LINEAR ALGEBRA AND ITS APPLICATIONS

ELSEVIER Linear Algebra and its Applications 282 (1998) 297–310

On the sensitivity of the SR decomposition [1]

Xiao-Wen Chang [2]

Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver B.C., Canada V6T 1Z4

Received 9 November 1997; accepted 20 April 1998

Submitted by V. Mehrmann

Abstract

First-order componentwise and normwise perturbation bounds for the SR decomposition are presented. The new normwise bounds are at least as good as previously known results. In particular, for the R factor, the normwise bound can be significantly tighter than the previous result.

Keywords: SR decomposition; Sensitivity; Condition estimate

1. Introduction

Let $A \in \mathbb{R}^{2n\times 2n}$, and let $P = [e_1, e_3, \ldots, e_{2n-1}, e_2, e_4, \ldots, e_{2n}]$ with e_k denoting the kth unit vector. Let

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

If all even leading principal submatrices of $PA^{\mathrm{T}}JAP^{\mathrm{T}}$ are nonsingular, then Bunse-Gerstner [4] showed that A can be factored as

$$A = SR \equiv \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}, \tag{1}$$

[1] Supported by Killam and NSERC postdoctoral fellowships.

[2] Current address: School of Computer Science, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7. Tel.: +1 514 398 8259; fax: +1 514 398 3883; e-mail: chang@cs.mcgill. ca.

PII: S0024-3795(98)10077-0

where S satisfies

$$S^{\mathrm{T}}JS = J,$$

and is called the *symplectic* matrix; R_{ij}, $i,j = 1,2$, are upper triangular, and $\mathrm{diag}(R_{21}) = 0$. This is called the SR decomposition. In order to make the factorization unique, we require

$$\mathrm{diag}(R_{11}) = |\mathrm{diag}(R_{22})|, \qquad \mathrm{diag}(R_{12}) = 0. \tag{2}$$

The existence and uniqueness of the SR decomposition satisfying Eq. (2) can easily be shown by following the idea of Theorem 3.8 in [4]. In this paper when we refer to the SR decomposition we assume that R satisfies Eq. (2). The SR decomposition is a useful tool in the computation of some optimal control problems. For more details, see for example [4,5,10].

Suppose ΔA is small enough that all even leading principal submatrices of $P(A+\Delta A)^{\mathrm{T}}J(A+\Delta A)P^{\mathrm{T}}$ are still nonsingular, so that $A+\Delta A$ has a unique SR decomposition

$$A + \Delta A = (S + \Delta S)(R + \Delta R).$$

The goal of the sensitivity analysis for the SR factorization is to determine a bound on $\|\Delta S\|$ (or $|\Delta S|$) and a bound on $\|\Delta R\|$ (or $|\Delta R|$) in terms of $\|\Delta A\|$ (or $|\Delta A|$).

The sensitivity analysis of the SR factorization has been considered by Bhatia [2], who gave first-order normwise perturbation bounds. In [2] it is assumed that $\mathrm{diag}(R_{11}) = \mathrm{diag}(R_{22})$ instead of the first equality in Eq. (2). But a simple example like

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

shows that such an SR decomposition may not exist even though all even leading principal submatrices of $PA^{\mathrm{T}}JAP^{\mathrm{T}}$ are nonsingular. However the perturbation bounds derived in [2] are correct if we require the first equality in Eq. (2) to hold. The purpose of this paper is to derive tighter first-order bounds.

Before proceeding, let us introduce some notation. Let $B = (b_{ij}) \in \mathbb{R}^{n\times n}$, we define the upper triangular matrix

$$\mathrm{up}(B) \equiv \begin{bmatrix} \frac{1}{2}b_{11} & b_{12} & \cdot & b_{1n} \\ 0 & \frac{1}{2}b_{22} & \cdot & b_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}b_{nn} \end{bmatrix}, \tag{3}$$

and use $\mathrm{sut}(B)$ to denote the strictly upper triangular part of B, i.e.,

$$\mathrm{sut}(B) \equiv \begin{bmatrix} 0 & b_{12} & b_{13} & \cdot & b_{1n} \\ 0 & 0 & b_{23} & \cdot & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & b_{n-1,n} \\ 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}. \tag{4}$$

For any

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \quad \text{with } B_{ij} \in \mathbb{R}^{n\times n} \quad (i,j = 1,2),$$

we define (b denotes "block")

$$\mathrm{bup}(B) \equiv \begin{bmatrix} \mathrm{sut}(B_{11}) & \mathrm{up}(B_{12}) \\ \mathrm{up}(B_{21}) & \mathrm{sut}(B_{22}) \end{bmatrix}. \tag{5}$$

The rest of this paper is organized as follows. In Section 2 we first derive expressions for $\dot{S}(0)$ and $\dot{R}(0)$ in the SR decomposition $A + tG = S(t)R(t)$, then use these expressions to derive the first-order componentwise and normwise perturbation bounds for R and S, respectively. In Section 3 we give numerical examples and suggest practical condition estimates. Finally we briefly summarize our findings in Section 4.

2. Main results

2.1. Rate of change of S and R

Here we derive, for later use, the basic results on how S and R change as A changes.

Theorem 1. *Given $A \in \mathbb{R}^{2n\times 2n}$. Suppose all even leading principal submatrices of $PA^{\mathrm{T}}JAP^{\mathrm{T}}$ are nonsingular and suppose A has the SR decomposition $A = SR$. Let $\Delta A \in \mathbb{R}^{2n\times 2n}$ satisfy $\Delta A = \epsilon G$. If ϵ is small enough that all even leading principal submatrices of $P(A + tG^{\mathrm{T}})J(A + tG)P^{\mathrm{T}}$ are still nonsingular for $|t| \leqslant \epsilon$, then $A + tG$ has a unique SR decomposition*

$$A + tG = S(t)R(t), \quad |t| \leqslant \epsilon, \tag{6}$$

which leads to:

$$\dot{S}(0) = GR^{-1} + SJ\,\mathrm{bup}(R^{-\mathrm{T}}G^{\mathrm{T}}JS + S^{\mathrm{T}}JGR^{-1}), \tag{7}$$

$$\dot{R}(0) = -J\,\mathrm{bup}(R^{-\mathrm{T}}G^{\mathrm{T}}JS + S^{\mathrm{T}}JGR^{-1})R. \tag{8}$$

In particular, $A + \Delta A$ has the SR decomposition

$$A + \Delta A = (S + \Delta S)(R + \Delta R) \tag{9}$$

with ΔR and ΔS *satisfying*:

$$\Delta S = \epsilon\dot{S}(0) + O(\epsilon^2), \tag{10}$$

$$\Delta R = \epsilon\dot{R}(0) + O(\epsilon^2). \tag{11}$$

Proof. Since for any $|t| \leqslant \epsilon$ all even leading principal submatrices of $P(A+tG)^{\mathrm{T}}J(A+tG)P^{\mathrm{T}}$ are nonsingular, $A+tG$ has the unique SR decomposition (6). Note that $R(0)=R$, $R(\epsilon)=R+\Delta R$, $S(0)=S$, and $S(\epsilon)=S+\Delta S$. When $t=\epsilon$, Eq. (6) becomes Eq. (9). Since $S(t)^{\mathrm{T}}JS(t)=J$, from Eq. (6) we have

$$(A+tG)^{\mathrm{T}}J(A+tG) = R(t)^{\mathrm{T}}JR(t).$$

Differentiating this at $t=0$ gives

$$A^{\mathrm{T}}JG + G^{\mathrm{T}}JA = R^{\mathrm{T}}J\dot{R}(0) + \dot{R}(0)^{\mathrm{T}}JR. \tag{12}$$

Premultiplying by $R^{-\mathrm{T}}$ and post-multiplying by R^{-1} on both sides of the above equation and using $A=SR$, we obtain

$$R^{-\mathrm{T}}\dot{R}(0)^{\mathrm{T}}J + J\dot{R}(0)R^{-1} = R^{-\mathrm{T}}G^{\mathrm{T}}JS + S^{\mathrm{T}}JGR^{-1}. \tag{13}$$

Now we want to use the special structure of R and $\dot{R}(0)$ to give an expression for $J\dot{R}(0)R^{-1}$ in Eq. (13). In order to do this, we write:

$$R(t) \equiv \begin{bmatrix} R_{11}(t) & R_{12}(t) \\ R_{21}(t) & R_{22}(t) \end{bmatrix}, \qquad \dot{R}(0) \equiv \begin{bmatrix} \dot{R}_{11}(0) & \dot{R}_{12}(0) \\ \dot{R}_{21}(0) & \dot{R}_{22}(0) \end{bmatrix},$$

$$\dot{R}(0)R^{-1} \equiv \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, \tag{14}$$

where all subblocks are $n \times n$ matrices. Note that for $|t| \leqslant \epsilon$, $PR(t)P^{\mathrm{T}}$ is a nonsingular *upper triangular* matrix with diagonal elements $(R_{11}(t))_{ii}$ and $(R_{22}(t))_{ii}$, $i=1,\ldots,n$, thus $(R_{11}(t))_{ii} = |(R_{22}(t))_{ii}| \neq 0$. It is easy to show that U_{ij}, $i,j=1,2$ are all upper triangular, $\operatorname{diag}(U_{12}) = \operatorname{diag}(U_{21}) = 0$, and

$$(U_{11})_{ii} = \frac{(\dot{R}_{11}(0))_{ii}}{(R_{11})_{ii}}, \quad (U_{22})_{ii} = \frac{(\dot{R}_{22}(0))_{ii}}{(R_{22})_{ii}}, \qquad i=1,\ldots,n. \tag{15}$$

Since for $|t| \leqslant \epsilon$ and $i=1,\ldots,n$

$$(R_{11}(t))_{ii} = |(R_{22}(t))_{ii}| = \operatorname{sgn}(R_{22}(t))_{ii} \cdot (R_{22}(t))_{ii},$$

by Taylor expansion theory we have

$$(R_{11} + t\dot{R}_{11}(0) + O(t^2))_{ii} = \operatorname{sgn}(R_{22}(t))_{ii} \cdot (R_{22} + t\dot{R}_{22}(0) + O(t^2))_{ii}. \tag{16}$$

Since $(R_{22}(t))_{ii}$ is a continuous function of t and $(R_{22}(t))_{ii} \neq 0$, we must have

$$\operatorname{sgn}(R_{22}(t))_{ii} = \operatorname{sgn}(R_{22})_{ii}, \quad i=1,\ldots,n. \tag{17}$$

Then from Eq. (16) we obtain

$$(\dot{R}_{11}(0))_{ii} = \mathrm{sgn}(R_{22})_{ii} \cdot (\dot{R}_{22}(0))_{ii}, \quad i = 1, \ldots, n.$$

By this and Eq. (15) we have

$$(U_{11})_{ii} = \frac{\mathrm{sgn}(R_{22})_{ii} \cdot (\dot{R}_{22}(0))_{ii}}{\mathrm{sgn}(R_{22})_{ii} \cdot (R_{22})_{ii}} = (U_{22})_{ii}, \quad i = 1, \ldots, n. \tag{18}$$

From Eq. (14) we have

$$R^{-\mathrm{T}}\dot{R}(0)^{\mathrm{T}}J + J\dot{R}(0)R^{-1} = \begin{bmatrix} -U_{21}^{\mathrm{T}} + U_{21} & U_{11}^{\mathrm{T}} + U_{22} \\ -U_{22}^{\mathrm{T}} - U_{11} & U_{12}^{\mathrm{T}} - U_{12} \end{bmatrix}.$$

It follows from Eq. (13), the structures of U_{ij}, $i, j = 1, 2$ and Eq. (18) that with the notation "bup" in Eq. (5)

$$J\dot{R}(0)R^{-1} = \mathrm{bup}(R^{-\mathrm{T}}G^{\mathrm{T}}JS + S^{\mathrm{T}}JGR^{-1}),$$

which gives Eq. (8).

Differentiating $A + tG = S(t)R(t)$ at $t = 0$ we have

$$G = \dot{S}(0)R + S\dot{R}(0),$$

so

$$\dot{S}(0) = GR^{-1} - S\dot{R}(0)R^{-1}.$$

Then Eq. (7) follows from this and Eq. (8).

Finally the Taylor expansions for $S(t)$ and $R(t)$ about $t = 0$ give Eqs. (10) and (11). □

2.2. Sensitivity analysis for R

From Eq. (8) it follows that

$$|\dot{R}(0)| \leqslant |J|\mathrm{bup}(|R^{-\mathrm{T}}| \cdot |G^{\mathrm{T}}| \cdot |J| \cdot |S| + |S^{\mathrm{T}}| \cdot |J| \cdot |G| \cdot |R^{-1}|)|R|.$$

Then by Eq. (11) and $\Delta A = \epsilon G$ we have the following componentwise bound

$$|\Delta R| \lesssim |J|\mathrm{bup}(|R^{-\mathrm{T}}| \cdot |\Delta A^{\mathrm{T}}| \cdot |J| \cdot |S| + |S^{\mathrm{T}}| \cdot |J| \cdot |\Delta A| \cdot |R^{-1}|)|R|.$$

Now we derive a normwise bound by using a similar approach to one of our approaches for the sensitivity analysis of the R in the QR factorization developed in [7]. Let $\mathscr{D}_{2n}$ be the set of all $2n \times 2n$ real positive definitive diagonal matrices. For any

$$D \equiv \mathrm{diag}(D^{(1)}, D^{(2)}) \equiv \mathrm{diag}(\delta_1^{(1)}, \ldots, \delta_n^{(1)}, \delta_1^{(2)}, \ldots, \delta_n^{(2)}) \in \mathscr{D}_{2n}, \tag{19}$$

let $R = D\bar{R}$. Note for any $B \in \mathbb{R}^{2n \times 2n}$ we have $\mathrm{bup}(BD) = \mathrm{bup}(B)D$. Hence if we define $B \equiv S^{\mathrm{T}}JG\bar{R}^{-1}$, then from Eq. (8) we have

$$\begin{aligned}\dot{R}(0) &= -J\,\mathrm{bup}(D^{-1}\bar{R}^{-\mathrm{T}}G^{\mathrm{T}}JSD + S^{\mathrm{T}}JG\bar{R}^{-1})\bar{R} \\ &= -J[\mathrm{bup}(B) - D^{-1}\mathrm{bup}(B^{\mathrm{T}})D]\bar{R}.\end{aligned} \tag{20}$$

To bound this we need the following lemma.

Lemma 2. *For any $B \in \mathbb{R}^{2n\times 2n}$ and $D \in \mathscr{D}_{2n}$,*

$$\varphi \equiv \|\mathrm{bup}(B) - D^{-1}\mathrm{bup}(B^{\mathrm{T}})D\|_F \leqslant \sqrt{1+\zeta_D^2}\,\|B\|_F, \tag{21}$$

where

$$\zeta_D \equiv \max\left\{\max_{i<j}\left\{\frac{\delta_j^{(1)}}{\delta_i^{(1)}}, \frac{\delta_j^{(2)}}{\delta_i^{(2)}}\right\}, \max_{i\leqslant j}\left\{\frac{\delta_j^{(1)}}{\delta_i^{(2)}}, \frac{\delta_j^{(2)}}{\delta_i^{(1)}}\right\}\right\}. \tag{22}$$

Proof. Let

$$B \equiv \begin{bmatrix} K & L \\ M & N \end{bmatrix}$$

with $K, L, M, N \in \mathbb{R}^{n\times n}$. Then

$$\begin{aligned}\varphi^2 &= \left\|\begin{bmatrix} \mathrm{sut}(K) - (D^{(1)})^{-1}\mathrm{sut}(K^{\mathrm{T}})D^{(1)} & \mathrm{up}(L) - (D^{(1)})^{-1}\mathrm{up}(M^{\mathrm{T}})D^{(2)} \\ \mathrm{up}(M) - (D^{(2)})^{-1}\mathrm{up}(L^{\mathrm{T}})D^{(1)} & \mathrm{sut}(N) - (D^{(2)})^{-1}\mathrm{sut}(N^{\mathrm{T}})D^{(2)} \end{bmatrix}\right\|_F^2 \\ &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(k_{ij} - \frac{\delta_j^{(1)}}{\delta_i^{(1)}}k_{ji}\right)^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(n_{ij} - \frac{\delta_j^{(2)}}{\delta_i^{(2)}}n_{ji}\right)^2 \\ &\quad + \left[\sum_{i=1}^{n}\frac{1}{4}\left(l_{ii} - \frac{\delta_i^{(2)}}{\delta_i^{(1)}}m_{ii}\right)^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(l_{ij} - \frac{\delta_j^{(2)}}{\delta_i^{(1)}}m_{ji}\right)^2\right. \\ &\quad \left. + \sum_{i=1}^{n}\frac{1}{4}\left(m_{ii} - \frac{\delta_i^{(1)}}{\delta_i^{(2)}}l_{ii}\right)^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(m_{ij} - \frac{\delta_j^{(1)}}{\delta_i^{(2)}}l_{ji}\right)^2\right] \\ &\equiv \varphi_1 + \varphi_2 + \varphi_3.\end{aligned}$$

By the Cauchy–Schwartz theorem,

$$\left(k_{ij} - \frac{\delta_j^{(1)}}{\delta_i^{(1)}}k_{ji}\right)^2 \leqslant (k_{ij}^2 + k_{ji}^2)\left[1 + \left(\frac{\delta_j^{(1)}}{\delta_i^{(1)}}\right)^2\right],$$

so

$$\varphi_1 \leqslant \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[1 + \left(\frac{\delta_j^{(1)}}{\delta_i^{(1)}}\right)^2\right](k_{ij}^2 + k_{ji}^2) \leqslant \left[1 + \max_{i<j}\left(\frac{\delta_j^{(1)}}{\delta_i^{(1)}}\right)^2\right]\|K\|_F^2.$$

Similarly,

$$\varphi_2 \leqslant \left[1 + \max_{i<j}\left(\frac{\delta_j^{(2)}}{\delta_i^{(2)}}\right)^2\right]\|N\|_F^2.$$

Now we bound φ_3

$$\begin{aligned}\varphi_3 \leqslant & \frac{1}{4}\sum_{i=1}^{n}\left[1 + \left(\frac{\delta_i^{(2)}}{\delta_i^{(1)}}\right)^2\right](l_{ii}^2 + m_{ii}^2) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[1+\left(\frac{\delta_j^{(2)}}{\delta_i^{(1)}}\right)^2\right](l_{ij}^2 + m_{ji}^2)\\ & + \frac{1}{4}\sum_{i=1}^{n}\left[1 + \left(\frac{\delta_i^{(1)}}{\delta_i^{(2)}}\right)^2\right](l_{ii}^2 + m_{ii}^2) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[1+\left(\frac{\delta_j^{(1)}}{\delta_i^{(2)}}\right)^2\right](l_{ji}^2 + m_{ij}^2)\\ \leqslant & \left[1 + \max_{i\leqslant j}\left\{\left(\frac{\delta_j^{(1)}}{\delta_i^{(2)}}\right)^2, \left(\frac{\delta_j^{(2)}}{\delta_i^{(1)}}\right)^2\right\}\right](\|L\|_F^2 + \|M\|_F^2).\end{aligned}$$

Thus with ζ_D in Eq. (22) we have

$$\varphi^2 \leqslant (1 + \zeta_D^2)(\|K\|_F^2 + \|L\|_F^2 + \|M\|_F^2 + \|N\|_F^2) = (1 + \zeta_D^2)\|B\|_F^2. \qquad \square$$

We can now bound $\dot{R}(0)$ in Eq. (20):

$$\begin{aligned}\|\dot{R}(0)\|_F &\leqslant \varphi\|\bar{R}\|_2 \leqslant \sqrt{1 + \zeta_D^2}\|S^{\mathrm{T}}\|_2\|J\|_2\|G\|_F\|\bar{R}^{-1}\|_2\|\bar{R}\|_2\\ &= \sqrt{1 + \zeta_D^2}\kappa_2(D^{-1}R)\|S\|_2\|G\|_F.\end{aligned} \tag{23}$$

Since this is true for any $D \in \mathscr{D}_{2n}$, we have:

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_F} \leqslant \kappa_R(A)\frac{\|G\|_F}{\|A\|_F}, \tag{24}$$

$$\kappa_R(A) \equiv \inf_{D\in\mathscr{D}_{2n}} \kappa_R(A,D), \tag{25}$$

$$\kappa_R(A,D) \equiv \sqrt{1 + \zeta_D^2}\kappa_2(D^{-1}R)\frac{\|S\|_2\|A\|_F}{\|R\|_F}. \tag{26}$$

Thus from the Taylor expansion (11) and $\Delta A = \epsilon G$ we obtain

$$\frac{\|\Delta R\|_F}{\|R\|_F} \lesssim \kappa_R(A)\frac{\|\Delta A\|_F}{\|A\|_F}. \tag{27}$$

Clearly $\kappa_R(A)$ can be regarded as a measure of the sensitivity of the R factor in the SR decomposition. Since a *condition number* as a function of matrix of a certain class has to be from a bound which is attainable to first-order for any matrix in the given class, we use a qualified term *condition estimate* when this criterion is not met. For general A the bound (24) (or the bound (27)) may not be attainable, i.e, for some A, we cannot find $G \neq 0$ such that the inequality in Eq. (24) becomes an equality. Therefore we say $\kappa_R(A)$ is a condition estimate for the R factor in the SR decomposition. Certainly we can use the so called matrix-vector equation approach developed by Chang [6] (see also [7]) to derive the condition number by Eq. (12), but it would be tedious.

If we take $D = I$ in Eq. (26), then $\zeta_D = 1$, and from Eq. (27) we obtain the following bound:

$$\frac{\|\Delta R\|_F}{\|R\|_F} \lesssim \kappa_R(A, I)\frac{\|\Delta A\|_F}{\|A\|_F} = \sqrt{2}\kappa_2(R)\frac{\|S\|_2\|A\|_F}{\|R\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}, \tag{28}$$

or

$$\|\Delta R\|_F \lesssim \sqrt{2}\kappa_2(R)\|S\|_2\|\Delta A\|,$$

which is due to Bhatia [2]. We see the new first-order bound Eq. (27) is at least as good as Eq. (28). Our analysis shows the sensitivity of R in the SR decomposition is dependent on the row scaling in $R = D\bar{R}$. If the ill conditioning of R is mostly due to the bad scaling of its rows, then the correct choice of D in $R = D\bar{R}$ can give $\kappa_2(\bar{R})$ very near one. If at the same time ζ_D is not large, then $\kappa_R(A, D)$ can be much smaller than $\kappa_R(A, I)$. So potentially the bound (28) can severely overestimate the true sensitivity. Let us give a simple example to illustrate this. Let

$$R = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & \epsilon & 0 & 0 \\ \hline 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & \epsilon \end{array}\right]$$

with very small positive ϵ. Take

$$D^{(1)} = D^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}.$$

Then $\zeta_D = 1$, and it is easy to obtain $\kappa_R(A, D)/\kappa_R(A, I) = O(\epsilon)$.

We will consider how to choose D in Eq. (26) for general case in Section 3. Certainly if R has good row scaling, Bhatia's condition estimate for the R factor will be as good as the new one. For example if R is an $2n \times 2n$ identity matrix, then it is easy to show $\kappa_R(A) = \kappa_R(A, I)$.

Since $S^TJS = J$, we have $JS = S^{-T}J$, which gives $\|S\|_2 = \|S^{-1}\|_2$. Since $A = SR$, we have $\|A\|_F \leqslant \|S\|_2\|R\|_F$. Thus from Eq. (28) we obtain the following weaker but simpler bound:

$$\frac{\|\Delta R\|_F}{\|R\|_F} \lesssim \sqrt{2}\kappa_2(S)\kappa_2(R)\frac{\|\Delta A\|_F}{\|A\|_F}.$$

2.3. Sensitivity analysis for S

From Eq. (7) we obtain

$$|\dot{S}(0)| \leqslant |G||R^{-1}| + |S||J|\text{bup}(|R^{-T}||G^T|J||S| + |S^T||J||G||R^{-1}|),$$

then from Eq. (10) and $A = \epsilon G$ we have the componentwise bound

$$|\Delta S| \lesssim |\Delta A||R^{-1}| + |S||J|\text{bup}(|R^{-T}||\Delta A^T||J||S| + |S^T||J||\Delta A||R^{-1}|).$$

Now we derive a normwise bound. Multiplying S^TJ on both sides of Eq. (7) and using $S^TJS = J$ gives

$$S^TJ\dot{S}(0) = S^TJGR^{-1} - \text{bup}(R^{-T}G^TJS + S^TJGR^{-1}). \tag{29}$$

For any D having the form of Eq. (19), let $S = \bar{S}D$. Then if we define $B \equiv \bar{S}^TJGR^{-1}$, we have from Eq. (29) that

$$\begin{aligned}\bar{S}^TJ\dot{S}(0) &= \bar{S}^TJGR^{-1} - \text{bup}(D^{-1}R^{-T}G^TJ\bar{S}D + \bar{S}^TJGR^{-1}) \\ &= B - \text{bup}(B - D^{-1}B^TD).\end{aligned} \tag{30}$$

In order to bound this we need the following lemma which is similar to Lemma 2.

Lemma 3. *For any $B \in \mathbb{R}^{2n\times 2n}$ and $D \in \mathscr{D}_{2n}$,*

$$\psi \equiv \|B - \text{bup}(B) - D^{-1}\,\text{bup}(B^T)D\|_F \leqslant \sqrt{1 + \zeta_D^2}\|B\|_F. \tag{31}$$

where ζ_D is defined by Eq. (22).

Proof. The proof is similar to that of Lemma 2, so we omit it. □

Now we can bound $\bar{S}^TJ\dot{S}(0)$ in Eq. (30):

$$\|\bar{S}^TJ\dot{S}(0)\|_F \leqslant \psi \leqslant \sqrt{1 + \zeta_D^2}\|\bar{S}\|_2\|G\|_F\|R^{-1}\|_2.$$

Thus

$$\begin{aligned}\|\dot{S}(0)\|_F = \|J\dot{S}(0)\|_F &= \|\bar{S}^{-T}\bar{S}^TJ\dot{S}(0)\|_F \\ &\leqslant \sqrt{1 + \zeta_D^2}\|\bar{S}^{-1}\|_2\|\bar{S}\|_2\|G\|_F\|R^{-1}\|_2 \\ &= \sqrt{1 + \zeta_D^2\kappa_2(SD^{-1})}\|R^{-1}\|_2\|G\|_F.\end{aligned} \tag{32}$$

Since this is true for any $D \in \mathscr{D}_{2n}$, we have:

$$\frac{\|\dot{S}(0)\|_F}{\|S\|_F} \leqslant \kappa_S(A)\frac{\|G\|_F}{\|A\|_F}, \tag{33}$$

$$\kappa_S(A) \equiv \inf_{D\in\mathscr{D}_{2n}} \kappa_S(A,D), \tag{34}$$

$$\kappa_S(A,D) \equiv \sqrt{1+\zeta_D^2\kappa_2(SD^{-1})}\frac{\|R^{-1}\|_2\|A\|_F}{\|S\|_F}. \tag{35}$$

Then from the Taylor expansion (10) and $\Delta A = \epsilon G$ we obtain

$$\frac{\|\Delta S\|_F}{\|S\|_F} \lesssim \kappa_S(A)\frac{\|\Delta A\|_F}{\|A\|_F}. \tag{36}$$

So $\kappa_S(A)$ is a condition estimate for the S factor in the SR decomposition.

If we take $D = I$ in Eq. (35), then $\zeta_D = 1$, and we obtain the following bound:

$$\frac{\|\Delta S\|_F}{\|S\|_F} \lesssim \kappa_S(A,I)\frac{\|\Delta A\|_F}{\|A\|_F} = \sqrt{2}\kappa_2(S)\frac{\|R^{-1}\|_2\|A\|_F}{\|S\|_F}\frac{\|\Delta A\|_F}{\|A\|_F} \tag{37}$$

or

$$\|\Delta S\|_F \lesssim \sqrt{2}\kappa_2(S)\|R^{-1}\|_2\|\Delta A\|_F,$$

which is due to Bhatia [2]. We see the new first-order bound (36) is at least as good as Eq. (37). But so far we have not found an example to show that $\kappa_S(A)$ can be arbitrarily smaller than $\kappa_S(A,I)$.

Using $\|A\|_F \leqslant \|S\|_F\|R\|_2$ we obtain from Eq. (37) the following weaker but simpler bound:

$$\frac{\|\Delta S\|_F}{\|S\|_F} \leqslant \sqrt{2}\kappa_2(S)\kappa_2(R)\frac{\|\Delta A\|_F}{\|A\|_F}.$$

3. Numerical experiments

In Section 2 we derived new condition estimates for R and S. Our perturbation results are tighter than previous results.

The optimization problems (25) and (34) are complicated. In practice we would like to choose D such that $\kappa_R(A,D)$ is a good approximation to the infimum $\kappa_R(A)$ and choose another D such that $\kappa_S(A,D)$ is a good approximation to the infimum $\kappa_S(A)$.

By a well-known result of van der Sluis [9], $\kappa_2(D^{-1}R)$ will be nearly minimal when the rows of $D^{-1}R$ are equilibrated. But this could lead to a large ζ_D in Eq. (22). So a reasonable compromise is to choose D to equilibrate R as far as possible in some sense while keeping $\zeta_D = 1$. There are four obvious possibilities for D:

- $\delta_1^{(1)} = \sqrt{\sum_{j=1}^{2n} r_{1j}^2}$,

$$\delta_i^{(1)} = \begin{cases} \sqrt{\sum_{j=1}^{2n} r_{ij}^2} & \text{if } \sqrt{\sum_{j=1}^{2n} r_{ij}^2} \leqslant \delta_{i-1}^{(1)}, \\ \delta_{i-1}^{(1)} & \text{otherwise}, \end{cases} \quad i = 2, \ldots, n,$$

$D^{(2)} = D^{(1)}$.

- $\delta_1^{(2)} = \sqrt{\sum_{j=1}^{2n} r_{n+1,j}^2}$,

$$\delta_i^{(2)} = \begin{cases} \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2} & \text{if } \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2} \leqslant \delta_{i-1}^{(2)}, \\ \delta_{i-1}^{(2)} & \text{otherwise}, \end{cases} \quad i = 2, \ldots, n,$$

$D^{(1)} = D^{(2)}$.

- $\delta_1^{(1)} = \max\left\{\sqrt{\sum_{j=1}^{2n} r_{1j}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+1,j}^2}\right\}$,

$$\delta_i^{(1)} = \begin{cases} \max\left\{\sqrt{\sum_{j=1}^{2n} r_{ij}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2}\right\} \\ \quad \text{if} \max\left\{\sqrt{\sum_{j=1}^{2n} r_{ij}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2}\right\} \leqslant \delta_{i-1}^{(1)}, \\ \delta_{i-1}^{(1)} \quad \text{otherwise}, \end{cases} \quad i = 2, \ldots, n,$$

$D^{(2)} = D^{(1)}$.

- $\delta_1^{(1)} = \min\left\{\sqrt{\sum_{j=1}^{2n} r_{1j}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+1,j}^2}\right\}$,

$$\delta_i^{(1)} = \begin{cases} \min\left\{\sqrt{\sum_{j=1}^{2n} r_{ij}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2}\right\}, \\ \quad \text{if } \min\left\{\sqrt{\sum_{j=1}^{2n} r_{ij}^2}, \sqrt{\sum_{j=1}^{2n} r_{n+i,j}^2}\right\} \leqslant \delta_{i-1}^{(1)}, \\ \delta_{i-1}^{(1)} \quad \text{otherwise}, \end{cases} \quad i = 2, \ldots, n,$$

$D^{(2)} = D^{(1)}$.

For the same reason we may use the corresponding column version of the above four methods with respect to S to scale the columns of S.

To illustrate our results and the scaling strategies above we present two sets of examples. The first set of matrices are $2n \times 2n$ `frank` matrices ($a_{ij} = 2n - j + 1, i \leqslant j$; $a_{i,i-1} = 2n - i + 1$; $a_{ij} = 0, i > j + 1$) and the second set of matrices are $2n \times 2n$ `pascal` matrices ($a_{i1} = 1$, $a_{1j} = 1$, $a_{ij} = a_{i-1,j} + a_{i,j-1}$), $n = 5, 6, 7$. Both are from The Test Matrix Toolbox for Matlab (Version 3.0) by Higham [8]. The Matlab program for computing the SR decomposition was provided by Peter Benner. The numerical results for Bhatia's condition estimates ($\kappa_R(A, I)$ and $\kappa_S(A, I)$) and our new condition estimates ($\kappa_R(A, D)$ and $\kappa_S(A, D)$) with four different choices of D for R and S are presented in Tables 1–3. In order to see whether our choice of D is good or not, we used Matlab function `fmins` to compute the local minima of $\kappa_R(A, D)$ and $\kappa_S(A, D)$ with respect to D by using the D determined above as initial points. The termination tolerance for both the variable and function is 10^{-4}, and the maximum iteration numbers for $n = 5, 6, 7$ are 2000, 2400 and 2800, respectively. The computed minima (opti$i$, $i = 1, 2, 3, 4$, corresponding to the different initial D obtained by our four different choices) are shown in Tables 1–3 too.

From Tables 1–3 we see for the R factor, Bhatia's condition estimate $\kappa_R(A, I)$ can be much larger than $\kappa_R(A, D)$ with D determined by any of the four choices. The latter is only slightly worse than the local minima computed by `fmins`. But for the S factor, Bhatia's condition estimate $\kappa_S(A, I)$ is almost the same as or slightly better than $\kappa_S(A, D)$ with D determined by the four choices. The computed local minima of $\kappa_S(A, D)$ are slightly better than $\kappa_S(A, I)$. For R, according to Tables 1–3 and our other numerical tests we do not see which choice of D is superior to others. But on average we find the third choice is preferable. For S, we suggest in practice using Bhatia's $\kappa_S(A, I)$ as the

Table 1
Condition estimates for test matrices of order 10

Method	Frank		Pascal	
	R	S	R	S
Bhatia	3.20×10^7	4.50×10^7	3.60×10^{11}	3.28×10^{11}
new1	1.51×10^4	4.50×10^7	5.17×10^6	3.28×10^{11}
opti1	1.44×10^4	3.22×10^7	2.64×10^6 [a]	2.47×10^{11} [a]
new2	1.49×10^4	4.50×10^7	1.37×10^7	3.46×10^{11}
opti2	1.44×10^4 [a]	3.22×10^7	2.52×10^6 [a]	2.50×10^{11}
new3	1.46×10^4	4.50×10^7	5.17×10^6	3.46×10^{11}
opti3	1.44×10^4	3.22×10^7	2.64×10^6 [a]	2.51×10^{11}
new4	1.49×10^4	4.50×10^7	1.37×10^7	3.28×10^{11}
opti4	1.44×10^4	3.22×10^7	2.52×10^6 [a]	2.47×10^{11} [a]

[a] The optimization algorithm stops after 2000 iterations.

Table 2
Condition estimates for test matrices of order 12

Method	Frank		Pascal	
	R	S	R	S
Bhatia	4.89×10^9	6.78×10^9	2.54×10^{14}	2.33×10^{14}
new1	2.09×10^5	6.78×10^9	2.93×10^8	2.33×10^{14}
opti1	1.97×10^5	4.80×10^9	1.23×10^8 [a]	1.79×10^{14} [a]
new2	2.06×10^5	6.78×10^9	8.85×10^8	2.63×10^{14}
opti2	1.98×10^5	4.80×10^9 [a]	1.31×10^8 [a]	1.75×10^{14}
new3	2.04×10^5	6.78×10^9	2.93×10^8	2.63×10^{14}
opti3	1.97×10^5	4.80×10^9	1.23×10^8 [a]	1.75×10^{14}
new4	2.07×10^5	6.78×10^9	8.85×10^8	2.33×10^{14}
opti4	1.98×10^5	4.80×10^9 [a]	1.31×10^8 [a]	1.79×10^{14} [a]

[a] The optimization algorithm stops after 2400 iterations.

Table 3
Condition estimates for test matrices of order 14

Method	Frank		Pascal	
	R	S	R	S
Bhatia	1.01×10^{12}	1.39×10^{12}	1.90×10^{17}	1.75×10^{17}
new1	3.32×10^6	1.39×10^{12}	1.75×10^{10}	1.75×10^{17}
opti1	3.15×10^6	9.81×10^{11}	6.99×10^9 [a]	1.34×10^{17} [a]
new2	3.24×10^6	1.39×10^{12}	5.93×10^{10}	2.46×10^{17}
opti2	3.15×10^6 [a]	9.81×10^{11}	8.16×10^9 [a]	1.33×10^{17} [a]
new3	3.26×10^6	1.39×10^{12}	1.75×10^{10}	2.46×10^{17}
opti3	3.16×10^6 [a]	9.81×10^{11}	6.99×10^9 [a]	1.33×10^{17} [a]
new4	3.27×10^6	1.39×10^{12}	5.93×10^{10}	1.75×10^{17}
opti4	3.18×10^6 [a]	9.81×10^{11} [a]	8.16×10^9 [a]	1.34×10^{17} [a]

[a] The optimization algorithm stops after 2800 iterations.

conditioning measure. Why is the effect of scaling on $\kappa_S(A, D)$ quite different from that on $\kappa_R(A, D)$? One explanation may be that R is mainly subjected to only a zero/nonzero structure constraint, but S has to be subjected to the constraint $S^TJS = J$. From the numerical experiments we also observe that S is more sensitive than R.

4. Summary

New first-order componentwise and normwise perturbation bounds have been presented for both R and S in the SR decomposition. The new condition estimates we derived are as follows:

- $\kappa_R(A) \equiv \inf_{D \in \mathcal{D}_{2n}} \kappa_R(A, D)$ for R,

where $\kappa_R(A,D) \equiv \sqrt{1+\zeta_D^2\kappa_2(D^{-1}R)}\ \|S\|_2\|A\|_F/\ \|R\|_F$
(see Eqs. (25) and (26)).

- $\kappa_S(A) \equiv \inf_{D\in\mathscr{D}_{2n}} \kappa_S(A,D)$ for S,

  where $\kappa_S(A,D) \equiv \sqrt{1+\zeta_D^2\kappa_2(SD^{-1})}\ \|R^{-1}\|_2\|A\|_F\ /\|S\|_F$
  (see Eqs. (34) and (35)).

When $D = I$, $\kappa_R(A,D)$ and $\kappa_S(A,D)$ become the condition estimates essentially obtained by Bhatia [2]. We have shown how to choose D in practice. Our numerical examples showed that $\kappa_R(A,D)$ with our choices of D can be significantly smaller than $\kappa_R(A,I)$. But they did not suggest that the corresponding results would hold for the S factor. Can $\kappa_S(A)$ be significantly smaller than $\kappa_S(A,I)$? This question is left for future study.

The techniques presented here could easily be applied to the HR decomposition (see for example [3,1]), and similar perturbation bounds could be obtained. But we chose not to do this here in order to keep the material and basic ideas as brief as possible.

Acknowledgements

The author wish to thank Peter Benner for providing the Matlab program (a modified version of the code developed by Gerd Banse) for computing the SR decomposition and helpful discussion on the uniqueness of the SR decomposition. Thanks to Chris Paige and Jim Varah for their helpful comments and suggestions.

References

[1] P. Benner, H. Fassbender, D. Watkins, Two connections between the SR and HR eigenvalue algorithms, Linear Algebra and Appl. 272 (1998) 17–32.
[2] R. Bhatia, Matrix factorizations and their perturbations, Linear Algebra and Appl. 197–198 (1994) 245–276.
[3] A. Bunse-Gerstner, An analysis of the HR algorithm for computing the eigenvalues of a matrix, Linear Algebra and Appl. 35 (1981) 155–178.
[4] A. Bunse-Gerstner, Matrix factorizations for symplectic QR-like methods, Linear Algebra and Appl. 83 (1986) 49–77.
[5] A. Bunse-Gerstner, V. Mehrmann, A symplectic QR-like algorithm for the solution of the real algebraic Riccati equation, IEEE Trans. Automat. Control. AC-31 (1986) 1104–1113.
[6] X.-W. Chang, Perturbation Analysis of Some Matrix Factorizations, Ph.D. thesis, Computer Science, McGill University, Montreal, Canada, 1997.
[7] X.-W. Chang, C.C. Paige, G.W. Stewart, Perturbation analyses for the QR factorization, SIAM J. Matrix Anal. Appl. 181 (1997) 775–791.
[8] N.J. Higham, The Test Matrix Toolbox for Matlab, version 3.0, Numerical Analysis Report No. 265, University of Manchester, Manchester, UK, 1995.
[9] A. van der Sluis, Condition numbers and equilibration of matrices, Numer. Math. 14 (1969) 14–23.
[10] D.S. Watkins, L. Elsner, Self-similar flows, Linear Algebra and Appl. 110 (1988) 213–242.