Wrocław University of Technology
Institute of Mathematics and Computer Science

Higham's scaled method for polar decomposition and numerical matrix-inversion

Andrzej Kiełbasiński, Paweł Zieliński, Krystyna Ziętak

Report I18/2007/P-045

Wrocław, July 2007

Higham's scaled method for polar decomposition and numerical matrix-inversion

Andrzej Kiełbasiński[1], Paweł Zieliński[2] and Krystyna Ziętak[2]

[1] *University of Warsaw, Institute of Applied Mathematics and Mechanics,*
02-097 Warsaw, Poland

[2] *Wrocław University of Technology,*
Institute of Mathematics and Computer Science, Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland

E-mail: pawel.zielinski@pwr.wroc.pl krystyna.zietak@pwr.wroc.pl

July 3, 2007

We present the theory (and illustrating experiments) of the numerical Higham's scaled method for computing the unitary factor of a nonsingular matrix. We show how the quality of the computed inverses of matrices influences the accuracy of the computed polar factorization. In particular: the inversion *via* GECP-factorization and reasonable scaling guarantee a good quality of the computed polar factors (with GEPP-factorization the computed unitary factor can be unacceptable). Some problems of scaling and switching criteria are discussed and experimentally investigated.

Keywords: polar decomposition of a matrix, Higham's method, rounding-errors analysis, numerical matrix-inversion

AMS subject classification: 65G50, 65F30

1 Introduction

We deal with the polar decomposition of a complex nonsingular matrix $A \in \mathbb{C}^{n\times n}$:

$$A = U_{\mathrm{A}} H_{\mathrm{A}}, \quad U_{\mathrm{A}} - \text{unitary}, \quad H_{\mathrm{A}} \in \mathcal{HPD}, \tag{1.1}$$

where $\mathcal{HPD}$ is the class of *Hermitian positive-definite matrices.* U_{A} is the unitary factor of A (the *orthogonal factor* of $A \in \mathbb{R}^{n\times n}$). Matrices $\{U_{\mathrm{A}}, H_{\mathrm{A}}\}$ are the *polar factors* of A.

The factorization (1.1) can be computed from SVD, the *singular value decomposition* of A. The iterative methods are alternative ways to compute (1.1), see for example [1, 3, 4, 5, 7].

In *Higham's scaled method* [4, 7], denoted by `HS` (referred also as *Newton's scaled method*), one constructs a sequence $\{X_k\}_{k=0}^{\infty}$ of matrices:

$$X_0 = A, \quad X_{k+1} = \frac{1}{2}\left(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H}\right), \quad \gamma_k > 0, \tag{1.2}$$

convergent to $U_{\rm A}$, the common unitary factor of all X_k. There are known several *theoretical* or *practical* rules of the choice of *scaling parameters* γ_k which increase the speed of convergence, see [3, 4, 7].

Let $\{\tilde{X}_k\}_{k=0}^{l}$ be the sequence of *iterates* computed in the *numerical* `HS`-*algorithm.* In all cases when this algorithm *converges* a *good unitarity* of the *computed unitary factor* $\tilde{U} \overset{\rm df}{=} \tilde{X}_l$ is achieved:

$$||\tilde{U}^H\tilde{U} - I||_2 \leqslant \varepsilon_0 \tag{1.3}$$

(all ε_s in this paper are of the size ν, the *computing precision*). We can now compute the *Hermitian factor* $\tilde{H}$ of A:

$$\tilde{B} := \tilde{U}^H * A; \qquad \tilde{H} := (\tilde{B} + \tilde{B}^H)/2. \tag{1.4}$$

The problem is: *whether* the *computed polar factors* $\{\tilde{U}, \tilde{H}\}$ of A are *acceptable*? That means: *whether* the following *relations* hold:

$$\tilde{H} \in \mathcal{HPD}, \qquad ||\tilde{U}\tilde{H} - A||_2 \leqslant \varepsilon_1 ||A||_2 \quad ? \tag{1.5}$$

In [8] we try to explain *how* it *happens* that the computed by numerical `HS` polar factors *are acceptable*? We reveal also *two main dangers*: the *poor quality* of the *computed inverses* and using of *too small scaling parameters.*

Our further research is presented in [9]. We explain there all *phenomena* we were able to perceive in *our experiments.* Therefore the *experimental results* play in [9] rather only the role of *illustrations.*

This paper is a *concise version* of [9]. We skip here the proofs, theorem 2.2, the estimation of the accuracy of experimental results and many detailed remarks. We concentrate on the *most important problem* of the *quality* of the *matrix-inversion* in the *numerical* `HS`-*algorithm.*

The theory is presented in sections 2 and 4. Section 3 explains *how* to *read experimental results.* Sections 5 and 6 present briefly some problems of *scaling* and *switching criteria.* For *final conclusions* see section 7.

We add the appendix presenting the proof of the `NC-Property` of the *inversion* by `B`-method, *via* `GECP`-factorization, see [2].

2 The theory of HS, the numerical Higham's method

Here and in all next sections HS means the *numerical* HS algorithm (to distinguish from (1.2), where the *theoretical* algorithm is defined).

Let $\tilde{X}_k$ be the *computed iterate* and X_k the *matrix satisfying* the *conditions* (2.3) below. Neither $\tilde{X}_k$ nor X_k here is identical with X_k in (1.2).

Let us define the following two functions

$$p : \mathbb{C}^{n\times n} \longrightarrow [n^{-1/2}, 1], \quad p(\mathbf{\Psi}) \overset{\mathrm{df}}{=} \begin{cases} 1, & \text{when } \mathbf{\Psi} = 0, \\ ||\mathbf{\Psi}||_2(||\mathbf{\Psi}||_F)^{-1}, & \text{otherwise,} \end{cases}$$

$$f : (0,\infty) \longrightarrow [1,\infty), \quad f(t) \overset{\mathrm{df}}{=} \frac{1}{2}\left(t + t^{-1}\right). \tag{2.1}$$

These *reserved* functions "produce" a series of derivate symbols $(f_k, p_k, p_+, \ldots)$ the values of f or p on concrete arguments.

We assume that the computations in HS are performed in the floating-point arithmetic with *precision* ν and that *neither* underflow *nor* overflow occurs.

The *epsilons* $(\varepsilon_0, \varepsilon_x, \ldots)$ are *modest multiples* of ν. Not all of them must be positive. We signal it writing, for example: $|\varepsilon'_k| \leqslant \varepsilon$. The only exceptions (see section 4) are "false epsilons" $(\breve{\varepsilon}_x, \breve{\varepsilon}_k, \ldots)$, the quantities which ought to be the *true epsilons* (and sometimes are) but – due to breaking of the basic assumption (2.3) – can be much larger than "a modest multiple of ν". Usually these false epsilons satisfy $|\breve{\varepsilon}| \ll 1$.

Let us formulate already now the following *general assumptions*:

$$\hat{\varepsilon}\ \mathrm{cond}_2(A) < 1, \quad \hat{\varepsilon} < \nu^{2/3} \lesssim 10^{-4}, \tag{2.2}$$

where $\hat{\varepsilon}$ is specified in (2.6), (2.3).

2.1 Main definitions and relations

Let us consider a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ and the sequence $\{\tilde{X}_k\}_{k=0}^{l}$ of matrices (1.2) computed in HS, $\tilde{X}_0 := A$.

Let γ_k be the *chosen scaling parameter* and G_k the *computed inverse* of $\tilde{X}_k$. We *assume* that *exists* a *nonsingular* matrix X_k *satisfying* the relations:

$$\tilde{X}_k = X_k - \mathbf{\Delta}_k, \qquad G_k = X_k^{-1} - \mathbf{\Delta}'_k, \tag{2.3}$$

where $||\mathbf{\Delta}_k||_F \leqslant \varepsilon_x ||X_k||_2$, $||\mathbf{\Delta}'_k||_F \leqslant \varepsilon_g ||X_k^{-1}||_2$. This defines (not uniquely) X_k for $k < l$. Let us extend it to $k = l$: $X_l = \tilde{X}_l$. The sequences $\{X_k\}$ and $\{\tilde{X}_k\}$ are *neighbour-sequences* and many important properties of X_k are close to these of $\tilde{X}_k$. We describe the HS-process in terms of the sequence $\{X_k\}$ since this sequence *imitates well* the relation (1.2), see below (2.5), (2.6).

The *assignment-statements*

$$G_k := \tilde{X}_k^{-1}, \quad \tilde{X}_{k+1} := \left(\tilde{X}_k * \gamma_k + G^H/\gamma_k\right)/2, \tag{2.4}$$

and (2.3) imply the *equalities*

$$X_{k+1} = Z_{k+1} + T_k, \quad Z_{k+1} \overset{\text{df}}{=} \frac{1}{2}\left(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H}\right) \tag{2.5}$$

and the *bound*

$$||T_k||_F \leqslant \hat{\varepsilon} f_k, \quad f_k \overset{\text{df}}{=} ||Z_{k+1}||_2, \quad \hat{\varepsilon} = 2\varepsilon_x + \varepsilon_g + 3\sqrt{n}\nu + O(\nu^2). \tag{2.6}$$

Let us consider the SVD of X_k:

$$X_k = P_k \text{diag}(\sigma_1^{(k)}, \dots, \sigma_n^{(k)}) Q_k^H, \quad P_k, Q_k \text{ unitary}$$

and define d_k, the *distance* of X_k *from* the *unitarity*:

$$d_k \overset{\text{df}}{=} \max_i |\sigma_i^{(k)} - 1| = \max\left\{\sigma_{\max}^{(k)} - 1, 1 - \sigma_{\min}^{(k)}\right\}, \tag{2.7}$$

$$\sigma_{\max}^{(k)} \overset{\text{df}}{=} \max_i\{\sigma_i^{(k)}\}, \quad \sigma_{\min}^{(k)} \overset{\text{df}}{=} \min_i\{\sigma_i^{(k)}\}. \tag{2.8}$$

The *efficiency* of HS depends on *how quickly* $\{d_k\}_{k=1}^l$ decrease, the *near-unitarity* of the *computed factor* $\tilde{U} = \tilde{X}_l$ depends on the *limiting accuracy* $d \overset{\text{df}}{=} \lim\ \sup d_k$ of the conceptional infinite sequence $\{d_k\}_{k=0}^\infty$. The *last iterate* $\tilde{X}_l$ constructed in HS should be the *first one* reaching the level $d_l \lesssim d$.

Let us define further quantities

$$c_k \overset{\text{df}}{=} \text{cond}_2(X_k) = \frac{\sigma_{\max}^{(k)}}{\sigma_{\min}^{(k)}}, \quad \gamma_k^{(\text{opt})} \overset{\text{df}}{=} \left(\sigma_{\max}^{(k)}\sigma_{\min}^{(k)}\right)^{-1/2},$$

$$\rho_k \overset{\text{df}}{=} \left(\gamma_k^{(\text{opt})}\gamma_k^{-1}\right)^{-2}, \quad \tau_k \overset{\text{df}}{=} \max\{\rho_k, \rho_k^{-1}\}. \tag{2.9}$$

The quantities ρ_k, τ_k "measure" the distance of γ_k from $\gamma_k^{(\text{opt})}$, the *optimal scaling parameter*.

In [9] we show the following relations, see (2.6),

$$f_k = f(\sqrt{c_k \tau_k}), \quad d_{k+1} = (1 - \varepsilon_k^*) f_k - 1, \quad |\varepsilon_k^*| \leqslant \hat{\varepsilon}, \tag{2.10}$$

$$\hat{\varepsilon} f_k < 1 \quad \textit{implies} \quad c_{k+1} \leqslant (1+\hat{\varepsilon}) f_k (1-\hat{\varepsilon} f_k)^{-1}. \tag{2.11}$$

The assumptions (2.2), *practical scaling* $(1,\infty)$-scaling [4] or (F)-scaling [7]) and appropriate *switching criteria* in HS guarantee that the sequence $\{f_k\}_{k=0}^{l-1}$ is *strictly decreasing* and the bounds $\hat{\varepsilon} f_k < 1, \tau_k < \sqrt{n}$ hold. We find ultimately in [8] that the bound (1.3) is satisfied with

$$\varepsilon_0 \approx 2d_l \leqslant \varepsilon' \overset{\text{df}}{=} \varepsilon_x + \varepsilon_g + 2\sqrt{n}\nu. \tag{2.12}$$

Remarks 2.1.

(i) In the case of the standard *double-precision* computations and HS with *practical scaling* in most cases $l \leqslant 10$ holds.

(ii) In some *special experiments* (see sections 4 and 5) we modify the *normal* HS-algorithm *introducing* (in a few initial steps only) *either* matrices G_k not satisfying (2.3) *or* scaling parameters γ_k much smaller than $\gamma_k^{(\text{opt})}$. But these modifications *neither* destroy the monotonic decrease of $\{f_k\}$ *nor* influence the final convergence of $\{\tilde{X}_k\}$. Hence the bounds (2.12) and (1.3) remain valid.

We need some *further notions* to discuss the *acceptability* (1.5) of the *computed polar factors* $\{\tilde{U}, \tilde{H}\}$.

Let the abbreviations AUF, APF mean: *approximate unitary factor, approximate polar factors*, respectively.

Let us consider any matrices $X, U \in \mathbb{C}^{n\times n}$, X-nonsingular , U-unitary. If $H_{ux} \overset{\text{df}}{=} \frac{1}{2}(U^H X + X^H U) \in \mathcal{HPD}$ then we will say that U is an AUF ($\{U, H_{ux}\}$ are APF) of X with *accuracy* (relative error):

$$\texttt{acc}(U,X) \overset{\text{df}}{=} \frac{||UH_{ux} - X||_F}{||X||_2}.$$

Let us fix now U as the *unitary factor* of $\tilde{U} = \tilde{X}_l = X_l$. Hence the *polar decomposition* of $\tilde{U}$ is, see (2.7), (2.12),

$$\tilde{U} = UH_u, \quad H_u \in \mathcal{HPD}, \quad ||\tilde{U} - U||_2 = d_l \lesssim \frac{1}{2}\varepsilon'. \tag{2.13}$$

Let now define for $k = 0, \ldots, l$ the following matrices and quantities:

$$H_k \overset{\text{df}}{=} \frac{1}{2}\left(U^H X_k + X_k^H U\right), \quad \delta_k \overset{\text{df}}{=} ||X_k - UH_k||_F ||X_k||_2^{-1}. \tag{2.14}$$

Evidently the following implication holds: $H_k \in \mathcal{HPD}$ *implies* $\delta_k = \texttt{acc}(U, X_k)$. In particular, see (2.13), $H_l = H_u \in \mathcal{HPD}$, $\delta_l = \texttt{acc}(U, X_l) = 0$.

The following lemma shows that the *properties* of the pair $\{H_0, \delta_0\}$ are *decisive* for the *acceptability* of the *computed polar factors* $\{\tilde{U}, \tilde{H}\}$.

Lemma 2.1. Let introduce the quantities $p_0 \overset{\mathrm{df}}{=} p(X_0 - UH_0)$, $\varepsilon_{\mathrm{I}} \overset{\mathrm{df}}{=} 2.5\varepsilon_x + \varepsilon_g + \nu m(\sqrt{n})$, where $m(t)$ is a modest polynomial in t (depending on the way of computing $\tilde{B}$ in (1.4)). If $(p_0\delta_0 + \varepsilon_{\mathrm{I}})\mathrm{cond}_2(A) < 1$ holds and $H_0 \in \mathcal{HPD}$ then the following relations hold:

$$\tilde{H} \in \mathcal{HPD}, \qquad \left| \frac{||A - \tilde{U}\tilde{H}||_2}{||A||_2} - p_0\delta_0 \right| \lesssim \varepsilon_{\mathrm{I}}.$$

Remark 2.2. Lemma 2.1 is valid only when G_0 satisfies (2.3).

Conclusion 2.1. The *computed polar factors* $\{\tilde{U}, \tilde{H}\}$ are *acceptable* **iff** $H_0 \in \mathcal{HPD}, \delta_0$ is of the order ν and A is sufficiently well-conditioned, since the following bounds hold: $|p_0\delta_0 - \varepsilon_{\mathrm{I}}| \leqslant ||A - \tilde{U}\tilde{H}||_2||A||_2^{-1} \leqslant p_0\delta_0 + \varepsilon_{\mathrm{I}}$.

In the next subsection we present an *explicit expression* of δ_k in terms of: $\delta_{k+1}, \rho_k, c_k, \hat{\varepsilon}$, see (2.17)-(2.20). This opens a chance for "theoretical transfer" from $\delta_l = 0$ to the *important quantity* δ_0.

We must be prepared that $\mathtt{acc}(U, \tilde{X}_k) \gtrsim \mathtt{acc}(U, \tilde{X}_{k+1})$ holds since the *rounding errors* in the computation of G_k and $\tilde{X}_{k+1}$, see (2.4), *can partly spoil* the *information* on $\tilde{X}_k$ *transferred* to $\tilde{X}_{k+1}$ (hence also to $\tilde{U} = \tilde{X}_l$). The same concerns the *neighbour-sequence* $\{X_k\}_{k=0}^{l}$: the *relation* $\delta_k \gtrsim \delta_{k+1}$ *can be expected*!

We should recognize *benign rounding errors* in (2.4) – such that δ_k is at most *only slightly larger* than δ_{k+1} – and *dangerous rounding errors* – such that $\delta_k \gg \delta_{k+1}$ *can occur.*

2.2 BIT, the backward-induction theorem

Let us introduce the matrix, see (2.5),$\mathbf{\Psi}_k \overset{\mathrm{df}}{=} UH_{k+1} - Z_{k+1}$ and the quantities, see (2.6), (2.1), (2.8),

$$\xi_k \overset{\mathrm{df}}{=} ||\mathbf{\Psi}_k||_2, \quad \vartheta_k \overset{\mathrm{df}}{=} ||\mathbf{\Psi}_k||_F f_k^{-1}, \quad r_k \overset{\mathrm{df}}{=} \frac{f_k}{f(\sigma_{\max}^{(k)}\gamma_k)}. \tag{2.15}$$

Theorem 2.1 (BIT). If the relations

$$\xi_k < 1, \qquad H_{k+1} \in \mathcal{HPD} \tag{2.16}$$

are satisfied then $\delta_k = \vartheta_k|\chi_k + \kappa_k\zeta_k|r_k,\ \zeta_k \overset{\text{df}}{=} (3\sqrt{2}+2)(2-\xi_k)^{-1}\xi_k$,

$$c_k\vartheta_k|\mu_k + \lambda_k\zeta_k|r_k < 1 \quad \textit{implies} \quad H_k \in \mathcal{HPD},$$

where $\chi_k, \mu_k, \kappa_k, \lambda_k$ are real numbers, *either* all equal zero *or* satisfying inequalities:

$$0 \leqslant \mu_k < \chi_k \leqslant 1, \quad |\kappa_k| < 1, \quad |\lambda_k| < 1. \tag{2.17}$$

Remark 2.3. Theorem 2.1 is valid also in cases when the matrices G_k, G_{k+1} are not satisfying (2.3).

Corollary 2.1. The quantity r_k, see (2.15), satisfies the relations

$$r_k = \max\left\{1, (c_k+\rho_k)(c_k\rho_k+1)^{-1}\right\} < \max\{1, \rho_k^{-1}\}\,. \tag{2.18}$$

If the matrices G_k, G_{k+1} satisfy (2.3) then

$$\xi_k = p'_k\left|\delta_{k+1}(1+\varepsilon'_k) + \varepsilon'_k\right| f_k, \quad p'_k \overset{\text{df}}{=} p(\mathbf{\Psi}_k), \quad |\varepsilon'_k| \leqslant \hat{\varepsilon}, \tag{2.19}$$

and – provided (2.16) holds –

$$\delta_k = \left|\delta_{k+1}(1+\varepsilon'_k) + \varepsilon'_k\right|\ |\chi_k + \kappa_k\zeta_k|\, r_k, \quad |\varepsilon'_k| \leqslant \hat{\varepsilon}. \tag{2.20}$$

This allows us to *simplify* the *backward-induction rule*: if $\xi_k \ll 1$ and $H_{k+1} \in \mathcal{HPD}$ holds then

$$\delta_k \approx \left|\delta_{k+1} + \varepsilon'_k\right| \chi_k r_k, \quad |\varepsilon'_k| \leqslant \hat{\varepsilon}, \quad \chi_k \in [0,1], \tag{2.21}$$

$$c_k(\delta_{k+1} + \hat{\varepsilon})(1 + 7\xi_k)r_k < 1 \quad \textit{implies} \quad H_k \in \mathcal{HPD}.$$

Remarks 2.4.

(i) In *double-precision computations* the approximate equality (2.21) *describes adequately* the behaviour of the sequence $\{\delta_k\}$, since in this case all $\{\xi_k\}$ are very small (the only exception can be ξ_0 when G_0 is not satisfying (2.3), see section 4).

(ii) With *optimal* or *practical scaling* the *relations* $\chi_k r_k \lesssim 1$ can be expected, see section 5. But in the *general case* the *rounding errors* in the computations of $\tilde{X}_{k+1}$ in (2.4) *can be dangerous* when $\rho_k \ll 1$ and $c_k \gg 1$ holds: this implies $r_k \gg 1$ (Theorem 2.2 in [9] shows that χ_k tends to decrease with ρ_k, but we can not expect that always $\chi_k r_k \lesssim 1$ holds, see section 5).

(iii) *Optimal* or *practical scaling* and *inverses* G_k, G_{k+1} *satisfying* (2.3) *guarantee* $\delta_k \lesssim \delta_{k+1} + \hat{\varepsilon}$. Hence in this case the *rounding errors* in *both operations* of (2.4) are *benign*.

(iv) If *any* of the matrices G_k, G_{k+1} is not satisfying (2.3) then the bound $\hat{\varepsilon}$ on $|\varepsilon'_k|$ in corollary 2.1 must be replaced with a much larger quantity: the *rounding errors* in the *computation* of *such inverse are dangerous*. We deal with such cases in section 4.

3 Introduction to examples of numerical tests

In sections 4 and 5 we present *examples* of *numerical tests* illustrating relevant fragments of the theory. All our tests were performed for matrices $A \in \mathbb{R}^{n\times n}$, $6 \leqslant n \leqslant 35$, in the IEEE standard *double-precision*, $\nu = \nu_d \approx 2.2 \times 10^{-16}$ (with cummulation of "inner products" on standard *extended-precision* variables, $\nu = \nu_e \approx 10^{-19}$).

In most cases we present the computed results with at least *two correct leading decimals*. The results marked with a *star* $(*)$ have probably *only one* correct leading decimal. In results with *exclamation mark* (!) even the first decimal is doubtful.

For each example we present the *matrix* A, the information on *matrix-inversion* and *scaling* in HS. We present also the quantity $\tilde{\Delta}_l \overset{\text{df}}{=} ||\tilde{U}^T\tilde{U} - I||_F, \tilde{U} = \tilde{X}_l$, and the result of the Cholesky-positivity test of $\tilde{H}$, see (1.4). Then we present for several iterations, $k = 0, 1, \ldots$ some of the computed quantities: $c_k, \rho_k, r_k, e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}, \hat{\delta}_k$ (eventually also some other auxiliary quantities), where

$$e_k^{(\mathrm{L})} \overset{\text{df}}{=} ||I - G_k\tilde{X}_k||_F w_k^{-1},\ e_k^{(\mathrm{R})} \overset{\text{df}}{=} ||I - \tilde{X}_kG_k||_F w_k^{-1},\ w_k \overset{\text{df}}{=} ||\tilde{X}_k||_2||G_k||_2, \tag{3.1}$$
$$\hat{\delta}_k \overset{\text{df}}{=} ||\tilde{X}_k - U\hat{H}_k||_F||\tilde{X}_k||_2^{-1}, \quad \hat{H}_k \overset{\text{df}}{=} \frac{1}{2}\left(U^T\tilde{X}_k + \tilde{X}_k^TU\right).$$

Remarks 3.1.

(i) Let $\tilde{p}_0 \overset{\text{df}}{=} p(\tilde{X}_0 - U\hat{H}_0)$. Then $\tilde{p}_0\hat{\delta}_0$ is *a close approximation* of $||A - \tilde{U}\tilde{H}||_2||A||_2^{-1}$.

(ii) $\hat{\delta}_k$ is a *close approximation* of δ_k, see (2.14), provided G_k is satisfying (2.3).

Example 3.1. In table 3.1 we present the computed results of the `HSTEST`-program (see section 3 in [9]) for the 10×10 matrix $A_1 = \text{tril}(\text{rand}(10))^8\text{rand}(U)$, see [2], $\Delta_9 = 5.14 \times 10^{-18}$, applying (F)-*scaling* and the `GEPP`-*matrix-inversion.* Matrix $\tilde{H}$ passed the positivity test.

Table 3.1

k	$c_k - 1$	ρ_k	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	$\hat{\delta}_k$
0	$8.74e + 14*$	$0.930*$	$3.10e - 17$	$8.72e - 09$	$5.12e - 09$
1	$1.66e + 06$	0.708	$3.28e - 17$	$1.96e - 15$	$1.19e - 15$
2	$7.56e + 02$	1.00	$5.90e - 17$	$7.52e - 16$	$4.09e - 16$
3	$1.19e + 01$	0.732	$1.07e - 16$	$1.44e - 16$	$2.68e - 16$
4	$1.17e + 00$	1.07	$2.97e - 16$	$2.95e - 16$	$2.80e - 16$
5	$8.38e - 02$	1.03	$5.08e - 16$	$5.16e - 16$	$3.43e - 16$
6	$1.51\mathrm{e} - 03$	1.00	$5.74e - 16$	$5.74e - 16$	$3.40e - 16$
7	$7.01\mathrm{e} - 07$	1.00	$5.35\mathrm{e} - 16$	$5.35e - 16$	$2.64e - 16$
8	$2.46e - 13$	1.00	$4.84e - 16$	$4.84e - 16$	$1.80e - 16$

Remarks 3.2.

(i) The value of $e_0^{(\mathrm{R})}$ shows that matrix G_0 is not satisfying (2.3).

(ii) The quantity $||A - \tilde{U}\tilde{H}||_2||A||_2^{-1}$, see (1.4) and remark 3.1 (i), cannot be smaller than $\hat{\delta}_0 n^{-1/2} \approx 1.62 \times 10^{-9}$. Hence the computed polar factors $\{\tilde{U}, \tilde{H}\}$ are *not acceptable.* It is the result of breaking the assumption (2.3) for $k = 0$, see section 4.

(iii) The results presented in table 3.1 for $k > 3$ are typical for all our experiments. In next examples we will present only the relevant part of experimental results.

4 The quality problem of the matrix-inversion in the `HS`-process

Some contemporary standard procedures *compute* the *inverses* from the *Gaussian triangular factorization* with *partial pivoting* (`GEPP`) of the *inverted matrix*, see [2]. Using *these procedures* in the `HS`-process *yields frequently* (but *not always*!) *acceptable results* (see example 3.1). The inversion *via triangular factorization* with *complete pivoting* (`GECP`) yields practically *always acceptable results* in `HS` with *practical* or *optimal scaling.*

We should *recognize* the *properties* of the *computed inverse* G_k of $\tilde{X}_k$ *not impending* the *good numerical behaviour* of the HS-process and *those properties* which *can seriously spoil* the *quality* of the *computed unitary factor* $\tilde{U}$ of A.

4.1 Properties of computed inverses

Let G be the computed inverse of the nonsingular matrix X. We introduce auxiliary quantities $x \overset{\mathrm{df}}{=} ||X||_2,\ g \overset{\mathrm{df}}{=} ||G||_2,\ c \overset{\mathrm{df}}{=} \mathrm{cond}_2(X) = x||X^{-1}||_2$ and consider the following four eventual properties of G:

$$||G - X^{-1}||_F \leqslant \varepsilon g c, \tag{4.1}$$

$$||GX - I||_F \leqslant \varepsilon g x, \tag{4.2}$$

$$||XG - I||_F \leqslant \varepsilon g x, \tag{4.3}$$

$$\exists \Delta', \Delta : \quad G + \Delta' = (X + \Delta)^{-1}, \quad ||\Delta'||_F \leqslant \varepsilon_g g, \quad ||\Delta||_F \leqslant \varepsilon_x x. \tag{4.4}$$

The same relations define the *properties of inversion procedures* as follows: Let $\mathbb{M}$ be a subset of nonsingular $n \times n$ matrices X. We say that an inversion algorithm `Inv` is *numerically stable* (`NS`) in $\mathbb{M}$ if for each $X \in \mathbb{M}$ the computed inverse G satisfies (4.1). In the same way:

– (4.2) defines the *left-residual stability* (`LRS`) of `Inv` in $\mathbb{M}$,

– (4.3) defines the *right-residual stability* (`RRS`) of `Inv` in $\mathbb{M}$,

– (4.4) defines the *numerical correctness* (`NC`) of `Inv` in $\mathbb{M}$.

We shall use the same notation: `NS`, `LRS`, `RRS`, `NC` for the properties (4.1)–(4.4) of the matrix G (no matter what is the “official property” in $\mathbb{M}$ of the algorithm which computed G).

We define also two *combined* properties of G:

$$\texttt{Alt} \overset{\mathrm{df}}{=} \texttt{LRS}\ \textit{or}\ \texttt{RRS}, \quad \texttt{Conj} \overset{\mathrm{df}}{=} \texttt{LRS}\ \textit{and}\ \texttt{RRS}. \tag{4.5}$$

Assuming $\varepsilon_x + \varepsilon_g + \varepsilon_x \varepsilon_g \leqslant \varepsilon$ and $\varepsilon x g < 1$ we find the following implications:

$$\texttt{NC} \Longrightarrow \texttt{Conj} \Longrightarrow \texttt{Alt} \Longrightarrow \texttt{NS} \tag{4.6}$$

and the bounds

$$||GX - I||_F \leqslant c||XG - I||_F, \quad ||XG - I||_F \leqslant c||GX - I||_F. \tag{4.7}$$

Let us note further that for small c, say $c \leqslant 10$, NS implies NC (for example: with $\varepsilon_x = 0, \varepsilon_g \leq 10\varepsilon$). Hence all listed properties of G can differ distinctly only when $c = \mathrm{cond}_2(X)$ is large.

Further definitions: We will say that G has LRS-Only-Property if G has the LRS-Property but has not the RRS-Property. In this case G has the Alt-Property (hence also NS-Property) but has *neither* the Conj-Property *nor* the NC-Property. In the same way, using the term: to have the Only-Property, we define other eventual *highest-properties* of G in the hierarchical system defined by (4.5), (4.6).

Let us note at last that the NC-Property is the *highest general quality* (expressed in *norms of matrices*) of an *inverse G computed* in a *constant finite precision.* According to the formulation of W. Kahan, see [6], in this NC-case: G is *a slightly wrong inverse* of a *slightly wrong matrix* X.

4.2 The W-conjecture

There are several versions of computing the inverse G from GEPP-triangular factorization of X, see [2], which are either *left-residual* - or *right-residual-stable* in a broad subset $\mathbb{M}$ of $n \times n$ matrices. Hence such GEPP-*inversion algorithms* guarantee the Alt-Property of computed inverses. For well-conditioned matrices X it means practically the Conj-Property of G. But also for badly conditioned matrices X we can check directly that *frequently both residuals* $||GX - I||_F, ||XG - I||_F$ *are small* (are bounded by εxg), see [2], [10]. That means that G has the Conj-Property in spite of (4.7) with large c.

J.H. Wilkinson *explained* this *phenomenon*, in [10, pp. 110-111], showing that the matrix G (computed *via* GEPP-factorization by A-method, see [2]) has the NC-Property provided the *triangular systems* – involved in the computation of G from GEPP – are *solved* to *high accuracy.* This happens *frequently* but *not always.* It seems probable that *this is the only reason why happens* the Conj-Property of inverses computed *via* GEPP-factorization. Let us express it as follows:

W-conjecture. If an inverse G computed *via* GEPP-factorization of X has the Conj-Property then, probably, G has also the (stronger) NC-Property.

The experiments of subsection 4.5 and all our experiments with GEPP-inversion seem to justify the W-conjecture.

4.3 `HS` with inverses not always satisfying (2.3)

In (2.3) we postulate in fact the `NC-property` (4.4) of all computed inverses G_k of $\tilde{X}_k, k = 0, \dots, l-1$. Hence , see remark 2.4 (iii), the `NC-property` of all $\{G_k\}$ is *sufficient* for *good behaviour* of the `HS`-process with *practical scaling.*

The problem is *whether* the *inverses* G_k *not satisfying* (2.3) *can spoil* (and *how much*?) the *quality* of the *computed unitary factor* $\tilde{U}$?

We will consider only the case of G_k with `Alt-Property` (this includes the `Conj-Property` and `NC-Property` as special subcases). Let us incorporate *these eventual deviations* (from the *normality* of (2.3)) into our general description of `HS`.

Let assume hence the relations

$$X_k = \tilde{X}_k + \mathbf{\Delta}_k, \quad X_k^{-1} = G_k + \mathbf{\Delta}'_k \tag{4.8}$$

and let us introduce the quantities (in general: *false epsilons*): $\check{\varepsilon}_x^{(k)} \overset{\text{df}}{=} ||\mathbf{\Delta}_k||_F ||X_k||_2^{-1}$, $\check{\varepsilon}_g^{(k)} \overset{\text{df}}{=} ||\mathbf{\Delta}'_k||_F ||X_k^{-1}||_2^{-1}$.

Let us assume further the relations: $\hat{c}_k \gg 1$, $\varepsilon \hat{c}_k \ll 1$, $\hat{c}_k \overset{\text{df}}{=} \text{cond}_2(\tilde{X}_k)$. We present below a *simplified version* of theorem 4.1 in [9], using an *approximate equality* $a \approx b$ (a, b nonnegative) meaning any of the following *three possibilities*:

$$|a-b| \leqslant O(\varepsilon), \ |a-b| \leqslant O(\varepsilon \hat{c}_k) \max\{a, b\}, \ |a-b| \leqslant O(\hat{c}_k^{-1}) \max\{a, b\}.$$

Theorem 4.1. The only minimizer $\mathbf{\Delta}_k$ of the linear functional

$$\varphi_k(\mathbf{\Delta}) \overset{\text{df}}{=} \max_{\mathbf{\Delta} \in \mathbb{C}^{n \times n}} \left\{ \frac{||\mathbf{\Delta}||_F}{||\tilde{X}_k||_2}, \frac{||\tilde{X}_k^{-1} - G_k - \tilde{X}_k^{-1} \mathbf{\Delta} \tilde{X}_k^{-1}||_F}{||G_k||_2} \right\}$$

defines in (4.8) the nonsingular matrix X_k and the matrix $\mathbf{\Delta}'_k$ such that the following relations hold:

$$c_k \overset{\text{df}}{=} \text{cond}_2(X_k) \approx \hat{c}_k, \quad \check{\varepsilon}_x^{(k)} \approx \check{\varepsilon}_g^{(k)} \approx \hat{\varphi}_k \overset{\text{df}}{=} \varphi_k(\mathbf{\Delta_k}).$$

Introducing the quantities, see (3.1):

$$\varepsilon_k^{(\text{A})} \overset{\text{df}}{=} \min\left\{ e_k^{(\text{L})}, e_k^{(\text{R})} \right\}, \quad \check{\varepsilon}_k^{(\text{A})} \overset{\text{df}}{=} \max\left\{ e_k^{(\text{L})}, e_k^{(\text{R})} \right\}, \quad e_k^{(c)} = \sqrt{e_k^{(\text{L})} e_k^{(\text{R})}},$$

we specify $\hat{\varphi}_k$ according to the assumed property of G_k:

(i) If G_k has the `Alt-Only-Property` then $\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{Alt})}$ where

$$\frac{1}{2}\check{\varepsilon}_k^{(\mathrm{A})} \lesssim \hat{\varphi}_k^{(\mathrm{Alt})} \lesssim \frac{1}{\sqrt{2}} e_k^{(c)} c_k^{1/2} \lesssim \frac{1}{\sqrt{2}} \varepsilon_k^{(\mathrm{A})} c_k. \tag{4.9}$$

(ii) If G_k has the `Conj-Only-Property` then $\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{Conj})}$ where

$$\frac{1}{2}\varepsilon_k^{(\mathrm{A})} \lesssim \hat{\varphi}_k^{(\mathrm{Conj})} \lesssim \frac{1}{\sqrt{2}} \varepsilon_k^{(c)} c_k^{1/2}, \quad \varepsilon_k^{(c)} \overset{\mathrm{df}}{=} e_k^{(c)}. \tag{4.10}$$

(iii) If G_k has the `NC-Property` then $\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{NC})}$ where $\hat{\varphi}_k^{(\mathrm{NC})} \lesssim \max\{\varepsilon_x, \varepsilon_g\}$.

If matrices G_k, G_{k+1} can have the `Alt` – or `Conj` – or `NC-Property` then the bound $\hat{\varepsilon}$ in the relevant relations of section 2 must be replaced with $\check{\varepsilon}_k^*$: $\check{\varepsilon}_k^* \overset{\mathrm{df}}{=} 2\hat{\varphi}_k + \hat{\varphi}_{k+1} + 3\sqrt{n}\nu$. But for important recursive formulas (2.20), (2.21) we should rather choose the *presentation exposing* the *potentially dominating terms*. For example, when $\xi_k \ll 1, H_{k+1} \in \mathcal{HPD}$ and G_k has the `Alt-Only-Property` or the `Conj-Only-Property`, let us choose the presentation:

$$\delta_k \approx \Big|\; |\varphi_k^* + \theta_k' \hat{\varphi}_{k+1}| + \theta_k'' |\delta_{k+1} + O(\nu)| \;\Big| \chi_k r_k, \quad \theta_k', \theta_k'' \in [-1, 1],$$

where $\varphi_k^* \overset{\mathrm{df}}{=} ||\Delta_k \gamma_k + \Delta_k'^H \gamma_k^{-1}||_F (2f_k)^{-1}$. Closer examination of the matrices Δ_k, Δ_k' shows that the following bounds (respectively) hold:

$$\hat{\varphi}_k^{(\mathrm{Alt})} \lesssim \varphi_k^* \lesssim \sqrt{2}\hat{\varphi}_k^{(\mathrm{Alt})} \quad \mathit{or} \quad \varphi_k^* \lesssim 2\hat{\varphi}_k^{(\mathrm{Conj})}. \tag{4.11}$$

Hence in the case of *distinctly* `Alt-Only-Property` of G_k (when $\check{\varepsilon}_k^{(\mathrm{A})} \gg \max\{\hat{\varphi}_{k+1}, \delta_{k+1}\}$) the *relation*

$$\delta_k \gtrsim \frac{1}{2}\check{\varepsilon}^{(\mathrm{A})} = \frac{1}{2}\max\{e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\} \tag{4.12}$$

is *inevitable*, see (4.9).

Conclusion 4.1. The *rounding errors* in the *computation* of G_k with `Alt-Only-` or `Conj-Only-Property` are *dangerous.*

4.4 Experiments with inverses G_k having the `Alt-Only-Property`

We apply here in `HS` the *practical-scaling* and the computation of the inverses G_k *via* `GEPP`-factorization of $\tilde{X}_k$ (versions `LRS` - or `RRS`-stable).

In example 3.1 we presented already such experiment with the 10×10 matrix $A_1 = \text{tril}(\text{rand}(10))^8\text{rand}(U))$, see [2].

Example 4.1. The matrices $\tilde{H}$ passed the positivity test only in examples (i), (ii) below. Matrices A_3, A_4 are defined in [2].

(i) The results for $A_2 = A_1^T, \tilde{\Delta}_9 = 6.2 \times 10^{-16}$ are presented in table 4.1.

Table 4.1

k	c_k	$e_k^{(\text{L})}$	$e_k^{(\text{R})}$	$\hat{\delta}_k$
0	$8.75e+14*$	$8.79e-09$	$3.25e-17*$	$5.45e-09$
1	$1.86e+06$	$5.57e-15$	$6.12e-17*$	$2.69e-15$
2	$2.96e+02$	$6.39e-16$	$3.46e-16$	$3.46e-16$

(ii) Table 4.2 includes the results for $n = 15, A_3 = \text{rand}(Q)\text{qr}(\text{vand}(15))$, $\tilde{\Delta}_{10} = 9.17 \times 10^{-16}$.

Table 4.2

k	c_k	$e_k^{(\text{L})}$	$e_k^{(\text{R})}$	$\hat{\delta}_k$
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	$1.15e+01$	$2.22e-16$	$3.01e-16$	$5.47e-16$

(iii) In table 4.3 we give the results for $n = 25, A_4 = \text{rand}(Q)\text{qr}(\text{vand}(25))$, $\tilde{\Delta}_{10} = 2.46 \times 10^{-15}$.

Table 4.3

k	c_k	$e_k^{(\text{L})}$	$e_k^{(\text{R})}$	$\hat{\delta}_k$
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	$5.26e+01$	$3.47e-16$	$6.38e-16$	$1.09e-15$

Remarks 4.1.

(i) For example (i) see remarks 3.1 (i), (ii).

(ii) In examples (ii), (iii) G_0 and G_1 have the `LRS-Only-Property`.

(iii) Notice that the relation (4.12) is clearly demonstrated for $k = 0$ in all tests of example 3.1 and example 4.1.

4.5 Experiments with inverses G_k having the `Conj-Only-Property`

We apply here in `HS` the *optimal-scaling* and the procedure `INVCONJ`(X) yielding (*via* `SVD` of X) the computed inverse G of X with `Conj-Property` (if *possible*: with `Conj-Only-Property`), see subsection 4.5 in [9]. We present below the experiments with matrices $A_s = P_s \mathrm{diag}(\sigma_j^{(s)}) Q_s^T \in \mathbb{R}^{n\times n}$ for $s = 5, 6, 7$ (P_s, Q_s orthogonal, random). In all these experiments the relative residuals $e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}$ are not exceeding 2.7×10^{-15}. Hence we present only the quantities $c_k, c_k^{1/2}, \hat{\delta}_k, m_k$, where m_k is the number of singular values $\{\hat{\sigma}_i^{(k)}\}$ of X_k *close* to $\hat{\alpha}_k \overset{\mathrm{df}}{=} \left(\hat{\sigma}_{\max}^{(k)} \hat{\sigma}_{\min}^{(k)}\right)^{1/2}$ (with $A_s \in \mathbb{R}^{n\times n}$ the *rounding errors* in G_k with `Conj-Only-Property` are *dangerous* only when $m_k \geqslant 2$ holds, see subsection 4.5 in [9]).

Examples 4.2. In experiments below all matrices $\tilde{H}$ passed the positivity test.

(i) In table 4.4 we present the results for $n = 6, \tilde{\Delta}_6 = 5.76 \times 10^{-16}$ and

$$\{\sigma_i^{(5)}\} = \{10^7, \sqrt{2\times 10^7}, 1, 1, \sqrt{5\times 10^{-8}}, 10^{-7}\}.$$

Table 4.4

k	c_k	$\sqrt{c_k}$	$\hat{\delta}_k$	m_k
0	$1.00e+14$	$1.00e+07$	$5.49e-10$	2
1	$5.06e+06$	$2.25e+03$	$1.01e-13$	2
2	$1.06e+03$	$3.26e+01$	$8.74e-16$	–

(ii) In table 4.5 we present the results for $n = 20, \tilde{\Delta}_6 = 1.99 \times 10^{-15}$ and

$$\{\sigma_i^{(6)}\} = \{10^{14}, 10^7, \ldots, 10^7, 1\}.$$

Table 4.5

k	c_k	$\sqrt{c_k}$	$\hat{\delta}_k$	m_k
0	$9.99e+13$	$1.00e+07$	$7.04e-09$	18
1	$5.17e+06$	$2.27e+03$	$1.72e-15$	$-$

(iii) In table 4.6 we present the results for $n = 20, \tilde{\Delta}_8 = 1.87 \times 10^{-15}$ and

$$\sigma_i^{(7)} = (10^{14/19})^{i-1} \qquad (i = 1, \dots, 20).$$

Table 4.6

k	c_k	$\sqrt{c_k}$	$\hat{\delta}_k$	m_k
0	$1.00e+14$	$1.00e+07$	$4.39e-10$	2
1	$3.61e+06$	$1.90e+03$	$1.31e-13$	2
2	$7.27e+02$	$8.50e+01$	$6.62e-15$	1
3	$1.35e+01$	$3.07e+00$	$2.10e-15$	$-$

Remark 4.2. The experimental results presented above are evidently consistent with the bounds (4.11), (4.10).

5 The problems of scaling

Assuming: $\xi_k \ll 1, H_{k+1} \in \mathcal{HPD}$ and G_k, G_{k+1} satisfying (2.3), we can use the simplified form of recursion, see (2.21),

$$\delta_k \approx |\delta_{k+1} + \varepsilon'_k| z_k, \quad z_k \overset{\mathrm{df}}{=} \chi_k r_k, \quad |\varepsilon'_k| \leqslant \hat{\varepsilon},$$

where: $\chi_k \leqslant 1, r_k = \max\{1, (c_k + \rho_k)(c_k\rho_k + 1)^{-1}\}, \rho_k = (\gamma_k/\gamma_k^{(\mathrm{opt})})^2$. If $\gamma_k \ll \gamma_k^{(\mathrm{opt})}$ and $c_k \gg 1$ then $r_k \gg 1$ holds. Though χ_k tends to decrease with ρ_k, see theorem 2.2 in [9], it can happen that also $z_k \gg 1$ holds, what implies $\delta_k \gg \delta_{k+1}$. That is *the problem of too small scaling parameters.*

This can happen in *one step*, but also in *several consecutive steps*, when $z_k > 1, z_{k-1} > 1, ...$ holds.

In HS with *practical scaling* $r_k < \sqrt{n}$ holds, hence the *danger* is *not very serious.* What's more: all *known experiments* seem to indicate that HS *with practical-scaling* is *immune* to the *danger* of too small scaling parameters: the *relation* $z_k \overset{\mathrm{df}}{=} \chi_k r_k \lesssim 1$ is *always observed.* Section 5 in [9] proposes an *explanation* for this phenomenon.

But for *drastically small scaling parameters* the *danger* of $z_k \gg 1$ *really exists*!

Example 5.1. For a random 10×10 matrix A_8 we apply the HS-process with GECP matrix-inversion and – essentially – (F)-scaling, introducing "artificially" very small γ_k for $k = 0, 2, 4$. The results are presented in table 5.1. We additionally compute the quantities $\hat{\chi}_k \overset{\mathrm{df}}{=} \hat{\delta}_k r_k^{-1}(\hat{\delta}_{k+1} + 10^{-16})^{-1}$ (probably lower bounds on χ_k). Matrix $\widetilde{H}$ passed the positivity test.

Table 5.1

k	c_k	ρ_k	r_k	$\hat{\delta}_k$	$\hat{\chi}_k$
0	$9.61e+14*$	$8.21e-05*$	$1.21e+04*$	$3.96e-13$	0.0013
1	$1.12e+09$	$1.12e+00$	1	$2.46e-14$	0.422
2	$1.17e+04$	$1.27e-04$	$5.13e+03$	$5.85e-14$	0.013
3	$5.17e+03$	$1.08e+00$	1	$6.71e-16$	0.647
4	$3.15e+01$	$3.25e-02$	$1.55e+01$	$8.36e-16$	0.154
5	$1.64e+01$	$1.37e+00$	1	$1.51e-16$	0.302

Remarks 5.1.

(i) Table 5.1 demonstrates the tendency of χ_k to decrease with ρ_k.

(ii) Very small ρ_k (hence large τ_k, (2.9)) retard the decreasing of $\{c_k\}$, see relations (2.10), (2.11).

(iii) Section 5 in [9] presents more examples of this type.

Another problem is the influence of scaling on the effectiveness of the HS-process. Both considered above ways of *practical scaling* have two advantages:

– for large n, say $n \geqslant 10$, the cost of computing of $\{\gamma_k\}$ is negligible (with respect to the cost of the matrix-inversion),

– there is a chance of accelerating the convergence when there are large gaps in the spectrum of singular values of A.

The following way of quasi-optimal scaling:

- choose positive quantities a_0, b_0 such that $a_0 < \sigma_j(A) < b_0$ holds,
- compute: $\mu_0 := b_0/a_0$, $\gamma_0^{(\mathrm{q})} := (a_0\sqrt{\mu_0})^{-1}$, and for $k > 0$

$$\mu_k := (\mu_{k-1}^{1/2} + \mu_{k-1}^{-1/2})/2, \qquad \gamma_k^{(\mathrm{q})} := \mu_k^{-1/2},$$

guarantees the first advantage for all n; however, it does not have the second advantage.

6 The switching criteria in HS

In our experiments, aimed to study the problems of sections 4 and 5, we tested additionally the *criteria* (proposed in [4], [7], [8]) for *accepting* the *last computed iterate* as the *computed unitary factor* $\tilde{U}$. We tested also the *criteria* (proposed in [4], [8]) for *switching* from $(1,\infty)$-*scaling* to *unscaled iterations.* Section 6 in [9] presents the details of these tests. One of the *conclusions* is presented in section 7 (iii).

7 Final conclusions

(i) Matrix-inversion in the HS-process should yield the computed inverse G of the matrix X (the inverse of the current iterate) satisfying the condition (2.3) (the NC-property). This property is warranted by the inversion *via* GECP-triangularization of X. Using in HS the standard inversion *via* GEPP, see [2], can fail, yielding for some special matrices A a poor unitary factor $\tilde{U}$. This will never occur for well-conditioned matrices A, say: $\text{cond}_2(A) \leqslant 10^2$.

(ii) Using in the HS-process a good matrix-inversion, see (i), and *either* (F)-scaling [7] *or* $(1,\infty)$-scaling [4] (with appropriate switch to unscaled iterations) practically guarantees good quality of the computed unitary factor $\tilde{U}$ of A (the same quality, as yields the unitary factor computed *via* SVD of A).

(iii) An appropriate stopping criterion in most cases guarantees that $\tilde{U} = \tilde{X}_l$ is the first iterate reaching the limiting accuracy. With the stopping criterion in [4] frequently one redundant step is performed.

(iv) The formal cost (the number of arithmetic operations) of the HS-process in the standard-double precision is at most of the same order as for SVD (is smaller for well-conditioned matrices or matrices with large gap in the spectrum of the singular values).

(v) Using in the HS-process scaling parameters $\{\gamma_k\}$ distinctly larger or smaller than the optimal ones, see relations (2.9) and (2.10), can spoil the convergence. Using $\{\gamma_k\}$ distinctly smaller is spoiling also the quality of $\tilde{U}$ as an approximate unitary factor of A. Practical scaling, see (iii), is not involving such impendency.

References

[1] Å. Björck, C. Bowie, An iterative algorithm for computing the best estimate of an orthogonal matrix, SIAM J. Numer. Anal. 8 (1971) 358–364.

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

[3] W. Gander, Algorithms for the polar decomposition, SIAM J. Sci. Stat. Comput. 11 (1990) 1102–1115.

[4] N.J. Higham, Computing the polar decomposition – with applications, SIAM J. Sci. Stat. Comput. 7 (1986) 1160–1174.

[5] N.J. Higham, *Functions of a Matrix: Theory and Computation*, Book in preparation.

[6] W. Kahan, A survey of error analysis, *Proc. IFIP Congr.* 71, vol. I, 220–226.

[7] 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.

[8] A. Kiełbasiński, K. Ziętak, Numerical behaviour of Higham's scaled method for polar decomposition, Numerical Algorithms 32 (2003), 105–140.

[9] A. Kiełbasiński, P. Zieliński, K. Ziętak, Numerical experiments with Higham's scaled method for polar decomposition, *Report* I18/2006/P-013, Wrocław Univ. of Technology, Inst. Math. Comput. Science, Wrocław, May 2006 (`http://www.im.pwr.wroc.pl/~zietak/reports/`).

[10] J.H. Wilkinson, *Rounding Errors in Algebraic Processes* (Her Majesty's Stationery Office, London, 1963).

[11] P. Zieliński, K. Ziętak, The polar decomposition – properties, applications and algorithms, Applied Mathematics, Annals of Polish Math. Soc. 38 (1995) 23–49.

A1 Numerical correctness of inverting matrices *via* GECP

.

Theorem A1.1. Let assume that the `GECP`-process for $n\times n$ matrix X yields the permutation matrices $P_{\rm L}, P_{\rm R}$, and lower and upper triangular matrices $L = \text{tril}(L) = [l_{ij}], R = \text{triu}(R) = [r_{ij}]$, respectively, such that the following relations hold:

$$P_{\rm L}(X+\boldsymbol{\Delta})P_{\rm R}^T = L\cdot R,\ ||\boldsymbol{\Delta}|| \leqslant \varepsilon_x||X||, \quad r_{ii} \neq 0, l_{ii} = 1 \text{ for every } i. \tag{A1.1}$$

Let G be the *inverse* of X, *computed* by the `B`-method *via* `GECP` of X (that means: from the factors $P_{\rm L}, P_{\rm R}, L, R$). Then the matrix G satisfies the relations:

$$G + \boldsymbol{\Delta}' = (X+\boldsymbol{\Delta})^{-1}, \quad ||\boldsymbol{\Delta}'|| \leqslant \varepsilon_g||G||, \tag{A1.2}$$

where $\varepsilon_x, \varepsilon_g$ are modest multiples of ν (the computing precision).

Proof. We will use here the ∞-norm of matrices: $||\cdot|| = ||\cdot||_\infty$. Not lessening the generality of considerations let assume $P_{\rm L} = I = P_{\rm R}$. Let introduce the matrices

$$D \overset{\rm df}{=} \text{diag}(r_{ii}), \quad U \overset{\rm df}{=} D^{-1}R = [u_{ij}]. \tag{A1.3}$$

The `GECP` process guarantees the relations for every i, j:

$$l_{ii} = u_{ii} = 1, \quad |l_{ij}| \leqslant 1, \quad |u_{ij}| \leqslant 1,$$

what implies the bounds:

$$||L|| \leqslant n, \quad ||U|| \leqslant n, \quad ||L^{-1}|| \leqslant 2^{n-1}, \quad ||U^{-1}|| \leqslant 2^{n-1}. \tag{A1.4}$$

Let present the `B`-method as following two assignment-statements:

$$V := R^{-1}, \quad G := V * L^{-1}. \tag{A1.5}$$

Let $\mathbf{v}_i^T, \mathbf{g}_i^T$ be the i-th rows of V and G, respectively. Row-wise implementation of (A1.5) amounts to solving the following triangular equations:

$$\mathbf{v}_i^T R \overset{!}{=} \mathbf{e}_i^T, \quad \mathbf{g}_i^T L \overset{!}{=} \mathbf{v}_i^T \quad (i = 1, \ldots, n),$$

where $\mathbf{e}_i^T$ is the i-th row of the identity matrix. The computed solutions $\mathbf{v}_i, \mathbf{g}_i$ of these equations satisfy the equalities

$$\mathbf{v}_i^T(R+\delta R_i) = \mathbf{e}_i^T, \quad \mathbf{g}_i^T(L+\delta L_i) = \mathbf{v}_i^T, \tag{A1.6}$$

where the *perturbation matrices* $\delta R_i, \delta L_i$ (equivalent to rounding-errors in the solving algorithms) are bounded:

$$|\delta R_i| \leq \nu c|R|, \quad |\delta L_i| \leq \nu c|L|. \tag{A1.7}$$

($c \approx 1$, if "inner products" are cumulated on higher-precision variable, otherwise $c = n$.)

Let rewrite the equalities (A1.6) in the form

$$\mathbf{v}_i^T R = \mathbf{e}_i^T (I + \mathbf{\Phi}_i)^{-1}, \quad \mathbf{g}_i^T (I + \mathbf{\Psi}_i) = \mathbf{v}_i^T L^{-1}, \tag{A1.8}$$

where, with $\delta U_i \overset{\text{df}}{=} D^{-1} \delta R_i$, see (A1.3),

$$\mathbf{\Phi}_i \overset{\text{df}}{=} R^{-1} \delta R_i = U^{-1} \delta U_i, \quad \mathbf{\Psi}_i \overset{\text{df}}{=} \delta L_i L^{-1}. \tag{A1.9}$$

(We assume $||\mathbf{\Phi}_i|| < \frac{1}{2}$, since $||\mathbf{\Phi}_i|| \leq \nu c n 2^{n-1}$, see (A1.4), is for large n practically always a severe overbound.)

All row-equalities (A1.8) can be presented in the matrix form:

$$VR = I - \hat{\mathbf{\Phi}}, \quad G + \mathbf{\Delta}_1' = V L^{-1}, \tag{A1.10}$$

where [using the equality $(I + \mathbf{\Phi})^{-1} = I - \mathbf{\Phi}(I + \mathbf{\Phi})^{-1}$] the i-th row of $\hat{\mathbf{\Phi}}$ is equal to $\mathbf{e}_i^T \mathbf{\Phi}_i (I + \mathbf{\Phi}_i)^{-1}$, and the i-th row of $\mathbf{\Delta}_1'$ is equal to $\mathbf{g}_i^T \mathbf{\Psi}_i$. From (A1.4), (A1.7), (A1.9) follow the bounds

$$||\hat{\mathbf{\Phi}}|| \leqslant \varepsilon_1 (1 - \varepsilon_1)^{-1}, \quad ||\mathbf{\Delta}_1|| \leqslant \varepsilon_1 ||G||, \quad \varepsilon_1 \overset{\text{df}}{=} \nu c n 2^{n-1}. \tag{A1.11}$$

From (A1.1), (A1.10), (A1.11) we obtain ultimately

$$G + \mathbf{\Delta}' = (X + \mathbf{\Delta})^{-1}, \quad ||\mathbf{\Delta}'|| < \nu c n 2^n ||G||, \tag{A1.12}$$

where $\mathbf{\Delta}' \overset{\text{df}}{=} \mathbf{\Delta}_1' + \hat{\mathbf{\Phi}}(I - \hat{\mathbf{\Phi}})^{-1}(G + \mathbf{\Delta}_1')$, what completes the proof. □

Remarks A.1.

(i) Relations (A1.1) are satisfied for any sufficiently well-conditioned matrix X. But simple modification of `GECP` guarantees (A1.1) (with ε_x being a modest multiple of ν) for any matrix $X \neq 0$. This allows us to apply the `HS`-process also for such matrices.

(ii) Let's note that $||\mathbf{\Delta}'|| \approx \max_i ||\mathbf{g}_i^T \delta L_i L^{-1} + \mathbf{e}_i^T U^{-1} \delta U_i G||$ and that the bounds (A1.4) on $||L^{-1}||$ and $||U^{-1}||$ are for larger n practically never approached. Hence in most cases (A1.12) is a severe *overbound* on $||\mathbf{\Delta}'||$. We can expect that ε_g in (A1.2) is practically always a modest multiple of ν.

(iii) In [10, pp. 110–111] Wilkinson proves the `NC`-property (A1.2) of the matrix G, computed *via* `GEPP` by the `A`-method, see [2], under assumptions that all involved triangular systems are *solved* to *high accuracy.* Since in the case of `GECP` this condition is always satisfied hence theorem A.1 is valid also for the `A`-method.