

Positive Polar Factors of Graded Matrices¹Ren-Cang Li²

October 2003

ABSTRACT

Let B be an $m \times n$ ($m \geq n$) complex (or real) matrix. It is known that there is a unique *polar decomposition* $B = QH$, where $Q^*Q = I$, the $n \times n$ identity matrix, and H is positive definite, provided B has full column rank. If B is perturbed to \tilde{B} , how do the polar factors Q and H change? This question has been investigated quite extensively, but most work so far was on how the perturbation changed the unitary polar factor Q , and very little on the positive polar factor H , except $\|H - \tilde{H}\|_F \leq \sqrt{2}\|B - \tilde{B}\|_F$ in the Frobenius norm, due to F. Kittaneh (*Comm. Math. Phys.*, 104 (1986), pp. 307–310), where \tilde{Q} and \tilde{H} are the corresponding polar factors of \tilde{B} . While this inequality of Kittaneh shows that H is always well-behaved under perturbations, it does not tell much about smaller entries of H in the case when H 's entries vary a great deal in magnitudes. This paper is intended to fill the gap by addressing the variations of H for a graded matrix $B = GS$, where S is a scaling matrix and usually diagonal (but may not be.). The elements of S can vary wildly, while G is well-conditioned. In cases as such, the magnitudes of H 's entries indeed often vary a lot and thus any bound on $\|H - \tilde{H}\|_F$ means little, if any thing, to the accuracy of \tilde{H} 's smaller entries. This paper proposes a new way to measure the errors in the H factor via bounding the scaled difference $(\tilde{H} - H)S^{-1}$, as well as how to accurately compute the factor when S is diagonal. Numerical examples are presented.

The results are also extended to the matrix square root of a graded positive definite matrix.

¹This report is available on the web at <http://www.ms.uky.edu/~rcli/>.

²Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rcli@ms.uky.edu.) This work was supported in part by the National Science Foundation CAREER award under Grant No. CCR-9875201.

Positive Polar Factors of Graded Matrices ^{*}

Ren-Cang Li [†]

October 2003

Abstract

Let B be an $m \times n$ ($m \geq n$) complex (or real) matrix. It is known that there is a unique *polar decomposition* $B = QH$, where $Q^*Q = I$, the $n \times n$ identity matrix, and H is positive definite, provided B has full column rank. If B is perturbed to \tilde{B} , how do the polar factors Q and H change? This question has been investigated quite extensively, but most work so far was on how the perturbation changed the unitary polar factor Q , and very little on the positive polar factor H , except $\|H - \tilde{H}\|_F \leq \sqrt{2}\|B - \tilde{B}\|_F$ in the Frobenius norm, due to F. Kittaneh (*Comm. Math. Phys.*, 104 (1986), pp. 307–310), where \tilde{Q} and \tilde{H} are the corresponding polar factors of \tilde{B} . While this inequality of Kittaneh shows that H is always well-behaved under perturbations, it does not tell much about smaller entries of H in the case when H 's entries vary a great deal in magnitudes. This paper is intended to fill the gap by addressing the variations of H for a graded matrix $B = GS$, where S is a scaling matrix and usually diagonal (but may not be.). The elements of S can vary wildly, while G is well-conditioned. In cases as such, the magnitudes of H 's entries indeed often vary a lot and thus any bound on $\|H - \tilde{H}\|_F$ means little, if any thing, to the accuracy of \tilde{H} 's smaller entries. This paper proposes a new way to measure the errors in the H factor via bounding the scaled difference $(\tilde{H} - H)S^{-1}$, as well as how to accurately compute the factor when S is diagonal. Numerical examples are presented.

The results are also extended to the matrix square root of a graded positive definite matrix.

1 Introduction

Let B be an $m \times n$ ($m \geq n$) complex matrix. It is known that there are Q with orthonormal column vectors, i.e., $Q^*Q = I$, and a unique positive semi-definite H such that

$$B = QH. \tag{1.1}$$

^{*}This material is based on work supported in part by the National Science Foundation CAREER award under Grant No. CCR-9875201.

[†]Department of Mathematics, University of Kentucky, Lexington, KY 40506. Email: rccli@ms.uky.edu.

Hereafter I denotes an identity matrix with appropriate dimensions which will be clear from the context or specified. The decomposition (1.1) is called the *polar decomposition* of B . If, in addition, B has full column rank, then Q is uniquely determined, too. In fact,

$$H = (B^*B)^{1/2}, \quad Q = B(B^*B)^{-1/2}, \quad (1.2)$$

where superscript “*” denotes conjugate transpose. The decomposition (1.1) can also be computed from the *singular value decomposition* (SVD) $B = U\Sigma V^*$ by

$$H = V\Sigma_1V^*, \quad Q = U_1V^*, \quad (1.3)$$

where $U = (U_1, U_2)$ and V are unitary, U_1 is $m \times n$, $\Sigma = \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix}$ and $\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_n)$ is nonnegative.

Assume now that B is perturbed to \tilde{B} , and

$$B = QH \quad \text{and} \quad \tilde{B} = \tilde{Q}\tilde{H} \quad (1.4)$$

are their polar decompositions, respectively. There are many published bounds on how much the two factor matrices Q and H may change with respect to the perturbation to B being *additive* or *multiplicative*. The additive perturbation refers to the situation when no assumption was made on how B was perturbed except possibly an assumption on the smallness of $\|\tilde{B} - B\|$ for some matrix norm $\|\cdot\|$. The multiplicative perturbation refers to the situation when $\tilde{B} = D_1^*BD_2$ such that D_1 and D_2 are assumed close to identity matrices. The perturbation for *one-sided scaling* case (or the *graded* case) can be translated into this kind [13]. By one-sided scaling we mean that B and \tilde{B} take the forms

$$B = GS, \quad \tilde{B} = \tilde{G}S \equiv (G + \Delta G)S, \quad (1.5)$$

where S is a scaling matrix and usually diagonal (but this is not necessary to some theorems below.) The elements of S can vary wildly. G has full column rank and usually better-conditioned than B itself, i.e., the ratio of G 's largest singular value over its smallest one is much smaller than that for B .

Much work was done in the past for the additive perturbation, e.g., [2, 4, 5, 8, 9, 11, 12, 14, 15, 16, 17, 19], except [13] which was for the perturbation of the other kind. Among them most attention was gone into how the unitary factor Q changed with respect to the two different kinds of perturbations and to the number fields – real *vs.* complex. Perhaps this is caused by that the study on the perturbation of H could be considered *complete*, owing to the following result [10, Theorem 2]:

$$\|H - \tilde{H}\|_F \leq \sqrt{2}\|B - \tilde{B}\|_F, \quad (1.6)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This paper motivated by the following example concerns the H factor for the graded case.

Example 1.1 Let

$$B = GS = \begin{pmatrix} 6 & -2 & 14 & -5 \\ 8 & 5 & -7 & -8 \\ -2 & -11 & 2 & -3 \\ 5 & -8 & -16 & 9 \end{pmatrix} \begin{pmatrix} 10^6 & & & \\ & 10^4 & & \\ & & 10^2 & \\ & & & 1 \end{pmatrix}. \quad (1.7)$$

We take ΔG a random complex matrix (in MATLAB): $10^{-5} * \mathbf{randn}(4)$. Then (1.6) implies $\|H - \tilde{H}\|_F \leq \sqrt{2}\|\Delta GS\|_F \approx 24.7393$. However, the true H is (only first 5 decimal digits in each entry are presented here)

$$H = \begin{pmatrix} 1.1358 \cdot 10^7 & 8.6928 \cdot 10^3 & -4.9320 \cdot 10^2 & -3.7828 \cdot 10^0 \\ 8.6928 \cdot 10^3 & 1.4603 \cdot 10^5 & 3.1908 \cdot 10^2 & -4.4827 \cdot 10^0 \\ -4.9320 \cdot 10^2 & 3.1908 \cdot 10^2 & 2.1691 \cdot 10^3 & -7.7287 \cdot 10^0 \\ -3.7828 \cdot 10^0 & -4.4827 \cdot 10^0 & -7.7287 \cdot 10^0 & 9.2121 \cdot 10^0 \end{pmatrix} \quad (1.8)$$

which, together with $\|H - \tilde{H}\|_F \leq 24.7393$, gives no assurance to the correctness of smaller entries in \tilde{H} ! Actually, all the entries of H are determined to high relative accuracy by the data as we shall see soon.

2 Main Result

With (1.4), write $\tilde{Q} = Q + \Delta Q$. We have

$$\begin{aligned} \tilde{Q}\tilde{H} &= \tilde{G}S, \\ \tilde{H}S^{-1} &= \tilde{Q}^*\tilde{G} \\ &= \tilde{Q}^*(G + \Delta G) \\ &= (Q + \Delta Q)^*G + \tilde{Q}^*\Delta G \\ &= Q^*G + \Delta Q^*G + \tilde{Q}^*\Delta G \\ &= HS^{-1} + \Delta Q^*G + \tilde{Q}^*\Delta G. \end{aligned}$$

Thus $(\tilde{H} - H)S^{-1} = \Delta Q^*G + \tilde{Q}^*\Delta G$, and

$$\|(\tilde{H} - H)S^{-1}\|_F \leq \|\Delta Q^*\|_F \|G\|_2 + \|\Delta G\|_F, \quad (2.1)$$

where $\|\cdot\|_2$ denotes the spectral norm. It is proved in [13] that

$$\|\Delta Q\|_F \leq \sqrt{\|(\Delta G)G^\dagger\|_F^2 + \left\|I - (I + (\Delta G)G^\dagger)^{-1}\right\|_F^2}, \quad (2.2)$$

$$\leq \sqrt{1 + \frac{1}{(1 - \|G^\dagger\|_2 \|\Delta G\|_2)^2}} \|G^\dagger\|_2 \|\Delta G\|_F, \quad (2.3)$$

where $G^\dagger = (G^*G)^{-1}G^*$ is the pseudo-inverse of G . A consequence of (2.1), (2.2), and (2.3) is the following theorem, which says up to the first order $\|(\tilde{H} - H)S^{-1}\|_F$ is bounded by $[\sqrt{2}\kappa(G) + 1]\|\Delta G\|_F$, where $\kappa(G) \stackrel{\text{def}}{=} \|G\|_2 \|G^\dagger\|_2$.

Theorem 2.1 *Let $B = GS$ and $\tilde{B} = \tilde{G}S$ be two $m \times n$ matrices having full column rank with the polar decompositions (1.4). S is $n \times n$ and nonsingular. If $\|\Delta G\|_2 \|G^\dagger\|_2 < 1$ then*

$$\|(\tilde{H} - H)S^{-1}\|_{\text{F}} \leq \sqrt{\|(\Delta G)G^\dagger\|_{\text{F}}^2 + \left\|I - (I + (\Delta G)G^\dagger)^{-1}\right\|_{\text{F}}^2} \|G\|_2 + \|\Delta G\|_{\text{F}}, \quad (2.4)$$

$$\leq \left(\sqrt{1 + \frac{1}{(1 - \|G^\dagger\|_2 \|\Delta G\|_2)^2}} \|G^\dagger\|_2 \|G\|_2 + 1 \right) \|\Delta G\|_{\text{F}}. \quad (2.5)$$

The scaled difference between H and \tilde{H} in Theorem 2.1 is measured by the Frobenius norm of $\|(\tilde{H} - H)S^{-1}\|_{\text{F}}$. But since the H factors are Hermitian, some readers may prefer some kind of symmetric scaling, i.e., two sided scaling. This can be easily done when S is diagonal. Let $|S|$ be the diagonal matrix obtained by taking entry-wise absolute value on S . Then $|S|^{-1/2}$ is well-defined, and $|S|^{-1/2}(\tilde{H} - H)|S|^{-1/2}$ is Hermitian and has the same eigenvalues as $(\tilde{H} - H)|S|^{-1}$. Therefore for diagonal S

$$\||S|^{-1/2}(\tilde{H} - H)|S|^{-1/2}\|_{\text{F}} \leq \|(\tilde{H} - H)|S|^{-1}\|_{\text{F}} = \|(\tilde{H} - H)S^{-1}\|_{\text{F}}. \quad (2.6)$$

More generally it can be proven that (2.6) holds for normal S with $|S|$ being interpreted as $(S^*S)^{1/2}$ which in the case of diagonal S is the same as taking entry-wise absolute values. We now outline a proof. Assume that S is normal and let $S = X\Lambda X^*$ be its eigen-decomposition, where X is unitary and Λ is diagonal. Then $|S| = X|\Lambda|X^*$ and $|S|^{-1/2} = X|\Lambda|^{-1/2}X^*$. We have

$$\begin{aligned} \||S|^{-1/2}(\tilde{H} - H)|S|^{-1/2}\|_{\text{F}} &= \||\Lambda|^{-1/2}X^*(\tilde{H} - H)X|\Lambda|^{-1/2}\|_{\text{F}} \\ &\leq \|X^*(\tilde{H} - H)X|\Lambda|^{-1}\|_{\text{F}} \\ &= \|X^*(\tilde{H} - H)X\Lambda^{-1}\|_{\text{F}} \\ &= \|(\tilde{H} - H)X\Lambda^{-1}X^*\|_{\text{F}} \\ &= \|(\tilde{H} - H)S^{-1}\|_{\text{F}}, \end{aligned}$$

as expected. So we have¹

Corollary 2.1 *To the conditions of Theorem 2.1 add this: S is normal (and thus diagonal S included). Then*

$$\begin{aligned} &\||S|^{-1/2}(\tilde{H} - H)|S|^{-1/2}\|_{\text{F}} \\ &\leq \sqrt{\|(\Delta G)G^\dagger\|_{\text{F}}^2 + \left\|I - (I + (\Delta G)G^\dagger)^{-1}\right\|_{\text{F}}^2} \|G\|_2 + \|\Delta G\|_{\text{F}}, \quad (2.7) \end{aligned}$$

$$\leq \left(\sqrt{1 + \frac{1}{(1 - \|G^\dagger\|_2 \|\Delta G\|_2)^2}} \|G^\dagger\|_2 \|G\|_2 + 1 \right) \|\Delta G\|_{\text{F}}. \quad (2.8)$$

¹One may also see that $\|S^{-1/2}(\tilde{H} - H)S^{-1/2}\|_{\text{F}} = \||S|^{-1/2}(\tilde{H} - H)|S|^{-1/2}\|_{\text{F}}$ for any square root S of a normal S . But then $S^{-1/2}(\tilde{H} - H)S^{-1/2}$ is no longer Hermitian.

We caution the reader that the bounds on $\| |S|^{-1/2}(\tilde{H} - H)|S|^{-1/2} \|_{\text{F}}$ in Corollary 2.1 can overestimate the errors in some entries of \tilde{H} much worse than the bounds on $\|(\tilde{H} - H)S^{-1}\|_{\text{F}}$ in Theorem 2.1 do, even though bounding $\| |S|^{-1/2}(\tilde{H} - H)|S|^{-1/2} \|_{\text{F}}$ seems to be more mathematically elegant. In a moment we shall revisit Example 1.1 to show our point. For now, let us analyze the case when S is diagonal. Suppose

$$\| |S|^{-1/2}(\tilde{H} - H)|S|^{-1/2} \|_{\text{F}} \leq \|(\tilde{H} - H)S^{-1}\|_{\text{F}} \leq \beta.$$

Write $S = \text{diag}(s_1, s_2, \dots)$, $H = (h_{ij})$, and $\tilde{H} = (\tilde{h}_{ij})$. Entry-wise, $\| |S|^{-1/2}(\tilde{H} - H)|S|^{-1/2} \|_{\text{F}} \leq \beta$ gives

$$|h_{ij} - \tilde{h}_{ij}| \leq \beta \sqrt{|s_i s_j|}, \quad (2.9)$$

and $\|(\tilde{H} - H)S^{-1}\|_{\text{F}} \leq \beta$ gives

$$|h_{ij} - \tilde{h}_{ij}| = |h_{ji} - \tilde{h}_{ji}| \leq \beta \min\{|s_i|, |s_j|\}. \quad (2.10)$$

If $|s_i| \gg |s_j|$ or $|s_i| \ll |s_j|$, (2.9) will be much less sharp than (2.10). Accordingly, the numbers of correct significant decimal digits defined as

$$-\log_{10}(|h_{ij} - \tilde{h}_{ij}|/|h_{ij}|)$$

in \tilde{H} are at least

$$-\log_{10}(\beta \sqrt{|s_i s_j|}/|h_{ij}|) \text{ by (2.9), or } -\log_{10}(\beta, \min\{|s_i|, |s_j|\})/|h_{ij}|) \text{ by (2.10).}$$

Example 1.1 (continued). Use the same ΔG . Note that S is diagonal and positive definite. Corollary 2.1 implies

$$\begin{aligned} \|S^{-1/2}(\tilde{H} - H)S^{-1/2}\|_{\text{F}} &\leq \|(\tilde{H} - H)S^{-1}\|_{\text{F}} \\ &\leq 1.4298 \cdot 10^{-4}, && \text{(by (2.7))} \\ &\leq 2.0756 \cdot 10^{-4}. && \text{(by (2.8))} \end{aligned}$$

Take $\beta = 1.4298 \cdot 10^{-4}$ in (2.9) and (2.10). Two sets of lower bounds on the numbers of correct significant decimal digits in \tilde{h}_{ij} obtained from using, respectively, $\|S^{-1/2}(\tilde{H} - H)S^{-1/2}\|_{\text{F}} \leq \beta$ and $\|(\tilde{H} - H)S^{-1}\|_{\text{F}} \leq \beta$ are, entry-wise,

$$\begin{pmatrix} 4.9 & 2.8 & 2.5 & 1.4 \\ 2.8 & 5.0 & 3.3 & 2.5 \\ 2.5 & 3.3 & 5.2 & 3.7 \\ 1.4 & 2.5 & 3.7 & 4.8 \end{pmatrix}, \quad \begin{pmatrix} 4.9 & 3.8 & 4.5 & 4.4 \\ 3.8 & 5.0 & 4.3 & 4.5 \\ 4.5 & 4.3 & 5.2 & 4.7 \\ 4.4 & 4.5 & 4.7 & 4.8 \end{pmatrix}.$$

These bounds yield two conclusions. First, all entries of \tilde{H} have at least 1 correct decimal digit and the diagonal entries have at least 4 correct decimal digits, much sharper than the bound (1.6) indicates. Second, $\|(\tilde{H} - H)S^{-1}\|_{\text{F}} \leq \beta$ can be potentially sharper than $\|S^{-1/2}(\tilde{H} - H)S^{-1/2}\|_{\text{F}} \leq \beta$ when it comes to off-diagonal entries.

3 Stable Computation of H

Given the perturbation theory we have developed in Section 2, how do we compute the factor H as accurately as predicted? It is not clear if this can be done for an arbitrary S . But we shall show it is always possible for diagonal S :

1. Compute SVD of $B \equiv GS = U\Sigma V^*$ by one-sided Jacobi [7];
 2. Set $Q = UV^*$, $W = Q^*G$, and then $H = WS$.
- (3.1)

It is also tempting to compute H as $V\Sigma V^*$. But as we shall show by Example 3.1 below, doing so can destroy the high relative accuracy in V and Σ delivered by the one-sided Jacobi method. The following theorem shows that (3.1) will deliver H with a small scaled error.

Theorem 3.1 *The computed \tilde{H} by (3.1) satisfies $\|(\tilde{H} - H)S^{-1}\|_F = \mathcal{O}(\epsilon_m \kappa(G)\|G\|_F)$.*

Proof: Denote all corresponding computed matrices by the same symbols except with tildes. It is proved in Algorithm 3.1, Theorems 3.1 and 3.2 and their proofs in [7] that the computed SVD $\tilde{U}\tilde{\Sigma}\tilde{V}^*$ is the exact SVD of a nearby \tilde{B} :

$$\tilde{B} = (I + E)B(I + F), \quad \text{satisfying} \quad \|E\|_F = \mathcal{O}(\epsilon_m \kappa(G)), \quad \|F\|_F = \mathcal{O}(\epsilon_m \kappa(G)),$$

where ϵ_m is the machine roundoff for the working precision. Note also hiding in the two $\mathcal{O}(\cdot)$ are some modest increasing functions of n . It can be seen that the computed \tilde{Q} , \tilde{W} , and \tilde{H} satisfy

$$\tilde{Q} = \tilde{U}\tilde{V}^* + E_1, \quad \tilde{W} = \tilde{Q}^*G + E_2, \quad \tilde{H} = (\tilde{W}S) \circ M,$$

where \circ denote the entry-wise Hadamard product,

$$\|E_1\|_F = \mathcal{O}(\epsilon_m), \quad \|E_2\|_F = \mathcal{O}(\epsilon_m\|G\|_F),$$

and each M 's (i, j) th entry $m_{ij} = 1 + \mathcal{O}(\epsilon_m)$ since S is diagonal. With those equations, we get

$$\begin{aligned} \tilde{H}S^{-1} &= \tilde{W} \circ M \\ &= (\tilde{Q}^*G + E_2) \circ M \\ &= [(\tilde{U}\tilde{V}^* + E_1)^*G + E_2] \circ M \\ &= [(\tilde{U}\tilde{V}^*)^*G] \circ M + (E_1^*G + E_2) \circ M. \end{aligned}$$

Now by [13, Theorem 1], we have $\tilde{U}\tilde{V}^* = Q + E_3$ satisfying

$$\|E_3\|_F = \mathcal{O}(\|E\|_F + \|F\|_F) = \mathcal{O}(\epsilon_m \kappa(G)).$$

Write $M = J + E_4$, where J 's entries are all 1's and $\|E_4\|_F = \mathcal{O}(\epsilon_m)$. Therefore noticing $Q^*G = HS^{-1}$, we have

$$\begin{aligned}\tilde{H}S^{-1} &= (Q^*G) \circ M + (E_3^*G + E_1^*G + E_2) \circ M \\ &= (HS^{-1}) \circ (J + E_4) + (E_3^*G + E_1^*G + E_2) \circ M \\ &= HS^{-1} + (Q^*G) \circ E_4 + (E_3^*G + E_1^*G + E_2) \circ M,\end{aligned}$$

which gives

$$\|(\tilde{H} - H)S^{-1}\|_F \leq \|(Q^*G) \circ E_4 + (E_3^*G + E_1^*G + E_2) \circ M\|_F = \mathcal{O}(\epsilon_m \kappa(G) \|G\|_F),$$

as expected. ■

Example 3.1 $n = 10$, and

$$G = \begin{pmatrix} 4.656 & 7.220 & 3.831 & 2.924 & 7.556 & 7.329 & 4.105 & 2.827 & 6.787 & 7.7860 \\ 4.187 & 2.644 & 5.986 & 9.774 & 7.660 & 7.170 & 1.042 & 2.240 & 5.235 & 6.1470 \\ 7.518 & 6.780 & 5.691 & 1.071 & 8.527 & 2.231 & 0.220 & 7.386 & 9.997 & 4.8770 \\ 3.829 & 7.801 & 2.164 & 1.158 & 1.907 & 4.992 & 7.945 & 9.693 & 7.529 & 2.1330 \\ 3.515 & 1.907 & 9.999 & 0.686 & 2.950 & 2.240 & 5.921 & 9.097 & 5.180 & 3.4390 \\ 0.871 & 4.511 & 4.406 & 3.102 & 4.731 & 1.793 & 3.100 & 9.608 & 6.400 & 6.9290 \\ 8.565 & 8.088 & 9.318 & 3.557 & 6.034 & 9.012 & 1.087 & 8.167 & 1.351 & 4.6020 \\ 1.351 & 5.289 & 3.025 & 3.591 & 1.210 & 2.123 & 8.771 & 4.650 & 8.002 & 9.6420 \\ 2.408 & 2.745 & 3.893 & 4.202 & 5.845 & 3.501 & 0.603 & 2.775 & 0.912 & 9.1680 \\ 0.231 & 2.726 & 6.898 & 9.243 & 3.813 & 5.065 & 0.594 & 8.415 & 5.140 & 1.9450 \end{pmatrix},$$

$$S = \text{diag}(10^3, 10^8, 10^5, 10^4, 1, 10^4, 10^9, 10^8, 10^3, 10^8).$$

Throughout computations, IEEE single precision was used and thus $\epsilon_m = 2^{-24} \approx 5.96 \times 10^{-8}$. We compared accuracy of numerical results via the one-sided Jacobi SVD code provided to us by Dr. Z. Drmač and the SVD code `sgesvd` by the QR algorithm from LAPACK [1]. The exact Q and H are obtained by Maple² with `Digits:=50`. The following table lists various errors by the two methods, with Q and H computed as in (3.1).

	$\ (\tilde{H} - H)S^{-1}\ _F$	$\ \tilde{H} - H\ _F$	$\ \tilde{Q} - Q\ _F$	$\max \tilde{\sigma} - \sigma /\sigma$
Jacobi	1.19e-5	8.83e+2	1.75e-6	4.55e-6
QR	8.67e+1	2.47e+3	2.00e+0	2.19e+1

Here the last column is for the maximum relative errors among all computed singular values. The 3rd column does not really suggest any accuracy advantage of the one-sided Jacobi SVD over `sgesvd` due to the inadequacy of $\|\tilde{H} - H\|_F$ in this graded example because it merely reflects errors in the largest entries of \tilde{H} . All errors associated with QR are unacceptably large, as expected; this is because SVD by QR can obscure all singular values by as much as $\mathcal{O}(\epsilon_m \|B\|_2)$ which in this example is bigger than the smallest

²<http://www.maplesoft.com/>

singular values, and consequently corresponding computed singular vectors are very much inaccurate.

We note in passing that H , if computed as $V\Sigma V^*$, can be very inaccurate even with highly relatively accurate V and Σ by one-sided Jacobi SVD. In fact for this example, we get $\|(\tilde{H} - H)S^{-1}\|_{\mathbb{F}} = 1.68\mathbf{e} + 1$ if doing so.

4 Extensions to the Matrix Square Root of a Positive Definite Matrix

There is a natural extension of the theory in Section 2 to the perturbation of the matrix square root of a positive definite matrix that allows some kind of symmetric scaling for better conditioning. By this we mean $A = S^*TS$, where S , as above, is a scaling matrix (and usually diagonal), and T is positive definite and well-conditioned, i.e., $\|T\|_2\|T^{-1}\|_2$ is of moderate magnitude. Then we have

$$A = S^*T^{1/2}T^{1/2}S = B^*B,$$

where B takes the form in (1.5) with $G = T^{1/2}$. Assume now that A is perturbed to $\tilde{A} = S^*\tilde{T}S$ such that $\Delta T \stackrel{\text{def}}{=} \tilde{T} - T$ is sufficient tiny. Then

$$A = S^*(T + \Delta T)S = S^*T^{1/2}(I + T^{-1/2}(\Delta T)T^{-1/2})T^{1/2}S = \tilde{B}^*\tilde{B},$$

where \tilde{B} takes the form in (1.5) with

$$\tilde{G} = (I + \hat{T})T^{1/2}, \quad \hat{T} \stackrel{\text{def}}{=} T^{-1/2}(\Delta T)T^{-1/2}.$$

We have

$$\begin{aligned} \Delta G &\stackrel{\text{def}}{=} \tilde{G} - G = [(I + \hat{T})^{1/2} - I]G, \\ (\Delta G)G^{-1} &= (I + \hat{T})^{1/2} - I, \\ I - (I + (\Delta G)G^{-1})^{-1} &= I - (I + \hat{T})^{-1/2}. \end{aligned}$$

Let $\delta_p = \|\hat{T}\|_p$, where $p = 2, \mathbb{F}$. It can be verified that

$$\begin{aligned} \|(I + \hat{T})^{1/2} - I\|_p &\leq \frac{\delta_p}{1 + \sqrt{1 - \delta_2}}, \\ \|I - (I + \hat{T})^{-1/2}\|_p &\leq \frac{\delta_p}{(1 + \sqrt{1 - \delta_2})\sqrt{1 - \delta_2}}. \end{aligned}$$

Let the polar decompositions of B and \tilde{B} be given as (1.4). Then

$$A^{1/2} = H, \quad \tilde{A}^{1/2} = \tilde{H}.$$

Now apply Theorem 2.1 to get

Theorem 4.1 Let $A = S^*TS$ and $\tilde{A} = S^*\tilde{T}S$ be two $n \times n$ positive definite matrices, where S is $n \times n$ and nonsingular, and let $\hat{T} \stackrel{\text{def}}{=} T^{-1/2}(\Delta T)T^{-1/2}$ and $\delta_p = \|\hat{T}\|_p$, where $p = 2, \text{F}$. If $\delta_2 < 1$ then

$$\|(\tilde{A}^{1/2} - A^{1/2})S^{-1}\|_{\text{F}} \leq \left(\sqrt{\frac{2 - \delta_2}{1 - \delta_2}} + 1 \right) \|T^{1/2}\|_2 \frac{\delta_{\text{F}}}{1 + \sqrt{1 - \delta_2}}, \quad (4.1)$$

$$\approx \frac{\sqrt{2} + 1}{2} \|T^{1/2}\|_2 \delta_{\text{F}}. \quad (4.2)$$

A corollary of this theorem similar to Corollary 2.1 can be stated by noting that for normal S ,

$$\| |S|^{-1/2}(\tilde{A}^{1/2} - A^{1/2})|S|^{-1/2} \|_{\text{F}} \leq \|(\tilde{A}^{1/2} - A^{1/2})S^{-1}\|_{\text{F}}.$$

But we shall omit the detail.

For the same reason as for the positive polar factor, Theorem 4.1 can provide a much sharper bound than the existing one [3, 20]

$$\|\tilde{A}^{1/2} - A^{1/2}\|_{\text{F}} \leq \frac{1}{\|\tilde{A}^{-1/2}\|_2^{-1} + \|A^{-1/2}\|_2^{-1}} \|\tilde{A} - A\|_{\text{F}}. \quad (4.3)$$

Example 4.1 Take $A = B^*B$ with B as in (1.7), i.e.,

$$T = \begin{pmatrix} 129 & 10 & -56 & -43 \\ 10 & 214 & 43 & -69 \\ -56 & 43 & 505 & -164 \\ -43 & -69 & -164 & 179 \end{pmatrix}, \quad S = \begin{pmatrix} 10^6 & & & \\ & 10^4 & & \\ & & 10^2 & \\ & & & 1 \end{pmatrix}.$$

Then $A^{1/2}$ is B 's positive polar factor which is given as in (1.8). Take $E = 10^{-5} * \text{randn}(4)$ and $\Delta T = E + E^* + EE^*$. Then (4.3) yields a bound

$$\|\tilde{A}^{1/2} - A^{1/2}\|_{\text{F}} \leq 1.4288 \cdot 10^6$$

which is too big to be of any use. However, our new bound (4.1) produces

$$\|(\tilde{A}^{1/2} - A^{1/2})S^{-1}\|_{\text{F}} \leq 1.17363 \cdot 10^{-5}.$$

But can we compute $A^{1/2}$ as accurately as predicted by Theorem 4.1? Indeed we can. Analogously to Section 3, we have the following algorithm that will deliver an highly accurately computed $A^{1/2}$.

1. Decompose $T = G^*G$ (e.g., Cholesky decomposition [6]);
 2. Compute polar decomposition $GS = QH$ by (3.1), and then return $A^{1/2} = H$.
- (4.4)

Step 2 of (4.4) is guaranteed to compute a highly accurate H for $\tilde{G}S$, given the computed \tilde{G} by Step 1. Thus it suffices to show that \tilde{G} is the Cholesky factor of a nearby T in order to convince ourselves that (4.4) will work. This has been done in [6, 18] from where one can conclude that the computed \tilde{G} indeed satisfies $\tilde{G}\tilde{G}^* = T + \Delta T$ with $\|\Delta T\|_{\text{F}} = \mathcal{O}(\epsilon_m \|A\|_{\text{F}})$.

Acknowledgement

I thank Dr. Z. Drmač for his kindness to send me his one-sided Jacobi SVD code and Mr. James Money for running numerical Example 3.1.

References

- [1] E. ANDERSON, Z. BAI, C. BISCHOF, J. DEMMEL, J. DONGARRA, J. D. CROZ, A. GREENBAUM, S. HAMMARLING, A. MCKENNEY, S. OSTROUCHOV, AND D. SORENSEN, *LAPACK Users' Guide*, SIAM, Philadelphia, 3rd ed., 1999.
- [2] A. BARRLUND, *Perturbation bounds on the polar decomposition*, BIT, 30 (1990), pp. 101–113.
- [3] R. BHATIA, *Some inequalities for norm ideals*, Comm. Math. Phys., 111 (1987), pp. 33–39.
- [4] F. CHATELIN AND S. GRATTON, *On the condition numbers associated with the polar factorization of a matrix*, Numer. Linear Algebra Appl., 7 (2000), pp. 337–354.
- [5] C.-H. CHEN AND J.-G. SUN, *Perturbation bounds for the polar factors*, J. Comp. Math., 7 (1989), pp. 397–401.
- [6] J. DEMMEL, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
- [7] J. W. DEMMEL, M. GU, S. C. EISENSTAT, I. SLAPNIČAR, K. VESELIĆ, AND Z. DRMAČ, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra and its Applications, 299 (1999), pp. 21–80.
- [8] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 1160–1174.
- [9] C. KENNEY AND A. J. LAUB, *Polar decomposition and matrix sign function condition estimates*, SIAM Journal on Scientific and Statistical Computing, 12 (1991), pp. 488–504.
- [10] F. KITTANEH, *Inequalities for the Schatten p -norm. III*, Comm. Math. Phys., 104 (1986), pp. 307–310.
- [11] R.-C. LI, *A perturbation bound for the generalized polar decomposition*, BIT, 33 (1993), pp. 304–308.
- [12] ———, *New perturbation bounds for the unitary polar factor*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 327–332.
- [13] ———, *Relative perturbation bounds for the unitary polar factor*, BIT, 37 (1997), pp. 67–75.
- [14] W. LI AND W. SUN, *Perturbation bounds for unitary and subunitary polar factors*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1183–1193.
- [15] ———, *New perturbation bounds for unitary polar factors*, SIAM J. Matrix Anal. Appl., (2003). to appear.
- [16] J.-Q. MAO, *The perturbation analysis of the product of singular vector matrices UV^H* , J. Comp. Math., 4 (1986), pp. 245–248.
- [17] R. MATHIAS, *Perturbation bounds for the polar decomposition*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 588–597.
- [18] J.-G. SUN, *Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations*, Linear Algebra and Its Applications, 173 (1992), pp. 77–97.

- [19] J.-G. SUN AND C.-H. CHEN, *Generalized polar decomposition*, Math. Numer. Sinica, 11 (1989), pp. 262–273. In Chinese.
- [20] J. L. VAN HEMMEN AND T. ANDO, *An inequality for trace ideals*, Comm. Math. Phys., 76 (1980), pp. 143–148.