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


Some relevant papers


Polar decomposition

\[ A = UH \]

\[ A \in \mathbb{C}^{n \times n}, \text{ nonsingular} \]

\[ U - \text{unitary}, \quad H - \text{Hermitian positive definite} \]

Generalized polar decomposition

\[ A = EH \]

\[ A \in \mathbb{C}^{m \times n} \]

\[ E - \text{subunitary}, \quad H - \text{Hermitian positive semidefinite} \]
Subunitary matrices

\[ \|Ex\|_2 = \|x\|_2, \quad x \in \text{range}(E^H) \]

Equivalent conditions:

- \( EE^H E = E \)
- \( E^H = E^\dagger \) 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 $A$

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

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

Polar decomposition

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

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

Generalized polar decomposition

$$A = EH$$

$$E = P\text{diag}(l_r, l_k, 0)Q^H, \quad r = \text{rank}(A)$$
Iterative Algorithms for $A = UH$

\[ X_0 = A, \quad \lim_{k \to \infty} X_k = U \]

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

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

\[ A = UH, \quad A_\Delta = U_\Delta H_\Delta = A + \Delta, \quad A, A_\Delta \text{ nonsingular} \]

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

\[ \|U - U_\Delta\| \leq \frac{2}{\sigma_{\text{min}}(A) + \sigma_{\text{min}}(A_\Delta)} \|\Delta\| \]

unitarily invariant norms

Ren-Cang Li 1995, Chatelin, Gratton 2000; Wen Li, Weiwei Sun 2002
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 $H$

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

$A$ complex or real, $m \geq n$

$$\text{cond}(A) = \frac{\sigma_1(A)}{\sigma_n(A)}$$
Perturbation of subunitary polar factors

\[ A = EU, \quad E \text{ – subunitary,} \quad r = \text{rank}(A) \]

\[ A + \Delta, \quad \text{rank}(A + \Delta) = r \]

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

Wen Li, Weiwei Sun 2002
Applications of polar factors $A = UH$

**Approximation by unitary matrices**

\[
\|A - U\| = \min_{Z - \text{unitary}} \|A - Z\| 
\]

*Fan, Hoffman 1955*

\[\| \cdot \| - \text{unitarily invariant}\]

**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

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

*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**

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

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 by positive definite matrices

\[ ||A - C|| = \min_{X \text{ positive}} ||A - X|| \]

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

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

where

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

q \text{ number } \sigma_j(A) \text{ 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

$$\|A - \tilde{E}\| \leq \|A - E\|, \text{ where } \tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H,$$
Theorem-cont.

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

for all subunitary matrices $E$ we have $||A - \hat{X}|| \leq ||A - E|| \leq ||A + \tilde{E}||$, where

$$\hat{X} = P \begin{bmatrix} l_q & 0 \\ 0 & 0 \end{bmatrix} Q^H, \quad \tilde{E} = P \begin{bmatrix} l_n \\ 0 \end{bmatrix} Q^H.$$
for all subunitary matrices $E$ of rank $r = \text{rank}(A)$ 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 matrices
Sun, Chen 1989 - Frobenius norm, subunitary matrices
Laszkiewicz, Ziętak 2006 - generalization
Theorem-cont.

for all subunitary matrices $E$ of rank $r = \text{rank}(A)$ 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

$$\hat{X} = P \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q^H, \quad \tilde{E} = P \begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^H.$$
Family of Gander methods
for computing orthonormal polar factor $\tilde{E}$ of rectangular $A$ of full rank $n$

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

$x_0 = A$, $f$ - 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

\[(s^2)^{\nu/2}(1 - s^2) = 0, \quad \nu = \frac{2 - f}{f - 1}\]

\[b = \sqrt{\frac{5 - 3f}{1 + f}}, \quad c = \sqrt{\frac{2 - f}{f}}\]

\[1 < f < 5/3, \quad [0, b), \quad (b, c), \quad (c, \infty)\]

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

\[ X_{k+1} = \frac{1}{2} \left( \gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H} \right), \quad X_0 = A \]

Optimal scaling: \( \gamma^{(\text{opt})}_k = \frac{1}{\sqrt{\sigma_{\text{max}}(X_k)\sigma_{\text{min}}(X_k)}} \)

Practical scaling: \( \gamma^{(1,\infty)}_k = 4 \sqrt{\frac{\|X_k^{-1}\|_1 \|X_k^{-1}\|_\infty}{\|X_k\|_1 \|X_k\|_\infty}} \)

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 a common unitary factor of all \( X_k, k = 0, 1, \ldots \)
- Fast reduction of \( \text{cond}_2(X_k) \):

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

where \( \rho_k = \frac{\gamma_{k,\text{opt}}}{\gamma_k} \)
Convergence of Higham’s method

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

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

Kenney, Laub 1992:
- Theoretically \( X_s = U \) 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 \( \gamma_k = 1 \)
  \[ \gamma_k^{(F)} = \sqrt{\frac{\|X_k^{-1}\|_F}{\|X_k\|_F}} \] satisfies
Convergence of Higham’s method

**Stop criterion:** \[ \| X_k - X_{k-1} \|_1 \leq \delta \| X_{k-1} \|_1 \]

**Switch criterion:** \[ \gamma_k^{(1,\infty)}, \quad \| X_k - X_{k-1} \|_1 \leq 0.01 \]

Kenney, Laub 1992:
- Theoretically, \( X_s = U \) 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 \( \gamma_k = 1 \)

\[
\gamma_k^{(F)} = \sqrt{\frac{\| X_{k-1}^{-1} \|_F}{\| X_k \|_F}} \quad \text{satisfies}
\]
Average time of computing the unitary polar factor $E$ (using cputime)
Average unitarity of the computed unitary polar factor $E$

- $\|E^HE-I\|_2$

- $\|E^HE-I\|_F$

- SVD
- Higham's method

Spectral norm vs. Frobenius norm for $E$ as a function of the order of the matrix.
### 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 \( \hat{X} = \frac{1}{2}E(E_C + I_n) \)

Krystyna Ziętek

Algorithms for polar decomposition and applications
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

\[ A = EH \quad (E \text{ orthonormal}) \]

**Stage 2:** computing unitary polar factor \( E_C \) of \( C = 2H - I \)

**Stage 3:** computing \( \hat{X} = \frac{1}{2}E(E_C + I_n) \)

Krystyna Ziętak

Algorithms for polar decomposition and applications
computing best subunitary approximant: average time
computing best subunitary approximant: average number of iterations
computing best subunitary approximant: average unitarity

![Graphs comparing different algorithms for polar decomposition](image-url)
Minimal rank approximation $A \in \mathbb{C}^{m \times n}$

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

$\delta$ 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

![Graph showing the time (seconds) vs. n - order of the matrix for different algorithms. The graph includes data points for SVD, SVD-bis, alg. IV, and alg. IV-bis.]
Rounding error analysis of Higham’s method

\[ X_{k+1} = \frac{1}{2} \left( \gamma_k X_k + \frac{1}{\gamma_k} X^{-H} \right) \]

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

\[ \hat{U} := X_I, \quad \hat{H} := \frac{1}{2} \left( \hat{U}^H A + A^H \hat{U} \right) \]

\[ ||\hat{U}^H \hat{U} - I|| \leq \varepsilon_1, \quad ||A - \hat{U} \hat{H}_A|| \leq \varepsilon_2 ||A|| \]

\( \hat{H}_A \) - positive-definite,
\( \varepsilon_i \) modest multiple of \( 2^{-t} \)
Model of inversion

Numerical correctness - NC property

\[ G - \text{numerically computed } X^{-1}: \quad G = (X + \Delta X)^{-1} + \Delta G \]

\[ \|\Delta X\| \leq \varepsilon_1 \|X\|, \quad \|\Delta G\| \leq \varepsilon_2 \|G\| \]

Remark:

In the proofs we use SVD of \( \tilde{X} = X + \Delta \)
Relative right and left residuals

\[ rr = \frac{\|XG - I\|}{\|X\| \|G\|}, \quad lr = \frac{\|GX - I\|}{\|X\| \|G\|} \]

\[ lr \leq \varepsilon \implies rr \leq \varepsilon \cond(X), \]
\[ rr \leq \varepsilon \implies lr \leq \varepsilon \cond(X) \]

\[ lr \leq \varepsilon \mathrm{ or } \ rr \leq \varepsilon \implies \text{numer. stability :} \]
\[ \|X^{-1} - G\| \leq \varepsilon \cond(X) \|G\| \]
NC property of computed inverse $G$

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

NC $\Rightarrow$ \textit{rr} and \textit{lr} small $\Rightarrow$ numer. stability

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

both \textit{rr} and \textit{lr} small $\Rightarrow$ NC property
Main lemma (backward induction)

Under some assumptions if

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

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

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

\[
Y = \gamma_k X_k, \quad X = X_{k+1} = \frac{1}{2} \left( Y + Y^{-H} \right)
\]
Matrix inversion should yield **NC property (GECP)**.

Using **GEPP** can fail for some $A$ - poor unitarity of unitary polar factor.

$\gamma_k$ distinctly smaller or large then optimal-ones can spoil convergence and quality computed unitary polar factor.

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).

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.**: $\beta_k \equiv ||X_k - G^H_k||_F \leq \sqrt{2^{1-t} n^{1/2}}$

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

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

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

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

left and right relative residuals are small for \( G_0 \)

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

\[ \tilde{H}_0 = \frac{1}{2} \left( \tilde{U}^T X_0 + X_0^T \tilde{U} \right) \text{ is PSD,} \quad \frac{\|X_0 - \tilde{U} \tilde{H}_0\|_F}{\|X_0\|_2} = \frac{\epsilon \sqrt{c}}{(\sqrt{2}p)} \]
Test matrices for both residuals small

\[ A = P \text{diag}(\sigma_j) Q^H, \quad P, Q \text{ random orthogonal} \]

\[ c_k = \text{cond}(X_k) \]

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

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

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

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

\[ \delta_2 = 1.742e - 15, \quad \delta_1 = 1.72e - 15, \quad \delta_0 = 7.04e - 09 \]
Scaling parameters

\[ \rho_k = \left( \frac{\gamma_k}{\gamma_k^{(\text{opt})}} \right)^2, \quad \gamma_k^{(\text{opt})} = \frac{1}{\sqrt{\sigma_{\max}(X_k)\sigma_{\min}(X_k)}} \]

\[ \delta_k = \frac{\| \tilde{X}_k - \tilde{U} \tilde{H}_k \|_F}{\| \tilde{X}_k \|_2} = \alpha_k (\chi_k + \beta_k) \]

- \( \rho_k \) too small are danger for accuracy
- but multipliers \( \chi_k \) can act soothingly!!!
Scaling parameters

\[ \rho_k = \left( \frac{\gamma_k}{\gamma_k^{(\text{opt})}} \right)^2, \quad \gamma_k^{(\text{opt})} = \frac{1}{\sqrt{\sigma_{\text{max}}(X_k)\sigma_{\text{min}}(X_k)}} \]

\[ \delta_k = \frac{\|\tilde{X}_k - \tilde{U}\tilde{H}_k\|_F}{\|\tilde{X}_k\|_2} = \alpha_k (\chi_k + \beta_k) \]

- \( \rho_k \) too small are danger for accuracy
- but multipliers \( \chi_k \) can act soothingly!!!
Scaling parameters

\[ \rho_k = \left( \frac{\gamma_k}{\gamma_k^{(\text{opt})}} \right)^2, \quad \gamma_k^{(\text{opt})} = \frac{1}{\sqrt{\sigma_{\text{max}}(X_k)\sigma_{\text{min}}(X_k)}} \]

\[ \delta_k = \frac{||\tilde{X}_k - \tilde{U}\tilde{H}_k||_F}{||\tilde{X}_k||_2} = \alpha_k(\chi_k + \beta_k) \]

- \( \rho_k \) too small are danger for accuracy
- but multipliers \( \chi_k \) can act soothingly!!!
Influence of $\rho_k$ and $\chi_k$ on accuracy of computed polar decomposition

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

$R$ – upper triangular random

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

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

(a) \( n = 20, \sigma_i = 2^i, A = P\Sigma Q^T \),
(b) \( n = 10, A = QR^8 \)
(c) \( n = 10, A = LR^8 \),
(d) \( n = 20, A - \text{Hilbert matrix} \)

\( P, Q - \text{random orth., } L, R - \text{random triang.} \)

<table>
<thead>
<tr>
<th></th>
<th>( \text{cond}_2(A) )</th>
<th></th>
<th>( \kappa(U) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>( 5.24 \times 10^5 )</td>
<td>(a)</td>
<td>( 3.33 \times 10^{-1} )</td>
</tr>
<tr>
<td>(b)</td>
<td>( 6.40 \times 10^{13} )</td>
<td>(b)</td>
<td>( 3.12 \times 10^9 )</td>
</tr>
<tr>
<td>(c)</td>
<td>( 2.17 \times 10^{14} )</td>
<td>(c)</td>
<td>( 6.84 \times 10^9 )</td>
</tr>
<tr>
<td>(d)</td>
<td>( 1.43 \times 10^{18} )</td>
<td>(d)</td>
<td>( 5.76 \times 10^{17} )</td>
</tr>
</tbody>
</table>
- **HS-G** - GEPP Gauss elimination
- **HS-QR** - QR decomposition
- **HS-QRP** - QR with column pivot.

### Numbers of iterations for HS-G

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_k^{(\text{opt})}$</th>
<th>$\gamma_k^{(1,\infty)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>8</td>
<td>6+2</td>
</tr>
<tr>
<td>(b)</td>
<td>9</td>
<td>7+3</td>
</tr>
<tr>
<td>(c)</td>
<td>9</td>
<td>7+3</td>
</tr>
<tr>
<td>(d)</td>
<td>10</td>
<td>8+2</td>
</tr>
</tbody>
</table>
\[ \frac{\|A - UH\|_F}{\|A\|_F} \]

<table>
<thead>
<tr>
<th>$\sigma_i = 2^i$</th>
<th>$n = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-G</td>
<td>$5.63 \times 10^{-16}$</td>
</tr>
<tr>
<td>HS-QR</td>
<td>$7.53 \times 10^{-16}$</td>
</tr>
<tr>
<td>HS-QRP</td>
<td>$8.64 \times 10^{-16}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A = QR^8$</th>
<th>$n = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-G</td>
<td>$2.34 \times 10^{-07}$</td>
</tr>
<tr>
<td>HS-QR</td>
<td>$1.64 \times 10^{-08}$</td>
</tr>
<tr>
<td>HS-QRP</td>
<td>$4.58 \times 10^{-16}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hilbert</th>
<th>$n = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-G</td>
<td>$1.59 \times 10^{-13}$</td>
</tr>
<tr>
<td>HS-QR</td>
<td>$8.35 \times 10^{-15}$</td>
</tr>
</tbody>
</table>
$A = LR^8$ and HS-G with $\gamma_k^{(1, \infty)}$

<table>
<thead>
<tr>
<th>$c_k$</th>
<th>$\delta_k$</th>
<th>$rr_k$</th>
<th>$lr_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{14}$</td>
<td>$1.5 \times 10^{-07}$</td>
<td>$8.9 \times 10^{-19}$</td>
<td>$1.6 \times 10^{-07}$</td>
</tr>
<tr>
<td>$10^6$</td>
<td>$4.0 \times 10^{-14}$</td>
<td>$1.7 \times 10^{-17}$</td>
<td>$2.1 \times 10^{-14}$</td>
</tr>
<tr>
<td>$10^2$</td>
<td>$5.9 \times 10^{-16}$</td>
<td>$1.8 \times 10^{-17}$</td>
<td>$1.4 \times 10^{-15}$</td>
</tr>
<tr>
<td>$10^1$</td>
<td>$1.8 \times 10^{-16}$</td>
<td>$3.5 \times 10^{-17}$</td>
<td>$7.3 \times 10^{-17}$</td>
</tr>
<tr>
<td>2</td>
<td>$2.1 \times 10^{-16}$</td>
<td>$9.2 \times 10^{-17}$</td>
<td>$9.2 \times 10^{-17}$</td>
</tr>
</tbody>
</table>

Computed Hermitian factor of the matrix $A$ is not positive definite!!!
HS-G iterations with
\[ \gamma_k = \gamma_k^{(1, \infty)} \] for \( k > 0 \), \( \gamma_0 = p \gamma_0^{(1, \infty)} \)

| \( p \)  | \( \sigma_j = 2^j \) | \( \frac{||A-UH||_F}{||A||_F} \) | \( \frac{||A-UH||_F}{||A||_F} \) | \( \frac{||A-UH||_F}{||A||_F} \) | \( \frac{||A-UH||_F}{||A||_F} \) |
|-------|-------------------|------------------|------------------|------------------|------------------|
| 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.