Numerical properties of Higham's method for polar decomposition

Andrzej Kiełbasiński (Warszawa, Poland)
Paweł Zieliński (Opole, Poland)
Krystyna Ziętak (Wrocłlaw, Poland)

$$A = UH$$

$A - n \times n$, complex, nonsingular

U - unitary

H - Hermitian positive definite

Algorithms

$$X_0 = A$$

$$\lim_{k \to \infty} X_k = U$$

$$H = U^H A = \frac{1}{2}(U^H A + A^H U)$$

Björck - Bowie 1971

Higham (Newton) 1986

Higham - Schreiber (Schulz) 1990

Gander (Halley) 1990

Higham - Papadimitriou (1994)

(parallel)

Singular value decomposition

$$A = P\Sigma Q^H, \quad n \times n$$

P, Q - unitary

$\Sigma = \text{diag}\,(\sigma_j)$

$$U = PQ^H, \quad H = Q\Sigma Q^H$$

$Higham$ 1986

$$X_0 = A$$

$$X_{k+1} = \frac{1}{2}(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H})$$

$\gamma_k - scaling\ parameters$

$$\gamma_k^{(opt)} = 1/\sqrt{\sigma_{max}(X_k)\sigma_{min}(X_k)}$$

$X_s = U$

s **number** *of* **distinct** $\sigma_j(A)$

Kenney and Laub

$$[\gamma_k^{(opt)}]^2 \leq \gamma_k \leq 1$$

$$Kenney,\ Laub\ 1992$$

$$\gamma_k^{(1,\infty)} = \sqrt[4]{\frac{||X_k^{-1}||_1\ ||X_k^{-1}||_\infty}{||X_k||_1\ ||X_k||_\infty}}$$

$$\gamma_k^{(F)} = \sqrt{\frac{||X_k^{-1}||_F}{||X_k||_F}}$$

$$\gamma_k = 1/\sqrt{a_k b_k}$$

$$0 < a_k \leq \sigma_j(X_k) \leq b_k$$

quasi-optimal $\gamma_k^{(q)}$

$$0 < a_0 \leq \sigma_j(A) \leq b_0$$

$$\gamma_0^{(q)} = \frac{1}{\sqrt{a_0 b_0}} \qquad \gamma_k^{(q)} = \frac{1}{\sqrt{\mu_k}}$$

$$\mu_0 = \frac{b_0}{a_0}, \quad \mu_{k+1} = \frac{1}{2}(\sqrt{\mu_k} + \frac{1}{\sqrt{\mu_k}})$$

$$\sigma_j(X_k) \in [1, \mu_k]$$

$$\mathrm{cond}_2(X_k) \leq \mu_k$$

$$\mu_{k+1} < \sqrt{\mu_k} < \mu_k$$

(**a**) $n = 20$, A - close to orthogonal matrix;

$\sigma_1 = 1.0001, \quad \sigma_{20} = 1;$

(**b**) $n = 20$,

$\sigma_i = 1$ for $i = 1, \ldots, 10$,

$\sigma_i = 2$ for $i = 11, \ldots 20$,

(**c**) $n = 20, \quad \sigma_i = i$

(**d**) $n = 20, \quad \sigma_i = i^4$

(**e**) $n = 20, \quad \sigma_i = 2^i$

(**f**) $n = 10, A = QR^8$

(**g**) $n = 10, A = LR^8$

(**h**) $n = 20, A$ - Hilbert matrix.

(f), (g) - Du Croz, Higham

condition numbers

	$\mathrm{cond}_2(A)$	$\kappa(U)$
(a)	1.0001	1.0
(b)	2	1.0
(c)	20	0.66
(d)	1.60×10^5	1.18×10^{-1}
(e)	5.24×10^5	3.33×10^{-1}
(f)	6.40×10^{13}	3.12×10^9
(g)	2.17×10^{14}	6.84×10^9
(h)	1.43×10^{18}	5.76×10^{17}

Herm. factor is well-conditioned

$$\mathrm{cond}(A) = ||A||_2 \; ||A^{-1}||_2$$

$$\kappa(U) = \frac{2}{\sigma_{n-1}(A) + \sigma_n(A)}$$

two smallest singular values

numbers of iterations for `HS-G`

	$\gamma_k^{(opt)}$	$\gamma_k^{(1,\infty)}$	$\gamma_k^{(q,o)}$	$\gamma_k^{(q,\infty)}$
(a)	3	1+2	1+2	1+2
(b)	3	3+2	3+2	4+3
(c)	6	5+2	4+3	5+2
(d)	8	6+2	6+2	6+2
(e)	8	6+2	6+2	6+2
(f)	9	7+3	7+2	7+2
(g)	9	7+3	7+3	6+3
(h)	10	8+2	9+2	8+2

stop criterion

$$||X_k - X_{k-1}||_1 \leq 10eps||X_{k-1}||_1$$

switch criterion

$$\gamma_k^{(1,\infty)}, \qquad ||X_k - X_{k-1}||_1 \leq 0.01$$

Error analysis of Higham's method

Acceptable factors from polar decomposition of A

$$||\hat{U}^H \hat{U} - I|| \le \varepsilon_0$$

$$\hat{H}_A := \frac{1}{2}(\hat{U}^H A + A^H \hat{U})$$

$\hat{H}_A$ - positive-definite

$$||A - \hat{U}\hat{H}_A|| \le \varepsilon_1 ||A||$$

$$X := \frac{1}{2}(Y + Y^{-H}) \to X_{k+1}$$

$$Y = \gamma_k X_k$$

Under some assumptions if an unitary matrix $\hat{U}$ and

$$H_X = \frac{1}{2}(\hat{U}^H X + X^H \hat{U})$$

are exact polar factors for a matrix close to X

$$X := \frac{1}{2}(Y + Y^{-H})$$

then $\hat{U}$ and

$$H_Y = \frac{1}{2}(\hat{U}^H Y + Y^H \hat{U})$$

are exact polar factors for a matrix close to Y.

Reverse induction

model of matrix inversion

G - numericaly computed Y^{-1}

$$G = \hat{Y}^{-1} + F, \quad \hat{Y} = Y + E$$

$$||E|| \leq \varepsilon_1 ||Y||, \quad ||F|| \leq \varepsilon_2 ||G||$$

right, left residuals

$$||YG - I|| \leq \varepsilon_3 ||Y||\ ||G||$$

$$||GY - I|| \leq \varepsilon_4 ||Y||\ ||G||$$

$$||E|| \leq \varepsilon_1 ||\hat{Y}||, \quad ||F|| \leq \varepsilon_2 ||\hat{Y}||$$

HS-G - Gauss elimination with partial pivoting

HS-QR - QR decomposition

HS-QRP - QR decomposition with column pivoting

$$X_{k+1} = \frac{1}{2}(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H})$$

$$X_k = Q_k R_k$$

$$X_{k+1} = \frac{1}{2} Q_k [\gamma_k R_k + \frac{1}{\gamma_k} R_k^{-H}]$$

$\gamma_k^{(1,\infty)} -$ $\quad R_k$ instead of X_k

	$\frac{\|\|A-UH\|\|_F}{\|\|A\|\|_F}$
(e) $\sigma_i = 2^i$	$n = 20$
HS-G	5.63×10^{-16}
HS-QR	7.53×10^{-16}
HS-QRP	8.64×10^{-16}
(f) $A = QR^8$	$n = 10$
HS-G	2.34×10^{-07}
HS-QR	1.64×10^{-08}
HS-QRP	4.58×10^{-16}
(g) $A = LR^8$	H was not positive-def. $n = 10$
HS-G	1.51×10^{-07}
HS-QR	2.44×10^{-08}
HS-QRP	5.29×10^{-16}
(h) Hilbert	$n = 20$
HS-G	1.59×10^{-13}
HS-QR	8.35×10^{-15}
HS-QRP	8.17×10^{-15}

$$\hat{H}_j = (1/2)(\hat{U}^T X_j + X_j^T \hat{U})$$

$$\alpha_j = ||X_j - \hat{U}\hat{H}_j||_F / ||X_j||_F$$

$$c_j = \mathrm{cond}_2(X_j)$$

$$r_k = \frac{||X_k G_k - I||_F}{||G_k||_F \; ||X_k||_F}, \quad l_k = \frac{||G_k X_k - I||_F}{||G_k||_F \; ||X_k||_F},$$

Additional results for matrix
$A = QR^8$ and `HS-QR` with $\gamma_k^{(R)}$

c_k	α_k	r_k	l_k
10^{13}	1.6×10^{-08}	4×10^{-19}	1.5×10^{-08}
10^{6}	4.2×10^{-16}	1.6×10^{-17}	1.2×10^{-18}
10^{2}	3.5×10^{-18}	1.5×10^{-17}	8.3×10^{-18}
8.81	4.4×10^{-16}	2.3×10^{-17}	1.8×10^{-17}
1.68	3.3×10^{-16}	2.8×10^{-17}	3.4×10^{-17}
1.03	3.4×10^{-16}	2.2×10^{-18}	3.2×10^{-18}

Additional results for matrix $A = LR^8$ and HS-G with $\gamma_k^{(1,\infty)}$

c_k	α_k	r_k	l_k
10^{14}	1.5×10^{-07}	8.9×10^{-19}	1.6×10^{-07}
10^{6}	4.0×10^{-14}	1.7×10^{-17}	2.1×10^{-14}
10^{2}	5.9×10^{-16}	1.8×10^{-17}	1.4×10^{-15}
10^{1}	1.8×10^{-16}	3.5×10^{-17}	7.3×10^{-17}
2	2.1×10^{-16}	9.2×10^{-17}	9.2×10^{-17}

REMARK. Computed Hermitian factor of the matrix of A is not positive definite.

$$Y = \gamma_k X_k,$$

$$G = \hat{Y}^{-1} + F$$

$$\hat{X} = \hat{Y}/\gamma$$

$$X := \frac{1}{2}(Y + Y^{-H})$$

$$\tilde{X} = \frac{1}{2}(\hat{Y} + \hat{Y}^{-H})$$

$$||X - \tilde{X}||_F \leq \varepsilon_3 ||\tilde{X}||_2$$

$$\sqrt{\rho} = \gamma/\gamma^{(opt)}(\hat{X})$$

$$C = \max\{\rho, 1/\rho\} \operatorname{cond}_2(\hat{Y})$$

Main lemma

If $\hat{U}$ - unitary, X, Y - given

$$\hat{Y} = Y + \Delta Y, \quad ||\Delta Y||_F \leq \varepsilon_1 ||\hat{Y}||_2$$

$$\tilde{X} = (1/2)(\hat{Y} + \hat{Y}^{-H})$$

$$H_X = \frac{1}{2}(\hat{U}^H X + X^H \hat{U}) \quad - psd$$

$$||X - \tilde{X}||_F \leq \varepsilon_2 ||\tilde{X}||_2$$

$$||X - \hat{U} H_X||_F \leq \varepsilon_3 ||X||_2$$

$$[\varepsilon_1 + \varepsilon_2 + \varepsilon_3(1 + \varepsilon_2)C] < 1$$

$$\epsilon_2 + \epsilon_3 < 0.004$$

$$\sigma_{min}(\hat{Y}) = \max\{1, \rho\} C^{-1/2}$$

$$\sigma_{max}(\hat{Y}) = \min\{1, \rho\} C^{1/2}$$

then

$$H_Y = \frac{1}{2}(\hat{U} Y^H + Y \hat{U}^H) \quad \text{is positive def.}$$

$$||Y - \hat{U} H_Y||_F \leq \varepsilon_4 ||Y||_2$$

$$\varepsilon_4 = \frac{\varepsilon_1}{1 - \varepsilon_1} + \frac{\varepsilon_2 + \varepsilon_3(1 + \varepsilon_2)}{\min\{1, \rho\}(1 - \varepsilon_1)} \times$$

$$[1 + 3(\varepsilon_2 + \varepsilon_3 + \varepsilon_2 \varepsilon_3)\sqrt{C}]$$

$$\hat{U} := X_l, \quad X_l - \text{computed by } \texttt{HS}$$

$$\hat{H}_k := \frac{1}{2}(\hat{U}^H X_k + X_k^H \hat{U})$$

$$k = l, l-1, \ldots, 0$$

Reverse induction