GENAUE EIGENWERTBERECHNUNG NICHTSINGULÄRER SCHIEFSYMMETRISCHER MATRIZEN

DISSERTATION
zur
Erlangung des Grades Dr. rer. nat.
des Fachbereiches Mathematik
der Fernuniversität –Gesamthochschule– Hagen

vorgelegt von
Eberhard Pietzsch
aus Bergheim, Erft

Hagen, 1993

Inhaltsverzeichnis

Einleitung **3**

1 Störungstheorie **10**

2 Jacobi-ähnliche Verfahren **21**
2.1 Strukturen bei mehrfachen Eigenwerten 22
2.2 Jacobi-ähnliche Verfahren für das Paar $(L^T L, J)$ 41
2.2.1 Explizites Verfahren 42
2.2.2 Der formale Algorithmus des expliziten Verfahrens 50
2.2.3 Implizites Verfahren 54
2.2.4 Der formale Algorithmus des impliziten Verfahrens 66
2.2.5 Bemerkungen zum impliziten Verfahren 75
2.3 Konvergenz 77
2.3.1 Über das Hadamardsche Maß 77
2.3.2 Globale Konvergenz 87

3 Fehler in berechneten Eigenwerten von $(L^T L, J)$ **91**
3.1 Über die Kondition der skalierten Matrix 91
3.1.1 Wachstumsschranken bei Skalierungen und Rotationen . . 92
3.1.2 Wachstumsschranken bei modifizierten Zyklen 98
3.2 Fehleranalyse des impliziten Verfahrens 101
3.2.1 Bemerkungen zum Skalarprodukt 104
3.2.2 Relative Fehlerschranken bei Skalierungen und Rotationen 108
3.2.3 Relative Fehlerschranken beim modifizierten Zyklus 134

4 Schiefsymmetrische Zerlegung $PSP^T = LJL^T$ **154**
4.1 Der Algorithmus der Zerlegung 156
4.2 Fehleranalyse 159
4.3 Zur Beschränktheit der skal. Kondition von $L^T L$ 162

5 Gesamtfehler in berechneten Eigenwerten **174**

6 Numerische Experimente **178**
6.1 Alternative Verfahren 178
6.1.1 Ein QR-Verfahren 178
6.1.2 Das Verfahren von Paardekooper 179
6.2 Numerische Resultate 179
6.2.1 Zur relativen Genauigkeit 182
6.2.2 Zum Konvergenzverhalten 200

Symbolverzeichnis **204**

Literatur **205**

Einleitung

In der vorliegenden Arbeit beschäftigen wir uns mit dem reellen nichtsingulären schiefsymmetrischen Eigenwertproblem. Nach Einführung in eine Störungstheorie werden wir ein reelles Verfahren zur Eigenreduktion vorstellen, mit dem wir nachstehendes Ziel verfolgen: Haben kleine relative Störungen der Matrixelemente im Sinne der Störungstheorie nur kleine relative Änderungen der Eigenwerte zur Folge, so haben die in endlicher Genauigkeit berechneten Eigenwerte kleine relative Fehler. In diesem Sinne ist das Adjektiv „genau“ im Titel zu verstehen. Die Bestimmung oberer Schranken für die relativen Fehler in den berechneten Eigenwerten ist weiterer zentraler Bestandteil vorliegender Arbeit.

Nichtsinguläre schiefsymmetrische Eigenwertprobleme entstehen beispielsweise aus physikalisch-technischen Problemstellungen. So führt die Betrachtung gyroskopischer Systeme auf das quadratische Eigenwertproblem

$$(\lambda^2 + \lambda X + Y)x = 0$$

mit schiefsymmetrischem nichtsingulärem X und symmetrischem positiv definitem Y. Mit $Y = ZZ^T$ ist das quadratische Eigenwertproblem äquivalent zum nichtsingulären schiefsymmetrischen Eigenwertproblem

$$\begin{bmatrix} 0 & Z^T \\ -Z & -X \end{bmatrix} y = \mu y \ .$$

Wir werden uns in vorliegender Arbeit auf nichtsinguläre Eigenwertprobleme beschränken. Bei singulären Eigenwertproblemen, z.B. solchen ungerader Ordnung, wäre die Betrachtung relativer Genauigkeit der in endlicher Arithmetik berechneten Eigenwerte nicht möglich.

Demmel, Kahan [12] untersuchten die relative Genauigkeit bei der Berechnung von Singulärwerten bidiagonaler Matrizen. Barlow, Demmel [4] verallgemeinerten die Resultate für skaliert diagonaldominante Matrizen und Demmel, Veselić [13] für symmetrische positiv definite Matrizen. Veselić [31] präsentierte ein Jacobiähnliches Verfahren für indefinite symmetrische Matrizen, von dem Slapničar [29] Genauigkeit im obigen Sinne untersuchte. Deichmöller [10] behandelte Verfahren und Genauigkeitsanalyse für verallgemeinerte Singulärwertprobleme. Die Betrachtung schiefsymmetrischer Matrizen in vorliegender Arbeit ist im Lichte

der zitierten Resultate zu sehen.

Das vorzustellende Verfahren basiert auf einer Grundidee von Veselić [31], [32] und besteht im Wesentlichen aus zwei Schritten:

1. Von der gegebenen nichtsingulären Matrix $S \in \mathbb{R}^{2n\times 2n}$, $S = -S^T$, $n \geq 2$ wird eine schiefsymmetrische Faktorisierung

$$PSP^T = LJL^T \quad , \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \quad , \quad I \in \mathbb{R}^{n\times n} \tag{0.1}$$

   mit vollständiger Pivotierung gemäß einer von Bunch [6] angegeben Methode vorgenommen, wobei die Permutationsmatrix P durch die Pivotierung bestimmt wird.

2. Der Faktor L wird so von rechts mit einem symplektischen T (d.h. $T^TJT = J$ bzw. äquivalent $TJT^T = J$) transformiert, daß $LT = U\Delta$ mit unitärem U und diagonalem positiv definitem $\Delta = E \oplus E$ gilt. Dann ist $PSP^T = LJL^T = LTJT^TL^T = U\Delta J\Delta U^T$, d.h. $U^TPSP^TU = \Delta J\Delta$. Die Quadrate der Spaltennormen der resultierenden Matrix $U\Delta$ sind also die positiven Imaginärteile der Eigenwerte von S, während die normierten Spalten die orthogonale Eigenvektormatrix von S bilden. Die Transformation mit T erfolgt in einem iterativen Prozeß.

Der erste Schritt ist eigentlich ein Schritt des LR-Verfahrens. Man kann ihn als Präkonditionierer auffassen. Er bildet die Grundlage, um im zweiten Schritt implizit das zu (S, I) äquivalente Eigenwertproblem des Paares (L^TL, J), d.h.

$$Kx = i\lambda Jx \quad , \quad K = L^TL = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \quad , \quad i^2 = -1 \tag{0.2}$$

zu lösen. Die gegebene schiefsymmetrische Matrix S wird also nicht, wie bei der Methode von Paardekooper [25], orthogonal transformiert. Vielmehr werden die Spalten des Faktors L in einem iterativen Prozeß orthogonalisiert. In diesem zweistufigen Vorgehen liegt gerade der Vorteil der vorgestellten Methode, denn der iterative Prozeß kann als eine stabile Diagonalisierung des positiv definiten K mittels Kongruenztransformationen aufgefaßt werden ([13], [29], [31]): $T^TKT = T^TL^TLT = \Delta U^TU\Delta = \Delta^2$. Darüber hinaus ist in einem Programm die Akkumulation der Transformationsmatrizen zur Bestimmung der Eigenvektoren von S nicht nötig; grundsätzlich wird auf einem einzigen zweidimensionalen Feld gearbeitet. Gegenüber dem Verfahren von Paardekooper spart man dadurch Speicherplatz und Rechenzeit. Wir werden in vorliegender Arbeit eine neue Jacobi-ähnliche Methode für den iterativen Prozeß des obigen zweiten Schrittes vorstellen.

Der zweite Schritt besteht aus einer Folge symplektischer Transformationen

$$L^{(m+1)} = L^{(m)} T^{(m)} \quad , \quad m = 1, 2, \ldots \tag{0.3}$$

mit $L^{(1)} = L$. Die Transformationen kann man wie bei Jacobi-Verfahren üblich zu *Zyklen* gruppieren. Wir werden zwei Arten solcher Zyklen unterscheiden. In *normalen Zyklen* gibt es ausschließlich Skalierungen und orthogonale Rotationen, die sich höchstens in Submatrizen der Ordnung 4 von der Einheitsmatrix unterscheiden. Jeweils einige aufeinanderfolgende Transformationen sind so aufeinander abgestimmt, daß ihre Konkatenation eine Submatrix

$$\begin{bmatrix} F_{pp}^{(m)} & F_{pq}^{(m)} & B_{pp}^{(m)} & B_{qp}^{(m)} \\ F_{pq}^{(m)} & F_{qq}^{(m)} & B_{pq}^{(m)} & B_{qq}^{(m)} \\ B_{pp}^{(m)} & B_{pq}^{(m)} & C_{pp}^{(m)} & C_{pq}^{(m)} \\ B_{qp}^{(m)} & B_{qq}^{(m)} & C_{pq}^{(m)} & C_{qq}^{(m)} \end{bmatrix} \quad , \quad 1 \le p \le n-1 \ , \quad p+1 \le q \le n \ , \tag{0.4}$$

von

$$K^{(m)} = L^{(m)^T} L^{(m)} = \begin{bmatrix} F^{(m)} & B^{(m)^T} \\ B^{(m)} & C^{(m)} \end{bmatrix}$$

diagonalisiert. In der Tat kann man durch geeignete Vortransformation von K sogar

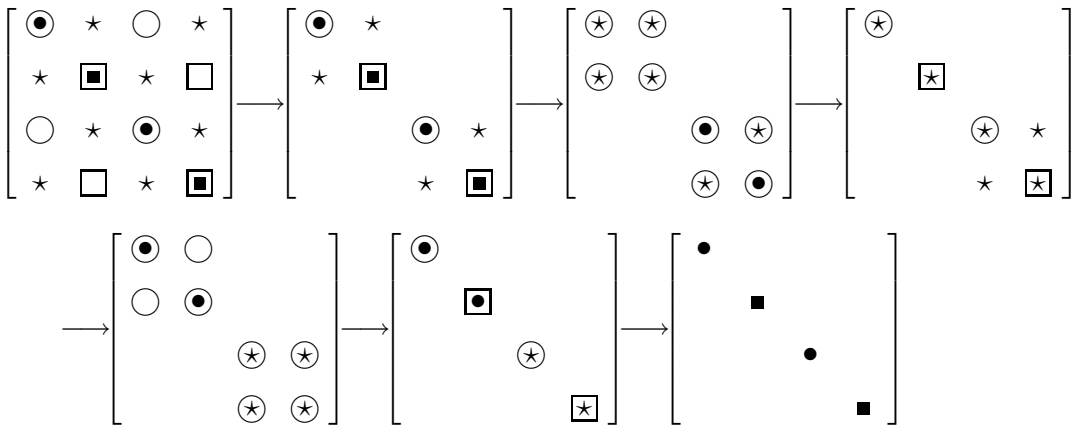
$$F_{pp}^{(m)} = C_{pp}^{(m)} \quad , \quad F_{qq}^{(m)} = C_{qq}^{(m)} \quad , \quad B_{pp}^{(m)} = B_{qq}^{(m)} = 0 \tag{0.5}$$

für sämtliche Submatrizen (0.4) und alle m annehmen. Die Diagonalisierung einer solchen Submatrix wollen wir hier kurz skizzieren. Dazu führen wir eine symbolische Schreibweise für Matrizen der Ordnung 4 ein. Läßt man die Umrahmungen von Positionen zunächst außer Acht, so repräsentiert das linke Symbol von

$$\begin{bmatrix} \circledast & \circledast & & \\ \circledast & \circledast & & \\ & & \odot & \circledast \\ & & \circledast & \odot \end{bmatrix} \longrightarrow \begin{bmatrix} \circledast & & & \\ & \boxed{\star} & & \\ & & \circledast & \star \\ & & \star & \boxed{\star} \end{bmatrix} .$$

eine blockdiagonale Matrix der Ordnung 4, während rechts außerdem der erste Block diagonal ist. Mit Sternen wird die Beliebigkeit von Matrixelementen symbolisiert, während gleiche andere Symbole (in der linken Matrix die beiden gefüllten Kreise) Gleichheit von Elementen darstellen. Mit dem Pfeil deuten wir den Übergang der linken in die rechte Matrix mittels symplektischer Transformation an. Dabei unterscheidet sich die Transformationsmatrix nur in den in der linken Matrix umrahmten Positionen von der Einheitsmatrix, und sie diagonalisiert die erste Blockdiagonale. Die in der rechten Matrix umrahmten Positionen deuten die anschließende Anwendung einer Skalierung an, wobei die Verschiedenheit der Rahmen (Kreis oder Quadrat) auf die Verschiedenheit der Parameterbestimmung hinweist. Mit dieser symbolischen Schreibweise können wir die Diagonalisierung

einer Submatrix (0.4) mit der Eigenschaft (0.5) auf sehr kompakte Weise veranschaulichen:



Die diagonale Matrix hat also wieder die Eigenschaft (0.5). Alle verwendeten nichtdiagonalen Transformationen sind orthogonal. Bei den Transformationen ist zu bedenken, daß Blockdiagonalmatrizen $U \oplus V$ nur für $V = U^{-T}$ symplektisch sind. Daher sind neben den orthogonalen Transformationen Skalierungen notwendig, die der Kontrolle der relevanten Kondition der transformierten Matrizen dienen.

Wir werden die globale Konvergenz der auf diese Weise entstehenden Folge der $K^{(m)}$ gegen eine Diagonalmatrix, d.h. von $L^{(m)}T^{(m)}$ gegen ein $U\Delta$, zeigen. Hat S mehrfache Eigenwerte, so wird sich i.A. jedoch keine asymptotisch quadratische Konvergenz einstellen. Dazu betrachten wir das numerische Beispiel

$$K = \begin{bmatrix} 1.0e+0 & -8.4e-6 & 1.6e-4 & 0 & -2.1e-5 & -8.6e-5 \\ -8.4e-6 & 1.0e+0 & 1.6e-5 & -2.1e-5 & 0 & 8.9e-5 \\ 1.6e-4 & 1.6e-5 & 1.0e+0 & -8.6e-5 & 8.9e-5 & 0 \\ 0 & -2.1e-5 & -8.6e-5 & 1.0e+0 & 8.4e-6 & -1.6e-4 \\ -2.1e-5 & 0 & 8.9e-5 & 8.4e-6 & 1.0e+0 & -1.6e-5 \\ -8.6e-5 & 8.9e-5 & 0 & -1.6e-4 & -1.6e-5 & 1.0e+0 \end{bmatrix}.$$

Gerundet auf eine Dezimalstelle ist hier $\|K - \mathrm{diag}(K)\|_2 = 2.1e-4$ und nach einem wie oben beschriebenen Zyklus gilt $\|\tilde{K} - \mathrm{diag}(\tilde{K})\|_2 = 4.4e-5$ für die transformierte gerundete Matrix

$$\tilde{K} = \begin{bmatrix} 1.0e+00 & 1.1e-16 & 4.1e-07 & -2.2e-17 & 2.1e-17 & -4.3e-05 \\ 1.1e-16 & 1.0e+00 & -1.1e-07 & 2.1e-17 & 4.8e-17 & 1.2e-05 \\ 4.1e-07 & -1.1e-07 & 1.0e+00 & -4.3e-05 & 1.2e-05 & -5.8e-17 \\ -2.2e-17 & 2.1e-17 & -4.3e-05 & 1.0e+00 & 5.6e-17 & -4.1e-07 \\ 2.1e-17 & 4.8e-17 & 1.2e-05 & 5.6e-17 & 1.0e+00 & 1.1e-07 \\ -4.3e-05 & 1.2e-05 & -5.8e-17 & -4.1e-07 & 1.1e-07 & 1.0e+00 \end{bmatrix}.$$

Also ist die Abweichung von der Diagonalmatrix in derselben Größenordnung geblieben. Ursache für die lineare Konvergenz ist die Struktur von K: es gilt $F-I=I-C$ und $B=B^T$. Wird zu Beginn eines Zyklus' eine derartige Struktur erkannt, so werden zur Vermeidung linearer Konvergenz statt eines normalen Zyklus' zuerst Transformationen

$$K' = T^T K T \quad , \quad K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \quad , \quad T = \begin{bmatrix} R^{-1} & 0 \\ 0 & R^T \end{bmatrix} \quad , \quad F = R^T R \tag{0.6}$$

und

$$K'' = T'^T K' T' \ , \quad K' = \begin{bmatrix} F' & B'^T \\ B' & C' \end{bmatrix} \quad , \quad T' = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \quad , \quad X = \frac{-1}{2}(B' + B'^T) \tag{0.7}$$

vorgenommen. Diese sind auf obige Struktur von K abgestimmt und führen bei vorliegendem Beispiel gerundet zu

$$K'' = \begin{bmatrix} 1.0e+00 & -5.9e-22 & -4.4e-21 & 0 & -7.8e-09 & 2.1e-10 \\ -5.9e-22 & 1.0e+00 & -1.1e-21 & 7.8e-09 & 0 & -2.0e-09 \\ -4.4e-21 & -1.1e-21 & 1.0e+00 & -2.1e-10 & 2.0e-09 & 0 \\ 0 & 7.8e-09 & -2.1e-10 & 1.0e+00 & 5.1e-09 & 2.0e-09 \\ -7.8e-09 & 0 & 2.0e-09 & 5.1e-09 & 1.0e+00 & -4.6e-10 \\ 2.0e-10 & -2.0e-09 & 0 & 2.0e-09 & -4.6e-10 & 1.0e+00 \end{bmatrix}$$

mit $\|K'' - \mathrm{diag}(K'')\| = 1.1e-8$, also zu quadratischer Konvergenz. Zeigen Submatrizen von K die oben beschriebenen Strukturen, so erstrecken sich auch die Transformationen (0.6), (0.7) nur auf diese Submatrizen. Auf den restlichen Elementen von K wird wie in einem normalen Zyklus verfahren. Einen Zyklus, in dem Transformationen (0.6), (0.7) vorkommen, werden wir *modifizierten Zyklus* nennen. Die Grundidee zu derartigen Modifikationen Jacobi-ähnlicher Verfahren stammt von Veselić und Hari, im vorliegenden Fall konkret von Veselić. Numerische Experimente bescheinigen diesem Vorgehen stets asymptotisch quadratische Konvergenz.

Wenden wir uns nun Aussagen zur Genauigkeit des zuvor skizzierten Verfahrens zu und nehmen an, wir rechneten mit endlicher Genauigkeit ε. Von jedem der beiden Verfahrensschritte werden wir eine Rückwärtsfehleranalyse vornehmen. Dabei werden wir für den ersten Schritt

$$\tilde{L} J \tilde{L}^T = PSP^T + F \quad , \quad |F| \ \leq \ 3n\,(|PSP^T| + |\tilde{L}|\ |J|\ |\tilde{L}|^T)\,\varepsilon \ ,$$

wenn $\tilde{L}$ der in endlicher Genauigkeit berechnete Faktor ist, zeigen und daraus für die jeweils geordneten Eigenwerte $\pm i\lambda_j$ von $PSP^T = LJL^T$ bzw. $\pm i\lambda'_j$ von $\tilde{L}J\tilde{L}^T$

$$\frac{|\lambda_j - \lambda'_j|}{|\lambda_j|} \ \leq \ \frac{12n^2\varepsilon}{\lambda_{\min}(M)} \ , \quad M = (D^{-1}\sqrt{-(\tilde{L}J\tilde{L}^T)^2}D^{-1}) \ , \quad D = \mathrm{diag}(\|\tilde{L}_{i\cdot}\|_2) \tag{0.8}$$

für alle j schließen. Aufgrund numerischer Experimente kann man moderate $\lambda_{\min}(M)$, also kleine maximale relative Fehler, erwarten.

Bei der Fehleranalyse des zweiten Verfahrensschrittes werden wir für jede einzelne endlich genaue Transformationen (0.3) eine Analyse der Art

$$\begin{tikzcd} L^{(m)} \arrow[r, "\text{Rechnung in}", "\text{endlicher Genauigkeit}"'] \arrow[d, "+\delta L^{(m)}"'] & \mathrm{fl}(L^{(m+1)}) \\ L^{(m)} + \delta L^{(m)} \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

vornehmen, sie also als das Ergebnis einer exakten Transformation einer gestörten Matrix darstellen und jeweils eine kleine obere Schranke für die Norm der skalierten Störung

$$\|\delta L_S^{(m)}\|_2 = \|\delta L^{(m)} {D^{(m)}}^{-1}\|_2 \quad , \quad D = \mathrm{diag}(\|L_{\cdot i}^{(m)}\|_2) \ .$$

angeben. Mit Hilfe der störungstheoretischen Aussage

$$(1-\eta)^2 \ \leq \ \frac{|\lambda_j'|}{|\lambda_j|} \ \leq \ (1+\eta)^2 \quad , \quad \eta = \frac{\|\delta L_S^{(m)}\|_2}{\sigma_{\min}(L^{(m)} {D^{(m)}}^{-1})} \tag{0.9}$$

für die geordneten Eigenwerte $\pm i\lambda_j$ von $S^{(m)} := L^{(m)} J {L^{(m)}}^T$ bzw. $\pm i\lambda_j'$ von $\tilde{S}^{(m)} := (L^{(m)} + \delta L^{(m)}) J ({L^{(m)}}^T + \delta {L^{(m)}}^T)$ haben wir dann bereits Schranken für die relativen Fehler in den Eigenwerten von $S^{(m)}$ gegenüber $S^{(m+1)}$. Die Eigenwerte von $\tilde{S}^{(m)}$ und $S^{(m+1)}$ sind nämlich gleich.

Eine günstige Abschätzung (0.9) können wir nur für moderate

$$\sigma^{(m)} := \frac{1}{\sigma_{\min}(L^{(m)} {D^{(m)}}^{-1})}$$

erwarten. Wir werden zeigen, daß $\sigma^{(1)}$ allein abhängig von der Ordnung $2n$ beschränkt ist. Außerdem ist $\sigma^{(1)}$ aufgrund der Präkonditionierungseigenschaft des ersten Verfahrensschrittes zumeist recht klein (≤ 100), wie numerische Experimente ergaben. Darüber hinaus gilt die Abschätzung $\sigma^{(m+1)} \leq \sqrt{2}\sigma^{(m)}$, die je nach vorgenommener Transformation noch verbessert werden kann. Wie in [13], [29] verwenden wir $1/\sigma_{\min}(L^{(m)} {D^{(m)}}^{-1})$ statt $1/\sigma_{\min}(L^{(m)})$ als Konditionsmaß, denn oft gilt $\sigma_{\min}(L^{(m)} {D^{(m)}}^{-1}) \gg \sigma_{\min}(L^{(m)})$.

Fassen wir alle im zweiten Verfahrensschritt entstehenden maximalen relativen Fehler additiv zusammen, so erhalten wir eine Schranke für den relativen Gesamtfehler in den berechneten Eigenwerten von $\tilde{L} J \tilde{L}^T$, den wir nur noch zum

maximalen relativen Fehler (0.8) zu addieren brauchen, um zum maximalen Gesamtfehler in den berechneten Eigenwerten von S zu gelangen. Trotz des recht hohen Aufwandes bei der Fehleranalyse des zweiten Verfahrensschrittes ist der dort entstehende Fehler zumeist gering im Vergleich zu dem des ersten Schrittes.

Das in dieser Arbeit vorgestellte Verfahren übertrifft QR-Verfahren [25] und das von Paardekooper [34] vorgestellte Jacobi-ähnliche Verfahren in der Genauigkeit der berechneten Eigenwerte häufig bei weitem. Besonders das Verfahren von Paardekooper liefert zumeist bei Matrizen mit Eigenwerten verschiedener Größenordnung keine brauchbaren Ergebnisse. Wir können die Vorteile unseres hier vorgestellten Verfahrens wie folgt zusammenfassen:

- Man kann eine obere Schranke für die maximalen relativen Fehler in den berechneten Eigenwerten angeben. QR-Verfahren sind zwar um etwa den Faktor 6 schneller als unser Verfahren, bieten jedoch kein derartiges Genauigkeitsverhalten. Es muß allerdings konstatiert werden, daß die Genauigkeit von QR-Verfahren oft besser als erwartet ausfällt.

- Unser Verfahren ist zumeist um ein vielfaches schneller als das Verfahren von Paardekooper. Das Verfahren von Paardekooper liefert bei Matrizen mit Eigenwerten verschiedener Größordnung zumeist katastrophal ungenaue Eigenwerte.

- Liegt $S = LJL^T$ bereits in faktorisierter Form vor, so gewinnt unser Verfahren gegenüber QR an Geschwindigkeit. Außerdem wird es genauer.

- Der iterative Teil unseres Verfahrens ist inhärent parallel.

Die vorliegende Arbeit ist folgendermaßen aufgebaut. Im nachfolgenden Kapitel 1 bringen wir einige Ergebnisse aus der Störungstheorie. In Kapitel 2 kommen wir dann zuerst zur Entwicklung des zweiten Verfahrensschrittes. Wir beschreiben zunächst die Strukturen, die zu den modifizierten Zyklen führen. Anschließend stellen wir eine explizite Version des Verfahrens vor. Sie dient dem besseren Verständnis der recht komplizierten impliziten Version. Die formellen Algorithmen beider Versionen sind in einem Pseudocode abgedruckt. Der Nachweis globaler Konvergenz der Verfahren schließt das Kapitel 2 ab. In Kapitel 3 nehmen wir eine Fehleranalyse des impliziten Verfahrens vor. In Kapitel 4 wenden wir uns schließlich dem ersten Verfahrensschritt, der schiefsymmetrischen Zerlegung und deren Fehleranalyse, zu. In Kapitel 5 fassen wir die Ergebnisse der Fehleranalysen zusammen und geben eine obere Schranke für den relativen Gesamtfehler in den in endlicher Genauigkeit berechneten Eigenwerten von S an. Schließlich setzen wir unser Verfahren im abschließenden Kapitel 6 der Konkurrenz mit QR-Verfahren und Paardekooperschem Verfahren aus. Die numerischen Ergebnisse untermauern die theoretischen Resultate der vorangehenden Kapitel.

Kapitel 1

Störungstheorie

In diesem Kapitel werden einige Ergebnisse aus der Störungstheorie angegeben. Sie sind Modifikationen der Ergebnisse von [4], [13], [28], [29], [33]. Zum Teil können die Ergebnisse der vorgenannten Arbeiten direkt für schiefsymmetrische Matrizen S übernommen werden, so sämtliche Ergebnisse aus [33], wenn man hermitesche Störungen der hermiteschen Matrix $iS = -iS^T$, $i^2 = -1$, betrachtet. Der Vollständigkeit halber seien einige grundlegende Ergebnisse vorstehender Arbeiten, angepaßt auf die hier behandelten Eigenwertprobleme, aufgeführt. Ergebnisse dieses Kapitels werden wir bei der Bestimmung von Schranken für berechnete Eigenwerte verwenden.

Satz 1.0.1 *Für schiefsymmetrisches $S \in \mathbb{R}^{2n\times 2n}$ oder $S \in \mathbb{R}^{(2n-1)\times(2n-1)}$ sei die positiv semidefinite Matrix*

$$|S| = \sqrt{-S^2} \tag{1.1}$$

gegeben. δS sei eine schiefsymmetrische Störung von S, so daß für ein $\eta < 1$ und alle $x \in \mathbb{C}^{2n}$ bzw. $x \in \mathbb{C}^{2n-1}$

$$|x^*\delta S x| \leq \eta\, x^* |S| x$$

gilt. Die Eigenwerte von S seien $\pm i\lambda_j$, $1 \leq j \leq n$, $i^2 = -1$, und $\lambda_0 = 0$ im Falle $S \in R^{(2n-1)\times(2n-1)}$ mit

$$0 \leq \lambda_1 \leq \ldots \leq \lambda_n \ ,$$

und die Eigenwerte von $S' = S + \delta S$ seien $\pm i\lambda'_j$, $1 \leq j \leq n$, und $\lambda'_0 = 0$ im Falle $S' \in R^{(2n-1)\times(2n-1)}$ mit

$$0 \leq \lambda'_1 \leq \ldots \leq \lambda'_n \ .$$

Dann haben iS und iS' dieselbe Trägheit und für die nichtverschwindenden λ_i gilt

$$1 - \eta \ \leq \ \frac{\lambda'_j}{\lambda_j} \ \leq \ 1 + \eta \quad , \quad 1 \leq j \leq n \ .$$

Beweis:
Anwendung von [33], (Th. 2.1) auf $H = iS$. Q.E.D.

Mit folgenem Satz wollen wir ein weiteres Ergebnis aus [33] zitieren. Die darin definierten Matrizen werden uns in Abschnitt 6.2 wieder begegnen.

Satz 1.0.2 *Sei $S = -S^T = D(Q+R)D \in \mathbb{R}^{2n\times 2n}$ nichtsingulär mit diagonalem, positiv definitem D und $Q = -Q^T = -Q^{-1}$, sowie $QD = DQ$ und $\|R\|_2 < 1$. Sei δS eine Störung von S mit $|\delta S_{jk}| \leq \varepsilon |S_{jk}|$, $1 \leq j,k \leq 2n$. Dann gilt für die Eigenwerte $\pm i\lambda_j$, $i^2 = -1$, mit*

$$0 < \lambda_1 \leq \ldots \leq \lambda_n$$

von S bzw. die Eigenwerte $\pm i\lambda'_j$ mit

$$0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

von $S + \delta S$

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \leq 2n\,\varepsilon\,\frac{1 + \|\,|R|\,\|_2}{1 - \|R\|_2} \quad , \quad j = 1 \ldots n \ , \tag{1.2}$$

falls die rechte Seite von (1.2) kleiner als 1 ist.

Beweis:
Anwendung von [33], (Th. 2.30) auf $H = iS, E = Q, N = iR$. Q.E.D.

Die Aussagen der letzten beiden Sätze sind auch dann gültig, wenn man statt schiefsymmetrischer Störungen von S allgemeine hermitesche Störungen von iS, $i^2 = -1$, betrachtet. Weil wir in dieser Arbeit ausschließlich reelle numerische Verfahren betrachten werden, brauchen wir auch nur rein imaginäre hermitesche Störungen von iS zu behandeln.

Wir betrachten anschließend noch kurz die Störung des Faktors L, wenn für S die Zerlegung

$$S = LJL^T$$

gilt. Dazu benötigen wir die unitären Hilfsmatrizen

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} I & iI \\ iI & I \end{bmatrix} \quad \text{und} \quad J_1 = I \oplus (-I) \ , \quad i^2 = -1 \ , \tag{1.3}$$

für die $U^*JU = iJ_1$ gilt, und die es uns ermöglicht, einen Teil der nachfolgenden Störungstheorie auf bereits bekannte Ergebnisse zurückzuführen.

Wir schicken ein Lemma voran:

Lemma 1.0.3 *Seien $A, B \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$. Dann sind die folgenden Aussagen äquivalent*

$$\|Az\|_2 \leq \alpha \|Bz\|_2 \quad \text{für alle } z \in \mathbb{C}^n \tag{1.4}$$
$$\|Ax\|_2 \leq \alpha \|Bx\|_2 \quad \text{für alle } x \in \mathbb{R}^n \tag{1.5}$$

Beweis:
(1.4) $\Rightarrow$ (1.5) ist trivial und für die umgekehrte Richtung gilt mit $z = u + iv$, $u, v \in \mathbb{R}^n$

$$\begin{aligned}
\|Az\|_2^2 &= z^*A^*Az = (u^T - iv^T)A^TA(u+iv) = u^TA^TAu + v^TA^TAv \\
&= \|Au\|_2^2 + \|Av\|_2^2 \\
&\leq \alpha(\|Bu\|_2^2 + \|Bv\|_2^2) = \alpha(u^TB^TBu + v^TB^TBv) = \alpha(z^*B^*Bz) \\
&= \alpha \|Bz\|_2^2
\end{aligned}$$

Q.E.D.

Satz 1.0.4 *Sei $L \in \mathbb{R}^{2n\times 2n}$ nichtsingulär und $0 \leq \eta < 1$, $i^2 = -1$. Für ein $\delta L \in \mathbb{R}^{2n\times 2n}$ sei*

$$L' = L + \delta L, \quad \text{wobei} \quad \|\delta Lx\|_2 \leq \eta \|Lx\|_2 \quad \text{für alle} \quad x \in \mathbb{R}^{2n} . \tag{1.6}$$

Dann haben $iS = iLJL^T$ und $iS' = iL'JL'^T$ dieselbe Trägheit. Für die Eigenwerte $\pm i\lambda_j$, $1 \leq j \leq n$ von S mit

$$0 < \lambda_1 \leq \ldots \leq \lambda_n$$

bzw. die Eigenwerte $\pm\lambda'_j$, $1 \leq j \leq n$ von S' mit

$$0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

gilt

$$(1-\eta)^2 \leq \frac{\lambda'_k}{\lambda_k} \leq (1+\eta)^2 \quad k = 1 \ldots n . \tag{1.7}$$

Dasselbe gilt, wenn statt S das Paar (L^TL, J) und statt S' das Paar (L'^TL', J) mit J aus (0.1) betrachtet wird.

Beweis:
Sei $G = LU$ mit U aus (1.3). Dann ist $-iS = -iLJL^T = -iLUiJ_1U^*L^* = GJ_1G^*$. Für $G' = L'U$ gilt analog $-iS' = G'J_1G'^\star$. Setzt man noch $\delta G = \delta LU$, so kann die Zeile (1.6) wegen 1.0.3 äquivalent als

$$G' = G + \delta G\,, \text{ wobei } \|\delta Gz\|_2 \leq \eta \|Gz\|_2 \quad \text{für alle } z \in \mathbb{C}^{2n}$$

ausgedrückt werden. Aus [33], (Th. 3.3) folgt nun (1.7). Weil die Eigenwerte von S bzw. S' mit denen des Paares (L^TL, J) bzw. (L'^TL', J) übereinstimmen, gilt auch die Aussage bezüglich dieser Paare. Q.E.D.

Satz 1.0.5 *Sei $L = L_S D \in I\!R^{2n\times 2n}$ nichtsingulär mit $D = diag(\|L_{\cdot i}\|_2)$. Sei $L' = L + \delta L$ mit $\delta L = \delta L_S D$, wobei $\eta = \|\delta L_S\|_2/\sigma_{\min}(L_S) < 1$ gelte. Dann sind die Voraussetzungen von Satz 1.0.4 erfüllt, also gilt seine Aussage.*

Beweis:
Für $x \in I\!R^{2n}$ gilt

$$\|\delta L x\|_2 = \|\delta L_S D x\|_2 \le \|\delta L_S\|_2 \, \|Dx\|_2 \le \|\delta L_S\|_2 \, \|L_S^{-1}\|_2 \, \|L_S D x\|_2 \ .$$

Wegen $\eta = \|\delta L_S\|_2 \, \|L_S^{-1}\|_2 = \|\delta L_S\|_2/\sigma_{\min}(L_S) < 1$ ist bereits alles gezeigt.

Q.E.D.

Wir betrachten nun in $I\!R^{2n\times 2n}$ das Eigenwertproblem (0.2). Zu K benötigen wir häufig auch die assoziierte skalierte Matrix, die wir mit

$$K_S = D^{-1} K D^{-1} = \begin{bmatrix} F_S & B_S^T \\ B_S & C_S \end{bmatrix} \quad \text{mit} \quad D = \mathrm{diag}(K_{jj}^{1/2}) \tag{1.8}$$

bezeichnen wollen. Definieren wir noch die *Kondition* $\kappa(X) = \|X\|_2 \, \|X^{-1}\|_2$, so gilt

Satz 1.0.6 *Sei $K = D K_S D \in I\!R^{2n\times 2n}$ symmetrisch und positiv definit mit $D = diag(\sqrt{K_{jj}})$, also $K_{S_{jj}} = 1, j = 1 \ldots 2n$. Sei $\delta K = D \delta K_S D$ eine symmetrische Störung mit $\|\delta K_S\|_2 < \lambda_{\min}(K_S)$. Die Eigenwerte des Problems $Kx = i\lambda Jx$ seien*

$$-\lambda_n \le \ldots \le -\lambda_1 < 0 < \lambda_1 \le \ldots \le \lambda_n$$

und die Eigenwerte des Problems $K'x = i\lambda Jx$ für $K' = K + \delta K$ seien

$$-\lambda'_n \le \ldots \le -\lambda'_1 < 0 < \lambda'_1 \le \ldots \le \lambda'_n \ .$$

Dann gilt für $1 \le j \le n$

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \le \frac{\|\delta K_S\|_2}{\lambda_{\min}(K_S)} \le \kappa(K_S) \cdot \|\delta K_S\|_2 \ . \tag{1.9}$$

Ist $|\delta K_{jk}| \le |K_{jk}|\varepsilon$ mit $\sqrt{2n}\|K_S\|_2\varepsilon \le \lambda_{\min}(K_S)$, dann gilt

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \le \sqrt{2n}\kappa(K_S) \cdot \varepsilon \ .$$

Ist $|\delta K_{jk}| \le \sqrt{K_{jj}K_{kk}}\varepsilon$ mit $2n\varepsilon \le \lambda_{\min}(K_S)$, dann gilt

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \le \frac{2n\varepsilon}{\lambda_{\min}(K_S)} \le 2n\kappa(K_S) \cdot \varepsilon \ .$$

Beweis:
Sei

$$T = \begin{bmatrix} \operatorname{diag}(\sqrt[4]{K_{j+n,j+n}/K_{jj}}) & 0 \\ 0 & \operatorname{diag}(\sqrt[4]{K_{jj}/K_{j+n,j+n}}) \end{bmatrix}$$

und $K_1 = T^T K T$. Für δK sei $\delta K_1 = T^T \delta K T$ und für $K' = K + \delta K$ sei $K_1' = T^T K' T$. Die Skalierung mit T dienet dazu, die Aussagen des Satzes auf bereits bekannte Ergebnisse zurückzuführen. Es gilt $K_{1S} = K_S$, $\delta K_{1S} = \delta K_S$ und die Zeile

$$K' = K + \delta K \quad \text{mit} \quad K = D K_S D \ , \quad \delta K = D \delta K_S D \ , \quad \|\delta K_S\|_2 < \lambda_{\min}(K_S)$$

können wir äquivalent durch

$$\begin{gathered} K_1' = K_1 + \delta K_1 \ , \\ K_1 = T^T D K_S D T \ , \quad \delta K_1 = T^T D \delta K_{1S} D T \ , \quad \|\delta K_{1S}\|_2 < \lambda_{\min}(K_{1S}) \end{gathered} \tag{1.10}$$

ausdrücken. Weil die Eigenwerte des Paares (K, J) bzw. (K', J) identisch sind mit den Eigenwerten des Paares (K_1, J) bzw. (K_1', J), genügt es, die Aussage des Satzes für $(K_1, J), (K_1', J)$ statt $(K, J), (K', J)$ zu zeigen. Für K_1, K_{1S}, K_1', K_{1S}' seien also die Voraussetzungen des Satzes erfüllt (d.h. $T = I$). Mit U, J_1 aus (1.3) gilt

$$K_1 x = i\lambda J x \Longleftrightarrow K_1 U y = i\lambda J U y \Longleftrightarrow U^* K_1 U y = -\lambda J_1 y \ .$$

Dabei ist für $H = U^* K_1 U$

$$H_{jj} = H_{j+n,j+n} = (K_{1jj} + K_{1j+n,j+n})/2 = K_{1jj} = K_{1j+n,j+n} \ , \quad j = 1 \ldots n.$$

Mit $H' = U^* K_1' U$ ist (1.10) daher äquivalent zu

$$\begin{gathered} H' = U^* K_1 U + U^* \delta K_1 U \ , \\ K_1 = D K_S D \ , \quad \delta K_1 = D \delta K_{1S} D \ , \quad \|\delta K_{1S}\|_2 < \lambda_{\min}(K_{1S}) \end{gathered}$$

und zu

$$\begin{gathered} H' = H + \delta H \ , \\ \delta H = U^* D \delta K_{1S} D U = D U^* \delta K_{1S} U D = D \delta H_S D \ , \quad \|\delta H_S\|_2 < \lambda_{\min}(H_S) \ . \end{gathered}$$

Die Eigenwerte der Paare (K_1, J) und (H, J_1) bzw. (K_1', J) und (H', J_1) sind also bis auf das Vorzeichen gleich und die Aussage des Satzes folgt aus [28], (Th. 1.2).
Q.E.D.

Im vorangehenden Satz ist der Fehler in den Eigenwerten in Abhängigkeit von $\kappa(K_S)$, der *skalierten Kondition*, dargestellt. Der Fehler kann auch abhängig von der Kondition $\kappa(K)$ dargestellt werden. Das gewählte Vorgehen ist jedoch oft günstiger. Nach [30] gilt

$$\kappa(K_S) \leq 2n \min_{D=\text{diag}} \kappa(DKD) \ ,$$

die gewählte Skalierung ist also „nahezu“ optimal. Es kann zwar $\kappa(K_S) \geq \kappa(K)$ gelten, numerische Experimente zeigen jedoch, daß wir für diejenigen $K = L^T L$, die aus unserer schiefsymmetrischen Zerlegung $S = LJL^T$ entstehen, zumeist $\kappa(K_S) \ll \kappa(K)$ erwarten können. Wir kommen in den Kapiteln 4, 6 darauf zurück.

Wir schließen dieses Kapitel mit einigen Aussagen über *skaliert diagonaldominante Matrizen* ab. Sie basieren auf Ergebnissen in [4]. Dazu benötigen wir

$$J_n = \oplus_{j=1}^n \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad , \quad \tilde{J}_n = \oplus_{j=1}^n \begin{bmatrix} 0 & \sigma_j \\ -\sigma_j & 0 \end{bmatrix} \quad , \quad \sigma_j \in \{-1, 1\} \ . \tag{1.11}$$

Definition 1.0.7 *Sei $\tilde{J}_n$ gemäß (1.11) und $\Delta = \oplus_{j=1}^n \delta_j I_2$ mit $\delta_j > 0$, $j = 1 \ldots n$. Eine schiefsymmetrische Matrix $S = \Delta \tilde{J}_n \Delta + R$ heißt $\gamma-$diagonaldominant ($\gamma-$d.d.), wenn es ein γ, $0 \leq \gamma < 1$, mit*

$$\|R\|_2 \leq \gamma \min_{1 \leq j \leq n} \delta_j^2$$

gibt. S heißt skaliert $\gamma-$diagonaldominant (skaliert $\gamma-$d.d.), wenn $\Delta^{-1} S \Delta^{-1}$ $\gamma-$d.d. ist.

Wir werden nun zeigen, daß skaliert $\gamma-$d.d. Matrizen ihre Eigenwerte gut bestimmen, also kleine Änderungen der Matrixelemente nur zu kleinen Änderungen der Eigenwerte führen. Dazu verwenden wir hilfsweise die Matrix

$$U_n = \bigoplus_{j=1}^n U_1 \quad , \quad U_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \quad , \quad i^2 = -1 \ . \tag{1.12}$$

Sie ist unitär mit

$$U_1^* J_1 U_1 = i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad , \quad J_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \ .$$

Satz 1.0.8 *Sei $S = -S^T \in \mathbb{R}^{2n \times 2n}$ skaliert $\gamma-$d.d., d.h. es sei $S = \Delta(\tilde{J}_n + R')\Delta$ mit $\Delta = \oplus_{j=1}^n \delta_j I_2$, $\delta_j > 0$, $j = 1 \ldots n$, und $\|R'\|_2 < \gamma$. Sei δS eine schiefsymmetrische Störung von S mit $\eta = \|\Delta^{-1} \delta S \Delta^{-1}\|_2 < (1-\gamma)/n$. $S + \delta S$ sei ebenfalls skaliert $\gamma-$d.d.. Dann sind S und $S + \delta S$ nichtsingulär und für die Eigenwerte $\pm i\lambda_j$ mit*

$$0 < \lambda_1 \leq \ldots \leq \lambda_n$$

von S bzw. die Eigenwerte $\pm i\lambda'_j$ mit

$$0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

von $S + \delta S$ gilt

$$1 - \frac{n\,\eta}{1-\gamma} \leq \frac{\lambda'_j}{\lambda_j} \leq (1 - \frac{n\,\eta}{1-\gamma})^{-1} \quad , \quad 1 \leq j \leq n \tag{1.13}$$

Beweis:
Sei U_n gemäß (1.12). Wir betrachten die hermitesche Matrix

$$H = U_n^* i S U_n = \Delta A \Delta \ .$$

Mit J_n gemäß (1.11) gilt

$$A = \Delta^{-1} U_n^* i S U_n \Delta^{-1} = U_n^* \Delta^{-1} i S \Delta^{-1} U_n = U_n^* (i \tilde{J}_n + i R') U_n = E + N \ ,$$

wobei

$$E = U_n^* i \tilde{J}_n U_n = \bigoplus_{j=1}^{n} \begin{bmatrix} -\sigma_j & 0 \\ 0 & \sigma_j \end{bmatrix} \quad , \quad N = i U_n^* R' U_n \ .$$

Nach Voraussetzung ist dabei $\|N\|_2 = \|R'\|_2 \leq \gamma < 1$. Weiter betrachten wir

$$\delta H = U_n^* i \delta S U_n \ .$$

Nach Voraussetzung ist

$$\|\Delta^{-1} \delta H \Delta^{-1}\|_2 = \|U_n^* \Delta^{-1} i \delta S \Delta^{-1} U_n\|_2 = \|\Delta^{-1} \delta S \Delta^{-1}\|_2 < \frac{1-\gamma}{n} \ .$$

Mit den soeben eingeführten Bezeichnungen folgt die Aussage des Satzes sofort aus [4], Proposition 4, denn die dortigen Ergebnisse gelten unverändert für hermitesche Matrizen. Dabei ist zu bedenken, daß die in [4] gemachte Voraussetzung, die Diagonale von N müsse verschwinden, entfallen kann. Aus dem Beweisgang in [4] ergibt sich leicht, daß (1.13) tatsächlich n enthält (und nicht die volle Ordnung $2n$). Q.E.D.

Analog zu [4], Th. 4, können wir den Faktor n in (1.13) unterdrücken. Es gilt

Satz 1.0.9 *Sei $S = -S^T \in \mathbb{R}^{2n \times 2n}$ skaliert $\gamma-$d.d., d.h. es sei $S = \Delta(\tilde{J}_n + R')\Delta$ mit $\Delta = \oplus_{j=1}^{n} \delta_j I_2$, $\delta_j > 0$, $j = 1 \ldots n$, und $\|R'\|_2 < \gamma$. Sei δS eine schiefsymmetrische Störung von S mit $\eta = \|\Delta^{-1} \delta S \Delta^{-1}\|_2$. Für alle ξ, $0 \leq \xi \leq 1$ sei $S + \xi \delta S$ skaliert $\gamma-$d.d.. Für die Eigenwerte $\pm i \lambda_j$ mit*

$$0 < \lambda_1 \leq \ldots \leq \lambda_n$$

von S bzw. die Eigenwerte $\pm i \lambda'_j$ mit

$$0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

von $S + \delta S$ gilt dann

$$\exp(\frac{-\eta}{1-\gamma}) \leq \frac{\lambda'_j}{\lambda_j} \leq \exp(\frac{\eta}{1-\gamma}) \quad , \quad 1 \leq j \leq n$$

Beweis:
Sei U_n gemäß (1.12), $H = U_n^* i S U_n = \Delta A \Delta$, sowie $\delta H = U_n^* i \delta S U_n$. Anwendung von [4], Th. 4, auf H und δH liefert sofort die Aussage des Satzes. Q.E.D.

Nachfolgender Satz besagt, daß Nebendiagonalelemente skaliert γ-d.d. schiefsymmetrischer Matrizen bereits Näherungen an die Imaginärteile der Eigenwerte sind.

Satz 1.0.10 *Die Matrix $S = -S^T \in \mathbb{R}^{2n\times 2n}$ mit $|S_{2k,2k-1}| \leq |S_{2k+2,2k+1}|$, $k = 1\ldots n-1$, sei skaliert $\gamma-$d.d.. S habe die Eigenwerte $\pm i\lambda_j$ mit*

$$0 < \lambda_1 \leq \ldots \leq \lambda_n \ .$$

Dann gilt

$$1-\gamma \leq \frac{\lambda_k}{|S_{2k,2k-1}|} \leq 1+\gamma \quad , \quad k = 1\ldots n \ .$$

Beweis:
Mit U_n gemäß (1.12) und $\tilde{J}_n$ gemäß (1.11) haben wir

$$H = P^T U_n^* i S U_n P \ ,$$

wobei P eine Permutationsmatrix sei, die zu

$$H_{11} \leq \ldots \leq H_{nn} < 0 < H_{n+1,n+1} \leq \ldots \leq H_{2n,2n}$$

führt. Wegen Definition 1.0.7 gilt

$$\begin{aligned}
H &= P^T U_n^* \Delta\Delta^{-1} i S \Delta^{-1} \Delta U_n P \\
&= P^T U_n^* \Delta i (\tilde{J}_n + \Delta^{-1} R \Delta^{-1}) \Delta U_n P \\
&= P^T \Delta (U_n^* i \tilde{J}_n U_n + \Delta^{-1} U_n^* i R U_n \Delta^{-1}) \Delta P \\
&= \Delta' P^T (\bigoplus_{j=1}^{n} \begin{bmatrix} -\sigma_j & 0 \\ 0 & \sigma_j \end{bmatrix} + \Delta^{-1} U_n^* i R U_n \Delta^{-1}) P \Delta' \quad , \quad \Delta' = P^T \Delta P \\
&= \Delta'(D+N)\Delta'
\end{aligned}$$

mit

$$D = P^T \bigoplus_{j=1}^{n} \begin{bmatrix} -\sigma_j & 0 \\ 0 & \sigma_j \end{bmatrix} P \quad , \quad N = P^T \Delta^{-1} U_n^* i R U_n \Delta^{-1} P \ .$$

Dabei ist

$$\|N\|_2 = \|\Delta^{-1} R \Delta^{-1}\|_2 \leq \|\Delta^{-1}\|_2^2 \ \gamma \min_{1\leq i\leq n} \delta_i^2 = \gamma \ .$$

H ist also $\gamma-$s.d.d. gemäß [4]. Anwendung von [4], Proposition 2, auf H liefert nun die Aussage des Satzes. Q.E.D.

Wir führen noch einen weiteren Begriff von Diagonaldominanz ein. Er wird erst in Kapitel 4 seine Nützlichkeit erweisen. Sei $S \in I\!R^{2n\times 2n}$ in m^2 Blöcke der Ordnung r mit $rm = 2n$ partitioniert:

$$S = (S^{[i,j]})_{1\le i,j\le m} \quad S^{[i,j]} = (S_{pq})_{\substack{(i-1)r+1\le p\le ir \\ (j-1)r+1\le q\le jr}} \quad . \tag{1.14}$$

Definition 1.0.11 *Eine schiefsymmetrische Matrix S mit der Partitionierung (1.14) heißt $\gamma - r-$diagonaldominant ($\gamma - r-$d.d.), wenn für ein γ, $0 \le \gamma < 1$, und alle j, $1 \le j \le m$,*

$$\sum_{i=1, i\ne j}^{m} \|S^{[i,j]}\|_2 \le \gamma \, \|S^{[j,j]}\|_2$$

ist. S heißt skaliert $\gamma-r-$diagonaldominant (skaliert $\gamma-r-$d.d.), wenn $\Delta^{-1}S\Delta^{-1}$ mit $\Delta = \oplus_{j=1}^{m}(\|S^{[j,j]}\|_2^{1/2} I_r)$ $\gamma - r-$d.d. ist.

Definition und Lemma 1.0.12 *Die für die partitionierte Matrix*

$$X \in I\!R^{n\times n} \quad , \quad X = (X^{[i,j]})_{1\le i,j\le m} \ , \quad n = mr \ , \quad X^{[i,j]} \in I\!R^{r\times r}$$

durch

$$|||X|||_2 = \max_{1\le j\le m} \sum_{i=1}^{m} \|X^{[i,j]}\|_2$$

definierte Abbildung ist eine Matrixnorm. Gilt $\|X^{[i,j]}\|_2 = \|X^{[j,i]}\|_2$ für alle i, j, $1 \le i, j \le m$, so ist $\|X\|_2 \le |||X|||_2$.

Beweis:
Die Eigenschaften

1. $|||X|||_2 \ge 0$
2. $|||X|||_2 = 0 \Longleftrightarrow X = 0$
3. $|||cX|||_2 = |c| \ |||X|||_2$, $c \in I\!R$
4. $|||X + Y|||_2 \le |||X|||_2 + |||Y|||_2$

werden trivialerweise erfüllt. Zu zeigen ist noch

5. $|||XY|||_2 \le |||X|||_2 \ |||Y|||_2$.

Es gilt

$$\begin{aligned}
|||XY|||_2 &= \max_{1\le j\le m}\sum_{i=1}^{m}\|(XY)^{[i,j]}\| = \max_{1\le j\le m}\sum_{i=1}^{m}\|\sum_{k=1}^{m}X^{[i,k]}Y^{[k,j]}\| \\
&\le \max_{1\le j\le m}\sum_{i=1}^{m}\sum_{k=1}^{m}\|X^{[i,k]}\|\,\|Y^{[k,j]}\| = \max_{1\le j\le m}\sum_{k=1}^{m}\|Y^{[k,j]}\|\sum_{i=1}^{m}\|X^{[i,k]}\| \\
&\le \max_{1\le j\le m}\sum_{k=1}^{m}\|Y^{[k,j]}\|\max_{1\le k\le m}\sum_{i=1}^{m}\|X^{[i,k]}\| = |||X|||_2\,|||Y|||_2
\end{aligned}$$

Nun zum Beweis der letzten Behauptung. Es ist mit entsprechender Partitionierung von Vektoren $x \in I\!R^{2n}$

$$\begin{aligned}
\|(Ax)^{[i]}\|_2 &= \|\sum_{j=1}^{m}A^{[i,j]}x^{[j]}\|_2 \\
&\le \sum_{j=1}^{m}\|A^{[i,j]}\|_2^{1/2}\|A^{[i,j]}\|_2^{1/2}\|x^{[j]}\|_2 \\
&\le \max_{1\le i\le m}((\sum_{j=1}^{m}\|A^{[i,j]}\|_2)^{1/2})\cdot(\sum_{j=1}^{m}\|A^{[i,j]}\|_2\|x^{[j]}\|_2^2)^{1/2}
\end{aligned}$$

und daher

$$\begin{aligned}
\|Ax\|_2 &= (\sum_{i=1}^{m}\|(Ax)^{[i]}\|_2^2)^{1/2} \\
&\le \max_{1\le i\le m}((\sum_{j=1}^{m}\|A^{[i,j]}\|_2)^{1/2})\cdot\sum_{i=1}^{m}(\sum_{j=1}^{m}\|A^{[i,j]}\|_2\|x^{[j]}\|_2^2)^{1/2} \\
&\le \max_{1\le i\le m}((\sum_{j=1}^{m}\|A^{[i,j]}\|_2)^{1/2})\cdot(\sum_{j=1}^{m}\|x^{[j]}\|_2^2\max_{1\le j\le m}\sum_{i=1}^{m}\|A^{[i,j]}\|_2)^{1/2} \\
&\le \max_{1\le i\le m}((\sum_{j=1}^{m}\|A^{[i,j]}\|_2)^{1/2})\max_{1\le j\le m}((\sum_{i=1}^{m}\|A^{[i,j]}\|_2)^{1/2})\|x\|_2\ ,
\end{aligned}$$

womit alles gezeigt ist. Q.E.D.

Lemma 1.0.13 *Eine schiefsymmetrische, skaliert $\gamma-2-$d.d. Matrix ist auch skaliert $\gamma-$d.d..*

Beweis:
Sei $\tilde{J}_n$ gemäß (1.11) und $S = \Delta(\tilde{J}_n + R)\Delta \in I\!R^{2n\times 2n}$. skaliert $\gamma-2-$d.d.. Dabei sei $R = (R^{[i,j]})_{1\le i,j\le n}$ mit $R^{[j,j]} = 0$, $j = 1\ldots n$. Dann ist

$$\|R\|_2 \le |||R|||_2 = \max_{1\le j\le n}\sum_{i=1}^{n}\|R^{[i,j]}\|_2 \le \gamma\ .$$

Q.E.D.

Auch die skaliert $\gamma - 2-$d.d. Matrizen bestimmen ihre Eigenwerte gut:

Korollar 1.0.14 *Die Sätze 1.0.8, 1.0.9, 1.0.10 behalten ihre Gültigkeit, wenn „skaliert $\gamma-$d.d.“ durch „skaliert $\gamma - 2-$d.d.“ und die Spektralnorm $\|\cdot\|_2$ durch die Norm $|||\cdot|||_2$ ersetzt werden.*

Beweis:
Triviale Anwendung des letzten Lemmas. Q.E.D.

Kapitel 2

Jacobi-ähnliche Verfahren

Ein Hauptanliegen vorliegender Arbeit ist die Vorstellung eines neuen Verfahrens zur Lösung des nichtsingulären schiefsymmetrischen Eigenwertproblems $Sx = \lambda x$. Den ersten Teil dieses Verfahrens, nämlich die schiefsymmetrische Zerlegung

$$PSP^T = LJL^T \quad , \quad P \text{ Permutationsmatrix} \quad , \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \quad , \quad I \in I\!R^{n \times n}$$

werden wir erst in Kapitel 4 behandeln. Im vorliegenden Kapitel gehen wir hingegen davon aus, der Faktor L der schiefsymmetrischen Zerlegung liege bereits vor. S könnte z.B. auch durch einen gegebenen Faktor L definiert sein. Wir betrachten dann die symmetrische, positiv definite Matrix $K = L^T L$. Es gibt symplektisches T, das K diagonalisiert[1], d.h.

$$T^T K T = T^T L^T L T = \Delta^2 \quad , \quad \Delta = E \oplus E \ , \quad E = \text{diag} \tag{2.1}$$

Setzt man $V = LT$, so ist $U = V\Delta^{-1}$ offensichtlich orthogonal. Wegen der äquivalenz von $T^T J T = J$ und $T J T^T = J$ gilt außerdem $PSP^T = LJL^T = LTJT^T L^T = U\Delta J \Delta U^T$, d.h. $U^T PSP^T U = \Delta J \Delta$. Die Quadrate der Spaltennormen von V sind also die positiven Imaginärteile der Eigenwerte von S, während die normierten Spalten orthogonale Eigenvektoren von S sind. Die Eigenwertprobleme (S, I) und (K, J) sind äquivalent.

[1] Der Existenzbeweis sei wie folgt skizziert: Es gibt ein orthogonales Q und ein positiv definites diagonales E mit

$$Q^T S Q = \Delta J \Delta \quad , \quad \Delta = E \oplus E \ .$$

Mit obiger Zerlegung von S haben wir daher $\Delta^{-1} Q^T P^T L J L^T P Q \Delta^{-1} = J$, also ist $\Delta^{-1} Q^T P^T L$ und seine Inverse symplektisch und es gilt

$$(\Delta^{-1} Q^T P^T L)^{-T} L^T L (\Delta^{-1} Q^T P^T L)^{-1} = \Delta^2 \ .$$

Im vorliegenden Kapitel werden wir iterative Verfahren zur Durchführung von (2.1) vorstellen. Zuvor werden wir jedoch in Abschnitt 2.1 Strukturen solcher K_S untersuchen, für die (K, J) mehrfache Eigenwerte hat, denn an diesen Strukturen werden wir einen Teil der Verfahren orientieren. Dann werden wir ein explizites Verfahren zur Diagonalisierung von K in (2.1) vorstellen. Nach der Entwicklung des Verfahrens geben wir den formalen Algorithmus an. Hauptsächlich werden wir uns aber einem äquivalenten impliziten Verfahren widmen, mit dem die Spalten von L in (2.1) orthogonalisiert werden. Auch vom impliziten Verfahren werden wir den recht komplizierten formalen Algorithmus aufschreiben. Die Darstellung des expliziten Verfahrens dient eigentlich nur dem besseren Verständnis des impliziten Verfahrens. Das implizite Verfahren basiert auf einer Idee von Veselić [31]. Es fand auch in [32] Verwendung. Ein Abschnitt über die globale Konvergenz der Verfahren schließt das Kapitel ab.

2.1 Strukturen bei mehrfachen Eigenwerten

Bereits in [36] wurde gezeigt, daß Matrizen A mit mehrfachen Eigenwerten und genügend kleinem $\|A - \mathrm{diag}(A)\|_2$ gewisse Strukturen zeigen. Auch [16] enthält derartige Aussagen für Matrizenpaare $(A, I_n \oplus (-I_m))$, A symmetrisch. Man kann analoge Aussagen auch für das Paar (K, J) zeigen. Weil wir in vorliegender Arbeit jedoch vorwiegend relative Aussagen über Eigenwerte formulieren wollen, betrachten wir in diesem Abschnitt Strukturaussagen für die skalierte Matrix K_S, wenn $\|K_S - I\|_2$ klein genug ist und (K, J) mehrfache Eigenwerte hat. Auf Strukturaussagen für K selbst verzichten wir hingegen aus Platzgründen. Anschließend werden wir strukturerhaltende Transformationen einführen und deren Eigenschaften betrachten. Einige Aussagen über Matrizenpaare (K, J) mit nur einem einzigen Eigenwert schließen den Abschnitt ab.

In diesem Abschnitt können wir z.T. ähnliche Beweistechniken wie in [16], [36] verwenden, obwohl die Verhältnisse in diesem Abschnitt zumeist komplizierter sind.

Sei

$$\begin{aligned} K &= (k_{ij}) \qquad (2.2) \\ &= \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \quad \text{mit} \quad B = (b_{ij}) \ , \ C = (c_{ij}) \ , \ F = (f_{ij}) \in \mathbb{R}^{n \times n} \ , \ n \geq 2 \end{aligned}$$

symmetrisch und positiv definit, wobei außerdem

$$f_{jj} = c_{jj} \quad \text{und} \quad f_{jj} \geq f_{j+1,j+1} \quad \text{für alle möglichen } j \ . \qquad (2.3)$$

gelte. Diese Eigenschaften lassen sich durch symplektische Skalierungen bzw. symplektische Permutation erzielen. In diesem Abschnitt werden wir kleine Buchstaben für Matrixelemente und große Buchstaben für Submatrizen verwenden. Die

skalierte Matrix K_S durch

$$\begin{aligned} K_S &= \begin{bmatrix} F_S & B_S^T \\ B_S & C_S \end{bmatrix} \\ &= \begin{bmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} K \begin{bmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \quad , \quad D = \mathrm{diag}(\sqrt{f_{jj}}) \end{aligned} \tag{2.4}$$

definiert. Die Eigenwerte $i\lambda_j = -i\lambda_{2n+1-j}$, $j = 1 \ldots 2n$, $i^2 = -1$, des Paares (K, J) seien gemäß

$$\begin{aligned} & \lambda_1 = \ldots = \lambda_{s_1} > \lambda_{s_1+1} = \ldots = \lambda_{s_2} > \ldots > \lambda_{s_{q-1}+1} = \ldots = \lambda_{s_q} > 0 \\ > \; & \lambda_{s_q+1} = \ldots = \lambda_{s_{q+1}} > \ldots > \lambda_{s_{p-1}+1} = \ldots = \lambda_{s_p} \end{aligned} \tag{2.5}$$

mit $s_q = n$ und $s_p = 2n$ geordnet. Aus Konsistenzgründen definieren wir zusätzlich

$$\lambda_{s_0} = \infty \quad , \quad \lambda_{s_{p+1}} = -\infty \quad , \quad s_0 = 0 \quad , \quad s_{p+1} = 2n+1 \ ,$$

und damit sei

$$n_r = s_r - s_{r-1} \, , \ 1 \le r \le p$$

die Vielfachheit des Eigenwertes $i\lambda_{s_r}$, $1 \le r \le p$. B_S, C_S, F_S und D seien gemäß $n_1 \ldots n_q$ partitioniert, also

$$F_S = \begin{bmatrix} F_{S11} & \cdots & F_{S1q} \\ \vdots & \ddots & \vdots \\ F_{Sq1} & \cdots & F_{Sqq} \end{bmatrix} \quad , \quad B_S, C_S \ \text{ analog,} \tag{2.6}$$

und es sei

$$K_r = \begin{bmatrix} F_{rr} & B_{rr}^T \\ B_{rr} & C_{rr} \end{bmatrix} \quad , \quad K_{Sr} = \begin{bmatrix} F_{Srr} & B_{S_{rr}}^T \\ B_{Srr} & C_{Srr} \end{bmatrix} \quad , \quad 1 \le r \le q \ .$$

Sei außerdem

$$\begin{aligned} \varepsilon_r &= \min\{\frac{|\lambda_{s_r} - \lambda_{s_{r+1}}|}{\max\{|\lambda_{s_r}|, |\lambda_{s_{r+1}}|\}} \ , \ \frac{|\lambda_{s_{r-1}} - \lambda_{s_r}|}{\max\{|\lambda_{s_{r-1}}|, |\lambda_{s_r}|\}}\} \\ \varepsilon &= \min_{1 \le r \le p} \varepsilon_r \ . \end{aligned} \tag{2.7}$$

Dann gilt im übrigen $\varepsilon \le \varepsilon_r \le 1$.

Lemma 2.1.1 *Seien K und K_S gemäß (2.2) bis (2.4) mit den Eigenschaften (2.5) bis (2.7) gegeben. Sei $\|K_S - I\|_2 \le \varepsilon/(\gamma + \varepsilon)$ für ein $\gamma > 2$. Dann gilt für jedes r, $1 \le r \le q$,*

$$\begin{aligned} & (\|F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2}\|_E^2 + \|B_{S_{rr}}^T - B_{Srr}\|_E^2)^{1/2} \\ \le \; & \frac{\gamma + \varepsilon_r}{2\varepsilon_r(\gamma - 2)} (\|F_{Srr} - C_{Srr}\|_E^2 + \|B_{S_{rr}}^T + B_{Srr}\|_E^2 \\ & \quad + 2 \sum_{j=1, j \ne r}^{q} (\|F_{Srj}\|_E^2 + \|C_{Srj}\|_E^2 + \|B_{Srj}\|_E^2 + \|B_{Sjr}\|_E^2)) \\ \le \; & \frac{\gamma + \varepsilon}{(\gamma - 2)\,\varepsilon} \, \|K_S - I\|_E^2 \ . \end{aligned}$$

Dasselbe gilt für $\lambda_{s_{p+1-r}}$ statt λ_{s_r}.

Beweis:
Die Beweistechnik orientiert sich an [36]. Für beliebiges r, $1 \leq r \leq q$, und $i^2 = -1$ betrachten wir die Matrix

$$\begin{aligned} \tilde{M}_{S_r} &= K_S + i\lambda_{s_r} \cdot \begin{bmatrix} 0 & D^{-2} \\ -D^{-2} & 0 \end{bmatrix} \\ &= \begin{bmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} (K + i\lambda_{s_r} J) \begin{bmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} . \end{aligned}$$

Sie hat Rang $2n-n_r$. Es gibt eine symplektische Permutationsmatrix $P_r = \hat{P}_r \oplus \hat{P}_r$, für die

$$P_r^T \tilde{M}_{S_r} P_r = \begin{bmatrix} F_{Srr} & F_{S_r'}^T & B_{S_{rr}}^T + i\lambda_{s_r} D_{rr}^{-2} & B_{S_r'}^T \\ F_{S_r'} & F_{S_r''} & B_{S_r''}^T & B_{S_r'''}^T + i\lambda_{s_r} \tilde{D}_{rr}^{-2} \\ B_{Srr} - i\lambda_{s_r} D_{rr}^{-2} & B_{S_r''} & C_{Srr} & C_{S_r'}^T \\ B_{S_r'} & B_{S_r'''} - i\lambda_{s_r} \tilde{D}_{rr}^{-2} & C_{S_r'} & C_{S_r''} \end{bmatrix}$$

gilt, wobei $D_{rr} \oplus \tilde{D}_{rr} = \hat{P}_r^T D \hat{P}_r$ ist. Mit Hilfe der unitären Matrix

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} I_n & iI_n \\ iI_n & I_n \end{bmatrix}$$

definieren wir

$$\begin{aligned} M_{Sr} &= U^* P_r^T \tilde{M}_{S_r} P_r U \\ &= \begin{bmatrix} \frac{1}{2}(F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2} + i(B_{S_{rr}}^T - B_{Srr})) & G_{S_r}^* \\ G_{S_r} & H_{S_r} \end{bmatrix} , \end{aligned}$$

die ebenso Rang $2n - n_r$ hat und deren Submatrizen uns nun interessieren. Ist H_{S_r} invertierbar, so gilt

$$\begin{aligned} M_{Sr} = & \begin{bmatrix} I_{n_r} & G_{S_r}^* H_{S_r}^{-1} \\ 0 & I_{2n-n_r} \end{bmatrix} \cdot \\ & \cdot \begin{bmatrix} \frac{1}{2}(F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2} + i(B_{S_{rr}}^T - B_{Srr})) - G_{S_r}^* H_{S_r}^{-1} G_{S_r} & 0 \\ G_{S_r} & H_{S_r} \end{bmatrix} \end{aligned}$$

und die obere Blockdiagonalmatrix ist hermitesch vom Rang 0, weil H_{S_r} Rang $2n - n_r$ hat, also ist

$$\frac{1}{2}(F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2} + i(B_{S_{rr}}^T - B_{Srr})) = G_{S_r}^* H_{S_r}^{-1} G_{S_r} .$$

Daraus folgt

$$\begin{aligned}
&\|F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2} + i(B_{S_{rr}}^T - B_{Srr})\|_E \\
&= (\|F_{Srr} + C_{Srr} - 2\lambda_{s_r} D_{rr}^{-2}\|_E^2 + \|B_{S_{rr}}^T - B_{Srr}\|_E^2)^{1/2} \\
&\leq 2\|G_{Sr}\|_E^2 \, \|H_{S_r}^{-1}\|_2
\end{aligned}$$

mit

$$\begin{aligned}
2\|G_{Sr}\|_E^2 &= \frac{1}{2}(\|F_{Srr} - C_{Srr}\|_E^2 + \|B_{Srr} + B_{S_{rr}}^T\|_E^2 \\
&\quad +2 \sum_{j=1, j\neq r}^{q} (\|F_{Srj}\|_E^2 + \|C_{Srj}\|_E^2 + \|B_{Srj}\|_E^2 + \|B_{Sjr}\|_E^2)) \; .
\end{aligned}$$

Also ist zum Beweis der ersten Ungleichung des Lemmas nur noch $\|H_{S_r}^{-1}\|_2 \leq (\gamma + \varepsilon_r)/(\varepsilon_r(\gamma - 2))$ zu zeigen. Dazu reicht es aus zu beweisen, daß die Beträge aller Eigenwerte von H_{S_r} größer oder gleich $\varepsilon_r(\gamma - 2)/(\gamma + \varepsilon_r)$ sind, womit wegen $\varepsilon > 0$ gleichzeitig auch die Invertierbarkeit von H_{S_r} gezeigt ist. Wir setzen nun

$$\tilde{k}_{jj} = \begin{cases} k_{jj} & , \quad 1 \leq j \leq n \\ -k_{2n+1-j,2n+1-j} & , \quad n+1 \leq j \leq 2n \; . \end{cases}$$

Für $\gamma > 2$ gilt nach Voraussetzung $\|K_S - I\|_2 \leq \varepsilon/(\gamma + \varepsilon) < 1$. Daher kann Satz 1.0.6 auf diag(K) und die Störung

$$\delta K = \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} (K_S - I) \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix}$$

angewendet werden und aus (1.9) folgt

$$\frac{|\tilde{k}_{jj} - \lambda_j|}{|\tilde{k}_{jj}|} \leq \frac{\|K_S - I\|_2}{\lambda_{\min}(I)} \leq \frac{\varepsilon}{\gamma + \varepsilon} \quad , \quad 1 \leq j \leq 2n \; . \tag{2.8}$$

Durch Umformung erhalten wir

$$\frac{|\lambda_j|}{|\tilde{k}_{jj}|} \geq \frac{\gamma}{\gamma + \varepsilon} \quad , \quad 1 \leq j \leq 2n \; , \tag{2.9}$$

und daraus

$$\frac{|\tilde{k}_{jj} - \lambda_j|}{|\lambda_j|} = \frac{|\tilde{k}_{jj} - \lambda_j|}{|\tilde{k}_{jj}|} \cdot \frac{|\tilde{k}_{jj}|}{|\lambda_j|} \leq \frac{\varepsilon}{\gamma + \varepsilon} \cdot \frac{\gamma + \varepsilon}{\gamma} = \frac{\varepsilon}{\gamma} \quad , \quad 1 \leq j \leq 2n \; . \tag{2.10}$$

Sei nun $\Omega = \{1, \ldots, 2n\}$ und

$$\Omega_l = \{j \in \Omega \mid \frac{|\tilde{k}_{jj} - \lambda_{s_l}|}{|\lambda_{s_l}|} \leq \frac{\varepsilon}{\gamma}\} \quad , \quad 1 \leq l \leq p \; .$$

Wir werden zeigen, daß die Ω_l disjunkt sind. Sei dazu $1 \le l, r \le p$, $l \ne r$. Wegen (2.5), (2.10) ist $\{s_{r-1}+1, \ldots, s_r\} \subset \Omega_r$ und nach (2.7) gilt

$$\frac{|\lambda_{s_r} - \lambda_{s_l}|}{\max\{|\lambda_{s_r}|, |\lambda_{s_l}|\}} \ge \max\{\varepsilon_r, \varepsilon_l\} \ .$$

Für $j \in \Omega_r$ ist daher

$$\begin{aligned} \frac{|\tilde{k}_{jj} - \lambda_{s_l}|}{|\lambda_{s_l}|} &\ge \frac{|\lambda_{s_r} - \lambda_{s_l}| - |\tilde{k}_{jj} - \lambda_{s_r}|}{\max\{|\lambda_{s_r}|, |\lambda_{s_l}|\}} \ge \frac{|\lambda_{s_r} - \lambda_{s_l}|}{\max\{|\lambda_{s_r}|, |\lambda_{s_l}|\}} - \frac{|\tilde{k}_{jj} - \lambda_{s_r}|}{|\lambda_{s_r}|} \\ &\ge \max\{\varepsilon_r, \varepsilon_l\} - \frac{\varepsilon}{\gamma} > \frac{2\varepsilon}{\gamma} - \frac{\varepsilon}{\gamma} = \frac{\varepsilon}{\gamma} \ , \end{aligned} \tag{2.11}$$

d.h. $j \notin \Omega_l$. Wir haben also $\Omega_r = \{s_{r-1}+1, \ldots, s_r\}$, $1 \le r \le q$. Seien $\eta_{S_j}^{(r)}$, $j \in \Omega \backslash \Omega_r$, die Eigenwerte von H_{S_r}. Sie können so geordnet werden, daß aus dem Satz von Weyl [20]

$$|\eta_{S_j}^{(r)} - (1 - \frac{\lambda_{s_r}}{\tilde{k}_{jj}})| \le \|H_{S_r} - \mathrm{diag}(H_{S_r})\|_2 \le \|M_{S_r} - \mathrm{diag}(M_{S_r})\|_2$$

folgt. Aus $\|M_{S_r} - \mathrm{diag}(M_{S_r})\|_2 = \|K_S - I\|_2$, was wir im nachfolgenden Lemma noch zeigen werden, ergibt sich dann

$$|\eta_{S_j}^{(r)} - (1 - \frac{\lambda_{s_r}}{\tilde{k}_{jj}})| \le \|K_S - I\|_2 \le \frac{\varepsilon}{\gamma + \varepsilon} \ .$$

Wegen $j \notin \Omega_r$ gibt es ein $l \ne r$ mit $j \in \Omega_l$ und aus (2.7), (2.8), (2.9) folgt

$$\begin{aligned} |\eta_{S_j}^{(r)}| &\ge |1 - \frac{\lambda_{s_r}}{\tilde{k}_{jj}}| - \frac{\varepsilon}{\gamma + \varepsilon} = \frac{|\tilde{k}_{jj} - \lambda_{s_l} + \lambda_{s_l} - \lambda_{s_r}|}{|\tilde{k}_{jj}|} - \frac{\varepsilon}{\gamma + \varepsilon} \\ &\ge \frac{|\lambda_{s_l} - \lambda_{s_r}|}{|\lambda_{s_l}|} \cdot \frac{|\lambda_{s_l}|}{|\tilde{k}_{jj}|} - \frac{|\tilde{k}_{jj} - \lambda_{s_l}|}{|\tilde{k}_{jj}|} - \frac{\varepsilon}{\gamma + \varepsilon} \\ &\ge \max\{\varepsilon_r, \varepsilon_l\} \frac{\gamma}{\gamma + \varepsilon} - \frac{\varepsilon}{\gamma + \varepsilon} - \frac{\varepsilon}{\gamma + \varepsilon} \\ &\ge \frac{\varepsilon_r}{\gamma + \varepsilon_r} \cdot (\gamma - 2) \ . \end{aligned}$$

Damit ist die erste Ungleichung des Lemmas bewiesen. Die zweite Ungleichung folgt aus

$$\begin{aligned} &\|F_{S_{rr}} - C_{S_{rr}}\|_E^2 + \|B_{S_{rr}}^T + B_{S_{rr}}\|_E^2 \\ &\qquad + 2 \sum_{j=1, j \ne r}^{q} (\|F_{S_{rj}}\|_E^2 + \|C_{S_{rj}}\|_E^2 + \|B_{S_{rj}}\|_E^2 + \|B_{S_{jr}}\|_E^2) \\ &\le 2\|F_{S_{rr}} - I\|_E^2 + 2\|C_{S_{rr}} - I\|_E^2 + 4\|B_{S_{rr}}\|_E^2 \end{aligned}$$

$$
\begin{aligned}
& +2\sum_{j=1,j\neq r}^{q}(\|F_{S_{rj}}\|_E^2+\|C_{S_{rj}}\|_E^2+\|B_{S_{rj}}\|_E^2+\|B_{S_{jr}}\|_E^2) \\
= \;& 2(\|F_{S_{rr}}-I\|_E^2+\sum_{j=1,j\neq r}^{q}\|F_{S_{rj}}\|_E^2+\|C_{S_{rr}}-I\|_E^2+\sum_{j=1,j\neq r}^{q}\|C_{S_{rj}}\|_E^2 \\
& +\sum_{j=1}^{q}(\|B_{S_{rj}}\|_E^2+\|B_{S_{jr}}\|_E^2)) \\
\leq \;& 2\|K_S-I\|_E^2 \; .
\end{aligned}
$$

Schließlich gelten die beiden Ungleichungen des Lemmas wegen $|\lambda_{s_{p+1-r}}|=|\lambda_{s_r}|$ auch für $\lambda_{s_{p+1-r}}$ statt λ_{s_r}. Q.E.D.

Folgendes Lemma tragen wir zur Vervollständigung vorstehenden Beweises noch nach.

Lemma 2.1.2 *Sei $K=(k_{ij})\in I\!R^{2n\times 2n}$ symmetrisch und positiv definit mit $k_{jj}=k_{j+n,j+n}$, $j=1\ldots n$. Sei $P=\hat{P}\oplus\hat{P}\in I\!R^{2n\times 2n}$ eine symplektische Permutationsmatrix. Seien*

$$
\Delta=\begin{bmatrix} 0 & \hat{\Delta} \\ -\hat{\Delta} & 0 \end{bmatrix} \; , \quad U=\frac{1}{\sqrt{2}}\begin{bmatrix} I & iI \\ iI & I \end{bmatrix}\in I\!R^{2n\times 2n}
$$

mit diagonalem $\hat{\Delta}$. Sei $\|\cdot\|$ eine unitär invariante Norm. Für

$$
\tilde{K}=U^*P^*(K-i\Delta)PU
$$

gilt dann

$$
\|K-diag(K)\|=\|\tilde{K}-diag(\tilde{K})\| \; .
$$

Beweis:
Es gilt

$$
\begin{aligned}
\tilde{K}-\mathrm{diag}(\tilde{K}) \;=\; & U^*P^*KPU-iU^*P^*\Delta PU \\
& -\mathrm{diag}(U^*P^*KPU)-i\mathrm{diag}(U^*P^*\Delta PU)
\end{aligned}
$$

Die Matrix $U^*P^*\Delta PU$ ist bereits diagonal. Wegen

$$
(P^*KP)_{jj}=(P^*KP)_{j+n,j+n} \quad , \quad j=1\ldots n \; ,
$$

gilt

$$
\mathrm{diag}(U^*P^*KPU)=\mathrm{diag}(P^*KP)=U^*\mathrm{diag}(P^*KP)U=U^*P^*\mathrm{diag}(K)PU \; .
$$

Daraus folgt

$$
\begin{aligned}
\|\tilde{K}-\mathrm{diag}(\tilde{K})\| \;&=\; \|U^*P^*KPU-\mathrm{diag}(U^*P^*KPU)\| \\
&=\; \|U^*P^*KPU-U^*P^*\mathrm{diag}(K)PU\|
\end{aligned}
$$

und wegen der unitären Invarianz der Norm ergibt sich daraus bereits die Aussage des Lemmas. Q.E.D.

Die Aussage von Lemma 2.1.1 ändert sich nicht, wenn dort statt K ein symplektisch permutiertes $P^T K P$ betrachtet wird, solange noch eine Zuordnung der Diagonalelemente zu Eigenwerten möglich ist. Dies bedeutet, daß die Voraussetzung (2.5) zu

$$\lambda_1 = \ldots = \lambda_{s_1} \quad , \quad \lambda_{s_1+1} = \ldots = \lambda_{s_2} \quad , \quad \ldots \quad , \quad \lambda_{s_{q-1}+1} = \ldots = \lambda_{s_q} > 0$$

mit $\lambda_j = -\lambda_{2n+1-j}$ abgeschwächt werden kann, wenn die Diagonalelemente von F und C geeignet den Eigenwerten zugeordnet werden können [16]. Dies wollen wir nicht in voller Allgemeinheit vertiefen. Eine Aussage in dieser Richtung liefert jedoch das nachstehende Korollar, auf das wir noch zurückkommen werden.

Korollar 2.1.3 *Sei $K' = D' K'_S D'$ symmetrisch und positiv definit mit*

$$K'_S = \begin{bmatrix} F'_S & B'^T_S \\ B'_S & C'_S \end{bmatrix} \quad , \quad D' = \begin{bmatrix} D_F & 0 \\ 0 & D_C \end{bmatrix}$$

und $F'_S, B'_S, C'_S \in \mathbb{R}^{n\times n}$ und diagonalen $D'_F, C'_F \in \mathbb{R}^{n\times n}$, wobei die Diagonalelemente K'_S gleich 1 seien. Es gebe ein symplektisches diagonales T sowie eine symplektische Permutationsmatrix P mit $K' = P^T T^T K T P$, und K erfülle die Voraussetzungen von Lemma 2.1.1. Partitioniert man K'_S genauso wie K_S und ist $P = Q \oplus Q$, $Q = \oplus_{r=1}^{q} Q_r$, $Q_r \in \mathbb{R}^{n_r \times n_r}$, so gilt für jedes r, $1 \le r \le q$,

$$\begin{aligned}
&(\|F'_{Srr} + C'_{Srr} - 2\lambda_{s_r}(D_{F_{rr}} D_{C_{rr}})^{-1}\|_E^2 + \|B'^T_{Srr} - B'_{Srr}\|_E^2)^{1/2} \\
&\le \frac{\gamma + \varepsilon_r}{2\varepsilon_r(\gamma - 2)} (\|F'_{Srr} - C'_{Srr}\|_E^2 + \|B'^T_{Srr} + B'_{Srr}\|_E^2 \\
&\qquad + 2 \sum_{j=1, j\neq r}^{q} (\|F'_{Srj}\|_E^2 + \|C'_{Srj}\|_E^2 + \|B'_{Srj}\|_E^2 + \|B'_{Sjr}\|_E^2)) \\
&\le \frac{\gamma + \varepsilon}{(\gamma - 2)\,\varepsilon} \, \|K'_S - I\|_E^2 \; .
\end{aligned}$$

Dasselbe gilt für $\lambda_{s_{p+1-r}}$ statt λ_{s_r}.

Beweis:
Es gilt $K'_S = P^T K_S P$ und daher für alle r, s, $1 \le r, s \le q$

$$F'_{Srs} = Q_r^T F_{Srs} Q_s \; , \quad B'_{Srs} = Q_r^T B_{Srs} Q_s \; , \quad C'_{Srs} = Q_r^T C_{Srs} Q_s \; .$$

Die Eigenwerte von (K, J) und (K', J) sind gleich. Und schließlich gilt $D_{Frr} = Q_r^T T_{rr} D_{rr} Q_r$, $D_{Crr} = Q_r^T T_{rr}^{-1} D_{rr} Q_r$, $r = 1 \ldots q$, wenn man T genauso wie die anderen Matrizen partitioniert. Daraus folgt $D_{F_{rr}} D_{C_{rr}} = Q_r^T D_{rr}^2 Q_r$. Die Aussage des Korollars ergibt sich nun durch Einsetzen in Lemma 2.1.1. Q.E.D.

Mit folgendem Satz vereinfachen wir die Aussage von Korollar 2.1.3. Unter Vernachlässigung der Details macht er die Struktur der skalierten Matrix K_S bei mehrfachen Eigenwerten von (K, J) und kleinem $\|K_S - I\|$ deutlich.

Satz 2.1.4 *K' erfülle die Voraussetungen von Korollar 2.1.3. Dann gilt*

$$\begin{aligned}
&\sum_{r=1}^{q}(\|F'_{Srr} + C'_{Srr} - 2I_{n_r}\|_E^2 + \|{B'_{Srr}}^T - B'_{Srr}\|_E^2) \\
&\quad\leq \sum_{r=1}^{q}(\|F'_{Srr} + C'_{Srr} - 2\lambda_{s_r}(D_{Frr}D_{Crr})^{-1}\|_E^2 + \|{B'_{Srr}}^T - B'_{Srr}\|_E^2) \\
&\quad\leq \frac{(\gamma+\varepsilon)^2}{(\gamma-2)^2\varepsilon^2}\,\|K'_S - I\|_E^4\ .
\end{aligned}$$

Beweis:
Die Gültigkeit der ersten Ungleichung ist offensichtlich. Die zweite Ungleichung folgt durch Anwendung von Korollar 2.1.3 aus

$$\begin{aligned}
&\sum_{r=1}^{q}(\|F'_{Srr} + C'_{Srr} - 2\lambda_{s_r}(D_{Frr}D_{Crr})^{-1}\|_E^2 + \|{B'_{Srr}}^T - B'_{Srr}\|_E^2) \\
&\quad\leq \sum_{r=1}^{q}\frac{(\gamma+\varepsilon_r)^2\,\|K'_S - I\|_E^2}{2\varepsilon_r^2(\gamma-2)^2}\cdot(\|F'_{Srr} - C'_{Srr}\|_E^2 + \|{B'_{Srr}}^T + B'_{Srr}\|_E^2 \\
&\qquad +2\sum_{j=1,j\neq r}^{q}(\|F'_{Srj}\|_E^2 + \|C'_{Srj}\|_E^2 + \|B'_{Srj}\|_E^2 + \|B'_{Sjr}\|_E^2)) \\
&\quad\leq \frac{(\gamma+\varepsilon)^2\,\|K'_S - I\|_E^2}{2\varepsilon^2(\gamma-2)^2}\cdot 2\sum_{r=1}^{q}(\|F'_{Srr} - I_{n_r}\|_E^2 + \|C'_{Srr} - I_{n_r}\|_E^2 + 2\,\|B'_{Srr}\|_E^2 \\
&\qquad + \sum_{j=1,j\neq r}^{q}(\|F'_{Srj}\|_E^2 + \|C'_{Srj}\|_E^2 + \|B'_{Srj}\|_E^2 + \|B'_{Sjr}\|_E^2)) \\
&\quad\leq \frac{(\gamma+\varepsilon)^2}{\varepsilon^2(\gamma-2)^2}\,\|K'_S - I\|_E^4\ .
\end{aligned}$$

Q.E.D.

Satz 2.1.4 gilt auch im Falle getrennter Eigenwerte von (K', J). Dann lautet seine Aussage

$$(\sum_{r=1}^{n}(\frac{f_{rr}'^{1/2}c_{rr}'^{1/2} - \lambda_r}{f_{rr}'^{1/2}c_{rr}'^{1/2}})^2)^{1/2} \leq \frac{\gamma+\varepsilon}{2\,(\gamma-2)\,\varepsilon}\,\|K'_S - I\|_E^2\ ,$$

wobei die f'_{rr}, c'_{rr} Diagonalelemente von K' sind. Ist also $\alpha = \|K'_S - I\|_E$ klein genug, so sind die relativen Fehler der skalierten Diagonalelemente von K' gegenüber den Beträgen der Eigenwerte von (K', J) von der Ordnung $\mathcal{O}(\alpha^2)$.

Bei Verfahren zur Berechnung der Eigenwerte von (K, J) wollen wir den Fall mehrfacher Eigenwerte besonders berücksichtigen. Sind mehrfache Eigenwerte vorhanden und erfüllt K die Voraussetzungen von Lemma 2.1.1, so sollen gerade die Submatrizen K_r transformiert werden, und zwar so, daß die Normen $\|F_{Srr} - I\|_E$, $\|C_{Srr} - I\|_E$ und $\|B_{Srr}\|_E$ nach diesen Transformationen von der Ordnung $\mathcal{O}(\omega^2)$ sind, wenn sie zuvor von der Ordnung $\mathcal{O}(\omega)$ mit einem geeigneten ω waren.

Wir werden nun Transformationen mit der soeben angedeuteten Wirkung einführen und deren Eigenschaften untersuchen. Zuvor jedoch geben weitere Folgerungen von Lemma 2.1.1 an. Es sind algorithmisch leicht nachprüfbare Eigenschaften von K, die uns später der Formulierung von Bedingungen dienen, unter denen vorgenannte Transformationen vorgenommen werden. Wir benötigen die folgenden beiden Hilfsaussagen.

Lemma 2.1.5 *Seien die Voraussetzungen von Korollar 2.1.3 erfüllt und sei $\gamma > 1 + \sqrt{2}$. Setzt man*

$$\varrho = \frac{\gamma + \varepsilon}{\sqrt{2}(\gamma - 2)\,\varepsilon} \|K'_S - I\|_E^2 \ ,$$

so gilt für alle r, $1 \le r \le q$, und alle i, j mit $s_{r-1} + 1 \le i, j \le s_r$

$$\frac{|f_{ii}^{1/2} c_{ii}^{1/2} - f_{jj}^{1/2} c_{jj}^{1/2}|}{f_{ii}^{1/2} c_{ii}^{1/2} + f_{jj}^{1/2} c_{jj}^{1/2}} \le \varrho \tag{2.12}$$

und

$$\frac{f_{ii}^{1/2} c_{ii}^{1/2}}{f_{jj}^{1/2} c_{jj}^{1/2}} \le \frac{1}{1 - 2\varrho} \ . \tag{2.13}$$

Beweis:
Aus $|f_{ii}^{1/2} c_{ii}^{1/2} - \lambda_{s_r}| \ge |f_{ii}^{1/2} c_{ii}^{1/2} - f_{jj}^{1/2} c_{jj}^{1/2}| - |f_{jj}^{1/2} c_{jj}^{1/2} - \lambda_{s_r}|$ und Korollar 2.1.3 folgt

$$\begin{aligned}
\frac{|f_{ii}^{1/2} c_{ii}^{1/2} - f_{jj}^{1/2} c_{jj}^{1/2}|}{f_{ii}^{1/2} c_{ii}^{1/2} + f_{jj}^{1/2} c_{jj}^{1/2}} &\le \frac{|f_{ii}^{1/2} c_{ii}^{1/2} - \lambda_{s_r}|}{f_{ii}^{1/2} c_{ii}^{1/2}} + \frac{|f_{jj}^{1/2} c_{jj}^{1/2} - \lambda_{s_r}|}{f_{jj}^{1/2} c_{jj}^{1/2}} \\
&\le \sqrt{2}\Big(\frac{(f_{ii}^{1/2} c_{ii}^{1/2} - \lambda_{s_r})^2}{f_{ii} c_{ii}} + \frac{(f_{jj}^{1/2} c_{jj}^{1/2} - \lambda_{s_r})^2}{f_{jj} c_{jj}}\Big)^{1/2} \\
&= \frac{1}{\sqrt{2}}\Big((2 - 2\frac{\lambda_{s_r}}{f_{ii}^{1/2} c_{ii}^{1/2}})^2 + (2 - 2\frac{\lambda_{s_r}}{f_{jj}^{1/2} c_{jj}^{1/2}})^2\Big)^{1/2} \\
&\le \frac{\gamma + \varepsilon}{\sqrt{2}(\gamma - 2)\,\varepsilon} \|K'_S - I\|_E^2
\end{aligned}$$

Damit ist (2.12) bewiesen. Nun zum Beweis von (2.13). Ist $f_{ii} c_{ii} \le f_{jj} c_{jj}$, so ist nichts zu zeigen. Andernfalls sei $\gamma > 1 + \sqrt{2}$. Dann ist $\varrho < 1/2$ und aus

$$1 - \frac{f_{jj}^{1/2} c_{jj}^{1/2}}{f_{ii}^{1/2} c_{ii}^{1/2}} = 2\frac{f_{ii}^{1/2} c_{ii}^{1/2} - f_{jj}^{1/2} c_{jj}^{1/2}}{2 f_{ii}^{1/2} c_{ii}^{1/2}} \le 2\frac{f_{ii}^{1/2} c_{ii}^{1/2} - f_{jj}^{1/2} c_{jj}^{1/2}}{f_{ii}^{1/2} c_{ii}^{1/2} + f_{jj}^{1/2} c_{jj}^{1/2}} \le 2\varrho$$

folgt schließlich (2.13). Q.E.D.

Lemma 2.1.6 *Seien die Voraussetzungen von Korollar 2.1.3 erfüllt und $\gamma \geq 3$. f'_{ii} und c'_{ii}, $1 \leq i \leq n$ seien die Diagonalelemente von K'. Dann gilt für r, $1 \leq r \leq q-1$, und alle i, j mit $i \leq s_r < j$*

$$1 + (\gamma - 2)\,\|K'_S - I\|_E < \frac{f'^{1/2}_{ii} c'^{1/2}_{ii}}{f'^{1/2}_{jj} c'^{1/2}_{jj}} \ .$$

Beweis:
Sei $1 \leq r \leq q-1$ und $i \leq s_r < j$. Wir verwenden die Bezeichnungen aus dem Beweis von Lemma 2.1.1. Permutiert und skaliert man K', so ergibt sich wie in (2.11) ein $\lambda_{s_l} \neq \lambda_{s_r}$ mit

$$\begin{aligned}
\frac{|f'^{1/2}_{ii} c'^{1/2}_{ii} - \lambda_{s_l}|}{\lambda_{s_l}} &\geq \frac{|\lambda_{s_r} - \lambda_{s_l}| - |f'^{1/2}_{ii} c'^{1/2}_{ii} - \lambda_{s_r}|}{\max\{\lambda_{s_r}, \lambda_{s_l}\}} \\
&\geq \frac{|\lambda_{s_r} - \lambda_{s_l}|}{\max\{\lambda_{s_r}, \lambda_{s_l}\}} - \frac{|f'^{1/2}_{ii} c'^{1/2}_{ii} - \lambda_{s_r}|}{\lambda_{s_r}} \\
&\geq \frac{\gamma\varepsilon}{\gamma} - \frac{\varepsilon}{\gamma} = \frac{(\gamma-1)\varepsilon}{\gamma} \ .
\end{aligned}$$

Wegen $f'^{1/2}_{ii} c'^{1/2}_{ii} > f'^{1/2}_{jj} c'^{1/2}_{jj}$ folgt daraus mit (2.9), (2.10)

$$\begin{aligned}
\frac{f'^{1/2}_{ii} c'^{1/2}_{ii}}{f'^{1/2}_{jj} c'^{1/2}_{jj}} - 1 &\geq \frac{f'^{1/2}_{ii} c'^{1/2}_{ii} - f'^{1/2}_{jj} c'^{1/2}_{jj}}{f'^{1/2}_{jj} c'^{1/2}_{jj}} \\
&\geq \left(\frac{|f'^{1/2}_{ii} c'^{1/2}_{ii} - \lambda_{s_l}|}{\lambda_{s_l}} - \frac{|f'^{1/2}_{jj} c'^{1/2}_{jj} - \lambda_{s_l}|}{\lambda_{s_l}}\right)\frac{\lambda_{s_l}}{f'^{1/2}_{jj} c'^{1/2}_{jj}} \\
&> \left(\frac{(\gamma-1)\varepsilon}{\gamma} - \frac{\varepsilon}{\gamma}\right)\frac{\gamma}{\gamma+\varepsilon} = \frac{(\gamma-2)\varepsilon}{\gamma+\varepsilon} \geq (\gamma-2)\,\|K'_S - I\|_E \ .
\end{aligned}$$

Q.E.D.

Satz 2.1.7 *Sei $\gamma \geq 30$. Seien die Voraussetzungen von Lemma 2.1.1 für dieses γ erfüllt und sei $1 \leq r \leq q$. Sei $\omega = \|K_S - I\|_E$, $\omega_r = \|K_{S_r} - I\|_E$ und*

$$\omega^2 = \frac{\varepsilon}{\gamma+\varepsilon}\,\omega_r \ .$$

Dann gilt

$$\frac{f_{ii}}{f_{jj}} \leq 1 + \frac{3\,\omega_r}{2\,\gamma} \quad , \quad s_{r-1} + 1 \leq i \leq j \leq s_r \tag{2.14}$$

$$Tr\,(F_{S_{rr}} - I)(C_{S_{rr}} - I) - \frac{1}{4}\,\|B_{S_{rr}} + B_{S_{rr}}^T\|_E^2 + 12\,\omega_r^3 \le -\frac{\omega_r^2}{10} \tag{2.15}$$

$$\frac{f_{ii}}{f_{jj}} > 1 + (\gamma - 2)\,\omega_r \quad , \quad i \le s_{r-1} < j \quad oder \quad i \le s_r < j \tag{2.16}$$

Beweis:
Zum Beweis von (2.14) wenden wir 2.1.5 an. Durch Ausrechnen erhält man nämlich

$$\frac{f_{ii}}{f_{jj}} \le \frac{1}{1 - 2\varrho} \le \frac{1}{1 - 2\omega_r/(\sqrt{2}(\gamma - 2))} \le 1 + \frac{1.5\,\omega_r}{\gamma} \ .$$

Die Ungleichung (2.15) erhält man aus

$$\begin{aligned}
&\mathrm{Tr}\,(F_{Srr} - I)(C_{Srr} - I) - \frac{1}{4}\|B_{Srr} + B_{Srr}^T\|_E^2 + 12\omega_r^3 \\
&= \frac{1}{2}(\|F_{Srr} + C_{Srr} - 2\,I\|_E^2 - \|F_{Srr} - I\|_E^2 - \|C_{Srr} - I\|_E^2) \\
&\quad -\|B_{Srr}\|_E^2 + \frac{1}{4}\|B_{Srr} - B_{Srr}^T\|_E^2 + 12\omega_r^3 \\
&= \frac{1}{2}(\|F_{Srr} + C_{Srr} - 2\,I\|_E^2 + \|B_{Srr} - B_{Srr}^T\|_E^2) - \frac{1}{4}\|B_{Srr} - B_{Srr}^T\|_E^2 \\
&\quad -\frac{1}{2}\|F_{Srr} - I\|_E^2 - \frac{1}{2}\|C_{Srr} - I\|_E^2 - \|B_{Srr}\|_E^2 + 12\omega_r^3 \\
&\le \frac{1}{2}\,\frac{(\gamma + \varepsilon)^2\,\omega^4}{(\gamma - 2)^2\varepsilon^2} - \frac{1}{2}\,\omega_r^2 + 12\,\omega_r^3 \le \frac{w_r^2}{2}\,(\frac{1}{(\gamma - 2)^2} - 1 + 24\,\omega_r) \\
&\le \frac{\omega_r^2}{2}\,(\frac{1}{(\gamma - 2)^2} - 1 + \frac{24}{\gamma + 1}) \le -0.11\,\omega_r^2 \ .
\end{aligned}$$

Schließlich folgt (2.16) sofort aus 2.1.6. Q.E.D.

Die Relationen (2.14), (2.15), (2.16) werden wir zur Formulierung numerischer Verfahren noch vereinfachen.

Wir führen nun Transformationen von K ein, die sich an der Struktur von K_S orientieren. Dazu betrachten wir wieder die Matrix K gemäß (2.2), (2.3) mit Eigenwerten gemäß (2.5), die wie in (2.6) partitioniert sei. R_r sei der Faktor der Cholesky-Zerlegung

$$R_r^T R_r = \frac{1}{\delta_r^2} F_{rr} \quad , \quad \delta_r = f_{t_{r-1}+1, t_{r-1}+1} \ , \quad r = 1 \ldots q \ ,$$

(δ_r wurde nur zur Abkürzung definiert) und es sei

$$K' = \begin{bmatrix} F' & B'^T \\ B' & C' \end{bmatrix} = T^T K T \quad , \quad T = \begin{bmatrix} R^{-1} & 0 \\ 0 & R^T \end{bmatrix} \ , \quad R = \bigoplus_{r=1}^{q} R_r \ . \tag{2.17}$$

K' und seine Submatrizen F', B', C' seien genauso wie K bzw. F, B, C partitioniert. Zur Definition einer zweiten Transformation sei

$$X_r = \frac{-1}{2\delta_r^2}\,(B'_{rr} + B'^T_{rr}) \quad , \quad r = 1 \dots q \ ,$$

und

$$K'' = \begin{bmatrix} F'' & B''^T \\ B'' & C'' \end{bmatrix} = T'^T K T' \quad , \quad T' = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \quad , \quad X = \bigoplus_{r=1}^{q} X_r \ . \tag{2.18}$$

Man prüft leicht $T^T J T = J$ und $T'^T J T' = J$ nach, T und T' sind also symplektisch. Die Matrizen K'', F'', B'', C'' seien ebenfalls wie K, F, B, C partitioniert. Dann läßt sich auch

$$F''_{rr} = F'_{rr} = \delta_r^2 I_{m_r} \quad , \quad B''_{rr} = \frac{1}{2}\,(B'_{rr} - B'^T_{rr}) \ , \quad r = 1 \dots q$$

leicht nachrechnen. Wir definieren noch

$$\Delta = \bigoplus_{r=1}^{q} \Delta_r \quad , \quad \Delta_r = (F'_{rr})^{1/2} = \delta_r I_{s_r}$$

und halten die Form der Matrizen fest, wie wir sie später verwenden wollen:

$$\begin{aligned} K &= \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} K_S \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} \quad , \quad D = \mathrm{diag}(f_{ii}^{1/2}) \\ K' &= \begin{bmatrix} \Delta & 0 \\ 0 & \Delta D' \end{bmatrix} K'_S \begin{bmatrix} \Delta & 0 \\ 0 & \Delta D' \end{bmatrix} \quad , \quad D' = \Delta^{-1}\mathrm{diag}(c_{ii}'^{1/2}) \\ K'' &= \begin{bmatrix} \Delta & 0 \\ 0 & \Delta D'' \end{bmatrix} K''_S \begin{bmatrix} \Delta & 0 \\ 0 & \Delta D'' \end{bmatrix} \quad , \quad D'' = \Delta^{-1}\mathrm{diag}(c_{ii}''^{1/2}) \ . \end{aligned}$$

Darüber hinaus setzen wir $\tilde{D} = \Delta^{-1} D$, und für die skalierten Matrizen schreiben wir auch

$$K_S = \begin{bmatrix} F_S & B_S^T \\ B_S & C_S \end{bmatrix} \quad , \quad K'_S = \begin{bmatrix} F'_S & B'^T_S \\ B'_S & C'_S \end{bmatrix} \quad , \quad K''_S = \begin{bmatrix} F''_S & B''^T_S \\ B''_S & C''_S \end{bmatrix} \ .$$

Die diagonalen Matrizen $\tilde{D}$, D, D', D'' und Δ seien genauso wie die anderen Matrizen partitioniert. Dann ist das $(1,1)$-Element jedes Blockes $\tilde{D}_{rr}$ gleich 1. Wir übernehmen dieselbe Partitionierung auch für K_S, K'_S und K''_S.

Liegen für K und K'' die Voraussetzungen des Satzes 2.1.4 vor, so haben wir mit einem geeigneten γ''

$$(\sum_{r=1}^{m} \|C''_{Srr} - I_{n_r}\|_E^2 + 4\|B''_{Srr}\|_E^2)^{1/2} \le \frac{(\gamma'' + \varepsilon)}{(\gamma'' - 2)\varepsilon}\ \|K''_S - I\|_E^2 \ . \tag{2.19}$$

War also $\omega = \|K_S - I\|_E$ vor dem Transformationspaar (2.17), (2.18) klein genug und ist auch $\|K_S'' - I\|_E = \mathcal{O}(\omega)$, so sind die Normen $\|K_{S_r}'' - I\|_E$ nach dem Transformationspaar von der Ordnung $\mathcal{O}(\omega^2)$. Dies werden wir mit Satz 2.1.8 präzisieren.

Satz 2.1.8 *Seien die Voraussetzungen von Lemma 2.1.1 erfüllt und sei $\gamma \geq 30$. Auf K werde das Transformationspaar (2.17), (2.18) angewendet. Dann gilt*

$$\|K_S' - I\|_E \leq \frac{\varepsilon}{\gamma' + \varepsilon} \quad \textit{für alle} \quad \gamma' \quad \textit{mit} \quad 1 \leq \gamma' \leq 0.25\,\gamma - 0.75\ . \tag{2.20}$$

K' erfüllt die Voraussetzungen von Korollar 2.1.3. Außerdem gilt

$$\|K_S'' - I\|_E \leq \frac{\varepsilon}{\gamma'' + \varepsilon} \quad \textit{für alle} \quad \gamma'' \quad \textit{mit} \quad 1 \leq \gamma'' \leq 0.08\,\gamma - 0.92\ . \tag{2.21}$$

Auch K'' erfüllt die Voraussetzungen von Korollar 2.1.3.

Beweis:
Der Beweis ist langwierig und erfolgt durch Ausrechnen. Wir verwenden, falls erforderlich, für alle auftretenden Matrizen dieselbe Partitionierung. Zur Abkürzung verwenden wir $\omega = \|K_S - I\|_E$, und wir bemerken vorab

$$\omega \leq \frac{\varepsilon}{\gamma + \varepsilon} \leq \frac{1}{31} \quad , \quad \|K_S^{-1}\|_2 \leq \frac{1}{1 - \|K_S - I\|_2} \leq \frac{31}{30}\ . \tag{2.22}$$

Wir verwenden die Bezeichnung $\|X\|_{E,\text{off}} = \|X - \text{diag}(X)\|_E$ (wobei $\text{diag}(X) = \text{diag}(x_{ii})$) und bemerken für allgemeine X, Y und diagonale Z noch

$$\|XY\|_{E,\text{off}} \leq \|XY\|_E \leq \|X\|_2 \|Y\|_E \quad , \quad \|ZX\|_{E,\text{off}} \leq \|Z\|_2 \|X\|_{E,\text{off}}\ .$$

Für $L_S = \tilde{D}^{-1} L$ ist $L_S L_S^T = \bigoplus_{r=1}^q F_{Srr}$ und nach [17] gilt

$$L_S = I + S \quad , \quad \|S\|_E \leq \frac{\sqrt{2}}{1 + \sqrt{1 - 2\,\omega}}\,\omega \leq 0.72\,\omega\ . \tag{2.23}$$

Außerdem gibt es ein M mit

$$L_S^{-1} = I - S + M \quad , \quad \|M\|_E \leq \|F_S^{-1}\|_2^{1/2} \|S\|_E^2 \leq 0.53\,\omega^2\ .$$

Δ kommutiert mit L und es gilt

$$\begin{aligned} K_S' &= \begin{bmatrix} \Delta^{-1} & 0 \\ 0 & \Delta^{-1} D'^{-1} \end{bmatrix} \begin{bmatrix} L^{-1} F L^{-T} & L^{-1} B^T L \\ L^T B L^{-T} & L^T C L \end{bmatrix} \begin{bmatrix} \Delta^{-1} & 0 \\ 0 & \Delta^{-1} D'^{-1} \end{bmatrix} \\ &= \begin{bmatrix} L_S^{-1} F_S L_S^{-T} & L_S^{-1} B_S^T \tilde{D}^2 L_S D'^{-1} \\ D'^{-1} L_S^T \tilde{D}^2 B_S L_S^{-T} & D'^{-1} L_S^T \tilde{D}^2 C_S \tilde{D}^2 L_S D'^{-1} \end{bmatrix}\ . \end{aligned}$$

Es gilt somit

$$\begin{aligned}\|K'_S - I\|_E^2 &= \|L_S^{-1} F_S L_S^{-T}\|_{E,\text{off}}^2 + 2\,\|D'^{-1} L_S^T \tilde{D}^2 B_S L_S^{-T}\|_E^2 \\ &\quad + \|D'^{-1} L_S^T \tilde{D}^2 C_S \tilde{D}^2 L_S D'^{-1}\|_{E,\text{off}}^2 \,.\end{aligned}$$

Für jeden der vorstehenden Summanden werden wir eine obere Schranke bestimmen. Als erstes haben wir

$$\begin{aligned}\|F'_S\|_{E,\text{off}} &= \|L_S^{-1} F_S L_S^{-T}\|_{E,\text{off}} \\ &= \|(I - S + M) F_S (I - S^T + M^T)\|_{E,\text{off}} \\ &\le \|F_S\|_{E,\text{off}} + 2\,\|F_S\|_2\,\|M - S\|_E + \|F_S\|_2\,\|M - S\|_E^2 \\ &\le \omega + (1+\omega)(2\,\omega\,(0.72 + 0.53\,\omega) + \omega^2\,(0.72 + 0.53\,\omega)^2) \\ &\le 2.54\,\omega \,. \end{aligned} \tag{2.24}$$

Nun schätzen wir den zweiten Summanden ab. Wegen des Cauchyschen Zwischenwertsatzes gilt $\lambda_{\min}(D'^2) \ge \lambda_{\min}(\Delta^{-1} C' \Delta^{-1})$ und daher unter Verwendung von $D^{-1} = \Delta^{-1} \tilde{D}^{-1}$

$$\begin{aligned}\|D'^{-1}\|_2 &\le \|(\Delta^{-1} C' \Delta^{-1})^{-1}\|_2^{1/2} = \|\Delta L_S^{-1} \tilde{D}^{-1} D^{-1} C_S^{-1} D^{-1} \tilde{D}^{-1} L_S^{-T} \Delta\|_2^{1/2} \\ &\le \|\tilde{D}^{-2}\|_2 \|F_S^{-1}\|_2^{1/2} \|C_S^{-1}\|_2^{1/2} \le \|\tilde{D}^{-2}\|_2 \|K_S^{-1}\|_2 \,,\end{aligned}$$

wobei wegen 2.1.5 für $\gamma \ge 30$

$$\|\tilde{D}^{-2}\|_2 \le \frac{1}{1 - \omega^2 \sqrt{2}(\gamma + \varepsilon)/((\gamma - 2)\varepsilon)} \le 1.01$$

gilt. Damit erhalten wir wegen $\|\tilde{D}\| \le 1$ und (2.22)

$$\begin{aligned}\|D'^{-1} L_S^T \tilde{D}^2 B_S L_S^{-T}\|_E &\le \|\tilde{D}^{-2}\|_2 \|K_S^{-1}\|_2 \|L_S\|_2 \|\tilde{D}^2\|_2 \|B_S\|_E \|L_S^{-1}\|_2 \\ &\le 1.01\,\frac{31}{30}\,(1 + \frac{0.72}{31})\,\frac{\omega}{\sqrt{2}}\,(1 + \frac{0.72}{31} + \frac{0.53}{31^2}) \\ &\le 0.78\,\omega \,.\end{aligned}$$

Schließlich bestimmen wir eine obere Schranke für den dritten Summanden. Es gilt

$$\begin{aligned}&\|D'^{-1} L_S^T \tilde{D}^2 C_S \tilde{D}^2 L_S D'^{-1}\|_{E,\text{off}} \\ &\quad\le \|D'^{-1} L_S^T \tilde{D}^2 (C_S - I) \tilde{D}^2 L_S D'^{-1}\|_E + \|D'^{-1} L_S^T \tilde{D}^4 L_S D'^{-1}\|_{E,\text{off}} \\ &\quad\le \|D'^{-2}\|_2 (\|F_S\|_2 \|\tilde{D}^4\|_2 \|C_S - I\|_E + \|L_S^T \tilde{D}^4 L_S\|_{E,\text{off}})\end{aligned}$$

und wegen

$$\begin{aligned}\|L_S^T \tilde{D}^4 L_S\|_{E,\text{off}} &= \|(I + S^T) \tilde{D}^4 (I + S)\|_{E,\text{off}} \le \|\tilde{D}^4\|_2 (2\|S\|_E + \|S\|_E^2) \\ &\le \omega\,(2 \cdot 0.72 + 0.72^2\,\omega) \le 1.46\,\omega\end{aligned}$$

ist mit (2.22

$$\begin{aligned}\|D'^{-1}L_S^T\tilde{D}^2C_S\tilde{D}^2L_SD'^{-1}\|_{E,\text{off}} &\leq 1.01^2\,\frac{31^2}{30^2}\cdot((1+\omega)\,\omega+1.46\,\omega)\\ &\leq 2.72\,\omega\ .\end{aligned}$$

Insgesamt haben wir also

$$\begin{aligned}\omega' = \|K'_S - I\|_E &\leq (2.54^2+2\cdot 0.78^2+2.72^2)^{1/2}\,\omega\\ &\leq 3.89\,\omega\ .\end{aligned}$$

Hinreichende Bedingung für $\omega' \leq \varepsilon/(\gamma'+\varepsilon)$ ist $\gamma' \leq 0.25\,\gamma - 0.75$. Damit ist (2.20) gezeigt. Für den ersten Teil des Satzes verbleibt zu beweisen, daß K' den Voraussetzungen von Korollar 2.1.3 genügt. Dabei verwenden wir für $X = (x_{ij})$ die Notation $\min(\text{diag}(X)) = \min_i(x_{ii})$ und $\max(\text{diag}(X)) = \max_i(x_{ii})$. Es reicht aus, für $r = 1\ldots q-1$

$$\max(\text{diag}(F'_{r+1,r+1})\cdot\text{diag}(C'_{r+1,r+1})) \leq \min(\text{diag}(F'_{rr})\cdot\text{diag}(C'_{rr})) \tag{2.25}$$

zu zeigen, weil dann die in Korollar 2.1.3 erwähnten Skalierungen und Permutationen existieren. Sei r, $1 \leq r \leq q$, fixiert. Wegen (2.23) ist $F_{Srr} = (I+S_{rr})(I+S_{rr}^T) = I + S_{rr} + S_{rr}^T + S_{rr}S_{rr}^T$ und deshalb

$$\|\text{diag}(S_{rr})\|_2 = \|-\frac{1}{2}\text{diag}(S_{rr}S_{rr}^T)\|_2 \leq \frac{1}{2}\|S_{rr}\|_2^2 \leq 0.26\,\omega^2\ .$$

Es gibt daher ein R_r mit

$$\begin{aligned}\text{diag}(F'_{rr})\cdot\text{diag}(C'_{rr}) &= \Delta_r^2\text{diag}(L_r^TC_{rr}L_r)\\ &= \delta_r^2\text{diag}((I+S_{rr}^T)\tilde{D}_{rr}C_{rr}\tilde{D}_{rr}(I+S_{rr}))\\ &= \delta_r^2(\text{diag}(C_{rr})+\delta_r^2\text{diag}(R_r))\ ,\end{aligned} \tag{2.26}$$

wobei wir $\|\text{diag}(R_r)\|_2$ noch abschätzen werden. Wegen 2.1.5 ist

$$\begin{aligned}&\|\text{diag}(R_r)\|_2\\ &\leq \|\delta_r^{-2}\text{diag}(-C_{rr}+\tilde{D}_{rr}C_{rr}\tilde{D}_{rr}+2\tilde{D}_{rr}C_{rr}\tilde{D}_{rr}S_{rr}+S_{rr}^T\tilde{D}_{rr}C_{rr}\tilde{D}_{rr}S_{rr})\|_2\\ &\leq \|\text{diag}(\tilde{D}_{rr}^2C_{Srr}\tilde{D}_{rr}^2-\tilde{D}_{rr}C_{Srr}\tilde{D}_{rr})\|_2\\ &\quad+\|\text{diag}(2\tilde{D}_{rr}^2(C_{Srr}-I+I)\tilde{D}_{rr}^2S_{rr}+S_{rr}^T\tilde{D}_{rr}^2(C_{Srr}-I+I)\tilde{D}_{rr}^2S_{rr})\|_2\\ &\leq \|\tilde{D}_{rr}^4-\tilde{D}_{rr}^2\|_2+2\,\|\text{diag}(\tilde{D}_{rr}^2(C_{Srr}-I)\tilde{D}_{rr}^2S_{rr})\|_2+2\,\|\text{diag}(\tilde{D}_{rr}^4S_{rr})\|_2\\ &\quad+\|\text{diag}(S_{rr}^T\tilde{D}_{rr}^2(C_{Srr}-I)\tilde{D}_{rr}^2S_{rr})\|_2+\|\text{diag}(S_{rr}^T\tilde{D}_{rr}^4S_{rr})\|_2\\ &\leq 2\,\varrho+2\,\omega\cdot 0.72\,\omega+2\cdot 0.26\,\omega^2+0.72^2\,\omega^3+0.72^2\,\omega^2\\ &\leq 2\,\varrho+2.50\,\omega^2\ .\end{aligned} \tag{2.27}$$

Mit Hilfe von 2.1.5, 2.1.6 schließen wir aus (2.26), (2.27)

$$\begin{aligned}
&\min(\mathrm{diag}(F'_{rr})\cdot\mathrm{diag}(C'_{rr})) - \max(\mathrm{diag}(F'_{r+1,r+1})\cdot\mathrm{diag}(C'_{r+1,r+1}))\\
&\quad\geq\ \delta_r^2(\delta_r^2(1-2\,\varrho)-\delta_r^2(2\,\varrho+2.50\,\omega^2))-\delta_{r+1}^2(\delta_{r+1}^2+\delta_{r+1}^2(2\,\varrho+2.50\,\omega^2))\\
&\quad\geq\ \delta_{r+1}^4(\frac{\delta_r^4}{\delta_{r+1}^4}(1-4\,\varrho-2.50\,\omega^2)-(1+2\,\varrho+2.50\,\omega^2))\\
&\quad\geq\ \delta_{r+1}^4((1+(\gamma-2)\,\omega)(1-\frac{2\sqrt{2}\,\omega}{\gamma-2}-2.50\,\omega^2)-(1+\frac{\sqrt{2}\,\omega}{\gamma-2}+2.50\,\omega^2))\\
&\quad\geq\ \omega\,\delta_{r+1}^4(\gamma-(2+\frac{3\sqrt{2}}{\gamma-2})-\omega(5.0+2\sqrt{2})-\omega^2\cdot 2.50\,(\gamma-2))\\
&\quad>\ 0\ .
\end{aligned}$$

Also ist (2.25) erfüllt und K' genügt den Voraussetzungen von Korollar 2.1.3. Wir wenden uns nun dem zweiten Teil des Satzes, also den Aussagen über K'', K''_S, zu. Es gilt

$$K''_S = \begin{bmatrix}\Delta^{-1} & 0\\ 0 & (\Delta D'')^{-1}\end{bmatrix}\begin{bmatrix}F' & B'^T+F'X\\ XF'+B' & XF'X+B'X+XB'^T+C'\end{bmatrix}\begin{bmatrix}\Delta^{-1} & 0\\ 0 & (\Delta D'')^{-1}\end{bmatrix}$$

mit X gemäß (2.18) und wegen der Kommutativität von Δ und X

$$\begin{aligned}
\|K''_S-I\|_E^2 \leq\ & \|F'_S\|_{E,\mathrm{off}}^2+2\|D''^{-1}(XF'_S+D'B'_S)\|_E^2\\
&+\|D''^{-1}(XF'_SX+XB'^T_SD'+D'B'_SX+D'C'_SD')D''^{-1}\|_{E,\mathrm{off}}^2\ .
\end{aligned}$$

Die Summanden werden wir wieder einzeln abschätzen. Als erstes haben wir $\|F''_S\|_{E,\mathrm{off}} = \|F'_S\|_{E,\mathrm{off}} \leq \omega'$. Für die anderen beiden Summanden müssen wir jedoch weiter ausholen. Es gilt

$$\begin{aligned}
\|D'^2\|_2 &= \|\mathrm{diag}(\Delta^{-1}C'\Delta^{-1})\|_2 = \|\mathrm{diag}(\Delta^{-1}L^TCL\Delta^{-1})\|_2\\
&= \|\mathrm{diag}(L_S^T\tilde{D}^2C_S\tilde{D}^2L_S)\|_2\\
&\leq \|\mathrm{diag}(L_S^T\tilde{D}^2(C_S-I)\tilde{D}^2L_S)\|_2+\|\mathrm{diag}(L_S^T\tilde{D}^4L_S)\|_2\\
&\leq \|F_S\|_2\|\tilde{D}^4\|_2\|C_S-I\|_2+\|F_S\|_2\\
&\leq (1+\omega)^2\leq 1.07\ .
\end{aligned}$$

Daraus folgt

$$\|X\|_E\leq\|D'\|_2\|B'_S\|_E\leq\sqrt{1.07}\,\frac{\omega'}{\sqrt{2}}\leq 0.74\,\omega'\ .$$

Es gilt außerdem

$$\begin{aligned}
\|D'^2-I\|_2 &= \|\mathrm{diag}(L_S^T\tilde{D}^2C_S\tilde{D}^2L_S)-I\|_2\\
&= \|\mathrm{diag}(L_S^T\tilde{D}^2(C_S-I)\tilde{D}^2L_S)+\mathrm{diag}(L_S^T\tilde{D}^4L_S)-I\|_2\\
&\leq \|F_S\|_2\|C_S-I\|_2+\|\mathrm{diag}(L_S^T(\tilde{D}^4-I)L_S)+\mathrm{diag}(L_S^TL_S)-I\|_2\\
&\leq \|F_S\|_2\|C_S-I\|_2+\|F_S\|_2\|\tilde{D}^2+I\|_2\|\tilde{D}^2-I\|_2\\
&\leq (1+\omega)(\omega+4\,\varrho)\ .
\end{aligned}$$

Aufgrund des Cauchyschen Zwischenwertsatzes ist weiterhin

$$\begin{aligned}\|D''^{-2}\|_2 &\leq \|(\mathrm{diag}(\Delta^{-1}(\bigoplus_{r=1}^{q} C''_{rr})\Delta^{-1}))^{-1}\|_2 \\ &\leq (1 - \|\mathrm{diag}(\Delta^{-1}(\bigoplus_{r=1}^{q} C''_{rr})\Delta^{-1}) - I\|_2)^{-1} \ .\end{aligned}$$

Dabei gilt wegen der Kommutativität von Δ und X und 2.1.5

$$\begin{aligned}&\|\mathrm{diag}(\Delta^{-1}(\bigoplus_{r=1}^{q} C''_{rr})\Delta^{-1}) - I\|_2 \\ &\leq \|X \bigoplus_{r=1}^{q}(F'_{S_{rr}})X\|_2 + 2\,\|\bigoplus_{r=1}^{q}(D'_{rr}B'_{S_{rr}})X\|_2 + \|D'^2 - I\|_2 \\ &\leq \|X\|_2^2 + 2\,\|X\|_2 \cdot 0.74\,\omega' + (1+\omega)(\omega + 4\,\varrho) \\ &\leq 0.069 \ ,\end{aligned}$$

also

$$\|D''^{-2}\|_2 \leq \frac{1}{1 - 0.069} \leq 1.08 \ .$$

Nun können wir endlich die beiden Abschätzungen

$$\begin{aligned}\|D''^{-1}(XF'_S + D'B'_S)\|_E &\leq \|D''^{-1}\|_2(\|X\|_E\|F'_S\|_2 + \|D'\|_2\|B'_S\|_E) \\ &\leq \sqrt{1.08}(0.74\,\omega'(1+\omega') + \sqrt{1.07}\frac{\omega'}{\sqrt{2}}) \\ &\leq 1.63\,\omega'\end{aligned}$$

und

$$\begin{aligned}&\|D''^{-1}(XF'_SX + XB'^T_SD' + D'B'_SX + D'C'_SD')D''^{-1}\|_{E,\mathrm{off}} \\ &\leq \|D''^{-2}\|_2(\|X\|_E^2\|F'_S\|_2 + 2\,\|D'\|_2\|X\|_E\|B'_S\|_E + \|D'^2\|_2\|C_S - I\|_E) \\ &\leq 1.08\,\omega'\,(0.74^2\,\omega'(1+\omega') + 2\sqrt{1.07}\cdot 0.74\,\frac{\omega'}{\sqrt{2}} + 1.07) \\ &\leq 1.39\,\omega'\end{aligned}$$

vornehmen, woraus wir zusammenfassend

$$\omega'' = \|K''_S - I\|_E \leq \omega'(1 + 2\cdot 1.63^2 + 1.39^2)^{1/2} \leq 2.88\,\omega'$$

erhalten. Hinreichende Bedingung für $\omega'' \leq \varepsilon/(\gamma''+\varepsilon)$ ist $\gamma'' \leq 0.08\,\gamma - 0.92$. Damit ist auch (2.21) bewiesen und es verbleibt zu zeigen, daß K'' den Voraussetzungen von Korollar 2.1.3 genügt. Dazu reicht es aus, für $1 \leq r \leq q-1$

$$\max \mathrm{diag}(C''_{r+1,r+1}) \leq \min \mathrm{diag}(C''_{rr}) \tag{2.28}$$

zu zeigen. Es gilt

$$\begin{aligned}
\operatorname{diag}(C''_{rr}) &= \operatorname{diag}(X_{rr}F'_{rr}X_{rr} + B'_{rr}X_{rr} + X_{rr}B'^T_{rr} + C'_{rr}) \\
&= \operatorname{diag}(X_{rr}F'_{rr}X_{rr}) + 2\operatorname{diag}(B'_{rr}X_{rr}) + \operatorname{diag}(C'_{rr}) \\
&= \operatorname{diag}(C'_{rr}) + \delta_r^2 \operatorname{diag}(R'_r)\ ,
\end{aligned}$$

wobei wegen $F'_{rr} = \delta_r^2 I_{n_r}$

$$\begin{aligned}
\|\operatorname{diag} R'_r\|_2 &\le \|\delta_r^{-2}\operatorname{diag}(X_{rr}F'_{rr}X_{rr})\|_2 + 2\,\|\delta_r^{-2}\operatorname{diag}(B'_{rr}X_{rr})\|_2 \\
&\le \|X_{rr}\|_E^2 + 2\,\|B'_{S_{rr}}\|_E \|D'_{rr}\|_2 \|X_{rr}\|_E \\
&\le 0.74^2\,\omega'^2 + 2\,\frac{\omega'}{\sqrt{2}}\sqrt{1.07}\cdot 0.74\,\omega' \le 1.64\,\omega'^2
\end{aligned}$$

gilt. Es folgt für $1 \le r \le q-1$

$$\begin{aligned}
&\min\operatorname{diag}(C''_{rr}) - \max\operatorname{diag}(C''_{r+1,r+1}) \\
&\ge \min\operatorname{diag}(C_{rr}) - \delta_r^2(2\,\varrho + 2.50\,\omega^2 + 1.64\,\omega'^2) \\
&\qquad - \max\operatorname{diag}(C_{i+1,i+1}) - \delta_{r+1}^2(2\,\varrho + 2.50\,\omega^2 + 1.64\,\omega'^2) \\
&\ge \delta_{r+1}^2\Big(\frac{\delta_r^2}{\delta_{r+1}^2}\,(1 - 2\,\varrho - 2\,\varrho - 2.50\,\omega^2 - 1.64\,\omega'^2) \\
&\qquad -(1 + 2\,\varrho + 2.50\,\omega^2 + 1.64\,\omega'^2)\Big) \\
&\ge \delta_{r+1}^2\Big((1 + (\gamma-2)\omega)(1 - 4\,\frac{\omega}{\sqrt{2}(\gamma-2)} - 2.50\,\omega^2 - 1.64\cdot 3.89^2\,\omega^2) \\
&\qquad -(1 + 2\,\frac{\omega}{\sqrt{2}(\gamma-2)} + 2.50\,\omega^2 + 1.64\cdot 3.89^2\,\omega^2)\Big) \\
&\ge \omega\,\delta_{r+1}^2\,\Big(\frac{-4}{\sqrt{2}(\gamma-2)} + (\gamma-2) - \frac{\sqrt{2}}{\gamma-2} \\
&\qquad -\omega\,(2\cdot 2.50 + 2\cdot 1.64\cdot 3.89^2 + 2\sqrt{2}) - \omega^2\,(\gamma-2)(2.50 + 1.64\cdot 3.89^2)\Big) \\
&> 0\ ,
\end{aligned}$$

womit schließlich alles gezeigt ist. Q.E.D.

Wir schließen diesen Abschnitt mit einigen Aussagen für den Fall, daß (K, J) nur ein einziges Eigenwertpaar besitzt, ab. In diesem Fall ist K nämlich bis auf einen Faktor symplektisch. Genauer gilt

Satz 2.1.9 *Hat (K, J) ausschließlich das Eigenwertpaar $\pm i\lambda$, $\lambda \in \mathbb{R}$, $\lambda \neq 0$, so ist $(1/\lambda)K$ symplektisch und es gibt ein symmetrisches X sowie ein nichtsinguläres R mit*

$$\begin{aligned}
K &= \lambda \begin{bmatrix} R^T & 0 \\ 0 & R^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & R^{-T} \end{bmatrix} \\
&= \lambda \begin{bmatrix} R^T R & R^T X R^{-T} \\ T^{-1} X R & R^{-1}(I + X^2)R^{-T} \end{bmatrix}
\end{aligned}$$

Beweis:
Es gibt ein symplektisches T mit $T^T K T = \lambda I$, also ist $\tilde{K} = (1/\lambda)K = T^{-T}T^{-1}$ symplektisch. Für

$$\tilde{K} = \begin{bmatrix} \tilde{F} & \tilde{B}^T \\ \tilde{B} & \tilde{C} \end{bmatrix} \quad , \quad \tilde{F} = \tilde{R}^T \tilde{R} \quad , \quad S = \begin{bmatrix} \tilde{R}^{-1} & 0 \\ 0 & \tilde{R}^T \end{bmatrix}$$

gilt

$$\tilde{K}' = \begin{bmatrix} I & \tilde{B}'^T \\ \tilde{B}' & \tilde{C} \end{bmatrix} = S^T \tilde{K} S, \tag{2.29}$$

wobei S und $\tilde{K}'$ symplektisch sind. Es folgt

$$J = \tilde{K}'^T J \tilde{K}' = \begin{bmatrix} \tilde{B}' - \tilde{B}'^T & \tilde{C} - (\tilde{B}'^T)^2 \\ \tilde{B}'^2 - \tilde{C}' & \tilde{B}'\tilde{C}' - \tilde{C}'\tilde{B}'^T \end{bmatrix}$$

und daher $\tilde{B}' = \tilde{B}'^T$ und $I + \tilde{B}'^2 = \tilde{C}'$, also

$$\tilde{K}' = \begin{bmatrix} I & \tilde{B}' \\ \tilde{B}' & I + \tilde{B}'^2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ \tilde{B}' & I \end{bmatrix} \begin{bmatrix} I & \tilde{B}' \\ 0 & I \end{bmatrix} \tag{2.30}$$

Aus (2.29), (2.30) folgt schließlich die Aussage des Satzes. Q.E.D.

Mit Hilfe von (2.19) kann man sich leicht plausibel machen, daß die r-Blöcke eines Paares (K, J) mit mehrfachen Eigenwerten bis auf einen kleinen Fehler die Eigenschaft des letzten Satzes haben. Die Matrizen des Transformationspaares (2.17), (2.18) haben nämlich die Struktur der Matrizen des letzten Satzes.

Im übrigen haben wir für Matrizen der Ordnung 4

Lemma 2.1.10 *Das Paar (K, J) mit $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \in \mathbb{R}^{4\times 4}$, $f_{ii} = c_{ii}$ und $b_{ii} = 0$, $i = 1, 2$, hat genau dann ausschließlich das Eigenwertpaar $\pm i\lambda$, $\lambda \in \mathbb{R}$, $\lambda \neq 0$, wenn gilt*

$$K = \begin{bmatrix} f & c & 0 & b \\ c & f & b & 0 \\ 0 & b & f & -c \\ b & 0 & -c & f \end{bmatrix} .$$

In diesem Falle ist $\lambda = \sqrt{f^2 - b^2 - c^2}$.

Beweis:
Zum Beweis der einen Richtung setzen wir zunächst voraus, daß (K, J) nur ein

Eigenwertpaar $\pm i\lambda$ habe. Wegen [26], (2.3.13) gilt dann

$$K = \begin{bmatrix} f & f_{12} & 0 & b_{21} \\ f_{12} & f & b_{12} & 0 \\ 0 & b_{12} & f & c_{12} \\ b_{21} & 0 & c_{12} & f \end{bmatrix}$$

mit

$$\begin{aligned} \det(JK - \lambda I) &= x^4 + 2x^2(f^2 - b_{12}b_{21} + f_{12}c_{12}) \\ &\quad + f^4 - f^2(b_{12}^2 + b_{21}^2 + f_{12}^2 + c_{12}^2) + (b_{12}b_{21} - f_{12}c_{12})^2 \ . \end{aligned}$$

Es gibt daher genau dann eine doppelte Nullstelle, wenn

$$(f^2 - b_{12}b_{21} + f_{12}c_{12})^2 = f^4 - f^2(b_{12}^2 + b_{21}^2 + f_{12}^2 + c_{12}^2) + (b_{12}b_{21} - f_{12}c_{12})^2$$

gilt. Dies formen wir äquivalent um:

$$\begin{aligned} & 2f^2(f_{12}c_{12} - b_{12}b_{21}) = -f^2(b_{12}^2 + b_{21}^2 + f_{12}^2 + c_{12}^2) \\ \iff & (f_{12} + c_{12})^2 + (b_{12} - b_{21})^2 = 0 \\ \iff & f_{12} = -c_{12} \wedge b_{12} = b_{21} \ . \end{aligned}$$

Die umgekehrte Richtung der Behauptung ist gezeigt, wenn man die Eigenwerte von (K, J) aus dem charakteristischen Polynom

$$x^4 + 2x^2(f^2 - b^2 - c^2) + (f^2 - b^2 - c^2)^2 = 0$$

von $-JK$ bestimmt. Die angegebenen λ ergeben sich gleichfalls aus vorstehendem Polynom. Q.E.D.

2.2 Jacobi-ähnliche Verfahren für $(\mathbf{L^T L}, \mathbf{J})$

In diesem Abschnitt werden zwei Jacobi-ähnliche Verfahren zur Lösung des Eigenwertproblems des Matrizenpaares (K, J) beschrieben. In ihrer Grundidee basieren sie auf einem in [31] beschriebenen Verfahren zur Lösung des Eigenwertproblems für definite Matrizenpaare.

Als erstes wird ein explizites Verfahren behandelt. Es besteht aus einer Folge symplektischer Konkruenztransformationen der symmetrischen positiv definiten Matrix K. Das Verfahren wird zunächst auf eine Weise beschrieben, die seine Struktur verdeutlichen soll. Die Details der Parameterbestimmungen spielen dabei nur eine untergeordnete Rolle. Auf die Bestimmung der Eigenvektoren von (K, J) wird nicht eingegangen. Der Beschreibung schließt sich der formale Algorithmus an.

Beim anschließend angegeben impliziten Verfahren wird nicht die Matrix K selbst, sondern ein Faktor L, $K = L^T L$, mit einer Folge symplektischer Matrizen von rechts multipliziert. Auch hier wird zuerst eine Beschreibung gegeben, der sich der formale Algorithmus anschließt. In exakter Arithmetik sind beide Verfahren äquivalent.

Zum besseren Verständnis wird bei der Beschreibung des expliziten Verfahrens auf Hinweise zu Laufzeit- und Speicheroptimierungen verzichtet. Im Unterabschnitt über das implizite Verfahren hingegen werden derartige Optimierungen berücksichtigt.

2.2.1 Explizites Verfahren

In diesem Unterabschnitt werden wir ein Verfahren zur Diagonalisierung von K in (2.1) entwickeln und im nachfolgenden Unterbschnitt den formalen Algorithmus angeben. Wir werden stets die Partitionierung

$$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \quad , \quad F, B, C \in \mathbb{R}^{n \times n} \ ,$$

für K und

$$K_S = \begin{bmatrix} F_S & B_S^T \\ B_S & C_S \end{bmatrix} \quad , \quad F_S, B_S, C_S \in \mathbb{R}^{n \times n}$$

für die auf 1-Diagonale skalierte Matrix verwenden.

Das Verfahren besteht aus einem endlichen vorbereitenden Teil (*Vorbereitung*) und einem iterativen Teil. In beiden Teilen wird K mit symplektischen Matrizen kongruent transformiert und die entstehende Folge von Transformationen wird K schließlich diagonalisieren.

2.2.1.1 Vorbereitung

Die Vorbereitung ist ein endlicher Prozeß. Sie wird zu Beginn des Verfahrens vorgenommen wird und dient dazu, K so zu transformieren, daß die Diagonalelemente von F und C übereinstimmen und die Diagonalelemente von B verschwinden. Dies wird nicht nur die Formulierung der anderen Teile des Algorithmus' erleichtern und die Beweisführung bei später nachfolgenden Sätzen vereinfachen, sondern hat auch numerische Vorteile, wie sich in Abschnitt 3.1 zeigen wird. Bei der Vorbereitung wird folgendermaßen verfahren:

K wird zuerst mit der symplektischen Matrix

$$V_1 = \begin{bmatrix} \operatorname{diag}(\sqrt[4]{C_{ii}/F_{ii}}) & 0 \\ 0 & \operatorname{diag}(\sqrt[4]{F_{ii}/C_{ii}}) \end{bmatrix} \tag{2.31}$$

transformiert (zum formellen Vorgehen siehe **1.1** von Verfahren 2.2.2). Der einfachen Darstellung halber werde dies mit der Notation $K := V_1^T K V_1$ beschrieben. Es gilt nun

$$F_{ii} = C_{ii} \quad , \quad i = 1 \dots n \tag{2.32}$$

und die Diagonalelemente von B können auf einfache Weise anulliert werden, indem K mit

$$V_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} I_n & I_n \\ -I_n & I_n \end{bmatrix} \tag{2.33}$$

transformiert wird: $K := V_2^T K V_2$ (**1.2** von Verfahren 2.2.2). Die Eigenschaft (2.32) wird dadurch jedoch zerstört und durch erneute Transformation mit

$$V_3 = \begin{bmatrix} \operatorname{diag}(\sqrt[4]{C_{ii}/F_{ii}}) & 0 \\ 0 & \operatorname{diag}(\sqrt[4]{F_{ii}/C_{ii}}) \end{bmatrix}$$

wiederhergestellt: $K := V_3^T K V_3$ (**1.3** von Verfahren 2.2.2). Die Vorbereitung ist damit beendet und es gilt

$$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \quad \text{mit} \quad F_{ii} = C_{ii} \ \text{ und } \ B_{ii} = 0 \quad , \quad i = 1 \dots n. \tag{2.34}$$

Das Verfahren ist so konzipiert, daß die Eigenschaft (2.34) während des sich nun anschließenden iterativen Teiles erhalten bleibt. Insbesondere werden die Transformationen so durchgeführt, daß die Diagonalelemente von B exakt Null bleiben.

2.2.1.2 Iterativer Teil

Der iterative Teil des Verfahrens besteht aus einer Folge von Zyklen. Zu Beginn jeden Zyklus' wird geprüft, ob (K, J) evtl. mehrfache Eigenwerte haben könnte, indem die Gültigkeit dreier aus Satz 2.1.7 abgeleiteter Bedingungen überprüft wird. Dazu wird K zuerst symplektisch permutiert, um die Diagonalelemente von F und C absteigend zu ordnen. Anschließend werden Submatrizen (Cluster) möglichst großer Ordnung ermittelt, die wir hier mit der Bezeichnung

$$\begin{aligned} K^{[l,u]} &= \begin{bmatrix} F^{[l,u]} & B^{[l,u]^T} \\ B^{[l,u]} & C^{[l,u]} \end{bmatrix} \\ &= \begin{bmatrix} D^{[l,u]} & 0 \\ 0 & D^{[l,u]} \end{bmatrix} \begin{bmatrix} F_S^{[l,u]} & B_S^{[l,u]^T} \\ B_S^{[l,u]} & C_S^{[l,u]} \end{bmatrix} \begin{bmatrix} D^{[l,u]} & 0 \\ 0 & D^{[l,u]} \end{bmatrix} \\ &= \begin{bmatrix} D^{[l,u]} & 0 \\ 0 & D^{[l,u]} \end{bmatrix} K_S^{[l,u]} \begin{bmatrix} D^{[l,u]} & 0 \\ 0 & D^{[l,u]} \end{bmatrix} \end{aligned} \tag{2.35}$$

mit

$$F^{[l,u]} = (F_{ij})_{l \le i,j \le u} \quad , \quad B^{[l,u]} = (B_{ij})_{l \le i,j \le u} \quad , \quad C^{[l,u]} = (C_{ij})_{l \le i,j \le u} \ ,$$

und

$$\mathrm{diag}(F_{S_{ii}}^{[l,u]}) = \mathrm{diag}(C_{S_{ii}}^{[l,u]}) = 1$$

versehen wollen, für die $\omega^{[l,u]} = \|K_S^{[l,u]} - I\|_E$ die Bedingungen

1. $\omega^{[l,u]} \leq 1/(10(u-l+1))$,
2. $F_{ll} \leq (1+\omega^{[l,u]})F_{uu}$ und
3. $\mathrm{Tr}\ (F_S^{[l,u]} - I)(C_S^{[l,u]} - I) - \|B_S^{[l,u]} + B_S^{[l,u]^T}\|_E^2/4 + 12\,\omega^{[l,u]^3} \leq -TOL^2/10$

erfüllt. Algorithmisch geschieht dies auf folgende Weise. Zuerst werden die Bedingungen für $l = 1$ und $u = l + 2 = 3$ geprüft. Dann wird der Index u solange um 1 erhöht, bis für $u+1$ eine der Bedingungen nicht mehr erfüllt ist, oder $u = n$ gilt. Ist eine der Bedingungen bereits für $u = l + 2$ nicht erfüllt, so wird l um 1 erhöht und die Prüfung für $u = l + 2$ fortgesetzt. Andernfalls werden die Indizes l, u des Clusters $K^{[l,u]}$ abgespeichert und die Prüfung für $l = u + 1$ fortgesetzt. Ist $l = n - 2$, so wird der Prozeß beendet. Auf diese Weise gilt stets $u - l \geq 3$.

Die Bedingungen sind algorithmisch leicht überprüfbar. Die auf diese Weise bestimmten Cluster haben nicht immer die maximal mögliche Ordnung, es werden insgesamt jedoch lediglich $\mathcal{O}(n^2)$ Operationen benötigt.

TOL ist eine Abbruchkonstante (s. 2.2.1.3) in der Größenordnung der Maschinengenauigkeit. Numerische Experimente zeigten, daß an dieser Stelle auch auch noch $TOL = 0$ gewählt werden kann. Zur Prüfung der dritten Bedingung müssen natürlich nicht alle Elemente von $K_S^{[l,u]}$ bestimmt werden, sondern z.B. nur diejenigen oberhalb der Diagonale.

Die drei obigen Bedingungen sind aus Satz 2.1.7 abgeleitet. Wegen Satz 2.1.7 ermitteln sie im Falle mehrfacher Eigenwerte gerade diejenigen Cluster, die den mehrfachen Eigenwerten zugeordnet werden können, falls die $\|K_S^{[l,u]} - I\|_E$ nicht bereits klein gegenüber $\|K_S - I\|_E$ sind.

Werden auf diese Weise Cluster ermittelt, so wird angenommen, (K, J) habe mehrfache Eigenwerte und K_S trage die in Kapitel 2.1 beschriebene Struktur. Beim nachfolgenden Zyklus handelt es sich dann um einen *modifizierten Zyklus*, andernfalls um einen *normalen* oder *unmodifizierten Zyklus*. In modifizierten Zyklen werden die ermittelten Cluster wie in (2.17), (2.18) transformiert.

Die drei obigen Bedingungen durchaus auch dann erfüllt sein, wenn keine mehrfachen Eigenwerte vorliegen, wie das Beispiel

$$K = \begin{bmatrix} I_3 & B^T \\ B & I_3 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1/50 & 0 & 0 \\ 0 & 1/100 & 0 \\ 0 & 0 & 1/200 \end{bmatrix}$$

für $TOL \leq 0.07$ zeigt. Anwendung des Transformationspaares (2.17), (2.18) diagonalisiert dieses K aber.

Wir werden in Kapitel 2.3 sehen, daß Transformationen (2.17), (2.18) der

nach obigen Bedingungen ermittelten Cluster stets zur Konvergenz beitragen, auch dann, wenn (K, J) tatsächliche keine mehrfachen Eigenwerte hat.

In den normalen Zyklen werden ausschließlich Diagonalisierungen von 4×4-Submatrizen vorgenommen. Auch in den modifizierten Zyklen werden außerhalb der ermittelten Cluster nur solche Diagonalisierungen vorgenommen. Wir werden zuerst das Vorgehen in den normalen Zyklen behandeln, also davon ausgehen, es gäbe keine Cluster (2.35).

Normaler Zyklus

Hier werden $n(n-1)/2$ Submatrizen

$$\hat{K} = \begin{bmatrix} \hat{F} & \hat{B}^T \\ \hat{B} & \hat{C} \end{bmatrix} = \begin{bmatrix} F_{pp} & F_{pq} & 0 & B_{qp} \\ F_{pq} & F_{qq} & B_{pq} & 0 \\ 0 & B_{pq} & F_{pp} & C_{pq} \\ B_{qp} & 0 & C_{pq} & F_{qq} \end{bmatrix} , \qquad \begin{array}{l} p = 1 \ldots n-1 \ , \\ q = p+1 \ldots n \end{array} \tag{2.36}$$

der Ordnung 4 von K so mit einer endlichen Folge elementarer symplektischer Transformationen diagonalisiert, daß für die diagonalisierte Submatrix wieder (2.34) erfüllt ist.

Die Reihenfolge der Paare (p, q) folgt einer zu Beginn des Verfahrens festgelegten *Pivotstrategie*. In vorliegender Arbeit kommt lediglich die bereits oben angedeutete *zeilenzyklische Pivotstrategie*

$$\begin{array}{cccc} (1,2) & (1,3) & \ldots & (1,n) \\ & (2,3) & \ldots & (2,n) \\ & & & \vdots \\ & & & (n-1,n) \end{array}$$

zum Einsatz. Auf die Darstellung *paralleler Strategien* (siehe z.B. [5], [27]) verzichten wir hier.

Die Transformationsmatrizen, deren Folge $\hat{K}$ diagonalisiert, unterscheiden sich höchstens an den mit $\hat{K}$ korrespondierenden Positionen von der Einheitsmatrix. Dort sind sie vom Typ

$$\hat{T}^{[1]} = \begin{bmatrix} t_1 & 0 & 0 & 0 \\ 0 & t_2 & 0 & 0 \\ 0 & 0 & 1/t_1 & 0 \\ 0 & 0 & 0 & 1/t_2 \end{bmatrix} ,$$

vom Typ

$$\hat{T}^{[2]} = \begin{bmatrix} cs_1 & sn_1 & 0 & 0 \\ -sn_1 & cs_1 & 0 & 0 \\ 0 & 0 & cs_1 & sn_1 \\ 0 & 0 & -sn_1 & cs_1 \end{bmatrix}$$

oder vom Typ

$$\hat{T}^{[3]} = \begin{bmatrix} cs_2 & 0 & sn_2 & 0 \\ 0 & cs_3 & 0 & sn_3 \\ -sn_2 & 0 & cs_2 & 0 \\ 0 & -sn_3 & 0 & cs_3 \end{bmatrix}$$

mit $sn_i^2 + cs_i^2 = 1$, $i = 1, 2, 3$, und $t_1, t_2 > 0$. Es handelt sich also ausschließlich um orthogonale Rotationen und Skalierungen. Bei der Diagonalisierung von $\hat{K}$ wird konkret folgendermaßen vorgegangen:

Zuerst wird $\hat{B}$ mit einer Transformation vom Typ $\hat{T}^{[3]}$ anulliert (zum formellen Vorgehen siehe **3.1** von Verfahren 2.2.2). Die Wirkung einer derartigen Transformation von K erkennt man leicht nach einer Vertauschung der zweiten und dritten Spalten und Zeilen von $\hat{K}$. Dann ergibt sich nämlich die permutierte Matrix

$$\begin{bmatrix} F_{pp} & 0 & F_{pq} & B_{qp} \\ 0 & F_{pp} & B_{pq} & C_{pq} \\ F_{pq} & B_{pq} & F_{qq} & 0 \\ B_{qp} & C_{pq} & 0 & F_{qq} \end{bmatrix}$$

und die Transformation von $\hat{K}$ mit einer Matrix vom Typ $\hat{T}^{[3]}$ ist äquivalent zur Transformation der letztgenannten Matrix mit

$$\begin{bmatrix} cs_p & sn_p & 0 & 0 \\ -sn_p & cs_p & 0 & 0 \\ 0 & 0 & cs_q & sn_q \\ 0 & 0 & -sn_q & cs_q \end{bmatrix}.$$

Man erkennt, daß die Anullierung von $\hat{B}$ äquivalent ist zur Singulärwertzerlegung von

$$\begin{bmatrix} F_{pq} & B_{qp} \\ B_{pq} & C_{pq} \end{bmatrix}$$

ist, wobei die Parameter gemäß [19] bestimmt werden. Man erkennt ferner, daß die Transformation weder die Diagonalelemente von $\hat{F}$ und $\hat{C}$ noch die bereits bestehenden Nullen in der Diagonale von $\hat{B}$ verändert.

Nach der Transformation gilt daher (der Einfachheit halber unter Verwendung

stets gleichbleibender Bezeichnungen auch für veränderte Elemente)

$$\hat{K} = \begin{bmatrix} \hat{F} & 0 \\ 0 & \hat{C} \end{bmatrix} = \begin{bmatrix} F_{pp} & F_{pq} & 0 & 0 \\ F_{pq} & F_{qq} & 0 & 0 \\ 0 & 0 & F_{pp} & C_{pq} \\ 0 & 0 & C_{pq} & F_{qq} \end{bmatrix}$$

und als nächstes wird eine Skalierung

$$\begin{bmatrix} \sqrt{F_{pp}} & 0 & 0 & 0 \\ 0 & \sqrt{F_{qq}} & 0 & 0 \\ 0 & 0 & 1/\sqrt{F_{pp}} & 0 \\ 0 & 0 & 0 & 1/\sqrt{F_{qq}} \end{bmatrix} \tag{2.37}$$

vom Typ $\hat{T}^{[1]}$ angewendet (**3.2** von Verfahren 2.2.2), um 1 als Diagonalelemente von $\hat{C}$ erhalten; nach der Skalierung ist daher

$$\hat{K} = \begin{bmatrix} \hat{F} & 0 \\ 0 & \hat{C} \end{bmatrix} = \begin{bmatrix} F_{pp} & F_{pq} & 0 & 0 \\ F_{pq} & F_{qq} & 0 & 0 \\ 0 & 0 & 1 & C_{pq} \\ 0 & 0 & C_{pq} & 1 \end{bmatrix}.$$

In Satz 3.1.3 werden wir sehen, daß die skalierte Kondition der so skalierten Matrix in den nachfolgenden Operationen nur ein geringes Wachstum erfährt. Die letztgenannte Matrix wird nun so mit einer orthogonalen Matrix vom Typ $\hat{T}^{[2]}$ transformiert (**3.3** von Verfahren 2.2.2), daß F_{pq} verschwindet. Die resultierende Matrix K wird erneut mit einer Transformation

$$\begin{bmatrix} 1/\sqrt{F_{pp}} & 0 & 0 & 0 \\ 0 & 1/\sqrt{F_{qq}} & 0 & 0 \\ 0 & 0 & \sqrt{F_{pp}} & 0 \\ 0 & 0 & 0 & \sqrt{F_{qq}} \end{bmatrix}$$

vom Typ $\hat{T}^{[1]}$ skaliert (**3.4** von Verfahren 2.2.2), diesmal um 1 als Diagonalelemente von $\hat{F}$ zu erhalten. Dies führt zu

$$\hat{K} = \begin{bmatrix} \hat{F} & 0 \\ 0 & \hat{C} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & C_{pp} & C_{pq} \\ 0 & 0 & C_{pq} & C_{qq} \end{bmatrix}.$$

Eine anschließende orthogonale Transformation vom Typ $\hat{T}^{[2]}$ anulliert C_{pq}, ohne

$\hat{F} = I$ zu verändern (**3.5** von Verfahren 2.2.2). Abschließende Skalierung mit

$$\begin{bmatrix} \sqrt[4]{C_{pp}} & 0 & 0 & 0 \\ 0 & \sqrt[4]{C_{qq}} & 0 & 0 \\ 0 & 0 & 1/\sqrt[4]{C_{pp}} & 0 \\ 0 & 0 & 0 & 1/\sqrt[4]{C_{qq}} \end{bmatrix}$$

(**3.6** von Verfahren 2.2.2) ergibt schließlich die diagonale Matrix

$$\hat{K} = \begin{bmatrix} \hat{F} & 0 \\ 0 & \hat{C} \end{bmatrix} = \begin{bmatrix} F_{pp} & 0 & 0 & 0 \\ 0 & F_{qq} & 0 & 0 \\ 0 & 0 & F_{pp} & 0 \\ 0 & 0 & 0 & F_{qq} \end{bmatrix} ,$$

mit der Eigenschaft (2.34).

Falls einzelne Außerdiagonalelemente von $\hat{K}$ bereits klein genug sind, so können die entsprechenden Transformationen weggelassen werden. Dies wird in Verfahren 2.2.2 nicht berücksichtigt, wohl aber beim impliziten Verfahren 2.2.4.

Modifizierter Zyklus

Bei einem modifizierten Zyklus werden zu jedem festgestellten Cluster (2.35) Transformationsmatrizen wie in (2.17), (2.18) bestimmt, die sich höchstens dort von der Einheitsmatrix unterscheiden, wo sie mit dem jeweiligen Cluster korrespondieren. Die Cluster $K^{[low[clu],upp[clu]]}$ seien für $clu = 1 \ldots num$ durchnummeriert mit entsprechenden Bezeichnungen für die Submatrizen. Entsprechend der Bezeichnung in (2.35) ist dabei *low* ein Ganzzahlvektor der unteren Indizes der Cluster und *upp* ein Ganzzahlvektor mit den entsprechenden obereren Indizes. Die erste Transformation (2.17) erhält man mit Hilfe der Cholesky-Zerlegungen

$$\frac{1}{F_{low[clu],low[clu]}} F^{[low[clu],upp[clu]]} = R_{clu}^T R_{clu} \quad , \quad clu = 1 \ldots num$$

und (**2.3.1** von Verfahren 2.2.2)

$$R = \bigoplus_{clu=1}^{num} (I_{low[clu]-upp[clu-1]-1} \oplus R_{clu}) \oplus I_{n-upp[clu]} ,$$

wobei $clu[0] = 0$. Die erste Transformation (**2.3.2** von Verfahren 2.2.2) lautet dann in abkürzender Notation

$$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T \quad , \quad T = \begin{bmatrix} R^{-1} & 0 \\ 0 & R^T \end{bmatrix} . \tag{2.38}$$

Anschließend wird für

$$X_{clu} = -\frac{1}{2F_{low[clu],low[clu]}} (B^{[low[clu],upp[clu]]} + B^{[low[clu],upp[clu]]^T}) , \quad clu = 1 \ldots num ,$$

und (**2.3.3** von Verfahren 2.2.2)

$$X = \bigoplus_{clu=1}^{num} (0_{low[clu]-\ upp[clu-1]-1} \oplus X_{clu}) \oplus I_{n-upp[clu]}$$

die zweite Transformation (2.18) gebildet, die wir abkürzend als

$$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T \quad , \quad T = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} . \tag{2.39}$$

darstellen (**2.3.4** von Verfahren 2.2.2). Weil die transformierten Submatrizen von B nun schiefsymmetrisch sind, gilt noch $B_{ii} = 0$, $i = 1 \ldots n$, so daß nur noch geeignete Skalierungen vorzunehmen sind, um die Eigenschaft (2.34) wiederherzustellen (**2.3.5** von Verfahren 2.2.2).

Man erkennt, daß die Transformationen (2.38) und (2.39) zusätzlichen Speicherplatz für L und X von etwa $n^2/2$ Elementen erfordern, wenn sie so wie hier geschildert durchgeführt werden. Eine Methode ohne bzw. mit minimalem zusätzlichen Speicherplatz wird im Unterabschnitt über das implizite Verfahren beschrieben.

Außerhalb der Cluster werden Transformationen angewendet, wie sie im Unterabschnitt über den normalen Zyklus behandelt wurden. Auch die dort festgelegte Pivotstrategie wird verwendet. Genauer werden Submatrizen $\hat{K}$ gemäß (2.36) diagonalisiert, für die nicht $low[clu] \leq p, q \leq upp[clu]$, clu $\in \{1 \ldots num\}$ gilt.

Von den modifizierten Zyklen erwarten wir den Eintritt asymptotisch quadratischer Konvergenz der Folge der iterierten symmetrischen Matrizen gegen eine Diagonalmatrix. Diese Erwartung ist in der Tat begründet. Sind nämlich alle Eigenwerte von (K, J) einfach, so wird man die asymptotisch quadratische Konvergenz genauso wie bei anderen Jacobi-ähnlichen Verfahren zeigen können (siehe z.B. [16]). Das Konvergenzverhalten basiert dann auf der Konvergenz der Folge der skalierten Transformationsmatrizen gegen die Einheitsmatrix. Weil dies im Falle mehrfacher Eigenwerte auch für die Transformationsmatrizen in (2.38), (2.39) gilt, können wir auch im Falle mehrfacher Eigenwerte asymptotisch quadratische Konvergenz erwarten.

2.2.1.3 Beendigung des Verfahrens

Das Verfahren wird beendet, wenn nach einem Zyklus das *Konvergenzkriterium*

$$\frac{|K_{ij}|}{\sqrt{K_{ii} K_{jj}}} < TOL \quad , \quad 1 \leq i, j \leq 2n \ , \quad i < j \tag{2.40}$$

mit einem geeigneten TOL, z.B. $TOL = 2n\,\varepsilon$ mit der Maschinengenauigkeit ε, erfüllt ist. Die Diagonalelemente von K sind dann Näherungen an die positiven

Imaginärteile der Eigenwerte von (K, J).

Wir verwenden also eine Abbruchbedingung wie in [13]. Sie ist konsistent mit Satz 1.0.6, wonach die relative Genauigkeit der Eigenwerte von der Kondition der skalierten Matrix $\kappa(K_S)$ abhängt. Das Verfahren ist also dann abzubrechen, wenn die Außerdiagonalelemente der skalierten Matrix klein genug sind.

Die Eigenvektoren von (K, J) könnten mittels Akkumulation der Transformationsmatrizen bestimmt werden. Dies erfordert etwa den doppelten Speicherplatz und etwa den zweifachen Rechenaufwand. Auf die Darstellung wurde verzichtet, weil uns die Eigenvektoren von (K, J) nicht interessieren.

2.2.2 Der formale Algorithmus des expliziten Verfahrens

In diesem Unterabschnitt geben wir ausschließlich das formelle Vorgehen beim expliziten Verfahren zur Bestimmung der Eigenwerte des Paares (K, J) an. Wir werden später darauf zurückkommen; vorwiegend dient es allerdings dem leichteren Verständnis eines später angegeben äquivalenten impliziten Verfahren.

/* The most important variables we use in the algorithm are stated as follows

name	**type**	**usage**	**comment**
TOL	scalar	input	stopping bound with $0 < TOL \ll 1$, e.g $TOL = 2n \cdot$ *macheps*, where *macheps* is the machine precision.
$\sigma_{\max}$	scalar	input	determines the minimal increase of the Hadamard measure in modified cycles, e.g. $\sigma_{\max} = -\ TOL^2/10$.
K	matrix	input/ output	symmetric positive definite matrix to be diagonalised; we write $$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix}, \quad F, B, C \in I\!R^{n \times n}.$$
T	matrix	inter- mediate	transformation matrix; T is the identity matrix except at the positions of a submatrix $\hat{T}$. In a computer program only components of $\hat{T}$ are stored. */

/* 1. preparation */
/* 1.1 scale K for $F_{pp} = C_{pp}, p = 1 \ldots n$ */
set $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T$ with $T = \mathrm{diag}(\sqrt[4]{C_{pp}/F_{pp}}) \oplus \mathrm{diag}(\sqrt[4]{F_{pp}/C_{pp}})$

/* 1.2 annihilate diagonal entries of B */

set $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T$ with $T = \frac{1}{\sqrt{2}} \begin{bmatrix} I & I \\ -I & I \end{bmatrix}$

/* 1.3 rescale K for $F_{pp} = C_{pp}, p = 1 \ldots n$, */

set $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T$ with $T = \mathrm{diag}(\sqrt[4]{C_{pp}/F_{pp}}) \oplus \mathrm{diag}(\sqrt[4]{F_{pp}/C_{pp}})$

/* now we have $F_{pp} = C_{pp}$ and $B_{pp} = 0$, $p = 1 \ldots n$ */

/* begin cycle */

WHILE $\max_{1 \le i \le 2n, 1 \le j \le i-1} |K_{ij}|/\sqrt{K_{ii}K_{jj}} \ge TOL$ DO

  /* 2.1 prepare for modification */

  apply symplectic permutation to K to get decreasing order
    of diagonal elements of F and C

  /* 2.2 determine clusters */

  $num = 1$ /* number of clusters */

  $l = 1$ /* lower index of first cluster */

A: IF $l + 1 < n$ THEN /* begin new cluster */

    $\sigma = (2F_{l+1,l}C_{l+1,l} - (B_{l+1,l} + B_{l,l+1})^2/2)/(F_{l+1,l+1}F_{ll})$

    $\bar{\omega} = 2(F_{l+1,l}^2 + C_{l+1,l}^2 + B_{l+1,l}^2 + B_{l,l+1}^2)/(F_{l+1,l+1}F_{ll})$

  ENDIF

  $u = l + 2$ /* upper index of actual cluster */

B: IF $u > n$ THEN

    $num = num - 1$ /* finish */

  ELSE

    $\sigma = \sigma + 2\sum_{i=l}^{u-1}(F_{ui}C_{ui})/(F_{uu}F_{ii}) - (\sum_{i=l}^{u-1}(B_{ui} + B_{iu})^2/(F_{uu}F_{ii}))/2$

    $\bar{\omega} = \bar{\omega} + 2\sum_{i=l}^{u-1}(F_{ui}^2 + C_{ui}^2 + B_{ui}^2 + B_{iu}^2)/(F_{uu}F_{ii})$

    $\omega = \sqrt{\bar{\omega}}$

    IF $10(u - l + 1)\,\omega \le 1 \wedge F_{ll} \le (1 + \omega)F_{uu} \wedge \sigma + 12\omega^3 \le \sigma_{\max}$ THEN

      IF $u = n$ THEN /* finish last cluster */

        $low[num] = l$, $upp[num] = u$

      ELSE /* increase upper index of actual cluster */

        $u = u + 1$, GOTO **B:**

      ENDIF

    ELSE

      IF $u = l + 2$ THEN /* restart cluster with increased lower index */

        $l = l + 1$, GOTO **A:**

      ELSE /* begin next cluster */

        $low[num] = l$, $upp[num] = u - 1$

        $num = num + 1$, $l = u$, GOTO **A:**

      ENDIF

    ENDIF

ENDIF
/* the extremal indices of clusters are now stored in integer vectors "*low*" and "*upp*" of length "*num*" */

/* 2.3 eventual modified cycle with rescaling */
IF *num*> 0 THEN
FOR *clu*= 1 TO *num* DO
/* 2.3.1 make first transformation matrix */
FOR $i = 1$ TO $upp[clu]-low[clu]+1$ DO
$R_{clu_{ii}} = (F_{low[clu]-1+i,low[clu]-1+i}/F_{low[clu],low[clu]} - \sum_{k=1}^{i-1} R^2_{clu_{ki}})^{1/2}$
FOR $j = i+1$ TO $upp[clu]-low[clu]+1$ DO
$R_{clu_{ij}} = (F_{low[clu]-1+j,low[clu]-1+i}/F_{low[clu],\mathrm{low}[clu]}$
$- \sum_{k=1}^{i-1} R_{clu_{kj}} R_{clu_{ki}})/R_{clu_{ii}}$
ENDFOR
ENDFOR
ENDFOR

/* 2.3.2 apply first transformation */
set $R = \oplus_{clu=1}^{num}(I_{low[clu]-upp[clu-1]-1} \oplus R_{clu}) \oplus I_{n-upp[num]}$ with $upp[0] = 0$
set $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T$, where $T = \begin{bmatrix} R^{-1} & 0 \\ 0 & R^T \end{bmatrix}$

/* 2.3.3 make second transformation matrix */
FOR $clu = 1$ TO *num* DO
FOR $i = 1$ TO $upp[clu]-low[clu]+1$ DO
FOR $j = i$ TO $upp[clu]-low[clu]+1$ DO
$X_{clu_{i,j}} = -(B_{low[clu]-1+i,low[clu]-1+j}$
$+B_{low[clu]-1+j,low[clu]-1+i})/(2F_{low[clu],low[clu]})$
IF $i \neq j$ THEN set $X_{clu_{ji}} = X_{clu_{ij}}$
ENDFOR
ENDFOR
ENDFOR

/* 2.3.4 apply second transformation */
set $X = \oplus_{clu=1}^{num}(0_{low[clu]-upp[clu-1]-1} \oplus X_{clu}) \oplus I_{n-upp[num]}$ with $upp[0] = 0$
set $K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = T^T K T$, where $T = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$

/* 2.3.5 scale K for $F_{pp} = C_{pp}$ within clusters */
FOR $clu = 1$ TO *num* DO
FOR $p =low[clu]$ TO $upp[clu]$ DO
set $K = T^T K T$ with $\hat{T} = \begin{bmatrix} t & 0 \\ 0 & t^{-1} \end{bmatrix}$, so that $F_{pp} = C_{pp}$

```
      ENDFOR
    ENDFOR
  ENDIF
```

/* 3. begin cyclic rotations */

FOR $p = 1$ TO $n - 1$ DO
  FOR $q = p + 1$ TO n DO
    rotate= TRUE
    FOR *clu* = 1 TO *num* DO
      IF *low*[*clu*] $\leq p, q \leq$*upp*[*clu*] THEN *rotate*= FALSE
    ENDFOR
    IF *rotate* THEN

/* 3.1 annihilate B_{pq} and B_{qp} */

IF $|B_{pq}|/(F_{pp}F_{qq})^{1/2} \geq TOL \vee |B_{qp}|/(F_{pp}F_{qq})^{1/2} \geq TOL$ THEN

compute SVD-parameters cs_1, sn_1, cs_2, sn_2,

so that $\begin{bmatrix} cs_1 & -sn_1 \\ sn_1 & cs_1 \end{bmatrix} \begin{bmatrix} F_{qp} & B_{qp} \\ B_{pq} & C_{qp} \end{bmatrix} \begin{bmatrix} cs_2 & sn_2 \\ -sn_2 & cs_2 \end{bmatrix} = \text{diag}$

set $K = T^T K T$, where $\hat{T} = \begin{bmatrix} cs_1 & 0 & sn_1 & 0 \\ 0 & cs_2 & 0 & sn_2 \\ -sn_1 & 0 & cs_1 & 0 \\ 0 & -sn_2 & 0 & cs_2 \end{bmatrix}$

ENDIF

IF $|F_{pq}|/(F_{pp}F_{qq})^{1/2} \geq TOL$ THEN

/* 3.2 scale K for $C_{pp} = C_{qq} = 1$ */

set $K = T^T K T$, where $\hat{T} = \text{diag}(C_{pp}^{1/2}, C_{qq}^{1/2}, C_{pp}^{-1/2}, C_{qq}^{-1/2})$

/* 3.3 annihilate F_{qp} and F_{pq} */

find orthogonal U, so that $U^T \cdot \begin{bmatrix} F_{pp} & F_{pq} \\ F_{qp} & F_{qq} \end{bmatrix} \cdot U = \text{diag}$,

set $K = T^T K T$, where $\hat{T} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$

ENDIF

IF $|C_{pq}|/(C_{pp}C_{qq})^{1/2} \geq TOL$ THEN

/* 3.4 scale K for $F_{pp} = F_{qq} = 1$ */

set $K = T^T K T$, where $\hat{T} = \text{diag}(F_{pp}^{-1/2}, F_{qq}^{-1/2}, F_{pp}^{1/2}, F_{qq}^{1/2})$

/* 3.5 annihilate C_{qp} and C_{pq} */

find orthogonal V, so that $V^T \cdot \begin{bmatrix} C_{pp} & C_{pq} \\ C_{qp} & C_{qq} \end{bmatrix} \cdot V = \text{diag}$,

set $K = T^T KT$, where $\hat{T} = \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix}$
ENDIF

/* 3.6 scale K for $F_{pp} = C_{pp}$ and $F_{qq} = C_{qq}$ */
set $K = T^T KT$ with $\hat{T} = \mathrm{diag}(t_{pp}, t_{qq}, t_{pp}^{-1}, t_{qq}^{-1})$, so that
$F_{pp} = C_{pp}$, $F_{qq} = C_{qq}$
ENDIF
ENDFOR
ENDFOR
ENDWHILE

2.2.3 Implizites Verfahren

In den vergangenen Unterabschnitten haben wir ein explizites Jakobi-ähnliches Verfahren für das Paar (K, J) entwickelt. Dies wird uns nun als Grundlage für ein äquivalentes, recht kompliziertes, implizites Verfahren dienen. Das Hauptanliegen vorliegender Arbeit gilt dem impliziten Verfahren.

Beim impliziten Verfahren [32], [31] geht man davon aus, daß bereits der Faktor L der symmetrischen positiv definiten Matrix $K = L^T L$ vorliegt. L wird in der Regel aus der Faktorisierung einer schiefsymmetrischen Matrix

$$S = LJL^T \tag{2.41}$$

entstanden sein (s. Kapitel 4). Diagonalisiert ein symplektisches T die Matrix K, also

$$T^T KT = T^T L^T LT = \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} = \Delta \ , D = \mathrm{diag} \ , \tag{2.42}$$

und setzt man $V = LT$, so ist $Q = V\Delta^{-1/2}$ offensichtlich orthogonal. Außerdem gilt

$$\begin{aligned} Q^T SQ &= \Delta^{-1/2} V^T LJL^T V \Delta^{-1/2} \\ &= \Delta^{-1/2} T^T L^T LT^T JTL^T LT \Delta^{-1/2} \\ &= \Delta^{1/2} J \Delta^{1/2} \ , \end{aligned}$$

d.h. die $\pm i \cdot \Delta_{jj}$ $(i^2 = -1)$ sind die Eigenwerte von S und die Spalten von Q sind Real- und Imaginärteile der Eigenvektoren von S.

Dies kann man sich für ein implizites Verfahren zunutze machen, bei dem man das Produkt $K = L^T L$ nicht explizit bildet, sondern lediglich den Faktor $L = L^{(1)}$ solange mit einer Folge symplektischer Matrizen $T^{(m)}$ von rechts multipliziert

$$L^{(m+1)} = L^{(m)} T^{(m)} \ ,$$

bis die Spalten eines $L^{(M)}$ hinreichend orthogonal zueinander sind und die Normen der Spalten j und $n+j$, $j = 1 \dots n$, übereinstimmen.

$$\Lambda = \operatorname{diag}\left(\|L_{\cdot 1}^{(M)}\|_E^2, \dots, \|L_{\cdot 2n}^{(M)}\|_E^2\right)$$

enthält dann Näherungswerte der positiven Imaginärteile der Eigenwerte von S und $L^{(M)}\Lambda^{-1/2}$ sind Näherungen an die Real- und Imaginärteile der Eigenvektoren von S. Dabei bezeichne $L_{\cdot j}^{(m)}$ die j-te Spalte von $L^{(m)}$.

Mit dem impliziten Verfahren spart man gegenüber einem expliziten Verfahren Rechenzeit und Speicherplatz, weil die Matrix der Eigenvektoren als Akkumulation der Transformationen nicht mitgeführt werden muß. Darüber hinaus kann man gegenüber dem expliziten Verfahren einen kleineren maximalen relativen Fehler in den berechneten Eigenwerte erwarten [13], [29].

Beim impliziten Verfahren werden die Transformationsmatrizen prinzipiell genauso bestimmt, wie beim expliziten Verfahren. Dabei ist die Anullierung eines Elementes K_{ij} beim expliziten Verfahren wegen $K_{ij} = \sum_{i=1}^{2n} L_{ki} L_{kj}$ identisch mit der Orthogonalisierung der Spalten i und j von L.

Es sei darauf hingewiesen, daß der in Verfahren 2.2.4 aufgeführte formale Algorithmus in einigen Details über die nachstehende Beschreibung hinausgeht, indem z.B. beim Algorithmus 2.2.4 darauf geachtet wird, daß einmal berechnete Elemente K_{ij} oder $K_{S_{ij}}$ nicht ein zweites Mal bestimmt werden. Darüber hinaus werden Divisionen möglichst durch Multiplikationen ersetzt. Diese Details bleiben bei der nachfolgenden Beschreibung unberücksichtigt.

Wir gehen im Folgenden davon aus, daß bereits der Faktor L von $K = L^T L$ vorliege. Der vorbereitende Teil unterscheidet sich etwas von der Vorbereitung des expliziten Verfahrens. Zu Beschleunigung des iterativen Teils des Verfahrens wollen wir nämlich schnelle Rotationen anwenden [22], wozu die Mitführung einer zusätzlichen Diagonalmatrix E notwendig wird.

2.2.3.1 Vorbereitung

Es ist nützlich, zu Beginn eines jeden Zyklus die Spalten von L auf 1 normiert zu haben und die tatsächlichen Normen der Spalten in einer weiteren Diagonalmatrix E (als Vektor) gespeichert zu haben. Die ersten Schritte der Vorbereitung besteht dann aus den Anweisungen (s. **1.1.1**, **1.1.2** von Verfahren 2.2.4)

FOR $i = 1$ TO $2n$ DO $E_i = (L_{\cdot i}^T L_{\cdot i})^{1/2}$

FOR $i = 1$ TO $2n$ DO $L_{\cdot i} = L_{\cdot i} E_i^{-1}$

Statt $E = \operatorname{diag}(E_{ii})$ werden wir durchgehend die Notation $E = \operatorname{diag}(E_i)$ verwenden, um daran zu erinnern, daß die Diagonalmatrix E als Vektor gespeichert

wird. Wir wollen wie beim expliziten Verfahren gleichbleibende Bezeichnungen auch für veränderte Variablen verwenden. Daher werden wir nun stets

$$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} = EL^T LE \tag{2.43}$$

für die zu diagonalisierende, positiv definite Matrix schreiben.

Die Transformation von K mit V_1 aus (2.31) zur Erzielung von $F_{ii} = C_{ii}$, d.h. $E_i = E_{n+i}$, wird beim impliziten Verfahren leicht durch die Schleife

FOR $i = 1$ TO n DO
  $E_i = E_{i+n} = E_i^{1/2} E_{i+n}^{1/2}$
ENDFOR

ausgedrückt (**1.1.3** von Verfahren 2.2.4). Beim expliziten Verfahren schließt sich nun die Anullierung der Diagonalelemente von B in K gemäß (2.43) an. Das Analogon beim impliziten Verfahren muß auch während des iterativen Teils regelmäßig, z.B. zu Beginn eines jeden Zyklus, wiederholt werden, weil man im Gegensatz zum expliziten Fall nicht davon ausgehen kann, daß die einmal orthogonalisierten Spaltenvektoren $L_{\cdot i}$ und $L_{\cdot n+i}$ während des gesamten iterativen Prozesses genügend orthogonal bleiben.

2.2.3.2 Iterativer Teil

Wie soeben dargelegt, wird LE zu Beginn eines jeden Zyklus mit V_2 gemäß (2.33) von rechts multipliziert. Die Multiplikation brauch dabei nur solche Paare von Spalten von L zu erfassen, die nicht bereits hinreichend orthogonal zueinander sind. Außerdem sollen die berührten Spalten von L anschließend wieder auf 1 normiert und auch $E_i = E_{i+n}$ wiederhergestellt werden. Beachtet man die Kommutativität von E und V_2, so können all diese Operationen in der einzigen Schleife

FOR $p = 1$ TO n DO
  IF $|L_{\cdot p}^T L_{\cdot p+n}| \geq TOL$ THEN
    $[L_{\cdot p} \;\; L_{\cdot p+n}] = [L_{\cdot p} - L_{\cdot p+n} \;\; L_{\cdot p} + L_{\cdot p+n}]$
    $(E_p, E_{p+n}) = (1/\sqrt{2})(E_p, E_{p+n})$
    $\delta_1 = (L_{\cdot p}^T L_{\cdot p})^{1/2}$
    $\delta_2 = (L_{\cdot n+p}^T L_{\cdot n+p})^{1/2}$
    $E_p = \delta_1 \, E_p$
    $E_{n+p} = \delta_2 \, E_{n+p}$
    $[L_{\cdot p}] = \delta_1^{-1} [L_{\cdot p}]$
    $[L_{\cdot n+p}] = \delta_2^{-1} [L_{\cdot n+p}]$
    $E_p = E_{p+n} = E_p^{1/2} E_{p+n}^{1/2}$
  ENDIF
ENDFOR

zusammengefaßt werden (**1.2** und **1.3** von Verfahren 2.2.4). Dabei ist TOL eine zuvor festgelegte Abbruchkonstante, auf die wir in 2.2.3.3 zurückkommen werden.

Anschließend (**2.1** von Verfahren 2.2.4) werden die Spalten von L und die zugehörigen Elemente von E permutiert, um die Elemente E_i, $1 \leq i \leq n$, absteigend zu ordnen. Entsprechend dem expliziten Verfahren gilt nun

$$L_{\cdot i}^T L_{\cdot n+i} = 0 \quad , \quad \|L_{\cdot i}\|_2 = 1 \quad , \quad E_i = E_{i+n} \quad , \quad i = 1 \ldots n$$

und

$$E_i \geq E_{i+1} \quad , \quad i = 1 \ldots n-1$$

Zur Entscheidung darüber, ob ein nachfolgender Zyklus modifiziert ist oder nicht, wird $L^T L$ analog zum expliziten Verfahren auf Cluster geprüft, die die drei Bedingungen von Seite 44 erfüllen (**2.2** von Verfahren 2.2.4). Sie lauten in der Notation des impliziten Verfahrens mit $L^{[l,u]} = [L_{\cdot l} \cdots L_{\cdot u} \quad L_{\cdot n+l} \cdots L_{\cdot n+u}]$ und $\omega^{[l,u]} = \|L^{[l,u]^T} L^{[l,u]} - I\|_E$

1. $\omega^{[l,u]} \leq 1/(10(u-l+1))$,
2. $E_l^2 \leq (1+\omega^{[l,u]}) E_u^2$ und
3. $\mathrm{Tr}\,([L_{\cdot l} \cdots L_{\cdot u}]^T [L_{\cdot l} \cdots L_{\cdot u}] - I)([L_{\cdot n+l} \cdots L_{\cdot n+u}]^T [L_{\cdot n+l} \cdots L_{\cdot n+u}] - I)$
   $-\|[L_{\cdot n+l} \cdots L_{\cdot n+u}]^T [L_{\cdot l} \cdots L_{\cdot u}] + [L_{\cdot l} \cdots L_{\cdot u}]^T [L_{\cdot n+l} \cdots L_{\cdot n+u}]\|_E^2/4$
   $+ 12\,\omega^{[l,u]^3} \leq -TOL^2/10$.

Normaler Zyklus

Wie beim expliziten Verfahren wird ein normaler Zyklus dann durchgeführt, wenn es keine Cluster gibt, die den o.g. drei Bedingungen genügen. Es werden dann Submatrizen $\hat{K}$ gemäß (2.36) von $K = L^T L$ diagonalisiert, so daß wir in diesem Unterabschnitt nur noch auf einige Aspekte einer solchen Diagonalisierung einzugehen brauchen.

Zuerst werden die Elemente B_{pq} und B_{qp} anulliert (**3.1** von Verfahren 2.2.4). Nach Berechnung dieser Elemente und F_{pq} sowie C_{pq} werden gemäß [19], (3.1a), (3.1b) orthogonale Parameter sn_p, cs_p, sn_q, cs_q zur Singulärwertzerlegung bestimmt:

$$\begin{bmatrix} cs_p & -sn_p \\ sn_p & cs_p \end{bmatrix} A \begin{bmatrix} cs_q & sn_q \\ -sn_q & cs_q \end{bmatrix} = \mathrm{diag} \tag{2.44}$$

mit

$$A = \frac{1}{\sqrt{F_{pp} F_{qq}}} \begin{bmatrix} E_p E_q L_{\cdot p}^T L_{\cdot q} & E_p E_{q+n} L_{\cdot q+n}^T L_{\cdot p} \\ E_q E_{p+n} L_{\cdot p+n}^T L_{\cdot q} & E_{p+n} E_{q+n} L_{\cdot p+n}^T L_{\cdot q+n} \end{bmatrix} .$$

Für $sn, cs \neq 0$ und $t = sn/cs$ ist

$$\begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} = cs \cdot \begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix} = sn \cdot \begin{bmatrix} 1/t & 1 \\ -1 & 1/t \end{bmatrix} ,$$

so daß schnelle Rotationen [22] angewendet werden können. Für $i = p, q$ ist

$$\begin{aligned}
&\begin{bmatrix} L_{\cdot i} & L_{\cdot i+n} \end{bmatrix} \begin{bmatrix} E_i & 0 \\ 0 & E_{i+n} \end{bmatrix} \begin{bmatrix} cs_i & sn_i \\ -sn_i & cs_i \end{bmatrix} \\
&= \begin{bmatrix} L_{\cdot i} & L_{\cdot i+n} \end{bmatrix} \begin{bmatrix} 1 & t_i E_i / E_{i+n} \\ -t_i E_{i+n}/E_i & 1 \end{bmatrix} \begin{bmatrix} cs_i E_i & 0 \\ 0 & cs_i E_{i+n} \end{bmatrix} && (2.45) \\
&= \begin{bmatrix} L_{\cdot i} & L_{\cdot i+n} \end{bmatrix} \begin{bmatrix} E_i/(|t_i| E_{i+n}) & \operatorname{sign}(sn_i) \\ -\operatorname{sign}(sn_i) & E_{i+n}/(|t_i| E_i) \end{bmatrix} \begin{bmatrix} |sn_i| E_{i+n} & 0 \\ 0 & |sn_i| E_i \end{bmatrix} && (2.46)
\end{aligned}$$

Weil die aus (2.44) entstehenden t nicht beschränkt sind, wird im Falle $|t| \leq 1$ die Multiplikation (2.45) gewählt, sonst (2.46). Dann liegt der mit E_i, E_{i+n} multiplizierte Wert zwischen $1/\sqrt{2}$ und 1, so daß sich die Wahrscheinlichkeit für Variablenunter- oder -überlauf verringert. Bei numerischen Experimenten kam es niemals zu Überlauf. Die trivialen Fälle $sn_i = 0$ oder $cs_i = 0$ werden gesondert behandelt. Die in (2.46) vorzunehmende Rechtsmultiplikation an $[L_{\cdot i} \;\; L_{\cdot i+n}]$ kann wegen $|t_i| \geq 1$ einer Permutation recht nahe kommen. Diese werden wir später (in **3.6** von Verfahren 2.2.4) durch Rückpermutation wieder aufheben. Zur geringfügigen Beschleunigung des Verfahrens könnte man die Rückpermutation auch unmittelbar an eine Transformation (2.46) anschließen. Dadurch wären jedoch die nachfolgenden Transformationen betroffen, worauf wir aus Transparenzgründen verzichten wollen.

An die Anullierung von B_{pq}, B_{qp} schließt sich wie beim expliziten Verfahren die Diagonalisierung von $\hat{F}$ und $\hat{C}$ an. Im formalen Algorithmus 2.2.4 wird hier besonders darauf geachtet, daß Elemente K_{ij} nicht unnötig mehrfach berechnet werden. Dies geschieht insbesondere durch Zwischenspeicherung bereits berechneter Elemente und durch Verwendung der Eigenschaft $F_{ii} = C_{ii}$. In Verfahren 2.2.4 werden dazu die boolschen Variablen $rotF$ und $rotC$ eingeführt. Außerdem werden Rotationen nur dann durchgeführt, wenn die zugehörigen Spalten von L nicht bereits hinreichend orthogonal sind. Bis auf diese Unterschiede, auf die wir jedoch nicht näher eingehen brauchen, verläuft die Diagonalisierung von $\hat{F}$, $\hat{C}$ analog zum expliziten Fall, wenn auch komplizierter als dort. An dieser Stelle soll daher lediglich die Skalierung (**3.2** in Verfahren 2.2.4) als Analogon zu (2.37) und die nachfolgende Diagonalisierung von $\hat{F}$ (**3.3** in 2.2.4) skizziert werden.

Zur Skalierung auf $C_{pp} = C_{qq} = 1$ wird

$$(E_{p+n} t_p, E_{q+n} t_q, E_{p+n}^{-1} t_p^{-1}, E_{q+n}^{-1} t_q^{-1})$$

mit

$$t_i = \frac{E_i}{E_{i+n}} (L_{\cdot i}^T L_{\cdot i})^{1/2} = (L_{\cdot i+n}^T L_{\cdot i+n})^{1/2} \quad , \quad i = p, q$$

bestimmt. Die Skalierung selbst brauch dann nur noch am diagonalen E zu erfolgen, ohne die Matrix L zu berühren.

Die anschließende Multiplikation mit einer orthogonalen Matrix erfordert die Bestimmung orthogonaler Parameter $sn, cs, t = sn/cs$ mit

$$\begin{bmatrix} cs & -sn \\ sn & cs \end{bmatrix} \begin{bmatrix} E_p^2 L_{\cdot p}^T L_{\cdot p} & E_p E_q L_{\cdot p}^T L_{\cdot q} \\ E_p E_q L_{\cdot p}^T L_{\cdot q} & E_q^2 L_{\cdot q}^T L_{\cdot q} \end{bmatrix} \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} = \text{diag} \ .$$

Weil stets $|t| \leq 1$ gilt, können schnelle Rotationen ähnlich wie in (2.45) verwendet werden.

Modifizierter Zyklus

Auch hier wird analog zum expliziten Verfahren verfahren: festgestellte Cluster werden gemäß (2.38), (2.39) transformiert. Über den Speicherplatz im Umfang von $4n^2$ Elementen zur Abspeicherung von L (sowie einiger Vektoren) hinaus wollen wir jedoch möglichst wenig zusätzlichen Speicher verwenden. Beim expliziten Verfahren ist weiterer Speicher von höchstens $n^2/2$ Elementen nötig (s. Seite 49). Vom nachfolgend geschilderten Vorgehen kann man bei dynamischer Speicherverwaltung zeigen, daß höchstens $n^2/4$ zusätzliche Speicherelemente benötigt werden, ohne daß die Fehleranalyse davon berührt wäre.

Wir betrachten zuerst eine Transformation gemäß (2.38) an einem einzelnen Cluster

$$K^{[l,u]} = \begin{bmatrix} F^{[l,u]} & B^{[l,u]^T} \\ B^{[l,u]} & C^{[l,u]} \end{bmatrix} \tag{2.47}$$

der Ordnung $2(u - l + 1)$. Die in (2.47) verwendete Notation wurde in (2.35) eingeführt. Dabei verwenden wir der Einfachheit halber die Bezeichnung

$$l = low[clu] \quad , \quad u = upp[clu] \ .$$

Unser Ziel ist es, $K^{[l,u]}$ so zu transformieren, daß die Transformierte von $F^{[l,u]}$ die Einheitsmatrix ist. Wir gehen dabei induktiv vor und nehmen an, daß für ein p, $l < p \leq u$, die erste Hauptuntermatrix $\hat{F}$ der Ordnung $p - l + 1$ von $F^{[l,u]}$ implizit bereits in der Form

$$\hat{F} = F_{ll} \begin{bmatrix} I_{p-l} & x \\ x^T & g \end{bmatrix} \quad , \quad g > 0$$

vorliege. (Zu Beginn, also für $p = l + 1$, bedeutet dies nicht anderes als $\hat{F} = \begin{bmatrix} F_{ll} & F_{l,l+1} \\ F_{l,l+1} & F_{l+1,l+1} \end{bmatrix}$.) In einem Schritt werden nun die Außerdiagonalelemente von $\hat{F}$ anulliert und dann p um 1 inkrementiert. Nachfolgend werden wir die zur Elimination der Außerdiagonalelemente, also der p-ten Zeile und Spalte, von $\hat{F}$ erforderlichen Operationen beschreiben. Wir verwenden die Bezeichnungen

$$\hat{F} = \hat{E}[L_{\cdot l} \cdots L_{\cdot p}]^T [L_{\cdot l} \cdots L_{\cdot p}] \hat{E} \ ,$$

$$\hat{L} = [L_{\cdot l} \cdots L_{\cdot p} \quad L_{\cdot n+l} \cdots L_{\cdot n+p}] \ ,$$

$$\hat{E} = \mathrm{diag}(\underbrace{E_l, \ldots, E_l}_{p-l}, E_p) \oplus \mathrm{diag}(E_l, \frac{E_{l+1}^2}{E_l}, \ldots, \frac{E_{p-1}^2}{E_l}, E_p) \ ,$$

wobei wir wegen **1.2.4** von Verfahren 2.2.4 noch $\|[L_{\cdot l}]\|_2 = \|[L_{\cdot p}]\|_2 = 1$ annehmen können. Es gilt dann $E_l^2 = F_{ll}$ und

$$\begin{aligned} F_{ll}^{-1}\hat{F} &= E_l^{-2} \begin{bmatrix} E_l^2 L_{\cdot l}^T L_{\cdot l} I_{p-l} & E_l [L_{\cdot l} \cdots L_{\cdot p-1}]^T L_{\cdot p} E_p \\ E_p L_{\cdot p}^T [L_{\cdot l} \cdots L_{\cdot p-1}] E_l & E_p^2 L_{\cdot p}^T L_{\cdot p} \end{bmatrix} \\ &= \begin{bmatrix} I_{p-l} & E_p/E_l r \\ E_p/E_l r^T & E_p^2/E_l^2 \end{bmatrix} \quad , \quad r = [L_{\cdot l} \cdots L_{\cdot p-1}]^T L_{\cdot p} \ . \end{aligned}$$

Nun wird die Cholesky-Zerlegung

$$F_{ll}^{-1}\hat{F} = \hat{R}^T \hat{R} \tag{2.48}$$

vorgenommen und das Produkt

$$\hat{L}\hat{E}\hat{T} \quad , \quad \hat{T} = \hat{R}^{-1} \oplus \hat{R}^T \tag{2.49}$$

gebildet. Mit der weiteren symplektischen Matrix

$$\hat{T}' = \mathrm{diag}(\underbrace{1, \ldots, 1}_{p-l}, E_l/E_p) \oplus \mathrm{diag}(\underbrace{1, \ldots, 1}_{p-l}, E_p/E_l)$$

haben wir

$$\hat{L}\hat{E}\hat{T} = \hat{L}\hat{E}\hat{T}\hat{E}^{-1}\hat{T}'^{-1}\hat{E}\hat{T}'$$

und das Produkt (2.49) kann durch die Multiplikation

$$\text{von} \quad \hat{L} \quad \text{mit} \quad \hat{E}\hat{T}\hat{E}^{-1}\hat{T}'^{-1} \tag{2.50}$$

und

$$\text{von} \quad \hat{E} \quad \text{mit} \quad \hat{T}' \tag{2.51}$$

ausgedrückt werden. $\hat{E}$ hat nun übereinstimmende führende Komponenten. Wir wenden uns nun einer Vereinfachung des Produktes (2.50) zu. Man kann leicht verifizieren, daß für $\hat{R}$ aus (2.48)

$$\hat{R} = \begin{bmatrix} I & x \\ 0 & \varrho \end{bmatrix} \quad , \quad \hat{R}^{-1} = \begin{bmatrix} I & -\varrho^{-1}x \\ 0 & \varrho^{-1} \end{bmatrix} \quad , \quad \varrho = (g - x^T x)^{1/2}$$

gilt. Daraus folgt

$$\hat{E}\hat{T}\hat{E}^{-1}\hat{T}'^{-1} = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \ ,$$

wobei mit $\tau = (1 - r^T r)^{1/2}$

$$\begin{aligned} R_1 &= \operatorname{diag}(\overbrace{E_l, \ldots, E_l}^{p-l}, E_p) \begin{bmatrix} I_{p-l} & -\varrho^{-1}x \\ 0 & \varrho^{-1} \end{bmatrix} \\ &\quad \cdot \operatorname{diag}(\underbrace{E_l^{-1}, \ldots, E_l^{-1}}_{p-l}, E_p^{-1}) \operatorname{diag}(\underbrace{1, \ldots, 1}_{p-l}, \frac{E_p}{E_l}) \\ &= \begin{bmatrix} I_{p-l} & -\varrho^{-1}x \\ 0 & E_p/(\varrho^{-1}E_l) \end{bmatrix} \\ &= \begin{bmatrix} I_{p-l} & -\tau^{-1}r \\ 0 & \tau^{-1} \end{bmatrix} \end{aligned} \tag{2.52}$$

und

$$\begin{aligned} R_2 &= \operatorname{diag}(E_l, \frac{E_{l+1}^2}{E_l}, \ldots, \frac{E_{p-1}^2}{E_l}, E_p) \begin{bmatrix} I_{p-l} & 0 \\ x^T & \varrho \end{bmatrix} \\ &\quad \cdot \operatorname{diag}(\frac{1}{E_l}, \frac{E_l}{E_{l+1}^2}, \ldots, \frac{E_l}{E_{p-1}^2}, \frac{1}{E_p}) \operatorname{diag}(1, \ldots, 1, \frac{E_l}{E_p}) \\ &= \begin{bmatrix} I_{p-l} & 0 \\ E_q(\operatorname{diag}(1/E_l, E_l/E_{l+1}^2, \ldots, E_l/E_{p-1}^2)x)^T & \varrho E_l/E_p \end{bmatrix} \\ &= \begin{bmatrix} I_{p-l} & 0 \\ (\operatorname{diag}(E_p^2/E_l^2, \ldots, E_p^2/E_{p-1}^2)r)^T & \tau \end{bmatrix} \end{aligned} \tag{2.53}$$

gilt. Die am Produkt (2.50) beteiligten Matrizen R_1, R_2 haben also eine recht einfache Gestalt. Auch das Produkt (2.51) ist einfach zu bilden, indem E_p durch E_l und E_{n+p} durch E_p^2/E_l ersetzt wird.

Die Skalierung mit $\hat{T}'$ hat nicht nur übereinstimmende führende Elemente von $\hat{E}$ zur Folge, sondern auch die Unabhängigkeit der Matrix R_1 von $\hat{E}$. Auch R_2 ist lediglich von den Quotienten der Elemente von $\hat{E}$ anhängig, die nach Voraussetzung nahe 1 sind.

Vorstehend erläuterte Operationen betrachten wir nun für alle p, $l < p \leq u$. Wir können alles in folgendem kurzen Algorithmus zusammenfassen:

FOR $p = l + 1$ TO u DO
  $r = [L_{\cdot l} \cdots L_{\cdot p-1}]^T L_{\cdot p}$
  $s = r^T r$
  $\tau = (1 - s)^{1/2}$
  $E' = \operatorname{diag}((E_p/E_l)^2, \ldots, (E_p/E_{p-1})^2)$
  $R_1 = \begin{bmatrix} I & -\tau^{-1}r \\ 0 & \tau^{-1} \end{bmatrix}$

$R_2 = \begin{bmatrix} I & 0 \\ (E'r)^T & \tau \end{bmatrix}$
$[L_{\cdot l} \cdots L_{\cdot p}] = [L_{\cdot l} \cdots L_{\cdot p}] R_1$
$[L_{\cdot n+l} \cdots L_{\cdot n+p}] = [L_{\cdot n+l} \cdots L_{\cdot n+p}] R_2$
ENDFOR
FOR $p = l + 1$ TO u DO
$E_p = E_l$
$E_{n+p} = E_{n+p}^2 / E_l$
ENDFOR

In einem Programm (siehe **2.3.1**, **2.3.2** von Verfahren 2.2.4) werden natürlich die Matrizen R_1, R_2 nicht explizit gebildet, weil sie nur in einer Zeile bzw. Spalte nichttrivial sind. Hier liegt gerade der Effekt der Speicherplatzersparnis gegenüber dem Verfahren 2.2.2. Zur Vermeidung von Variablenunter- und Überlauf werden noch kleine Modifikationen des obigen Programmstückes vorgenommen, die sämtlich dem vollständigen Algorithmus 2.2.4 zu entnehmen sind.

Jeder Cluster $K^{[l,u]}$, der Kandidat für die soeben geschilderte Transformation ist, genügt nach Voraussetzung der Ungleichung

$$\|K_S^{[l,u]} - I\|_E \leq \frac{1}{10(u-l+1)} \ ,$$

wenn exakte Arithmetik zugrunde gelegt wird. $\|F_S^{[l,u]} - I\|_E$ kann man auf dieselbe Weise abschätzen, so daß nach [17]

$$\|R^{[l,u]} - I\|_E \leq \frac{1}{10(u-l+1)} \quad , \quad F_S^{[l,u]} = R^{[l,u]^T} R^{[l,u]} \tag{2.54}$$

folgt. In vorstehendem Algorithmus wird $R^{[l,u]}$ jedoch nicht direkt ermittelt, sondern wegen (2.48) ist $R^{[l,u]}$ das Produkt verschiedener, sukzessive gebildeter Faktoren $\hat{R}$. Werden alle Rechnungen in Fließpunktarithmetik vorgenommen, so kann die Ungleichung (2.54) nur mit zusätzlichen (großen) Fehlertermen nachgewiesen werden, obwohl sie zumeist sogar exakt gilt. Die rechte Seite von (2.54) ist asymptotisch nämlich viel größer als die linke Seite. Weil wir die Gültigkeit von (2.54) für die Fehleranalyse benötigen, wird (2.54) bis auf einen kleinen relativen Fehler erzwungen, indem die Transformationen (2.38), (2.39) für den aktuellen Cluster unterbrochen werden, sobald (2.54) nicht mehr erfüllt sein sollte. Dies ist gleichbedeutend mit einer Verkleinerung des Clusters. Das Vorgehen fließt in **2.3.1** (Variable *sum*) von Algorithmus 2.2.4 ein, was wir kurz erläutern wollen. Wir führen für den Moment die Bezeichnungen

$$r^{[p]} = [L_{\cdot l} \cdots L_{\cdot p-1}]^T [L_{\cdot p}] \ \text{ und } \ \tau^{[p]} = (1 - r^{[p]^T} r^{[p]})^{1/2}$$

für $l+1 \le p \le \tilde{u} \le u$ ein. Dann gilt

$$\begin{aligned}
&\|R^{[l,\tilde{u}]} - I\|_E^2 \\
&= \|(\prod_{p=l+1}^{\tilde{u}} \begin{bmatrix} I_{p-1} & 0 & 0 \\ r^{[p]^T} & \tau^{[p]} & 0 \\ 0 & 0 & I \end{bmatrix}) - I\|_E^2 = \sum_{p=l+1}^{\tilde{u}} (r^{[p]^T} r^{[p]} + (1-\tau^{[p]})^2) \\
&= \sum_{p=l+1}^{\tilde{u}} (r^{[p]^T} r^{[p]} + 1 - 2\tau^{[p]} + 1 - r^{[p]^T} r^{[p]}) = 2 \sum_{p=l+1}^{\tilde{u}} \frac{1-\tau^{[p]^2}}{1+\tau^{[p]}} \\
&= 2 \sum_{p=l+1}^{\tilde{u}} \frac{r^{[p]^T} r^{[p]}}{1+\tau^{[p]}} \ .
\end{aligned}$$

Ein Vergleich mit dem Algorithmus auf Seite 61 zeigt, daß $sum = \|R^{[l,\tilde{u}]} - I\|_E^2$ leicht mit einem kleinen relativen Fehler durch Einfügen von

$sum = sum + 2s/(1+\tau)$

in die erste Schleife bestimmt werden kann (mit $sum = 0$ vor Schleifenbeginn). Gilt dann für ein $p < u$ bereits $sum > 1/(10(p-l+1))^2$, so wird $u = p$ gesetzt und die Schleife abgebrochen. Auf diese Weise können wir für jeden Cluster die Bedingung (2.54) bis auf einen kleinen relativen Fehler sicherstellen. Auf **2.3.1** von Verfahren 2.2.4 wird verwiesen. Man kann erwarten, daß die Schleife tatsächlich kaum jemals abgebrochen wird. Numerische Experimente bestätigen dies.

Anschließend betrachten wir die Transformation (2.39) an einem einzelnen Cluster

$$K^{[l,u]} = \begin{bmatrix} F^{[l,u]} & B^{[l,u]^T} \\ B^{[l,u]} & C^{[l,u]} \end{bmatrix} \quad , \quad F^{[l,u]} = F_{ll} \cdot I \tag{2.55}$$

der Ordnung $2(u-l+1)$. Die in (2.55) verwendete Notation wurde in (2.35) definiert. Zur Abkürzung verwenden wir wieder die Bezeichnung

$$l = low[clu] \quad , \quad u = upp[clu] \ .$$

Nachfolgend werden wir die erforderlichen Operationen zur Erzielung eines schiefsymmetrischen $B^{[l,u]}$ beschreiben. Nach Durchführung von **2.3.2** von Verfahren 2.2.4 sei

$$K^{[l,u]} = E^{[l,u]} L^{[l,u]^T} L^{[l,u]} E^{[l,u]}$$

mit

$$L^{[l,u]} = [L_{\cdot l} \cdots L_{\cdot u} \ L_{\cdot n+l} \cdots L_{\cdot n+u}] \quad , \quad E^{[l,u]} = E_l I_{u-l+1} \oplus \mathrm{diag}(E_{n+l}, \ldots, E_{n+u}) \ .$$

Nun wird die Matrix

$$T = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \quad , \quad X = -\frac{1}{2F_{ll}} (B^{[l,u]} + B^{[l,u]^T}) \tag{2.56}$$

gebildet und $L^{[l,u]}E^{[l,u]}$ von rechts mit T multipliziert, was wegen $L^{[l,u]}E^{[l,u]}T = L^{[l,u]}(E^{[l,u]}TE^{[l,u]^{-1}})E^{[l,u]}$ gleichbedeutend ist mit der Multiplikation

$$\text{von} \quad L^{[l,u]} \quad \text{mit} \quad T' = E^{[l,u]}TE^{[l,u]^{-1}} \ , \tag{2.57}$$

wobei $E^{[l,u]}$ unverändert bestehen bleibt. Es gilt

$$T' = E^{[l,u]}TE^{[l,u]^{-1}} = \begin{bmatrix} I & U \\ 0 & I \end{bmatrix} \quad , \quad U = E_l X \mathrm{diag}(\frac{1}{E_{n+l}}, \ldots, \frac{1}{E_{n+u}}) \ , \tag{2.58}$$

d.h.

$$\begin{aligned} U_{qp} &= -\frac{1}{2E_l E_{n+p}}(B^{[l,u]}_{qp} + B^{[l,u]}_{pq}) \\ &= -\frac{1}{2E_l E_{n+p}}(E_{n+q}[L_{\cdot n+q}]^T[L_{\cdot p}]E_l + E_{n+p}[L_{\cdot n+p}]^T[L_{\cdot q}]E_l) \\ &= -\frac{1}{2}(\frac{E_{n+q}}{E_{n+p}}[L_{\cdot n+q}]^T[L_{\cdot p}] + [L_{\cdot n+p}]^T[L_{\cdot q}]) \end{aligned} \tag{2.59}$$

Bei der Formulierung eines Algorithmus für die Transformationen (2.39) gehen wir methodisch etwas anders als bei (2.38) vor. Führte man nämlich Transformationen (2.57) sukzessive für Hauptuntermatrizen der Ordnung $p-l+1$, $p = l \ldots u$, von U durch, so hätten bereits transformierte Spalten $[L_{\cdot n+q}]$, $l \leq q \leq p$, Einfluß auf alle nachfolgenden Transformationsmatrizen. Man kann zeigen, daß in diesem Fall die Normen $\|T - I\|_2$ der in Fließpunktarithmetik berechneten T und damit auch die Normen $\|[L_{\cdot n+p}]\|_2$ nur dann klein genug bleiben, wenn $\|K_S^{[l,u]} - I\|_2$ bereits für die Praxis unrealistisch klein ist. Außerdem wäre die in endlicher Arithmetik bestimmte Matrix X aus (2.58) nicht symmetrisch.

Das Hilfsmittel des Abbruches der Transformationen (also der Verkleinerung des Clusters), falls $\|T - I\|_2$ zu groß wird, steht uns hier im Gegensatz zum Vorgehen bei den Transformationen (2.38) nicht zur Verfügung, weil dann die Konvergenz des Verfahrens nicht sichergestellt werden könnte.

Wir beschreiten also einen anderen Weg. Für $p = l \ldots u$ werden sämtliche Elemente U_{qp} und U_{pq}, $q = p \ldots u$, gemäß (2.59) berechnet und anschließend wird mit

$$\begin{bmatrix} I & [U_{\cdot p} \ \overbrace{0 \cdots 0}^{u\text{-}p}] \\ 0 & I \end{bmatrix}$$

transformiert. Die bei dieser Transformation noch nicht benötigten Elemente U_{pq}, $q > p$, werden zunächst zwischengespeichert und erst bei späteren Transformationen verwendet. Die Transformation (2.39) können wir nun in folgendem kurzen Stück Algorithmus zusammenfassen:

FOR $p = l$ TO u DO
  FOR $q = l$ TO $p - 1$ DO $u_q = U_{qp}$
  $u_p = -[L_{\cdot p}]^T [L_{\cdot n+p}]$
  FOR $q = p + 1$ TO u DO
    $y = [L_{\cdot n+q}]^T [L_{\cdot p}]$
    $z = [L_{\cdot q}]^T [L_{\cdot n+p}]$
    $u_q = -(y(E_{n+q}/E_{n+p}) + z)/2$
    $U_{pq} = -(y + z(E_{n+p}/E_{n+q}))/2$
  ENDFOR
  $[L_{\cdot n+p}] = [L_{\cdot n+p}] + \sum_{q=l}^{u} u_q [L_{\cdot q}]$
ENDFOR

Für das im vorstehenden Algorithmus zur Zwischenspeicherung verwendete Feld U benötigt man höchstens n^2 Speicherplätze. Das Feld ist jedoch während des gesamten Vorganges nur teilweise besetzt. Man kann zeigen, daß man bei Matrizen der vollen Ordnung $2n$ unter Verwendung dynamischer Speicherverwaltung mit etwa $n^2/4$ Speicherplätzen auskommt.

Das geschilderte Vorgehen hat den Vorteil, daß sämtliche Elemente U_{qp} mit Daten berechnet werden, wie sie zu Beginn des Transformationsvorganges bestehen. Bei der Fehleranalyse werden wir diese Tatsache mit Gewinn ausnutzen. Formell wird wie in **2.3.3**, **2.3.4** von Verfahren 2.2.4 beschrieben, vorgegangen. Auch die Fehleranalyse bezieht sich darauf.

Abschluß eines Zyklus

Zum Abschluß eines jeden Zyklus werden die Spalten von L nochmals auf eins normiert und E entsprechend skaliert, damit das Kriterium für die Beendigung des Verfahrens und die Bedingungen für einen modifizierten Zyklus leichter geprüft werden können (**3.7** von Verfahren 2.2.4).

2.2.3.3 Beendigung des Verfahrens

Das Verfahren wird beendet, wenn nach einem Zyklus das *Konvergenzkriterium*

$$|\sum_{k=1}^{2n} L_{ki} L_{kj}| < TOL \quad , \quad 1 \leq i, j \leq 2n \ , \quad i < j \tag{2.60}$$

mit einem geigneten TOL, z.B. $TOL = 2n \cdot \varepsilon$ mit der Maschinengenauigkeit ε, erfüllt ist. Die E_{ii}^2 sind dann Näherungen an die positiven Imaginärteile der Eigenwerte von S gemäß (2.41) und die Spalten von L sind Näherungen an die Eigenvektoren.

Wir verwenden also eine Abbruchbedingung wie in [13]. Sie ist konsistent mit Satz 1.0.6, wonach die relative Genauigkeit der Eigenwerte nicht von der

Kondition $\kappa(K)$, sondern von der Kondition der skalierten Matrix $\kappa(K_S)$ abhängt. Das Verfahren ist also dann abzubrechen, wenn die Außerdiagonalelemente der skalierten Matrix klein genug sind.

2.2.4 Der formale Algorithmus des impliziten Verfahrens

Hier geben wir nun das formelle Vorgehen beim impliziten Verfahren zur Lösung des Eigenwertproblems $(L^T L, J)$ an.

/* The most important variables we use in the algorithm are stated as follows

name	**type**	**usage**	**comment**
TOL	scalar	input	stopping bound with $0 < TOL \ll 1$, e.g $TOL = 2n \cdot$ *macheps*, where *macheps* is the machine precision.
$\sigma_{\max}$	scalar	input	determines the minimal increase of the Hadamard measure in modified cycles, e.g. $\sigma_{\max} = -\ TOL^2/10$.
E	vector	output	diagonal matrix stored as vector. Initially, $E_i = 1$, $i = 1 \ldots 2n$.
L	matrix	input/ output	on input, $K = \mathrm{diag}(E) L^T L \mathrm{diag}(E)$ is the positive definite matrix to be diagonalised; on output, L holds the eigenvectors of the pair (K, J). We write $$K = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix}, \quad F, B, C \in I\!R^{n \times n} .$$
ν	vector	intermediate	each component ν_i holds the square of the 2-norm of the i-th column of L with a small relative error. ν is used to reduce the computational count.
T	matrix	intermediate	transformation matrix; T is the identity matrix except at the positions of a submatrix $\hat{T}$. In a computer program only components of $\hat{T}$ are stored.
U	matrix	intermediate	$U \in I\!R^{n \times n}$ is used to store a scaled matrix X, where $\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$ is the second transformation in modified cycles. */

/* 1. preparation */
/* 1.1 scale L, compute and scale E for $F_{pp} = C_{pp}, p = 1 \ldots n$ */
/* 1.1.1 */

FOR $p = 1$ TO $2n$ DO $E_p = (\sum_{k=1}^{2n} L_{kp}^2)^{1/2}$
/* 1.1.2 */
FOR $p = 1$ TO $2n$ DO $[L_{\cdot p}] = (E_p^{-1}) \cdot [L_{\cdot p}]$
/* 1.1.3 */
FOR $p = 1$ TO n DO $E_p = E_{n+p} = E_p^{1/2} E_{n+p}^{1/2}$
FOR $p = 1$ TO $2n$ DO $\nu_p = 1$
/* begin cycle */
WHILE $\max_{1 \le i \le 2n, 1 \le j \le i-1} |\sum_{k=1}^{2n} L_{ki} L_{kj}| \ge TOL$ DO
  FOR $p = 1$ TO n DO
    IF $|\sum_{k=1}^{2n} L_{kp} L_{k,n+p}| \ge TOL$ THEN
      /* 1.2 annihilate diagonal entries of B, rescale L */
      /* 1.2.1 */
      $[L_{\cdot p} \;\; L_{\cdot n+p}] = [L_{\cdot p} - L_{\cdot n+p} \;\; L_{\cdot p} + L_{\cdot n+p}]$
      /* 1.2.2 */
      $(E_p, E_{n+p}) = (1/\sqrt{2}) \cdot (E_p, E_{n+p})$
      /* 1.2.3 */
      $\delta_1 = (\sum_{k=1}^{2n} L_{kp}^2)^{1/2}$
      $\delta_2 = (\sum_{k=1}^{2n} L_{k,n+p}^2)^{1/2}$
      $(E_p, E_{n+p}) = (\delta_1 E_p, \delta_2 E_{n+p})$
      /* 1.2.4 */
      $[L_{\cdot p}] = (\delta_1^{-1}) \cdot [L_{\cdot p}]$
      $[L_{\cdot n+p}] = (\delta_2^{-1}) \cdot [L_{\cdot n+p}]$
      /* 1.3 rescale E for $F_{pp} = C_{pp}$ */
      $E_p = E_{n+p} = E_p^{1/2} E_{n+p}^{1/2}$
    ENDIF
  ENDFOR $\{p\}$

  /* now we have $K = EL^T LE = \begin{bmatrix} F & B^T \\ B & C \end{bmatrix}$ with $B_{pp} = 0$ and
    $F_{pp} = C_{pp}, p = 1 \ldots n$, and $\|L_{\cdot p}\|_2 = 1, p = 1 \ldots n$ */

  /* 2.1 prepare for modification */
  set $L = LP, E = P^T EP$ with symplectic permutation matrix $P = \hat{P} \oplus \hat{P}$,
    so that $E_p \ge E_{p+1}$, $p = 1 \ldots n-1$

  /* 2.2 determine clusters */
  $num = 1$ /* number of clusters */
  $l = 1$ /* lower index of first cluster */
A: IF $l + 1 < n$ THEN /* begin new cluster */
    $\bar{\sigma} = \sum_{k=1}^{2n} L_{k,l+1} L_{kl}$
    $\hat{\sigma} = \sum_{k=1}^{2n} L_{k,n+l+1} L_{k,n+l}$
    $\bar{\beta} = \sum_{k=1}^{2n} L_{k,n+l+1} L_{kl}$

```
      β̂ = Σ_{k=1}^{2n} L_{k,n+l} L_{k,l+1}
      σ = 2σ̄σ̂ − (β̄ + β̂)²/2
      ω̄ = σ̄² + σ̂² + β̄² + β̂²
   ENDIF
   u = l + 2                                  /* upper index of actual cluster */
B: IF u > n THEN
      num=num−1                                                    /* finish */
   ELSE
      FOR i = l TO u − 1 DO
         σ̄ = Σ_{k=1}^{2n} L_{ku} L_{ki}
         σ̂ = Σ_{k=1}^{2n} L_{k,n+u} L_{k,n+i}
         β̄ = Σ_{k=1}^{2n} L_{k,n+u} L_{ki}
         β̂ = Σ_{k=1}^{2n} L_{k,n+i} L_{ku}
         σ = σ + 2σ̄σ̂ − (β̄ + β̂)²/2
         ω̄ = ω̄ + σ̄² + σ̂² + β̄² + β̂²
      ENDFOR {i}
      ω = (2 ω̄)^{1/2}
      IF (10(u − l + 1)) ω ≤ 1 ∧ E_l ≤ √(1 + ω) E_u ∧ σ + 12 ω³ ≤ σ_max THEN
         IF u = n THEN                                /* finish last cluster */
            low[num]= l
            upp[num]= u
         ELSE                          /* increase upper index of actual cluster */
            u = u + 1
            GOTO B:
         ENDIF
      ELSE
         IF u = l + 2 THEN       /* restart cluster with increased lower index */
            l = l + 1
            GOTO A:
         ELSE                                          /* begin next cluster */
            low[num]= l
            upp[num]= u − 1
            num=num+1
            l = u
            GOTO A:
         ENDIF
      ENDIF
   ENDIF
    /* the extremal indices of clusters are now stored in integer vectors
       "low" and "upp" of length "num" */

   /* 2.3 eventual modified cycle with rescaling */
```

```
FOR clu = 1 TO num DO

    /* 2.3.1 make part of first transformation matrix */
    sum = 0
    FOR p = low[clu] + 1 TO upp[clu] DO
        FOR q = low[clu] TO p − 1 DO r_q = Σ_{k=1}^{2n} L_kq L_kp
        s = Σ_{q=low[clu]}^{p−1} r_q^2
        τ = (1 − s)^{1/2}
        sum = sum +2s/(1 + τ)
        IF sum > 1/(10(p − low[clu] + 1))^2 THEN
            upp[clu] = p − 1, GOTO C:      /* break to guarantee error anal. */
        ENDIF

        /* 2.3.2 apply part of first transformation */
        FOR q = low[clu] TO p − 1 DO
            IF r_q ≠ 0 THEN [L_·p] = [L_·p] − r_q · [L_·q]
        ENDFOR {q}
        IF τ ≠ 1 THEN [L_·p] = τ^{−1}[L_·p]
        FOR q = low[clu] TO p − 1 DO
            IF r_q ≠ 0 THEN [L_·n+q] = [L_·n+q] + r_q(E_p/E_q)^2 · [L_·n+p]
        ENDFOR {q}
        [L_·n+p] = τ[L_·n+p]
    ENDFOR {p}
C:  FOR p = low[clu] + 1 TO upp[clu] DO
        E_p = E_low[clu]
        E_n+p = E_n+p(E_n+p/E_low[clu])
    ENDFOR {p}

    /* 2.3.3 make part of second transformation matrix */
    FOR p = low[clu] TO upp[clu] DO
        FOR q = low[clu] TO p − 1 DO
            u_q = U_qp              /* U holds parameters for later transformation */
        ENDFOR {q}
        u_p = − Σ_{k=1}^{2n} L_kp L_k,n+p
        FOR q = p + 1 TO upp[clu] DO
            y = Σ_{k=1}^{2n} L_k,n+q L_kp
            z = Σ_{k=1}^{2n} L_kq L_k,n+p
            u_q = −(y(E_n+q/E_n+p) + z)/2
            U_pq = −(y + z(E_n+p/E_n+q))/2
        ENDFOR {q}

        /* 2.3.4 apply part of second transformation */
```

```
      FOR q = low[clu] TO upp[clu] DO
         IF u_q ≠ 0 THEN [L_{·n+p}] = [L_{·n+p}] + u_q[L_{·q}]
      ENDFOR {q}
   ENDFOR {p}
ENDFOR {clu}

/* 2.3.5 scale E for F_pp = C_pp within clusters */
FOR clu = 1 TO num DO
   FOR p = low[clu] TO upp[clu] DO
      δ_1 = Σ_{k=1}^{2n} L_{kp}^2
      δ_2 = Σ_{k=1}^{2n} L_{k,n+p}^2
      (ν_p, ν_{n+p}) = (δ_1, δ_2)
      δ = δ_1^{1/4}/δ_2^{1/4}
      (E_p, E_{n+p}) = (E_p^{1/2} E_{n+p}^{1/2}/δ, E_p^{1/2} E_{n+p}^{1/2} δ)
   ENDFOR {q}
ENDFOR {clu}

/* 3. begin cyclic rotations */
/* Now we use variables f_pp, f_pq, f_qq, b_pq, b_qp, c_pp, c_pq, c_qq. These variables
   are not components of arrays but scalars, i.e. they should be understood
   as fpp, fpq, etc. . The variables are scaled entries of the matrix K;
   we have F_pp = E_p^2 f_pp, F_pq = E_p E_q f_pq etc. */
FOR p = 1 TO n − 1 DO
   FOR q = p + 1 TO n DO
      rotate= TRUE
      FOR clu = 1 TO num DO
         IF low[clu] ≤ p, q ≤upp[clu] THEN rotate= FALSE
      ENDFOR {clu}
      IF rotate THEN
         f_pp = ν_p
         f_qq = ν_q
         reperm = FALSE

         /* 3.1 annihilate B_pq and B_qp */
         b_pq = Σ_{k=1}^{2n} L_{k,n+p} L_{kq}
         b_qp = Σ_{k=1}^{2n} L_{k,n+q} L_{kp}
         IF (|b_pq| E_{n+p}/(E_p f_pp^{1/2} f_qq^{1/2}) ≥ TOL)
            ∨ (|b_qp| E_{n+q}/(E_q f_pp^{1/2} f_qq^{1/2}) ≥ TOL) THEN
            /* compute parameters for SVD decomposition of A as follows */
            f_pq = Σ_{k=1}^{2n} L_{kp} L_{kq}
            c_pq = Σ_{k=1}^{2n} L_{k,n+p} L_{k,n+q}
```

$A = \frac{1}{(f_{pp}^{1/2} f_{qq}^{1/2})} \begin{bmatrix} f_{pq} & b_{qp} E_{n+q}/E_q \\ b_{pq} E_{n+p}/E_p & c_{pq}(E_{n+p} E_{n+q})/(E_p E_q) \end{bmatrix}$
$a = A_{21}^2 + A_{22}^2$
$b = A_{11}^2 + A_{12}^2$
$c = A_{11} A_{21} + A_{12} A_{22}$
IF $c = 0$ THEN
  $t_p = 0$
ELSE
  $\vartheta = (b - a)/(2\,c)$
  IF $|\vartheta| > macheps^{-1/2}$ THEN
    $t_p = 1/(2\,|\vartheta|)$
  ELSE
    $t_p = 1/(|\vartheta| + \sqrt{1 + \vartheta^2})$
  ENDIF
  IF $\vartheta < 0$ THEN $t_p = -t_p$
ENDIF
$h_1 = A_{11} - A_{21} t_p$
$h_2 = A_{22} + A_{12} t_p$
$g_1 = A_{22} t_p - A_{12}$
$g_2 = A_{21} + A_{11} t_p$
IF $|h_1| > |h_2| \ \vee \ (|h_1| = |h_2| \ \wedge \ |g_1| \geq |g_2|)$ THEN
  $g = g_1$
  $h = h_1$
ELSE
  $g = g_2$
  $h = h_2$
ENDIF
$perm = (g \neq 0 \ \wedge \ h = 0)$
IF $g = 0 \ \vee h = 0$ THEN $t_q = 0$ ELSE $t_q = g/h$
FOR $i \in \{p, q\}$ DO
  IF $t_i \neq 0$ THEN
    IF $|t_i| > macheps^{-1/2}$ THEN
      $w_i = |t_i|$
    ELSE
      $w_i = \sqrt{1 + t_i^2}$
    ENDIF
    IF $|t_i| \leq 1$ THEN
      $\begin{bmatrix} T_{ii} & T_{i,n+i} \\ T_{n+i,i} & T_{n+i,n+i} \end{bmatrix} = \begin{bmatrix} 1 & t_i(E_i/E_{n+i}) \\ -t_i(E_{n+i}/E_i) & 1 \end{bmatrix}$
      $f_{ii} = f_{ii} \cdot w_i^2$
      $(E_i, E_{n+i}) = (E_i, E_{n+i})/w_i$
    ELSE

$$\begin{bmatrix} T_{ii} & T_{i,n+i} \\ T_{n+i,i} & T_{n+i,n+i} \end{bmatrix} = \begin{bmatrix} (E_i/E_{n+i})/|t_i| & \operatorname{sign}(t_i) \\ -\operatorname{sign}(t_i) & (E_{n+i}/E_i)/|t_i| \end{bmatrix}$$

$f_{ii} = f_{ii}(E_i/E_{n+i})^2(w_i/t_i)^2$
$(E_i, E_{n+i}) = |t_i/w_i| \cdot (E_{n+i}, E_i)$
$reperm = (i = q)$
ENDIF
ENDIF
$[L_{\cdot i} \;\; L_{\cdot n+i}] = [L_{\cdot i} \;\; L_{\cdot n+i}] \begin{bmatrix} T_{ii} & T_{i,n+i} \\ T_{n+i,i} & T_{n+i,n+i} \end{bmatrix}$
ENDFOR $\{i\}$
IF *perm* THEN
$[L_{\cdot q} \;\; L_{\cdot n+q}] = [-L_{\cdot n+q} \;\; L_{\cdot q}]$
$f_{qq} = E_q^2 f_{qq}/E_{n+q}^2$
$(E_q, E_{n+q}) = (E_{n+q}, E_q)$
$(\nu_q, \nu_{n+q}) = (\nu_{n+q}, \nu_q)$
ENDIF
ENDIF

$f_{pq} = \sum_{k=1}^{2n} L_{kp} L_{kq}$
$rotF = (|f_{pq}|/(f_{pp}^{1/2} f_{qq}^{1/2}) \geq TOL)$
IF *rotF* THEN

/* 3.2 scale E for $C_{pp} = C_{qq} = 1$ */
$\delta_1 = f_{pp}^{1/2}(E_p/E_{n+p})$
$\delta_2 = f_{qq}^{1/2}(E_q/E_{n+q})$
$(E_p, E_q, E_{n+p}, E_{n+q}) = (E_p E_{n+p}\delta_1, E_q E_{n+q}\delta_2, \delta_1^{-1}, \delta_2^{-1})$

/* 3.3 annihilate F_{qp} and F_{pq} */
$a = E_p^2 f_{pp}$
$b = E_q^2 f_{qq}$
$c = E_p E_q f_{pq}$
$\vartheta = (b - a)/(2\,c)$
IF $|\vartheta| > macheps^{-1/2}$ THEN
$t = 1/(2|\vartheta|)$
ELSE
$t = 1/(|\vartheta| + \sqrt{1 + \vartheta^2})$
ENDIF
IF $(c \geq 0 \;\wedge\; b < a) \;\vee\; (c < 0 \;\wedge\; b > a)$ THEN $t = -t$
$T = \oplus_{i=0}^{1} \begin{bmatrix} 1 & t(E_{p+i\cdot n}/E_{q+i\cdot n}) \\ -t(E_{q+i\cdot n}/E_{p+i\cdot n}) & 1 \end{bmatrix}$
$[L_{\cdot p} \;\; L_{\cdot q} \;\; L_{\cdot n+p} \;\; L_{\cdot n+q}] = [L_{\cdot p} \;\; L_{\cdot q} \;\; L_{\cdot n+p} \;\; L_{\cdot n+q}] \cdot T$
$(E_p, E_q, E_{n+p}, E_{n+q}) = (E_p, E_q, E_{n+p}, E_{n+q})/\sqrt{1 + t^2}$

/* compute new scaled elements of C */
/* we have $c_{pp}E_{n+p}^2 + c_{qq}E_{n+q}^2 = 2$, and we can compute
c_{pp}, c_{qq} with low relative errors as follows */
$c_{pq} = \sum_{k=1}^{2n} L_{k,n+p}L_{k,n+q}$
IF $t \cdot c_{pq} \geq 0$ THEN
$c_{pp} = \sum_{k=1}^{2n} L_{k,n+p}^2$
$c_{qq} = (2 - c_{pp}E_{n+p}^2)/E_{n+q}^2$
ELSE
$c_{qq} = \sum_{k=1}^{2n} L_{k,n+q}^2$
$c_{pp} = (2 - c_{qq}E_{n+q}^2)/E_{n+p}^2$
ENDIF

ELSE
/* compute scaled elements of C */
$c_{pq} = \sum_{k=1}^{2n} L_{k,n+p}L_{k,n+q}$
$c_{pp} = f_{pp}(E_p/E_{n+p})^2$
$c_{qq} = f_{qq}(E_q/E_{n+q})^2$
ENDIF

$rotC = (|c_{pq}|/(c_{pp}^{1/2}c_{qq}^{1/2}) \geq TOL)$
IF $rotC$ THEN

/* 3.4 scale E for $F_{pp} = F_{qq} = 1$ */
IF $rotF$ THEN
$\delta_1 = (\sum_{k=1}^{2n} L_{kp}^2)^{1/2}$
$\delta_2 = (\sum_{k=1}^{2n} L_{kq}^2)^{1/2}$
ELSE
$\delta_1 = f_{pp}^{1/2}$
$\delta_2 = f_{qq}^{1/2}$
ENDIF
$(E_p, E_q, E_{n+p}, E_{n+q}) = (\delta_1^{-1}, \delta_2^{-1}, E_pE_{n+p}\delta_1, E_qE_{n+q}\delta_2)$

/* 3.5 annihilate C_{qp} and C_{pq} */
$a = E_{n+p}^2 c_{pp}$
$b = E_{n+q}^2 c_{qq}$
$c = E_{n+p}E_{n+q}c_{pq}$
$\vartheta = (b - a)/(2\,c)$
IF $|\vartheta| > macheps^{-1/2}$ THEN
$t = 1/(2|\vartheta|)$
ELSE
$t = 1/(|\vartheta| + \sqrt{1 + \vartheta^2})$

ENDIF
IF $(c \geq 0 \ \wedge \ b < a) \ \vee \ (c < 0 \ \wedge \ b > a)$ THEN $t = -t$
$T = \oplus_{i=0}^{1} \begin{bmatrix} 1 & t(E_{p+i\cdot n}/E_{q+i\cdot n}) \\ -t(E_{q+i\cdot n}/E_{p+i\cdot n}) & 1 \end{bmatrix}$
$[L_{\cdot p} \ \ L_{\cdot q} \ \ L_{\cdot n+p} \ \ L_{\cdot n+q}] = [L_{\cdot p} \ \ L_{\cdot q} \ \ L_{\cdot n+p} \ \ L_{\cdot n+q}] \cdot T$
$(E_p, E_q, E_{n+p}, E_{n+q}) = (E_p, E_q, E_{n+p}, E_{n+q})/\sqrt{1+t^2}$
ENDIF

/* **3.6 scale** E **for** $F_{pp} = C_{pp}$ **and** $F_{qq} = C_{qq}$**, repermute** */
IF $rotC$ THEN
$\delta_1 = 1/E_p^2$
$\delta_2 = \sum_{k=1}^{2n} L_{k,n+p}^2$
$\delta_3 = 1/E_q^2$
$\delta_4 = \sum_{k=1}^{2n} L_{k,n+q}^2$
ELSE
$\delta_2 = c_{pp}$
$\delta_4 = c_{qq}$
IF $rotF$ THEN
$\delta_1 = \sum_{k=1}^{2n} L_{kp}^2$
$\delta_3 = \sum_{k=1}^{2n} L_{kq}^2$
ELSE
$\delta_1 = f_{pp}$
$\delta_3 = f_{qq}$
ENDIF
ENDIF
$(E_p, E_{n+p}) = (E_p^{1/2} E_{n+p}^{1/2}) \cdot (\delta_2^{1/4}/\delta_1^{1/4}, \delta_1^{1/4}/\delta_2^{1/4})$
$(E_q, E_{n+q}) = (E_q^{1/2} E_{n+q}^{1/2}) \cdot (\delta_4^{1/4}/\delta_3^{1/4}, \delta_3^{1/4}/\delta_4^{1/4})$
IF $reperm$ THEN
$[L_{\cdot q} \ \ L_{\cdot n+q}] = [-L_{\cdot n+q} \ \ L_{\cdot q}]$
$(E_q, E_{n+q}) = (E_{n+q}, E_q)$
$(\nu_p, \nu_q, \nu_{n+p}, \nu_{n+q}) = (\delta_1, \delta_4, \delta_2, \delta_3)$
ELSE
$(\nu_p, \nu_q, \nu_{n+p}, \nu_{n+q}) = (\delta_1, \delta_3, \delta_2, \delta_4)$
ENDIF
ENDIF
ENDFOR $\{q\}$
ENDFOR $\{p\}$

/* **3.7 rescale** L**, scale** E **for** $F_{pp} = C_{pp}, p = 1 \ldots n$ */
FOR $p = 1$ TO $2n$ DO
$E_p = \nu_p^{1/2} E_p$
$[L_{\cdot p}] = (\nu_p^{-1/2})[L_{\cdot p}]$

$\nu_p = 1$
ENDFOR $\{p\}$
FOR $p = 1$ TO n DO $E_p = E_{n+p} = E_p^{1/2} E_{n+p}^{1/2}$
ENDWHILE

/* 4. finish */
/* 4.1 compute imaginary parts of eigenvalues */
FOR $p = 1$ TO n DO $E_p = E_{n+p} = E_p^2$
/* The positive imaginary parts of the eigenvalues of LJL^T are now stored in $E_p, p = 1 \ldots n$, while the eigenvectors are stored in L */

2.2.5 Bemerkungen zum impliziten Verfahren

Die Verwendung der schnellen Rotationen (2.45), (2.46) macht es erforderlich, die Diagonalmatrix E im Zuge jeder Rotation mit einem sn oder cs, $1/\sqrt{2} \leq |sn|, |cs| \leq 1$ zu multiplizieren. Dies kann insbesondere bei großen Matrizen oder bei Matrizen mit sehr großen oder sehr kleinen Eigenwerten zu Unter- oder -überlauf führen. Ein Strategie, dies zu verhindern, wird in [1] dargestellt, in den hier vorgestellten Algorithmen jedoch nicht verwendet.

Bei der Prüfung auf Cluster werden Normen von Submatrizen berechnet. Werden dann aber keine Cluster festgestellt, so sind diese Berechnungen vergeblich, weil die Ergebnisse nicht mehr benötigt werden. Weil diese überflüssigen Berechnungen lediglich $\mathcal{O}(n^2)$ Operationen je Zyklus benötigen, sind sie nur von geringer Bedeutung.

Anders ist es mit der Bestimmung der Matrixelemente $L_{\cdot i}^T L_{\cdot j}$ innerhalb der Cluster. Diese werden nämlich zum Teil mehrfach berechnet: erstmals bei der Bestimmung der Clustergröße und ein zweites Mal bei der Durchführung der Transformation (2.38). Man kann dies mit folgender Änderung des Algorithmus' vermeiden: bei der Bestimmung der Clustergröße wird die Ordnung von Submatrizen von K, die die Bedingungen 1. bis 3. von Seite 57 erfüllen, sukzessive um 1 erhöht. Schließt sich an die Erhöhung der Ordnung unmittelbar die Transformation (2.38) für eine Zeile und Spalte an, so können die soeben berechneten Elemente $L_{\cdot i}^T L_{\cdot j}$ auch für die Bestimmung der Transformationsmatrix herangezogen werden. Darauf wurde aber verzichtet, um die Transparenz des Verfahrens zu erhalten.

Ist T die Transformationsmatrix gemäß (2.39), so wird L von rechts mit

$$T' = \begin{bmatrix} I & U \\ 0 & I \end{bmatrix} = ETE^{-1}$$

transformiert. Im Teil **2.3.3**, **2.3.4** von Verfahren 2.2.4 werden die Elemente von U oberhalb der Diagonalen nach ihrer Berechnung zunächst zwischengespeichert,

weil sie erst später zur Transformation benötigt werden. Bei der hier angewendeten Methode werden dafür n^2 Speicherplätze verwendet. Notwendig wären jedoch lediglich $n^2/4$ Speicherplätze, falls $n = 2k, k \in I\!N$ ist bzw. $(n^2-1)/4$ Speicherplätze, falls $n = 2k+1, k \in I\!N$ die halbe Ordnung von L ist. Auf die Verwendung der speichersparenden Methode wurde zur Erhaltung der Transparenz des Verfahrens gleichfalls verzichtet.

Der soeben angesprochene Teil **2.3.3**, **2.3.4** von Verfahren 2.2.4 kann ohne Verwendung zusätzlichen Speichers in der Größenordnung $\mathcal{O}(n^2)$ alternativ auch so formuliert werden:

/* 2.3.3 make part of second transformation matrix */
FOR $p =$*low*[*clu*] TO $p =$*upp*[*clu*] DO
  FOR $q =$*low*[*clu*] TO $p-1$ DO
    $y = \sum_{k=1}^{2n} L_{k,n+q} L_{kp}$
    $z = \sum_{k=1}^{2n} L_{kq} L_{k,n+p}$
    $u_q = -(y(E_{n+q}/E_{n+p}) + z)/2$
    $v_q = -(y + z(E_{n+p}/E_{n+q}))/2$
  ENDFOR
  $u_p = -\sum_{k=1}^{2n} L_{kp} L_{k,n+p}$

  /* 2.3.4 apply part of second transformation */
  FOR $q =$*low*[*clu*] TO p DO
    set $[L_{\cdot n+q}] = [L_{\cdot n+q}] + u_q[L_{\cdot p}]$
  ENDFOR
  FOR $q =$*low*[*clu*] TO $p-1$ DO
    set $[L_{\cdot n+p}] = [L_{\cdot n+p}] + v_q[L_{\cdot q}]$
  ENDFOR
ENDFOR

Bei dieser Methode werden bereits transformierte Spalten von L im weiteren Verlauf wieder zur Bestimmung von Transformationsparametern herangezogen. Die Änderung der Normen der Spalten von L kann in diesem Fall nur unzureichend kontrolliert werden, was die Fehleralanyse wesentlich erschwert. Auf diese Methode wurde daher verzichtet. Die in **2.3.3**, **2.3.4** zu Anwendung gekommene Methode hat zwar den Nachteil, zusätzlich $\mathcal{O}(n^2)$ Speicherplätze zu benötigen, jedoch den Vorteil, daß alle Informationen zur Bestimmung der Transformationsparameter bereits zu Beginn der Transformation zur Verfügung stehen.

Wir können noch festhalten, daß bei einer Matrix der Ordnung $2n$ für einen normalen Zyklus ohne Prüfung der Konvergenzbedingung im Wesentlichen maximal $23\,n^3$ Multiplikationen benötigt werden. Das Verfahren von Paardekooper [25] benötigt unter Verwendung schneller Rotationen hingegen $16n^3$ Multiplikationen, wenn auch die Eigenvektoren berechnet werden. Numerische Experimente zeigen, daß das Verfahren von Paardekooper dennoch zumeist um etwa Faktor 10 langsa-

mer ist, als das implizite Verfahren. Eine Ursache dafür ist, daß beim impliziten Verfahren auf Rotationen verzichtet werden kann, wenn Außerdiagonalelemente bereits im relativen Sinne klein sind. Dies ist beim Paardekooper-Verfahren nicht der Fall (siehe auch Abschnitt 6.2). Außerdem kommt dem neuen Verfahren die in FORTRAN besonders günstige spaltenweise Verarbeitung zugute. Zum Vergleich sei angeführt, daß das in [31] beschriebene (und in [32] nochmals dargestellte) implizite Verfahren für jeden Zyklus $20\,n^3$ Multiplikationen benötigt, wenn schnelle Rotationen verwendet werden und die Diagonalelemente stets durch Skalarprodukte bestimmt werden.

2.3 Konvergenz

In diesem Abschnitt widmen wir uns der Konvergenz der vorgestellten Jacobiähnlichen Verfahren. Die Verfahren sind hinsichtlich ihres Konvergenzverhaltens äquivalent, so daß wir uns auf das explizite Verfahren beschränken können. Als Hilfsmittel werden wir das Hadamardsche Maß verwenden, das wir im nächsten Unterabschnitt erführen werden. Wir werden sehen, daß das explizite Verfahren in exakter Arithmetik für beliebige symmetrische positiv definite Eingangsmatrizen K terminiert, also die Folge der transformierten Matrizen K stets gegen eine Diagonalmatrix konvergiert.

2.3.1 Über das Hadamardsche Maß

In diesem Unterabschnitt werden wir positiv definite Matrizen einer gewissen positiven reellen Zahl zuordnen. Diese Zahl nennen wir Hadamardsches Maß der Matrix, obwohl es sich im strengen mathematischen Sinne nicht um ein Maß handelt. Das Hadamardsche Maß wurde bereits in [13], [31] eingeführt und in [31] für einen Konvergenzbeweis herangezogen. Auch uns wird es in Unterabschnitt 2.3.2 dazu dienen, die globale Konvergenz der in Unterbschnitt 2.2.1, 2.2.3 beschriebenen Verfahren zu beweisen. Maßgeblich für den Konvergenzbeweis ist das Wachstum des Hadamardschen Maßes der positiv definiten Matrix nach jeder Transformation.

Definition und Lemma 2.3.1 *Sei $K \in \mathbb{R}^{2n\times 2n}$ symmetrisch und positiv definit. Das Hadamardsche Maß*

$$\mathcal{H}(K) = \frac{\det(K)}{\prod_{i=1}^{2n} K_{ii}}$$

von K hat folgende Eigenschaften:

1. *$\mathcal{H}(K) \leq 1$ und $\mathcal{H}(K) = 1$ genau dann, wenn K diagonal ist.*

2. *$\mathcal{H}(K) = \mathcal{H}(DKD)$ für beliebiges nichtsinguläres, diagonales D.*

3. *Sei $K = DK_SD$ mit einer Diagonalmatrix D und $K_{S_{ii}} = 1$, $i = 1 \ldots 2n$. Dann gilt*
$$\lambda_{\min}(K_S) \geq \frac{\mathcal{H}(K)}{e} = \frac{\det K_S}{e} = \frac{\mathcal{H}(K_S)}{e} \ ,$$
wobei $e = \exp(1)$.

4. *Die symplektische Matrix $T \in \mathbb{R}^{2n\times 2n}$ unterscheide sich nur in einer symplektischen Submatrix $\hat{T} \in \mathbb{R}^{2m\times 2m}$ von der Einheitsmatrix. Die zu $\hat{T}$ korrespondierende Submatrix von K sei $\hat{K}$ und es sei $\hat{K}' = \hat{T}^T\hat{K}\hat{T}$. Dann gilt*
$$\mathcal{H}(T^TKT) = \mathcal{H}(K)\frac{\mathcal{H}(\hat{K}')}{\mathcal{H}(\hat{K})} \ .$$

5. *Ist unter den Voraussetzungen von 3. zusätzlich $\hat{K}'$ diagonal, so gilt*
$$\mathcal{H}(T^TKT) = \frac{\mathcal{H}(K)}{\mathcal{H}(\hat{K})} \ .$$

6. *Für jede Hauptuntermatrix $\hat{K} \in \mathbb{R}^{2m\times 2m}$ von $K \in \mathbb{R}^{2n\times 2n}$ gilt $\mathcal{H}(K) \leq \mathcal{H}(\hat{K})$.*

Beweis:
Zu 1., 2. und 3. siehe [13]. Auch die 4. Aussage wird ähnlich zu [13] bewiesen. Es gilt

$$\begin{aligned}
\mathcal{H}(T^TKT) &= \frac{\det(T^TKT)}{\prod_{i=1}^{2n} K'_{ii}} \\
&= \frac{\det(K)}{\prod_{i=1}^{2n} K_{ii}} \ \frac{\prod_{i=1}^{2m} \hat{K}_{ii}}{\prod_{i=1}^{2m} \hat{K}'_{ii}} \\
&= \mathcal{H}(K) \ \frac{\det(\hat{T}^T\hat{K}\hat{T})}{\det(\hat{K})} \ \frac{\prod_{i=1}^{2m} \hat{K}_{ii}}{\prod_{i=1}^{2m} \hat{K}'_{ii}} \\
&= \mathcal{H}(K) \ \frac{\mathcal{H}(\hat{K}')}{\mathcal{H}(\hat{K})} \ .
\end{aligned}$$

Die 5. Aussage ist eine Kombination von 1. und 4. Zum Beweis der 6. Aussage nehmen wir o.B.d.A.
$$K = \begin{bmatrix} \hat{K} & \bar{K} \\ \bar{K}^T & \tilde{K} \end{bmatrix}$$

an. Aus der Fischerschen Ungleichung [20] $\det K \leq \det \hat{K} \det \tilde{K}$ folgt

$$\begin{aligned} \mathcal{H}(K) &= \frac{\det K}{\prod_{i=1}^{2n} K_{ii}} \leq \frac{\det \hat{K}}{\prod_{i=1}^{2m} \hat{K}_{ii}} \frac{\det \tilde{K}}{\prod_{i=1}^{2n-2m} \tilde{K}_{ii}} \\ &\leq \mathcal{H}(\hat{K})\mathcal{H}(\tilde{K}) \leq \mathcal{H}(\hat{K}) \end{aligned}$$

Q.E.D.

Transformiert man also eine Submatrix von $K \in I\!R^{2n\times 2n}$, so ist die Änderung des Hadamardschen Maßes der transformierten Matrix gleich der Änderung des Hadamardschen Maßes der transformierten Submatrix. Aus 1. und 5. entnimmt man, was dies für die Diagonalisierung einer Submatrix bedeutet. Wir betrachten nun die Wirkung der Transformationen in modifizierten Zyklen auf das Hadamardsche Maß und können uns auf diejenigen Submatrizen beschränken, denen die jeweilige Transformation gilt. Zur Erzielung einer gewissen Bezeichnungsökonomie werden wir die Submatrizen mit den vollen Matrizen identifizieren. Wir werden also auf einige Partitionierungen verzichten und hier wieder große Buchstaben für Matrixelemente verwenden.

Sei

$$\begin{aligned} K &= \begin{bmatrix} F & B^T \\ B & C \end{bmatrix} \\ &= \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} K_S \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix}, \quad K_S = \begin{bmatrix} F_S & B_S^T \\ B_S & C_S \end{bmatrix}, \quad D = \operatorname{diag}(F_{ii}^{1/2}) \end{aligned} \tag{2.61}$$

mit

$$F_{ii} = C_{ii} \quad , \quad F_{ii} \geq F_{i+1,i+1} \text{ für alle möglichen } i$$

eine symmetrische und positiv definite Submatrix, für die die drei Bedingungen von Seite 44 erfüllt sind. Wir betrachten die Transformationen (2.38) und (2.39). Sei

$$\begin{aligned} K' &= \begin{bmatrix} F' & B'^T \\ B' & C' \end{bmatrix} \\ &= \begin{bmatrix} D'_F & 0 \\ 0 & D'_C \end{bmatrix} K'_S \begin{bmatrix} D'_F & 0 \\ 0 & D'_C \end{bmatrix}, \quad K'_S = \begin{bmatrix} F'_S & B'^T_S \\ B'_S & C'_S \end{bmatrix}, \end{aligned}$$

mit $D'_F = \operatorname{diag}(F'^{1/2}_{ii})$, $D'_C = \operatorname{diag}(C'^{1/2}_{ii})$ die gemäß (2.38) transformierte Matrix und

$$\begin{aligned} K'' &= \begin{bmatrix} F'' & B''^T \\ B'' & C'' \end{bmatrix} \\ &= \begin{bmatrix} D''_F & 0 \\ 0 & D''_C \end{bmatrix} K''_S \begin{bmatrix} D''_F & 0 \\ 0 & D''_C \end{bmatrix}, \quad K''_S = \begin{bmatrix} F''_S & B''^T_S \\ B''_S & C''_S \end{bmatrix}, \end{aligned}$$

mit $D''_F = \mathrm{diag}(F''^{1/2}_{ii})$, $D''_C = \mathrm{diag}(C''^{1/2}_{ii})$ die gemäß (2.39) transformierte Matrix. Bei den Transformationen gilt natürlich jetzt $num = 1$. Wir haben als erstes das

Lemma 2.3.2 *Die Matrix K aus (2.61) werde gemäß (2.38), (2.39) transformiert. Dann gilt für die transformierte Matrix K'' und $c > 1$*

$$\mathcal{H}(K'') > c \cdot \mathcal{H}(K) \iff \prod_{i=1}^{2n} K_{ii} > c \cdot \prod_{i=1}^{2n} K''_{ii} \ . \tag{2.62}$$

Beweis:
Aus

$$\mathcal{H}(K'') = \frac{\det K''}{\prod_{i=1}^{2n} K''_{ii}} = \frac{\det K}{\prod_{i=1}^{2n} K''_{ii}} \frac{\prod_{i=1}^{2n} K_{ii}}{\prod_{i=1}^{2n} K_{ii}} = \mathcal{H}(K) \frac{\prod_{i=1}^{2n} K_{ii}}{\prod_{i=1}^{2n} K''_{ii}}$$

folgt die Aussage unmittelbar durch Einsetzen. Q.E.D.

Wir wollen nun zeigen, daß die drei Bedingungen von Seite 44 hinreichend für ein Wachstum des Hadamardschen Maßes ist. Dazu nehmen wir noch $n \geq 3$ an, was keine Beschränkung der Allgemeinheit ist, weil K im Falle $n \leq 2$ exakt diagonalisiert werden kann, wie sich in Abschnitt 2.2 zeigte. Definieren wir noch

$$\tilde{S} = (\tilde{S}_{ij}) \quad \text{mit} \quad \tilde{S}_{ij} = \begin{cases} F_{S_{ij}} & , \quad i > j \\ 0 & , \quad i \leq j \ , \end{cases} \tag{2.63}$$

und

$$\hat{S} = (\hat{S}_{ij}) \quad \text{mit} \quad \hat{S}_{ij} = \begin{cases} C_{S_{ij}} & , \quad i > j \\ 0 & , \quad i \leq j \ , \end{cases} \tag{2.64}$$

so gelangen wir zunächst zu folgender Hilfsaussage.

Lemma 2.3.3 *Sei F_S gemäß (2.61) mit $\alpha = \|F_S - I\|_E \leq 1/30$ gegeben und $F_S = L_S L_S^T$, $L_S = I + S$, ihre Cholesky-Zerlegung. Für die Matrix $\tilde{S}$ sei gemäß (2.63) gilt dann $\|S - \tilde{S}\|_E < 0.366\,\alpha^2$.*

Beweis:
Nach [17] ist

$$\|S\|_E \leq (\sqrt{2}\alpha)/(1 + \sqrt{1 - 2\alpha}) < 0.72\,\alpha \ .$$

Definiert man

$$R = (R_{ij}) \quad \text{mit} \quad R_{ij} = \begin{cases} (SS^T)_{ij} & , \quad i > j \\ (SS^T)_{ij}/2 & , \quad i = j \\ 0 & , \quad i < j, \end{cases}$$

so folgt

$$\begin{aligned} I + \tilde{S} + \tilde{S}^T &= F_S = (I + S)(I + S^T) = I + S + S^T + SS^T \\ &= I + (S + R) + (S + R)^T \end{aligned}$$

und deshalb $\tilde{S} = S + R$, $\|R\|_E \leq \|SS^T\|_E/\sqrt{2} < 0.366\,\alpha^2$. Q.E.D.

Wir sind nun in der Lage, den gewünschten Satz zu formulieren.

Satz 2.3.4 *Erfüllt K aus (2.61) die drei Bedingungen von Seite 44 und wird K gemäß (2.38), (2.39) transformiert, so gilt für die transformierte Matrix K''*

$$\mathcal{H}(K'') > \frac{10}{10 - TOL^2}\,\mathcal{H}(K)\ .$$

Beweis:
Für den einfachen, aber langwierigen Beweis setzen wir $F_{11} = 1$ voraus. Weil dies stets durch Multiplikation der Matrix mit der Konstanten $1/F_{11}$ erzielt werden kann, ist dies keine Einschränkung der Allgemeinheit. Wir halten zunächst einige Resultate fest, die sich aus den Bedingungen von Seite 44 ergeben (das dortige $\omega^{[l,u]}$ ist hier ω):

$$\begin{aligned}
\|D^{-2}\|_2 &= \|(I - (I - D^2))^{-1}\|_2 \leq \frac{1}{1 - \|D^2 - I\|_2} \leq 1 + \omega \\
\|D^{-2} - I\|_2 &\leq \|D^{-2}\|_2 \|D^2 - I\|_2 \leq \omega \\
\|D^{-4} - I\|_2 &\leq \|D^{-2} - I\|_2 \|D^{-2} + I\|_2 \leq \|D^{-2} - I\|_2 \|D^{-2}\|_2 \|D^2 + I\|_2 \\
&\leq 2\,\omega(1 + \omega)\ .
\end{aligned}$$

Setzt man $C''_{ii} = D^4_{ii} + \mu_i$, $i = 1 \ldots n$, so ist

$$\begin{aligned}
\prod_{i=1}^{2n} K''_{ii} &= \prod_{i=1}^{n} C''_{ii} = \prod_{i=1}^{n} (D^4_{ii} + \mu_i) = \prod_{i=1}^{n} D^4_{ii} \cdot \prod_{i=1}^{n} (1 + \frac{\mu_i}{D^4_{ii}}) \\
&= \prod_{i=1}^{2n} K_{ii} \cdot \prod_{i=1}^{n} (1 + \frac{\mu_i}{D^4_{ii}})
\end{aligned}$$

und (2.62) ist äquivalent zu

$$p_n := \prod_{i=1}^{n} (1 + \frac{\mu_i}{D^4_{ii}}) < \frac{1}{c}\ . \tag{2.65}$$

Durch Anwendung der binomischen Formel erhält man

$$p_n = 1 + \sum_{i=1}^{n} \frac{\mu_i}{D^4_{ii}} + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{\mu_i \mu_j}{D^4_{ii} D^4_{jj}} + r \quad , \quad |r| \leq \sum_{i=3}^{n} \binom{n}{i} \max_{1\leq k\leq n} |\frac{\mu_k}{D^4_{kk}}|^i\ . \tag{2.66}$$

Um zur Aussage des Satzes zu gelangen, bestimmen wir eine obere Schranke des linken Ausdruckes in (2.66). Dazu bemerken wir vorab, daß es nach [17] eine untere Dreiecksmatrix S gibt mit

$$L_S = I + S \quad , \quad \|S\|_E \leq \frac{\sqrt{2}\omega}{1 + \sqrt{1 - 2\omega}} < 0.72\,\omega\ .$$

In Lemma 2.3.3 wurde $\|R\|_E \leq 0.366\,\omega^2$ für $R = \tilde{S} - S$ bewiesen. Außerdem ist $\mathrm{diag}(\tilde{S}_{ii}) = 0$ und es gilt $\|\tilde{S}\|_E = \|F_S - I\|_E/\sqrt{2} \leq \omega/\sqrt{2}$ und ebenso $\|\hat{S}\|_E \leq \omega/\sqrt{2}$. Zuerst schätzen wir $|\mu_i/D_{ii}^4|$, $i = 1 \ldots n$, ab. Es gilt

$$\frac{\mu_i}{D_{ii}^4} = \frac{1}{D_{ii}^4}((X^2)_{ii} + 2(B'X)_{ii} + (L^TCL)_{ii} - D_{ii}^4)\ . \tag{2.67}$$

Außerdem ist

$$\begin{aligned}\|B'\|_E &= \|L^TBL^{-T}\|_E \leq \|B\|_E\|L^T\|_2\|L^{-T}\|_2 \leq \|B\|_E\frac{\sqrt{1+\omega}}{\sqrt{1-\omega}} \\ &\leq \|B_S\|_E\frac{\sqrt{1+\omega}}{\sqrt{1-\omega}}\ .\end{aligned}$$

Damit können bereits wir den ersten Teil von (2.67) abschätzen:

$$\begin{aligned}\frac{1}{D_{ii}^4}((X^2)_{ii} + 2(B'X)_{ii}) &\leq (1+\omega)^2(\|X\|_E^2 + 2\,\|B'X\|_E) \\ &\leq (1+\omega)^2(\|B'\|_E^2 + \|B'(B' + B'^T)\|_E) \\ &\leq (1+\omega)^2 \cdot 3\,\frac{1+\omega}{1-\omega}\,\frac{\omega^2}{2} < 1.61\,\omega^2\ .\end{aligned}$$

Weiter haben wir wegen $C_S = I + \hat{S} + \hat{S}^T$

$$\begin{aligned}|(L^TCL)_{ii} - D_{ii}^4| &= |((I+S^T)D^2(I+\hat{S}+\hat{S}^T)D^2(I+S))_{ii} - D_{ii}^4| \\ &= |2(D^4S)_{ii} + (S^TD^4S)_{ii} \\ &\quad +((I+S^T)D^2(\hat{S}+\hat{S}^T)D^2(I+S))_{ii}| \\ &\leq 2|R_{ii}| + \|S\|_E^2 + |(2\hat{S}^TS + 2S^T\hat{S}S)_{ii}| \\ &\leq 2\cdot 0.366\,\omega^2 + 0.72^2\,\omega^2 + 2\frac{\omega}{\sqrt{2}}\cdot 0.72\,\omega + 2\cdot 0.72^2\,\omega^2\,\frac{\omega}{\sqrt{2}} \\ &< 2.30\,\omega^2\ ,\end{aligned}$$

so daß folgt

$$\frac{|\mu_i|}{D_{ii}^4} \leq 1.61\,\omega^2 + 2.30\,\omega^2\,(1+\omega)^2 \leq 4.07\,\omega^2\ . \tag{2.68}$$

Damit können wir $|r|$ nun abschätzen:

$$\begin{aligned}|r| &\leq \sum_{i=3}^{n}\binom{n}{i}(4.07\,\omega^2)^i \\ &\leq \sum_{i=3}^{n}\frac{n!}{(n-i)!\,n^i}\,\frac{1}{i!}\left(\frac{4.07\omega}{10}\right)^i \leq \sum_{i=3}^{\infty}\frac{1}{i!}\,(0.407\,\omega)^i \\ &\leq (0.407\,\omega)^3\sum_{i=0}^{\infty}\frac{1}{i!}(0.407\,\omega)^i \leq 0.407^3\,\exp(\frac{0.407}{30})\,\omega^3 \\ &< 0.069\,\omega^3\ .\end{aligned} \tag{2.69}$$

Als nächstes soll der Term $\sum_{i=1}^{n} \mu_i / D_{ii}^4$ vereinfacht werden. Mit (2.67) gilt

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\mu_i}{D_{ii}^4} &= \mathrm{Tr}\ D^{-4}X^2 + 2\mathrm{Tr}\ D^{-4}B'X + \mathrm{Tr}\ (D^{-2}L^TCLD^{-2} - I) \\
&= \frac{1}{4}\mathrm{Tr}\ D^{-4}(2B' + 2B'B'^T) - \mathrm{Tr}\ D^{-4}B'^2 - \mathrm{Tr}\ D^{-4}B'B'^T \\
&\quad + \mathrm{Tr}\ (D^{-2}(I+S^T)D^2C_SD^2(I+S)D^{-2} - I) \\
&= -\frac{1}{2}\mathrm{Tr}\ D^{-4}B'^2 - \frac{1}{2}\|D^{-2}B'\|_E^2 \\
&\quad + 2\mathrm{Tr}\ D^{-2}S^TD^2C_S + \mathrm{Tr}\ D^{-2}S^TD^2C_SD^2SD^{-2}
\end{aligned} \tag{2.70}$$

Um die Summanden von (2.70) einzeln zu untersuchen, soll vorab bemerkt werden, daß es eine Matrix M gibt mit

$$(I+S)^{-1} = I - S + M \quad , \quad \|M\|_E \le \|F_S^{-1}\|_2^{1/2}\|S\|_E^2 < 0.53\,\omega^2 \ ,$$

und daß $|\mathrm{Tr}\ XY| \le \|X\|_E\|Y\|_E$ gilt. Damit folgt

$$\begin{aligned}
\mathrm{Tr}\ D^{-4}B'^2 &= \mathrm{Tr}\ D^{-4}(I+S^T)B^2(I-S^T+M^T) \\
&= \mathrm{Tr}\ (D^{-4} - I + I)B^2 \\
&\quad + \mathrm{Tr}\ D^{-4}(B^2(-S^T+M^T) + S^TB^2(I+S^T)^{-1}) \\
&= \mathrm{Tr}\ B^2 + r_1
\end{aligned}$$

mit

$$\begin{aligned}
|r_1| &\le \|D^{-4} - I\|_2\|B\|_E^2 + (1+\omega)^2(\|B\|_E^2(\|S\|_E + \|M\|_E) \\
&\quad + \|S\|_E\|B\|_E^2\|(I+S^T)^{-1}\|_2) \\
&\le 2(1+\omega)\omega\frac{\omega^2}{2} + (1+\omega)^2(\frac{\omega^2}{2}(0.72\,\omega + 0.53\,\omega^2) + 0.72\,\omega\frac{\omega^2}{2\sqrt{1-\omega}}) \\
&\le \omega^3\,(1 + \omega + \frac{(1+\omega)^2}{2}(0.72 + \frac{0.53}{30}) + \frac{0.72}{2\sqrt{1-\omega}}) \\
&< 1.80\,\omega^3 \ .
\end{aligned}$$

Weiterhin ist

$$\begin{aligned}
\mathrm{Tr}\ B^2 &= \mathrm{Tr}\ D^2B_SD^2B_S = \mathrm{Tr}\ (I + D^2 - I)B_S(I + D^2 - I)B_S \\
&= \mathrm{Tr}\ B_S^2 + 2\mathrm{Tr}\ (D^2 - I)B_S^2 + \mathrm{Tr}\ (D^2 - I)B_S(D^2 - I)B_S \\
&= \mathrm{Tr}\ B_S^2 + r_2
\end{aligned}$$

mit

$$\begin{aligned}
|r_2| &\le 2\,\|D^2 - I\|_2\|B_S'\|_E^2 + \|D^2 - I\|_2^2\|B_S\|_E^2 \le 2\,\frac{\omega}{1+\omega}\,\frac{\omega^2}{2} + \frac{\omega^2}{(1+\omega)^2}\,\frac{\omega^2}{2} \\
&< 0.99\,\omega^3 \ ,
\end{aligned}$$

so daß wir schließlich

$$\text{Tr } D^{-4}B'^2 = \text{Tr } B_S^2 + \tilde{r}_1 \quad , \quad |\tilde{r}_1| \leq |r_1| + |r_2| < 2.79\,\omega^3 \tag{2.71}$$

erhalten. Nun betrachten wir den zweiten Summanden in (2.70). Es gilt

$$\begin{aligned} \|D^{-2}B'\|_E &= \|D^{-2}(I+S^T)D^2B_S(DD^{-1})(I-S^T+M^T)\|_E \\ &= \|B_S\|_E + r_3 \end{aligned}$$

mit

$$\begin{aligned} |r_3| &\leq \|B_S(-S^T+M^T)\|_E + \|D^{-2}S^TD^2B_S(I+S^T)^{-1}\|_E \\ &\leq \frac{\omega}{\sqrt{2}}(0.72\,\omega + 0.53\,\omega^2) + (1+\omega)\,0.72\,\omega\frac{\omega}{\sqrt{2}\sqrt{1-\omega}} \\ &\leq \omega^2\,(\frac{0.72}{\sqrt{2}} + \frac{0.53}{30\sqrt{2}} + \frac{0.72\cdot 31}{30\sqrt{2}\sqrt{29/30}}) \\ &< 1.06\,\omega^2\ . \end{aligned}$$

Es folgt

$$\|D^{-2}B'\|_E^2 = \|B_S\|_E^2 + r_4 \tag{2.72}$$

mit

$$\begin{aligned} |r_4| &\leq |r_3|^2 + 2\,|r_3|\,\|B_S\|_E \leq 1.06^2\,\omega^4 + 2\cdot 1.06\,\frac{\omega^3}{\sqrt{2}} \\ &< 1.54\,\omega^3\ . \end{aligned}$$

Schließlich betrachten wir den dritten Summanden von (2.70). Wegen $0 = \text{Tr } \tilde{S} = \text{Tr } (S+R)$ gilt für diesen Summanden

$$\begin{aligned} \text{Tr } D^{-2}S^TD^2C_S &= \text{Tr } (D^{-2}S^TD^2(I+\hat{S}+\hat{S}^T)) = \text{Tr } S + \text{Tr } D^{-2}S^TD^2\hat{S} \\ &= -\text{Tr } R + \text{Tr } D^{-2}(\tilde{S}^T - R^T)D^2\hat{S} \\ &= -\frac{1}{2}\text{Tr } (\tilde{S}\tilde{S}^T - 2\tilde{S}R^T + RR^T) \\ &\quad + \text{Tr } (D^{-2}-I+I)\tilde{S}^T(D^2-I+I)\hat{S} - \text{Tr } D^{-2}R^TD^2\hat{S} \\ &= -\frac{1}{2}\text{Tr } \tilde{S}\tilde{S}^T + \text{Tr } \tilde{S}^T\hat{S} + r_5 \end{aligned} \tag{2.73}$$

mit

$$\begin{aligned} |r_5| &\leq \|\tilde{S}\|_E\|R\|_E + \frac{1}{2}\|R\|_E^2 + (\|D^{-2}-I\|_2 + \|D^2-I\|_2 \\ &\quad + \|D^{-2}-I\|_2\|D^2-I\|_2)\|\tilde{S}\|_E\|\hat{S}\|_E + \|D^{-2}\|_2\|R\|_E\|\hat{S}\|_E \\ &\leq \frac{\omega}{\sqrt{2}}\cdot 0.366\,\omega^2 + \frac{1}{2}\,0.366^2\omega^4 + (\omega + \frac{\omega}{1+\omega} + \frac{\omega^2}{1+\omega})\,\frac{\omega^2}{2} \end{aligned}$$

$$
\begin{aligned}
& +(1+\omega)\,0.366\,\omega^2\frac{\omega}{\sqrt{2}} \\
\leq\ & \omega^3\,(\frac{0.366}{\sqrt{2}}+\frac{0.366^2}{60}+\frac{1}{2}\,(1+\frac{30}{31}+\frac{1}{31})+\frac{31}{30\sqrt{2}}\,0.366 \\
<\ & 1.53\,\omega^3\ .
\end{aligned}
$$

Schließlich gilt für den letzten Summanden in (2.70)

$$
\begin{aligned}
& \mathrm{Tr}\ D^{-2}S^TD^2C_SD^2SD^{-2} \\
=\ & \mathrm{Tr}\ D^{-2}S^TD^4SD^{-2}+\mathrm{Tr}\ D^{-2}S^TD^2(C_S-I)D^2SD^{-2} \\
=\ & \mathrm{Tr}\ D^{-4}S^T(D^4-I+I)S+r_6\ .
\end{aligned}
$$

Dabei ist

$$
\begin{aligned}
|r_6| &\leq \|D^{-4}\|_2\|S\|_E^2\|C_S-I\|_E \leq (1+\omega)^2\,0.72^2\,\omega^2\,\frac{\omega}{1+\omega} \\
&< 0.56\,\omega^3
\end{aligned}
$$

und weiter

$$
\begin{aligned}
& \mathrm{Tr}\ D^{-2}S^TD^2C_SD^2SD^{-2} \\
=\ & \mathrm{Tr}\ D^{-4}S^T(D^4-I)S+\mathrm{Tr}\ (D^{-4}-I+I)S^TS+r_6 \\
=\ & \mathrm{Tr}\ D^{-4}S^T(D^4-I)S+\mathrm{Tr}\ (D^{-4}-I)S^TS \\
& \quad +\mathrm{Tr}\ (\tilde{S}^T-R^T)(\tilde{S}-R)+r_6 \\
=\ & \mathrm{Tr}\ \tilde{S}^T\tilde{S}+\mathrm{Tr}\ D^{-4}S^T(D^4-I)S+\mathrm{Tr}\ (D^{-4}-I)S^TS \\
& \quad -2\mathrm{Tr}\ \tilde{S}^TR+\mathrm{Tr}\ R^TR+r_6 \\
=\ & \mathrm{Tr}\ \tilde{S}^T\tilde{S}+r_6+r_7
\end{aligned}
\tag{2.74}
$$

mit

$$
\begin{aligned}
|r_7| &\leq \|D^{-4}\|_2\|S\|_E^2\|D^4-I\|_2+\|D^{-4}-I\|_2\|S\|_E^2+2\,\|\tilde{S}\|_E\|R\|_E+\|R\|_E^2 \\
&\leq (1+\omega)^2\cdot 0.72^2\,\omega^2\,\frac{2\omega}{1+\omega}+2\,\omega(1+\omega)0.72^2\,\omega^2 \\
&\quad +\frac{2}{\sqrt{2}}\,\omega\,0.366\,\omega^2+0.366^2\,\omega^4 \\
&\leq \omega^3\,(2\,\frac{31}{30}\,2\cdot 0.72^2+\sqrt{2}\cdot 0.366+\frac{0.366}{30}) \\
&< 2.68\,\omega^3\ .
\end{aligned}
$$

Zur Bestimmung von $\sum_{i=1}^n \mu_i/D_{ii}^4$ fassen wir (2.71), (2.72), (2.73) und (2.74) zusammen und erhalten

$$
\begin{aligned}
\sum_{i=1}^n \frac{\mu_i}{D_{ii}^4} &= -\frac{1}{2}\mathrm{Tr}\ B_S^2-\frac{1}{2}\mathrm{Tr}\ B_SB_S^T-\mathrm{Tr}\ \tilde{S}\tilde{S}^T+2\mathrm{Tr}\ \tilde{S}^T\hat{S}+\mathrm{Tr}\ \tilde{S}^T\tilde{S}+r' \\
&= -\frac{1}{4}\mathrm{Tr}\ (B_S+B_S^T)^2+2\mathrm{Tr}\ \tilde{S}^T\hat{S}+r'
\end{aligned}
\tag{2.75}
$$

mit

$$\begin{aligned} |r'| &\leq \omega^3(2.79/2 + 1.54/2 + 2 \cdot 1.53 + 0.56 + 2.68) \\ &< 8.47\,\omega^3 \ . \end{aligned}$$

Es folgt

$$\begin{aligned} |\sum_{i=1}^{n} \frac{\mu_i}{D_{ii}^4}| &\leq \frac{1}{4}\|B_S + B_S^T\|_E^2 + 2\|\tilde{S}\|_E\|\hat{S}\|_E + |r'| \\ &\leq \frac{1}{4}(2\,\omega)^2 + 2\,(\frac{\omega^2}{2}) + |r'| \leq \omega^2\,(2 + \frac{8.47}{30}) \\ &< 2.29\,\omega^2 \ . \end{aligned}$$

Dies und (2.68) können wir zur Bestimmung einer Schranke von

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{\mu_i \mu_j}{D_{ii}^4 D_{jj}^4}$$

verwenden:

$$\begin{aligned} |\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{\mu_i \mu_j}{D_{ii}^4 D_{jj}^4}| &\leq \sum_{i=1}^{n} |\frac{\mu_i}{D_{ii}^4}|\,|\sum_{j=1}^{n} \frac{\mu_j}{D_{jj}^4} - \frac{\mu_i}{D_{ii}^4}| \\ &\leq \sum_{i=1}^{n} |\frac{\mu_i}{D_{ii}^4}|\,|\sum_{j=1}^{n} \frac{\mu_j}{D_{jj}^4}| + \sum_{i=1}^{n} \frac{\mu_i^2}{D_{ii}^8} \\ &\leq 2.29\,\omega^2 \cdot n \cdot 4.07\,\omega^2 + 4.07^2 \cdot n \cdot \omega^4 \\ &\leq \omega^3\,(\frac{2.29 \cdot 4.07}{10} + \frac{4.07^2}{10}) \\ &< 2.59\,\omega^3 \ . \end{aligned} \tag{2.76}$$

Aus (2.69),(2.75) und (2.76) ergibt sich schließlich eine Abschätzung von p_n in (2.66)

$$p_n = 1 - \frac{1}{4}\mathrm{Tr}\ (B_S + B_S^T)^2 + 2\mathrm{Tr}\ \tilde{S}^T\hat{S} + \hat{r}$$

mit

$$\begin{aligned} \hat{r} &< \omega^3\,(0.069 + 8.47 + 2.59) \\ &< 11.13\,\omega^3 \ . \end{aligned}$$

Gilt also

$$-\frac{1}{4}\mathrm{Tr}\ (B_S + B_S^T)^2 + 2\mathrm{Tr}\ \tilde{S}^T\hat{S} + 12\,\omega^3 \leq \frac{1}{c} - 1 =: -\frac{TOL^2}{10}\ ,$$

so ist

$$\mathcal{H}(K'') > c \cdot \mathcal{H}(K) = \frac{1}{1 - TOL^2/10}\,\mathcal{H}(K)$$

wegen (2.65) erfüllt, was wegen $2\mathrm{Tr}\ \tilde{S}^T\hat{S} = \mathrm{Tr}\ (F_S - I)(C_S - I)$ die Aussage des Satzes ist. Q.E.D.

Das in 2.2.2 bzw. 2.2.4 beschriebene Verfahren zur Lösung des Eigenwertproblems (K, J) besteht aus einer Folge von Diagonalisierungen von Submatrizen bzw. von Transformationen (2.38), (2.39). Bei all diesen Operationen wächst also das Hadamardsche Maß der transformierten Matrix. Dies werden wir im nachfolgenden Unterabschnitt zum Beweis der globalen Konvergenz des Verfahrens heranziehen.

2.3.2 Globale Konvergenz

In diesem Unterabschnitt wenden wir uns der Konvergenz des Verfahrens 2.2.2 zu, also der Konvergenz der Folge der positiv definiten, symmetrischen Matrizen K gegen eine Diagonalmatrix, zu. Wir werden die *globale Konvergenz* des Verfahrens, also die Konvergenz für beliebige positiv definite Eingabematrizen K zeigen, falls in exakter Arithmetik gerechnet wird. Die Konvergenz des impliziten Verfahrens 2.2.4 ist äquivalent zu derjenigen des expliziten Verfahrens, so daß wir uns aus Vereinfachungsgründen zunächst auf das explizite Verfahren beschränken können.

Die Idee des Konvergenzbeweises entstammt [31] und besteht darin, daß das Hadamardsche Maß der betrachteten Matrix nach jeder Transformation stärker als eine vorgegebene Konstante wächst und aus der Beschränktheit des Maßes die Konvergenz des Verfahrens folgt.

Vorab erwähnen wir noch, daß es sich beim Verfahren 2.2.2 um ein *zyklisches Verfahren mit Schwellenstrategie* handelt: die Elemente an jeder Position (i, j), $i \neq j$, werden zyklisch für eine Transformation vorgesehen, falls sie (skaliert) nicht bereits kleiner als eine vorgegebene Konstante sind.

Lemma 2.3.5 *Sei $TOL > 0$ eine Abbruchkonstante (Schwelle). Wird in exakter Arithmetik gerechnet, so terminiert das Verfahren 2.2.2 nach endlich vielen Zyklen. Es gilt dann*

$$\frac{|K_{ij}|}{\sqrt{K_{ii}K_{jj}}} < TOL \quad , \quad 1 \leq i, j \leq 2n \ , \quad i \neq j \tag{2.77}$$

Beweis:
Jeder Zyklus besteht aus einer endlichen Folge von Diagonalisierungen von Submatrizen der Ordnung 4 oder von Transformationspaaren (2.38), (2.39).

Wir werden zuerst die diagonalisierenden Transformationen T betrachten. Deren nichttrivialer Teil sei $\hat{T}$. Sie werden nur bei solchen Submatrizen $\hat{K}$ angewendet, die ein Element $\hat{K}_{ij}$, $i \neq j$, mit $|\hat{K}_{ij}|/(\hat{K}_{ii}\hat{K}_{jj})^{1/2} \geq TOL$ haben. Wegen der

Aussagen 5., 6. von 2.3.1 (siehe auch [31]) bestimmt

$$\mathcal{H}(T^T KT) \geq \mathcal{H}(K)\ \frac{\hat{K}_{ii}\hat{K}_{jj}}{\hat{K}_{ii}\hat{K}_{jj} - \hat{K}_{ij}^2} \geq \mathcal{H}(K)\ \frac{1}{1 - TOL^2}\ .$$

das Mindestwachstum des Hadamardschen Maßes.

Nun behandeln wir das Transformationspaar (2.38), (2.39). Es wird nur an solchen Submatrizen vorgenommen, für die die drei Bedingungen von Seite 44 erfüllt sind. Bezeichnen wir mit K'' die transformierte Matrix, so folgt aus Satz 2.3.4

$$\mathcal{H}(K'') > \mathcal{H}(K)\ \frac{10}{10 - TOL^2}\ .$$

Die Folge der Hadamardschen Maße der transformierten Matrizen wächst also nach jeglicher Transformation mindestens um $10/(10-TOL^2)$. Nach 2.3.1 ist die Folge jedoch beschränkt, so daß nach einer endlichen Zahl von Zyklen keine Transformation mehr vorgenommen wird. Dann gibt es also keine Submatrix mehr, für alle drei Bedingungen von Seite 44 erfüllt wäre. Weil insbesondere auch keine Diagonalisierungen mehr durchgeführt werden, ist für alle $1 \leq i, j \leq n$, $i \neq j$ die Abbruchbedingung (2.77) erfüllt. Q.E.D.

Man erkennt, daß in der 3. Bedingung von Seite 44 statt $-TOL^2/10$ auch $-c \cdot TOL^2$ mit einem kleinen $c > 0$ verwendet werden könnte. In numerischen Tests führte aber $c \leq 1/10$ und sogar $c = 0$ zu guten Ergebnissen.

Satz 2.3.6 *Die positiven Eigenwerte des Problems*

$$Kx = i\lambda Jx \quad , \quad K = K^{[0]} \in I\!R^{2n\times 2n}$$

mit symmetrischem positiv definitem K seien

$$\lambda_1 \geq \ldots \geq \lambda_n > 0\ .$$

Sei

$$(TOL_k)\ ,\quad k = 1, 2, \ldots \quad mit \quad TOL_1 < \frac{1}{4n} \tag{2.78}$$

eine positive monotone Nullfolge von Abbruchkonstanten. Für $k = 1, 2, \ldots$ werde jeweils der Algorithmus 2.2.2 in exakter Arithmetik mit der Abbruchkonstanten TOL_k, der Eingabematrix $K^{[k-1]}$ und der Ausgabematrix $K^{[k]}$ durchgeführt. Sei

$$K_S^{[k]} = D^{[k]^{-1}} K^{[k]} D^{[k]^{-1}} \quad , \quad D^{[k]} = diag(K_{ii}^{[k]^{1/2}})\ .$$

Dann gilt

$$\lim_{k\to\infty} \|K_S^{[k]} - I\|_E = 0 \tag{2.79}$$

und

$$\lim_{k\to\infty} F_{ii}^{[k]} = \lambda_i \quad , \quad i = 1 \ldots n\ . \tag{2.80}$$

Beweis:
Wegen Lemma 2.3.5 ist für jedes k

$$|K_{S_{ij}}^{[k]}| = \frac{|K_{ij}^{[k]}|}{\sqrt{K_{ii}^{[k]} K_{jj}^{[k]}}} < TOL_k \quad , \quad 1 \leq i,j \leq 2n \ , \quad i \neq j$$

und deshalb

$$\|K_S^{[k]} - I\|_E \leq \sqrt{2\,n\,(2n-1)}\ TOL_k \leq 2\,n\ TOL_k \ . \tag{2.81}$$

Es folgt

$$\lim_{k\to\infty} \|K_S^{[k]} - I\|_E \leq 2\,n \lim_{k\to\infty} TOL_k = 0 \ ,$$

was die Aussage (2.79) ist. Um (2.80) zu zeigen, sei

$$\delta K^{[k]} = D^{[k]^2} - K^{[k]} \quad , \quad \delta K_S^{[k]} = D^{[k]^{-1}}(D^{[k]^2} - K^{[k]})D^{[k]^{-1}} \ .$$

Dann ist wegen (2.78) und (2.81)

$$\|\delta K_S^{[k]}\|_2 = \|I - K_S^{[k]}\|_2 \leq 2\,n\ TOL_k < \frac{1}{2} \ .$$

Wegen

$$\begin{aligned} \lambda_{\min}(K_S^{[k]}) &= \|K_S^{[k]^{-1}}\|_2^{-1} = \|(I - (I - K_S^{[k]}))^{-1}\|_2^{-1} \\ &\geq (\frac{1}{1 - \|I - K_S^{[k]}\|})^{-1} \geq 1 - 2\,n\ TOL_k \\ &> \frac{1}{2} \end{aligned}$$

ist

$$\|\delta K_S^{[k]}\| < \lambda_{\min}(K_S^{[k]}) \ .$$

Gemäß Algorithmus 2.2.2 sind die Diagonalelemente von $D^{[k]}$ für alle $k \geq 1$ wie die λ_i absteigend geordnet, und für jedes $k \geq 1$ haben $(K^{[k]}, J)$ und $(K^{[0]}, J)$ dieselben Eigenwerte. Die Voraussetzungen von Satz 1.0.6 sind daher für alle $k \geq 1$ erfüllt und aus (1.9) folgt

$$\frac{|\lambda_i - D_{ii}^{[k]^2}|}{\lambda_i} \leq \frac{\|\delta K_S^{[k]}\|_2}{\lambda_{\min}(K_S^{[k]})} < 4\,n\ TOL_k \ ,$$

also gilt für alle i, $1 \leq i \leq n$

$$\lim_{k\to\infty} |\lambda_i - D_{ii}^{[k]^2}| = \lambda_i \cdot 4\,n \lim_{k\to\infty} TOL_k = 0 \ ,$$

was die Aussage (2.80) ist. Q.E.D.

Bei einer vorgegeben Nullfolge von Abbruchkonstanten konvergiert die Folge der Ausgabematrizen des Verfahrens 2.2.2 also gegen eine feste Diagonalmatrix, deren Elemente die Imaginärteile der Eigenwerte von (K, J) sind, falls in exakter Arithmetik gerechnet wird. Für das implizite Verfahren bedeutet dies, daß die Folge der Elemente der ausgegebenen Diagonalmatrizen E gegen die Eigenwerte konvergiert. Die ausgegebenen Matrizen L, deren Spalten zu 1 normiert sind, sind Näherungen an die Real- bzw. Imaginärteile der Eigenvektoren der schiefsymmetrischen Matrix $S = LJL^T$.

Kapitel 3

Fehler in den berechneten Eigenwerten von $(\mathbf{L}^{\mathrm{T}}\mathbf{L}, \mathbf{J})$

Bei den Verfahren 2.2.2 bzw. 2.2.4 werden symplektische Transformationen

$$K^{(m+1)} = T^{(m)^T} K^{(m)} T^{(m)} \quad , \quad m = 1, 2, \ldots$$

bzw.

$$L^{(m+1)} = L^{(m)} T^{(m)} \quad , \quad m = 1, 2, \ldots$$

vorgenommen. Wird in endlicher Arithmetik gerechnet, so sind wir an Schranken für die maximalen relativen Fehler in den mit Verfahren 2.2.2 bzw. 2.2.4 berechneten Eigenwerten interessiert. In diesem Kapitel werden wir die Grundlagen zur Bestimmung derartiger Schranken für das implizite Verfahren legen. Mit Hilfe der Ergebnisse aus Kapitel 1 werden wir eine Rückwärtsfehleranalyse des impliziten Verfahrens vornehmen. Auf eine Fehleranlyse des expliziten Verfahrens verzichten wir.

In Kapitel 1 (s. Satz 1.0.5, Satz 1.0.6) sind die Fehlerschranken abhängig von einer Störung der skalierten iterierten Matrizen und von der Kondition $\kappa(K_S^{(m)}) = \kappa^2(L_S^{(m)})$ der skalierten iterierten Matrizen (*skalierte Kondition*) dargestellt worden. In Abschnitt 3.1 interessieren wir uns für die Folge der skalierten Konditionen der iterierten Matrizen. In Abschnitt 3.2 werden wir die recht aufwendige Fehleranalyse des impliziten Verfahrens vornehmen, indem wir den bei jeder einzelnen Transformation entstehenden Fehler auf eine Störung der jeweils skalierten untransformierten Matrix zurückführen. Durch Akkumulation werden wir dann in Kapitel 5 Schranken für den Gesamtfehler in den berechneten Eigenwerten von $(L^T L, J)$ bestimmen können.

3.1 Über die Kondition der skalierten Matrix

Ist $K^{(1)} = L^T L$ und L aus der der schiefsymmetrischen Zerlegung $S = LJL^T$ entstanden, so ist die Kondition $\kappa(K_S^{(1)})$ meist klein, was wir in Abschnitt 4.3

und Kapitel 6.2 vertiefen werden.

In diesem Abschnitt werden wir zeigen, daß die in Verfahren 2.2.2 bzw. 2.2.4 vorgenommenen Transformationen im schlechtesten Fall nur zu einem geringen Wachtum der skalierten Kondition führen. Die Konvergenz der $K_S^{(m)}$ gegen die Einheitsmatrix hat die Konvergenz der skalierten Kondition gegen 1 zur Folge; zumeist wird sich daher sogar ein Sinken der skalierten Kondition einstellen. Man kann also tatsächlich kleine $\kappa(K_S^{(m)})$ erwarten.

In Unterabschnitt 3.1.1 behandeln wir orthogonale Transformationen und Skalierungen und in Abschnitt 3.1.2 Transformationen bei modifizierten Zyklen. Die meisten der in Unterabschnitt 3.1.1 dargestellten Ergebnisse sind Modifikationen oder Verbesserungen der Resultate in [13]. Die Betrachtungen für explizite und implizite Verfahren sind wegen $\kappa^2(L_S) = \kappa(L_S^T L_S)$ äquivalent, so daß wir in diesem Abschnitt der Einfachkeit halber das explizite Verfahren behandelt werden.

Wir werden bei der Fehleranalyse (Abschnitt 3.2) sehen, daß das Wachstum von $\lambda_{\min}(K_S^{(m)})$ größere Bedeutung hat, als das Wachstum der eigentlichen skalierten Kondition $\kappa(K_S^{(m)})$, denn $\lambda_{\max}(K_S^{(m)}) \leq 2\,n$, falls $K^{(m)} \in I\!R^{2n\times 2n}$. Eine Schranke der skalierten Kondition ist also stets bekannt, wenn nur eine von $1/\lambda_{\min}(K_S^{(m)})$ bekannt ist. Statt der skalierten Kondition werden wir also vorwiegend das Verhalten von $\lambda_{\min}(K_S^{(m)})$ untersuchen. In diesem Abschnitt können wir exakte Rechnung voraussetzen.

3.1.1 Wachstumsschranken bei Skalierungen und Rotationen

Wir werden nun Schranken für das Wachstum der skalierten Kondition nach Skalierungen und orthogonalen Rotationen angegeben, wie sie in Teil **1** und **3** von Algorithmus 2.2.2 vorgenommen werden. Das Hauptergebnis ist Satz 3.1.5, in dem eine derartige Wachstumsschranke bei der Diagonalisierung einer Submatrix der Ordnung 4 bestimmt wird.

Das folgende einfache Lemma stellen wir voran.

Lemma 3.1.1 *Die symmetrische positiv definite Matrix $K = DK_SD$ mit $D = diag(K_{ii}^{1/2})$ sei von gerader Ordnung. Sei S diagonal und nichtsingulär und $T = S \oplus S^{-T}$. Für $K' = T^T K T = D' K'_S D'$ mit $D' = diag({K'_{ii}}^{1/2})$ gilt dann $K'_S = K_S$.*

Beweis:
trivial. Q.E.D.

Auch die orthogonalen Transformationen aus **3.1** von Algorithmus 2.2.2 ändern die skalierte Kondition nicht. Genauer gilt

Satz 3.1.2 *Sei* $K = DK_SD \in I\!R^{2n\times 2n}$, $D = diag(K_{ii}^{1/2})$ *mit* $K_{ii} = K_{n+i,n+i}$, $K_{i,n+i} = K_{n+i,i} = 0$, $i = 1\dots n$. *Sei*

$$K' = T^TKT = D'K'_SD' \quad , \quad D' = diag(K'^{\,1/2}_{ii})$$

für

$$T = \begin{bmatrix} diag(cs_i) & diag(sn_i) \\ -diag(sn_i) & diag(cs_i) \end{bmatrix} \quad mit \quad sn_i^2 + cs_i^2 = 1,\ i = 1\dots n.$$

Dann stimmen die Eigenwerte von K_S *mit denen von* K'_S *überein.*

Beweis:
Aus $K_{ii} = K_{n+i,n+i}$, $K_{i,n+i} = K_{n+i,i} = 0, i = 1\dots n$ folgt $K'_{ii} = K'_{n+i,n+i}$, $K'_{i,n+i} = K'_{n+i,i} = 0, i = 1\dots n$. Also gilt $D_{ii} = D_{i+n,i+n} = D'_{ii} = D'_{i+n,i+n}, i = 1\dots n$. Daraus folgt

$$K'_S = D'^{-1}T^TDK_SDTD'^{-1} = D'^{-1}DT^TK_STDD'^{-1} = T^TK_ST,$$

woraus sich schließlich die Aussage des Satzes ergibt. Q.E.D.

Zur Behandlung anderer orthogonaler Rotationen ist folgender Satz nützlich. Er besagt, daß beliebige othogonale Ebenenrotationen von Matrizen mit gleichen Diagonalelementen keinen ungünstigen Einfluß auf die skalierte Kondition haben. Derartige Transformationen verwenden wir in **3.3** von Verfahren 2.2.2 bzw. 2.2.4 (siehe auch Seite 47).

Satz 3.1.3 *Sei* $A = DA_SD \in I\!R^{n\times n}$ *symmetrisch und positiv definit mit* $D = diag(A_{ii}^{1/2})$. *Für fixierte* j, k, $1 \le j, k \le n$ *sei* $A_{jj} = A_{kk}$. *Eine orthogonale Matrix* $U \in I\!R^{n\times n}$ *unterscheide sich nur an den Positionen* (j, j), (j, k), (k, j), (k, k) *von der Einheitsmatrix. Dann gilt für* $A' = U^TAU = D'A'_SD'$, $D' = diag(A'^{\,1/2}_{ii})$

$$\lambda_{\min}(A'_S) \ \ge\ \frac{\lambda_{\min}(A_S)}{\min\{2, \lambda_{\max}(A_S)\}} \ .$$

Beweis:
Wir verwenden die Bezeichnung $\lambda_{\min}(X)$ für den kleinsten Eigenwert einer allgemeinen Matrix X. Es ist

$$A'_S = D'^{-1}U^TDA_SDUD'^{-1} = Z^TA_SZ \quad , \quad Z = DUD'^{-1}$$

und deshalb

$$\begin{aligned} \lambda_{\min}(A'_S) &= \|A'^{-1}_S\|_2^{-1} \ge \|Z^{-T}Z^{-1}\|_2^{-1}\|A_S^{-1}\|_2^{-1} \\ &= \|Z^{-T}Z^{-1}\|_2^{-1}\lambda_{\min}(A_S) \ . \end{aligned}$$

Wegen $D_{ii} = D'_{ii}, i \notin \{j,k\}$ und $D_{jj} = D_{kk}$ gilt weiter

$$\begin{aligned} \|Z^{-T}Z^{-1}\|_2 &= \lambda_{\max}(Z^{-T}Z^{-1}) = \lambda_{\max}(D^{-1}UD'^2U^TD^{-1}) \\ &= \max\{1, A_{jj}^{-1}D'^2_{jj}, A_{jj}^{-1}D'^2_{kk}\} \leq \min\{2, \lambda_{\max}(A_S)\} \ , \end{aligned}$$

weil $D'^2_{jj}, D'^2_{kk} \leq 2A_{jj}$ ist. Daraus folgt die Aussage des Satzes. Q.E.D.

Dies ziehen wir zum Beweis einer zu [13], Proposition 6.1, ähnlichen Aussage heran.

Satz 3.1.4 *Sei*

$$K^{(0)} = \begin{bmatrix} F^{(0)} & B^{(0)^T} \\ B^{(0)} & C^{(0)} \end{bmatrix} \in I\!R^{2n\times 2n} \quad , \quad C_{ii}^{(0)} = 1 \ , \quad i = 1\ldots n \ .$$

symmetrisch und positiv definit. Die Matrix $K^{(m)}$ entstehe durch $m \leq n/2$ nicht überlappende (und daher kommutative) Transformationen von $K^{(0)}$ mit symplektischen, orthogonalen Matrizen $T^{(r)} = U^{(r)} \oplus U^{(r)}, r = 1\ldots m$, wobei sich jedes $U^{(r)}$ höchstens in einer Submatrix der Ordnung 2 von der Einheitsmatrix unterscheide und jede Transformation $T^{(r)}$ ein Außerdiagonalelement $F_{ij}^{(r-1)} = F_{ij}^{(0)}$ von $F^{(r-1)}$ anulliert. Dann gilt mit $K^{(r)} = D^{(r)}K_S^{(r)}D^{(r)}$, $D^{(r)} = diag(\sqrt{K_{ii}^{(r)}})$

$$\lambda_{\min}(K_S^{(m)}) \ \geq \ \frac{\lambda_{\min}(K_S^{(0)})}{\min\{2, \lambda_{\max}(K_S^{(0)})\}}.$$

Es genügt, $C_{ii}^{(0)} = 1$ dort zu verlangen, wo $U = U^{(1)} \cdot \ldots \cdot U^{(m)}$ nichttrivial ist.

Beweis:
Es werde o.B.d.A. die Folge der Elemente an den Positionen

$$(1,2), (3,4), \ldots, (2m-1, 2m)$$

anulliert. Für $T = T^{(1)} \cdot \ldots \cdot T^{(m)}$ gilt dann

$$K_S^{(m)} = D^{(m)^{-1}}T^TD^{(0)}K_S^{(0)}D^{(0)}TD^{(m)^{-1}} = Z^TK_S^{(0)}Z \quad \text{mit} \quad Z = D^{(0)}TD^{(m)^{-1}}.$$

Z hat dieselbe Blockstruktur wie T. Führen wir für allgemeine X die Bezeichnung

$$X_p = \begin{bmatrix} X_{pp} & X_{p,p+1} \\ X_{p+1,p} & X_{p+1,p+1} \end{bmatrix}$$

ein, so gilt

$$(Z^{-T}Z^{-1})_p = \begin{cases} F_{S_p}^{(0)} & \text{für } p = 1, 3, \ldots, 2\,m-1 \\ T_pD_p^{(m)^2}T_p^T & \text{für } p = n+1, n+3, \ldots, n+2\,m-1 \ . \end{cases} \tag{3.1}$$

Die erste Zeile von (3.1) gilt wegen $Z_p^T F_{S_p}^{(0)} Z_p = I$ und die zweite Zeile gilt wegen $C_{ii}^{(0)} = 1$. An allen anderen Positionen, die von (3.1) nicht erfaßt werden, entspricht Z der Einheitsmatrix. Wegen

$$\begin{aligned} \lambda_{\min}(K_S^{(m)}) = \|{K_S^{(m)}}^{-1}\|_2^{-1} &\geq \|Z^{-T}Z^{-1}\|_2^{-1}\|{K_S^{(0)}}^{-1}\|_2^{-1} \\ &= \|Z^{-T}Z^{-1}\|_2^{-1}\lambda_{\min}(K_S^{(0)}) \end{aligned}$$

brauchen wir nur noch eine obere Schranke von $\|Z^{-T}Z^{-1}\|_2$ zu bestimmen. Aus Satz 3.1.3 und (3.1) folgt

$$\begin{aligned} \|Z^{-T}Z^{-1}\|_2 &= \lambda_{\max}(Z^{-T}Z^{-1}) \\ &= \max_{1\leq p\leq m}\{\lambda_{\max}(F_{S_p}^{(0)}), \lambda_{\max}(T_p {D_{p+n}^{(m)}}^2 T_p^T)\} \\ &\leq \min\{2, \lambda_{\max}(K_S^{(0)})\}. \end{aligned}$$

Daraus folgt die Aussage des Satzes. Q.E.D.

Ein weiteres Analogon zu [13], (Prop. 6.1) mit verbesserten Schranken ist

Satz 3.1.5 *Sei $K^{(0)} \in I\!R^{2n\times 2n}$ mit $K_{ii}^{(0)} = K_{i+n,i+n}^{(0)}$ und $K_{i,i+n}^{(0)} = K_{i+n,i}^{(0)} = 0$, $i = 1\ldots n$, symmetrisch und positiv definit. Die Matrix $K^{(m)}$ entstehe durch $m \leq n/2$ nicht überlappende (und daher kommutative) Transformationen von $K^{(0)}$ mit symplektischen Matrizen $T^{(r)}, r = 1\ldots m$, wobei sich jedes $T^{(r)}$ höchstens in einer Submatrix der Ordnung 4 von der Einheitsmatrix unterscheide und nach dem Algorithmus 2.2.2 gebildet werde. Dann gilt mit $K^{(r)} = D^{(r)}K_S^{(r)}D^{(r)}$, $D_r = diag(\sqrt{K_{ii}^{(r)}})$*

$$\lambda_{\min}(K_S^{(m)}) \geq \frac{\lambda_{\min}(K_S^{(0)})}{\min\{2, \lambda_{\max}(K_S^{(0)})\}} \tag{3.2}$$

$$\frac{\kappa(K_S^{(m)})}{\kappa(K_S^{(0)})} \leq \min\{\kappa(K_S^{(0)}), 2n-1\} \tag{3.3}$$

$$\frac{\kappa(K_S^{(i+1)})}{\kappa(K_S^{(i)})} \leq 4 \tag{3.4}$$

Beweis:
Wir verwenden die Beweistechnik aus [13]. Mit den Bezeichnungen

$$X_p = \begin{bmatrix} X_{pp} & X_{p,p+1} \\ X_{p+1,p} & X_{p+1,p+1} \end{bmatrix} \quad \text{und} \quad K_{[p]}^{(0)} = \begin{bmatrix} F_p^{(0)} & {B_p^{(0)}}^T \\ B_p^{(0)} & C_p^{(0)} \end{bmatrix}$$

(dabei ist X eine allgemeine Matrix) und analogen Bezeichnungen für die skalierten Matrizen $K_{S_{[p]}}^{(0)}$ werde o.B.d.A. die Folge der Submatrizen $K_{[p]}^{(0)}$, $p =$

$1, 3, \ldots 2m-1$, mit der Folge der Transformationen $T^{(1)}, \ldots, T^{(m)}$ diagonalisiert. Dann gilt für $T = T^{(1)} \cdot \ldots \cdot T^{(m)}$

$$K_S^{(m)} = Z^T K_S^{(0)} Z \quad \text{mit} \quad Z = D^{(0)} T D^{(m)^{-1}}.$$

Z hat dieselbe Blockstruktur wie T. Der nichttriviale Teil von $T^{(p)}$ werde mit $T_{[p]}$ bezeichnet und diese Bezeichnung werde für Z mit $Z_{[p]}$ und für die anderen Matrizen analog übernommen, Wegen $I = K_{S_{[p]}}^{(m)} = Z_{[p]}^T K_{S_{[p]}}^{(0)} Z_{[p]}$ gilt

$$K_{S_{[p]}}^{(0)} = Z_{[p]}^{-T} Z_{[p]}^{-1} \tag{3.5}$$

und wegen

$$\begin{aligned} \det(K_S^{(m)} - \lambda I) = 0 \quad &\Longleftrightarrow \quad \det(Z^T K_S^{(0)} Z - \lambda I) = 0 \\ &\Longleftrightarrow \quad \det(K_S^{(0)} - \lambda Z^{-T} Z^{-1}) = 0 \end{aligned} \tag{3.6}$$

folgt

$$\lambda_{\min}(K_S^{(m)}) = \min_{x \neq 0} \frac{x^T K_S^{(0)} x}{x^T Z^{-T} Z^{-1} x} \geq \frac{\min_{\|x\|=1} x^T K_S^{(0)} x}{\max_{\|x\|=1} x^T Z^{-T} Z^{-1} x} = \frac{\lambda_{\min}(K_S^{(0)})}{\|Z^{-T} Z^{-1}\|_2} \ . \tag{3.7}$$

Jede Matrix $T_{[p]}$ wird nach Algorithmus 2.2.2 als eine Folge von Skalierungen $S_{[p]}^i$ und orthogonalen Transformationen $U_{[p]}^i$ gebildet:

$$T_{[p]} = U_{[p]}^{3.1} S_{[p]}^{3.2} U_{[p]}^{3.3} S_{[p]}^{3.4} U_{[p]}^{3.5} S_{[p]}^{3.6} \quad \text{und} \quad T_{[p]}^T K_{[p]}^{(0)} T_{[p]} = D_{[p]}^{(m)^2},$$

wobei die oberen Indizes die Nummern der Anweisungen in Algorithmus 2.2.2 bezeichnen, also z.B. $U_{[p]}^{3.1}$ die orthogonale Matrix bezeichnet, mit der $K_{[p]}^{(0)}$ in Teil **3.1** von Algorithmus 2.2.2 transformiert wird. Wir setzen nun $T_{[p]} = U_{[p]}^{3.1} X_{[p]}$. Dann ist $X_{[p]}$ ein Produkt von Blockdiagonalmatrizen, die sich jeweils höchstens in einer Hauptuntermatrix der Ordnung 2 von der Einheitsmatrix unterscheiden. Es folgt

$$\begin{aligned} K_{S_{[p]}}^{(0)} = Z_{[p]}^{-T} Z_{[p]}^{-1} &= (D_{[p]}^{(0)} U_{[p]}^{3.1} X_{[p]} D_{[p]}^{(m)^{-1}})^{-T} (D_{[p]}^{(0)} U_{[p]}^{3.1} X_{[p]} D_{[p]}^{(0)^{-1}})^{-1} \\ &= D_{[p]}^{(0)^{-1}} U_{[p]}^{3.1} X_{[p]}^{-T} D_{[p]}^{(m)} D_{[p]}^{(m)} X_{[p]}^{-1} U_{[p]}^{3.1^T} D_{[p]}^{(0)^{-1}} \end{aligned}$$

und wegen der Kommutativität von $D_{[p]}^{(0)}$ und $U_{[p]}^{3.1}$ ist $K_{S_{[p]}}^{(0)}$ orthogonal ähnlich zu

$$\hat{K}_{S_{[p]}}^{(0)} = D_{[p]}^{(0)^{-1}} X_p^{-T} D_{[p]}^{(m)} D_{[p]}^{(m)} X_p^{-1} D_{[p]}^{(0)^{-1}}.$$

Diese Matrix hat 1 als Diagonalelemente und das rechts stehende Produkt besteht aus Blockdiagonalmatrizen, die sich jeweils höchtens in einer Submatrix der Ordnung 2 von der Einheitsmatrix unterscheiden. Also gilt

$$\lambda_{\max}(K_{S_{[p]}}^{(0)}) = \lambda_{\max}(\hat{K}_{S_{[p]}}^{(0)}) \leq 2$$

und

$$\begin{aligned} \|Z^{-T}Z^{-1}\|_2 &= \lambda_{\max}(Z^{-T}Z^{-1}) = \max_{1\le p\le m} \lambda_{\max}(K_{S_{[p]}}^{(0)}) \\ &\le \min\{2, \lambda_{\max}(K_S^{(0)})\}. \end{aligned} \tag{3.8}$$

Mit (3.7) folgt daraus (3.2). Aus (3.6) folgt außerdem

$$\begin{aligned} \lambda_{\max}(K_S^{(m)}) &= \max_{x\neq 0} \frac{x^T K_S^{(0)} x}{x^T Z^{-T} Z^{-1} x} \le \frac{\max_{\|x\|=1} x^T K_S^{(0)} x}{\min_{\|x\|=1} x^T Z^{-T} Z^{-1} x} \\ &\le \frac{\lambda_{\max}(K_S^{(0)})}{\lambda_{\min}(K_S^{(0)})} = \kappa(K_S^{(0)}) \end{aligned}$$

und daher wegen (3.2)

$$\kappa(K_S^{(m)}) \le \frac{\kappa(K_S^{(0)})}{\lambda_{\min}(K_S^{(m)})} \le \kappa^2(K_S^{(0)}) \ ,$$

was der erste Teil von (3.3) ist. Der zweite Teil von (3.3) ergibt sich mit (3.2) aus

$$\begin{aligned} \frac{\kappa(K_S^{(m)})}{\kappa(K_S^{(0)})} &= \frac{\lambda_{\max}(K_S^{(m)})\lambda_{\min}(K_S^{(0)})}{\lambda_{\min}(K_S^{(m)})\lambda_{\max}(K_S^{(0)})} \le \frac{\lambda_{\max}(K_S^{(m)})\lambda_{\min}(K_S^{(0)})}{\lambda_{\max}(K_S^{(0)})} \cdot \frac{\lambda_{\max}(K_S^{(0)})}{\lambda_{\min}(K_S^{(0)})} \\ &= \lambda_{\max}(K_S^{(m)}) \le 2n-1 \end{aligned}$$

Zeigt man noch $\lambda_{\max}(K_S^{(i+1)}) \le 2\lambda_{\max}(K_S^{(i)})$, so folgt (3.4), denn mit (3.2) gilt dann

$$\frac{\kappa(K_S^{(i+1)})}{\kappa(K_S^{(i)})} \le 2\frac{\lambda_{\min}(K_S^{(i)})}{\lambda_{\min}(K_S^{(i+1)})} \le 4.$$

Es genügt also, $\lambda_{\max}(K_S^{(1)}) \le 2\lambda_{\max}(K_S^{(0)})$ zu zeigen. Die zugehörige Transformation anulliere o.B.d.A das $(1,2)$-Element von $K^{(0)}$. Schreibt man

$$K_S^{(0)} = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \quad \text{mit} \quad A_{11} = K_{S_1}^{(0)} \in I\!R^{2\times 2},$$

so daß

$$K_S^{(1)} = Z^T K_S^{(0)} Z = \begin{bmatrix} A'_{11} & A'_{12} \\ A'^T_{12} & A'_{22} \end{bmatrix} \quad \text{mit} \quad A'_{11} = K_{S_1}^{(1)} = I,$$

so gibt es einen Einheitsvektor $x^T = [x_1^T \ x_2^T]$ in konformer Partitionierung mit

$$\lambda_{\max}(K_S^{(1)}) = \frac{x_1^T A_{11} x_1 + 2x_1^T A_{12} x_2 + x_2^T A_{22} x_2}{x_1^T A_{11} x_1 + x_2^T x_2}.$$

Schreibt man $\xi = x_1^T x_1$, $1-\xi = x_2^T x_2$, $x_1^T A_{11} x_1 = \tau_1 \xi$, $x_2^T A_{22} x_2 = \tau_2(1-\xi)$, so gilt wegen $2x_1^T A_{12} x_2 \leq x_1^T A_{11} x_1 + x_2^T A_{22} x_2$

$$\begin{aligned}\lambda_{\max}(K_S^{(1)}) &\leq 2\frac{x_1^T A_{11} x_1 + x_2^T A_{22} x_2}{x_1^T A_{11} x_1 + x_2^T x_2} = 2\frac{\tau_1\xi + \tau_2(1-\xi)}{\tau_1\xi + 1 - \xi} \\ &= 2(1 + \frac{(\xi-1)(1-\tau_2)}{\xi(\tau_1 - 1) + 1}).\end{aligned}$$

Der letztgenannte Term nimmt als Funktion von ξ sein Maximum an den Rändern des Intervalles $[0,1]$ an und es gilt $\lambda_{\max}(K_S^{(1)}) \leq 2\tau_2 \leq 2\lambda_{\max}(K_S^{(0)})$, was für (3.4) noch zu zeigen war. Q.E.D.

Jede einzelne der Transformationen aus Teil **1** und Teil **3** von Algorithmus 2.2.2 führt also höchstens zu einem Wachstum von $1/\lambda_{\min}(K_S^{(m)})$ um Faktor 2. Darüber hinaus ist die skalierte Kondition gegenüber Skalierungen (auch denen in **2.3.5** von Verfahren 2.2.2) invariant. Sogar die Akkumulation mehrerer Transformationen zur Diagonalisierung nicht überlappender 4×4-Submatrizen führt höchstens zu einem Wachstum von $1/\lambda_{\min}(K_S^{(m)})$ um Faktor 2. Dies wäre sicher nicht der Fall, wenn die Diagonalelemente von B nicht verschwinden würden.

Parallele Pivotstrategien ([5], [27]) haben also ein kleineres maximales Wachstum von $1/\lambda_{\min}(K_S^{(m)})$ zu Folge, als die zeilenzyklische Pivotstrategie (s. (2.36)). In der Praxis ist der Unterschied zumeist jedoch vernachlässigbar.

3.1.2 Wachstumsschranken bei modifizierten Zyklen

Im diesem Unterabschnitt bestimmen wir Schranken für das Wachstum der skalierten Kondition, wenn die Transformationen des modifizierten Zyklus angewendet werden. Dies sind die Transformationen (2.38) und (2.39). Wir verwenden die Notation des Abschnittes 2.1.

Satz 3.1.6 *Sei K symmetrisch und positiv definit. Für Submatrizen (2.35) seien die drei Bedingungen von Seite 44 erfüllt. K werde gemäß (2.38), (2.39) transformiert. Dann gilt mit $\bar{\omega} = \max_{1\leq r\leq m} \omega_r$*

$$\begin{aligned}\lambda_{\min}(K_S') &\geq \frac{1-\bar{\omega}}{(1+\bar{\omega})^4}\,\lambda_{\min}(K_S) \qquad (3.9)\\ &\geq 0.84\,\lambda_{\min}(K_S)\\ \lambda_{\min}(K_S'') &\geq (\frac{3}{2}\frac{(1+\bar{\omega})^3}{(1-\bar{\omega})^3}\bar{\omega}^2 + \frac{(1+\bar{\omega})^4}{(1-\bar{\omega})^2})^{-1}\\ &\quad \cdot(1 + \frac{\sqrt{1+\bar{\omega}}}{\sqrt{2}\sqrt{1-\bar{\omega}}}\bar{\omega})^{-2} \cdot \lambda_{\min}(K_S') \qquad (3.10)\\ &\geq 0.77\,\lambda_{\min}(K_S') \geq 0.64\,\lambda_{\min}(K_S)\ .\end{aligned}$$

Beweis:
Wir verwenden für alle beteiligten Matrizen dieselbe Partitionierung. Wie in Abschnitt 2.1 schreiben wir $\tilde{D} = \Delta^{-1} D$ und $L = \tilde{D} L_S$. Dann ist

$$\|\tilde{D}\|_2 \leq 1 \quad , \quad \|\tilde{D}^{-2}\|_2 \leq 1 + \bar{\omega}$$

und für alle r, $1 \leq r \leq m$,

$$\|L_{Sr}\|_2^2 = \|F_{Srr}\|_2 \leq 1 + \bar{\omega} \quad , \quad \|L_{S_r}^{-1}\|_2^2 = \|F_{Srr}^{-1}\|_2 \leq \frac{1}{1 - \bar{\omega}} \ .$$

Um (3.9) zu zeigen, schreiben wir

$$K_S' = Z^T K_S Z \quad , \quad Z = \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} L^{-T} & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \Delta^{-1} & 0 \\ 0 & \Delta^{-1} D'^{-1} \end{bmatrix} \ .$$

Es folgt

$$\lambda_{\min}(K_S') = \|(Z^T K_S Z)^{-1}\|_2^{-1} \geq \|Z^{-1}\|_2^{-2} \, \lambda_{\min}(K_S) \tag{3.11}$$

und wir werden nun eine obere Schranke von $\|Z^{-1}\|_2^2$ bestimmen. Es gilt

$$\|Z^{-1}\|_2^2 \leq \max_{1 \leq r \leq m} \max\{\|\Delta_r L_r^T D_{rr}^{-1}\|_2^2 \ , \ \|D_{rr}' \Delta_r L_r^{-1} D_{rr}^{-1}\|_2^2\} \ .$$

Wegen $\Delta_r^{-1} L_r^{-1} D_{rr} F_{Srr} D_{rr} L_r^{-T} \Delta_r^{-1} = I$ ist dabei

$$\|\Delta_r L_r^T D_{rr}^{-1}\|_2^2 = \|F_{Srr}\|_2 \leq 1 + \bar{\omega} \quad , \quad 1 \leq r \leq m \ . \tag{3.12}$$

Weiterhin haben wir

$$\begin{aligned} \|D_{rr}' \Delta_r L_r^{-1} D_{rr}^{-1}\|_2^2 &= \|D_{rr}' \Delta_r L_{S_{rr}}^{-1} \tilde{D}_{rr}^{-2} \Delta_r^{-1}\|_2^2 \\ &\leq \|D_{rr}'^2\|_2 \|L_{S_{rr}}^{-1}\|_2^2 \|\tilde{D}_{rr}^{-2}\|_2^2 \end{aligned}$$

und wegen

$$\begin{aligned} \|D_{rr}'\|_2^2 \leq \frac{1}{\delta_r^2} \|C_{rr}'\|_2 &\leq \frac{1}{\delta_r^2} \|L_r^T C_{rr} L_r\|_2 \leq \|\tilde{D}_{rr}^4\|_2 \|C_{Srr}\|_2 \|F_{Srr}\|_2 \\ &\leq (1 + \bar{\omega})^2 \end{aligned} \tag{3.13}$$

ist

$$\|D_{rr}' \Delta_r L_r^{-1} D_{rr}^{-1}\|_2^2 \leq \frac{(1 + \bar{\omega})^4}{1 - \bar{\omega}} \ . \tag{3.14}$$

(3.9) folgt nun aus (3.11), (3.12), (3.14). Zum Beweis von (3.10) gehen wir ähnlich vor. Wir schreiben

$$K_S'' = Z'^T K_S' Z' \ , \quad Z' = \begin{bmatrix} \Delta & 0 \\ 0 & \Delta D' \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} \Delta^{-1} & 0 \\ 0 & \Delta^{-1} D''^{-1} \end{bmatrix} \ ,$$

wobei X mit Δ kommutiert. Es folgt

$$\lambda_{\min}(K''_S) = \|(Z'^T K'_S Z')^{-1}\|_2^{-1} \geq \|Z'^{-1}\|_2^{-2}\, \lambda_{\min}(K'_S) \tag{3.15}$$

und wir schätzen $\|Z'^{-1}\|_2^2$ nach oben ab:

$$\|Z'^{-1}\|_2^2 \leq \max_{1\leq r\leq m} \left\{\max\{1\ ,\ \|D'^{-1}_{rr}\|_2^2 \|D''_{rr}\|_2^2\} \cdot (1+\|X_r\|_2)^2\right\}\ .$$

Dabei ist für alle r, $1 \leq r \leq m$,

$$\begin{aligned}
\|X_r\|_2 \leq \frac{1}{\delta_r^2}\|B'_{rr}\|_2 &= \frac{1}{\delta_r^2}\|L_r^T B_{rr} L_r^{-T}\|_2 \\
&= \frac{1}{\delta_r^2}\|L_{S_r}\tilde{D}_{rr}\Delta_r\tilde{D}_{rr}B_{S_{rr}}\tilde{D}_{rr}\Delta_r\tilde{D}_{rr}^{-1}L_{S_r}^{-1}\|_2 \\
&\leq \|\tilde{D}_{rr}\|_2^2\|F_{S_{rr}}\|_2^{1/2}\|B_{S_{rr}}\|_2\|F_{S_{rr}}^{-1}\|_2^{1/2} \\
&\leq \frac{\sqrt{1+\bar{\omega}}}{\sqrt{2}\sqrt{1-\bar{\omega}}}\,\bar{\omega}\ .
\end{aligned} \tag{3.16}$$

Außerdem ist

$$\begin{aligned}
\|D'^{-1}_{rr}\|_2^2 &\leq \|(\Delta_r^{-2}L_r^T C_{rr} L_r)^{-1}\|_2 = \|L_r^{-1}C_{rr}^{-1}L_r^{-T}\Delta_r^2\|_2 \\
&\leq \|L_{S_r}^{-1}\|_2^2\|\tilde{D}_{rr}^{-4}\|_2\|C_{S_{rr}}^{-1}\|_2 \leq \frac{(1+\bar{\omega})^2}{(1-\bar{\omega})^2}
\end{aligned}$$

und wegen (3.13), (3.16) gilt

$$\begin{aligned}
\|D''_{rr}\|_2^2 &\leq \frac{1}{\delta_r^2}\|X_r\Delta_r^2 X_r + B'_{rr}X_r + X_r B'^T_{rr} + C'_{rr}\|_2 \\
&\leq \|X_r\|_2^2 + \frac{2}{\delta_r^2}\,\|B'_{rr}\|_2\|X_r\|_2 + \frac{1}{\delta_r^2}\,\|C'_{rr}\|_2 \leq \frac{3}{\delta_r^4}\,\|B'_{rr}\|_2^2 + (1+\bar{\omega})^2 \\
&\leq 3\,\bar{\omega}^2\,\frac{1+\bar{\omega}}{2(1-\bar{\omega})} + (1+\bar{\omega})^2\ .
\end{aligned}$$

Insgesamt haben wir also

$$\|Z'^{-1}\|_2^2 \leq \left(\frac{3}{2}\,\frac{(1+\bar{\omega})^3}{(1-\bar{\omega})^3}\,\bar{\omega}^2 + \frac{(1+\bar{\omega})^4}{(1-\bar{\omega})^2}\right)\cdot\left(1+\frac{\sqrt{1+\bar{\omega}}}{\sqrt{2}\sqrt{1-\bar{\omega}}}\,\bar{\omega}\right)^2\ .$$

Mit Hilfe von (3.15) folgt daraus (3.10). Die konkreten Abschätzungen in (3.9) und (3.10) erhält man durch Einsetzen von $\bar{\omega} \leq 1/30$. Q.E.D.

Ist $\max \omega^{[l,u]}$ der Bedingungen von Seite 44 also klein genug, so führen die Transformationen (2.38), (2.39) nur zu einem geringen Wachstum der skalierten Kondition. Weil die Transformationen nur wenig von der Einheitsmatrix abweichen, war dies allerdings auch zu erwarten. Mit Satz 3.1.6 nahmen wir eine Quantifizierung des Wachstums vor.

3.2 Fehleranalyse des impliziten Verfahrens

In diesem Abschnitt werden wir die Fehleranalyse des Verfahrens 2.2.4 vornehmen. Dazu nehmen wir an, der Algorithmus 2.2.4 werde von einem Rechner in endlicher Genauigkeit ε ausgeführt. Konvergenz setzen wir voraus. Unter Rechnen in endlicher Genauigkeit ε verstehen wir die Anwendung den folgenden Modells für die elementaren Rechenoperationen. Ist $\mathrm{fl}(*)$ das in endlicher Genauigkeit ε erzielte Ergebnis einer solchen Operation $(*)$, so existieren ε_i mit $|\varepsilon_i| \leq \varepsilon$, so daß

$$\begin{aligned}
\mathrm{fl}(a \pm b) &= a(1+\varepsilon_1) \pm b(1+\varepsilon_2) \\
\mathrm{fl}(a \cdot b) &= a \cdot b(1+\varepsilon_3) \\
\mathrm{fl}(a/b) &= a/b(1+\varepsilon_4) \\
\mathrm{fl}(\sqrt{a}) &= \sqrt{a}(1+\varepsilon_5)\ .
\end{aligned}$$

Andere Rechenoperationen als diese elementaren werden nicht vorgenommen. Dabei wird $\varepsilon \ll 1$ i.A. die *Maschinengenauigkeit* sein, also die kleinste positive Zahl x, für die in endlicher Genauigkeit $1 + x > 1$ gilt.[1] Das Modell entspricht demjenigen in [13] und ist eine leichte Verallgemeinerung des Modells mit $\mathrm{fl}(a \pm b) = (a \pm b)(1 + \varepsilon_1)$. Es kann auch bei Rechnern ohne Guard-Digit (z.B. Cray) angewendet werden.

Wir nehmen an, daß Addition oder Subtraktion von 0 sowie Multiplikation mit oder Division durch ± 1 exakt durchgeführt werden. Wir nehmen weiter an, daß gewisse Werte wie z.B. $1/\sqrt{2}$ als Konstanten gespeichert sind und daher nur einen relativen Fehler von höchstens ε aufweisen. Daß in der weitverbreiteten IEEE-Arithmetik [3] manche Operationen, z.B. Multiplikation mit und Divisionen durch 2-er Potenzen, exakt sind, werden wir hingegen nicht berücksichtigen.

Variablenunterlauf wird wie in [13] beim verwendeten Modell nicht berücksichtigt. Daher gelten die Aussagen der Fehleranalyse unter der Bedingung, daß kein Unterlauf eintritt.

Die in diesem Abschnitt häufig verwendete Fomulierung „in erster Ordnung von ε“ bedeutet, daß die jeweils gemeinte (Un-)Gleichung exakt gilt, wenn das vorkommende ε durch $\varepsilon(1+c\,\varepsilon)$ mit einer mäßigen Konstante $c \in \mathbb{R}$ ersetzt wird. Kommt ε in der gemeinten (Un-)Gleichung nicht vor, so gilt sie nach Multiplikation mit $1 + c\,\varepsilon^2$ mit einem mäßigen $c \in \mathbb{R}$ exakt. Häufig wird die Konstante c von der Ordnung der betrachteten Matrizen oder Vektoren abhängen.

Die im Folgenden verwendeten, mit einem Index oder einem Indexpaar versehenen Werte ε_i bzw. ε_{ij} sind im Betrag sämtlich durch ε beschränkt ($|\varepsilon_i|, |\varepsilon_{ij}| \leq \varepsilon$). In erster Ordnung von ε werden darüber hinaus üblicherweise Näherungen wie $\sqrt{1+2k\varepsilon} = 1 + k\varepsilon$ oder, falls $\circ$ eine Multiplikation oder eine Division ist,

[1] Je nach Rundungsverfahren des Rechners (Compilers) kann ε auch die größte positive Zahl sein, für die in endlicher Genauigkeit $1 + x = 1$ gilt.

$(1 + k\varepsilon_1) \circ (1 + l\varepsilon_2) = 1 + (k + l)\varepsilon_3$, $k, l \in I\!N$, benutzt.

Zur Durchführung der Fehleranalyse werden wir zu jeder einzelnen, beim Verfahren 2.2.4 an L, E vorgenommenen Operation die Abschätzung einer Störung der Eingangsdaten bestimmen, unter der das Fließpunktergebnis in exakter Arithmetik erzielt worden wäre. Unter Verwendung der Ergebnisse von Kapitel 1 werden wir dann relative Fehlerschranken für die Eigenwerte der berechneten Matrizen angeben.

Wir gehen folgendermaßen vor. Nach einigen weiteren Vorbemerkungen wird im nachfolgenden Unterabschnitt eine kurze Fehleranalyse des Skalarproduktes vorgenommen. Darauf folgen die Analyse der im Verfahren 2.2.4 vorgenommenen Skalierungen und orthogonalen Rotationen sowie die Analyse der Transformationen im modifizierten Zyklus. Wir setzen durchgehend $n \geq 2$ voraus.

Sei $K = K^{(1)} = D^{(1)} K_S^{(1)} D^{(1)} \in I\!R^{2n \times 2n}$, $n \geq 2$, diejenige Matrix, für die das Eigenwertproblem

$$Kx = i\lambda Jx \quad , \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \quad , \quad i^2 = -1$$

gelöst werden soll und sei

$$K^{(m)} = D^{(m)} K_S^{(m)} D^{(m)} = D^{(m)} L_S^{(m)^T} L_S^{(m)} D^{(m)} = E^{(m)} L^{(m)^T} L^{(m)} E^{(m)} \quad , \tag{3.17}$$

wobei $E^{(m)}$ ein Diagonalmatrix ist und $L^{(m)} E^{(m)}$ aus $L^{(m-1)} E^{(m-1)}$ durch Anwendung einer einzelnen symlektischen Rechtsmultiplikation in Fließpunktarithmetik entsteht. Es sei stets (exakt) $D^{(m)} = \operatorname{diag}(\|L_{\cdot i}^{(m)}\|_2)$, also sind die euklidischen Normen der Spaltenvektoren von $L_S^{(m)}$ sämtlich 1. Wie schon in vorangehenden Abschnitt werden wir auch hier manchmal die Bezeichnungen

$$K^{(m)} = L^{(m)^T} L^{(m)} = \begin{bmatrix} F^{(m)} & B^{(m)^T} \\ B^{(m)} & C^{(m)} \end{bmatrix}$$

und

$$K_S^{(m)} = L_S^{(m)^T} L_S^{(m)} = \begin{bmatrix} F_S^{(m)} & B_S^{(m)^T} \\ B_S^{(m)} & C_S^{(m)} \end{bmatrix}$$

für $m \geq 1$ verwenden. Zur Erleichterung der Rückwärtsfehleranalyse werden wir zuerst darstellen, wie eine Störung von $E^{(m)}$ als Störung von $L^{(m)}$ dargestellt werden kann.

Lemma 3.2.1 *Sei $L = L_S \tilde{D} \in I\!R^{2n \times 2n}$ mit $\tilde{D} = diag(\|[L_{\cdot i}]\|_2)$. Seien $E, \delta E \in I\!R^{2n \times 2n}$ diagonal und E nichtsingulär mit $\delta E = \delta E_S E$. Sei $\varrho = rg(\delta E)$. Dann gibt es δL mit $L(E + \delta E) = (L + \delta L)E$ und für $\delta L_S = \delta L \tilde{D}^{-1}$ gilt*

$$\|\delta L_S\|_2 \leq \sqrt{\varrho}\, \|\delta E_S\|_2$$

Beweis:
Es ist $L(E+\delta E) = (L + L\delta E E^{-1})E$. Deshalb gilt für $\delta L = L\delta E E^{-1}$

$$\|\delta L_S\|_2 = \|L\delta E E^{-1}\tilde{D}^{-1}\|_2 = \|L\tilde{D}^{-1}\delta E_S\|_2 \le \|L_S\tilde{I}\|_2\,\|\delta E_S\|_2 \ ,$$

wobei $\tilde{I}$ eine Diagonalmatrix ist mit $\tilde{I}_{ii} = 0$, falls $\delta E_{ii} = 0$, und $\tilde{I}_{ii} = 1$ sonst. Wegen $\|L_S\tilde{I}\|_2 \le \sqrt{\varrho}$ folgt die Aussage des Lemmas. Q.E.D.

Wie sich die Fehler in den Eigenwerten von $(K^{(m)}, J)$ mit $K^{(m)}$ gemäß (3.17) für aufeinanderfolgende m akkumulieren, beschreiben wir mit nachfolgendem Lemma, dessen Aussage wir mehrfach verwenden werden.

Lemma 3.2.2 *Seien $L^{(1)}, E^{(1)} \in \mathbb{R}^{2n\times 2n}$, gegeben und für $1 \le m \le M-1$ entstehen $L^{(m+1)}, E^{(m+1)}$ durch Anwendung jeweils einer Rechtsmultiplikation in Fließpunktarithmetik aus den Matrizen $L^{(m)}, E^{(m)}$, d.h. es gelte folgendes Diagramm*

$$\begin{tikzcd}
L^{(m)}, E^{(m)} \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L^{(m)}"'] & L^{(m+1)}, E^{(m+1)} \\
L^{(m)} + \delta L^{(m)}, E^{(m)} \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der vertikale Pfeil, daß $L^{(m)} + \delta L^{(m)}$ durch Addition einer Störung aus $L^{(m)}$ hervorgeht. Der diagonale Pfeil bedeutet, daß $L^{(m+1)}, E^{(m+1)}$ als das Ergebnis einer Fließpunktoperation auch durch Anwendung einer exakten Rechtsmultiplikation mit einem symplektischen $T^{(m)}$ aus $(L^{(m)} + \delta L^{(m)})E^{(m)}$ entsteht, also $(L^{(m)} + \delta L^{(m)})E^{(m)}T^{(m)} = L^{(m+1)} \cdot E^{(m+1)}$. Seien

$$-\lambda_n^{(m)} \le \ldots - \lambda_1^{(m)} < 0 < \lambda_1^{(m)} \le \ldots \le \lambda_n^{(m)}$$

die Eigenwerte des Problems

$$(E^{(m)}{L^{(m)}}^T L^{(m)} E^{(m)})x = i\lambda J x$$

für $1 \le m \le M$. In erster Ordnung von ε sei

$$\frac{|\lambda_j^{(m)} - \lambda_j^{(m+1)}|}{\lambda_j^{(m)}} \le l_m\,\varepsilon \quad , \quad 1 \le j \le n \ , \quad 1 \le m \le M \ . \tag{3.18}$$

Dann gilt in erster Ordnung von ε

$$\frac{|\lambda_j^{(1)} - \lambda_j^{(M)}|}{\lambda_j^{(1)}} \le \sum_{m=1}^{M-1} l_m\,\varepsilon \quad , \quad 1 \le j \le n \ . \tag{3.19}$$

Beweis:
Für $1 \leq j \leq n$ gilt

$$\begin{aligned}
\frac{|\lambda_j^{(1)} - \lambda_j^{(M)}|}{\lambda_j^{(1)}} &= \frac{|\sum_{m=1}^{M-1}(\lambda_j^{(m)} - \lambda_j^{(m+1)})|}{\lambda_j^{(1)}} \\
&= \frac{1}{\lambda_j^{(1)}} \sum_{m=1}^{M-1} \lambda_j^{(m)} \frac{|\lambda_j^{(m)} - \lambda_j^{(m+1)}|}{\lambda_j^{(m)}} \\
&\leq \sum_{m=1}^{M-1} \prod_{k=1}^{m-1} (\frac{\lambda_j^{(k+1)}}{\lambda_j^{(k)}})\, l_m\, \varepsilon \ .
\end{aligned}$$

Wegen (3.18) ist $\lambda_j^{(k+1)}/\lambda_j^{(k)} = 1 + \mathcal{O}(\varepsilon)$ für $1 \leq k \leq M-1$ und deshalb $\prod_{k=1}^{m-1} \lambda_j^{(k+1)}/\lambda_j^{(k)} = 1 + \mathcal{O}(\varepsilon)$. Daraus folgt bereits die Aussage des Lemmas. Q.E.D.

3.2.1 Bemerkungen zum Skalarprodukt

Beim untersuchten Verfahren werden häufig Skalarprodukte verwendet. Es ist deshalb nützlich, zunächst Schranken für die Fließpunktfehler dieser Operation zu bestimmen. Es wird sich zeigen, daß Skalarprodukte i.A. einen geringeren maximalen absoluten Fehler aufweisen, wenn sie nicht nach dem üblichen Verfahren, sondern nach einem modifizierten Algorithmus (binary splitting) berechnet werden. Es gilt

Lemma 3.2.3 *Sei $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Die Berechnung der Summe $s = \sum_{i=1}^n x_i$ mit dem Algorithmus*
$s = 0$
FOR $i = 1$ *TO* n *DO* $s = s + x_i$
führt bei Rechnung in endlicher Genauigkeit ε zum Ergebnis

$$fl(s) = s + r \quad , \quad |r| \leq (n-1)\varepsilon \sum_{i=1}^n |x_i| \leq (n-1)\sqrt{n}\varepsilon \|x\|_2.$$

Beweis:
(Siehe auch [37], [35].) Die Aussage des Lemmas gilt trivialerweise für $n = 1$. Für $n \geq 2$ gilt:

$$\begin{aligned}
\mathrm{fl}(s) &= (1 + (n-1)\varepsilon_1)x_1 + \sum_{i=2}^n (1 + (n-i+1)\varepsilon_i)x_i \\
&= s + r
\end{aligned}$$

mit

$$\begin{aligned} |r| &= |(n-1)\varepsilon_1 x_1 + \sum_{i=2}^{n}(n-i+1)\varepsilon_i x_i| \\ &\leq (n-1)\varepsilon|x_1| + \sum_{i=2}^{n}(n-i+1)\varepsilon|x_i| \\ &\leq (n-1)\varepsilon\sum_{i=1}^{n}|x_i|. \end{aligned}$$

Der Rest der Aussage folgt aus der Cauchy-Schwarzschen Ungleichung. Q.E.D.

Daraus ergibt sich für das Skalarprodukt:

Lemma 3.2.4 *Seien $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in I\!R^n$. Die Berechnung des Skalarproduktes $\sigma = x^T y = \sum_{i=1}^{n} x_i y_i$ mit dem Algorithmus*
$\sigma = 0$
FOR $i = 1$ TO n DO $\sigma = \sigma + x_i y_i$
führt bei Rechnung in endlicher Genauigkeit ε zum Ergebnis

$$fl(x^T y) = x^T y + r \quad , \quad |r| \leq n\varepsilon\sum_{i=1}^{n}|x_i|\,|y_i| \leq n\varepsilon\|x\|_2\|y\|_2.$$

Beweis:
(Siehe auch [37], [35].) Die Aussage des Lemmas ist für $n = 1$ trivialerweise erfüllt. Für $n \geq 2$ ist der Algorithmus hinsichtlich der Fehleranalyse äquivalent zu
$\sigma = 0$
FOR $i = 1$ TO n DO $t_i = x_i y_i$
FOR $i = 1$ TO n DO $\sigma = \sigma + t_i$
und weil die t_i bis auf einen relativen Fehler ε_i, $|\varepsilon_i| \leq \varepsilon$ gebildet werden können, folgt die Aussage aus Lemma 3.2.3. Q.E.D.

Der Vollständigkeit halber folgt noch eine Aussage über den maximalen Fehler bei der Berechnung der Euklidischen Norm eines Vektors.

Lemma 3.2.5 *Sei $x = (x_1, \ldots, x_n) \in I\!R^n$. Die Berechnung der Euklidischen Norm $\|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2}$ mit dem Algorithmus*
$\sigma = 0$
FOR $i = 1$ TO n DO $\sigma = \sigma + x_i^2$
$\sigma = \sqrt{\sigma}$
führt bei Rechnung in endlicher Genauigkeit ε zum Ergebnis

$$fl(\|x\|_2) = \|x\|_2 + r \quad , \quad |r| \leq (\frac{n}{2} + 1)\varepsilon\|x\| \ .$$

Die Euklidische Norm wird also mit einem relativen Fehler berechnet.

Beweis:
Mit Hilfe von Lemma 3.2.4 gilt

$$\begin{aligned} \mathrm{fl}(\|x\|_2) &= (1+\varepsilon_1)(\mathrm{fl}(x^T x))^{1/2} \quad , |\varepsilon_1| \le \varepsilon \\ &= (1+\varepsilon_1)((1+\varepsilon_2)x^T x)^{1/2} \quad , |\varepsilon_2| \le n\,\varepsilon \\ &= (1+\varepsilon_3)\,\|x\|_2 \quad , |\varepsilon_3| \le \varepsilon + (n/2)\varepsilon \ . \end{aligned}$$

Q.E.D.

Kleinere Fehlerschranken können durch Verwendung modifizierter Algorithmen bestimmt werden. Es gilt folgender Satz, der ohne Beweis bereits in [9] enthalten ist.

Satz 3.2.6 *Sei $x = (x_1, \ldots, x_n) \in I\!R^n$. Die Berechnung der Summe $s = \sum_{i=1}^n x_i$ mit dem Algorithmus (binary splitting)*

```
k = n
WHILE k > 1 DO
   od = (k IS odd)
   k = ⌈k/2⌉
   IF od THEN
      FOR j = 1 TO k − 1 DO x_j = x_{2j−1} + x_{2j}
      x_k = x_{2k−1}
   ELSE
      FOR j = 1 TO k DO x_j = x_{2j−1} + x_{2j}
ENDWHILE
s = x_1
```

führt bei Rechnung in endlicher Genauigkeit ε zum Ergebnis

$$\mathit{fl}(s) = s + r \quad , \quad |r| \le \lceil \log_2 n \rceil \varepsilon \sum_{i=1}^{n} |x_i| \le \lceil \log_2 n \rceil \sqrt{n}\varepsilon \|x\|_2$$

Beweis:
Es sei vorab bemerkt, daß die Bildung von $\sum_{i=1}^n x_i$ aufgrund exakter Addition von 0 zum selben maximalen Fehler führt, wie

$$\sum_{i=1}^{2^{\lceil \log_2 n \rceil}} x_i' \quad \text{mit} \quad x_i' = \begin{cases} x_i & , \quad i \le n \\ 0 & , \quad \text{sonst.} \end{cases}$$

Der Satz kann daher mit vollständiger Induktion über m mit $n = 2^m$ gezeigt werden. Für $m = 0$ ist die Aussage des Satzes trivialerweise erfüllt. Für $m \to m+1$ ist zu beachten, daß mit $n_1 = 2^{m+1}$ die WHILE-Schleife des Algorithmus' einmal häufiger durchlaufen wird, als im Falle $n = 2^m$. Es gilt daher für den Induktionsschluß:

$$\mathrm{fl}(s) = \mathrm{fl}(\sum_{i=1}^{2^{m+1}} x_i) = \mathrm{fl}(\sum_{i=1}^{2^m} y_i)$$

mit

$$y_i = \mathrm{fl}(x_{2i} + x_{2i-1}) = x_{2i}(1+\varepsilon_{2i}) + x_{2i-1}(1+\varepsilon_{2i-1}) \ ,$$

also in erster Ordnung von ε

$$\mathrm{fl}(\sum_{i=1}^{2^{m+1}} x_i) = \sum_{i=1}^{2^m} y_i + r \quad , \quad |r| \le m\varepsilon \sum_{i=1}^{2^m} |y_i| \le m\varepsilon \sum_{i=1}^{2^{m+1}} |x_i| \ .$$

Wegen

$$\sum_{i=1}^{2^m} y_i = \sum_{i=1}^{2^{m+1}} x_i + r' \quad , \quad |r'| \le \varepsilon \sum_{i=1}^{2^{m+1}} |x_i|$$

folgt schließlich

$$\mathrm{fl}(\sum_{i=1}^{2^{m+1}} x_i) = \sum_{i=1}^{2^{m+1}} x_i + r_1 \quad , \quad |r_1| \le (m+1)\varepsilon \sum_{i=1}^{2^{m+1}} |x_i|.$$

Q.E.D.

Daraus ergibt sich für das Skalarprodukt:

Korollar 3.2.7 *Seien $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in I\!R^n$. Die Berechnung des Skalarproduktes $\sigma = x^T y = \sum_{i=1}^n x_i y_i$ mit dem Algorithmus*

```
FOR j = 1 TO n DO x_j = x_j · y_j
k = n
WHILE k > 1 DO
   od = (k IS odd)
   k = ⌈k/2⌉
   IF od THEN
      FOR j = 1 TO k − 1 DO x_j = x_{2j−1} + x_{2j}
      x_k = x_{2k−1}
   ELSE
      FOR j = 1 TO k DO x_j = x_{2j−1} + x_{2j}
ENDWHILE
σ = x_1
```

führt bei Rechnung in endlicher Genauigkeit ε zum Ergebnis

$$fl(x^T y) = x^T y + r$$

mit

$$|r| \le (\lceil \log_2 n \rceil + 1)\varepsilon \sum_{i=1}^n |x_i||y_i| \le (\lceil \log_2 n \rceil + 1)\varepsilon \|x\|_2 \|y\|_2$$

Beweis:
Weil die zuerst gebildeten Produkte zu einem relativen Fehler kleiner ε führen, folgt die Aussage aus Satz 3.2.6. Q.E.D.

Auch hier wollen wir auf die Herleitung des offensichtlichen maximalen relativen Fehlers bei der Berechnung der Euklidischen Norm eines Vektors verzichten.

Numerische Tests eines zu Korollar 3.2.7 äquivalenten, laufzeitoptimierten Programmes zur Berechnung von $\sigma = x^T y$ ergaben je nach Speicherverwaltung recht unterschiedliche Differenzen im Laufzeitverhalten gegenüber Lemma 3.2.4. Für $n \times m$-Matrizen mit $n = 50$ bis zu einigen hundert und $m = 10^3$ bis $m = 3 \cdot 10^4$, deren Spaltennormen berechnet wurden, ergaben sich längere Verarbeitungszeiten gegenüber dem Standardalgorithmus von etwa Faktor zwei. Hingegen konnte für Vektoren der Länge $n = 10^5$ keine meßbare Differenz im Laufzeitverhalten gegenüber dem Standardverfahren festgestellt werden. Größere Vektorlängen bis $n = 1.5 \cdot 10^6$ ergaben hingegen wieder einen höheren Zeitaufwand von bis zu 30%.

Bei den nachfolgenden Analysen werden wir die algorithmische Bestimmung von Skalarprodukten gemäß Korollar 3.2.7 nicht weiterverfolgen. Würde man Korollar 3.2.7 verwenden, so hätte dies nämlich lediglich einen größeren Einfluß auf die Aussage von Satz 3.2.8 und einigen Voraussetzungen bei der Analyse des modifizierten Zyklus (siehe dort). Insgesamt wären die erzielbaren Verbesserungen in den Aussagen nicht praxisrelevant.

3.2.2 Relative Fehlerschranken bei Skalierungen und Rotationen

In diesem Unterabschnitt betrachten wir die Teile **2**, **3** und **4** des Verfahrens 2.2.4. Wir werden zuerst die Fehleranalyse der Vorbereitung (Teil **2**) vornehmen. Wir benötigen dazu mehrere Sätze, deren Zuordnung zu den Anweisungen in Verfahren 2.2.4 wir der besseren Übersicht halber voranstellen wollen:

Nummer der Operation in 2.2.4	Satz für relativen Fehler
1.1.1	3.2.8 mit $\Delta = I, \Delta' = E$, $M = \{1 \dots 2n\}$ und (3.21)
1.1.2	3.2.9 mit $\Delta = E$, $M = \{1 \dots 2n\}$
1.1.3	3.2.10 mit $\Delta = E$, $M = \{1 \dots n\}$, $\tilde{\Delta} = I$

Wir beginnen mit

Satz 3.2.8 *Seien $L, \Delta = diag(\Delta_p) \in I\!R^{2n \times 2n}$ gegeben. Es werde folgender Algorithmus in endlicher Genauigkeit ε ausgeführt:*

let $M \subset \{1 \dots 2n\}$
FOR $p \in M$ DO
$\quad \Delta'_p = (\sum_{k=1}^{2n} L_{kp}^2)^{1/2} \cdot \Delta_p$
ENDFOR

Dann gilt folgendes Diagramm:

$$\begin{tikzcd} L, \Delta \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(\Delta') \\ L + \delta L, \Delta \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(\Delta')$ durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus den Eingangsmatrizen L, Δ entsteht. Der diagonale Pfeil bedeutet, daß $fl(\Delta')$ auch durch exakte Berechnung mit Hilfe desselben Algorithmus' aus Matrizen $L+\delta L, \Delta$ entsteht. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $\delta L = \delta L_S D$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \leq (n+2)\sqrt{|M|}\,\varepsilon \tag{3.20}$$

beschränkt. Gilt $\Delta_p = 1$ für alle $p \in M$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \leq (n+1)\sqrt{|M|}\,\varepsilon \tag{3.21}$$

beschränkt.

Beweis:
Sei $p \in M$. Wegen Lemma 3.2.5 gibt es ein $\tilde{\varepsilon}$ mit $|\tilde{\varepsilon}| \leq (n+2)\,\varepsilon$ und

$$\mathrm{fl}(\Delta'_p) = (1+\tilde{\varepsilon})\|[L_{\cdot p}]\|_2 \Delta_p = \|(1+\tilde{\varepsilon})[L_{\cdot p}]\|_2 \Delta_p = \|[L_{\cdot p}] + \tilde{\varepsilon}[L_{\cdot p}]\|_2 \Delta_p.$$

Für jedes $p \in M$ gibt es also eine Störung $[\delta L_{\cdot p}] = \tilde{\varepsilon}[L_{\cdot p}]$ mit $\|[\delta L_{\cdot p}]\|_2 \leq (n+2)\|[L_{\cdot p}]\|_2\,\varepsilon$, so daß $\mathrm{fl}(\Delta'_p)$ Ergebnis der exakten Durchführung des obigen Algorithmus' ist. Angewendet auf alle $p \in M$ gibt es daher eine Matrix δL, so daß das endlich genaue Ergebnis Δ' Ergebnis der exakten Durchführung des Algorithmus' für die Matrix $L + \delta L$ ist, wobei

$$\delta L = [\delta L_{\cdot 1} \cdots L_{\cdot 2n}] \quad , \quad \|\delta L_{\cdot p}\|_2 = \begin{cases} \leq (n+2)\|L_{\cdot p}\|_2 \varepsilon & p \in M \\ = 0 & \text{sonst} \end{cases}$$

ist. Durch Skalierung folgt

$$\begin{aligned} \|\delta L_S\|_2 &= \|\delta L D^{-1}\|_2 = \|[\delta L_{S\cdot 1} \cdots \delta L_{S\cdot 2n}]\|_2 \\ &\leq (\sum_{p \in M} \|\delta L_{S\cdot p}\|_2^2)^{1/2} \\ &\leq (n+2)\sqrt{|M|}\varepsilon \end{aligned}$$

Die Aussage (3.21) gilt wegen der fehlerfreien Multiplikation mit 1. Q.E.D.

Satz 3.2.9 *Seien $L, \Delta = diag(\Delta_p) \in I\!R^{2n\times 2n}$ mit $\Delta_p > 0, p = 1 \ldots 2n$, gegeben. Es werde folgender Algorithmus in endlicher Genauigkeit ε ausgeführt:*

let $M \subset \{1 \ldots 2n\}$
FOR $p \in M$ DO
$h = 1/\Delta_p$
FOR $k = 1$ TO $2n$ DO $L'_{kp} = h \cdot L_{kp}$
ENDFOR

Dann gilt folgendes Diagramm:

$$\begin{tikzcd} L, \Delta \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L') \\ L + \delta L, \Delta \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L')$ durch Fließpunkt-Berechnung mit Hilfe des obigen Programmes aus den Eingangsmatrizen L und Δ entsteht. Der diagonale Pfeil bedeutet, daß $fl(L')$ auch durch exakte Berechnung mit Hilfe desselben Algorithmus' aus der Matrix $L + \delta L$ entsteht. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $\delta L = \delta L_S D$ so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \;\leq\; 2\sqrt{|M|}\;\varepsilon \tag{3.22}$$

beschränkt.

Beweis:
Sei zunächst $p \in M$ fest. Dann gibt es für alle $k, 1 \leq k \leq 2n$, ein $\tilde{\varepsilon}_k$ mit $|\tilde{\varepsilon}_k| \leq 2\varepsilon$, so daß gilt

$$\mathrm{fl}(L'_{kp}) = (1 + \tilde{\varepsilon}_k) L_{kp} \Delta_p^{-1} = L_{kp}\Delta_p^{-1} + r_k.$$

Für $r = (r_1, \ldots, r_{2n})^T$ ist dann $\|r\| \leq 2\varepsilon \|[L_{\cdot p}]\|_2 \Delta_p^{-1}$ und damit

$$\begin{aligned} \mathrm{fl}([L'_{\cdot p}]) &= [L_{\cdot p}] \cdot \Delta_p^{-1} + r \\ &= ([L_{\cdot p}] + \Delta_p r)\Delta_p^{-1} \\ &= ([L_{\cdot p}] + [\delta L_{\cdot p}])\Delta_p^{-1} \end{aligned}$$

mit

$$\|[\delta L_{\cdot p}]\|_2 = \Delta_p \|r\|_2 \leq 2\varepsilon \|[L_{\cdot p}]\|_2.$$

Division durch $D_p = \|[L_{\cdot p}]\|_2$ bringt

$$\|[\delta L_{S\cdot p}]\|_2 = \|[\delta L_{\cdot p}] \cdot D_p^{-1}\|_2 \leq 2\varepsilon.$$

Anwendung auf alle $p \in M$ führt schließlich auf die Aussage des Satzes. Q.E.D.

Kann die Skalierung direkt an der Diagonalmatrix durchgeführt werden, so wird folgender Satz verwendet:

Satz 3.2.10 *Seien* $\Delta = diag(\Delta_p) \in I\!R^{2n\times 2n}$ *mit* $\Delta_p > 0, p = 1\ldots 2n$, *und* $\tilde{\Delta} = diag(\tilde{\Delta}_p) \in I\!R^{n\times n}$ *mit* $\tilde{\Delta}_p \neq 0, p = 1\ldots n$. *Es werde der Algorithmus*

let $M \subset \{1,\ldots,n\}$
FOR $p \in M$ *DO*
$\quad \Delta'_p = \Delta_p^{1/2}\Delta_{p+n}^{1/2}\tilde{\Delta}_p$
$\quad \Delta'_{p+n} = \Delta_p^{1/2}\Delta_{p+n}^{1/2}/\tilde{\Delta}_p$
ENDFOR

in endlicher Genauigkeit ε *durchgeführt. Dann gilt folgendes Diagramm:*

$$\begin{tikzcd}
\Delta, \tilde{\Delta} \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta\Delta"'] & fl(\Delta') \\
\Delta + \delta\Delta, \tilde{\Delta} \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(\Delta')$ *durch Fließpunkt-Berechnung mit Hilfe des obiges Programmes aus den Eingangsmatrizen* $\Delta, \tilde{\Delta}$ *entsteht. Der diagonale Pfeil bedeutet, daß* $fl(\Delta')$ *auch durch exakte Berechnung mit Hilfe desselben Algorithmus' aus* $\Delta + \delta\Delta, \tilde{\Delta}$ *entsteht. Schließlich bedeutet der senkrechte Pfeil, daß* $\Delta + \delta\Delta$ *durch Addition einer diagonalen Störung aus* Δ *hervorgeht. Für die diagonale Matrix* $\Delta + \delta\Delta$ *gilt in erster Ordnung von* ε

$$(\Delta + \delta\Delta)_{pp} = \Delta_p(1 + \delta_p) \quad , \quad |\delta_p| \leq 4\,\varepsilon \quad , \quad p \in M \vee p \in n + M \ ,$$

also $\|\delta\Delta_S\|_2 = \|\delta\Delta \cdot \Delta^{-1}\|_2 \leq 4\,\varepsilon$.

Beweis:
Sei $p \in M$ und $\circ$ eine Multiplikation oder eine Division. Dann gilt

$$\begin{aligned}
\mathrm{fl}(\Delta'_p) &= (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4)\Delta_p^{1/2}\Delta_{p+n}^{1/2} \circ \tilde{\Delta}_p \\
&= ([(1+4\varepsilon_5)\Delta_p][(1+4\varepsilon_6)\Delta_{p+n}])^{1/2} \circ \tilde{\Delta}_p \\
&= ((\Delta_p + \delta\Delta_p)(\Delta_{p+n} + \delta\Delta_{p+n}))^{1/2} \circ \tilde{\Delta}_p
\end{aligned}$$

mit $|\delta\Delta_p| \leq 4\varepsilon|\Delta_p|$ sowie $|\delta\Delta_{p+n}| \leq 4\varepsilon|\Delta_{p+n}|$. Die Aussage des Satzes folgt nun mit Division durch Δ_p bzw. Δ_{p+n}. Q.E.D.

Wir fassen nun die vorstehenden Ergebnisse zusammen und bestimmen mit Hilfe der Störungstheorie den maximalen relativen Fehler in den Eigenwerten des von der Vorbereitung erzeugten Matrizenpaares.

Satz 3.2.11 *Die Matrix $L \in \mathbb{R}^{2n\times 2n}$ werde gemäß Operationen* **1.1.1**, **1.1.2**, **1.1.3**, *von Verfahren 2.2.4 in endlicher Genauigkeit ε bearbeitet. Dabei sei $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $(n+1)\sqrt{2n}\,\varepsilon < \sigma_{\min}(L_S)$. Die daraus entstehenden Matrizen seien L' und E'. Dann gilt für die Eigenwerte*

$$-\lambda_n \le \ldots \le -\lambda_1 < 0 < \lambda_1 \le \ldots \le \lambda_n$$

bzw.

$$-\lambda'_n \le \ldots \le -\lambda'_1 < 0 < \lambda'_1 \le \ldots \le \lambda'_n$$

des Problems $L^T Lx = i\lambda Jx$ bzw. des Problems $E'L'^T L'E'x = i\lambda Jx$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \le \frac{(2\,n+14)\sqrt{2n}}{\sigma_{\min}(L_S)}\,\varepsilon\ , \quad k = 1\ldots n \tag{3.23}$$

Beweis:
Wir werden die Matrizen, die nach den im Satz genannten Operationen entstanden sind, mit einem oberen Index versehen. Z.B. seien $L^{1.1.3} = L'$, $E^{1.1.3} = E'$ die zuletzt entstandenen Matrizen. Die Störungen werden wir analog bezeichnen. Dann gilt

wegen	die Aussage
3.2.8	$\|\delta L_S\|_2 \le (n+1)\sqrt{2n}\,\varepsilon$
3.2.9	$\|\delta L_S^{1.1.1}\|_2 \le 2\sqrt{2n}\,\varepsilon$
3.2.10, 3.2.1	$\|\delta L_S^{1.1.2}\|_2 \le 4\sqrt{2n}\,\varepsilon$

Nach Satz 1.0.5 und Lemma 3.2.2 folgt für $1 \le k \le n$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \le \frac{2}{\sigma_{\min}(L_S)}\sqrt{2n}(n+7)\,\varepsilon\ ,$$

also die Aussage des Satzes. Q.E.D.

Als nächstes werden wir uns der Analyse der Operationen **1.2** und **1.3** zuwenden. Wir stellen wieder eine Tabelle mit den Zuordnungen der jeweiligen Sätze zu den Anweisungen in Verfahren 2.2.4 voran.

Nummer der Operation in 2.2.4	Satz für relativen Fehler
1.2.1	3.2.12
1.2.2	3.2.13 mit $\Delta = E$, $\tilde{\Delta} = (1/\sqrt{2})I$
1.2.3	3.2.8 mit $\Delta = E$
1.2.4	3.2.9 mit $\Delta = \mathrm{diag}(\delta_i)$
1.3	3.2.10 mit $\Delta = E$, $\tilde{\Delta} = I$

Ein Teil der Analyse ist also bereits erledigt. Es fehlen die Aussagen der folgenden beiden Sätze.

Satz 3.2.12 *Sei $L \in \mathbb{R}^{2n\times 2n}$ mit $\|[L_{\cdot i}]\|_2 = 1 + \mathcal{O}(\varepsilon), i = 1\ldots 2n$. Es sei $M \subset \{1\ldots n\}$ und $\tilde{I} = diag(\tilde{i}_{kk})$ mit $\tilde{i}_{kk} = 1$ für $k \in M$ und $\tilde{i}_{kk} = 0$ sonst. Es werde die Matrixmultiplikation*

$$L' = L \cdot \begin{bmatrix} I & \tilde{I} \\ \tilde{I} & I \end{bmatrix}$$

in endlicher Genauigkeit ε mittels folgenden Programmes durchgeführt:

FOR $p \in M$ DO
  FOR $k = 1$ TO $2n$ DO
    $L'_{kp} = L_{kp} - L_{k,p+n}$
    $L'_{k,p+n} = L_{kp} + L_{k,p+n}$
  ENDFOR
ENDFOR

Dann gilt folgendes Diagramm

$$\begin{tikzcd} L \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L') \\ L + \delta L \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L')$ durch Fließpunkt-Berechnung mit Hilfe des obigen Programmes aus der Eingangsmatrix L entsteht. Der diagonale Pfeil bedeutet, daß $fl(L')$ auch durch exakte Berechnung mit Hilfe desselben Algorithmus' aus der Matrix $L + \delta L$ entsteht. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $\delta L = \delta L_S D$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L\|_2 \;=\; \|\delta L_S\|_2 \;\leq\; 2\sqrt{2|M|}\;\varepsilon \tag{3.24}$$

beschränkt.

Beweis:
Sei o.B.d.A. $M = \{1\ldots m\}$ und sei p, $1 \leq p \leq m$ fest. Für alle k, $1 \leq k \leq 2n$, gibt es dann $\varepsilon_k, \varepsilon'_k, \varepsilon''_k, \varepsilon'''_k$ mit $|\varepsilon_k|, |\varepsilon'_k|, |\varepsilon''_k|, |\varepsilon'''_k| \leq \varepsilon$, so daß

$$\mathrm{fl}(L'_{kp}) = L_{kp}(1 + \varepsilon_k) - L_{k,p+n}(1 + \varepsilon'_k)$$

und

$$\mathrm{fl}(L'_{k,p+n}) = L_{kp}(1 + \varepsilon''_k) + L_{k,p+n}(1 + \varepsilon'''_k)\ .$$

Daraus folgt mit der Minkowskischen Ungleichung bis auf einen kleinen relativen Fehler

$$\mathrm{fl}(L'_{kp}) = L_{kp} - L_{k,p+n} + X_{kp} \quad , \quad \|X_{\cdot p}\|_2 \leq \varepsilon(\|L_{\cdot p}\|_2 + \|L_{\cdot p+n}\|_2) = 2\varepsilon$$

und

$$\mathrm{fl}(L'_{k,p+n}) = L_{kp} + L_{k,p+n} + X_{k,p+n} \quad , \quad \|X_{\cdot p+n}\|_2 \leq \varepsilon(\|L_{\cdot p}\|_2 + \|L_{\cdot p+n}\|_2) = 2\varepsilon \ ,$$

weil der Term $1 + \mathcal{O}(\varepsilon)$ jeweils ignoriert werden kann. Also gilt

$$\begin{aligned}
[L'_{\cdot p} \ \ L'_{\cdot p+n}] &= [L_{\cdot p} \ \ L_{\cdot p+n}] \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} + [X_{\cdot p} \ \ X_{\cdot p+n}] \\
&= ([L_{\cdot p} \ \ L_{\cdot p+n}] + \frac{1}{2}[X_{\cdot p} \ \ X_{\cdot p+n}] \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}) \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \\
&= ([L_{\cdot p} \ \ L_{\cdot p+n}] + [Y_{\cdot p} \ \ Y_{\cdot p+n}]) \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}
\end{aligned}$$

und es folgt

$$\|Y_{\cdot p}\|_2 \leq \frac{1}{2}(\|[X_{\cdot p}]\|_2 + \|[X_{\cdot p+n}]\|_2) \leq \frac{1}{2}(2\varepsilon + 2\varepsilon) = 2\varepsilon$$

und analog

$$\|Y_{\cdot p+n}\|_2 \leq 2\varepsilon$$

Also können wir mit $\delta L = [Y_{\cdot 1} \ldots Y_{\cdot 2n}]$ wegen

$$\|\delta L_S\|_2^2 = \|\delta L\|_2^2 \leq \|\delta L\|_E^2 = 2m(2\varepsilon)^2$$

auf $\|\delta L\|_2 = \|\delta L_S\|_2 \leq 2\sqrt{2m}\varepsilon$ schließen. Q.E.D.

Satz 3.2.13 *Sei $\Delta = diag(\Delta_p) \in \mathbb{R}^{m \times m}$. Es sei eine Diagonalmatrix $\tilde{\Delta}$ exakt berechnet und dann gespeichert. Dann führt für $M \subset \{1, \ldots, m\}$ die Bearbeitung des Programmes*

FOR $p \in M$ DO $\Delta'_p = \tilde{\Delta}_p \circ_p \Delta_p$

wobei $\circ_p$ eine Multiplikation oder eine Division ist, bei Rechnung in endlicher Genauigkeit ε auf folgendes Diagramm

$$\begin{tikzcd}
\Delta \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta\Delta"'] & fl(\Delta') \\
\Delta + \delta\Delta \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß fl(Δ') durch Fließpunkt-Berechnung mit Hilfe des obigen Programmes aus der Eingangsmatrix Δ entsteht. Der diagonale Pfeil bedeutet, daß fl(Δ') auch durch exakte Berechnung mit Hilfe desselben Algorithmus' aus der Matrix $\Delta + \delta\Delta$ entsteht. Schließlich bedeutet der senkrechte Pfeil, daß $\Delta + \delta\Delta$ durch Addition einer Störung aus Δ hervorgeht. Schreibt man mit $\delta\Delta = \delta\Delta_S\Delta$, so ist $\delta\Delta_S$ in erster Ordnung von ε durch

$$\|\delta\Delta_S\|_2 \;\leq\; \varepsilon \tag{3.25}$$

beschränkt.

Beweis:
Für alle $p, 1 \leq p \leq m$ gilt $\mathrm{fl}(\Delta'_p) = (1+\varepsilon_p)\tilde{\Delta}_p \circ_p \Delta_p = \tilde{\Delta}_p \circ_p (\Delta_p + \varepsilon_p\Delta_p)$. Also ist $|\delta\Delta_p| \leq \varepsilon\Delta_p$ für alle p und daher $\|\delta\Delta_S\|_2 \leq \varepsilon$. Q.E.D.

Wir fassen die Ergebnisse wieder zusammen und bestimmen mit Hilfe der Störungstheorie den maximalen relativen Fehler in den Eigenwerten des von **1.2** und **1.3** erzeugten Matrizenpaares.

Satz 3.2.14 *Die Matrizen $L, E \in \mathbb{R}^{2n\times 2n}$ werden gemäß den Operationen* **1.2** *und* **1.3** *von Verfahren 2.2.4 in endlicher Genauigkeit ε bearbeitet. Dabei sei $L = L_S D$ mit $D = diag(\|L_{.i}\|_2)$ und $(2\,n+4)\sqrt{n}\,\varepsilon < \sigma_{\min}(L_S)$. Die nach* **1.3** *entstandenen Matrizen seien L' und E'. Dann gilt für die Eigenwerte*

$$-\lambda_n \leq \ldots \leq -\lambda_1 < 0 < \lambda_1 \leq \ldots \leq \lambda_n$$

bzw.

$$-\lambda'_n \leq \ldots \leq -\lambda'_1 < 0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

des Problems $L^T Lx = i\lambda Jx$ bzw. des Problems $E'L'^T L'E'x = i\lambda Jx$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \leq \frac{(\sqrt{2}(2\,n+20)+4)\sqrt{2n}}{\sigma_{\min}(L_S)}\,\varepsilon\;, \quad k = 1\ldots n \tag{3.26}$$

Beweis:
Wie im Beweis von Satz 3.2.11 werden wir die nach den betrachteten Operationen von Algorithmus 2.2.4 jeweils entstandenen Matrizen mit einem oberen Index versehen. Z.B. seien $L^{1.3} = L'$, $E^{1.3} = E'$ die abschließend entstandenen Matrizen. Für die Störungen und die skalierten Matrizen werden wir entsprechend verfahren. Dann gilt

wegen	die Aussage
3.2.12	$\|\delta L_S\|_2 \leq 2\sqrt{2n}\,\varepsilon$
3.2.13,3.2.1	$\|\delta L_S^{1.2.1}\|_2 \leq 2\sqrt{2n}\,\varepsilon$
3.2.8	$\|\delta L_S^{1.2.2}\|_2 \leq (n+2)\sqrt{2n}\,\varepsilon$
3.2.9	$\|\delta L_S^{1.2.3}\|_2 \leq 2\sqrt{2n}\,\varepsilon$
3.2.10,3.2.1	$\|\delta L_S^{1.2.4}\|_2 \leq 4\sqrt{2n}\,\varepsilon$

Bei der Schranke von $\|\delta L_S^{1.2.1}\|_2$ haben wir berücksichtigt, daß der Wert $1/\sqrt{2}$ als Konstante gespeichert ist und daher nur einen relativen Fehler von höchstens ε hat.

Wegen Lemma 3.1.1 und Satz 3.1.4 beeinflußt nur die Operation **1.2.1** den kleinsten Singulärwert von L_S, d.h. es gilt $1/\sigma_{\min}(L_S^{1.2.1}) \leq \sqrt{2}/\sigma_{\min}(L_S)$ und $\sigma_{\min}(L_S^{1.2.1}) = \sigma_{\min}(L_S^{1.2.2}) = \ldots = \sigma_{\min}(L_S^{1.3})$. Nach Satz 1.0.5 und Lemma 3.2.2 folgt für $k = 1, \ldots, n$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda_k'|}{\lambda_k} \leq \frac{2}{\sigma_{\min}(L_S)} \sqrt{2n}(2 + \sqrt{2}(2 + (n+2) + 2 + 4))\,\varepsilon\ .$$

Daraus folgt die Behauptung. Q.E.D.

Wir könnten eine günstigere Abschätzung der Fehler in den Eigenwerten gewinnen, wenn wir berücksichtigten, daß die betrachteten Operationen evtl. nicht für alle $p = 1, \ldots n$ durchgeführt werden, und wenn wir im Beweis des letzten Satzes statt der der oberen Schranke für $1/\sigma_{\min}(L_S^{1.2.1})$ den tatsächlichen Wert berücksichtigten. Zum Erhalt einer gewissen Übersichtlichkeit wollen wir darauf aber verzichten.

Wir werden nun mit der Analyse der Diagonalisierung einer 4×4-Submatrix gemäß den Operationen **3.1** bis **3.6** von Verfahren 2.2.4 fortfahren. Auf den Teil **2** werden wir erst im Unterabschnitt 3.2.3 eingehen.

Zuerst stellen wir wieder die Zuordnung der Operationen zu den entsprechenden Sätzen tabellarisch dar:

Nummer der Operation in 2.2.4	Satz für relativen Fehler
3.1	3.2.15 bzw. 3.2.16 bzw. 3.2.17
3.2	3.2.18
3.3	3.2.19
3.4	3.2.18 mit Umbenennung der Variablen
3.5	3.2.19 mit Umbenennung der Variablen
3.6	3.2.10 mit $M = \{p, q\}$

Folgender Satz behandelt die Analyse von **3.1**. Bemerkenswert ist, daß die Fehler bei der Berechnung der Parameter $\tilde{t}_1 = \mathrm{fl}(t_1) = \tan\varphi$ und $\tilde{t}_2 = \mathrm{fl}(t_2) = \tan\psi$ nach [19], die die Singulärwertzerlegung von A bestimmen, nicht in die Analyse eingehen.

Die Fallunterscheidung im Algorithmus, die von der Größenordnung von $\tilde{t}_1$, $\tilde{t}_2$ anhängt, wird ausschließlich zur Vermeidung von Variablenunter- oder -überlauf vorgenommen. Ihre Ursache liegt nicht in der Fehleranalyse und hat kaum einen

Einfluß auf die Fehleranalyse. Sieht man von dieser Fallunterscheidung ab, so ist die Fehleranalyse unabhängig von der Größenordnung der $\tilde{t}_i$, denn aufgrund der Konstruktion des Algorithmus' kommutiert die Skalierung auf 1-Diagonale mit der Transformationsmatrix. Insofern unterscheidet sich die hier gegebene Fehleranalyse von der in [13].

In **3.1** gibt es eine Programmverzweigung je nach Größenordnung von $\tilde{t}_1$, $\tilde{t}_2$. Wir werden beispielhaft nur einen der Fälle behandeln, nämlich den Fall $\tilde{t}_1 \leq 1 < \tilde{t}_2$. Außerdem beschränken wir uns auf den Fall $|\tilde{t}_2| < macheps^{-1/2}$. Andernfalls werden zwar etwas bessere Fehlerschranken erzielt, zur Vereinfachung wollen wir jedoch eine einheitliche Behandlung aller Fälle sicherstellen.

Durch geeignete Programmierung könnte man im übrigen den Fall $1 < |\tilde{t}_1|, |\tilde{t}_2|$ ganz vermeiden. Dies ist für unsere Zwecke jedoch ohne Belang.

Satz 3.2.15 *Sei $1 \leq p < q \leq n$. Sei $LE = L_S D \in \mathbb{R}^{2n \times 2n}$ mit*

$$D = diag(\|[(LE)_{\cdot i}]\|_2) \quad , \quad \tilde{D} = diag(\|[L_{\cdot i}]\|_2) \ ,$$

also $\tilde{D}E = D$. Sei außerdem

$$D_p = D_{p+n}(1 + \mathcal{O}(\varepsilon)) \quad , \quad D_q = D_{q+n}(1 + \mathcal{O}(\varepsilon)) \tag{3.27}$$

und

$$L^T_{S_{\cdot p}} L_{S_{\cdot p+n}} = \mathcal{O}(\varepsilon) \quad , \quad L^T_{S_{\cdot q}} L_{S_{\cdot q+n}} = \mathcal{O}(\varepsilon) \ . \tag{3.28}$$

Es werde folgender Algorithmus in endlicher Genauigkeit ε für vorgegebene Werte $\tilde{t}_1$, $\tilde{t}_2$ mit $|\tilde{t}_1| \leq 1 < |\tilde{t}_2|$ durchgeführt:

$w_1 = \sqrt{1 + \tilde{t}_1^2}$
$sn_2 = \tilde{t}_2 / \sqrt{1 + \tilde{t}_2^2}$
$\tau = \ sign(\tilde{t}_2)$
$\tau_1 = \tilde{t}_1 E_p / E_{p+n}$
$\tau_1' = \tilde{t}_1 E_{p+n} / E_p$
$\tau_2 = E_q / (\tilde{t}_2 E_{q+n})$
$\tau_2' = E_{q+n} / (\tilde{t}_2 E_q)$
FOR $k = 1$ *TO* $2n$ *DO*
  $L'_{kp} = L_{kp} - \tau_1' L_{k,p+n}$
  $L'_{k,p+n} = \tau_1 L_{kp} + L_{k,p+n}$
  $L'_{kq} = \tau\,(\tau_2 L_{kq} - L_{k,q+n})$
  $L'_{k,q+n} = \tau\,(L_{kq} + \tau_2' L_{k,q+n})$
ENDFOR
$E'_p = E_p / w_1$
$E'_{p+n} = E_{p+n} / w_1$
$E'_q = |sn_2| E_{q+n}$
$E'_{q+n} = |sn_2| E_q$

Dann gilt folgendes Diagramm:

$$\begin{tikzcd} L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L'), fl(E') \\ L+\delta L, E \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L')$ und $fl(E')$ durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß eine symplektische Matrix T mit $(L+\delta L)ET = fl(L') \cdot fl(E')$ existiert. Die Matrizen $(L+\delta L)E$ und $fl(L') \cdot fl(E')$ gehen also durch exakte Rechtsmultiplikation mit einer symplektischen Matrix ineinander über. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S \tilde{D}$ und $\delta L = \delta L_S \tilde{D}$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \leq 26\ \varepsilon \tag{3.29}$$

beschränkt.

Beweis:
Wir verwenden die Beweistechnik aus [29], (Th. 3.3.3), erhalten insgesamt jedoch günstigere Abschätzungen. Die Werte $\tilde{t}_1$, $\tilde{t}_2$ seien gegeben. Wir nehmen o.B.d.A. $\tilde{t}_2 > 1$ an, weil das Vorzeichen bei der Fehleranalyse ohne Belang ist. In exakter Rechnung sei $w_1 = \sqrt{1+\tilde{t}_1^2}$ sowie $sn_2 = |sn_2| = \tilde{t}_2/\sqrt{1+\tilde{t}_2^2}$ und daher

$$\begin{aligned} \tilde{w}_1 &= \mathrm{fl}(w_1) \\ &= (1+\varepsilon_1)\sqrt{(1+\varepsilon_2)+(1+\varepsilon_3)(1+\varepsilon_4)\tilde{t}_1^2} \\ &= (1+\tilde{\varepsilon}_1)\sqrt{1+\tilde{t}_1^2} \end{aligned}$$

mit $|\tilde{\varepsilon}_1| \leq 2\varepsilon$ und auf ähnliche Weise

$$\widetilde{sn}_2 = \mathrm{fl}(sn_2) = (1+\tilde{\varepsilon}_2)\tilde{t}_2/\sqrt{1+\tilde{t}_2^2}$$

mit $|\tilde{\varepsilon}_2| \leq 3\varepsilon$. Weiter ist

$$\tilde{\tau}_1 = \mathrm{fl}(\tau_1) = (1+\hat{\varepsilon}_1)\tilde{t}_1 E_p/E_{p+n} \quad , \quad \tilde{\tau}_1' = \mathrm{fl}(\tau_1') = (1+\hat{\varepsilon}_1')\tilde{t}_1 E_{p+n}/E_p$$

mit $|\hat{\varepsilon}_1|$, $|\hat{\varepsilon}_1'| \leq 2\varepsilon$ und

$$\tilde{\tau}_2 = \mathrm{fl}(\tau_2) = (1+\hat{\varepsilon}_2)E_q/(\tilde{t}_2 E_{q+n}) \quad , \quad \tilde{\tau}_2' = \mathrm{fl}(\tau_2') = (1+\hat{\varepsilon}_2')E_{q+n}/(\tilde{t}_2 E_q)$$

mit $|\hat{\varepsilon}_2|$, $|\hat{\varepsilon}_2'| \leq 2\varepsilon$. Daraus folgt für alle $k, 1 \leq k \leq 2n$, mit $|\varepsilon_i| \leq \varepsilon$, $i = 1 \ldots 6$,

$$\begin{aligned} \mathrm{fl}(L_{kp}') &= (1+\varepsilon_1)L_{kp} - (1+\varepsilon_2)(1+\varepsilon_3)(1+\hat{\varepsilon}_1')\tilde{t}_1 L_{k,p+n}E_{p+n}/E_p \\ &= L_{kp} - \tilde{t}_1 L_{k,p+n}E_{p+n}/E_p + X_{kp} \\ \mathrm{fl}(L_{k,p+n}') &= (1+\varepsilon_4)(1+\varepsilon_5)(1+\hat{\varepsilon}_1)\tilde{t}_1 L_{kp}E_p/E_{p+n} + (1+\varepsilon_6)L_{k,p+n} \\ &= \tilde{t}_1 L_{kp}E_p/E_{p+n} + L_{k,p+n} + X_{k,p+n} \end{aligned}$$

mit

$$\|[X_{\cdot p}]\|_2 \leq \varepsilon \tilde{D}_p + 4\varepsilon \left|\tilde{t}_1\right| \tilde{D}_{p+n} E_{p+n}/E_p$$

und

$$\|[X_{\cdot p+n}]\|_2 \leq 4\varepsilon \left|\tilde{t}_1\right| \tilde{D}_p E_p/E_{p+n} + \varepsilon \tilde{D}_{p+n}.$$

Daraus ergibt sich

$$\begin{aligned}
&[\mathrm{fl}(L'_{\cdot p}) \;\; \mathrm{fl}(L'_{\cdot p+n})] \\
&= [L_{\cdot p} \;\; L_{\cdot p+n}] \begin{bmatrix} 1 & \tau_1 \\ -\tau'_1 & 1 \end{bmatrix} + [X_{\cdot p} \;\; X_{\cdot p+n}] \\
&= ([L_{\cdot p} \;\; L_{\cdot p+n}] + [X_{\cdot p} \;\; X_{\cdot p+n}] \begin{bmatrix} 1 & \tau_1 \\ -\tau'_1 & 1 \end{bmatrix}^{-1}) \begin{bmatrix} 1 & \tau_1 \\ -\tau'_1 & 1 \end{bmatrix} \\
&= ([L_{\cdot p} \;\; L_{\cdot p+n}] + [Y_{\cdot p} \;\; Y_{\cdot p+n}]) \begin{bmatrix} 1 & \tau_1 \\ -\tau'_1 & 1 \end{bmatrix} .
\end{aligned}$$

Dabei ist

$$[Y_{\cdot p} \;\; Y_{\cdot p+n}] = \frac{1}{1+\tau_1\tau'_1}[X_{\cdot p} + \tau'_1 X_{\cdot p+n} \quad -\tau_1 X_{\cdot p} + X_{\cdot p+n}]$$

und deshalb wegen $\tau_1\tau'_1 = \tilde{t}_1^2$

$$\begin{aligned}
\|[Y_{\cdot p}]\|_2 &= \frac{1}{1+\tilde{t}_1^2}\|[X_{\cdot p} + \tau'_1 X_{\cdot p+n}]\|_2 \\
&\leq \frac{1}{1+\tilde{t}_1^2}(\|[X_{\cdot p}]\|_2 + |\tau'_1|\,\|[X_{\cdot p+n}]\|_2) \\
&\leq \frac{1}{1+\tilde{t}_1^2}(\varepsilon\tilde{D}_p + 4\varepsilon|\tilde{t}_1|\tilde{D}_{p+n}E_{p+n}/E_p \\
&\qquad + |\tilde{t}_1|E_{p+n}/E_p(4\varepsilon|\tilde{t}_1|\tilde{D}_p E_p/E_{p+n} + \varepsilon\tilde{D}_{p+n})) \; .
\end{aligned}$$

Weil

$$E_p\tilde{D}_p = D_p = D_{p+n}(1+\mathcal{O}(\varepsilon)) = E_{p+n}\tilde{D}_{p+n}(1+\mathcal{O}(\varepsilon)) \tag{3.30}$$

ist, folgt bis auf einen kleinen relativen Fehler

$$\|[Y_{\cdot p}]\|_2 \leq \varepsilon\tilde{D}_p \frac{4\tilde{t}_1^2 + 5|\tilde{t}_1| + 1}{1+\tilde{t}_1^2}.$$

Der letztgenannte Bruch ist in $\left|\tilde{t}_1\right| \leq 1$ monoton wachsend. Es gilt also

$$\|[Y_{\cdot p}]\|_2 \leq 5\tilde{D}_p\varepsilon$$

und analog auch

$$\|[Y_{\cdot p+n}]\|_2 \leq 5\tilde{D}_{p+n}\varepsilon \; .$$

Außerdem gilt alle $k, 1 \leq k \leq 2n$

$$\begin{aligned}
\mathrm{fl}(L'_{kq}) &= (1+\varepsilon_1)(1+\varepsilon_2)(1+\hat{\varepsilon}_2)L_{kq}E_q/(\tilde{t}_2 E_{q+n}) - (1+\varepsilon_3)L_{k,q+n} \\
&= L_{kq}E_q/(\tilde{t}_2 E_{q+n}) - L_{k,q+n} + X_{kq} \\
\mathrm{fl}(L'_{k,q+n}) &= (1+\varepsilon_4)L_{kq} + (1+\varepsilon_5)(1+\varepsilon_6)(1+\hat{\varepsilon}'_2)L_{k,q+n}E_{q+n}/(\tilde{t}_2 E_q) \\
&= L_{kq} + L_{k,q+n}E_{q+n}/(\tilde{t}_2 E_q) + X_{k,q+n}
\end{aligned}$$

mit

$$\|[X_{\cdot q}]\|_2 \leq 4\varepsilon \tilde{D}_q E_q/(\tilde{t}_2 E_{q+n}) + \varepsilon \tilde{D}_{q+n}$$

und

$$\|[X_{\cdot q+n}]\|_2 \leq \varepsilon \tilde{D}_q + 4\varepsilon \tilde{D}_{q+n} E_{q+n}/(\tilde{t}_2 E_q).$$

Daraus ergibt sich

$$\begin{aligned}
&[\mathrm{fl}(L'_{\cdot q}) \quad \mathrm{fl}(L'_{\cdot q+n})] \\
&= [L_{\cdot q} \quad L_{\cdot q+n}] \begin{bmatrix} \tau_2 & 1 \\ -1 & \tau'_2 \end{bmatrix} + [X_{\cdot q} \quad X_{\cdot q+n}] \\
&= ([L_{\cdot q} \quad L_{\cdot q+n}] + [X_{\cdot q} \quad X_{\cdot q+n}] \begin{bmatrix} \tau_2 & 1 \\ -1 & \tau'_2 \end{bmatrix}^{-1}) \begin{bmatrix} \tau_2 & 1 \\ -1 & \tau'_2 \end{bmatrix} \\
&= ([L_{\cdot q} \quad L_{\cdot q+n}] + [Y_{\cdot q} \quad Y_{\cdot q+n}]) \begin{bmatrix} \tau_2 & 1 \\ -1 & \tau'_2 \end{bmatrix}
\end{aligned}$$

Dabei ist

$$[Y_{\cdot q} \quad Y_{\cdot q+n}] = \frac{1}{1+\tau_2\tau'_2}[\tau'_2 X_{\cdot q} + X_{\cdot q+n} \quad -X_{\cdot q} + \tau_2 X_{\cdot q+n}]$$

und deshalb wegen $\tau_2\tau'_2 = \tilde{t}_2^2$

$$\begin{aligned}
\|[Y_{\cdot q}]\|_2 &= \frac{1}{1+\tilde{t}_2^2}\|[\tau'_2 X_{\cdot q} + X_{\cdot q+n}]\|_2 \\
&\leq \frac{1}{1+\tilde{t}_2^2}(\tau'_2\|[X_{\cdot q}]\|_2 + \|[X_{\cdot q+n}]\|_2) \\
&\leq \frac{1}{1+\tilde{t}_2^2}(E_{q+n}/(\tilde{t}_2 E_q)(4\varepsilon \tilde{D}_q E_q/(\tilde{t}_2 E_{q+n}) + \varepsilon \tilde{D}_{q+n}) \\
&\qquad + \varepsilon \tilde{D}_q + 4\varepsilon \tilde{D}_{q+n} E_{q+n}/(\tilde{t}_2 E_q))
\end{aligned}$$

Weil

$$E_q \tilde{D}_q = D_q = D_{q+n}(1+\mathcal{O}(\varepsilon)) = E_{q+n}\tilde{D}_{q+n}(1+\mathcal{O}(\varepsilon)) \tag{3.31}$$

ist, folgt bis auf einen kleinen relativen Fehler

$$\|[Y_{\cdot q}]\|_2 \leq \varepsilon \tilde{D}_q \frac{\tilde{t}_2^2 + 5\tilde{t}_2 + 4}{\tilde{t}_2^2(1+\tilde{t}_2^2)}$$

Der letztgenannte Bruch ist in $\tilde{t}_2 \geq 1$ monoton fallend. Das Maximum wird daher bei $\tilde{t}_1 = 1$ mit dem Funktionswert 5 angenommen. Es gilt also

$$\|[Y_{\cdot q}]\|_2 \leq 5\tilde{D}_q\varepsilon$$

und analog auch

$$\|[Y_{\cdot q+n}]\|_2 \leq 5\tilde{D}_{q+n}\varepsilon \ .$$

Insgesamt folgt mit den Abkürzungen

$$\hat{Y} = [Y_{\cdot p}\ Y_{\cdot q}\ Y_{\cdot n+p}\ Y_{\cdot n+q}] \quad , \quad \hat{\tilde{D}} = \mathrm{diag}(\tilde{D}_p, \tilde{D}_q, \tilde{D}_{n+p}, \tilde{D}_{n+q})$$

bis auf einen kleinen relativen Fehler

$$\|\hat{Y}\hat{\tilde{D}}^{-1}\|_2 \leq 20\,\varepsilon \ . \tag{3.32}$$

Weiter gilt

$$\mathrm{fl}(E'_p) = \mathrm{fl}(E_p/\tilde{w}_1) = (1+\varepsilon_1)(1+\tilde{\varepsilon}_1)E_p/w_1 = (E_p + \delta E_p)/w_1$$

mit $|\delta E_p| \leq 3\varepsilon E_p$ sowie analog

$$\mathrm{fl}(E'_{p+n}) = \mathrm{fl}(E_{p+n}/\tilde{w}_1) = (E_{p+n} + \delta E_{p+n})/\tilde{w}_1$$

mit $|\delta E_{p+n}| \leq 3\varepsilon E_{p+n}$. Außerdem ist

$$\begin{aligned} \mathrm{fl}(E'_q) &= \mathrm{fl}(\widetilde{sn}_2 \cdot E_{q+n}) \\ &= (1+\varepsilon_1)(1+\tilde{\varepsilon}_2)sn_2E_{q+n} \\ &= sn_2(E_{q+n} + \delta E_{q+n}) \end{aligned}$$

mit $|\delta E_{q+n}| \leq 4\varepsilon E_{q+n}$ und analog

$$\mathrm{fl}(E'_{q+n}) = \mathrm{fl}(\widetilde{sn}_2 \cdot E_q) = sn_2(E_q + \delta E_q)$$

mit $|\delta E_q| \leq 4\varepsilon E_q$. Kürzen wir

$$\hat{E} = \mathrm{diag}(E_p, E_q, E_{n+p}, E_{n+q}) \quad , \quad \delta\hat{E} = \mathrm{diag}(\delta E_p, \delta E_q, \delta E_{n+p}, \delta E_{n+q})$$

ab, so gilt

$$\hat{E} + \delta\hat{E} = \hat{E}(I + \delta\hat{E}_S) \quad , \quad \|\delta\hat{E}_S\|_2 = \|\delta\hat{E}\hat{E}^{-1}\|_2 \leq 4\varepsilon$$

und daher mit $\hat{L} = [L_{\cdot p}\ L_{\cdot q}\ L_{\cdot n+p}\ L_{\cdot n+q}]$

$$[\mathrm{fl}(L'_{\cdot p})\ \ \mathrm{fl}(L'_{\cdot q})\ \ \mathrm{fl}(L'_{\cdot p+n})\ \ \mathrm{fl}(L'_{\cdot q+n})] \cdot \mathrm{diag}(\mathrm{fl}(E'_p), \mathrm{fl}(E'_q), \mathrm{fl}(E'_{p+n}), \mathrm{fl}(E'_{q+n}))$$

$$
\begin{aligned}
&= (\hat{L}+\hat{Y}) \cdot \begin{bmatrix} 1 & 0 & \tau_1 & 0 \\ 0 & \tau_2 & 0 & 1 \\ -\tau_1' & 0 & 1 & 0 \\ 0 & -1 & 0 & \tau_2' \end{bmatrix} \cdot \hat{E}(I+\delta\hat{E}_S)\mathrm{diag}(cs_1, sn_2, cs_1, sn_2) \\
&= (\hat{L}+\hat{Y})\hat{E} \cdot \begin{bmatrix} 1 & 0 & \tilde{t}_1 & 0 \\ 0 & 1/\tilde{t}_2 & 0 & 1 \\ -\tilde{t}_1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1/\tilde{t}_2 \end{bmatrix} \cdot \mathrm{diag}(cs_1, sn_2, cs_1, sn_2)(I+\delta\hat{E}_S) \\
&= (\hat{L}+\hat{Y})\hat{E}(I+\Delta) \cdot \begin{bmatrix} cs_1 & 0 & sn_1 & 0 \\ 0 & cs_2 & 0 & sn_2 \\ -sn_1 & 0 & cs_1 & 0 \\ 0 & -sn_2 & 0 & cs_2 \end{bmatrix}. \qquad (3.33)
\end{aligned}
$$

Ist U die orthogonale, symplektische Matrix aus (3.33), so ist $\Delta = U\delta\hat{E}_S U^T$. Den Ausdruck (3.33) wollen wir abschließend noch so umformen, daß nur noch eine Störung an $\hat{L}$ auftritt. Es gilt

$$
\begin{aligned}
(\hat{L}+\hat{Y})\hat{E}(I+\Delta) &= (\hat{L}+\hat{Y})\hat{E}(I+\Delta)\hat{E}^{-1}\hat{E} \\
&= (\hat{L}+\hat{L}\hat{E}\Delta\hat{E}^{-1}+\hat{Y}+\hat{Y}\hat{E}\Delta\hat{E}^{-1})\hat{E} \\
&= (\hat{L}+\delta\hat{L})\hat{E}
\end{aligned}
$$

und

$$
\begin{aligned}
\|\delta\hat{L}_S\|_2 &= \|\delta\hat{L}\hat{\tilde{D}}^{-1}\|_2 \\
&\leq \|\hat{L}\hat{\tilde{D}}^{-1}\hat{\tilde{D}}\hat{E}U\delta\hat{E}_S U^T\hat{E}^{-1}\hat{\tilde{D}}^{-1}\|_2 \\
&\quad + \|\hat{Y}\hat{\tilde{D}}^{-1}\|_2 + \|\hat{Y}\hat{\tilde{D}}^{-1}\hat{\tilde{D}}\hat{E}U\delta\hat{E}_S U^T\hat{E}^{-1}\hat{\tilde{D}}^{-1}\|_2 \,.
\end{aligned}
$$

Wegen (3.30), (3.31) gilt

$$
\|\hat{\tilde{D}}\hat{E}U(\hat{\tilde{D}}\hat{E})^{-1}\|_2 = 1 + \mathcal{O}(\varepsilon)
$$

und deshalb in erster Ordnung von ε

$$
\begin{aligned}
&\|\hat{L}\hat{\tilde{D}}^{-1}\hat{\tilde{D}}\hat{E}U\delta\hat{E}_S U^T\hat{E}^{-1}\hat{\tilde{D}}^{-1}\|_2 \\
&\quad \leq \|\hat{L}\hat{\tilde{D}}^{-1}\|_2\,\|\hat{\tilde{D}}\hat{E}U(\hat{\tilde{D}}\hat{E})^{-1}\|_2\,\|\delta\hat{E}_S\|_2\,\|\hat{\tilde{D}}\hat{E}U^T(\hat{\tilde{D}}\hat{E})^{-1}\|_2 \\
&\quad = \|\hat{L}\hat{\tilde{D}}^{-1}\|_2\,\|\delta\hat{E}_S\|_2 \,.
\end{aligned}
$$

Wegen (3.28) gilt $\|\hat{L}\hat{\tilde{D}}^{-1}\|_2 = \|(\hat{L}\hat{\tilde{D}}^{-1})^T\hat{L}\hat{\tilde{D}}^{-1}\|_2^{1/2} \leq \sqrt{2}$, so daß mit (3.32) in erster Ordnung von ε schließlich folgt

$$
\begin{aligned}
\|\delta L_S\|_2 = \|\delta\hat{L}_S\|_2 &\leq \sqrt{2}\,\|\delta\hat{E}_S\|_2 + \|\hat{Y}\hat{\tilde{D}}^{-1}\|_2 \\
&\leq \sqrt{2}\cdot 4\,\varepsilon + 20\,\varepsilon \\
&\leq 26\,\varepsilon
\end{aligned}
$$

Q.E.D.

Die beiden weiteren zur Analyse der Transformation **3.1** von Verfahren 2.2.4 notwendigen Sätze sind

Satz 3.2.16 *Sei $1 \leq p < q \leq n$. Sei $LE = L_S D \in \mathbb{R}^{2n \times 2n}$ mit*

$$D = diag(\|[(LE)_{\cdot i}]\|_2) \quad , \quad \tilde{D} = diag(\|[L_{\cdot i}]\|_2) \ ,$$

also $\tilde{D}E = D$. Sei außerdem $D_p = D_{p+n}(1 + \mathcal{O}(\varepsilon))$, $D_q = D_{q+n}(1 + \mathcal{O}(\varepsilon))$, $L_{\cdot p}^T L_{\cdot p+n} = \mathcal{O}(\varepsilon)$ und $L_{\cdot q}^T L_{\cdot q+n} = \mathcal{O}(\varepsilon)$. Es werde folgender Algorithmus in endlicher Genauigkeit ε für vorgegebene Werte $\tilde{t}_1$, $\tilde{t}_2$ mit $|\tilde{t}_1|, |\tilde{t}_2| \leq 1$ durchgeführt:

$w_1 = \sqrt{1 + \tilde{t}_1^2}$
$w_2 = \sqrt{1 + \tilde{t}_2^2}$
$\tau_1 = \tilde{t}_1 E_p / E_{p+n}$
$\tau_1' = \tilde{t}_1 E_{p+n} / E_p$
$\tau_2 = \tilde{t}_2 E_q / E_{q+n}$
$\tau_2' = \tilde{t}_2 E_{q+n} / E_q$
FOR $k = 1$ *TO* $2n$ *DO*
  $L_{kp}' = L_{kp} - \tau_1' L_{k,p+n}$
  $L_{k,p+n}' = \tau_1 L_{kp} + L_{k,p+n}$
  $L_{kq}' = L_{kq} - \tau_2' L_{k,q+n}$
  $L_{k,q+n}' = \tau_2 L_{kq} + L_{k,q+n}$
ENDFOR
$E_p' = E_p / w_1$
$E_{p+n}' = E_{p+n} / w_1$
$E_q' = E_q / w_2$
$E_{q+n}' = E_{q+n} / w_2$

Dann gilt folgendes Diagramm:

$$\begin{tikzcd}
L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L'), fl(E') \\
L + \delta L, E \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L')$ und $fl(E')$ durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß eine symplektische Matrix T mit $(L + \delta L)ET = fl(L') \cdot fl(E')$ existiert. Die Matrizen $(L + \delta L)E$ und $fl(L') \cdot fl(E')$ gehen also durch exakte Rechtsmultiplikation mit einer symplektischen Matrix ineinander über. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition

einer Störungen aus L hervorgeht. Schreibt man $L = L_S \tilde{D}$ und $\delta L = \delta L_S \tilde{D}$, so ist δL_S durch

$$\|\delta L_S\|_2 \leq 26\ \varepsilon \tag{3.34}$$

beschränkt.

Beweis:
Analog zu Satz 3.2.15. Wir könnten zwar eine etwas bessere Schranke für $\|\delta E_S\|_2$ bestimmen, wollen darauf jedoch verzichten, um später alle drei Typen von Rotationen in **3.1** von Verfahren 2.2.4 einheitlich behandeln zu können. Q.E.D.

Satz 3.2.17 *Sei $1 \leq p < q \leq n$. Sei $LE = L_S D \in \mathbb{R}^{2n \times 2n}$ mit*

$$D = diag(\|[(LE)_{\cdot i}]\|_2) \quad , \quad \tilde{D} = diag(\|[L_{\cdot i}]\|_2) \ ,$$

also $\tilde{D}E = D$. Sei außerdem $D_p = D_{p+n}(1 + \mathcal{O}(\varepsilon))$, $D_q = D_{q+n}(1 + \mathcal{O}(\varepsilon))$, $L_{\cdot p}^T L_{\cdot p+n} = \mathcal{O}(\varepsilon)$ und $L_{\cdot q}^T L_{\cdot q+n} = \mathcal{O}(\varepsilon)$. Es werde folgender Algorithmus in endlicher Genauigkeit ε für vorgegebene Werte $\tilde{t}_1$, $\tilde{t}_2$ mit $1 < |\tilde{t}_1|, |\tilde{t}_2|$ durchgeführt:

$sn_1 = \tilde{t}_1 / \sqrt{1 + \tilde{t}_1^2}$
$sn_2 = \tilde{t}_2 / \sqrt{1 + \tilde{t}_2^2}$
$\tau = \ sign(\tilde{t}_1)$
$\tau' = \ sign(\tilde{t}_2)$
$\tau_1 = E_p / (\tilde{t}_1 E_{p+n})$
$\tau_2' = E_{p+n} / (\tilde{t}_1 E_p)$
$\tau_2 = E_q / (\tilde{t}_2 E_{q+n})$
$\tau_2' = E_{q+n} / (\tilde{t}_2 E_q)$
FOR $k = 1$ *TO* $2n$ *DO*
  $L'_{kp} = \tau\,(\tau_1 L_{kp} - L_{k,p+n})$
  $L'_{k,p+n} = \tau\,(L_{kp} + \tau_1' L_{k,p+n})$
  $L'_{kq} = \tau'\,(\tau_2 L_{kq} - L_{k,q+n})$
  $L'_{k,q+n} = \tau'\,(L_{kq} + \tau_2' L_{k,q+n})$
ENDFOR
$E'_p = |sn_1| E_{p+n}$
$E'_{p+n} = |sn_1| E_p$
$E'_q = |sn_2| E_{q+n}$
$E'_{q+n} = |sn_2| E_q$

Dann gilt folgendes Diagramm:

$$\begin{tikzcd}
L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L'), fl(E') \\
L + \delta L, E \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L')$ und $fl(E')$ durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß eine symplektische Matrix T mit $(L + \delta L)ET = fl(L') \cdot fl(E')$ existiert. Die Matrizen $(L + \delta L)E$ und $fl(L') \cdot fl(E')$ gehen also durch exakte Rechtsmultiplikation mit einer symplektischen Matrix ineinander über. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S\tilde{D}$ und $\delta L = \delta L_S\tilde{D}$, so ist δL_S durch

$$\|\delta L_S\|_2 \leq 26\ \varepsilon \tag{3.35}$$

beschränkt.

Beweis:
Analog zu Satz 3.2.15. Q.E.D.

Wir werden nun den Teil **3.2** von Verfahren 2.2.4 behandeln, wobei wieder die Fehler in den berechneten $\tilde{\delta}_i = \mathrm{fl}(\delta_i)$ keine Rolle spielen.

Satz 3.2.18 *Sei $1 \leq p < q \leq n$. Sei $E \in \mathbb{R}^{2n\times 2n}$ nichtsinguläre Diagonalmatrix. Für gegebene Werte $\tilde{\delta}_1, \tilde{\delta}_2$ werde folgender Algorithmus in endlicher Genauigkeit ε durchgeführt:*
$E'_p = \tilde{\delta}_1 E_p E_{p+n}$
$E'_q = \tilde{\delta}_2 E_q E_{q+n}$
$E'_{p+n} = \tilde{\delta}_1^{-1}$
$E'_{q+n} = \tilde{\delta}_2^{-1}$
Dann gilt folgendes Diagramm:

$$\begin{tikzcd} E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta E"'] & fl(E') \\ E + \delta E \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(E')$ durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus der E E entsteht. Der diagonale Pfeil bedeutet, daß eine symplektische Matrix T existiert mit $(E + \delta E)T = fl(E')$ gilt. Die Matrizen $E + \delta E$ und $fl(E')$ gehen also durch exakte Rechtsmultiplikation mit einer symplektischen Matrix ineinander über. Schließlich bedeutet der senkrechte Pfeil, daß $E + \delta E$ durch Addition einer Störung aus E hervorgeht. Schreibt man $\delta E = \delta E_S E$, so ist δE_S durch

$$\|\delta E_S\|_2 \quad \leq \quad 2\ \varepsilon \tag{3.36}$$

beschränkt.

Beweis:
Es sei vorab bemerkt, daß $T = \mathrm{diag}(\tilde{\delta}_1 E_{p+n}, \tilde{\delta}_2 E_{q+n}, \tilde{\delta}_1^{-1} E_{p+n}^{-1}, \tilde{\delta}_2^{-1} E_{q+n}^{-1})$ symplektisch ist. Es gilt bis auf einen kleinen relativen Fehler

$$\begin{aligned} &\mathrm{fl}(\mathrm{diag}(E'_p, E'_q, E'_{p+n}, E'_{q+n})) \\ &= (\mathrm{diag}(E_p, E_q, E_{p+n}, E_{q+n}) \\ &\quad + \mathrm{diag}((\varepsilon_1 + \varepsilon_2)E_p, (\varepsilon_3 + \varepsilon_4)E_q, \varepsilon_5 E_{p+n}, \varepsilon_6 E_{q+n})) \cdot T \end{aligned}$$

Daraus folgt unmittelbar die Aussage. Q.E.D.

Satz 3.2.18 beschreibt unter geeigneter Umbenennung der Variablen auch die Fehleranalyse des Teiles **3.4** von Verfahren 2.2.4. Mit folgendem Satz analysieren wir Teil **3.3** von 2.2.4.

Satz 3.2.19 *Sei $1 \le p < q \le n$. Sei $LE = L_S D \in \mathbb{R}^{2n \times 2n}$ mit*

$$D = diag(\|[(LE)_{\cdot i}]\|_2) \quad , \quad \|[L_{S \cdot i}]\|_2 = 1 \ , \quad i = 1 \ldots 2n \ ,$$

mit einer diagonalen, positiv defititen Matrix E. Darüber hinaus sei $L = L_S \tilde{D}$ mit $\tilde{D} = diag(\|[L_{\cdot i}]\|_2)$, also $\tilde{D}E = D$. Dabei sei $D_{p+n} = D_{q+n}(1 + \mathcal{O}(\varepsilon))$. Es werde folgender Algorithmus in endlicher Genauigkeit ε durchgeführt:

$a = E_p^2 \sum_{k=1}^{2n} L_{kp}^2$
$b = E_q^2 \sum_{k=1}^{2n} L_{kq}^2$
$c = E_p E_q \sum_{k=1}^{2n} L_{kp} L_{kq}$
$\vartheta = (b - a)/(2c)$
$t = 1/(|\vartheta| + \sqrt{1 + \vartheta^2})$
IF $(c \ge 0 \ \wedge \ b < a) \ \vee \ (c < 0 \ \wedge \ b > a)$ *THEN* $t = -t$
$\tau_1 = tE_p/E_q$
$\tau_2 = -tE_q/E_p$
$\tau'_1 = tE_{p+n}/E_{q+n}$
$\tau'_2 = -tE_{q+n}/E_{p+n}$
FOR $k = 1$ *TO* $2n$ *DO*
  $L'_{kp} = L_{kp} + \tau_2 L_{kq}$
  $L'_{kq} = \tau_1 L_{kp} + L_{kq}$
  $L'_{k,p+n} = L_{k,p+n} + \tau'_2 L_{k,q+n}$
  $L'_{k,q+n} = \tau'_1 L_{k,p+n} + L_{k,q+n}$
ENDFOR
$diag(E'_p, E'_q, E'_{p+n}, E'_{q+n}) = diag(E_p, E_q, E_{p+n}, E_{q+n})/\sqrt{1 + t^2}$

Dann gilt folgendes Diagramm:

$$\begin{tikzcd} L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L'), fl(E') \\ L + \delta L, E \arrow[ur, "\text{exakt}"'] & \end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß fl(L') und fl(E') durch Fließpunkt-Berechnung mit Hilfe obigen Programmes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß eine symplektische Matrix T existiert, so daß $(L+\delta L)ET = \mathrm{fl}(L')\cdot\mathrm{fl}(E')$ gilt. Die Matrizen $(L+\delta L)E$ und fl$(L')\cdot$fl(E') gehen also durch exakte Rechtsmultiplikation mit einer symplektischen Matrix ineinander über. Schließlich bedeutet der senkrechte Pfeil, daß $L+\delta L$ durch Addition einer Störung aus L hervorgeht. δL_S ist in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \quad \leq \quad 50\ \varepsilon \tag{3.37}$$

beschränkt. Dasselbe gilt, wenn a, b auf andere Weise mit einem kleinen relativen Fehler berechnet werden, z. B. durch
$a = (E_{p+n}^2/E_p^2)\sum_{k=1}^{2n} L_{k,p+n}^2$
$b = (E_{q+n}^2/E_q^2)\sum_{k=1}^{2n} L_{k,q+n}^2$

Beweis:
Wir verwenden die Beweistechnik aus [13] bzw. [29]. Die wesentlichen Unterschiede zu [13] bestehen in der Verwendung schneller Rotationen im Algorithmus sowie darin, daß zwei orthogonale Rotationen der Ordnung 2 angewendet werden statt nur einer. Dennoch ist die in Störung (3.37) günstiger als die in [13] angegebene. Sie ist auch günstiger, als die Ergebnisse in [29] erwarten lassen.

Für die im Programm genannten a, b, c gilt

$$\begin{aligned}
\mathrm{fl}(a) &= (1+\varepsilon_a)a \quad , \quad \varepsilon_a = \mathcal{O}(\varepsilon) \\
\mathrm{fl}(b) &= (1+\varepsilon_b)b \quad , \quad \varepsilon_b = \mathcal{O}(\varepsilon) \\
\mathrm{fl}(c) &= c + \varepsilon_c D_p D_q \quad , \quad |\varepsilon_c| \leq (2n+2)\,\varepsilon\ .
\end{aligned}$$

Damit ist

$$\begin{aligned}
\mathrm{fl}(\vartheta) &= (1+\varepsilon_1)\frac{(1+\varepsilon_2)\mathrm{fl}(b) - (1+\varepsilon_3)\mathrm{fl}(a)}{(1+\varepsilon_4)2\mathrm{fl}(c)} \\
&= \frac{(1+\varepsilon_1)(1+\varepsilon_3)}{1+\varepsilon_4}\cdot\frac{\widetilde{\mathrm{fl}(b)} - \mathrm{fl}(a)}{2\mathrm{fl}(c)} \\
&= (1+\varepsilon_\vartheta)\tilde{\vartheta}
\end{aligned}$$

wobei wir

$$\widetilde{\mathrm{fl}(b)} = \frac{1+\varepsilon_2}{1+\varepsilon_3}\mathrm{fl}(b) = (1+\tilde{\varepsilon}_b)\mathrm{fl}(b) \quad , \quad |\tilde{\varepsilon}_b| \leq 2\varepsilon$$

und

$$\tilde{\vartheta} = \frac{\widetilde{\mathrm{fl}(b)} - \mathrm{fl}(a)}{2\mathrm{fl}(c)} \quad , \quad |\varepsilon_\vartheta| \leq 3\varepsilon$$

gesetzt haben. Weiter ist mit $\sigma = \operatorname{sign}((\mathrm{fl}(b) - \mathrm{fl}(a))/\mathrm{fl}(c))$

$$\begin{aligned}
\tilde{t} &= \mathrm{fl}(t) \\
&= \sigma\,(1+\varepsilon_1)\Big((1+\varepsilon_2)(1+\varepsilon_\vartheta)|\tilde{\vartheta}| \\
&\qquad +(1+\varepsilon_3)(1+\varepsilon_4)\sqrt{(1+\varepsilon_5)+(1+\varepsilon_6)(1+\varepsilon_7)(1+\varepsilon_\vartheta)^2\tilde{\vartheta}^2}\Big)^{-1} \\
&= (1+\varepsilon_t)\frac{\sigma}{|\tilde{\vartheta}|+\sqrt{1+\tilde{\vartheta}^2}} \quad , \quad |\varepsilon_t| \le 7\varepsilon
\end{aligned} \tag{3.38}$$

und mit $\tau = \sqrt{1+\tilde{t}^2}$ ist

$$\tilde{\tau} = \mathrm{fl}(\tau) = \sqrt{1+\tilde{t}^2}(1+\varepsilon_\tau) \quad , \quad |\varepsilon_\tau| \le 2\varepsilon. \tag{3.39}$$

Schließlich sind

$$\begin{aligned}
\tilde{\tau}_1 &= \mathrm{fl}(\tau_1) = (1+\varepsilon_{12})\tilde{t}E_p/E_q \\
\tilde{\tau}_2 &= \mathrm{fl}(\tau_2) = (1+\varepsilon_{21})\tilde{t}E_q/E_p \\
\tilde{\tau}_1' &= \mathrm{fl}(\tau_1') = (1+\varepsilon_{34})\tilde{t}E_{p+n}/E_{q+n} \\
\tilde{\tau}_2' &= \mathrm{fl}(\tau_2') = (1+\varepsilon_{43})\tilde{t}E_{q+n}/E_{p+n}
\end{aligned}$$

mit $|\varepsilon_{12}|, |\varepsilon_{21}|, |\varepsilon_{34}|, |\varepsilon_{43}| \le 2\varepsilon$. Daraus folgt für alle $k, 1 \le k \le 2n$

$$\begin{aligned}
\mathrm{fl}(L_{kp}') &= (1+\varepsilon_1)L_{kp} - (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_{21})\frac{\tilde{t}E_q}{E_p}L_{kq} \\
&= L_{kp} - \frac{\tilde{t}E_q}{E_p}L_{kq} + X_{kp}
\end{aligned}$$

mit

$$\|[X_{\cdot p}]\|_2 \le \varepsilon\tilde{D}_p + 4\varepsilon\frac{\left|\tilde{t}\right| E_q}{E_p}\tilde{D}_q$$

und analog

$$\mathrm{fl}(L_{kq}') = \frac{\tilde{t}E_p}{E_q}L_{kp} + L_{kq} + X_{kq}$$

mit

$$\|[X_{\cdot q}]\|_2 \le 4\varepsilon\frac{\left|\tilde{t}\right| E_p}{E_q}\tilde{D}_p + \varepsilon\tilde{D}_q \ .$$

Ähnlich ist

$$\mathrm{fl}(L_{k,p+n}') = L_{k,p+n} - \frac{\tilde{t}E_{q+n}}{E_{p+n}}L_{k,q+n} + X_{k,p+n}$$

mit

$$\|[X_{\cdot p+n}]\|_2 \le \varepsilon\tilde{D}_{p+n} + 4\varepsilon\frac{\left|\tilde{t}\right| E_{q+n}}{E_{p+n}}\tilde{D}_{q+n}$$

sowie

$$\mathrm{fl}(L'_{k,q+n}) = \frac{\tilde{t}E_{p+n}}{E_{q+n}} L_{k,p+n} + L_{k,p+n} + X_{k,q+n}$$

mit

$$\|[X_{\cdot q+n}]\|_2 \leq 4\varepsilon \frac{\left|\tilde{t}\right| E_{p+n}}{E_{q+n}} \tilde{D}_{p+n} + \varepsilon \tilde{D}_{q+n} \ .$$

Nun wird $\mathrm{fl}(E')$ betrachtet. Wegen (3.39) haben wir in erster Ordnung von ε

$$\begin{aligned}
&\mathrm{fl}(\begin{bmatrix} E'_p & 0 & 0 & 0 \\ 0 & E'_q & 0 & 0 \\ 0 & 0 & E'_{p+n} & 0 \\ 0 & 0 & 0 & E'_{q+n} \end{bmatrix}) \\
&= \frac{1+\varepsilon_\tau}{\tau} \cdot \begin{bmatrix} (1+\varepsilon_1)E_p & 0 & 0 & 0 \\ 0 & (1+\varepsilon_2)E_q & 0 & 0 \\ 0 & 0 & (1+\varepsilon_3)E_{p+n} & 0 \\ 0 & 0 & 0 & (1+\varepsilon_4)E_{q+n} \end{bmatrix} \\
&= \tau^{-1} \cdot \begin{bmatrix} E_p & 0 & 0 & 0 \\ 0 & E_q & 0 & 0 \\ 0 & 0 & E_{p+n} & 0 \\ 0 & 0 & 0 & E_{q+n} \end{bmatrix} (I + \delta\hat{E}_S) = \tau^{-1} \cdot \hat{E}(I + \delta\hat{E}_S) \qquad (3.40)
\end{aligned}$$

mit $\hat{E} = \mathrm{diag}(E_p, E_q, E_{p+n}, E_{q+n})$ und $\delta\hat{E}_S = \mathrm{diag}(\delta_p, \delta_q, \delta_{p+n}, \delta_{q+n})$. Dabei ist $\|\delta\hat{E}_S\|_2 \leq 3\varepsilon$. Wir definieren noch die exakt orthogonale Hilfsmatrix

$$U = \tau^{-1} \begin{bmatrix} 1 & \tilde{t} \\ -\tilde{t} & 1 \end{bmatrix} ,$$

und zur Abkürzung schreiben wir

$$\begin{aligned}
X &= [X_{\cdot p} \ \ X_{\cdot q} \ \ X_{\cdot p+n} \ \ X_{\cdot q+n}] \\
\hat{L} &= [L_{\cdot p} \ \ L_{\cdot q} \ \ L_{\cdot p+n} \ \ L_{\cdot q+n}] \\
\hat{\tilde{D}} &= \mathrm{diag}(\|L_{\cdot p}\|_2, \|L_{\cdot q}\|_2, \|L_{\cdot p+n}\|_2, \|L_{\cdot q+n}\|_2)
\end{aligned}$$

sowie

$$\hat{T} = \begin{bmatrix} 1 & \tau_1 \\ \tau_2 & 1 \end{bmatrix} \oplus \begin{bmatrix} 1 & \tau'_1 \\ \tau'_2 & 1 \end{bmatrix} .$$

Nun soll eine obere Schranke von $\|\delta\hat{L}_S\|_2 = \|\delta\hat{L}\hat{\tilde{D}}^{-1}\|_2$ gefunden werden, wobei $\delta\hat{L}$ durch

$$(\hat{L}\hat{T} + X)\hat{E}(I + \delta\hat{E}_S) \cdot \tau^{-1} = (\hat{L} + \delta\hat{L})\hat{E} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$$

definiert wird. Wir bestimmen $\delta\hat{L}$ aus

$$\begin{aligned}
&(\hat{L}T + X)\hat{E}(I + \delta\hat{E}_S)\cdot\tau^{-1} \\
&= (\hat{L}\hat{T} + \hat{L}\hat{T}\delta\hat{E}_S + X + X\delta\hat{E}_S)\hat{E}\cdot\tau^{-1} \\
&= (\hat{L} + \hat{L}\hat{T}\delta\hat{E}_S\hat{T}^{-1} + X\hat{T}^{-1} + X\delta\hat{E}_S\hat{T}^{-1})\hat{T}\hat{E}\cdot\tau^{-1}\ .
\end{aligned}$$

Dabei ist bereits $\hat{T}\hat{E}\cdot\tau^{-1} = \hat{E}\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$ und für $\delta\hat{L} = [Y_{\cdot p}\ \ Y_{\cdot q}\ \ Y_{\cdot p+n}\ \ Y_{\cdot q+n}]$ gilt

$$\begin{aligned}
[Y_{\cdot p}\ \ Y_{\cdot q}] = \frac{1}{1+\tilde{t}^2}([L_{\cdot p}\ \ L_{\cdot q}]\cdot &\begin{bmatrix} 1 & \frac{\tilde{t}E_p}{E_q} \\ -\frac{\tilde{t}E_q}{E_p} & 1 \end{bmatrix}\cdot\begin{bmatrix} \delta_p & -\delta_p\frac{\tilde{t}E_p}{E_q} \\ \delta_q\frac{\tilde{t}E_q}{E_p} & \delta_q \end{bmatrix} \\
&+[X_{\cdot p} + \frac{\tilde{t}E_q}{E_p}X_{\cdot q}\ \ -\frac{\tilde{t}E_p}{E_q}X_{\cdot p} + X_{\cdot q}] \\
&+[\delta_p X_{\cdot p} + \frac{\tilde{t}E_q}{E_p}\delta_q X_{\cdot q}\ \ -\frac{\tilde{t}E_p}{E_q}\delta_p X_{\cdot p} + \delta_q X_{\cdot q}]).
\end{aligned}$$

Bis auf einen kleinen relativen Fehler haben wir

$$\begin{aligned}
\|[Y_{\cdot p}]\|_2 &= \frac{1}{1+\tilde{t}^2}\|\delta_p([L_{\cdot p}] - \frac{\tilde{t}E_q}{E_p}[L_{\cdot q}]) \\
&\quad +\delta_q\frac{\tilde{t}E_q}{E_p}(\frac{\tilde{t}E_p}{E_q}[L_{\cdot p}] + [L_{\cdot q}]) + [X_{\cdot p}] + \frac{\tilde{t}E_q}{E_p}[X_{\cdot q}]\|_2 \\
&\le \frac{1}{1+\tilde{t}^2}(3\varepsilon(\tilde{D}_p + \frac{|\tilde{t}|E_q}{E_p}\tilde{D}_q) + 3\varepsilon\frac{|\tilde{t}|E_q}{E_p}(\frac{|\tilde{t}|E_p}{E_q}\tilde{D}_p + \tilde{D}_q) \\
&\quad +\varepsilon\tilde{D}_p + 4\varepsilon\frac{|\tilde{t}|E_q}{E_p}\tilde{D}_q + \frac{|\tilde{t}|E_q}{E_p}(4\varepsilon\frac{|\tilde{t}|E_p}{E_q}\tilde{D}_p + \varepsilon\tilde{D}_q)) \\
&= \frac{\varepsilon\tilde{D}_p}{1+\tilde{t}^2}(7\tilde{t}^2 + 11|\tilde{t}|\frac{D_q}{D_p} + 4)
\end{aligned}$$

Auf analoge Weise erhält man unter Beachtung von $D_{p+n}/D_{q+n} = 1 + \mathcal{O}(\varepsilon)$ bis auf einen kleinen relativen Fehler

$$\begin{aligned}
\|[Y_{\cdot q}]\|_2 &\le \frac{\varepsilon\tilde{D}_q}{1+\tilde{t}^2}(7\tilde{t}^2 + 11\left|\tilde{t}\right|\frac{D_p}{D_q} + 4) \\
\|[Y_{\cdot p+n}]\|_2 &\le \frac{\varepsilon\tilde{D}_{p+n}}{1+\tilde{t}^2}(7\tilde{t}^2 + 11\left|\tilde{t}\right|\frac{D_{q+n}}{D_{p+n}} + 4) = \frac{\varepsilon\tilde{D}_{p+n}}{1+\tilde{t}^2}(7\tilde{t}^2 + 11\left|\tilde{t}\right| + 4) \\
\|[Y_{\cdot q+n}]\|_2 &\le \frac{\varepsilon\tilde{D}_{q+n}}{1+\tilde{t}^2}(7\tilde{t}^2 + 11\left|\tilde{t}\right|\frac{D_{p+n}}{D_{q+n}} + 4) = \frac{\varepsilon\tilde{D}_{q+n}}{1+\tilde{t}^2}(7\tilde{t}^2 + 11\left|\tilde{t}\right| + 4).
\end{aligned}$$

Wir nehmen o.B.d.A. $x = D_q/D_p \le 1$ an. Andernfalls betrachten wir die permutierte Matrix. Für ein noch näher zu bestimmendes $\bar{x}$ mit z.B. $1/10 \le \bar{x} \le 9/10$

sei zunächst $x \le \bar{x}$. Dann ist

$$\|[Y_{\cdot p}]\|_2 \le \varepsilon \tilde{D}_p \frac{7\tilde{t}^2 + 11\bar{x}\left|\tilde{t}\right| + 4}{1+\tilde{t}^2}$$

und weil der Bruch auf der rechten Seite eine monoton wachsende Funktion von $\left|\tilde{t}\right|$ ist für alle $\bar{x}$, gilt

$$\|[Y_{\cdot p}]\|_2 \le \varepsilon \tilde{D}_p \frac{11(1+\bar{x})}{2} \text{ , da } \left|\tilde{t}\right| \le 1. \tag{3.41}$$

Weiter ist

$$\|[Y_{\cdot q}]\|_2 \le \varepsilon \tilde{D}_q \frac{7\tilde{t}^2 + 11|\tilde{t}|/\bar{x} + 4}{1+\tilde{t}^2}.$$

Wegen (3.38) gilt

$$\begin{aligned}
\left|\tilde{t}\right| &\le (1+\varepsilon_t)\frac{1}{2|\tilde{\vartheta}|} \\
&= (1+\varepsilon_t)\frac{|\mathrm{fl}(c)|}{\left|\widetilde{\mathrm{fl}(b)} - \mathrm{fl}(a)\right|} \\
&= (1+\varepsilon_t)\frac{|c+\varepsilon_c D_p D_q|}{|(1+\tilde{\varepsilon}_b)(1+\varepsilon_b)b - (1+\varepsilon_a)a|} \\
&= (1+\varepsilon_t)\frac{\left|c/D_p^2 + \varepsilon_c D_q/D_p\right|}{\left|(1+\tilde{\varepsilon}_b)(1+\varepsilon_b)b/D_p^2 - (1+\varepsilon_a)a/D_p^2\right|} \\
&= (1+\varepsilon_t)\frac{|c/(D_p D_q) + \varepsilon_c|\, x}{|(1+\tilde{\varepsilon}_b)(1+\varepsilon_b)x^2 - (1+\varepsilon_a)|}
\end{aligned}$$

mit

$$(1+\tilde{\varepsilon}_b)(1+\varepsilon_b)x^2 - (1+\varepsilon_a) = (x^2-1)(1+\varepsilon_g) \quad , \quad \varepsilon_g = \frac{(\tilde{\varepsilon}_b + \varepsilon_b)x^2 - \varepsilon_a}{x^2-1} = \mathcal{O}(\varepsilon),$$

so daß

$$\left|\tilde{t}\right| \le (1+\varepsilon_t)\frac{|c/(D_p D_q) + \varepsilon_c|\, x}{|x^2-1|\,(1+\varepsilon_g)}$$

folgt. Mit $|c/(D_p D_q) + \varepsilon_c| \le 1 + \mathcal{O}(\varepsilon)$ folgt schließlich

$$\left|\tilde{t}\right| \le \frac{\bar{x}}{1-\bar{x}^2}(1+\mathcal{O}(\varepsilon)).$$

Damit ist dann bis auf einen kleinen relativen Fehler

$$\|[Y_{\cdot q}]\|_2 \le \varepsilon \tilde{D}_q \frac{7\tilde{t}^2 + 11/(1-\bar{x}^2) + 4}{1+\tilde{t}^2} \le \varepsilon \tilde{D}_q (4 + \frac{11}{1-\bar{x}^2}), \tag{3.42}$$

weil der mittlere Bruch eine monoton fallende Funktion in $\left|\tilde{t}\right|$ ist für alle $\bar{x}$. Weiter ist

$$\|[Y_{\cdot p+n}]\|_2 \leq \varepsilon \tilde{D}_{p+n} \frac{7\tilde{t}^2 + 11\left|\tilde{t}\right| + 4}{1+\tilde{t}^2} \leq 11\varepsilon \tilde{D}_{p+n} \tag{3.43}$$

$$\|[Y_{\cdot q+n}]\|_2 \leq \varepsilon \tilde{D}_{q+n} \frac{7\tilde{t}^2 + 11\left|\tilde{t}\right| + 4}{1+\tilde{t}^2} \leq 11\varepsilon \tilde{D}_{q+n}, \tag{3.44}$$

weil die mittleren Brüche jeweils monoton wachsende Funktionen in $\left|\tilde{t}\right|$ sind. Addition von (3.41), (3.42), (3.43), (3.44) ergibt

$$\|\delta \hat{L}_S\|_2 \leq \varepsilon(\frac{11(1+\bar{x})}{2} + 4 + \frac{11}{1-\bar{x}^2} + 22) = \varepsilon(\frac{63}{2} + \frac{11}{2}\bar{x} + \frac{11}{1-\bar{x}^2}). \tag{3.45}$$

Sei nun $x > \bar{x}$. Dann ist

$$\begin{aligned}
\|[Y_{\cdot p}]\|_2 &\leq \varepsilon \tilde{D}_p \frac{7\tilde{t}^2 + 11\left|\tilde{t}\right| + 4}{1+\tilde{t}^2} \leq 11\varepsilon \tilde{D}_p \\
\|[Y_{\cdot q}]\|_2 &\leq \varepsilon \tilde{D}_q \frac{7\tilde{t}^2 + 11\left|\tilde{t}\right|/\bar{x} + 4}{1+\tilde{t}^2} \leq \varepsilon \tilde{D}_q(\frac{11}{2} + \frac{11}{2\bar{x}}) \\
\|[Y_{\cdot p+n}]\|_2 &\leq 11\varepsilon \tilde{D}_{p+n} \\
\|[Y_{\cdot q+n}]\|_2 &\leq 11\varepsilon \tilde{D}_{q+n},
\end{aligned}$$

weil die mittleren Terme der ersten beiden Ungleichungen als Funktionen von $|\tilde{t}|$ monoton wachsen für alle $\bar{x}$. Also gilt im Falle $x > \bar{x}$ durch Addition

$$\|\delta \hat{L}_S\|_2 \leq \varepsilon(\frac{77}{2} + \frac{11}{2\bar{x}}). \tag{3.46}$$

Ein für beide Abschätzungen (3.45), (3.46) optimales $\bar{x}$ ist $\bar{x} = 0.52$. In beiden Fällen (also für $x \leq \bar{x}$ und für $x > \bar{x}$) gilt

$$\|\delta L_S\|_2 = \|\delta \hat{L}_S\|_2 \leq 50\varepsilon.$$

Q.E.D.

Im vorstehenden Satz haben wir ausschließlich den Fall $|\vartheta| \leq macheps^{-1/2}$ behandelt. Der andere Fall $|\vartheta| > macheps^{-1/2}$ zeichnet sich gegenüber dem behandelten durch einen kleineren Fehler im berechneten $\tilde{t} = \mathrm{fl}(t)$ aus. Obwohl dann auch die Abschätzung (3.37) etwas verbessert werden kann, werden wir aus Vereinfachungsgründen auf ein eigenes Resultat verzichten.

Wir fassen die Ergebnisse wieder zusammen und geben mit folgendem Satz den maximalen relativen Fehler in den berechneten Eigenwerten des nach einer Diagonalisierung erzeugten Matrizenpaares an.

Satz 3.2.20 *Die Matrizen $L, E \in \mathbb{R}^{2n\times 2n}$ werden gemäß den Operationen* **3.1** *bis* **3.6** *von Verfahren 2.2.4 in endlicher Genauigkeit ε bearbeitet. Dabei sei $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $100\,\varepsilon < \sigma_{\min}(L_S)$. Die nach* **3.6** *entstandenen Matrizen seien L' und E'. Dann gilt für die Eigenwerte*

$$-\lambda_n \le \ldots \le -\lambda_1 < 0 < \lambda_1 \le \ldots \le \lambda_n$$

bzw.

$$-\lambda'_n \le \ldots \le -\lambda'_1 < 0 < \lambda'_1 \le \ldots \le \lambda'_n$$

des Problems $L^T L x = i\lambda J x$ bzw. des Problems $E' L'^T L' E' x = i\lambda J x$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \le \frac{160 + 156\sqrt{2}}{\sigma_{\min}(L_S)}\,\varepsilon\ , \quad k = 1\ldots n \tag{3.47}$$

Beweis:
Wir werden wieder die nach den betrachteten Operationen von Verfahren 2.2.4 jeweils entstandenen Matrizen mit einem oberen Index versehen. Z.B. seien $L^{3.6} = L'$, $E^{3.6} = E'$ die abschließend entstandenen Matrizen. Für die Störungen und die skalierten Matrizen werden wir entsprechend verfahren. Dann gilt

wegen	die Aussage
3.2.15,3.2.16,3.2.17	$\|\delta L_S\|_2 \le 26\,\varepsilon$
3.2.18,3.2.1	$\|\delta L_S^{3.1}\|_2 \le 2\sqrt{4}\,\varepsilon$
3.2.19	$\|\delta L_S^{3.2}\|_2 \le 50\,\varepsilon$
3.2.18,3.2.1	$\|\delta L_S^{3.3}\|_2 \le 2\sqrt{4}\,\varepsilon$
3.2.19	$\|\delta L_S^{3.4}\|_2 \le 50\,\varepsilon$
3.2.10,3.2.1	$\|\delta L_S^{3.5}\|_2 \le 12\sqrt{4}\,\varepsilon$

Bei der Abschätzung von $\|\delta L_S^{3.5}\|_2$ haben wir berücksichtigt, daß $x^{1/4}/y^{1/4}$ mit einem relativen Rückwärtsfehler von höchstens $8\,\varepsilon$ berechnet wird.

Nach Lemma 3.1.1, Satz 3.1.2 und Satz 3.1.5 (3.2) erhöhen höchstens die Operationen **3.3** oder **3.5** den kleinsten Singulärwert des skalierten L, und zwar insgesamt maximal um Faktor $\sqrt{2}$. Es ist also

$$1/\sigma_{\min}(L_S^{3.3}), 1/\sigma_{\min}(L_S^{3.4}), 1/\sigma_{\min}(L_S^{3.5}) \le \sqrt{2}/\sigma_{\min}(L_S)\ .$$

Mit Satz 1.0.5 und Lemma 3.2.2 folgt für $k = 1, \ldots n$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \le \frac{2}{\sigma_{\min}(L_S)}\,(26 + 4 + 50 + \sqrt{2}(4 + 50 + 24))\,\varepsilon\ .$$

Daraus folgt die Behauptung. Q.E.D.

Wir schließen diesen Unterabschnitt mit der Zusammenfassung der Rückwärtsfehler der Skalierung **3.7** ab.

Satz 3.2.21 *Die Matrizen $L, E \in \mathbb{R}^{2n\times 2n}$ werden gemäß Operation* **3.7** *von Verfahren 2.2.4 in endlicher Genauigkeit ε bearbeitet. Dabei sei $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$ und $(n+2)\sqrt{2n}\,\varepsilon < \sigma_{\min}(L_S)$. Die daraus entstehenden Matrizen seien L' und E'. Dann gilt für die Eigenwerte*

$$-\lambda_n \leq \ldots \leq -\lambda_1 < 0 < \lambda_1 \leq \ldots \leq \lambda_n$$

bzw.

$$-\lambda'_n \leq \ldots \leq -\lambda'_1 < 0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

des Problems $L^T Lx = i\lambda Jx$ bzw. des Problems $E'L'^T L'E'x = i\lambda Jx$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \leq \frac{(2\,n+16)\sqrt{2n}}{\sigma_{\min}(L_S)}\,\varepsilon\ , \quad k = 1\ldots n \tag{3.48}$$

Beweis:
Analog zu Satz 3.2.11. Q.E.D.

3.2.3 Relative Fehlerschranken beim modifizierten Zyklus

In diesem Unterabschnitt wird die Fehleranalayse eines modifizierten Zyklus in Verfahren 2.2.4 vorgenommen. Dabei gehen wir folgendermaßen vor. Zunächst wird die erste Transformation (**2.3.1**, **2.3.2** von Verfahren 2.2.4) in einer Form niedergeschrieben, die für die Analyse in diesem Unterabschnitt besser geeignet ist. Dann wird untersucht, wie diese Transformation in Fließpunktarithmetik die betroffenen Spalten der Matrix L ändert. Daran schließt sich die eigentliche Fehleranalyse an. Genauso werden wir beim zweiten Teil der Modifikation (**2.3.3**, **2.3.4** von Verfahren 2.2.4) verfahren. Zuerst wird die Transformation in einer günstigeren Form niedergeschrieben und dann in Fließpunktarithmetik ihre Wirkung auf die betroffenen Spalten von L untersucht. Auch hier folgt zuletzt die eigentliche Fehleranalyse.

Die Fehleranalyse der beiden Transformationen erfordert einige zusätzliche Anforderungen an die Maschinengenauigkeit ε. Ist m die halbe Ordnung des Clusters und $\mu = 1/(10m)$, so wird $\varepsilon < \mu/(2n)$ verlangt. Diese Voraussetzung ist eher beweistechnischer Natur, denn man kann sie in der Praxis stets durch Verkleinerung des Clusters erreichen. Ursache für die Voraussetzung ist der absolute Fließpunktfehler (s. Lemma 3.2.4) bei der Berechnung von Skalarprodukten. Wird statt des Standardverfahrens nach Lemma 3.2.4 zur Bestimmung des Skalarproduktes das Verfahren des Binary Splitting aus Korollar 3.2.7 verwendet, so kann auch die Voraussetzung $\varepsilon < \mu/(2n)$ durch $(\lceil \log_2 2n \rceil + 1)\varepsilon < \mu$ ersetzt werden, was

in der Praxis ohne Veränderung der Clustergröße selbst für große n und ε erfüllt sein dürfte. In der Praxis wird aber i.A. die Verwendung von Binary Splitting nicht notwendig sein, so daß wir darauf nicht näher eingehen werden.

Für Matrizen kleiner Ordnung reicht die Forderung $\varepsilon < \mu/(2m)$ noch nicht aus, vielmehr wird für einige Aussagen noch $\varepsilon < \mu/50$ vorausgesetzt. Auch dies ist eine praxisgerechte Forderung, die die Gültigkeit der Fehleranalyse nicht wesentlich einschränkt. Zusammengefaßt werden wir also $\varepsilon < \min\{\mu/(2n), \mu/50\}$ verlangen.

Wir betrachten nun den Teil **2.3.1**, **2.3.2** von Verfahren 2.2.4. Weil wir nur einen einzigen Cluster zu untersuchen brauchen, nehmen wir o.B.d.A. $num = 1$ und $low[clu] = 1$ sowie $3 \leq m = upp[clu] \leq n$ an. Deshalb sind höchstens die Spalten $1, \ldots, m$ und $n+1, \ldots, n+m$ der Matrix $L = [L_{\cdot 1} \cdots L_{\cdot 2n}]$ von der Transformation betroffen. Nachfolgend wird der untersuchte Programmteil nochmals mit eindeutigen Variablenbezeichnungen niedergeschrieben.

$[L_{\cdot 1}^{[1]}] = [L_{\cdot 1}]$, $[L_{\cdot n+1}^{[1]}] = [L_{\cdot n+1}]$, $sum = 0$
FOR $p = 2$ TO m DO
  FOR $q = p$ TO m DO $[L_{\cdot q}^{[p-1]}] = [L_{\cdot q}]$, $[L_{\cdot n+q}^{[p-1]}] = [L_{\cdot n+q}]$
  $r^{[p]} = [L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}]^T \cdot [L_{\cdot p}^{[p-1]}]$
  $\tau^{[p]} = (1 - {r^{[p]}}^T r^{[p]})^{1/2}$
  sum=sum+$2\,({r^{[p]}}^T r^{[p]})/(1 + \tau^{[p]})$
  IF $sum > 1/(10p)^2$ THEN
    $m = m - 1$
    GOTO FINISH
  ENDIF
  $[L_{\cdot 1}^{[p]} \cdots L_{\cdot p}^{[p]}] = [L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p}^{[p-1]}] \cdot \left[\begin{array}{c|c} I_{p-1} & -r^{[p]}/\tau^{[p]} \\ \hline 0 & \tau^{[p]^{-1}} \end{array}\right]$
  $[L_{\cdot n+1}^{[p]} \cdots L_{\cdot n+p}^{[p]}] = [L_{\cdot n+1}^{[p-1]} \cdots L_{\cdot n+p}^{[p-1]}]$
  $\cdot \left[\begin{array}{c|c} I_{p-1} & 0 \\ \hline {r^{[p]}}^T \operatorname{diag}(E_p^2/E_1^2, \ldots, E_p^2/E_{p-1}^2) & \tau^{[p]} \end{array}\right]$
ENDFOR
FINISH:
FOR $p = 2$ TO m DO
  $E'_p = E_1$, $E'_{n+p} = E_{n+p}^2/E_1$
ENDFOR

Vorstehendes Programm dient also lediglich der Einführung einer eindeutigen Notation für die in diesem Unterabschnitt verwendeten Variablen. Hinsichtlich des Fehlerverhaltens ist es äquivalent zum Programmteil **2.3.1**, **2.3.2** von Verfahren 2.2.4. Algorithmus 2.2.4 ist einer Analyse weniger zugänglich, weil dort Variablen in jedem Durchlauf überschrieben werden und deshalb für mehrere

verschiedene Objekte verwendet werden. In vorstehendem Programm hingegen wird für jedes Objekt eine eigene Variable verwendet. In diesem Unterabschnitt nehmen wir also die Fehleranalyse der Teile **2.3.1**, **2.3.2** von Algorithmus 2.2.4 unter Verwendung der Notation vorstehenden Programmstückes vor. Wir nehmen o.B.d.A. an, daß die Bedingung in der IF-Abfrage nie erfüllt ist, die Cluster der modifizierten Transformation also nicht verkleinert werden.

Bei der Bildung der Cluster in **2.3.1** von Verfahren 2.2.4 werden abhängig vom Verhältnis verschiedener Fließpunktwerte zueinander Programmverzweigungen vorgenommen. Ein Vergleich $\mathrm{fl}(a) \leq \mathrm{fl}(b)$ ist i.A. natürlich nicht äquivalent zu $a \leq b$. Ist $\mathrm{fl}(a) \leq \mathrm{fl}(b)$, so kann man über a und b selbst nur schwächere Aussagen treffen. Wir werden anschließend einige Folgerungen aus derartigen Fließpunktvergleichen ziehen. Zuerst jedoch geben wir Abschätzungen für die tatsächliche Norm der Spalten von L an, wenn diese in endlicher Genauigkeit auf Norm 1 skaliert wurden.

Lemma 3.2.22 *Der Vektor $v \in \mathbb{R}^k$ werde mit dem Programm*
$\sigma = 0$
FOR $i = 1$ *TO* k *DO* $\sigma = \sigma + v_i^2$
FOR $i = 1$ *TO* k *DO* $v_i' = v_i/\sqrt{\sigma}$
in endlicher Genauigkeit ε normiert. In erster Ordnung von ε gilt dann

$$\|fl(v')\|_2 = 1 + e \quad , \quad |e| \leq (k/2 + 2)\,\varepsilon \ .$$

Beweis:
Nach Lemma 3.2.5 gilt in erster Ordnung von ε

$$\mathrm{fl}(\sqrt{\sigma}) = \|v\|_2 + r \quad , \quad |r| \leq (k/2 + 1)\|v\|_2\,\varepsilon \ ,$$

so daß

$$\mathrm{fl}(v_i') = v_i/\sqrt{\sigma} + r_i' \quad , \quad |r_i'| \leq (k/2 + 2)(|v_i|/\sqrt{\sigma})\,\varepsilon$$

folgt. Daraus folgt bereits die Aussage. Q.E.D.

Vor der mit obigem Programmstück beschriebenen Transformation werden die Spalten von L zu 1 normiert (s. **1.2** von Verfahren 2.2.4). Nach der Normierung gilt also

$$\|[L_{\cdot i}]\|_2 = 1 + \lambda_i \quad \text{mit} \quad |\lambda_i| \leq (n + 2)\,\varepsilon \quad , \quad i = 1 \ldots 2n \ . \tag{3.49}$$

Eine ähnliche Aussage können wir über Skalarprodukte verschiedener Spalten von L formulieren.

Lemma 3.2.23 *Seien $u, v \in \mathbb{R}^k$ mit $\|u\|_2, \|v\|_2 = 1 + \mathcal{O}(\varepsilon)$ und sei $x > 3$. Ist $fl(\sqrt{2(u^T v)^2}) \leq fl(1/x)$, so gilt tatsächlich in erster Ordnung von ε lediglich*

$$|u^T v| \leq 1/(\sqrt{2}x) + (k + 1)\,\varepsilon \ .$$

Beweis:
Ist erster Ordnung von ε ist mit $|\varepsilon_i| \leq \varepsilon$

$$\begin{aligned} \mathrm{fl}(\sqrt{2(u^T v)^2}) &= (1+\varepsilon_1)\sqrt{(1+\varepsilon_2)(1+\varepsilon_3)\,2\,(u^T v + k\,\varepsilon_4\,\|u\|_2\,\|v\|_2)^2} \\ &= (1+2\,\varepsilon_5)\sqrt{2}(|u^T v| + k\,\varepsilon_4)\ . \end{aligned}$$

Also folgt

$$\begin{aligned} |u^T v| &\leq \frac{1}{\sqrt{2}}(1+2\,\varepsilon)\mathrm{fl}(\sqrt{2(u^T v)^2}) + k\,\varepsilon \leq \frac{1}{\sqrt{2}\,x}(1+3\,\varepsilon) + k\,\varepsilon \\ &\leq \frac{1}{\sqrt{2}\,x} + (k+1)\,\varepsilon \end{aligned}$$

Q.E.D.

Werden die Cluster gemäß Teil **2.2** von Verfahren 2.2.4 gebildet, so gilt innerhalb jedes Clusters $\mathrm{fl}(|[L_{\cdot i}]^T[L_{\cdot j}]|) \leq \mathrm{fl}(1/(10m)) = \mathrm{fl}(\mu)$ für $i \neq j$, es gilt also in erster Ordnung von ε tatsächlich

$$|[L_{\cdot i}]^T[L_{\cdot j}]| \leq \frac{\mu}{\sqrt{2}} + (2n+1)\,\varepsilon \quad , \quad i \neq j\ . \tag{3.50}$$

Dies ist allerdings eine recht pessimistische Abschätzung. Schließlich wollen wir eine weitere Aussage dieser Art für die Norm von L zeigen.

Lemma 3.2.24 *Sei $m \geq 1$ und*

$$M, N \in I\!R^{2n\times m} \quad , \quad \|[M_{\cdot i}]\|_2, \|[N_{\cdot i}]\|_2 = 1 + \mathcal{O}(\varepsilon)\ , \quad i = 1\ldots m\ .$$

Sei in endlicher Genauigkeit ε

$$\begin{aligned} &\mathit{fl}(\|[M\ \ N]^T[M\ \ N] - I\|_E) \\ &\quad = \mathit{fl}((2\textstyle\sum_{i=1}^m \sum_{j=1}^{i-1}((M_{\cdot i}^T M_{\cdot j})^2 + (M_{\cdot i}^T N_{\cdot j})^2 + (N_{\cdot i}^T M_{\cdot j})^2 + (N_{\cdot i}^T N_{\cdot j})^2))^{1/2}) \\ &\quad \leq \mathit{fl}(1/(10m))\ , \end{aligned} \tag{3.51}$$

so gilt tatsächlich in erster Ordnung von ε lediglich

$$\|[M\ \ N]^T[M\ \ N] - I\|_E \leq \frac{1}{10m} + m(4n+1)\,\varepsilon\ .$$

Beweis:
Allgemein gilt für $\alpha = (\alpha_i) = (\mathrm{fl}(x_i^T y_i)) \in I\!R^k$, $x_i, y_i \in I\!R^{2n}$, $\|x\|_2, \|y\|_2 = 1 + \mathcal{O}(\varepsilon)$ ähnlich wie in Lemma 3.2.4, 3.2.5 in erster Ordnung von ε

$$\mathrm{fl}((2\sum_{i=1}^k \alpha_i^2)^{1/2}) = (2\sum_{i=1}^k \alpha_i^2)^{1/2}(1+c\varepsilon) \quad , \quad |c| \leq 1 + \frac{k+1}{2}\ ,$$

also ist

$$\mathrm{fl}((2\sum_{i=1}^{k}\alpha_i^2)^{1/2}) \geq (2\sum_{i=1}^{k}\alpha_i^2)^{1/2}(1-(1+\frac{k+1}{2})\varepsilon) \ .$$

Weiterhin ist

$$(2\sum_{i=1}^{k}\alpha_i^2)^{1/2} = (2\sum_{i=1}^{k}(x_i^T y_i)^2)^{1/2} + r' \quad , \quad |r'| \leq \sqrt{2}\cdot 2n\cdot\sqrt{k}\varepsilon \ ,$$

so daß folgt

$$\mathrm{fl}((2\sum_{i=1}^{k}\alpha_i^2)^{1/2}) \geq [(2\sum_{i=1}^{k}(x_i^T y_i)^2)^{1/2} - 2\sqrt{2}n\sqrt{k}\varepsilon][1-(1+\frac{1}{2}(k+1))\varepsilon] \ .$$

Dies wenden wir auf (3.51) an ($k = 4m(m-1)/2$) und bedenken dabei noch $\mathrm{fl}(1/(10m)) \leq (1+\varepsilon)/(10m)$. In erster Ordnung von ε gilt dann

$$[\|[M\ N]^T[M\ N]-I\|_E - 4n\sqrt{m(m-1)}\varepsilon][1-(1+\frac{1}{2}(2m(m-1)+1))\varepsilon] \leq \frac{1+\varepsilon}{10m} \ ,$$

d.h.

$$\begin{aligned}
[\|[M\ N]^T[M\ N]-I\|_E &\leq \frac{1}{10m}(1+\varepsilon)[1+(\frac{3}{2}+m(m-1))\varepsilon] \\
&\quad +4n\sqrt{m(m-1)}\,\varepsilon \\
&= \frac{1}{10m} + [\frac{1+3/2+m(m-1)}{10m} + 4n\sqrt{m(m-1)}]\,\varepsilon \\
&\leq \frac{1}{10m} + m\,[\frac{5}{20m^2} + \frac{1}{10} + 4n]\,\varepsilon \\
&\leq \frac{1}{10m} + m(4n+1)\,\varepsilon
\end{aligned}$$

Daraus folgt bereits die Aussage. Q.E.D.

Wird die Ordnung eines Clusters in endlicher Genauigkeit anhand der Bedingung (3.51) festgestellt, so gilt für den Cluster die in Lemma 3.2.24 genannte exakte Ungleichung. Die exakte Ungleichung kann gegenüber der in endlicher Genauigkeit geprüften durchaus große relative Fehler aufweisen.

Mit folgendem Satz können wir nun die Wirkung der ersten Transformation auf die betroffenen Spalten von L angeben.

Satz 3.2.25 *Sei $n \geq m \geq 3$ und $L, diag(E_i) \in \mathbb{R}^{2n\times 2n}$. Sei ε die Maschinengenauigkeit und für $\mu = 1/(10m)$ sei bis auf einen kleinen relativen Fehler*

$$\varepsilon \leq \min\{\frac{\mu}{2n}, \frac{\mu}{50}\} \ . \tag{3.52}$$

Sei $E_i = E_{n+i}, i = 1 \dots m$ *und* $E_i \geq E_{i+1}, i = 1 \dots m-1$. *Weiter sei bis auf einen kleinen relativen Fehler*

$$\|[L_{\cdot i}]\|_2 \leq 1 + \lambda_i \quad , \quad |\lambda_i| \leq (n+2)\,\varepsilon \ , \quad i = 1 \dots m, n+1 \dots n+m, \tag{3.53}$$

$$\begin{aligned} &\|[L_{\cdot 1} \cdots L_{\cdot m} \ L_{\cdot n+1} \cdots L_{\cdot n+m}]^T [L_{\cdot 1} \cdots L_{\cdot m} \ L_{\cdot n+1} \cdots L_{\cdot n+m}] - I\|_E \\ &\quad \leq \mu + m(4n+1)\,\varepsilon \ , \end{aligned} \tag{3.54}$$

und

$$E_1^2 \leq (1+\mu) E_m^2 \ . \tag{3.55}$$

Die Operationen **2.3.1**, **2.3.2** *von Verfahren 2.2.4 werden für einen Cluster der Ordnung* $2m$ *in endlicher Genauigkeit* ε *durchgeführt. Unter Verwendung der Notation des Programmstückes von Seite 135 gilt nach Beendigung bis auf einen kleinen relativen Fehler für alle* berechneten $[L_{\cdot i}^{[p]}]$, $p \in \{2, \dots, m\}$, $i \in \{1, \dots, p, n+1, \dots, n+p\}$,

$$[L_{\cdot i}^{[p]}] = [L_{\cdot i}] + [l_{\cdot i}] \quad , \quad \|[l_{\cdot i}]\|_2 \leq \begin{cases} 0 & i = 1 \\ 2\,\mu & i = 2 \dots p \\ 0.5\sqrt{\mu} & i = n+1 \dots n+p \ . \end{cases} \tag{3.56}$$

Außerdem gilt

$$|[L_{\cdot n+q}^{[m]}]^T [L_{\cdot p}^{[m]}]| \leq 4\,\mu \quad , \quad p, q \in \{2, \dots, m\} \ . \tag{3.57}$$

Beweis:
Zuerst wird (3.56) betrachtet, was lediglich für $p = m$ zu zeigen ist. Die Gleichung für $\|[l_{\cdot 1}]\|_2$ ergibt sich unmittelbar aus dem Verfahren, denn die erste Spalte von L wird nicht verändert. Wir werden nun eine Schranke für $\|[l_{\cdot p}]\|_2$ für $2 \leq p \leq m$ bestimmen. Wird Algorithmus 2.2.4 ausgeführt und verwenden wir die Notation von Seite 135, so gilt aufgrund der evtl. Clusterverkleinerung (IF-Abfrage mit Variable *sum*) und exakter Ganzzahlrechnung

$$\mathrm{fl}(\tilde{r}^{[p]^T} \tilde{r}^{[p]}) \leq \mu^2 (1+\varepsilon) \ , \tag{3.58}$$

wobei $\tilde{r}^{[p]} = \mathrm{fl}(r^{[p]})$. Weil nach Lemma 3.2.4 in erster Ordnung von ε

$$\mathrm{fl}(\tilde{r}^{[p]^T} \tilde{r}^{[p]}) = \tilde{r}^{[p]^T} \tilde{r}^{[p]} + \varrho \quad , \quad |\varrho| \leq (p-1) \tilde{r}^{[p]^T} \tilde{r}^{[p]}\,\varepsilon$$

ist, folgt in erster Ordnung von ε

$$\tilde{r}^{[p]^T} \tilde{r}^{[p]} \leq \mathrm{fl}(\tilde{r}^{[p]^T} \tilde{r}^{[p]}) + |\varrho| \leq \mu^2\,(1+\varepsilon) + (p-1) \tilde{r}^{[p]^T} \tilde{r}^{[p]}\,\varepsilon \ ,$$

d.h.

$$\tilde{r}^{[p]^T} \tilde{r}^{[p]} \leq \mu^2\,(1+\varepsilon)(1+(p-1)\,\varepsilon) \ .$$

Also ist

$$\|\tilde{r}^{[p]}\|_2 \leq \mu\,(1+\frac{p}{2}\,\varepsilon) \leq \mu\,(1+\frac{\mu}{4})\ . \tag{3.59}$$

Für $\tilde{\tau}^{[p]} = \mathrm{fl}(\tau^{[p]}) = \mathrm{fl}((1-{r^{[p]}}^T r^{[p]})^{1/2})$ gilt wegen (3.58)

$$(1-\mu^2)(1-2\,\varepsilon) \leq (1-\mu^2)^{1/2}(1-2\,\varepsilon) \leq \tilde{\tau}^{[p]} \leq 1\ , \tag{3.60}$$

außerdem ist

$$1 \leq \mathrm{fl}({\tau^{[p]}}^{-1}) \leq (1-\mu^2)^{-1/2}(1+3\,\varepsilon) \leq (1+\mu^2)(1+3\,\varepsilon)\ . \tag{3.61}$$

Damit gilt gemäß Algorithmus 2.2.4 in erster Ordnung von ε

$$\begin{aligned} L_{kp}^{[p]} = \mathrm{fl}(L_{kp}^{[p]}) &= \mathrm{fl}({\tau^{[p]}}^{-1}(L_{kp}^{[p-1]} - \sum_{q=1}^{p-1} L_{kq}^{[p-1]}\tilde{r}_q^{[p]})) \\ &= \mathrm{fl}({\tau^{[p]}}^{-1})(L_{kp}^{[p-1]} - \sum_{q=1}^{p-1} L_{kq}^{[p-1]}\tilde{r}_q^{[p]}) + Y_k^{[p]} \end{aligned}$$

mit

$$|Y_k^{[p]}| \leq \mathrm{fl}({\tau^{[p]}}^{-1})(|L_{kp}^{[p-1]}| + \sum_{q=1}^{p-1} |L_{kq}^{[p-1]}||\tilde{r}_q^{[p]}|)p\,\varepsilon\ . \tag{3.62}$$

Also ist $L_{kp}^{[p]} = L_{kp}^{[p-1]} + l_{kp}$ mit

$$|l_{kp}| \leq |(\mathrm{fl}({\tau^{[p]}}^{-1}) - 1)L_{kp}^{[p-1]}| + |\mathrm{fl}({\tau^{[p]}}^{-1})\sum_{q=1}^{p-1} L_{kq}^{[p-1]}\tilde{r}_q^{[p]}| + |Y_k^{[p]}|\ . \tag{3.63}$$

Wegen (3.59), (3.61) und (3.62) ist in erster Ordnung von ε

$$\begin{aligned} &|(\mathrm{fl}({\tau^{[p]}}^{-1}) - 1)L_{kp}^{[p-1]}| + |Y_k^{[p]}| \\ &\leq (3\,\varepsilon + \mu^2(1+3\,\varepsilon))|L_{kp}^{[p-1]}| + (1+\mu^2)(|L_{kp}^{[p-1]}| + \sum_{q=1}^{p-1} |L_{kq}^{[p-1]}||\tilde{r}_q^{[p]}|)p\,\varepsilon \\ &\leq (\mu^2 + (3+3\,\mu^2+(1+\mu^2)p)\,\varepsilon)|L_{kp}^{[p-1]}| \\ &\quad +(1+\mu^2)\mu(1+\mu/4)(\sum_{q=1}^{p-1} {L_{kq}^{[p-1]}}^2)^{1/2}p\,\varepsilon. \end{aligned}$$

Mit (3.61) und (3.63) erhalten wir daraus einen Ansatz für eine obere Schranke von $\|[l_{\cdot p}]\|_2$. Es gilt in erster Ordnung von ε

$$\begin{aligned} \|[l_{\cdot p}]\|_2 \ \leq\ & (\mu^2 + 3(1+\mu^2)\,\varepsilon + (1+\mu^2)p\,\varepsilon)\|[L_{\cdot p}]\|_2 \\ & +(1+\mu^2)\mu(1+\mu/4)\|[L_{\cdot 1}^{[p-1]}\cdots L_{\cdot p-1}^{[p-1]}]\|_E\,p\,\varepsilon \\ & +(1+\mu^2)(1+3\,\varepsilon)\,\|[L_{\cdot 1}^{[p-1]}\cdots L_{\cdot p-1}^{[p-1]}]\|_2\,\|\tilde{r}^{[p]}\|_2\ . \end{aligned}$$

Dabei ist $[L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}] = [L_{\cdot 1} \cdots L_{\cdot p-1}] + [l_{\cdot 1} \cdots l_{\cdot p-1}]$ und wegen Lemma 3.2.24 sowie (3.52) ist

$$\begin{aligned} \|[L_{\cdot 1} \cdots L_{\cdot p-1}]\|_2 &\leq (1 + \|[L_{\cdot 1} \cdots L_{\cdot p-1}]^T [L_{\cdot 1} \cdots L_{\cdot p-1}] - I\|_2)^{1/2} \\ &\leq (1 + \mu + p(4n+1)\,\varepsilon)^{1/2} \\ &\leq 1.12\ . \end{aligned}$$

Verwenden wir noch

$$\|[L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}]\|_E \leq \sqrt{p}\|[L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}]\|_2$$

und (3.49), so folgt wegen (3.52) in erster Ordnung von ε schließlich

$$\begin{aligned} \|[l_{\cdot p}]\|_2 &\leq \mu\ [(\mu + \frac{3(1+\mu^2)}{50} + \frac{1+\mu^2}{2})(1 + (n+2)\,\varepsilon) \\ &\quad +(1+\mu^2)\,(1+\frac{\mu}{4})(1.12 + \|[l_{\cdot 1} \cdots l_{\cdot p-1}]\|_2)\sqrt{p}\frac{p\mu}{2n} \\ &\quad +(1+\mu^2)(1.12 + \|[l_{\cdot 1} \cdots l_{\cdot p-1}]\|_2)\,(1+\frac{\mu}{4})] \\ &\leq \mu\ [1.78\ +\ 1.04\,(\sum_{q=1}^{p-1} \|[l_{\cdot q}]\|_2^2)^{1/2}]\ . \end{aligned}$$

Die letztgenannte obere Schranke für die $\|[l_{\cdot p}]\|_2$ wird also von der Folge (a_p) mit $a_1 = 0$ und

$$a_p = \mu(1.78 + 1.04\,(\sum_{q=1}^{p-1} a_q^2)^{1/2}) \quad , \quad p \geq 2$$

majorisiert. Für $p \geq 2$ ist $a_p \geq a_{p-1}$ äquivalent zu $(\sum_{q=1}^{p-1} a_q^2)^{1/2} \geq (\sum_{q=1}^{p-2} a_q^2)^{1/2}$ und dies bedeutet, daß die Folge (a_p) monoton wächst. Sie wird daher von der Folge (b_p) mit $b_1 = 0$ und

$$b_p = \mu(1.78 + 1.04\,\sqrt{m} b_{p-1}) \quad , \quad p \geq 2$$

majorisiert. Eine obere Schranke der b_p erhalten wir für $p \geq 2$ aus

$$\begin{aligned} b_p &= \sum_{q=0}^{p-2} (1.04\,\mu\sqrt{m})^q (1.78\,\mu) \\ &= 1.78\,\mu\,\frac{1 - (1.04\,\mu\sqrt{m})^{p-1}}{1 - 1.04\,\mu\sqrt{m}} \\ &\leq \mu\,\frac{1.78}{1 - 1.04\,\mu\sqrt{m}} \\ &\leq 1.99\,\mu\ . \end{aligned}$$

Für $2 \le p \le m$ ist also $\|[l_{\cdot p}]\|_2 \le 1.90\,\mu$. Anschließend werden wir eine Schranke von $\|[l_{\cdot j}]\|_2$ für $j > n$ bestimmen. Sei dazu $2 \le p \le m$ und $1 \le q < p$. In erster Ordnung von ε gilt gemäß Algorithmus 2.2.4

$$
\begin{aligned}
L_{k,n+q}^{[p]} &= \mathrm{fl}(L_{k,n+q}^{[p]}) \qquad (3.64)\\
&= \mathrm{fl}(L_{k,n+q}^{[p-1]} + r_q^{[p]}(E_p/E_q)^2 L_{k,n+p}^{[p-1]})\\
&= (1+\varepsilon_1^{[p-1]})L_{k,n+q}^{[p-1]} + (1+\varepsilon_2^{[p-1]})\tilde{r}_q^{[p]}(E_p/E_q)^2 L_{k,n+p}^{[p-1]})\\
&\qquad \text{mit} \quad |\varepsilon_1^{[p-1]}| \le \varepsilon \quad , \quad |\varepsilon_2^{[p-1]}| \le 5\,\varepsilon\\
&= (1+\varepsilon_1^{[p-1]})\left((1+\varepsilon_1^{[p-2]})L_{k,n+q}^{[p-2]} + (1+\varepsilon_2^{[p-2]})\tilde{r}_q^{[p-1]}(E_{p-1}/E_q)^2 L_{k,n+p}^{[p-2]}\right)\\
&\quad +(1+\varepsilon_2^{[p-1]})\tilde{r}_q^{[p]}(E_p/E_q)^2 L_{k,n+p}^{[p-1]}\\
&\qquad \text{mit} \quad |\varepsilon_1^{[p-2]}| \le \varepsilon \quad , \quad |\varepsilon_2^{[p-2]}| \le 5\,\varepsilon\\
&= (1+\hat{\varepsilon}_1^{[p-2]})L_{k,n+q}^{[p-2]} + (1+\hat{\varepsilon}_2^{[p-2]})\tilde{r}_q^{[p-1]}(E_{p-1}/E_q)^2 L_{k,n+p}^{[p-2]}\\
&\quad + (1+\hat{\varepsilon}_2^{[p-1]})\tilde{r}_q^{[p]}(E_p/E_q)^2 L_{k,n+p}^{[p-1]}\\
&\qquad \text{mit} \quad |\hat{\varepsilon}_1^{[p-2]}| \le 2\,\varepsilon \quad , \quad |\hat{\varepsilon}_2^{[p-2]}| \le 6\,\varepsilon \quad , \quad |\hat{\varepsilon}_2^{[p-1]}| \le 5\,\varepsilon\\
&\ \vdots\\
&= (1+\hat{\varepsilon}_1^{[q]})L_{k,n+q}^{[q]} + \sum_{i=q+1}^{p}(1+\hat{\varepsilon}_2^{[i-1]})\tilde{r}_q^{[i]}(E_i/E_q)^2 L_{k,n+p}^{[i-1]}\\
&\qquad \text{mit} \quad |\hat{\varepsilon}_1^{[q]}| \le (p-q)\,\varepsilon \quad , \quad |\hat{\varepsilon}_2^{[i-1]}| \le (5+(p-i))\,\varepsilon \le (m+3)\,\varepsilon\ .
\end{aligned}
$$

Beachten wir noch $L_{k,n+q}^{[q]} = (1+\tilde{\varepsilon})\tilde{\tau}^{[q]} L_{k,n+q}$ mit $|\tilde{\varepsilon}| \le \varepsilon$, so können wir

$$L_{k,n+q}^{[p]} = L_{k,n+q} + l_{k,n+q}^{[p]}$$

mit

$$
\begin{aligned}
\|[l_{\cdot n+q}^{[p]}]\|_2 \ &\le\ |(1-(p-q+1)\,\varepsilon)\tilde{\tau}^{[q]} - 1|\ \|[L_{\cdot n+q}]\|_2\\
&\quad + \sum_{i=q+1}^{p}(1+(m+3)\,\varepsilon)|\tilde{r}_q^{[i]}(E_i/E_q)^2|\ \|L_{\cdot n+p}\|_2
\end{aligned}
$$

schließen (es ist $[L_{\cdot n+p}] = [L_{\cdot n+p}^{[i-1]}]$ für $i \le p$). Wegen (3.60) ist dabei in erster Ordnung von ε

$$
\begin{aligned}
|(1-(p-q+1)\,\varepsilon)\tilde{\tau}^{[q]} - 1| \ &\le\ |(1-(p-q+1)\,\varepsilon)(1-\mu^2)(1-2\,\varepsilon) - 1|\\
&\le\ (p-q+3)\,\varepsilon + \mu^2(1+(p-q+3)\,\varepsilon)\\
&\le\ \mu^2 + (1+\mu^2)(n+2)\,\varepsilon\\
&\le\ 0.11\,\sqrt{\mu}\ .
\end{aligned}
$$

Weiterhin haben wir wegen (3.59) in erster Ordnung von ε

$$\sum_{i=q+1}^{p}(1+(m+3)\,\varepsilon)|\tilde{r}_q^{[i]}(E_i/E_q)^2|$$

$$\begin{aligned}
&\leq (1+(m+3)\,\varepsilon)\sum_{i=q+1}^{p}|\tilde{r}_q^{[i]}| \leq (1+(m+3)\,\varepsilon)\sqrt{p-q}(\sum_{i=q+1}^{p}(\tilde{r}_q^{[i]})^2)^{1/2} \\
&\leq (1+(m+3)\,\varepsilon)\sqrt{p-q}\mu(1+\mu/4) \leq 0.33\,\sqrt{\mu}\ .
\end{aligned}$$

Es folgt

$$\begin{aligned}
\|[l_{\cdot n+q}^{[p]}]\|_2 &\leq 0.11\,\sqrt{\mu}\,\|[L_{\cdot n+q}]\|_2 + 0.33\,\sqrt{\mu}\,\|L_{\cdot n+p}\|_2 \\
&\leq 0.11\,\sqrt{\mu}(1+(n+2)\,\varepsilon) + 0.33\,\sqrt{\mu}(1+(n+2)\,\varepsilon) \\
&\leq 0.45\,\sqrt{\mu}
\end{aligned}$$

Wegen $[l_{\cdot n+q}] = [l_{\cdot n+q}^{[m]}]$ ist (3.56) damit bewiesen. Es bleibt noch (3.57) zu zeigen. Dazu formulieren wir (3.64) zunächst um. Es gibt Diagonalmatrizen

$$\bar{D} = \mathrm{diag}(1+\bar{\varepsilon}_k) \in I\!R^{2n\times 2n} \quad , \quad |\bar{\varepsilon}_k| \leq m\,\varepsilon$$

und

$$\tilde{D}^{[i]} = \mathrm{diag}(1+\tilde{\varepsilon}_k^{[i]}) \in I\!R^{2n\times 2n} \quad , \quad i=q,\ldots,m-1\ , \quad |\tilde{\varepsilon}_k^{[i]}| \leq (m+3)\,\varepsilon\ ,$$

mit

$$[L_{\cdot n+q}^{[m]}] = \tilde{r}^{[q]}\bar{D}[L_{\cdot n+q}] + \sum_{i=q+1}^{m}\tilde{r}_q^{[i]}(E_i/E_q)^2\tilde{D}^{[i-1]}[L_{\cdot n+m}]\ .$$

Daraus folgt mit $[L_{\cdot p}^{[m]}] = [L_{\cdot p}] + [l_{\cdot p}]$

$$\begin{aligned}
|[L_{\cdot n+q}^{[m]}]^T[L_{\cdot p}^{[m]}]| &\leq |\tilde{r}^{[q]}[L_{\cdot n+q}]^T\bar{D}[L_{\cdot p}]| + |\tilde{r}^{[q]}[L_{\cdot n+q}]^T\bar{D}[l_{\cdot p}]| \\
&\quad + \sum_{i=q+1}^{m}|\tilde{r}_q^{[i]}(E_i/E_q)^2|\ |[L_{\cdot n+m}]^T\tilde{D}^{[i-1]}[L_{\cdot p}^{[m]}]|\ .
\end{aligned}$$

Dabei ist wegen (3.50) in erster Ordnung von ε

$$\begin{aligned}
|[L_{\cdot n+q}]^T\bar{D}[L_{\cdot p}]| &\leq |[L_{\cdot n+q}]^T[L_{\cdot p}]| + |[L_{\cdot n+q}]^T(\bar{D}-I)[L_{\cdot p}]| \\
&\leq \frac{\mu}{\sqrt{2}} + (2n+1)\,\varepsilon + \|\bar{D}-I\|_2 \\
&\leq \mu\,(\frac{1}{\sqrt{2}} + 1 + \frac{\mu}{50} + \frac{1}{500}) \leq 1.71\,\mu\ ,
\end{aligned}$$

$$\begin{aligned}
|[L_{\cdot n+q}]^T\bar{D}[l_{\cdot p}]| &\leq |[L_{\cdot n+q}]^T[l_{\cdot p}]| + \|\bar{D}-I\|_2\|[l_{\cdot p}]\|_2 \\
&\leq (1+(n+2)\,\varepsilon + \|\bar{D}-I\|_2)\|[l_{\cdot p}]\| \\
&\leq (1+(m+n+2)\,\varepsilon)\cdot 2\,\mu \\
&\leq 2.07\,\mu
\end{aligned}$$

sowie

$$\begin{aligned}
|[L_{\cdot n+m}]^T \tilde{D}^{[i-1]}[L_{\cdot p}^{[m]}]| &\leq |[L_{\cdot n+m}]^T \tilde{D}^{[i-1]}[L_{\cdot p}]| + |[L_{\cdot n+m}]^T \tilde{D}^{[i-1]}[l_{\cdot p}]| \\
&\leq |[L_{\cdot n+m}]^T [L_{\cdot p}]| + \|\tilde{D}^{[i-1]} - I\|_2 + \|\tilde{D}^{[i-1]}\|_2 \|[l_{\cdot p}]\|_2 \\
&\leq \frac{\mu}{\sqrt{2}} + (2n+1)\,\varepsilon + (m+3)\,\varepsilon + (1 + (m+3)\,\varepsilon) \cdot 2\,\mu \\
&\leq 4.33\,\mu\ ,
\end{aligned}$$

also

$$\begin{aligned}
|[L_{\cdot n+q}]^T [L_{\cdot p}^{[m]}]| &\leq 1.71\,\mu + 2.07\,\mu + \sum_{i=q+1}^{m} |\tilde{r}_q^{[i]}| \cdot 4.33\,\mu \\
&\leq \mu\,(1.71 + 2.07 + \sqrt{m}(\sum_{i=q+1}^{m} (\tilde{r}_q^{[i]})^2)^{1/2})\ .
\end{aligned}$$

Es gilt $\sum_{i=q+1}^{m} (\tilde{r}_q^{[i]})^2 \leq (1+\varepsilon)/(10m)^2$, also folgt

$$|[L_{\cdot n+q}^{[m]}]^T [L_{\cdot p}^{[m]}]| \leq \mu\,(3.78 + \sqrt{m}\,\frac{1+\varepsilon}{10m}) \leq 4\,\mu\ .$$

Damit ist auch (3.57) gezeigt. Q.E.D.

Wir kennen jetzt exakte Abschätzungen der Veränderungen, die das in endlicher Genauigkeit durchgeführte Programmstück **2.3.1**, **2.3.2** von Verfahren 2.2.4 an den betroffenen Spalten von L hervorruft. Diese Kenntnis werden wir nun bei der Fehleranalyse der Transformation verwenden. Auch bei der Untersuchung der Transformation **2.3.3**, **2.3.4** von Verfahren 2.2.4 werden wir noch einmal darauf zurückkommen.

Satz 3.2.26 *Die Voraussetzungen des Satzes 3.2.25 seien erfüllt. Der Teil* **2.3.1**, **2.3.2** *von Verfahren 2.2.4 werde in endlicher Genauigkeit ε durchgeführt. Dann gilt folgendes Diagramm*

$$\begin{tikzcd}
L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L^{[m]}), fl(E') \\
L + \delta L, E \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L^{[m]}), fl(E')$ durch Fließpunkt-Berechnung mit Hilfe des Programmes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß $(L + \delta L) \cdot E$ durch exakte Rechtsmultiplikation mit einer symplektischen Matrix in $fl(L^{[m]}) \cdot fl(E')$ übergeht. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht.

Schreibt man $L = L_S diag(\|L_i\|_2)$ und $\delta L = \delta L_S diag(\|L_i\|_2)$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \leq 4.5\, m\, \sqrt{m}\, \varepsilon \tag{3.65}$$

beschränkt.

Beweis:
Wir verwenden die Notation

$$\hat{E} = \mathrm{diag}(E_1, \ldots, E_m) \oplus \mathrm{diag}(E_1, \ldots, E_m)$$

und

$$\hat{L} = [L_{\cdot 1} \cdots L_{\cdot m} \quad L_{\cdot n+1} \cdots L_{\cdot n+m}] \ .$$

Bezeichnungen $\hat{L}^{[m]}$ und $\hat{E}'$ werden analog angewendet. Außerdem wenden wir die Notation des Programmstückes von Seite 135 an. Es werden mehrere Aussagen aus dem Beweis von Satz 3.2.25 verwendet. Sei $1 \leq p \leq m$ und

$$[L_{\cdot 1}^{[p]} \cdots L_{\cdot p-1}^{[p]}] = \mathrm{fl}([L_{\cdot 1}^{[p]} \cdots L_{\cdot p}^{[p]}]) = \mathrm{fl}([L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p}^{[p-1]}] \cdot \begin{bmatrix} I & -r^{[p]}/\tau^{[p]} \\ 0 & \tau^{[p]^{-1}} \end{bmatrix}) \ ,$$

wobei

$$r^{[p]} = [L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}]^T [L_{\cdot p}^{[p-1]}] \quad , \quad \tau^{[p]} = (1 - r^{[p]^T} r^{[p]})^{1/2} \ .$$

Es gilt $\mathrm{fl}(L_{ki}^{[q]}) = L_{ki}^{[q-1]}$, $q = 1, \ldots, p-1$, und mit $\tilde{r}^{[p]} = \mathrm{fl}(r^{[p]})$, $\tilde{\tau}^{[p]} = \mathrm{fl}(\tau^{[p]})$ ist

$$\begin{aligned} \mathrm{fl}(L_{kp}^{[p]}) &= \mathrm{fl}(\tau^{[p]^{-1}}(L_{kp}^{[p-1]} - \sum_{q=1}^{p-1} r_q^{[p]} L_{kq}^{[p-1]})) \\ &= \tilde{\tau}^{[p]^{-1}}(L_{kp}^{[p-1]} - \sum_{q=1}^{p-1} \tilde{r}_q^{[p]} L_{kq}^{[p-1]}) + Y_{kp} \ , \end{aligned}$$

wobei wegen Lemma 3.2.4 gilt

$$|Y_{kp}| \leq \tilde{\tau}^{[p]^{-1}}(|L_{kp}^{[p-1]}| + \sum_{q=1}^{p-1} |\tilde{r}_q^{[p]}|\, |L_{kq}^{[p-1]}|)(p+2)\, \varepsilon \ .$$

Wir verwenden nun (3.59) und (3.61) aus dem Beweis von 3.2.25. Danach gilt $\|\tilde{r}^{[p]}\|_2 \leq \mu(1+\mu/4)$ und bis auf einen kleinen relativen Fehler $1 \leq \tilde{\tau}^{[p]^{-1}} \leq 1+\mu^2$. In erster Ordnung von ε ist daher nach Satz 3.2.25

$$\begin{aligned} \|[Y_{\cdot p}]\|_2 &\leq (1+\mu^2)(\|[L_{\cdot p}]\|_2 + \|\tilde{r}^{[p]}\|_2\ \|[L_{\cdot 1}^{[p-1]} \cdots L_{\cdot p-1}^{[p-1]}]\|_E)(p+2)\, \varepsilon \\ &\leq (1+\mu^2)(1+\mu(1+\mu/4)(\sum_{q=1}^{p-1} \|[L_{\cdot q}] + [l_{\cdot q}]\|_2^2)^{1/2})(p+2)\, \varepsilon \\ &\leq (1+\mu^2)(1+\mu(1+\mu/4)(1+2\,\mu)\sqrt{p-1})(p+2)\, \varepsilon \\ &\leq 1.07\,(p+2)\, \varepsilon \ . \end{aligned} \tag{3.66}$$

Sei weiterhin

$$\begin{aligned}
&[L^{[p]}_{\cdot n+1} \cdots L^{[p]}_{\cdot n+p}] = \mathrm{fl}([L^{[p]}_{\cdot n+1} \cdots L^{[p]}_{\cdot n+p}]) \\
&= \mathrm{fl}([L^{[p-1]}_{\cdot n+1} \cdots L^{[p-1]}_{\cdot n+p}] \cdot \begin{bmatrix} I & 0 \\ r^{[p]T}\mathrm{diag}(E_p^2/E_1^2, \ldots, E_p^2/E_{p-1}^2) & \tau^{[p]} \end{bmatrix}) \ .
\end{aligned}$$

Gemäß Verfahren 2.2.4 ist dann $\mathrm{fl}(L^{[m]}_{k,n+m}) = \mathrm{fl}(\tau^{[m]} L_{k,n+m})$, also

$$\mathrm{fl}([L^{[m]}_{\cdot n+m}]) = \tilde{\tau}^{[m]}[L_{\cdot n+m}] + [Y_{\cdot n+m}] \ , \quad \|[Y_{\cdot n+m}]\|_2 \le \tilde{\tau}^{[m]} \|[L_{\cdot n+m}]\|_2 \, \varepsilon \le \varepsilon \ . \tag{3.67}$$

Für $q < m$ verwenden wir (3.64) aus dem Beweis von Satz 3.2.25, woraus wir

$$\mathrm{fl}([L^{[m]}_{\cdot n+q}]) = \tilde{\tau}^{[q]}[L_{\cdot n+q}] + \sum_{i=q+1}^{m} \tilde{r}_q^{[i]} (\frac{E_i}{E_q})^2 [L_{\cdot n+m}] + [Y_{\cdot n+q}]$$

mit

$$\begin{aligned}
\|[Y_{\cdot n+q}]\|_2 &\le (m-q+1)\tilde{\tau}^{[q]} \|[L_{\cdot n+q}]\|_2 \, \varepsilon \\
&\quad + (m+3) \sum_{i=q+1}^{m} |\tilde{r}_q^{[i]}| \, |(\frac{E_i}{E_q})^2 \|[L_{\cdot n+m}]\|_2 \, \varepsilon \\
&\le m\,\varepsilon + (m+3)\sqrt{m-q} (\sum_{i=q+1}^{m} \tilde{r}_q^{[i]^2})^{1/2} \varepsilon
\end{aligned}$$

schließen. Aufgrund der evtl Clusterverkleinerung im Programm (IF-Abfrage mit Variable *sum*) gilt bis auf einen kleinen relativen Fehler

$$(\sum_{i=q+1}^{m} \tilde{r}_q^{[i]^2})^{1/2} \le \mu^2 \ ,$$

so daß folgt

$$\begin{aligned}
\|[Y_{\cdot n+q}]\|_2 &\le m\,\varepsilon + (m+3)\sqrt{m-q}\,\mu^2\,\varepsilon \le m\,\varepsilon(1 + (1+\frac{3}{m})\frac{1}{100\sqrt{m}m}) \\
&\le 1.01\, m\,\varepsilon \ .
\end{aligned} \tag{3.68}$$

Zusammengefaßt haben wir in (3.66), (3.67) und (3.68) die maximalen Fehler angegeben, die die in endlicher Genauigkeit ε berechneten neuen Spalten von $\hat{L}$ gegenüber den neuen Spalten haben, wenn diese in exakter Arithmetik berechnet worden wären. Wir werden nun ein symplektisches T und ein δL konstruieren, die den Forderungen des Satzes genügen. Sei

$$\hat{R}_1^{[p]} = \begin{bmatrix} I_{p-1} & -\tilde{r}^{[p]}/\tilde{\tau}^{[p]} \\ 0 & \tilde{\tau}^{[p]^{-1}} \end{bmatrix} \oplus I_{m-p} \ ,$$

$$\hat{R}_2^{[p]} = \begin{bmatrix} I_{p-1} & 0 \\ \tilde{r}^{[p]^T}\mathrm{diag}(E_p^2/E_1^2, \ldots, E_p^2/E_{p-1}^2) & \tilde{\tau}^{[p]} \end{bmatrix} \oplus I_{m-p}$$

und damit

$$\hat{T} = \prod_{p=2}^{m} (\hat{R}_1^{[p]} \oplus \hat{R}_2^{[p]}) \ .$$

Weiter sei

$$\hat{S} = \mathrm{diag}(E_1/E_1, \ldots, E_1/E_m) \oplus \mathrm{diag}(E_1/E_1, \ldots, E_m/E_1) \ .$$

Wir werden wir zeigen, daß $T = \hat{E}^{-1}\hat{T}\hat{E}\hat{S}$ symplektisch ist (die mit $\tilde{\cdot}$ versehenen Werte sind dabei Fließpunktwerte). Weil $\hat{S}$ trivialerweise bereits selbst symplektisch ist, ist dies nur noch für $\hat{E}^{-1}\hat{T}\hat{E}$ zu zeigen. Aus

$$\begin{aligned} &\mathrm{diag}(E_1^{-1}, \ldots, E_m^{-1}) \prod_{p=2}^{m} (\hat{R}_1^{[p]}) \ \mathrm{diag}(E_1, \ldots, E_m) \\ &= (\mathrm{diag}(E_1, \ldots, E_m) \prod_{p=2}^{m} (\hat{R}_1^{[p]^{-T}}) \ \mathrm{diag}(E_1^{-1}, \ldots, E_m^{-1}))^{-T} \\ &= (\prod_{p=2}^{m} \begin{bmatrix} I_{p-1} & 0 \\ E_p\tilde{r}^{[p]^T}\mathrm{diag}(E_1^{-1}, \ldots, E_{p-1}^{-1}) & \tilde{\tau} \end{bmatrix})^{-T} \end{aligned} \tag{3.69}$$

und

$$\begin{aligned} &\mathrm{diag}(E_1^{-1}, \ldots, E_m^{-1}) \prod_{p=2}^{m} (\hat{R}_2^{[p]}) \ \mathrm{diag}(E_1, \ldots, E_m) \\ &= \prod_{p=2}^{m} \begin{bmatrix} I_{p-1} & 0 \\ E_p\tilde{r}^{[p]^T}\mathrm{diag}(E_1^{-1}, \ldots, E_{p-1}^{-1}) & \tilde{\tau}^{[p]} \end{bmatrix} \ , \end{aligned} \tag{3.70}$$

folgt aber unmittelbar, daß $\hat{E}^{-1}\hat{T}\hat{E}$ als die direkte Summe von (3.69) und (3.70) symplektisch ist. Als nächstes bestimmen wir ein δL, das der Gleichung

$$\mathrm{fl}(\hat{L}^{[m]})\mathrm{fl}(\hat{E}') = (\hat{L} + \delta\hat{L})\hat{E}T = (\hat{L} + \delta\hat{L})\hat{T}\hat{E}\hat{S} \tag{3.71}$$

und skaliert der Ungleichung 3.65 genügt. Nach **2.3.2** von Verfahren 2.2.4 gilt

$$\mathrm{fl}(\hat{E}') = \Delta\hat{E}\hat{S} \quad , \quad \Delta = \mathrm{diag}(1 + \varepsilon_i) \ , \quad |\varepsilon_i| \leq 2\,\varepsilon, i = 1 \ldots m \ .$$

Also ist (3.71) äquivalent zu

$$\mathrm{fl}(\hat{L}^{[m]})\Delta = (\hat{L} + \delta\hat{L})\hat{T} \ . \tag{3.72}$$

Wir haben zu Beginn bereits gezeigt, daß Verfahren 2.2.4 $\mathrm{fl}(\hat{L}^{[m]}) = \hat{L}\hat{T} + Y$ liefert, wobei auch die Normen der Spalten von Y bereits bestimmt worden sind.

Es gilt

$$\begin{aligned}
\mathrm{fl}(\hat{L}^{[m]})\Delta &= \hat{L}\hat{T}\Delta + Y\Delta \\
&= (\hat{L}\hat{T}\Delta\hat{T}^{-1} + Y\Delta\hat{T}^{-1})\hat{T} \\
&= (\hat{L} + \hat{L}(\hat{T}\Delta\hat{T}^{-1} - I) + Y\Delta\hat{T}^{-1})\hat{T} \\
&= (\hat{L} + \delta\hat{L})\hat{T}\ .
\end{aligned}$$

Auf diese Weise haben wir also ein $\delta\hat{L}$ definiert, das (3.72) genügt und für das bis auf einen kleinen relativen Fehler

$$\begin{aligned}
\|\delta\hat{L}_S\|_2 &= \|\delta\hat{L}\,\mathrm{diag}(\|[\hat{L}_{\cdot i}]\|_2)^{-1}\|_2 \\
&= \|\delta\hat{L}\|_2 \\
&= \|\hat{L}(\hat{T}\Delta\hat{T}^{-1} - I) + Y\Delta\hat{T}^{-1}\|_2 \\
&\leq \|\hat{L}\|_2\|\hat{T}\Delta\hat{T}^{-1} - I\|_2 + \|Y\Delta\hat{T}^{-1}\|_2
\end{aligned}$$

gilt. Dabei ist $\|\hat{L}\|_2 \leq \sqrt{2m}$. Weiter ist

$$\|\hat{T}\Delta\hat{T}^{-1} - I\|_2 = \|\hat{T}(\Delta - I)\hat{T}^{-1}\|_2 \leq \|\hat{T}\|_2\|\hat{T}^{-1}\|_2\|\Delta - I\|_2$$

und wir können mit

$$\bar{R}^{[p]} = \begin{bmatrix} I_{p-1} & 0 \\ \tilde{r}^{[p]^T} & \tilde{\tau}^{[p]} \end{bmatrix}$$

bis auf einen kleinen relativen Fehler

$$\begin{aligned}
\|\hat{T}\|_2 &= \max\{\|\prod_{p=2}^{m}\hat{R}_1^{[p]}\|_2, \|\prod_{p=2}^{m}\hat{R}_2^{[p]}\|_2\} \\
&\leq \max\{\|(\prod_{p=2}^{m}\hat{R}_1^{[p]^{-T}})^{-T}\|_2, \\
&\qquad \|\mathrm{diag}(E_1^2,\ldots,E_m^2)\prod_{p=2}^{m}(\bar{R}^{[p]})\,\mathrm{diag}(E_1^{-2},\ldots,E_m^{-2})\|_2\} \\
&\leq \max\{(1 - \|I - \prod_{p=2}^{m}\bar{R}^{[p]}\|_E)^{-1}, \frac{E_1^2}{E_m^2}(1 + \|I - \prod_{p=2}^{m}\bar{R}^{[p]}\|_E)\} \\
&\leq \max\{(1-\mu)^{-1}, (1+\mu)^2\} \\
&= (1+\mu)^2
\end{aligned}$$

sowie

$$\begin{aligned}
\|\hat{T}^{-1}\|_2 &\leq \|(\hat{E}T\hat{S}^{-1}\hat{E}^{-1})^{-1}\|_2 \leq \|\hat{E}\hat{S}\|_2\|\hat{E}^{-1}\|_2\|T^{-1}\|_2 \\
&= \|\hat{E}\hat{S}\|_2\|\hat{E}^{-1}\|_2\|T\|_2 \leq \|\hat{E}\hat{S}\|_2\|\hat{E}^{-1}\|_2\|\hat{E}^{-1}\hat{T}\hat{E}\hat{S}\|_2 \\
&\leq \|(\hat{E}\hat{S})^2\|_2\|\hat{E}^{-2}\|_2\|\hat{T}\|_2 \leq E_1^2/E_m^2(1+\mu)^2 \leq (1+\mu)^3
\end{aligned}$$

abschätzen. Also ist in erster Ordnung von ε

$$\|\hat{L}\|_2\|\hat{T}\Delta\hat{T}^{-1} - I\|_2 \leq \sqrt{2m} \cdot 2\,(1+\mu)^5\,\varepsilon \leq 3.34\sqrt{m}\,\varepsilon\ . \tag{3.73}$$

Schließlich bestimmen wir noch eine obere Schranke von $\|Y\Delta\hat{T}^{-1}\|_2$. Wegen (3.66), (3.67) und (3.68) gilt in erster Ordnung von ε

$$\begin{aligned} \|Y\Delta\hat{T}^{-1}\|_2 &\leq \|Y\|_2\|\hat{T}^{-1}\|_2 \\ &\leq (1.07\sqrt{m}(m+2)\,\varepsilon + \varepsilon + 1.01\,\sqrt{m}\,m\,\varepsilon)(1+\mu)^3 \\ &\leq 3.30\,m\,\sqrt{m}\,\varepsilon\ . \end{aligned} \tag{3.74}$$

Addition von (3.73) und (3.74) ergibt

$$\|\delta L_S\|_2 = \|\delta\hat{L}_S\|_2 \leq m\,\sqrt{m}\,(3.34/m + 3.30)\,\varepsilon \leq 4.5m\sqrt{m}\varepsilon\ ,$$

also die Aussage des Satzes. Q.E.D.

Wir können nun die zweite Transformation des modifizierten Zyklus, also die Transformation **2.3.3**, **2.3.4** von Verfahren 2.2.4 behandeln. Auch hier ist es nützlich, das relevante Programmstück nochmals auf eine Weise niederzuschreiben, die der Analyse zugänglicher ist. Dabei gehen wir davon aus, daß die Eingangsmatrizen L und E vom Programmstück auf Seite 135 unter den Voraussetzungen des Satzes 3.2.25 erzeugt wurden. Dann gelten für die Spalten $[L_{\cdot i}] = [L_{\cdot i}^{[m]}]$ die Aussagen von Satz 3.2.25 und es ist $E_1 = \ldots = E_m = E_{n+1} \geq \ldots \geq E_{n+m}$ sowie bis auf einen kleinen relativen Fehler $E_{n+m} \geq (1+\mu)^{-1}E_{n+1}$. Wir setzen o.B.d.A. wieder *num*= 1 und *low*[*clu*] = 1 sowie $3 \leq m = upp[clu] \leq n$ voraus. Das zu untersuchende Programmstück lautet dann

```
FOR p = 1 TO m DO
   U_pp = - Σ_{k=1}^{2n} L_kp L_{k,n+p}
   FOR q = p + 1 TO m DO
      y_pq = Σ_{k=1}^{2n} L_{k,n+q} L_kp
      z_pq = Σ_{k=1}^{2n} L_kq L_{k,n+p}
      U_pq = -(y_pq + (E_{n+p}/E_{n+q}) z_pq)/2
      U_qp = -((E_{n+q}/E_{n+p}) y_pq + z_pq)/2
   ENDFOR
ENDFOR
FOR p = 1 TO m DO
   FOR q = 1 TO m DO
      [L'_{·n+p}] = [L_{·n+p}] + U_qp [L_{·q}]
   ENDFOR
ENDFOR
```

Es sei nochmals angemerkt, daß E nicht verändert wird. Hinsichtlich der Fehleranalayse ist vorstehendes Programmstück äquivalent zu den Teilen **2.3.3**, **2.3.4** von Verfahren 2.2.4. Vorstehende Darstellung hat den Vorteil, daß zuerst die Matrix U vollständig explizit gebildet wird und anschließend die Transformation durchgeführt wird. $[L_{\cdot i}]$ für $i \leq m$ und E bleiben unverändert bestehen.

Satz 3.2.27 *Die Matrizen L und $E = diag(E_1 \dots E_{2n})$ seien unter Anwendung des Satzes 3.2.25 entstanden, die Aussagen von Satz 3.2.25 seien also erfüllt und es sei $E_1 = E_m = E_{n+1} \geq \dots \geq E_{n+m}$ sowie bis auf einen kleinen relativen Fehler $E_{n+m} \geq (1+\mu)^{-1} E_{n+1}$. Der Teil* **2.3.3**, **2.3.4** *von Verfahren 2.2.4 werde für einen Cluster der Ordnung $2m$ in endlicher Genauigkeit ε durchgeführt. Dann gilt folgendes Diagramm*

$$\begin{tikzcd}
L, E \arrow[r, "\text{Fließpunkt}", "\text{Berechnung}"'] \arrow[d, "+\delta L"'] & fl(L'), E \\
L + \delta L, E \arrow[ur, "\text{exakt}"'] &
\end{tikzcd}$$

Dabei bedeutet der horizontale Pfeil, daß $fl(L'), E$ durch Fließpunkt-Berechnung mit Hilfe des Programmstückes aus den Eingangsmatrizen L, E entstehen. Der diagonale Pfeil bedeutet, daß $(L + \delta L) \cdot E$ durch exakte Rechtsmultiplikation mit einer symplektischen Matrix in $fl(L') \cdot E$ übergeht. Schließlich bedeutet der senkrechte Pfeil, daß $L + \delta L$ durch Addition einer Störung aus L hervorgeht. Schreibt man $L = L_S diag(\|L_{\cdot i}\|_2)$ und $\delta L = \delta L_S diag(\|L_{\cdot i}\|_2)$, so ist δL_S in erster Ordnung von ε durch

$$\|\delta L_S\|_2 \leq 2.9\, m \sqrt{m}\, \varepsilon \tag{3.75}$$

beschränkt.

Beweis:
Wie angekündigt, wird nicht das Programmstück **2.3.3**, **2.3.4** von Verfahren 2.2.4 analysiert, sondern das äquivalente Programmstück auf Seite 149. Sei $\hat{U} = (\hat{U})_{1 \leq p,q \leq m}$ und für $1 \leq p, q \leq m$ sei

$$\hat{U}_{qp} = -\frac{1}{2}\left(\frac{E_{n+q}}{E_{n+p}} \mathrm{fl}([L_{\cdot n+q}]^T [L_{\cdot p}]) + \mathrm{fl}([L_{\cdot q}]^T [L_{\cdot n+p}])\right) . \tag{3.76}$$

Dann ist $\tilde{U}_{qp} = \mathrm{fl}(U_{qp}) = \hat{U}_{qp} + r_{qp}$ mit

$$|r_{qp}| \leq \varepsilon\left(2\frac{E_{n+q}}{E_{n+p}} |\mathrm{fl}([L_{\cdot n+q}]^T [L_{\cdot p}])| + |\mathrm{fl}([L_{\cdot q}]^T [L_{\cdot n+p}])|\right) .$$

Wegen Lemma 3.2.4, (3.52) und (3.57) gilt

$$\begin{aligned}
|\mathrm{fl}([L_{\cdot n+q}]^T[L_{\cdot p}])| &\leq |[L_{\cdot n+q}]^T[L_{\cdot p}]| + n\,\varepsilon\,\|[L_{\cdot n+q}]\|_2\|[L_{\cdot p}]\|_2 \\
&\leq 4\,\mu + n\,\varepsilon(1+\frac{\sqrt{\mu}}{2})(1+2\,\mu) \\
&\leq 4\,\mu(1+\frac{n}{4\,\mu}\,\frac{\mu}{2\,n}\,(1+\frac{\sqrt{\mu}}{4})(1+2\,\mu) \leq 4.59\,\mu
\end{aligned}$$

und analog $|\mathrm{fl}([L_{\cdot q}]^T[L_{\cdot n+p}])| \leq 4.59\,\mu$. Es folgt

$$\begin{aligned}
|r_{qp}| &\leq \varepsilon\,(2\frac{E_{n+q}}{E_{n+p}}+1)\cdot 4.59\,\mu \\
&\leq 14.08\,\mu\,\varepsilon
\end{aligned}$$

und bis auf einen kleinen relativen Fehler

$$\begin{aligned}
|\tilde{U}_{qp}| &\leq \frac{1}{2}\,(\frac{E_{n+q}}{E_{n+p}}|\mathrm{fl}([L_{\cdot n+q}]^T[L_{\cdot p}])| + |\mathrm{fl}([L_{\cdot q}]^T[L_{\cdot n+p}])| \\
&\leq \frac{1}{2}((1+\mu)+1)\cdot 4.59\,\mu \leq 4.67\,\mu\ .
\end{aligned}$$

Es folgt

$$\begin{aligned}
&\mathrm{fl}(L'_{k,n+p}) \\
&= \mathrm{fl}(L_{k,n+p} + \sum_{q=1}^{m} L_{kq}U_{qp}) \\
&= L_{k,n+p} + \sum_{q=1}^{m} L_{kq}\tilde{U}_{qp} + W'_{kp}\ ,\ |W'_{kp}| \leq \varepsilon\,(m+1)(|L_{k,n+p}| + \sum_{q=1}^{m}|L_{kq}\tilde{U}_{qp}|)\ .
\end{aligned}$$

Wegen

$$\sum_{q=1}^{m} L_{kq}\tilde{U}_{qp} = \sum_{q=1}^{m} L_{kq}\hat{U}_{qp} + \sum_{q=1}^{m} L_{kq}r_{qp}$$

folgt

$$\mathrm{fl}(L'_{k,n+p}) = L_{k,n+p} + \sum_{q=1}^{m} L_{kq}\hat{U}_{qp} + W''_{kp}$$

und wegen (3.76) ist in erster Ordnung von ε

$$\begin{aligned}
|W''_{kp}| &\leq \varepsilon\,(m+1)(|L_{k,n+p}| + \sum_{q=1}^{m}|L_{kq}\tilde{U}_{qp}|) + 14.08\,\mu\,\varepsilon\sum_{q=1}^{m}|L_{kq}| \\
&\leq \varepsilon\,(m+1)|L_{k,n+p}| + \mu\,\varepsilon((m+1)\cdot 4.67 + 14.08)\sum_{q=1}^{m}|L_{kq}|\ .
\end{aligned}$$

Also ist

$$[\mathrm{fl}(L'_{\cdot n+p})] = [L_{\cdot n+p}] + \sum_{q=1}^{m} [L_{\cdot q}]\hat{U}_{qp} + [W''_{\cdot p}]$$

mit

$$\|[W''_{\cdot p}]\|_2 \leq \varepsilon\,(m+1)\|[L_{\cdot n+p}]\|_2 + \varepsilon\,\mu\,(4.67\,m + 18.75)\sqrt{m}\|[L_{\cdot 1}\cdots L_{\cdot m}]\|_E \ .$$

Wegen (3.56) ist $\|[L_{\cdot 1}\cdots L_{\cdot m}\|_E^2 \leq m\,(1+2\,\mu)^2$, so daß folgt

$$\begin{aligned} \|[W''_{\cdot p}]\|_2 &\leq \varepsilon\,[(m+1)(1+\sqrt{\mu}/2) + \mu\,(4.67\,m + 18.75)\,m\,(1+2\,\mu)] \\ &\leq 2.62\,m\,\varepsilon \ . \end{aligned}$$

In erster Ordnung von ε ist daher

$$\begin{aligned} [\mathrm{fl}(L'_{\cdot n+1})\cdots\mathrm{fl}(L'_{\cdot n+m})] &= \mathrm{fl}([L_{\cdot n+1}\cdots L_{\cdot n+m}] + [L_{\cdot 1}\cdots L_{\cdot m}]U) \\ &= [L_{\cdot n+1}\cdots L_{\cdot n+m}] + [L_{\cdot 1}\cdots L_{\cdot m}]\hat{U} + [W''_{\cdot 1}\cdots W''_{\cdot m}] \end{aligned}$$

mit

$$\|[W''_{\cdot 1}\cdots W''_{\cdot m}]\|_2 \leq \sqrt{m}\max_{1\leq p\leq m}\|[W''_{\cdot p}]\|_2 \leq 2.62\sqrt{m}\,m\,\varepsilon \ .$$

Wir werden nun ein geeignetes symplektisches T und daraus ein δL konstruieren, das den Behauptungen des Satzes genügt. Es gilt

$$\mathrm{fl}(L') = LT' + W \quad , \quad T' = \begin{bmatrix} I & \hat{U} \\ 0 & I \end{bmatrix} \ , \quad W = [\underbrace{0}_{n}\ W''_{\cdot 1}\cdots W''_{\cdot m}] \ .$$

Wir werden zeigen, daß $T = E^{-1}T'E$ symplektisch ist. Es gilt

$$\begin{aligned} E^{-1}T'E &= \begin{bmatrix} I & (1/E_1)\hat{U}\mathrm{diag}(E_{n+1}\ldots E_{n+m}) \\ 0 & I \end{bmatrix} \ , \quad \text{denn} \quad E_1 = \ldots = E_m \\ &= \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} , \end{aligned}$$

wobei

$$X_{qp} = \frac{E_{n+p}}{E_{n+1}}\hat{U}_{qp} = -\frac{1}{2E_{n+1}}(E_{n+q}\mathrm{fl}([L_{\cdot n+q}]^T[L_{\cdot p}]) + E_{n+p}\mathrm{fl}([L_{\cdot q}]^T[L_{\cdot n+p}])) = X_{pq}$$

ist. X ist also symmetrisch und T symplektisch. Aus diesem T gewinnen wir nun ein δL, das (3.75) genügt. Es gilt

$$\begin{aligned} &\mathrm{fl}(L')E = (L+\delta L)ET \\ &\iff LT' + W = (L+\delta L)ETE^{-1} \iff L + WT'^{-1} = L + \delta L \\ &\iff \delta L = WT'^{-1} \iff \delta L = W \quad , \text{denn } T'^{-1} = \begin{bmatrix} I & -U \\ 0 & I \end{bmatrix} \\ &\iff \delta L_S = W\mathrm{diag}(\|[L_{\cdot i}]\|_2^{-1}) \ . \end{aligned}$$

Wir haben also ein δL bestimmt, das der Abschätzung

$$\begin{aligned}\|\delta L_S\|_2 &\leq \|[W''_{\cdot 1}\cdots W''_{\cdot m}]\|_2 \max_{1\leq i\leq m}(\|[L_{\cdot n+i}]\|_2^{-1}) \leq 2.62\, m\sqrt{m}\,\varepsilon\cdot(1-\sqrt{\mu}/2)^{-1}\\ &\leq 2.89\, m\sqrt{m}\,\varepsilon\end{aligned}$$

genügt und für das deshalb die Aussage des Satzes gilt. Q.E.D.

Wie im vorangehenden Unterabschnitt fassen wir den bei der Operation **2.3** entstehenden maximalen Rückwärtsfehler zusammen.

Satz 3.2.28 *Die Matrizen $L, E \in I\!R^{2n\times 2n}$ werden gemäß* **2.3** *von Verfahren 2.2.4 in endlicher Genauigkeit ε bearbeitet transformiert. Der nichttriviale Teil der beteiligten Transformationsmatrizen sei jeweils m und es werde lediglich ein Cluster transformiert. Es sei $\max\{7.5\sqrt{m}\varepsilon, 4.5\, m\sqrt{m}\varepsilon\} < \sigma_{\min}(L_S)$ und $L = L_S D$ mit $D = diag(\|L_{\cdot i}\|_2)$. Die nach der Transformation entstandenen Matrizen seien L' und E'. Dann gilt für die Eigenwerte*

$$-\lambda_n \leq \ldots \leq -\lambda_1 < 0 < \lambda_1 \leq \ldots \leq \lambda_n$$
$$-\lambda'_n \leq \ldots \leq -\lambda'_1 < 0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

des Problems $EL^TLEx = i\lambda Jx$ bzw. des Problems $E'L'^TL'E'x = i\lambda Jx$ in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \leq \frac{(16\, m + 15)\sqrt{m}}{\sigma_{\min}(L_S)}\,\varepsilon\ , \quad k = 1\ldots n \tag{3.77}$$

Beweis:
Wegen Satz 3.2.10 und wegen des Wachstums der skalierten Kondition gemäß Satz 3.1.6 wird $4\sqrt{2m}\cdot 1.2\cdot 1.1\,\varepsilon \leq 7.5\,\varepsilon < \sigma_{\min}(L_S)$ vorausgesetzt; wegen Satz 3.2.26 ist $4.5\, m\sqrt{m}\,\varepsilon < \sigma_{\min}(L_S)$ vorauszusetzen. Wir werden wieder die nach den betrachteten Operationen von Verfahren 2.2.4 jeweils entstandenen Matrizen mit einem oberen Index versehen. Z.B. seien $L^{2.3.5} = L'$, $E^{2.3.5} = E'$ die abschließend entstandenen Matrizen. Für die Störungen und die skalierten Matrizen werden wir entsprechend verfahren. Dann gilt

wegen	die Aussage
3.2.26	$\|\delta L_S\|_2 \leq 4.5\, m\sqrt{m}\varepsilon$
3.2.27	$\|\delta L_S^{2.3.2}\|_2 \leq 2.9\, m\sqrt{m}\varepsilon$
3.2.10,3.2.1	$\|\delta L_S^{2.3.5}\|_2 \leq 4\sqrt{2\, m}\,\varepsilon$

Aus Satz 3.1.6 folgt wegen $\omega \leq 1/30$ nach Wurzelbildung $1/\sigma_{\min}(L_S^{2.3.2}) \leq 1.1/\sigma_{\min}(L_S)$ und $1/\sigma_{\min}(L'_S) \leq 1.2/\sigma_{\min}(L_S^{2.3.2})$. Wegen Satz 1.0.5 und Lemma 3.2.2 folgt in erster Ordnung von ε

$$\frac{|\lambda_k - \lambda'_k|}{\lambda_k} \leq \frac{2}{\sigma_{\min}(L_S)}\sqrt{m}(4.5\, m + 1.1\,(2.9\, m + 1.2\cdot 4\sqrt{2}))\,\varepsilon\ .$$

Daraus folgt die Behauptung. Q.E.D.

Kapitel 4

Schiefsymmetrische Zerlegung $\mathbf{PSP^T = LJL^T}$

Soll zur Lösung des schiefsymmetrischen Eigenwertproblems

$$Sx = \lambda x \tag{4.1}$$

für nichtsinguläres schiefsymmetrisches $S \in \mathbb{R}^{2n\times 2n}$ ein (implizites) Jacobi-ähnliches Verfahren gemäß Kapitel 2 angewendet werden, so ist zunächst eine Zerlegung

$$PSP^T = LJL^T \quad , \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \tag{4.2}$$

notwendig, wobei P eine vorzeichenbehaftete Permutationsmatrix ist und L eine permutierte untere Dreiecksmatrix. Die Vorzeichen in P könnte man vermeiden, wenn man etwas höhere Rechenzeit in Kauf nähme. Die Fehleranalyse wäre dann jedoch dieselbe.

Im nächsten Abschnitt dieses Kapitels wird ein Verfahren zur Zerlegung (4.2) beschrieben. Dabei handelt es sich um eine bereits in [6] eingeführte, auf [7] basierende Methode. Im daran anschließenden Abschnitt nehmen wir die Fehleranalyse des Verfahrens vor. Ein Abschnitt über die Beschränktheit der Kondition der Matrix

$$K_S = D^{-1}L^T L D^{-1} \quad , \quad D = \operatorname{diag}(\|L_{\cdot i}\|_2) \tag{4.3}$$

beenden das Kapitel.

Bevor wir im Detail auf den Algorithmus der Zerlegung eingehen, soll er zur Verdeutlichung des Vorgehens in groben Zügen skizziert werden. Statt der Zerlegung (4.2) wird zuerst die Zerlegung

$$PSP^T = L' J_n L'^T \quad , \quad \text{wobei} \quad J_k = \bigoplus_{i=1}^{k} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad , \quad k \in \mathbb{N} \tag{4.4}$$

mit einer vorzeichenbehafteten Permutationsmatrix P, die durch die Pivotierung bestimmt wird, und einer unteren Dreiecksmatrix L' vorgenommen. Wir werden nun einen Schritt der Zerlegung (4.4) betrachten. Sei $\bar{P}$ eine vorzeichenbehaftete Permutationsmatrix mit

$$\bar{P}S\bar{P}^T = \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} \quad , \quad A = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix} \quad , \quad a > 0 \ , \tag{4.5}$$

wobei

$$a = \max_{i,j} |S_{ij}| \ . \tag{4.6}$$

Weil S nichtsingulär ist, kann tatsächlich $a > 0$ gefordert werden, und es gibt stets ein geeignetes $\bar{P}$. Es gilt

$$\bar{P}S\bar{P}^T = \bar{L} \begin{bmatrix} J_1 & 0 \\ 0 & S_1 \end{bmatrix} \bar{L}^T \tag{4.7}$$

mit

$$\bar{L} = \begin{bmatrix} B & 0 \\ N & I_{2n-2} \end{bmatrix} \quad , \quad B = a^{1/2} I_2 \ , \quad N = -a^{-1/2} M J_1 \quad , \quad S_1 = S' - N J_1 N^T \ . \tag{4.8}$$

Rekursive Anwendung von (4.7) führt zur Zerlegung (4.4), wobei P das Produkt der $\bar{P}$ ist. Für die durch die Permutation

$$\begin{pmatrix} 1 & 2 \cdots & n & n+1 & n+2 & \cdots & 2n \\ 1 & 3 \cdots & 2n-1 & 2 & 4 & \cdots & 2n \end{pmatrix}$$

definierte Permutationsmatrix $\hat{P}$ gilt

$$\hat{P}^T J_n \hat{P} = J \ , \tag{4.9}$$

also kann man die Zerlegung (4.2) mit einer Vertauschung von Spalten gemäß

$$L = L' \hat{P} \tag{4.10}$$

gewinnen.

Wählt man $\bar{P}$ in einem Schritt (4.5) so, daß (4.6) gilt, so erfolgt die Suche nach einer geeigneten Permutationsmatrix in der Zeit $\mathcal{O}(n^2)$, in der gesamten Zerlegung also in der Zeit $\mathcal{O}(n^3)$. Zu dieser Wahl von $\bar{P}$ gibt es jedoch Alternativen. Wählt man wie in [6] z.B. $\bar{P}$ im k-ten Schritt, $1 \leq k \leq n$, so, daß a das betragsgrößte Element der $2k-1$- und $2k$-ten Spalten von S wird, so kann die Pivotsuche bereits in der Zeit $\mathcal{O}(n)$ pro Schritt, also in der Zeit $\mathcal{O}(n^2)$ in der gesamten Zerlegung, erfolgen.

Die Rechenzeit der Zerlegung macht nur einen kleinen Teil der Rechenzeit des gesamten Verfahrens zur Lösung des Eigenwertproblems (4.1) aus, so daß die *vollständige Pivotsuche* gemäß (4.6) vorgezogen wird, zumal dann die Beschränktheit der Kondition von K_S aus (4.3) gezeigt werden kann und sich gemäß [6] das kleinste maximale Wachstum der Matrixelemente beim Übergang von S zu S_1 einstellt. Nach [6] ist das maximale Elementwachstum (der Quotient von betragsgrößtem Element in allen Zerlegungen und $\max_{i,j}(|S_{ij}|)$) im Falle vollständiger Pivotsuche durch

$$1.8\ \sqrt{2n}(2n)^{(\ln(2n))/4}$$

beschränkt.

4.1 Der Algorithmus der Zerlegung

In diesem Abschnitt geben wir den vollständigen Algorithmus der Zerlegung (4.2) in einem Pseudocode an. Im nächsten Abschnitt werden wir die Fehleranalyse dieses Algorithmus' bei Rechnung in endlicher Arithmetik vornehmen. Die Matrix $S = -S^T \in \mathbb{R}^{2n\times 2n}$ sei im Feld S gespeichert und der Ganzzahlvektor P der Länge $2n$ wird Informationen zur Permutation gemäß (4.2) aufnehmen. Mit γ und ϱ seien je ein temporärer Ganzzahl- bzw. Fließpunkt-Hilfsvektor der Länge $2n$ bereitgestellt. Die Permutatiomsmatrix P wird durch einen gleichnamigen Indexvektor repräsentiert. Wir können den Algorithmus der Zerlegung durch die Prozedur

```
PROCEDURE ljlt(input,output: S, P; γ, ϱ)
   FOR i = 1 TO 2n DO P_i = i
   FOR k = 1 TO n DO
      pivot(S, k, p, q, σ)                          /* search pivot element */
      IF σ = 0 THEN
         WRITE('matrix singular')
         STOP
      ELSE
         perm(S, k, p, q, σ < 0, P)                 /* permute matrix */
         newcol(S, k, (−S_{2k,2k−1})^{1/2})         /* compute new columns */
         update(S, k)                               /* compute reduced matrix */
      ENDIF
   ENDFOR
   makefactor(S, γ, ϱ)                              /* final permutation */
END ljlt
```

ausdrücken, die im Wesentlichen aus einigen weiteren Prozeduraufrufen besteht. Letztere werden nachfolgend beschrieben. In der Prozedur pivot wird die Position eines Elementes gemäß (4.6) bestimmt:

```
PROCEDURE pivot(input:S, k; output:p, q, σ)
   η = 0
   FOR j = 2k − 1 TO 2n − 1 DO
      FOR i = j + 1 TO 2n DO
         IF |S_ij| > |η| THEN
            η = S_ij
            p = i
            q = j
         ENDIF
      ENDFOR
   ENDFOR
   σ = sign(η)
END pivot
```

Mit der Prozedur perm werden die Zeilen und Spalten von S nun so permutiert, daß das Element an der Position (p, q) an die Position $(2k, 2k-1)$ rückt und dort negatives Vorzeichen trägt. P wird entsprechend modifiziert. Dabei werden wir zur Abkürzung die Notation $x \longleftarrow\!\oplus\!\longrightarrow y$ bzw. $x \longleftarrow\!\ominus\!\longrightarrow y$ für die Programmzeilen

$$\begin{array}{lcl} \eta = y & & \eta = -y \\ y = x & \text{bzw.} & y = -x \\ x = \eta & & x = \eta \end{array}$$

verwenden. Damit lautet die Prozedur

```
PROCEDURE perm(input,output: S; input: k, p, q,negative; input,output: P)
   FOR i = 1 TO 2k − 2 DO S_{2k−1,i} ←⊕→ S_qi
   FOR i = 2k TO q − 1 DO S_{i,2k−1} ←⊖→ S_qi
   FOR i = q + 1 TO 2n DO S_{i,2k−1} ←⊕→ S_iq
   P_{2k−1} ←⊕→ P_q
   S_{q,2k−1} = −S_{q,2k−1}
   IF negative THEN
      FOR i = 1 TO 2k − 1 DO S_{2k,i} ←⊕→ S_pi
      FOR i = 2k + 1 TO p − 1 DO S_{i,2k} ←⊖→ S_pi
      FOR i = p + 1 TO 2n DO S_{i,2k} ←⊕→ S_ip
      P_2k ←⊕→ P_p
   ELSE
      FOR i = 1 TO 2k − 1 DO S_{2k,i} ←⊖→ S_pi
      FOR i = 2k + 1 TO p − 1 DO S_{i,2k} ←⊕→ S_pi
      FOR i = p + 1 TO 2n DO S_{i,2k} ←⊖→ S_ip
      P_2k ←⊖→ P_p
   ENDIF
```

```
    S_{p,2k} = -S_{p,2k}
END perm
```

Man könnte auf das Vorzeichen in P auch verzichten, wenn man andere Permutationen verwenden würde. Auch auf die Prozedur perm könnte man zur Laufzeitoptimierung verzichten und statt dessen in permutierter Weise auf die Elemente von S zugreifen. Aus Vereinfachungsgründen werden wir darauf nicht näher eingehen. Mit der Prozedur newcol wird der neue Inhalt der Spalten $2k-1$ und $2k$ von S, also der Submatrix N in (4.8), bestimmt.

```
PROCEDURE newcol(input,output: S; input:k, c)
    S_{2k,2k-1} = 0
    S_{2k-1,2k-1} = c
    S_{2k,2k} = c
    c' = 1/c
    FOR i = 2k + 1 TO 2n DO
        η = S_{i,2k} · c'
        S_{i,2k} = -S_{i,2k-1} · c'
        S_{i,2k-1} = η
    ENDFOR
END newcol
```

In update wird die Matrix S_1 gemäß (4.7), (4.8) gebildet.

```
PROCEDURE update(input,output: S; input: k)
    FOR j = 2k + 1 TO 2n - 1 DO
        FOR i = j + 1 TO 2n DO
            S_{ij} = S_{ij} - S_{i,2k-1} S_{j,2k} + S_{i,2k} S_{j,2k-1}
        ENDFOR
    ENDFOR
END update
```

Nach Abschluß von update sind die Elemente in und unterhalb der Diagonalen von S die Elemente des Faktors. In der Prozedur makefactor werden die Elemente oberhalb der Diagonalen von S gelöscht und die Permutation (4.10) wird vorgenommen, um schließlich den Faktor L zu erhalten.

```
PROCEDURE makefactor(input,output: S; h, v)
    FOR j = 2 TO 2n DO
        FOR i = 1 TO j - 1 DO S_{ij} = 0
        h_i = 0
    ENDFOR
    k = 2
```

```
A: h_k = 1
   FOR i = 1 TO 2n DO v_i = S_ik
B: IF k ≤ n THEN l = 2k − 1 ELSE l = 2(k − n)
   IF h_l = 0 THEN
      FOR i = 1 TO 2n DO S_ik = S_il
      k = l
      h_k = 1
      GOTO B
   ELSE
      FOR i = 1 TO 2n DO S_ik = v_i
   ENDIF
   FOR i = 2 TO 2n − 1 DO
      IF h_i = 0 THEN
         k = i
         GOTO A
      ENDIF
   ENDFOR
END makefactor
```

Nach Abschluß von makefactor ist die Matrix L im Feld S gespeichert. Der Vektor P bestimmt die vorzeichenbehaftete Permutationsmatrix P gemäß (4.2). P wird bei der Bestimmung der Eigenvektoren von S noch einmal benötigt.

4.2 Fehleranalyse

In diesem Abschnitt nehmen wir die Fehleranalyse der in Zerlegung aus 4.1 vor. Mit der Notation $|X|_{ij} = (|X_{ij}|)$ gilt

Satz 4.2.1 *Sei $S = -S^T \in \mathbb{R}^{2n\times 2n}$. Wird der Algorithmus aus 4.1 in endlicher Genauigkeit ε durchgeführt und bricht er nicht wegen einer berechneten singulären Zwischenmatrix ab, so gilt in erster Ordnung von ε für den berechneten Faktor $\tilde{L} = fl(L)$*

$$\begin{aligned} \tilde{L} J \tilde{L}^T &= PSP^T + F \\ |F| &\leq 3n\,(|PSP^T| + |\tilde{L}|\ |J|\ |\tilde{L}|^T)\,\varepsilon\ . \end{aligned} \tag{4.11}$$

Dabei ist P eine vorzeichenbehaftete Permutationsmatrix.

Beweis:
Wie in [29] wird die Beweistechnik aus [18] zur Fehleranalyse der Gauß-Elimination verwendet. Die obige Prozedur makefactor, die u.a. die Permutation (4.10) vornimmt, wird in exakter Arithmetik durchgeführt. Mit $\hat{P}$ aus (4.9) ist daher (4.11) äquivalent zu

$$\begin{aligned} \tilde{L}' J_n \tilde{L}'^T &= PSP^T + F \\ |F| &\leq 3\,n\ (|PSP^T| + |\tilde{L}'|\ |J_n|\ |\tilde{L}'|^T)\ \varepsilon\ , \end{aligned} \tag{4.12}$$

wobei $\tilde{L}' = \mathrm{fl}(L')$ die in Fließpunktarithmetik berechnete Matrix in und unterhalb der Diagonale des Feldes S vor Eintritt in die Prozedur makefactor ist. Statt (4.11) wird (4.12) mit vollständiger Induktion über n gezeigt.

Für $n = 1$ ist die Aussage offensichtlich erfüllt. Zum Beweis des Induktionsschrittes betrachten wir

$$\begin{aligned} P_0 S P_0^T &= \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} \quad , \quad A = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix} \quad , \quad a > 0 \\ &= \begin{bmatrix} B & 0 \\ N & I \end{bmatrix} \begin{bmatrix} J_1 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} B^T & N^T \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} BJ_1B^T & BJ_1N^T \\ NJ_1B^T & NJ_1N^T + S_1 \end{bmatrix} \end{aligned}$$

mit einer vorzeichenbehafteten Permutationsmatrix P_0. Dabei gilt in exakter Arithmetik

$$\begin{aligned} B &= a^{1/2} I \\ N &= -a^{-1/2} M J_1 \\ S_1 &= S' - N J_1 N^T \end{aligned}$$

und gemäß betrachtetem Algorithmus in endlicher Arithmetik

$$\begin{aligned} \tilde{B} &= \mathrm{fl}(B) = \mathrm{fl}(a^{1/2} I) = B + \delta B \quad , \quad |\delta B| \le |B|\, \varepsilon \\ \tilde{N} &= \mathrm{fl}(N) = \mathrm{fl}(-(a^{-1/2}) M J_1) = N + \delta N \quad , \quad |\delta N| \le 3\, |N|\, \varepsilon \ . \end{aligned}$$

Auch $\tilde{S}_1 = \mathrm{fl}(S_1)$ soll bestimmt werden. Für $\tilde{N} = [\tilde{N}_{\cdot 1} \ \ \tilde{N}_{\cdot 2}]$ ist

$$(\tilde{N} J_1 \tilde{N}^T)_{ij} = \tilde{N}_{i1} \tilde{N}_{j2} - \tilde{N}_{i2} \tilde{N}_{j1}$$

und daher in endlicher Arithmetik

$$\begin{gathered} \mathrm{fl}((\tilde{N} J_1 \tilde{N})_{ij}) = (\tilde{N} J_1 \tilde{N})_{ij} + R_{ij} \ , \\ |R_{ij}| \le (|\tilde{N}_{i1} \tilde{N}_{j2}| + |\tilde{N}_{i2} \tilde{N}_{j1}|)\ 2\, \varepsilon\, (1 + \mathcal{O}(\varepsilon)) \ , \end{gathered}$$

also in erster Ordnung von ε

$$\mathrm{fl}(\tilde{N} J_1 \tilde{N}) = \tilde{N} J_1 \tilde{N} + R \quad , \quad |R| \le 2\ |\tilde{N}|\ |J_1|\ |\tilde{N}|^T\ \varepsilon \ .$$

Es folgt

$$\begin{gathered} \tilde{S}_1 = \mathrm{fl}(S' - \tilde{N} J_1 \tilde{N}) = S' - \tilde{N} J_1 \tilde{N} + \delta S_1 \ , \\ |\delta S_1| \le 3\ (|S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T)\ \varepsilon \ . \end{gathered} \tag{4.13}$$

Daraus ergibt sich auch

$$|\tilde{S}_1| \le (1 + 3\, \varepsilon)\ (|S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T) \ .$$

Nach Induktionsvoraussetzung wird von $\tilde{S}_1$ in endlicher Arithmetik ein Faktor $\tilde{L}_1'$ berechnet, so daß

$$\begin{aligned} \tilde{L}_1' J_{n-1} \tilde{L}_1'^T &= P_1 \tilde{S}_1 P_1^T + F_1 \ , \\ |F_1| \ &\leq \ 3\,(n-1)\ (|P_1 \tilde{S}_1 P_1^T| + |\tilde{L}_1'|\ |J_{n-1}|\ |\tilde{L}_1'|^T)\ \varepsilon \ , \end{aligned} \tag{4.14}$$

mit einer vorzeichenbehafteten Permutationsmatrix P_1 gilt. Damit haben wir

$$\begin{aligned} \tilde{S} \ &:= \ \begin{bmatrix} \tilde{B} & 0 \\ \tilde{N} & P_1^T \tilde{L}_1' \end{bmatrix} J_n \begin{bmatrix} \tilde{B}^T & \tilde{N}^T \\ 0 & \tilde{L}_1'^T P_1 \end{bmatrix} \\ &= \ \begin{bmatrix} \tilde{B} \tilde{J}_1 \tilde{B}^T & \tilde{B} J_1 \tilde{N}^T \\ \tilde{N} J_1 \tilde{B}^T & \tilde{N} J_1 \tilde{N}^T + P_1^T \tilde{L}_1' J_{n-1} \tilde{L}_1'^T P_1 \end{bmatrix} \ , \end{aligned}$$

wobei wegen (4.13) und der Induktionsvoraussetzung (4.14)

$$\begin{aligned} &\tilde{N} J_1 \tilde{N}^T + P_1^T \tilde{L}_1' J_{n-1} \tilde{L}_1'^T P_1 \\ &\quad = \ \tilde{N} J_1 \tilde{N}^T + S' - \tilde{N} J_1 \tilde{N}^T + \delta S_1 + P_1^T F_1 P_1 \\ &\quad = \ S' + F' \end{aligned}$$

mit

$$\begin{aligned} &|F'| \\ &\quad \leq \ |\delta S_1| + |P_1^T F_1 P_1| \\ &\quad \leq \ 3\,(|S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T)\,\varepsilon \\ &\qquad\quad +3\,(n-1)\,((1+3\,\varepsilon)(|S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T) + |P_1^T \tilde{L}_1'|\ |J_{n-1}|\ |P_1^T \tilde{L}_1'|^T)\,\varepsilon \\ &\quad \leq \ 3\,n\,(|S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T + |P_1^T \tilde{L}_1'|\ |J_{n-1}|\ |P_1^T \tilde{L}_1'|^T)\,\varepsilon \end{aligned}$$

in erster Ordnung von ε gilt. In erster Ordnung von ε folgt wegen $n \geq 2$

$$\tilde{S} = \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} + F_1'$$

mit

$$\begin{aligned} &|F_1'| \\ &\quad \leq \ \begin{bmatrix} 2\,|B|\ |J_1|\ |B|^T\,\varepsilon & 4\,|B|\ |J_1|\ |N|^T\,\varepsilon \\ 4\,|N|\ |J_1|\ |B|^T\,\varepsilon & |F'| \end{bmatrix} \\ &\quad \leq \ 3n \begin{bmatrix} |\tilde{B}|\ |J_1|\ |\tilde{B}|^T & |\tilde{B}|\ |J_1|\ |\tilde{N}|^T \\ |\tilde{N}|\ |J_1|\ |\tilde{B}|^T & |S'| + |\tilde{N}|\ |J_1|\ |\tilde{N}|^T + |P_1^T \tilde{L}_1'|\ |J_{n-1}|\ |P_1^T \tilde{L}_1'|^T \end{bmatrix} \varepsilon \\ &\quad \leq \ 3n\,(\begin{bmatrix} |A| & |-M|^T \\ |M| & |S'| \end{bmatrix} + \begin{bmatrix} |\tilde{B}| & 0 \\ |\tilde{N}| & |P_1^T \tilde{L}_1'| \end{bmatrix} J_n \begin{bmatrix} |\tilde{B}|^T & |\tilde{N}|^T \\ 0 & |P_1^T \tilde{L}_1'|^T \end{bmatrix})\ \varepsilon \ . \end{aligned}$$

Für

$$\tilde{L}' = \begin{bmatrix} \tilde{B} & 0 \\ P_1 \tilde{N} & \tilde{L}_1' \end{bmatrix} \quad , \quad P' = I_2 \oplus P_1 \quad , \quad P = P' P_0$$

ist also

$$\begin{aligned}
\tilde{L}' J_n \tilde{L}'^T &= P' \begin{bmatrix} \tilde{B} & 0 \\ \tilde{N} & P_1 \tilde{L}'_1 \end{bmatrix} J \begin{bmatrix} \tilde{B}^T & \tilde{N}^T \\ 0 & \tilde{L}'^T_1 P_1^T \end{bmatrix} P'^T \\
&= P' \begin{bmatrix} \tilde{B} J_1 \tilde{B}^T & \tilde{B} J_1 \tilde{N}^T \\ \tilde{N}_1 J \tilde{B}^T & \tilde{N} J_1 \tilde{N}^T + P_1 \tilde{L}'_1 J_{n-1} \tilde{L}'^T_1 P_1^T \end{bmatrix} P'^T \\
&= P' \left(\begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} + F'_1 \right) P'^T \\
&= P' (P_0 S P_0^T + F'_1) P'^T \\
&= PSP^T + F
\end{aligned}$$

mit

$$\begin{aligned}
|F| &\leq |P' F'_1 P'^T| \leq 3\,n\,(|P' P_0 S P_0^T P'^T| + |\tilde{L}'|\ |J_n|\ |\tilde{L}'|^T)\,\varepsilon \\
&\leq 3\,n\,(|PSP^T| + |\tilde{L}'|\ |J_n|\ |\tilde{L}'|^T)\,\varepsilon\ ,
\end{aligned}$$

womit der Induktionsschritt gezeigt ist. Q.E.D.

Der Beweis von Satz 4.2.1 verläuft über weite Strecken ähnlich zum Beweis einer analogen Aussage in [29] (Th. 4.2.1 und nachfolgende Bemerkungen) für symmetrische indefinite Zerlegungen im Falle von 1×1-Pivots. Die Schranke (4.11) ist jedoch um den Faktor 2 günstiger als die entsprechende Fehlerabschätzung in [29]. Ursache hierfür ist im Wesentlichen, daß beim Algorithmus 4.1 in jedem Schritt zwei Spalten verarbeitet werden.

4.3 Zur Beschränktheit der skalierten Kondition von $L^T L$

In diesem Abschnitt betrachten wir die Matrix K_S gemäß (4.3). Wir werden zeigen, daß ihre Kondition beschränkt ist, wobei die Schranke ausschließlich von der Dimension abhängt. Anschließend werden wir sog. $\gamma - r-$diagonaldominante schiefsymmetrische Matrizen einführen und für die wichtige Untermenge der $\gamma - 2-$diagonaldominanten Matrizen recht kleine obere Schranken für die Kondition $\kappa(K_S)$ abgeben.

Satz 4.3.1 *Wird die Zerlegung (4.2) von $S = -S^T \in I\!R^{2n \times 2n}$ gemäß Algorithmus 4.1 in exakter Arithmetik durchgeführt, so ist die Kondition*

$$\kappa(K_S) = \frac{\lambda_{\max}(K_S)}{\lambda_{\min}(K_S)}$$

von K_S aus (4.3) durch

$$\kappa(K_S) \leq n\ \frac{1 + 2\,n}{8}\ 9^n \tag{4.15}$$

beschränkt.

Beweis:
Wir verwenden die Beweistechnik aus [29]. Sei

$$\bar{P}S\bar{P}^T = \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} = \hat{\bar{L}} \begin{bmatrix} A & 0 \\ 0 & S_1 \end{bmatrix} \hat{\bar{L}}^T \quad , \quad A = aJ_1 \ , \quad a > 0 \tag{4.16}$$

mit J_1 gemäß (4.2), wobei die vorzeichenbehaftete Permutationsmatrix $\bar{P}$ gemäß Algorithmus 4.1 so gewählt wird, daß

$$a = \max_{i>j} |S_{ij}|$$

gilt. Wegen $|M_{ij}| \leq a$ ist dann

$$\hat{\bar{L}} = \begin{bmatrix} I_2 & 0 \\ -a^{-1}MJ_1 & I_{2n-2} \end{bmatrix} \quad , \quad |\hat{\bar{L}}_{ij}| \leq 1 \ , \quad 1 \leq i,j \leq 2n \ . \tag{4.17}$$

Rekursive Anwendung von (4.16) liefert die Zerlegung

$$PSP^T = \hat{L}'\hat{A}\hat{L}'^T = LJL^T \ , \tag{4.18}$$

wobei mit gewissen $\alpha_i > 0$ und $\alpha = \oplus_{i=1}^n (\alpha_i^{1/2} I_2)$

$$\begin{aligned} \hat{A} &= \bigoplus_{i=1}^n (\alpha_i J_1) \\ &= \bigoplus_{i=1}^n (\alpha_i^{1/2} I_2) \cdot \hat{P}J\hat{P}^T \cdot \bigoplus_{i=1}^n (\alpha_i^{1/2} I_2) \\ &= \alpha \hat{P} J \hat{P}^T \alpha \end{aligned}$$

und

$$L = \hat{L}'\alpha\hat{P}$$

gilt. Verwenden wir die Bezeichnung

$$(X^T X)_S = D_X^{-1}(X^T X)D_X^{-1} \quad , \quad D_X = \mathrm{diag}(\|X_{\cdot i}\|_2)$$

für nichtsinguläres X, so folgt

$$\begin{aligned} K_S &= (L^T L)_S \\ &= (\hat{P}^T \alpha \hat{L}'^T \hat{L}' \alpha \hat{P})_S \\ &= \hat{P}^T (\hat{L}'^T \hat{L}')_S \hat{P} \\ &= \hat{P}\hat{D}'^{-1}\hat{L}'^T\hat{L}'\hat{D}'^{-1}\hat{P} \quad , \quad \hat{D}' = \mathrm{diag}(\|\hat{L}'_{\cdot i}\|_2) \ . \end{aligned} \tag{4.19}$$

Zur Bestimmung einer Schranke von $\kappa(K_S)$ werden wir nun $\|K_S\|_2$ und $\|K_S^{-1}\|$ abschätzen. Die Schranke

$$\|K_S\|_2 \leq 2\,n \tag{4.20}$$

ist wegen $|K_{S_{ij}}| \leq 1$ trivial. Weiter folgt aus (4.19)

$$\|K_S^{-1}\|_2 = \|\hat{D}'\hat{L}'^{-1}\hat{L}'^{-T}\hat{D}'\|_2 \ . \tag{4.21}$$

Wegen (4.17) gilt

$$\hat{D}'_{2i-1}\,,\ \hat{D}'_{2i} \leq (1+2(n-i))^{1/2} \quad , \quad i = 1\ldots n \ . \tag{4.22}$$

Wir definieren nun

$$E_k = (1)_{1\leq i\leq 2k\,,\ 1\leq j\leq 2} \in I\!R^{2k\times 2} \quad , \quad E = E_1 \ . \tag{4.23}$$

Wenn man $\hat{L}'$ als Produkt von Matrizen (4.17) betrachtet, gilt

$$|\hat{L}'^{-1}| \leq \check{L} \ ,$$

wobei

$$\begin{aligned} \check{L} &= \prod_{i=1}^{n-1} \begin{bmatrix} I_{2(n-i-1)} & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & E_i & I_{2i} \end{bmatrix} \\ &= (\check{L}^{[i,j]})_{1\leq i,j\leq n} \end{aligned} \tag{4.24}$$

mittels

$$\check{L}^{[i,j]} = \begin{cases} 0_2 & , \ i<j \\ I_2 & , \ i=j \\ E(I+E)^{i-j-1} & , \ i>j \end{cases} \tag{4.25}$$

in 2×2-Blöcke partitioniert werden kann. Aus (4.21), (4.22) und (4.25) folgt

$$\begin{aligned} \|K_S^{-1}\|_2 &\leq \|\hat{D}'\check{L}\check{L}^T\hat{D}'\|_2 \\ &\leq \mathrm{Tr}\ \hat{D}'\check{L}\check{L}^T\hat{D}' \\ &= \sum_{i=1}^{n}(1+2\,(n-i))(2+\sum_{j=1}^{i-1}\mathrm{Tr}\ (E(I+E)^{2(i-j-1)}E)) \ . \end{aligned} \tag{4.26}$$

Dabei ist

$$\begin{aligned} &\sum_{j=1}^{i-1}\mathrm{Tr}\ (E(I+E)^{2(i-j-1)}E) \\ &= \sum_{j=1}^{i-1}\|E(I+E)^{i-j-1}\|_E^2 = \sum_{j=1}^{i-1}\|E\sum_{k=0}^{i-j-1}\begin{pmatrix} i-j-1 \\ k \end{pmatrix}E^k\|_E^2 \\ &= \sum_{j=1}^{i-1}\|E\sum_{k=0}^{i-j-1}\begin{pmatrix} i-j-1 \\ k \end{pmatrix}2^{k-1}E\|_E^2 \\ &= \|E\|_E^2\sum_{j=1}^{i-1}(\sum_{k=0}^{i-j-1}\begin{pmatrix} i-j-1 \\ k \end{pmatrix}2^k)^2 \\ &= 4\sum_{j=1}^{i-1}3^{2(i-j-1)} = 4\sum_{j=0}^{i-2}9^j = 4\,\frac{9^{i-1}-1}{8} \end{aligned}$$

und daher

$$\begin{aligned} \|K_S^{-1}\|_2 &\leq \sum_{i=1}^{n}(1+2\,(n-i))(2+\frac{9^{i-1}-1}{2}) \leq \frac{1+2\,n}{2}\sum_{i=0}^{n-1}9^i = \frac{1+2\,n}{2}\;\frac{9^n-1}{8} \\ &\leq \frac{1+2\,n}{16}\;9^n\;. \end{aligned} \tag{4.27}$$

Aus (4.20) und (4.27) folgt die Aussage des Satzes. Q.E.D.

Ein Vergleich von (4.15) mit der entsprechenden Schranke

$$2\,n\,(1+30\,n)\cdot 3.781^{4\,n}$$

für indefinite Zerlegungen der Ordnung $2n$ nach [29], (Th. 4.4.2) zeigt, daß in (4.15) sowohl die Mantisse günstiger ist, als auch der Exponent nur halb so groß ist. Eine ähnliche Aussage wurde auch in [13] gezeigt. Es ist bemerkenswert, daß trotz beliebiger Kondition von LJL^T diejenige von $(L^T L)_S$ beschränkt ist.

Wir wollen noch kurz auf diejenigen Matrizen eingehen, für die (4.15) (nahezu) erreicht wird. Verfolgt man den Beweis von Satz 4.3.1, so stellt man fest, daß nur die Abschätzungen (4.26) und (4.27) nicht erreichbar sind. Die dort entstehenden Fehler sind jedoch in ihrer Größenordnung vernachlässigbar. Aus (4.18) und (4.24) erkennt man ferner, daß die Schranke (4.15) für

$$S = \hat{L}'\Delta J_n \Delta \hat{L}'^T$$

mit beliebigen

$$\Delta = \bigoplus_{i=1}^{n}(\delta_i I_2) \quad , \quad \delta_i \geq \delta_{i+1}\;, \quad i = 1\ldots n-1$$

sowie

$$\hat{L}' = \prod_{i=1}^{n-1}\begin{bmatrix} I_{2(i-1)} & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & -E_{n-i} & I_{2(n-i)} \end{bmatrix}$$

(E_k gemäß (4.23)) nahezu erreicht wird. Diese Matrizen S haben die in 2×2-Submatrizen partitionierte Form

$$S = (S^{[i,j]})_{1\leq i,j\leq n}$$

mit

$$S^{[i,j]} = \begin{cases} \delta_j J_1 & , \; i = j \\ \delta_j \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} & , \; i > j \\ \delta_i \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} & , \; i < j \end{cases} \tag{4.28}$$

und $\delta_k \geq \delta_{k+1}, k = 1 \ldots n-1$. Ist S vom Typ (4.28), so kann man leicht zeigen, daß die Schranke (4.15) auch für Matrizen des Typs

$$\hat{J}S\hat{J} \quad , \quad \hat{J} = \bigoplus_{i=1}^{n} \hat{J}_i \ , \quad \hat{J}_i = I_2 \ \text{ oder } \ \hat{J}_i = J_1 \tag{4.29}$$

erreicht wird.

Wir kommen nun noch einmal zu den diagonaldominanten Matrizen gemäß Kapitel 1 zurück. Wir sind nun an Aussagen zur Zerlegung solcher Matrizen interessiert. Dazu betrachten wir beispielhaft die $\gamma-$d.d. Matrix

$$S = \begin{bmatrix} J_1 & -M^T \\ M & S' \end{bmatrix} = \begin{bmatrix} 0 & 1 & -\alpha & 0 & \alpha & 0 \\ -1 & 0 & 0 & \alpha & 0 & \alpha \\ \alpha & 0 & 0 & 1 & -2\alpha & 0 \\ 0 & -\alpha & -1 & 0 & 0 & -2\alpha \\ \alpha & 0 & 2\alpha & 0 & 0 & 1 \\ 0 & -\alpha & 0 & 2\alpha & -1 & 0 \end{bmatrix} \quad , \quad J_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} .$$

S ist $1-$d.d. zum Beispiel für $\alpha = 0.4$, nicht aber $\gamma - 2-$d.d.. Führen wir einen Schritt der Zerlegung (4.6), (4.7) durch, d.h.

$$S = \begin{bmatrix} I & 0 \\ MJ_1^{-1} & I \end{bmatrix} \begin{bmatrix} J_1 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} I & J_1^{-T}M^T \\ 0 & I \end{bmatrix} ,$$

so ist die reduzierte Matrix

$$S_1 = S' - MJ_1M^T = \begin{bmatrix} 0 & 1-\alpha^2 & 0 & -2\alpha-\alpha^2 \\ -1+\alpha^2 & 0 & -2\alpha+\alpha^2 & 0 \\ 0 & 2\alpha-\alpha^2 & 0 & 1-\alpha^2 \\ 2\alpha+\alpha^2 & 0 & -1+\alpha^2 & 0 \end{bmatrix}$$

nicht γ-d.d., denn für $\alpha = 0.4$ ist $2\alpha + \alpha^2 = 0.96 > 0.84 = 1 - \alpha^2$. Die bei der Zerlegung einer γ-d.d. Matrix entstehenden reduzierten Matrizen sind also i.A. nicht mehr γ-d.d.. Anders ist es mit den $\gamma - r-$d.d. Matrizen. Für diese haben wir als erstes folgenden

Satz 4.3.2 *Die Matrix*

$$S = -S^T = \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} = (S^{[i,j]})_{1 \leq i,j \leq m} \quad , \quad rm = 2n \ , \quad S^{[i,j]} \in I\!R^{r \times r}$$

mit $A = S^{[1,1]} = aQ$, $a \in I\!R \backslash \{0\}$, $Q^{-1} = Q^T = -Q$, *sei* $\gamma - r-$*d.d.* . *Die durch*

$$S = \begin{bmatrix} I & 0 \\ MA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} I & A^{-T}M^T \\ 0 & I \end{bmatrix}$$

definierte reduzierte Matrix T, *also* $T = S' + MA^{-1}M^T$, *ist dann gleichfalls* $\gamma - r-$*d.d.*.

Beweis:
(siehe auch [18], 3.4.10) T sei wie S partitioniert:

$$T = (T^{[i,j]})_{1\le i,j\le m-1} \quad , \quad T^{[i,j]} \in I\!R^{r\times r} \ .$$

Wegen $\|A^{-1}\|_2 = \|a^{-1}Q^T\|_2 = a^{-1}$ gilt für alle j, $1 \le j \le m-1$,

$$\begin{aligned}
\sum_{i=1,i\neq j}^{m-1} \|T^{[i,j]}\|_2 &= \sum_{i=1,i\neq j}^{m-1} \|S^{[i+1,j+1]} + S^{[i+1,1]}A^{-1}S^{[j+1,1]^T}\|_2 \\
&\le \sum_{i=1,i\neq j}^{m-1} \|S^{[i+1,j+1]}\|_2 + \frac{1}{a}\|S^{[j+1,1]}\|_2 \sum_{i=1,i\neq j}^{m-1} \|S^{[i+1,1]}\|_2 \\
&\le \gamma\,\|S^{[j+1,j+1]}\|_2 - \|S^{[1,j+1]}\|_2 \\
&\quad + \frac{1}{a}\,\|S^{[1,j+1]}\|_2\,(\underbrace{\gamma}_{<1}\,\underbrace{\|S^{[1,1]}\|_2}_{=a} - \|S^{[j+1,1]}\|_2) \\
&\le \gamma\,\|S^{[j+1,j+1]}\|_2 - \frac{1}{a}\,\|S^{[j+1,1]}\|_2^2 \ . \qquad (4.30)
\end{aligned}$$

Wegen

$$T^{[j,j]} = S^{[j+1,j+1]} + S^{[j+1,1]}A^{-1}S^{[j+1,1]^T}$$

und (4.30) folgt für alle j, $1 \le j \le m-1$,

$$\begin{aligned}
\gamma\,\|T^{[j,j]}\|_2 &\ge \gamma\,\|S^{[j+1,j+1]}\|_2 - \frac{\gamma}{a}\,\|S^{[j+1,1]}\|_2^2 \\
&\ge \gamma\,\|S^{[j+1,j+1]}\|_2 - \frac{1}{a}\|S^{[j+1,1]}\|_2^2 \\
&\ge \sum_{i=1,i\neq j}^{m-1} \|T^{[i,j]}\|_2 \ ,
\end{aligned}$$

womit der Satz bewiesen ist. Q.E.D.

Für skaliert diagonaldominante Matrizen können wir nur eine schwächere Aussage zeigen:

Satz 4.3.3 *Die Matrix*

$$S = -S^T = \begin{bmatrix} A & -M^T \\ M & S' \end{bmatrix} = (S^{[i,j]})_{1\le i,j\le m} \quad , \quad rm = 2n \ , \quad S^{[i,j]} \in I\!R^{r\times r}$$

mit $A = S^{[1,1]} = aQ$, $a \in I\!R\backslash\{0\}$, $Q^{-1} = Q^T = -Q$, *sei skaliert* $\gamma - r-d.d.$ *mit* $\gamma < 1/\sqrt{2}$. *Die durch*

$$S = \begin{bmatrix} I & 0 \\ MA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} I & A^{-T}M^T \\ 0 & I \end{bmatrix} \qquad (4.31)$$

definierte reduzierte Matrix T, *also* $T = S' + MA^{-1}M^T$, *ist dann skaliert* $\gamma' - r-d.d.$ *mit* $\gamma' = \gamma/\sqrt{1-\gamma^2}$.

Beweis:
Durch Einsetzen von

$$T^{[i,j]} = S^{[i+1,j+1]} + S^{[i+1,1]}A^{-1}S^{[j+1,1]^T}$$

werden wir

$$\sigma_j := \sum_{i=1,i\neq j}^{m-1} \|T^{[i,i]}\|_2^{-1/2}\|T^{[j,j]}\|_2^{-1/2}\|T^{[i,j]}\|_2 \leq \gamma'$$

für alle j, $1 \leq j \leq m-1$, zeigen. Für alle j gilt

$$\begin{aligned}
&\|S^{[j+1,j+1]} + S^{[j+1,1]}A^{-1}S^{[j+1,1]^T}\|_2 \\
&\geq \|S^{[j+1,j+1]}\|_2 - \underbrace{\frac{1}{a}\|S^{[j+1,1]}\|_2^2\|S^{[j+1,j+1]}\|_2^{-1}}_{=:\alpha_j\leq\gamma^2}\|S^{[j+1,j+1]}\|_2 \\
&= (1-\alpha_j)\|S^{[j+1,j+1]}\|_2 \ .
\end{aligned}$$

Daraus folgt

$$\begin{aligned}
\sigma_j &= \sum_{i=1,i\neq j}^{m-1} (\|S^{[i+1,i+1]} + S^{[i+1,1]}A^{-1}S^{[i+1,1]^T}\|_2^{-1/2} \\
&\qquad \cdot\|S^{[j+1,j+1]} + S^{[j+1,1]}A^{-1}S^{[j+1,1]^T}\|_2^{-1/2} \\
&\qquad \cdot\|S^{[i+1,j+1]} + S^{[i+1,1]}A^{-1}S^{[j+1,1]^T}\|_2) \\
&\leq \sum_{i=1,i\neq j}^{m-1} ((1-\alpha_i)^{-1/2}(1-\alpha_j)^{-1/2}\|S^{[i+1,i+1]}\|_2^{-1/2}\|S^{[j+1,j+1]}\|_2^{-1/2} \\
&\qquad \cdot\|S^{[i+1,j+1]} + S^{[i+1,1]}A^{-1}S^{[j+1,1]^T}\|_2) \\
&\leq (1-\alpha_j)^{-1/2}(1-\gamma^2)^{-1/2}(\sum_{i=1,i\neq j}^{m-1} \|S^{[i+1,i+1]}\|_2^{-1/2}\|S^{[j+1,j+1]}\|_2^{-1/2}\|S^{[i+1,j+1]}\|_2 \\
&\qquad +\sqrt{\alpha_j}\sum_{i=1,i\neq j}^{m-1} \frac{1}{\sqrt{a}}\|S^{[i+1,1]}\|_2\|S^{[i+1,i+1]}\|_2^{-1/2}) \\
&\leq (1-\alpha_j)^{-1/2}(1-\gamma^2)^{-1/2}(\gamma - \frac{1}{\sqrt{a}}\|S^{[j+1,j+1]}\|_2^{-1/2}\|S^{[j+1,1]}\|_2 \\
&\qquad \sqrt{\alpha_j}(\underbrace{\gamma - \frac{1}{\sqrt{a}}\|S^{[j+1,j+1]}\|_2^{-1/2}\|S^{[j+1,1]}\|_2)}_{=\sqrt{\alpha_j}}) \\
&\leq \frac{\gamma-\alpha_j}{\sqrt{1-\alpha_j}}\ \frac{1}{\sqrt{1-\gamma^2}} \leq \frac{\gamma}{\sqrt{1-\gamma^2}} = \gamma' \ .
\end{aligned}$$

Q.E.D.

Wir werden zwei Sätze über obere Schranken von $\kappa(K_S)$ bei der Zerlegung von skaliert $\gamma - 2-$d.d. S zeigen.

Satz 4.3.4 *Sei $S = -S^T \in I\!R^{2n\times 2n}$, $n \geq 2$, skaliert $\gamma - 2-$d.d. mit $\gamma < 1/\sqrt{n-1}$. S sei gemäß (4.2) mit vollständiger Pivotwahl (also nach dem in 4.1 geschilderten Algorithmus) zerlegt, d.h. $PSP^T = LJL^T$. Für*

$$K_S = D^{-1}L^T L D^{-1} \quad , \quad D = diag(\|L_{\cdot i}\|_2)$$

gilt dann

$$\kappa(K_S) = \|K_S\|_2 \|K_S^{-1}\|_2 \leq \frac{(1+\sqrt{n-1}\,\gamma)^4}{(1-\sqrt{n-1}\,\gamma)^2}$$

Beweis:
Sei $PSP^T = L' J_n L'^T$ die Zerlegung gemäß (4.4). Setzen wir $D' = \mathrm{diag}(L'_{ii})$ und $\bar{L}' = L'D'^{-1}$, so gilt mit (4.9), (4.10)

$$\begin{aligned} K_S &= D^{-1}L^T L D^{-1} = D^{-1}\hat{P}^T L'^T L' \hat{P} D^{-1} \\ &= D^{-1}\hat{P}^T D' \bar{L}'^T \bar{L}' D' \hat{P} D^{-1} \\ &= \bar{D}'^{-1}\hat{P}^T \bar{L}'^T \bar{L}' \hat{P} \bar{D}'^{-1} \quad , \quad \bar{D}'^{-1} = \hat{P}^T D' \hat{P} D^{-1} \ . \end{aligned}$$

Man kann also o.B.d.A. für S von vorneherein annehmen, die Zerlegung (4.4) führe zu einen L' mit $L'_{ii} = 1$, $i = 1 \ldots 2n$. Nehmen wir dafür die Zerlegung

$$PSP^T = L' J_n L'^T$$

gemäß (4.4) vor, so gilt

$$L' = \prod_{i=1}^{n-1} L'_j \quad , \quad L'_j = \begin{bmatrix} I_{2j-2} & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & X_j & I_{2n-2j} \end{bmatrix} \ .$$

Setzen wir

$$X_j = \begin{bmatrix} X_j^{[j+1]} \\ \vdots \\ X_j^{[n]} \end{bmatrix} \quad , \quad X_j^{[i]} \in I\!R^{2\times 2}$$

für $j = 1 \ldots n-1$, so haben wir

$$\begin{array}{lll} \text{für } j=1 & : & \sum_{i=2}^n \|X_1^{[i]}\|_2 \leq \gamma \\ \text{für } j=2 & : & \sum_{i=3}^n \|X_2^{[i]}\|_2 \leq \frac{\gamma}{\sqrt{1-\gamma^2}} < \frac{\gamma\sqrt{n-1}}{\sqrt{n-2}} < \frac{1}{\sqrt{n-2}} \\ \text{für } j=3 & : & \sum_{i=4}^n \|X_3^{[i]}\|_2 \leq \frac{\gamma\sqrt{n-1}}{\sqrt{n-3}} < \frac{1}{\sqrt{n-3}} \ , \end{array}$$

und allgemein

$$\sum_{i=j+1}^{n} \|X_j^{[i]}\|_2 \leq \frac{\gamma\sqrt{n-1}}{\sqrt{n-j}} \leq \sqrt{n-1}\,\gamma \quad , \quad j = 1 \ldots n-1 \ .$$

Verwenden wir die übliche Bezeichnung

$$(Z^T Z)_S = \Delta^{-1}(Z^T Z)\Delta^{-1} \quad , \quad \Delta = \mathrm{diag}(\|Z_{\cdot i}\|_2)$$

für skalierte positiv definite Matrizen, so ist

$$\begin{aligned} K_S &= (L^T L)_S = (\hat{P}^T L'^T L' \hat{P})_S = \hat{P}^T (L'^T L')_S \hat{P} \\ &= \hat{P}^T D'^{-1} L'^T L' D'^{-1} \hat{P} \quad , \quad D' = \mathrm{diag}(\|L'_{\cdot i}\|_2) \ , \end{aligned}$$

und daher

$$\kappa(K_S) \leq \|L'\|_2^2 \|L'^{-1}\|_2^2 \|D'\|_2^2 \|D'^{-1}\|_2^2 \ .$$

Wir werden nun obere Schranken für jeden der vorstehenden Faktoren bestimmen. Es gilt für $1 \leq j \leq n-1$

$$D'_{2j-1,2j-1} = \|L_{\cdot 2j-1}\|_2 \leq 1 + \sum_{i=j+1}^{n} \|X_j^{[i]}\|_2 \leq 1 + \sqrt{n-1}\,\gamma$$

und weil dieselbe Abschätzung für $D'_{2j,2j}$ gilt, ist

$$\|D'\|_2 \leq 1 + \sqrt{n-1}\,\gamma \ .$$

Trivialerweise ist

$$\|D'^{-1}\|_2 \leq 1$$

und weiter

$$\|L'\|_2 \leq |||L'|||_2 = \max_{1 \leq j \leq n-1} (1 + \sum_{i=j+1}^{n} \|X_j^{[i]}\|_2) \leq 1 + \sqrt{n-1}\,\gamma \ ,$$

sowie

$$\|L'^{-1}\|_2 \leq \frac{1}{1 - \|L' - I\|_2} \leq \frac{1}{1 - \max_{1 \leq j \leq n-1} \sum_{i=j+1}^{n} \|X_j^{[i]}\|_2} \leq \frac{1}{1 - \sqrt{n-1}\,\gamma} \ ,$$

also insgesamt

$$\kappa(K_S) \leq \frac{(1 + \sqrt{n-1}\,\gamma)^4}{(1 - \sqrt{n-1}\,\gamma)^2}$$

Q.E.D.

Die pessimistische Aussage des letzten Satzes ist Resultat der Pivotstrategie bei der Zerlegung. Orientierte sich die Pivotwahl an einer skalierten Matrix, so könnte man im letzten Satz $\sqrt{n-1}$ durch 1 ersetzen. Darüber hinaus können wir sogar zeigen

Satz 4.3.5 *Sei $S = -S^T \in \mathbb{R}^{2n\times 2n}$ nichtsingulär und skaliert $\gamma - 2-$d.d., d.h. es existiere ein diagonales, positiv definites $\tilde{D} = \oplus_{i=1}^n \tilde{d}_i I_2$, so daß $\tilde{D}^{-1} S \tilde{D}^{-1}$ $\gamma - 2-$d.d. ist. Dann gibt es für die Zerlegung (4.2) eine Pivotstrategie, die*

$$P\tilde{D}P^T = \bigoplus_{i=1}^n d'_i I_2 \quad , \quad d'_i \geq d'_{i+1} \ , \quad i = 1 \ldots n-1$$

zur Folge hat und für die gilt

$$\kappa(K_S) \leq \frac{(1+\mu\gamma)^4}{(1-\mu\gamma)^2} \quad , \quad \mu = \max_{1\leq j\leq n-1} \frac{d'_{j+1}}{d'_j} \ .$$

Beweis:
Nach Voraussetzung gibt es ein $\tilde{D}$, so daß

$$\tilde{S} = \tilde{D}^{-1} S \tilde{D}^{-1} \quad , \quad \tilde{D} = \bigoplus_{i=1}^n \tilde{d}_i I_2 \ , \quad \tilde{d}_i > 0$$

$\gamma - 2-$d.d. ist. $\tilde{S}$ werde gemäß (4.4) mit vollständiger Pivotierung in

$$P\tilde{S}P^T = L' J_n L'^T$$

zerlegt. Setzt man $\bar{L}' = L'\Delta^{-1}$, $\Delta = \mathrm{diag}(L'_{ii})$, so ist $P\tilde{S}P^T = \bar{L}'\Delta J_n \Delta \bar{L}'^T$ und es folgt

$$\begin{aligned} PSP^T &= P\tilde{D}P^T \bar{L}'\Delta J_n \Delta \bar{L}'^T P\tilde{D}P^T \\ &= D'\bar{L}'\Delta J_n \Delta \bar{L}'^T D' \quad , \quad D' = P\tilde{D}P^T \ . \end{aligned}$$

Die Pivotwahl erfolge so, daß

$$D' = \bigoplus_{i=1}^n d'_i I_2 \quad , \quad d'_i \geq d'_{i+1} \ , \quad i = 1 \ldots n-1$$

gilt. Mit $\hat{P}$ gemäß (4.9) ist $S = P^T D' \bar{L}' \Delta \hat{P} J \hat{P}^T \Delta \bar{L}'^T D' P$ und es folgt

$$K = \hat{P}^T \Delta \bar{L}'^T D' P P^T D' \bar{L}' \Delta \hat{P} = \hat{P}^T \Delta \bar{L}'^T D'^2 \bar{L}' \Delta \hat{P} \ . \tag{4.32}$$

Wegen

$$D'\bar{L}' = \bar{L}'' D'$$

mit

$$\bar{L}'' = (\bar{L}''^{[i,j]})_{1\leq i,j\leq n} \quad , \quad \bar{L}''^{[i,j]} = \begin{cases} I_2 & i = j \\ \frac{d'_i}{d'_j}\bar{L}'^{[i,j]} & i > j \\ 0_2 & i < j \end{cases}$$

folgt aus (4.32)

$$\begin{aligned} K_S &= (\hat{P}^T \Delta \bar{L}'^T D'^2 \bar{L}' \Delta \hat{P})_S = \hat{P}^T (\bar{L}'^T D'^2 \bar{L}')_S \hat{P} = \hat{P}^T (\bar{L}''^T \bar{L}'')_S \hat{P} \\ &= \hat{P}^T \bar{D}''^{-1} \bar{L}''^T \bar{L}'' \bar{D}''^{-1} \hat{P} \quad , \quad \bar{D}'' = \mathrm{diag}(\|\bar{L}''_{\cdot i}\|_2) \ , \end{aligned}$$

wobei wir wieder die Bezeichnung

$$(Z^T Z)_S = D_Z^{-1} (Z^T Z) D_Z^{-1} \quad , \quad D_Z = \mathrm{diag}(\|Z_{\cdot i}\|_2)$$

verwendet haben. Zur Bestimmung einer oberen Schranke von $\kappa(K_S)$ werden wir $\|K_S\|$ und $\|K_S^{-1}\|$ einzeln abschätzen. Es gilt für $1 \le j \le n$

$$\begin{aligned} \bar{D}''_{2j-1,2j-1} &= \|\bar{L}''_{2j-1}\|_2 \le 1 + \sum_{i=j+1}^{n} \frac{d'_i}{d'_j} \|\bar{L}'^{[i,j]}\|_2 \\ &\le 1 + \frac{d'_{j+1}}{d'_j} \sum_{i=j+1}^{n} \|\bar{L}'^{[i,j]}\|_2 \le 1 + \frac{d'_{j+1}}{d'_j} \gamma \end{aligned}$$

und analog

$$\bar{D}''_{2j} \le 1 + \frac{d'_{j+1}}{d'_j} \gamma \ ,$$

also

$$1 \le \|\bar{D}''\|_2 \le 1 + \mu\gamma \quad , \quad \mu = \max_{1 \le j \le n-1} \frac{d'_{j+1}}{d'_j} \ .$$

Weiterhin ist

$$\bar{L}'' = I + X \quad , \quad X = (X^{[i,j]})_{1 \le i,j \le n} = \begin{cases} 0_2 & i \le j \\ \frac{d'_i}{d'_j} \bar{L}'^{[i,j]} & i > j \end{cases}$$

und daher mit der Norm aus 1.0.12

$$\|\bar{L}''\|_2 \le 1 + \|X\|_2 \le 1 + |||X|||_2 \le 1 + \mu\gamma$$

sowie

$$\|\bar{L}''^{-1}\|_2 \le \frac{1}{1 - \|X\|} \le \frac{1}{1 - \mu\gamma} \ .$$

Zusammenfassend ist also

$$\kappa(K_S) \le \|\bar{D}''\|_2^2 \, \|\bar{D}''^{-1}\|_2^2 \, \|\bar{L}''\|_2^2 \, \|\bar{L}''^{-1}\|_2^2 \le \frac{(1+\mu\gamma)^4}{(1-\mu\gamma)^2} \quad , \quad \mu = \max_{1 \le j \le n-1} \frac{d'_{j+1}}{d'_j}$$

Q.E.D.

Für $\gamma - 2-$d.d. Matrizen konnten wir also mit Satz 4.3.4 eine obere Schranke von $\kappa(K_S)$ angeben, die ausschließlich von γ und der Ordnung abhängt. Mit Satz

4.3.5 konnten wir diese Aussage verschärfen. Etwas vergröbernd kann man sagen, daß die Schranke für die skalierte Kondition um so kleiner werden kann, je stärker S skaliert ist. Welche Pivotwahl zu einer minimalen skalierten Kondition führt, scheint noch ein offenes Problem zu sein.

Numerische Experimente (s. Abschnitt 6.2) zeigen, daß $\kappa(K_S)$ auch sonst meist klein ist (≤ 100). Die Größenordnung der Schranke (4.15) wird zumeist nur von Matrizen S des Typs (4.29) erreicht.

Kapitel 5

Der Gesamtfehler in den berechneten Eigenwerten

In diesem kurzen Kapitel fassen wir nun alle Ergebnisse der Fehleranalyse zusammen und bestimmen eine obere Schranke für den relativen Fehler der vom Verfahren in endlicher Arithmetik berechneten Eigenwerte einer nichtsingulären schiefsymmetrischen Matrix S, wenn zuerst die Zerlegung $S = LJL^T$ gemäß Abschnitt 4.1 vorgenommen wird, und anschließend die Eigenwerte des Paares $(L^T L, J)$ mit dem Verfahren 2.2.4 berechnet werden.

Auch hier bedeutet die Formulierung „in erster Ordnung von ε“ wieder, daß die gemeinte Ungleichung bis auf Multiplikation mit $1 + c\varepsilon$ mit einem mäßigen $c \geq 0$ gilt. Die Konvergenz des Verfahrens 2.2.4 auch in endlicher Arithmetik wird vorausgesetzt.

Satz 5.0.6 *Sei $S \in \mathbb{R}^{2n\times 2n}$ nichtsingulär und schiefsymmetrisch mit den Eigenwerten $\pm i\lambda_j$, $j = 1 \ldots n$, wobei*

$$0 < \lambda_1 \leq \ldots \leq \lambda_n \ .$$

Sei $\tilde{L} = L^{(1)}$ der mit Algorithmus 4.1 in endlicher Genauigkeit ε berechnete Faktor der Zerlegung von S und sei

$$\hat{L} = D^{-1}\tilde{L} \quad , \quad D = diag\,(\|\tilde{L}_{i\cdot}\|_2) \ .$$

Seien $\pm i\lambda'_j$, $1 \leq j \leq n$, mit

$$0 < \lambda'_1 \leq \ldots \leq \lambda'_n$$

die mittels Algorithmus 2.2.4 in endlicher Genauigkeit berechneten Eigenwerte des Paares $(\tilde{L}^T\tilde{L}, J)$, so gilt in erster Ordnung von ε

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \ \leq \ \varrho = \varrho_Z + \varrho_J \quad , \quad 1 \leq j \leq n \ ,$$

wobei

$$\varrho_Z \leq \frac{12n^2\varepsilon}{\lambda_{\min}(D^{-1}\|\tilde{L}J\tilde{L}^T\|D^{-1})} \tag{5.1}$$

$$\begin{aligned}\varrho_J \leq\ & (\frac{(2\,n+14)\sqrt{2n}}{\sigma_{\min}(L_S^{(1)})} + \sum_{z=1}^{Z}\frac{(\sqrt{2}(2\,n+20)+4)\sqrt{2n}}{\sigma_{\min}(L_S^{<z>})} \\ & + \sum_{r=1}^{R}\frac{160+156\sqrt{2}}{\sigma_{\min}(L_S^{(r)})} + \sum_{z=1}^{Z}\sum_{i=1}^{num(z)}\frac{(16m_i+15)\sqrt{m_i}}{\sigma_{\min}(L_S^{<z>})} \\ & + \sum_{z=1}^{Z}\frac{(2n+16)\sqrt{2n}}{\sigma_{\min}(L_S^{<z>})} + \sqrt{n/2} + (2n)^2)\,\varepsilon + 2n\,TOL\ . \end{aligned} \tag{5.2}$$

Dabei sei $\varrho_Z < 1$ und $\|\cdot\|$ ist aus Satz 1.0.1. Weiterhin sei m_i die Ordnung des nichttrivialen Teils einer Transformation in einem modifizierten Zyklus und

$$\varepsilon \leq \min\{\frac{1}{20\,m_i\,n}, \frac{1}{500\,m_i}, \frac{\sigma_{\min}(L_S^{(r)})}{100}, \frac{\sigma_{\min}(L_S^{<z>})}{4.5\,n\sqrt{n}}\}$$

für alle r, z und alle m_i. Außerdem sei R die Anzahl der Diagonalisierungen von 4×4-Submatrizen und Z die Zahl der Zyklen. Schließlich sei

$$L_S^{(r)} = L^{(r)}\,diag(\|L_{\cdot i}^{(r)}\|_2^{-1})$$

die skalierte Matrix vor der r-ten Diagonalisierung einer 4×4-Submatrix und

$$L_S^{<z>} = L^{<z>}\,diag(\|L_{\cdot i}^{<z>}\|_2^{-1})$$

sei die skalierte Matrix vor dem z-ten Zyklus.

Beweis:
Wir folgen der Beweistechnik in [29]. Wir können o.B.d.A annehmen, daß die in Satz 4.2.1 erwähnte Permutationsmatrix die Einheitsmatrix ist. Dann gilt nach Satz 4.2.1 für den berechneten Faktor $\tilde{L}$ der Zerlegung

$$\tilde{L}J\tilde{L}^T = S + \delta S \quad , \quad |\delta S| \leq 3n(|S| + |\tilde{L}|\,|J|\,|\tilde{L}|^T)\,\varepsilon\ .$$

Wegen

$$|S| \leq |LJL^T| + |\delta S| \leq |L|\,|J|\,|L|^T + |\delta S|$$

folgt in erster Ordnung von ε

$$|\delta S| \leq 6n|\tilde{L}|\,|J|\,|\tilde{L}|^T\varepsilon\ .$$

Für beliebiges $0 \neq x \in \mathbb{R}^{2n\times 2n}$ gilt daher

$$\begin{aligned}
|x^*\delta S x| &\leq |x|^T\,|\delta S|\,|x| \\
&\leq 6n\,|x|^T\,|\tilde{L}|\,|J|\,|\tilde{L}|^T\,|x|\,\varepsilon \\
&\leq 6n\,|x|^T\,D|\hat{L}|\,|J|\,|\hat{L}|^T D\,|x|\,\varepsilon \\
&\leq 6n\,|x|^T DED|x|\,\varepsilon \quad \text{mit} \quad E = (1)_{ij} \\
&\leq 12n^2\,|x|^T D^2|x|\,\varepsilon \\
&\leq 12n^2\,x^* D^2 x \cdot \frac{x^*|\tilde{L}J\tilde{L}^T|x}{x^*|\tilde{L}J\tilde{L}^T|x}\,\varepsilon \\
&\leq \frac{12n^2\,\varepsilon}{\lambda_{\min}(D^{-1}|\tilde{L}J\tilde{L}^T|D^{-1})}\,x^*|\tilde{L}J\tilde{L}^T|x\ .
\end{aligned}$$

Wegen Satz 1.0.1 folgt daraus (5.1), wenn nur Fehler in erster Ordnung von ε betrachtet werden. Die Aussage (5.2) folgt aus Lemma 3.2.2 mit Hilfe der in Abschnitt 3.2 ermittelten Zwischenergebnisse. Die ersten vier Summanden stammen aus der Anwendung der Sätze 3.2.11, 3.2.14, 3.2.20, 3.2.28 und 3.2.21. Der Term $\sqrt{n/2}\,\varepsilon$ hat seine Ursache im relativen Rückwärtsfehler von $\varepsilon/2$ des Teiles **4.** von Verfahren 2.2.4 und Anwendung von Satz 1.0.5, Lemma 3.2.1. Weil analog zu Lemma 3.2.23 aus der Bedingung für *TOL* lediglich

$$|\sum_{k=1}^{2n} L_{ki}L_{ki}| \leq TOL + 2n\,\varepsilon \quad , \quad 1 \leq i,j \leq 2\,n$$

folgt, entstehen die Terme $(2n)^2$ und $2n{\cdot}TOL$ aus der Anwendung von Satz 1.0.6 (1.9).

Die angegebenen Voraussetzungen für ε ergeben sich aus der Anwendung vorgenannter Sätze. Q.E.D.

Die Abschätzung (3.47) könnte, wie bereits in den vorausgehenden Unterabschnitten angedeutet, verbessert werden, wenn z.B.

- verschiedene Einzelfälle des Verfahrens 2.2.4 besonders behandelt würden,
- die Fälle *rotF=FALSE* oder *rotC=FALSE* des Verfahrens 2.2.4 besonders berücksichtigt würden,
- die tatsächlichen $\sigma_{\min}(L_S)$ berücksichtigt würden und nicht die pessimistischen Schranken aus Kapitel 1 oder
- berücksichtigt würde, wenn einzelne Transformationen die Identität sind.

Diese Verbesserungen könnten in ein Programm, mit dem 2.2.4 implementiert wird, aufgenommen werden. Die hier dargestellte Fehleranalyse würden diese

Fälle jedoch nur unnötig verkomplizieren.

Über die Aussage von Satz 5.0.6 hinaus erwarten wir sogar einen von n unabhängigen Fehler, weil ein Faktor n in einem Fehlerterm meist aus einer Vektoroperation entsteht. Bei Vektoroperationen kann man erwarten, daß sich die Fehler gegenseitig nahezu aufheben. Genauer erwarten wir

$$\frac{|\lambda_j - \lambda_j'|}{\lambda_j} \leq \varrho_E \quad , \quad \varrho_E = \left(\frac{1}{\lambda_{\min}(D^{-1}|\tilde{L}J\tilde{L}^T|D^{-1})} + \frac{1}{\sigma_{\min}(\tilde{L}_S)}\right) \varepsilon \ , \quad j = 1 \ldots n \ . \tag{5.3}$$

Numerische Experimente (siehe Abschnitt 6.2) werden diese Erwartung begründen.

Den erwarteten Fehler (5.3) wollen wir möglichst mit Matrizen ausdrücken, die im Laufe des Verfahrens anfallen. Dazu ziehen wir nochmals die unitären Matrizen

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} I & iI \\ iI & I \end{bmatrix} \quad , \quad J' = I \oplus (-I) \ , \quad i^2 = -1$$

heran, für die $U^* J U = iJ'$ gilt. Weiterhin sei $\tilde{L} = L^{(1)}$ der in endlicher Genauigkeit berechnete Faktor der Zerlegung von S. Dieser Faktor wird im iterativen Teil des Verfahrens mit einer Folge symplektischer Matrizen von rechts transformiert. Seien $V = L^{(R)}$ und E die Matrizen, die zum Abschluß des iterativen Verfahrensteils entstanden sind. Wir wissen bereits, daß E bis auf einen kleinen relativen Fehler die Diagonalmatrix der positiven Imaginärteile der Eigenwerte von $\tilde{L}J\tilde{L}^T$ ist. Weiter nehmen wir an, daß V bis auf einen kleinen Fehler die Eigenvektormatrix ist, also

$$V^T \tilde{L} J \tilde{L}^T V \approx JE \ .$$

Dann gilt

$$\begin{aligned} |\tilde{L}J\tilde{L}^T| &\approx |VJEV^T| = |iVE^{1/2}U^*J'UE^{1/2}V^T| \\ &= |iVU^*E^{1/2}J'E^{1/2}UV^T| = VU^*EUV^T \\ &= VEV^T \ . \end{aligned}$$

Berechnet man noch $D = \mathrm{diag}(\|\tilde{L}_{i\cdot}\|_2)$, so läßt sich

$$M = D^{-1}|\tilde{L}J\tilde{L}^T|D^{-1} \tag{5.4}$$

mit Hilfe von Matrizen ausdrücken, die im Laufe des Verfahrens gewonnen werden.

Man kann M als fast optimal skaliert annehmen, d.h.

$$\max(\mathrm{diag}(M))/\min(\mathrm{diag}(M)) \quad \text{ist moderat.}$$

Eine Begründung werden wir in Abschnitt 6.2 geben. Man könnte also den ersten Summanden in (5.3) durch $1/\lambda_{\min}(|\tilde{L}J\tilde{L}^T|_S) \approx 1/\lambda_{\min}(|S|_S)$ ersetzen, wenn man noch $S \approx \tilde{L}J\tilde{L}^T$ annimmt.

Kapitel 6

Numerische Experimente

6.1 Alternative Verfahren

In diesem Abschnitt werden wir kurz auf existierende Verfahren zur Lösung des Eigenwertproblems schiefsymmetrischer Matrizen eingehen. In Abschnitt 6.2 werden wir diese dann experimentell mit dem in vorliegender Arbeit eingeführten impliziten Verfahren vergleichen.

6.1.1 Ein QR-Verfahren

Bereits in [34] wurde ein QR-Verfahren zur Lösung des schiefsymmetrischen Eigenwertproblems vorgestellt. Es wurde unter Beibehaltung der grundsätzlichen Schritte Tridiagonalisierung, Lösung des Eigenwertproblems einer Bidiagonalmatrix, Bildung der Eigenvektormatrix weitgehend modifiziert. Folgende Variation des QR-Verfahrens wurde implementiert.

1. Die schiefsymmetrische Matrix S wird tridiagonalisiert, d.h. es wird eine orthogonale Matrix U bestimmt, für die

   $$U^T S U = T$$

   mit tridiagonalem T gilt. Dies geschieht mit einem Givens-Verfahren unter Verwendung schneller Rotationen. Die Matrix U wird dabei nicht explizit berechnet, sondern es werden Informationen zur Berechnung von U im selben Feld gespeichert, das auch S enthält (siehe [18], 5.1.8 ff.).

2. Das tridiagonale T wird mit $\hat{P}$ gemäß (4.9) permutiert. Es gilt dann

   $$\hat{P}^T T \hat{P} = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}$$

   und A ist eine obere Bidiagonalmatrix.

3. Die Singulärwerte von A werden bestimmt, d.h. es werden orthogonale Matrizen P, Q bestimmt, für die $Q^T AP$ diagonal und positiv ist. Die Singulärwerte von A sind die positiven Imaginärteile der Eigenwerte von S. Zur Singulärwertzerlegung wird die in [12] vorgestellte und in LAPACK [2] implementierte Routine `DBDSQR` verwendet.

4. Zur Bestimmung der transponierten Matrix der Eigenvektoren von S wird abschließend noch das Produkt

$$V^T = \begin{bmatrix} P^T & 0 \\ 0 & Q^T \end{bmatrix} \hat{P}^T U^T$$

gebildet.

6.1.2 Das Verfahren von Paardekooper

Aus der Literatur ist auch ein Jacobi-Verfahren für schiefsymmetrische Matrizen bekannt. Es wurde von Paardekooper [25] vorgeschlagen. Die Grundidee verwendet die Tatsache, daß man schiefsymmetrische Matrizen der Ordnung 4 mit 4 Ebenenrotationen auf Murnaghansche Form transformieren kann. Das Paardekoopersche Verfahren besteht nun darin, zyklisch Submatrizen der Ordnung 4 auf Murnaghansche Form zu transformieren. Nach einigen Zyklen hat auch die bearbeitete Matrix Murnaghansche Form und die Eigenwerte können abgelesen werden.

6.2 Numerische Resultate

In diesem Abschnitt stellen wir mit Hilfe numerischer Resultate Vergleiche verschiedener Verfahren zur Lösung des schiefsymmetrischen Eigenwertproblems an. Dabei gehen wir folgendermaßen vor. Zuerst wird die technische Umgebung der Experimente dargestellt. Anschließend wird beschrieben, welche Programme implementiert wurden, und es werden einige Anmerkungen zu den Verfahren, der konkreten Kompilation und zu möglichen Optimierungen gemacht. In zwei eigenen Unterbschnitten stellen wir dann die Ergebnisse hinsichtlich der erzielten Genauigkeit bzw. des Konvergenzverhaltens dar.

Die Experimente wurden auf einem Rechner IBM RS/6000 Modell 53H unter AIX 3.1 durchgeführt. Zur Erzeugung der Testmatrizen und zur Testanordnung wurde PRO-MATLAB 3.5 verwendet. Die Verfahren zur Lösung der Eigenwertprobleme wurden in FORTRAN implementiert. Die Maschinengenauigkeiten beim verwendeten XL-FORTRAN-2.0-Compiler betragen

2^{-23}	$\approx$	$1.19 \cdot 10^{-7}$	bei einfacher Genauigkeit (`single`)
2^{-52}	$\approx$	$2.22 \cdot 10^{-16}$	bei doppelter Genauigkeit (`double`)
2^{-104}	$\approx$	$4.93 \cdot 10^{-32}$	bei vierfacher Genauigkeit (`real*16`) .

PRO-MATLAB verwendet gleichfalls doppelte Genauigkeit. Variablenunter- bzw. -überlauf wird etwa bei $10^{\pm 38}$ bei einfacher bzw. $10^{\pm 308}$ bei doppelter und vierfacher Genauigkeit erreicht.

Folgende Verfahren wurden implementiert:

`yssqr`	QR-Verfahren, wie in 6.1.1 geschildert.
`yprd`	Verfahren von Paardekooper (s. 6.1.2). Das in [25] angegebene ALGOL-Programm arbeitet fehlerhaft und wurde modifiziert.
`xjssif`	Jacobi-Verfahren, wie es in vorliegender Arbeit vorgestellt wurde mit Zerlegung der schiefsymmetrischen Matrix mit vollständiger Pivotierung gemäß Kapitel 4 und anschließender Anwendung der Verfahrens 2.2.4 auf den Faktor.

Programme, deren vorstehende Namen den Buchstaben `y` enthalten, wurden in einfacher (`sssqr` bzw. `sprd`) und doppelter (`dssqr` bzw. `dprd`) Genauigkeit implementiert. Programme, die den Buchstaben `x` enthalten, wurden in einfacher (`sjssif`), doppelter (`djssif`) und vierfacher (`qjssif`) Genauigkeit implementiert. Außerdem wurde noch eine Version (`dsjssif`) des Jacobi-Verfahrens mit Zerlegung in doppelter und iterativem Teil in einfacher Genauigkeit implementiert.

Zur Kompilation sei noch folgendes angemerkt.

- Alle FORTRAN-Programme wurden mit der Option `-O` kompiliert, die der Optimierung dient.

- Bei den Programmen in einfacher Genauigkeit wurde außerdem die Kompilieroption `-qrndsngl`, verwendet, die sicherstellt, daß auch Zwischenergebnisse auf einfache Genauigkeit gerundet werden und somit ein fairer Vergleich zwischen den verschieden genauen Programmen möglich ist.

- Beim verwendeten Compiler unterliegt die Verwendung vierfacher Genauigkeit (`real*16`) einigen Einschränkungen. So kann sie nur für die Grundrechenarten verwendet werden. Die Belegung von Konstanten mit Werten vierfacher Genauigkeit kann nur durch Einlesen der Werte aus einer Datei geschehen. Korrektes Rechnen in vierfacher Genauigkeit konnte anhand von Vergleichen vierfach genauer Programme (Paardekooper-Verfahren und implizites Jacobi-Verfahren) sichergestellt werden.

Zu den Implementationen sei noch folgendes vermerkt:

- Die Bestimmung der Clustergröße in modifizierten Zyklen wurde etwas vereinfacht, indem die rechte Seite der dritten Bedingung von Seite 57 zu Null gesetzt wurde. Damit erreicht man, daß das Hadamardsche Maß bei Anwendung der Modifikation wächst, ohne jedoch ein Mindestwachstum vorzugeben. Dies entspricht $\sigma_{\max} = 0$ im Verfahren 2.2.4, was in der Praxis zu guten Ergebnisses führte.

- Die Programme `xjssif` wurden entgegen Verfahren 2.2.4 bereits dann abgebrochen, wenn $n(n-1)/2$ aufeinanderfolgende Submatrizen der Ordnung 4 aufgrund kleiner relativer Außerdiagonalelemente nicht mehr diagonalisiert wurden. Operationen innerhalb der modifizierten Zyklen wurden dabei entsprechend berücksichtigt. Auf diese Weise kann Rechenzeit eingespart werden.

- Die Rechenzeiten wurden mit der eingebauten Funktion `mclock()` gemessen.

- Liest man in Programmen vierfacher Genauigkeit die verwendeten Konstanten aus einer Datei ein, so werden diese auch als vierfach genaue Werte gespeichert.

- Die BLAS1-Routinen [24] wurden wegen Verwendung vierfach genauer Programme modifiziert.

Schließlich könnte man noch folgende Laufzeitoptimierungen vornehmen:

- Neben der als Vektor gespeicherten Diagonalmatrix E könnten auch die Normen der Spalten von L in einem eigenen Vektor gespeichert werden [32]. Damit könnte man deren häufige Neuberechnung einsparen. Die Fehleranalyse wäre dann allerdings nicht mehr gültig, weil die Vektorelemente und die Spaltennormen von L nicht immer bis auf einen kleinen relativen Fehler übereinstimmen.

- In FORTRAN werden die Größen der verwendeten Felder bereits zur Kompilationszeit festgelegt. Die Ordnung der Felder, in denen die Matrizen gespeichert werden, wurde in allen Programmen auf 400 gesetzt. Die Verarbeitungsgeschwindigkeit könnte sicherlich gesteigert werden, wenn die Ordnung der Felder diejenige der dort gespeicherten Matrizen nur wenig überschreitet.

- Wird die im iterativen Teil von `xjssif` bearbeitete Matrix

$$L = [L' \; L''] \text{ in der Form } \begin{bmatrix} L' \\ L'' \end{bmatrix}$$

  abgespeichert, so könnte aufgrund der FORTRAN-typischen Speicherstruktur sicherlich noch eine Geschwindigkeitssteigerung erzielt werden.

- Auf parallele Pivotstrategien bei den Jacobi-Verfahren wurde verzichtet. Ihre Verwendung könnte aufgrund der Ergebnisse in Abschnitt 3.1 zumeist zu einem kleineren maximalen relativen Fehler in den berechneten Eigenwerten führen. In der Praxis spielt die Pivotstrategie jedoch hinsichtlich der Genauigkeit der erzielten Ergebnisse keine Rolle. Wegen besserer Cache-Nutzung könnten parallele Strategien höhere Geschwindigkeiten der Programme erbringen.

- Bei der Implementation der modifizierten Zyklen von Algorithmus 2.2.4 könnten BLAS3-Routinen [15],[14],[24] verwendet werden, was zu höherer Geschwindigkeit auf Kosten von Speicherplatz und evtl. Genauigkeit führen dürfte. Darauf haben wir jedoch verzichtet.

- Die in den nachfolgenden Unterabschnitten aufgeführten Resultate erfordern eine Reihe von Messungen während des Programmlaufes. Verzichtet man auf diese Messungen, so wird sicherlich effizienterer Code entstehen.

In den beiden nachfolgenden Unterabschnitten präsentieren wir die numerischen Resultate hinsichtlich der erzielten relativen Genauigkeit bzw. des Konvergenzverhaltens des impliziten Verfahrens.

6.2.1 Zur relativen Genauigkeit

Wir widmen uns hier zuerst den erzielten Genauigkeiten der numerischen Verfahren.

In jedem der nachfolgenden Unterabschnitte stellen wir jeweils die bei einem bestimmten Matrizentyp erzielten Resultate dar. Jeder Unterabschnitt ist folgendermaßen gegliedert. Zuerst wird beschrieben, welcher Typ von Matrizen betrachtet wird und ggf. ein Verfahren zur ihrer Erzeugung angegeben. Dann werden die Rechenzeiten aller Algorithmen und die Anzahl der Operationen der Jacobi-Verfahren angegeben. Die Rechenzeiten werden nur in den ersten beiden der nachfolgenden Unterabschnitten angegeben. Danach werden die maximalen relativen Fehler aller Verfahren gegenüber dem vierfach genauen Programm `qjssif` angegeben und ausgewertet. Eine Betrachtung verschiedener „Konditionszahlen“ schließt sich an. Sie liefern eine Begründung der zuvor geschilderten relativen Fehler. Schließlich werden noch die erwarteten relativen Fehler den tatsächlichen relativen Fehlern gegenübergestellt.

Skaliert blockdiagonaldominante Matrizen

In diesem Unterabschnitt betrachten wir solche Matrizen, die Satz 1.0.2 genügen (siehe auch [33]). Matrizen dieser Art der Ordnung $2n$ wurden auf folgende Weise bestimmt:

1. Bestimmung eines $\eta \in \{2, 5, 8, 15, 17\}$. Generierung eines „zufälligen“ natürlichen b, $2 \leq b \leq n/4$, der Anzahl von Blocks.

2. Generierung von $\delta_i = \exp(\eta(2r_i - 1))$, $i = 1 \ldots b$, mit „zufälligen“ $r_i \in [0, 1]$. Generierung „zufälliger“ natürlicher Zahlen n_i, $i = 1 \ldots b$, mit $\sum_{i=1}^{b} n_i = n$. Bildung von Diagonalmatrizen $\Delta_i = \delta_i I_{2n_i}$ und $\Delta = \oplus_{i=1}^{b} \Delta_i$.

3. Generierung orthogonaler, schiefsymmetrischer Matrizen Q_i, $i = 1 \ldots b$, der jeweiligen Ordnung $2n_i$ und Bestimmung von $Q = \oplus_{i=1}^{b} Q_i$.

4. Generierung einer schiefsymmetrischen Störung E der Ordnung $2n$ und von der Norm $\|E\|_2 = 1/2$.

5. Bildung der Testmatrix $S = P^T \Delta (Q + E) \Delta P$, wobei P eine „zufällige“ Permutationsmatrix ist.

Für jedes $\eta \in \{2, 5, 8, 15, 17\}$ und $n = 10$ wurden 30 Testmatrizen, für $n = 25$, $n = 50$ wurden 20 Matrizen, und für $n = 100$ wurden 10 Testmatrizen generiert, insgesamt also 400 Matrizen.

Zuerst stellen wir die erzielten Rechenzeiten der doppelt genauen Programme tabellarisch dar. Bei den angegebenen Zeiten sind Eingabe- und Ausgabeoperationen nicht berücksichtigt. Zur Darstellung von T nichtnegativen Testergebnissen x_t, $t = 1 \ldots T$, verwenden wir das arithmetische Mittel

$$\mu_a = \frac{1}{T} \sum_{t=1}^{T} x_t$$

und die arithmetische Streuung

$$\sigma_a = (\frac{1}{T} \sum_{t=1}^{T} x_t^2 - (\frac{1}{T} \sum_{t=1}^{T} x_t)^2)^{1/2} \ .$$

Folgende Rechenzeiten wurden erzielt:

Rechenzeiten in 1/100 sec. bei skaliert blockdiagonaldominanten Matrizen					
	Verf.	μ_a	σ_a	min	max
n=10	`dssqr`	3	0.5	2	5
150 Tests	`dprd`	20	1.2	17	23
	`djssif`	6	1.2	3	10
n=25	`dssqr`	9	1.2	7	16
100 Tests	`dprd`	316	6.9	300	333
	`djssif`	49	10.2	33	80
n=50	`dssqr`	44	3.8	37	56
100 Tests	`dprd`	2567	326	290	2714
	`djssif`	313	84	125	541
n=100	`dssqr`	289	16	267	330
50 Tests	`dprd`	18050	9142	2574	23965
	`djssif`	2122	457	1321	3246

Die Rechenzeiten aller Verfahren fallen etwas mit wachsendem η und eine weitere Auswertung für $n = 100$ ergab, daß der Quotient der jeweiligen Rechenzeiten von `djssif` und `dssqr` zwischen etwa 7.7 ($\eta = 2$) und 6.4 ($\eta = 17$) lag. Das QR-Verfahren war überdies recht unempfindlich gegenüber mehrfachen Eigenwerten.

In nachfolgender Tabelle wird die Anzahl der Operationen bei den Jacobi-Verfahren zusammengestellt. Beim Paardekooper-Verfahren wird neben der Anzahl der Zyklen auch die Zahl der Rotationen angegeben. Bei dem in dieser Arbeit vorgestellten impliziten Verfahren wird neben der Zahl der Zyklen die Anzahl der Diagonalisierungen von 4×4-Submatrizen angegeben, wobei auch Transformationen im Rahmen der modifizierten Zyklen entsprechend berücksichtigt werden. Auch hier werden die Ergebnisse für alle η zusammengefaßt.

		Zahl der Zyklen				Zahl der Rotationen bzw. 4×4-Diagonalisierungen			
	Verf.	μ_a	σ_a	min	max	μ_a	σ_a	min	max
n=10	`dprd`	5	1.0	3	7	812	163	482	1202
	`djssif`	7	1.1	5	11	205	51	92	425
n=25	`dprd`	6	0.9	4	9	7131	994	4738	9766
	`djssif`	8	1.3	6	13	1455	361	842	2586
n=50	`dprd`	8	1.0	4	11	37573	4690	16822	49650
	`djssif`	10	1.7	6	14	6613	2169	1931	12806
n=100	`dprd`	10	0.9	8	11	180149	14794	155146	204978
	`djssif`	11	1.7	8	16	27587	7227	14368	44814

Die recht große Streubreite der Rechenzeiten bei `djssif` spiegelt sich hier in einer ebenso großen Streubreite der Zahl der Zyklen und Diagonalisierungen wider. Die asymptotisch quadratische Konvergenz trat also häufig erst recht spät ein. Resultate zur asymptotisch quadratischen Konvergenz werden weiter unten zusammengefaßt.

Als nächstes betrachten wir die maximalen relativen Fehler der berechneten Eigenwerte aller Verfahren gegenüber den von `qjssif` in vierfacher Genauigkeit berechneten Werten. Sind also λ_i, $i = 1 \dots n$, die positiven Imaginärteile der von `qjssif` berechneten Eigenwerte, und λ_i' die auf dieselbe Weise geordneten und von einem anderen Verfahren berechneten Eigenwerte, so bestimmen wir für jedes Verfahren und jede Testmatrix

$$\varrho = \max_{1 \le i \le n} \frac{|\lambda_i - \lambda_i'|}{\lambda_i}$$

und stellen diese maximalen relativen Fehler in nachfolgender Tabelle dar. Die wesentliche Aussagekraft liegt hier bei den Exponenten den jeweiligen Werte. Zur Darstellung von T positiven Testergebnissen x_t, $t = 1 \dots T$, verwenden wir hier daher das geometrische Mittel

$$\mu_g = (\prod_{t=1}^{T} x_t)^{1/T} = \exp(\frac{1}{T} \sum_{t=1}^{T} \log(x_t))$$

und die geometrische Streuung

$$\sigma_g = \exp(\ (\frac{1}{T} \sum_{t=1}^{T} (\log x_t)^2 - (\frac{1}{T} \sum_{t=1}^{T} \log x_t)^2)^{1/2}\ )\ .$$

Weil wir erwarten, daß die Ordnung der Matrix in der Praxis keine Rolle spielt (was man auch der Tabelle entnehmen kann), fassen wir die Tabelleneinträge für alle oben angegebenen n zusammen. Angaben über die Kondition $\kappa(S) = \|S\|_2 \|S^{-1}\|_2$ der Testmatrizen wurden gleichfalls in die Tabelle aufgenommen.

Maximale relative Fehler gegenüber `qjssif`					
	Verf.	μ_g	σ_g	min	max
$\eta = 2$	`sssqr`	6.7e−07	1.8e+00	2.1e−07	4.0e−06
80 Tests	`dssqr`	1.5e−15	1.6e+00	4.0e−16	5.5e−15
	`sprd`	1.0e−06	2.2e+00	2.1e−07	6.4e−06
	`dprd`	1.8e−15	1.8e+00	4.3e−16	7.7e−15

	sjssif	4.3e−06	2.1e+00	1.3e−06	2.6e−05
	djssif	1.3e−14	2.9e+00	2.2e−15	4.6e−13
	dsjssif	2.3e−06	1.6e+00	9.7e−07	6.3e−06
	$\kappa(S)$	3.3e+02	5.3e+00	1.6e+00	3.0e+03
$\eta = 5$	sssqr	1.9e−06	7.4e+00	2.5e−07	1.3e−02
80 Tests	dssqr	4.5e−15	7.9e+00	7.0e−16	2.2e−11
	sprd	1.1e−06	2.5e+00	2.5e−07	2.1e−05
	dprd	2.1e−15	2.1e+00	6.3e−16	1.4e−14
	sjssif	3.6e−06	1.8e+00	1.0e−06	2.2e−05
	djssif	9.6e−15	2.2e+00	2.8e−15	7.0e−14
	dsjssif	2.2e−06	1.7e+00	7.4e−07	7.5e−06
	$\kappa(S)$	7.1e+05	8.5e+02	1.0e+00	5.0e+08
$\eta = 8$	sssqr	8.4e−06	5.2e+01	3.4e−07	9.4e−01
80 Tests	dssqr	1.7e−14	4.7e+01	5.2e−16	2.8e−09
	sprd	4.8e−06	6.4e+01	1.7e−07	1.0e+05
	dprd	3.7e−15	4.6e+00	6.1e−16	2.6e−12
	sjssif	3.3e−06	1.7e+00	1.2e−06	1.3e−05
	djssif	8.0e−15	2.0e+00	2.1e−15	4.8e−14
	dsjssif	2.0e−06	1.5e+00	7.7e−07	5.5e−06
	$\kappa(S)$	5.8e+09	2.3e+02	1.0e+00	6.6e+13
$\eta = 15$	sssqr	3.3e−04	2.3e+03	2.7e−07	6.0e+06
80 Tests	dssqr	4.8e−13	1.1e+03	5.6e−16	8.1e−03
	sprd	5.5e−03	3.1e+04	4.0e−07	1.2e+13
	dprd	1.5e−13	3.2e+02	7.9e−16	1.0e−06
	sjssif	3.0e−06	2.1e+00	6.3e−07	2.1e−05
	djssif	6.9e−15	1.9e+00	2.6e−15	6.0e−14
	dsjssif	1.9e−06	1.7e+00	6.7e−07	9.9e−06
	$\kappa(S)$	1.5e+18	6.5e+05	1.6e+00	5.5e+25
$\eta = 17$	sssqr	5.0e−04	1.2e+04	3.3e−07	1.1e+12
80 Tests	dssqr	1.2e−12	1.3e+04	5.6e−16	4.3e+03
	sprd	4.3e−02	8.6e+04	2.8e−07	7.5e+11
	dprd	5.1e−13	1.7e+03	8.4e−16	5.0e−03
	sjssif	2.7e−06	1.9e+00	7.3e−07	2.6e−05
	djssif	6.8e−15	1.8e+00	1.9e−15	4.1e−14
	dsjssif	1.7e−06	1.6e+00	7.1e−07	7.7e−06
	$\kappa(S)$	1.1e+21	2.8e+06	1.2e+00	1.9e+29

Wie erwartet, liefern die impliziten Jacobi-Verfahren `xjssif` kleine relative Fehler, und zwar offensichtlich unabhängig von der Ordnung. Die mit QR-Verfahren berechneten Eigenwerte weisen unerwartet häufig ziemlich kleine relative Fehler auf. Dieses Phänomen tritt auch bei anderen Testmatrizen ein.

Mit nachfolgender Tabelle begründen wir die am Schluß von Abschnitt 5 geäußerten Erwartungen. Wir betrachten neben $|S|_S$ die Matrizen M gemäß (5.4) und $\tilde{L}_S = L_S^{(1)}$ gemäß Satz 5.0.6. Die Vermutungen

- $\max(\mathrm{diag}(M)) \approx \min(\mathrm{diag}(M))$
- $1/\lambda_{\min}(|S|_S) \approx 1/\lambda_{\min}(M)$

finden wir durch nachfolgende Tabelle bestätigt. Die Tabelle enthält auch Angaben über $\sigma_{\min}(\tilde{L}_S)$. Wir fassen die Testergebnisse für alle η und alle n zusammen und verwenden wieder geometrisches Mittel und geometrische Streuung, die in diesem Fall allerdings nahezu den arithmetischen Vergleichswerten entsprechen.

Eigenschaften von M, $|S|_S$ und $\tilde{L}$ bei `djssif` (400 Tests)

	μ_g	σ_g	min	max
$\max_i(M_{ii})/\min_i(M_{ii})$	7.7e+00	1.8e+00	1.6e+00	6.6e+01
$\lambda_{\min}(M)$	2.0e−01	1.6e+00	3.8e−02	6.7e−01
$\lambda_{\min}(\|S\|_S)$	7.8e−01	1.2e+00	5.6e−01	1.0e+00
$\lambda_{\min}(M)/\lambda_{\min}(\|S\|_S)$	2.6e−01	1.6e+00	5.9e−02	7.6e−01
$\sigma_{\min}(\tilde{L}_S)$	2.3e−01	1.6e+00	4.7e−02	6.6e−01

Man könnte eine garantierte obere Schranke für den relativen Fehler der berechneten Eigenwerte angeben, indem man gemäß (5.2) nach jeder Diagonalisierung einer 4×4-Submatrix sowie jeder Transformation der Modifikation die Berechnung von $1/\sigma_{\min}(L_S^{(i)})$ mit einem Aufwand von etwa $8n^3/6$ Operationen (Cholesky-Zerlegung) vornimmt. Dieser Aufwand würde aber die Komplexität des Verfahrens erhöhen. Wir wollen uns daher mit oberen Schranken von $1/\sigma_{\min}(L_S^{(i)})$ zufrieden geben und diese aus bereits berechneten Zwischenresultaten gewinnen. Derartige Abschätzungen werden wir nun betrachten. Sei

$$\tau_1 = \prod_{j=1}^{2n} \tilde{L}_{S_{jj}} = \prod_{j=1}^{2n} L_{S_{jj}}^{(1)} = \sqrt{\mathcal{H}(\tilde{L}_S^T \tilde{L}_S)} \tag{6.1}$$

und

$$\tau_{i+1} = \frac{\tau_i}{\sqrt{\det(\hat{K}_{S_i})}} \quad , \quad i \geq 1 \ .$$

wobei $\hat{K}_{S_i}$ die im i-ten Schritt diagonalisierte 4×4-Submatrix ist. Nach 2.3.1 gilt dann

$$\frac{1}{\sigma_{\min}(L_S^{(i)})} \leq \frac{\sqrt{e}}{\tau_i} =: sh_i \quad , \quad e = \exp(1) \ . \tag{6.2}$$

Auf diese Weise haben wir monoton fallende obere Schranken sh_i. Bei modifizierten Zyklen ist sh entsprechend zu erneuern. Wir geben noch weitere Schranken

an. Nach Satz 3.1.4 bzw. Satz 3.1.5 gilt

$$\frac{1}{\sigma_{\min}(L_S^{(i+1)})} \leq \frac{\sqrt{\lambda_{\max}(\hat{K}_{S_i})}}{\sigma_{\min}(L_S^{(i)})} =: se_i \ ,$$

wobei $\hat{K}_{S_i}$ die im i-ten Schritt diagonalisierte 4×4-Submatrix ist. Bei Transformationen im modifizierten Zyklus kann man wegen Satz 3.1.6 eine ähnliche Abschätzung angeben. Mit Hilfe des berechneten $1/\sigma_{\min}(L_S^{(1)})$ erhalten wir also monoton wachsende Schranken se_i. Mit nachfolgender Tabelle machen wir Angaben über den numerisch ermittelten Wert sh_1. Außerdem geben wir die kleinste Nummer r der 4×4-Diagonalisierung, für die $se_i \geq sh_i$, $i \geq r$, gilt, an, sowie se_r. Angaben zu $se_1 = 1/\lambda_{\min}(\tilde{L}_S)$ sind bereits in obiger Tabelle enthalten. Die Skalierung η ist hier belanglos, so daß Werte für verschiedene η zusammengefaßt werden.

Verhalten von $1/\sigma_{\min}(L_S^{(i)})$ im Programm `djssif`

		μ_a	σ_a	min	max
$n = 10$	sh_1	3.6e+02	6.7e+02	7.7e+00	4.9e+03
150 Tests	r	22	9	1	41
	se_r	3.0e+01	3.2e+01	4.0e+00	1.5e+02
$n = 25$	sh_1	4.7e+11	3.3e+12	1.1e+02	2.3e+13
100 Tests	r	200	43	111	266
	se_r	7.0e+06	4.4e+07	1.2e+01	3.2e+08
$n = 50$	sh_1	1.3e+26	6.5e+26	1.1e+05	3.3e+27
100 Tests	r	841	184	279	1054
	se_r	9.0e+16	4.4e+17	2.8e+02	2.3e+18
$n = 100$	sh_1	2.1e+71	1.1e+72	6.8e+08	5.4e+72
50 Tests	r	3315	799	1549	4572
	se_r	3.2e+48	1.6e+49	1.7e+04	8.0e+49

Man erkennt, daß die Schätzwerte höchstens für kleine n brauchbar sind. Die Rotation r befindet sich stets innerhalb des ersten Zyklus.

Abschließend betrachten wir noch die Quotienten von tatsächlichem maximalen relativen Fehler ϱ und erwartetem relativen Fehler ϱ_E gemäß (5.3), wobei wir die Testergebnisse für alle n und alle η wieder zusammenfassen können. Die nachfolgende Tabelle zeigt, daß der erwartete Fehler in der Tat realistisch ist. Hier werden nur Ergebnisse der Programme `sjssif` und `djssif` ausgewertet.

Quotienten tatsächlicher maximaler rel. Fehler und erwarteter maximaler rel. Fehler (400 Tests)				
Verf.	μ_g	σ_g	min	max
`sjssif`	3.0e+00	1.7e+00	8.0e−01	2.7e+01
`djssif`	4.2e+00	1.9e+00	1.1e+00	1.8e+02

Matrizen mit vorbestimmter Kondition von $|S|_S$

In diesem Unterabschnitt betrachten wir solche schiefsymmetrischen Matrizen, deren Eigenschaften hinsichtlich der Störungstheorie man bereits bei ihrer Generierung vorbestimmen kann. Wir generierten Matrizen der Ordnung $2n$ auf folgende Weise:

1. Bestimmung von $\delta, \eta > 0$ und Bildung von $D = \mathrm{diag}(\exp(\delta(2r_i - 1))) \in \mathbb{R}^{2n\times 2n}$ sowie $\Lambda = \mathrm{diag}(\exp(\eta(2s_i - 1))) \in \mathbb{R}^{2n\times 2n}$, wobei $r_i, s_i \in [0, 1]$ „zufällig“ und absteigend geordnet sind.

2. Anwendung eines „Anti-Jacobi“-Verfahrens nach Veselić auf $A_0 = \Lambda$. Dies besteht aus einer Folge orthogonaler Ebenenrotationen $A_{m+1} = X_m^T A_m X_m$, wobei die X_m auf folgende Weise bestimmt werden. Seien

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad , \quad \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix}$$

die Pivotsubmatrizen von A_m bzw. X_m. Dann ist

$$cs = \frac{1}{\sqrt{1+t^2}} \qquad sn = cs \cdot t \quad , \quad t = \frac{\mathrm{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}} \quad , \quad \zeta = \frac{2c}{b-a} \ .$$

Die Folge der auf diese Weise erzeugten Matrizen A_m konvergiert gegen eine Matrix A mit übereinstimmenden Diagonalelementen, d.h. $\kappa(A) = \kappa(A_S)$. Die Konvergenz ist sehr langsam, jedoch bereits nach wenigen (ca. 2) Zyklen gilt $\max(\mathrm{diag}(A))/\min(\mathrm{diag}(A)) \leq 10$.

3. Bestimmung von $|H| = DAD$.

4. Berechnung der Eigenwertzerlegung von $|H|$ mit dem Verfahren `dsyevj` gemäß [29], d.h. Bestimmung eines orthogonalen V und eines diagonalen Δ' mit $V^T|H|V = \Delta'$. Die Diagonalelemente von Δ' seien absteigend geordnet.

5. Bildung der Diagonalmatrix Δ, deren Diagonalelemente die geometrischen Mittel benachbarter Diagonalelemente von Δ' sind, d.h.

$$\Delta_{2i,2i} = \Delta_{2i-1,2i-1} = \sqrt{\Delta'_{2i-1,2i-1}\Delta'_{2i,2i}} \quad , \quad i = 1 \ldots n \ .$$

6. Bestimmung der schiefsymmetrischen Matrix $S = PVJ_n\Delta V^T P^T$, wobei P noch eine „zufällige“ Permutationsmatrix ist.

Unter Zuhilfenahme von

$$U_n = \frac{1}{\sqrt{2}} \bigoplus_{i=1}^{n} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \quad , \quad J'_n = \bigoplus_{i=1}^{n} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad , \quad i^2 = -1$$

mit $U_n^* J_n U_n = iJ'_n$ haben wir nun

$$\begin{aligned} |S| &= |PVJ_n\Delta V^T P^T| = |iPV\Delta^{1/2} U_n J'_n U_n^* \Delta^{1/2} V^T P^T| \\ &= |iPVU_n\Delta^{1/2} J'_n \Delta^{1/2} U_n^* V^T P^T| = PVU_n\Delta U_n^* V^T P^T \\ &= PV\Delta V^T P^T \ . \end{aligned}$$

Wir können annehmen, daß obiger Schritt 5. die Eigenwerte von Δ' nicht stark (d.h. nicht in ihrer Größenordnung) verändert, so daß wir

$$\lambda_{\min}(|S|_S) = \lambda_{\min}((V\Delta V^T)_S) \approx \lambda_{\min}((V\Delta' V^T)_S) = \lambda_{\min}(|H|_S) = \lambda_{\min}(A) \tag{6.3}$$

erhalten. Die unten aufgeführten numerischen Resultate bestätigen diese Annahme. Die Eigenwerte von A sind jedoch nach Schritt 1. bekannt. (6.3) gilt natürlich auch für $\lambda_{\max}(\cdot)$ und $\kappa(\cdot)$. Es wurden numerische Tests für

$$\eta = 2 \quad , \quad \delta \in \{2, 5, 8, 15, 17\} \quad , \quad n \in \{10, 25, 50, 100\}$$

vorgenommen, und zwar 30 Tests für $n = 10$, jeweils 20 Tests für $n = 25$, $n = 50$ und 10 Tests für $n = 100$. Insgesamt wurden also 400 Matrizen dieses Typs generiert.

Bei der Darstellung der Testergebnisse gehen wir analog zum vorhergehenden Unterabschnitt vor. Zuerst geben wir einen Überblick über die Rechenzeiten, wobei die Ergebnisse für verschiedene δ zusammengefaßt werden.

Rechenzeiten in 1/100 sec.
bei Matrizen mit vorbestimmtem $|S|_S$

	Verf.	μ_a	σ_a	min	max
n=10	`dssqr`	3	0.6	1	4
150 Tests	`dprd`	20	1.2	17	25
	`djssif`	3	1.0	1	7
n=25	`dssqr`	9	1.0	7	11
100 Tests	`dprd`	316	7.6	294	335
	`djssif`	31	9.8	19	52
n=50	`dssqr`	43	3.6	38	53
100 Tests	`dprd`	2610	31	2501	2683
	`djssif`	214	69	125	355
n=100	`dssqr`	290	20	266	345
50 Tests	`dprd`	16745	9686	2580	23827
	`djssif`	1650	478	992	2489

Das QR-Verfahren `dssqr` war für $n = 100$ etwa 4.9 bis 6.4 mal schneller als `djssif` mit einem Mittel bei etwa 5.4.

Anschließend geben wir analog zum vorhergehenden Unterabschnitt einen Überblick über die Zahl der Operationen, zusammengefaßt für alle δ.

		Zahl der Zyklen				Zahl der Rotationen bzw. 4×4-Diagonalisierungen			
	Verf.	μ_a	σ_a	min	max	μ_a	σ_a	min	max
n=10	`dprd`	4	0.9	3	6	663	152	426	1050
	`djssif`	4	0.9	3	6	100	34	53	196
n=25	`dprd`	6	0.7	5	8	6766	806	5070	8430
	`djssif`	5	1.2	4	8	894	336	501	1636
n=50	`dprd`	8	1.0	6	10	36454	4362	28276	48328
	`djssif`	7	1.5	5	10	4498	1707	2389	7739
n=100	`dprd`	10	1.1	8	11	175794	19570	143446	204416
	`djssif`	8	1.3	6	11	21952	7432	12281	34523

Als nächstes betrachten wir wieder die maximalen relativen Fehler der berechneteten Eigenwerte gegenüber den mit `qjssif` berechneten Werten. Wie im vorangehenden Unterabschnitt werden hier das goemetrische Mittel μ_g und die geometrische Streuung σ_g verwendet. Wir fassen wieder die Ergebnisse für alle n zusammen.

Maximale relative Fehler gegenüber `qjssif`					
	Verf.	μ_g	σ_g	min	max
$\delta = 2$ 80 Tests	`sssqr`	9.0e−07	1.8e+00	2.9e−07	3.3e−06
	`dssqr`	2.3e−15	1.9e+00	6.5e−16	7.5e−15
	`sprd`	1.8e−06	2.5e+00	3.4e−07	1.3e−05
	`dprd`	3.1e−15	2.4e+00	7.6e−16	3.7e−14
	`sjssif`	5.4e−06	2.6e+00	9.9e−07	3.1e−05
	`djssif`	1.3e−14	2.9e+00	2.6e−15	7.4e−14
	`dsjssif`	3.1e−06	2.2e+00	8.7e−07	1.3e−05
	$\kappa(S)$	2.9e+03	2.4e+00	1.5e+02	9.5e+03
$\delta = 5$ 80 Tests	`sssqr`	2.9e−06	6.0e+00	2.1e−07	2.8e−04
	`dssqr`	6.0e−15	5.7e+00	5.5e−16	6.4e−13
	`sprd`	2.4e−05	8.9e+00	5.0e−07	2.8e−02

	dprd	3.2e−14	7.0e+00	7.2e−16	2.2e−12
	sjssif	4.0e−06	2.3e+00	9.6e−07	3.3e−05
	djssif	9.1e−15	2.6e+00	2.3e−15	5.8e−14
	dsjssif	2.2e−06	2.1e+00	5.7e−07	9.5e−06
	$\kappa(S)$	1.0e+08	4.0e+00	1.4e+06	5.5e+08
$\delta = 8$	sssqr	4.6e−06	1.1e+01	2.0e−07	1.9e−03
80 Tests	dssqr	1.1e−14	1.1e+01	6.7e−16	1.4e−11
	sprd	7.5e−02	5.0e+01	3.7e−05	2.4e+01
	dprd	5.7e−12	4.0e+01	1.8e−15	3.5e−09
	sjssif	3.6e−06	2.3e+00	8.6e−07	5.3e−05
	djssif	8.3e−15	2.6e+00	2.0e−15	6.1e−14
	dsjssif	1.8e−06	1.9e+00	5.4e−07	8.3e−06
	$\kappa(S)$	4.2e+12	5.9e+00	3.4e+10	9.0e+13
$\delta = 15$	sssqr	1.5e−05	3.3e+01	1.8e−07	1.4e−01
80 Tests	dssqr	3.8e−14	2.8e+01	6.4e−16	5.0e−11
	sprd	1.1e+06	1.9e+04	2.4e−02	2.6e+16
	dprd	1.6e−05	5.2e+03	1.9e−13	5.5e+03
	sjssif	3.1e−06	2.3e+00	7.1e−07	4.9e−05
	djssif	6.8e−15	2.6e+00	1.5e−15	4.7e−14
	dsjssif	1.4e−06	1.9e+00	3.4e−07	1.4e−06
	$\kappa(S)$	8.9e+23	5.2e+02	1.3e+18	7.8e+25
$\delta = 17$	sssqr	1.7e−05	2.8e+01	1.7e−07	2.4e−02
80 Tests	dssqr	4.3e−14	2.5e+01	8.2e−16	5.6e−11
	sprd	3.1e+08	1.2e+06	9.3e−03	2.0e+20
	dprd	6.5e−04	4.5e+04	4.6e−12	4.8e+07
	sjssif	2.6e−06	2.2e+00	5.9e−07	2.6e−05
	djssif	6.3e−15	2.7e+00	8.4e−16	3.6e−14
	dsjssif	1.3e−06	1.9e+00	3.1e−07	4.0e−06
	$\kappa(S)$	1.2e+27	1.1e+02	6.2e+20	2.0e+29

Auch wird wird deutlich, daß die mit QR-Verfahren berechneten Eigenwerte unerwartet kleine relative Fehler haben. Das QR-Verfahren liefert trotz hoher Kondition oft noch brauchbare Ergebnisse. Hingegen sind die Ergebnisse des Paardekooper-Verfahrens meist unbrauchbar.

Rechenzeiten und relative Genauigkeiten sind also beim Paardekooper-Verfahren unbrauchbar. Wir werden es in den nachfolgenden Unterabschnitten daher nicht mehr zu Vergleichszwecken heranziehen.

Anschließend betrachten wir wieder Eigenschaften der Matrizen $|S|_S$, M gemäß (5.4) und $\tilde{L} = L_S^{(1)}$ gemäß Satz 5.0.6. Der nachfolgenden Tabelle kann man auch entnehmen, daß die Testmatrizen die erwarteten Eigenschaften haben, denn

ihre Kondition ist hoch, während die Kondition von $|S|_S$ zumeist klein ist.

Eigenschaften von M, $\|S\|_S$ und $\tilde{L}$ bei `djssif`(400 Tests)				
	μ_g	σ_g	min	max
$\max_i(M_{ii})/\min_i(M_{ii})$	3.4e+00	2.0e+00	1.0e+00	2.1e+01
$\lambda_{\min}(M)$	6.3e−02	3.3e+00	5.4e−07	3.4e−01
$\lambda_{\min}(\|S\|_S)$	1.1e−01	3.7e+00	6.4e−07	4.4e−01
$\lambda_{\min}(M)/\lambda_{\min}(\|S\|_S)$	5.8e−01	1.5e+00	1.9e−01	1.0e+00
$\sigma_{\min}(\tilde{L}_S)$	3.5e−01	1.8e+00	6.0e−02	9.2e−01

Wie im vorangehenden Unterabschnitt betrachten wir nun das Verhalten von $1/\sigma_{\min}(L_S^{(i)})$.

Verhalten von $1/\sigma_{\min}(L_S^{(i)})$ im Programm `djssif`					
		μ_a	σ_a	min	max
$n = 10$	sh_1	5.8e+00	6.8e+00	1.8e+00	4.9e+01
150 Tests	r	17	12	1	44
	se_r	3.8e+00	2.6e+00	1.7e+00	1.9e+01
$n = 25$	sh_1	1.1e+05	5.5e+05	5.5e+00	3.9e+06
100 Tests	r	100	31	35	172
	se_r	1.6e+03	6.0e+03	3.6e+00	4.2e+04
$n = 50$	sh_1	4.8e+16	1.2e+16	3.1e+03	5.4e+16
100 Tests	r	537	98	350	698
	se_r	6.3e+10	2.1e+11	1.5e+02	1.1e+12
$n = 100$	sh_1	6.9e+41	3.0e+42	1.6e+12	1.5e+43
50 Tests	r	2428	287	1801	2870
	se_r	4.4e+31	2.1e+32	1.7e+07	1.1e+33

Die damit bestimmbaren Schätzwerte sind auch hier allenfalls für kleine n brauchbar, wenngleich die hier erzielten Ergebnisse zumeist etwas günstiger sind als die im vorangehenden Unterabschnitt.

Als letztes folgt auch hier eine Tabelle über das Verhältnis von tatsächlichem zu erwartetem Fehler. Es werden nur Ergebnisse von `sjssif` und `djssif` ausgewertet.

Quotienten tatsächlicher maximaler rel. Fehler und erwarteter maximaler rel. Fehler (400 Tests)				
Verf.	μ_g	σ_g	min	max
`sjssif`	1.5e+00	4.5e+00	4.0e−06	1.2e+01
`djssif`	1.9e+00	4.9e+00	5.8e−06	1.2e+01

Wie im vorangehenden Unterabschnitt sind die erwarteten Fehler in der Tat realistisch oder sogar größer als die tatsächlichen Fehler.

Tridiagonale Matrizen

Kleine relative Störungen schiefsymmetrischer tridiagonaler Matrizen führen zu kleinen relativen Änderungen der Eigenwerte [12]. Schiefsymmetrische tridiagonale Matrizen bestimmen also ihre Eigenwerte unabhängig von der Kondition gut.

Das in Abschnitt 6.1 beschriebene QR-Verfahren bestimmt die Eigenwerte solcher Matrizen mit kleinen relativen Fehlern [12]. Dies erwarten wir auch vom hier eingeführten impliziten Verfahren. In diesem Unterabschnitt präsentieren wir numerische Resultate zu diesen Matrizen. Die Matrizen wurden auf folgende Weise erzeugt:

1. Bestimmung der Ordnung $2n$ und eines $\eta \in \{2, 5, 8, 15, 17\}$.
2. Generierung von $D = \mathrm{diag}(\delta_i)$ mit $\delta_i = \exp(\eta(2r_i - 1))$, $i = 1 \ldots 2n - 1$, und „zufälligen“ $r_i \in [0, 1]$.
3. Generierung von $E = \mathrm{diag}(\sigma_i)$ mit „zufälligen“ $\sigma_i \in \{-1, 1\}$, $i = 1 \ldots 2n-1$.
4. Generierung der Testmatrix S mit $S_{i,i+1} = -S_{i+1,i} = D_i E_i$, $i = 1 \ldots 2n - 1$, und $S_{ij} = 0$ sonst.

Für jedes $\eta \in \{2, 5, 8, 15, 17\}$ und $n = 10$ wurden 30 Testmatrizen, für $n = 25$, $n = 50$ wurden 20 Matrizen, und für $n = 100$ wurden 10 Testmatrizen generiert, insgesamt also 400 Matrizen.

Die Kondition der auf diese Weise erzeugten Matrizen wird sehr groß und bei den einfach genauen Programmen tritt häufig Unter- oder -überlauf ein. Wir verzichten daher in diesem Unterabschnitt auf die Darstellung der Ergebnisse der einfach genauen Programme.

Wie angekündigt, werden wir die Ergebnisse der Paardekooper-Verfahren in diesem Unterabschnitt nicht mehr aufführen. Es sei hier nur erwähnt, daß diese Verfahren hinsichtlich der erzielten Genauigkeit keine brauchbaren Ergebnisse liefern.

Auf die Angabe der Rechenzeiten verzichten wir hier auch. Nachfolgend geben wir sogleich die Zahl der erforderlichen 4×4-Diagonalisierungen beim Verfahren `djssif` an. Mit Hilfe der entsprechenden Angaben in den vorangehenden Unterabschnitten könnten Schätzwerte für die Rechenzeiten bestimmt werden.

	Zahl der Zyklen				4×4-Diagonalisierungen			
n	μ_a	σ_a	min	max	μ_a	σ_a	min	max
10	3	0.9	2	5	61	42	13	159
25	4	0.9	3	6	293	294	46	996
50	4	1.0	3	6	799	971	105	3678
100	4	1.0	3	6	1807	2417	224	8936

Bei Matrizen dieser Art sollte natürlich das QR-Verfahren das Verfahren der Wahl sein, weil die Matrix nicht trianguliert zu werden brauch und somit die Rechenzeit vernachlässigbar ist. Außerdem werden die Eigenwerte mit hoher relativer Genauigkeit berechnet. Dies kommt in der nachfolgenden Aufstellung der erzielten maximalen relativen Fehler in den berechneten Eigenwerten auch zum Ausdruck. Wir verzichten hier wie angekündigt auf die Ergebnisse der einfach genauen Programme.

Maximale relative Fehler gegenüber `qjssif`					
	Verf.	μ_g	σ_g	min	max
$\eta = 2$	`dssqr`	2.3e−15	2.0e+00	6.9e−16	3.7e−14
80 Tests	`djssif`	5.0e−15	1.6e+00	1.8e−15	1.6e−14
	$\kappa(S)$	9.2e+04	1.7e+03	3.5e+01	8.4e+22
$\eta = 5$	`dssqr`	1.5e−15	1.5e+00	5.4e−16	4.2e−15
80 Tests	`djssif`	3.5e−15	1.9e+00	9.6e−16	2.8e−14
	$\kappa(S)$	4.6e+11	1.6e+06	7.7e+02	1.7e+27
$\eta = 8$	`dssqr`	1.2e−15	1.5e+00	5.6e−16	1.4e−14
80 Tests	`djssif`	2.8e−15	2.1e+00	7.0e−16	2.6e−14
	$\kappa(S)$	1.6e+20	4.0e+12	2.2e+04	1.8e+80
$\eta = 15$	`dssqr`	7.5e−16	1.5e+00	2.7e−16	2.1e−15
80 Tests	`djssif`	2.0e−15	2.0e+00	5.1e−16	1.4e−14
	$\kappa(S)$	1.5e+39	1.4e+27	1.4e+10	2.9e+169
$\eta = 17$	`dssqr`	7.0e−16	1.4e+00	2.7e−16	1.6e−15
80 Tests	`djssif`	1.6e−15	1.8e+00	5.2e−16	1.3e−14
	$\kappa(S)$	4.7e+39	9.9e+20	1.2e+12	2.4e+97

Wir erwarteten vom impliziten Jacobi-Verfahren kleine relative Fehler in den berechneten Eigenwerten. Diese Erwartung wird also bestätigt. Das etwas unerwartete Verhalten der Kondition ist ohne Belang.

Wir betrachten nun Eigenschaften der Matrizen $|S|_S$, M gemäß (5.4) und $\tilde{L} = L_S^{(1)}$ gemäß Satz 5.0.6. Wir erwähnen hier jedoch nur die Resultate für $\eta = 2$ und $n \in \{10, 25, 50\}$ sowie für $\eta \in \{5, 8\}$ und $n = 10$. Bereits für $\eta \in \{5, 8\}$ sind die Ergebnisse nicht brauchbar. Die Darstellung der Resultate für andere η, n ist somit nicht sinnvoll und wurde unterlassen. Da die tatsächlichen maximalen relativen Fehler klein sind, zeigt dies deutlich, daß die Fehleranalyse der schiefsymmetrischen Zerlegung dieser Art von Matrizen verbessert werden sollte.

Eigenschaften von M, $\lvert S\rvert_S$ und $\tilde{L}$ bei `djssif`(400 Tests)					
		μ_g	σ_g	min	max
$\eta = 2$	$\max_i(M_{ii})/\min_i(M_{ii})$	1.8e+00	1.2e+00	1.4e+00	2.9e+00
$n = 10$	$\lambda_{\min}(M)$	8.2e−03	8.2e+00	6.3e−05	1.7e−01
50 Tests	$\lambda_{\min}(\lvert S\rvert_S)$	9.8e−03	8.4e+00	7.6e−05	2.0e−01
	$\lambda_{\min}(M)/\lambda_{\min}(\lvert S\rvert_S)$	8.4e−01	1.1e+00	7.2e−01	9.8e−01
	$\sigma_{\min}(\tilde{L}_S)$	5.5e−01	1.1e+00	3.9e−01	7.1e−01
$\eta = 2$	$\max_i(M_{ii})/\min_i(M_{ii})$	2.4e+00	1.2e+00	1.8e+00	2.9e+00
$n = 25$	$\lambda_{\min}(M)$	2.8e−05	4.6e+01	1.1e−08	3.8e−03
30 Tests	$\lambda_{\min}(\lvert S\rvert_S)$	3.4e−05	4.4e+01	1.6e−08	4.0e−03
	$\lambda_{\min}(M)/\lambda_{\min}(\lvert S\rvert_S)$	8.3e−01	1.1e+00	7.1e−01	9.8e−01
	$\sigma_{\min}(\tilde{L}_S)$	4.5e−01	1.1e+00	3.6e−01	5.4e−01
$\eta = 2$	$\max_i(M_{ii})/\min_i(M_{ii})$	2.7e+00	1.1e+00	2.2e+00	3.4e+00
$n = 50$	$\lambda_{\min}(M)$	3.6e−05	4.3e+01	4.5e−09	1.5e−03
30 Tests	$\lambda_{\min}(\lvert S\rvert_S)$	4.4e−05	4.4e+01	5.1e−09	2.0e−03
	$\lambda_{\min}(M)/\lambda_{\min}(\lvert S\rvert_S)$	8.3e−01	1.1e+00	7.0e−01	9.7e−01
	$\sigma_{\min}(\tilde{L}_S)$	4.0e−01	1.1e+00	2.9e−01	5.2e−01
$\eta = 5$	$\max_i(M_{ii})/\min_i(M_{ii})$	1.5e+00	1.3e+00	1.0e+00	2.1e+00
$n = 10$	$\lambda_{\min}(M)$	1.9e−05	9.3e+02	3.0e−14	1.9e−02
50 Tests	$\lambda_{\min}(\lvert S\rvert_S)$	2.1e−05	9.2e+02	3.1e−14	2.1e−02
	$\lambda_{\min}(M)/\lambda_{\min}(\lvert S\rvert_S)$	9.1e−01	1.1e+00	7.1e−01	1.0e+00
	$\sigma_{\min}(\tilde{L}_S)$	6.7e−01	1.2e+00	4.4e−01	8.9e−01
$\eta = 8$	$\max_i(M_{ii})/\min_i(M_{ii})$	1.4e+00	1.3e+00	1.0e+00	2.2e+00
$n = 10$	$\lambda_{\min}(M)$	2.3e−07	1.4e+04	4.6e−17	1.8e−01
50 Tests	$\lambda_{\min}(\lvert S\rvert_S)$	2.8e−07	1.1e+04	6.4e−16	2.2e−01
	$\lambda_{\min}(M)/\lambda_{\min}(\lvert S\rvert_S)$	8.5e−01	1.7e+00	5.1e−02	1.0e+00
	$\sigma_{\min}(\tilde{L}_S)$	7.1e−01	1.2e+00	5.1e−01	9.7e−01

Bei tridiagonalen Matrizen sind während des Verfahrens anfallende Meßwerte durchaus häufig zur Bestimmung garantierter oberer Schranken für den relativen Fehler in den Eigenwerten verwendbar, wie nachfolgende Tabelle über die Entwicklung von $1/\sigma_{\min}(L_S^{(i)})$ zeigt. Tridiagonale Matrizen unterscheiden sich hierin also von den in den vergangenen unterabschnitten behandelten Matrizen. Wir fassen die Ergebnisse für alle n und alle η zusammen. Die angegeben Werte wachsen leicht mit n.

Verhalten von $1/\sigma_{\min}(L_S^{(i)})$ im Programm `djssif`				
	μ_a	σ_a	min	max
sh_1	1.2e+04	1.0e+05	1.8e+00	1.2e+06
r	85	95	1	561
se_r	8.0e+01	5.4e+02	1.7e+00	7.1e+03

Eine Aufstellung der Quotienten der tatsächlichen relativen Fehler und den erwarteten relativen Fehler geben wir hier nicht, weil die in der vorletzten Tabelle dargelegten Resultate von $\lambda_{\min}(M)$ und $\sigma_{\min}(\tilde{L}_S)$ zu unrealistisch hohen erwarteten Fehlern führen würden. Allenfalls für kleine η und kleine n ergäben sich brauchbare Schätzwerte.

Azyklische Matrizen

Matrizen, deren Graph azyklisch ist, heißen auch selbst azyklisch. Azyklische Matrizen bestimmen nach [11] ihre Singulärwerte unabhängig von der Kondition gut, d.h. kleine relative Störungen der nichtverschwindenden Elemente der Matrix führen zu kleinen relativen Änderungen der Singulärwerte. Sei A azyklisch und nichtsingulär. Dann sind die positiven Imaginärteile der Eigenwerte von

$$S = \begin{bmatrix} 0 & A \\ -A^T & 0 \end{bmatrix}$$

genau die positiven Singulärwerte von A. Kleine relative Störungen der Elemente von S führen also zu kleinen relativen Änderungen der Eigenwerte von S. Ein Bisektionsverfahren zur genauen Eigenwertbestimmung wurde in [12] angegeben.

Auch vom hier vorgeschlagenen impliziten Verfahren erwarten wir eine hohe relative Genauigkeit in den berechneten Eigenwerten. Resultate werden in diesem Unterabschnitt angegeben. Auf einen direkten Vergleich mit dem in [12] angegebenen Verfahren wurde verzichtet. Das dortige Bisektionsverfahren wurde nicht implementiert. Die relativen Fehler vergleichen wir wie in den vorangehenden Unterabschnitten mit dem beschriebenen QR-Verfahren.

Folgende Typen azyklischer Matrizen wurden generiert und getestet:

1. Bestimmung der Ordnung $2n$ und eines $\eta \in \{2, 5, 8, 15, 17\}$.
2. Generierung der Vektoren $D \in R^n$, $E \in R^{n-1}$ mit $D_i = \exp(\eta(2r_i - 1))$, $i = 1 \ldots n$, und $E_i = \exp(\eta(2s_i - 1))$, $i = 1 \ldots n-1$. Dabei seien $r_i, s_i \in [0, 1]$ „zufällig“.
3. Bestimmung der Matrix $A \in \mathbb{R}^{n \times n}$ mit $A_{ii} = D_i$, $i = 1 \ldots n$, $A_{in} = E_i$, $i = 1 \ldots n-1$, und $A_{ij} = 0$ sonst.
4. Bestimmung der Testmatrix

$$S = P^T \begin{bmatrix} 0 & A \\ -A^T & 0 \end{bmatrix} P \ ,$$

wobei P noch eine „zufällige“ Permutationsmatrix ist.

Für jedes $\eta \in \{2, 5, 8, 15, 17\}$ und $n = 10$ wurden 30 Testmatrizen, für $n = 25$, $n = 50$ wurden 20 Matrizen, und für $n = 100$ wurden 10 Testmatrizen generiert, insgesamt also 400 Matrizen.

Auch hier verzichten wir auf die Angabe von Rechenzeiten. Der Vollständigkeit halber geben wir lediglich eine Aufstellung der benötigten Zahl der 4×4-Diagonalisierungen des Programms `djssif` an. Schätzwerte für die Rechenzeiten könnten anhand der Angaben in den ersten beiden Unterabschnitten bestimmt werden. Auf Angaben für die einfach und vierfach genauen Verfahren verzichten wir hier. Wie angekündigt, werden Angaben über das Paardekooper-Verfahren nicht mehr aufgenommen.

	Zahl der Zyklen				4×4-Diagonalisierungen			
n	μ_a	σ_a	min	max	μ_a	σ_a	min	max
10	4	1.0	2	6	83	45	18	184
25	4	1.1	3	6	606	359	142	1350
50	5	1.0	3	7	2594	1669	552	5919
100	5	1.4	3	9	11177	7819	2554	26743

Nachfolgend betrachten wir die relativen Fehler in den berechneten Eigenwerten gegenüber den vierfach genauen Ergebnissen. Angaben über die Kondition der Testmatrizen sind in die Tabelle aufgenommen.

Maximale relative Fehler gegenüber `qjssif`					
	Verf.	μ_g	σ_g	min	max
$\eta = 2$	`sssqr`	8.6e−07	2.0e+00	2.6e−07	4.1e−06
80 Tests	`dssqr`	2.0e−15	1.9e+00	5.9e−16	9.5e−15
	`sjssif`	3.4e−06	2.1e+00	7.5e−07	1.9e−05
	`djssif`	1.1e−14	2.7e+00	2.0e−15	7.6e−14
	`dsjssif`	1.8e−06	1.8e+00	6.8e−07	1.0e−05
	$\kappa(S)$	3.2e+02	4.7e+00	1.6e+01	9.2e+03
$\eta = 5$	`sssqr`	6.4e−07	1.6e+00	2.3e−07	2.2e−06
80 Tests	`dssqr`	1.7e−15	1.7e+00	6.4e−16	7.4e−15
	`sjssif`	1.9e−06	2.1e+00	5.1e−07	1.7e−05
	`djssif`	6.4e−15	2.3e+00	6.8e−16	3.3e−14
	`dsjssif`	9.4e−07	1.7e+00	2.6e−07	5.0e−06
	$\kappa(S)$	2.7e+05	2.4e+01	1.4e+03	3.2e+08
$\eta = 8$	`sssqr`	5.1e−07	1.6e+00	2.1e−07	2.1e−06
80 Tests	`dssqr`	1.6e−15	1.6e+00	5.4e−16	4.5e−15

	`sjssif`	1.1e−06	2.5e+00	1.2e−07	1.5e−05
	`djssif`	5.2e−15	2.4e+00	1.2e−15	2.6e−14
	`dsjssif`	5.4e−07	1.9e+00	1.4e−07	5.4e−06
	$\kappa(S)$	4.8e+08	1.3e+02	1.1e+05	2.8e+13
$\eta = 15$ 80 Tests	`sssqr`	3.8e−07	1.7e+00	1.1e−07	2.0e−06
	`dssqr`	1.2e−15	1.5e+00	3.9e−16	3.2e−15
	`sjssif`	7.5e−07	2.8e+00	8.8e−08	1.3e−05
	`djssif`	3.8e−15	2.4e+00	5.2e−16	2.6e−14
	`dsjssif`	3.2e−07	1.7e+00	8.8e−08	1.1e−06
	$\kappa(S)$	2.7e+16	3.0e+03	1.3e+10	3.2e+25
$\eta = 17$ 80 Tests	`sssqr`	3.6e−07	1.6e+00	1.0e−07	9.5e−07
	`dssqr`	1.1e−15	1.6e+00	2.2e−16	3.5e−15
	`sjssif`	5.9e−07	2.7e+00	8.1e−08	8.6e−06
	`djssif`	3.2e−15	2.5e+00	6.2e−16	2.5e−14
	`dsjssif`	3.2e−07	2.0e+00	1.3e−07	4.5e−06
	$\kappa(S)$	1.5e+18	1.4e+04	3.3e+08	1.9e+28

Insgesamt lieferte das QR-Verfahren genaue Resultate. Aber auch die Ergebnisse der impliziten Verfahren haben kleine relative Fehler.

Anschließend betrachten wir wieder Eigenschaften der Matrizen $\|S\|_S$, M gemäß (5.4) und $\tilde{L} = L_S^{(1)}$ gemäß Satz 5.0.6.

Eigenschaften von M, $\|S\|_S$ und $\tilde{L}$ bei `djssif`(400 Tests)				
	μ_g	σ_g	min	max
$\max_i(M_{ii})/\min_i(M_{ii})$	2.3e+00	2.0e+00	1.0e+00	2.6e+01
$\lambda_{\min}(M)$	1.5e−05	1.8e+03	8.7e−15	1.0e+00
$\lambda_{\min}(\|S\|_S)$	2.1e−05	1.9e+03	1.5e−14	1.0e+00
$\lambda_{\min}(M)/\lambda_{\min}(\|S\|_S)$	7.1e−01	1.4e+00	2.0e−01	1.0e+00
$\sigma_{\min}(\tilde{L}_S)$	5.2e−01	4.7e+00	1.7e−01	1.0e+00

Nun stellen wir zusammen, wie sich $1/\sigma_{\min}(L_S^{(i)})$ verhalten hat, wobei wie die Werte für alle n und alle η zusammenfassen.

Verhalten von $1/\sigma_{\min}(L_S^{(i)})$ bei `djssif`				
	μ_a	σ_a	min	max
sh_1	3.2e+01	3.50+02	1.6e+00	5.0e+03
r	69	257	1	3915
se_r	7.5e+00	3.0e+01	1.0e+00	3.9e+02

Diese Werte sind also durchaus zur Bestimmung garantierter oberer Schranken für die maximalen relativen Fehler in den berechneten Eigenwerten verwendbar.

Die Werte wachsen leicht mit n. Bei anderen Typen von Matrizen (siehe die ersten beiden Unterabschnitte) waren die Werte unbrauchbar.

Schließlich folgt auch hier eine Tabelle über das Verhältnis von tatsächlichem zu erwartetem Fehler.

Quotienten tatsächlicher maximaler rel. Fehler und erwarteter maximaler rel. Fehler (400 Tests)				
Verf.	μ_g	σ_g	min	max
`sjssif`	1.5e−04	2.6e+03	1.2e−13	5.2e+00
`djssif`	3.5e−04	2.0e+03	1.8e−13	8.8e+00

Die tatsächlichen Fehler sind hier also zumeist weitaus kleiner als die erwarteten maximalen relativen Fehler. Dies liegt vor allem am ungünstigen $\lambda_{\min}(M)$. Man kann erwarten, daß die Störungstheorie dieser Art von Matrizen noch verbessert werden kann.

6.2.2 Zum Konvergenzverhalten

In diesem Unterabschnitt betrachten wir ausschließlich das in dieser Arbeit vorgestellte implizite Verfahren. Wir widmen uns dem Konvergenzverhalten seines iterativen Teils.

Hinsichtlich des Konvergenzverhaltens sind explizites und implizites Verfahren äquivalent. Alle Aussagen drücken wir daher in Termen des expliziten Verfahrens aus. Sei L der Faktor der schiefsymmetrischen Zerlegung von $S \in \mathbb{R}^{2n\times 2n}$, also $S = -S^T = LJL^T$. Außerdem sei $K = L^T L$ die symmetrische, positiv definite Matrix, die im iterativen Teil mit einer Folge symplektischer Transformationen diagonalisiert wird. Sei $K^{(1)} = K$ und $K^{(i)}$, $i > 1$, die transformierte Matrix vor dem i-ten Zyklus. In jedem Zyklus werden $n(n-1)/2$ Diagonalisierungen von Submatrizen der Ordnung 4 vorgenommen, wobei die Transformationen in modifizierten Zyklen entsprechend mitzuzählen sind. Sei weiterhin $K^{(i)} = D^{(i)} K_S^{(i)} D^{(i)}$, $D^{(i)} = \operatorname{diag}((K_{jj}^{(i)})^{1/2})$, die jeweils skalierte Matrix.

Wir erwarten asymptotisch quadratische Konvergenz der skalierten Matrizen zur Einheitsmatrix I, genauer

$$\lim_{i\to\infty} \frac{\|K_S^{(i)} - I\|_E}{\|K_S^{(i-1)} - I\|_E^2} \le c \tag{6.4}$$

mit einem moderaten c. Einen Beweis werden wir nicht bringen. Wegen der endlichen Rechnergenauigkeit kann (6.4) natürlich nur für möglichst große i plausibel gemacht werden. Insbesondere erwarten wir von den modifizierten Zyklen, daß

sie dem Verfahren auch bei mehrfachen Eigenwerten schließlich zu quadratischer Konvergenz verhelfen.

Zur Untersuchung des Verhaltens verwenden wir Resultate des vierfach genauen Programms `qjssif`. Wir bezeichnen die Abbruchkonstanten der Programme `djssif` und `sjssif` mit TOL_d bzw. TOL_s. Solange $\max_{i\neq j} |K_{S_{ij}}^{(k)}| \geq TOL_d$ bzw. $\geq TOL_s$ gilt, sind die Zwischenergebnisse der verschieden genauen Programme (nahezu) gleich. Die verschiedenen Genauigkeiten sind hier ohne Belang. Weil bei den Programmen geringerer Genauigkeit die Berechnung früher abgebrochen wird, kann man an ihren Ergebnissen das asymptotische Konvergenzverhalten nur recht unzureichend beobachten. Wir beobachten daher die Resultate von `qjssif` und verwenden diese zur Beurteilung des Konvergenzverhaltens der Programme geringerer Genauigkeit. Sei

$$k_s = \min_{\|K_S^{(i)}-I\|_E \leq TOL_s} \{i \geq 1\} \quad , \quad k_d = \min_{\|K_S^{(i)}-I\|_E \leq TOL_d} \{i \geq 1\} \ .$$

Für $\varrho = 1, 2$ geben dann die Werte

$$c_s^{[\varrho]} = \frac{\|K_S^{(k_s)} - I\|_E}{\|K_S^{(k_s-1)} - I\|_E^{\varrho}} \quad , \quad c_d^{[\varrho]} = \frac{\|K_S^{(k_d)} - I\|_E}{\|K_S^{(k_d-1)} - I\|_E^{\varrho}}$$

Auskunft über das lineare bzw. quadratische Konvergenzverhalten der einfach bzw. doppelt genauen Programme.

In den nachfolgenden Unterabschnitten geben wie die Resultate wieder, die wir bei skaliert blockdiagonaldominanten Matrizen und bei Matrizen mit vorbestimmter Kondition von $|S|_S$ erhielten.

Skaliert blockdiagonaldominante Matrizen

Die in diesem Unterabschnitt betrachteten Matrizen werden genauso erzeugt, wie die entsprechenden Matrizen im vorangehenden Abschnitt (s. Seite 183). Die Norm der Störungsmatrix E variieren wir allerdings zwischen den drei Werten $0, 10^{-11}$ und $1/2$. Da die Skalierung hier ohne Belang ist, beschränken wir uns auf $\eta = 2$.

Zuerst sei $\|E\|_2 = 0$. Dies sind Matrizen, deren Eigenwerte in einfacher Genauigkeit als mehrfach erscheinen und in doppelter Genauigkeit sehr nahe beieinander liegen (Cluster). Für die Cluster ist der Fehler bei der Generierung der Testmatrizen verantwortlich. Von `sjssif` können wir asymptotisch quadratische Konvergenz erwarten, während in der letzten Phase vor dem Programmabbruch von `djssif` nur noch lineare Konvergenz zu erwarten ist. Die oben definierten Werte $c_s^{[\varrho]}$ und $c_d^{[\varrho]}$ haben sich, zusammengefaßt für alle 80 Testmatrizen der Ordnung $n \in \{10, 25, 50, 100\}$, so verhalten:

Konvergenzverhalten von `qjssif` (80 Tests)				
	μ_g	σ_g	min	max
$c_s^{[1]}$	4.1e−12	7.1e+00	2.1e−13	1.8e−08
$c_s^{[2]}$	1.6e−09	2.8e+01	1.7e−11	5.6e−04
$c_d^{[1]}$	7.6e−02	2.6e+00	1.9e−03	6.0e−01
$c_d^{[2]}$	1.2e+14	2.5e+00	8.5e+12	1.7e+15

Sei nun $\|E\|_2 = 10^{-11}$. Diese Matrizen erscheinen dem einfach genauen Programm `sjssif` wie solche mit mehrfachen Eigenwerten. Für das doppelt genaue Programm erscheinen sie hingegen wie solche mit verschiedenen Eigenwerten. Für das einfach genaue Programm `sjssif` erwarten wir also asymptotisch quadratische Konvergenz.

Konvergenzverhalten von `qjssif` (80 Tests)				
	μ_g	σ_g	min	max
$c_s^{[1]}$	2.8e−07	3.9e+01	1.4e−10	4.7e−05
$c_s^{[2]}$	6.5e−03	9.0e+02	2.0e−08	2.3e+02
$c_d^{[1]}$	6.3e−05	1.6e+01	4.4e−08	9.2e−03
$c_d^{[2]}$	1.4e+10	1.0e+01	1.1e+07	1.3e+12

Während die Erwartung für das einfach genaue Programm eintrat, befand sich das doppelt genaue Programm zum Zeitpunkt des Abbruchs gerade auf dem Übergang von linearer zu quadratischer Konvergenz, wie ein Vergleich der Werte $c_d^{[1]}$ und $c_d^{[2]}$ mit der Tabelle für $\|E\|_2 = 0$ zeigt.

Schließlich sei $\|E\|_2 = 1/2$. Diese Matrizen haben mehrfache Eigenwerte. Wir erwarten daher bei `sjssif` und `djssif` asymptotisch quadratische Konvergenz. Die erzielten Ergebnisse entsprechen den Erwartungen. Auch die größten Werte sind noch moderat.

Konvergenzverhalten von `qjssif` (80 Tests)				
	μ_g	σ_g	min	max
$c_s^{[1]}$	5.1e−06	2.5e+01	3.4e−10	8.5e−03
$c_s^{[2]}$	5.6e−01	6.9e+01	2.3e−03	6.9e+02
$c_d^{[1]}$	6.0e−13	3.6e+02	1.5e−18	1.8e−08
$c_d^{[2]}$	1.8e−01	1.7e+02	4.0e−08	3.9e+03

Matrizen mit vorbestimmter Kondition von $|S|_S$

Diese Matrizen werden genauso erzeugt wie die entsprechenden Matrizen des letzten Abschnittes (s. Seite 189). Weil die Skalierung belanglos ist, beschränken wir uns hier auf den Fall $\delta = \eta = 2$.

Die Matrizen haben mehrfache Eigenwerte. Wir erwarten daher stets asymptotisch quadratische Konvergenz. Die Resultate, zusammengefaßt für alle 80 Testmatrizen der Ordnung $n \in \{10, 25, 50, 100\}$, spiegeln die Erwartung wider:

Konvergenzverhalten von `qjssif` (80 Tests)				
	μ_g	σ_g	min	max
$c_s^{[1]}$	5.7e−07	7.3e+01	5.5e−14	1.8e−03
$c_s^{[2]}$	1.5e−02	2.2e+01	9.9e−09	5.6e+01
$c_d^{[1]}$	1.8e−13	2.2e+02	3.6e−19	2.4e−09
$c_d^{[2]}$	2.9e−03	1.9e+02	7.7e−09	3.8e+01

Symbolverzeichnis

Symbol	Bemerkung	erstmals verwendet auf Seite
$X \oplus Y$	direkte Summe der Matrizen X, Y	11
$\bigoplus_{i=1}^{n} X_i$	direkte Summe der Matrizen $X_i, i = 1 \ldots n$	15
$\lambda_{\max}(X)$	größter Eigenwert von X	92
$\lambda_{\min}(X)$	kleinster Eigenwert von X	13
$\sigma_{\min}(X)$	kleinster Singulärwert von X	13
$0, (0_n)$	Nullmatrix (der Ordnung n)	49
$I, (I_n)$	Einheitsmatrix (der Ordnung n)	15
J_n	$J_n = \bigoplus_{i=1}^{n} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$	15
J	$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$, I von geeigneter Ordnung	4
$\lVert \cdot \rVert_2$, $\lVert \cdot \rVert_E$	Spektralnorm, Euklidische Norm	11, 11
$\lvert X \rvert$	Matrix $(\lvert X \rvert)_{ij} = \lvert X_{ij} \rvert$	159
$\lfloor X \rfloor$	$\lfloor X \rfloor = \sqrt{XX^*}$	10
$\det(X)$	Determinante der Matrix X	41
exp	Exponentialfunktion	16
diag	Diagonalmatrix	21
$\mathrm{diag}(x_i)$	$A = \mathrm{diag}(x_i) \Longleftrightarrow A_{ij} = x_i$, falls $i = j$ und $A_{ij} = 0$ sonst	13
$\mathrm{diag}(X)$	$A = \mathrm{diag}(X) \Longleftrightarrow A_{ij} = X_{ij}$, falls $i = j$ und $A_{ij} = 0$ sonst	34
$\mathrm{fl}(\cdot)$	Fließpunktergebnis einer Operation	101
$\mathcal{H}(X)$	Hadamardsches Maß	77
$X_{\cdot i}$ $(X_{i\cdot})$	i-ter Spaltenvektor (Zeilenvektor) von X	13
$[X_{\cdot 1} \cdots X_{\cdot n}]$	Matrix bestehend aus den Vektoren $X_{\cdot i}$	57
$\lceil x \rceil$	kleinste natürliche Zahl $\geq x$	106
$\prod_{i=1}^{n} X_i$	Produkt von Zahlen oder Matrizen X_i für wachsende i	77
$\kappa(A)$	Kondition $\kappa(A) = \lVert A \rVert_2 \lVert A^{-1} \rVert_2$	13
$\mathrm{rg}(X)$	Rang von X	102
$\mathrm{Tr}\ (X)$	Spur der Matrix X	31
X_S	skalierte Matrix: $X_S = D^{-1} X D^{-1}, D = \mathrm{diag}(X_{ii}^{1/2})$	13

Literaturverzeichnis

[1] Anda, A.A.; Park, H.: Fast Plane Rotation Algorithms with Dynamic Scaling; Computer Science Department, University of Mennesity, Mineapolis, MN 55455

[2] Anderson, E.; Bai, Z.; et al.: LAPACK User's Guide, Society for Industrial and Applied Mathematics (SIAM) Publication, Philadelphia

[3] ANSI/IEEE, New York. IEEE Standard for Binary Floating Point Arithmetik. Ausgabe 754-1985, 1985

[4] Barlow, J.; Demmel, J.: Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices, SIAM J. Numer. Anal., Vol. 27, No. 3, June 1990, pp. 762–791

[5] Brent, R.P.; Luk, F.T.: The Solution of Singular-Value and Symmetric Eigenvalue Problems on Multiprocessor Arrays, Dept. of Computer Science, Cornell University Ithaca, New York, 1983

[6] Bunch, James R.: A Note on the Stable Decomposition od Skew-Symmetric Matrices, Mathematics of Computation, Vol. 38, No. 158, April 1982, pp. 475–479

[7] Bunch, J.R.; Parlett, B.N.: Direct Methods for Solving Symmetric Indefinite Systems of Linear Equations, SIAM J. Numer. Anal., Vol. 8, No. 4, Dec. 1971, pp. 639–655

[8] Bunse-Gerstner, Angelika: Symplectic QR-like Methods; Habilitationsschrift, Universität Bielefeld 1986

[9] Caprani, O.: Roundoff Error in Floating-Point Summation, BIT 15 (1975), pp. 5–9

[10] Deichmöller, Anette: Über die Berechnung verallgemeinerter singulärer Werte mittels Jacobi-ähnlicher Verfahren; Dissertation, Fernuniversität Hagen, 1991

[11] Demmel, J.; Gragg, W.: On computing accurate singular values and eigenvalues of acyclic matrices; unveröffentlichtes Manuskript, zum Erscheinen in Linear Algebra and its Applications vorgesehen

[12] Demmel, J.; Kahan, W.: Accurate Singular Values of Bidiagonal Matrices, SIAM Journal of Scientific and Statistical Computing, Vol. 11, No. 5, Sept. 1990, pp.873–912

[13] Demmel, J.; Veselić, K.: Jacobi's method is more accurate than QR, SIAM Journal on Matrix Analysis and its Applications, Vol. 13, No. 8, pp. 1204–1245, Okt. 1992

[14] Dongarra, J.; Du Croz, J.; Duff, I.S.; Hammarling, S.: A Set of Level 3 Basic Linear Algebra Subprograms, Argonne National Laboratory Report, ANL-MCS-TM-88

[15] Dongarra, J.; Du Croz, J.; Hammarling, S.; Hanson, R.J.: An Extended Set of FORTRAN Basic Linear Algebra Subprograms, ACM Transactions on Mathematical Software, 14, pp. 1–17

[16] Drmač, Z.; Hari, V.: On Quadratic Convergence Bounds for the J-symmetric Jacobi Method, Fakultät für Mathematik, Universität Zagreb, Kroatien; zum Erscheinen in Numerische Mathematik vorgesehen

[17] Drmač, Z.; Omladić, M.; Veselić, K.: On the Perturbation of the Cholesky Decomposition, Seminarberichte aus dem Fachbereich Mathematik, Fernuniversität Hagen, Nr. 44, 1992

[18] Golub, G.H.; van Loan, C.F.: Matrix Computations; Johns Hopkins University Press, Baltimore, 1991

[19] Hari,V.; Veselić, K.: On Jacobi Methods for Singular Value Decomposition; SIAM Journal on Scientific and Statistical Computing 8 (1987), pp. 741–754

[20] Horn, R.A.; Johnson, C.R.: Matrix Analysis; Cambridge University Press 1985

[21] Jacobi, C.G.J.: Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen; J. Reine Angew. Math., 30, pp. 51–94, 1846

[22] Jacobs, D.: The State of the Art in Numerical Analysis, Academic Press, London, New York, San Francisco 1977

[23] Lancaster,P.; Tismenetsky, M.: The Theory of Matrices; 2. Edition with Appl.; Academic Press, Orlando, London 1985

[24] Lawson, C.L.; Hanson, R.J.; Kincaid, D.R.; Krogh, F.T.: Basic Linear Algebra Subprograms for FORTRAN Usage, ACM Transactions on Mathematical Software, 5, pp. 308–323

[25] Paardekooper, M.H.C.: An Eigenvalue Algorithm for Skew-Symmetric Matrices; Num. Math. 17, 1971, pp. 189–202

[26] Pietzsch, Eberhard: Jacobi-ähnliche Verfahren für Hamiltonsche Matrizen, Fernuniversität Hagen 1990, Diplomarbeit

[27] Sameh, Ahmed H.: On Jacobi an Jacobi-Like Algorithms for a Parallel Computer; Mathematics of Computation, Vol. 25, No. 115, July 1971, pp. 579–590

[28] Slapničar, Ivan: Accurate symmetric eigenreduction by a Jacobi Method, Fernuniversität, Semesterberichte, preprint, 1991

[29] Slapničar, Ivan: Accurate Symmetric Eigenreduction by a Jacoby Method, Dissertation, Fernuniversität Hagen, 1992

[30] Van der Sluis, A.: Condition Numbers and Equilibration of Matrices, Numerische Mathematik, Vol. 14, 1969, pp. 14–23

[31] Veselić, Krešimir: A Jacobi Eigenreduction Algorithm for Definite Matrix Pairs; Numerische Mathematik 64 (1993), pp. 241–269

[32] Veselić, K.; Hari, V.: A Note on a One-Sided Jacobi Algorithm; Numerische Mathematik 56 (1989), pp. 627–633

[33] Veselić, K.; Slapničar, I.: Floating-Point perturbations of Hermitian matrices, zum Erscheinen in Linear Algebra and its Applications vorgesehen

[34] Ward, R.C.; Gray, L.J.: Eigensystem Computation for Skew-Symmetric Matrices and a Class of Symmetric Matrices, ACM Transactions of Mathematical Software, Vol. 4, No. 3, Sept. 1978, pp. 278–285

[35] Wilkinson, J.H.: The Algebraic Eigenvalue Problem; Oxford University Press 1965

[36] Wilkinson, J.H.: Almost Diagonal Matrices with Multiple or Close Eigenvalues; Linear Algebra and its Applications, Nr.1, 1968, pp. 1–12

[37] Wilkinson, J.H.: Rundungsfehler; Springer, Berlin, Heidelberg, New York 1969

[38] Wilkinson, J.H.; Reinsch, C.: Handbook for Automatic Computation - Vol. 2: Linear Algebra; Springer, Berlin, Heidelberg, New York 1971