal of Numerical Analysis* 1981; **1**: 127-133.

[23] Hoffmann W, Parlett BN. A new proof of global convergence for the tridiagonal QL algorithm. *Journal of Numerical Analysis* 1978; **15**: 929-937.

[24] Wang TL, Gragg WB. Convergence of the Shifted QR Algorithm for Unitary Hessenberg Matrices. *Mathematics of Computation* 2001; **71**: 1473-1496.

[25] Stewart GW. *Afternotes goes to graduate school*. SIAM: Philadelphia, PA, 1987.

[26] Ruess M. Eine Methode zur vollständigen Bestimmung der Eigenzustände großer Profilmatrizen. *Ph.D. Thesis*, Technische Universität Berlin, Berlin, 2005.

[27] Calvetti D, Reichel L, Sorensen DC. An Implicitly Restarted Lanczos Method for Large Symmetric Eigenvalue Problems. *Electronic Transactions on Numerical Analysis* 1994; **2**: 1-21.