Algorithms for polar decomposition and applications

Krystyna Ziętak

Wrocław University of Technology,
Institute of Mathematics and Computer Science

coauthors:
Andrzej Kiełbasiński (Warsaw),
Beata Laszkiewicz (Wrocław),
Paweł Zieliński (Wrocław)

Basic papers

1. **PZ, KZ**, The polar decomposition - properties, applications and algorithms, *Applied Mathematics, Annals of Polish Math Soc.* 38 (1995), 23–49.
2. **AK, KZ**, Numerical behaviour of Higham's scaled method for polar decomposition, *Numerical Algorithms* 32 (2003), 105–140.
3. **BL, KZ**, Approximation of matrices and family of Gander methods for polar decomposition, *BIT Numer. Math.*, on line first 3 May 2006.
4. **AK, PZ, KZ**, Numerical experiments with Higham's scaled method for polar decomposition, *Numerical Algorithms*, submitted.

Some relevant papers

1. **W. Gander**, Algorithms for polar decomposition, *SIAM J. Sci. Stat. Comput.* 11 (1990), 1102–1115.
2. **N.J. Higham**, Computing the polar decomposition - with applications, *SIAM J. Sci. Stat. Comput.* 7 (1986), 1160–1173.
3. **Ch. Kenney, A.J. Laub**, On scaling Newton's method for polar decomposition and the matrix sign function, *SIAM J. Matrix Anal. Appl.* 13 (1992), 688–706.
4. **P.J. Maher**, Partially isometric approximation of positive operators, *Illinois J. Math.* 33 (1989), 227–243.

Polar decomposition

$$A = UH$$

$$A \in \mathbb{C}^{n\times n}, \quad \text{nonsingular}$$

U - unitary, H - Hermitian positive definite

Generalized polar decomposition

$$A = EH$$

$$A \in \mathbb{C}^{m\times n}$$

E - subunitary, H - Hermitian positive semidefinite

Subunitary matrices

$$||Ex||_2 = ||x||_2, \qquad x \in \mathrm{range}(E^H)$$

Equivalent conditions:

- $EE^HE = E$
- $E^H = E^\dagger$ Moore-Penrose inverse
- EE^H is an orthogonal projector

Outline

1. Perturbation bounds for polar factors
2. Applications of polar factors
3. Family of Gander methods
4. Higham's scaled method
5. Algorithms for approximation by subunitary matrices
6. Algorithms for smaller rank approximation
7. Higham's method - rounding error analysis
8. Numerical experiments

Singular value decomposition of A

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

$$P, Q \text{ - unitary }, \qquad \Sigma = \mathrm{diag}\,(\sigma_j)$$

Polar decomposition

$$A = UH = (PQ^H)(Q\Sigma Q^H)$$

If $\mathrm{rank}(A) = n$ then U is unique

Generalized polar decomposition

$$A = EH$$

$$E = P\mathrm{diag}(I_r, I_k, 0)Q^H, \qquad r = \mathrm{rank}(A)$$

Iterative Algorithms for $A = UH$

$$X_0 = A, \qquad \lim_{k\to\infty} X_k = U$$

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

Björck - Bowie 1971, Higham (Newton) 1986,
Higham - Schreiber (Schulz iterations) 1990,
Gander (Halley) 1990,
Higham - Papadimitriou (parallel) 1994,
Higham, Mackey, Tisseur - 2004
(structure preserving in matrix group)

Perturbation bounds of polar factors

Higham 1986, Barrlund 1989;
Kenney, Laub 1991, Mathias 1993
Ren-Cang Li 1995, Chatelin, Gratton 2000;
Wen Li, Weiwei Sun 2002

$$A = UH, \quad A_\Delta = U_\Delta H_\Delta = A + \Delta, \qquad A, A_\Delta \quad \textbf{nonsingular}$$

$$||H - H_\Delta||_F \leq \sqrt{2}||\Delta||_F$$

$$||U - U_\Delta|| \leq \frac{2}{\sigma_{\min}(A) + \sigma_{\min}(A_\Delta)} \, ||\Delta||$$

unitarily invariant norms

Absolute condition numbers

Unitary polar factor U

$$\kappa(U) = \lim_{\delta \to 0} \sup_{||\Delta||_F \leq \delta} \frac{||U_A - U_{A+\Delta}||_F}{\delta}$$

$$\kappa(U) = \frac{1}{\sigma_n(A)}$$

A complex and $m \geq n$;

A real and $m > n$

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

A real and $m = n$

two smallest $\sigma_j(A)$)

Absolute condition numbers

Hermitian polar factor H

$$\frac{\sqrt{2(1+\mathrm{cond}(A)^2)}}{1+\mathrm{cond}(A)}$$

A complex or real, $m \geq n$

$$\mathrm{cond}(A) = \sigma_1(A)/\sigma_n(A)$$

Perturbation of subunitary polar factors

$$A = EU, \qquad E - \text{subunitary}, \quad r = \text{rank}(A)$$

$$A + \Delta, \qquad \text{rank}(A + \Delta) = r$$

$$||E_A - E_{A+\Delta}||_F \leq \frac{2}{\sigma_r(A) + \sigma_r(A + \Delta)} ||\Delta||_F$$

Wen Li, Weiwei Sun 2002

Applications of polar factors $A = UH$

Approximation by unitary matrices

$$||A - U|| = \min_{Z-unitary} ||A - Z||$$

Fan, Hoffman 1955

$|| \cdot || -$ *unitarily invariant*

Orthogonal Procrustes problem

$$||A - BU||_F \leq ||A - BZ||_F \leq ||A + BU||_F$$

Z unitary

Applications of polar factors $A = UH$

Approximation by unitary matrices

$$||A - U|| = \min_{Z-unitary} ||A - Z||$$

Fan, Hoffman 1955

$|| \cdot || - unitarily\ invariant$

Orthogonal Procrustes problem

$$||A - BU||_F \leq ||A - BZ||_F \leq ||A + BU||_F$$

Z unitary

Applications of polar factors $A = UH$

Approximation by positive definite matrices

$$||A - C|| = \min_{X-positive} ||A - X||$$

If A - Hermitian then $C = \frac{1}{2}(A + H)$ where $A = UH$ (unitarily invariant norm)

Positive definite square root $B^{1/2}$

$$B = LL^H, \quad \text{(Cholesky)}, \quad L = UH \quad \text{(polar decomposition)}$$

$$B^{1/2} = H$$

Higham 1986

Applications of polar factors $A = UH$

Approximation by positive definite matrices

$$||A - C|| = \min_{X-positive} ||A - X||$$

If A - Hermitian then $C = \frac{1}{2}(A + H)$ where $A = UH$ (unitarily invariant norm)

Positive definite square root $B^{1/2}$

$$B = LL^H, \quad \text{(Cholesky)}, \quad L = UH \quad \text{(polar decomposition)}$$

$$B^{1/2} = H$$

Higham 1986

Approximation of $A \in \mathbb{C}^{m\times n}$ by subunitary matrices

$$A = P\Sigma Q^H,$$

$r = \mathrm{rank}(A)$,

q number $\sigma_j(A)$ bigger or equal to $\frac{1}{2}$

Theorem

Let $A \in \mathbb{C}^{m\times n}$ and let $||\cdot||$ be arbitrary unitarily invariant norm. Then

- *for all orthonormal matrices E, $E^H E = I$, we have $||A - \tilde{E}|| \leq ||A - E||$, where $\tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H$,*

Theorem-cont.

- *for all subunitary matrices E of rank $r = \mathrm{rank}(A)$ we have $||A - \hat{E}|| \le ||A - E||$, where $\hat{E} = P \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^H$*
- *for all subunitary matrices E we have $||A - \hat{X}|| \le ||A - E|| \le ||A + \tilde{E}||$, where*

$$\hat{X} = P \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q^H, \quad \tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H.$$

Ky Fan, Hoffman 1955 - unitary matrices
Maher 1989 - c_p norms, subunitary
Sun, Chen 1989 - Frobenius norm, subunitary
Laszkiewicz, Ziętak 2006 - generalization

Theorem-cont.

- *for all subunitary matrices E of rank $r = \mathrm{rank}(A)$ we have* $||A - \hat{E}|| \leq ||A - E||$, *where* $\hat{E} = P \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^H$
- *for all subunitary matrices E we have* $||A - \hat{X}|| \leq ||A - E|| \leq ||A + \tilde{E}||$, *where*

$$\hat{X} = P \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q^H, \quad \tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H.$$

Ky Fan, Hoffman 1955 - unitary matrices
Maher 1989 - c_p norms, subunitary
Sun, Chen 1989 - Frobenius norm, subunitary
Laszkiewicz, Ziętak 2006 - generalization

Theorem-cont.

- *for all subunitary matrices E of rank $r = \mathrm{rank}(A)$ we have $||A - \hat{E}|| \leq ||A - E||$, where $\hat{E} = P \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^H$*
- *for all subunitary matrices E we have $||A - \hat{X}|| \leq ||A - E|| \leq ||A + \tilde{E}||$, where*

$$\hat{X} = P \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q^H, \quad \tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H.$$

Ky Fan, Hoffman 1955 - unitary matrices
Maher 1989 - c_p norms, subunitary
Sun, Chen 1989 - Frobenius norm, subunitary
Laszkiewicz, Ziętak 2006 - generalization

Family of Gander methods

for computing orthonormal polar factor $\tilde{E}$ of rectangular A of full rank n

$$X_{k+1} = X_k \left((2f-3)I + X_k^H X_k\right) \left((f-2)I + f X_k^H X_k\right)^{-1}$$

$X_0 = A, \quad f$ - parameter, $f \neq 1$

$f = 1$ Björck, Bowie

$f = 2$ unscaled Higham's method

X_k tends to $\tilde{E}$ (orthonormal polar factor) , but for some f not for every A

Properties of Gander's method

Newton's method for scalar equation

$$(s^2)^{\nu/2}(1 - s^2) = 0, \qquad \nu = \frac{2 - f}{f - 1}$$

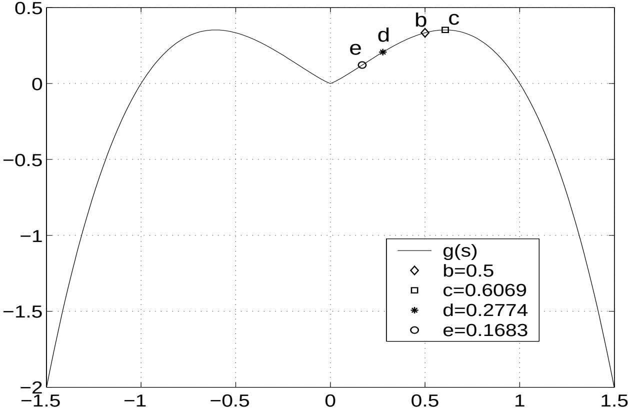
$$b = \sqrt{\frac{5 - 3f}{1 + f}}, \qquad c = \sqrt{\frac{2 - f}{f}}$$

$$1 < f < 5/3, \qquad [0, b), \qquad (b, c), \qquad (c, \infty)$$

For $f = 19/13$ we have $b = 1/2$. If, for example, A has some singular values in (b, c) then the sequence X_k can not tend to $\hat{X}$ in some cases.

an error in Gander's paper

$$g(s) = (s^2)^{\nu/2}(1 - s^2) = 0, \quad \nu = \frac{2-f}{f-1}, \quad f = \frac{19}{13}$$



Higham's method, 1986

$$X_{k+1} = \frac{1}{2}(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H}), \qquad X_0 = A$$

Optimal scaling: $\gamma_k^{(opt)} = \frac{1}{\sqrt{\sigma_{max}(X_k)\sigma_{min}(X_k)}}$

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

Interpretation (for $\gamma_k = 1$):

Newton's method applied to scalar equation $1 - s^2 = 0$ with initial point $s_0 = \sigma_j(A)$

Theoretical properties of Higham method

$$X_0 = A = UH$$

- U is common unitary factor of all X_k, $k = 0, 1, \ldots$
- Fast reduction of $\mathrm{cond}_2\,(X_k)$:

$$\mathrm{cond}_2(X_{k+1}) \leq \max\left\{\rho_k, \frac{1}{\rho_k}\right\}\sqrt{\mathrm{cond}_2(X_k)}$$

where $\quad \rho_k = \frac{\gamma_k}{\gamma_k^{(opt)}}$

Convergence of Higham's method

stop criterion: $||X_k - X_{k-1}||_1 \leq \delta ||X_{k-1}||_1$

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

Kenney, Laub 1992:

- Theoreticaly $X_s = U$ where s number of distinct $\sigma_j(A)$
- If $\left(\gamma_k^{(opt)}\right)^2 \leq \gamma_k \leq 1$ then faster convergence than for $\gamma_k = 1$

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

Convergence of Higham's method

stop criterion: $||X_k - X_{k-1}||_1 \leq \delta ||X_{k-1}||_1$

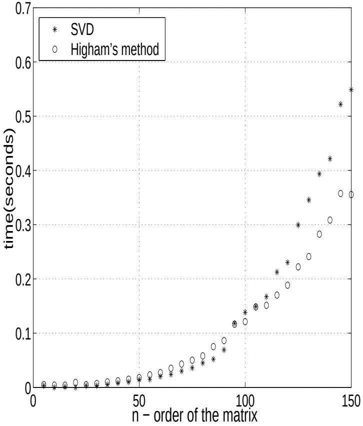
switch criterion: $\gamma_k^{(1,\infty)}, \qquad ||X_k - X_{k-1}||_1 \leq 0.01$

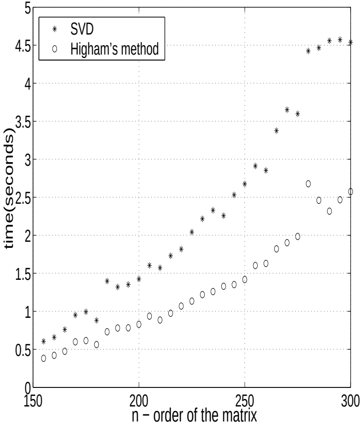
Kenney, Laub 1992:

- Theoreticaly $X_s = U$ where s number of distinct $\sigma_j(A)$
- If $\left(\gamma_k^{(opt)}\right)^2 \leq \gamma_k \leq 1$ then faster convergence than for $\gamma_k = 1$

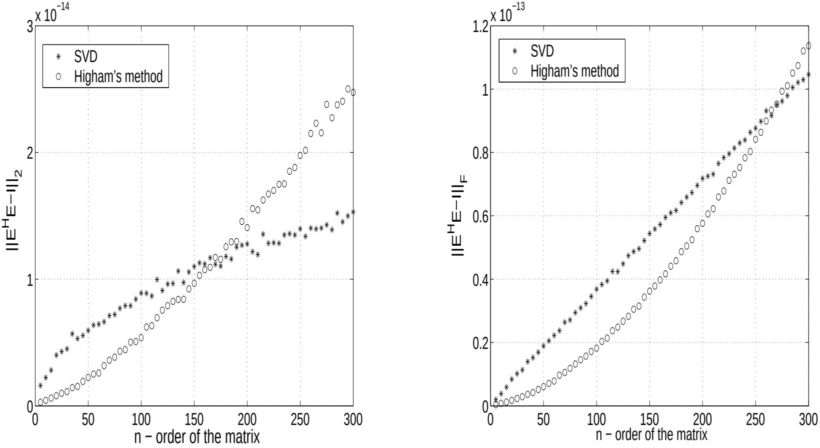
$$\gamma_k^{(F)} = \sqrt{\frac{||X_k^{-1}||_F}{||X_k||_F}} \qquad \textit{satisfies}$$

Average time of computing the unitary polar factor E (using `cputime`)





Average unitarity of the computed unitary polar factor E



Approximation by subunitary matrices

Algorithm I:

$\hat{X}$ is computed directly from the SVD of A

Algorithm II:

$\hat{X}$ is the limit of the sequence $X_k, X_0 = A$, generated by Gander's method with $f = 19/13$

Algorithm III:

Stage 1: computing orthonormal polar decomposition $A = EH$ (E orthonormal)
Stage 2: computing unitary polar factor E_C of $C = 2H - I$
Stage 3: computing $\hat{X} = \frac{1}{2}E(E_C + I_n)$

Approximation by subunitary matrices

Algorithm I:

$\hat{X}$ is computed directly from the SVD of A

Algorithm II:

$\hat{X}$ is the limit of the sequence $X_k, X_0 = A$, generated by Gander's method with $f = 19/13$

Algorithm III:

Stage 1: computing orthonormal polar decomposition $A = EH$ (E orthonormal)
Stage 2: computing unitary polar factor E_C of $C = 2H - I$
Stage 3: computing $\hat{X} = \frac{1}{2}E(E_C + I_n)$

Approximation by subunitary matrices

Algorithm I:

$\hat{X}$ is computed directly from the `SVD` of A

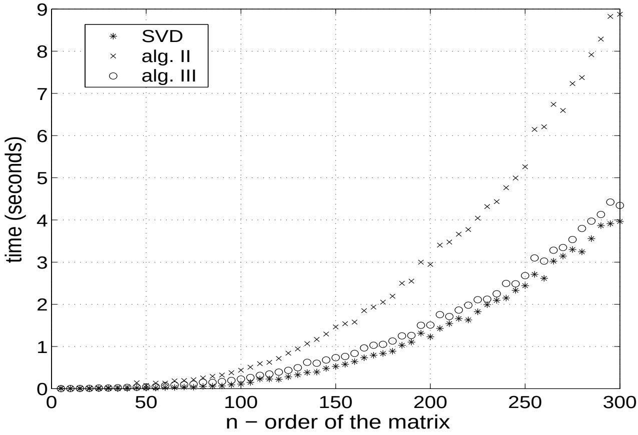
Algorithm II:

$\hat{X}$ is the limit of the sequence $X_k, X_0 = A$, generated by Gander's method with $f = 19/13$

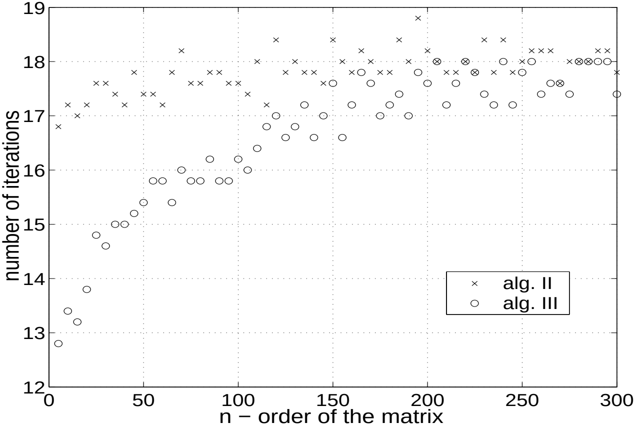
Algorithm III:

Stage 1: computing orthonormal polar decomposition $A = EH$ (E orthonormal)
Stage 2: computing unitary polar factor E_C of $C = 2H - I$
Stage 3: computing $\hat{X} = \frac{1}{2}E(E_C + I_n)$

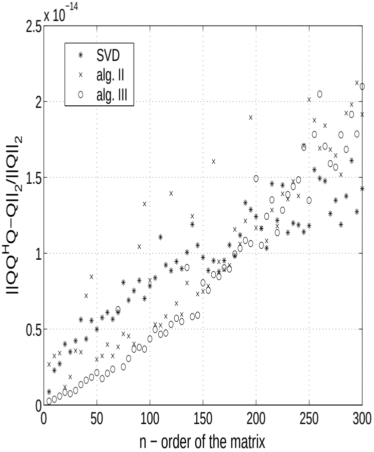
computing best subunitary approximant: average time

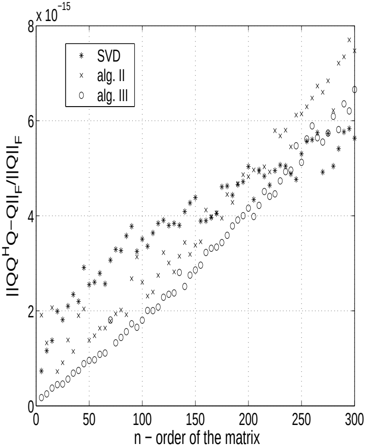


computing best subunitary approximant: average number of iterations



computing best subunitary approximant: average unitarity





Minimal rank approximation $A \in \mathbb{C}^{m\times n}$

$$\min_{B \text{ minimal rank}} ||A - B||_2 < \delta,$$

δ given, Golub 1968

Algorithm IV

- computing Hermitian polar factor H of A
- computing unitary polar factor E_D of $D = H - \delta I$
- computing $\hat{B} = \frac{1}{2}A(E_D + I)$

Minimal rank approximation $A \in \mathbb{C}^{m\times n}$

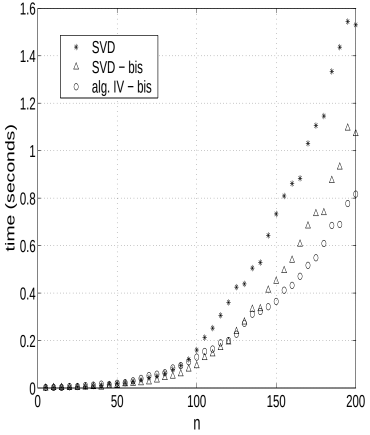
Algorithm IV-bis

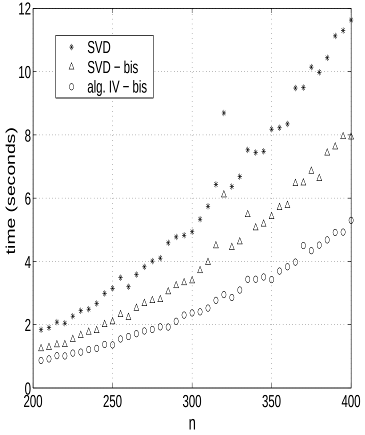
- computing unitary polar factor E of $A^H A - \delta^2 I$
- computing $\hat{B} = \frac{1}{2}A(E + I)$

- **SVD**: computing $\hat{B}$ by means `SVD` applied to A
- **SVD-bis**: computing $\hat{B}$ by means `SVD` applied to $A^H A$

Numerical tests for rectangular A, $2n \times n$

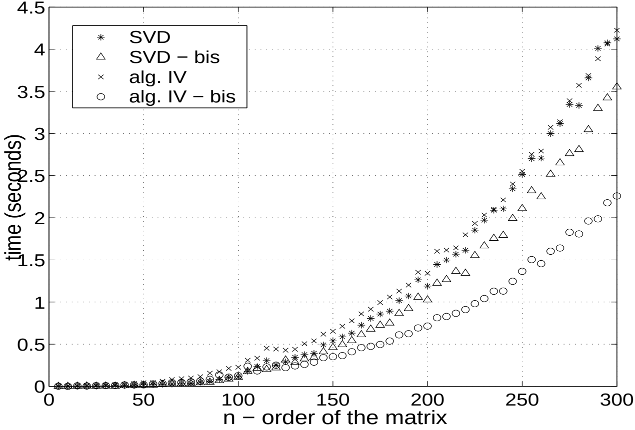
minimal rank approximant: average time





Numerical tests for square A

average time of computing minimal rank approximant



Rounding error analysis of Higham's method

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

Acceptable polar factors U and H of A computed in fl, ($\mu = 2^{-t}$) (A nonsingular)

$$\hat{U} := X_l, \qquad \hat{H} := \frac{1}{2}\left(\hat{U}^H A + A^H \hat{U}\right)$$

$$||\hat{U}^H \hat{U} - I|| \le \varepsilon_1, \qquad ||A - \hat{U}\hat{H}_A|| \le \varepsilon_2 ||A||$$

$\hat{H}_A$ - positive-definite,
ε_i modest multiple of 2^{-t}

Numerical correctness - **NC property**

G - numericaly computed X^{-1}: $\quad G = (X + \Delta X)^{-1} + \Delta G$

$$||\Delta X|| \leq \varepsilon_1 ||X||, \quad ||\Delta G|| \leq \varepsilon_2 ||G||$$

Remark:

In the proofs we use **SVD** of $\widetilde{X} = X + \Delta$

Relative right and left residuals

$$rr = \frac{||XG - I||}{||X||\ ||G||}, \qquad lr = \frac{||GX - I||}{||X||\ ||G||}$$

$$lr \le \varepsilon \quad \Rightarrow \quad rr \le \varepsilon \operatorname{cond}(X),$$
$$rr \le \varepsilon \quad \Rightarrow \quad lr \le \varepsilon \operatorname{cond}(X)$$

$lr \le \varepsilon$ **or** $rr \le \varepsilon \Rightarrow$ **numer. stability** :

$$||X^{-1} - G|| \le \varepsilon \operatorname{cond}(X)\ ||G||$$

NC property of computed inverse G

$$G = (X + \Delta X)^{-1} + \Delta G$$

NC $\Rightarrow$ rr and lr small $\Rightarrow$ **numer. stability**

Wilkinson's conjecture for inversion via GEPP (1962):

both rr **and** lr **small** $\Rightarrow$ NC **property**

Main lemma (backward induction)

Under some assumptions if

- $\widetilde{U}, \widetilde{H}_{k+1}$ are acceptable polar factors of $\widetilde{X}_{k+1}$,
- G_k (computed inverse) has NC property

then $\widetilde{U}, \widetilde{H}_k$ are acceptable polar factors for $\widetilde{X}_k$, where

$$\widetilde{H}_k := \frac{1}{2}\left(\widetilde{U}^H \tilde{X}_k + \tilde{X}_k^H \widetilde{U}\right)$$

Interpretation of main lemma

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

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

are exact polar factors for a matrix close to X then $\hat{U}$ and

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

are exact polar factors for a matrix close to Y.

$$Y = \gamma_k X_k, \qquad X = X_{k+1} = \frac{1}{2}\left(Y + Y^{-H}\right)$$

Conclusions from rounding error analysis and experiments (Higham's method)

1. Matrix inversion should yield **NC property** (**GECP**).
2. Using **GEPP** can fail for some A - poor unitarity of unitary polar factor.
3. γ_k distinctly smaller or large then optimal-ones can spoil convergence and quality computed unitary polar factor.
4. If we apply $\gamma_k^{(1,\infty)}$ or $\gamma_k^{(F)}$ then practically good matrix inversion guarantees good quality of computed polar factor (if A is not too ill conditioned).
5. With stopping criterion proposed by Higham frequently one redundant iteration is performed.

Stopping criteria

- **Higham**: $||X_{k+1} - X_k||_1 \leq \delta_n ||X_k||_1$ for $\delta_n = 2^{2-t}$
- **AK. KZ.**: $\beta_k \equiv ||X_k - G_k^H||_F \leq \sqrt{2^{1-t} n^{1/2}}$

achieving acceptable limiting accuracy

Switching to unscaled iterations

- **Higham**: $||X_k - X_{k-1}||_1 \leq 0.01$
- **AK, KZ**: $\gamma_k^{(1,\infty)}$ and $\beta_k \leq 1.5$ or $\beta_k \geq \beta_{k-1}$

cautiousness

Example: smallness of both residuals is not sufficient property of computed inverse

$$X_0 = \mathrm{diag}(c, \sqrt{c}, \sqrt{c}, 1), \quad c = \mathrm{cond}_2(X_0) \quad \gamma_0 = \gamma^{(opt)}(X_0) = \frac{1}{\sqrt{c}}$$

$$X_1 = U_1 H_1 \ \text{without rounding errors for}\, G_0,$$

where $G_0 = X_0^{-1} + \epsilon\sqrt{c}(e_2 e_3^T - e_3 e_2^T) \quad (\epsilon \approx 2^{-t})$

left and right relative residuals are small for $\mathbf{G_0}$!!!

but exact orthogonal factor $\tilde{U} = U_1$ of X_1 is not good for X_0

$$\widetilde{H}_0 = \frac{1}{2}\left(\widetilde{U}^T X_0 + X_0^T \widetilde{U}\right) \quad \text{is PSD,} \quad \frac{||X_0 - \widetilde{U}\widetilde{H}_0||_F}{||X_0||_2} = \frac{\epsilon\sqrt{c}}{(\sqrt{2}p)}$$

Test matrices for both residuals small

$$A = P\mathrm{diag}(\sigma_j)Q^H, \qquad P, Q \quad \text{random orthogonal}$$

$$c_k = \mathrm{cond}(X_k)$$

m_k number singular values of X_k close to $\frac{1}{\gamma_k^{(opt)}}$

$$n = 20, \quad m_0 = 18, \qquad \{\sigma_j\} = \{10^{14}, 10^7, 10^7, \ldots, 10^7, 1\}$$

$$\delta_k = \frac{||X_k - \tilde{U}H_k||_F}{||X_k||_F}, \qquad G_k = X_k + \Delta \text{ "computed" inverse}$$

$$c_2 = 1.07, \quad c_1 = 5.17e+06, \quad c_0 = 9.99e+13$$

$$\delta_2 = 1.742e-15, \quad \delta_1 = 1.72e-15, \quad \delta_0 = 7.04e-09$$

Scaling parameters

$$\rho_k = \left(\frac{\gamma_k}{\gamma_k^{(\mathrm{opt})}}\right)^2, \qquad \gamma_k^{(\mathrm{opt})} = \frac{1}{\sqrt{\sigma_{\max}(X_k)\sigma_{\min}(X_k)}}$$

$$\delta_k = \frac{||\tilde{X}_k - \tilde{U}\tilde{H}_k||_F}{||\tilde{X}_k||_2} = \alpha_k(\chi_{\mathbf{k}} + \beta_k)$$

- ρ_k too small are danger for accuracy
- but multipliers $\chi_{\mathbf{k}}$ can act soothingly!!!

Scaling parameters

$$\rho_k = \left(\frac{\gamma_k}{\gamma_k^{(\mathrm{opt})}}\right)^2, \qquad \gamma_k^{(\mathrm{opt})} = \frac{1}{\sqrt{\sigma_{\max}(X_k)\sigma_{\min}(X_k)}}$$

$$\delta_k = \frac{||\tilde{X}_k - \tilde{U}\tilde{H}_k||_F}{||\tilde{X}_k||_2} = \alpha_k(\chi_{\mathbf{k}} + \beta_k)$$

- ρ_k too small are danger for accuracy
- but multipliers $\chi_{\mathbf{k}}$ can act soothingly!!!

Scaling parameters

$$\rho_k = \left(\frac{\gamma_k}{\gamma_k^{(\mathrm{opt})}}\right)^2, \qquad \gamma_k^{(\mathrm{opt})} = \frac{1}{\sqrt{\sigma_{\max}(X_k)\sigma_{\min}(X_k)}}$$

$$\delta_k = \frac{||\tilde{X}_k - \tilde{U}\tilde{H}_k||_F}{||\tilde{X}_k||_2} = \alpha_k(\chi_{\mathbf{k}} + \beta_k)$$

- ρ_k too small are danger for accuracy
- but multipliers $\chi_{\mathbf{k}}$ can act soothingly!!!

Influence of ρ_k and χ_k on accuracy of computed polar decomposition

$$n = 10, \quad A = \mathrm{tril}(\mathrm{rand}(10))^8 \mathrm{rand}(R)$$

$$R - \text{upper triangular random}$$

k	c_k	ρ_k	δ_k	$\hat{\chi}_k$
0	$8.75e+14$	$8.27e-04$	$5.82e-13$	0.078
1	$4.35e+08$	$1.19e-03$	$6.09e-15$	0.036
2	$2.65e+05$	$1.11e-03$	$1.90e-14$	0.026
3	$6.00e+03$	$9.44e-04$	$7.96e-15$	0.041
4	$1.24e+03$	$1.12e+00$	$1.16e-16$	0.431
5	$1.51e+01$	$9.26e-01$	$1.69e-16$	0.720

- inverses computed by means of **GECP**
- special scaling parameters distinctly smaller than $\gamma_k^{(\mathrm{opt})}$ only in several initial iterations

Test matrices

(a) $n = 20, \sigma_i = 2^i,\ A = P\Sigma Q^T$,

(b) $n = 10, \mathbf{A} = \mathbf{QR^8}$

(c) $n = 10, \mathbf{A} = \mathbf{LR^8}$,

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

P, Q - random orth., L, R - random triang.

Conditions numbers

	$\mathrm{cond}_2(A)$
(a)	5.24×10^5
(b)	6.40×10^{13}
(c)	2.17×10^{14}
(d)	1.43×10^{18}

	$\kappa(U)$
(a)	3.33×10^{-1}
(b)	3.12×10^9
(c)	6.84×10^9
(d)	5.76×10^{17}

- **HS-G** - GEPP Gauss elimination
- **HS-QR** - *QR* decomposition
- **HS-QRP** - *QR* with column pivot.

Numbers of iterations for `HS-G`

	$\gamma_k^{(opt)}$	$\gamma_k^{(1,\infty)}$
(a)	8	6+2
(b)	9	7+3
(c)	9	7+3
(d)	10	8+2

$$\frac{||A - UH||_F}{||A||_F}$$

$\sigma_i = 2^i$	$n = 20$
HS-G	5.63×10^{-16}
HS-QR	7.53×10^{-16}
HS-QRP	8.64×10^{-16}

$A = QR^8$	$n = 10$
HS-G	2.34×10^{-07}
HS-QR	1.64×10^{-08}
HS-QRP	4.58×10^{-16}

Hilbert	$n = 20$
HS-G	1.59×10^{-13}
HS-QR	8.35×10^{-15}

$$A = LR^8 \text{ and } \texttt{HS-G} \text{ with } \gamma_k^{(1,\infty)}$$

c_k	δ_k	rr_k	lr_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}

Computed Hermitian factor of the matrix A
is not positive definite!!!

HS-G iterations with
$\gamma_k = \gamma_k^{(1,\infty)}$ for $k > 0$, $\gamma_0 = p\gamma_0^{(1,\infty)}$

	$\sigma_j = 2^j$		$A = QR^8$	
p	$\frac{\|\|A-UH\|\|_F}{\|\|A\|\|_F}$	iter	$\frac{\|\|A-UH\|\|_F}{\|\|A\|\|_F}$	iter
$1/20$	$2.792e-14$	7+2	$1.371e-6$	7+3
$1/10$	$1.008e-14$	6+3	$1.261e-6$	7+3
$1/5$	$3.599e-15$	7+2	$9.725e-7$	7+3
1	$5.633e-16$	6+2	$2.343e-7$	7+3
5	$5.201e-16$	6+3	$1.882e-8$	7+2
10	$4.892e-16$	6+3	$4.990e-9$	7+3

Remark: The notation $7+3$ means that 7 iteration was performed with scaling and 3 without scaling.