Beiträge zur Numerischen Mathematik
9 (1981), 13–19

Zur numerischen Auflösung der Matrizengleichung $AX^2 - I = 0$

Götz Alefeld

1. Die Aufgabe, eine Matrix X zu bestimmen, für die $AX^2 - I = 0$ gilt, wurde in der Literatur wiederholt behandelt. Siehe dazu insbesondere Albrecht [2, 3], Elsner [7], Laasonen [8], Liebl [10], Ljusternik-Sobolew [11] und Stummel-Hainer [13]. Gewöhnlich wird dabei vorausgesetzt, daß A symmetrisch und positiv definit ist oder zumindest positive Eigenwerte besitzt bzw. daß mit einer Matrix B, für die $\|B\| < 1$ ist, die Darstellung $A = I - B$ gilt. Als Spezialfall ist die betrachtete Aufgabe auch in [6] enthalten.

In dieser Arbeit betrachten wir die gleiche Problemstellung unter der Voraussetzung, daß A von monotoner Art ist. Eine Matrix $A = (a_{ij})$ heißt nach Collatz [5] von monotoner Art, wenn aus $Ax \geqq 0$ die Ungleichung $x \geqq 0$ folgt. Dies ist äquivalent mit $A^{-1} \geqq 0$. Die Halbordnung zwischen Vektoren bzw. Matrizen ist dabei und im folgenden elementweise erklärt. Spezielle Matrizen von monotoner Art sind die M-Matrizen. Eine Matrix A heißt M-Matrix, wenn $a_{ij} \leqq 0$, $i \neq j$, und $A^{-1} \geqq 0$ ist. (Siehe z. B. Varga [14].)

Anwendungen für die oben beschriebene Problemstellung ergeben sich beispielsweise bei der numerischen Behandlung der allgemeinen Eigenwertaufgabe

$$(\lambda A - B)\, x = 0$$

mit einer symmetrischen Matrix B und einer Matrix A, die von monotoner Art ist. Bezeichnet man eine Lösung der Gleichung $AX^2 - I = 0$ mit $A^{-1/2}$, so erhält man mit $y = A^{1/2}x$, $C = A^{-1/2}BA^{-1/2}$ das symmetrische Eigenwertproblem

$$(\lambda I - C)\, y = 0.$$

Wir geben in dieser Arbeit ein Verfahren an, welches eine in der elementweisen Halbordnung monoton wachsende Folge liefert, die unter geeigneten Bedingungen an die Startwerte gegen eine nichtsinguläre Matrix X_* mit $AX_*^2 - I = 0$ konvergiert. Entsprechende Aussagen sind für das Schulzsche Verfahren zur Matrizeninvertierung von Albrecht [1] bewiesen worden.

Ohne auf Einzelheiten einzugehen, erwähnen wir, daß die nachfolgenden Ergebnisse auch in der Menge der linearen beschränkten Operatoren eines Banachraumes in sich mit einer geeigneten Halbordnung bewiesen werden können.

2. Es sei A eine zunächst beliebige nichtsinguläre Matrix. Die Bestimmung einer Matrix X_*, für die $AX_*^2 - I = 0$ gilt, ist dann äquivalent mit der Bestimmung

einer nichtsingulären Matrix X, für die

$$F(X) \equiv A - (X^{-1})^2 = 0 \tag{0}$$

gilt. Die Anwendung des Newton-Verfahrens

$$F(X_i) + F'(X_i)\,(X_{i+1} - X_i) = 0 \tag{1}$$

auf diese Gleichung führt wegen

$$F'(X)\,W = (X^{-1})^2\,WX^{-1} + X^{-1}W(X^{-1})^2 \tag{2}$$

auf die Iterationsvorschrift

$$X_{i+1}X_i + X_iX_{i+1} = 2X_i^2 + X_i^2(I - AX_i^2), \qquad i = 0, 1, 2, \ldots \tag{3}$$

Ist X_0 nichtsingulär und mit A vertauschbar, $AX_0 = X_0A$, so liefert das Newton-Verfahren (3) eine Folge $\{X_i\}$, die der Vorschrift

$$X_{i+1} = X_i + \frac{1}{2}\,X_i(I - AX_i^2), \qquad i = 0, 1, 2, \ldots, \tag{4}$$

genügt. Alle X_i sind mit A und untereinander vertauschbar.

Die Möglichkeit, diese Vorschrift zur Auflösung der Gleichung $AX^2 - I = 0$ zu verwenden, wurde in [8] — allerdings unter anderen Voraussetzungen — verworfen.

Lemma 1. *Es gelte* $A^{-1} \geqq 0$,[1]) $X_0 \geqq 0$, $I - AX_0^2 \geqq 0$, *und* X_0 *sei mit* A *vertauschbar. Dann gilt für die nach* (4) *berechneten Iterierten:*

a) $I - AX_i^2 \geqq 0$, $i \geqq 0$.

b) $0 \leqq X_0 \leqq X_1 \leqq \cdots \leqq X_i \leqq X_{i+1} \leqq \cdots$.

c) *Die Folge* $\{X_i^2\}$ *ist konvergent.*

d) *Für* $X_0 \geqq \varkappa I$, $\varkappa > 0$ *ist die Folge* $\{X_i\}$ *konvergent.*

e) *Ist* X_0 *nichtsingulär, so gilt* $\lim\limits_{i\to\infty} X_i^2 = A^{-1}$ *und im Falle* $X_0 \geqq \varkappa I$, $\varkappa > 0$, *auch* $\lim\limits_{i\to\infty} X_i = X_*$ *mit* $AX_*^2 - I = 0$.

Beweis. Zu a). Es gelte $I - AX_i^2 \geqq 0$, was nach Voraussetzung für $i = 0$ richtig ist. Dann gilt aufgrund der Vertauschbarkeit von A mit allen X_i

$$I - AX_{i+1}^2 = \left(I - \frac{1}{4}\,AX_i^2\right)(I - AX_i^2)^2. \tag{5}$$

Wegen $I - \frac{1}{4}\,AX_i^2 \geqq \frac{1}{4}\,I - \frac{1}{4}\,AX_i^2 = \frac{1}{4}\,(I - AX_i^2) \geqq 0$ gilt daher auch $I - AX_{i+1}^2 \geqq 0$.

Zu b). Wegen $X_i \geqq 0$, was nach Voraussetzung für $i = 0$ richtig ist, und wegen a) folgt $X_{i+1} = X_i + \frac{1}{2}\,X_i(I - AX_i^2) \geqq X_i$.

---

[1]) Es wird in diesem Lemma nicht vorausgesetzt, daß A eine M-Matrix ist.

Zu c). Nach a) gilt $I - AX_i^2 \geqq 0$, $i \geqq 0$, also wegen $A^{-1} \geqq 0$ die Ungleichung $X_i^2 \leqq A^{-1}$, $i \geqq 0$, d. h., die Folge $\{X_i^2\}$ ist durch A^{-1} nach oben beschränkt. Wegen b) gilt $X_{i+1}^2 \geqq X_i X_{i+1} \geqq X_i^2$, d. h., die Folge $\{X_i^2\}$ ist auch monoton wachsend, somit also konvergent.

Zu d). Mit $X_0 \geqq \varkappa I$, $\varkappa > 0$, gilt nach b) $X_i \geqq X_0 \geqq \varkappa I$, also $\frac{1}{\varkappa} X_i \geqq I$, und damit $\frac{1}{\varkappa} X_i^2 \geqq X_i$, d. h., die Folge $\{X_i\}$ ist auch beschränkt und damit konvergent.

Zu e). Wir zeigen zunächst, daß $\varrho(I - AX_0^2) < 1$ gilt (ϱ bezeichnet den Spektralradius). Dazu zerlegen wir A in $A = M - N$ mit $M = (X_0^{-1})^2$, $N = (X_0^{-1})^2 - A$. Wegen $M^{-1}N = I - AX_0^2 \geqq 0$, $NM^{-1} = I - AX_0^2 \geqq 0$ und $X_0^2 \geqq 0$ ist die angegebene Zerlegung von A schwachregulär (siehe Definition 2.4.15 in [12]). Da nach Voraussetzung $A^{-1} \geqq 0$ ist, gilt nach 2.4.17 in [12]

$$\varrho(M^{-1}N) = \varrho(I - AX_0^2) < 1.$$

Somit gibt es eine natürliche Matrixnorm ($\|I\| = 1$), für die

$$\|I - AX_0^2\| < 1$$

gilt.

Es gelte nun

$$\|I - AX_i^2\| < 1$$

für ein $i \geqq 0$, was — wie gerade gezeigt — für $i = 0$ richtig ist. Dann gilt auch

$$\left\|I - \frac{1}{4} AX_i^2\right\| = \left\|\frac{3}{4} I + \frac{1}{4} (I - AX_i^2)\right\| < \frac{3}{4} + \frac{1}{4} = 1.$$

Damit folgt aus (5)

$$\|I - AX_{i+1}^2\| \leqq \|I - AX_i^2\|^2, \qquad i \geqq 0, \tag{6}$$

und

$$\|I - AX_i^2\| \leqq \|I - AX_0^2\|^{2^i}, \qquad i \geqq 0.$$

Somit erhält man mit (5)

$$\|A^{-1} - X_{i+1}^2\| = \left\|A^{-1}\left(I - \frac{1}{4} AX_i^2\right)(I - AX_i^2)^2\right\| \leqq \|A^{-1}\| \, \|I - AX_0^2\|^{2^{i+1}},$$

also

$$\lim_{i \to \infty} X_i^2 = A^{-1}.$$

Wir haben unter d) mit der Voraussetzung $X_0 \geqq \varkappa I$, $\varkappa > 0$, bereits die Konvergenz $\lim_{i \to \infty} X_i = X_*$ gezeigt. Damit gilt dann auch

$$\lim_{i \to \infty} X_i^2 = X_*^2 = A^{-1},$$

also

$$AX_*^2 - I = 0. \blacksquare$$

Im nächsten Lemma wollen wir zeigen, daß die in Lemma 1 benötigten Voraussetzungen bei M-Matrizen sehr einfach realisiert werden können.

Lemma 2. *Es sei A eine M-Matrix und $X_0 = \varkappa I$ mit*

$$0 < \varkappa \leqq \left(\max_{1 \leqq i \leqq n} a_{ii}\right)^{-1/2}.$$

Dann ist $X_0 \geqq 0$, X_0 nichtsingulär, X_0 mit A vertauschbar, und es gilt $I - AX_0^2 \geqq 0$.

Beweis. Es sei $A = (a_{ij})$ zerlegt in $A = D - B$, wobei D den Diagonalanteil bzw. B den außerhalb der Diagonale stehenden Anteil von A bezeichnet. Dann ist

$$I - AX_0^2 = I - \varkappa^2(D - B) = I - \varkappa^2 D + \varkappa^2 B \geqq 0,$$

falls $1 - \varkappa^2 a_{ii} \geqq 0$, $i = 1(1)\, n$, also

$$0 \leqq \varkappa \leqq \left(\max_{1 \leqq i \leqq n} a_{ii}\right)^{-1/2}$$

gilt. Schließt man $\varkappa = 0$ aus, so ist X_0 nichtsingulär, und alle anderen Behauptungen sind offensichtlich erfüllt.

3. Beim numerischen Rechnen, d. h. bei der Durchführung auf einer Rechenanlage mit Gleitpunktdarstellung der Maschinenzahlen, wird aufgrund der Rundungsfehler während der Rechnung im allgemeinen die Vertauschbarkeit der nach (4) berechneten X_i mit A und untereinander zerstört. Es muß daher untersucht werden, wie sich das Iterationsverfahren (4) in der Nähe von $A^{-1/2}$ für nicht notwendig mit A vertauschbare Startwerte X_0 verhält. Wir können diese Untersuchung in Analogie zu LIEBL [10] durchführen: Setzen wir

$$G(X) = X + \frac{1}{2} X(I - AX^2),$$

so ergibt sich für die Frechet-Ableitung an der Stelle $A^{-1/2}$ die $n^2 \times n^2$-Matrix

$$G'(A^{-1/2}) = \frac{1}{2}\,[I \times I - A^{1/2} \times A^{-1/2}].$$

Dabei bezeichnet $X \times Y$ für $n \times n$-Matrizen $X = (x_{ij})$ und $Y = (y_{ij})$ das Tensorprodukt, welches definiert ist als

$$X \times Y = \begin{pmatrix} x_{11}Y & \dots & x_{1n}Y \\ \dots & \dots & \dots \\ x_{n1}Y & \dots & x_{nn}Y \end{pmatrix}.$$

Siehe dazu z. B. LANCASTER [9].

Wir zeigen, daß die Eigenwerte von $G'(A^{-1/2})$ sicherlich nicht für alle Matrizen A von monotoner Art dem Betrage nach kleiner als 1 sind. Das bedeutet, daß es in beliebiger Nähe von $A^{-1/2}$ Matrizen X_0 gibt, für die das Verfahren (4) nicht gegen $A^{-1/2}$ konvergiert.

Dazu setzen wir speziell voraus, daß A eine symmetrische M-Matrix ist. Solche Matrizen heißen Stieltjes-Matrizen. Sie sind positiv definit. Siehe dazu VARGA [14]. Für die Orthogonalmatrix U gelte

$$UAU^T = \Lambda = \operatorname{diag}(\lambda_i).$$

Λ enthält in der Diagonale die Eigenwerte von A. Um die Eigenwerte von $G'(A^{-1/2})$ zu bestimmen, wenden wir auf $G'(A^{-1/2})$ eine Ähnlichkeitstransformation mit der Matrix $U \times U$ an. Dies führt auf eine Diagonalmatrix mit den Diagonalelementen

$$\lambda_{i,k} = \frac{1}{2}\left(1 - \sqrt{\frac{\lambda_i}{\lambda_k}}\right), \qquad i, k = 1(1)\, n.$$

Liegen also die Eigenwerte von A nicht sehr nahe zusammen, so sind nicht alle $|\lambda_{ik}| < 1$.

Eine Möglichkeit, die eventuell auftretende numerische Instabilität in der Nähe von $A^{-1/2}$ zu umgehen, besteht in der Durchführung des Verfahrens (3), welches ja lokal konvergent ist und sich gegenüber Rundungsfehlern gutartig verhält. Allerdings ist der Aufwand pro Schritt gegenüber (4) erheblich. Siehe dazu etwa [4].

Einen im allgemeinen sehr guten Startwert für (3) findet man durch Anwendung von (4) wie in dieser Arbeit beschrieben. (4) sollte abgebrochen werden, sobald die Monotonie verletzt ist oder in $I - AX_0^2$ mindestens ein Element negativ ist oder die nach (6) theoretisch bestehende Ungleichung für eine Norm nicht mehr erfüllt ist.

4. Abschließend sollen noch einige numerische Ergebnisse angegeben werden. Die Beispiele wurden auf der Rechenanlage CDC 6500 (Mantissenlänge 48 bit) der Zentraleinheit Rechenzentrum der Technischen Universität Berlin gerechnet. Für die Programmierung danke ich Herrn Dipl.-Math. H. SCHWANDT.

1. Die Matrix

$$A = \begin{pmatrix} 4 & 0 & -1 & -1 & -1 \\ -2 & 5 & 0 & -1 & -1 \\ -1 & -2 & 6 & -1 & 0 \\ 0 & -2 & -1 & 7 & -2 \\ -2 & -1 & 0 & -1 & 5 \end{pmatrix}$$

besitzt außerhalb der Hauptdiagonale nichtpositive Elemente und ist strengdiagonaldominant. Somit ist $A^{-1} \geqq 0$. (Siehe z. B. [5].) Als Startmatrix für das Verfahren (4) können wir nach Lemma 2 die Matrix $X_0 = \frac{1}{\sqrt{7}} I$ wählen. Es gilt für die numerische Rechnung

$$X_{i+1} \geqq X_i \qquad \text{für} \quad i = 0(1)6,$$

$$I - AX_i^2 \geqq 0 \quad \text{für} \quad i = 0(1)7,$$

$$\|I - AX_{i+1}\|_\infty \leqq \|I - AX_i^2\|_\infty^2 \quad \text{für} \quad i = 0(1)6.$$

Es ist

$$\|I - AX_8^2\|_\infty \approx 0.28_{10^{-13}},$$

$$\|I - AX_9^2\|_\infty \approx 0.21_{10^{-13}}.$$

2. $A = I - B$ mit $B = (b_{ij})$, $b_{ij} = \begin{cases} 0, & i = j, \\ 0.246, & i \neq j, \end{cases} \quad 1 \leqq i, j \leqq n.$

Mit der gleichen Begründung wie in Beispiel 1 ist $A^{-1} \geqq 0$. Es gilt für die numerische Rechnung

$$X_{i+1} \geqq X_i \qquad \text{für} \quad i = 0(1)10,$$

$$I - AX_i^2 \geqq 0 \qquad \text{für} \quad i = 0(1)10,$$

$$\|I - AX_{i+1}^2\|_\infty \leqq \|I - AX_i^2\|^2 \quad \text{für} \quad i = 0(1)9.$$

Zur Demonstration listen wir noch $r := \|I - AX_i^2\|_\infty$ für $i = 0(1)19$ auf:

i	0	1	2	3	4	5	6
r	0.984	0.965	0.921	0.834	0.665	0.406	0.140

i	7	8	9	10	11	12
r	$0.154_{10^{-1}}$	$0.177_{10^{-3}}$	$0.236_{10^{-7}}$	$0.142_{10^{-9}}$	$0.554_{10^{-9}}$	$0.217_{10^{-8}}$

i	13	14	15	16	17	18	19
r	$0.848_{10^{-8}}$	$0.332_{10^{-7}}$	$0.130_{10^{-6}}$	$0.508_{10^{-6}}$	$0.199_{10^{-5}}$	$0.777_{10^{-5}}$	$0.104_{10^{-4}}$

Die Norm $\|I - AX_i^2\|_\infty$ nimmt monoton bis $i = 10$ ab. Die Vertauschbarkeit der X_i mit A und untereinander wird in der Nähe von $A^{-1/2}$ zerstört. Von $i = 11$ ab nimmt r wieder zu.

Literatur

[1] ALBRECHT, J., Bemerkungen zum Iterationsverfahren von Schulz zur Matrizeninversion, ZAMM **41** (1961), 262–263; **44** (1964), 148.

[2] ALBRECHT, J., Quadratisch konvergente Iterationsverfahren zur Berechnung von $A^{1/2}$ und A^{-1}. Moderne Methoden der numerischen Mathematik, Tagung Clausthal 1975, ISNM **32** (1976).

[3] ALBRECHT, J., Bemerkungen zu Iterationsverfahren zur Berechnung von $A^{1/2}$ und A^{-1}, ZAMM **57** (1977), T262–T263.

[4] BARTELS, R. H., and G. W. STEWART, Algorithm 432, Solution of the Matrix Equation $AX + XB = C$, Communications of the ACM **15** (1972), 820–826.

[5] Collatz, L., Funktionalanalysis und numerische Mathematik, Springer-Verlag, Berlin—Heidelberg—New York 1964.
[6] Dennis, J. E., J. F. Traub and R. P. Weber, Algorithms for solvents of matrixpolynomials, SIAM J. Numer. Anal. **15** (1978), 523—533.
[7] Elsner, L., Iterative Verfahren zur Lösung der Matrizengleichung $X^2 - A = 0$, Bul. Inst. Pol. Iaşi XVI(XX) (1970), 15—24.
[8] Laasonen, P., On the iterative solution of the matrix equation $AX^2 - I = 0$, MTAC **12** (1958), 109—116.
[9] Lancaster, P., Theory of Matrices, Academic Press, New York and London 1969.
[10] Liebl, P., Einige Bemerkungen zur numerischen Stabilität von Matrizeniterationen, Apl. Mat. **10** (1965), 249—254.
[11] Ljusternik, L. A., und W. I. Sobolew, Elemente der Funktionalanalysis, Akademie-Verlag, Berlin 1965 (Übersetzung aus dem Russischen).
[12] Ortega, J. M., and W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York 1970.
[13] Stummel, F., und K. Hainer, Praktische Mathematik, Teubner Studienbücher, Stuttgart 1971.
[14] Varga, R. S., Matrix Iterative Analysis, Prentice-Hall Inc., Englewood Cliffs, N. J., 1962.

Manuskripteingang: 25. 10. 1978

VERFASSER:

Prof. Dr. Götz Alefeld, Fachbereich 3/Mathematik, H 64 der Technischen Universität Berlin (West)