Mathematica Pannonica
16/2 (2005), 231–240

NEWTON ITERATION FOR THE CLOSEST NORMAL MATRIX OF ORDER TWO

Lajos **László**

Department of Numerical Analysis, Eötvös University Budapest, H–1117 Budapest, Pázmány Péter sétány 1/C, Hungary

Received: September 2005

MSC 2000: 15 A 60; 15 A 57, 65 F 30

Keywords: Normal matrix, Newton's method, order of convergence, fixed point, Taylor series.

Abstract: Instead of the existing Jacobi-type algorithms we propose a Newton-like iteration for the closest normal matrix. For second order matrices a convergent iteration will be obtained. Representations of the iterates by help of Chebyshev moments and central binomial coefficients also are discussed.

1. Introduction

All the iterative methods for solving the approximation problem

$$\nu(A) = \inf\{\|A - X\| : X^*X = XX^*,\ X \in \mathbb{C}^{n\times n}\},$$

where $A \in \mathbb{C}^{n\times n}$ is a given complex matrix of order n, and $\|\cdot\| = \|\cdot\|_F$ stands for the Frobenius norm, are based on the Jacobi-idea, see e.g. [1], [5], [2], and the references in them. (Some of them discuss the closely related eigenproblem.) At the same time, the closest unitary matrix can be computed via the Newton method as

$$X_0 = A, \quad X_{k+1} = \frac{1}{2}\Big(X_k + (X_k^*)^{-1}\Big), \quad k = 0, 1, \ldots$$

cf. [4], where further interesting iterations are found. The simplicity of this method leads us to find a similar procedure for the best normal approximation problem, as well.

In Sections 2 and 3 we develop a Newton-like approach for second order matrices, while in this first and also the last sections we discuss matrices of arbitrary order, for our hope is that there exists a Newton-type iteration for them as well. Note that the existing formula for $n = 2$ makes use of the eigenvalues of the matrix – while the iteration below is a rational process!

To define the iteration, it is natural to start with $X_0 = A$ and to write

$$X_{k+1} = X_k - f'(X_k)^{-1} f(X_k),$$

where $f(X) = X^*X - XX^*$. The main trouble is that the inverse here does not exist. However, the set of tangential directions

$$\begin{aligned}\mathcal{L}(X) &= \{Y : f'(X)Y = f(X)\} \\ &= \{Y : X^*Y - YX^* + Y^*X - XY^* = X^*X - XX^*\}\end{aligned}$$

calculated in Ruhe [5] is well defined. Obviously $Y = X/2$ belongs to $\mathcal{L}(X)$, therefore it suffices to examine the homogeneous system

$$\mathcal{L}^{hom}(X) = \{Y : X^*Y - YX^* + Y^*X - XY^* = 0\}.$$

Our point is to find a linearly independent system $(Y_k^{(i)})$ in $\mathcal{L}^{hom}(X_k)$ and to determine the coefficients $(c_k^{(i)})$ in the representation

$$X_{k+1} = X_k/2 + \sum_i c_k^{(i)} Y_k^{(i)}$$

from the requirement

$$\|X_{k+1} - A\| \to \min .$$

For instance, here are two sequences from $\mathcal{L}^{hom}(X)$:

1. $I,\ X^*,\ (X^*)^2,\ (X^*)^3, \dots,$ the Krylov sequence, and
2. $i\,X,\ i\,XX^*X,\ i\,XX^*XX^*X, \dots,\quad i = \sqrt{-1}.$

Members of the first list can be multiplied by any complex, while those from the second by any real number only. This may motivate choosing the Krylov sequence, which certainly works for second order matrices.

2. The case $n = 2$

It is known [3] that the best normal approximant to a given $A \in \mathbb{C}^{2\times 2}$ is

$$BN(A) = \frac{1}{2}(A + zA^*) + \frac{1}{4}\text{trace}(A - zA^*)I, \quad z = \text{sign}(\lambda_1 - \lambda_2)^2,$$

where λ_1, λ_2 are the eigenvalues of A, and $\text{sign}(x) = \frac{x}{|x|}$ for all $0 \neq x \in \in \mathbb{C}$. (If $\lambda_1 = \lambda_2$, then $|z| = 1$ is arbitrary and $BN(A)$ is not unique. This case will be excluded from the discussion.) In the algorithm to be developed, we ask for a pure rational process, i.e. we must not make use of neither the eigenvalues nor the sign function.

Observe that $BN(A)$ is a linear combination of the matrices $\{I, A, A^*\}$; and that in general, a matrix (of arbitrary order) of the form $\gamma_0 I + \gamma_1 A + \gamma_2 A^*$ is normal if and only if $|\gamma_1| = |\gamma_2|$, while this equality, of course, does not need to hold for the *iterates* below.

Assume without loss of generality that our matrix is of the form

$$A = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}, \quad \alpha\beta \neq 0.$$

Any square matrix can be brought to such zero-diagonal form ([6], p. 77) by a translation $A \to A - \gamma I$, and a subsequent unitary similarity, under both of which the best normal approximation problem is invariant. Then, as a direct calculation shows,

$$\text{(1)} \qquad \text{BN}(A) = \frac{|\alpha| + |\beta|}{2} \begin{pmatrix} 0 & \text{sign}(\alpha) \\ \text{sign}(\beta) & 0 \end{pmatrix}.$$

Let $X_0 = A = \left(\begin{smallmatrix} 0 & x_0 \\ y_0 & 0 \end{smallmatrix}\right)$, $X_k = \left(\begin{smallmatrix} 0 & x_k \\ y_k & 0 \end{smallmatrix}\right)$, and seek X_{k+1} in the form of

$$\text{(2)} \qquad X_{k+1} = \frac{1}{2}X_k + r_k X_k^*$$

omitting the identity because of the special zero-diagonal form of X_k.

The requirement $\|X_{k+1} - A\| \to \min$ yields

$$x_{k+1} = \frac{x_k}{2} + r_k \overline{y}_k, \quad y_{k+1} = \frac{y_k}{2} + r_k \overline{x}_k$$

with the overline denoting complex conjugate transpose, and

$$\text{(3)} \qquad r_k = \frac{\text{trace}(AX_k) - \text{trace}(X_k^2)/2}{\|X_k\|^2} = \frac{\alpha y_k + \beta x_k - x_k y_k}{|x_k|^2 + |y_k|^2}.$$

It will be proved that the matrix sequence (X_k) converges to (1). Let

$$x_k = \frac{\alpha}{|\alpha|}u_k, \quad y_k = \frac{\beta}{|\beta|}v_k,$$

then the iteration assumes $u_0 = |\alpha|$, $v_0 = |\beta,|$ and

$$
(4)\qquad
\begin{aligned}
u_{k+1} &= \frac{u_k}{2} + \overline{v}_k \frac{|\alpha|v_k + |\beta|u_k - u_kv_k}{|u_k|^2 + |v_k|^2},\\
v_{k+1} &= \frac{v_k}{2} + \overline{u}_k \frac{|\alpha|v_k + |\beta|u_k - u_kv_k}{|u_k|^2 + |v_k|^2}.
\end{aligned}
$$

It is easy to see by induction that $u_0, v_0 \in \mathbb{R}$ implies $u_k, v_k \in \mathbb{R}$ for all k. However, more is true. Denote to this for brevity

$$a = |\alpha|, \quad b = |\beta|$$

and assume $a < b$.

Lemma 1. *$(u_k, v_k) \in [a,b]^2$ implies $(u_{k+1}, v_{k+1}) \in [a,b]^2$.*

Proof. First we prove the lower bound for u_{k+1}. Omitting the subscripts we have to show that

$$a \le \frac{u}{2} + v\,\frac{av + bu - uv}{u^2 + v^2}$$

holds for all u, v with $a \le u, v \le b$, which is a consequence of

$$(u^2 + 2bv) - (v^2 + 2au) = u(u-a) + v(b-v) + a(b-u) + b(v-a) \ge 0.$$

The upper bound for u_{k+1}, i.e.

$$\frac{u}{2} + v\,\frac{av + bu - uv}{u^2 + v^2} \le b$$

follows from the equivalent inequality

$$(u + 2b - 2a)\left(v - \frac{bu}{u + 2b - 2a}\right)^2 + \\ + \frac{u^2}{u + 2b - 2a}\left((b-a)(3b-a) - (u-a)^2\right) \ge 0.$$

The bounds for v_{k+1} are shown analogously. ◊

This means that for the function $G = (g_1, g_2) : \mathbb{R}^2 \to \mathbb{R}^2$ with

$$g_1(u,v) = \frac{u}{2} + v\frac{av + bu - uv}{u^2 + v^2}, \quad g_2(u,v) = \frac{v}{2} + u\frac{av + bu - uv}{u^2 + v^2}$$

it holds that

$$G([a,b]^2) \subset [a,b]^2,$$

i.e. the first condition of Banach' fixed point theorem is satisfied. However, G is not a contraction with respect to the most known norms, hence we proceed in a different way, decomposing the plane iteration into two scalar problems. To this we will need the bounds for the function

$$r(u,v) = \frac{av + bu - uv}{u^2 + v^2}, \quad (u,v) \in [a,b]^2.$$

Lemma 2. *We have*

$$\frac{a}{2b} \le r(u,v) \le \frac{b}{2a}.$$

Proof. Observe that, the partial derivative r_v of r with respect to v is negative on $[a,b]^2$ due to

$$(u^2+v^2)^2\, r_v(u,v) = (a-u)(u^2+v^2) - 2v(av + bu - uv) < 0,$$

and the positiveness of the numerator of r:

$$av + bu - uv = (b-v)(u-a) + ab \ge ab > 0.$$

Hence minimum occurs for $v = b$ with

$$r(u,v) \ge r(u,b) = \frac{ab}{u^2+b^2} \ge \frac{a}{2b},$$

while maximum is attained at $v = a$ with

$$r(u,v) \le r(u,a) = \frac{a^2 + u(b-a)}{u^2+a^2} \le r(a,a) = \frac{b}{2a}. \; \Diamond$$

After this we can prove the main result.

Theorem 1. *The iteration* (4) *is convergent, and*

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} v_k = \frac{a+b}{2}.$$

Proof. Dividing the first equation by the second gives

$$\frac{u_{k+1} - v_{k+1}}{u_{k+1} + v_{k+1}} = q_k \frac{u_k - v_k}{u_k + v_k}, \quad q_k = \frac{1 - 2r_k}{1 + 2r_k},$$

where $r_k = r(u_k, v_k)$. Since

$$|q_k| \le \frac{b-a}{b+a} < 1$$

by virtue of the bounds for r_k, we have

$$\lim_{k\to\infty} \frac{u_k - v_k}{u_k + v_k} = 0,$$

and at the same time

$$\lim_{k\to\infty} (u_k - v_k) = 0.$$

On the other hand, for the error $e_k = u_k + v_k - (a+b)$ we obtain

$$e_{k+1} = (u_k - v_k)\, \frac{u_k^2 - v_k^2 + 2bv_k - 2au_k}{2(u_k^2 + v_k^2)},$$

showing – since the factor at $(u_k - v_k)$ is bounded –, that

$$\lim_{k\to\infty}(u_k+v_k)=a+b$$

also holds to prove the theorem. ◊

Remark 1. This means, that the original iteration (2–3) is convergent, too.

Remark 2. The quantity $t=b/a$ plays the role of "condition number" for the problem. Indeed, $t=1$ is "ideal," (the matrix at issue is normal), while $t\gg 1$ results in a poor convergence. In case of $a=0$ convergence does not hold at all.

Remark 3. Before this fairly simple proof, we had experimented with other methods, for instance by "recomplexizing" the pair (u_k,v_k), and also by using Taylor expansions, to which we devote the next section.

As for the first, notice that the definitions

$$z_k=\frac{1+i}{2}u_k+\frac{1-i}{2}v_k,\quad c=\frac{1+i}{2}a+\frac{1-i}{2}b$$

result in an equivalent complex *scalar* iteration

$$z_{k+1}=\frac{1}{2}\left(z_k+\frac{\Re[(2c-z_k)z_k]}{z_k}\right),\quad z_0=c.$$

If both z_k and c are real, then $z_{k+1}=c$. What is more, the following is valid:

$$\lim_{k\to\infty}z_k=\begin{cases}\Re(c), & \text{if } |\Re(c)|\ge|\Im(c)|\\ i\Im(c), & \text{else.}\end{cases}$$

Observe that, while this statement holds true for $c\in\mathbb{C}$ arbitrary, in our case of $c=\frac{a+b}{2}+i\frac{a-b}{2}$ we have always to do with case 1.

3. Series expansions

3.1. Chebyshev-moments

Go back to the original variables x_k,y_k, and substitute this time

$$x_k=\alpha\xi_k,\quad y_k=\beta\eta_k$$

to obtain

$$\xi_{k+1}=\frac{\xi_k}{2}+t^2\eta_k\rho_k,\quad \eta_{k+1}=\frac{\eta_k}{2}+\xi_k\rho_k$$

with

$$\rho_k=\frac{\xi_k+\eta_k-\xi_k\eta_k}{\xi_k^2+t^2\eta_k^2}.$$

The advantage of this change of variables is that only one parameter –

the condition number $t = b/a$ – is present, enabling us to expand the iterates into a formal power series with respect to t.

Calculations with Maple suggest that the functions

$$\xi(t) = \lim_{k\to\infty} \xi_k(t) \quad \text{and} \quad \eta(t) = \lim_{k\to\infty} \eta_k(t)$$

can be expanded as

$$\xi(t) = 1 + \frac{1}{4}(t^2-1) - \frac{1}{16}(t^2-1)^2 + \frac{1}{32}(t^2-1)^3 - \\ - \frac{5}{256}(t^2-1)^4 + \frac{7}{512}(t^2-1)^5 + \ldots$$

and

$$\eta(t) = 1 - \frac{1}{4}(t^2-1) + \frac{3}{16}(t^2-1)^2 - \frac{5}{32}(t^2-1)^3 - \\ - \frac{35}{256}(t^2-1)^4 - \frac{63}{512}(t^2-1)^5 + \ldots,$$

where we can recognize the even moments

$$\mu_{2k}^{(1)} = \int_{-1}^{1} \frac{s^{2k}}{\sqrt{1-s^2}} ds, \quad \mu_{2k}^{(2)} = \int_{-1}^{1} s^{2k}\sqrt{1-s^2} ds$$

of the powers for the Chebyshev weights of the first and second kind

$$\left\{\mu_{2k}^{(1)} : k \geq 0\right\} = 2\pi\left\{\frac{1}{2}, \frac{1}{4}, \frac{3}{16}, \frac{5}{32}, \ldots\right\},$$

and

$$\left\{\mu_{2k}^{(2)} : k \geq 0\right\} = 2\pi\left\{\frac{1}{4}, \frac{1}{16}, \frac{1}{32}, \frac{5}{256}, \ldots\right\}.$$

In possession of these expansions, their formal sums can be calculated.

Lemma 3.

$$\xi(t) = 1 + \frac{t^2-1}{2\pi}\sum_{k=0}^{\infty} \mu_{2k}^{(2)}(1-t^2)^k = \frac{1+t}{2},$$

$$\eta(t) = 1 + \frac{1}{2\pi}\sum_{k=0}^{\infty} \mu_{2k}^{(1)}(1-t^2)^k = \frac{1+t}{2t}.$$

Proof. Using the integral representation of the moments and changing the order of the summation and integration (if possible), the formulas

$$\int_{-1}^{1} \frac{\sqrt{1-s^2}ds}{1+s^2(t^2-1)} = \frac{\pi}{t+1}, \quad \int_{-1}^{1} \frac{ds}{[1+s^2(t^2-1)]\sqrt{1-s^2}} = \frac{\pi}{t}$$

yield the result. ◇

Unfortunately, the manipulation of the sum is not allowed for any $t > 1$. Since $|\mu_k^{(1)}| \leq 1$, $|\mu_k^{(2)}| \leq 1$, it is certainly correct for $|t^2 - 1| < 1$, i.e. for $t \in (1, \sqrt{2})$ – a quite limited interval. To enlarge the convergence interval, we look for another change of variables.

3.2. Central binomial coefficients

Let us introduce a new variable τ by help of

$$t^2 = \frac{1 + 2\tau}{1 - 2\tau}, \qquad \tau = \frac{1}{2}\frac{t^2 - 1}{t^2 + 1} = \frac{1}{2}\frac{b^2 - a^2}{b^2 + a^2},$$

where τ ranges in $(0, 1/2)$. The Taylor expansion yields now

$$\xi(t) = \hat{\xi}(\tau) = 1 + \tau + \tau^2 + 2\tau^3 + 3\tau^4 + 6\tau^5 + 10\tau^6 + 20\tau^7 + \dots,$$

$$\eta(t) = \hat{\eta}(\tau) = 1 - \tau + \tau^2 - 2\tau^3 + 3\tau^4 - 6\tau^5 + 10\tau^6 - 20\tau^7 + \dots,$$

where the numbers can be recognized as the (modified) central binomial coefficients $\binom{n}{\lfloor n/2 \rfloor}$ with $\lfloor x \rfloor = \text{floor}(x)$. Then we have a result analogous to the above.

Lemma 4. The sums of the above series exist for $0 < \tau < \frac{1}{2}$, and

$$\hat{\xi}(\tau) = 1 + \tau \sum_{n=0}^{\infty} \binom{n}{\lfloor n/2 \rfloor} \tau^n = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1 + 2\tau}{1 - 2\tau}} = \frac{1 + t}{2},$$

$$\hat{\eta}(\tau) = 1 - \tau \sum_{n=0}^{\infty} \binom{n}{\lfloor n/2 \rfloor} (-\tau)^n = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1 - 2\tau}{1 + 2\tau}} = \frac{1 + t}{2t}.$$

Proof. By virtue of the known [7] formula

$$1 + 2\tau + 3\tau^2 + 6\tau^3 + 10\tau^4 + \dots = \frac{1 - 4\tau^2 - \sqrt{1 - 4\tau^2}}{2\tau^2(2\tau - 1)}$$

valid for $|\tau| < \frac{1}{2}$, the series assume

$$\hat{\xi}(\tau) = 1 + \tau + \tau^2 \frac{1 - 4\tau^2 - \sqrt{1 - 4\tau^2}}{2\tau^2(2\tau - 1)},$$

$$\hat{\eta}(\tau) = 1 - \tau + \tau^2 \frac{1 - 4\tau^2 - \sqrt{1 - 4\tau^2}}{2\tau^2(-2\tau - 1)},$$

and the formulas follow immediately by the connection between τ and t. ◊

4. An example – and the order of convergence

We apply the method (2–3) for the matrix of Ruhe [5]

$$A = \begin{pmatrix} 0.7616 + 1.2296i & -1.4740 - .4577i \\ -1.6290 - 2.6378i & 0.1885 - 0.8575i \end{pmatrix}$$

with closest normal matrix

$$BN(A) = \begin{pmatrix} 1.1449 + 0.8324i & -2.0841 - 0.9957i \\ -1.0695 - 2.0473i & -0.1948 - 0.4603i \end{pmatrix}.$$

The first six iteration errors $e_k = \|A - X_k\|$ are shown in the left hand table, while the three columns on the right with contain the quantities e_{k+1}/e_k^p for $p = 1$, $p = \frac{1+\sqrt{5}}{2}$, and $p = 2$, respectively:

0.5105
0.0902
0.0097
2.6432e-004
7.4437e-007
5.5709e-011

0.367189	0.2995	0.2641
0.176712	0.2678	0.3462
0.107263	0.4744	1.1890
0.027317	0.4801	2.8230
0.002816	0.4580	10.654
0.000075	0.4588	100.54

Remark 4. The figures indicate a superlinear convergence: the middle column on the right makes one to think that $p = \frac{1+\sqrt{5}}{2}$ is the exact order of convergence. Additional calculations in Maple show that, expanding $\xi_k(\tau)$ into Taylor series around $\tau = 0$, the first $2f_k$ terms of $\xi_k(\tau)$ and $\xi(\tau)$ coincide, where (f_i) is a Fibonacci sequence with initial values $f_0 = 1$, $f_1 = 1$. This confirms the guess, and, with regard to the well known property of the Newton iteration, raises the question: "why not quadratical?"

5. The case of general matrices

In this last section we briefly discuss the possibility of constructing a Newton-like iteration converging to the closest normal matrix of an arbitrary order. As we have seen, one has to characterize the subspace $\mathcal{L}^{hom}(X)$ (by determining a suitable basis), to choose a subspace in it (by dropping the "superfluous" basis elements), and to find the good coefficients ensuring convergence.

In case of $n = 3$ we tried the choice $\{I, X^*, (X^*)^2\}$ – a natural continuation of $\{I, X^*\}$ for $n = 2$ –, however couldn't find to them appropriate coefficients. It seems that the Krylov sequence does not suffice to this aim for $n \geq 3$. There are, however, further "generic" sequences of $O(n)$ members in $\mathcal{L}^{hom}(X)$, e.g.

$$X^{-1}(X^*X^{-1})^{k-1} + ((X^*)^{-1}X)^k(X^*)^{-1}, \quad k = 1, 2, \ldots$$

and

$$(X^*)^{-1}X^k(X^*)^{-1} + \sum_{j=1}^{k}(X^*)^{j-1}X^{-1}(X^*)^{k-j}, \quad k = 1, 2, \ldots$$

The special cases of both sequences for $k = 1$ coincide, and the matrix $Y = \varphi(X)$ obtained has an interesting property: φ is idempotent, or, in other words, it holds the *inversion formula:*

$$Y = X^{-1} + (X^*)^{-1}X(X^*)^{-1},$$

$$X = Y^{-1} + (Y^*)^{-1}Y(Y^*)^{-1},$$

which can be proved using the Sherman–Morrison–Woodbury formula for perturbed matrices.

We close the section by observing that discussing the subspace $\mathcal{L}^{hom}(X)$ is – according to the above – interesting in its own right, while it may contribute to finding a new iteration related to the best normal approximation problem.

References

[1] EBERLEIN, P. J.: A Jacobi-like method for automatic computation of eigenvalues and eigenvectors of an arbitrary matrix, *J. Soc. Indust. Appl. Math.* **10** (1) (1962), 74–88.

[2] GABRIEL, R.: Minimization of the Frobenius norm of a complex matrix using planar similarities, *Applied Numerical Mathematics* **40** (2002), 391–414.

[3] GABRIEL, R.: Zur besten normalen Approximation komplexer Matrizen in der Euklidischen Norm, *Mathematische Zeitschrift* **200** (1989), 591–600.

[4] HIGHAM, N. J.: Matrix Nearness Problems and Applications, in: Applications of Matrix Theory, M. J. C. Gover and S. Barnett (Eds.), pp. 1–27, Oxford University Press, Cambridge, 1985.

[5] RUHE, A.: Closest normal matrix finally found!, *BIT* **27** (1987), 585–598.

[6] HORN, R. A. and JOHNSON, C. R.: Matrix Analysis, Cambridge Univ. Press, 1985.

[7] http://mathworld.wolfram.com/CentralBinomialCoefficient.html