Divide-et-Impera-Verfahren für das verallgemeinerte Eigenwertproblem

Dissertation

zur
Erlangung des Grades eines Dr.rer.nat.
des Fachbereichs Mathematik
der Fernuniversität-Gesamthochschule-Hagen

vorgelegt von
Daniela Keller
aus Herdecke

Herdecke 1994

UNIVERSITÄTSBIBLIOTHEK STUTTGART

Berichterstatter: Prof. Dr. Kresimir Veselič
Prof. Dr. Hansjörg Linden

Tag der mündlichen Prüfung: 29. August 1994

Inhaltsverzeichnis

1. Einleitung 1
2. Zur Theorie des verallgemeinerten Eigenwertproblems 8
   2.1. Grundlegende Begriffe und Aussagen 8
   2.2. Vollständige Lösung von Eigenwertproblemen und simultane Diagonalisierung von Matrizenpaaren 16
   2.3. Das positiv definite verallgemeinerte Eigenwertproblem 25
3. Einige numerische Verfahren zum verallgemeinerten Eigenwertproblems . 29
   3.1. Potenzmethode, inverse Iteration und Rayleighquotientenverfahren . 29
   3.2. Rayleigh-artige Verfahren 30
      3.2.1. Die Rayleigh-Ritz-Approximation und die Unterraum-Iteration
      3.2.2. Koordinatenrelaxation
   3.3. Der QZ-Algorithmus 32
   3.4. Das verallgemeinerte Jacobi-Verfahren 33
   3.5. Weitere Verfahren 34
      3.5.1. Bisektionsverfahren
      3.5.2. Lanczos-Algorithmus
   3.6. Verfahren zur Bestimmung des Definitheitsintervalls 35
      3.6.1. Ein stochastisches Verfahren
      3.6.2. Ein Koordinatenrelaxationsverfahren zum Test auf positive Definitheit
4. Der Divide-et-Impera-Algorithmus für ein Matrizenpaar (A, J) mit Tridiagonalmatrix A 41
   4.1. Das eigentliche Verfahren 41
      4.1.1. Der Divide-Schritt
      4.1.2. Der Impera-Schritt
      4.1.3. Die Lösung des verallgemeinerten Eigenwertproblems $(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$
         4.1.3.1. Minuseigenwerte, Pluseigenwerte und das Definitheitsintervall
         4.1.3.2. Eigenschaften der Funktion w
         4.1.3.3. Deflation
         4.1.3.4. Bestimmung der Eigenvektoren von $(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$
      4.1.4. Die Zerlegung des Eigenwertproblems $A\mathbf{x} = \lambda J \mathbf{x}$
      4.1.5. Testen auf positive Definitheit

4.2. Die Nullstellensuche 59
4.2.1. Die Bestimmung der Nullstellen der Funktion w
4.2.2. Das quadratische Einschließungsverfahren
4.2.3. Praktisches Vorgehen
4.2.3.1. Ein lineares Einschließungsverfahren
4.2.3.2. Verschiebungsstrategien
4.3. J-Orthogonalität von Eigenvektoren und Aussagen zur Genauigkeit des Divide-et-Impera-Verfahrens 72
4.4. Ausblick: Divide-et-Impera-Verfahren für nichtsymmetrische gewöhnliche Eigenwertprobleme 78
5. Zur Reduktion der Matrix A auf Tridiagonalgestalt 81
5.1. Teilweise Reduktion auf Tridiagonalgestalt 81
5.1.1. Die modifizierte Householder-Transformationsmethode
5.2. Statistische Untersuchungen zur Reduktion von Matrizen auf Tridiagonalform mittels verallgemeinerter Householder-Transformationen 93
6. Divide-et-Impera-Techniken für das verallgemeinerte Eigenwertproblem $Ax = \lambda Jx$ 103
6.1. Der Fall einer berandeten Diagonalmatrix 104
6.1.1. Der Deflationsschritt
6.1.2. Berechnung der Eigenvektoren
6.1.3. Die Eigenwerte als Nullstellen der Funktion w
6.1.4. Das Einschließungsverfahren
6.1.4.1. Einschließung der Nullstellen von w, die im Intervall $(-\infty, \alpha_1)$ liegen
6.1.4.2. Berechnung der Schranken M_u und M_o
6.1.4.3. Ein lineares Eingrenzungsverfahren
6.2. Ein zusammengesetztes Divide-et-Impera-Verfahren für ineinandergeschachtelte berandete Tridiagonalmatrizen 119
Anhang A. 126
A.1. Numerische Ergebnisse 126
A.2. Überblick über die Computerprogramme 145
A.3. Literaturverzeichnis 148

Zusammenfassung

Bekanntlich hat das verallgemeinerte Eigenwertproblem $A\,x = \lambda B\,x$, wobei A und B reelle symmetrische n×n-Matrizen sind, zumindest dann n Eigenvektoren, wenn es positiv definit ist, d.h. für geeignete reelle Koeffizienten α, β die Linearkombination $\alpha A + \beta B$ positiv definit wird. Äquivalent dazu ist die Existenz einer (nichtsingulären) Eigenvektormatrix Q, die sowohl A als auch B auf Diagonalgestalt D_A bzw. D_B transformiert:

$$Q^T A\,Q = D_A, \qquad Q^T B\,Q = D_B$$

In dieser Arbeit wird angenommen, daß B die Gestalt einer Diagonalmatrix J hat, deren Diagonalelemente +1 oder −1 sind. (Dies ist für nichtsinguläres B durch eine Transformation immer erreichbar, die symmetrische Matrix A geht dabei gleichzeitig in eine symmetrische Matrix A_{mod} über.) Für eine Eigenvektormatrix Q gilt dann statt der üblichen Orthogonalität die Beziehung $Q^T J\,Q = J$.

Auf der Basis des für das Standardeigenwertproblem bekannten Divide-et-Impera-Algorithmus von Bunch, Nielsen, Sorensen u.a. für eine Tridiagonalmatrix und dem Verfahren von Wilkinson für eine berandete Diagonalmatrix werden zwei neue Algorithmen für die entsprechenden verallgemeinerten Eigenwertprobleme entwickelt. Die Nullstellensuche, die einen wesentlichen Teil in beiden Algorithmen darstellt, besteht aus einem speziell hierfür konstruierten neuen Eingrenzungsverfahren und dem bekannten Pegasus-Verfahren.

Mit einer Modifikation des üblichen Householderverfahrens zur Reduktion einer Matrix auf Tridiagonalgestalt wird das Paar (A_{mod}, J) auf die Form (T, J) transformiert: dabei ist T eine Tridiagonalmatrix, wenn während des Transformationsalgorithmus keine Division durch null oder durch eine betragsmäßig sehr kleine Zahl verlangt wird, sonst eine Ineinanderschachtelung von berandeten Tridiagonalmatrizen.

Durch eine Kombination unserer beiden Algorithmen ist es möglich, Paare (T, J) mit T Ineinanderschachtelung berandeter Tridiagonalmatrizen auf positive Definitheit zu untersuchen und gegebenenfalls alle Eigenwerte und Eigenvektoren zu berechnen.

Weiß man von vornherein noch nicht, ob ein Matrizenpaar (A, J) bzw. (T, J) positiv definit ist, so kann man die hier vorgestellten Algorithmen als Testverfahren verwenden; sie entdecken nicht positiv definite Paare im allgemeinen recht frühzeitig.

1. Einleitung

Unter einem *verallgemeinerten Eigenwertproblem* versteht man die Aufgabenstellung, zu quadratischen Matrizen A, B einen Vektor $\mathbf{x} \neq \mathbf{0}$ und eine Zahl λ zu finden, die der Beziehung

$$(1.1) \qquad \mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}$$

genügen. Der Vektor $\mathbf{x}$ heißt *Eigenvektor* zum *Eigenwert* λ.

Ist die Matrix B in der Eigenwertgleichung nichtsingulär, so läßt sich das verallgemeinerte Eigenwertproblem durch Invertieren von B theoretisch immer auf ein gewöhnliches zurückführen:

$$(1.2) \qquad \mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}$$

Abgesehen von dem Rechenaufwand, der für die Invertierung von B nötig ist, und den bei schlecht konditioniertem B zu erwartenden Rechenungenauigkeiten, sprechen auch andere Gründe gegen diesen Weg: Häufig sind A und B dünn besetzte Matrizen oder haben eine spezielle Struktur, z.B. Tridiagonalform; die invertierte Matrix $\mathbf{B}^{-1}$ - und damit auch $\mathbf{B}^{-1}\,\mathbf{A}$ - wird im allgemeinen aber vollbesetzt sein. Auch die Symmetrie von A und B wird nicht zu einer Symmetrie von $\mathbf{B}^{-1}\,\mathbf{A}$ führen. Ein Verfahren, das ohne Invertierung auskommt, ist daher für das verallgemeinerte Eigenwertproblem sehr wünschenswert.

Auf verallgemeinerte Eigenwertprobleme stößt man z.B. bei linearen, zeitinvarianten Systemen in der Physik und in der Regelungstechnik, wenn man die zugrundeliegenden linearen Differentialgleichungssysteme

$$(1.3) \qquad \mathbf{L}\mathbf{y} = \mathbf{R}\mathbf{y}$$

(L, R lineare Matrix-Differentialoperatoren bzgl. der Zeit t mit konstanten Koeffizienten) mit Hilfe des Exponentialansatzes $\mathbf{y} \equiv \mathbf{y}(t) = \mathbf{x}\,e^{\lambda t}$ zu lösen versucht ("übergedämpfte lineare Systeme").
Falls L und R nicht bereits Differentialoperatoren erster Ordnung sind, kann dies in der üblichen Art durch Einführung neuer Variablen für die Ableitungen $\dot{y}$, $\ddot{y}$, $\dddot{y}$, ... erreicht werden. Wir dürfen somit immer ein Differentialgleichungssystem der Gestalt

$$(1.4) \qquad \left[\,\mathbf{A}_1 + \mathbf{B}_1 \frac{d}{dt}\,\right]\mathbf{y}(t) = \left[\,\mathbf{A}_2 + \mathbf{B}_2 \frac{d}{dt}\,\right]\mathbf{y}(t)$$

voraussetzen. Dieses geht bei dem angeführten Exponentialansatz in die Matrixbeziehung

$$\left[\,\mathbf{A}_1 - \mathbf{A}_2\,\right]\mathbf{x} = \lambda\left[\,-\mathbf{B}_1 + \mathbf{B}_2\,\right]\mathbf{x}\,,$$

also in $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ mit $\mathbf{A} := \mathbf{A}_1 - \mathbf{A}_2$, $\mathbf{B} := -\mathbf{B}_1 + \mathbf{B}_2$ über.

Bei gekoppelten Systemen in der Mechanik kann B häufig als eine Verallgemeinerung des positiv definiten "Massenoperators"

$$(1.5) \qquad M = \begin{bmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_n \end{bmatrix}$$

interpretiert werden, wobei $m_1, \ldots, m_n$ die (natürlich positiven) Massen der n wechselwirkenden Teilchen bedeuten.

In der Regelungstechnik ist das zeitliche Verhalten linearer, zeitinvarianter Systeme in der Form

$$L\, y_{Zustand} = R\, y_{Eingang}$$

beschreibbar. Hängt die Eingangsgröße $y_{Eingang}$ durch eine Rückkopplung

$$y_{Eingang} = S\, y_{Zustand}$$

wieder von der Zustandsgröße $y_{Zustand}$ ab, so erhalten wir ein Differentialgleichungssystem

$$(1.6) \qquad L\, y_{Zustand} = RS\, y_{Zustand}$$

und damit ein verallgemeinertes Eigenwertproblem, wenn man mit dem Exponentialansatz in diese Differentialgleichung geht.

Ein weiteres Gebiet, in dem verallgemeinerte Eigenwertprobleme auftreten können, ist mit der Regelungstechnik verwandt, nämlich die Analyse elektrischer Netzwerke. Auch hier entstehen durch ohmsche, kapazitative und induktive Komponenten (näherungsweise) lineare Systeme, deren zeitliches Verhalten durch Differentialgleichungen der angegebenen Art beschrieben werden kann.

Es ist ziemlich einfach zu beweisen, daß das verallgemeinerte Eigenwertproblem $Ax = \lambda Bx$ im Fall eines positiv definiten, symmetrischen B und eines symmetrischen A n linear unabhängige Eigenvektoren hat. Eine etwas schwächere hinreichende Bedingung ist die positive Definitheit des *Paares* (A, B). (Das Paar (A, B) heißt positiv definit, wenn für geeignete reelle α, β die Linearkombination $\alpha A + \beta B$ eine positiv definite Matrix ist, vgl. Abschnitt 2.3 und Parlett [46], Golub, van Loan [29]).

Auch wenn die Bedingung der positiven Definitheit eines Matrizenpaares in der Praxis häufig erfüllt ist, so muß doch betont werden, daß sie keineswegs für die Lösbarkeit des zugehörigen Eigenwertproblems notwendig ist. Der QZ-Algorithmus z.B. kann auch auf den Fall eines nicht positiv definiten B angewendet werden, selbst die Symmetrie von A und B ist nicht notwendig. Nachteilig beim QZ-Algorithmus ist aber, daß von vornherein im Komplexen gearbeitet werden muß und auch im Fall eines positiv definiten Paares die Eigenvektoren nicht automatisch reell sind.

Für symmetrisches B kann ein verallgemeinertes Eigenwertproblem $Ax = \lambda Bx$ relativ schnell auf ein Problem der Gestalt $A_{mod}x = \lambda Jx$, J Diagonalmatrix mit Einsen, Nullen und Minuseinsen auf der Diagonalen, transformiert werden. Ist B zusätzlich nichtsingulär, so läßt sich $J = \mathrm{diag}(\pm 1)$ erreichen. (Ein Algorithmus, der *ohne* Lösung des Eigenwertproblems für B auskommt, wurde von Brebner und Gard [12] angegeben.) Für das Eigenwertproblem $Ax = \lambda Jx$ stellt das verallgemeinerte Jacobi-Verfahren (Veselič [63]) bei kleineren Dimensionen eine gute Möglichkeit dar, das Paar (A_{mod}, J) auf positive Definitheit zu testen und gegebenenfalls die Eigenwerte und Eigenvektoren zu berechnen. Der wesentliche Vorteil des Jacobi-Verfahrens ist seine Genauigkeit, ein Nachteil aber die geringe Geschwindigkeit (zumindest beim nicht-parallelen Jacobi-Algorithmus) und die Zerstörung einer speziellen Gestalt der Matrix A_{mod} (etwa Bänderstruktur oder dünn besetzt) im Verlauf der Rechnung.

In dieser Arbeit wird als erstes das von Bunch, Nielsen und Sorensen [13] für das normale Eigenwertproblem $Tx = \lambda x$ entwickelte Divide-et-Impera-Verfahren auf den Fall $Tx = \lambda Jx$ verallgemeinert. Wesentliche Voraussetzung in [13] wie zunächst auch bei uns ist die *Tridiagonalität* der Matrix C. Unser Verfahren bricht irgendwann ab - und zwar meist ziemlich früh - wenn das Paar (T, J) *nicht* positiv definit ist, andernfalls liefert es alle Eigenwerte und Eigenvektoren.

Beim Divide-et-Impera-Verfahren wird das Ausgangsproblem in niedrigerdimensionale Teilprobleme zerlegt, die man direkt lösen kann (*Divide-Schritt*). Aus diesen Lösungen wird dann schrittweise die Gesamtlösung zusammengesetzt (*Impera-Schritt*). Bei unserem Verfahren teilen wir das gegebene Eigenwertproblem solange auf, bis wir direkt lösbare Eigenwertprobleme der Dimension 2 oder 1 erhalten.

Während des Impera-Schrittes müssen wir Eigenwertprobleme lösen, bei denen die Matrix A eine Diagonalmatrix mit Rang-Eins-Störung ist. Ein solches Eigenwertproblem kann zu einem Nullstellenproblem für eine spezielle rationale Funktion w umgeschrieben werden: die Nullstellen von w sind gerade die gesuchten Eigenwerte.

Bunch, Nielsen und Sorensen haben in ihrer Arbeit [13] ein Nullstellenverfahren vorgestellt, das auf einer rationalen Approximation der Funktion w beruht. Die dabei verwendeten Funktionen sind wegen der speziellen Gestalt der Nullstellenfunktion im Fall J = I (streng) monoton, zwischen je zwei Polstellen von w liegt genau eine Nullstelle. Im Fall unseres verallgemeinerten Divide-et-Impera-Verfahrens kann diese Monotonieeigenschaft nicht mehr garantiert werden. Auch die Verteilung der Nullsstellen ist für $J \neq I$ anders: alle Nullstellen liegen zwischen dem kleinsten und größten Pol von w, und es gibt ein ausgezeichnetes In-

tervall, das sogenannte Definitheitsintervall, in dem im Fall eines positiv definiten Problems zwei Nullstellen enthalten sind. Das obige Nullstellenverfahren ist daher nicht übertragbar. Aus diesem Grund habe ich ein Eingrenzungsverfahren entwikkelt, das das Intervall, in dem die jeweilige Nullstelle oder die zwei Nullstellen liegen müssen, immer weiter verkleinert. Nicht selten ist nach den vier durchgeführten Eingrenzungsschritten (oder schon früher) die Nullstelle bereits gefunden, sonst hat man zumindest ein sehr gutes Startintervall für das Pegasus-Verfahren (eine verbesserte Variante der Regula Falsi), mit dem dann die Nullstelle abschließend berechnet wird. (Falls Nullstelle und Polstelle fast zusammenfallen, wird zusätzlich noch das Bisektionsverfahren als ein äußerst sicheres Verfahren herangezogen.) Dieses Eingrenzungsverfahren nutzt die Tatsache aus, daß w eine sehr spezielle rationale Funktion ist, bei der man genau weiß, wieviele Nullstellen in jedem Intervall liegen - leicht modifiziert, eignet sich das Verfahren aber auch zur Bestimmung von Nullstellen einer beliebigen rationalen Funktion.

Der Fall einer Tridiagonalmatrix A ist nicht nur deshalb interessant, weil solche Matrizen z.B. bei der Diskretisierung von Differentialoperatoren erster oder zweiter Ordnung entstehen, sondern auch, weil jede symmetrische Matrix durch sog. Householdertransformationen auf diese Gestalt gebracht werden kann. Diese Transformationen werden nun auf den Fall eines verallgemeinerten Eigenwertproblems so erweitert, daß die Matrix J (bis auf eventuelle Vertauschungen) unverändert bleibt (vgl. Bunse-Gerstner [15], Steinhoff [54]). Anders als beim normalen Eigenwertproblem kann es jedoch passieren, daß bei diesem Algorithmus durch null oder eine sehr kleine Zahl dividiert werden müßte. Ideen von Dax und Kaniel [20] aufgreifend, lassen wir dann solche kritischen Zeilen und Spalten stehen und machen mit der Transformation der restlichen Matrix weiter. Auf diese Weise bekommen wir statt einer Tridiagonalmatrix eine "Ineinanderschachtelung von berandeten Tridiagonalmatrizen".

Um ein Gefühl dafür zu bekommen, wie häufig Zeilen und Spalten bei der modifizierten Householdermethode *nicht* transformiert werden sollten, wurden hierzu statistische Untersuchungen durchgeführt. Im Einklang mit den angestellten theoretischen Überlegungen trat dieser Fall besonders dann häufig auf, wenn die Matrix J ungefähr gleich viele +1 und -1 hat, die auf der Diagonale einigermaßen gleichmäßig verteilt sind.

"Berandete *Diagonal*matrizen" hat Wilkinson in seinem Buch "The Algebraic Eigenvalue Problem" [68] behandelt. Die Lösung des Eigenwertproblems $Cx = \lambda x$ für eine symmetrische, berandete Diagonalmatrix C erfolgt formal ähnlich wie beim Divide-et-Impera-Verfahren: Nach vorausgegangener Deflation sind die Eigenwerte die Nullstellen einer speziellen rationalen Funktion. Die Eigenvektoren zu einfachen Eigenwerten können wie beim Divide-et-Impera-Verfahren direkt durch eine Formel angegeben werden.

Dieses Verfahren habe ich auf ein Problem der Form $Cx = \lambda Jx$ mit J gleich $\mathrm{diag}(\pm 1)$ verallgemeinert. Im Vergleich zum Standardverfahren von Wilkinson ist

die Berechnung der Eigenvektoren bei mehrfachen Eigenwerten etwas aufwendiger, da die berechneten Vektoren noch J-orthogonalisiert werden müssen. Die Probleme bei der Nullstellensuche sind vergleichbar mit denen beim verallgemeinerten Divide-et-Impera-Verfahren: auch hier kann ein Eingrenzungsverfahren entwickelt werden, das analoge Eigenschaften wie das bereits vorgestellte hat. Dieses verallgemeinerte Wilkinson-Verfahren läßt sich leicht auf den Fall erweitern, daß C eine Ineinanderschachtelung von berandeten Diagonalmatrizen ist; theoretisch könnte man es auch auf eine vollbesetzte Matrix C anwenden, aber gut geeignet ist es sicherlich nur für Matrizen mit wenigen Rändern.

Eine Kombination des verallgemeinerten Divide-et-Impera-Verfahrens und des verallgemeinerten Wilkinson-Verfahrens erlaubt jetzt die Lösung des Eigenwertproblems $A_{mod}x = \lambda Jx$, wobei A_{mod} eine Ineinanderschachtelung von berandeten Tridiagonalmatrizen ist. Zunächst werden Rang-Eins-Abspaltungen wie im Divide-Schritt des Divide-et-Impera-Verfahrens durchgeführt:

$$A_{mod} = A_T - \sum_i |b_i| u_i u_i^T \qquad (1.7)$$

Die Matrix A_T besteht aus entlang der Diagonalen ineinandergeschachtelten Blökken, die Tridiagonalmatrizen oder berandete Tridiagonalmatrizen sind, die Matrix J wird ebenfalls passend dazu in Blöcke aufgeteilt. Die Teileigenwertprobleme von (A_T, J) können für einen Tridiagonalblock direkt mit dem verallgemeinerten Divide-et-Impera-Verfahren gelöst werden; im Fall eines berandeten Tridiagonalblocks bestimmt man zunächst die vollständige Lösung des inneren verallgemeinerten Tridiagonaleigenwertproblems, dann gelangt man mittels geeigneter Transformationen auf ein verallgemeinertes Eigenwertproblem mit berandeter Diagonalmatrix. Dieses Problem löst man mit unserem modifizierten Wilkinson-Verfahren. Die Lösungen der Teileigenwertprobleme werden anschließend durch die Impera-Schritte miteinander verbunden, wodurch sich am Ende die Lösung des Gesamteigenwertproblems ergibt. Eine solche Lösung existiert sicherlich dann, wenn das Paar (A_{mod}, J) positiv definit ist; unser Verfahren stoppt, sobald ein Teilproblem diese Bedingung verletzt.

Das verallgemeinerte Divide-et-Impera-Verfahren liefert meist sehr gute Ergebnisse. Bei einigen sehr speziellen Matrizen (wie z.B. den sogenannten Wilkinson-Matrizen) kann es allerdings zu einem teilweisen Verlust der (J)-Orthogonalität der Eigenvektoren kommen. Dies ist ein schon für das Standard-Divide-Verfahren bekanntes Phänomen. (Es tritt auf, wenn zwei Nullstellen der rationalen Funktion w fast vollständig mit dem zwischen ihnen liegenden Pol übereinstimmen.) Mit dieser Schwierigkeit könnte man wahrscheinlich fertigwerden, indem man die folgenden, beim Standard-Divide-Verfahren bereits erprobten Modifikationen benutzt:

1. Kritische Terme in der Nullstellensuche und der Eigenvektorformel werden mit höherer Genauigkeit ausgewertet (vgl. Sorensen, Tang [53]).

2. Berechnung der Eigenvektoren mit inverser Iteration (vgl. Gates [27])

3. Bestimmung der Eigenvektoren mittels einer alternativen Formel (vgl. Gu und Eisenstat [32])

In dieser Arbeit wird nur kurz auf diese Verfahren eingegangen, in der aktuellen Implementierung des Algorithmus wird die Möglichkeit 1 durch Verwendung von Single/Double-Variablen getestet. Ein Vergleich der drei Lösungsansätze für unser verallgemeinertes Divide-et-Impera-Verfahren wäre sicher interessant.

Theoretische Fehlerabschätzungen und Genauigkeitsaussagen sind für das verallgemeinerte Divide-et-Impera-Verfahren noch nicht durchgeführt worden. Die bekannten Ergebnisse aus den Arbeiten von Bunch, Nielsen, Sorensen [13], Cuppen [19] und Sorensen, Tang [53] lassen sich nicht direkt übertragen: Die Eigenvektoren sind im Spezialfall $J = I$ orthogonal, während wir mit J-orthogonalen Vektoren arbeiten, deren euklidische Norm keineswegs 1 sein muß; desweiteren kann für $J = I$ die Ableitung der Nullstellenfunktion an den Nullstellen nach unten abgeschätzt werden, was für $J \neq I$ nicht mehr möglich ist, da die Ableitung der Nullstellenfunktion dort theoretisch beliebige Werte aus $\mathbb{R}$ annehmen kann. Gerade diese Abschätzung der Ableitung ist aber für die Beweise besonders wichtig.

Da wir in unserem Divide-et-Impera-Verfahren das Eigenwertproblerm so lange aufteilen, bis wir Teilprobleme der Dimension 1 oder 2 erhalten, ist das Nichterfülltsein der positiven Definitheit oft schon bei niedrigdimensionalen Teilproblemen erkennbar. Eine etwas andere - für eine rekursive Programmierung günstigere - Möglichkeit bestünde darin, nur so lange aufzuteilen, bis man Eigenwertprobleme hat, auf die sich z.B. das verallgemeinerte Jacobi-Verfahren (vgl. Veselič [63]) oder das QZ-Verfahren (vgl. Golub/van Loan [29]) anwenden lassen.

Unklar ist, ob und wie das verallgemeinerte Divide-et-Impera-Verfahren auf den Fall eines nicht positiv definiten Paares (T, J) verallgemeinerbar ist. -

Diese Arbeit baut sich folgendermaßen auf:

Im Kapitel 2 beschäftigen wir uns mit der theoretischen Lösbarkeit des verallgemeinerten Eigenwertproblems $A\,x = \lambda\, B\, x$ und leiten hierzu einige prinzipielle Ergebnisse ab. Die meisten dieser Aussagen dürften in etwa bekannt sein, sind aber so noch nicht zusammengefaßt worden.

Im Kapitel 3 stellen wir zunächst bekannte Verfahren zur Lösung des symmetrischen verallgemeinerten Eigenwertproblems $A\,x = \lambda\, B\, x$ vor; die meisten dieser Methoden verlangen die positive Definitheit von B. Das verallgemeinerte Jacobi-Verfahren dagegen kommt mit der positiven Definitheit des Paares (A, J) aus. Sehr universell ist das QZ-Verfahren, wenn es im Komplexen formuliert wird; es versucht, für beliebige Matrizenpaare (A, B) die Eigenwertbestimmung durchzuführen.

Unabhängig von dem Divide-et-Impera-Verfahren, das den Hauptteil dieser Arbeit bildet, wurden zwei einfachere Algorithmen entwickelt, die jedoch das Paar (A, B) nur auf positive Definitheit testen und fallweise ein μ aus $\mathbb{R}$ liefern, für das $A-\mu B$ positiv definit wird. Diese Algorithmen werden kurz in Abschnitt 3.6 skizziert. Der erste stellt eine Art Monte-Carlo-Optimierung dar, der zweite beruht auf einem Theorem von Kovač-Veselič [38] über Spurminimierung. Kennt man ein solches μ, so kann man zu einem Paar (A, B') mit einem positiv definiten B' der Gestalt $B' := A - \mu B$ übergehen und das Eigenwertproblem $Ax = \lambda B'x$ mit einem der dafür geeigneten Verfahren lösen.

Im Kapitel 4 - das den Kern dieser Arbeit bildet - stellen wir ausführlich unser verallgemeinertes Divide-et-Impera-Verfahren für das Paar (T, J) - mit T Tridiagonalmatrix und $J = \mathrm{diag}(\pm 1)$ - vor.

Im Kapitel 5 beschreiben wir die (evtl. nur teilweise) Reduktion einer Matrix auf Tridiagonalgestalt mittels modifizierter Householder-Transformationen. Um ein Gefühl zu bekommen, wie häufig der Fall einer unvollständigen Reduktion auftritt, führen wir neben theoretischen statistischen Untersuchungen auch Simulationen unter Verwendung von Pseudozufallszahlen durch.

In Kapitel 6 wird zunächst ein Algorithmus zur Lösung des speziellen Eigenwertproblems $C\,x = \lambda\, J\, x$ - mit C berandete Diagonalmatrix - beschrieben. Eine Kombination dieses Algorithmus mit dem verallgemeinerten Divide-et-Impera-Verfahren aus Kapitel 4 erlaubt uns dann, Eigenwertprobleme $T\,x = \lambda\, J\, x$ auch dann zu lösen, wenn T eine Ineinanderschachtelung berandeter Tridiagonalmatrizen ist.

Ein Anhang bringt eine systematische Zusammenstellung und Diskussion aller gewonnenen Computer-Ergebnisse. Diese wurden auf PCs mit den INTEL-Prozessoren 80386/387 und 80486, auf SUN-RISC-Workstations und auf dem Siemens-Vektorrechner SNI S-600/20 erhalten. Die Programme sind in ihrer ursprünglichen Form in Turbo-Pascal (für große Matrizen ist der "Protected Mode" in Version 7.0 notwendig) geschrieben und wurden später in C übertragen. Die wesentlichen Teile des Turbo-Pascal-Programms werden in einem weiteren Anhang präsentiert.

2. Zur Theorie des verallgemeinerten Eigenwertproblems

2.1 Grundlegende Begriffe und Aussagen

Es wird geprüft, wie weit sich die bekannten Begriffe und Aussagen zum normalen Eigenwertproblem auf verallgemeinerte Eigenwertprobleme erweitern lassen. Das neue - beim normalen Eigenwertproblem nicht auftretende - Phänomen "singulärer Büschel" wird ausführlicher untersucht.

Im Hinblick auf die später nur im Reellen funktionierenden Algorithmen werden wir für das *verallgemeinerte Eigenwertproblem*

(2.1) $\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}$

in dieser Arbeit stets den Vektorraum $\mathbb{R}^n$ (über dem Körper $\mathbb{R}$ der reellen Zahlen) annehmen; der Eigenwert soll aus dem zugrundegelegten Körper, also bei uns reell sein. Die Übertragung auf den $\mathbb{C}^n$ über dem Körper $\mathbb{C}$ der komplexen Zahlen macht aber größtenteils keine Schwierigkeiten.
Anders als beim "normalen" Eigenwertproblem

(2.1') $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$,

das dem Fall B = I, I n-dimensionale Einheitsmatrix, entspricht, ist sogar eine Verallgemeinerung auf den Fall nichtquadratischer Matrizen möglich, denn die Gleichung (2.1) macht auch noch Sinn, wenn A und B (m, n)-Matrizen sind. Dies soll jedoch in dieser Arbeit nicht weiterverfolgt werden.

Natürlich kann im Fall einer nichtsingulären Matrix B das verallgemeinerte Eigenwertproblem theoretisch immer durch Multiplikation beider Seiten mit B^{-1} auf ein gewöhnliches zurückgeführt werden. Es wurde aber bereits betont, daß dies in der Regel nicht empfehlenwert ist: vor allem in der Mechanik sind die Matrizen A und B symmetrisch, die Matrix B^{-1} A wird jedoch normalerweise keinesfalls symmetrisch sein. Wie das folgende Lemma zeigt, sind (nichtsymmetrische) gewöhnliche Eigenwertprobleme sogar äquivalent zu verallgemeinerten Eigenwertproblemen mit nichtsingulärem B:

Lemma 2.1: *Jede (reelle) Matrix M kann als Produkt $M = AB^{-1}$ mit A, B symmetrisch und B invertierbar geschrieben werden.*

Beweis: (Eine Beweisskizze mit etwas anderen Überlegungen findet sich im Buch von Parlett [46])
Mit Hilfe einer reellen, nichtsingulären Transformationsmatrix F kann M auf Jordan-Normalform M_{Jordan} gebracht werden:

$M_{Jordan} = F^{-1} M F$ mit

$$(2.2)\quad M_{Jordan} = \begin{bmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_m \end{bmatrix} \quad \text{und} \quad J_r = \begin{bmatrix} \lambda_r & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_r \end{bmatrix} .$$

Ein "Jordankästchen" J_r der Dimension k läßt sich aber durch Vertauschen der Spalten in eine symmetrische Matrix verwandeln:

$$\begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & 1 & \\ & & \ddots & \ddots \\ & & \lambda & 1 \\ 0 & & & \lambda \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 & \cdots & 1 & \lambda \\ \vdots & \vdots & & \lambda & 0 \\ \vdots & 1 & & 0 & \vdots \\ 1 & \lambda & & \vdots & \vdots \\ \lambda & 0 & \cdots & 0 & 0 \end{bmatrix}$$

Diese Umkehrung der Spaltenreihenfolge kann aber auch durch eine Matrix-Operation erzeugt werden, denn es gilt für beliebige Vektoren $\mathbf{a}_1, \dots, \mathbf{a}_k \in \mathbb{R}^k$:

$$\left[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k\right] \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & & 1 & 0 \\ \vdots & & \iddots & & \vdots \\ 0 & 1 & & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} = \left[\mathbf{a}_k, \mathbf{a}_{k-1}, \dots, \mathbf{a}_1\right] .$$

Zusammensetzen der Kästchen zur ursprünglichen Matrix führt zu

$$M_{Jordan} = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix} = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{bmatrix} \begin{bmatrix} P_1 & & & \\ & P_2 & & \\ & & \ddots & \\ & & & P_m \end{bmatrix}$$

$$\text{mit} \quad A_r = \begin{bmatrix} 0 & & & 1 & \lambda_r \\ & & \iddots & \iddots & \\ & \iddots & \iddots & & \\ 1 & \iddots & & & 0 \\ \lambda_r & & & & \end{bmatrix} \quad \text{und} \quad P_r = \begin{bmatrix} 0 & & & & 1 \\ & & & \iddots & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & 0 \end{bmatrix} .$$

Die Matrix $P = \begin{bmatrix} P_1 & & 0 \\ & \ddots & \\ 0 & & P_r \end{bmatrix}$ ist symmetrisch und hat als Determinante den

Wert +1 oder −1, also existiert die Inverse.

Die Matrix $A = \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_r \end{bmatrix}$ ist ebenfalls symmetrisch.

Somit gilt insgesamt

$$M = F\, M_{Jordan}\, F^{-1} = F\, A\, P\, F^{-1} = F\, A\, F^T (F^T)^{-1}\, P\, F^{-1}$$

$$= A'\, B_1' \quad \text{mit} \quad A' = F\, A\, F^T \quad \text{und} \quad B_1' = (F^{-1})^T\, P\, F^{-1}.$$

Die Matrizen A', B_1' sind offensichtlich symmetrisch. Das gleiche gilt auch für die Inverse $B' := B_1'^{-1}$. Damit haben wir die angegebene Darstellung $M = A'\, B'^{-1}$ gewonnen.

Q.E.D.

Gehen wir bei der hier definierten Faktorisierung nicht von M, sondern von M^T aus, so erhalten wir eine entsprechende Darstellung in umgekehrter Reihenfolge:

$$M = (M^T)^T \equiv (A'\, B'^{-1})^T = (B'^{-1})^T\, A'^T = (B'^T)^{-1}\, A'^T \equiv B^{-1} A.$$

Ein (beliebiges nichtsymmetrisches) Eigenwertproblem $A\,\mathbf{x} = \lambda\,\mathbf{x}$ kann daher stets als ein *symmetrisches verallgemeinertes Eigenwertproblem* aufgefaßt werden:

$$(2.3) \qquad A\,\mathbf{x} = \lambda\,\mathbf{x} \iff A_2^{-1} A_1\,\mathbf{x} = \lambda\,\mathbf{x} \iff A_1\,\mathbf{x} = \lambda\, A_2\,\mathbf{x}$$

mit A_1, A_2 symmetrisch.

Aus dem Lemma (2.1) folgt auch sofort, daß wir bei einem verallgemeinerten Eigenwertproblem die Matrix auf der rechten Seite (oder auf der linken Seite) immer als symmetrisch annehmen dürfen:

$$(2.4) \qquad A\,\mathbf{x} = \lambda\, B\,\mathbf{x} \iff A\,\mathbf{x} = \lambda\, B_1^{-1} B_2\,\mathbf{x} \iff B_1\, A\,\mathbf{x} = \lambda\, B_2\,\mathbf{x},$$

wobei hier B_2 symmetrisch ist.

Ein über das nichtsymmetrische, gewöhnliche Eigenwertproblem hinausgehender Gesichtspunkt tritt in Erscheinung, falls die Matrix B auf der rechten Seite singulär ist: Dann ist auch die symmetrische Matrix B_2 singulär, und eine Diagonalisierung von B_2 durch eine orthogonale Matrix U führt (evtl. nach geeigneter Permutation) auf die Darstellung

$$B_2 := U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} U^T \quad \text{mit} \quad D = \operatorname{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda_1, \dots, \lambda_m \neq 0.$$

Durch Multiplikation dieser Diagonalmatrix mit

$$S_1 = \begin{bmatrix} \frac{1}{\sqrt{|\lambda_1|}} & & & 0 \\ & \ddots & & \\ & & \frac{1}{\sqrt{|\lambda_m|}} & \\ 0 & & & I_{n-m} \end{bmatrix}, \quad S_2 = \begin{bmatrix} \frac{\operatorname{sgn}(\lambda_1)}{\sqrt{|\lambda_1|}} & & & 0 \\ & \ddots & & \\ & & \frac{\operatorname{sgn}(\lambda_m)}{\sqrt{|\lambda_m|}} & \\ 0 & & & I_{n-m} \end{bmatrix},$$

von links bzw. von rechts - "sgn" bedeutet die übliche Vorzeichenfunktion - bekommen wir schließlich

$$A\,\mathbf{x} = \lambda\, B\,\mathbf{x} \iff B_1\, A\,\mathbf{x} = \lambda\, B_2\,\mathbf{x} \iff U^T B_1\, A\, (U\, U^T)\,\mathbf{x} = \lambda\, U^T B_2 (U\, U^T)\,\mathbf{x}$$

$$\iff U^T B_1\, A\, U\, U^T \mathbf{x} = \lambda \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} U^T \mathbf{x}$$

$$\iff \underbrace{S_1\, U^T B_1\, A\, U\, S_2}_{=:\,\tilde{A}}\, S_2^{-1}\, U^T \mathbf{x} = \lambda \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} S_2^{-1}\, U^T \mathbf{x}.$$

Da $S_2^{-1}\, U^T \mathbf{x}$ für $\mathbf{x} \in \mathbb{R}^n$ den ganzen $\mathbb{R}^n$ durchläuft, ist ein verallgemeinertes Eigenwertproblem immer einem der Form

$$(2.5) \qquad \tilde{A}\,\mathbf{x} = \lambda \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} \mathbf{x}$$

äquivalent, wobei das m gleich dem Rang von B ist.

Wir werden im nächsten Abschnitt sehen, daß für $m < n$ im allgemeinen nicht mehr n Eigenvektoren gefunden werden können.

Vom Standpunkt der Numerik ist bemerkenswert, daß der Fall eines singulären B häufig beseitigt werden kann. Für ein beliebiges σ gilt nämlich offensichtlich:

$$(2.6) \qquad A\,\mathbf{x} = \lambda\, B\,\mathbf{x} \iff (A + \lambda\sigma A)\,\mathbf{x} = \lambda\,(B + \sigma A)\,\mathbf{x}$$

$$\iff A\,\mathbf{x} = \lambda\, \frac{1}{1 + \lambda\sigma}\,(B + \sigma A)\,\mathbf{x}$$

Sofern nicht $B + \sigma A$ für *jedes* σ singulär ist, kann für ein geeignetes σ_0 die Matrix $(B + \sigma_0 A)$ verhältnismäßig risikolos hinsichtlich der Genauigkeit invertiert und das neue (äquivalente) Eigenwertproblem $A\,\mathbf{x} = \lambda\, B'\,\mathbf{x}$ betrachtet

werden. (Ist A invertierbar, so mag auch das Eigenwertproblem $\lambda^{-1}\,\mathbf{x} = A^{-1}\,B\,\mathbf{x}$ für $\lambda \neq 0$ untersucht werden. Der Eigenwert $\lambda = 0$ darf bei nichtsingulärem A natürlich nicht vorkommen!)

Faßt man die angestellten Überlegungen zusammen, so erkennt man, daß - anders als beim üblichen Eigenwertproblem - ein verallgemeinertes Eigenwertproblem meist in ein symmetrisches transformiert werden kann:
Falls $A + \sigma_0 B$ für irgendein σ_0 invertierbar ist, betrachtet man das nichtsinguläre Problem

$$A\,\mathbf{x} = \lambda\,B'\,\mathbf{x} \quad \text{mit} \quad B' = A + \sigma_0 B.$$

Dieses ist aber äquivalent zu $B'^{-1}\,A\,\mathbf{x} = \lambda\,\mathbf{x}$.
Die Matrix $M := B'^{-1}\,A$ kann nun nach Lemma 2.1 in der Form $M = B_1^{-1}\,A_1$ mit symmetrischen Matrizen A_1, B_1 geschrieben werden. Mit diesen geht das Eigenwertproblem in die Form $A_1\,\mathbf{x} = \lambda\,B_1\,\mathbf{x}$ eines *symmetrischen* verallgemeinerten Eigenwertproblems über.

Die entscheidende Bedingung, daß die Schar $A + \sigma B$ nicht für *alle* $\sigma \in \mathbb{R}$ singulär ist, kann man symmetrischer formulieren:
Ein Matrizen*paar* (A, B) heißt *singulär*, wenn das "Büschel" $\{\alpha A + \beta B \mid \alpha, \beta \in \mathbb{R}\}$ ausschließlich aus singulären Matrizen besteht, sozusagen ein *singuläres Büschel* ist, sonst heißt es nichtsingulär.
(Ist $\alpha A + \beta B$ invertierbar, so auch $A + \sigma B$ mit $\sigma := \beta/\alpha$, sofern $\alpha \neq 0$. Im Sonderfall $\alpha = 0$ muß B selbst invertierbar sein. Dann kann aber

$$P(\sigma) := \det\left(A + \sigma B\right) = \sigma^n\left[\det\left(\tfrac{1}{\sigma}A + B\right)\right],$$

das ein Polynom n-ten Grades darstellt, nicht identisch null sein, sonst würde dies auch für $P(\sigma)/\sigma^n = \det\left(\frac{1}{\sigma}A + B\right)$ und damit für den Limes $\sigma \to \infty$, also für $\det(B)$ gelten.)

Falls $\operatorname{Ker} A \cap \operatorname{Ker} B \neq 0$, so ist $B + \sigma A$ stets singulär. Da aber dieser gemeinsame Unterraum - zumindest bei symmetrischen Matrizen A, B - abspaltbar ist (s. später), erweist sich dieser Fall als nicht so kritisch. Leider kann es aber auch passieren, daß A und B nur den Nullvektor als gemeinsamen Kern haben, und trotzdem enthält die Matrizenschar

$$\{B + \sigma A \mid \sigma \in \mathbb{R}\} \quad (\text{bzw. allgemeiner das "Büschel" } \{\alpha A + \beta B \mid \alpha, \beta \in \mathbb{R}\})$$

keine nichtsinguläre Matrix, wie die folgenden Beispiele zeigen.
Im nichtsymmetrischen Fall ist dies bereits für n = 2 möglich:
Sei

$$A := \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \quad B := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Dann ist $\operatorname{Ker} A = \left\{ \lambda\begin{pmatrix} 1 \\ 1 \end{pmatrix} \,\middle|\, \lambda \in \mathbb{R} \right\}$ und $\operatorname{Ker} B = \left\{ \lambda\begin{pmatrix} 0 \\ 1 \end{pmatrix} \,\middle|\, \lambda \in \mathbb{R} \right\}$,

also $\operatorname{Ker} A \cap \operatorname{Ker} B = \mathbf{0}$. Andererseits gilt

$$\det(A + \sigma B) = \begin{vmatrix} 1+\sigma & -1 \\ 0 & 0 \end{vmatrix} = 0,$$

d.h. $A + \sigma B$ ist für alle σ singulär.

Im symmetrischen Fall liefern die folgenden (3x3)-Matrizen ein Beispiel für den gleichen Sachverhalt:

$$A := \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \qquad B := \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Wie man schnell nachprüft, gilt nämlich

$$\operatorname{Ker} A = \left\{ \lambda\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \,\middle|\, \lambda \in \mathbb{R} \right\}, \quad \operatorname{Ker} B = \left\{ \lambda\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \,\middle|\, \lambda \in \mathbb{R} \right\}$$

und damit $\operatorname{Ker} A \cap \operatorname{Ker} B = \mathbf{0}$.
Wegen

$$\det(A + \sigma B) = \begin{vmatrix} 1+\sigma & 2 & 1 \\ 2 & 3-\sigma & 1 \\ 1 & 1 & 0 \end{vmatrix} = 1\cdot(2 - 3 + \sigma) - 1\cdot(1 + \sigma - 2) = 0$$

ist aber $A + \sigma B$ immer singulär. -

Das Phänomen einer stets singulären Matrizenschar $A + \sigma B$ ist auch aus einem anderen Grund interessant:
Wie im Fall des gewöhnlichen Eigenwertproblems kann das *charakteristische Polynom*

$$(2.7) \qquad P(\lambda) := \det(A - \lambda B)$$

(mit B statt der Einheitsmatrix I) zur Lösung des Eigenwertproblems verwendet werden. Eine Zahl λ ist genau dann Eigenwert, wenn sie Nullstelle des charakteristischen Polynoms ist.
Der Grad des charakteristischen Polynoms muß jedoch nicht mehr gleich n sein:

Satz 2.2: *Für beliebige quadratische Matrizen A, B gilt: grad P ≤ Rang(B)*

Beweis: Wir haben in Lemma 2.1 gesehen, daß wir $B = B_1^{-1}\,B_2$ schreiben und B_2 durch nichtsinguläre Matrizen F_1, F_2 auf die Blockdiagonalgestalt

$$\operatorname{diag}(I_m, 0) \text{ mit } m = \operatorname{Rang}(B_2) = \operatorname{Rang}(B)$$

bringen können.

Also gilt:

$$\begin{aligned}\det(A - \lambda B) &= \left(\det(B_1)\right)^{-1} \det(B_1 A - \lambda B_2) \\ &= \left(\det(B_1)\right)^{-1} \det\Big(F_1^{-1} \big(\underbrace{F_1 B_1 A F_2^{-1}}_{=:A'} - \lambda \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}\big) F_2\Big) \\ &= \left(\det(B_1)\right)^{-1} \left(\det(F_1)\right)^{-1} \det(F_2) \det\Big(A' - \lambda \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}\Big) .\end{aligned}$$

Aus

$$\begin{vmatrix} a'_{11} - \lambda & a'_{12} & \cdots & a'_{1n} \\ & \ddots & & \\ \vdots & & a'_{mm} - \lambda & \vdots \\ & & & a'_{m+1,m+1} \\ a'_{n1} & a'_{n2} & \cdots & \ddots \; a'_{nn} \end{vmatrix} = (-\lambda)^m a'_{m+1,m+1} * \ldots * a'_{nn} + O(\lambda^{m-1})$$

liest man sofort die Richtigkeit der Aussage ab.

Q.E.D.

Nach den vorhergehenden Gegenbeispielen kann es also passieren, daß das charakteristische Polynom identisch verschwindet und damit jedes λ Eigenwert ist. Es ist aber auch der Fall nicht ausgeschlossen, daß P gleich einer nichtverschwindenden Konstante ist: in diesem Fall existiert überhaupt kein Eigenwert!

Nach diesen doch etwas überraschenden Fakten kommen wir jetzt zu einigen Eigenschaften, die bereits vom üblichen Eigenwertproblem bekannt sind.

Satz 2.3: *Es seien $\mathbf{x}_1, \ldots, \mathbf{x}_r$ Eigenvektoren zu verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_r$. Falls Ker B = **0** gilt, sind $\mathbf{x}_1, \ldots, \mathbf{x}_r$ linear unabhängig.*

Beweis:
Es seien (o. E.) $\mathbf{x}_1, \ldots, \mathbf{x}_k$ $(k < r)$ linear unabhängig, aber $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$ linear abhängig. Dann gilt

$$\mathbf{x}_{k+1} = \alpha_1 \mathbf{x}_1 + \ldots + \alpha_k \mathbf{x}_k .$$

Wenden wir auf beide Seiten den Operator $(A - \lambda_{k+1} B)$ an, so erhalten wir

$$\begin{aligned}\mathbf{0} &= \alpha_1 \left(A - (\lambda_1 + \lambda_{k+1} - \lambda_1) B\right) \mathbf{x}_1 + \ldots + \alpha_k \left(A - (\lambda_k + \lambda_{k+1} - \lambda_k) B\right) \mathbf{x}_k \\ &= \alpha_1 (\lambda_1 - \lambda_{k+1}) B \mathbf{x}_1 + \ldots + \alpha_k (\lambda_k - \lambda_{k+1}) B \mathbf{x}_k \\ &= B \left[\alpha_1 (\lambda_1 - \lambda_{k+1}) \mathbf{x}_1 + \ldots + \alpha_k (\lambda_k - \lambda_{k+1}) \mathbf{x}_k\right] .\end{aligned}$$

Da nach Voraussetzung der Kern von B nur aus dem Nullvektor besteht, gilt

$$\alpha_1 (\lambda_1 - \lambda_{k+1}) \mathbf{x}_1 + \ldots + \alpha_k (\lambda_k - \lambda_{k+1}) \mathbf{x}_k = \mathbf{0} .$$

Die Faktoren $(\lambda_i - \lambda_{k+1})$, $i = 1, \ldots, k$ sind ungleich null, mindestens ein α-Koeffizient ist ebenfalls ungleich null, also sind $\mathbf{x}_1, \ldots, \mathbf{x}_k$ linear abhängig im Gegensatz zur Annahme.

Q.E.D.

Falls B einen nichttrivialen Kern hat, ist es dagegen möglich, daß ein und derselbe Vektor Eigenvektor zu *unendlich vielen* Eigenwerten ist:
Wenn nämlich Ker A $\cap$ Ker B $\neq$ **0** ist, so gilt für jedes nichtverschwindende $\mathbf{x}$ aus Ker A $\cap$ Ker B

$$A\mathbf{x} = \mathbf{0} \quad \text{und} \quad B\mathbf{x} = \mathbf{0}, \quad \text{d.h. } A\mathbf{x} = \lambda B \mathbf{x} \quad \text{für } \textit{alle } \lambda \in \mathbb{R} \text{ (oder } \mathbb{C}).$$

Vom üblichen *symmetrischen* Eigenwertproblem weiß man, daß Eigenvektoren zu verschiedenen Eigenwerten nicht nur linear unabhängig, sondern darüber hinaus auch orthogonal sind. Die Übertragung auf verallgemeinerte symmetrische Eigenwertprobleme lautet:

Satz 2.4: *Zwei Eigenvektoren $\mathbf{x}_1$, $\mathbf{x}_2$ zu zwei verschiedenen Eigenwerten des symmetrischen Eigenwertproblems $A x = \lambda B \mathbf{x}$ sind "B-orthogonal", d.h. es gilt*

$$(\mathbf{x}_1, \mathbf{x}_2)_B = 0 ,$$

wobei $(\mathbf{x}, \mathbf{y})_B$ das (im allgemeinen) nicht positiv definite "Skalarprodukt" $\mathbf{x}^T B \mathbf{y} \equiv (\mathbf{x}, B\mathbf{y})$ bedeutet.

Beweis:
Aus

$$(\mathbf{x}_1, A\mathbf{x}_2) = (\mathbf{x}_1, \lambda_2 B \mathbf{x}_2) = \lambda_2 (\mathbf{x}_1, B\mathbf{x}_2)$$

und

$$(\mathbf{x}_1, A\mathbf{x}_2) = (A^T \mathbf{x}_1, \mathbf{x}_2) = (A\mathbf{x}_1, \mathbf{x}_2) = (\lambda_1 B \mathbf{x}_1, \mathbf{x}_2) = \lambda_1 (\mathbf{x}_1, B\mathbf{x}_2)$$

folgt durch Subtraktion

$$(\lambda_2 - \lambda_1)(\mathbf{x}_1, B\mathbf{x}_2) = 0, \text{ also } (x_1, B\mathbf{x}_2) = 0. \qquad \text{Q.E.D.}$$

Da - wie bereits erwähnt - ein nichtsymmetrisches verallgemeinertes Eigenwertproblem meist auf ein symmetrisches zurückgeführt werden kann, ist die Aussage von Satz 2.4 recht nützlich.

2.2 Vollständige Lösung von Eigenwertproblemen und simultane Diagonalisierung von Matrixpaaren

Es wird gezeigt, daß die für das gewöhnliche Eigenwertproblem gültige Äquivalenz zwischen vollständiger Lösung des Eigenwertproblems und Diagonalisierbarkeit der Matrix beim verallgemeinerten Eigenwertproblem nur unter Zusatzvoraussetzungen - wie Nichtsingularität von B oder Symmetrie von A und B - richtig ist.
Für die theoretische Formulierung erweist sich der Begriff der Diagonalisierbarkeit als günstig.

Beim gewöhnlichen Eigenwertproblem $A\mathbf{x} = \lambda\mathbf{x}$ im $\mathbb{R}^n$ ist derjenige Fall besonders interessant, in dem n linear unabhängige Eigenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ existieren. Faßt man diese n (Spalten-) Vektoren zu einer Matrix X zusammen und bildet mit den Eigenwerten $\lambda_1, \dots, \lambda_n$ eine Diagonalmatrix $\Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, so gilt

$$A\,X = X\,\Lambda.$$

Da die Matrix X nach Voraussetzung maximalen Rang hat, ist sie invertierbar, und es gilt

$$X^{-1}\,A\,X = \Lambda,$$

d.h. A kann durch die Matrizen $F := X^{-1}$, $G := X$ diagonalisiert werden.

Wir werden jetzt eine Übertragung auf das verallgemeinerte Eigenwertproblem versuchen.

Definition 2.5: *Das verallgemeinerte Eigenwertproblem $A\mathbf{x} = \lambda B\mathbf{x}$, A, B nxn-Matrizen, heiße vollständig lösbar, wenn es n linear unabhängige Eigenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ gibt.*

Da der Grad des charakteristischen Polynoms $P(\lambda) \equiv \det(A - \lambda B)$ kleiner als n sein kann, ist es für die vollständige Lösbarkeit des verallgemeinerten Eigenwertproblems nicht mehr ausreichend, wenn bei allen Eigenwerten geometrische und algebraische Vielfachheit übereinstimmen, d.h. kein Eigenwert defektiv ist. Deswegen wollen wir hier die Begriffe "defektiver Eigenwert" und "defektive Matrix" vermeiden.

Fassen wir die Spaltenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ zur "Eigenvektormatrix" X zusammen, so bedeutet vollständige Lösbarkeit das Bestehen der Matrix-Gleichung

$$A\,X = B\,X\,\Lambda, \tag{2.8}$$

wobei $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ die "Eigenwertmatrix" darstellt.

Ist A symmetrisch, so sind die Eigenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ beim Standardeigenwertproblem orthogonal bzw. können im Entartungsfall orthogonal gewählt werden; nach Normierung der Eigenvektoren wird X zu einer orthogonalen Matrix, weshalb wir auch

$$X^T\,A\,X = \Lambda$$

schreiben dürfen: A ist der Diagonalmatrix Λ "unitär" (orthogonal)-äquivalent.

Die Orthogonalitätseigenschaft der Eigenvektormatrix beim (gewöhnlichen) symmetrischen Eigenwertproblem kann in folgender Weise auf das verallgemeinerte Eigenwertproblem übertragen werden:

Wir nehmen dabei zur Vereinfachung an, daß B nichtsingulär ist und damit auf eine Diagonalmatrix J transformiert werden kann, deren Diagonalelemente +1 oder -1 sind, so daß $J^2 = I$ erfüllt ist. Für Eigenvektoren $\mathbf{x}_i$, $\mathbf{x}_j$ zu verschiedenen Eigenwerten λ_i, λ_j gilt

$$(\mathbf{x}_i, J\mathbf{x}_j) \equiv \mathbf{x}_i^T J\,\mathbf{x}_j = 0.$$

Für Eigenvektoren zu gleichen Eigenwerten können wir - ähnlich wie beim Gram-Schmidt-Verfahren - eine Linearkombination verwenden, um J-Orthogonalität zu erreichen. Da nach dem Sylvesterschen Trägheitssatz die Anzahl der Plus - und Minuseinsen gleich bleibt, gilt

$$(\mathbf{x}_i, J\mathbf{x}_i) \equiv \mathbf{x}_i^T J\,\mathbf{x}_i > 0 \quad \text{oder} \quad < 0$$

und nach eventueller Permutation und Normierung für die Eigenvektormatrix X:

$$X^T J\,X = J, \text{ woraus durch Multiplikation } J\,X^T J\,X = J^2 = I \tag{2.9}$$

folgt. Die nichtsinguläre Matrix $X^* := J X^T J$ ist also invers zu X oder anders ausgedrückt, X ist *"J-orthogonal"* :

$$\langle \mathbf{u}, X\mathbf{v} \rangle \equiv \mathbf{u}^T J\,X\mathbf{v} = \mathbf{u}^T J\,X J\,J\mathbf{v} = (J X^T J\,\mathbf{u})^T J\,\mathbf{v} = \langle (J X^T J)\mathbf{u}, \mathbf{v} \rangle = \langle X^*\mathbf{u}, \mathbf{v} \rangle$$

Wie beim nichtsymmetrischen Standardeigenwertproblem kann eine praktikable hinreichende Bedingung für die vollständige Lösbarkeit des verallgemeinerten Eigenwertproblems nicht gegeben werden, wobei aus den bekannten Gründen die Symmetrie nicht mehr entscheidend ist. (Natürlich gilt, daß im Fall einer nichtsingulären Matrix B n unterschiedliche Nullstellen des charakteristischen Polynoms zwangsläufig zu n linear unabhängigen Eigenvektoren führen müssen, dies läßt sich jedoch nicht direkt aus den Matrizen A und B ablesen!)

Offensichtlich gilt aber der folgende einfache und wichtige Sachverhalt:

Lemma 2.6: *Notwendig für die vollständige Lösung des verallgemeinerten Eigenwertproblems ist das Erfülltsein der folgenden Bedingung:*

$$\mathit{Bild}(A) \subseteq \mathit{Bild}(B)$$

Beweis:

Aus $A = B\,X\,\Lambda\,X^{-1}$ folgt

$$\mathrm{Bild}(A) = \mathrm{Bild}(B\,X\,\Lambda\,X^{-1}) \equiv \{B\,X\,\Lambda\,X^{-1}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} \subseteq \{B\,\mathbf{y} \mid \mathbf{y} \in \mathbb{R}^n\}$$
$$= \mathrm{Bild}(B).$$

Q.E.D.

Der Diagonalisierungsbegriff bei einer Matrix wird nun folgendermaßen verallgemeinert:

Definition 2.7: *Zwei (n,n)-Matrizen A, B heißen simultan diagonalisierbar, wenn es nichtsinguläre (n,n)-Matrizen F, G gibt, so daß*

$$F\,A\,G \quad \text{und} \quad F\,B\,G$$

Diagonalmatrizen werden.

Für numerische Zwecke ist dieser Begriff im allgemeinen weniger nützlich als der der vollständigen Lösung des Eigenwertproblems. Er ist aber - wie wir gleich sehen werden - wegen der symmetrischen Rolle von A und B eleganter und außerdem auch auf beliebig viele Matrizen erweiterbar.

Ist es möglich, A und B gemeinsam auf Diagonalgestalt zu transformieren, so kann man - wie in Abschnitt 2.1 erläutert wurde - durch eine weitere Transformation erreichen, daß B (oder A) die folgende einfache Form annimmt:

$$\begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}$$

Häufig schränkt man die Klasse der zugelassenen Transformationen ein:
Eine mögliche Einschränkung stellt die Bedingung $F = G^{-1}$ dar. Eine andere ist bei symmetrischen Matrizen naheliegend: $F = G^T$. In diesem Fall liefern die Transformationen

$$A \longrightarrow G^T A\,G, \qquad B \longrightarrow G^T B\,G$$

wieder symmetrische Matrizen. (Beim üblichen symmetrischen Eigenwertproblem sind beide Bedingungen erfüllt, da F orthogonal ist; dies wird beim verallgemeinerten symmetrischen Eigenwertproblem aber wegen der "B-Orthogonalität" von X nicht mehr richtig sein.)

In den wichtigsten Fällen liefert eine vollständige Lösung des Eigenwertproblems Transformationsmatrizen F, G zur simultanen Diagonalisierbarkeit.

Satz 2.8: *Es sei B nichtsingulär und $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ eine vollständige Lösung des Eigenwertproblems $A\,\boldsymbol{x} = \lambda\,B\,\boldsymbol{x}$. Dann bewirkt*

$$F := X^{-1} B^{-1}, \quad G := X$$

eine simultane Diagonalisierung von A und B.

Beweis: Unter Verwendung von $A\,X = B\,X\,\Lambda$ erhält man direkt:

$$F\,A\,G = X^{-1} B^{-1} A\,X = X^{-1} B^{-1} B\,X\,\Lambda = \Lambda$$

$$F\,B\,G = X^{-1} B^{-1} B\,X = I$$

A ist also in die Eigenwertmatrix Λ und B in die Einheitsmatrix transformiert worden. Q.E.D.

Wir betrachten als nächstes den wichtigen Fall symmetrischer Matrizen A, B, wobei B auch singulär sein darf:

Satz 2.9: *Es seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ eine vollständige Lösung des Eigenwertproblems $A\boldsymbol{x} = \lambda B\boldsymbol{x}$ und A, B symmetrisch. Dann gibt es eine Matrix W, so daß $W^T A\,W$ und $W^T B\,W$ diagonal sind.*

Beweis:
Da wir auch singuläre B zulassen, müssen wir anders als beim letzten Satz vorgehen:
Aus $A\,X = B\,X\,\Lambda$ folgt einerseits durch Multiplikation

$$X^T A\,X = X^T B\,X\,\Lambda,$$

andererseits durch Transposition $X^T A = \Lambda\,X^T B$ und damit

$$X^T A\,X = \Lambda\,X^T B\,X.$$

Dies bedeutet, daß $B' := X^T B\,X$ mit Λ vertauscht:

$$B'\,\Lambda = \Lambda\,B'$$

Außerdem ist B' symmetrisch. Nach einem bekannten Satz der Matrizentheorie existiert somit eine orthogonale Matrix U, die Λ *und* B' gleichzeitig auf Diagonalgestalt bringt:

$$U^T B'\,U = B_{diag}, \quad U^T \Lambda\,U = \Lambda_{diag}$$

(Λ bleibt natürlich bis auf Permutation der Eigenwerte erhalten!)

Für $A' := X^T A\,X$ gilt $A' = B'\,\Lambda$ und damit

$$U^T A'\,U = U^T B'\,U\,U^T \Lambda\,U = B_{diag}\,\Lambda_{diag}.$$

A' ist durch U also ebenfalls in Diagonalgestalt überführt worden.
Insgesamt haben wir für A und B die Transformationen

$$A \longrightarrow U^T X^T A\,X\,U = (X\,U)^T A\,(X\,U),$$

$$B \longrightarrow U^T X^T B\,X\,U = (X\,U)^T B\,(X\,U)$$

durchgeführt, die Matrix $W := X\,U$ leistet demgemäß das Gewünschte. Q.E.D.

Da X nur B-orthogonal, aber nicht orthogonal im üblichen Sinn ist, wird W^T - anders als beim Standardeigenwertproblem - im allgemeinen ungleich W^{-1} sein.

Satz 2.9 ist insofern allgemeiner als Satz 2.8, als wir für nichtsinguläres B gemäß den Überlegungen nach Lemma 2.1 auf ein symmetrisches verallgemeinertes Eigenwertproblem transformieren können.

Wenn weder die Voraussetzungen von Satz 2.8 (Nichtsingularität von B bzw. schwächer: Nichtsingularität des Büschels (A, B)) noch die von Satz 2.9 (Symmetrie von A und B) gegeben sind, dann kann es sehr wohl passieren, daß ein Paar (A, B) *nicht* simultan diagonalisierbar ist, *obwohl* das zugehörige Eigenwertproblem vollständig lösbar ist:

Gegenbeispiel:

Es sei

$$A := \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad B := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

also das Paar (A, B) ein singuläres Büschel.

Dann ist das Eigenwertproblem $A\,\mathbf{x} = \lambda\, B\,\mathbf{x}$ vollständig lösbar, denn für

$$X := \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}, \quad \Lambda := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

gilt

$$A\,X = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

und

$$B\,X\,\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

also $A\,X = B\,X\,\Lambda$.

Trotzdem ist (A, B) *nicht* simultan diagonalisierbar:
Um dies zu beweisen, nehmen wir an, es gäbe zwei nichtsinguläre Matrizen

$$F := \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \quad \text{und} \quad G := \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

mit der Eigenschaft, daß F A G und F B G diagonal sind.
Aus

$$F\,B\,G = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}\begin{bmatrix} g_{11} & g_{12} \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} f_{11}\,g_{11} & f_{11}\,g_{12} \\ f_{21}\,g_{11} & f_{21}\,g_{12} \end{bmatrix} \overset{!}{=} \begin{bmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{bmatrix}$$

liest man die folgenden Bedingungen ab:

(1) $f_{11}\,g_{12} = 0$ (2) $f_{21}\,g_{11} = 0$

Aus

$$F\,A\,G = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{11} \\ f_{21} & f_{21} \end{bmatrix}\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

$$= \begin{bmatrix} f_{11}(g_{11} + g_{21}) & f_{11}(g_{12} + g_{22}) \\ f_{21}(g_{11} + g_{21}) & f_{21}(g_{12} + g_{22}) \end{bmatrix} \overset{!}{=} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}$$

ergeben sich die weiteren Bedingungen:

(3) $f_{11}\,(g_{12} + g_{22}) = 0$ (4) $f_{21}\,(g_{11} + g_{21}) = 0$

Es sei zunächst $f_{11} \neq 0$. Dann gilt nach (1), daß $g_{12} = 0$ und damit nach (3) $f_{11}\,g_{22} = 0$, also $g_{22} = 0$. Dann hat G aber die Gestalt

$$G = \begin{bmatrix} * & 0 \\ * & 0 \end{bmatrix} \quad \text{und ist singulär.}$$

Es bleibt also der Fall $f_{11} = 0$ übrig. $f_{21} = 0$ würde zur Singularität von F führen, also muß $f_{21} \neq 0$ sein, was nach (2) $g_{11} = 0$ impliziert und nach (4) $g_{21} = 0$. Dann wäre G wiederum singulär. Q.E.D.

Die Frage, wann aus der simultanen Diagonalisierbarkeit die vollständige Lösung des Eigenwertproblems folgt, läßt sich restlos beantworten:

Satz 2.10: *Es seien A, B simultan diagonalisierbar.*
Das verallgemeinerte Eigenwertproblem $A\,\mathbf{x} = \lambda\, B\,\mathbf{x}$ ist genau dann vollständig lösbar, wenn Bild(A) $\subseteq$ Bild (B) gilt.

Beweis:

Die Notwendigkeit der Bedingung Bild(A) $\subseteq$ Bild (B) ist bereits in Lemma (2.6) festgestellt worden. Wir zeigen jetzt, daß sie hinreichend ist:
Es seien F, G die behaupteten Transformationsmatrizen. Dann gilt

$$F\,A\,G = \mathrm{diag}(a_1, \ldots, a_n), \quad F\,B\,G = \mathrm{diag}(b_1, \ldots, b_n).$$

Aus der Voraussetzung Bild(A) $\subseteq$ Bild(B) folgt Bild(F A G) $\subseteq$ Bild(F B G); dies bedeutet aber, daß das Diagonalelement a_i verschwinden muß, wenn das entsprechende b_i null ist.
Wir setzen jetzt

$$(2.10) \qquad \lambda_i = \begin{cases} \dfrac{a_i}{b_i} & \text{falls } b_i \neq 0 \\ \text{beliebig} & \text{sonst} \end{cases}$$

Damit gilt in jedem Fall $a_i = \lambda_i\, b_i$.

Die Vektoren

$$\mathbf{x}_i := G\,\mathbf{e}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ni} \end{bmatrix}, \quad i = 1, \dots, n$$

sind dann linear unabhängige Eigenvektoren des verallgemeinerten Eigenwertproblems:

$$\begin{aligned} A\,\mathbf{x}_i &= A\,G\,\mathbf{e}_i = F^{-1}\,F\,A\,G\,\mathbf{e}_i = F^{-1}\,\mathrm{diag}(a_1, \dots, a_n)\,\mathbf{e}_i = F^{-1}\,a_i\,\mathbf{e}_i \\ &= F^{-1}\,\lambda_i\,b_i\,\mathbf{e}_i = \lambda_i\,F^{-1}\,\mathrm{diag}(b_1, \dots, b_n)\,\mathbf{e}_i \\ &= \lambda_i\,F^{-1}\,\mathrm{diag}(b_1, \dots, b_n)\,G^{-1}\,G\,\mathbf{e}_i \\ &= \lambda_i\,B\,G\,\mathbf{e}_i = \lambda_i\,B\,\mathbf{x}_i. \end{aligned}$$

Q.E.D.

Aus dem letzten Satz ersieht man den Vorzug einer Formulierung mit der simultanen Diagonalisierbarkeit:

1. Dem Fall $a_i \neq 0$, $b_i = 0$ ("1/0") entspricht überhaupt kein echter Eigenwert, sondern nur formal ein "unendlich" großer. (Vom numerischen Gesichtspunkt wird daher der Fall $|b_i|$ sehr klein, $|a_i|$ nicht klein, kritisch sein.)
2. Dem Fall $a_i = 0$, $b_i = 0$ ("0/0") kann jede beliebige Zahl λ als i-ter Eigenwert zugeordnet werden.

Da ein nichtsinguläres B als Bild stets den gesamten Vektorraum hat, ist die Bedingung Bild(A) $\subseteq$ Bild(B) in Satz 2.10 trivialerweise erfüllt. Kombination mit Satz 2.8 liefert:

> *Falls B nichtsingulär ist (schwächer: das Paar (A, B) kein singuläres Büschel bildet), so kann das Eigenwertproblem* $A\mathbf{x} = \lambda B\mathbf{x}$ *genau dann vollständig gelöst werden, wenn die Matrizen A und B simultan diagonalisierbar sind.*

Dagegen reicht die Symmetrie von A und B allein nicht aus, um die Äquivalenz von vollständiger Lösbarbarkeit des Eigenwertproblems und simultaner Diagonalisierbarkeit zu gewährleisten: Zwar folgt aus Satz 2.8 aus der vollständigen Lösung des Eigenwertproblems die simultane Diagonalisierbarkeit, die Umkehrung ist aber nicht richtig. Man nehme

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

dann sind diese Matrizen - da bereits diagonal - simultan diagonalisierbar. Wegen Bild(A) = $\{(x, y, 0) \mid x, y \in \mathbb{R}\}$ und Bild(B) = $\{(x, 0, 0) \mid x \in \mathbb{R}\}$ ist jedoch die für die vollständige Lösung des Eigenwertproblems notwendige Bedingung aus Lemma 2.6 nicht erfüllt.

Wir wollen jetzt noch kurz untersuchen, inwieweit im Fall einer singulären Matrix B eine Reduktion auf ein "kleineres" Eigenwertproblem möglich ist. Man könnte vermuten, daß in dieser Situation die Abspaltung eines geeigneten Unterraums so möglich ist, daß anschließend ein reduziertes Eigenwertproblem $A'\mathbf{x} = \lambda B'\mathbf{x}$ mit nichtsingulärem B' entsteht. Entgegen den Andeutungen in Bunse-Gerstner [17] ist aber (im nichtsymmetrischen Fall) eine vollständige Diagonalisierung des ursprünglichen Problems nicht mehr gewährleistet:

Nach unseren früheren Überlegungen dürfen wir (durch Faktorisierung von B) davon ausgehen, daß B symmetrisch ist. Durch eine entsprechende Transformation W (Drehung und anschließende Skalierung) der linken und rechten Seite geht das Paar (A, B) schließlich in die Form

$$A' := W^T A\,W = \begin{bmatrix} A_1 & \mathbf{a} \\ \mathbf{a}^T & \alpha \end{bmatrix}, \qquad B' := W^T B\,W = \begin{bmatrix} J_1 & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}$$

über mit $\mathbf{a}, \mathbf{a} \in \mathbb{R}^{n-1}$, $\alpha \in \mathbb{R}$, A_1 und J_1 (n-1, n-1)-Matrizen. (Der Übersichtlichkeit halber wollen wir nur einen eindimensionalen Unterraum abspalten.)
Das weitere Vorgehen hängt davon ab, ob der Koeffizient α null ist oder nicht:

<u>Fall 1:</u> $\alpha \neq 0$
Hier können wir durch Gauß-artige Transformationen, die zunächst auf die n-te Spalte und anschließend auf die n-te Zeile angewendet werden, erreichen, daß die Koeffizienten $a_{1,n}, \dots, a_{n-1,n}, a_{n,1}, \dots, a_{n,n-1}$ zu null werden; die Matrix B' bleibt dabei unverändert:

$$A'' := \begin{bmatrix} A_1' & \mathbf{0} \\ \mathbf{0}^T & +1 \end{bmatrix}, \qquad B'' \equiv B' := \begin{bmatrix} J_1 & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}$$

Da die Bedingung Bild(A) $\subseteq$ Bild(B) offensichtlich nicht erfüllt ist (der Vektor (0,.....,0,1) wäre Eigenvektor zum Eigenwert "∞" !), läßt sich das "große" Eigenwertproblem nicht vollständig lösen, auch wenn dies für das "kleinere"

$$A_1'\,\mathbf{u} = \lambda\,J_1\,\mathbf{u}$$

der Fall sein könnte. Dagegen reduziert sich die simultane Diagonalisierbarkeit wegen der erfolgreichen Zerlegung in invariante Unterräume auf die des reduzierten Problems (A_1', J_1).

Fall 2: $\alpha = 0$
Hier haben wir das Paar

$$(2.11) \qquad A' := \begin{bmatrix} A_1 & \mathbf{a} \\ \mathbf{a}^T & 0 \end{bmatrix}, \qquad B' := \begin{bmatrix} J_1 & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}$$

zu behandeln. Für die vollständige Lösbarkeit des Gesamteigenwertproblems ist wieder Bild(A) $\subseteq$ Bild(B) notwendig, was $\mathbf{a} = \mathbf{0}$ erfordert. In diesem Fall haben wir den Eigenvektor $\mathbf{x}_n^T = (0, \dots, 0, 1)$ zum Eigenwert "0/0", also zu beliebigem $\lambda \in \mathbb{R}$. Die restlichen Eigenvektoren

$$\begin{pmatrix} \mathbf{u}_1 \\ v_1 \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{u}_{n-1} \\ v_{n-1} \end{pmatrix} \qquad (\mathbf{u}_i \in \mathbb{R}^{n-1},\ v_i \in \mathbb{R})$$

müssen $A_1 \mathbf{u}_i + \mathbf{a} v_i = \lambda_i J_1 \mathbf{u}_i$ erfüllen, was man mit dem Ansatz $v_i := 0$ zu lösen versuchen wird.

Von sehr speziellen Fällen abgesehen, wird also für $\alpha = 0$ die vollständige Lösung des "großen" Eigenwertproblems nicht möglich sein, auch wenn sie für (A_1, J_1) erreicht werden kann.

Etwas besser sieht die Situation bei symmetrischen Matrizen A, B aus. Hier wird man die Transformation so durchführen, daß die Symmetrie gewahrt bleibt. Die neuen Matrizen A', B' haben folglich die Gestalt

$$A' \equiv \begin{bmatrix} A_1 & \mathbf{a} \\ \mathbf{a}^T & \alpha \end{bmatrix}, \qquad B' \equiv \begin{bmatrix} B_1 & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}.$$

Falls α ungleich null ist, können wir $\mathbf{a}$ und $\mathbf{a}^T$ zum Verschwinden bringen. Anders als im unsymmetrischen Fall müssen wir aber jetzt darauf achten, daß die Symmetrie von A und B nicht zerstört wird: Es sei

$$\mathbf{c}^T := \left(-\frac{a_{1n}}{\alpha}, \dots, -\frac{a_{n-1,n}}{\alpha}\right) \quad \text{und} \quad F := \begin{bmatrix} I & \mathbf{c} \\ \mathbf{0}^T & 1 \end{bmatrix}. \quad \text{Dann gilt:}$$

$$F A F^T = \begin{bmatrix} I & \mathbf{c} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} A_1 & \mathbf{a} \\ \mathbf{a}^T & \alpha \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{c}^T & 1 \end{bmatrix} = \begin{bmatrix} A_1 + \mathbf{c}\,\mathbf{a}^T & \mathbf{a} + \alpha \mathbf{c} \\ \mathbf{a}^T & \alpha \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{c}^T & 1 \end{bmatrix}$$

$$= \begin{bmatrix} A_1' & \mathbf{0} \\ \mathbf{a}^T & \alpha \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{c}^T & 1 \end{bmatrix} = \begin{bmatrix} A_1' & \mathbf{0} \\ \mathbf{a}^T + \alpha \mathbf{c}^T & \alpha \end{bmatrix} = \begin{bmatrix} A_1' & \mathbf{0} \\ \mathbf{0}^T & \alpha \end{bmatrix}$$

Das neue Problem läßt sich also für $\alpha \neq 0$ reduzieren, für $\alpha = 0$ jedoch nur unter den oben angegebenen Einschränkungen.

Wichtig ist der Sonderfall eines gemeinsamen Unterraumes U von Ker A und Ker B. Nach einer eventuellen Umnumerierung kommt man auf die Gestalt (2.11) mit $\mathbf{a} = \mathbf{0}$. Eine vollständige Zerlegung ist zumindest im Fall eines symmetrischen A möglich.-

2.3 Das positiv definite verallgemeinerte Eigenwertproblem

Anders als beim normalen Eigenwertproblem reicht beim verallgemeinerten Eigenwertproblem die Symmetrie der Matrizen nicht für die vollständige Lösung aus. Hinreichende (aber keineswegs notwendige) Bedingung ist die Positiv-Definitheit von B bzw. allgemeiner des Paares (A, B).

Wie bereits bemerkt, läßt sich ein beliebiges Eigenwertproblem $A\mathbf{x} = \lambda B \mathbf{x}$ mit Ausnahme des Falls, daß das charakteristische Polynom $P(\lambda) = \det(A - \lambda B)$ identisch null ist ("singuläre Büschel"), immer auf ein äquivalentes mit symmetrischen Matrizen A, B transformieren.
Sehr häufig wird B sogar positiv definit sein, etwa wenn diese Matrix den "Massen-Operator" bei einem gekoppelten mechanischen System darstellt. In diesem Fall ist das verallgemeinerte Eigenwertproblem stets lösbar:

Satz 2.11: ***Es seien A, B symmetrische Matrizen und B positiv definit. Dann ist das Eigenwertproblem $A\mathbf{x} = \lambda B \mathbf{x}$ vollständig lösbar bzw. die beiden Matrizen A und B sind mit zwei Matrizen W, W^T simultan diagonalisierbar.***

Beweis: (Einen entsprechenden Satz mit Beweis findet man auch in Parlett [46], wir geben den einfachen Beweis der Vollständigkeit halber hier wieder.)
Es existiert eine orthogonale Transformation U, die B in Diagonalgestalt transformiert:

$$U^T B U = B_{diag} = \mathrm{diag}(b_1, \dots, b_n) \quad \text{mit } b_1, \dots, b_n > 0.$$

Gleichzeitig geht A in eine (wiederum) symmetrische Matrix

$$A' := U^T A U$$

über. Multiplizieren wir jetzt B_{diag} von links und von rechts mit der nichtsingulären Diagonalmatrix

$$S := \mathrm{diag}\left(\frac{1}{\sqrt{b_1}}, \dots, \frac{1}{\sqrt{b_n}}\right),$$

so wird B_{diag} in die Einheitsmatrix überführt, $A'' = S A' S$ bleibt wegen $S^T = S$ symmetrisch und kann mit Hilfe einer weiteren orthogonalen Matrix V diagonalisiert werden:

$$V^T A'' V = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$$

Wegen $V^T I V = V^T V = I$ ändert sich das transformierte B nicht mehr. Die zusammengesetzten Transformationsmatrizen

$$F := V^T S U^T, \quad G := U S V = (V^T S U^T)^T = F^T$$

bewirken damit

$$F A G = \mathrm{diag}(\lambda_1, \dots, \lambda_n), \quad F B G = I.$$

Wegen $\det B \neq 0$ gilt Bild(B) = $\mathbb{R}^n$. Damit ist nach Satz 2.10 auch das verallgemeinerte Eigenwertproblem $A\mathbf{x} = \lambda B \mathbf{x}$ vollständig lösbar.

Q.E.D.

Bezüglich der simultanen Diagonalisierbarkeit können die Rollen von A, B im letzten Satz natürlich vertauscht werden, d.h. A positiv definit ist ebenfalls hinreichend. Für das verallgemeinerte Eigenwertproblem stimmt dies jedoch nicht mehr ganz, weil dort das Problem unendlicher Eigenwerte (Fall "1/0") auftreten kann.
Wenn B (positiv) semidefinit ist, so kann man - wie am Schluß von Abschnitt 2.2 kurz skizziert - versuchen, den Nullraum von B "abzuspalten" und das durch Deflation verkürzte Eigenwertproblem $A'\mathbf{x} = \lambda B'\mathbf{x}$ mit positiv definitem B' zu lösen. Die Fortsetzung auf das Ausgangsproblem ist zumindest im Fall $\operatorname{Ker} A \subseteq \operatorname{Ker} B$ immer möglich, allgemein haben wir jedoch die gleiche Situation wie in Abschnitt 2.2.

Die Voraussetzung eines positiv definiten B (oder A) kann nicht unwesentlich abgeschwächt werden:

Definition 2.12: *Ein Paar (A, B) von symmetrischen Matrizen A, B heißt positiv definit, wenn es Koeffizienten $\alpha, \beta \in \mathbb{R}$ gibt, so daß die Linearkombination $\alpha A + \beta B$ positiv definit ist. (Man spricht dann auch von einem positiv definiten Büschel.)*

Es gibt verschiedene äquivalente Definitionen für die Positiv-Definitheit von Paaren, eine interessante wurde von A. Berman und A. Ben-Israel [9] für Matrizen der Dimension $n \geq 3$ gegeben:

(A, B) ist positiv definit, wenn $(\mathbf{x}, A\mathbf{x}) = (\mathbf{x}, B\mathbf{x}) = 0$ nur für $\mathbf{x} = \mathbf{0}$ möglich ist.

Das Beispiel

$$A := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

zeigt, daß die Bedingung $n \geq 3$ nicht abschwächbar ist: $(\mathbf{x}, A\mathbf{x}) = (\mathbf{x}, B\mathbf{x}) = 0$ kann nur für $\mathbf{x} = \mathbf{0}$ möglich sein. Andererseits ist das Paar (A, B) nicht positiv definit, denn

$$\alpha A + \beta B = \begin{bmatrix} \alpha & \beta \\ \beta & -\alpha \end{bmatrix}$$

ist für keine Wahl von α, β positiv definit, wie man z.B. aus dem Hurwitz-Kriterium erkennt. (Übrigens ist im Komplexen das Paar (A, B) sehr wohl diagonalisierbar, seine Eigenwerte sind +i, -i.)

Eine weitere - für beliebiges n gültige - Definition der positiven Definitheit von A, B lautet:

Wenn X eine positiv semidefinite Matrix ist, so ist $tr(AX) = tr(BX) = 0$ nur für $X = 0$ möglich.

(Interpretiert man die Spurbildung als inneres Produkt zwischen Matrizen, d.h. $\langle X, Y \rangle := tr(XY)$, dann bedeutet dies, daß es keine positiv semidefinite Matrix gibt, die "senkrecht" auf A, B steht.)
Die letzte Definition ist in Hinblick auf Spurminimierungsaussagen nützlich; wir werden im folgenden aber immer mit der ursprünglichen Definition für positive Definitheit von Paaren auskommen.

Mit Hilfe von Satz 2.11 ist jetzt die folgende wichtige Aussage auf elementare Weise beweisbar :

Satz 2.13: *Es seien A, B symmetrische Matrizen, die ein positiv definites Büschel bilden. Dann sind A, B simultan diagonalisierbar.*

Beweis:
Wir können uns auf den Fall $\alpha A + \beta B$ positiv definit mit $\alpha \neq 0$, $\beta \neq 0$ beschränken, denn sonst dürfen wir direkt Satz (2.11) heranziehen.
Für das Paar $(A, \alpha A + \beta B)$ gibt es nach dem erwähnten Satz eine nichtsinguläre Matrix W, so daß $W^T A W$ und $W^T (\alpha A + \beta B) W$ diagonal werden. Die entsprechenden Diagonalmatrizen bezeichnen wir mit A_{diag}, C_{diag}.
Es gilt

$$\begin{aligned} W^T B W &= \frac{1}{\beta} W^T \left(\alpha A + \beta B - \alpha A\right) W \\ &= \frac{1}{\beta} W^T \left(\alpha A + \beta B\right) W - \frac{\alpha}{\beta} W^T A W \\ &= \frac{1}{\beta} C_{diag} - \frac{\alpha}{\beta} A_{diag}, \end{aligned}$$

d.h. B wird ebenfalls auf Diagonalgestalt transformiert.

Q.E.D.

Obwohl in vielen praktischen Anwendungen das Paar (A, B) ein positiv definites Büschel bildet, so kann doch nicht übersehen werden, daß diese Bedingung eine wesentliche und keineswegs notwendige Einschränkung darstellt:
Stellt man sich vor, daß ein Paar (A, B) simultan auf Diagonalgestalt gebracht ist, so bedeutet dies für die Diagonalelemente $a_1, \ldots, a_n, b_1, \ldots, b_n$ das *gleichzeitige* Erfülltsein der folgenden Bedingungen für ein geeignetes Paar (α, β) von Koeffizienten:

$$\begin{aligned} \alpha a_1 + \beta b_1 &> 0 \\ &\vdots \\ \alpha a_n + \beta b_n &> 0 \end{aligned}$$

Da einer der beiden Koeffizienten ungleich null sein muß und (o.B.d.A.) als positiv angenommen werden kann, sind die folgenden Bedingungen äquivalent für ein geeignetes ρ:

$$a_1 + \rho b_1 > 0, \quad \ldots \quad, a_n + \rho b_n > 0.$$

Für $b_i \neq 0$ bedeutet dies, daß eine Ungleichung der Art

$$\rho > -\frac{a_i}{b_i} \; (b_i > 0) \quad \text{oder} \quad \rho < -\frac{a_i}{b_i} \; (b_i < 0)$$

besteht. Für i = 1 haben wir noch keine Einschränkung; für die weiteren zufällig erhaltenen, unabhängigen $a_2, \dots, a_n, b_2, \dots, b_n$ wird die Wahrscheinlichkeit, daß ein solches ρ existiert, von der Größenordnung $(1/2)^{n-1}$ sein. -

Im weiteren Verlauf dieser Arbeit wird nur der Fall eines symmetrischen positiv definiten Paares (A, B) behandelt. Dabei wird zusätzlich vorausgesetzt, daß B von vornherein nicht singulär ist, was nach unseren früheren Überlegungen (genauso wie die Symmetrie) keine allzu wesentliche Einschränkung ist. Die Matrix B können wir uns durch eine unitäre Transformation V auf Diagonalgestalt $B' := \mathrm{diag}(b_1, \dots, b_n)$ gebracht denken. Um die Symmetrie von $A' = V^T A V$ nicht zu zerstören, wenden wir jetzt auf A' und B' die Matrix

$$S := \mathrm{diag}\left(\frac{1}{\sqrt{|b_1|}}, \dots, \frac{1}{\sqrt{|b_n|}}\right)$$

von links und rechts an. Das ursprüngliche B wird damit insgesamt zu

$$S \; \mathrm{diag}(b_1, \dots, b_n) \; S = J := \mathrm{diag}(j_1, \dots, j_n) \text{ mit } j_i = \pm 1,$$

$S^T V A V S$ bleibt wegen $S^T = S$ eine symmetrische Matrix.
Dies bedeutet, daß wir uns immer auf den Fall A symmetrisch, $B \equiv J$ beschränken dürfen. Dabei muß die Diagonalisierung von B nicht wirklich ausgeführt werden - es können die für den Gauß-Algorithmus üblichen Zeilen- und Spaltenoperationen gleichzeitig auf A und B angewendet werden, bis B die angestrebte J-Form erreicht hat. Durch Permutationen ist für J auch die speziellere Form

$$J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$$

erreichbar. Diese wird aber in dieser Arbeit - anders als in Kovač-Striko und Veselič [38] nicht vorausgesetzt.
Für die hier zu behandelnden Divide-et-Impera-Verfahren ist entscheidend, daß A - unter Beibehaltung der Form von B (bzw. J) - in Tridiagonalgestalt oder Tridiagonalgestalt mit Rändern transformiert werden kann. Auf eventuelle numerische Probleme wird in Kapitel 5 ausführlich eingegangen werden.

Weitere Aussagen über Matrizenbüschel findet man etwa in Berman und Ben Israel [9]; Stewart stellt in [56] Gershgorin-Theoreme für das verallgemeinerte Eigenwertproblem vor. Mit der Störungstheorie für das verallgemeinerte Eigenwertproblem beschäftigen sich u.a. Barlow und Demmel [6], Chu [18], Stewart [55], [58].

3. Einige numerische Verfahren zum verallgemeinerten Eigenwertproblem

In diesem Kapitel stellen wir zunächst die üblichsten Verfahren zur Lösung des verallgemeinerten Eigenwertproblems $A\mathbf{x} = \lambda B\mathbf{x}$ dar und diskutieren ihre Vor- und Nachteile. Unter ihnen besticht das QZ-Verfahren durch seine Universalität und Schnelligkeit, das verallgemeinerte Jacobi-Verfahren durch seine Genauigkeit und die implizit vorhandene Möglichkeit, die Positivdefinitheit eines Paares zu überprüfen. Zum Schluß stellen wir zwei neue Verfahren vor, die auf einfache Weise testen, ob das Paar (A, B) positiv definit ist.

Viele der in den letzten Jahrzehnten entwickelten Verfahren zur Lösung des Standardeigenwertproblems $A\mathbf{x} = \lambda\mathbf{x}$ lassen sich auch auf das verallgemeinerte Eigenwertproblem $A\mathbf{x} = \lambda B\mathbf{x}$, A, B symmetrisch, übertragen, wenn die Matrix B bestimmte Bedingungen erfüllt: die meisten dieser Verfahren verlangen die ***positive Definitheit der Matrix B,*** die wir deshalb in diesem Kapitel meist voraussetzen werden.

3.1 Potenzmethode, inverse Iteration und Rayleighquotientenverfahren

Im Fall des Standardeigenwertproblems $A\mathbf{x} = \lambda\mathbf{x}$ bietet sich die ***Potenzmethode*** zur Berechnung des betragsgrößten Eigenwertes und des zugehörigen Eigenvektors an, die ***inverse Iteration*** zur Bestimmung des betragskleinsten Eigenwertes und seines Eigenvektors. Möchte man den nächsten Eigenwert zu einer Eigenwertnäherung $\sigma \in \mathbb{R}$ erhalten, so kann man die inverse Iteration mit der um σ verschobene Matrix $A - \sigma$ durchführen.
Wenn wir diese beiden Methoden auf ein Eigenwertproblem der Gestalt $B^{-1}A\mathbf{x} = \lambda\mathbf{x}$ anwenden und die entstehenden Gleichungen geeignet umformen, so daß B^{-1} nicht direkt gebildet werden muß, erhalten wir die folgenden Algorithmen für das verallgemeinerte Eigenwertproblem $A\mathbf{x} = \lambda B\mathbf{x}$:

Sei $\mathbf{x}_0$ ein geeignet gewählter Startvektor.

Potenzmethode:

$$\text{(3.1)} \quad \begin{aligned} &\text{Bestimme } \mathbf{y}_{k+1} \text{ aus } B\mathbf{y}_{k+1} = A\mathbf{x}_k. \\ &\text{Normiere: } \mathbf{x}_{k+1} := \frac{\mathbf{y}_{k+1}}{\mathbf{y}_{k+1}^T B\,\mathbf{y}_{k+1}}. \end{aligned}$$

Inverse Iteration mit Verschiebung:

σ sei eine Eigenwertnäherung, die mit keinem Eigenwert des Eigenwertproblems $A\mathbf{x} = \lambda B\mathbf{x}$ exakt übereinstimmt.

$$\text{(3.2)} \quad \begin{aligned} &\text{Bestimme } \mathbf{y}_{k+1} \text{ aus } (A - \sigma B)\mathbf{y}_{k+1} = B\mathbf{x}_k. \\ &\text{Normiere: } \mathbf{x}_{k+1} := \frac{\mathbf{y}_{k+1}}{\mathbf{y}_{k+1}^T B\,\mathbf{y}_{k+1}}. \end{aligned}$$

Die Vektoren $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ bilden dann in beiden Fällen eine Folge von Näherungen für den gesuchten Eigenvektor $\mathbf{x}$.

Die Konvergenzaussagen zu diesen beiden Verfahren entsprechen denen im Fall B = I, vgl. dazu (Parlett [46], S. 60 - 62); falls die Folge konvergiert, ist die Konvergenz linear.

Bei beiden Methoden muß ein Gleichungssystem gelöst werden. Für große Matrizen A und B hängt die Anwendbarkeit der beiden Verfahren davon ab, ob diese Gleichungssysteme z.B. wegen einer speziellen Gestalt der beteiligten Matrizen gut lösbar sind.

Eine gute Wahl für die Verschiebung σ in der Methode der inversen Iteration ist der verallgemeinerte *Rayleighquotient*

$$(3.3) \qquad \rho(\mathbf{x}) := \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T B \mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

Auf diese Weise erhält man das sogenannte *Rayleighquotientenverfahren:*

$$(3.4) \qquad \begin{aligned} &\text{Sei } \mathbf{x}_0 \text{ ein geeignet gewählter Startvektor.} \\ &\text{Man bestimme } \sigma_k := \frac{\mathbf{x}_k^T A \mathbf{x}_k}{\mathbf{x}_k^T B \mathbf{x}_k}. \\ &\text{Man berechne } \mathbf{y}_{k+1} \text{ mit } (A - \sigma_k B)\mathbf{y}_{k+1} = B\mathbf{x}_k. \\ &\text{Man normiere: } \mathbf{x}_{k+1} := \frac{\mathbf{y}_{k+1}}{\mathbf{y}_{k+1}^T B \mathbf{y}_{k+1}}. \end{aligned}$$

Dieses Verfahren konvergiert entweder (kubisch) direkt gegen den durch den Startvektor festgelegten Eigenvektor oder oszilliert zwischen zwei Eigenvektoren (vgl. Parlett [46], S. 76, S. 320).

3.2 Rayleigh-artige Verfahren

Die folgenden Verfahren verwenden den Rayleighquotienten oder das Rayleighquotientenverfahren als Hilfsmittel:

3.2.1 Die Rayleigh-Ritz-Approximation und die Unterraum-Iteration

Zu einem gegebenen Unterraum S^m des $\mathbb{R}^n$ möchte man die "beste" Menge von Eigenvektornäherungen aus S^m für Eigenvektoren des verallgemeinerten Eigenwertproblems $A\mathbf{x} = \lambda B\mathbf{x}$ bestimmen.
Wie im Fall des Standardeigenwertproblems gelingt dies durch die Bildung sogenannter *Rayleigh-Ritz-Approximationen:*

Sei dabei S eine $n \times m$-Matrix, deren Spalten eine Basis des S^m bilden.

Dann sind die Paare $(\vartheta_i, S\mathbf{g}_i)$, $i = 1, .., m$, die den Eigenwertgleichungen

$$(3.5) \qquad (S^T A S)\,\mathbf{g}_i = \vartheta_i\,(S^T B S)\,\mathbf{g}_i, \quad i = 1, \dots, m$$

genügen, die gesuchten Rayleigh-Ritz-Approximationen.

Im Buch von Parlett [46] (s. S. 322) wird angegeben, in welchem Sinne diese Paare optimal sind.

Auch die *Unterraum-Iteration* läßt sich auf das verallgemeinerte Eigenwertproblem übertragen, was nicht verwunderlich ist, denn ihre grundlegende Idee besteht darin, die Blockversion der inversen Iteration (inklusive Verschiebungen) mit Rayleigh-Ritz-Approximationen in jedem Schritt zu kombinieren, und beide Verfahren sind auf das verallgemeinerte Eigenwertproblem anwendbar.
Eine schöne Zusammenfassung dieses Verfahrens findet man in Parlett [46], S. 326. Dort werden auch die Probleme des Verfahrens angesprochen: die Effizienz dieser Methode hängt nicht unwesentlich von der Wahl der Blockgröße ab - sie sollte größer gewählt werden als die Anzahl der gewünschten Eigenwerte und Eigenvektoren - nur wie groß, ist vor Beginn des Programms leider schwierig zu bestimmen.

3.2.2 Koordinatenrelaxation

Die *Koordinatenrelaxation* berechnet den kleinsten Eigenwert des Matrizenpaares (A, B). Im Fall sehr großer Matrizen A, B ist sie eine sehr wichtige Alternative zur Rayleigh-Quotienten-Iteration, da sie ohne das Lösen linearer Gleichungssysteme auskommt - es werden nur Matrix-Vektor-Multiplikationen ausgeführt. Sie ist empfehlenswert für dünn besetzte Matrizen sowie allgemein für alle Matrizen mit einer besonders gut für Matrix-Vektor-Multiplikationen geeigneten Gestalt. Dafür muß man sich mit einer im allgemeinen nicht sehr großen Konvergenzgeschwindigkeit abfinden.

Der Koordinatenrelaxation liegt eine Minimierung des Rayleigh-Quotienten zugrunde. Für einen von null verschiedenen Vektor $\mathbf{x}$ ist der verallgemeinerte Rayleigh-Quotient $\rho(\mathbf{x}) := (\mathbf{x}^T A\mathbf{x})/(\mathbf{x}^T B\mathbf{x})$ größer oder gleich dem kleinsten Eigenwert von (A, B). Dieser Vektor $\mathbf{x}$ wird in der Komponente j (j durchläuft zyklisch $\{1, \dots, n\}$) so verändert, daß der neue Vektor $\mathbf{y}$ den Rayleigh-Quotienten über der Menge $\{\alpha\mathbf{x} + \beta\mathbf{e}_j \mid (\alpha, \beta) \in \mathbb{R}^2 \setminus \{\mathbf{0}\}\}$, $\mathbf{e}_j$ j-ter Einheitsvektor, minimiert. Zu lösen sind dann verallgemeinerte Eigenwertprobleme der Dimension 2x2, deren Lösungen man direkt bestimmen kann. (Die vollständige Beschreibung des Algorithmus mit Erläuterungen und Bemerkungen findet sich z.B. im Buch von Bunse-Gerstner [17], S. 285-287.)
Will man Aussagen für mehrere kleinste Eigenwerte und Eigenvektoren erhalten, so verwendet man die sogenannte *simultane Gruppenkoordinatenrelaxation*, bei der gleichzeitig mit mehreren Vektoren gearbeitet wird und Veränderungen jeweils in mehreren Komponenten zugelassen sind.

3.3 Der QZ-Algorithmus

Mit Hilfe des *QZ-Algorithmus* (der Erweiterung des *QR-Algorithmus* auf das verallgemeinerte Eigenwertproblem) kann man alle Eigenwerte (und - falls gewünscht - auch die Eigenvektoren) des reellen nxn-Matrizenpaares (A, B) bestimmen. Dabei sind auch nichtsymmetrische Matrizen erlaubt, und B darf singulär sein. Diese sehr geringen Voraussetzungen an das Matrizenpaar (A, B) unterscheiden das QZ-Verfahren deutlich von den anderen hier beschriebenen Methoden. Da es nicht nur vielseitig anwendbar ist, sondern auch schnell, kann man es als *das* Standard-Verfahren zur Eigenwertbestimmung (auf nicht-parallelen Rechnern) ansehen. (Welche Bedeutung ihm zukommt, erkennt man auch an seiner Behandlung in der Literatur: im Buch von Golub/van Loan [29] z.B. nimmt die Beschreibung des QR- und des QZ-Algorithmus den weitaus größten Teil der Kapitel über Eigenwerte ein, ja die gesamte Theorie scheint um ihn herum aufgebaut zu sein...)

Die theoretische Grundlage des QZ-Verfahrens ist die *verallgemeinerte Schur-Zerlegung* nach Moler und Stewart (vgl. Golub/van Loan [29], Theorem 7.7.1): komplexe Matrizen A, B können mittels unitärer Transformationen gleichzeitig auf obere Dreiecksgestalt transformiert werden. Bei reellen Matrizen A, B ist eine (reelle) Transformation auf obere Dreiecksmatrizen im allgemeinen nicht mehr möglich: A kann mittels orthogonaler Matrizen nur noch auf "Quasidreiecksgestalt" transformiert werden, B wird wieder zu einer Dreiecksmatrix (*"verallgemeinerte reelle Schur-Zerlegung"*, vgl. hierzu Golub/van Loan [29] , Theorem 7.7.2).

Das QZ-Verfahren besteht jetzt aus den folgenden wesentlichen Schritten:

1. Man reduziere A auf obere Hessenbergform und B auf obere Dreiecksform mittels orthogonaler Transformationen.
2. Es wird ein Deflationsschritt in den folgenden Situationen durchgeführt:
   (i) Ist $a_{k+1,k} = 0$ für ein k, so teilt man das Eigenwertproblem in zwei unabhängige kleinere Teilprobleme auf.
   (ii) Gilt dagegen $b_{kk} = 0$, so kann man eine Null in Position (n, n-1) von A erzeugen und das zugehörige 1x1-Problem $a_{nn}\, x' = \lambda\, 0\, x'$ abspalten.
3. Während des eigentlichen QZ-Prozesses wird A so durch eine Folge von Iterationen in eine obere Quasi-Dreiecksmatrix transformiert, daß B seine (obere Dreiecks-) Gestalt behält. (Wichtig ist hierbei auch die Verwendung geeigneter Verschiebungsstrategien.)

Eine gute und übersichtliche Darstellung der wesentlichen Ideen für den Fall komplexer Matrizen A, B findet sich im Artikel von Kaufman [35].
Für symmetrische Paare (A, B) wird die Symmetrie im Verlauf des oben beschriebenen Algorithmus zerstört, weswegen auch Parlett [46] andere Verfahren für diesen Fall vorzieht. Golub schlägt hier eine Kombination aus Cholesky-Zerlegung und symmetrischem QR-Algorithmus vor, wenn B noch zusätzlich positiv definit ist. (Als weitere Literatur s. etwa Bunse-Gerstner [16], Kaufman [36], Moler und Stewart [42], Brebner und Grad [12] sowie Ward [67].)

3.4 Das verallgemeinerte Jacobi-Verfahren

Parlett beschreibt in seinem Buch [46], wie man das Jacobi-Verfahren für das Standardeigenwertproblem direkt für das Paar (A, B) verallgemeinern kann, wenn B positiv definit ist. Ein Vorteil dieses verallgemeinerten Jacobi-Verfahrens liegt darin, daß - im Gegensatz zum QZ-Verfahren etwa - die Symmetrie der beteiligten Matrizen nicht zerstört wird. Als Anwendungsgebiet für dieses Jacobi-Verfahren schlägt Parlett kleinere Matrizen A und B vor, die diagonaldominant sind.
Wegen seiner Genauigkeitseigenschaften und der Möglichkeit einer Parallelisierung ist das Interesse an diesem Verfahren mittlerweile wieder sehr groß, man vgl. Demmel, Veselič [21] (Genauigkeitsaussagen), Sameh [48] (ein paralleles Jacobi-Verfahren).
Das verallgemeinerte Jacobi-Verfahren ist nicht nur anwendbar, wenn B positiv definit ist, sondern auch für positiv definite *Paare* (A, B), vgl. Veselič [63], Hari und Slapničar [52]. Nun weiß man aber für ein Paar (A, B) nicht im voraus, ob es positiv definit ist. Veselič reduziert das Paar (A, B) auf die Gestalt (A', J) mit $J = \mathrm{diag}(I_m, -I_{n-m})$. Nach Veselič konvergiert für solche Paare (A', J) der verallgemeinerte Jacobi-Algorithmus oder entdeckt die Indefinitheit des Paares.

Die Jacobi-Methode besteht aus einer iterativen Anwendung von Kongruenztransformationen der speziellen Form

$$C = \begin{bmatrix} c_{rr} & c_{rs} \\ c_{sr} & c_{ss} \end{bmatrix} \quad \text{(die nicht dargestellten Elemente sind die der Einheitsmatrix)}$$

auf die Matrix A: $A' = C^T A C$, wobei die Transformationsmatrix C die Bedingung $C^T J C = J$ erfüllt.

Dabei werden für $1 \le r, s \le m$ oder $m+1 \le r, s \le n$ die üblichen trigonometrischen Transformationen ausgeführt, in den übrigen Fällen sogenannte hyperbolische Transformationen:

$$(3.6) \quad \begin{bmatrix} c_{rr} & c_{rs} \\ c_{sr} & c_{ss} \end{bmatrix} = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix} \quad \text{bzw.} \quad \begin{bmatrix} c_{rr} & c_{rs} \\ c_{sr} & c_{ss} \end{bmatrix} = \begin{bmatrix} \cosh y & \sinh y \\ \sinh y & \cosh y \end{bmatrix}$$

Die Parameter x bzw. y werden dabei so gewählt, daß das Element an der Position (r, s) der transformierten Matrix A' zu null gemacht wird, d.h.

$$\tan 2x = \frac{2a_{rs}}{a_{rr} - a_{ss}} \quad \text{bzw.} \quad \tanh 2y = -\frac{2a_{rs}}{a_{rr} + a_{ss}} .$$

Dabei muß die Bedingung $|a_{rr} + a_{ss}| > 2\,|a_{rs}|$ erfüllt sein. Wenn sie verletzt ist, bricht der Algorithmus zusammen, aber das geschieht für positiv definite Paare (A, J) niemals (vgl. Veselič [63], Theorem 2.1). Im Fall positiv definiter Paare konvergiert das verallgemeinerte Jacobi-Verfahren bei Verwendung der zeilenzyklischen Strategie (Veselič [63], Theorem 2.3). Ist das Paar (A', J) nicht positiv definit, so deckt die Jacobi-Methode die Indefinitheit in endlich vielen Schritten auf. (Veselič [63], Theorem 2.4; zu weiteren Aussagen über Jacobi-Algorithmen vgl. auch Veselič [61], [62] und [65]).

3.5 Weitere Verfahren

Neben dem QZ-Algorithmus, Jacobi-und Rayleigh-artigen Verfahren sind noch eine Reihe anderer Verfahren entwickelt worden, die aber insgesamt nicht soviel Bedeutung erlangt haben. (Schöne Übersichten über Lösungsverfahren für das verallgemeinerte Eigenwertproblem stellen etwa die Artikel von Bathe, Wilson [7] und Stewart [57] dar.) Zwei dieser Methoden, das *Bisektionsverfahren* und der *Lanczos-Algorithmus*, werden hier kurz beschrieben, da das Interesse an ihnen in letzter Zeit wieder gestiegen ist.

3.5.1 Bisektionsverfahren

Mit dem Aufkommen der Parallelrechner hat das Bisektionsverfahren wieder an Bedeutung gewonnen, da es gut parallelisierbar ist. Dem Verfahren liegen die folgenden Ideen zugrunde (vgl. Bunse-Gerstner [17], S. 270-273, Veselič [60]):
Für symmetrische Matrizen A, B (mit B positiv definit) kann man mit Hilfe des Trägheitssatzes von Sylvester beliebig kleine Intervalle konstruieren, die Eigenwerte des Paares (A, B) enthalten: Für festes $\mu \in \mathbb{R}$ gibt der Trägheitsindex von $A - \mu B$ an, wieviele Eigenwerte von (A, B) kleiner als, gleich oder größer als μ sind. Vergleicht man diese Zahlen für zwei verschiedene Werte μ_1 und μ_2 mit $\mu_1 < \mu_2$, so kann man feststellen, wieviele Eigenwerte im Intervall $[\mu_1, \mu_2]$ liegen. Beim Bisektionsverfahren rücken dann diese Intervallgrenzen - etwa durch fortgesetztes Halbieren des Intervalls - zusammen.
Zur Bestimmung des Trägheitsindex läßt sich die Cholesky-Zerlegung $A - \mu B = L\,D\,L^T$ verwenden, wobei L eine untere Dreiecksmatrix mit Einheitsdiagonale ist und D eine Diagonalmatrix; D ist kongruent zu $A - \mu B$, die Vorzeichen der Diagonalelemente von D bestimmen also den Trägheitsindex.

Dieses Verfahren ist gut geeignet für schwach besetzten Bandmatrizen A, B, wenn keine hohe Genauigkeit verlangt wird oder wenn nur sehr wenige Eigenwerte in einem gegebenen Intervall gesucht werden.
(Im Spezialfall B = I und A Tridiagonalmatrix benötigt man übrigens keine Cholesky-Zerlegung, sondern nutzt die "Sturmfolgeneigenschaft" aus, vgl. Golub/van Loan [29]).

3.5.2 Lanczos-Algorithmus

Beim Standardeigenwertproblem $A\mathbf{x} = \lambda\mathbf{x}$ stellen *Lanczos-Algorithmen* eine interessante Alternative zu den Standardverfahren dar, wenn es um die Bestimmung der Eigenwerte einer großen, spärlich besetzten, symmetrischen Matrix A geht. Das Lanczos-Verfahren führt eine partielle Tridiagonalisierung der Matrix A durch; dabei erhält man Informationen über die "äußersten" Eigenwerte von A meist schon lange bevor die Tridiagonalisierung vollständig ist. Daher ist diese Methode besonders geeignet zur Berechnung einiger weniger größter oder kleinster Eigenwerte von A.
Bei der Grundversion des Lanczos-Algorithmus kann es zum Verlust der Orthogonalität zwischen den von der Iteration produzierten Lanczos-Vektoren kommen. Die Version des Lanczos-Algorithmus mit selektiver Orthogonalisierung wird mit dieser Hauptschwierigkeit des Lanczos-Verfahrens fertig.

Eine Verallgemeinerung des Lanczos-Algorithmus auf das verallgemeinerte Eigenwertproblem mit positiv definitem B ist möglich (vgl. Parlett [46], S. 323 - 325), während der Fall eines nichtsymmetrischen einfachen Eigenwertproblems große Schwierigkeiten bereitet (vgl. Golub/van Loan [29], S. 501 - 503).
Als weitere Literatur sei auf Borri, Mantegazza [11] und Scott [50] verwiesen; Scott kombiniert Rayleigh-Ritz-Approximationen mit dem Lanczos-Verfahren.

3.6 Verfahren zur Bestimmung des Definitheitsintervalls

Wenn ein Paar (A, B) positiv definit ist, so gibt es (nach Definition) $\alpha, \beta \in \mathbb{R}$ mit $\alpha A + \beta B$ positiv definit. Kennt man diese $\alpha, \beta \in \mathbb{R}$, so kann man vom Paar (A, B) zum Paar (A, B') übergehen mit $B' := \alpha A + \beta B$ und dann das Eigenwertproblem für (A, B') lösen. Wie wir in den Abschnitten 3.1 bis 3.5 gesehen haben, gibt es für solche Paare mit B' positiv definit eine Reihe attraktiver Verfahren zur Bestimmung einiger oder aller Eigenwerte von (A, B'). Nur leider kennt man solche Koeffizienten α, β im allgemeinen nicht.

Mit den im folgenden beschriebenen einfachen Verfahren wird festgestellt, ob das Paar (A, B) überhaupt positiv definit sein kann. Dabei suchen wir o.B.d.A. nach einem $\mu \in \mathbb{R}$, so daß $A - \mu B$ positiv definit ist; solche μ liegen im sog. Definitheitsintervall (vgl. Kapitel 4).

3.6.1 Ein stochastisches Verfahren

Aus der Definition für positive Definitheit läßt sich direkt eine Vorschrift zur Bestimmung von μ ableiten:

$$\mathbf{x}^T (A - \mu B)\mathbf{x} > 0 \iff \mathbf{x}^T A\mathbf{x} > \mu\, \mathbf{x}^T B\mathbf{x} \quad \text{für alle } \mathbf{x} \neq \mathbf{0}$$

Für $\mathbf{x}^T B\mathbf{x} > 0$ muß daher

$$\mu < \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}},$$

für $\mathbf{x}^T B\mathbf{x} < 0$ entsprechend

$$\mu > \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$$

gelten. Wir setzen jetzt

$$(3.7) \qquad \mu_+ := \inf_{\mathbf{x}^T B\mathbf{x} > 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}, \qquad \mu_- := \sup_{\mathbf{x}^T B\mathbf{x} < 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}} .$$

(Gibt es keine derartigen $\mathbf{x}$ für μ_+ bzw. μ_-, so nehmen wir $\mu_+ := \infty$ bzw. $\mu_- := -\infty$.)

Gilt $\mu_- < \mu_+$, so ist für $\mu_- < \mu < \mu_+$ die Matrix $A - \mu B$ positiv definit. Damit ist die Suche nach einem geeigneten μ auf zwei Extremwertprobleme zurückgeführt. Da hier *globale* Extrema verlangt werden, scheiden aber die üblichen numerischen Optimierungsalgorithmen aus, weil diese - von einem Startpunkt ausgehend - in einem kleinen Bereich ein *lokales* Minimum suchen. Geeignet sind dagegen stochastische Algorithmen.

Sehr einfach zu programmieren - aber wenig effizient - ist die "blinde Monte-Carlo-Methode", bei der

$$(3.8) \qquad \rho(\mathbf{x}) := \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}} \qquad \text{(der verallgemeinerte Rayleighquotient)}$$

auf einer Folge von Zufallsvektoren $\mathbf{x}_1, \mathbf{x}_2, \ldots$ getestet und jede eingetretene "Verbesserung" registriert wird.

Wesentlich günstiger sind Kombinationen von stochastischen mit analytischen Elementen. Man wählt zunächst als Anfangspunkte die Einheitsvektoren, später zufällig erzeugte Punkte. Von einem solchen Startpunkt $\mathbf{x}$ geht man jetzt in einer Richtung $\mathbf{u}$ weiter: $\mathbf{x}' := \mathbf{x} + t\mathbf{u}$, $t \in \mathbb{R}$. Dieses $\mathbf{u}$ kann man (z.B.) wahlweise nach einer der folgenden Vorschriften bestimmen:

(i) Gradientenmethode:

$$(3.9) \qquad \mathbf{u} := (\mathbf{x}^T B\mathbf{x}) A\mathbf{x} - (\mathbf{x}^T A\mathbf{x}) B\mathbf{x}$$

(ii) Man nehme den Index i, für den $|(A\mathbf{x})_i|$ maximal wird. Als weiteren Index j wähle man per Zufall einen, für den die Komponente $|(B\mathbf{x})_j|$ nicht zu klein wird. Der durch

$$(3.10) \qquad u_i := 1, \quad u_j := -\frac{(B\mathbf{x})_i}{(B\mathbf{x})_j} \quad \text{und} \quad u_k := 0 \quad \text{für } k \neq i, j$$

definierte Vektor ist orthogonal zu $B\mathbf{x}$, es gilt deshalb

$$\rho(\mathbf{x} + t\mathbf{u}) = \frac{\mathbf{x}^T A\mathbf{x} + 2t\, \mathbf{u}^T A\mathbf{x} + t^2\, \mathbf{u}^T A\mathbf{u}}{\mathbf{x}^T B\mathbf{x} + t^2\, \mathbf{u}^T B\mathbf{u}} .$$

Zumindest für genügend kleine $|t|$ wird jetzt die angestrebte Erhöhung oder Erniedrigung von $\rho(\mathbf{x} + t\mathbf{u})$ erreicht.

(iii) Man erzeuge per Zufall einen normierten Richtungsvektor $\mathbf{u}$. Stellt der Wert $\rho(\mathbf{x} + t\mathbf{u})$ für eine kleine Schrittweite $t > 0$ eine Verbesserung dar, so akzeptiere man diese Richtung, andernfalls probiere man es mit der entgegengesetzten Richtung $-\mathbf{u}$. Bringt auch diese keine Verbesserung, erzeuge man eine neue Zufallsrichtung. Hat man eine geeignete Richtung gefunden, so laufe man in diese Richtung, aber jetzt mit doppelter Schrittweite. Erst wenn eine Verschlechterung eintritt, gehe man mit halbierter Schrittweite wieder etwas zurück.

Das Akzeptieren eines neuen Punktes kann mit der Methode des "adiabatischen Abkühlens" (adiabatic annealing) ebenfalls vom Zufall abhängig gemacht werden: hier werden - mit abnehmender Wahrscheinlichkeit - auch ungünstigere Richtungen verfolgt, um nicht in einem lokalen Minimum hängen zu bleiben (vgl. K. Keller [37]).

Stochastische Optimierungsverfahren ganz anderer Art sind die sogenannten genetischen Algorithmen. In ihrer üblichen Formulierung (vgl. [37]) scheinen sie bei dem hier gestellten Problem jedoch nur auf Parallelrechnern konkurrenzfähig zu sein.

3.6.2 Ein Koordinatenrelaxationsverfahren zum Test auf positive Definitheit

Grundlage dieses Verfahrens ist einerseits das Koordinatenrelaxationsverfahren von Schwarz [49] (*"The Method of Coordinate Overrelaxation for $(A - \lambda B)\mathbf{x} = \mathbf{0}$"*), dessen einzelne Schritte hier modifiziert werden, und andererseits die theoretischen Erkenntnisse aus der Arbeit von Kovač- Striko/Veselič [38] (*"Trace Minimization and Definiteness of Symmetric Pencils"*). Sie zeigen dort, daß die positive Definitheit eines Paares (A, B) symmetrischer (n,n)-Matrizen mit B indefinit in engem Zusammenhang steht mit der Existenz des Minimums von

$$\text{Spur}(\pm X^T A X) \text{ unter der Nebenbedingung } X^T B X = J_1$$

bei Variation von X, wobei J_1 eine gegebene Diagonalmatrix der Ordnung $\leq n$ mit $J_1^2 = I$ ist.

Anders als in Abschnitt 3.6.1 ist jetzt die (äquivalente) Formulierung als Optimierungsaufgabe mit Nebenbedingung günstiger. Zu bestimmen sind also die folgenden Extrema:

$$(3.11) \qquad \begin{array}{ll} \min\ \mathbf{x}^T A\mathbf{x} & \text{unter der Nebenbedingung } \mathbf{x}^T B\mathbf{x} = 1 \\ \max\ \mathbf{x}^T A\mathbf{x} & \text{unter der Nebenbedingung } \mathbf{x}^T B\mathbf{x} = -1 \end{array} .$$

Statt über beliebige $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{x}^T B\, \mathbf{x} = 1$ zu minimieren bzw. mit $\mathbf{x}^T B\, \mathbf{x} = -1$ zu maximieren, betrachten wir wie bei der Koordinatenrelaxation Vektoren

$$\mathbf{x}' = \varphi\, \mathbf{x} + \gamma\, \mathbf{y}\,, \quad \mathbf{x}, \mathbf{y} \text{ Startvektoren,}$$

und suchen (φ, γ) mit

$$\mathbf{x}'^T A\, \mathbf{x}' \overset{!}{=} \min \quad \text{unter der Nebenbedingung } \mathbf{x}'^T B\, \mathbf{x}' = 1$$

(3.12) bzw.

$$\mathbf{x}'^T A\, \mathbf{x}' \overset{!}{=} \max \quad \text{unter der Nebenbedingung } \mathbf{x}'^T B\, \mathbf{x}' = -1.$$

Ausmultiplizieren und Ordnen führt uns auf ein verallgemeinertes 2x2-Eigenwertproblem, wenn wir die Extrema mit Hilfe Lagrange'scher Multiplikatoren bestimmen:

$$f(\varphi, \gamma) := \mathbf{x}^T A\, \mathbf{x}' = \varphi^2\, \mathbf{x}^T A\, \mathbf{x} + 2\, \gamma\, \varphi\, \mathbf{y}^T A\, \mathbf{x} + \gamma^2\, \mathbf{y}^T A\, \mathbf{y} \overset{!}{=} \text{Extremum}$$

unter der Nebenbedingung

$$g(\varphi, \gamma) := \mathbf{x}'^T B\, \mathbf{x}' = \varphi^2\, \mathbf{x}^T B\, \mathbf{x} + 2\, \gamma\, \varphi\, \mathbf{y}^T B\, \mathbf{x} + \gamma^2\, \mathbf{y}^T B\, \mathbf{y} = 1 \quad (\text{oder } -1)$$

führt dann zu

$$\begin{bmatrix} \mathbf{x}^T A\, \mathbf{x} & \mathbf{x}^T A\, \mathbf{y} \\ \mathbf{y}^T A\, \mathbf{x} & \mathbf{y}^T B\, \mathbf{y} \end{bmatrix} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x}^T B\, \mathbf{x} & \mathbf{x}^T B\, \mathbf{y} \\ \mathbf{y}^T B\, \mathbf{x} & \mathbf{y}^T B\, \mathbf{y} \end{bmatrix} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix} \tag{3.13}$$

mit $\mathbf{x}'^T B\, \mathbf{x}' = 1$ (oder -1).

Der Grundalgorithmus zum Testen des Paares (A, B) auf positive Definitheit und zur Bestimmung des Definitheitsintervalls erhält jetzt die folgende Gestalt:

1. Man wähle einen Startvektor $\mathbf{x}_1$ mit $\mathbf{x}_1^T B\, \mathbf{x}_1 = 1$ und einen Startvektor $\mathbf{x}_2$ mit $\mathbf{x}_2^T B\, \mathbf{x}_2 = -1$.
   Man nehme als Startintervall $I_{Start} := (-\infty, +\infty)$.

2. Man bestimme Iterationsvektoren $\mathbf{y}_1$ und $\mathbf{y}_2$.
   Die typische Wahl ist $\mathbf{y} = \mathbf{e}_j$ mit $\mathbf{e}_j$ j-ter Einheitsvektor, j = 1, ..., n. (Eine andere Möglichkeit wären Zufallsvektoren oder Gradientenvektoren).

3. Man löse das 2x2-Eigenwertproblem

$$\begin{bmatrix} \mathbf{x}^T A\, \mathbf{x} & \mathbf{x}^T A\, \mathbf{y} \\ \mathbf{y}^T A\, \mathbf{x} & \mathbf{y}^T A\, \mathbf{y} \end{bmatrix} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x}^T B\, \mathbf{x} & \mathbf{x}^T B\, \mathbf{y} \\ \mathbf{y}^T B\, \mathbf{x} & \mathbf{y}^T B\, \mathbf{y} \end{bmatrix} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix},$$

d.h.

$$A_{klein} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix} = \lambda\, B_{klein} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix}$$

für $\mathbf{x} = \mathbf{x}_1$ und für $\mathbf{x} = \mathbf{x}_2$.

Für $\mathbf{x} = \mathbf{x}_1$ (bzw. $\mathbf{x} = \mathbf{x}_2$) wähle man den kleinsten (bzw. größten) Eigenwert mit Eigenvektor $\mathbf{z} \in \mathbb{R}^2$ aus, für den

$$\mathbf{z}^T B_{klein}\, \mathbf{z} = 1 \quad (\text{bzw. } \mathbf{z}^T B_{klein}\, \mathbf{z} = -1)$$

gilt; wir bezeichnen ihn mit λ_{Plus} (bzw. λ_{Minus}).

Das Verfahren bricht ab, wenn das 2x2-Problem nicht im Reellen lösbar ist. Existieren λ_{Plus} und λ_{Minus} und ist λ_{Minus} kleiner als λ_{Plus}, so setzen wir

$$I_{Defint} := (\lambda_{Minus}, \lambda_{Plus})$$

und bilden den Durchschnitt von I_{Start} und I_{Defint}:

$$I_{Start} := I_{Start} \cap I_{Defint}$$

Wenn das neue I_{Start} leer ist oder $\lambda_{plus} \le \lambda_{minus}$, so höre man ebenfalls auf.

4. Man definiere je einen neuen Startvektor $\mathbf{x}_1^{neu}$, $\mathbf{x}_2^{neu}$ nach der Vorschrift

$$\mathbf{x}_i^{neu} := (\mathbf{x}_i, \mathbf{y}_i)\, \mathbf{z}_i = \varphi_i\, \mathbf{x}_i + \gamma_i\, \mathbf{y}_i\,,$$

   wobei $\mathbf{z}_i$, i = 1, 2 die zuvor ausgerechneten Eigenvektoren sind.
   Für die so erzeugten Vektoren $\mathbf{x}_i^{neu}$, i = 1, 2 gilt:

$$(\mathbf{x}_i^{neu})^T\, B\, \mathbf{x}_i^{neu} = \pm 1$$

5. Solange der Algorithmus nicht abbricht und das Definitheitsintervall nicht genau genug bestimmt ist, führe man die Schritte 2, 3 und 4 durch.

Für eine praktische Implementierung benötigt man neben einer geeigneten Abbruchbedingung noch eine Überprüfung, ob tatsächlich globale Minima (bzw. Maxima) gefunden wurden oder nur stationäre Punkte. (Hierzu und allgemein zu verallgemeinerten Koordinatenrelaxationsverfahren vgl. man Gose [30] und Quick, Matlala und Miller [47].)

Für den Fall hermitescher Matrizen sei auf den Artikel "Finding a Positive Definite Linear Combination of Two Hermitian Matrices" von Moon, Crawford [44] verwiesen. Interessant hierzu ist auch die Arbeit "An Algorithm for Rotating a Symmetrical Generalized Eigenvalue Problem" von Moon [43].

Da man aber meist nicht nur das Definitheitsintervall bestimmen will, sondern das Eigenwertproblem $A\,\mathbf{x} = \lambda\, B\,\mathbf{x}$ vollständig lösen möchte, wäre ein Verfahren günstiger, das in seinen einzelnen Schritten Tests auf positive Definitheit durchführt, abbricht, wenn das Paar (A, B) nicht positiv definit ist, und andernfalls alle Eigenwerte und Eigenvektoren berechnet.

Für kleine bis mittlere Matrizenpaare ist das verallgemeinerte Jacobi-Verfahren solch ein Verfahren (vgl. Abschnitt 3.4), für kleine bis sehr große Matrizen - vor allem, wenn A tridiagonal ist - wird es das in dieser Arbeit beschriebene, sehr schnelle Divide-et-Impera-Verfahren sein.

4. Der Divide-et-Impera-Algorithmus für ein Matrizenpaar (A, J) mit Tridiagonalmatrix A

Wir verallgemeinern das Standard-Divide-et-Impera-Verfahren (zum Eigenwertproblem $T\mathbf{x} = \lambda\mathbf{x}$, T symmetrische Tridiagonalmatrix) auf den Fall $T\mathbf{x} = \lambda J\mathbf{x}$ mit J Diagonalmatrix mit 1 und -1 auf der Diagonalen. Wie das verallgemeinerte Jacobi-Verfahren findet unser verallgemeinertes Divide-et-Impera-Verfahren für den Fall eines positiv definiten Paares (T, J) immer eine Lösung, andernfalls entdeckt es das Verletztsein der positiven Definitheit. Da das Verfahren auf Tridiagonalmatrizen zugeschnitten ist, können - anders als beim Jacobi-Verfahren - Matrizen sehr großer Dimension behandelt werden.

In diesem Kapitel wollen wir uns mit einem "Divide-et-Impera"-Verfahren zur Lösung des verallgemeinerten Eigenwertproblems befassen. Dieses Prinzip des "Teile-und-herrsche" war schon den alten Römer sehr geläufig, und als "Divide-and-Conquer" wurde es *das* Herrschaftsmittel der Briten in Indien. In der Informatik versteht man darunter ein algorithmisches Lösungsverfahren, das sich in die beiden folgenden Schritte gliedert:

Divide-Schritt: Das Problem wird in zwei (oder mehr) möglichst gleichgroße Teilprobleme derselben Art wie das Originalproblem aufgespalten.
Diese kleineren Teilprobleme werden unabhängig voneinander gelöst.

Impera-Schritt : Die Lösungen der Teilprobleme werden zu einer Lösung des Gesamtproblems zusammengefügt.

Ein typisches Divide-et-Impera-Verfahren ist der bekannte Sortieralgorithmus *Quicksort*.

4.1 Das eigentliche Verfahren

Durch Rang-Eins-Abspaltungen wird das Gesamteigenwertproblem $T\mathbf{x} = \lambda J\mathbf{x}$ in zwei- und eindimensionale Eigenwertprobleme zerlegt (Divide-Schritt). Nachdem sie gelöst sind, werden je zwei dieser kleinen Eigenwertprobleme zusammen mit der entsprechenden Rang-Eins-Störung im Impera-Schritt behandelt. Die Eigenwerte dieses so entstandenen Eigenwertproblems ergeben sich als Nullstellen einer speziellen rationalen Funktion. Die Eigenvektoren erhält man anschließend durch eine endliche Zahl arithmetischer Operationen. Durch Zusammensetzen zu immer größeren Matrizen gelangt man schließlich zur Lösung des ursprünglichen Eigenwertproblems.

In der Matrizentheorie wurde ein Divide-et-Impera-Algorithmus zur Bestimmung aller Eigenwerte und Eigenvektoren einer symmetrischen Tridiagonalmatrix $A \in M^{n,n}(\mathbb{R})$ von Bunch, Nielsen, Sorensen und Cuppen angegeben, vgl. Bunch, Nielsen, Sorensen [13], Cuppen [19] und Golub [28]. Dabei wird die Matrix A zerlegt in

$$A = \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} + \rho\, \mathbf{b}\mathbf{b}^T$$

mit quadratischen Matrizen T_1, T_2 der Dimensionen m_1, m_2 ($m_1 + m_2 = n$), $\rho \in \mathbb{R}$ und $\mathbf{b} \in \mathbb{R}^n$.

Jetzt löst man zunächst die kleineren Eigenwertprobleme

$$T_1 \mathbf{x} = \lambda \mathbf{x} \quad \text{und} \quad T_2 \mathbf{x} = \lambda \mathbf{x}\,.$$

(Wir werden dies rekursiv durch Zurückführung auf zwei- oder eindimensionale Eigenwertprobleme tun; man könnte auch für genügend kleine Matrizen eines der Standard-Verfahren wie z.B Jacobi-oder QZ-Verfahren benutzen.)

Danach enthält die Diagonalmatrix $D_1 \in M^{m_1,m_1}(\mathbb{R})$ die Eigenwerte von T_1, die Diagonalmatrix $D_2 \in M^{m_2,m_2}(\mathbb{R})$ die von T_2. Die zugehörigen Eigenvektoren faßt man in zwei orthogonalen Matrizen Q_1 bzw. Q_2 aus $M^{m_1,m_1}(\mathbb{R})$ bzw. $M^{m_2,m_2}(\mathbb{R})$ zusammen.

Nach diesem Divide-Schritt kann man A in der Form

$$A = Q(D + \rho\, \mathbf{z}\mathbf{z}^T)\, Q^T$$

darstellen mit

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}, \quad \mathbf{z} = Q^T\mathbf{b}.$$

Da die Eigenwerte von A und von $D + \rho\, \mathbf{z}\mathbf{z}^T$ dieselben sind, berechnet man jetzt im Impera-Schritt die Eigenwerte und Eigenvektoren von $D + \rho\, \mathbf{z}\mathbf{z}^T$. Ist $\mathbf{x}$ ein Eigenvektor von $D + \rho\, \mathbf{z}\mathbf{z}^T$, so ist $Q\mathbf{x}$ ein Eigenvektor von A.

Im folgenden beschäftigen wir uns mit dem verallgemeinerten Eigenwertproblem

(4.1) $\qquad A\mathbf{x} = \lambda J \mathbf{x}\,.$

Dabei soll A eine symmetrische Tridiagonalmatrix aus $M^{n,n}(\mathbb{R})$ sein, die keineswegs positiv definit zu sein hat, und J eine Diagonalmatrix mit Plus- oder Minuseinsen auf der Diagonalen.

Wir wissen zunächst nicht, ob das Paar (A, J) positiv definit ist. Wir wollen jetzt einen verallgemeinerten Divide-et-Impera-Algorithmus (kurz mit VDI-Algorithmus bezeichnet) entwickeln, der dies entscheidet und im Fall der Positiv-Definitheit alle Eigenwerte und Eigenvektoren dieses Eigenwertproblems berechnet.

O.B.d.A. bezeichnen wir im folgenden ein Matrizenpaar (A, J) als positiv definit, wenn es ein $\mu \in \mathbb{R}$ gibt, so daß die Linearkombination $A - \mu J$ positiv definit ist.

4.1.1 Der Divide-Schritt:

Wie bei dem zuvor skizzierten Divide-et-Impera–Verfahren für das einfache Eigenwertproblem zerlegen wir die Matrix A in

$$(4.2) \quad A = \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} - \rho\mathbf{b}\mathbf{b}^T = T - \rho\mathbf{b}\mathbf{b}^T \quad \text{mit} \quad T = \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix},$$

$T_1 \in M^{m_1,m_1}(\mathbb{R})$, $T_2 \in M^{m_2,m_2}(\mathbb{R})$, $m_1 + m_2 = n$, $1 \le m_1 \le n-1$, $\mathbf{b} \in \mathbb{R}^n$ und $\rho \in \mathbb{R}^+$.

(Die genauere Aufteilung von n in $m_1 + m_2$ wird in Abschnitt 4.1.4 diskutiert.)

Dabei gehen wir folgendermaßen vor:

Zunächst schreiben wir A in der Form

$$A = \left[\begin{array}{cccccc|cccc} a_1 & b_1 & & & & & & & & \\ b_1 & a_2 & b_2 & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & \ddots & \ddots & \ddots & & & & \\ & & & & & a_m & b_m & & & \\ \hline & & & & & b_m & a_{m+1} & \ddots & & \\ & & & & & & \ddots & \ddots & \ddots & \\ & & & & & & & \ddots & \ddots & a_{n-1} \; b_{n-1} \\ & & & & & & & & & b_{n-1} \; a_n \end{array}\right]$$

$$= \left[\begin{array}{c|c} A_1 & b_m \\ \hline b_m & A_2 \end{array}\right].$$

Ist eines der b_i, $i = 1, \ldots, n-1$, gleich null, so zerfällt das Eigenwertproblem direkt in zwei Teilprobleme, die wir dann getrennt lösen können. Wir gehen also im folgenden davon aus, daß *alle* b_i, $i = 1, \ldots, n-1$ von null verschieden sind.

Wenn wir A an der Stelle $m = m_1$ aufteilen, ändern wir nur die Elemente a_m, a_{m+1}, b_m, um zur Matrix

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$$

zu gelangen.

Es genügt offensichtlich, sich die Aufteilung an der folgenden 4 × 4- Matrix zu verdeutlichen:

$$\left[\begin{array}{cc|cc} a_{m-1} & b_{m-1} & 0 & 0 \\ b_{m-1} & a_m & b_m & 0 \\ \hline 0 & b_m & a_{m+1} & b_{m+1} \\ 0 & 0 & b_{m+1} & a_{m+2} \end{array}\right]$$

$$= \left[\begin{array}{cc|cc} a_{m-1} & b_{m-1} & 0 & 0 \\ b_{m-1} & a_m + |b_m| & 0 & 0 \\ \hline 0 & 0 & a_{m+1} + |b_m| & b_{m+1} \\ 0 & 0 & b_{m+1} & a_{m+2} \end{array}\right]$$

$$- |b_m| \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 0 & 1 & -\operatorname{sgn} b_m & 0 \\ \hline 0 & -\operatorname{sgn} b_m & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

$$= \left[\begin{array}{cc|cc} a_{m-1} & b_{m-1} & 0 & 0 \\ b_{m-1} & a_m + |b_m| & 0 & 0 \\ \hline 0 & 0 & a_{m+1} + |b_m| & b_{m+1} \\ 0 & 0 & b_{m+1} & a_{m+2} \end{array}\right]$$

$$- |b_m| \begin{bmatrix} 0 \\ 1 \\ -\operatorname{sgn} b_m \\ 0 \end{bmatrix} \left[\, 0,\ 1,\ -\operatorname{sgn} b_m,\ 0 \,\right] .$$

Allgemein ist also eine Darstellung $T - \rho \mathbf{b}\mathbf{b}^T$ mit

$$\rho = |b_m| \quad \text{und} \quad \mathbf{b} = \begin{array}{c} \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ -\operatorname{sgn} b_m \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \leftarrow \text{m-te Position} \\ \leftarrow \text{m+1-te Position} \\ \\ \\ \\ \\ \\ \end{array}$$

möglich.

T_1 stimmt bis auf das m-te Diagonalelement mit A_1, T_2 bis auf das erste Diagonalelement mit A_2 überein.

Beim Standardeigenwertproblem ist die Zerlegung in die zwei Terme

$$\begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} \quad \text{und} \quad \textit{Rang-Eins-Störung} \quad \rho \mathbf{b}\mathbf{b}^T$$

verhältnismäßig frei wählbar, Cuppen [19] wählt z.B. für $\mathbf{b}^T$ den Vektor $(0, \ldots, 0, 1, 1, 0, \ldots, 0)$ und $\rho = b_m$.

Um die Positivdefinitheit der Teilprobleme zu gewährleisten, ist beim verallgemeinerten Eigenwertproblem die Zerlegung etwas eingeschränkt. Die hier vorgestellte Wahl für das verallgemeinerte Eigenwertproblem ergab sich aus dem folgenden Grund: Unsere Zerlegung $A = T - \rho \mathbf{b}\mathbf{b}^T$ darf auf keinen Fall dazu führen, daß das Paar (T, J) nicht positiv definit ist, während (A, J) es doch ist. Denn in diesem Fall wäre nicht mehr sichergestellt, daß die "kleinen" Eigenwertprobleme

$$T_1 \mathbf{x} = \lambda J_1 \mathbf{x}, \quad T_2 \mathbf{x} = \lambda J_2 \mathbf{x} \quad \text{mit} \quad J = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}$$

lösbar sind. Unsere Wahl von ρ und $\mathbf{b}$ sichert die Vererbung der positiven Definitheit von (A, J) auf (T, J), da die Diagonalelemente von T im Vergleich zu A nur vergrößert werden und die Vergrößerung eines Diagonalelements einer positiv definiten Matrix die positive Definitheit nicht zerstört.

Wenn also eines der Matrixpaare (T_1, J_1) oder (T_2, J_2) nicht positiv definit ist, dann wissen wir, daß auch das Ausgangspaar (A, J) nicht positiv definit gewesen ist.

Nach dieser Aufteilung von A in $T - \rho \mathbf{b}\mathbf{b}^T$ behandeln wir jetzt die "kleinen" Eigenwertprobleme

$$(4.3) \quad T_1 \mathbf{x} = \lambda J_1 \mathbf{x} \quad \text{und} \quad (4.4) \quad T_2 \mathbf{x} = \lambda J_2 \mathbf{x} .$$

Diese Eigenwertprobleme sind lösbar, wenn die jeweiligen Matrizenpaare positiv definit sind. Mit Q_1 und Q_2 bezeichnen wir die Matrizen der J_1- bzw. J_2-normierten Eigenvektoren der Eigenwertprobleme (4.3) und (4.4). Die Eigenwerte stehen auf der Diagonalen der Diagonalmatrizen $D_1 J_1$ bzw. $D_2 J_2$. Dann gilt:

$$(4.5) \quad Q_1^T T_1 Q_1 = D_1 , \quad Q_1^T J_1 Q_1 = J_1$$

$$(4.6) \quad Q_2^T T_2 Q_2 = D_2 , \quad Q_2^T J_2 Q_2 = J_2$$

Dies können wir zusammenfassen zu

$$(4.7) \quad Q^T T Q = D \quad \text{und} \quad Q^T J Q = J \quad \text{mit} \quad Q = \left[\begin{array}{c|c} Q_1 & 0 \\ \hline 0 & Q_2 \end{array}\right], \quad D = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & D_2 \end{array}\right] .$$

Damit ist unser Divide-Schritt (inklusive der Diagonalisierung der Teilprobleme) beendet.

4.1.2 Der Impera-Schritt

Wie wir jetzt gleich sehen werden, kann man das Eigenwertproblem

$$A\mathbf{x} = (T - \rho\mathbf{b}\mathbf{b}^T)\,\mathbf{x} = \lambda J\mathbf{x}$$

schreiben als

(4.8) $\quad W(D - \rho\mathbf{z}\mathbf{z}^T)W^T\mathbf{x} = \lambda WJW^T\mathbf{x}$

mit geeignet gewähltem $W \in M^{n,n}(\mathbb{R})$ und $\mathbf{z} \in \mathbb{R}^n$ und der Diagonalmatrix D aus dem Divide-Schritt. Die Eigenwerte von (A, J) und die Eigenwerte des Paares ($D - \rho\mathbf{z}\mathbf{z}^T$, J) sind die gleichen, die Eigenvektoren des Paares (A, J) lassen sich leicht aus denen von ($D - \rho\mathbf{z}\mathbf{z}^T$, J) berechnen.
Aus der Formel (4.7) folgt nämlich die folgende Darstellung für T und J :

(4.9) $\quad T = (Q^T)^{-1}DQ^{-1}\,,\; J = (Q^T)^{-1}JQ^{-1}$

Nun ist wegen $Q^TJQ = J$ gerade

(4.10) $\quad Q^{-1} = JQ^TJ$ und $Q^{-T} = JQJ$

und damit

(4.11) $\quad (JQ)^{-1} = JQ^T$ und $(JQ)^{-T} = QJ.$

Einsetzen in (4.9) ergibt für T und J die Darstellungen

(4.12) $\quad T = Q^{-T}DQ^{-1} = JQJ\;D\;JQ^TJ = (JQ)D(JQ)^T$

(4.12') $\quad J = (JQ)J(JQ)^T$.

Das Eigenwertproblem

$$(T - \rho\mathbf{b}\mathbf{b}^T)\mathbf{x} = \lambda J\mathbf{x}$$

können wir jetzt schreiben als

$$(JQ)[D - \rho(JQ)^{-1}\mathbf{b}\mathbf{b}^T(JQ)^{-T}]\,(JQ)^T\mathbf{x} = \lambda(JQ)J(JQ)^T\mathbf{x}$$

bzw. mit Formel (4.11) als

$$(JQ)[D - \rho(JQ^T)\mathbf{b}\mathbf{b}^T(QJ)](JQ)^T\mathbf{x} = \lambda(JQ)J(JQ)^T\mathbf{x}\,,$$

was

(4.13) $\quad (JQ)[D - \rho\mathbf{z}\mathbf{z}^T](JQ)^T\mathbf{x} = \lambda(JQ)J(JQ)^T\mathbf{x}$

und damit (4.8) mit $\mathbf{z} := JQ^T\mathbf{b}$ und $W := JQ$ liefert.

Da das Eigenwertproblem

$$A\mathbf{x} = (T - \rho\mathbf{b}\mathbf{b}^T)\mathbf{x} = \lambda J\mathbf{x}$$

und das Problem

$$(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y} \quad \text{mit } \mathbf{y} := (JQ)^T\mathbf{x}$$

dieselben Eigenwerte haben, lösen wir im folgenden das Eigenwertproblem

(4.14) $\quad (D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}.$

Hat man die Eigenvektoren $\mathbf{y}$ des Eigenwertproblems (4.14) berechnet, so erhält man die Eigenvektoren $\mathbf{x}$ des ursprünglichen Problems $A\mathbf{x} = \lambda J\mathbf{x}$ aus

(4.15) $\quad \mathbf{x} = (JQ)^{-T}\mathbf{y} = QJ\mathbf{y}$ gemäß Identität (4.11).

Wenn wir das vereinfachte Eigenwertproblem (4.14) gelöst und mit Hilfe von (4.15) alle Eigenvektoren $\mathbf{x}$ des ursprünglichen Eigenwertproblems berechnet haben, ist der Impera-Schritt beendet.

4.1.3 Die Lösung des verallgemeinerten Eigenwertproblems $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$

Die Hauptaufgabe besteht nun in der Lösung des verallgemeinerten Eigenwertproblems

(4.16) $\quad (D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$.

(Man beachte, daß bei unserer Konstruktion ρ immer größer oder gleich null ist.)

Im folgenden gehen wir jetzt von $\rho > 0$ aus, denn für $\rho = 0$ haben wir das Eigenwertproblem $D\mathbf{y} = \lambda J\mathbf{y}$, dessen Lösung wir direkt ablesen können. Desweiteren setzen wir zunächst die zwei Eigenschaften

(a) alle Komponenten z_i von $\mathbf{z} \in \mathbb{R}^n$ sind ungleich null
(b) alle Eigenwerte des Paares (D, J) sind verschieden

voraus.

Diese Bedingungen müssen keineswegs immer erfüllt sein. Über $\mathbf{z}$ wissen wir nur, daß $\mathbf{z} = JQ^T\mathbf{b}$ ist, Nullkomponenten sind also durchaus möglich, das Paar (D, J) kann natürlich auch Eigenwerte mit Vielfachheit größer eins haben. Wie wir gleich sehen werden, kann man in diesen Fällen das ursprüngliche Problem durch eine jeweils geeignete Deflation auf eines mit niedrigerer Dimension zurückführen, das (a) und (b) sicherstellt. (Die Deflation verringert außerdem den Rechenaufwand!)

Das Eigenwertproblem $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$ läßt sich auch schreiben als

(4.17) $\quad (D - \lambda J)\mathbf{y} = \rho\,\mathbf{z}(\mathbf{z}^T\mathbf{y})$.

Für die weitere Rechnung ist es entscheidend, daß die Matrix $D - \lambda J$ invertierbar ist für alle Eigenwerte λ von $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$:

Lemma 4.1: *Für jeden Eigenwert λ des Eigenwertproblems $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$ gilt unter den Voraussetzungen (a) und (b):*
(i) $\quad \mathbf{z}^T\mathbf{y} \neq 0.$
(ii) $\quad$ die Matrix $D - \lambda J$ ist invertierbar.

Beweis:

Diese Aussage ergibt sich analog zu der Rechnung im Artikel von Cuppen [19] (Beweis von Theorem 2.1). Wir wiederholen den Beweis hier der Vollständigkeit halber und um zu zeigen, daß die Matrix J sich noch kaum auswirkt.

Wir nehmen an, daß die Matrix $D - \lambda J$ nicht invertierbar ist. Aus der Singularität von $D - \lambda J$ und der Voraussetzung (b) folgt, daß es ein $i \in \{1,\dots,n\}$ gibt mit $\lambda = d_i j_i$. Hieraus ergibt sich für (4.17)

$$0 = (d_i - \lambda j_i) y_i = ((D - \lambda J)\mathbf{y})_i = \rho(\mathbf{z}^T\mathbf{y}) z_i .$$

Nun sind nach Voraussetzung alle z_i's ungleich null und $\rho > 0$. Also muß $\mathbf{z}^T\mathbf{y} = 0$ sein. Hieraus folgt aber

$$((D - \lambda J)\mathbf{y})_j = 0 \text{ für } \textit{alle } j = 1,\dots,n.$$

Da alle Eigenwerte nach Voraussetzung (b) verschieden sind, gilt $y_j = 0$ für alle $j = 1,\dots,n$ mit $j \neq i$. Deshalb ist $0 = \mathbf{z}^T\mathbf{y} = z_i y_i$. Die Komponente y_i muß ungleich null sein, denn sonst könnte $\mathbf{y}$ kein Eigenvektor sein. Wir erhalten damit $z_i = 0$. Dies ist ein Widerspruch zur Voraussetzung (a). ↯

Die Matrix $D - \lambda J$ muß also invertierbar sein. In diesem Fall kann $\mathbf{z}^T\mathbf{y}$ nicht null sein, wie man aus der Gleichung (4.17) sieht: Wäre $\mathbf{z}^T\mathbf{y} = 0$, so folgte $(D - \lambda J)\mathbf{y} = \mathbf{0}$ und damit $\mathbf{y} = \mathbf{0}$. ↯

Q.E.D.

Unter den Voraussetzungen (a) und (b) können wir das *Eigenwertproblem*

$$(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$$

auf ein *Nullstellenproblem* zurückführen:

Aus

$$(D - \lambda J)\mathbf{y} = \rho(\mathbf{z}^T\mathbf{y})\mathbf{z}$$

folgt

$$\mathbf{y} = \rho(\mathbf{z}^T\mathbf{y})(D - \lambda J)^{-1}\mathbf{z} ,$$

da die Matrix $D - \lambda J$ invertierbar ist und $\mathbf{z}^T\mathbf{y}$ eine Zahl. Multiplikation der Gleichung mit dem Vektor $\mathbf{z}^T$, der keine Nullkomponenten hat, liefert

$$\mathbf{z}^T\mathbf{y} = \rho(\mathbf{z}^T\mathbf{y})\mathbf{z}^T(D - \lambda J)^{-1}\mathbf{z} .$$

Da $\mathbf{z}^T\mathbf{y} \neq 0$ ist, erhalten wir

$$1 = \rho\mathbf{z}^T(D - \lambda J)^{-1}\mathbf{z} .$$

Die Eigenwerte von $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J\mathbf{y}$ sind also gerade die Nullstellen der durch

$$(4.18) \quad w(\lambda) := 1 - \rho\mathbf{z}^T(D - \lambda J)^{-1}\mathbf{z} = 1 - \rho\sum_{i=1}^{n}\frac{z_i^2}{d_i - \lambda j_i}$$

definierten Funktion w.

4.1.3.1 Minuseigenwerte, Pluseigenwerte und das Definitheitsintervall

Die Eigenwerte α_i, $i = 1,\dots,n$, des Paares (T, J) sind dieselben wie die von (D, J) und daher gegeben durch

$$\alpha_i = d_i j_i , \; i = 1,\dots,n .$$

Sei jetzt p die Anzahl der Plus-Einsen und q die der Minus-Einsen auf der Diagonalen von $J = \mathrm{diag}(j_1,\dots,j_n)$ mit $p + q = n$. Da das Paar (T, J) bzw. (D, J) positiv definit ist und alle Eigenwerte verschieden sind, können wir die α_i, $i = 1,\dots,n$, - wie im Artikel von Kovač-Striko und Veselič [38] beschrieben - folgendermaßen anordnen:

$$(4.19) \quad \alpha_1^- < \alpha_2^- < \dots < \alpha_{q-1}^- < \alpha_q^- < \alpha_1^+ < \alpha_2^+ < \dots < \alpha_p^+$$

Dabei bezeichnet man $\alpha_1^-, \dots, \alpha_q^-$ als *Minuseigenwerte*, denn für ihre zugehörigen Eigenvektoren $\mathbf{u}$ gilt $\mathbf{u}^T J\mathbf{u} < 0$, und $\alpha_1^+, \dots, \alpha_p^+$ als *Pluseigenwerte* wegen $\mathbf{u}^T J\mathbf{u} > 0$. Das Intervall, wo Minus- und Pluseigenwerte aufeinandertreffen, in unserer Schreibweise das Intervall $[\alpha_q^-, \alpha_1^+]$, nennt man das *Definitheitsintervall* des Paares (T, J) bzw. von (D, J). Für jedes $\mu \in (\alpha_q^-, \alpha_1^+)$ ist die Matrix $T - \mu J$ bzw. $D - \mu J$ positiv definit.

4.1.3.2 Eigenschaften der Funktion w

Über den Verlauf einer Funktion der Gestalt

$$F(x) = 1 - \rho\sum_{j=1}^{n}\frac{a_j}{\beta_j - x}$$

läßt sich für beliebige $\beta_j, a_j \in \mathbb{R}$ zunächst wenig aussagen. Man weiß nur, wo die Polstellen liegen und daß

$$\lim_{x\to-\infty} F(x) = \lim_{x\to\infty} F(x) = 1$$

gilt. Über die Lage der reellen Nullstellen und ihre Anzahl sind allgemein aber keine genaueren theoretischen Aussagen möglich (vgl. hierzu auch die entsprechenden Bemerkungen in Abschnitt 4.4).

Für unsere Funktion w haben wir unter der Voraussetzung der positiven Definitheit von (A, J) zum Glück weitere Informationen, etwa über die Anordnung der α_i und j_i, $i = 1, \ldots, n$. Unter Berücksichtigung von $\alpha_i = d_i j_i$ können wir daher w folgendermaßen darstellen :

$$(4.20) \qquad w(x) = 1 - \rho \left(\sum_{\substack{i=1 \\ (j_i=+1)}}^{n} \frac{z_i^2(+1)}{d_i - x} + \sum_{\substack{i=1 \\ (j_i=-1)}}^{n} \frac{z_i^2(-1)}{-d_i - x} \right)$$

$$= 1 + \rho \sum_{j=1}^{q} \frac{z_j^2}{\alpha_j - x} - \rho \sum_{j=q+1}^{n} \frac{z_j^2}{\alpha_j - x}$$

In der ersten Summe stehen die Minuseigenwerte $\alpha_j = \alpha_j^-$, $j = 1, \ldots, q$, in der zweiten die Pluseigenwerte $\alpha_{q+1} = \alpha_1^+, \ldots, \alpha_n = \alpha_p^+$.

Wir unterscheiden nun drei Fälle:

1. Fall: $p = n$:

In diesem Fall ist $q = 0$ und damit $J = I$. Es liegt also das Standardeigenwertproblem

$$(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda \mathbf{y}$$

vor.

Die Funktion w mit

$$w(x) = 1 - \rho \sum_{j=1}^{n} \frac{z_j^2}{\alpha_j - x}$$

ist auf jedem einzelnen der Intervalle, auf denen sie definiert ist, streng monoton fallend, alle Nullstellen von w liegen im Intervall $[\alpha_1 - \rho \mathbf{z}^T\mathbf{z}, \alpha_n]$, vgl. etwa Bunch, Nielsen, Sorensen [13].

2. Fall: $q = n$

Jetzt ist $J = -I$, es liegt wieder ein Standardeigenwertproblem vor:

$$(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = -\lambda \mathbf{y} \quad \text{bzw.} \quad (-D + \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda \mathbf{y}$$

Die Funktion w mit

$$w(x) = 1 + \rho \sum_{j=1}^{n} \frac{z_j^2}{\alpha_j - x}$$

ist streng monoton wachsend auf jedem einzelnen Intervall, auf dem sie definiert ist, alle Nullstellen liegen im Intervall $[\alpha_1, \alpha_n + \rho \mathbf{z}^T\mathbf{z}]$.

Weitere Ergebnisse über die Funktion w im Fall $J = I$ findet man in den Arbeiten von Bunch, Nielsen, Sorensen [13] und Cuppen [19].

Neu ist der Fall, bei dem es sowohl Pluseinsen als auch Minuseinsen gibt:

3. Fall: $0 < q < n$:

Die Funktion w hat jetzt die Gestalt

$$w(x) = 1 + \rho \sum_{j=1}^{q} \frac{z_j^2}{\alpha_j - x} - \sum_{j=q+1}^{n} \frac{z_j^2}{\alpha_j - x} .$$

Wenn α_i, $i = 1, \ldots, q$, Minuseigenwert ist, gilt:

$$\lim_{x \to \alpha_i -} w(x) = +\infty \quad \text{und} \quad \lim_{x \to \alpha_i +} w(x) = -\infty$$

Für die Pluseigenwerte α_i, $i = q+1, \ldots, n$ ergibt sich umgekehrt:

$$\lim_{x \to \alpha_i -} w(x) = -\infty \quad \text{und} \quad \lim_{x \to \alpha_i +} w(x) = +\infty$$

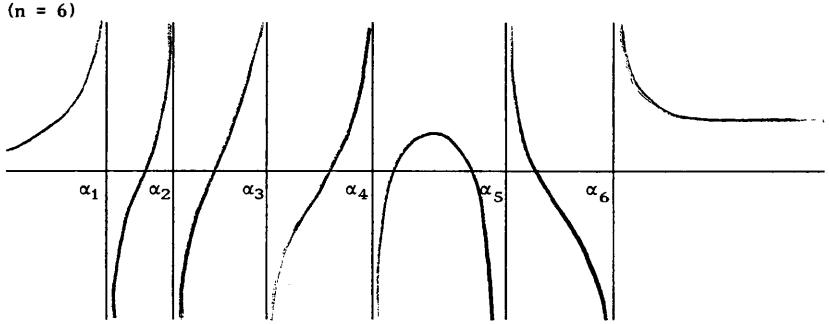
Da

$$w'(x) = \rho \sum_{j=1}^{q} \frac{z_j^2}{(\alpha_j - x)^2} - \rho \sum_{j=q+1}^{n} \frac{z_j^2}{(\alpha_j - x)^2}$$

aus Termen unterschiedlichen Vorzeichens besteht, können wir keine Aussage über die Monotonie von w in den Intervallen (α_i, α_{i+1}), $i = 1, \ldots, n-1$ machen.

Im wesentlichen muß die Funktion w für ein positiv definites Paar folgenden Verlauf haben :

(n = 6)



In dieser Skizze sind α_1, α_2, α_3, α_4 Minuseigenwerte und α_5, α_6 Pluseigenwerte.

Wegen des Vorzeichenwechsels von w zwischen je zwei Minuseigenwerten und zwischen je zwei Pluseigenwerten haben wir in $[\alpha_1, \alpha_n]$ auf jeden Fall $(q-1) + (p-1) = n-2$ Nullstellen. Im Definitheitsintervall können jetzt zwei, eine

oder keine Nullstelle liegen. Wenn das Paar ($D - \rho \mathbf{z}\,\mathbf{z}^T$, J) positiv definit ist, gibt es ein Definitheitsintervall, welches Teilmenge des alten Definitheitsintervalls $[\alpha_q\ ,\ \alpha_{q+1}\ ]$ ist. Wir untersuchen jetzt die Funktion w im Definitheitsintervall $[\alpha_q\ ,\ \alpha_{q+1}]$ genauer:

Zunächst stellt man sofort fest, daß die Funktion w konkav ist, denn

$$w''(x) = \rho \sum_{j=1}^{q} \frac{2z_j^2}{(\alpha_j - x)^3} - \rho \sum_{j=q+1}^{n} \frac{2z_j^2}{(\alpha_j - x)^3}$$

ist kleiner null, da die erste Summe negativ ist und die zweite positiv.

Außerdem nimmt w auf $I := [\alpha_q\ ,\ \alpha_{q+1}]$ sein Maximum an :

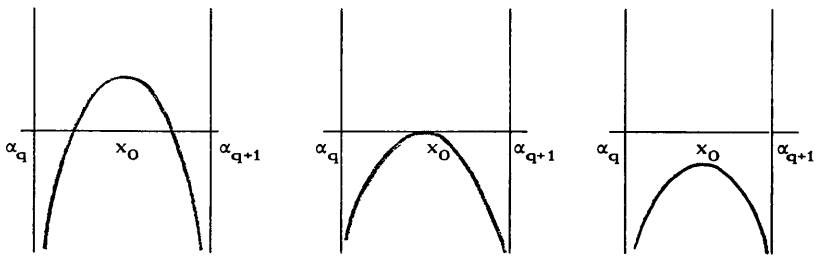
Da w" auf I kleiner null ist, ist die Funktion w' dort streng monoton fallend. Wegen

$$\operatorname{sgn}\,(\lim_{x \to \alpha_q^+} w'(x)) = +1 \quad \text{und} \quad \operatorname{sgn}\,(\lim_{x \to \alpha_{q+1}^-} w'(x)) = -1$$

hat w' einen Vorzeichenwechsel in I und damit genau eine Nullstelle wegen der Monotonie. Einsetzen dieser Nullstelle x_0 in w" liefert $w''(x_0)$ kleiner null und damit, daß x_0 die Maximumstelle von w in I ist.

Ist $w(x_0)$ positiv, so hat w zwei Nullstellen in I, für $w(x_0) = 0$ liegt eine doppelte Nullstelle vor und für $w(x_0) < 0$ keine. Das Definitheitsintervall für das Paar ($D - \rho \mathbf{z}\mathbf{z}^T$, J) existiert also nur für $w(x_0) > 0$. Ist $w(x_0) = 0$, so ist das Paar positiv semidefinit, für $w(x_0) < 0$ indefinit.

Wenn wir also im Definitheitsintervall nicht genau zwei reelle Nullstellen finden, so ist das Paar ($D - \rho \mathbf{z}\mathbf{z}^T$, J) nicht positiv definit und damit das Paar (A, J) auch nicht.



(positiv definites Paar) (semidefinites Paar) (indefinites Paar)

(Vergleiche hierzu auch den allgemeinen Satz 4.2.)

4.1.3.3 Deflation

Die gerade vorausgegangenen Rechnungen wurden unter den Voraussetzungen gemacht, daß

(a) alle Komponenten von $\mathbf{z} \in \mathbb{R}^n$ ungleich null

und

(b) alle Eigenwerte des Paares (D, J) verschieden sind.

Wir wollen jetzt untersuchen, was die Verletzung von (a) oder (b) bedeutet.

Zu a):

Wie wirkt sich $z_k = 0$ für ein $k \in \{1,\dots,n\}$ auf das Eigenwertproblem

$$(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$$

aus?

Wir können das Eigenwertproblem mit $z_k = 0$ schreiben als

$$\begin{bmatrix} d_1 - \rho z_1^2 & , z_1 z_2 & , \dots 0, \dots & z_1 z_n \\ z_1 z_2 & , d_2 - \rho z_2^2 & , \dots 0, \dots & z_2 z_n \\ \vdots & & & \\ 0 & , \quad 0 & , \dots d_k, \dots & 0 \\ \vdots & & & \\ z_1 z_n & , z_2 z_n & , \dots 0, \dots & d_n - \rho z_n^2 \end{bmatrix} \mathbf{y} = \lambda \begin{bmatrix} j_1 & & & & & \\ & j_2 & & & & \\ & & \ddots & & & \\ & & & j_k & & \\ & & & & \ddots & \\ & & & & & j_n \end{bmatrix} \mathbf{y}$$

Hieraus ersieht man sofort, daß $\lambda_k := d_k j_k$ Eigenwert zum Eigenvektor $\mathbf{e}_k$ ist:

$$(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{e}_k = d_k \mathbf{e}_k = \lambda_k j_k \mathbf{e}_k$$

Den k-ten Eigenwert und seinen Eigenvektor haben wir somit direkt erhalten. Nun gehen wir zu dem kleineren Eigenwertproblem über, das durch Streichen der k-ten Zeile und Spalte in $D - \rho \mathbf{z}\mathbf{z}^T$ und J entsteht.

Dies führt man für alle weiteren Indizes k mit $z_k = 0$, $k \in \{1,\dots,n\}$ durch. Für das am Ende entstandene Eigenwertproblem ist jetzt Voraussetzung (a) erfüllt.

Die Deflation für $z_k = 0$ kann also genauso wie im Fall eines einfachen Eigenwertproblems durchgeführt werden, vgl. Bunch, Nielsen und Sorensen [13], Cuppen [19].

Auch die Deflation für mehrfache Eigenwerte des Problems $D\mathbf{y} = \lambda \mathbf{y}$ läßt sich leicht auf unser verallgemeinertes Eigenwertproblem übertragen, da Plus- und Minuseigenwerte getrennt voneinander sind.

Ein mehrfacher Eigenwert α ist also immer Pluseigenwert oder immer Minuseigenwert. Daher können wir auch hier wieder eine Deflation durchführen:

Der Eigenwert α von (T,J) bzw. (D,J) habe die Vielfachheit $r \geq 2$ und sei entweder nur Pluseigenwert oder nur Minuseigenwert. $V \in M^{n,r}(\mathbb{R})$ sei die Matrix der Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_r$ zu α. Dann gilt für $j = 1, \ldots, r$:

$$\mathbf{v}_j^T J \mathbf{v}_j = \begin{cases} +1 \text{ , wenn } \alpha \text{ Pluseigenwert} \\ -1 \text{ , wenn } \alpha \text{ Minuseigenwert} \end{cases}$$

Man teile die Matrix Q aller Eigenvektoren des Paares (T, J) so auf und permutiere D und J derart, daß die Situation

$$Q = (V, Q_R) \text{ , } \quad J = \begin{bmatrix} J_V & 0 \\ 0 & J_R \end{bmatrix} = \begin{bmatrix} \pm I & 0 \\ 0 & J_R \end{bmatrix}$$

mit $V \in M^{n,r}(\mathbb{R})$, $Q_R \in M^{n,n-r}(\mathbb{R})$, $J_V \in M^{r,r}(\mathbb{R})$, $J_R \in M^{n-r,n-r}(\mathbb{R})$ vorliegt.

Da für ein positiv definites verallgemeinertes Eigenwertproblem die Eigenvektoren J-orthonormal sind, gilt

$$Q^T J Q = \begin{bmatrix} V^T \\ Q_R^T \end{bmatrix} J \begin{bmatrix} V, & Q_R \end{bmatrix} = \begin{bmatrix} V^T J V \text{ , } V^T J Q_R \\ Q_R^T J V \text{ , } Q_R^T J Q_R \end{bmatrix} = \begin{bmatrix} \pm I & 0 \\ 0 & J_R \end{bmatrix} = J .$$

Der Vektor $\mathbf{z}$ war gegeben durch $\mathbf{z} = J Q^T \mathbf{b}$. Teilen wir $\mathbf{z}$ auf in $\begin{bmatrix} \mathbf{z}_V \\ \mathbf{z}_R \end{bmatrix}$ mit $\mathbf{z}_V \in \mathbb{R}^r$ und $\mathbf{z}_R \in \mathbb{R}^{n-r}$, so können wir schreiben:

$$\mathbf{z} = J Q^T \mathbf{b} = \begin{bmatrix} \mathbf{z}_V \\ \mathbf{z}_R \end{bmatrix} = \begin{bmatrix} J_V & 0 \\ 0 & J_R \end{bmatrix} \begin{bmatrix} V^T \\ Q_R^T \end{bmatrix} \mathbf{b} = \begin{bmatrix} \pm V^T \mathbf{b} \\ J_R Q_R^T \mathbf{b} \end{bmatrix}$$

Unser Ziel ist nun, in $\mathbf{z}_V \in \mathbb{R}^r$ möglichst viele Nullen zu erzeugen, um eine Deflation - wie in (a) beschrieben - durchführen zu können. Aus diesem Grund wählen wir uns eine geeignete *Householdermatrix* $H \in M^{r,r}(\mathbb{R})$ mit

$$H \mathbf{z}_V = -\sigma \mathbf{e}_1 \text{ , } \sigma = \| \mathbf{z}_V \|_2 \text{ , } \quad \mathbf{e}_1 \in \mathbb{R}^r \text{ , } \quad H H^T = H^T H = I.$$

Die Matrix H hat die Gestalt $H = I - \gamma \mathbf{u}\mathbf{u}^T$ mit geeignet gewähltem Vektor $\mathbf{u}$ und $\gamma \in \mathbb{R}$. (Für genauere Informationen über Householdermatrizen sei auf die Bücher von Parlett [46] und Bunse-Gerstner [17]) verwiesen.)

Danach definieren wir $V_1 := V H^T$. Für dieses V_1 gilt:

(i) $\quad V_1^T \mathbf{b} = H V^T \mathbf{b} = H(\pm \mathbf{z}_V) = \pm(-\sigma \mathbf{e}_1) = \mp \sigma \mathbf{e}_1$

(ii) $\quad V_1^T J V_1 = H V^T J V H^T = H(\pm I) H^T = \pm H H^T = \pm I$

(iii) $\quad V_1^T J Q_R = H V^T J Q_R = 0$

Man kann jetzt ohne weiteres die Matrix Q durch $Q' := (V_1 \text{ , } Q_R)$ ersetzen, denn auch für Q' gilt

$$Q'^T J Q' = J \quad \text{und} \quad Q'^T T Q' = D.$$

Wenn wir den Vektor $\mathbf{z}$ mit Hilfe von Q' berechnen, so stellen wir fest, daß er r-1 Nullkomponenten hat:

$$\mathbf{z} = J Q'^T \mathbf{b} = \begin{bmatrix} \pm V_1^T \mathbf{b} \\ J_R Q_R^T \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mp \sigma \mathbf{e}_1 \\ J_R Q_R^T \mathbf{b} \end{bmatrix}$$

Die Deflation aus Teil (a) verkleinert dann das Eigenwertproblem um r-1 Zeilen und Spalten.

Wie man sich leicht überlegt, zerstört die Deflation die positive Definitheit nicht.

4.1.3.4 Bestimmung der Eigenvektoren von $(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$

In Abschnitt 4.1.3 haben wir gesehen, daß die Eigenvektoren proportional zu $(D - \lambda J)^{-1} \mathbf{z}$ sind (vgl. Formel 4.17). Ein (J-normierter) Eigenvektor $\mathbf{y}$ des Eigenwertproblems $(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$ zum Eigenwert λ berechnet sich also wie folgt:

$$(4.21) \qquad \mathbf{y} = \frac{(D - \lambda J)^{-1} \mathbf{z}}{\left| \mathbf{z}^T (D - \lambda J)^{-1} J (D - \lambda J)^{-1} \mathbf{z} \right|^{1/2}}$$

$$= \left(\frac{z_i}{d_i - \lambda j_i} \right)_{i=1,\ldots,n} \frac{1}{\left| \mathbf{z}^T (D - \lambda J)^{-1} J (D - \lambda J)^{-1} \mathbf{z} \right|^{1/2}}$$

Dabei gilt $\mathbf{y}^T J \mathbf{y} = \pm 1$.

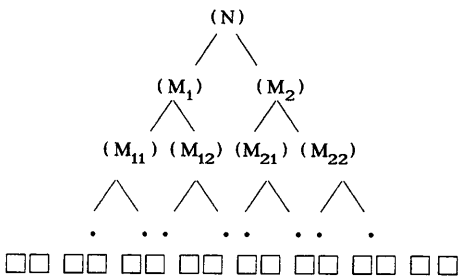
Wegen $\mathbf{x} = Q J \mathbf{y}$ (vgl. Formel 4.15) kann man jetzt schnell die Eigenvektoren des Eigenwertproblems $A\mathbf{x} = \lambda J \mathbf{x}$ berechnen.

4.1.4 Die Zerlegung des Eigenwertproblems $Ax = \lambda Jx$

In unserem Algorithmus bleiben wir nicht bei der Zerlegung des Eigenwertproblems $Ax = \lambda Jx$ in nur zwei Teilprobleme stehen, die durch eine Rang-Eins-Störung verbunden werden, sondern teilen diese kleineren Probleme weiter auf, bis wir zu 2x2- und 1x1-Problemen gelangen, die durch explizite Formeln gelöst werden.
Insgesamt entsteht der folgende Algorithmus:

Divide-et-Impera-Verfahren zur Lösung des Eigenwertproblems $Ax = \lambda Jx$ mit tridiagonaler Matrix A

1. Man zerlege das Eigenwertproblem $Ax = \lambda Jx$ in genügend kleine Teilprobleme:
   Dabei stehe im folgenden Baum
   (N) für das Eigenwertproblem $Ax = \lambda Jx$,
   (M) für ein Teilproblem der Dimension M,
   $\wedge$ für die Rang-Eins- Störung, die je zwei Teilprobleme verbindet,
   $\square$ für die "direkt" lösbaren "kleinen" Eigenwertprobleme der untersten Stufe.



2. Man löse die Teilprobleme der untersten Baumstufe direkt: dies könnte bei den 2x2 - Problemen mithilfe der charakteristischen Gleichung geschehen. Tests mit kritischen Beispielen haben jedoch gezeigt, daß es numerisch günstiger ist, *eine* geeignet umgeschriebene verallgemeinerte Jacobi-Rotation zu verwenden.

3. Solange man noch nicht bei der Wurzel (N), dem ursprünglichen Eigenwertproblem, angekommen ist, verbinde man je zwei Teilprobleme über die zugehörige Rang-Eins-Störung und löse das entstandene größere Problem der Gestalt $(D-\rho zz^T)x = \lambda Jx$ mit Hilfe eines Nullstellenverfahrens. Aus den so erhaltenen Eigenwerten und Eigenvektoren bestimme man die Lösung des aktuellen Teilproblems (M_i).

Wie man solch einen Problembaum geeignet aufstellt, hängt auch von der zur Verfügung stehenden Hardware und der verwendeten Programmiersprache ab. So würde man zunächst eine rekursive Programmierung als natürlich ansehen, mit deren Hilfe ein möglichst ausgeglichener Baum erzeugt wird. Da aber der rekursive Aufruf von Unterprogrammen sehr speicherplatzaufwendig ist und auch die Rechenzeit vergrößert, wurde eine nichtrekursive Formulierung verwendet. Diese hat den Vorteil, daß sie leichter parallelisierbar ist; außerdem stellt sie sehr früh fest, ob ein Teilproblem nicht positiv definit ist (vgl. nächsten Abschnitt). (Zu einem parallelen Divide-et-Impera-Verfahren vgl. Lo, Philippe und Sameh [41].)

4.1.5 Testen auf positive Definitheit

Für ein positiv definites Paar (T, J) berechnet das verallgemeinerte Divide-et-Impera-Verfahren alle Eigenwerte und Eigenvektoren. Man kann das Verfahren allerdings auch verwenden, wenn man nicht weiß, ob das Paar (T, J) positiv definit ist.
Die Aufspaltung in kleinere Teilprobleme - wir gehen hinunter bis zu Teilproblemen der Dimension 2 und 1 - zerstört in der von uns beschriebenen Weise die positive Definitheit nicht. Ist also eines der kleinen Teilprobleme nicht positiv definit, so kann auch das Gesamtproblem (T, J) nicht positiv definit sein. Der Algorithmus meldet, daß (T, J) nicht positiv definit ist.
Sind alle 2x2-Eigenwertprobleme gelöst und alle 2x2-Paare (T_{2x2}, J_{2x2}) positiv definit, so berechnet man den Durchschnitt aller ihrer Definitheitsintervalle. Das tatsächliche Definitheitsintervall ist Teilmenge dieses Durchschnittes; ist der Durchschnitt leer, so kann das Problem (T, J) nicht mehr positiv definit sein.
Durch Zusammenfassen zweier Teileigenwertprobleme gelangen wir im folgenden Schritt zu einem höherdimensionalen Teilproblem, also etwa von zwei 2x2-Teilproblemen zu einem 4x4-Problem. Die Eigenwerte dieses neuen Teilproblems berechnen wir mit Hilfe des im nächsten Abschnitt beschriebenen Nullstellenverfahrens. Wenn im Definitheitsintervall nicht genau zwei reelle Nullstellen liegen, so ist das aktuelle Teilproblem (T_{part}, J_{part}) nicht mehr positiv definit, und wir beenden das Verfahren (s. der folgender Satz 4.2).
Auf diese Art und Weise arbeitet man sich durch unseren Problembaum, bis der Algorithmus mit der Meldung "Paar nicht positiv definit" abbricht oder das Gesamteigenwertproblem gelöst ist.
Bei den meisten Testbeispielen erkennt das Verfahren das Verletztsein der positiven Definitheit sehr früh, oft schon bei der Lösung der 2x2-Teilprobleme. Diese werden mit verallgemeinerten Jacobi-Transformationen gelöst (vgl. Veselič [63]), was einerseits sehr genau ist und andererseits direkt auf positive Definitheit testet.

Jetzt überlegen wir uns noch, daß aus der Existenz zweier Nullstellen im Definitheitsintervall (α_q, α_{q+1}) die positive Definitheit des Paares $(D - \rho zz^T, J)$ folgt:

Wir betrachten das positiv definite Paar (D, J), $J = \mathrm{diag}(-I_q, +I_p)$, $p + q = n$, $D = \mathrm{diag}(d_i)$, das nach eventueller Deflation die Eigenwerte $\alpha_1 < \alpha_2 < \ldots < \alpha_q < \alpha_{q+1} < \ldots < \alpha_n$ und das Definitheitsintervall (α_q, α_{q+1}) hat. Die Eigenwerte des Paares $(D - \rho \mathbf{z}\mathbf{z}^T, J)$, $\rho > 0$, $\mathbf{z} \in \mathbb{R}^n$ mit lauter von null verschiedenen Komponenten, sind gerade die Nullstellen der Funktion w (vgl. Formel (4.20)), für die gilt (vgl. S 52):
$\alpha_1 < \lambda_1 < \alpha_2 < \lambda_2 < \ldots < \lambda_{q-1} < \alpha_q < \alpha_{q+1} < \lambda_{q+1} < \ldots < \lambda_n < \alpha_n$. Im Definitheitsintervall (α_q, α_{q+1}) können eine, zwei oder keine Nullstelle liegen.

Satz 4.2: *Es gelten die oben beschriebenen Voraussetzungen. Außerdem habe die Funktion w in (α_q, α_{q+1}) zwei verschiedene Nullstellen $\lambda_q < \lambda_{q+1}$. Dann ist das Paar $(D - \rho \mathbf{z}\mathbf{z}^T, J)$ positiv definit und $(\lambda_q, \lambda_{q+1})$ das zugehörige Definitheitsintervall.*

Beweis:
Zu den n verschiedenen Eigenwerte $\lambda_1, \ldots \lambda_n$ bestimmen wir die zugehörigen Eigenvektoren gemäß Formel (4.21) und fassen sie zur Matrix $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$ zusammen. Für die in *dieser* Reihenfolge definierte Matrix Q gilt $Q^T J Q = J$, denn:

Nach Formel (4.21) ist $\mathbf{q}_i = \frac{1}{C_1}\left(\frac{z_k j_k}{\alpha_k - \lambda_i}\right)_{k=1,\ldots,n}$ mit Normierungsfaktor C_1,

derart, daß $\mathbf{q}_i^T J \mathbf{q}_i = \pm 1$. Wenn wir uns diesen Ausdruck genauer ansehen, stellen wir fest, daß $\mathbf{q}_i^T J \mathbf{q}_i = -1$ für $i = 1, \ldots q$ und $\mathbf{q}_i^T J \mathbf{q}_i = +1$ für $i = q+1, \ldots, n$ der Fall ist:

$$\mathbf{q}_i^T J \mathbf{q}_i = \frac{1}{C_1^2} \sum_{k=1}^{n} \frac{z_k^2 j_k}{(\alpha_k - \lambda_i)^2} = \frac{1}{C_1^2}\left(\frac{1}{\rho}\left(\rho \sum_{k=1}^{n} \frac{z_k^2 j_k}{(\alpha_k - \lambda_i)^2}\right)\right)$$
$$= C_2\left(-w'(\lambda_i)\right) \quad \text{mit} \quad C_2 = \frac{1}{C_1^2 \rho} > 0.$$

Nun gilt $w'(\lambda_i) > 0$ für $i = 1, \ldots, q-1$, da λ_i einzige (und einfache) Nullstelle in (α_i, α_{i+1}) ist und somit w in einer kleinen Umgebung um λ_i streng monoton wachsend sein muß. Analog ergibt sich $w'(\lambda_i) < 0$ für $i = q+2, \ldots n$. Aus Konkavität und Polverhalten der Funktion w im Definitheitsintervall sowie der Existenz zweier Nullstellen können wir $w'(\lambda_q) > 0$ und $w'(\lambda_{q+1}) < 0$ folgern. Damit ergeben sich für die Eigenvektoren $\mathbf{q}_i$ die oben behaupteten Vorzeichen, und $\lambda_1, \ldots, \lambda_q$ sind gerade die Minuseigenwerte und $\lambda_{q+1}, \ldots \lambda_n$ die Pluseigenwerte des Paares $(D - \rho \mathbf{z}\mathbf{z}^T, J)$.

Wenn wir dieses Paar jetzt mit Hilfe von Q diagonalisieren, erhalten wir
$Q^T(D - \rho \mathbf{z}\mathbf{z}^T)Q = \mathrm{diag}(\lambda_i j_i) =: D'$, $Q^T J Q = J$.
Damit ist $(D - \rho \mathbf{z}\mathbf{z}^T, J)$ genau dann positiv definit, wenn (D', J) es ist.
(D', J) ist genau dann positiv definit, wenn es ein $\mu \in \mathbb{R}$ gibt mit $\lambda_i j_i - \mu j_i > 0$ für $i = 1, \ldots, n$.
Dies bedeutet aber $-\lambda_i + \mu > 0$ für $i = 1, \ldots, q$ und $\lambda_i - \mu > 0$ für $i = q+1, \ldots, n$.
Hieraus erhalten wir $\lambda_q < \mu < \lambda_{q+1}$ wegen der Anordnung der Eigenwerte.
Damit ist $(\lambda_q, \lambda_{q+1})$ Definitheitsintervall von $(D - \rho \mathbf{z}\mathbf{z}^T, J)$.

Q.E.D.

4.2 Die Nullstellensuche

Die für unser Divide-et-Impera-Verfahren notwendige Nullstellensuche geschieht mithilfe eines neuen Einschließungsverfahrens, das uns für jede Nullstelle genau ein Intervall liefert. Die eigentliche Nullstellenberechnung erfolgt mit der bekannten Pegasus-Methode, einer Variante der Regula Falsi. Damit unterscheidet sich unser Vorgehen von dem für den Fall J = I aus der Literatur bekannten Verfahrensweisen. (Wegen gewisser fehlender Monotonieeigenschaften sind diese nicht geeignet.) Außerdem kann man unser Eingrenzungsverfahren noch so erweitern, daß es auf allgemeinere rationale Funktionen anwendbar ist.

4.2.1 Die Bestimmung der Nullstellen der Funktion w

Die Nullstellen der Funktion w mit

$$w(x) = 1 + \rho \sum_{j=1}^{q} \frac{z_j^2}{\alpha_j - x} - \rho \sum_{j=q+1}^{m} \frac{z_j^2}{\alpha_j - x}$$

müssen möglichst genau berechnet werden, um die Teilprobleme der Gestalt

$$(D_T - \rho \mathbf{z}\mathbf{z}^T)\mathbf{x} = \lambda J_T \mathbf{x} \quad \text{mit} \quad D_T = \mathrm{diag}(d_1, \ldots, d_m), \quad J_T = \mathrm{diag}(j_1, \ldots, j_m),$$

$$\alpha_i := d_i j_i, \; i = 1, \ldots, m, \quad q = \text{Anzahl der Minuseinsen in } J_T, \quad \rho > 0,$$

mit ausreichender Genauigkeit lösen zu können. Wegen der insgesamt sehr großen Anzahl zu bestimmender Nullstellen ist außerdem ein möglichst schnelles Verfahren entscheidend. Durch eine vorausgegangene Deflation wurde dafür gesorgt, daß alle α_i, $i = 1, \ldots, m$ verschieden sind und alle z_i, $i = 1, \ldots, m$ ungleich null.
Für den Fall $J = -I$ (bzw. $J = I$) gibt es ein schönes Nullstellenverfahren, das die spezielle Gestalt von w berücksichtigt. Man schreibt die Funktion w als $w(x) = 1 + \varphi(x) + \psi(x)$ mit zwei monotonen und konvexen Funktionen φ und ψ und führt eine rationale Approximation durch, s. Bunch, Nielsen und Sorensen [13] für eine ausführliche Beschreibung.
Dieses Verfahren ist für $J \neq I$ nicht anwendbar, denn die Monotonie- und Konvexitätseigenschaften sind jetzt nicht mehr gegeben, außerdem weiß man für das sogenannte Definitheitsintervall zunächst noch nicht, ob überhaupt eine Nullstelle existiert.
Wir werden daher bei der Berechnung der Nullstellen einen anderen Weg einschlagen. Zunächst werden wir mit Hilfe eines *Einschließungsverfahrens* das Intervall verkleinern, in dem die Nullstellen liegen müssen, wenn sie existieren. Dann folgt die eigentliche Nullstellensuche mit Hilfe des *Pegasus-Verfahrens*, vgl. (Engeln-Müllges [26] und Dowell, Jarratt [25]).

4.2.2 Das quadratische Einschließungsverfahren

Das im folgenden beschriebene Einschließungsverfahren könnte theoretisch auf eine beliebige Funktion h mit

$$h(x) = \frac{1}{\rho} - \sum_{i=1}^{m} \frac{a_i}{\alpha_i - x} \quad , \rho, a_i, \alpha_i \in \mathbb{R}$$

angewendet werden. Dabei sollen die α_i, $i = 1,\dots,m$, in aufsteigender Reihenfolge angeordnet sein mit $\alpha_1 < \alpha_2 < \dots < \alpha_m$. Die $a_i \in \mathbb{R}$ können positiv oder negativ sein, nur die Null ist nicht zugelassen. Wo die (höchstens m) reellen Nullstellen von h liegen und wie viele es genau sind, ist zunächst nicht klar. Das soll jetzt das Einschließungsverfahren feststellen. Wir schauen uns dieses Einschließungsverfahren hier für unsere spezielle Funktion w an mit

$$(4.22) \quad w(x) := \frac{1}{\rho} - \sum_{i=1}^{m} \frac{z_i^2 j_i}{\alpha_i - x} = \frac{1}{\rho} + \sum_{i=1}^{q} \frac{z_i^2}{\alpha_i - x} - \sum_{i=q+1}^{m} \frac{z_i^2}{\alpha_i - x}$$

Für w wissen wir bereits, wo die Nullstellen x_i, $i = 1, \dots, m$, liegen müssen:

Im Fall $q = 0$ gibt es genau m Nullstellen mit

$$\alpha_1 - \rho \mathbf{z}^T\mathbf{z} < x_1 < \alpha_1 < x_2 < \alpha_2 < \dots < x_m < \alpha_m \quad ,$$

für $q = m$ gilt

$$\alpha_1 < x_1 < \alpha_2 < x_2 < \dots < \alpha_m < x_m < \alpha_m + \rho \mathbf{z}^T\mathbf{z} \; .$$

Für den besonders interessanten Fall $1 \le q \le m-1$ erhalten wir

$$\alpha_1 < x_1 < \alpha_2 < x_2 < \dots < x_{q-1} < \alpha_q$$

und

$$\alpha_{q+1} < x_{q+2} < \alpha_{q+2} < \dots < x_m < \alpha_m .$$

Wie bereits erwähnt, können im Definitsheitsintervall $(\alpha_q , \alpha_{q+1})$ zwei, eine oder keine Nullstelle liegen.
Um schnell festzustellen, ob das zugrundeliegende Eigenwertproblem positiv definit ist, sollte man in diesem Fall zunächst mit der Nullstellensuche im Definitheitsintervall beginnen. Wenn es keinen positiven Wert im Definitheitsintervall gibt und damit auch keine zwei Nullstellen, kann man den Algorithmus mit der Meldung "Paar nicht positiv definit" abbrechen.

Nullstellensuche in einem inneren Intervall:

Wir betrachten zunächst den Fall $1 \le k \le m-1$, d.h. x liegt im Intervall $[\alpha_1, \alpha_m]$. Die Funktion w schreiben wir in der Form

$$(4.23) \quad w(x) = \underbrace{-\left(\frac{z_k^2 j_k}{\alpha_k - x} + \frac{z_{k+1}^2 j_{k+1}}{\alpha_{k+1} - x} \right)}_{=: \; - \; h_1(x)} + \underbrace{\frac{1}{\rho} - \sum_{\substack{i=1 \\ i \ne k, k+1}}^{m} \frac{z_i^2 j_i}{\alpha_i - x}}_{=: \; h_2(x)} \; .$$

Wir haben eine Nullstelle x_N von w genau dann gefunden, wenn gilt:

$$(4.23') \quad h_1(x_N) = h_2(x_N)$$

Wir suchen als erstes ein Intervall $(x_u , x_o) \subset (\alpha_k , \alpha_{k+1})$, in dem die gesuchte Nullstelle liegt (falls sie existiert).
Beim Start ist $x_u := \alpha_k$ und $x_o := \alpha_{k+1}$. Im Einschließungsschritt werden neue Ränder x_u , x_o berechnet. Mit ihnen kann man den Einschließungsschritt wiederholen und bekommt auf diese Weise ein Iterationsverfahren. (Im Computerprogramm wird der Einschließungsschritt pro Intervall maximal dreimal durchgeführt.)

Mit M_+ bezeichnen wir eine (möglichst kleine) obere Schranke für h_2 im Intervall (x_u , x_o), mit M_- eine (möglichst große) untere Schranke für h_2 im selben Intervall. Für $x \in \mathbb{R}$ und $\alpha_k \le x_u < x < x_o \le \alpha_{k+1}$ bekommen wir solche Schranken M_+ und M_- folgendermaßen:

$$h_2(x) = \frac{1}{\rho} - \sum_{\substack{i=1 \\ i \ne k, k+1}}^{m} \frac{z_i^2 j_i}{\alpha_i - x}$$

$$= \frac{1}{\rho} + \sum_{\substack{i<k \\ j_i=+1}} \frac{z_i^2}{x - \alpha_i} - \sum_{\substack{i<k \\ j_i=-1}} \frac{z_i^2}{x - \alpha_i} - \sum_{\substack{i>k+1 \\ j_i=+1}} \frac{z_i^2}{\alpha_i - x} + \sum_{\substack{i>k+1 \\ j_i=-1}} \frac{z_i^2}{\alpha_i - x}$$

$$\le \frac{1}{\rho} + \sum_{\substack{i<k \\ j_i=+1}} \frac{z_i^2}{x_u - \alpha_i} - \sum_{\substack{i<k \\ j_i=-1}} \frac{z_i^2}{x_o - \alpha_i} - \sum_{\substack{i>k+1 \\ j_i=+1}} \frac{z_i^2}{\alpha_i - x_u} + \sum_{\substack{i>k+1 \\ j_i=-1}} \frac{z_i^2}{\alpha_i - x_o}$$

$$= \frac{1}{\rho} - \sum_{i \ne k, k+1} \frac{z_i^2 j_i}{\alpha_i - \xi_i} =: M_+$$

$$\text{mit} \quad \xi_i := \begin{cases} x_u & \text{für } j_i = +1 \\ x_o & \text{für } j_i = -1 \end{cases} .$$

Analog bestimmen wir jetzt M_- und erhalten

$$M_- := \frac{1}{\rho} - \sum_{i \neq k,k+1} \frac{z_i^2 j_i}{\alpha_i - \xi_i} \quad \text{mit} \quad \xi_i := \begin{cases} x_o & \text{für } j_i = +1 \\ x_u & \text{für } j_i = -1 \end{cases}$$

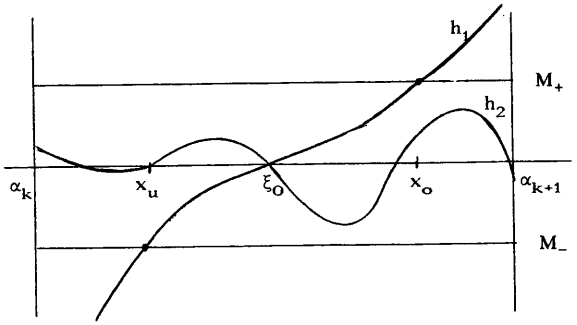
Nach diesen Vorbereitungen kommen wir zum eigentlichen <u>Einschließungsverfahren:</u>
Wir suchen ein Intervall $(x_u, x_o) \subset (\alpha_k, \alpha_{k+1})$, in dem die Nullstelle von w sicher liegt. Beim Start ist $x_u = \alpha_k$, $x_o = \alpha_{k+1}$. Dabei müssen wir drei Fälle unterscheiden:

<u>1.Fall: $j_k = j_{k+1} = +1$</u>

In diesem Fall ist die Funktion h_1 mit

$$h_1(x) = \frac{z_k^2}{\alpha_k - x} + \frac{z_{k+1}^2}{\alpha_{k+1} - x}$$

streng monoton wachsend auf (x_u, x_o). Die Funktion h_2 hat keine Pole in (α_k, α_{k+1}) und damit auch nicht in (x_u, x_o). Da h_1 streng monoton wachsend ist und als stetige Funktion auf (α_k, α_{k+1}) alle Werte zwischen $+\infty$ und $-\infty$ annimmt, gibt es mindestens eine Stelle ξ_0, wo sich h_1 und h_2 schneiden und w somit eine Nullstelle hat. Aus den Informationen, die wir bereits über w haben, wissen wir, daß es in diesem Fall genau eine Nullstelle im Intervall (α_k, α_{k+1}) gibt.



Jetzt bestimmen wir ein $x'_u > x_u$ wie folgt: Es gibt keine Nullstelle der Funktion w in (α_k, α_{k+1}) für diejenigen x, die

$$h_1(x) < M_-$$

erfüllen.
x'_u soll jetzt Lösung der Gleichung

$$h_1(x'_u) = M_-$$

sein.
Diese (quadratische) Gleichung hat zwei Nullstellen ξ_+ und ξ_-, von denen eine innerhalb und eine außerhalb des Intervalls (α_k, α_{k+1}) liegt. Von den beiden berechneten Nullstellen wählen wir nun diejenige aus, die in (x_u, x_o) liegt:

$$x'_u := \begin{cases} \xi_- & \text{für } \xi_- \in (x_u, x_o) \\ \xi_+ & \text{sonst} \end{cases} .$$

Dann setzen wir $x_u := x'_u$ und bestimmen ein M_+ auf dem neuen Intervall (x_u, x_o): Es existiert keine Nullstelle für diejenigen x mit

$$h_1(x) > M_+ .$$

Die Grenze x'_o bekommen wir durch Lösen der Gleichung

$$h_1(x'_o) = M_+.$$

Wieder gibt es zwei Lösungen ξ_+ und ξ_-, von denen wir diejenige als x'_o nehmen, die im Intervall (x_u, x_o) liegt. Wir setzen $x_o := x'_o$ und haben damit ein neues Intervall (x_u, x_o) bekommen, mit dem wir den Einschließungsschritt wiederholen können.

<u>2. Fall: $j_k = j_{k+1} = -1$</u>

In diesem Fall ist die Funktion h_1 mit

$$h_1(x) = \frac{-z_k^2}{\alpha_k - x} + \frac{-z_{k+1}^2}{\alpha_{k+1} - x}$$

streng monoton fallend. Sie hat keine Nullstelle für diejenigen x aus dem Intervall (α_k, α_{k+1}), die

$$h_1(x) > M_+$$

erfüllen.
Wie im Fall 1 hat die Gleichung

$$h_1(x'_u) = M_+$$

zwei Lösungen, von denen wir diejenige als neue Grenze x_u auswählen, die in (x_u, x_o) liegt. Auf dem neuen Intervall (x_u, x_o) wird jetzt M_- bestimmt.

Die Funktion w hat keine Nullstelle für die x mit

$$h_1(x) < M_- .$$

Die neue Intervallgrenze x'_o ist nun die Lösung der Gleichung

$$h_1(x'_o) = M_- ,$$

die in (x_u, x_o) liegt.

3. Fall: $j_k = -1$ und $j_{k+1} = +1$ (d.h. $k = q$)

Die Funktion h_1 hat in diesem Fall die folgende Gestalt:

$$h_1(x) = \frac{-z_k^2}{\alpha_k - x} + \frac{z_{k+1}^2}{\alpha_{k+1} - x} .$$

Für $x \in (\alpha_k, \alpha_{k+1})$ ist h_1 immer größer oder gleich null und wegen $h_1''(x) > 0$ konvex.

Die durch

$$h_2(x) = \frac{1}{\rho} + \sum_{j=1}^{k-1} \frac{z_j^2}{\alpha_j - x} - \sum_{j=k+2}^{m} \frac{z_j^2}{\alpha_j - x}$$

definierte Funktion h_2 stellt auf (α_k, α_{k+1}) wegen $h_2''(x) < 0$ eine konkave Funktion dar.

Wir wissen in diesem Fall, dem Fall des "Definitheitsintervalls", noch nicht, ob es eine, zwei oder keine Nullstelle gibt. In diesem Intervall ist w eine konkave Funktion, die dort ihr Maximum x_E annimmt. Wir unterscheiden drei Fälle:

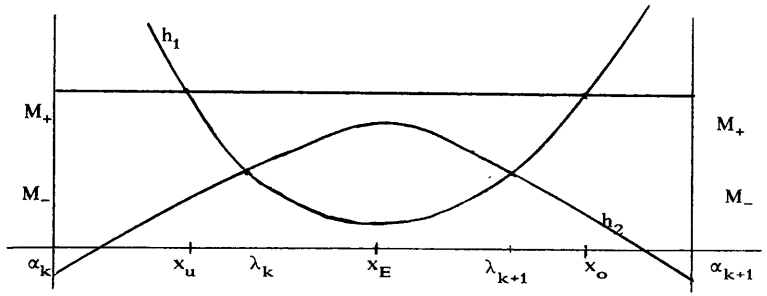
(a) $w(x_E) > 0$:
Dann gibt es zwei Nullstellen λ_k und λ_{k+1} in $(\alpha_k , \alpha_{k+1})$ und damit ein neues Definitheitsintervall $(\lambda_k , \lambda_{k+1}) \subset (\alpha_k , \alpha_{k+1})$.

(b) $w(x_E) = 0$:
In diesem Fall ist das Zwischeneigenwertproblem positiv semidefinit, und wir beenden die Rechnung.

(c) $w(x_E) < 0$:
Es gibt keine Nullstelle von w in $(\alpha_k , \alpha_{k+1})$, und das Zwischeneigenwertproblem ist nicht positiv definit. Auch in diesem Fall beenden wir das Programm.

Man könnte die Extremwertstelle x_E direkt ausrechen, aber das würde eine Nullstellensuche für die Ableitung von w bedeuten, was Rechenzeit kostet. Im Computerprogramm suchen wir daher nur nach einem Wert x_E, für den $w(x_E)$ positiv ist. Gibt es solch ein x_E nicht, so existiert auch kein Extremum von w mit positivem Funktionswert.

Nach der Berechnung von x_E mit $w(x_E) > 0$ wissen wir, daß es im Intervall (α_k , x_E) und im Intervall (x_E , α_{k+1}) jeweils eine Nullstelle gibt.



Die Gesamtstrategie für den Einschließungsschritt für das Definitheitsintervall sieht jetzt folgendermaßen aus:

1. Man bestimme ein möglichst kleines M_+. Ist M_+ gleich null, so kann $w(x)$ nicht mehr positiv sein und die Positivdefinitheitsvoraussetzung ist nicht gegeben.
   Falls möglich, löse man die Gleichung
   $$h_1(x) = M_+ .$$
   Hat sie keine oder nur eine Nullstelle, so besitzt auch w keine Nullstelle im Definitheitsintervall, und wir beenden das Programm. Andernfalls wählen wir die beiden Nullstellen als x_u und x_o .
2. Im Intervall (A , B) mit A:= x_u und B:= x_o suchen wir einen Wert x_E mit $w(x_E) > 0$. (Im Computerprogramm wird dies mit einem "Goldenen-Schnitt"-Verfahren durchgeführt.) Existiert ein solches x_E nicht, so wird das Programm beendet.
3. Auf dem Intervall (x_u, x_E) bestimmen wir M_+ und lösen die Gleichung
   $$h_1(x'_u) = M_+ .$$
   Die Lösung x'_u aus diesem Intervall übernimmt jetzt die Rolle von x_u: $x_u := x'_u$.
   Auf (x_u, x_E) berechnen wir M_- und suchen die Lösung der Gleichung
   $$h_1(x_o) = M_- .$$
   Sie übernimmt die Rolle von x_E.
   Nach diesem Schritt haben wir ein Intervall (x_{u1} , x_{o1}) erhalten, in dem die Nullstelle liegen muß.

Diesen Schritt wiederholen wir ein- bis zweimal.

Mit dem erhaltenen Intervall gehen wir dann in das eigentliche Nullstellenverfahren. (Im Computerprogramm ist es das Pegasus-Verfahren.)

4. Den Schritt 3 führen wir jetzt für das Intervall (x_E, x_o) aus.

Behandlung der Randintervalle:

Bis jetzt haben wir nach Nullstellen im Intervall (α_1, α_m) gesucht. Unter der Voraussetzung der positiven Definitheit des aktuellen Paares gibt es nur im Fall $q = 0$ oder $q = m$ Nullstellen von w, die außerhalb von (α_1, α_m) liegen. Wie wir schon bei der Untersuchung der Funktion w gesehen haben, befinden sich m-1 Nullstellen in (α_1, α_m). Für die verbliebene Nullstelle gilt:

Sie liegt im Fall $q = 0$ im Intervall $[\alpha_1 - \rho \mathbf{z}^T\mathbf{z}, \alpha_1)$ und im Fall $q = m$ im Intervall $(\alpha_m, \alpha_m + \rho \mathbf{z}^T\mathbf{z}]$ (vgl. Abschnitt 4.1.3.2).

Auch für diese beiden Fälle lohnt sich das Einschließungsverfahren, um Intervallgrenzen für die Nullstellensuche zu bekommen, die ein Stück entfernt vom Pol liegen.

Der Fall q = 0:

In diesem Fall sind alle $j_i = 1$ für $i = 1,\dots,m$. h_1 hat daher die Gestalt

$$h_1(x) = \frac{z_1^2}{\alpha_1 - x} + \frac{z_2^2}{\alpha_2 - x}$$

und ist im Intervall $\Omega := [\alpha_1 - \rho \mathbf{z}^T\mathbf{z}, \alpha_1)$ positiv, streng monoton wachsend und konvex.

Die Funktion h_2 mit

$$h_2(x) = \frac{1}{\rho} - \sum_{i=3}^{m} \frac{z_i^2}{\alpha_i - x}$$

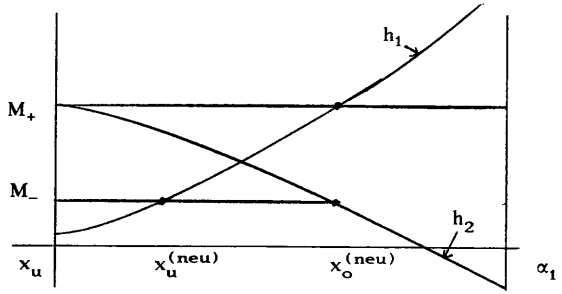
ist auf Ω streng monoton fallend und konkav.
Da die Funktion w in Ω genau eine Nullstelle hat, müssen sich h_1 und h_2 in genau einem Punkt schneiden.
x_u ist beim Start gerade gleich $\alpha_1 - \rho \mathbf{z}^T\mathbf{z}$, x_o gleich α_1, und für M_+ wählen wir

$$M_+ := h_2(x_u).$$

Da $h_1(x_u) \le M_+$ gilt und h_1 streng monoton gegen $+\infty$ wächst, hat die Gleichung

$$h_1(x) = M_+$$

in (x_u, α_1) genau eine Lösung. Sie wird unser neues x_o.



Als M_- nehmen wir

$$M_- := h_2(x_o).$$

Die Gleichung

$$h_1(x) = M_-$$

muß auf $[x_u, x_o]$ keine Lösung haben. Dies ist der Fall, wenn

$$h_2(x_o) < h_1(x_u)$$

ist. Dann bleibt die untere Intervallgrenze x_u unverändert, andernfalls wird sie durch die Lösung der Gleichung ersetzt.

Auf dem so entstandene neuen Intervall $[x_u, x_o]$ kann man den Einschliessungsschritt jetzt wiederholen.

Der Fall q = m:

Dieser Fall ist analog zum Fall $q = 0$:
Alle j_i sind gleich -1 für $i = 1,\dots,m$. Daher haben h_1 und h_2 das folgende Aussehen:

$$h_1(x) = \frac{-z_{m-1}^2}{\alpha_{m-1} - x} + \frac{-z_m^2}{\alpha_m - x}$$

und

$$h_2(x) = \frac{1}{\rho} + \sum_{i=1}^{m-2} \frac{z_i^2}{\alpha_i - x} .$$

Auf dem Intervall $(\alpha_m, x_o]$ mit $x_o = \alpha_m + \rho \mathbf{z}^T\mathbf{z}$ ist h_1 positiv, streng monoton fallend und konvex, h_2 streng monoton wachsend und konkav. Wie im Fall q = 0 sieht man, daß die Gleichung

$$h_1(x) = M_+$$

mit

$$M_+ := h_2(x_o)$$

genau eine Lösung hat. Diese Lösung wählen wir als neues x_u.

Dann lösen wir die Gleichung

$$h_1(x) = M_-$$

für

$$M_- := h_2(x_u) .$$

Hat die Gleichung keine Lösung, so bleibt x_o unverändert, sonst wird es durch die Lösung ersetzt.

4.2.3 Praktisches Vorgehen

Nachdem man mit dem gerade beschriebenen Eingrenzungsverfahren ein sehr gutes Startintervall gefunden hat, wird die endgültige Nullstellenbestimmung mit dem Pegasus-Verfahren durchgeführt. Der Vorteil des Pegasus-Verfahrens (einer Variante der Regula Falsi) liegt darin, daß es - im Gegensatz etwa zum Newton-Verfahren - ein Einschließungsverfahren ist (man kann die Nullstelle also nicht verlieren) und deutlich schneller als die üblichen Regula-Falsi-Verfahren arbeitet. (Liegt etwa die Nullstelle nahe an einem Intervallende, so hält im Gegensatz zum Pegasus-Verfahren die Regula Falsi den einen Einschließungspunkt immer fest.)
Theoretisch sollte man auf diese Art und Weise alle Nullstellen finden. In der Praxis muß man aber noch einige Vorkehrungen treffen, damit der Algorithmus auch diejenigen Nullstellen entdeckt, die äußerst nahe an den Polen der Funktion w liegen. Wichtig ist dabei eine geeignete Verschiebungsstrategie; daher erweitern wir die im Artikel von Bunch, Nielsen und Sorensen [13] beschriebene Vorgehensweise auf unser Nullstellenproblem und betrachten nicht den tatsächlichen Wert von x, sondern den Abstand zum nächsten Pol. Desweiteren ist zu bemerken, daß die Nullstellen bereits im Einschließungsschritt öfter so genau bestimmt werden, daß sich das anschließende Pegasus-Verfahren erübrigt. Bei besonders kritischen Beispielen (Nullstellen und Polstellen fallen fast zusammen) wird das Pegasus-Verfahren mit dem langsameren, aber sehr sichere Bisektionsverfahren ergänzt.

4.2.3.1 Ein lineares Einschließungsverfahren

In Fällen, wo eine Nullstelle fast vollständig mit einem Pol übereinstimmt, kann das quadratische Eingrenzungsverfahren zu ungenaue Werte liefern, weil dort quadratische Gleichungen durch Wurzelziehen gelöst werden müssen. Um auch in diesen Ausnahmefällen ein geeignetes Startintervall zu erhalten, wird dem quadratischen Eingrenzungsverfahren noch ein lineares vorgeschaltet, das ausschließlich mit linearen Gleichungen arbeitet.

Das lineare Einschließungsverfahren benutzt ähnliche Ideen wie das quadratische Verfahren. Es setzt jedoch voraus, daß die beiden Intervallgrenzen des Intervalls, in dem die Nullstellensuche stattfinden soll, Pole unserer Funktion w sind; im Gegensatz zum quadratischen Einschließungsverfahren darf es daher nur auf den Intervallen $[\alpha_k, \alpha_{k+1}]$, k = 1, ... , m-1 (und auch nur einmal) angewendet werden. In einigen extremen Fällen hat man trotzdem nach diesem ersten Eingrenzungsschritt die Nullstelle oder die Nullstellen bereits gefunden.

Das Verfahren läßt sich folgendermaßen zusammenfassen:
Zur Nullstellensuche im Intervall $[\alpha_k, \alpha_{k+1}]$ setzen wir $\tilde{\alpha} := (\alpha_k + \alpha_{k+1})/2$ und bestimmen ein $x_u \in (\alpha_k, \tilde{\alpha}]$ und ein $x_o \in [\tilde{\alpha}, \alpha_{k+1})$ so, daß die eine oder im Fall des Definitheitsintervalles für ein positives Matrizenpaar beide Nullstellen im Intervall $[x_u, x_o]$ liegen:

Bestimmung der unteren Intervallgrenze x_u:

Wir schreiben die Funktion w als

$$\frac{w(x)}{\rho} = -\frac{j_k z_k^2}{\alpha_k - x} + \frac{1}{\rho} - \sum_{i \neq k} \frac{j_i z_i^2}{\alpha_i - x} = -h_1(x) + h_2(x)$$

$$\text{mit} \quad h_1(x) := \frac{j_k z_k^2}{\alpha_k - x} \quad \text{und} \quad h_2(x) := \frac{1}{\rho} - \sum_{i \neq k} \frac{j_i z_i^2}{\alpha_i - x} .$$

1. Fall: $j_k = 1$:
Wie beim quadratischen Eingrenzungsverfahren schätzt man h_2 geeignet ab; wir bestimmen hier eine untere Schranke S_u mit $h_2(x) \geq S_u$ im Intervall $[\alpha_k, \tilde{\alpha}]$. Für die Nullstelle x muß gelten:

$-h_1(x) + h_2(x) = 0$, also $h_1(x) = h_2(x)$, woraus $S_u \leq h_1(x)$ folgt, sofern $x \in [\alpha_k, \tilde{\alpha}]$.

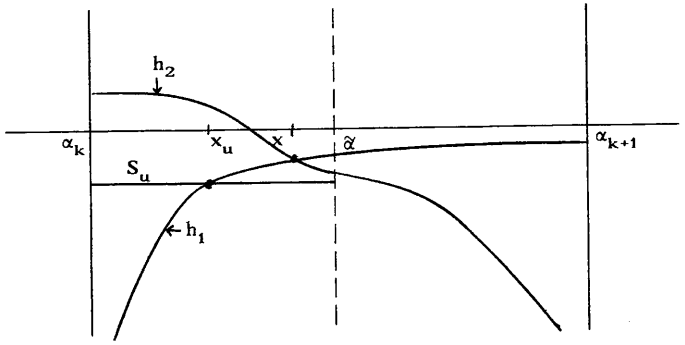
Wir suchen jetzt ein x_u aus $(\alpha_k, \tilde{\alpha}]$, das $S_u = h_1(x_u)$ erfüllt.

Gibt es ein solches x_u, muß

$$\frac{z_k^2 j_k}{\alpha_k - x_u} = S_u, \text{ also } x_u = \alpha_k - \frac{z_k^2 j_k}{S_u} \quad \text{für } S_u \neq 0$$

gelten, sonst ist $x_u := \tilde{\alpha}$ zu setzen.
Da h_1 monoton wächst, ist klar, daß keine Nullstelle im Intervall (α_k, x_u) sein kann.

Situation im Fall $j_k = +1$:



Analog verfahren wir für den zweiten Fall, also für $j_k = -1$.

Mit fast gleichen Überlegungen wie in 4.2.2 erhalten wir die folgenden Formeln für S_u und x_u:

$$S_u := \frac{1}{\rho} - \sum_{i \neq k} \frac{z_i^2 j_i}{\alpha_i - \alpha_i'} \quad \text{mit} \quad \alpha_i' := \begin{cases} \tilde{\alpha} & \text{wenn } \operatorname{sgn}(j_i) = \operatorname{sgn}(j_k) \\ \alpha_k & \text{sonst} \end{cases}$$

$$x_u := \begin{cases} \tilde{\alpha}, & \text{falls } S_u = 0 \\ \alpha_k - \dfrac{z_k^2 j_k}{S_u} & \text{sonst} \end{cases}$$

Erhält man im Fall $S_u \neq 0$ einen Wert außerhalb von $(\alpha_k, \tilde{\alpha})$, so ist x_u auf $x_u := \tilde{\alpha}$ abzuändern.

Bestimmung der oberen Intervallgrenze x_o:

1. Fall: $j_{k+1} = 1$:
Auch hier spalten wir w/ρ in der Form $w(x)/\rho = -h_1(x) + h_2(x)$ auf, allerdings jetzt mit

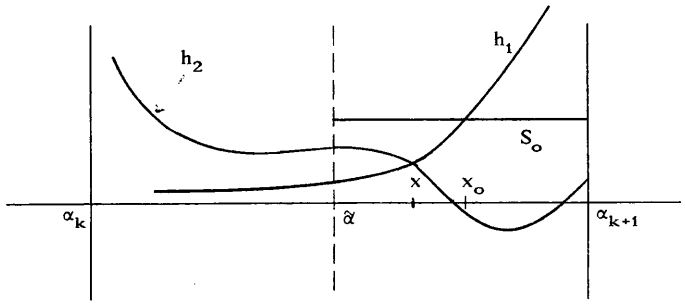
$$h_1(x) := \frac{j_{k+1} z_{k+1}^2}{\alpha_{k+1} - x} \quad \text{und} \quad h_2(x) := \frac{1}{\rho} - \sum_{i \neq k+1} \frac{j_i z_i^2}{\alpha_i - x}.$$

Nun schätzen wir h_2 im Intervall $[\tilde{\alpha}, \alpha_{k+1}]$ durch eine geeignete Schranke S_o nach oben ab. Für die Nullstelle x muß

$$-h_1(x) + h_2(x) = 0 \quad \text{gelten und damit} \quad S_o \geq h_1(x) \quad \text{für } x \in [\tilde{\alpha}, \alpha_{k+1}].$$

Wir berechnen, falls es existiert, ein $x_o \in (\tilde{\alpha}, \alpha_{k+1})$ mit $S_o = h_1(x_o)$.

Situation im Fall $j_{k+1} = 1$:



Analog verfahren wir im 2. Fall $j_{k+1} = -1$.

Die Formeln für S_o und x_o sind jetzt gegeben durch:

$$S_o := \frac{1}{\rho} - \sum_{i \neq k+1} \frac{z_i^2 j_i}{\alpha_i - \alpha_i'} \quad \text{mit} \quad \alpha_i' := \begin{cases} \tilde{\alpha}, & \text{wenn } \operatorname{sgn}(j_i) = \operatorname{sgn}(j_{k+1}) \\ \alpha_{k+1} & \text{sonst} \end{cases}$$

$$x_o := \begin{cases} \tilde{\alpha}, & \text{falls } S_o = 0 \\ \alpha_{k+1} - \dfrac{z_{k+1}^2 j_{k+1}}{S_o} & \text{sonst} \end{cases}$$

Liegt das im Fall $S_o \neq 0$ berechnete x_o außerhalb des Intervalls $[\tilde{\alpha}, \alpha_{k+1}]$, so ist $x_o := \tilde{\alpha}$ zu setzen.

4.2.3.2 Verschiebungsstrategien

Durch die Verwendung von Verschiebungen beim Eingrenzungsverfahren, bei der Nullstellensuche als auch bei der Berechnung eines Eigenvektors **y** des Eigenwertproblems $(D - \rho \mathbf{z}\mathbf{z}^T)\mathbf{y} = \lambda J \mathbf{y}$ mittels der Formel (4.21) kann man die Genauigkeit der Ergebnisse deutlich erhöhen, da man dadurch die Auslöschung wesentlicher Stellen größtenteils verhindert.
Dabei arbeitet man mit Nennern der Gestalt $\alpha_i - \lambda$; zu Genauigkeitsverlusten beim Bilden dieser Differenzen kommt es, wenn sich λ und die Polstelle α_i kaum unterscheiden. Um dies zu vermeiden, verwenden wir anstelle von α_i und λ von vornherein die verschobenen Größen der Gestalt $\alpha_i - \mathit{shift} := \alpha_i^{neu}$, $\lambda - \mathit{shift} := x$. Liegt etwa der im Nullstellenverfahren zu berechnende Eigenwert λ nahe bei α_j, so wählt man $\mathit{shift} := \alpha_j$. Dies hat den Vorteil, daß man statt mit der dann kritischen Differenz $\alpha_j - \lambda$ mit

$$-x = (\alpha_j - \mathit{shift}) - (\lambda - \mathit{shift}),$$

d.h. mit dem Abstand zum kritischen Pol, arbeiten kann.
Die Verschiebungen werden dabei so gewählt, daß immer derjenige Pol als Bezugspunkt genommen wird, der der Nullstelle am nächsten ist.

4.3 J-Orthogonalität von Eigenvektoren und Aussagen zur Genauigkeit des Divide-et-Impera-Verfahrens

Die für $J = I$ bekannten Fehlerabschätzungen lassen sich nicht auf den Fall $J \neq I$ übertragen.
Normalerweise ist das Divide-et-Impera-Verfahren sowohl für $J = I$ als auch für $J \neq I$ ein recht genaues Verfahren. Bei einigen sehr speziellen Beispielen wie den sogenannten Wilkinsonmatrizen kann es allerdings zum Verlust der (J-) Orthogonalität der Eigenvektoren kommen. Dies ist ein für den Fall $J = I$ bekanntes Phänomen. Wir stellen drei Methoden aus der Literatur vor, die aufzeigen, wie man mit diesem Problem fertigwerden kann. Die zugehörigen Algorithmen (nicht aber die Theorie) lassen sich alle auf den Fall $J \neq I$ verallgemeinern.

Wie die Testbeispiele zeigen (vgl. den Abschnitt über Computerergebnisse), ist das Divide-et-Impera-Verfahren sowohl für $J = I$ als auch für $J \neq I$ ein schnelles und genaues Verfahren. Allerdings kann es bei einigen ziemlich speziellen Beispielen (etwa Wilkinson-Matrizen mit $J = I$, Wilkinson-Matrizen mit $J \neq I$, ein Diagonalelement von J ist gleich -1) zum Verlust der Orthogonalität bzw. der J-Orthogonalität von Eigenvektoren kommen. Dieses Verhalten ist für die bekannten Implementierungen des Divide-et-Impera-Verfahrens

zu $J = I$ bereits in der Literatur behandelt. Solche kritischen Fälle können auftreten, wenn sich die Eigenwerte in einem Teilschritt des Verfahrens fast überhaupt nicht mehr von den Polstellen der zugrundeliegenden Nullstellenfunktion w unterscheiden; dies ist etwa bei sehr kleinen **z**-Komponenten der Fall. Setzt man bei der Deflation nach kleinen **z**-Komponenten die Deflationsschranke etwa von einem Wert in Größenordnung der Maschinengenauigkeit auf Wurzel aus der Maschinengenauigkeit hinauf, so erhält man wieder orthogonale Vektoren - durch die größere Anzahl von Deflationen wird das Verfahren sogar schneller, aber der Preis dafür ist eine insgesamt geringere Genauigkeit der Eigenvektoren.
Nun sind aber sehr kleine **z**-Komponenten nicht generell gefährlich - was auch ein großes Problem wäre, da dies sehr häufig auftritt. Zum Verlust der Orthogonalität kommt es im Fall $J = I$ erst, wenn für zwei Eigenwerte λ_i, λ_{i+1} gilt: $\lambda_i \approx$ Pol Nr. $i \approx \lambda_{i+1}$. (Diese wesentliche Erkenntnis findet man in der Arbeit von Kevin Gates [27]). Im Fall $J \neq I$ kann ein Eigenwert, der fast vollständig mit einer Polstelle übereinstimmt, die gleichzeitig ein Randpunkt des Definitheitsintervalls ist, ebenfalls zu nicht ausreichend J-orthogonalen Eigenvektoren führen. (Dazu müssen die zugehörigen **z**-Komponenten gar nicht so winzig sein, vgl. Computerergebnisse für Wilkinson-Matrizen mit $J \neq I$).
Mit dem Verlust der Orthogonalität und mit Genauigkeitsabschätzungen haben sich im Fall $J = I$ zahlreiche Autoren beschäftigt. Im folgenden wird versucht, die wesentlichen Ergebnisse und Ideen kurz vorzustellen und zu erläutern, wie die Situation für $J \neq I$ aussieht.

In der Arbeit von Bunch, Nielsen, Sorensen [13] wird u.a. die Empfindlichkeit der Eigenvektoren des Eigenwertproblems

$$(4.24) \qquad (D + \rho \mathbf{z}\mathbf{z}^T)\mathbf{x} = \lambda \mathbf{x}$$

untersucht; je besser die Nullstellen der Funktion f mit

$$(4.25) \qquad f(\lambda) = 1 + \rho \sum_{i=1}^{n} \frac{z_i^2}{d_i - \lambda}$$

berechnet werden, umso weniger unterscheidet sich der berechnete Eigenvektor **x** zum Eigenwert λ von dem tatsächlichen (vgl. Theorem 6 in [13]).
Es stellt sich heraus, daß Aussagen über die Ableitung der Nullstellenfunktion wesentlich für die Genauigkeitsanalyse sind. Nach Cuppen bestimmt eine untere Schranke für die Größe der Ableitung $f'(\lambda)$ die Genauigkeit, mit der Eigenwerte λ als Nullstellen von f berechnet werden können (vgl. Theorem 3.1 und Korollar 3.1 in Cuppen [19]). Als Abschätzung für f' findet Cuppen

$$(4.26) \qquad \rho f'(\lambda_i) \geq 1$$

Bezüglich der Orthogonalität und der Genauigkeit braucht man eine möglichst kleine obere Schranke für f', wie aus den Lemmata 3.2 und 3.3 in [19] für die Eigenvektoren $\mathbf{y}_i$ hervorgeht:

$$|\mathbf{y}_i^T \mathbf{y}_k| \lesssim \frac{2\,L\,\varepsilon}{|\lambda_i - \lambda_k|}\left(2 + \sqrt{\rho \max(f'(\lambda_i), f'(\lambda_k))}\right)$$

(4.27) (Lemma 3.2 aus Cuppen [19])

$$\|(D + \rho\,\mathbf{z}\mathbf{z}^T - \lambda_i I)\mathbf{y}_i\|_2 \lesssim L\,\varepsilon\sqrt{2 + 8\,\rho\,f'(\lambda_i)}$$

(Lemma 3.3 aus Cuppen [19])

Dabei ist $L = 2\|T\|_\infty$ ($\|\ \|_\infty$ Zeilensummennorm und T die ursprüngliche Tridiagonalmatrix in $T\mathbf{x} = \lambda\,\mathbf{x}$) und kann durch $L \geq \rho$ und $L \geq \max_{i,j} |d_j - \lambda_i|$ abgeschätzt werden; ε ist die Maschinengenauigkeit.

Cuppen hat für f' folgende obere Abschätzung erhalten:

$$\rho\,f'(\lambda_i) \lesssim \frac{8}{\delta^2\zeta^2} \text{ (Approximation erster Ordnung für kleine } \delta \text{ und } \zeta)$$

Hier wird δ vom Abstand zweier benachbarter Polstellen und der Größe $2\|T\|_\infty$ bestimmt:

$$|d_{i+1} - d_i| > 2\,\delta\,\|T\|_\infty$$

und ζ durch die Größenordnung der **z**-Komponenten z_i: $|z_i| > \zeta$.

Korollar 3.1 aus [19] besagt dann, daß "die Eigenwerte λ_i aus $f(\lambda) = 0$ mit fast vollständiger Genauigkeit bzgl. des größten Eigenwertes der Matrix T berechnet werden können" (d.h. der Fehler bei der Nullstellenbestimmung ist von der gleichen Größenordnung wie der Fehler bei der Bestimmung des betragsgrößten Eigenwertes.)

Während im Fall J = I insbesondere die Abschätzung (4.26) sehr einfach zu erhalten ist (sie folgt fast direkt mit Hilfe der Cauchy-Schwarz'schen Ungleichung), sind geeignete Abschätzungen im Fall J ≠ I nicht bekannt und wären auch sehr unwahrscheinlich. Die Nullstellenfunktion w hat nämlich wegen der wechselnden Vorzeichen einen komplizierteren Verlauf, insbesondere sind die üblichen Monotoneigenschaften für eine Funktion der Art

$$w(\lambda) = 1 - \rho \sum_{i=1}^{n} \frac{z_i^2\, j_i}{\alpha_i - \lambda}$$

nicht gegeben (vgl. Abschnitt 4.1.3.2)

Aus den hier dargestellten Ideen und Ergebnissen ersieht man, daß es im Fall J ≠ I nicht leicht sein wird, allgemeine Genauigkeitsaussagen für das Divide-et-Impera-Verfahren zu machen und eine vollständige Fehleranalyse durchzuführen (für eine vollständige Fehleranalyse im Fall J = I s. auch Barlow [5]).

Um mit den Schwierigkeiten "nichtorthogonaler" Eigenvektoren fertigzuwerden, schlägt Cuppen eine Reorthogonalisierung der Eigenvektoren vor und macht Abschätzungen, welche Fehler dabei auftreten können. Bei diesem Vorgehen sind die Eigenvektoren danach zwar orthogonal, aber es kommen zusätzliche Fehler hinzu, die sich auf die Genauigkeit insgesamt auswirken.

Eine solche Reorthogonalisierung ist auch im Fall J ≠ I möglich, wenn man ein verallgemeinertes Gram-Schmidt-J-Orthogonalisierungsverfahren verwendet; da die während des Algorithmus auftretenden Teilprobleme positiv definit sind, tritt der Fall $\mathbf{y}^T J\,\mathbf{y} = 0$ für einen Eigenvektor **y** nie auf, und das Verfahren erzeugt J-orthogonale Vektoren.

Eine andere Idee im Fall J = I war, teilweise mit höherer Genauigkeit zu arbeiten, insbesondere bei der Funktionsauswertung der Funktion f (s. Formel 4.25) und bei der Bestimmung der Eigenwerte, vgl. etwa (Kahan [34]). Kahan wies in seiner Arbeit auch darauf hin, daß die Deflation zu mehrfachen Eigenwerten unkritisch ist. Den Durchbruch schafften Sorensen und Tang mit ihrer Arbeit *"On the Orthogonality of Eigenvectors Computed by Divide-and-Conquer Techniques"* (Sorensen, Tang [53]). Sie zeigten, daß im Fall J = I die Orthogonalität der Eigenvektoren garantiert werden kann, wenn man bei der Nullstellenbestimmung doppelt so genau wie sonst arbeitet (da man in den Standardprogrammiersprachen numerische Operationen bereits mit "doppelter Genauigkeit" ausführt, d.h. einen Datentyp wie "extended", "double" hat, müßte man die Nullstellen mit "vierfacher Genauigkeit" berechen.) Bei diesem Vorgehen kann man für J = I sicher sein, orthogonale Eigenvektoren zu erhalten, der Nachteil dieser Methode ist die Maschinenabhängigkeit der Implementierung des Algorithmus und natürlich auch die höhere Rechenzeit.

Die theoretische Grundlage liefert dabei das Lemma 4.7 aus Dongarra und Sorensen [24], wo Abschätzungen für Eigenvektoren **u**, **v** zu vom Algorithmus berechneten Näherungen λ', μ' der exakten Eigenwerten λ, μ angegeben werden:

Es gelte:

$d_i - \lambda' = (d_i - \lambda)(1 + \vartheta_i)$, $d_i - \mu' = (d_i - \mu)(1 + \eta_i)$, $i = 1, \dots n$ und $|\vartheta_i|, |\eta_i| \leq \varepsilon \ll 1$.

Dann folgt für die Eigenvektoren **u**, **v**:

$$|\mathbf{u}^T \mathbf{v}| = \left|\mathbf{u}^T \operatorname{diag}\left(\frac{\vartheta_i + \eta_i + \vartheta_i\,\eta_i}{(1 + \vartheta_i)(1 + \eta_i)}\left(\frac{f'(\lambda)\,f'(\mu)}{f'(\lambda')\,f'(\mu')}\right)^{1/2}\right)_{i=1,\dots,n} \mathbf{v}\right|$$

$$\leq \varepsilon\,(2 + \varepsilon)\left(\frac{1 + \varepsilon}{1 - \varepsilon}\right)^2$$

Das Lemma zeigt, daß man numerische Orthogonalität sicherstellen kann, wenn man die Größen $d_i - \lambda$, $i = 1, \dots, n$ mit hoher relativer Genauigkeit zu berechnen in der Lage ist.

Wichtig für dieses Lemma und die weiterführenden Überlegungen in Sorensen und Tang [53] ist, daß $f'(\lambda)$ gut nach unten abschätzbar ist und nie null wird.
Tests haben gezeigt, daß doppelte Genauigkeit bei der Nullstellenberechnung auch sehr sinnvoll im Fall $J \neq I$ ist. Der Beweis, daß sie auch immer genügt, dürfte schwierig sein - eine direkte Übertragung der Überlegungen von Sorensen/Tang auf $J \neq I$ ist nicht möglich, da aus den zuvor erwähnten Gründen keine einfachen Abschätzungen für w und w' existieren.

Eine völlig andere Möglichkeit, orthogonale Vektoren im Fall $J = I$ zu erhalten, zeigt Gates in seiner Dissertation auf (Gates [27]). Er stellt fest, daß die inverse Iteration (geeignet implementiert) relativ schnell bei Tridiagonalmatrizen ist, solange keine Eigenvektoren für numerisch mehrfache Eigenwerte berechnet werden. Beim Divide-et-Impera-Verfahren werden numerisch mehrfache Eigenwerte mittels Deflation behandelt, und daher gibt es keine Orthogonalitäts- oder Genauigkeitsprobleme für die Eigenvektoren zu mehrfachen Eigenwerten. Kritisch sind dagegen nahe beieinanderliegende Eigenwerte $\lambda_i \approx$ Pol Nr. $i \approx \lambda_{i+1}$, die numerisch nicht mehrfach sind.
Für ein Eigenwertproblem der Gestalt

$$(4.28) \qquad T\,\mathbf{x} = \left(\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + \rho\,\mathbf{z}\mathbf{z}^T \right) \mathbf{x} = \lambda\,\mathbf{x}$$

bestimmt Gates ebenfalls die Eigenwerte mittels der Nullstellengleichung $f(\lambda) = 0$, nachdem Deflation für die mehrfachen Eigenwerte durchgeführt wurde. Die noch zu bestimmenden Eigenvektoren berechnet er mittels inverser Iteration, wobei er direkt mit der Matrix T arbeitet und dadurch eine Matrix-Matrix-Multiplikation im Vergleich zum Standard-Divide-Verfahren spart. Den Aufwand seines Verfahrens bezeichnet er als in etwa vergleichbar mit dem von Sorensen/Tang beschriebenen mit verdoppelter Stellenlänge.

Inverse Iteration ist auch im Fall $J \neq I$ möglich. Wichtig dabei ist, daß die vorkommenden Gleichungssysteme gut lösbar sind.
Dieses Vorgehen stellt aber eine deutliche Modifikation des ursprünglichen Divide-et-Impera-Verfahrens dar und wird daher hier nicht verfolgt.

Gu und Eisenstat [32] fanden im Fall $J = I$ eine weitere Möglichkeit, orthogonale Eigenvektoren zu erhalten. Sie schauten sich dazu die bekannten Rechnungen für die Rang-1 - Teilprobleme näher an:

Zu einer gegebenen Diagonalmatrix $D = \mathrm{diag}(d_1, d_2, \dots, d_n)$ mit $d_1 < d_2 < \dots < d_n$ und einem reellen Vektor $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$ mit $z_j > 0$ finde man die Spektralzerlegung

$$D + \mathbf{z}\mathbf{z}^T = Q\,\Lambda\,Q^T \quad \text{mit} \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

Bei den üblichen Divide-et-Impera-Verfahren (analog auch in dieser Arbeit) werden die Eigenvektoren $\mathbf{q}_i$ mithilfe von $\mathbf{z}$ und λ_i nach der Formel

$$(4.29) \qquad \mathbf{q}_i = \left(\frac{z_1}{d_1 - \lambda_i}, \dots, \frac{z_n}{d_n - \lambda_i} \right)^T \Big/ \sqrt{\sum_j \left(\frac{z_j}{d_j - \lambda_i} \right)^2}$$

bestimmt. Es ist leicht erkennbar, daß kleine Veränderungen in λ_i große Veränderungen in diesen Eigenvektoren bewirken. Der Algorithmus liefert aber keine exakten Lösungen λ_i, sondern nur Näherungen λ'_i; dies kann sich sehr ungünstig auf die Qualität der Eigenvektoren auswirken. Wenn man allerdings ein $\mathbf{z}'$ finden könnte, so daß λ'_i, $i = 1, \dots, n$ die *exakten* Eigenwerte einer *neuen* Rang-1-Modifikation der Gestalt $D + \mathbf{z}'\mathbf{z}'^T$ sind, so erhält man mit (4.29) und $\mathbf{z}'$ statt $\mathbf{z}$ einen genau berechneten Eigenvektor zu λ'_i.
Wegen

$$D + \mathbf{z}\mathbf{z}^T = (D + \mathbf{z}'\mathbf{z}'^T) + (\mathbf{z} - \mathbf{z}')\mathbf{z}^T + \mathbf{z}(\mathbf{z} - \mathbf{z}')^T - (\mathbf{z} - \mathbf{z}')(\mathbf{z} - \mathbf{z}')^T$$

weicht $D + \mathbf{z}\mathbf{z}^T$ von $D + \mathbf{z}'\mathbf{z}'^T$ in erster Näherung linear ab.

Gu und Eisenstat rechnen dann aus, daß solch ein $\mathbf{z}'$ unter den angegebenen Voraussetzungen immer existiert und geben die Formel für $\mathbf{z}'$ an.

Eine Verallgemeinerung dieser Idee auf den Fall $J \neq I$ ist direkt möglich, vgl. Veselič [64]. Zu einem gegebenen positiv definiten Paar (A, B) bestimmt er einen Vektor $\mathbf{z}$, so daß das Paar $(A \pm \mathbf{z}\mathbf{z}^T, B)$ vorgegebene Eigenwerte hat, die gewissen Durchwebungseigenschaften genügen. Er beschreibt sowohl den Fall, wo $(A \pm \mathbf{z}\mathbf{z}^T, B)$ ebenfalls positiv definit ist (vgl. Theorem 1 in [64]) als auch den, wo die positive Definitheit verlorengeht (vgl. Theorem 2 in [64]). In beiden Fällen ist die Berechnung des Vektors $\mathbf{z}$ möglich.

Wir erhalten also die für unser Eigenwertproblem nötigen Ergebnisse:
Der Vektor $\mathbf{z}'$ kann immer berechnet werden, das mit diesem $\mathbf{z}'$ gebildete Matrizenpaar $(D - \rho\,\mathbf{z}'\mathbf{z}'^T, J)$ ist ebenfalls noch positiv definit.
Die Komponenten von $\mathbf{z}'$ erhält man in unserem Fall mithilfe der Formel

$$(4.30) \qquad z'^2_i = -j_i \frac{\prod_{j=1}^{n} (\lambda_j - \alpha_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\alpha_j - \alpha_i)} .$$

Anders als bei Gu und Eisenstat muß hier noch das Vorzeichen fixiert werden.

4.4 Ausblick: Divide-et-Impera-Verfahren für nichtsymmetrische gewöhnliche Eigenwertprobleme

Es werden kurz einige theoretische Ergebnisse und Verfahren zur Lösung des nichtsymmetrischen, tridiagonalen Eigenwertproblems vorgestellt und die typischen Schwierigkeiten aufgezeigt, die einem lauffähigen Algorithmus im Wege stehen.

Neben dem Divide-et-Impera-Verfahren für ein tridiagonales Standard-Eigenwertproblem wurden Divide-et-Impera-Techniken auch für verwandte Problemstellungen verwendet. So behandelt Arbenz in [2] das Eigenwertproblem für eine Bandmatrix; Gragg und Reichel studieren in [31] unitäre und orthogonale Eigenwertprobleme, Thompson sucht in [59] nicht nur Eigenwerte, sondern auch Singulärwerte.
Arbenz, Gander und Golub stellen fest, wie man die Eigenwerte und Eigenvektoren einer Matrix $A + V V^T$ erhält, wenn A eine symmetrische Matrix mit bekannter Spektralzerlegung ist und $V V^T$ eine positiv semidefinite Matrix von niedrigem Rang (vgl. [3]). In einer weiteren Arbeit [4] untersuchen Arbenz und Golub auch den hermiteschen Fall.
In anderen Arbeiten werden die Divide-et-Impera-Techniken zum Teil deutlich modifiziert und andersartige Divide-et-Impera-Algorithmen entwickelt. So arbeiten etwa Bini und Pan (Bini, Pan und Pan [10]) nicht mit der üblichen Nullstellenfunktion, sondern direkt mit charakteristischen Polynomen, K. Li und T. Li kombinieren in [39] sogar Homotopiemethoden mit Divide-et-Impera-Methoden. Beattie und Ribbens [8] verwenden die Weinstein-Aronszajn-Theorie, um ein Verfahren für das verallgemeinerte Eigenwertproblem $T\mathbf{x} = \lambda S\,\mathbf{x}$ mit zwei Tridiagonalmatrizen T und S zu entwickeln.

Ein weiterer Forschungsschwerpunkt sind im Moment Divide-et-Impera-Verfahren für das nichtsymmetrische Eigenwertproblem (vgl. auch Dongarra und Sidani [23]).
Wie wir in Kapitel 2 diskutiert haben, können symmetrische verallgemeinerte Eigenwertprobleme mit nichtsingulärem J in ein nichtsymmetrisches Eigenwertproblem

$$(4.31) \qquad A_{Trid}\mathbf{x} = \lambda\,\mathbf{x}$$

umgeschrieben werden. Zumindest im positiv definiten Fall ist das nicht sinnvoll. Aber auch im allgemeinen Fall scheint ein Übergang von einem verallgemeinerten symmetrischen Eigenwertproblem auf ein gewöhnliches, aber nicht symmetrisches Problem nicht empfehlenswert: E. Jessup zeigt in ihrer Arbeit *"A Case Against a Divide and Conquer Approach to the Nonsymmetric Eigenvalue Problem"* ([33]) die auftretenden Probleme auf.

Um die Schwierigkeiten eines nichtsymmetrischen Divide-et-Impera-Verfahrens im direkten Vergleich mit dem symmetrischen Verfahren zu verdeutlichen, wendet sie das symmetrische Verfahren auf das nichtsymmetrische Tridiagonaleigenwertproblem an; dabei setzt sie voraus, daß die Untermatrizen, die durch das Aufteilen entstehen, diagonalisierbar sind.
Ein wesentlicher Unterschied zum symmetrischen Fall ergibt sich schon beim Aufspalten der Startmatrix: Im Gegensatz zum symmetrischen Fall spaltet Jessup eine Rang-2-Matrix ab:

$$(4.32) \qquad A_{Trid} = \begin{bmatrix} T_1 & & \\ & \alpha_m & \\ & & T_2 \end{bmatrix} + \text{Rang-2-Störung},$$

wobei die Diagonalelemente von A_{Trid} die α_i sind. Die Eigenwerte erhält man wieder als Nullstellen einer speziellen rationalen Funktion g. Da diese Funktion aus der Spektralzerlegung von T_1 und T_2 entsteht, kann sie komplexe Koeffizienten haben. Die komplexen Nullstellen von g treten als konjugierte Paare auf; im Gegensatz zum symmetrischen Fall, wo die Nullstellen die Pole "durchweben", gibt es keine offensichtliche Beziehung zwischen der Lage der Pole und der Nullstellen von g.
Jessup macht einige Vorschläge zur Nullstellensuche (insbesondere ein Verfahren, das auf der Cauchy'schen Integralformel beruht) und weist auf die Probleme mit diesen Methoden hin; es fehlt ein speziell an die Funktion g angepaßtes Verfahren (vgl. Abschnitt 5.1 in [33]).
Als äußerst kritisch sieht Jessup die mögliche Instabilität der Divide-et-Impera-Methode an: Auch für eine bzgl. des Eigenwertproblems gut konditionierte Originalmatrix kann eine schlechtkonditionierte Untermatrix auf jeder Stufe des "Updating" erzeugt werden. So ist zu befürchten, daß selbst kleine Fehler beim Zerlegen, der Deflation oder dem Impera-Schritt zu großen Fehlern in der berechneten Spektralzerlegung führen. Da die Originalmatrix beim Impera-Schritt nie verwendet wird, gibt es keine Möglichkeit, solche Fehler zu korrigieren.

Während Jessup hauptsächlich die Schwierigkeiten mit einem nichtsymmetrischen Divide-et-Impera-Verfahren aufzeigt, haben Loice Adams und Peter Arbenz in ihrer Arbeit *"Towards a Divide and Conquer Algorithm for the Real Nonsymmetric Eigenvalue Problem"* (Adams und Arbenz [1]) versucht, die theoretischen Grundlagen für einen solchen Algorithmus zu entwickeln.
Sie haben einen denkbaren Algorithmus skizziert, aber noch nicht implementiert, und auch kein brauchbares Nullstellenverfahren angegeben. Die durch die endliche Rechengenauigkeit auftretenden Probleme wie numerische Vielfachheit von Eigenwerten oder numerisch linear unabhängige Eigenvektoren zu nahe beieinanderliegende Eigenwerte sind ebenfalls noch nicht behandelt worden.

Sie betrachten in ihrer Arbeit zunächst das Eigenwertproblem

(4.33) $A' \mathbf{x} = (A + U V^T) \mathbf{x} = \lambda \mathbf{x}, \quad U, V \in \mathbb{R}^{n \times r}, \ \mathrm{rang}(U) = \mathrm{rang}(V) = r$

und bestimmen die Eigenwerte und Eigenvektoren von A' unter der Voraussetzung, daß man alle Eigenwerte und eine orthonormale Basis von zugehörigen Eigenräumen von A kennt.
Die Autoren demonstrieren ihre theoretischen Überlegungen am Fall eines nichtsymmetrischen Tridiagonaleigenwertproblems

(4.34) $$T' = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + \mathbf{u}\mathbf{v}^T, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^n,$$

wobei nicht vorausgesetzt wird, daß T' diagonalisierbar ist.
Wenn $\beta_1, \ldots, \beta_{n-1}$ die obere Nebendiagonale von T' ist und $\gamma_2, \ldots, \gamma_n$ die untere Nebendiagonale und wir die Matrix T' an der Stelle β_k, γ_{k+1} aufspalten, dann ist die Nullstellenfunktion $Z(\lambda)$ für die Eigenwerte von T', die keine Eigenwerte von T sind, durch

(4.35) $$Z(\lambda) = 1 + \frac{\beta_k}{w} \mathbf{e}_k^T \left(T_1 - \lambda\right)^{-1} \mathbf{e}_k + w\, \gamma_{k+1}\, \mathbf{e}_1^T \left(T_2 - \lambda\right)^{-1} \mathbf{e}_1 = 0$$

gegeben. (Das w wird bei der Abspaltung geeignet gewählt.)
Die Ausdrücke $\left(T_1^{-1} - \lambda\right) \mathbf{e}_k$ und $\left(T_2^{-1} - \lambda\right) \mathbf{e}_1$ müssen dabei - etwa durch Gauß'sche Elimination mit partieller Pivotisierung - ausgewertet werden. (Man beachte, daß man im symmetrischen Fall statt $(T_i - \lambda)$ nur $(D_i - \lambda)$ mit Diagonalmatrizen D_i, $i = 1, 2$ hat !)

Auf der Grundlage dieser Überlegungen wird dann ein rekursives Divide-et-Impera-Verfahren für das Eigenwertproblem $T' \mathbf{x} = \lambda \mathbf{x}$ beschrieben.

Insgesamt scheint der Weg zu einem brauchbaren nichtsymmetrischen Divide-et-Impera-Verfahren noch weit zu sein. Wenn das Eigenwertproblem in der Form $A\mathbf{x} = \lambda B\mathbf{x}$ mit symmetrischen Matrizen A, B gegeben ist, sollte man meiner Meinung nach nicht wie oben beschrieben zu einem nichtsymmetrischen Problem übergehen, sondern stattdessen das Divide-et-Impera-Verfahren für den positiv definiten Fall auf den Fall allgemeiner symmetrischer Paare (A, B) ausdehnen. Die auftretenden Probleme beim symmetrischen verallgemeinerten Eigenwertproblem dürften insgesamt wohl leichter zu bewältigen sein als beim nichtsymmetrischen gewöhnlichen Eigenwertproblem. Eine geeignete Implementierung eines solchen verallgemeinerten Divide-et-Impera-Verfahrens ist eine reizvolle Aufgabe für die Zukunft.

5. Zur Reduktion der Matrix A auf Tridiagonalgestalt

Wir transformieren ein Matrizenpaar (A, J), A symmetrische, vollbesetzte Matrix und J Diagonalmatrix mit +1 und -1 auf der Diagonalen, mittels verallgemeinerter Householdertransformationen auf ein Paar (T_{mod}, J'), wobei J' (bis auf Vertauschungen) dieselbe Gestalt wie J hat und T_{mod} entweder eine symmetrische Tridiagonalmatrix ist oder eine symmetrische Matrix, die bis auf wenige vollbesetzte Zeilen und Spalten tridiagonal ist ("Ineinanderschachtelung berandeter Tridiagonalmatrizen"). Auf die Transformation dieser Zeilen und Spalten wurde während des Tridiagonalisierungsprozesses verzichtet, da ihre Reduktion die Division durch eine spezielle kleine Größe (oder gar durch null) verlangt hätte. Um festzustellen, wie häufig solche kritischen Fälle auftreten, führen wir mit Zufallsmatrizen eine statistische Untersuchung durch.

Das im vorherigen Kapitel beschriebene Divide-et-Impera-Verfahren überprüft Matrizenpaare (T, J) mit T Tridiagonalmatrix, $J = \mathrm{diag}(j_i)$ mit $j_i = +1$ oder -1, auf ihre positive Definitheit und löst für positiv definite Paare das zugehörige Eigenwertproblem. Wir wollen in diesem und im nächsten Kapitel untersuchen, wie man im Fall eines Eigenwertproblems $A\mathbf{x} = \lambda J\mathbf{x}$, wobei A jetzt eine vollbesetzte Matrix ist und J wie oben angegeben, vorgehen kann. (Im Kapitel 2 hatten wir ja bereits gesehen, daß man von einem Eigenwertproblem $A'\mathbf{x} = \lambda B\mathbf{x}$, B nichtsingulär, zu einem Eigenwertproblem der Gestalt $A\mathbf{x} = \lambda J\mathbf{x}$ übergehen kann. Daher beschränken wir uns im folgenden auf den Fall B = J.)

In diesem Kapitel werden wir eine Reduktion des Paares (A, J) auf (T_{mod}, J') durchführen, wobei die Matrix T_{mod} entweder eine Tridiagonalmatrix oder eine Tridiagonalmatrix mit einigen wenigen zusätzlich besetzten Zeilen und Spalten ist (also eine "Ineinanderschachtelung berandeter Tridiagonalmatrizen"). Im darauffolgenden Kapitel wird gezeigt, in welcher Weise unser Divide-et-Impera-Algorithmus auf (verallgemeinerte) Eigenwertprobleme mit Matrizen dieses Typs anwendbar ist.

5.1 Teilweise Reduktion auf Tridiagonalgestalt

Wir beschreiben hier die zur teilweisen Tridiagonalisierung des Paares (A, J) notwendigen verallgemeinerten Householdertransformationen und zeigen auf, wann eine solche Transformation nicht möglich oder nicht sinnvoll ist. Desweiteren geben wir den vollständigen Transformationsalgorithmus in Struktogrammform an.

Das Eigenwertproblem $A\mathbf{x} = \lambda J\mathbf{x}$ bzw. das äquivalente $C\mathbf{x} = \lambda \mathbf{x}$ mit $C := JA$ kann *theoretisch* immer in ein Eigenwertproblem der Gestalt

$$T\mathbf{z} = \lambda \mathbf{z}$$

umgewandelt werden mit T Tridiagonalmatrix. (Man vergleiche hierzu Theorem 1 aus Bunse-Gerstner, "On the Similarity Transformation to Tridiagonal Form" [15]).

Um die zu C ähnliche Tridiagonalmatrix T zu berechnen, bieten sich die folgenden Möglichkeiten an :

1. der Lanczos-Prozeß (Golub, van Loan [29], Steinhoff [54])

2. die Transformation von C auf untere Hessenbergform und anschließende weitere Elimination zur Reduktion auf Tridiagonalgestalt (Wilkinson [68])

3. Reduktion von C auf Tridiagonalgestalt mittels sukzessiver elementarer Ähnlichkeitstransformationen (Wilkinson [68])

Diese Verfahren sind jedoch nicht immer stabil. Bunse-Gerstner untersucht in [14], [15] und [16] den Zusammenhang zwischen den drei Verfahren und zeigt auf, wann sie scheitern können: Bei allen drei Verfahren wählt man zwei beliebige Startvektoren und konstruiert davon ausgehend eine Transformationsmatrix, die dann die gegebene Matrix C auf Tridiagonalgestalt bringt. Die beiden Startvektoren bilden dabei die erste Spalte bzw. erste Zeile in der Transformationsmatrix bzw. in ihrer Inversen. Aus Satz 6.1.2 in [14] (dem "Eindeutigkeitssatz") folgt, daß die so gebildeten Transformationsmatrizen der drei Verfahren bei *gleichen Startvektoren* bis auf Skalierung gleich sind. Alle Verfahren zur Tridiagonalisierung müssen abbrechen, wenn das Verfahren die Division durch einen verschwindenden Ausdruck verlangt. Wird ein solcher Ausdruck fast null, so können große Rundungsfehler auftreten, wenn er im Verhältnis zu den zu dividierenden Größen sehr klein ist. Aus dem Korollar 6.1.10 in [14] folgt, daß man dann nur die Möglichkeit hat, die Gesamttransformationsmatrix in der ersten Spalte oder in ihrer Inversen in der ersten Zeile zu ändern und den Prozeß neu zu starten. Allerdings hat man auch dann keine Garantie, daß die Tridiagonalisierung jetzt vollständig bzw. mit nicht zu großen Rundungsfehlern durchführbar ist. Die Auswahl des Verfahrens spielt keine Rolle: Bei gleichen Startvektoren brechen alle drei Prozesse nach Korollar 6.1.10 im entsprechenden Teilschritt ab, bzw. es treten große Rundungsfehler auf.

Wir werden daher im folgenden das Eigenwertproblem $\mathbf{A}\mathbf{x} = \lambda \mathbf{J}\mathbf{x}$ mit Hilfe von *modifizierten Householdermatrizen* transformieren. Dabei muß die entstehende Matrix T_{mod} aus dem äquivalenten Eigenwertproblem

$$T_{mod}\,\mathbf{z} = \lambda\, J'\mathbf{z} \tag{5.1}$$

nicht unbedingt mehr eine Tridiagonalmatrix sein: Für die Zeilen bzw. Spalten, wo die Transformation mittels einer modifizierten Householdermatrix nicht oder nur zu ungenau möglich ist, verzichten wir auf die Durchführung der Transformation und gehen zur nächsten Zeile und Spalte über. Nach Durchführung des Algorithmus ist T entweder eine Tridiagonalmatrix oder eine Ineinanderschachtelung von "berandeten" Tridiagonalmatrizen.
(Die Idee, Zeilen und Spalten bei einer solchen Transformation stehenzulassen, stammt von Dax und Kaniel, vgl. [20].)

Für Dimension von A gleich 8 und Scheitern der Transformation in Zeile und Spalte 3 und 5 erhält man z.B. die folgende (symmetrische) Matrix:

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ t_{21} & t_{22} & t_{23} & 0 & 0 & 0 & 0 & 0 \\ 0 & t_{32} & t_{33} & t_{34} & t_{35} & t_{36} & t_{37} & t_{38} \\ 0 & 0 & t_{43} & t_{44} & t_{45} & 0 & 0 & 0 \\ 0 & 0 & t_{53} & t_{54} & t_{55} & t_{56} & t_{57} & t_{58} \\ 0 & 0 & t_{63} & 0 & t_{65} & t_{66} & t_{67} & 0 \\ 0 & 0 & t_{73} & 0 & t_{75} & t_{76} & t_{77} & t_{78} \\ 0 & 0 & t_{83} & 0 & t_{85} & 0 & t_{87} & t_{88} \end{bmatrix}$$

5.1.1 Die modifizierte Householder-Transformationsmethode

Wie z.B. im Buch von Parlett ("The symmetric eigenvalue problem", Abschnitt 7. 4 : "Explicit Reduction of a full matrix", [46]) beschrieben, kann man jede symmetrische Matrix A durch eine Folge von einfachen orthogonalen Ähnlichkeitstransformationen, die Nullelemente Zeile für Zeile und damit auch Spalte für Spalte erzeugen, in eine Tridiagonalmatrix T transformieren. Geeignete Transformationsmatrizen sind z.B. die sogenannten Householdermatrizen, die wir im folgenden auch verwenden werden. Zur Reduktion der ersten Zeile und Spalte von A schreibt man sich A als

$$A =: A^{(1)} = \begin{bmatrix} a_{11} & \mathbf{a}_1^T \\ \mathbf{a}_1 & M_1 \end{bmatrix} \qquad \text{mit } \mathbf{a}_1^T = (a_{12}, \dots, a_{1n}).$$

Man bestimmt eine orthogonale (n-1, n-1)- Matrix H_1, die $H_1^T \mathbf{a}_1 = \alpha_1 \mathbf{e}_1$ erfüllt mit $\alpha_1 \in \mathbb{R}$.
Eine Ähnlichkeitstransformation der Form

$$A_{mod} = A^{(1)}_{mod} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & H_1^T \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{a}_1^T \\ \mathbf{a}_1 & M_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & H_1 \end{bmatrix} = \begin{bmatrix} a_{11} & \mathbf{a}_1^T H_1 \\ H_1^T \mathbf{a}_1 & H_1^T M_1 H_1 \end{bmatrix}$$

liefert eine Matrix $A^{(1)}_{mod}$, die tridiagonal in ihrer ersten Zeile und Spalte ist:

$$A^{(1)}_{mod} = \begin{bmatrix} a_{11} & \alpha_1 & 0 \dots 0 \\ \alpha_1 & & \\ 0 & & H_1^T M_1 H_1 \\ \vdots & & \\ 0 & & \end{bmatrix}$$

Wenn man H_1 ausgesucht hat, wird $A^{(2)} = H_1^T M_1 H_1$ direkt berechnet.
Im zweiten Schritt wird dieselbe Technik auf

$$A^{(2)} = \begin{bmatrix} a_{22}^{(2)} & \mathbf{a}_2^T \\ \mathbf{a}_2 & M_2 \end{bmatrix}, \quad A^{(2)} \in M^{(n-1,n-1)}(\mathbb{R})$$

angewendet. Man wählt eine orthogonale $(n-2) \times (n-2)$-Matrix H_2 aus mit $H_2^T \mathbf{a}_2 = \alpha_2 \mathbf{e}_1$ und berechnet $A_3 = H_2^T M_2 H_2$. Die Ähnlichkeitstransformationen im zweiten Schritt zerstören dabei nicht die im ersten Schritt erzeugten Nullen.
Dieser Prozeß wird fortgesetzt, die Dimension der noch nicht tridiagonalen Untermatrix wird immer kleiner, und nach $(n-2)$ Schritten ist die Endmatrix T erreicht.

Im Fall eines Paares (A, J) müssen wir versuchen, A auf Tridiagonalgestalt zu bringen, ohne die Gestalt von J zu zerstören.
Wie gerade für den Fall einer symmetrischen Matrix A beschrieben, teilen wir jetzt A und J auf in

$$A = \begin{bmatrix} a_{11} & \mathbf{a}^T \\ \mathbf{a} & M \end{bmatrix}, \quad J = \begin{bmatrix} j_1 & \mathbf{0}^T \\ \mathbf{0} & J_M \end{bmatrix} \quad \text{mit } \mathbf{a}^T = (a_{12}, a_{13}, \dots, a_{1n}).$$

Dann wählen wir eine Transformationsmatrix $H_1 \in M^{n-1,n-1}(\mathbb{R})$ mit

(5.2) $\quad H_1^T \mathbf{a} = \alpha_1 \mathbf{e}_1$ und $H_1^T J_M H_1 = J_M$,

wenn dies möglich ist. Die Matrix H_1^T ist dabei eine Matrix vom Householdertyp der folgenden Gestalt (Bunse-Gerstner [16] und Steinhoff [54]) :

(5.3) $$H_{house} = H_{house}(\mathbf{w}) = I - \frac{2\mathbf{w}\mathbf{w}^T J_M}{\mathbf{w}^T J_M \mathbf{w}} \quad \text{für } \mathbf{w}^T J_M \mathbf{w} \neq 0$$

mit passend zu bestimmendem $\mathbf{w}$.

Diese Matrizen haben bekannterweise die folgenden Eigenschaften:

1. Sie sind J_M - hermitesch, d.h. $J_M H_{house} J_M = H_{house}^T$

2. Ihr Quadrat ist die Einheitsmatrix, woraus folgt:
$$H_{house}^{-1} = H_{house}$$

3. Es gilt $H_{house}^T J_M H_{house} = J_M$.

Ist $\mathbf{a}^T = (a_{12}, 0, \dots, 0)$, so wählt man $H_1^T = I$.

Im anderen Fall bestimmt man ein $\mathbf{w}$ aus

$$H_{house}\, \mathbf{a} = \alpha_1 \mathbf{e}_1 ,$$

was allerdings nur möglich ist, wenn die beiden folgenden Bedingungen erfüllt sind:

(5.4) $\quad \mathbf{a}^T J_M \mathbf{a} \neq 0$

(5.5) $\quad \operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) = \operatorname{sgn}(J_M)_{11} = \operatorname{sgn}(j_2)$.

Die Berechnung der Matrix H_{house} wird im gleich folgenden Struktogramm beschrieben. Wie man nachrechnen kann, ist die Bedingung $\mathbf{w}^T J_M \mathbf{w} \neq 0$ dann automatisch erfüllt.

Wir unterscheiden jetzt die folgenden Fälle:

1. $\underline{\mathbf{a}^T J_M \mathbf{a} = 0}$:

   In diesem Fall ist die Bestimmung einer Matrix H_{house} nicht möglich.
   Wir verzichten auf eine Transformation.

2. $\underline{\mathbf{a}^T J_M \mathbf{a} \neq 0}$ und $\underline{\operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) = \operatorname{sgn}(j_2)}$:

   Die Transformation ist durchführbar, und wir berechnen die Matrix $H^T = H_{house}$.

3. $\underline{\mathbf{a}^T J_M \mathbf{a} \neq 0}$ und $\underline{\operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) \neq \operatorname{sgn}(j_2)}$:

   Für $J_M = I$ oder $J_M = -I$ kann dieser Fall nie auftreten.
   J_M enthalte also jetzt p Pluseinsen $(p < n-1)$ und q Minuseinsen $(q < n-1)$ mit $p+q = \dim(J_M)$.
   Dann ist

$$\operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) = \operatorname{sgn}\Big(\sum_{i=2}^{n} a_{1i}^2 j_i\Big) = \operatorname{sgn}\Big(\sum_{j_i=+1} a_{1i}^2 - \sum_{j_i=-1} a_{1i}^2\Big).$$

   Ist $\operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) = 1$ und $j_2 = -1$, so muß es mindestens ein $s \in \{2, \dots, n\}$ geben mit $j_s = +1$, gilt $\operatorname{sgn}(\mathbf{a}^T J_M \mathbf{a}) = -1$ und $j_2 = +1$, so existiert ein $s \in \{2, \dots, n\}$ mit $j_s = -1$.
   Wir bestimmen ein solches s (im Algorithmus nehmen wir das erste geeignete aus $\{2, \dots, n\}$) und eine Matrix H_{house} mit

$$H_{house}\, \mathbf{a} = \alpha_s \mathbf{e}_s \quad \text{und setzen } H^T := H_{house} .$$

Nach der Transformation

$$\left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H^T \end{array}\right] \left[\begin{array}{c|c} a_{11} & a_{12} \dots a_{1s} \dots a_{1n} \\ \hline a_{21} & \\ \vdots & \\ a_{s1} & M \\ \vdots & \\ a_{n1} & \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H \end{array}\right] = \left[\begin{array}{c|c} a_{11} & 0 \dots \alpha_s \dots 0 \\ \hline 0 & \\ \vdots & \\ \alpha_s & H^T M H \\ \vdots & \\ 0 & \end{array}\right],$$

und

$$\left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H^T \end{array}\right] \left[\begin{array}{c|c} j_1 & \mathbf{0}^T \\ \hline \mathbf{0} & J_M \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H \end{array}\right] = \left[\begin{array}{c|c} j_1 & \mathbf{0}^T \\ \hline \mathbf{0} & J_M \end{array}\right],$$

da $H^T J_M H = J_M$ nach Konstruktion ist, führen wir noch eine Vertauschung der 2-ten und der s-ten Zeile und Spalte durch und erhalten

$$A_{mod} = P \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H^T \end{array}\right] \left[\begin{array}{c|c} a_{11} & \mathbf{a}^T \\ \hline \mathbf{a} & M \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & H \end{array}\right] P = \left[\begin{array}{c|c} a_{11} & \alpha_s \quad 0 \dots 0 \\ \hline \alpha_s & \\ 0 & \\ \vdots & M_{mod} \\ 0 & \end{array}\right],$$

$$J' = \mathrm{diag}(j_1, j_s, \dots, j_2, \dots, j_n),$$

wobei P die Permutationsmatrix zur Vertauschung der 2-ten und der s-ten Zeile und Spalte ist.

Struktogramm zur Berechnung der Householdermatrix

Wir befinden uns dabei im r-ten Schritt des Householdertransformationsalgorithmus, und die Bedingung (5.4) $\mathbf{a}^T J_M \mathbf{a} \neq 0$ ist als erfüllt vorausgesetzt. In diesem Fall wird eine verallgemeinerte Householdertransformation durchgeführt und die zugehörige Householdermatrix H_{house} berechnet sich entsprechend der folgenden Vorschrift:

Block 1: Bildung der Householdertransformationsmatrix H_{house}

$$\alpha := \sum_{i=r+1}^{n} a_{ir}^2 j_i$$

$\alpha j_{r+1} > 0$

Ja	Nein
(* keine Permutation *) Permutation := false	(* Permutation notwendig *) Permutation := true
$\alpha := -\sqrt{\alpha j_{r+1}}$	$s := r+1$
$a_{r+1,r} \neq 0$	Suche Index $s \geq r+1$ mit $\alpha j_s > 0$
Ja: $\alpha := \mathrm{sgn}(a_{r+1,r})\,\alpha$ / Nein:	$\alpha := -\sqrt{\alpha j_s}$
$w_1 := 1 - \frac{a_{r+1,r}}{\alpha}$	$a_{s,r} \neq 0$
$i = 2, \dots, n-r$: $w_i := -\frac{1}{\alpha} a_{r+i,r}$	Ja: $\alpha := \mathrm{sgn}(a_{sr})\,\alpha$ / Nein:
$\gamma := j_{r+1} / \lvert w_1 \rvert$	$w_{s-r} := 1 - \frac{a_{sr}}{\alpha}$
	$i = 1, \dots, n-r$: $i \neq s-r$: Ja: $w_i := -\frac{a_{r+i,r}}{\alpha}$ / Nein:
	$\gamma := j_s / \lvert w_{s-r} \rvert$

$$H_{house} := I - \gamma \mathbf{w} \mathbf{w}^T \mathrm{diag}(j_{r+1}, \dots, j_n)$$

Die Matrizen A_{mod} und J' übernehmen jetzt die Rollen von A und J.

Wie im Fall J gleich I werden einmal erzeugte Nullen nicht mehr zerstört, wenn sie durch Householderartige Transformationen entstanden sind.
Wird z.B. die k-te Zeile und Spalte wegen $\mathbf{a}^T J_M \mathbf{a} = 0$ nicht transformiert, so wirken sich alle folgenden Householderartigen Transformationen auf die Koeffizienten dieser Zeile und Spalte aus : Hat das nichtmodifizierte $\mathbf{a}$ die Gestalt

$$\mathbf{a}^T = (a_{k,k+1}, \dots, a_{kn}),$$

so ändern sich im k+j-ten Schritt die Komponenten

$$\left(a_{k,k+1+j}, \ldots, a_{kn}\right).$$

Wichtig für einen brauchbaren Householder-Transformations-Algorithmus ist, daß man die jeweiligen Untermatrizen $H^T M H$ gut berechnen kann, d.h. möglichst ohne alle Matrizenmultiplikationen $H^T M$, $(H^T M)H$ vollständig durchzuführen. Die Ideen aus dem Buch von Parlett [46] für eine symmetrische Matrix A und Householdermatrizen H lassen sich direkt auf unsere Householderartigen Matrizen übertragen:

Mit

$$H_{house}(\mathbf{w}) = I - \frac{2}{\mathbf{w}^T J_M \mathbf{w}} \mathbf{w}\mathbf{w}^T J_M =: I - \gamma \mathbf{w}\mathbf{w}^T J_M \quad \text{mit } \gamma := \frac{2}{\mathbf{w}^T J_M \mathbf{w}}$$

erhalten wir

$$H_{house}\ M\ H_{house}^T = \left(I - \gamma \mathbf{w}\mathbf{w}^T J_M\right) M \left(I - \gamma J_M \mathbf{w}\mathbf{w}^T\right)$$

$$= M - \gamma \mathbf{w}\mathbf{w}^T J_M M - \gamma M J_M \mathbf{w}\mathbf{w}^T + \gamma^2 \mathbf{w}\mathbf{w}^T J_M M J_M \mathbf{w}\mathbf{w}^T.$$

Wir setzen

(5.6) $\quad \mathbf{p} := \gamma M J_M \mathbf{w}$.

Dann ist

$$H_{house}\, M\, H_{house}^T = M - \mathbf{w}\mathbf{p}^T - \mathbf{p}\mathbf{w}^T + \gamma \mathbf{w}\mathbf{w}^T J_M \mathbf{p}\mathbf{w}^T .$$

Da $\beta := \gamma \mathbf{w}^T J_M \mathbf{p} \in \mathbb{R}$ ist, kann man den Ausdruck $\gamma \mathbf{w}\mathbf{w}^T J_M \mathbf{p}\mathbf{w}^T$ auf die restlichen Summanden wie folgt verteilen:

Mit

(5.7) $\quad \mathbf{q} := \mathbf{p} - \mathbf{w}\,\frac{\beta}{2}$

erhalten wir

(5.8) $\quad H_{house}^T\, M\, H_{house} = M - \mathbf{w}\mathbf{q}^T - \mathbf{q}\mathbf{w}^T.$

Im Algorithmus wird dann zunächst $\mathbf{p}$, dann β, dann $\mathbf{q}$ und daraus $H_{house}\, M\, H_{house}^T$ gebildet.

Der Algorithmus für die teilweise Tridiagonalisierung des Paares (A,J) mit A vollbesetzte Matrix und $J = \mathrm{diag}(j_i)$ mit $j_i \in \{-1,+1\}$ lautet jetzt:

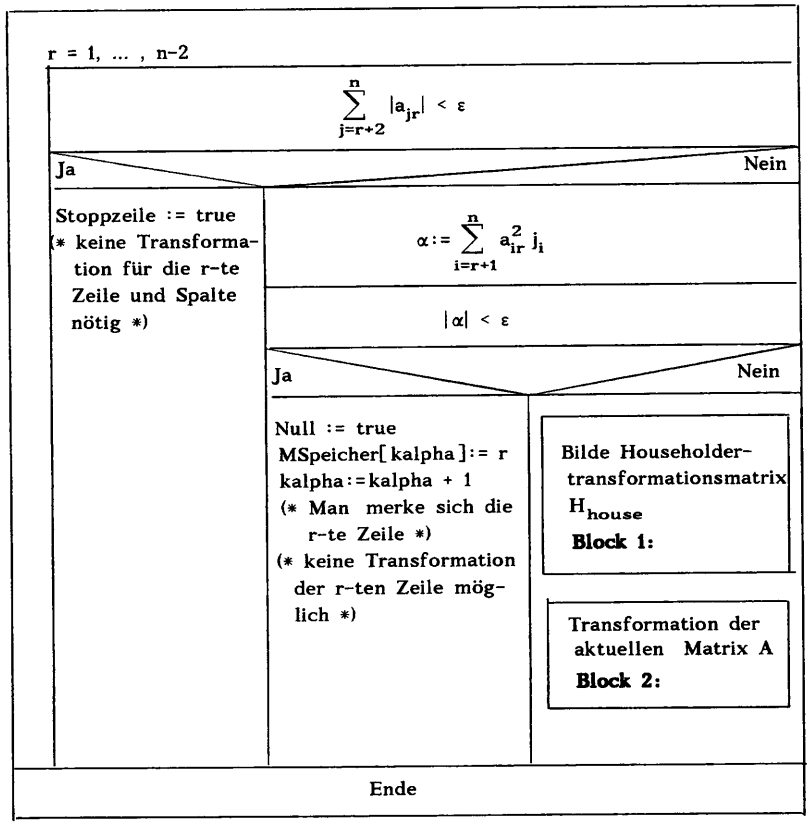
Algorithmus 5.1: Teilweise Tridiagonalisierung eines Paares (A,J)

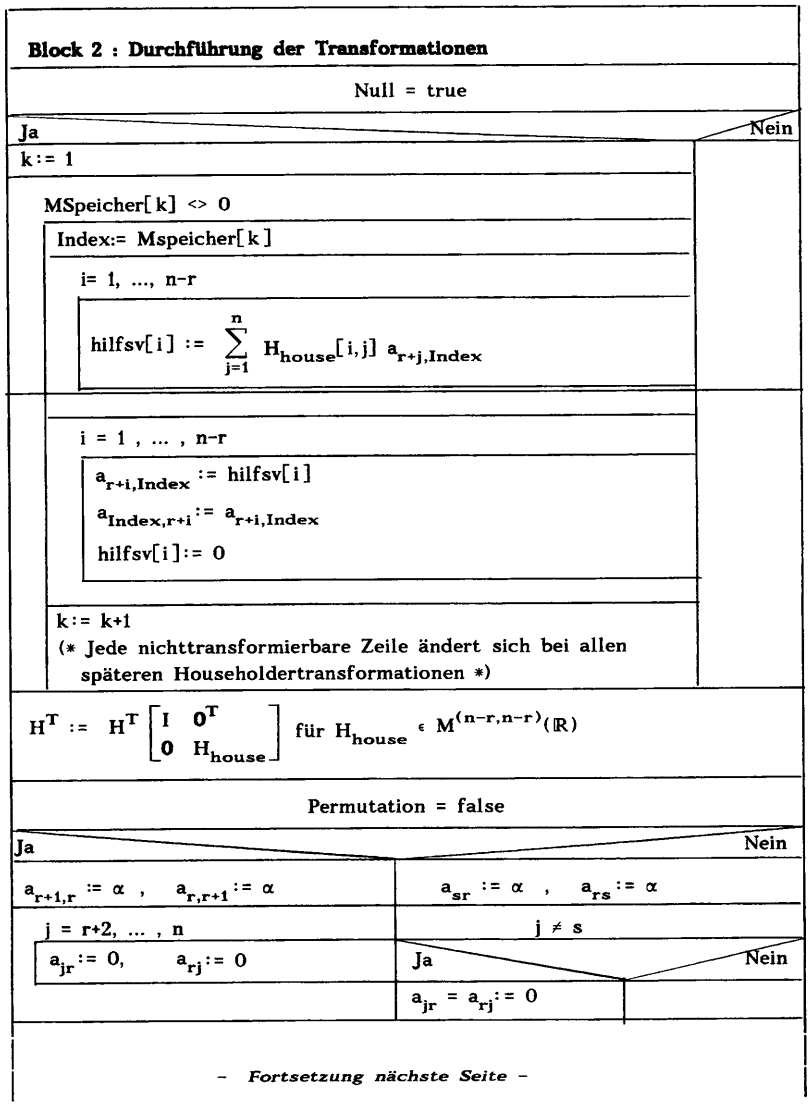
Bezeichnungen:

ε:	ε steht für eine kleine positive reelle Zahl (vgl. Bemerkung am Ende diese Abschnittes)
A:	vollbesetzte n×n- Matrix mit den Elementen a_{ij}
J:	J ist Diagonalmatrix mit +1 und -1 auf der Diagonalen
n:	n ist die Dimension der Matrizen
H^T:	Das ist die Transformationsmatrix
Stoppzeile:	Stoppzeile = true bedeutet, daß keine Transformation für die aktuelle Zeile und Spalte nötig ist
Null:	Null = true zeigt an, daß die Größe $\lvert\alpha\rvert < \varepsilon$ ist und damit α als null betrachtet wird.
MSpeicher:	Man merke sich die Zeilen und Spalten, wo wegen α = 0 keine Transformation stattgefunden hat. Diese Zeilen und Spalten müssen bei allen folgenden Householdertransformationen mittransformiert werden.
Permutation:	Permutation = true bedeutet, daß Zeilen- und Spaltenvertauschungen durchzuführen sind.

Vorbelegungen:

$H^T := I$
kalpha := 1
Null := false





- *Fortsetzung nächste Seite* -

$\mathbf{p} := \gamma\ (a_{ij})_{\substack{i=r+1,\dots,n \\ j=r+1,\dots,n}} \cdot \mathrm{diag}(j_{r+1}, \dots, j_n) \cdot \mathbf{w}$
$\beta := \gamma\ \mathbf{w}^T \cdot \mathrm{diag}(j_{r+1}, \dots, j_n) \cdot \mathbf{p}/2$
$\mathbf{q} := \mathbf{p} - \mathbf{w}\beta$
$(a_{ij})_{\substack{i=r+1,\dots,n \\ j=r+1,\dots,n}} := (a_{ij})_{\substack{i=r+1,\dots,n \\ j=r+1,\dots,n}} - \mathbf{w}^T\mathbf{q} - \mathbf{q}\mathbf{w}^T$
Permutation = true
Ja / Nein
Vertausche in den Matrizen A, H^T und in der Matrix J die r+1-te und die s-te Zeile und Spalte

Wenn wir den Algorithmus zur teilweisen Tridiagonalisierung auf das Paar (A,J) angewendet haben, ist J in ein J' = diag(± 1) übergegangen, wobei sich die Anzahl der Plus-und Minuseinsen nicht geändert hat, und A in T_{mod}, wobei für T_{mod} die folgenden Fälle auftreten können:

1. T_{mod} ist Tridiagonalmatrix
2. Eine Zeile und Spalte von T_{mod} weicht von der Tridiagonalgestalt ab
3. Mehrere Zeilen und Spalten von T_{mod} sind vollbesetzt

In den meisten Beispielen wird Fall 1 auftreten. Wie oft Fall 2 oder gar Fall 3 möglich sind, hängt auch davon ab, ab wann der Computer den Ausdruck $\alpha := \mathbf{a}^T J_M \mathbf{a}$ als null ansieht. Wie schon erwähnt, können bei kleinem $|\mathbf{a}^T J_M \mathbf{a}|$ starke Rundungsfehler auftreten. Um diese Fehler zu vermeiden, könnte man auf eine householderartige Transformation der Zeilen und Spalten mit kleinem $|\mathbf{a}^T J_M \mathbf{a}|$ verzichten.
Im folgenden Abschnitt werden wir dazu statistische Untersuchungen durchführen, die uns verraten, wie wahrscheinlich ein kleines $|\mathbf{a}^T J_M \mathbf{a}|$ ist. Dabei werden wir feststellen, daß betragsmäßig kleine α insbesondere für Matrizenpaare auftreten, bei denen die Matrix J gleich viele +1 und -1 auf der Diagonalen enthält. Bei unseren Testbeispielen werden wir α als null ansehen, wenn es kleiner als ein vorgegebens ε ist. Dabei werden wir ε von recht verschiedener Größe wählen, etwa ε = Maschinengenauigkeit oder sogar $\varepsilon = 0.6$ (vgl. die Ergebnisse im Anhang).
Hat man das Eigenwertproblem $T_{mod}\mathbf{y} = \lambda J'\mathbf{y}$ lösen können, so ergeben sich die Eigenvektoren von $A\mathbf{x} = \lambda J\mathbf{x}$ aus

(5.9) $$\mathbf{x} = J H^T J' \mathbf{y}$$

mit J Startmatrix, J' während des Algorithmus modifiziertes J.

5.2 Statistische Untersuchungen zur Reduktion von Matrizen auf Tridiagonalform mittels verallgemeinerter Householder-Transformationen

Wir untersuchen hier, wie häufig eine verallgemeinerte Householdertransformation während des Tridiagonalisierungsprozesses unmöglich ist oder zu ungenaue Ergebnisse liefert. Mithilfe von wahrscheinlichkeitstheoretischen Methoden stellen wir zunächst fest, daß diese kritischen Fälle auftreten können, wenn +1 und -1 sich auf der Diagonalen von J ungefähr abwechseln; andernfalls sind sie recht unwahrscheinlich. Um diese aus einem (vereinfachten) Modell erhaltenen theoretischen Ergebnisse zu untermauern, führen wir zusätzlich Monte-Carlo-Simulationen durch.

Bei der Überführung einer symmetrischen Matrix A in Tridiagonalgestalt durch Anwendung von (verallgemeinerten) Householder-Transformationen - bei gleichzeitiger Erhaltung der Matrix J - spielen die Größen

(5.10) $$z_i := \sum_{j=i+1}^{n} \left(a_{ij}^{(i-1)}\right)^2 J_{jj}\ ,\ i = 1,\dots,n-2$$

$\left((a_{ij}^{(k)})_{ij}\right.$ Matrix nach Ausführung der k-ten Householder-Reduktion$\left.\right)$

eine entscheidende Rolle. Je kleiner die Größen $|z_1|, |z_2|, \dots$, sind, umso grösser werden im allgemeinen die Fehler sein, die durch die Verwendung der transformierten Matrix entstehen.
Viele Beispiele zeigten, daß der Wert

$$z := \min_i |z_i|$$

öfter sehr viel kleiner als die Größenordnung der Matrix-Koeffizienten ist, so daß eine vollständige Reduktion mit Householder-Transformationen nicht zu empfehlen ist. In diesem Abschnitt soll mit Hilfsmitteln der Statistik abgeschätzt werden, wie "wahrscheinlich" derartig kleine Werte von z sind. Dazu machen wir das folgende Modell:
Die Koeffizienten a_{ij}, i, j = 1,...,n der Matrix A sollen zufällige unabhängige Größen sein, die alle einer bestimmten Verteilung mit Erwartungswert μ, Streuung σ genügen. (Dabei muß natürlich $a_{ij} = a_{ji}$ gelten !)
Der Wert z_1 ergibt sich nun als Realisierung der *Zufallsvariablen*

(5.11) $$Z_1 := \sum_{j=2}^{n} A_{1j}^2 J_{jj}\ ,$$

wobei die Zufallsvariablen $A_{12}, \dots, A_{1n}$ alle den gleichen Erwartungswert μ und die gleiche Streuung σ haben. (A_{1j} liefert den zufälligen Wert von a_{1j} !)
In ähnlicher Weise sind auch

(5.12) $$Z_2 := \sum_{j=3}^{n} A_{2j}^2 J_{jj}\ ,\ \dots\ ,\ Z_{n-2} := \sum_{j=n-2}^{n} A_{n-2,j}^2 J_{jj}$$

Zufallsvariablen. Allerdings ist wegen der bereits durchgeführten Householdertransformationen nicht mehr explizit bekannt, welcher Verteilung die Zufallsvariablen $A_{ij}^{(i-1)}$, $j = 3,\dots,n$ für $i \geq 2$ genügen. (Die entsprechenden Koeffizienten $a_{ij}^{(i-1)}$ gehen ja in ziemlich komplizierter Weise aus den Koeffizienten $a_{ij}^{(0)} \equiv a_{ij}$ hervor!)

Als vereinfachtes Modell nehmen wir jetzt die Verteilung der Zufallsvariablen

$$(5.13) \qquad Z := \sum_{j=1}^{p} X_j^2 - \sum_{j=p+1}^{p+q} X_j^2 ,$$

wobei $X_1, \dots, X_{p+q}$ unabhängige Realisierungen einer kontinuierlichen Zufallsvariablen X (mit Dichtefunktion f_X) sein sollen. (Dementsprechend haben die X_i die gleiche Verteilung wie X und natürlich auch gleiche Erwartungswerte und Streuungen.)

Setzen wir $Y_j := X_j^2$ für $j \leq p$ und $Y_j := -X_j^2$ für $j > p$, so ist Z als Summe der unabhängigen Zufallsvariablen $Y_1, \dots, Y_{p+q}$ gegeben:

$$Z := \sum_{j=1}^{p+q} Y_j$$

Dabei haben $Y_1, \dots, Y_p$ die gleiche Dichtefunktion f_Y, die Dichtefunktion von $Y_{p+1}, \dots, Y_{p+q}$ geht durch Spiegelung aus der von f_Y hervor:

$$f_{Y_j} = f_Y^{\vee} \quad (\text{d. h. } f_{Y_j}^{\vee}(y) = f_Y(-y)) \text{ für } j > p$$

Im Prinzip läßt sich die Dichtefunktion f_Z von Z aus f_Y wegen

$$f_{Y_1+Y_2}(z) = (f_{Y_1} * f_{Y_2})(z) \equiv \int_{-\infty}^{\infty} f_{Y_1}(z-y)\, f_{Y_2}(y)dy$$

durch mehrfache Faltung berechnen:

$$f_Z = f_{Y_1} * f_{Y_2} * \dots * f_{Y_{p+q}} = \underbrace{(f_Y * f_Y * \dots * f_Y)}_{\text{p-mal}} * \underbrace{(f_Y^{\vee} * \dots * f_Y^{\vee})}_{\text{q-mal}}$$

$$= (f_Y * f_Y * \dots * f_Y) * (f_Y * \dots * f_Y)^{\vee}$$

Tatsächlich durchführen lassen sich diese Faltungen jedoch nur in wenigen Fällen. (Etwa, wenn $X_1, \dots, X_p$ normalverteilt mit Erwartungswert 0 und Streuung 1 sind, so genügt

$$X_1^2 + \dots + X_n^2$$

der χ_n^2 -Verteilung.)

Wenn p+q nicht zu klein ist (also $p+q \gtrsim 10$), so kann die Verteilung von Z jedoch angenähert durch den zentralen Grenzwertsatz bestimmt werden:

Z ist nämlich asymptotisch $N(\mu, \sigma^2)$ - verteilt mit

$$\mu = E(Y_1 + \dots + Y_{p+q}) = E(Y_1) + \dots + E(Y_{p+q}) = p\, E(Y_1) + q\, E(Y_{p+1})$$

$$= p\, E(X^2) + q\, E(-X^2)$$

$$= (p - q)\, E(X^2)$$

und

$$\sigma^2 = Var(Y_1 + \dots + Y_{p+q}) = Var(Y_1) + \dots + Var(Y_{p+q})$$

$$= p\, Var(X^2) + q\, Var(-X^2) = p\, Var(X^2) + q\, Var(X^2)$$

$$= (p + q)\, Var(X^2)$$

$$= (p + q)\left[E(X^4) - (E(X^2))^2\right].$$

Die Verteilung von Z ist also asymptotisch allein durch das zweite Moment μ_2 und das vierte Moment μ_4 von X gegeben:

$$\mu = (p - q)\, \mu_2$$

$$\sigma^2 = (p + q)\left[\mu_4 - (\mu_2)^2\right],$$

wobei

$$\mu_k \equiv E(X^k) \equiv \int_{-\infty}^{\infty} x^k f_X(x)\, dx.$$

Die Wahrscheinlichkeit, daß Z betragsmäßig kleiner als eine Schranke α ist, läßt sich jetzt näherungsweise leicht berechnen:

$$(5.14) \qquad P(-\alpha < Z < \alpha) = P\Big(\frac{-\alpha-\mu}{\sigma} < \underbrace{\frac{Z-\mu}{\sigma}}_{:= Z^*,\ N(0,1)\text{-verteilt}} < \frac{\alpha-\mu}{\sigma}\Big)$$

$$= \Phi\Big(\frac{\alpha-\mu}{\sigma}\Big) - \Phi\Big(\frac{-\alpha-\mu}{\sigma}\Big)$$

Hierbei bedeutet Φ die Gauß'sche Fehlerfunktion entsprechend der Definition

$$(5.15) \qquad \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt .$$

Die Wahrscheinlichkeit wird am größten, wenn $\mu = 0$ ist, d.h. gleich viele "+" und "-" -Vorzeichen in der Summe auftreten.

In diesem Fall ist Z symmetrisch um 0 verteilt, und es gilt:

$$(5.16)\qquad P(-\alpha < Z < \alpha) = \Phi\left(\frac{\alpha}{\sigma}\right) - \Phi\left(-\frac{\alpha}{\sigma}\right) = 2\,\Phi\left(\frac{\alpha}{\sigma}\right) - 1$$

Zum Beispiel erhalten wir aus der Normalverteilungstabelle im Fall, daß X^2 die Streuung 1 hat, für $\alpha := 0.1$ und $p = q = 10$ den folgenden Wert:

$$2\,\Phi\left(\frac{\alpha}{\sigma}\right) - 1 = 2\,\Phi\left(\frac{\alpha}{\sqrt{p+q}}\right) - 1 = 2\,\Phi\left(\frac{0.1}{\sqrt{20}}\right) - 1 \approx 2\cdot 0.509 - 1 = 0.018$$

Da uns hier nur sehr kleine Argumente $\frac{\alpha}{\sigma}$ interessieren, können wir auf die Normalverteilungstabelle verzichten und stattdessen den Mittelwertsatz anwenden:

$$\Phi\left(\frac{\alpha}{\sigma}\right) - \Phi\left(-\frac{\alpha}{\sigma}\right) \approx \Phi'(0)\left(\frac{\alpha}{\sigma} - \left(-\frac{\alpha}{\sigma}\right)\right) = \frac{1}{\sqrt{2\pi}}\, e^{0}\, \frac{2\alpha}{\sigma}$$

$$= \frac{1}{\sqrt{2\pi}}\,\frac{2\alpha}{\sqrt{(p+q)\,R}} \approx 0.8\,\frac{\alpha}{\sqrt{(p+q)\,R}}$$

mit

$$R = E(X^4) - \left(E(X^2)\right)^2$$

(Für unser vorhergehendes Beispiel erhalten wir mit dieser Näherung: $0.8\,\frac{0.1}{\sqrt{20}} \approx 0.018$)

Wesentlich unwahrscheinlicher ist ein Z-Wert in der Nähe von null, wenn p ungleich q ist, da die Dichtefunktion von Z ihr Maximum dann an der Stelle

$$(p - q)\,E(X^2)$$

hat. Nach dem Mittelwertsatz erhält man

$$(5.17)\qquad \begin{aligned} P(-\alpha < Z < \alpha) &\approx \Phi\left(-\frac{\mu}{\sigma} + \frac{\alpha}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma} - \frac{\alpha}{\sigma}\right) \\ &\approx \Phi'\left(-\frac{\mu}{\sigma}\right)\cdot 2\,\frac{\alpha}{\sigma} = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\mu^2}{2\sigma^2}}\cdot\frac{2\alpha}{\sigma}\,. \end{aligned}$$

Nur in dem Fall, daß $\mu^2 \ll \sigma^2$, also

$$\frac{(p-q)^2}{p+q} \ll \frac{E(X^4)}{(E(X^2))^2} - 1$$

gilt, liefert der Ausdruck $\exp(-\mu^2/(2\sigma^2))$ einen wesentlich von null verschiedenen Vorfaktor. Bei den wichtigsten Verteilungen - z.B. bei Normalverteilung und Gleichverteilung - ist $E(X^4)$ von der gleichen Größenordnung wie $(E(X^2))^2$. Wir erhalten dann wieder $p \approx q$ als Bedingung, das heißt, daß in der Summe die Anzahl der '+' prozentual gleich (oder fast gleich) der Anzahl der '-' sein muß.

In Annäherung an die (analytisch zu komplizierten) Verhältnisse bei der Durchführung mehrerer Householder-Transformationen diskutieren wir jetzt die Verteilung der Zufallsvariablen Z'(A), die durch

$$(5.18)\qquad Z'(A) := \min_{i=1,\dots,n-2} |\,Z_i\,|$$

mit

$$Z_i := \sum_{j=i+1}^{n} A_{ij}^2\, J_{jj}\quad,\ i = 1, \dots, n-2$$

definiert ist. Da $Z_1, \dots, Z_{n-2}$ - und damit auch $|Z_1|, \dots, |Z_{n-2}|$ - unabhängig sind, gilt

$$P(Z' \geq \alpha) = P(|Z_1| \geq \alpha)\,P(|Z_2| \geq \alpha)\,\dots\,P(|Z_{n-2}| \geq \alpha)$$

und damit durch Übergang zu den Gegenwahrscheinlichkeiten:

$$(5.19)\qquad \begin{aligned} P(Z' < \alpha) &= 1 - P(Z' \geq \alpha) \\ &= 1 - P(|Z_1| \geq \alpha)\,P(|Z_2| \geq \alpha)\,\dots\,P(|Z_{n-2}| \geq \alpha) \\ &= 1 - (1 - P(|Z_1| < \alpha))\,\dots\,(1 - P(|Z_{n-2}| < \alpha)) \\ &= P(|Z_1| < \alpha) + \dots + P(|Z_{n-2}| < \alpha) \\ &\quad - P(|Z_1| < \alpha)P(|Z_2| < \alpha) - \dots - P(|Z_{n-3}| < \alpha)P(|Z_{n-2}| < \alpha) \\ &\quad + \dots + (-1)^{n-2}\,P(|Z_1| < \alpha)\dots P(|Z_{n-2}| < \alpha) \end{aligned}$$

In dem letzten Ausdruck auf der rechten Seite dürfen wir quadratische oder höhere Terme vernachlässigen, weil die Wahrscheinlichkeiten $P(|Z_1| < \alpha)$, ,,, , $P(|Z_{n-2}| < \alpha)$ für α nahe bei null alle recht klein sind. Unter Verwendung von (5.17) erhalten wir dann näherungsweise:

$$(5.20)\qquad \begin{aligned} P(Z' < \alpha) &\approx P(|Z_1| < \alpha) + \dots + P(|Z_{n-2}| < \alpha) \\ &\approx \frac{2\alpha}{\sqrt{2\pi}} \sum_{i=1}^{n-2} e^{-\mu_i^2/2\sigma_i^2}\,\frac{1}{\sigma_i} \end{aligned}$$

mit

$$\mu_i = E(Z_i),\ \sigma_i = \sqrt{Var(Z_i)}\,.$$

(Es soll nicht verschwiegen werden, daß die Normalverteilungsannahme nur für diejenigen Z_i ganz gerechtfertigt ist, die Summe *vieler* unabhängiger Zufallsvariablen sind - dies ist für die letzten Z_i's natürlich nicht der Fall. Da es hier aber nur um eine grobe Abschätzung geht, ist der zusätzliche Fehler nicht besonders bedeutsam.)

Wegen

(5.21) $\mu_i = (n_i(+) - n_i(-)) \cdot E(X^2)$,

wobei $n_i(+)$, $n_i(-)$ die Anzahl der "+" bzw. "-"-Vorzeichen in der Summe von Z_i ist, werden betragsmäßig kleine Werte von Z' besonders dann häufig auftreten, wenn die Vorzeichenfolge in J zumindest gegen das Ende hin alternierend ist:

..... + - + - + - +

(Wechselnde Vorzeichen über die gesamte Diagonale sind aber nicht unbedingt notwendig, entscheidend ist nur, daß ab einem gewissen Index insgesamt gleich viel "+" und "-" Vorzeichen auftreten.)
Besteht dagegen J aus zwei unterschiedlich großen Blöcken von "+" und "-", so sind kleine Z'- Werte äußerst unwahrscheinlich.
Schreibt man den ersten Teil der Formel (5.19) mit Hilfe der jeweiligen Verteilungsfunktion um, wobei man

(5.22) $P(|Z_i| < z) = P(Z_i < z) - P(Z_i < -z) = F_{Z_i}(z) - F_{Z_i}(-z)$

für $z > 0$ benutzt, so erhält man

$$F_{Z'}(z) = 1 - \left(1 - F_{Z_1}(z) + F_{Z_1}(-z)\right) \dots \left(1 - F_{Z_{n-2}}(z) + F_{Z_{n-2}}(-z)\right)$$

für $z > 0$ und null sonst. Durch Differenzieren nach z läßt sich nun die Dichtefunktion von Z' gewinnen:

$$\begin{aligned} f_{Z'}(z) &= \left(f_{Z_1}(z) + f_{Z_1}(-z)\right)\left(1 - F_{Z_2}(z) + F_{Z_2}(-z)\right)\dots\left(1 - F_{Z_{n-2}}(z) + F_{Z_{n-2}}(-z)\right) + \dots \\ &\quad \dots + \left(1 - F_{Z_1}(z) + F_{Z_1}(-z)\right)\dots\left(1 - F_{Z_{n-2}}(z) + F_{Z_{n-2}}(-z)\right)\left(f_{Z_{n-2}}(z) + f_{Z_{n-2}}(-z)\right) \\ &= \left(\prod_{i=1}^{n-2}\left(1 - F_{Z_i}(z) + F_{Z_i}(-z)\right)\right)\sum_{i=1}^{n-2}\frac{f_{Z_i}(z) + f_{Z_i}(-z)}{1 - F_{Z_i}(z) + F_{Z_i}(-z)} \end{aligned} \tag{5.23}$$

In Wirklichkeit müssen wir natürlich die Householder-Transformationen durchführen und die Größen a_{2j}, a_{3j}, ... durch die mittels der Transformation aus den ursprünglichen Größen hervorgehenden Koeffizienten $a_{2j}^{(1)}$, $a_{3j}^{(2)}$, ... ersetzen, d.h. statt Z'(A) die Zufallsvariable

$$Z(A) := \min_{i=1,\dots,n-2} \left| \sum_{j=i+1}^{n} \left(A_{ij}^{(j)}\right)^2 J_{jj} \right|$$

betrachten.
Da die Zufallsvariablen $A_{ij}^{(k)}$ (i, j = 2, ... , n-2) aber aus den Zufallsvariablen $A_{ij}^{(k-1)}$ (i, j = 1, ... , n) entstehen, sind sie im allgemeinen nicht mehr ganz unabhängig, und diese Voraussetzung des zentralen Grenzwertsatzes ist nicht länger erfüllt.
Zur Untersuchung von Z wurde deshalb eine Monte-Carlo-Simulation durchgeführt: Die Koeffizienten der Ausgangsmatrix wurden durch Transformation

aus in [0, 1] gleichverteilten Zufallszahlen so bestimmt, daß sie einer der drei folgenden Verteilungen genügen:

(5.24a) *Gleichverteilung* in $[-\sqrt{3}, \sqrt{3}]$: $f_X(x) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{für } |x| < \sqrt{3} \\ 0 & \text{sonst} \end{cases}$

(5.24b) *Standardnormalverteilung*: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

(5.24c) (transformierte) *Studentverteilung* mit Freiheitsgrad 3:

$$f_X(x) = \frac{2}{\pi} \frac{1}{(1 + x^2)^2}$$

In (5.24a) und (5.24c) wurde entsprechend der Formel

$$X^* := \frac{X - E(X)}{\sigma_X}$$

so transformiert, daß wie in (5.24b) neben $E(X) = 0$ auch $\sigma = 1$ gilt ("Standardisierung"). Aus der allgemeinen Transformationsformel

(5.25) $f_{g(X)}(x) = f_X(g^{-1}(x)) \frac{1}{g'(g^{-1}(x))}$

folgt für $g(x) = x^2$ im Fall einer symmetrischen Verteilung:

(5.26) $f_{X^2}(x) = \begin{cases} \frac{1}{2\sqrt{x}} [f_X(\sqrt{x}) + f_X(-\sqrt{x})] = \frac{1}{\sqrt{x}} f_X(\sqrt{x}) & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$

Dies liefert bei unseren drei Verteilungen:

(5.27a) $f_{X^2}(x) = \begin{cases} \frac{1}{\sqrt{x}} \frac{1}{2\sqrt{3}} & \text{für } 0 < x < 3 \\ 0 & \text{sonst} \end{cases}$

(5.27b) $f_{X^2}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$

(5.27c) $f_{X^2}(x) = \begin{cases} \frac{2}{\pi} \frac{1}{\sqrt{x}} \frac{1}{(1 + x)^2} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$

Als Erwartungswert für $Y := X^2$ erhalten wir in allen drei Fällen wegen $E(X) = 0$ und $E(Y) = E(X^2) = E(X^2) - (E(X))^2 = Var(X) = \sigma^2 = 1$ den Wert 1.
Für X^2 existiert die Streuung nur bei den ersten beiden Verteilungen. Mit der allgemeinen Formel

$$\mathrm{Var}(Y) \equiv E(Y^2) - (E(Y))^2 = E(X^4) - (E(X))^2 = \int_{-\infty}^{\infty} x^4 f_X(x)\,dx - 1$$

bekommt man im ersten Testfall

$$\mathrm{Var}(Y) = \int_{-\sqrt{3}}^{\sqrt{3}} x^4 \frac{1}{2\sqrt{3}}\,dx - 1 = \frac{1}{2\sqrt{3}}\, 2\,\frac{(\sqrt{3})^5}{5} - 1 = \frac{9}{5} - 1 = \frac{4}{5}$$

also $\sigma_{X^2} = \frac{2}{\sqrt{5}} \approx 0.8944$

und im zweiten Testfall

$$\mathrm{Var}(Y) = \int_{-\infty}^{\infty} x^4 \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx - 1$$

$$= 2\int_0^{\infty} (2y)^2 \frac{1}{\sqrt{2\pi}} e^{-y} \frac{\sqrt{2}}{2\sqrt{y}}\,dy - 1 \qquad \text{mit } x^2 = 2y$$

$$= \frac{4}{\sqrt{\pi}} \int_0^{\infty} y^{2-\frac{1}{2}} e^{-y}\,dy - 1$$

$$= \frac{4}{\sqrt{\pi}}\,\Gamma(\tfrac{5}{2}) - 1 = \frac{4}{\sqrt{\pi}}\,\frac{3}{2}\,\frac{1}{2}\,\Gamma(\tfrac{1}{2}) - 1 = \frac{3}{\sqrt{\pi}}\sqrt{\pi} - 1 = 2,$$

also $\sigma_{X^2} = \sqrt{2} \approx 1.41.$

Bei der Normalverteilung haben wir deshalb mit einer etwa um die Hälfte größeren Streuung zu rechnen.
Der Grund, weshalb wir als dritte Verteilung die Student-Verteilung mit Freiheitsgrad 3 genommen haben, liegt darin, daß zwar für X die Streuung existiert, nicht aber für X^2. Deshalb ist der zentrale Grenzwertsatz überhaupt nicht anwendbar. Die Computer-Ergebnisse zeigen jedoch, daß außer einer Verbreiterung der Verteilung kein Unterschied zu den Fällen (1) und (2) zu erkennen ist.

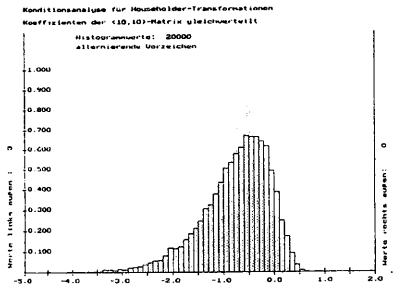
Da Z' und Z für Dimension $n \le 70$ überwiegend in einem Bereich von 10^{-4} bis 10^{1} variieren, wurde die Verteilung der dekadisch logarithmierten Größen untersucht, d.h. die Größen $\lg Z'$ und $\lg Z$. Bei dieser Transformation verändert sich eine Dichtefunktion gemäß der Formel

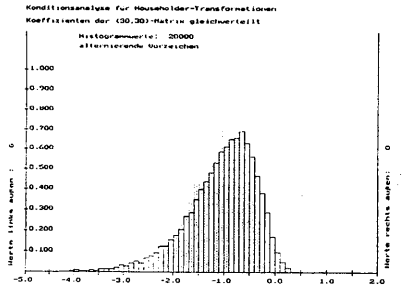
$$(5.28) \qquad f_{\lg X}(x) = f_X\left(e^{x \ln 10}\right) e^{x \ln 10} \ln 10 = f_X(10^x)\, 10^x \ln 10.$$

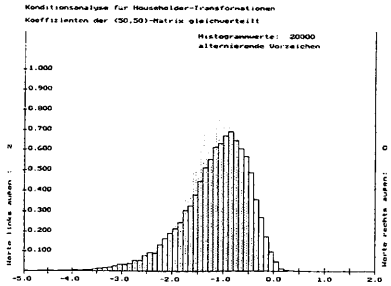
Für das Histogramm von Z' wurden in den Testbeispielen punktierte Flächen, für das von Z ausgezogene Rechtecke verwendet. Damit die Gesamtfläche unter der Histogrammkurve die Fläche 1 entsprechend der Normierungsbedingung hat, geben nicht die Höhen, sondern die *Flächen* die relativen Häufigkeiten an. (Um die relative Häufigkeit zu einem Intervall zu erhalten, ist also die Höhe noch mit der Länge des Intervalls zu multiplizieren!)

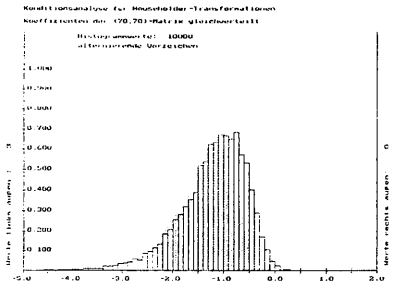
Wie von unseren theoretischen Überlegungen her zu erwarten war, kommen bei nicht gleichmäßig verteilten "+" und "-"- Vorzeichen problematisch kleine Z bzw. Z'- Werte ziemlich selten vor. Anders ist es im Fall alternierender Vorzeichen, wo auch Werte kleiner als 10^{-4} auftreten. Es zeigt sich bei allen untersuchten Dimensionen, daß die Verteilung von Z nur ein wenig gegenüber der Verteilung von Z' verschoben ist. Unser grobes Modell beschreibt also das tatsächliche Verhalten schon recht gut. Statistisch gesehen kann man sagen, daß die Householdertransformationen oft zu etwas kleineren z-Werten führen. Insgesamt ist direkt abzulesen, daß ein beträchtlicher Prozentsatz der z-Werte (mehr als 20 %) kleiner als 0.1 ist. Etwa 5% der Werte ist sogar kleiner als 0.01.

Eine auf 486-PCs vorgenommene Monte-Carlo-Simulation mit ca. 10000 bis 30000 Versuchen je Beispiel führte zu folgenden Ergebnissen:

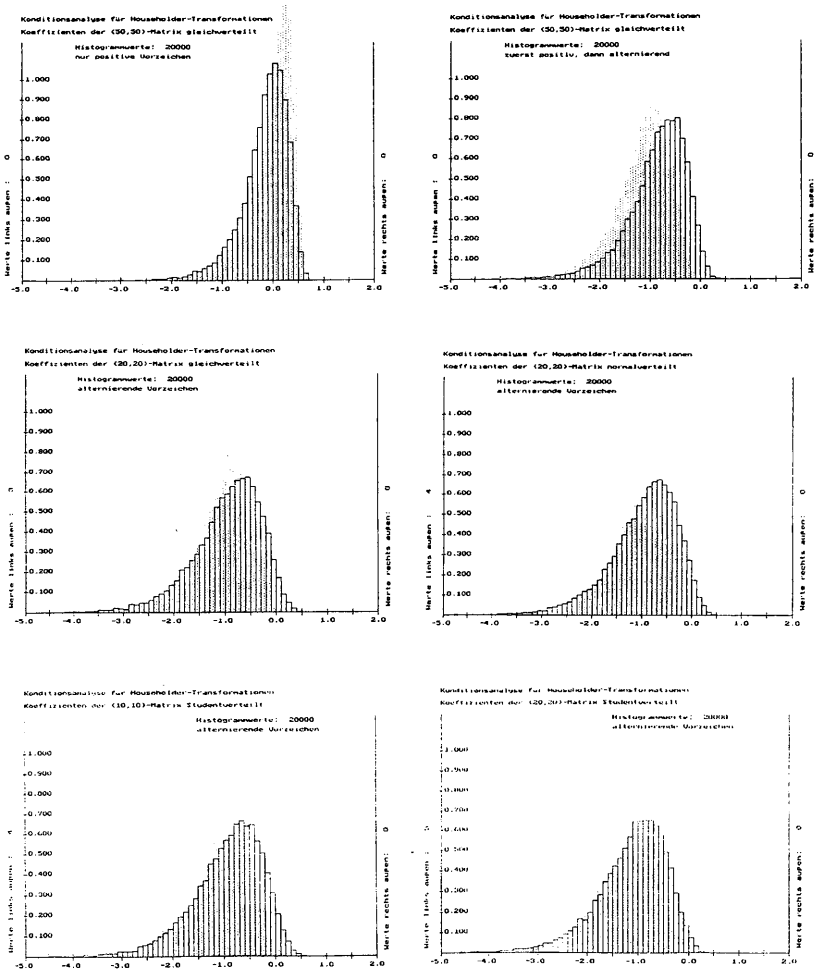








Weder lg Z, lg Z' noch Z, Z' genügen einer ungefähren Normalverteilung. Im Gegenteil, für große n tritt die "Schiefheit" (Asymmetrie) der Verteilung immer deutlicher hervor.
Im Hinblick auf die drei verwendeten Verteilungen für die Koeffizienten der Matrix A ergibt sich, daß die entsprechenden Verteilungen von lg Z nur in der Breite variieren. Dabei liefert die Normalverteilung eine geringfügig breitere Verteilung als die Gleichverteilung, die Studentverteilung eine deutlich breitere. Diese Ergebnisse sind qualitativ im Einklang mit unseren theoretischen Überlegungen.



6. Divide-et-Impera-Techniken für das verallgemeinerte Eigenwertproblem $Ax = \lambda J x$

Wir entwickeln hier einen Algorithmus, der auf Divide-et-Impera-Techniken aufbaut, um ein Eigenwertproblem $T'\mathbf{x} = \lambda J \mathbf{x}$ zu lösen, bei dem die Matrix T' bis auf einige vollbesetzte Zeilen und Spalten eine Tridiagonalmatrix ist und $J = \mathrm{diag}(\pm 1)$. Er besteht aus einer Kombination des verallgemeinerten Divide-et-Impera-Verfahrens aus Kapitel 4 mit einem Algorithmus zur Behandlung verallgemeinerter Eigenwertprobleme für berandete Diagonalmatrizen. Wenn das Paar (T', J) positiv definit ist, berechnet diese neue Methode alle Eigenwerte und Eigenvektoren, andernfalls entdeckt sie das Verletztsein dieser Bedingung.

Wir stellen uns jetzt vor, daß wir mit Hilfe des Tridiagonalisierungsalgorithmus aus Kapitel 5 das ursprüngliche Eigenwertproblem auf die Gestalt $A\mathbf{x} = \lambda J \mathbf{x}$ transformiert haben, wobei A Tridiagonalmatrix oder Ineinanderschachtelung von berandeten Tridiagonalmatrizen ist und die Gestalt von J bis auf Vertauschungen der Reihenfolge der Diagonalelemente erhalten geblieben ist.
In Kapitel 4 hatten wir das Divide-et-Impera-Verfahren für ein verallgemeinertes Eigenwertproblem mit A Tridiagonalmatrix entwickelt. Wir müssen jetzt noch einen Algorithmus für den Fall finden, daß A nur teilweise Tridiagonalgestalt hat.
Ist etwa T_{mod} aus einer Tridiagonalmatrix und einer vollbesetzten ersten Zeile und Spalte aufgebaut, so schreiben wir T_{mod} in der folgenden Form:

$$T_{mod} = \left[\begin{array}{c|cccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \hline a_{21} & a_{22} & a_{23} & & \\ a_{31} & a_{32} & a_{33} & \ddots & \\ \vdots & & \ddots & \ddots & a_{n-1,n} \\ a_{n1} & & & a_{n-1,n} & a_{n,n} \end{array}\right] = \left[\begin{array}{c|c} a_{11} & \mathbf{a}^T \\ \hline \mathbf{a} & \mathrm{Tridiag}(A) \end{array}\right]$$

Wenn wir das Eigenwertproblem

$$\mathrm{Tridiag}(A)\ \mathbf{y} = \lambda\ \mathrm{diag}(j_2, \ldots, j_n)\mathbf{y}$$

mit Hilfe des Divide-et-Impera-Algorithmus lösen können und dabei G die zugehörige Eigenvektormatrix zu den Eigenwerten $\gamma_2, \ldots, \gamma_n$ ist, dann erhalten wir mit

$$\left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & G^T \end{array}\right] \left[\begin{array}{c|c} a_{11} & \mathbf{a}^T \\ \hline \mathbf{a} & \mathrm{Tridiag}(A) \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & G \end{array}\right] = \left[\begin{array}{c|c} a_{11} & \mathbf{a}^T G \\ \hline G^T\mathbf{a} & G^T\ \mathrm{Tridiag}(A)\, G \end{array}\right]$$

$$= \left[\begin{array}{c|c} a_{11} & \mathbf{a}^T G \\ \hline G^T\mathbf{a} & \mathrm{diag}(\gamma_2,\dots,\gamma_n) \end{array}\right]$$

eine berandete Diagonalmatrix. Daher lösen wir in Abschnitt 6.1 zunächst dieses Problem, um dann im Abschnitt 6.2 den gesamten Algorithmus vorzustellen.

6.1 Der Fall einer berandeten Diagonalmatrix

Wir verallgemeinern den Algorithmus von Wilkinson zur Lösung des Eigenwertproblems $C\mathbf{x} = \lambda\mathbf{x}$ für eine "berandete" Diagonalmatrix auf den Fall $C\mathbf{x} = \lambda J\mathbf{x}$, $J = \mathrm{diag}(\pm 1)$. Auch in diesem Fall sind die gesuchten Eigenwerte Nullstellen einer speziellen rationalen Funktion. Um diese zu bestimmen, modifizieren wir das in Kapitel 4 vorgestellte Einschließungsverfahren.

Es sei C eine symmetrische berandete Diagonalmatrix der folgenden Gestalt:

$$C = \left[\begin{array}{c|cccc} \gamma_0 & c_1 & c_2 & \dots & c_{n-1} \\ \hline c_1 & \gamma_1 & 0 & \dots & 0 \\ c_2 & 0 & \gamma_2 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ c_{n-1} & 0 & 0 & \dots & \gamma_{n-1} \end{array}\right] \quad \text{und} \quad J = \mathrm{diag}(\delta_0,\dots,\delta_{n-1}) \text{ mit } \delta_0, \delta_i \in \{-1,1\},\ i = 1, \dots, n-1$$

Wir wollen das Paar (C,J) auf positive Definitheit untersuchen und gegebenenfalls alle Eigenwerte und Eigenvektoren des Eigenwertproblems

(6.1) $\qquad C\mathbf{x} = \lambda J\mathbf{x}$

berechnen.

Für J = I ist das Paar (C,J) immer positiv definit, und ein Algorithmus zur Lösung des Eigenwertproblems $C\mathbf{x} = \lambda\mathbf{x}$ findet sich im Buch von Wilkinson [68]. Die Eigenwerte werden nach vorausgegangener Deflation als Nullstellen einer speziellen rationalen Funktion berechnet, das Vorgehen ähnelt also dem Lösen des Eigenwertproblems $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{x} = \lambda\mathbf{x}$ bzw. $(D - \rho\mathbf{z}\mathbf{z}^T)\mathbf{x} = \lambda J\mathbf{x}$. Wir schauen uns diesen Algorithmus gleich für beliebiges J mit $J = \mathrm{diag}(\pm 1)$ an. Dabei werden wir im folgenden voraussetzen, daß das Paar (Γ,Δ) mit $\Gamma = \mathrm{diag}(\gamma_1,\dots,\gamma_{n-1})$ und $\Delta = (\delta_1,\dots,\delta_{n-1})$ mit $\delta_i \in \{+1,-1\}$ positiv definit ist. Ist nämlich das Paar (Γ,Δ) nicht positiv definit, so kann nach dem Hurwitz-Kriterium auch das Paar (C,J) nicht positiv definit sein.

6.1.1 Der Deflationsschritt

Wir führen eine Deflation durch, wenn eine der beiden folgenden Bedingungen erfüllt ist :

a) Mindestens ein c_i, $i = 1,\dots,n-1$ ist gleich null

b) Das Paar (Γ,Δ) mit $\Gamma = \mathrm{diag}(\gamma_1,\dots,\gamma_{n-1})$ und $\Delta = (\delta_1,\dots,\delta_{n-1})$ hat mehrfache Eigenwerte.

Zunächst untersuchen wir den Fall a):

Ist etwa $c_k = 0$ für ein $k \in \{1,\dots,n-1\}$, dann ist $\lambda_{k+1} := \gamma_k\delta_k$ ein Eigenwert von $C\mathbf{x} = \lambda J\mathbf{x}$ und der k+1-te Einheitsvektor $\mathbf{e}_{k+1}$ der zugehörige Eigenvektor:

$$C\mathbf{e}_{k+1} = \left[\begin{array}{c|c} \gamma_0 & \mathbf{c}^T \\ \hline \mathbf{c} & \mathrm{diag}(\gamma_1,\dots,\gamma_{n-1}) \end{array}\right] \mathbf{e}_{k+1} = \gamma_k\mathbf{e}_{k+1} = (\gamma_k\delta_k)J\mathbf{e}_{k+1}$$

Durch Streichen der k-ten Zeile und Spalte kann man jetzt zu dem kleineren Eigenwertproblem

$$C'\mathbf{x}' = \left[\begin{array}{c|cccccccc} \gamma_0 & c_1 & c_2 & \dots & c_{k-1} & c_{k+1} & \dots & c_{n-1} \\ \hline c_1 & \gamma_1 & & & & & & \\ c_2 & & \gamma_2 & & & & & \\ \vdots & & & \ddots & & & & \\ c_{k-1} & & & & \gamma_{k-1} & & & \\ c_{k+1} & & & & & \gamma_{k+1} & & \\ \vdots & & & & & & \ddots & \\ c_{n-1} & & & & & & & \gamma_{n-1} \end{array}\right] \mathbf{x}' = \lambda J'\mathbf{x}'$$

mit

$$J' = \mathrm{diag}(\delta_0,\delta_1,\dots,\delta_{k-1},\delta_{k+1},,\delta_{n-1})$$

übergehen, denn:

Ist $\mathbf{x}'^T = (x_1,\dots,x_k,x_{k+2},\dots,x_n)$ ein Eigenvektor des kleineren Problems, so definiert $\mathbf{x}^T := (x_1,\dots,x_k,0,x_{k+2},\dots,x_n)$ einen Eigenvektor des großen Problems $C\mathbf{x} = \lambda J\mathbf{x}$.

Die Deflation zerstört auch hier die positive Definitheit nicht.

Diesen Deflationsschritt führen wir für alle Nullkomponenten von $\mathbf{c}^T$ durch. Danach haben wir pro Nullkomponente einen Eigenwert und Eigenvektor des Eigenwertproblems $C\mathbf{x} = \lambda J\mathbf{x}$ berechnet und ein Eigenwertproblem $C'\mathbf{x} = \lambda J'\mathbf{x}$ erhalten, dessen Dimension kleiner oder gleich der des ursprünglichen Pro-

blems ist und das im Vektor $\mathbf{b}$ keine Nullen mehr hat, wobei $\mathbf{b}^T = (b_1,\dots,b_m)$ die von null verschiedenen Komponenten von $\mathbf{c}$ enthält ($m \le n-1$) und die β_i gerade die zu den Nichtnullkomponenten gehörenden Diagonalelemente sind sowie die $j_1,\dots,j_m$ die zugehörigen Elemente aus J.

Ab jetzt habe C die Gestalt

$$C = \begin{bmatrix} \gamma_0 & b_1 & \cdots & b_m \\ b_1 & \beta_1 & & \\ \vdots & & \ddots & \\ b_m & & & \beta_m \end{bmatrix} \quad \text{mit } b_1,\dots,b_m \neq 0 \quad \text{und} \quad J = \begin{bmatrix} j_0 & & & \\ & j_1 & & \\ & & \ddots & \\ & & & j_m \end{bmatrix}.$$

Im Moment können $\beta_i j_i$, $i \in \{1, 2, \dots, m\}$ noch mehrfache Eigenwerte des Paares $\big(\operatorname{diag}(\beta_1,\dots,\beta_m),\ \operatorname{diag}(j_1,\dots,j_m)\big)$ sein.

Das charakteristische Polynom des Eigenwertproblems $C\mathbf{x} = \lambda J\mathbf{x}$ bekommen wir analog zu den Überlegungen bei Wilkinson [68] durch direkte Berechnung von

$$\det(C - \lambda J) = \det \begin{bmatrix} \gamma_0 - \lambda j_0 & b_1 & \cdots & b_m \\ b_1 & \beta_1 - \lambda j_1 & & \\ \vdots & & \ddots & \\ b_m & & & \beta_m - \lambda j_m \end{bmatrix}.$$

Das charakteristische Polynom ist dann gegeben durch

$$p(\lambda) = (\gamma_0 - \lambda j_0) \prod_{i=1}^{m} (\beta_i - \lambda j_i) - \sum_{j=1}^{m} b_j^2 \prod_{i \neq j} (\beta_i - \lambda j_i).$$

Nun überlegen wir uns, was im Fall b) passiert:

Wenn $\beta_i j_i$ ein Eigenwert der Vielfachheit r_i des Eigenwertproblems

$$\operatorname{diag}(\beta_1,\dots,\beta_m)\mathbf{y} = \lambda\,(j_1,\dots,j_m)\mathbf{y}$$

ist, so kommt der Ausdruck

$$(\beta_i - \lambda j_i)^{r_i - 1}$$

in jedem Summanden von p vor. Da die Eigenwerte von (C,J) die Nullstellen von p sind, muß also $\beta_i j_i$ ein Eigenwert der Vielfachheit $r_i - 1$ von (C,J) sein.

Der Deflationsschritt im Fall b) besteht jetzt darin, daß wir die mehrfachen Nullstellen von p herauskürzen:

Unter den $\beta_i j_i$ gebe es t verschiedene Werte, die wir o.B.d.A. mit $\beta_1 j_1, \dots, \beta_t j_t$ bezeichnen und die die Vielfachheit $r_1, r_2, \dots, r_t$ mit $r_1 + r_2 + \dots + r_t = m$ haben. Falls $t = m$ ist, so sind alle $\beta_i j_i$ einfach.

In $p(\lambda)$ kommt der Faktor F mit

$$F := \prod_{i=1}^{t} (\beta_i - \lambda j_i)^{r_i - 1}$$

vor, der die Eigenwerte von $C\mathbf{x} = \lambda J\mathbf{x}$ der Vielfachheit $r_i - 1$ enthält.

Wenn wir die Gleichung $p(\lambda) = 0$ durch F teilen, sehen wir, daß die noch zu berechnenden Eigenwerte die Nullstellen der Funktion

$$(6.2) \qquad w(\lambda) := (\gamma_0 - \lambda j_0) - \sum_{i=1}^{t} z_i^2 (\beta_i - \lambda j_i)^{-1}$$

sind, wobei jedes z_i^2 die Summe der r_i Werte b_j^2 ist, die zu $\beta_i j_i$ gehören. Daher ist $z_i^2 > 0$ für alle $i = 1, \dots, t$.

Die Nullstellensuche für die t+1 Nullstellen von w besprechen wir im übernächsten Abschnitt.

6.1.2 Berechnung der Eigenvektoren

Jetzt berechnen wir die Eigenvektoren zu den Eigenwerten von $C\mathbf{x} = \lambda J\mathbf{x}$:

$$\begin{bmatrix} \gamma_0 & b_1 & \cdots & b_m \\ b_1 & \beta_1 & & \\ \vdots & & \ddots & \\ b_m & & & \beta_m \end{bmatrix} \mathbf{x} = \lambda J \begin{bmatrix} j_0 & & & \\ & j_1 & & \\ & & \ddots & \\ & & & j_m \end{bmatrix} \mathbf{x}$$

Wir setzen dabei voraus, daß alle $b_j \neq 0$ sind für $j = 1,\dots,m$.

1. Fall: Eigenvektoren zu mehrfachen Eigenwerten:

Wie wir gerade gesehen haben, ist ein r-facher Eigenwert $\lambda = \beta_l j_l$ von

$$\big(\operatorname{diag}(\beta_1,\dots,\beta_m)\big)\mathbf{y} = \lambda\big(\operatorname{diag}(j_1,\dots,j_m)\big)\mathbf{y}$$

ein r-1-facher Eigenwert von $C\mathbf{x} = \lambda J\mathbf{x}$.

Wegen

$$\beta_l - \lambda j_l = 0,\ \beta_{l+1} - \lambda j_{l+1} = 0,\ \dots,\ \beta_{l+r-1} - \lambda j_{l+r-1} = 0$$

hat das Gleichungssystem

$$(C - \lambda J)\mathbf{x} = \mathbf{0}$$

zur Berechnung des Eigenvektors $\mathbf{x} = (x_0, x_1, \dots, x_m)$ zum Eigenwert $\lambda = \beta_l j_l$ die Gestalt

$$(6.3) \qquad (\gamma_0 - \lambda j_0)x_0 + \sum_{i=1}^{m} b_i x_i = 0$$

$$(6.4) \qquad b_k x_0 + (\beta_k - \lambda j_k)x_k = 0 \quad \text{für } k \in \{1,\dots,m\}\setminus\{l,l+1,\dots,l+r-1\}$$

$$(6.5) \qquad b_k x_0 = 0 \quad \text{für } k \in \{l,l+1,\dots,l+r-1\}.$$

Aus (6.5) folgt $x_0 = 0$ und aus (6.4) dann wegen $(\beta_k - \lambda j_k) \neq 0$ gerade $x_k = 0$ für $k \in \{1,\dots,m\}\setminus\{l,l+1,\dots,l+r-1\}$. Dann hat (6.3) die Gestalt

$$\sum_{i=l}^{l+r-1} b_i x_i = 0 \quad \text{oder} \quad x_l = -\sum_{i=l+1}^{l+r-1} b_i x_i \frac{1}{b_l}.$$

Damit sind die Komponenten $x_{l+1},\dots,x_{l+r-1}$ frei wählbar.

Zum Eigenwert $\lambda = \beta_l j_l$ gehören jetzt r-1 Eigenvektoren $\mathbf{x}_1,\dots,\mathbf{x}_{r-1}$.
Wählen wir nacheinander für $k = l+1,\dots,l+r-1$

$x_k := 1$

$x_j := 0$ für $j \in \{l+1,\dots,l+r-1\}\setminus\{k\}$

und berechnen

$x_l = -b_k / b_l$,

so erhalten wir r-1 linear unabhängige Eigenvektoren $\mathbf{x}_1,\dots,\mathbf{x}_{r-1}$ mit der J-Norm

$$\mathbf{x}^T J \mathbf{x} = j_k + (b_k/b_l)^2 j_l .$$

Aus der positiven Definitheit des Paares $\big(\mathrm{diag}(\beta_i),\mathrm{diag}(j_i)\big)$ folgt, daß für einen mehrfachen Eigenwert $\lambda = \beta_l j_l$ alle zugehörigen $j_l,\dots,j_{l+r-1}$ gleich sein müssen. $\mathbf{x}^T J \mathbf{x}$ ist für diese Eigenvektoren $\mathbf{x}$ also immer ungleich 0.

Es ist daher möglich, die Eigenvektoren zu λ mit Hilfe eines verallgemeinerten Gram-Schmidt-Orthogonalisierungsverfahren folgendermaßen zu J-orthonormieren:

Verallgemeinertes Gram-Schmidt-Orthogonalisierungsverfahren:

Gegeben seien r-1 linear unabhängige Vektoren $\mathbf{x}_1, \dots, \mathbf{x}_{r-1}$.
Die zu berechnenden J-orthonormalen Vektoren bezeichen wir mit $\mathbf{q}_1,\dots,\mathbf{q}_{r-1}$.
Wir J-normieren den Vektor $\mathbf{x}_1$ und setzen $\mathbf{q}_1 := \mathbf{x}_1$.
Den j+1-ten Vektor $\mathbf{q}'_{j+1}$ bestimmen wir dann mit der Formel

$$\mathbf{q}'_{j+1} := \mathbf{x}_{j+1} - j_1 \sum_{i=1}^{j} (\mathbf{q}_i^T J \mathbf{x}_{j+1})\, \mathbf{q}_i .$$

$\mathbf{q}'_{j+1}$ ist jetzt J-orthogonal zu $\mathbf{q}_1,\dots,\mathbf{q}_j$.
Den gesuchten Vektor $\mathbf{q}_{j+1}$ erhalten wir durch J-Normierung von $\mathbf{q}'_{j+1}$.

2. Fall: Eigenvektoren zu "einfachen" Eigenwerten:

Ist $\lambda = \beta_l j_l$ ein einfacher Eigenwert von $C\mathbf{x} = \lambda J\mathbf{x}$, so hat das Gleichungssystem zur Bestimmung des Eigenvektors $\mathbf{x}$ zum Eigenwert λ die Gestalt

$$(6.6) \qquad (\gamma_0 - \lambda j_0)x_0 + \sum_{i=1}^{m} b_i x_i = 0$$

$$(6.7) \qquad b_k x_0 + (\beta_k - \lambda j_k)x_k = 0 \ , \ k=1,\dots,m .$$

Wegen $(\beta_k - \lambda j_k) \neq 0$ folgt aus (6.7):

$$x_k = -\frac{b_k}{\beta_k - \lambda j_k} x_0 \quad \text{für } k = 1,\dots,m.$$

Einsetzen in (6.6) liefert

$$\Big((\gamma_0 - \lambda j_0) - \sum_{i=1}^{m} \frac{b_i^2}{\beta_i - \lambda j_i}\Big)x_0 = 0 \ , \quad \text{d.h. } x_0 w(\lambda) = 0 \quad \text{(vgl. Formel 6.2).}$$

Für den Eigenwert λ ist $w(\lambda)$ natürlich gleich null, x_0 ist somit frei wählbar.
Mit $x_0 := 1$ hat der Eigenvektor $\mathbf{x}$ zum Eigenwert λ die Gestalt

$$(6.8) \qquad \mathbf{x} = \begin{bmatrix} 1 \\ \Big(-\dfrac{b_k}{\beta_k - \lambda j_k}\Big)_{k=1,\dots,m} \end{bmatrix}.$$

Unter der Voraussetzung der positiven Definitheit des Paares (C,J) gilt $\mathbf{x}^T J \mathbf{x} \neq 0$ und $\mathbf{x}$ kann J-normiert werden.

6.1.3 Die Eigenwerte als Nullstellen der Funktion w

Die Dimension des Eigenwertproblems $C\mathbf{x} = \lambda J\mathbf{x}$ nach den beiden Deflationsschritten bezeichnen wir mit m+1. Die jetzt noch zu berechnenden m+1 Eigenwerte sind Nullstellen der Funktion

$$(6.9) \qquad w(x) = (\gamma_0 - x j_0) - \sum_{i=1}^{m} \frac{z_i^2}{\beta_i - x j_i} = (\gamma_0 j_0 - x) j_0 - \sum_{i=1}^{m} \frac{z_i^2 j_i}{\beta_i j_i - x}$$

$$= (\alpha_0 - x) j_0 - \sum_{i=1}^{m} \frac{z_i^2 j_i}{\alpha_i - x}$$

$$= (\alpha_0 - x) j_0 + \sum_{i=1}^{q} \frac{z_i^2}{\alpha_i - x} - \sum_{i=q+1}^{m} \frac{z_i^2}{\alpha_i - x}$$

mit $\alpha_0 := \gamma_0 j_0$, $\alpha_i := \beta_i j_i$ für $i = 1,\dots,m$ und $q :=$ Anzahl der Minuseinsen in $\mathrm{diag}(j_1,\dots,j_m)$.

Die α_i, $i = 1,\dots,m$, sind dabei die Eigenwerte des als positiv definit vorausgesetzten Paares $\left(\mathrm{diag}(\beta_1,\dots,\beta_m), \mathrm{diag}(j_1,\dots,j_m)\right)$ und können daher in der Form

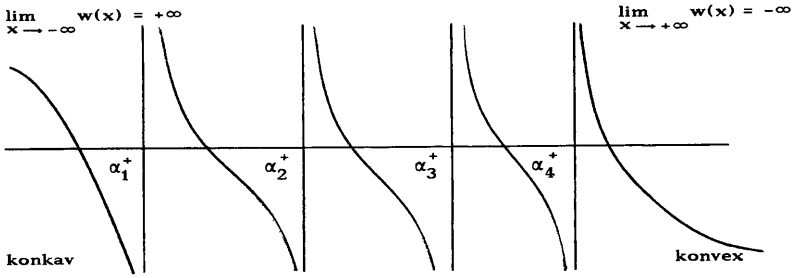
$$\alpha_1^- < \alpha_2^- < \dots < \alpha_q^- < \alpha_{q+1}^+ < \dots < \alpha_m^+$$

angeordnet werden, wobei α_j^- für die "Minuseigenwerte" steht und α_i^+ für die "Pluseigenwerte".

Die Lage der Nullstellen hängt von q und j_0 ab:

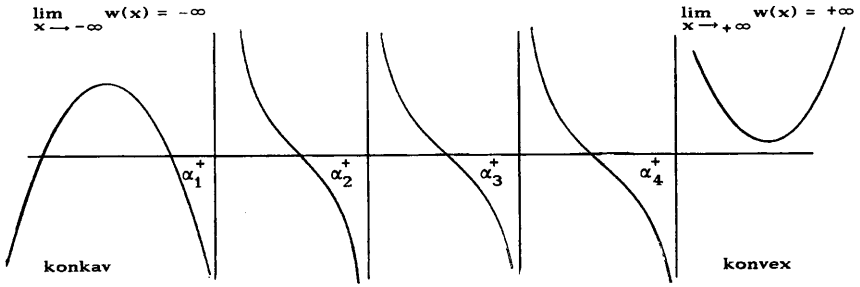
1. $q = 0$, $j_0 = 1$:

Die Funktion w ist auf jedem der Teilintervalle $(-\infty, \alpha_1)$, (α_1, α_2), ... , (α_{m-1}, α_m), $(\alpha_m, +\infty)$ streng monoton fallend und hat dort genau eine Nullstelle.



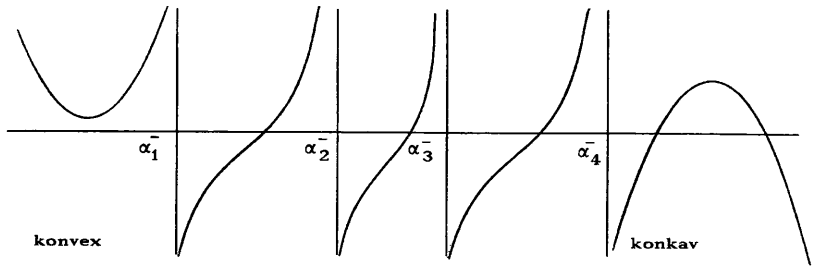
2. $q = 0$, $j_0 = -1$:

Die Funktion w ist nicht mehr streng monoton auf den Teilintervallen des Definitionsbereichs. In den Intervallen (α_1, α_2), ... , (α_{m-1}, α_m) liegt jeweils genau eine Nullstelle, im Intervall Defint := $(-\infty, \alpha_1)$ können eine, zwei oder keine Nullstelle liegen. Das Paar (C,J) ist nicht positiv definit, wenn es in Defint weniger als zwei Nullstellen gibt.



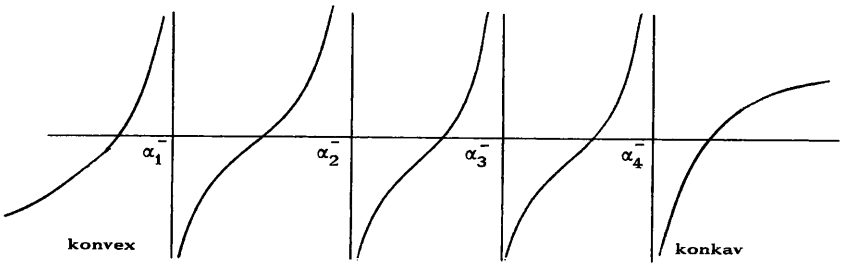
3. $q = m$, $j_0 = +1$:

Auch hier ist die Funktion w nicht monoton auf den Teilintervallen ihres Definitionsbereichs. In den Intervallen (α_1, α_2), ... , (α_{m-1}, α_m) liegt jeweils genau eine Nullstelle, im Intervall Defint := $(\alpha_m, +\infty)$ können eine, zwei oder keine Nullstellen liegen. Das Paar (C,J) ist wiederum nicht positiv definit, wenn es in Defint weniger als zwei Nullstellen gibt.



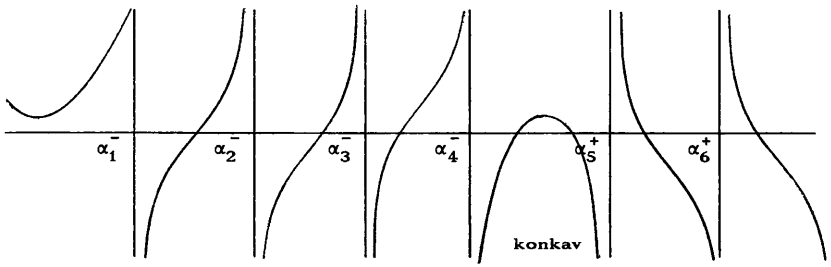
4. $q = m$, $j_0 = -1$:

Die Funktion w ist hier streng monoton wachsend auf den Teilintervallen $(-\infty, \alpha_1)$, ... , (α_{m-1}, α_m), $(\alpha_m, +\infty)$ ihres Definitionsbereichs. In jedem dieser Intervalle liegt genau eine Nullstelle.



5. $1 \le q < m$, $j_0 = +1$

In diesem Fall ist die Funktion w im allgemeinen nicht mehr monoton auf ihren Definitionsintervallen. In jedem dieser Intervalle (α_1, α_2) ... , (α_{q-1}, α_q), $(\alpha_{q+1}, \alpha_{q+2})$, ... , (α_{m-1}, α_m), $(\alpha_m, +\infty)$ liegt eine Nullstelle. Im Intervall Defint := (α_q, α_{q+1}) können eine, zwei oder keine Nullstelle liegen:



Das Paar (C,J) ist nicht positiv definit, wenn es in Defint weniger als zwei Nullstellen gibt.

6. $1 \le q < m$, $j_0 = -1$:

In jedem der Intervalle $(-\infty, \alpha_1)$, (α_1, α_2), ... , (α_{q-1}, α_q), $(\alpha_{q+1}, \alpha_{q+2})$, ..., (α_{m-1}, α_m) liegt eine Nullstelle. Wie im Fall 5 können im Definitheitsintervall Defint := (α_q, α_{q+1}) eine, zwei oder keine Nullstelle liegen.

Die Anzahl der Nullstellen im Definitheitsintervall entscheidet über die positive Definitheit des Paares (C,J). Wegen

$$w''(x) = \sum_{i=1}^{q} \frac{2 z_i^2}{(\alpha_i - x)^3} - \sum_{i=q+1}^{m} \frac{2 z_i^2}{(\alpha_i - x)^3} < 0 \quad \text{(vgl. Fall 2, 3, 5, 6)}$$

ist die Funktion w im Definitheitsintervall konkav, man kann daher durch die Berechnung ihres Extremums feststellen, ob es in diesem Intervall 2, 1 oder keine Nullstelle gibt: Liegt für das Extremum x_E der Fall $w(x_E) > 0$ vor, so hat w zwei Nullstellen in Defint; andernfalls können wir die Rechnung mit der Meldung "Das Paar (C,J) ist nicht positiv definit" beenden.

6.1.4 Das Einschließungsverfahren

Das Einschließungsverfahren aus dem Divide-et-Impera-Programm (vgl. Abschnitt 4.2) kann mit den folgenden Änderungen zur Nullstelleneingrenzung für die hier zugrundeliegende Funktion w angewendet werden:

a) Für $x \in (\alpha_1, \alpha_m)$ wählen wir die Hilfsfunktionen h_1 und h_2 ähnlich wie in Abschnitt 4.2.1 beschrieben aus, nur wird jetzt der lineare Term $(\alpha_0 - x)j_0$ zu h_2 hinzugeschlagen. Die Eingrenzungsschritte erfolgen dann analog zum in 4.2 beschriebenen Verfahren.

b) In den Fällen, wo Nullstellen in den Randintervallen $(-\infty, \alpha_1)$ oder $(\alpha_m, +\infty)$ liegen, müssen h_1 und h_2 dieser Situation entsprechend gewählt werden. Wir beschreiben die Wahl von h_1 und h_2 sowie die Durchführung des Eingrenzungsschrittes beispielhaft für das Intervall $(-\infty, \alpha_1)$ im gleich folgenden Abschnitt. Die Ergebnisse für $(\alpha_m, +\infty)$ ergeben sich dann analog.

c) Die Schranken M_u und M_o aus Abschnitt 4.2 ändern sich wegen des linearen Terms in unserer aktuellen Funktion w ebenfalls; die zugehörigen Formeln werden am Ende dieses Abschnitts angegeben.

6.1.4.1 Einschließung der Nullstellen von w, die im Intervall $(-\infty, \alpha_1)$ liegen

Wie wir gerade gesehen haben, besitzt die Funktion w unter der Voraussetzung der positiven Definitheit des Paares $\left(\mathrm{diag}(\beta_1, \ldots, \beta_m),\ \mathrm{diag}(j_1, \ldots, j_m)\right)$ außerhalb des Intervalls (α_1, α_m) mindestens eine Nullstelle.

Auf dem Intervall $(-\infty, \alpha_1)$ schreiben wir w als

$$w(x) = -\left(-(\alpha_0 - x)j_0 + \frac{z_1^2 j_1}{\alpha_1 - x}\right) - \sum_{i=2}^{m} \frac{z_i^2 j_i}{\alpha_i - x}$$

$$=: \; -h_1(x) + h_2(x)$$

mit

$$h_1(x) := -(\alpha_0 - x)j_0 + \frac{z_1^2 j_1}{\alpha_1 - x}$$

und

$$h_2(x) := -\sum_{i=2}^{m} \frac{z_i^2 j_i}{\alpha_i - x} .$$

Wir haben eine Nullstelle gefunden, wenn

$$h_1(x_N) = h_2(x_N)$$

gilt.

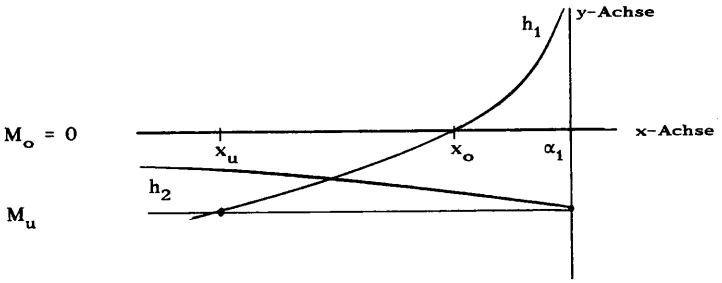
Auf dem Intervall $D := (-\infty, \alpha_1)$ unterscheiden wir die folgenden Fälle:

1. Fall: $q = 0$, $j_0 = +1$, d.h. $\mathrm{sgn}(j_0) = \mathrm{sgn}(j_1) = 1$:

h_1 mit $h_1(x) = -\left((\alpha_0 - x) - \dfrac{z_1^2}{\alpha_1 - x}\right)$ ist streng monoton wachsend auf dem Intervall D und konvex,

h_2 mit $h_2(x) = -\sum_{i=2}^{m} \dfrac{z_i^2}{\alpha_i - x}$ ist kleiner null auf D, streng monoton fallend und konkav.

(Eine obere Schranke für h_2 wäre also z.B. $M_o = 0$, eine untere $M_u \le h_2(\alpha_1)$.)



Das Lösen der Gleichung
$h_1(x) = M_u$
liefert ein $x_u \in \mathbb{R}$.

Das Lösen der Gleichung
$h_1(x) = M_o$
liefert ein $x_o \in \mathbb{R}$.

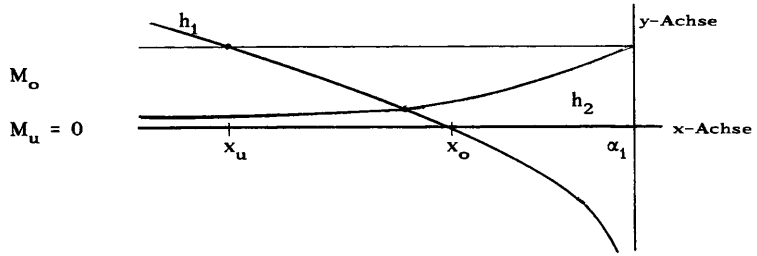
Außerhalb des Intervalls $[x_u, x_o]$ kann keine Nullstelle von w liegen.

2. Fall: $q = m$, $j_0 = -1$, d.h. $\mathrm{sgn}(j_0) = \mathrm{sgn}(j_1) = -1$:

Die Funktion h_1 mit $h_1(x) = \left((\alpha_0 - x) - \dfrac{z_1^2}{\alpha_1 - x}\right)$ ist auf D streng monoton fallend und konkav,

die Funktion h_2 mit $h_2(x) = +\sum_{i=2}^{m} \dfrac{z_i^2}{\alpha_i - x}$ ist größer null auf D, streng monoton wachsend und konvex.

(Eine untere Schranke für h_2 wäre also z.B. $M_u = 0$, eine obere $M_o \ge h_2(\alpha_1)$.)



Das Lösen der Gleichung
$h_1(x) = M_o$
liefert ein $x_u \in \mathbb{R}$

Das Lösen der Gleichung
$h_1(x) = M_u$
liefert ein $x_o \in \mathbb{R}$

Außerhalb des Intervalls $[x_u, x_o]$ kann keine Nullstelle von w liegen.

3. Fall: $0 < q < m$, $j_0 = -1$, d. h. $\mathrm{sgn}(j_0) = \mathrm{sgn}(j_1) = -1$:

Die Funktion h_1 hat dieselbe Gestalt wie im Fall 2 und ist daher streng monoton fallend. Die Funktion h_2 dagegen ist jetzt (für $q \ge 2$) weder positiv noch negativ:

$$h_2(x) = +\sum_{i=2}^{q} \frac{z_i^2}{\alpha_i - x} - \sum_{i=q+1}^{m} \frac{z_i^2}{\alpha_i - x}$$

Die Skizze unterscheidet sich von der aus Fall 2 nun dadurch, daß M_u nicht null ist und h_2 im allgemeinen nicht monoton ist und auch negative Werte annehmen kann. Die Bestimmung des Intervalls $[x_u, x_o]$ erfolgt analog.

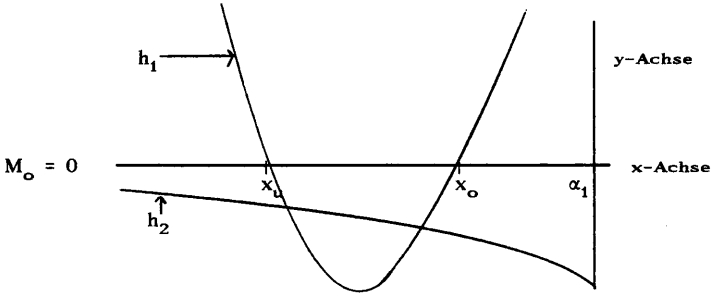
4. Fall: $q = 0$, $j_0 = -1$, d.h. $\mathrm{sgn}(j_0) \ne \mathrm{sgn}(j_1)$:

Jetzt befinden wir uns im Fall des Definitheitsintervalls, in dem eine, zwei oder keine reelle Nullstelle liegen können. Wie bei unserem früheren Einschließungsverfahren beschrieben, führen wir nun zunächst einen Einschliessungsschritt durch, um ein Intervall $[x_u, x_o]$ zu finden, in dem sicher die beiden Nullstellen von w liegen, wenn sie existieren. Anschließend suchen wir in diesem Intervall das Extremum bzw. ein x_E mit positivem Funktionswert $w(x_E)$. Die weitere Behandlung des Definitheitsintervalls erfolgt dann genauso wie in unserem früheren Einschließungsverfahren in Abschnitt 4.2.

Die Funktion h_1 mit $h_1(x) = (\alpha_0 - x) + \frac{z_1^2}{\alpha_1 - x}$ ist auf D konvex,

h_2 mit $h_2(x) = -\sum_{i=2}^{m} \frac{z_i^2}{\alpha_i - x}$ ist auf D kleiner null, streng monoton fallend und konkav.

(Eine obere Schranke für h_2 wäre z.B. $M_o = 0$.)



Wenn die Gleichung $h_1(x) = M_o$ zwei Lösungen x_u und x_o hat, so bestimmen diese das Intervall $[x_u, x_o]$, in dem die Nullstellen von w liegen müssen, falls diese existieren. Gibt es nur eine oder keine Nullstelle, so beenden wir das Programm mit der Meldung "Das Paar ist nicht positiv definit".

In den restlichen Fällen liegt für positiv definite Matrizenpaare keine Nullstelle in $(-\infty, \alpha_1)$.

Die Ergebnisse für das Intervall $D =]\alpha_m, +\infty[$ ergeben sich analog.

6.1.4.2 Berechnung der Schranken M_u und M_o

Wenn wir uns in einem der Intervalle (α_k, α_{k+1}) mit $k = 1, \ldots m-1$, befinden, so können wir für $x \in [x_u, x_o] \subseteq (\alpha_k, \alpha_{k+1})$ die Schrankem M_u und M_o wie folgt bestimmen:

$$M_o := (\alpha_0 - \xi_0) j_0 - \sum_{i \neq k,k+1}^{m} \frac{z_i^2 j_i}{\alpha_i - \xi_i}$$

mit

$$\xi_i := \begin{cases} x_u & \text{für } j_i = +1 \\ x_o & \text{für } j_i = -1 \end{cases} \quad \text{und} \quad \xi_0 := \begin{cases} x_u & \text{für } j_0 = +1 \\ x_o & \text{für } j_0 = -1 \end{cases},$$

$$M_u := (\alpha_0 - \xi_0) j_0 - \sum_{i \neq k,k+1}^{m} \frac{z_i^2 j_i}{\alpha_i - \xi_i}$$

mit

$$\xi_i := \begin{cases} x_o & \text{für } j_i = +1 \\ x_u & \text{für } j_i = -1 \end{cases} \quad \text{und} \quad \xi_0 := \begin{cases} x_o & \text{für } j_0 = +1 \\ x_u & \text{für } j_0 = -1 \end{cases}.$$

Im Fall des Intervalls $(-\infty, \alpha_1)$ gehen wir folgendermaßen vor:

Wegen

$$\begin{aligned} \max_{x<\alpha_1} h_2(x) &= \max_{x<\alpha_1} \left(\sum_{i=2}^{q} \frac{z_i^2}{\alpha_i - x} - \sum_{i=q+1}^{m} \frac{z_i^2}{\alpha_i - x} \right) \\ &\leq \max_{x<\alpha_1} \left(\sum_{i=2}^{q} \frac{z_i^2}{\alpha_i - x} \right) \\ &\leq \sum_{i=2}^{q} \frac{z_i^2}{\alpha_i - \alpha_1} \end{aligned}$$

wählen wir

$$M_o := \sum_{i=2}^{q} \frac{z_i^2}{\alpha_i - \alpha_1} \quad \text{für } q \geq 2 \text{ , sonst } M_o := 0.$$

Eine Abschätzung nach unten bekommen wir aus

$$\min_{x<\alpha_1} h_2(x) \geq \min_{x<\alpha_1} \left(-\sum_{\substack{i=q+1 \\ i\geq 2}}^{m} \frac{z_i^2}{\alpha_i - x} \right) \geq -\sum_{\substack{i=q+1 \\ i\geq 2}}^{m} \frac{z_i^2}{\alpha_i - \alpha_1}.$$

Für $\max(2, q+1) \leq m$ setzen wir $M_u := -\sum_{\substack{i=q+1 \\ i\geq 2}}^{m} \frac{z_i^2}{\alpha_i - \alpha_1}$, sonst $M_u := 0$.

6.1.4.3 Ein lineares Eingrenzungsverfahren

Wegen möglicher Ungenauigkeiten beim Lösen quadratischer Gleichungen wird auch diesem quadratischen Eingrenzungsverfahren ein lineares vorgeschaltet. Das neue Intervall $[x_u, x_o]$, in dem die zu berechnende Nullstelle oder die zwei Nullstellen liegen, ergibt sich folgendermaßen:

linkes Randintervall:

$$x_u := \min\left(\alpha_1 - 1,\ \alpha_0 - \sum_{j=1}^{m} \frac{z_j^2}{|\alpha_j - \alpha_1 + 1|}\right)$$

$$x_o := \max\left(\alpha_1 - 1,\ \alpha_1 - \frac{z_1^2}{|\alpha_1 - \alpha_0| + 1 + \sum_{j>1} \frac{z_j^2}{|\alpha_j - \alpha_1|}}\right)$$

rechtes Randintervall:

$$x_o := \max\left(\alpha_m + 1,\ \alpha_0 + \sum_{j=1}^{m} \frac{z_j^2}{|\alpha_j - \alpha_m - 1|}\right)$$

$$x_u := \min\left(\alpha_m + 1,\ \alpha_m + \frac{z_m^2}{|\alpha_m - \alpha_0| + 1 + \sum_{j<m} \frac{z_j^2}{|\alpha_j - \alpha_m|}}\right)$$

Intervall $]\alpha_i, \alpha_{i+1}[$, $i = 1, \ldots, m-1$:

$$x_u := \min\left(\frac{\alpha_i + \alpha_{i+1}}{2},\ \alpha_i + \frac{z_i^2}{\frac{\alpha_{i+1} - \alpha_i}{2} + |\alpha_0 - \alpha_i| + M_{i,u}}\right)$$

mit

$$M_{i,u} := \sum_{j \le i-1} \frac{z_j^2}{|\alpha_j - \alpha_i|} + \sum_{j \ge i+1} \frac{z_j^2}{\left|\alpha_j - \frac{\alpha_{i+1} + \alpha_i}{2}\right|}$$

und

$$x_o := \max\left(\frac{\alpha_i + \alpha_{i+1}}{2},\ \alpha_{i+1} - \frac{z_{i+1}^2}{\frac{\alpha_{i+1} - \alpha_i}{2} + |\alpha_0 - \alpha_{i+1}| + M_{i,o}}\right)$$

mit

$$M_{i,o} := \sum_{j \ge i+2} \frac{z_j^2}{|\alpha_j - \alpha_{i+1}|} + \sum_{j \le i} \frac{z_j^2}{\left|\alpha_j - \frac{\alpha_{i+1} + \alpha_i}{2}\right|}$$

6.2 Ein zusammengesetztes Divide-et-Impera-Verfahren für ineinandergeschachtelte berandete Tridiagonalmatrizen

Das Eigenwertproblem $T'\mathbf{x} = \lambda J\mathbf{x}$, mit $J = \mathrm{diag}(\pm 1)$ und T' Ineinanderschachtelung berandeter Tridiagonalmatrizen, kann so in Teilprobleme zerlegt werden, daß das verallgemeinerte Divide-et-Impera-Verfahren aus Kapitel 4 oder der Algorithmus für berandete Diagonalmatrizen aus Abschnitt 6.1 auf diese Teilprobleme anwendbar ist. Dazu zerlegen wir T' in Tridiagonalmatrizen und in Matrizen, die bis auf die erste Zeile und Spalte tridiagonal sind. Sobald zwei solche Teilprobleme gelöst sind, werden sie mit einem Impera-Schritt ähnlich wie beim verallgemeinerten Divide-et-Impera-Verfahren zusammengefaßt.

Wir schauen uns jetzt genauer an, wie wir eine Matrix A, die aus ineinandergeschachtelten berandeten Tridiagonalmatrizen besteht, aufteilen müssen, um Algorithmen vom Divide-et-Impera-Typ auf das Eigenwertproblem $A\mathbf{x} = \lambda J\mathbf{x}$ anwenden zu können. Mit *Divimp* bezeichnen wir den verallgemeinerten Divide-et-Impera-Algorithmus für das Paar (A,J) mit A Tridiagonalmatrix (vgl. Kapitel 4), mit *Randdiag* den entsprechenden für eine berandete Diagonalmatrix C und eine Matrix J (s. Abschnitt 6.1).

Zunächst betrachten wir den Fall, daß A in einen *tridiagonalen Teil* und eine *berandete Tridiagonalmatrix* zerlegbar ist:

$$A = \left[\begin{array}{cccccc|c|ccccc}
a_1 & b_1 & & & & & & & & & & \\
b_1 & \ddots & \ddots & & & & & & & & & \\
 & \ddots & \ddots & b_{k-1} & & & & & & & & \\
 & & b_{k-1} & a_k & b_k & & & & & & & \\
 & & & b_k & a_{k+1} & c_{k+1,k+2} & c_{k+1,k+3} & \cdots & c_{k+1,n} & & & \\
\hline
 & & & & c_{k+2,k+1} & & & & & & & \\
 & & & & c_{k+3,k+1} & & & & & & & \\
 & & & & \vdots & & & \text{Tridiagonalmatrix} & & & & \\
 & & & & c_{n,k+1} & & & T_2 & & & &
\end{array}\right]$$

$$
= \left[\begin{array}{cccc|c}
a_1 & b_1 & & & \\
b_1 & \ddots & \ddots & & \\
 & \ddots & \ddots & b_{k-1} & \\
 & & b_{k-1} & a_k + |b_k| & \\
\hline
 & & & & \begin{array}{c|ccc} a_{k+1} + |b_k| & c_{k+1,k+2} & c_{k+1,k+3} \cdots & c_{k+1,n} \\ \hline c_{k+2,k+1} & & & \\ c_{k+3,k+1} & & T_2 & \\ \vdots & & & \\ c_{n,k+1} & & & \end{array}
\end{array}\right]
$$

$$
- |b_k| \, \mathbf{u}\mathbf{u}^T
$$

$$
=: \left[\begin{array}{c|c} T_1 & \\ \hline & R_1 \end{array}\right] - |b_k| \, \mathbf{u}\mathbf{u}^T \quad \text{mit } \mathbf{u}^T := (0 \ldots 1, -\operatorname{sgn}(b_k), 0 \ldots 0),
$$

wobei die Nichtnullelemente in $\mathbf{u}$ an den Positionen k und k+1 stehen.

Dazu passend wird die Matrix J zerlegt:

$$
J = \left[\begin{array}{c|c} J_{T_1} & \\ \hline & J_{R_1} \end{array}\right]
$$

Aus der positiven Definitheit des Paares (A, J) folgt die positive Definitheit des Paares

$$
A_{mod} := \left[\begin{array}{c|c} T_1 & \\ \hline & R_1 \end{array}\right], \quad J = \left[\begin{array}{c|c} J_{T_1} & \\ \hline & J_{R_1} \end{array}\right],
$$

denn es gilt:

$$
\begin{aligned}
(\mathbf{x}, (A_{mod} - \mu J)\mathbf{x}) &= (\mathbf{x}, (A - \mu J)\mathbf{x}) + (\mathbf{x}, |b_k| \, \mathbf{u}\mathbf{u}^T\mathbf{x}) \\
&= (\mathbf{x}, (A - \mu J)\mathbf{x}) + |b_k| \, (\mathbf{x}, \mathbf{u}) \, \mathbf{u}^T\mathbf{x} \\
&= (\mathbf{x}, (A - \mu J)\mathbf{x}) + |b_k| \, (\mathbf{x}, \mathbf{u})^2 \geq (\mathbf{x}, (A - \mu J)\mathbf{x})
\end{aligned}
$$

Aus der positiven Definitheit des Paares (A_{mod}, J) folgt die der Teilpaare (T_1, J_{T_1}) und (R_1, J_{R_1}).

Im folgenden lösen wir jetzt die Teileigenwertprobleme

$$
(6.10) \qquad T_1 \mathbf{y} = \lambda J_{T_1} \mathbf{y}
$$

und

$$
(6.11) \qquad R_1 \mathbf{y} = \lambda J_{R_1} \mathbf{y} .
$$

Ist eines der Paare (T_1, J_{T_1}) und (R_1, J_{R_1}) nicht positiv definit, so kann auch (A, J) nicht positiv sein. Im anderen Fall gibt es Eigenvektormatrizen Q_1, Q_2 und Diagonalmatrizen D_1, D_2 mit

$$
(6.12) \qquad \begin{aligned} Q_1^T T_1 Q_1 &= D_1, & Q_1^T J_{T_1} Q_1 &= J_{T_1} \\ Q_2^T R_1 Q_2 &= D_2, & Q_2^T J_{R_1} Q_2 &= J_{R_1}. \end{aligned}
$$

Analog zum Impera-Schritt beim Divide-et-Impera-Algorithmus geht man vom Eigenwertproblem

$$
A\mathbf{x} = \left(\left[\begin{array}{c|c} T_1 & O \\ \hline O & R_1 \end{array}\right] - |b_k| \, \mathbf{u}\mathbf{u}^T \right) \mathbf{x} = \lambda J \mathbf{x}
$$

über zum Eigenwertproblem

$$
\left(\left[\begin{array}{c|c} D_1 & O \\ \hline O & D_2 \end{array}\right] - |b_k| \, \mathbf{z}\mathbf{z}^T \right) \mathbf{x}' = \lambda J \mathbf{x}'
$$

mit

$$
\mathbf{z} = J Q^T \mathbf{u} \quad \text{und} \quad \mathbf{x}' = Q^{-1}\mathbf{x} = (JQ)^T \mathbf{x} \quad \text{mit } Q := \left[\begin{array}{c|c} Q_1 & O \\ \hline O & Q_2 \end{array}\right]
$$

und löst es, wenn das Paar positiv definit ist, auf die im Kapitel 4 beschriebene Art und Weise.

Wir schauen uns jetzt die Behandlung der Teilprobleme an:

$$
(6.10') \qquad T_1 \mathbf{y} = \lambda J_{T_1} \mathbf{y}
$$

mit der Tridiagonalmatrix T_1 kann direkt mit dem Divide-et-Impera-Algorithmus aus Kapitel 4 gelöst werden.

Für das (verallgemeinerte) Eigenwertproblem

$$
(6.11') \qquad R_1 \mathbf{y} = \lambda J_{R_1} \mathbf{y}
$$

benötigen wir den Divide-et-Impera-Algorithmus *Divimp* und den Algorithmus *Randdiag* für ein Paar (C,J) mit berandeter Diagonalmatrix C:

Wir schreiben R_1 als

$$\left[\begin{array}{c|c} a_{k+1} + |b_k| & c^T \\ \hline c & T_2 \end{array}\right]$$

mit T_2 Tridiagonalmatrix und $c^T := (c_{k+1,k+2}, \dots, c_{k+1,n})$.

J_{R_1} teilen wir auf in

$$\left[\begin{array}{c|c} j_{k+1} & \\ \hline & J_{T_2} \end{array}\right] .$$

Ist das Paar (R_1, J_{R_1}) positiv definit, so auch (T_2, J_{T_2}).

<u>1.Schritt:</u> Wir lösen das Eigenwertproblem

$$T_2 y = \lambda J_{T_2} y$$

mit Hilfe des Algorithmus *Divimp* und erhalten entweder die Meldung "Das Paar ist nicht positiv definit" oder eine Eigenvektormatrix V_2 und eine Diagonalmatrix D_{T_2} mit

$$V_2^T T_2 V_2 = D_{T_2} \quad \text{und} \quad V_2^T J_{T_2} V_2 = J_{T_2} .$$

Damit läßt sich R_1 schreiben als

$$R_1 = \left[\begin{array}{c|c} a_{k+1} + |b_k| & c^T \\ \hline c & V_2^{-T} D_{T_2} V_2^{-1} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^{-T} \end{array}\right] \left[\begin{array}{c|c} a_{k+1} + |b_k| & c^T V_2 \\ \hline V_2^T c & D_{T_2} \end{array}\right] \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^{-1} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^{-T} \end{array}\right] \left[\begin{array}{c|c} a_{k+1} + |b_k| & c'^T \\ \hline c' & D \end{array}\right] \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^{-1} \end{array}\right]$$

mit

$$c'^T := c^T V_2 \quad \text{und} \quad D := D_{T_2} .$$

(Man beachte, daß man V_2 wegen $V_2^{-1} = J V_2^T J$ nicht wirklich invertieren muß.)

Dann hat das Eigenwertproblem

$$\left[\begin{array}{c|c} a_{k+1} + |b_k| & c'^T \\ \hline c' & D \end{array}\right] y' = \lambda J_{R_1} y'$$

dieselben Eigenwerte wie

$$R_1 y = \lambda J_{R_1} y ,$$

die Eigenvektoren lassen sich aus

$$y' = \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^{-1} \end{array}\right] y$$

berechnen.

<u>2.Schritt:</u> Wir lösen das Eigenwertproblem

$$C y' = \lambda J_{R_1} y'$$

für die berandete Diagonalmatrix C mit

$$\left[\begin{array}{c|c} a_{k+1} + |b_k| & c'^T \\ \hline c' & D \end{array}\right]$$

mit Hilfe des *Randdiag*-Algorithmus. Meldet er, daß das Paar (C, J_R) nicht positiv definit ist, so ist auch (R_1, J_{R_1}) und damit (A,J) nicht positiv definit. Sonst erhalten wir Matrizen V_1 und D_C mit

$$V_1^T C V_1 = D_C \quad \text{und} \quad V_1^T J_{R_1} V_1 = J_{R_1} .$$

Damit läßt sich C darstellen als

$$C = V_1^{-T} D_C V_1^{-1} .$$

Insgesamt gilt:

$$V_1^T \underbrace{\left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2^T \end{array}\right] R_1 \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2 \end{array}\right]}_{= C} V_1 = V_1^T C V_1 = D_C$$

Mit

$$Q_2 := \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & V_2 \end{array}\right] V_1 \quad \text{und} \quad D_2 := D_C$$

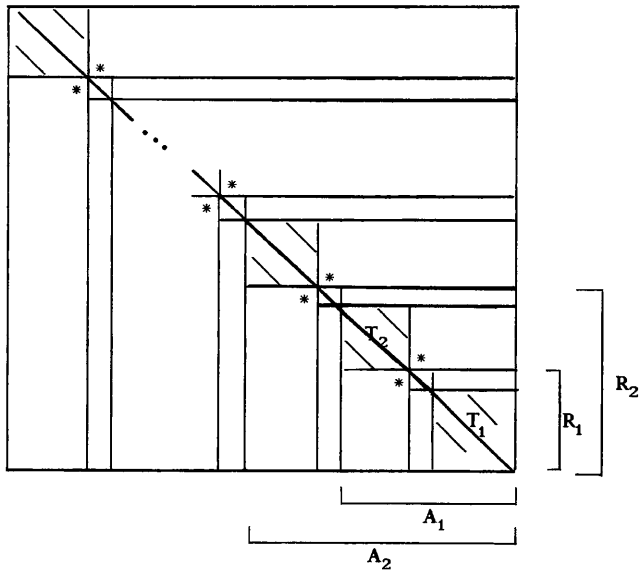
haben wir also

$$Q_2^T R_1 Q_2 = D_2 , \quad Q_2^T J_{R_1} Q_2 = J_{R_1}$$

erhalten.

Im Fall einer *ineinandergeschachtelten berandeten Tridiagonalmatrix* A geht man folgendermaßen vor:

Sei A gegeben durch:



Wir beginnen in der rechten unteren Ecke mit der größten Matrix, die aus einer Tridiagonalmatrix T_1 und einer berandeten Tridiagonalmatrix R_1 besteht. (Dabei darf T_1 auch leer sein.) Wie man solch eine Matrix in

$$A_1 = \left[\begin{array}{c|c} T_1 & \\ \hline & R_1 \end{array}\right] - |b_k| \mathbf{u}\mathbf{u}^T$$

aufspaltet und das Eigenwertproblem $A_1 \mathbf{y} = \lambda J_1 \mathbf{y}$ behandelt, haben wir gerade

gesehen. Wir können also das Eigenwertproblem

$$A_1 \mathbf{y} = \lambda J_1 \mathbf{y}$$

vollständig lösen, wenn das Paar (A_1, J_1) positiv definit ist, oder brechen andernfalls mit der Meldung "Das Paar ist nicht positiv definit" ab.

Die Matrix A_1 übernimmt die Rolle von T_2, wir bilden durch Hinzunahme des nächsten Randes eine neue Matrix R_1 und nehmen die nächste Tridiagonalmatrix als T_1. Das so entstandene Eigenwertproblem

$$A_2 \mathbf{y} = \lambda J_2 \mathbf{y}$$

wird nun behandelt. Entweder sind wir danach fertig ('Abbruch' oder die gesamte Matrix wurde behandelt) oder nehmen A_2 als T_2, bilden mit dem Rand die Matrix R_1 und wählen ein neues T_1 u.s.w.

Wie wir in Kapitel 5 gesehen hatten, ist die vollständige Tridiagonalisierung einer (vollbesetzten) Matrix A unter Beibehaltung der Gestalt von J oft nicht numerisch stabil, wenn J gleich viele +1 und -1 enthält. In diesen Fällen ist unsere teilweise Reduzierung auf Tridiagonalgestalt mit "Rändern" vorzuziehen. Das dabei entstehende verallgemeinerte Eigenwertproblem mit einer Ineinanderschachtelung berandeter Tridiagonalmatrizen kann jetzt mit dem gerade beschriebenen Verfahren gut gelöst werden. Ergebnisse für diese Methode finden sich im Anhang.—

ANHANG

Numerische Ergebnisse

1. Ergebnisse für das Divide-et-Impera-Verfahren

Der verallgemeinerte Divide-et-Impera-Algorithmus liefert für positiv definite Matrizenpaare meist sehr gute Ergebnisse. Wie beim Standard-Verfahren für den Fall J gleich I kann allerdings die (J-)Orthogonalität der Eigenvektoren verlorengehen, wenn die Funktion w, mit deren Hilfe wir die Eigenwerte berechnen, zwei Nullstellen hat, die fast vollständig mit dem zwischen ihnen liegenden Pol übereinstimmen.

1. Vorbemerkungen und Bezeichnungen

Die in dieser Arbeit vorgestellten Programme wurden an IBM-kompatiblen PCs in der Programmiersprache Turbo-Pascal entwickelt. Bei älteren Versionen von Turbo-Pascal kann man mit unserem verallgemeinerten Divide-et-Impera-Algorithmus problemlos Matrizenpaare bis zur Dimension N=72 behandeln, wenn man Variablen für reelle Zahlen als *extended* vereinbart; dann rechnet der PC mit ca. 19 Stellen hinter dem Komma. Stellt man vollbesetzte Matrizen innerhalb des Programms mithilfe von Zeigern dar oder arbeitet mit geringerer Genauigkeit, so sind auch höhere Dimensionen möglich. Ab der Version Turbo-Pascal 7.0 wird die Größe der Matrizen nur noch durch den vorhandenen Hauptspeicher eingeschränkt, bei einem Hauptspeicher von 4 MB sind Matrizen bis zur Dimension 256 noch rechenbar, bei 16 MB kommt man bis zur Dimension 500. Da die Rechenzeit aber stark zunimmt und eigentlich bei Dimensionen über hundert keine Phänome auftreten, die nicht schon vorher beobachtet werden konnten, wurden viele Beispiele mit Dimension kleiner als hundert gewählt.

Um festzustellen, wie sich der Algorithmus bei hohen Dimensionen verhält, wurde das Turbo-Pascal-Programm in ein C-Programm konvertiert, das auf Sun-Sparc-Workstations und Siemens-Großrechnern lauffähig ist — allerdings mit geringerer Genauigkeit (ca. 15 Stellen hinter dem Komma). Die Beispiele hoher Dimension wurden mit einer SUN-Sparc10 gerechnet.

Bezeichnungen:

Bei der Beschreibung der Ergebnisse verwenden wir die folgenden Bezeichnungen und Abkürzungen:

1. *maxF* steht für den maximalen Fehler bei der Eigenwert-Eigenvektor-Berechnung:
$$maxF := \max_{(\lambda,x)} ||Ax - \lambda Jx||_{max}, \quad x\ EV\ zum\ EW\ \lambda$$
2. *JOrth* gibt an, wie (J-)orthogonal die Eigenvektoren sind:
$$JOrth := \max_{i,j} \left| \left(Q^T\, J\, Q\ -\ J\right)_{ij} \right|$$
Dabei bezeichnet Q die Eigenvektormatrix.
3. Die Konstante C steht für eine reelle Zahl zwischen 0 und 10.
4. N bezeichnet die Dimension des Eigenwertproblems.
5. *DefSumme* und *DefGSumme* beschreiben das Deflationsverhalten:
DefSumme ist die Summe aller vorkommenden Deflationen,
DefGSumme die gewichtete Summe aller Deflationen:
$$DefGSumme = \sum_M Deflationsanzahl * M$$
mit M Dimension des aktuellen Teilproblems.

Die im folgenden gemachten Laufzeitangaben sollen aufzeigen, welche Änderungen im Laufzeitverhalten bei verschiedenen Dimensionen und bei unterschiedlichen Beispielarten auftreten. Natürlich sind diese Angaben abhängig vom verwendeten Rechnertyp und den Ausgaben, die man auf Bildschirm und Festplatte macht. Das hier zugrundegelegte Programm schreibt etwa Informationen über das Deflationsverhalten und ähnliches für statistische Zwecke auf die Festplatte, bereits der Verzicht hierauf führt zu einem schnelleren Programm.

2. Eigenwertprobleme mit J = I

Der in dieser Arbeit vorgestellte Divide-et-Impera-Algorithmus unterscheidet sich von dem Algorithmus von Bunch, Nielsen, Sorensen und Cuppen auch im Fall $J = I$ — wir verwenden das von uns entwickelte Eingrenzungsverfahren zur Bestimmung eines Einschließungsintervalls für die Nullstellen, wir teilen das Eigenwertproblem so lange auf, bis wir auf der untersten Stufe nur noch 2x2- und 1x1-Eigenwertprobleme zu lösen haben. Daher wurden zunächst die bekannten Testbeispiele für das Divide-et-Impera-Verfahren mit unserem Algorithmus gerechnet.

1.2.1 Zufallsmatrizen

Es wurden Matrizen A mit Pseudozufallszahlen in [0, 1] auf Diagonale und Nebendiagonale erzeugt sowie mit Pseudozufallszahlen in [-1, 1] auf der Diagonalen, [-f, f] auf der Nebendiagonalen. Dabei wurde die Startzufallszahl für den Zufallszahlengenerator festgesetzt, damit die Beispiele reproduzierbar sind. Die hierzu gerechneten Beispiele lieferten sehr gute Ergebnisse. Als interessantes Beispiel schauen wir uns das Beispiel mit f = 0.1 und Dimension N = 72 an: Im letzten Schritt des Divide-et-Impera-Verfahrens, der Lösung des 72x72-Problems, kommt es zu 40 Deflationen nach z-Komponenten, d.h. 40 der z_i aus

$$w(x) = 1 - \rho \sum_{i=1}^{N} \frac{z_i^2}{\alpha_i - x}$$

sind kleiner als $C * 10^{-20}$ (mit C=3.29)! Trotzdem haben diese Deflationen keine nachteiligen Folgen für die Genauigkeit:

$$JOrth \le C * 10^{-18}, \quad maxF \le C * 10^{-19}$$

1.2.2 Matrizen mit spezieller Bauart

Matrizen mit konstanter Diagonale und Nebendiagonale

1. Matrix A mit: $diag[i] = 2, \quad ndiag[i] = 1, \quad i = 1, \ldots, N$
2. Matrix A mit: $diag[i] = 4, \quad ndiag[i] = 1, \quad i = 1, \ldots, N$

Typisch für diese Beispiele ist das Fehlen von Deflation nach kleinen z-Komponenten und die große Anzahl von Deflationen nach mehrfachen Eigenwerten in den niedrigdimensionalen Teilproblemen.

Die Ergebnisse sind auch bei diesen Beispielen sehr gut:

$$N = 200: \quad JOrth \le C * 10^{-17}, \quad maxF \le C * 10^{-19}, \quad Laufzeit\ (PC\ 386/87) \le 1512\ sec$$

Matrizen mit teilweise kleinen Einträgen in Diagonale und Nebendiagonale

1. Matrix A mit:

$$diag[i] = \begin{cases} 2 & , i = 1, \dots, 4 \text{ und } i = 9, \dots, N, \\ 2 * 10^{-15} & , i = 5, \dots, 8 \end{cases}$$
$$ndiag[i] = \begin{cases} 1 & , i = 1, \dots, 4 \text{ und } i = N-1, N \\ 10^{-15} & , i = 5, \dots, N-2 \end{cases}$$

2. Matrix A mit:

$$diag[i] = 2, \; i = 1, \dots, N$$
$$ndiag[i] = \begin{cases} 1 & , i = 1, \dots, 4 \text{ und } i = 9, \dots, N, \\ 10^{-15} & , i = 5, \dots, 8 \end{cases}$$

Bei diesen Beispielen finden Deflationen nach Eigenwerten und nach z-Komponenten statt. Während das erste Beispiel noch recht gutartig ist,

$$N = 200: \quad Orth \le C * 10^{-18}, \quad maxF \le C * 10^{-17}, \quad Laufzeit\,(PC386/87) \le 1212\,sec$$

gehört das zweite Beispiel zu den kritischen Fällen, die nicht mehr mit der strengen Deflationsschranke $C*10^{-20}$ gerechnet werden können. Schon für das Beispiel der Dimension N = 6 sind die Eigenvektoren nicht mehr ausreichend orthogonal. Dies liegt daran, daß während der Nullstellensuche Eigenwerte berechnet werden, die fast gleich sind und fast vollständig mit dem Pol übereinstimmen:

$$\lambda_1 = 1.99999999999999942$$
$$\lambda_2 = 2.00000000000000058$$
$$Pol = 2.00000000000000000$$

Zwar sind die Eigenwerte noch gut berechnet worden, die mit Hilfe der direkten Formel erhaltenen Eigenvektoren des Teilproblems sind aber fast gleich und damit natürlich nicht orthogonal.

Wählt man nun stattdessen die Deflationsschranke $3.29*10^{-11}$, so erhält man noch gerade akzeptable Ergebnisse: $JOrth \le C * 10^{-18}$, $maxF \le C * 10^{-15}$ $(N = 4, \dots, 72)$. Hierbei treten Deflationen nach mehrfachen Eigenwerten insbesondere in den Zwischenschritten auf, bei den Deflationen nach z-Komponenten kommt es zu bis zu 8 Deflationen pro Teilproblem.

3. Matrix A mit: $diag[i] = i * 10^{-8}$, $ndiag[i] = 1$, $i = 1, \dots, N$.

Bei diesen Matrizen tritt überhaupt keine Deflation auf. Bemerkenswert ist, daß die Nullstellen in diesem Beispiel oft schon durch das Eingrenzungsverfahren (sogar schon durch die Prozedur *Ersteingrenzen*) gefunden werden. Die Ergebnisse sind ebenfalls wieder sehr gut:

$$N = 200: \quad Orth \le C * 10^{-17}, \quad maxF \le C * 10^{-19}, \quad Laufzeit\,(PC386/87) \le 1670\,sec$$

4. Matrix A mit:

$$diag[i] = \begin{cases} 2 & , i = 1, \dots, 4 \\ 2 * 10^{-15} & , i = 5, \dots, N \end{cases}$$
$$ndiag[i] = \begin{cases} 1 & , i = 1, \dots, 4 \text{ und } i = N-1, N \\ 1 * 10^{-15} & , i = 5, \dots, N-2 \end{cases}$$

5. Matrix A mit:

$$diag[i] = \begin{cases} 2 & , i = 1, \dots, 4 \text{ und } i = 9, \dots, N, \\ 2 * 10^{-10} & , i = 5, \dots, 8 \end{cases}$$
$$ndiag[i] = \begin{cases} 1 & , i = 1, \dots, 4 \text{ und } i = N-1, N \\ 10^{-10} & , i = 5, \dots, N-2 \end{cases}$$

Dies sind Beispiele mit Deflationen nach Eigenwerten vor allem in den niedrigerdimensionalen Teilproblemen und einigen Deflationen nach z-Komponenten.

Für die erste Matrix erhalten wir

$$N = 200: \quad Orth \le C * 10^{-17}, \quad maxF \le C * 10^{-15}, \quad Laufzeit\,(PC386/87) \le 1840\,sec,$$

für die zweite dagegen

$$N = 200: \quad Orth \le C * 10^{-17}, \quad maxF \le C * 10^{-11}, \quad Laufzeit\,(PC386/87) \le 1745\,sec$$

1.2.3 Matrizen mit bekannten Eigenwerten

Bei den folgenden Beispielen sind die Eigenwerte bekannt. Unser Algorithmus berechnet die Eigenwerte so genau, daß die vom Programm ausgegebenen Eigenwerte fast vollständig mit den tatsächlichen Eigenwerten übereinstimmen.

1. Matrix A mit:

$$diag[i] = 0, \quad ndiag[i] = \sqrt{i(N-i)}, \; i = 1, \dots, N$$

und den Eigenwerten: $-N+1, -N+3, -N+5, \dots, N-3, N-1.$

Für diese Matrix erhalten wir z.B. als Ergebnis:

$$N = 72: \quad Orth \le C * 10^{-18}, \quad maxF \le C * 10^{-17}$$

2. Matrix A mit:

$$diag[1] = a - b, \quad diag[N] = a + b,$$
$$diag[i] = a, \; i = 2, \dots, N-1$$
$$ndiag[i] = b, \quad i = 1, \dots, N$$

und den Eigenwerten: $\lambda_k = a + 2b \cos \frac{(2k-1)\pi}{2N}$

Hier ergibt sich: $N = 72: \quad Orth \le C * 10^{-18}, \quad maxF \le C * 10^{-19}$

Bei beiden Beispielen kommt keine Deflation nach z-Komponenten vor, die Matrizen in (1.) haben ab N=7 Deflationen nach Eigenwerten für die Teilprobleme niedrigerer Dimension, die Matrizen aus (2.) besitzen nur für die Dimensionen N=16, 32, 64 eine wesentliche Anzahl dieser Deflationen im letzten Impera-Schritt (Reduktion des Nullstellenproblems von Dimension N auf N/2). Interessant an diesen drei Beispielen ist außerdem, daß der größte Teil der Nullstellen bei der abschließenden Nullstellensuche sofort im Eingrenzungsschritt gefunden wird (7, 15, 31 Stück sogar schon mithilfe der Prozedur *Ersteingrenzen*).

1.2.4 Wilkinson-Matrizen

Die Wilkinson-Matrizen WPlus und WMinus

Die sog. *Wilkinson-Matrizen* hat Wilkinson [68] in seinem Buch "The Algebraic Eigenvalue Problem" vorgestellt. Es sind Beispiele, bei denen im allg. die Eigenwerte noch gut berechnet werden, die Eigenvektoren aber die Orthogonalität verlieren können.

Die Wilkinson-Matrizen WPlus und WMinus der Dimension *N:= 2M+1* sind folgendermaßen definiert:.

WPlus:

$$diag[i] = \begin{cases} M+1-i & , i = 1, \dots, M+1 \\ i-M-1 & , i = M+2, \dots, N \end{cases}$$
$$ndiag[i] = 1, \qquad i = 1, \dots, N$$

WMinus:

$$diag[i] = M+1-i, \; i = 1, \dots, N$$
$$ndiag[i] = 1, \qquad i = 1, \dots, N$$

Die Wilkinson-Matrizen WMinus lassen sich vom Divide-et-Impera-Algorithmus sehr gut lösen. Deflation nach *z*-Komponenten findet ab der Dimension N = 33 statt. Ein typisches Beispiel ist N = 71:

22 Deflationen im 64 x 64 - Teilproblem,

43 Deflationen im 71 x 71 - Teilproblem.

Im letzten Schritt muß man also statt 71 Nullstellen nur 28 suchen!

Ganz anders ist die Situation bei den Wilkinson-Matrizen WPlus. Sie sind typische Beispiele dafür, daß der Divide-et-Impera-Algorithmus zwar noch die Eigenwerte recht gut berechnen kann, aber nicht mehr alle Eigenvektoren ausreichend orthogonal sind, da Nullstellen unserer Funktion w mit dem zwischen ihnen liegendem Pol zusammenfallen.

Mit der Deflationskonstanten $C*10^{-20}$ erhalten wir noch einige akzeptable Ergebnisse, etwa für $N = 29: \; Jorth = C*10^{-17}, \; maxF = C*10^{-16}$, bei anderen ist dagegen die Orthogonalität fast vollständig verlorengegangen ($N = 33: \; Jorth = C*10^{-3}, \; N = 35: \; Jorth = C*10^{-2}, \; N = 37: \; Jorth = C*10^{-1}$), obwohl die Eigenwerte gut berechnet werden: $maxF = C*10^{-18}$.

Ändert man jetzt die Abbruchkonstante für die Deflation nach kleinen *z*-Komponenten auf $3.29*10^{-11}$, so erhält man als schlechteste Ergebnisse Resultate in der Größenordnung der Deflationskonstanten: $Jorth \leq C*10^{-10}$ oder $Jorth \leq C*10^{-11}$. Die Abweichung von der Orthogonalität ist also deutlich geringer, der maximale Fehler bei der Eigenwert-Eigenvektor-Berechnung *maxF* liegt meist bei $C*10^{-12}$.

Zusammengesetzte Wilkinson-Matrizen

Recht beliebte Testmatrizen sind auch die folgenden zusammengesetzten Wilkinson-Matrizen. Die Dimension ist dabei eine durch 21 teilbare Zahl. Dann setzt man *Anzahl = N/21*. Die Nebendiagonalelemente im Vektor *ndiag* sind alle gleich 1, die Diagonalelemente berechnen sich wie folgt:

zusammengesetzte WMinus-Matrizen:

Für j = 1, ... , Anzahl bestimme man:

$$diag[i+(j-1)*21] = M+1-i, \; i = 1, \dots, N$$

Am Beispiel der Dimension N=63 kann man die Schwierigkeiten mit dieser Sorte von Matrizen gut erkennen:

Für N = 63 erhält man mit der Abbruchschranke $3.29*10^{-20}$ nur 4 Deflationen nach *z*-Komponenten beim 63 x 63 - Teilproblem. Das Nullstellenverfahren muß also jetzt 59 Nullstellen der Funktion

$$w(x) = 1 - \rho \sum_{i=1}^{59} \frac{z_i^2}{\alpha_i - x}$$

finden. Dies ist nicht einfach, denn 38 Koeffizienten der Nullstellenfunktion sind äußerst klein: $0 < z_i^2 \leq C*10^{-24}$. Die Nullstellen werden trotzdem noch recht gut berechnet. Vergleicht man sie mit den Polstellen der Funktion *w*, so stellt man fest, daß der weitaus größte Teil dieser Eigenwerte sehr nahe an den Polen liegt (dies gilt für 50 von 63 Eigenwerten, aber nur 4 sind durch Deflation entstanden!). Von diesen Eigenwerten liegen wiederum 30 Stück paarweise sehr nahe beieinander. Dies erklärt das schlechte Orthogonalitätsverhalten der Eigenvektormatrix: $Jorth = C*10^{-7}$, während $maxF = C*10^{-18}$ gilt.

Durch Herabsetzen der Abbruchschranke auf $3.29*10^{-11}$ kann man die Orthogonalität der Eigenvektormatrix deutlich verbessern; man erhält $Jorth = C*10^{-19}$, während $maxF = C*10^{-11}$ gilt. Dabei steigt die Anzahl der Deflationen deutlich an: 42 Stück beim 63 x 63 - Teilproblem im Gegensatz zu 4 vorher.

Diese Beispiele wurden für N=21, 42, 63, ... gerechnet. Ab N=42 beginnt der Verlust der Orthogonalität (mit $JOrth \geq 2*10^{-7}$), während der maximale Fehler $maxF \leq C*10^{-14}$ ist. Für N >168 kommt es zum vollständigen Verlust der Orthogonalität.

zusammengesetzte WPlus-Matrizen:

Für j = 1, ... , Anzahl bestimme man:

$$diag[i+(j-1)*21] = \begin{cases} M+1-i & , i = 1, \dots, M+1 \\ i-M-1 & , i = M+2, \dots, N \end{cases}$$

Diese Matrizen sind etwas gutartiger, aber auch bei ihnen kann es zum Verlust der Orthogonalität kommen. Man erhält für N=21, 42 ,... , 252 mit der Deflationsschranke $C*10^{-20}$ die folgenden Ergebnisse:

$$C*10^{-11} \leq JOrth \leq C*10^{-4}$$
$$C*10^{-18} \leq maxF \leq C*10^{-13}$$

3. Eigenwertprobleme mit J ungleich I

1.3.1 Zufallsmatrizen

Da ein mit Zufallszahlen erzeugtes Paar meist *nicht* positiv definit sein wird, muß man spezielle Konstruktionen entwickeln, die die Positiv-Definitheit gewährleisten Wir verwenden hier ein positiv definites Paar (A, J), wobei A eine spezielle Tridiagonalmatrix ist und J Diagonalmatrix mit +1 und —1 auf der Diagonalen. Dabei ist

$$A = L^T L + \epsilon I + \kappa J$$

mit I Einheitsmatrix und

$$L = \begin{pmatrix} a_1 & 0 & . & . & . & . \\ b_1 & a_2 & 0 & . & . & . \\ 0 & b_2 & a_3 & 0 & . & . \\ . & 0 & b_3 & . & . & . \\ . & . & . & . & . & 0 \\ . & . & . & 0 & b_{n-1} & a_n \end{pmatrix} \text{ untere Dreiecksmatrix}$$

und

$$a_i = \left| g_1 \left(2\,\xi_1^{(i)} - 1 \right) \right|,\ i = 1, ..., n$$
$$b_i = g_2 \left(2\,\xi_2^{(i)} - 1 \right),\ i = 1, ..., n-1$$

mit $\xi_1^{(i)}$, $\xi_2^{(i)}$ Folge (gleichverteilter) Zufallszahlen in [0, 1], $\epsilon, \kappa, g_1, g_2$ vorgegebene Zahlen.

Zufallsmatrizen	Dimension	JOrth	maxF	DefSumme	DefGSumme	Laufzeit (PC 386/87)
g1=1, g2=1 kappa =2 J=diag(+1,-1,+1,-1,...) eps=1E-2	N=128, N=256	$\leq C * 10^{-18}$ $\leq C * 10^{-17}$	$\leq C * 10^{-18}$ $\leq C * 10^{-17}$	50 283	5824 57792	400 sec 2404 sec
wie oben, aber: eps=1E-8	N=128, N=256	$\leq C * 10^{-16}$ $\leq C * 10^{-15}$	$\leq C * 10^{-17}$ $\leq C * 10^{-17}$	48 270	5568 54976	403 sec 2428 sec
wie oben, aber: eps=1E-16	N=128, N=256	$\leq C * 10^{-15}$ $\leq C * 10^{-14}$	$\leq C * 10^{-18}$ $\leq C * 10^{-17}$	48 270	5568 54976	401 sec 2427 sec
g1=2, g2=4 kappa=6 J=diag(+1,-1,+1,-1,...) eps=1E-2	N=128, N=256	$\leq C * 10^{-18}$ $\leq C * 10^{-17}$	$\leq C * 10^{-17}$ $\leq C * 10^{-16}$	81 367	9408 3392	359 sec 2276 sec
wie oben, aber: J=diag(-1,+1,-1,+1,...)	N=128, N=256	$\leq C * 10^{-18}$ $\leq C * 10^{-17}$	$\leq C * 10^{-17}$ $\leq C * 10^{-16}$	81 367	9408 3392	359 sec 2276 sec

1.3.2 Matrizenpaare aus der Physik

Es wird das Paar (A, J) erzeugt, wobei A spezielle Tridiagonalmatrix und J Diagonalmatrix mit *diag(J) = (+1, −1, +1, −1, ...)* ist. A hat dabei die folgende Gestalt:

$$A = \begin{pmatrix} 0 & l_1 & 0 & . & . & . & . & 0 \\ l_1 & d_1 & r_1 & 0 & . & . & . & . \\ 0 & r_1 & 0 & l_2 & . & . & . & . \\ . & 0 & l_2 & d_2 & r_2 & . & . & . \\ . & . & . & r_2 & . & . & . & . \\ . & . & . & . & . & . & r_{m-1} & 0 \\ . & . & . & . & . & r_{m-1} & 0 & l_m \\ 0 & . & . & . & . & 0 & l_m & d_m \end{pmatrix}$$

mit

$$l_j = (m+1)\sqrt{\frac{m-j+2}{m-j+1}},\quad r_j = (m+1)\sqrt{\frac{m-j}{m-j+1}}$$

und $m = N/2$ mit N Dimension von A. (Sinnvollerweise sollte N gerade sein.)

Dieses Matrizenpaar ergibt sich aus der Linearisierung eines quadratischen Eigenwertproblems, welches die gedämpften Schwingungen eines Systems von m Massen beschreibt.

Für $d = d_j$, $j = 1, ..., m$ ist die Dämpfung proportional, und die Eigenwerte sind exakt bekannt:

$$\lambda_{2j-1,2j} = \frac{1}{2}\left(-d \pm \sqrt{d^2 - 16(m+1)^2 sin^2 \frac{j\pi}{2(m+1)}} \right) \text{ mit } j = 1, ..., m$$

Für das Beispiel in der folgenden Tabelle wird eine Größe $a := 4(M+1) + 6$ bestimmt und die von null verschiedenen Diagonalelemente von A, die d_j, gleich a gesetzt. Das entstehende Eigenwertproblem hat dann nur reelle Eigenwerte. Durch eine andere Wahl von a können auch Eigenwertprobleme mit einem oder mehreren komplexen Eigenwerten erzeugt werden.

Die Ergebnisse zu diesen Beispielmatrizenpaaren sind sehr gut, wie man der folgenden Tabelle entnehmen kann:

Eingabe-unit	Dimension	JOrth	maxF	z-Deflations-konstante	Laufzeit auf einen PC486
a:= 4(K+1)+6 mit K:= N/2	300 500	$\leq C * 10^{-17}$	$\leq C * 10^{-16}$	3.29E-20	1291 sec. 3702 sec.

1.3.3 Wilkinson-Matrizen mit J ungleich I

Die Wilkinson-Matrizen *WPlus* gehören, wie wir bereits gesehen haben, zu den für das Divide-Verfahren besonders kritischen Matrizen, da es bei vielen Beispielen schon in den Dimensionen von 5 bis 73 zum Verlust der Orthogonalität zwischen Eigenvektoren kommt. Um zu testen, wie sich das Verfahren bei solch kritischen Matrizen für J ungleich I verhält, wurde die Matrix I an einer Stelle durch eine -1 abgeändert und erhält den Namen J. (Da schon die ursprüngliche Wilkinson-Matrix nicht positiv definit ist — sie hat einen negativen Eigenwert und sonst nur

positive Eigenwerte — gibt es auch Beispiele mit J ungleich I, die nicht mehr positiv definit sind.)

a) Beispiele mit J_{11} = -1 und Beispiele mit J_{NN} = -1:

Das Verhalten dieser Matrizenpaare gleicht dem der Beispiele im Fall J=I, die Ergebnisse sind für die verschiedenen Dimensionen unterschiedlich gut, aber es kann zum vollständigen Verlust der J-Orthogonalität kommen:

$$N = 127: \ maxF \leq C * 10^{-18}, \quad JOrth \leq C * 10^{-17}$$
$$N = 201: \ maxF \leq C * 10^{-10}, \quad JOrth \leq C * 10^{-16}$$

N = 101	**J = I**	**J[1,1] = -1**	**J[N,N] = -1**
JOrth	$C * 10^{-2}$	$C * 10^{-2}$	$C * 10^{-2}$
maxF	$C * 10^{-16}$	$C * 10^{-16}$	$C * 10^{-16}$
Deflationen	M = 64: 12 Defl. M= 37: 11 Defl. M = 101: 26 EW-Defl., 27 Defl.	M = 32: 1Defl. M = 64: 12 Defl. M = 37: 11 Defl. M = 101: 14 EW-Defl., 39 Defl.	M = 64: 12 Defl. M= 37: 11 Defl. M = 101: 14 EW-Defl., 39 Defl.
Laufzeit: PC386	180.32 sec	181.31 sec.	180.76 sec.

b) Alle Diagonalelemente von J sind 1, bis auf eines mit Wert -1:

Wir legen eine Dimension N fest und lassen die eine Minuseins alle Positionen auf der Diagonale von J durchlaufen. Wenn wir dabei eine Dimension N wählen, für die schon im Fall J=I die Orthogonalität fast verlorengeht, so erhalten wir für einige i kritische Eigenwerte im Definitheitsintervall, wenn eine oder beide Nullstellen mit den Polstellen, die die Grenzen des Intervalls bei der Nullstellensuche bilden, übereinstimmen. Die Übereinstimmung ist teilweise so groß, daß in der Prozedur zur Eigenvektorberechnung die Meldung "Division durch null" erscheint und Turbo Pascal die Rechnung abbricht. Wie im Fall J = I kann durch Wahl einer weniger strengen Deflationsschranke ein brauchbares Ergebnis erzielt werden.

4. Beispiele mit Datentyp Single

Wir hatten gesehen, daß es bei einigen ausgesuchten Beispielen zum Verlust der Orthogonalität bzw. der J-Orthogonalität kommen kann. Eine Möglichkeit, mit diesem Problem fertigzuwerden, ist die Verwendung von doppelter Rechengenauigkeit in den kritischen Rechenschritten der Nullstellensuche und Eigenvektorberechnung.

Dazu haben wir diese Beispiele mit Datentyp *single* berechnet und anschließend mit *single* und *double* bei den kritischen Rechenschritten.

Bei der Single-Double-Version unseres Algorithmus kam es in keinem dieser Beispiele mehr zu einem Verlust der J-Orthogonalität.

Zur Veranschaulichung sei hier das Beispiel von Wilkinson-Matrizen mit J ungleich I angegeben:

Beispiel	**single**	**single-double**
N=51: J hat bis auf eine -1 nur +1 auf der Diagonalen	$C * 10^{-7} \leq JOrth \leq C * 10^{-2}$ $C * 10^{-6} \leq maxF \leq C * 10^{-4}$	$JOrth \leq C * 10^{-7}$ $C * 10^{-6} \leq maxF \leq C * 10^{-5}$
N=7 bis 119, N ungerade, $JJ_{1,1}$ = -1	$C * 10^{-7} \leq JOrth \leq C * 10^{-2}$ $C * 10^{-7} \leq maxF \leq C * 10^{-3}$	$JOrth \leq C * 10^{-7}$ $C * 10^{-6} \leq maxF \leq C * 10^{-4}$

5. Annähern an die Indefinitheit und der nicht positiv definite Fall

Gegeben sei ein Paar (T, J) mit T Tridiagonalmatrix, deren Elemente "zufällig" erzeugt wurden. Man bestimmt zunächst Zahlen a_{unten} und a_{oben}, für die das Paar (T — a_{unten}*I, J) positiv definit ist und das Paar (T — a_{oben}*I, J) nicht.

Ausgehend vom Intervall [a_{unten}, a_{oben}] wird der Mittelpunkt a des aktuellen Intervalls berechnet und T — a*I gebildet. Dann wird das Paar (T —a*I, J) auf positive Definitheit untersucht und dem Ergebnis entsprechend das Intervall [a_{unten}, a_{oben}] modifiziert: Ist das Paar positiv definit, so setzt man a_{unten} := a, sonst a_{oben} := a.

Dies wiederholt man solange, bis die Intervallbreite eine gegebene Konstante unterschreitet. In der folgenden Tabelle bedeutet *Start-Def* das Definitheitsintervall zu Beginn der Iteration und *End-Def* das Intervall am Ende der Iteration.

	Definitheitsintervall	**N**	**JOrth**	**MaxF**
Start-Def	$[-1.9168 * 10^{-6}, 0.43]$	8	$C * 10^{-19}$	$C * 10^{-19}$
End-Def	$[0.219417484..., 0.219417494...]$		$C * 10^{-6}$	$C * 10^{-19}$
Start-Def	$[0.01133989, 0.001696683]$	16	$C * 10^{-19}$	$C * 10^{-19}$
End-Def	$[-0.007030721..., -0.0070307208...]$		$C * 10^{-7}$	$C * 10^{-17}$
Start-Def	$[-3.393 * 10^{-9}, 1.0112 * 10^{-4}]$	32	$C * 10^{-17}$	$C * 10^{-18}$
End-Def	$[2.005074 * 10^{-5}, 2.005089 * 10^{-5}]$		$C * 10^{-6}$	$C * 10^{-16}$
Start-Def	$[-4.13 * 10^{-6}, 5.67 * 10^{-8}]$	64	$C * 10^{-18}$	$C * 10^{-18}$
End-Def	$[-2.945538... * 10^{-6}, -2.945533... * 10^{-6}]$		$C * 10^{-6}$	$C * 10^{-16}$
Start-Def	$[-2.85 * 10^{-7}, 6.46 * 10^{-5}]$	72	$C * 10^{-17}$	$C * 10^{-17}$
End-Def	$[2.679924... * 10^{-5}, 2.679928... * 10^{-5}]$		$C * 10^{-9}$	$C * 10^{-15}$
Start-Def	$[-8.83 * 10^{-6}, 2.490 * 10^{-19}]$	100	$C * 10^{-15}$	$C * 10^{-18}$
End-Def	$[-5.7062352 * 10^{-6}, 5.7062351 * 10^{-6}]$		$C * 10^{-7}$	$C * 10^{-18}$
Start-Def	$[-1.26 * 10^{-9}, -7.12 * 10^{-21}]$	200	$C * 10^{-15}$	$C * 10^{-17}$
End-Def	$[-5.3418... * 10^{-11}, -5.3403 * 10^{-11}]$		$C * 10^{-11}$	$C * 10^{-17}$

Die Matrizenpaare aus der Physik aus 1.3.2 können gut als Beispiele für Tests auf positive Definitheit eines Matrizenpaares verwendet werden; je nach Wahl des Parameters a haben diese Paare nur reelle oder auch reelle und komplexe Eigenwerte, die man direkt angeben kann:

Für $a = 44 sin \frac{\pi}{22}$ sind sehr viele Eigenwerte des Problems komplex; der Divide-et-Impera-Algorithmus entdeckt für die Dimension N = 4 bis 72, N gerade, bereits bei den 2x2-Problemen, daß das gegebene Matrizenpaar nicht positiv definit ist.
Für $a = 4 * (\frac{N}{2} + 1) - 6$ sind nur einige Eigenwerte komplex:

2x2-Paar nicht pos.def.	kein Definitheitsintervall für 2x2-Paare	kein positiver Wert im Definitheitsintervall beim Teilproblem der Dim. M
N = 4	kein Beispiel	M = 4: für N = 6 bis 10, 16 M = 8: für N = 12, 14, 18 bis 42 M = 16: für N = 44 bis 72

Für verschiedene Dimensionen N wurden Beispiele gerechnet, bei denen a aus einem Intervall $[a_{komplex}, a_{reell}]$ gewählt wird mit $a := (a_{komplex}, a_{reell})/2$; für $a_{komplex}$ hat das zugehörige Eigenwertproblem auch komplexe Eigenwerte, für a_{reell} nur reelle. Je nachdem, ob für das so berechnete a auch komplexe Eigenwerte oder nur reelle vorliegen, ersetzt a dann einen der Randpunkte des Intervalls. Für das so bestimmte neue Intervall legt man wieder a fest, usw. Durch dieses "Intervallhalbierungsverfahren" entsteht ein immer kleineres Intervall für a, Eigenwertprobleme mit nur reellen Eigenwerten und die mit auch komplexen Eigenwerten nähern sich immer stärker an.

Das Divide-et-Impera-Verfahren entdeckt die Fälle mit komplexen Eigenwerten, je kleiner das Intervall für a wird, desto später, schließlich erst beim abschließenden NxN-Teilproblem. Das Definitheitsintervall für die Beispiele mit reellen Eigenwerten wird immer kleiner, man arbeitet sich immer mehr auf einen doppelten Eigenwert im Definitheitsintervall vor, dabei geht die J-Orthogonalität allmählich verloren (was auch nicht anders zu erwarten ist, da sich die Eigenwerte im Definitheitsintervall nicht mehr genügend unterscheiden, um J-orthogonale Eigenvektoren liefern zu können: $C * 10^{-18} \leq maxF \leq C * 10^{-13}$, JOrth steigt an bis zum Verlust der J-Orthogonalität).

Dieses Intervallhalbierungsverfahren endet, wenn sich Anfangs- und Endpunkt des Intervalls für a im Rahmen der möglichen Genauigkeit nicht mehr unterscheiden.

Bei Zufallsmatrizen *A* und *J=diag(+1,-1,+1,-1,...)* wird meist schon sehr früh entdeckt, ob das Paar *(A, J)* nicht positiv definit ist:

N	Versuchsanzahl	2x2-Paar nicht pos.def.	Definitheitsintervall ex. nicht für 2x2-Paare	kein positiver Wert im Definitheitsintervall beim Teilproblem der Dim. M
40	50	43 Beispiele	4 Beispiele	M=4: 3 Beispiele
72	50	46 Beispiele	4 Beispiele	

Bei den Beispielen mit fester Startzahl für den Zufallszahlengenerator und wachsender Dimension ergibt sich folgendes Verhalten:

2x2-Paar nicht pos.def.	kein Definitheitsintervall für 2x2-Paare	kein positiver Wert im Definitheitsintervall beim Teilproblem der Dim. M
N = 16 bis 72	kein Beispiel	M = 4: für die Dimensionen N = 4 bis 15

6. Ergebnisse für "große" Matrizen

Für die Dimensionen N = 16, 64, 256 wurden je 400 Beispiele mit Pseudozufallszahlen auf einer Sun-SPARC10 gerechnet , 160 Beispiele für N = 512 sowie 20 Beispiele für N = 1000 (man beachte, daß die Rechengenauigkeit geringer als beim PC ist, nämlich nur etwa 16 Stellen):

Dimension	JOrth	maxF	Defsumme D	DefGSumme DG	Zeit in sec.
N = 16	$C * 10^{-16} \leq JOrth \leq C * 10^{-13}$	$C * 10^{-16} \leq maxF \leq C * 10^{-12}$	0	0	$0 \leq Zeit \leq 1$
N= 64	$C * 10^{-15} \leq JOrth \leq C * 10^{-11}$	$C * 10^{-15} \leq maxF \leq C * 10^{-12}$	$1 \leq D \leq 25$	$64 \leq DG \leq 1600$	$2 \leq Zeit \leq 4$
N = 256	$C * 10^{-15} \leq JOrth \leq C * 10^{-12}$	$C * 10^{-15} \leq maxF \leq C * 10^{-11}$	$316 \leq D \leq 424$	$61728 \leq DG \leq 76416$	$93 \leq Zeit \leq 100$
N = 512	$C * 10^{-14} \leq JOrth \leq C * 10^{-11}$	$C * 10^{-14} \leq maxF \leq C * 10^{-11}$	$1103 \leq D \leq 1263$	$349568 \leq DG \leq 384544$	$806 \leq Zeit \leq 840$
N = 1000	$C * 10^{-14} \leq JOrth \leq C * 10^{-12}$	$C * 10^{-14} \leq maxF \leq C * 10^{-12}$	$3153 \leq D \leq 3358$	$1617552 \leq DG \leq 1654408$	$6290 \leq Zeit \leq 6323$

7. Konditionszahlen

Für eine diagonalisierbare Matrix A hängt die Empfindlichkeit der Eigenwerte gegenüber Störungen von der Matrix der Eigenvektoren ab, mit der A auf Diagonalgestalt gebracht wird (vgl. Satz von Bauer und Fike, etwa im Buch von Bunse-Gerstner [17]). Aus der Abschätzungen im Satz von Bauer und Fike leitet man die folgende Definition der Eigenwert-Konditionszahl einer diagonalisierbaren Matrix A ab:

$$K_2(A) := \min_{T \epsilon M^{n,n}(\mathbb{C}),\, TAT^{-1}\ diagonal} \|T\|_2 \|T^{-1}\|_2$$

Für normale Matrizen gilt $K_2(A) = 1$, da sie mit unitären Matrizen auf Diagonalgestalt transformiert werden können. (Statt der hier verwendeten euklidischen Norm können auch andere Normen verwendet werden.)

Bei dem in dieser Arbeit behandelten Eigenwertproblem $Ax = \lambda Jx$ (mit positiv definitem Paar (A, J)) sind die Eigenvektormatrizen nicht mehr orthogonal, sondern nur noch J-orthogonal, d.h. für eine Eigenvektormatrix gilt: $Q^T JQ = J$. Nach dem Korollar A.2 aus [63] haben die das Paar (A, J) diagonalisierenden, J-orthogonalen Matrizen Q alle dieselbe euklidische Konditionszahl

$$K_2(A,J) = ||Q||_2||Q^{-1}||_2 = ||Q||_2||JQ^TJ||_2 \leq ||Q||_2^2$$

und für alle anderen diagonalisierenden Matrizen ist die Kondition schlechter, vgl. [65].

Die Berechnung von $K_2(A,J)$ ist recht aufwendig, aber wir können stattdessen die folgende Abschätzung für die euklidische Kondition von Q verwenden (vgl. Veselić [63], Theorem A.3):

$$||Q||_2||Q^{-1}||_2 \leq K_{cond}(Q) := n + 2\frac{Tr\,A - \left(\sum_{i=1}^{m}\lambda_i^+ - \sum_{i=1}^{n-m}\lambda_i^-\right)}{Laenge\ des\ Definitheitsintervalls}$$

Unser Divide-et-Impera-Verfahren teilt das Ausgangsproblem bekanntlich in Teilprobleme auf und setzt diese schrittweise wieder zusammen. Dabei müssen Eigenwertprobleme der Dimensionen M ≤ N gelöst werden. Für die zugehörigen Eigenvektormatrizen Q berechnen wir $||Q||_1^2 = \max||q_i||^2$ und $K_{cond}(Q)$ und bestimmen jeweils den maximalen Wert (*maximal ...*) davon. Außerdem merken wir uns das Ergebnis für den abschließenden Fall M = N (*final ...*). Die Quotienten

$$\zeta_E := \frac{maximal\ K_{cond}(Q)}{final\ K_{cond}(Q)}, \quad \zeta_1 := \frac{maximal\ ||Q||_1^2}{final\ ||Q||_1^2}.$$

stellen dann ein Maß für die Stabilität der Eigenvektoren da (vgl. Veselić [63]). (Sie zeigen an, wenn im Verlauf des Verfahrens in einem Zwischenschritt die Kondition größer wird als die Kondition der abschließend berechneten Eigenvektormatrix; ein solches Verhalten deutet auf mangelnde Stabilität des Algorithmus zur Eigenwert-Eigenvektor-Berechnung hin. Bei einem guten Verfahren sollten die obigen Werte in der Größenordnung von 1 sein, da der größte Wert eigentlich im NxN-Schritt angenommen werden müßte.)

Weiter bestimmen wir das Maximum über alle Eigenvektoren der Teilprobleme des Quotienten aus J-Norm und euklidischer Norm

$$\eta := \max_{Eigenvektor\ q} \frac{q^TJq}{q^Tq}.$$

Beispiel	ζ_E	ζ_1	η
Zufallsmatrizen			
N = 100:	1.1864	1	1
N = 200:	1.3236	1.042	1
Wilkinsonmatrizen mit J <> I N = 100: (kritisches Beispiel)	8	1.00000004	1
Matrizenpaare aus der Physik			
N = 100:	1	1.284	1
N = 200:	1	1.396	1

2. Ergebnisse für den verallgemeinerten Wilkinson-Algorithmus

Das verallgemeinerte Wilkinson-Verfahren haben wir für Matrizenpaare (A, J) getestet, bei denen die Matrix A eine Diagonalmatrix mit einem oder mit drei Rändern ist. Positiv definite Beispiele erhalten wir zum Beispiel, wenn wir

$$A = HH^T + \varepsilon I + \kappa J, \ \varepsilon > 0, \ \kappa\epsilon\Re\backslash\{0\}, \quad J = diag(+1,-1,+1,-1,...)$$

bilden, wobei die Matrix H eine Diagonalmatrix mit einem oder drei Rändern ist, deren von null verschiedene Elemente Zufallszahlen sind.

	N	**JOrth**	**maxF**
ein Rand an Position1, $\kappa = 2$, $\epsilon = 0.01$	5 bis 69	$JOrth \leq C * 10^{-18}$	$maxF \leq C * 10^{-18}$
ein Rand an Position 2 $\kappa = 2$, $\epsilon = 0.01$	5 bis 69	$JOrth \leq C * 10^{-18}$	$maxF \leq C * 10^{-18}$
drei Ränder, an den Positionen 1, 2, 3 $\kappa = 2$, $\epsilon = 0.01$	5 bis 69 ohne 50, 59, 64-67, 69 50, 59, 64-67, 69	$JOrth \leq C * 10^{-18}$	$maxF \leq C * 10^{-18}$ $maxF \leq C * 10^{-17}$

3. Ergebnisse für den zusammengesetzten Divide-et-Impera-Algorithmus

Bei den Beispielen für das zusammengesetzte Divide-et-Impera-Verfahren haben wir uns auf die Behandlung von berandeten Tridiagonalmatrizen mit *einem* Rand beschränkt; die wesentlichen Ideen des Programms werden bereits für diesen einfachen Fall deutlich, und eine Realisierung auf dem PC ist trotz einiger Speicherplatzproblemen gerade noch möglich.

Das zusammengesetzte Divide-et-Impera-Verfahren liefert sehr gute Ergebnisse, wenn man es auf positiv definite Paare (A_{TR}, J_{TR}) mit A_{TR} berandete Tridiagonalmatrix mit einem Rand, J_{TR} Diagonalmatrix mit +1 und —1 auf der Diagonalen anwendet. Als typische Testpaare wurden die folgendermaßen konstruierten Matrizenpaare (A_{TR}, J_{TR}) ausgewählt:

$$A = H * H^T + \varepsilon * I + \kappa * J, \ \kappa\epsilon\Re\backslash\{0\}, \ \varepsilon > 0,$$

mit H Bidiagonalmatrix mit Rand, z.B. für N = 5 und Rand an Position 1:

$$H = \begin{pmatrix} d_1 & b_2 & b_3 & b_4 & b_5 \\ 0 & d_2 & c_3 & 0 & 0 \\ 0 & 0 & d_3 & c_4 & 0 \\ 0 & 0 & 0 & d_4 & c_5 \\ 0 & 0 & 0 & 0 & d_5 \end{pmatrix}$$

Die Elemente von H sind Pseudozufallszahlen aus (-1,+1), $J = diag(+1,-1,+1,...)$.
Man erhält folgende Ergebnisse:
a) A_{TR} Tridiagonalmatrix mit Rand an Position 1, $J = diag(+1,-1,+1,...)$:

$$JOrth \leq C * 10^{-18}, \quad maxF \leq C * 10^{-17}$$

b) A_{TR} Tridiagonalmatrix mit Rand an Position N div 2, $J = diag(+1,-1,+1,...)$:

$$JOrth \leq C * 10^{-17}, \quad maxF \leq C * 10^{-17}$$

Besonders interessant sind Eigenwertprobleme (A_{TR}, J_{TR}), die durch *teilweise* Tridiagonalisierung eines Paares (A, J) mit A symmetrische, vollbesetzte Matrix und J = diag($\pm$1) entstehen. Für unsere Tests wurden solche positiv definiten Paare (A, J) folgendermaßen konstruiert:
Zunächst erzeugt man eine beliebige symmetrische Matrix $M = (m_{ij})_{i,j=1,...,N}$, berechnet $Mmax := \max_{i,j} |m_{ij}|$ und definiert A':= M + λ*I mit $\lambda := n * Mmax + \varepsilon$, $\varepsilon > 0$ beliebig klein.
Die so erzeugte Matrix A' ist positiv definit. Um ein positiv definites Matrizenpaar zu erhalten, bei dem A im allg. nicht positiv definit ist, setzt man $A := A' + \beta * J$, $\beta \epsilon \Re \backslash \{0\}$, $J = diag(I, -I)$ mit Dimension von I gleich

$$dim = \begin{cases} \frac{N-1}{2} + 1, & N ungerade \epsilon \\ \frac{N}{2}, & N gerade \epsilon \end{cases}.$$

(Bei den unten beschriebenen Beispielen wird ϵ = 0.01 und β = 2 gewählt.)
Durch Änderung des ersten Konstruktionsschrittes des Matrizenpaares (A, J) bekommt man Beispiele, die während des Tridiagonalisierungsvorganges für die erste Zeile und Spalte einen kritischen Wert annehmen :

$$|\alpha| := \left| \sum_{i=2}^{N} a_{1i}^2 * j_i \right| \approx 0, \quad aber\ a_{1i} \neq 0, \ i = 2, ..., N$$

Man erzeugt einfach eine beliebige symmetrische Matrix M mit

$$\alpha_M := \sum_{i=2}^{N} m_{1i}^2 * j_i \approx 0.$$

Dann gilt auch für die Endmatrix A , daß α ungefähr gleich 0 ist.
Bei solchen Matrizen ist es sinnvoll, die erste Zeile und Spalte nicht zu transformieren; der Tridiagonalisierungsprozeß erzeugt dann eine Tridiagonalmatrix mit Rand an Position 1.
Die Frage ist nun, wann man α als fast null ansieht. Beispiele zeigen, daß bereits α-Werte der Größenordnung 10^{-10} äußerst kritisch sind, selbst ein α der Größenordnung 0.01 produziert noch bemerkbare Rechenungenauigkeiten. Um ganz sicher zu gehen, sollte man Zeilen und Spalten nicht mehr während des Tridiagonalisierungsprozesses transformieren, wenn $\alpha \leq 0.5$ ist. (In den folgenden Beispielen wurde 0.6 als Schranke gewählt.)

1. Zufallsmatrizen mit $\alpha_M = C*10^{-10}$:

Bringt man das Paar (A, J) auf Tridiagonalgestalt, so sind bereits die Ergebnisse des Tridiagonalisierungsprozesses so schlecht, daß eine vollständige Tridiagonalisierung nicht sinnvoll erscheint

(z.B. gilt für N = 20, wenn wir mit H die Transformationsmatrix für die Tridiagonalisierung bezeichnen:

$$Jmaxfehler := \max_{i,j} \left| \left(H^T J_{TR} H - J\right)_{ij} \right| = C * 10^{-7}$$

$$Amaxfehler := \max_{i,j} \left| \left(H^T A_{TR} H - A\right)_{ij} \right| = C * 10^{8}$$

Führt man nur eine teilweise Tridiagonalisierung durch, so daß die erste Zeile und Spalte von A nicht transformiert wird (eine Zeile und die zugehörige Spalte bleibe erhalten, wenn das zugehörige $\alpha \leq 10^{-5}$ ist), so erhält man berandete Tridiagonalmatrizen mit einem Rand an Position 1.

	(relativ) gute Ergebnisse	"erträgliche" Ergebnisse	schlechte Ergebnisse
Dimension N	5-10, 12, 13, 19-23, 29, 31-37, 41	14, 28, 40, 50	11, 24, 25, 27, 30, 38, 39, 42-49
Tridiagonalisierung	$Amaxfehler \leq C*10^{-11}$ $JJmaxfehler \leq C * 10^{-17}$	$Amaxfehler \leq C * 10^{-13}$ $JJmaxfehler \leq C * 10^{-17}$	$C * 10^{-13} \leq Amaxfehler \leq C * 10^{-7}$ $C * 10^{-18} \leq JJmaxfehler \leq C * 10^{-14}$
zusammengesetztes Divide-et-Impera	$maxf \leq C * 10^{-11}$ $JOrth \leq C * 10^{-13}$	$maxf \leq C * 10^{-10}$ $JOrth \leq C * 10^{-14}$	$C * 10^{-10} \leq maxf \leq C * 10^{-5}$ $C * 10^{-15} \leq JOrth \leq C * 10^{-11}$
Gesamtprobe	$\leq C * 10^{-11}$	$= C * 10^{-10}$	$C * 10^{-9} \leq G \leq C * 10^{-5}$

Das schlechteste Ergebnis ergibt sich für N = 42:

Tridiagonalisierung	**zusammengesetztes Divide-et-Impera-V.**	**Gesamtprobe**
$Amaxfehler = C * 10^{-8}$	$maxF = C * 10^{-5}$	$C * 10^{-5}$
$JJmaxfehler = C * 10^{-14}$	$JOrth = C * 10^{-11}$	

Eine Erklärung dafür findet man, wenn man sich die Größen α in den einzelnen Rechenschritten ansieht: dabei befinden wir uns im Schritt Nr. i, wenn die Transformation der i-ten Zeile und Spalte ansteht; aufgelistet werden hier nur die "interessanten" Schritte:
Schritt 1: $\alpha = C * 10^{-10}$: Diese Zeile und Spalte werden *nicht* transformiert
Schritt 6: $\alpha = C * 10^{-4}$: Durchführung der Transformation
Schritt 26: $\alpha = C * 10^{-3}$: Durchführung der Transformation
Schritt 28: $\alpha = C * 10^{-3}$: Durchführung der Transformation
Schritt 30: $\alpha = C * 10^{-3}$: Durchführung der Transformation

Treten Größen α der Größenordnung $C * 10^{-4}$ oder auch $C * 10^{-3}$ während des Tridiagonalisierungsprozesses auf, so muß bereits mit relativ ungenauen Endergebnissen gerechnet werden.

2. Zufallsmatrizen mit α_M= C*0.001:

Die so erzeugten Beispiele sind ebenfalls kritische Beispiele, wenn man (A, J) vollständig tridiagonalisiert; zwar liegt die kritische Größe für α für die erste Zeile und Spalte nur bei C*0.001, aber im dritten Schritt zur Transformation der dritten Zeile und Spalte treten je nach Dimension α–Werte der Größenordnung $C * 10^{-5}$ oder gar $C * 10^{-6}$ auf.

Schon ab der Dimension N = 14 sind die Tridiagonalmatrizen A_{TR} stark verfälscht; das verallgemeinerte Divide-et-Impera-Verfahren kann von diesen ungenauen Daten ausgehend keine guten Eigenwert- und Eigenvektornäherungen mehr berechnen.

Um gute Ergebnisse bei der teilweisen Tridiagonalisierung zu erhalten, verzichten wir jetzt auf die Transformation einer Zeile und Spalte, wenn $\alpha \leq 0.01$ ist. Dies hat nicht nur Auswirkungen auf die erste Zeile und Spalte von A_{TR}, sondern auch auf die weiteren Schritte:

Dazu schaue man sich z.B. das Beispiel mit N = 19 an:

	vollständige Tridiagonalisierung	**teilweise Tridiagonalisierung: keine Transformation für $\alpha \leq 0.01$:**
Schritt 1:	$\alpha \approx -4.001 * 10^{-3}$	$\alpha \approx -4.001 * 10^{-3}$
Schritt 3:	$\alpha \approx 1.027 * 10^{-5}$	—
Schritt 13:	$\alpha \approx -5.558 * 10^{-3}$	—
Schritt 15:	$\alpha \approx -8.72 * 10^{-2}$	—
Schritt 17:	$\alpha \approx 5.27 * 10^{-3}$	—
	$Amaxfehler = C * 10^{-7}$	$Amaxfehler = C * 10^{-15}$
	$JJmaxfehler \leq C * 10^{-14}$	$JJmaxfehler \leq C * 10^{-18}$

Durch Verzicht auf die Transformation der ersten Zeile und Spalte werden auch die folgenden Transformationen für die restliche Matrix weniger kritisch.

Welche Ergebnisse ergeben sich, wenn man Zeile und Spalte während des Tridiagonalisierungsprozesses stehen läßt, wenn $\alpha < 0.01$ ist?

Jetzt kann es passieren, daß mehrere Zeilen und Spalten nicht transformiert werden; dies ist der Fall für N = 11, 14, 24, 25, 27, 28, 29, 30, 32, 34, 38, 39, 40, 42, 43, 45–48.

	(relativ) gute Ergebnisse	**erträgliche Ergebnisse**	**schlechte Ergebnisse**
Dimension N	5-10, 12,13, 15-23, 26, 31, 33, 35-37,41	50	44, 49
Tridiagonalisierung	$Amaxfehler \leq C*10^{-13}$ $JJmaxfehler \leq C * 10^{-17}$	$Amaxfehler \leq C * 10^{-13}$ $JJmaxfehler = C * 10^{-18}$	$Amaxfehler = C*10^{-12}$ $JJmaxfehler = C * 10^{-17}$

zusammengesetztes Divide-et-Impera-V.	$maxf \leq C * 10^{-11}$ $JOrth \leq C * 10^{-15}$	$maxf = C * 10^{-10}$ $JOrth = C * 10^{-14}$	$maxf = C * 10^{-9}$ $JOrth \leq C * 10^{-14}$
Gesamtprobe	$C * 10^{-17} \leq G \leq C * 10^{-11}$	$= C * 10^{-10}$	$= C * 10^{-9}$

3. Zufallsmatrizen:

Die Matrix A wurde aus einer Matrix M mit Pseudozufallszahlen als Elementen so konstruiert, daß das Paar (A, J) positiv definit ist. Bei diesen Beispielen würde man zunächst erwarten, daß der Fall eines kleinen α während des Tridiagonalisierungsprozesses so gut wie nie auftritt.

Sehr kleine α—Werte (α kleiner C*10^{-10}) sind in der Tat unwahrscheinlich, aber α—Werte der Größenordnung C*0.01 und C*0.001 sind relativ häufig, für die Dimension N = 33 erhalten wir $\alpha = 10^{-5}$ im 5. und 17. Transformationsschritt. Dies hat natürlich Auswirkungen auf die Genauigkeit der Lösungen des Eigenwertproblems.

Für N gleich 31 und 41 sieht man die unterschiedlichen Ergebnisse, je nachdem, ob eine vollständige Tridiagonalisierung durchgeführt wurde oder bei $\alpha < 0.01$ oder schon bei $\alpha < 0.6$ auf Transformationen verzichtet wurde. (Mit *RandNR* bezeichnen wir die Position des Randes.)

N = 31	**vollständige Tridiagonalisierung**	**Abbruch bei 0.01** RandNr. = 27	**Abbruch bei 0.6** RandNR. = 5
AmaxF.	$C * 10^{-15}$	$C * 10^{-15}$	$C * 10^{-16}$
JJmaxF.	$C * 10^{-17}$	$C * 10^{-17}$	$C * 10^{-18}$
maxF	$C * 10^{-13}$	$C * 10^{-14}$	$C * 10^{-15}$
JOrth	$C * 10^{-15}$	$C * 10^{-15}$	$C * 10^{-16}$
GesamtF.	$C * 10^{-12}$	$C * 10^{-14}$	$C * 10^{-15}$

N = 41	**vollständige Tridiagonalisierung**	**Abbruch bei 0.01** RandNR. = 5	**Abbruch bei 0.6** RandNR. = 3
AmaxF.	$C * 10^{-10}$	$C * 10^{-10}$	$C * 10^{-16}$
JJmaxF.	$C * 10^{-16}$	$C * 10^{-16}$	$C * 10^{-18}$
maxF	$C * 10^{-10}$	$C * 10^{-10}$	$C * 10^{-15}$
JOrth	$C * 10^{-14}$	$C * 10^{-13}$	$C * 10^{-16}$
GesamtF.	$C * 10^{-10}$	$C * 10^{-10}$	$C * 10^{-15}$

N = 70	**vollständige Tridiagonalisierung**	**Abbruch bei 0.1** RandNr. = 5
AmaxF.	$C * 10^{-12}$	$C * 10^{-13}$
JJmaxF.	$C * 10^{-17}$	$C * 10^{-17}$

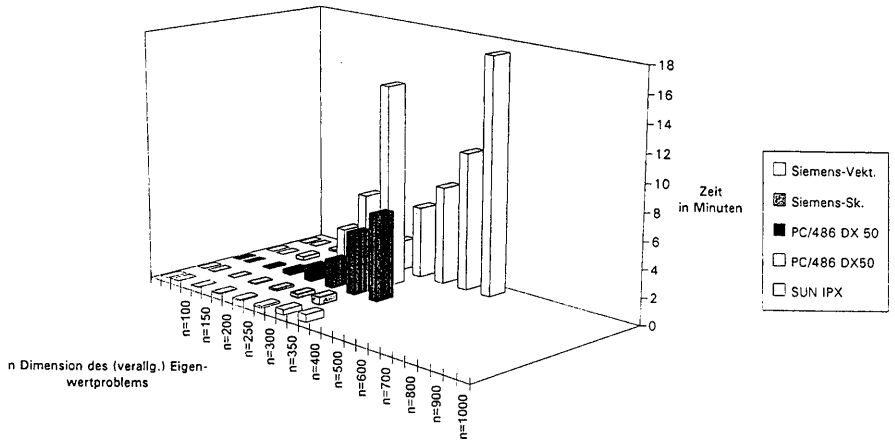
maxF	$C * 10^{-10}$	$C * 10^{-14}$
JOrth	$C * 10^{-15}$	$C * 10^{-14}$
GesamtF.	$C * 10^{-10}$	$C * 10^{-13}$

N = 80	**vollständige Tridiagonalisierung**	**Abbruch bei 0.1** RandNr. = 5
AmaxF.	$C * 10^{-11}$	$C * 10^{-14}$
JJmaxF.	$C * 10^{-16}$	$C * 10^{-18}$
maxF	$C * 10^{-10}$	$C * 10^{-13}$
JOrth	$C * 10^{-14}$	$C * 10^{-16}$
GesamtF	$C * 10^{-10}$	$C * 10^{-13}$

Vergleicht man das Laufzeitverhalten bei vollständiger Tridiagonalisierung und anschließendem verallgemeinerten Divide-et-Impera-Verfahren mit dem bei teilweiser Tridiagonalisierung und zusammengesetztem Divide-et-Impera-Verfahren, so erhält man etwa:

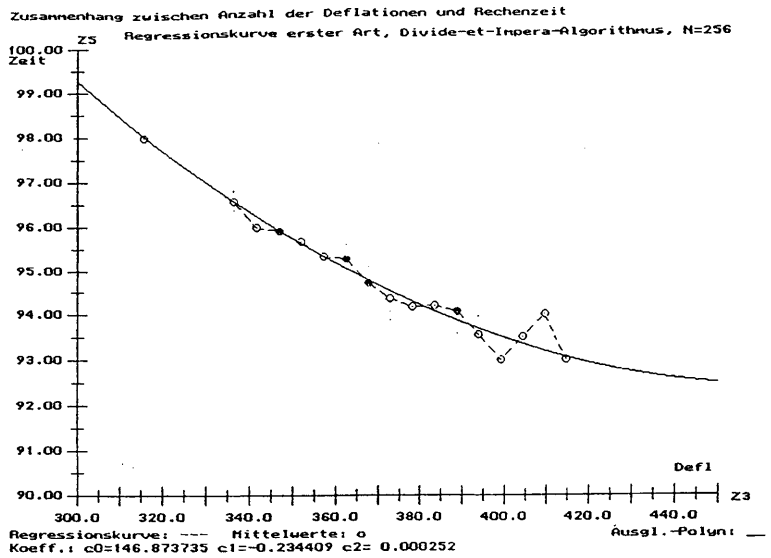
N	**vollständige Tridiagonalisierung**	**teilweise Tridiagonalisierung** (1 Zeile und Spalte bleibt stehen)	**Divide-et-Impera**	**zusammen-gesetztes Divide-et-Impera**
40	58	56	24	42
60	289	280	55	115
80	911	893	151	239
100	2215	2171	242	556

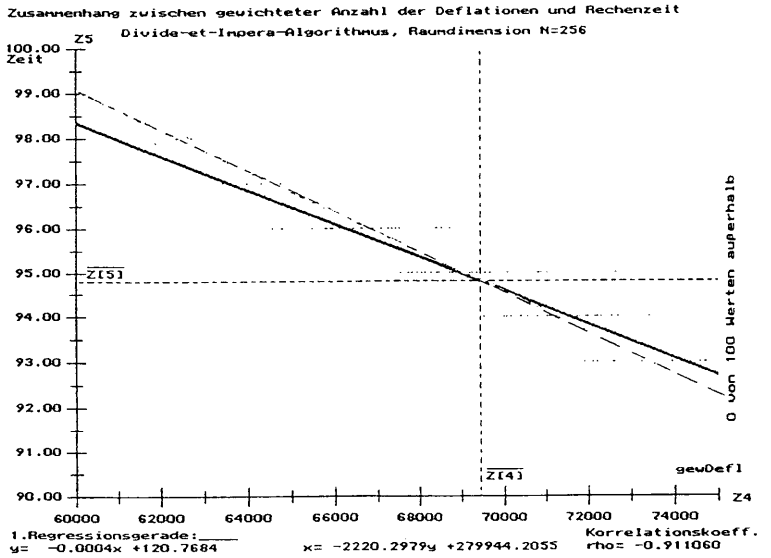
Laufzeitverhalten



Laufzeitverhalten

	PC/486 DX50 (Turbo Pascal) (extended)	PC/486 DX 50 (Laufzeit optimiert) (GNU C double)		SUN IPX (user = real) (GNU C double)		Siemens-Sk. (real/user) (C real)	Siemens-Vekt. (real/user) (C / VP real)
n=100	0:28,17 min	0:10 min		0:13,2 min		0:02,5 min/0:50 min	0:02 min/ 0:21,65 min
n=150	1:54,68 min	0:37 min		0:44 min		0:07,7 min/2:22 min	0:04,44 min/ 0:50,9 min
n=200	3:30,47 min	1:11 min		1:18 min		0:15 min/5:44 min	0:08 min/ 1:20,66 min
n=250	6:26 min	2:06 min		2:17 min		0:27 min/8:22 min	0:14,2 min/2:26 min
n=300	14:59,24 min	4:58 min		5:21 min		0:56 min/13:00 min	0:39,74 min/5:42 min
n=350		6:37 min		7:13 min			0:57,3 min/6:33 min
n=400				10:01 min			
n=500				17:31 min			
n=600							
n=700							
n=800							
n=900							
n=1000							





Überblick über die Computerprogramme

Aufbau der Computerprogramme

Wir geben einen kurzen Überblick über den Aufbau des verallgemeinerte Divide-et-Impera-Programms, des Programms zum verallgemeinerten Wilkinson-Algorithmus und zum zusammengesetzten Divide-et-Impera-Verfahren.

1. Das verallgemeinerte Divide-et-Impera-Verfahren

Das verallgemeinerte Divide-et-Impera-Verfahren
Berechnet werden im Fall eines positiv definiten Paares (A,J) alle Eigenwerte und Eigenvektoren, andernfalls erhält man die Meldung: "Paar nicht positiv definit".

Eingabe des Matrizenpaares (A, J):
A ist Tridiagonalmatrix, J Diagonalmatrix mit Eins und Minuseins auf der Diagonalen.
Eingelesen werden die Dimension N, die Diagonale von A, die Nebendiagonale von A, die Diagonale von J

Divide-Schritt:
Abspalten der Rang-Eins-Störungen
Lösen der 2- und 1-dimensionalen Teilprobleme

Impera-Schritt:
Zusammenfassen der Teilprobleme zu größeren Einheiten:
1. *Durchführung der Deflation*
—Eigenwertdeflation:
verallgemeinerte Householdertransformationen
— Deflation nach kleinen z-Komponenten
2. *Berechnung der Eigenwerte und Eigenvektoren*
—Shift-Strategien
—Suche nach einem positiven Funktionswert im Definitheitsintervall
—Ersteingrenzen
—quadratisches Eingrenzungsverfahren
—Nullstellensuche mit Pegasus
—bei kritischen Beispielen Nullstellensuche mit Bisektion
3. Berechnung des Eigenvektors zu jeder gefundenen Nullstelle
4. *Probe auf J-Orthogonalität:*
$JOrth := \max_{i,j} |(Q^T J Q - J)[i,j]|$

Probe:
Man berechne den maximalen Fehler $\max_{(\lambda,x)} ||Ax - \lambda Jx||_{max}$

2. Das verallgemeinerte Wilkinson-Verfahren

Der verallgemeinerte Wilkinson-Algorithmus
Berechnet werden im Fall eines positiv definiten Paares (A,J) alle Eigenwerte und Eigenvektoren, andernfalls erhält man die Meldung: "Paar nicht positiv definit."

Eingabe des Matrizenpaares (A, J):
A ist Diagonalmatrix mit maximal drei Rändern, J Diagonalmatrix mit +1, -1 auf der Diagonalen.
Eingelesen werden die Dimension N, die Positionen der Ränder, die Diagonale von A, die Ränder, die Diagonale von J

Hauptprogramm:
Beginnend mit der kleinsten berandeten Diagonalmatrix in der rechten unteren Ecke der Matrix A und der zugehörigen J-Matrix lösen wir nacheinander Eigenwertprobleme für **berandete Diagonalmatrizen**, bis alle Eigenwerte und Eigenvektoren für das Gesamtproblem bestimmt sind oder ein Teilproblem nicht mehr positiv definit ist.
Das Hauptprogramm ruft dazu die Prozeduren **Stoerung** und **Update** auf.

Probe:
Man berechne $\max_{(j,x)} ||Ax - \lambda Jx||_{max}$

In der Prozedur **Stoerung** werden — bis auf Punkt 1.1 — dieselben Schritte wie beim Impera-Schritt des verallgemeinerten Divide-et-Impera-Verfahrens durchgeführt, allerdings mit einer anderen rationalen Funktion zur Nullstellensuche, was sich z.B. deutlich auf die Eingrenzungsverfahren auswirkt.

Die Prozedur Stoerung: Lösen eines verallgemeinerten Eigenwertproblems für eine berandete Diagonalmatrix A und J Diagonalmatrix mit +1 und -1 auf der Diagonalen

1. *Durchführung der Deflation*
—Eigenwertdeflation
verallgemeinertes Gram-Schmidt-Orthogonalisierungsverfahren zur J-Orthogonalisierung von Eigenvektoren
—Deflation nach kleinen z-Komponenten
2. *Berechnung der Eigenwerte als Nullstellen einer rationalen Funktion, Berechnung der Eigenvektoren*
3. *Probe auf J-Orthogonalität*

Die Prozedur Update
Die Matrix A wird im Verlauf des Programms modifiziert. Am Anfang enthält sie die Diagonalmatrix mit drei Rändern. Nach der Lösung des ersten Teilproblems wird A zu einer Diagonalmatrix mit nur noch zwei Rändern, dann zu einer Matrix mit einem Rand. Bei diesen Transformationen werden auch die Ränder verändert.

3. Das zusammengesetzte Divide-et-Impera-Verfahren

Dieses Verfahren besteht aus der Kombination des verallgemeinerten Divide-et- Impera-Algorithmus und der verallgemeinerten Wilkinson-Methode.Der Impera-Schritt im folgenden Fall 3 entspricht dem Impera-Schritt des verallgemeinerten Divide- et- Impera-Verfahrens. Die Probe berechnet wie vorher den maximalen Fehler bei der Eigenwert-Eigenvektorbestimmung.

Das zusammengesetzte Divide-et-Impera-Verfahren
Berechnet werden im Fall eines positiv definiten Paares (A,J) alle Eigenwerte und Eigenvektoren, andernfalls erhält man die Meldung: "Paar nicht positiv definit".

Eingabe des Matrizenpaares (A, J):
A ist Tridiagonalmatrix, berandete Tridiagonalmatrix oder Ineinanderschachtelung von berandeten Tridiagonalmatrizen (realisiert wurde hier nur der Fall mit einem Rand)

1. Fall: A Tridiagonalmatrix:
Aufruf des verallgemeinerten Divide-et-Impera-Verfahrens

2. Fall: A berandete Tridiagonalmatrix:
2.1 *Aufruf des verallgemeinerten Divide-et-Impera-Verfahrens für die innere Tridiagonalmatrix und die zugehörige J-Matrix*
2.2 *Modifikation des Randes*
2.3 *Aufruf des verallgemeinerten Wilkinson-Verfahrens für die entstandene berandete Diagonalmatrix und die zugehörige J-Matrix*
2.4 *Berechnung der Eigenvektormatrix für das Gesamtproblem.*

3. Fall: A Ineinanderschachtelung von berandeten Tridiagonalmatrizen (ein Rand):
3.1 *Abspalten einer Rang-Eins-Störung und Zerlegen des Problems in zwei Teilprobleme: Das erste Teilproblem ist ein verallgemeinertes Eigenwertproblem für eine Tridiagonalmatrix und J, das zweite Teilproblem ist eines für eine berandete Tridiagonalmatrix und J.*
3.2 *Lösen des ersten Teilproblems:*
Aufruf des verallgemeinerten Divide-et-Impera-Verfahrens
3.3 *Lösen des zweiten Teilproblems:*
Man führe die Schritte aus **Fall 2** *aus.*
3.4 *Impera-Schritt: Zusammenfassen der Ergebnisse der beiden Teilprobleme*

Probe
wird für alle drei Fälle durchgeführt

Literaturverzeichnis

[1] L. Adams and P. Arbenz, *Towards a Divide and Conquer Algorithm for the Real Nonsymmetric Eigenvalue Problem*, Bericht der ETH Zürich, Abteilung Informatik, Institut für Wissenschaftliches Rechnen, (1991).

[2] P. Arbenz, *Divide and Conquer Algorithms for the Bandsymmetric Eigenvalue Problem*, Bericht der ETH Zürich, Abteilung Informatik, Institut für Wissenschaftliches Rechnen, (1991).

[3] P. Arbenz, W. Gander, and G. Golub, *Restricted Rank Modification of the Symmetric Eigenvalue Problem: Theoretical Considerations*, Linear Algebra Appl., 104 (1988), pp. 75–95.

[4] P. Arbenz and G. Golub, *On the Spectral Decomposition of Hermitian Matrices Modified by Low Rank Perturbations with Applications*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 40–58.

[5] J. Barlow, *Error Analysis of Update Methods for the Symmetric Eigenvalue Problem*, Pennsylvania State University, Bericht des Department of Computer Science, PA 16802, (1991).

[6] J. Barlow and J. Demmel, *Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices*, Technical Report 421, (1988).

[7] K.-J. Bathe and E. L. Wilson, *Solution Methods for Eigenvalue Problems in Structural Mechanics*, International Journal for Numerical Methods in Engineering, 6 (1973), pp. 213–226.

[8] C. Beattie and C. Ribbens, *Parallel Solution of a Generalized Symmetric Eigenvalue Problem*. Department of Mathematics, Department of Computer Science, Virginia Polytechnic Institute & State University, Blacksburg.

[9] A. Berman and A. Ben-Israel, *A Note on Pencils of Hermitian or Symmetric Matrices*, SIAM J. Appl. Math., 21 (1971).

[10] D. Bini, P. Pan, and V.Pan, *Practical Improvement of the Divide-and-Conquer Algorithms*, Computing, 48 (1992), pp. 109–123.

[11] M. Borri and P. Mantegazza, *Efficient Solution of Quadratic Eigenproblems Arising in Dynamic Analysis of Structures*, Computer Meth. Appl. Mech. Eng., 12 (1977), pp. 19–31.

[12] M. Brebner and J. Grad, *Eigenvalues of $Ax=\lambda Bx$ for Real Symmetric Matrices A and B Computed by Reduction to a Pseudosymmetric Form and the HR Process*, Linear Algebra and its Applications, 43 (1982), pp. 99–118.

[13] J. Bunch, G. Nielsen, and D. Sorensen, *Rang-One Modification of the Symmetric Eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[14] A. Bunse-Gerstner, *Der HR-Algorithmus zur numerischen Berechnung der Eigenwerte einer Matrix*, PhD thesis, Fakultät für Mathematik der Universität Bielefeld, 1978.

[15] ——, *On the Similarity Transformation to Tridiagonal Form*, Linear and Multilinear Algebra, 12 (1982), pp. 51–56.

[16] ——, *An Algorithm for the Symmetric Generalized Eigenvalue Problem*, Linear Algebra and its Applications, 58 (1984), pp. 43–68.

[17] A. Bunse-Gerstner and W. Bunse, *Numerische lineare Algebra*, Teubner Studienbücher Mathematik, 1985.

[18] K.-W. E. Chu, *Exclusion Theorems and the Perturbation Analysis of the Generalized Eigenvalue Problem*, Siam J. Numer. Anal., 24 (1987).

[19] J. Cuppen, *A Divide and Conquer Method for the Symmetric Tridiagonal Eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[20] A. Dax and S. Kaniel, *Pivoting Techniques for Symmetric Gaussian Elimination*, Numer. Math., 28 (1977), pp. 221–242.

[21] J. Demmel and K. Veselić, *Jacobi's Method is More Accurate than QR*. Courant Institute preprint 1989, to appear in SIAM J. Math. Anal. Appl.

[22] J. Dongarra, S. Hammarling, and D. Sorensen, *Block Reduction of Matrices to Condensed Form for Eigenvalue Computations*, J. Comp. Appl. Math, 27 (1989), pp. 215–227.

[23] J. Dongarra and M. Sidani, *A Divide and Conquer Method for the Nonsymmetric Eigenvalue Problem*, tech. rep., Department of Computer Science, University of Tennessee, 1991.

[24] J. Dongarra and D. Sorensen, *A Fully Parallel Algorithm for the Symmetric Eigenvalue Problem*, SIAM J. Sci. Stat. Comput., 8 (1987).

[25] M. Dowell and P. Jarratt, *The "Pegasus" Method for Computing the Root of an Equation*, BIT, 12 (1972), pp. 503–508.

[26] G. Engeln-Müllges and F. Reuter, *Numerik-Algorithmen mit Fortran 77-Programmen*, 7.Auflage, BI Wissenschaftsverlag, 1993.

[27] K. Gates, *Divide and Conquer Methods for the Symmetric Tridiagonal Eigenvalue Problem*, tech. rep., 1991.

[28] G. Golub, *Some Modified Matrix Eigenvalue Problems*, SIAM Review, 15 (1973), pp. 318–334.

[29] G. Golub and van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, second ed., 1989.

[30] G. Gose, *Zur Konvergenz der Koordinatenrelaxation für $Ax = \lambda Bx$*, ZAMM, 57 (1977), pp. 591–596.

[31] W. Gragg and L. Reichel, *A Divide and Conquer Method for Unitary and Orthogonal Eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[32] M. Gu and S. Eisenstat, *A Stable and Efficient Algorithm for the Rank-one Modification of the Symmetric Eigenproblem*, tech. rep., Yale University, Department of Computer Science, 1992. Research Report YALEU/DCS/RR-916.

[33] E. Jessup, *A Case Against a Divide and Conquer Approach to the Nonsymmetric Eigenvalue Problem*, tech. rep., Department of Computer Science, University of Colorado, 1992.

[34] W. Kahan, *Rank-1 Perturbed Diagonal's Eigensystem*. unpublished manuscript, July 1989.

[35] L. Kaufman, *The LZ Algorithm to Solve the Generalized Eigenvalue Problem for Complex Matrices*, ACM Transactions on Mathematical Software, 1 (1975), pp. 271–281.

[36] ——, *Some Thoughts on the QZ Algorithm for Solving the Generalized Eigenvalue Problem*, ACM Transactions on Mathematical Software, 3 (1977), pp. 65–75.

[37] K. Keller, *Pseudozufallszahlen und Monte-Carlo-Methoden*, Skript des Rechenzentrums der RWTH Aachen, 1993.

[38] E. Kovač-Striko and K. Veselić, *Trace Minimization and Definiteness of Symmetric Pencils*, tech. rep., Fernuniversität Hagen, 1992.

[39] K. Li and T.-Y. Li, *An Algorithm for Symmetric Tridiagonal Eigenproblems: Divide and Conquer with Homotopy Continuation*, SIAM Journal on Scientific Computing, 14 (1993).

[40] R. Li, *Solving Secular Equations Stably and Efficiently*. Departement of Mathematics, University of California at Berkeley, California 94720, 1993.

[41] S.-S. Lo, B. Philippe, and A. Sameh, *A Multiprocessor Algorithm for the Symmetric Tridiagonal Eigenvalue Problem*, SIAM J. Sci. STAT. COMPUT., 8 (1987), pp. 579–590.

[42] C. Moler and G. Stewart, *An Algorithm for Generalized Matrix Eigenvalue Problems*, SIAM J. Numer. Anal, 10 (1973).

[43] Y. Moon, *An Algorithm for Rotating a Symmetrical Generalized Eigenvalue Problem*, Dept. of Computer Science, The Chinese University of Hong Kong.

[44] Y. Moon and C. Crawford, *Finding a Positive Definite Linear Combination of Two Hermitean Matrices*, Linear Algebra and Its Applications, 51 (1983), pp. 37–48.

[45] D. O'Leary and G. Stewart, *Computing the Eigenvalues and Eigenvectors of Symmetric Arrowhead Matrices*, J. Comput. Phys., 90 (1990), pp. 497–505.

[46] B. Parlett, *The Symmetric Eigenvalue Problem*, no. N.J.07632, Prentice-Hall, Inc., Englewood Cliffs, 1980.

[47] R. Quick, J. Matlala, and H. Miller, *An Optimized Iterative Scheme for Obtaining the Dominant Eigenpairs of the Generalized Eigenvalue Problem*, tech. rep., Theoretical Physics Division and Numerical and Applied Mathematics Division, National Research Institute of Mathematical Science of the CSIR, Pretoria, South Africa, 1986.

[48] A. Sameh, *On Jacobi and Jacobi-like Algorithms for a Parallel Computer*, Math. Comp. 25, (1971), pp. 579–590.

[49] H. Schwarz, *The Method of Coordinate Overrelaxation for $(A-\lambda B)x=0$*, Numer. Math., 23 (1974), pp. 135–151.

[50] D. Scott, *Solving Sparse Symmetric Generalized Eigenvalue Problems without Factorization*, Numer. Math., 23 (1974), pp. 135–151.

[51] ——, *The Advantages of Inverted Operators in Rayleigh-Ritz-Approximations*, Siam J. SCI. STAT. COMPUT., 3 (1982).

[52] I. Slapnicar and V. Hari, *On the Quadratic Convergence of the Falk-Langemeyer Method*, Siam J. Math. Anal., 12 (1991), pp. 84–114.

[53] D. Sorensen and P. T. P. Tang, *On the Orthogonality of Eigenvectors Computed by Divide and Conquer Techniques*, Siam J. Numer. Anal., 28 (1991), pp. 1752–1775.

[54] Steinhoff, *Ein Beitrag zur Lösung des pseudosymmetrischen Eigenwertproblems*, PhD thesis, Fakultät für Mathematik der Fernuniversität Hagen, 1983.

[55] G. Stewart, *On the Sensitivity of the Eigenvalue Problem $Ax=\lambda Bx$*, Siam J. Numer. Anal., 9 (1972), pp. 669–686.

[56] ——, *Gershgorin Theory for the Generalized Eigenvalue Problem $Ax=\lambda Bx$*, Math. Comp., 29 (1975), pp. 600–606.

[57] ——, *A Bibliographical Tour of the Large Sparse Generalized Eigenvalue Problem*, Academic Press, New York, 1976.

[58] ——, *Perturbation Bounds for the Definite Generalized Eigenvalue Problem*, Linear Algebra and Its Applications, 23 (1979), pp. 69–85.

[59] R. Thompson, *The Behaviour of Eigenvalues and Singular Values under Perturbations of Restricted Rank*, Linear Algebra and Its Applications, 13 (1976), pp. 69–78.

[60] K. Veselić, *On Real Eigenvalues of Real Tridiagonal Matrices*. Ergebnisbericht der UNIDO, 1970.

[61] ——, *Some Convergent Jacobi-like Procedures for Diagonalizing J-symmetric Matrices*, Numer.Math., 27 (1976), pp. 67–75.

[62] ——, *Global Jacobi Method for a Symmetric Indefinite Problem $Sx = \lambda Tx$*, Comp. Meth. Appl. Mech. Engrg., 38 (1983), pp. 273–290.

[63] ——, *A Jacobi Eigenreduction Algorithm For Definite Matrix Pairs*, Numer. Math., 64 (1993), pp. 241–269.

[64] ——, *On Rank-one Perturbation of Definite Matrix Pairs*. Fernuniversität Hagen, 58084 Hagen, 1993.

[65] ——, *On the Condition of J-Orthonormal Eigenvectors*. Fernuniversität Hagen, 58084 Hagen, 1993.

[66] K. Veselić and V. Hari, *A Note on a One Sided Jacobi Algorithm*, Numer.Math, 56 (1989), pp. 627–633.

[67] R. Ward, *The Combination Shift QZ Algorithm*, SIAM J. Numer. Anal., 12 (1975), pp. 835–853.

[68] Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

Ich möchte mich an dieser Stelle bei Herrn Prof. Dr. K. Veselič für die Anregung zu diesem Thema und für seine hilfreichen Hinweise und Ratschläge bedanken. Mein besonderer Dank gilt meinem Mann, Dr. Klaus Keller, für seine ständige Bereitschaft zu weiterführenden Diskussionen.

Lebenslauf

1. Angaben zur Person:

Name:	Keller, Daniela, geb. Wilke
Geburtsdatum:	12. 08. 1957
Staatsangehörigkeit	deutsch
Familienstand:	verheiratet
Eltern:	Karl-Heinz Wilke Maria Wilke, geb. Geile

2. Bildungsweg:

1967 - 1976	neusprachliches Gymnasium in Hattingen/Ruhr Abitur
1976 - 1985	Mathematik-Informatik-Studium an der Universität Dortmund Diplom in Mathematik

3. Berufliche Tätigkeit:

1985 - 1990	wissenschaftliche Angestellte an der FernUniversität Hagen am Lehrstuhl für Mathematische Physik
seit 1990	Programmierer an der FernUniversität Hagen im Lehrgebiet Praktische Informatik I