Aplikace matematiky

Friedrich Ludwig Bauer

Numerische Abschätzung und Berechnung von Eigenwerten nichtsymmetrischer Matrizen

Aplikace matematiky, Vol. 10 (1965), No. 2, 178–189

Persistent URL: http://dml.cz/dmlcz/102945

NUMERISCHE ABSCHÄTZUNG UND BERECHNUNG VON EIGENWERTEN NICHTSYMMETRISCHER MATRIZEN

F. L. Bauer

Eigenwertaufgaben mit nichtsymmetrischen Matrizen kommen glücklicherweise in der Praxis nicht so häufig vor wie diejenigen mit symmetrischen, insbesondere mit symmetrischen, positiv definiten Matrizen[1]). Wir sagen glücklicherweise, denn die numerische Behandlung des nichtsymmetrischen Eigenwertproblems ist weitaus schwieriger; man darf sagen, daß in voller Allgemeinheit noch kein absolut befriedigendes Verfahren angegeben werden kann.

Wir werden versuchen, einige Gründe für diese Situation anzugeben und damit, dem Charakter des Symposiums entsprechend, die Problematik aufzuzeigen, sowie einige mögliche Wege zur Lösung. Im ersten Abschnitt werden wir untersuchen, wie sich der Fortfall kennzeichnender Eigenschaften symmetrischer Matrizen auswirkt. Im zweiten Abschnitt werden wir im Zusammenhang mit der Empfindlichkeit der Eigenwerte gegen Störungen der Matrix einige grundsätzliche Überlegungen zur Genauigkeitsuntersuchung in der Numerischen Mathematik wiedergeben. Im dritten Abschnitt werden wir das Problem behandeln, aus guten Approximationen für die Eigenvektoren strenge Abschätzungen für die Eigenwerte zu erhalten. Im letzten Abschnitt schließlich werden wir auf die Approximation der Eigenwerte durch ein Newtonverfahren eingehen und zeigen, daß die sogenannte *LR*-Transformation mit Nullpunktverschiebung eng mit dieser Aufgabe zusammenhängt.

1. BESONDERHEITEN DES EIGENWERTPROBLEMS MIT NICHTSYMMETRISCHEN MATRIZEN

Symmetrische Matrizen besitzen drei Eigenschaften, die für die numerische Behandlung des Eigenwertproblems von großer Bedeutung sind: es treten nur reelle Eigenwerte auf, es existiert ein vollständiges System gegenseitig orthogonaler Eigenvektoren und das Eigenwertproblem ist stets optimal konditioniert. Keine der drei Eigenschaften ist für nichtsymmetrische Matrizen gewährleistet, insbesondere fehlen

[1]) Wir beschränken uns auf den Fall reeller Matrizen, die Übertragung auf das Komplexe ist sinngemäß zu verstehen.

die beiden letzten Eigenschaften für nicht-normale Matrizen (wo $A^H \cdot A \neq A \cdot A^H$). Es liegt auf der Hand, daß dadurch die Bestimmung von Eigenwerten erheblich schwieriger wird.

Bei jedem Lösungsverfahren der allgemeinen Eigenwertaufgabe muß auf das mögliche Vorkommen von komplexen Eigenwerten Rücksicht genommen werden, die bei reellen Matrizen in konjugiert komplexen Paaren (also betragsgleich) auftreten. Das geschieht in der Regel durch die Bestimmung reeller quadratischer Faktoren des charakteristischen Polynoms[2]). Beispielsweise ist die Verwendung Sturmscher Ketten zur Steuerung der Eigenwertsuche, die für tridiagonale, symmetrische Matrizen ein überaus schnelles Bisektionsverfahren liefert (Givens [1], Wilkinson [2]), nicht mehr möglich.

Eigenvektoren, die zueinander sehr kleine Winkel bilden, spannen gewisse Richtungen schlechter auf als andere, da nämlich beim Zusammensetzen solcher Richtungen aus Eigenvektoren notwendig führende Ziffern verloren gehen. Dadurch wird die Fehlerempfindlichkeit erhöht, was sich auch in der Größenordnung der Elemente des zu einem vollen Eigenvektorsystem inversen Systems zeigt, also etwa eines Rechts- und eines Linkseigenvektorsystems. Zusammengehörige Rechts- und Linkseigenvektoren x_i, y_i^H können fast orthogonal sein. Das bedeutet, daß

$$K_i = \frac{\|y_i^H\|^D \cdot \|x_i\|}{|y_i^H x_i|},$$

der reziproke Cosinus zwischen diesen beiden Vektoren, sehr viel größer als eins ausfallen kann ($\|\,.\,\|$ ist eine Vektornorm und $\|\,.\,\|^D$ ist die dazu duale Norm)[3]). Ist also das System der Rechtseigenvektoren normiert, so kann die Länge der Vektoren des inversen Systems beliebig groß werden. Sind schließlich zusammengehörige Rechts- und Linkseigenvektoren exakt orthogonal, so liegt der Fall höherer Elementarteiler vor. Es existiert dann kein vollständiges Eigenvektorsystem mehr und das Minimalpolynom bekommt mehrfache Nullstellen.

Den oben erwähnten Ausdruck K_i bezeichnet man als die Kondition eines einfachen Eigenwertes. Aus der infinitesimalen Beziehung

$$\delta\lambda_i = y_i^H \delta A \, x_i$$

[2]) Für Verfahren von Bernoullischem Typus bedeutet das keine Schwierigkeit, z.B. für die Verfahren der Treppeniteration [3] oder der Bi-Iteration [3]. Diese Verfahren erfordern aber wegen ihrer langsamen, nämlich nur linearen Konvergenz konvergenzbeschleunigende Maßnahmen, wie Nullpunktverschiebung, siehe Abschnitt 4. Verfahren von Graeffeschem Typus wurden ebenfalls angegeben (Rutishauser und Bauer [4], Bauer [5]), sie führen auf kompliziertere Algorithmen und haben wenig Anklang gefunden.

[3]) Für die selbstduale euklidsche Norm sei $X = \{x_i\}$ ein vollständiges System von Rechtseigenvektoren, $M = X^H \cdot X$ ist dann die Matrix der $\cos(x_i, x_k)$, mit auf eins normierter Diagonale. Die Diagonalelemente der Inversen M^{-1} geben über die gegenseitige Lage von Rechts- und Linkseigenvektoren Aufschluß: es ist

$$|M^{-1}|_{ii} = \frac{\|y_i^H\|^2 \cdot \|x_i\|^2}{|y_i^H x_i|^2}.$$

folgt nämlich

$$|\delta\lambda_i| \leqq K_i \cdot \|\delta A\| \tag{1}$$

wobei $\|\delta A\|$ eine mit der Vektornorm $\|\,.\,\|$ verträgliche Matrixnorm von δA bedeutet[4]). Für symmetrische Matrizen hat jeder Eigenwert die bestmögliche Kondition eins, und das gilt auch bei mehrfachen Eigenwerten. Die Abschätzung (1) behält bei einfachen Eigenwerten auch für endliche Änderungen δA ihren Sinn. 1960 habe ich mit C. T. Fike darüber folgenden Satz bewiesen [6]:

Satz 1. *Hat A nur einfache Eigenwerte und ist λ_i ein solcher mit einem Rechtseigenvektor x_i und einem Linkseigenvektor y_i^H, so gibt es eine Konstante k_i derart, daß der Kreis*

$$M_i = \left\{ z : |z - \lambda_i| \leqq \frac{\|y_i^H\|^D \, \|x_i\|}{|y_i^H x_i|} \cdot \frac{\|B - A\|}{1 - k_i \|B - A\|} \right\}$$

genau einen Eigenwert der Matrix B enthält, falls nur $n.k_i\|B - A\| < 1$, nämlich

$$k_i = \sum_{j \neq i} \left[\left(\frac{\|y_i^H\|^D \|x_i\|}{|y_i^H x_i|} + \frac{\|y_j^H\|^D \, \|x_j\|}{|y_j^H x_j|} \right) \Big/ \, |\lambda_i - \lambda_j| \right];$$

k_i wird umso größer je näher an λ_i ein anderer Eigenwert liegt.

Ein einfacher Eigenwert λ_i der Matrix A wird also durch eine Abänderung der Matrix umso empfindlicher gestört, je größer seine Kondition $K_i = 1/[\cos(y_i^H, x_i)]$ ist. Es ist klar, daß in dieser Situation Rundungsfehler, die bei einer vorbereitenden Transformation der zu untersuchenden Matrix auftreten, die Eigenwerte erheblich stören können. Aber auch Fehler, mit denen die Matrixelemente als Eingangsdaten behaftet sind, z. B. Meßfehler oder Fehler, die auf vorhergehenden Rechnungen beruhen, bewirken solche erheblichen Unsicherheiten in den Eigenwerten. Bevor wir auf diesen Umstand im grundsätzlichen näher eingehen, sei noch angemerkt, daß häufig nahe zusammenliegende reelle Eigenwerte durch solche Einflüsse ins komplexe Gebiet hinaustreten.

2. ZUR AUFGABENSTELLUNG DER GENAUIGKEITSUNTERSUCHUNG IN DER NUMERISCHEN MATHEMATIK

„Gegeben sei eine Matrix A, deren Eigenwerte (numerisch) zu berechnen sind". Wir greifen diesen landläufigen Satz als Beispiel heraus, um zu zeigen, welche verschiedenen Bedeutungen er haben kann.

Das Problem legt zunächst eine Abbildung des (linearen) Raums der Matrizen A vom Grad n, der Eingangsdaten, in die nicht-geordneten n-tupel von Eigenwerten, die Resultate, fest. Allgemein können wir annehmen, daß ein Datenraum $\mathfrak{D}$ gegeben ist

[4]) Wird die Operatornorm (least upper bound norm) verwendet, so ist die Ungleichung für geeignetes δA sogar scharf.

und eine (eindeutige) Abbildung f in einem Resultatraum $\mathfrak{R}$, $\mathfrak{D} \to \mathfrak{R}$, wobei $\mathfrak{D}$ und $\mathfrak{R}$ durch cartesische Produkte oder auch nicht-geordnete n-tupel von reellen Zahlen dargestellt wird. Den Fall komplexer Zahlen führen wir der Einfachheit halber auf Paare reeller Zahlen zurück.

Ein numerischer Prozeß ist nun eine Abbildung von $\mathfrak{D}$ in $\mathfrak{R}$ durch eine endliche Abfolge von Speziesoperationen und vergleichsbedingten Fallunterscheidungen. Jeder numerische Prozeß ist eine Abbildung f, und es gibt zu jedem numerischen Prozeß viele äquivalente, das heißt die gleiche Abbildung bewirkende Prozesse. Zu einem Problem, d. h. für eine Abbildung f, einen äquivalenten numerischen Prozeß f' zu finden, ist die erste Aufgabe der Numerischen Mathematik, und diese Aufgabe hat oft viel zu viele Lösungen, oft aber auch keine Lösung. Wenn es mehrere Lösungen gibt, setzt ein Verfahrensvergleich ein bezüglich Aufwand und Güte, worauf wir noch zu sprechen kommen werden.

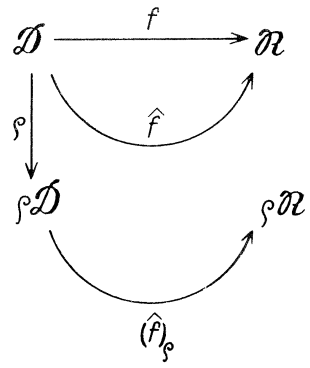

Abb.

Für die Aufgabe, die Eigenwerte einer n-reihigen Matrix zu bestimmen, gibt es keinen numerischen Prozeß, außer wenn man den Datenraum $\mathfrak{D}$ zu sehr, z. B. auf Diagonalmatrizen, einschränkt. Man ist dann gezwungen, eine numerische Methode $\hat{f}$ zu suchen und zu benützen, die f „approximiert", also ein „Näherungsverfahren" einzuschlagen. Aber auch für Aufgaben, die an sich durch einen numerischen Prozeß f gelöst werden können, greift man fast immer zu einem Prozeß, der f zwar möglichst nahekommen soll, aber fast nie mit f übereinstimmt. Man rechnet nämlich nicht mit reellen Zahlen und nicht einmal mit rationalen Zahlen in voller Allgemeinheit, sondern aus praktischen Gründen mit rationalen Zahlen fester maximaler Stellenzahl (als großer Fortschritt ist zu werten, daß in den großen elektronischen Maschinen seit einem Jahrzehnt die Kommalage noch mitgeführt wird). Es wird demgemäß eine als Rundung bezeichnete Abbildung ϱ der reellen Zahlen auf die genannte Teilmenge vorgenommen wenn immer das nötig werden sollte, das heißt womöglich in den Ausgangsdaten und nach den meisten Speziesoperationen. Der eigentliche numerische Prozeß ist die Abbildung von $\mathfrak{D}$ auf $\varrho\mathfrak{D}$ und sodann eine Abbildung f_ϱ bzw. $(\hat{f})_\varrho$ von $\varrho\mathfrak{D}$ in $\varrho\mathfrak{R}$.

Da nun $\hat{f}$ sowie f_ϱ oder auch $(\hat{f})_\varrho$ mit f nicht übereinstimmt, wird es notwendig, eine Fehlerabschätzung in $\mathfrak{R}$ zu verlangen und zu diesem Zweck eine Abstandsfunktion $d(x, y)$ auf $\mathfrak{R}$ einzuführen. Offensichtlich wird man bestrebt sein, f so zu wählen, daß $f_{\text{num}} = (\hat{f})_\varrho$ das Resultat nicht zu falsch wiedergibt, das heißt daß

$$\sup_{A \in \mathfrak{D}} d(f_{\text{num}} A, fA)$$

nicht zu groß wird. Oft wird eine Resultatgenauigkeit vorgeschrieben. Die Aufgabe lautet dann:

a) Für vorgegebenes A und gegebenes c ein $R = f_{\text{num}}A$ anzugeben derart, daß

$$d(R, fA) \leqq c\,.$$

Bezeichnet $\mathfrak{U}_c[R]$ den Resultatspielraum $\{X : d(R, X) \leqq c\}$, so lautet die Forderung $fA \in \mathfrak{U}_c[R]$.

Andererseits sind auch die Ausgangsdaten oft nicht genau bekannt; sie können als Meßwerte mit Meßfehlern oder als Resultate vorhergegangener Rechnungen mit numerisch bedingten Fehlern behaftet sein. Von den vorliegenden Ausgangsdaten A ist dann lediglich bekannt, daß sie bezüglich einer Abstandsfunktion $d'(a, b)$ nicht zu weit von den Sollwerten Y abweichen,

$$d'(A, Y) \leqq c'\,.$$

Mit $\mathfrak{U}_{c'}[A] = \{Y : d(A, Y) \leqq c'\}$ sei der Datenspielraum bezeichnet, den dieAngabe von A und c' läßt. Gelingt es, ein $R = f_{\text{num}}A$ derart anzugeben, daß $R = fY$ mit $Y \in \mathfrak{U}_{c'}[A]$, so könnte R sogar das exakte Resultat zu den wahren Ausgangsdaten Y sein, auf jeden Fall ist R ein zu akzeptierendes Resultat hinsichtlich des Spielraums, der für die Eingangsdaten gelassen ist. Die Aufgabe lautet hier (sie wurde kürzlich von Prager [7] für lineare Gleichungssysteme studiert):

b) Für vorgegebenes A und gegebenes c', ein $R = f_{\text{num}}A$ so anzugeben, daß

$$\exists Y \in \mathfrak{U}_{c'}[A] : f(Y) = R\,.$$

Offenbar sind diese beiden Aufgaben nur Spezialfälle der folgenden:

Für vorgegebenes A und gegebenes c', das den Datenspielraum festlegt, sowie gegebenes c, das den Resultatspielraum festlegt, ein Resultat $R = f_{\text{num}}A$ so anzugeben, daß

$$\exists Y \in \mathfrak{U}_{c'}[A] : f(Y) \in \mathfrak{U}_c[R]\,.$$

Für $c = 0$ ergibt sich Fall b), für $c' = 0$ Fall a). Besteht nun für die Abbildung f eine Lipschitzbedingung

$$d(fA, fY) \leqq l\,.\,d(A, Y)\,,$$

so gilt offenbar für den wahren Fehler

$$d(R, fA) \leqq d(R, fY) + d(fA, fY) \leqq c + l\,.\,c'\,.$$

Bei hinreichend großem l kann ein erheblicher wahrer Fehler trotz kleinem c, c' zustandekommen, oder umgekehrt kann ein erheblicher wahrer Fehler mit kleinem c, c' verträglich werden. Ein Resultat $f_{\text{num}}A$ ist jedoch zu akzeptieren, wenn es mit einem c', das der Genauigkeit der Eingangsdaten entspricht, und mit einem c, das der Genauigkeit der mitgeführten Stellen entspricht, verträglich ist. Abgesehen

von dem Fall, daß die Eingangsdaten exakt durch ganze Zahlen gegeben sind, kann auch von c' nicht erwartet werden, daß es kleiner ist als es der Genauigkeit der mitgeführten Stellen entspricht. Zusammenfassend können wir sagen:

Ein numerischer Prozeß f_{num}, der mit $c \leqq \varepsilon$ und $c' \leqq \varepsilon$, wo ε der Genauigkeit der mitgeführten Stellen entspricht, verträglich ist[5]), ist als gutartig zu bezeichnen. Ein gutartiger Prozeß leistet das, was man von einem nichtganzzahligen numerischen Prozeß überhaupt verlangen darf.

Auf das Eigenwertproblem mit nichtsymmetrischer Matrix angewandt, bedeuten diese Überlegungen, daß von einem numerischen Prozeß nicht erwartet werden soll, daß er besser arbeitet als es der problembedingten Empfindlichkeit der Eigenwerte entspricht. Die Anforderungen, die man an numerische Verfahren für das nichtsymmetrische Eigenwertproblem stellt, werden damit auf ein vernünftiges Maß reduziert.

3. ABSCHÄTZUNGEN BEIM EIGENWERTPROBLEM

Collatz folgend, fasse ich eine Berechnung grundsätzlich als die Bestimmung zweier, den zu berechnenden Wert einschließender Schranken auf. Wenn demnach ein numerisches Verfahren nur *einen* Resultatwert liefert, (und viele Verfahren leisten zunächst nicht mehr), so muß diese Rechnung ergänzt werden entweder

(1) durch eine a priori Fehleranalysis, die von den Eigenschaften des Rechenprozesses und der verwendeten Maschine ausgeht und den in Abschnitt 2 eingeführten Resultatspielraum $\mathfrak{U}_c(R)$ abschätzt (dabei ist es unerheblich, ob die Abschätzung womöglich unter Verwendung der Resultate a posteriori ausgewertet wird)

oder

(2) durch eine a posteriori Abschätzung, die allein auf den erhaltenen Resultaten beruht und von dem verwendeten numerischen Verfahren unabhängig ist.

Mit Punkt (1) kann ich mich hier nicht befassen, weil man dazu auf jede Methode (und manchmal auf jede Maschine) einzeln eingehen müßte. Die a priori Abschätzung ist im Prinzip immer möglich, jedoch im allgemeinen sehr mühsam und erst für wenige Grundprozesse der numerischen Mathematik durchgeführt. Wilkinson [8] gibt Beispiele detaillierter Untersuchungen. Die reine a priori Fehleranalysis dient in erster Linie dem Verfahrensvergleich und der grundsätzlichen Kenntnis des Verfahrens. Sie kann beispielsweise dazu führen, den durch das Verfahren bestimmten numerischen Prozeß als gutartig im Sinne von Abschnitt 2 nachzuweisen.

Abschätzungen gemäß (2) haben den Vorzug, auch vom numerischen Prozeß unabhängig zu sein (Abschätzungen gemäß (1) setzen überdies stets störungsfreies Arbeiten der Rechenanlage voraus). Jedoch ist ihre Verwendungsmöglichkeit dadurch

[5]) In der Praxis wird man sich auch noch mit $c \leqq k\varepsilon$, $c' \leqq k\varepsilon$ zufriedengeben, wo k eine kleine ganze Zahl ist, bei Matrixprozessen sogar mit $k = O(\sqrt{n})$, wo n der Grad der Matrix ist.

beschränkt, daß sie erst für wenige Probleme in einer für die Praxis hinreichend scharfen Weise existieren. Für das Eigenwertproblem beruht die Anwendung dieser Abschätzungen hauptsächlich darauf, zunächst für einige oder (sofern wir den Fall höherer Elementarteiler ausschließen) für alle Eigenvektoren gute Annäherungen zu bestimmen.

Im Falle symmetrischer Matrizen hat man Einschließungssätze, die einen Eigenwert λ_i umso besser eingrenzen, je besser die verwendete Näherung $\tilde{x}_i$ des zugehörigen Eigenvektors ist, zum Beispiel

$$|\lambda_i - \frac{\tilde{x}_i^T A \tilde{x}_i}{\tilde{x}_i^T x_i}| \leqq \sqrt{\left[\frac{\tilde{x}_i^T A^2 \tilde{x}_i}{\tilde{x}_i^T x_i} - \left(\frac{\tilde{x}_i^T A \tilde{x}_i}{\tilde{x}_i^T x_i}\right)^2\right]}.$$

Im Fall nichtsymmetrischer Matrizen muß eine Abschwächung in der rechten Seite solcher Eingrenzungen in Kauf genommen werden. Zusammen mit A. S. Householder habe ich gezeigt [9], daß die Einschließung

$$|\lambda_i - \gamma| \leqq \nu(A) \cdot \frac{\|(A - \gamma I)\, \tilde{x}_i\|}{\|\tilde{x}_i\|}$$

gilt mit beliebiger absoluter Norm. Dabei ist $\nu(A) = \inf_P \operatorname{lub}(P) \cdot \operatorname{lub}(P^{-1})$, wo $AP = PA$, $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, wo also P ein vollständiges Eigenvektorsystem ist, und $\operatorname{lub}(.)$ die zugehörige Operatornorm[6]). Für $\nu(A) = 1$ und $\gamma = (x_i^T A x_i)/(x_i^T x_i)$ erhält man die obige Einschließung. Naheliegend ist nun, $\gamma = \tilde{x}_i^T[(A + A^T)/2]\, \tilde{x}_i$ zu wählen. Steht aber neben einer Näherung $\tilde{x}_i$ für einen Rechtseigenvektor auch eine Näherung $\tilde{y}_i^T$ für einen Linkseigenvektor zur Verfügung, so sollte man erwarten, daß $(\tilde{y}_i^T A \tilde{x}_i)/(\tilde{y}_i \tilde{x}_i)$ den Eigenwert λ_i gut approximiert, und sollte dementsprechend nach Abschätzungen für $|\lambda_i - (\tilde{y}_i^T A \tilde{x}_i)/(\tilde{y}_i^T \tilde{x}_i)|$ suchen, in denen der Faktor $\nu(A)$ nicht auftritt. Anscheinend sind solche Abschätzungen nicht bekannt. Für die Verwendung der üblichen Ausschließungssätze ist die Situation ebenfalls ungünstig. Je besser die Näherungen $\{\tilde{x}_i\}$ an ein vollständiges Rechtseigenvektorsystem sind, desto stärker wird die auf die Basis $\{\tilde{x}_i\}$ ähnlich transformierte Matrix $\tilde{A}$ ein Überwiegen der Hauptdiagonale zeigen. Die Anwendung etwa des Gerschgorinschen Satzes kann also enge Schranken liefern. Tatsächlich kommt aber die Transformation auf A numerisch ohne die Invertierung des Systems $\{\tilde{x}_i\}$ nicht aus, und dieses System kann schlecht konditioniert sein, $\nu(A)$ ist gerade die Kondition. Bei der Berechnung der Abschätzung ist also, vom Aufwand abgesehen, schon größte numerische Sorgfalt geboten, und es ist zu erwarten, daß in den erzielten Abschätzungen ebenfalls eine Verschlech-

---

[6]) $\nu(A)$ kann als die Kondition von A bzgl. aller Eigenwerte angesehen werden, insbesondere gilt

$$\nu(A) \geqq \frac{\|y_i^H\|^D \, \|x_i\|}{\|y_i^H x_i\|} \quad \text{für alle } i\,.$$

terung um den Faktor $v(A)$ eintritt. Von besonderem Übel ist wieder, daß die Größenordnung von $v(A)$ durch die Kondition des am schlechtesten konditionierten Eigenwertes diktiert wird.

Die Gewinnung angenäherter Eigenvektoren kann (wenn sie nicht, wie bei den Verfahren von Bernoullischem Typ, im Verfahren selbst begründet liegt) stets aus angenäherten Eigenwerten durch Wielandtsche inverse Iteration erfolgen, wobei nach Wilkinson auf Pivotwahl Wert zu legen ist. Andererseits ist über Abschätzungen für Eigenvektoren, etwa durch Angabe von Einschließungsregeln, im allgemeinen kaum etwas bekannt. Es ist noch nicht klar, wie sich Ansätze von Wielandt und Falk auch auf den nichtsymmetrischen Fall brauchbar erweitern lassen.

4. NEWTONVERFAHREN ZUR EIGENWERTBERECHNUNG

Es verbleibt noch zu diskutieren, ob es zuverlässige, das heißt unter allen Umständen konvergente Iterationsverfahren zur Eigenwertbestimmung gibt, die auch rasch genug konvergieren. Wir werden zeigen, daß dieses Problem nicht schwieriger ist als das Problem der Nullstellenbestimmung von Polynomen, daß es nämlich Newtonverfahren dafür gibt.

Es sei $f(\lambda)$ eine Funktion der (komplexen) Variablen λ, deren Nullstellen zu bestimmen sind. Das allgemeine Newton-Verfahren k-ter Ordnung, $k \geqq 2$, nach Schröder [10] benützt eine Folge von Iterationen $x_{i+1} = x_i + \Phi_k(x_i)$, wobei

$$\frac{1}{\Phi_k(\lambda)} = \frac{1}{k-1} \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln \left(\frac{\mathrm{d}}{\mathrm{d}\lambda} \right)^{k-2} \frac{g(\lambda)}{f(\lambda)} =$$

$$= \frac{1}{k-1} \left[\left(\frac{\mathrm{d}}{\mathrm{d}\lambda} \right)^{k-1} \frac{g(\lambda)}{f(\lambda)} \right] \Big/ \left[\left(\frac{\mathrm{d}}{\mathrm{d}\lambda} \right)^{k-2} \frac{g(\lambda)}{f(\lambda)} \right].$$

Das Iterationsverfahren konvergiert von k-ter Ordnung in einer hinreichend kleinen Umgebung einfacher Nullstellen von $f(\lambda)/g(\lambda)$. Für unsere Zwecke ist $f(\lambda)$ das charakteristische Polynom einer Matrix A. Für $g(\lambda)$ kann ein beliebiges Polynom vom Grad $n-1$ oder geringer gewählt werden. Es ist vorteilhaft, wenn $g(\lambda)$ mit $f(\lambda)$ Nullstellen gemeinsam hat, diese Nullstellen werden dann beseitigt. Die Wahl $g(\lambda) = f'(\lambda)$ wird bei mehrfachen Nullstellen häufig verwendet. Am günstigsten ist es, wenn $g(\lambda)$ vom Grad $n-1$ ist und alle bis auf eine Nullstelle mit $f(\lambda)$ gemeinsam hat. Dann liefert bereits der erste Korrekturschritt diese Nullstelle.

Von einem Newtonverfahren zur Annäherung der Eigenwerte einer Matrix soll man erwarten, daß es sofort einen Eigenwert liefert, wenn die Matrix eine Diagonalmatrix ist. Dies wird dann erreicht, wenn $g(\lambda)$ das charakteristische Polynom eines $(n-1)$-reihigen Hauptminors der Matrix ist.

Um nun zur Bestimmung von Newton-Schröder-Korrekturen für die Eigenwerte einer Matrix A zu kommen, betrachten wir die Resolvente

$$(A - \lambda I)^{-1} = \frac{1}{f(\lambda)} G(\lambda),$$

wo $f(\lambda)$ das charakteristische Polynom ist und $G(\lambda)$ eine Matrix, deren Elemente Polynome in λ sind vom Grad $n - 1$ höchstens. Nun gilt

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(A - \lambda I)^{-k} = k \,.\, (A - \lambda I)^{-(k+1)}$$

und somit

$$(A - \lambda I)^{-k} = \frac{1}{(k-1)!}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^{k-1} \frac{G(\lambda)}{f(\lambda)}.$$

Deshalb ist

$$\frac{1}{\Phi_k(\lambda)} = \frac{[(A - \lambda I)^{-k}]_{ij}}{[(A - \lambda I)^{-(k-1)}]_{ij}}$$

eine Newtonkorrektur k-ter Ordnung mit $g(\lambda) = [G(\lambda)]_{ij}$. Für $i = j = n$ ist $g(\lambda)$ das charakteristische Polynom des $(n-1)$-reihigen Hauptminors von A.

Die Funktion $\Phi_k(\lambda)$ braucht nicht explizit bestimmt zu werden, es genügt, wenn $\Phi_k(x_i)$ numerisch berechnet wird, das heißt wenn

$$\frac{[(A - x_i I)^{-(k-1)}]_{nn}}{[(A - x_i I)^{-k}]_{nn}}$$

berechnet wird. Wird zur Abkürzung $A - x_i I = \mathfrak{A}$ gesetzt, und ist $\det(\mathfrak{A}) \neq 0$, so sei $\mathfrak{A}^k = \mathfrak{L}_k \mathfrak{R}_k$ die Dreieckszerlegung von $\mathfrak{A}^k$, etwa mit auf eins normierter Diagonale von $\mathfrak{L}_k$. Dann ist $\mathfrak{A}^{-k} = \mathfrak{R}_k^{-1} \mathfrak{L}_k^{-1}$, und $[\mathfrak{A}^{-k}]_{nn} = [(A - x_i I)^{-k}]_{nn} = [\mathfrak{R}_k^{-1}]_{nn} = 1/[\mathfrak{R}_k]_{nn}$. Wir haben also

$$\Phi_k(x_i) = \frac{[\mathfrak{R}_k]_{nn}}{[\mathfrak{R}_{k-1}]_{nn}} = [\mathfrak{R}_k(\mathfrak{R}_{k-1})^{-1}]_{nn}.$$

Setzen wir

$$\mathfrak{R}_k(\mathfrak{R}_{k-1})^{-1} = R_k \quad \text{und} \quad (\mathfrak{L}_{k-1})^{-1}\,\mathfrak{L}_k = L_k,$$

so gilt

$$\begin{aligned} R_k L_k &= \mathfrak{R}_k \mathfrak{R}_{k-1}^{-1} \mathfrak{L}_{k-1}^{-1} \mathfrak{L}_k = \\ &= \mathfrak{L}_k^{-1} \mathfrak{L}_k \mathfrak{R}_k \mathfrak{R}_{k-1}^{-1} \mathfrak{L}_{k-1}^{-1} \mathfrak{L}_k = \\ &= \mathfrak{L}_k^{-1} \mathfrak{A} \mathfrak{L}_k \end{aligned}$$

und

$$\begin{aligned} L_{k+1} R_{k+1} &= \mathfrak{L}_k^{-1} \mathfrak{L}_{k+1} \mathfrak{R}_{k+1} \mathfrak{R}_k^{-1} = \\ &= \mathfrak{L}_k^{-1} \mathfrak{L}_{k+1} \mathfrak{R}_{k+1} \mathfrak{R}_k^{-1} \mathfrak{L}_k^{-1} \mathfrak{L}_k = \\ &= \mathfrak{L}_k^{-1} \mathfrak{A} \mathfrak{L}_k, \end{aligned}$$

also

$$R_k L_k = L_{k+1} R_{k+1} .$$

Beginnend mit $\mathfrak{A} = L_1 R_1$, ergibt sich R_k durch den Algorithmus der LR-Transformation von Rutishauser [11]

$$R_i L_i = L_{i+1} R_{i+1}$$

nach k Schritten. Somit ist

$$\Phi_k(x_i) = [R_k]_{nn} = [L_{k+1} R_{k+1}]_{nn} = [\mathfrak{A}_{k+1}]_{nn}$$

wobei $\mathfrak{A}_{k+1} = R_k L_k = L_{k+1} R_{k+1}$ die nach k Schritten der LR-Transformation aus $\mathfrak{A} = A - x_i I$ entstehende, zu $A - x_i I$ ähnliche Matrix ist.

Die Bestimmung der Newtonkorrektur führt also in natürlicher Weise auf die LR-Transformation. Umgekehrt wird die LR-Transformation mit Nullpunktverschiebung zu einem Newton-Verfahren k-ter Ordnung, wenn als Korrektur das (n, n)-Element der k-fach transformierten Matrix benützt wird. Für $k = 1$ wurde dieses Verfahren schon häufig verwendet; es zeigt sich also, daß man höhere Approximationen bekommt, wenn man die Korrektur erst nach k Schritten, $k \geqq 2$, anwendet. Die Durchführbarkeit der LR-Transformation muß hier natürlich vorausgesetzt werden.

Die LR-Transformation hat aber auch die Eigenschaft, die transformierte Matrix mehr und mehr auf Dreiecksgestalt zu bringen. Damit wird $g(\lambda) = |G(\lambda)|_{nn}$ mehr und mehr einem Teiler von $f(\lambda)$ angenähert. Die quantitative Untersuchung dieses Sachverhalts ergibt folgendes:

Es sei wieder $g(\lambda)$ das charakteristische Polynom des $(n-1)$-reihigen Hauptminors von $\mathfrak{A}_1 = \mathfrak{A}$ und $g^*(\lambda)$ ebendasselbe für $\mathfrak{A}_2$. Es sei also $\mathfrak{A}$ partitioniert,

$$\mathfrak{A}_1 = \mathfrak{A} = \begin{pmatrix} W & u \\ v^T & \xi \end{pmatrix} = \begin{pmatrix} L & 0 \\ v^T R^{-1} & 1 \end{pmatrix} \cdot \begin{pmatrix} R & L^{-1}u \\ 0 & \xi - v^T W^{-1} u \end{pmatrix},$$

es ergibt sich dann sofort, daß

$$\mathfrak{A}_2 = \begin{pmatrix} L^{-1}(W + uv^T W^{-1})\, L & L^{-1}u \\ (\xi - v^T W^{-1} u)\, v^T R^{-1} & \xi - v^T W^{-1} u \end{pmatrix}$$

wird, die Existenz von W^{-1} vorausgesetzt. Dann ist also

$$g(\lambda) = \det(W - \lambda I)$$

und

$$g^*(\lambda) = \det(W + uv^T W^{-1} - \lambda I),$$

sowie

$$f(\lambda) = g(\lambda) \cdot [(\xi - \lambda) - v^T (W - \lambda I)^{-1} u] .$$

Somit ist

$$\begin{aligned}\frac{g^*(\lambda)}{g(\lambda)} &= \det\left((W-\lambda I)^{-1}\left(W+uv^TW^{-1}-\lambda I\right)\right) = \\ &= \det\left(I+(W-\lambda I)^{-1}uv^TW^{-1}\right) = \\ &= 1+v^TW^{-1}(W-\lambda I)^{-1}u\end{aligned}$$

oder

$$\begin{aligned}\lambda g^*(\lambda) &= g(\lambda)\,.\,\left[\lambda+v^T(W-\lambda I)^{-1}u-v^TW^{-1}u\right] = \\ &= g(\lambda)\,.\,(\xi-v^TW^{-1}u)-f(\lambda) = \\ &= g(\lambda)\frac{f(0)}{g(0)}-f(\lambda)\,.\end{aligned}$$

Hat also $g(\lambda)/f(\lambda)$ die Partialbruchzerlegung

$$\frac{g(\lambda)}{f(\lambda)} = \sum\frac{e_\mu}{\lambda_\mu-\lambda}\,,$$

so gilt

$$\frac{g^*(\lambda)}{f(\lambda)} \sim \sum\frac{e_\mu/\lambda_\mu}{\lambda_\mu-\lambda}\,.$$

Es werden also beim Übergang von $g(\lambda)/f(\lambda)$ nach $g^*(\lambda)/f(\lambda)$ die Gewichte in der Partialbruchzerlegung mit $1/\lambda_\mu$ multipliziert, die Gewichte der der Null benachbarten Nullstelle werden also am stärksten angehoben.

Selbstverständlich kann

$$\left[(A-x_iI)^{-k}\right]_{nn} = e_n^T(A-x_iI)^{-k}\,e_n$$

auch als Skalarprodukt von $\tilde{y}^T = e_n^T(A-x_iI)^{-k_1}$ mit $\tilde{x} = (A-x_iI)^{-k_2}\,e_n$ $(k_1+k_2=k)$ berechnet werden, wobei beide Vektoren etwa durch Wielandtsche gebrochene Iteration bestimmt werden und Annäherungen an einen Eigenvektor sind. Es gilt dann

$$\Phi_k(x_i) = \frac{\tilde{y}^T(A-\tilde{x}_iI)\,\tilde{x}}{\tilde{y}^T\tilde{x}}$$

und

$$x_{i+1} = x_i+\Phi_k(x_i) = \frac{\tilde{y}^TA\tilde{x}}{\tilde{y}^T\tilde{x}}\,,$$

die Newton-Näherung hat also die Form eines unsymmetrischen Rayleigh-Quotienten. An solche Quotienten Einschließungssätze anzuhängen, wäre wiederum höchst wünschenswert.

Zur Annäherung quadratischer Faktoren – was für den Fall des Auftretens konjugiert-komplexer Paare von Eigenwerten unerläßlich ist – kann ein zweidimensionales Newtonverfahren verwendet werden. Naheliegend ist zu vermuten, daß die

zweireihigen Minoren, gebildet aus den vorletzten und letzten Zeilen und Spalten der LR-Transformierten, dabei eine Rolle spielen, und zwar daß die charakteristischen Polynome dieser zweireihigen Matrizen sich einem quadratischen Faktor nähern[7]). Die Einzelheiten sind derzeit Gegenstand von Arbeiten an unserem Institut. Wenn man noch durch geeignete Kunstgriffe, wie Francis [12] es für die QR-Transformation getan hat, den expliziten Durchgang durch das Komplexe vermeidet, verbleibt für die Bestimmung der Eigenwerte nichtsymmetrischer Matrizen lediglich noch die bekannte Schwierigkeit, mit Newtonverfahren globale Konvergenz zu erreichen.

In Verbindung mit der LR-Transformation kommt es dabei darauf an, Kriterien zu finden, die die Wirksamkeit der Newton-Korrektur als konvergenzbeschleunigende Maßnahme garantieren. Daß im übrigen die LR-Zerlegung mit Pivotwahl geschehen muß, um Instabilitäten zu vermeiden, dürften sich von selbst verstehen.

Literaturverzeichnis

[1] *W. Givens:* Numerical computation of the characteristic values of a real symmetric matrix. Oak Ridge National Laboratory, ORNL – 1574 (1954).

[2] *J. H. Wilkinson:* Calculation of the eigenvalues of a symmetric tridiagonal matrix by the method of bisection. Numer. Math. 4, 362–367 (1962).

[3] *F. L. Bauer:* Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. Z. Angew. Math. Phys. 8, 214–235 (1957).

[4] *H. Rutishauser* und *F. L. Bauer:* Détermination des vecteurs propres d'une matrice par une méthode iterative avec convergence quadratique. Comptes Rendus 240, 1680 (1955).

[5] *F. L. Bauer:* Das Verfahren der abgekürzten Iteration. Z. Angew. Math. Phys. 7, 17–32 (1956).

[6] *F. L. Bauer* und *C. T. Fike:* Norms and exclusions theorems. Numer. Math. 2, 137–141 (1960).

[7] *W. Prager* und *W. Oettli:* Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand side. Erscheint in Numer. Math.

[8] *J. H. Wilkinson:* Rounding errors in algebraic processes. Her Majesty's Stationary Office, London 1963.

[9] *F. L. Bauer* und *A. S. Householder:* Moments and characteristic roots. Numer. Math. 2, 42–53 (1960).

[10] *E. Schröder:* Über unendlich viele Algorithmen zur Auflösung der Gleichungen. Math. Ann. 2, 317–365 (1870).

[11] *H. Rutishauser:* Report on the solution of eigenvalue problems with the LR-transformation. Nat. Bureau Standards Appl. Math. Ser. 49, 47–81 (1958).

[12] *J. G. F. Francis:* The QR-transformation. A unitary analogue to the LR-transformation. Comput. J. 4, 265–271 (1961, 1962).

Prof. Dr. *F. L. Bauer,* Mathematisches Institut der Technischen Hochschule, 8 München 2, Arcisstr. 21, Deutschland.

[7]) In der Praxis wird dies, allerdings bisher ohne eingehende theoretische Begründung, häufig benützt.