Numerical Matrix Inversion

Krystyna Ziętak

Wrocław University of Technology,
Institute of Mathematics and Computer Science

coauthors:
Andrzej Kiełbasiński (Warsaw)
Paweł Zieliński (Wrocław)

Outline

- Models of matrix inversion
- Higham's method for polar decomposition
- Numerical experiments
- Rounding error analysis

Models of matrix inversion

Let G – computed inverse of nonsingular X

Numerical correctness – NC property

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

$$||\Delta X|| \leq \varepsilon_x ||X||, \quad ||\Delta G|| \leq \varepsilon_g ||G||$$

Numerical correct algorithm for A^{-1}

Byers, Xu (2008) – rounding error of bidiagonal reduction-based algorithm

- compute $A = UBV^H$ where U, V unitary, B bidiagonal
- solve $BY = U^H$
- compute $G = VY$.

too expensive

Models of matrix inversion – continuation

Numerical stability

$$||G - X^{-1}||_F \leqslant \varepsilon ||X||_2 ||G||_2$$

Left and right residual stability

$$||GX - I||_F \leqslant \varepsilon ||X||_2 ||G||_2$$

$$||XG - I||_F \leqslant \varepsilon ||X||_2 ||G||_2$$

Combined properties – **ALT** and **CONJ**

$\texttt{Alt} \stackrel{\text{df}}{=} \text{LRS}$ <u>**or**</u> RRS, **(left or right residual)**

$\texttt{Conj} \stackrel{\text{df}}{=} \text{LRS}$ <u>**and**</u> RRS, **(left and right residual)**

$$\texttt{NC} \Longrightarrow \texttt{Conj} \Longrightarrow \texttt{Alt} \Longrightarrow \texttt{NS}$$

$$||GX - I||_F \leqslant ||X||_2||G||_2||XG - I||_F$$
$$||XG - I||_F \leqslant ||X||_2||G||_2||GX - I||_F$$

$$\mathrm{cond}(X) = ||X||\ ||X^{-1}||$$

Remark. For small $||X||_2||G||_2$, say $\leqslant 10$, NS implies NC. Hence all listed properties of G can differ distinctly only when $\mathrm{cond}(X)$ is large.

Artificial example

$$X = \mathrm{diag}(c, \sqrt{c}, 1), \qquad G = X^{-1} + \Delta, \qquad |\Delta| \le Z$$

$$c > 1, \quad \varepsilon c \ll 1, \quad \varepsilon' = \frac{\varepsilon}{1 - \varepsilon c}$$

For the properties NC, LRS, RRS, Conj of G we obtain the following **upper bounds Z on elements of** $|\Delta|$:

$$Z_{\mathrm{NS}} = \varepsilon' \begin{bmatrix} c & c & c \\ c & c & c \\ c & c & c \end{bmatrix}, \quad Z_{\mathrm{LRS}} = \varepsilon' \begin{bmatrix} 1 & \sqrt{c} & c \\ 1 & \sqrt{c} & c \\ 1 & \sqrt{c} & c \end{bmatrix},$$

$$Z_{\mathrm{RRS}} = \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ \sqrt{c} & \sqrt{c} & \sqrt{c} \\ c & c & c \end{bmatrix}, \quad Z_{\mathrm{Conj}} = \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ 1 & \sqrt{c} & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix}.$$

Artificial example – continuation

Numerical correctness – NC property

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

$$||\Delta X|| \le \varepsilon_x ||X||, \quad ||\Delta G|| \le \varepsilon_g ||G||$$

$\varepsilon_x + \varepsilon_g + \varepsilon_x \varepsilon_g \le \varepsilon$

$$Z_{\mathrm{NC}} = \frac{\varepsilon_x}{1 - \varepsilon_x c} \begin{bmatrix} c^{-1} & c^{-1/2} & 1 \\ c^{-1/2} & 1 & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix} + \frac{\varepsilon_g}{1 - \varepsilon c} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$Z_{\mathrm{NC}} < \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix}.$$

Inverting matrix

J.J. Du Croz, N.J. Higham, Stability of methods for matrix inversion, *IMA J. Numer. Anal.* 12 (1992), 1–19.

$A = LU, \qquad X = A^{-1}$

Method A: solve
$Ax_j = e_j, \quad j = 1, \dots, n.$

Method B: compute U^{-1} and solve
$XL = U^{-1}$.

Method C: solve $UXL = I$.

Method D: compute L^{-1} and U^{-1},
form $A^{-1} = U^{-1}L^{-1}$.

GEPP – Gauss elimination with partial pivoting

Wilkinson

Wilkinson explained that computed *via* `GEPP` inverse G has **NC**-property (numerical correctenss) provided the triangular systems involved in `GEPP` are solved to high accuracy.

This happens frequently, but not always.

A-method

GAUSS with complete pivoting

G inverse of X computed by B-method *via* GECP

Then there exist Δ and Δ' such that

$$G + \Delta' = (X + \Delta)^{-1}, \qquad ||\Delta'|| \leq \varepsilon_g ||G||, \quad ||\Delta|| \leq \varepsilon_x ||X||$$

where ε_g is **practically** modest multiple computing precision (theoretically it depends on 2^n, n order of X).

AK, PZ, KZ, Higham's scaled method for polar decomposition and numerical matrix-inversion, *Report* P–045, Wrocław, July 2007.

QR with column pivoting

Polar decomposition

$$A = UH$$

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

U - unitary, H - Hermitian positive definite

Higham's method for polar decomposition

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

γ_k – scaling parameters

Interpretation (for $\gamma_k = 1$):

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

N.J. Higham, Computing the polar decomposition - with applications, *SIAM J. Sci. Stat. Comput.* 7 (1986), 1160–1173.

Optimal scaling:

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

Practical scaling

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

R. Byers, H. Xu, A new scaling for Newton's iteration for the polar decompositionand its backward stability, *SIAM J. Matrix Anal. Appl.* 30 (2008), 822–834.

New scaling

Let $a \leq \lambda_j(A) \leq b$ and $f(t) = (t + t^{-1})/2$.

$$\gamma_0 = \frac{1}{\sqrt{ab}}, \quad \gamma_2 = \sqrt{\frac{2\sqrt{ab}}{a+b}}, \quad \gamma_k = \frac{1}{\sqrt{f(\gamma_k)}}$$

Kiełbasiński 1996–1998

W-conjecture

W-conjecture

If computed *via* **GEPP** inverse G of X has **CONJ**-property, then G has, probably, stronger property **NC**.

Our numerical experiments with Higham's method for computing U from polar decomposition $A = UH$ seem to justify **W**-conjecture.

Purpose of numerical experiments

G computed inverse of X

Alt-only property:
$||XG - I||$ or $||GX - I||$ small, but not both

CONJ-only-property:
$||XG - I||$ and $||GX - I||$ small, but condition $G + \Delta G = (X + \Delta X)^{-1}$ is not satisfied.

Remark. Rounding errors in computation of G with **ALT-only** or **CONJ-only** are **dangerous** in Higham's method for polar decomposition.

Numerical experiments with Higham's method

Double sweep-process

- In the first sweep we compute X_k for $k = 0, 1, \ldots, l-1$.
- $\widetilde{U} = X_l$, computed in the first sweep, will be used in the second sweep for computing

$$\delta_k = \frac{||X_k - \widetilde{U}H_k||_F}{||X_k||_2}, \qquad H_k = \frac{1}{2}\left(\widetilde{U}^T X_k + X_k^T \widetilde{U}\right)$$

- In the second sweep we compute also

$$c_k = \mathrm{cond}_2(X_k) \qquad \text{or} \quad c_k - 1$$

$$e_k^{(L)} = \frac{||I - G_k X_k||_F}{||X_k||_2 ||G_k||_2}, \quad e_k^{(R)} = \frac{||I - X_k G_k||_F}{||X_k||_2 ||G_k||_2}.$$

Examples – Alt-Only property

GEPP, $n = 10$

$A = L^8 R$, L, R – random lower, upper triangular

k	$c_k - 1$	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	δ_k
0	$8.74e+14*$	$3.10e-17$	$8.72e-09$	$5.12e-09$
1	$1.66e+06$	$3.28e-17$	$1.96e-15$	$1.19e-15$
2	$7.56e+02$	$5.90e-17$	$7.52e-16$	$4.09e-16$
3	$1.19e+01$	$1.07e-16$	$1.44e-16$	$2.68e-16$
4	$1.17e+00$	$2.97e-16$	$2.95e-16$	$2.80e-16$
5	$8.38e-02$	$5.08e-16$	$5.16e-16$	$3.43e-16$
6	$1.51e-03$	$5.74e-16$	$5.74e-16$	$3.40e-16$
7	$7.01e-07$	$5.35e-16$	$5.35e-16$	$2.64e-16$
8	$2.46e-13$	$4.84e-16$	$4.84e-16$	$1.80e-16$

Examples - continuation: ALT-only

$n = 15, \quad A = \mathrm{rand}(Q)\mathrm{qr}(\mathrm{vand}(15))$

k	c_k	$e_k^{(L)}$	$e_k^{(R)}$	δ_k
0	$1.58e+13$	$3.68e-17*$	$3.91e-14$	$2.13e-14$
1	$1.11e+06$	$8.92e-17*$	$1.65e-14$	$8.23e-15$
2	$4.82e+02$	$1.38e-16$	$1.21e-15$	$7.12e-16$
3	$1.15e+01$	$2.22e-16$	$3.01e-16$	$5.47e-16$

$n = 25, \quad A = \mathrm{rand}(Q)\mathrm{qr}(\mathrm{vand}(25))$

k	c_k	$e_k^{(L)}$	$e_k^{(R)}$	δ_k
0	$1.87e+18!$	$2.93e-17*$	$1.39e-10$	$8.55e-11$
1	$4.25e+08$	$8.65e-17*$	$1.67e-12$	$7.67e-13$
2	$1.10e+04$	$1.15e-16$	$6.69e-15$	$3.75e-15$
3	$5.26e+01$	$3.47e-16$	$6.38e-16$	$1.09e-15$

Conj-only property, $A = P\operatorname{diag}(\sigma_j)Q^H$

$$e_k^{(\mathrm{L})}, \quad e_k^{(\mathrm{R})} \le 2.7 \times 10^{-15}$$

m_k – number of singular values of X_k close to $\sqrt{\sigma_1(X_k)\sigma_n(X_k)}$

$n = 6$

$\{\sigma_j\} = \{10^7, \sqrt{2 \times 10^7}, 1, 1, \sqrt{5 \times 10^{-8}}, 10^{-7}\}$

k	c_k	δ_k	m_k
0	$1.00e+14$	$5.49e-10$	2
1	$5.06e+06$	$1.01e-13$	2
2	$1.06e+03$	$8.74e-16$	–

$n = 20$

$\{\sigma_j\} = \{10^{14}, 10^7, \dots, 10^7, 1\}$

k	c_k	δ_k	m_k
0	$9.99e+13$	$7.04e-09$	18
1	$5.17e+06$	$1.72e-15$	–

Comments

- Higham's method with **GEPP** can fail, yielding for some special matrices A a poor unitary factor U. This will never occur for well-conditioned A.

- Using γ_k distinctly smaller than $\gamma_k^{(\mathrm{opt})}$ is spoiling quality of computed U:

$$\rho_k = \frac{\gamma_k}{\gamma_k^{(\mathrm{opt})}}.$$

Higham's method – rounding error analysis

- Kiełbasiński, Ziętak, **Numer. Algor. 2003**
- Byers, Xu, **SIMAX 2008**

Comparison

- Byers and Xu apply the same model of matrix inversion as AK and KZ.
- Byers and Xu – first order error analysis.
- AK and KZ – Wilkinson's analysis.

Byers and Xu apply

$$\widehat{X}_k = X_k + 0(\varepsilon), \qquad \widehat{X}_k - \text{computed}$$

under assumption

$$c(n)\mathrm{cond}_2(A)\varepsilon < 1, \quad \varepsilon - \text{machine epsilon}$$

Doubts: $0(\varepsilon^2)$ can be skipped???

$0(\varepsilon)$ depends on $\varepsilon[\mathrm{cond}(A)]^{3/2}$ for $k = 1$

$0(\varepsilon) >> 1$, Proof not completed???

Open question: who is in this picture?

Answer

Krystyna Ziętak

Thank you for your attention!!!