TAIWANESE JOURNAL OF MATHEMATICS
Vol. 20, No. 2, pp. 333–363, April 2016
DOI: 10.11650/tjm.20.2016.5398
This paper is available online at http://journal.tms.org.tw

Mixed, Componentwise Condition Numbers and Small Sample Statistical Condition Estimation for Generalized Spectral Projections and Matrix Sign Functions

Wei-Guo Wang, Chern-Shuh Wang*, Yi-Min Wei and Peng-Peng Xie

Abstract. The explicit expressions of mixed and componentwise condition numbers for the spectral projections, generalized spectral projections and sign functions for matrices and regular matrix pairs are derived. The condition numbers improve some known results of the normwise type and reveal the structured perturbations. Statistical condition estimation is applied to these problems which can be calculated efficiently.

1. Introduction

We consider the mixed and componentwise condition numbers for the generalized spectral projections and the sign functions for matrices and regular matrix pairs. It is well known that the spectral projections and matrix sign functions play an important role in the perturbation theory for eigenvalue problems [6,9,16,28,34]. Applying the theory of condition number developed by Rice [25], the explicit expressions of the normwise case have been well investigated in [30,31]. A nice survey of the matrix sign function is presented in [18]. The survey introduces some historical background, perturbation theory, and applications in control problem and in eigenproblems. The computation of matrix sign functions can be adopted for finding invariant subspaces corresponding to certain eigenvalues and hence can be applied to solve Riccati and Sylvester equations as well [1,2,12,14,18,22,33]. The perturbation analysis of the matrix sign function has been studied in [1,2,12,18,24,29].

Received October 7, 2014, accepted November 16, 2015.
Communicated by Ming-Chih Lai.
2010 *Mathematics Subject Classification.* 15A18, 15A22, 65F15.
Key words and phrases. Mixed, Componentwise, Condition number, Generalized spectral projections, Structured perturbations, Matrix sign functions, Matrix regular pair, Statistical condition estimation.
Partial work was finished when W. Wang visited the Shanghai Key Laboratory of Contemporary Applied Mathematics of Fudan University.
W. Wang is supported by the National Natural Science Foundation of China under grant 11371333, Shandong Province Natural Science Foundation Grant ZR2013AM025, and the Fundamental Research Funds for the Central Universities under grant 201362030.
Y. Wei is supported by the National Natural Science Foundation of China under grant 11271084.
P. Xie is supported by the National Natural Science Foundation of China under grant 11271084.
*Corresponding author.

Let $A \in \mathbb{C}^{n\times n}$ and $U \in \mathbb{C}^{n\times n}$ be a unitary matrix. The Schur decomposition [11] is

$$A = U \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix} U^H, \quad A_{11} \in \mathbb{C}^{m\times m}, \quad (m < n). \tag{1.1}$$

Assume that the eigenvalues $\lambda(A_{11}) \cap \lambda(A_{22}) = \emptyset$, then the Sylvester equation

$$A_{11}X - XA_{22} = -A_{12}, \tag{1.2}$$

has a unique solution X.

Notice that

$$A = U \begin{bmatrix} I_m & X \\ \mathbf{0} & I_{n-m} \end{bmatrix} \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix} \begin{bmatrix} I_m & X \\ \mathbf{0} & I_{n-m} \end{bmatrix}^{-1} U^H = S \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix} S^{-1}, \tag{1.3}$$

where

$$S = U \begin{bmatrix} I_m & X \\ \mathbf{0} & I_{n-m} \end{bmatrix} \equiv [S_1, S_2], \quad T = S^{-1} \equiv \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}, \quad (S_1 \in \mathbb{C}^{n\times m},\ T_1 \in \mathbb{C}^{m\times n}). \tag{1.4}$$

Let the range spaces $\mathcal{R}(S_1)$ and $\mathcal{R}(S_2)$ be the invariant subspaces of A corresponding to $\lambda(A_{11})$ and $\lambda(A_{22})$, respectively. The spectral projection of A associated with $\lambda(A_{11})$ is defined by [31]

$$P = S \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} S^{-1} = U \begin{bmatrix} I_m & -X \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^H, \tag{1.5}$$

i.e., the spectral projection P is the projection onto $\mathcal{R}(S_1)$ along $\mathcal{R}(S_2)$.

If $\lambda(A_{11}) \in \mathbb{C}^-$ (eigenvalues in the open left complex plane), $\lambda(A_{22}) \in \mathbb{C}^+$ (eigenvalues in the open right complex plane) in (1.1), then the matrix sign function $\mathrm{sign}(A)$ can be computed by [1, 2, 18, 29]

$$\mathrm{sign}(A) = U \begin{bmatrix} -I_m & Y \\ \mathbf{0} & I_{n-m} \end{bmatrix} U^H, \tag{1.6}$$

where $Y \in \mathbb{C}^{m\times(n-m)}$ is the solution of Sylvester equation: $A_{11}Y - YA_{22} = -2A_{12}$.

Let $A = ZJZ^{-1}$ be the Jordan decomposition of A [11], where

$$J = \begin{bmatrix} J_1 & \mathbf{0} \\ \mathbf{0} & J_2 \end{bmatrix}, \qquad J_1 \in \mathbb{C}^{m\times m}, \quad J_2 \in \mathbb{C}^{(n-m)\times(n-m)}, \quad (\lambda(J_1) \in \mathbb{C}^-, \quad \lambda(J_2) \in \mathbb{C}^+).$$

It is well known that the matrix sign function $\mathrm{sign}(A)$ can also be characterized by [2,12,29]

$$\mathrm{sign}(A) = Z \begin{bmatrix} -I_m & \mathbf{0} \\ \mathbf{0} & I_{n-m} \end{bmatrix} Z^{-1}. \tag{1.7}$$

Let the regular matrix pair (A, B) of order n have the generalized Schur decomposition [11]

$$A = Q \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix} Z^H, \quad B = Q \begin{bmatrix} B_{11} & B_{12} \\ \mathbf{0} & B_{22} \end{bmatrix} Z^H, \tag{1.8}$$

where $\lambda(A_{11}, B_{11}) \in \mathbb{C}^+$, $\lambda(A_{22}, B_{22}) \in \mathbb{C}^-$, in which we denote $\lambda(A, B)$ the set of all eigenvalues of a regular matrix pair (A, B) (i.e., $\det(A + \lambda B) \not\equiv 0$ for $\lambda \in \mathbb{C}$) [11]. Two matrix sign functions of (A, B) can be defined as follows [1, 32, 33]:

$$\mathrm{sign}_L(A, B) = Q \begin{bmatrix} I_m & -2G \\ \mathbf{0} & -I_{n-m} \end{bmatrix} Q^H, \quad \mathrm{sign}_R(A, B) = Z \begin{bmatrix} I_m & -2H \\ \mathbf{0} & -I_{n-m} \end{bmatrix} Z^H,$$

where G and H are determined by the generalized Sylvester equations [7]

$$A_{11}H - GA_{22} = -A_{12}, \quad B_{11}H - GB_{22} = -B_{12}. \tag{1.9}$$

It is easy to see that

$$\mathrm{sign}_L(A, B) = \mathrm{sign}(AB^{-1}), \quad \mathrm{sign}_R(A, B) = \mathrm{sign}(B^{-1}A).$$

The normwise condition numbers can often overestimate the actual condition since the structures of the perturbation are ignored [26]. The mixed and componentwise condition numbers [10] have been presented for Moore-Penrose inverse and linear least squares problems [5], the total least squares [35], Tikhonov regularization [4], the generalized Sylvester equation [23] and symmetric algebraic Riccati equations [36]. In this paper, we investigate the mixed and componentwise condition numbers of the spectral projection, generalized spectral projection and two matrix sign functions. While in practical computation, it may be difficult to estimate these condition numbers. Kenney and Laub [17] developed the small-sample statistical condition estimates for general matrix functions, which is especially useful for the case where Fréchet derivative is known implicitly. We apply the idea of statistical condition estimation to the spectral projections and matrix sign functions.

In Sections 3-6, we derive the mixed and componentwise condition numbers of the spectral projection, generalized spectral projection and matrix sign functions and obtain computable upper bounds. The statistical condition estimation will be introduced in Section 7. In Section 8, we report some numerical results.

Throughout this paper, we use $\mathbb{C}^{m\times n}$ to denote the set of $m\times n$ complex matrices, $\mathbb{R}^{m\times n}$ the set of $m\times n$ real matrices. A^T denotes the transpose of a matrix A, $\overline{A}$ the conjugate of a matrix A, and $A^H=\overline{A}^T$, the conjugate transpose of A. I stands for the identity matrix, and $\mathbf{0}$ denotes the null matrix. $\|\cdot\|_F$ represents the Frobenius norm, $\|\cdot\|_{\max}$ for the largest one of the absolute values of the entries, and $\|\cdot\|_2$ the spectral matrix norm or the Euclidean vector norm. $\mathcal{R}(A)$ is the column space of A. $\lambda(A)$ denotes the set of all eigenvalues of a square matrix A, and $\lambda(A,B)$ the set of all eigenvalues of a regular matrix pair (A,B) [11]. For $A=[a_1,a_2,\ldots,a_n]=(a_{ij})\in\mathbb{R}^{m\times n}$ and a matrix $B=(b_{ij})\in\mathbb{R}^{p\times q}$, $A\otimes B=(a_{ij}B)\in\mathbb{R}^{mp\times nq}$ is a Kronecker product, and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A)=[a_1^T,a_2^T,\ldots,a_n^T]^T$ (see [11,15,27] for properties of the Kronecker product and vec operation). For the normal distribution with mean μ and variance σ, we denote $\mathbb{N}(\mu,\sigma^2)$.

2. Preliminaries

To deal with mixed and componentwise condition numbers, the following definition of "distance" function will be helpful. For any points $a,b\in\mathbb{R}^n$, we define $\frac{a}{b}=(c_1,c_2,\ldots,c_n)^T$ with

$$c_i=\begin{cases}a_i/b_i, & \text{if } b_i\neq 0,\\ 0, & \text{if } a_i=b_i=0,\\ \infty, & \text{otherwise.}\end{cases}$$

Then we define

$$d(a,b)=\left\|\frac{a-b}{b}\right\|_\infty=\max_{i=1,2,\ldots,n}\left\{\frac{|a_i-b_i|}{|b_i|}\right\}.$$

Note that if $d(a,b)<\infty$,

$$d(a,b)=\min\left\{v\geq 0\mid |a_i-b_i|\leq v|b_i|,\ \text{for } i=1,2,\ldots,n\right\}.$$

The distance of two matrices can be defined as

$$d(A,B)=d(\mathrm{vec}(A),\mathrm{vec}(B)).$$

It is easy to know that $\|\mathrm{vec}(A)\|_\infty=\|A\|_{\max}$.

We recall the definition in [5]. For $\varepsilon>0$, we denote $B^0(a,\varepsilon)=\{x\mid d(x,a)\leq\varepsilon\}$. For a partial function $F\colon\mathbb{R}^p\to\mathbb{R}^q$, hereafter, $\mathrm{Dom}(F)$ denotes the domain of F.

Definition 2.1. Let $F\colon\mathbb{R}^p\to\mathbb{R}^q$ be a continuous mapping defined on an open set $\mathrm{Dom}(F)\subset\mathbb{R}^p$ such that $0\notin\mathrm{Dom}(F)$. Let $a\in\mathrm{Dom}(F)$ such that $F(a)\neq 0$.

(i) The mixed condition number of F at a is defined by

$$m(F;a)=\lim_{\varepsilon\to 0}\sup_{\substack{x\in B^0(a,\varepsilon)\\ x\neq a}}\frac{\|F(x)-F(a)\|_\infty}{\|F(a)\|_\infty}\frac{1}{d(x,a)}.$$

(ii) Suppose that $F(a) = (f_1(a), f_2(a), \ldots, f_q(a))^T$ is such that $f_j(a) \neq 0$ for $j = 1, 2, \ldots, q$. Then the componentwise condition number of F at a is

$$c(F; a) = \lim_{\varepsilon \to 0} \sup_{\substack{x \in B^0(a,\varepsilon) \\ x \neq a}} \frac{d(F(x), F(a))}{d(x, a)}.$$

The explicit expressions of the mixed and componentwise condition numbers of F at a are determined by the following lemma.

Lemma 2.2. [10] *Suppose F is Fréchet differentiable at a. Then*

(a) *If $F(a) \neq 0$, then* $m(F; a) = \dfrac{\|F'(a)\,\mathrm{Diag}(a)\|_\infty}{\|F(a)\|_\infty} = \dfrac{\||F'(a)|\,|a|\|_\infty}{\|F(a)\|_\infty}$

(b) *If $(F(a))_i \neq 0$ for $i = 1, 2, \ldots, q$, then* $c(F; a) = \left\|\mathrm{Diag}(F(a))^{-1} F'(a)\,\mathrm{Diag}(a)\right\|_\infty = \left\|\frac{|F'(a)||a|}{|F(a)|}\right\|_\infty$, *where* $\mathrm{Diag}(a)$ *is a $p \times p$ diagonal matrix with $a_1, a_2, \ldots, a_p$ on the diagonal.*

Note that in the rest of this paper, once we deal with componentwise condition numbers, we assume that the solution has non-zero components.

3. The spectral projection

Let A be perturbed to $\widetilde{A} = A + \Delta A$, and the spectral projection P be perturbed to $\widetilde{P} = P + \Delta P$, respectively.

Let

$$\Omega = I_{n-m} \otimes A_{11} - A_{22}^T \otimes I_m, \quad \Gamma = A_{11}^T \otimes I_{n-m} - I_m \otimes A_{22}. \tag{3.1}$$

If $\|\Delta A\|_F$ is sufficiently small, then the explicit expression of $\mathrm{vec}(\Delta P)$ can be approximated by [31],

$$\mathrm{vec}(\Delta P) \approx \Phi\,\mathrm{vec}(\Delta A), \tag{3.2}$$

where

$$\Phi = (T_2^T \otimes S_1)\Omega^{-1}(S_2^T \otimes T_1) + (T_1^T \otimes S_2)\Gamma^{-1}(S_1^T \otimes T_2). \tag{3.3}$$

Now we define the mapping

$$\psi_P\colon\ \mathrm{vec}(A) \to \mathrm{vec}(P),$$

where P is the spectral projection [31].

The mixed and componentwise condition numbers for the spectral projection P can be defined as follows:

$$m_P(A) = \lim_{\varepsilon\to 0} \sup_{\left\|\frac{\Delta A}{A}\right\|_{\max}\le\varepsilon} \frac{\|\Delta P\|_{\max}}{\|P\|_{\max}} \frac{1}{\left\|\frac{\Delta A}{A}\right\|_{\max}},$$

$$c_P(A) = \lim_{\varepsilon\to 0} \sup_{\left\|\frac{\Delta A}{A}\right\|_{\max}\le\varepsilon} \frac{1}{\left\|\frac{\Delta A}{A}\right\|_{\max}} \left\|\frac{\Delta P}{P}\right\|_{\max}.$$

Here $\frac{B}{A}$ is an entrywise division defined by $\frac{B}{A} := \operatorname{unvec}\left(\frac{\operatorname{vec}(B)}{\operatorname{vec}(A)}\right)$.

Remark 3.1. $\operatorname{unvec}(a)$ is an operator which transforms a vector into a matrix with appropriate orders.

The main result of this section is the following theorem. It provides explicit expressions for the mixed and componentwise condition numbers for the spectral projection.

Theorem 3.2. *Let $\|\Delta A\|_F$ be sufficiently small. Using the notations above, we have*

$$m_P(A) = \frac{\||\Phi|\operatorname{vec}(|A|)\|_\infty}{\|\operatorname{vec}(P)\|_\infty}, \tag{3.4}$$

$$c_P(A) = \left\|\frac{|\Phi|\operatorname{vec}(|A|)}{\operatorname{vec}(P)}\right\|_\infty. \tag{3.5}$$

Proof. It follows from (3.2) that

$$\psi'_P(A) = \Phi.$$

From (a) of Lemma 2.2, we obtain

$$m_P(A) = m(\psi_P; a) = \frac{\||\psi'_P(a)|\,|a|\|_\infty}{\|\psi_P(a)\|_\infty} = \frac{\||\Phi|\operatorname{vec}(|A|)\|_\infty}{\|\operatorname{vec}(P)\|_\infty},$$

and

$$c_P(A) = c(\psi_P; a) = \left\|\frac{|\Phi|\,|a|}{|\psi_P(a)|}\right\|_\infty = \left\|\frac{|\Phi|\operatorname{vec}(|A|)}{\operatorname{vec}(P)}\right\|_\infty. \qquad \square$$

Theorem 3.2 presents explicit expressions for the condition numbers $m_P(A)$ and $c_P(A)$. But they will not be easy to compute due to their dependance on the Kronecker products. In order to obtain the easier computable upper bounds, we list the following lemma.

Lemma 3.3. *For any matrices M, N, P, Q, R, and S with dimensions making the following well defined*

$$(M\otimes N)P(Q\otimes R)\operatorname{vec}(S), \quad N\left(\operatorname{vec}^{-1}(P\operatorname{vec}(RSQ^T))\right)M^T,$$

let $\mathfrak{Y} = \operatorname{unvec}(P\operatorname{vec}(RSQ^T))$. Then we have

$$(M\otimes N)P(Q\otimes R)\operatorname{vec}(S) = \operatorname{vec}\left(N\mathfrak{Y}M^T\right).$$

Proof. Using the Kronecker product, this lemma is easy to prove. □

The following corollary displays upper bounds which are not difficult to estimate for these condition numbers.

Corollary 3.4. *In the hypothesis of Theorem* 3.2*, we have*

$$m_P(A) \le \frac{\left\| |S_1|\,\mathfrak{Y}_1\,|T_2| + |S_2|\,\mathfrak{Y}_2\,|T_1| \right\|_{\max}}{\|P\|_{\max}}, \tag{3.6}$$

$$c_P(A) \le \left\| \frac{|S_1|\,\mathfrak{Y}_1\,|T_2| + |S_2|\,\mathfrak{Y}_2\,|T_1|}{|P|} \right\|_{\max}, \tag{3.7}$$

where $\mathfrak{Y}_1 = \operatorname{unvec}\left(\left|\Omega^{-1}\right| \operatorname{vec}(|T_1|\,|A|\,|S_2|)\right)$, $\mathfrak{Y}_2 = \operatorname{unvec}\left(\left|\Gamma^{-1}\right| \operatorname{vec}(|T_2|\,|A|\,|S_1|)\right)$.

Proof. Using Theorem 3.2 and Lemma 3.3, it is obvious that

$$\begin{aligned}
|\Phi| \operatorname{vec}(|A|) &= \left|(T_2^T \otimes S_1)\Omega^{-1}(S_2^T \otimes T_1) + (T_1^T \otimes S_2)\Gamma^{-1}(S_1^T \otimes T_2)\right| \operatorname{vec}(|A|) \\
&\le \left|(T_2^T \otimes S_1)\right| \left|\Omega^{-1}\right| \left|(S_2^T \otimes T_1)\right| \operatorname{vec}(|A|) \\
&\quad + \left|(T_1^T \otimes S_2)\right| \left|\Gamma^{-1}\right| \left|(S_1^T \otimes T_2)\right| \operatorname{vec}(|A|) \\
&\le \left|(T_2^T \otimes S_1)\right| \left|\Omega^{-1}\right| \operatorname{vec}(|T_1||A||S_2|) + \left|(T_1^T \otimes S_2)\right| \left|\Gamma^{-1}\right| \operatorname{vec}(|T_2||A||S_1|) \\
&= \left|(T_2^T \otimes S_1)\right| \operatorname{vec}(\mathfrak{Y}_1) + \left|(T_1^T \otimes S_2)\right| \operatorname{vec}(\mathfrak{Y}_2) \\
&= \operatorname{vec}(|S_1|\,\mathfrak{Y}_1\,|T_2|) + \operatorname{vec}(|S_2|\,\mathfrak{Y}_2\,|T_1|).
\end{aligned}$$

Taking norms (and dividing by P), the corollary follows. □

4. The generalized spectral projection

Let (A, B) be a regular matrix pair, U and V be unitary matrices. The generalized Schur decomposition [11] is

$$A = V \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix} U^H, \quad B = V \begin{bmatrix} B_{11} & B_{12} \\ \mathbf{0} & B_{22} \end{bmatrix} U^H, \tag{4.1}$$

where $A_{11}, B_{11} \in \mathbb{C}^{m\times m}$ $(m < n)$ [11]. Suppose that $\lambda(A_{11}, B_{11}) \cap \lambda(A_{22}, B_{22}) = \emptyset$, then the generalized Sylvester equation [7, 23]

$$\begin{aligned}
A_{11}M - NA_{22} &= -A_{12}, \\
B_{11}M - NB_{22} &= -B_{12},
\end{aligned} \tag{4.2}$$

has a unique solution (M, N). Let

$$\begin{aligned} S &= U\begin{bmatrix} I_m & M \\ \mathbf{0} & I_{n-m} \end{bmatrix} \equiv [S_1, S_2], \qquad T = S^{-1} \equiv \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}, \qquad (S_1, T_1^T \in \mathbb{C}^{n\times m}), \\ Q &= V\begin{bmatrix} I_m & N \\ \mathbf{0} & I_{n-m} \end{bmatrix} \equiv [Q_1, Q_2], \qquad R = Q^{-1} \equiv \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \qquad (Q_1, R_1^T \in \mathbb{C}^{n\times m}). \end{aligned} \tag{4.3}$$

Then we obtain

$$Q^{-1}AS = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix}, \quad Q^{-1}BS = \begin{bmatrix} B_{11} & \mathbf{0} \\ \mathbf{0} & B_{22} \end{bmatrix}, \tag{4.4}$$

which indicates that $\mathcal{R}(S_1)$ and $\mathcal{R}(Q_1)$ are the right and left deflating subspaces of (A, B) corresponding to $\lambda(A_{11}, B_{11})$ [27,30].

The generalized spectral projections P_r and P_l of (A, B) associated with $\lambda(A_{11}, B_{11})$ are defined by [27,30]

$$\begin{aligned} P_r &= S\begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} S^{-1} = U\begin{bmatrix} I_m & -M \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^H, \\ P_l &= Q\begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} Q^{-1} = V\begin{bmatrix} I_m & -N \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^H, \end{aligned} \tag{4.5}$$

i.e., the generalized spectral projections P_r and P_l are the projections onto $\mathcal{R}(S_1)$ and $\mathcal{R}(Q_1)$ along $\mathcal{R}(S_2)$ and $\mathcal{R}(Q_2)$, respectively.

Let the matrix pair (A, B) be perturbed to $(\widetilde{A}, \widetilde{B}) = (A, B) + (\Delta A, \Delta B)$, and the generalized spectral projections P_r be perturbed to $\widetilde{P_r} = P_r + \Delta P_r$ and P_l to $\widetilde{P_l} = P_l + \Delta P_l$.

Assume that $\|(\Delta A, \Delta B)\|_F$ is sufficiently small. Let

$$\Omega = B_{22}^T \otimes A_{11} - A_{22}^T \otimes B_{11}, \quad \Gamma = A_{11}^T \otimes B_{22} - B_{11}^T \otimes A_{22}, \tag{4.6}$$

$$\begin{bmatrix} -(B_{22}^T \otimes I_m)\Omega^{-1} & (A_{22}^T \otimes I_m)\Omega^{-1} \\ -(I_{n-m} \otimes B_{11})\Omega^{-1} & (I_{n-m} \otimes A_{11})\Omega^{-1} \end{bmatrix} \equiv \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \tag{4.7}$$

and

$$\begin{bmatrix} -(B_{11}^T \otimes I_{n-m})\Gamma^{-1} & (A_{11}^T \otimes I_{n-m})\Gamma^{-1} \\ -(I_m \otimes B_{22})\Gamma^{-1} & (I_m \otimes A_{22})\Gamma^{-1} \end{bmatrix} \equiv \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}. \tag{4.8}$$

Then the explicit expressions of $\mathrm{vec}(\Delta P_r)$ and $\mathrm{vec}(\Delta P_l)$ are approximated in [30],

$$\mathrm{vec}(\Delta P_r) \approx [\Phi_{11}, \Phi_{12}] \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \mathrm{vec}(\Delta B) \end{bmatrix}, \tag{4.9}$$

and

$$\operatorname{vec}(\Delta P_l) \approx [\Phi_{21}, \Phi_{22}] \begin{bmatrix} \operatorname{vec}(\Delta A) \\ \operatorname{vec}(\Delta B) \end{bmatrix}, \tag{4.10}$$

where

$$\Phi_{1j} = (T_2^T \otimes S_1) D_{1j} (S_2^T \otimes R_1) - (T_1^T \otimes S_2) C_{1j} (S_1^T \otimes R_2), \quad (j = 1, 2), \tag{4.11}$$

and

$$\Phi_{2j} = (R_2^T \otimes Q_1) D_{2j} (S_2^T \otimes R_1) - (R_1^T \otimes Q_2) C_{2j} (S_1^T \otimes R_2), \quad (j = 1, 2). \tag{4.12}$$

Now we define the mappings

$$\psi_{P_r} \colon \operatorname{vec}(A, B) \to \operatorname{vec}(P_r), \quad \psi_{P_l} \colon \operatorname{vec}(A, B) \to \operatorname{vec}(P_l),$$

where P_r and P_l are the generalized spectral projections.

The following main result provides explicit expressions for the mixed and componentwise condition numbers for the generalized spectral projections.

Theorem 4.1. *Let $\|(\Delta A, \Delta B)\|_F$ be sufficiently small. Using the notations above, we have*

$$m_{P_r}(A) = \frac{\left\| |[\Phi_{11}, \Phi_{12}]| \begin{bmatrix} \operatorname{vec}(|A|) \\ \operatorname{vec}(|B|) \end{bmatrix} \right\|_\infty}{\|\operatorname{vec}(P_r)\|_\infty}, \quad m_{P_l}(A) = \frac{\left\| |[\Phi_{21}, \Phi_{22}]| \begin{bmatrix} \operatorname{vec}(|A|) \\ \operatorname{vec}(|B|) \end{bmatrix} \right\|_\infty}{\|\operatorname{vec}(P_l)\|_\infty}, \tag{4.13}$$

$$c_{P_r}(A) = \left\| \frac{|[\Phi_{11}, \Phi_{12}]| \begin{bmatrix} \operatorname{vec}(|A|) \\ \operatorname{vec}(|B|) \end{bmatrix}}{\operatorname{vec}(P_r)} \right\|_\infty, \quad c_{P_l}(A) = \left\| \frac{|[\Phi_{21}, \Phi_{22}]| \begin{bmatrix} \operatorname{vec}(|A|) \\ \operatorname{vec}(|B|) \end{bmatrix}}{\operatorname{vec}(P_l)} \right\|_\infty. \tag{4.14}$$

Proof. It follows from (4.9) that $\psi'_{P_r}(A, B) = [\Phi_{11}, \Phi_{12}]$. From (a) of Lemma 2.2, we obtain

$$m_{P_r}(A, B) = m(\psi_{P_r}; a) = \frac{\left\| \left| \psi'_{P_r}(a) \right| |a| \right\|_\infty}{\|\psi_{P_r}(a)\|_\infty} = \frac{|[\Phi_{11}, \Phi_{12}]| \begin{bmatrix} \operatorname{vec}(|A|) \\ \operatorname{vec}(|B|) \end{bmatrix}_\infty}{\|\operatorname{vec}(P_r)\|_\infty},$$

and

$$c_{P_r}(A,B) = c(\psi_{P_r};a) = \left\| \frac{|\Phi|\,|a|}{|\psi_{P_r}(a)|} \right\|_\infty = \left\| \frac{|[\Phi_{11},\Phi_{12}]| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \end{bmatrix}}{\mathrm{vec}(P_r)} \right\|_\infty .$$

In a similar way, the remainder part of the theorem follows. □

Theorem 4.1 presents explicit expressions for the mixed and componentwise condition numbers for the generalized spectral projections. The following corollary gives upper bounds which are convenient to compute for these condition numbers.

Corollary 4.2. *Based on the hypothesis of Theorem* 4.1*, we have*

$$m_{P_r}(A) \le \frac{\| |S_1|(\mathfrak{Y}_{11}+\mathfrak{Y}_{12})|T_2| + |S_2|(\mathfrak{Z}_{11}+\mathfrak{Z}_{12})|T_1| \|_{\max}}{\|P_r\|_{\max}}, \tag{4.15}$$

$$m_{P_l}(A) \le \frac{\| |Q_1|(\mathfrak{Y}_{21}+\mathfrak{Y}_{22})|R_2| + |Q_2|(\mathfrak{Z}_{21}+\mathfrak{Z}_{22})|R_1| \|_{\max}}{\|P_l\|_{\max}}, \tag{4.16}$$

$$c_{P_r}(A) \le \left\| \frac{|S_1|(\mathfrak{Y}_{11}+\mathfrak{Y}_{12})|T_2| + |S_2|(\mathfrak{Z}_{11}+\mathfrak{Z}_{12})|T_1|}{|P_r|} \right\|_{\max}, \tag{4.17}$$

$$c_{P_l}(A) \le \left\| \frac{|Q_1|(\mathfrak{Y}_{21}+\mathfrak{Y}_{22})|R_2| + |Q_2|(\mathfrak{Z}_{21}+\mathfrak{Z}_{22})|R_1|}{|P_l|} \right\|_{\max}, \tag{4.18}$$

where $\mathfrak{Y}_{ij} = \mathrm{unvec}(Y_{ij})$, $\mathfrak{Z}_{ij} = \mathrm{unvec}(Z_{ij})$, $(i,j=1,2)$, *with*

$$\begin{bmatrix} Y_{11}+Y_{12} \\ Y_{21}+Y_{22} \end{bmatrix} = \begin{bmatrix} |D_{11}| & |D_{12}| \\ |D_{21}| & |D_{22}| \end{bmatrix} \begin{bmatrix} \mathrm{vec}(|R_1|\,|A|\,|S_2|) \\ \mathrm{vec}(|R_1|\,|B|\,|S_2|) \end{bmatrix},$$

$$\begin{bmatrix} Z_{11}+Z_{12} \\ Z_{21}+Z_{22} \end{bmatrix} = \begin{bmatrix} |C_{11}| & |C_{12}| \\ |C_{21}| & |C_{22}| \end{bmatrix} \begin{bmatrix} \mathrm{vec}(|R_2|\,|A|\,|S_1|) \\ \mathrm{vec}(|R_2|\,|B|\,|S_1|) \end{bmatrix}.$$

Proof. The inequalities can be proved similarly as Corollary 3.4. □

5. The matrix sign function

Let $A = XJX^{-1}$ be the Jordan decomposition of $A \in \mathbb{C}^{n\times n}$. Moreover, let $\Delta A \in \mathbb{C}^{n\times n}$ be

$$\Delta A = XFX^{-1}, \quad F = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \quad F_{11} \in \mathbb{C}^{m\times m}.$$

If $\|\Delta A\|$ is sufficiently small for a consistent norm $\|\cdot\|$, then it holds [29]

$$\operatorname{sign}(A+\Delta A)=\operatorname{sign}(A)+X\begin{bmatrix}\mathbf{0} & L_1\\ -2K_1 & \mathbf{0}\end{bmatrix}X^{-1}+\mathscr{O}(\|\Delta A\|^2),$$

where K_1 and L_1 are determined by two Sylvester equations

$$K_1J_1-J_2K_1=F_{21},\quad J_1L_1-L_1J_2=-2F_{12}.$$

Denote

$$\Delta_{\operatorname{sign}}(A)=\operatorname{sign}(A+\Delta A)-\operatorname{sign}(A),\quad L_{\operatorname{sign}}(A,\Delta A)=X\begin{bmatrix}\mathbf{0} & L_1\\ -2K_1 & \mathbf{0}\end{bmatrix}X^{-1}.$$

Let $X=[X_1,X_2]$, $Y=X^{-H}=[Y_1,Y_2]$, where $X_1,Y_1\in\mathbb{C}^{n\times m}$. From [29, Eqn. (3.26)], we know that

$$\operatorname{vec}(L_{\operatorname{sign}}(A,\Delta A))=\Psi\operatorname{vec}(\Delta A), \tag{5.1}$$

where

$$\begin{aligned}\Psi=-2\big[&(\overline{Y}_1\otimes X_2)(J_1^T\otimes I_{n-m}-I_m\otimes J_2)^{-1}(X_1^T\otimes Y_2^H)\\&+(\overline{Y}_2\otimes X_1)(I_{n-m}\otimes J_1-J_2^T\otimes I_m)^{-1}(X_2^T\otimes Y_1^H)\big].\end{aligned}$$

Now we define the mapping

$$\psi_{\operatorname{sign}}\colon\ \operatorname{vec}(A)\to\operatorname{vec}(\operatorname{sign}(A)),$$

where $\operatorname{sign}(A)$ is the matrix sign function.

The mixed and componentwise condition numbers of the matrix sign function $\operatorname{sign}(A)$ are defined as follows:

$$\begin{aligned}m_{\operatorname{sign}}(A)&=\lim_{\varepsilon\to0}\sup_{\left\|\frac{\Delta A}{A}\right\|_{\max}\le\varepsilon}\frac{\|\Delta_{\operatorname{sign}}(A)\|_{\max}}{\|\operatorname{sign}(A)\|_{\max}}\frac{1}{\left\|\frac{\Delta A}{A}\right\|_{\max}},\\ c_{\operatorname{sign}}(A)&=\lim_{\varepsilon\to0}\sup_{\left\|\frac{\Delta A}{A}\right\|_{\max}\le\varepsilon}\frac{1}{\|\frac{\Delta A}{A}\|_{\max}}\left\|\frac{\Delta_{\operatorname{sign}}(A)}{\operatorname{sign}(A)}\right\|_{\max}.\end{aligned}$$

The following theorem derives explicit expressions for the condition numbers for the matrix sign function.

Theorem 5.1. *Let $\|\Delta A\|_F$ be sufficiently small. Then*

$$m_{\operatorname{sign}}(A)=\frac{\||\Psi|\operatorname{vec}(|A|)\|_\infty}{\|\operatorname{vec}(\operatorname{sign}(A))\|_\infty}, \tag{5.2}$$

$$c_{\text{sign}}(A) = \left\| \frac{|\Psi|\,\text{vec}(|A|)}{\text{vec}(\text{sign}(A))} \right\|_\infty, \tag{5.3}$$

where

$$\begin{aligned}\Psi = -2\big[&(\overline{Y}_1 \otimes X_2)(J_1^T \otimes I_{n-m} - I_m \otimes J_2)^{-1}(X_1^T \otimes Y_2^H) \\ &+ (\overline{Y}_2 \otimes X_1)(I_{n-m} \otimes J_1 - J_2^T \otimes I_m)^{-1}(X_2^T \otimes Y_1^H)\big].\end{aligned}$$

Proof. It follows from (5.1) that $\psi'_P(A) = \Psi$. From (a) of Lemma 2.2, we obtain

$$m_{\text{sign}}(A) = m(\psi_{\text{sign}}; a) = \frac{\left\| \left|\psi'_{\text{sign}}(a)\right| |a| \right\|_\infty}{\|\psi_{\text{sign}}(a)\|_\infty} = \frac{\||\Psi|\,\text{vec}(|A|)\|_\infty}{\|\text{vec}(\text{sign}(A))\|_\infty},$$

and

$$c_{\text{sign}}(A) = c(\psi_{\text{sign}}; a) = \left\| \frac{|\Psi|\,|a|}{|\psi_{\text{sign}}(a)|} \right\|_\infty = \left\| \frac{|\Psi|\,\text{vec}(|A|)}{\text{vec}(\text{sign}(A))} \right\|_\infty. \qquad \square$$

Theorem 5.1 reveals explicit expressions for the condition numbers $m_{\text{sign}}(A)$ and $c_{\text{sign}}(A)$. The following corollary presents computable upper bounds for these condition numbers.

Corollary 5.2. *Based on the hypothesis of Theorem* 5.1*, we have*

$$m_{\text{sign}}(A) \le \frac{\left\|2(|X_2|\,\mathfrak{W}_1\,|Y_1^H| + |X_1|\,\mathfrak{W}_2\,|Y_2^T|)\right\|_{\max}}{\|\text{sign}(A)\|_{\max}}, \tag{5.4}$$

$$c_{\text{sign}}(A) \le \left\| \frac{2(|X_2|\,\mathfrak{W}_1\,|Y_1^H| + |X_1|\,\mathfrak{W}_2\,|Y_2^T|)}{|\,\text{sign}(A)|} \right\|_{\max}, \tag{5.5}$$

where

$$\begin{aligned}\mathfrak{W}_1 &= \text{unvec}\left(\left|(J_1^T \otimes I_{n-m} - I_m \otimes J_2)^{-1}\right| \text{vec}(|Y_2^H|\,|A|\,|X_1|)\right), \\ \mathfrak{W}_2 &= \text{unvec}\left(\left|(I_{n-m} \otimes J_1 - J_2^T \otimes I_m)^{-1}\right| \text{vec}(|Y_1^H|\,|A|\,|X_2|)\right).\end{aligned}$$

Proof. Using Theorem 5.1 and Lemma 3.3, it is easy to see that

$$\begin{aligned}|\Psi|\,\text{vec}(|A|) &= \big|-2\big[(\overline{Y}_1 \otimes X_2)(J_1^T \otimes I_{n-m} - I_m \otimes J_2)^{-1}(X_1^T \otimes Y_2^H) \\ &\quad + (\overline{Y}_2 \otimes X_1)(I_{n-m} \otimes J_1 - J_2^T \otimes I_m)^{-1}(X_2^T \otimes Y_1^H)\big]\big|\,\text{vec}(|A|) \\ &\le 2\big[\left|(\overline{Y}_1 \otimes X_2)(J_1^T \otimes I_{n-m} - I_m \otimes J_2)^{-1}\right| \text{vec}(|Y_2^H|\,|A|\,|X_1|) \\ &\quad + \left|(\overline{Y}_2 \otimes X_1)(I_{n-m} \otimes J_1 - J_2^T \otimes I_m)^{-1}\right| \text{vec}(|Y_1^H|\,|A|\,|X_2|)\big] \\ &\le 2\left[\left|\overline{Y}_1 \otimes X_2\right| \text{vec}(\mathfrak{W}_1) + \left|\overline{Y}_2 \otimes X_1\right| \text{vec}(\mathfrak{W}_2)\right] \\ &= 2\left(|X_2|\,\mathfrak{W}_1\,|Y_1^H| + |X_1|\,\mathfrak{W}_2\,|Y_2^T|\right).\end{aligned}$$

Taking norms (and dividing by sign(A)), it proves the corollary. $\square$

6. The sign functions of regular matrix pairs

The following theorem is a direct result of [3, Theorem 3.1].

Theorem 6.1. *Let the regular matrix pair (A, B) have the generalized Schur decomposition* (1.8)*. Moreover, let $E, F \in \mathbb{C}^{n\times n}$ be expressed by*

$$E = Q\begin{bmatrix} E_{11} & E_{12} \\ \mathbf{0} & E_{22} \end{bmatrix} Z^H, \quad F = Q\begin{bmatrix} F_{11} & F_{12} \\ \mathbf{0} & F_{22} \end{bmatrix} Z^H. \tag{6.1}$$

If $\|[E, F]\|$ is sufficiently small for a consistent norm $\|\cdot\|$, then

$$\begin{aligned} \operatorname{sign}_L(A+E, B+F) &= \operatorname{sign}_L(A, B) + 2Q\begin{bmatrix} GP_1 & -G_1 \\ P_1 & -P_1G \end{bmatrix} Q^H + \mathscr{O}(\|[E, F]\|^2), \\ \operatorname{sign}_R(A+E, B+F) &= \operatorname{sign}_R(A, B) + 2Z\begin{bmatrix} HK_1 & -H_1 \\ K_1 & -K_1H \end{bmatrix} Z^H + \mathscr{O}(\|[E, F]\|^2), \end{aligned} \tag{6.2}$$

where G, H, P_1, K_1 and G_1, H_1 are determined by two generalized Sylvester equations (1.9) *and*

$$\begin{aligned} P_1A_{11} - A_{22}K_1 &= E_{21}, \\ P_1B_{11} - B_{22}K_1 &= F_{21}, \end{aligned} \tag{6.3}$$

and

$$\begin{aligned} A_{11}H_1 - G_1A_{22} &= G(E_{22} - P_1A_{12}) - (E_{11} + A_{12}K_1)H - E_{12}, \\ B_{11}H_1 - G_1B_{22} &= G(F_{22} - P_1B_{12}) - (F_{11} + B_{12}K_1)H - F_{12}, \end{aligned} \tag{6.4}$$

respectively.

The above theorem shows that the Fréchet derivatives $L_{\operatorname{sign}_L}((A, B), (E, F))$ and $L_{\operatorname{sign}_R}((A, B), (E, F))$ for the two matrix sign functions of (A, B) can be expressed by

$$\begin{aligned} L_{\operatorname{sign}_L}((A, B), (E, F)) &= 2Q\begin{bmatrix} GP_1 & -G_1 \\ P_1 & -P_1G \end{bmatrix} Q^H, \\ L_{\operatorname{sign}_R}((A, B), (E, F)) &= 2Z\begin{bmatrix} HK_1 & -H_1 \\ K_1 & -K_1H \end{bmatrix} Z^H. \end{aligned} \tag{6.5}$$

Let $Q = [Q_1, Q_2]$, $Z = [Z_1, Z_2]$, where $Q_1, Z_1 \in \mathbb{C}^{n\times m}$. Then

$$\begin{aligned} L_{\operatorname{sign}_L}((A, B), (E, F)) &= 2\left(Q_1GP_1Q_1^H - Q_1G_1Q_2^H + Q_2P_1Q_1^H - Q_2P_1GQ_2^H\right), \\ L_{\operatorname{sign}_R}((A, B), (E, F)) &= 2\left(Z_1HK_1Z_1^H - Z_1H_1Z_2^H + Z_2K_1Z_1^H - Z_2K_1HZ_2^H\right). \end{aligned}$$

So we have

$$\begin{aligned}\operatorname{vec}(L_{\operatorname{sign}_L}((A,B),(E,F))) &= 2\left[\overline{Q_1}\otimes(Q_1G)+\overline{Q_1}\otimes Q_2-(\overline{Q_2}G^T)\otimes Q_2\right]\operatorname{vec}(P_1)\\ &\quad -2(\overline{Q_2}\otimes Q_1)\operatorname{vec}(G_1),\\ \operatorname{vec}(L_{\operatorname{sign}_R}((A,B),(E,F))) &= 2\left[\overline{Z_1}\otimes(Z_1H)+\overline{Z_1}\otimes Z_2-(\overline{Z_2}H^T)\otimes Z_2\right]\operatorname{vec}(K_1)\\ &\quad -2(\overline{Z_2}\otimes Z_1)\operatorname{vec}(H_1).\end{aligned} \tag{6.6}$$

Denote

$$T=\begin{bmatrix}I_{n-m}\otimes A_{11} & -A_{22}^T\otimes I_m\\ I_{n-m}\otimes B_{11} & -B_{22}^T\otimes I_m\end{bmatrix},\quad S=\begin{bmatrix}A_{11}^T\otimes I_{n-m} & -I_m\otimes A_{22}\\ B_{11}^T\otimes I_{n-m} & -I_m\otimes B_{22}\end{bmatrix},$$

$$T^{-1}=\begin{bmatrix}T_1\\ T_2\end{bmatrix},\quad S^{-1}=\begin{bmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{bmatrix},$$

where $T_{11},S_{11}\in\mathbb{C}^{m(n-m)\times m(n-m)}$. Thus from (6.3) and (6.4), we have

$$\begin{bmatrix}\operatorname{vec}(P_1)\\ \operatorname{vec}(K_1)\end{bmatrix}=S^{-1}\begin{bmatrix}\operatorname{vec}(E_{21})\\ \operatorname{vec}(F_{21})\end{bmatrix}=\begin{bmatrix}S_{11}(Z_1^T\otimes Q_2^H)\\ S_{21}(Z_1^T\otimes Q_2^H)\end{bmatrix}\operatorname{vec}(E)+\begin{bmatrix}S_{12}(Z_1^T\otimes Q_2^H)\\ S_{22}(Z_1^T\otimes Q_2^H)\end{bmatrix}\operatorname{vec}(F)$$

and

$$\begin{aligned}\begin{bmatrix}\operatorname{vec}(H_1)\\ \operatorname{vec}(G_1)\end{bmatrix}&=T^{-1}\begin{bmatrix}\operatorname{vec}(G(E_{22}-P_1A_{12})-(E_{11}+A_{12}K_1)H-E_{12})\\ \operatorname{vec}(G(F_{22}-P_1B_{12})-(F_{11}+B_{12}K_1)H-F_{12})\end{bmatrix}\\ &=T^{-1}M\operatorname{vec}(E)+T^{-1}N\operatorname{vec}(F)\\ &=\begin{bmatrix}T_1\\ T_2\end{bmatrix}M\operatorname{vec}(E)+\begin{bmatrix}T_1\\ T_2\end{bmatrix}N\operatorname{vec}(F).\end{aligned}$$

We take

$$M=\begin{bmatrix}Z_2^T\otimes(GQ_2^H)-(H^TZ_1^T)\otimes Q_1^H-Z_2^T\otimes Q_1^H-(A_{12}^T\otimes G)S_{11}(Z_1^T\otimes Q_2^H)-(H^T\otimes A_{12})S_{21}(Z_1^T\otimes Q_2^H)\\ -(B_{12}^T\otimes G)S_{11}(Z_1^T\otimes Q_2^H)-(H^T\otimes B_{12})S_{21}(Z_1^T\otimes Q_2^H)\end{bmatrix},$$

$$N=\begin{bmatrix}-(A_{12}^T\otimes G)S_{12}(Z_1^T\otimes Q_2^H)-(H^T\otimes A_{12})S_{22}(Z_1^T\otimes Q_2^H)\\ Z_2^T\otimes(GQ_2^H)-(H^TZ_1^T)\otimes Q_1^H-Z_2^T\otimes Q_1^H-(B_{12}^T\otimes G)S_{12}(Z_1^T\otimes Q_2^H)-(H^T\otimes B_{12})S_{22}(Z_1^T\otimes Q_2^H)\end{bmatrix}.$$

Hence

$$\begin{aligned}\operatorname{vec}(L_{\operatorname{sign}_L}((A,B),(E,F)))&=[L_1,L_2]\begin{bmatrix}\operatorname{vec}(E)\\ \operatorname{vec}(F)\end{bmatrix},\\ \operatorname{vec}(L_{\operatorname{sign}_R}((A,B),(E,F)))&=[L_3,L_4]\begin{bmatrix}\operatorname{vec}(E)\\ \operatorname{vec}(F)\end{bmatrix},\end{aligned} \tag{6.7}$$

where

$$L_1 = 2\left[\overline{Q_1}\otimes(Q_1G) + \overline{Q_1}\otimes Q_2 - (\overline{Q_2}G^T)\otimes Q_2\right] S_{11}(Z_1^T\otimes Q_2^H) - 2(\overline{Q_2}\otimes Q_1)T_2M,$$
$$L_2 = 2\left[\overline{Q_1}\otimes(Q_1G) + \overline{Q_1}\otimes Q_2 - (\overline{Q_2}G^T)\otimes Q_2\right] S_{12}(Z_1^T\otimes Q_2^H) - 2(\overline{Q_2}\otimes Q_1)T_2N,$$
$$L_3 = 2\left[\overline{Z_1}\otimes(Z_1H) + \overline{Z_1}\otimes Z_2 - (\overline{Z_2}H^T)\otimes Z_2\right] S_{21}(Z_1^T\otimes Q_2^H) - 2(\overline{Z_2}\otimes Z_1)T_1M,$$
$$L_4 = 2\left[\overline{Z_1}\otimes(Z_1H) + \overline{Z_1}\otimes Z_2 - (\overline{Z_2}H^T)\otimes Z_2\right] S_{22}(Z_1^T\otimes Q_2^H) - 2(\overline{Z_2}\otimes Z_1)T_1N.$$

Similarly, we obtain the mixed and componentwise condition numbers for the sign functions of regular matrix pairs.

Theorem 6.2. *Let $\|(E,F)\|_F$ be sufficiently small. Using the notations above, we have*

$$m_{\mathrm{sign}_L}((A,B)) = \frac{\left\| |[L_1,L_2]| \begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\end{bmatrix}\right\|_\infty}{\|\mathrm{vec}(\mathrm{sign}_L((A,B)))\|_\infty}, \quad c_{\mathrm{sign}_L}((A,B)) = \left\| \frac{|[L_1,L_2]| \begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\end{bmatrix}}{\mathrm{vec}(\mathrm{sign}_L((A,B)))}\right\|_\infty, \tag{6.8}$$

$$m_{\mathrm{sign}_R}((A,B)) = \frac{\left\| |[L_3,L_4]| \begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\end{bmatrix}\right\|_\infty}{\|\mathrm{vec}(\mathrm{sign}_R((A,B)))\|_\infty}, \quad c_{\mathrm{sign}_R}((A,B)) = \left\| \frac{|[L_3,L_4]| \begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\end{bmatrix}}{\mathrm{vec}(\mathrm{sign}_R((A,B)))}\right\|_\infty. \tag{6.9}$$

7. Statistical condition estimation

Since an estimate of the condition number that is correct to within a factor 10 is usually acceptable [13], we can often tolerate errors in the estimate up to a factor of 10 or a factor that is a little bit bigger. In fact, we are interested in the magnitude of an error bound, not a precise value. Statistical condition estimate (SCE) proposed by Kenney and Laub [17] is an efficient method to estimate the condition numbers [8, 19–21].

7.1. Brief review of SCE

Suppose that $f\colon \mathbb{R}^p \to \mathbb{R}$ is at least twice continuously differentiable. Denote the gradient of f at $x \in \mathbb{R}^p$ by $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_p}\right)^T$, then using the first order of Taylor expansion of f at x along d, we have

$$f(x+\delta d) = f(x) + \delta \nabla f(x)^T d + \mathscr{O}(\delta^2),$$

where $\delta \in \mathbb{R}$ is small and $d \in \mathbb{R}^p$ has unit 2-norm. It shows that the norm of the gradient can measure the local sensitivity of f approximately. If d is selected uniformly

and randomly from the unit sphere S_{p-1} in $\mathbb{R}^p$, which is denoted by $d \in U(S_{p-1})$. Then from [17], the expected value of the condition estimator $\nu = \frac{\left|(\nabla f(x))^T d\right|}{\omega_p}$ satisfies that

$$\boldsymbol{E}(\nu) = \|\nabla f(x)\|_2 ,$$

where ω_p is the Wallis factor

$$\omega_p = \begin{cases} 1, & \text{for } p \equiv 1, \\ \frac{2}{\pi}, & \text{for } p \equiv 2, \\ \frac{1\cdot 3\cdot 5\cdots(p-2)}{2\cdot 4\cdot 6\cdots(p-1)}, & \text{for odd } p > 2, \\ \frac{2}{\pi}\frac{2\cdot 4\cdot 6\cdots(p-2)}{1\cdot 3\cdot 5\cdots(p-1)}, & \text{for even } p > 2, \end{cases}$$

and for $\gamma > 1$ we have

$$\text{Prob}\left(\frac{\|\nabla f(x)\|_2}{\gamma} \le \nu \le \gamma \|\nabla f(x)\|_2\right) \ge 1 - \frac{2}{\pi\gamma} + \mathscr{O}\left(\frac{1}{\gamma^2}\right).$$

Therefore, we can use the absolute value

$$\left|\frac{f(x+\delta d) - f(x)}{\delta \omega_p}\right|$$

as a first order condition estimator, which can estimate $\|\nabla f(x)\|_2$ with high probability for the function f at x. In practice, the Wallis factor can be approximated accurately [17] by

$$\omega_p \approx \sqrt{\frac{2}{\pi\left(p - \frac{1}{2}\right)}}. \tag{7.1}$$

In situations where we need more reliability, we use more function evaluations to get different values $\nu^{(1)}, \nu^{(2)}, \ldots, \nu^{(m)}$ corresponding to independently randomly generated vectors $d^{(1)}, d^{(2)}, \ldots, d^{(m)} \in U(S_{p-1})$ and then take the average

$$\nu(m) \equiv \frac{\nu^{(1)} + \nu^{(2)} + \cdots + \nu^{(m)}}{m}.$$

(Here we use superscripts in parentheses to distinguish between different vectors)

This is the so called "averaged small-sample statistical method" and we can show that [17] for $\gamma > 1$

$$\text{Prob}\left(\frac{\|\nabla f(x)\|_2}{\gamma} \le \nu(m) \le \gamma \|\nabla f(x)\|_2\right) \ge 1 - \frac{1}{m!}\left(\frac{2m}{\pi\gamma}\right)^m + \mathscr{O}\left(\frac{1}{\gamma^{m+1}}\right).$$

Thus $\nu(m)$ is a mth-order condition estimator.

Compared with this method, the subspace statistical method can give sharper estimates. Firstly, select k vectors from $U(S_{p-1})$ and find mutually orthonormal vectors $d_1, d_2, \ldots, d_k$ by using a Gram-Schmidt procedure or a QR decomposition [11]. Thus the norm of the projection of $\nabla f(x)$ onto the span of $d_1, d_2, \ldots, d_k$ is $(\left|\nabla f(x)^T d_1\right|^2 + \left|\nabla f(x)^T d_2\right|^2 + \cdots + \left|\nabla f(x)^T d_k\right|^2)^{1/2}$. From [17], we know that

$$\boldsymbol{E}\left(\frac{\omega_k}{\omega_p}\sqrt{\left|\nabla f(x)^T d_1\right|^2 + \left|\nabla f(x)^T d_2\right|^2 + \cdots + \left|\nabla f(x)^T d_k\right|^2}\right) = \left\|\nabla f(x)\right\|_2.$$

Therefore, we can define the subspace condition estimator as

$$\xi(k) = \frac{\omega_k}{\omega_p}\sqrt{\left|\nabla f(x)^T d_1\right|^2 + \left|\nabla f(x)^T d_2\right|^2 + \cdots + \left|\nabla f(x)^T d_k\right|^2}.$$

As shown in [17], these condition estimators give better results than the averaged statistical estimators and are analytically very tractable.

From [17, Theorem 3.3], we find

$$\begin{aligned}
&\mathrm{Prob}\left(\frac{\left\|\nabla f(x)\right\|_2}{\gamma} \le \xi(2) \le \gamma\left\|\nabla f(x)\right\|_2\right) \approx 1 - \frac{\pi}{4\gamma^2},\\
&\mathrm{Prob}\left(\frac{\left\|\nabla f(x)\right\|_2}{\gamma} \le \xi(3) \le \gamma\left\|\nabla f(x)\right\|_2\right) \approx 1 - \frac{32}{3\pi^2\gamma^3},\\
&\mathrm{Prob}\left(\frac{\left\|\nabla f(x)\right\|_2}{\gamma} \le \xi(4) \le \gamma\left\|\nabla f(x)\right\|_2\right) \approx 1 - \frac{81\pi^2}{512\gamma^4}.
\end{aligned}$$

These estimates are generally very accurate for $\gamma \ge 10$.

7.2. SCE for the spectral projection

Take $\Delta A = \delta\widehat{A}$, where $\left\|\widehat{A}\right\|_F = 1$ and δ is sufficiently small. Denote

$$S^{-1}\widehat{A}S = \begin{bmatrix} \Delta A_{11} & \Delta A_{12} \\ \Delta A_{21} & \Delta A_{22} \end{bmatrix}, \quad \Delta A_{11} \in \mathbb{C}^{m\times m}.$$

Obviously we can see that $\left\|\Delta A_{ij}\right\|_F \le \left\|S^{-1}\widehat{A}S\right\|_F \le \left\|S^{-1}\right\|_F \left\|S\right\|_F$, which implies that $\left\|\Delta A_{ij}\right\|_F = \mathscr{O}(1)$. From [31], we know that

$$\Delta P = S\begin{bmatrix} EF(I_m - EF)^{-1} & -E - EF(I_m - EF)^{-1}E \\ F + FEF(I_m - EF)^{-1} & -F(I_m - EF)^{-1}E \end{bmatrix} S^{-1},$$

where E and F are the solutions to the two Sylvester equations:

$$\begin{aligned}
A_{11}E - EA_{22} &= -\delta\Delta A_{12} + \delta E\Delta A_{22} - \delta\Delta A_{11}E + \delta E\Delta A_{21}E,\\
A_{22}F - FA_{11} &= -\delta\Delta A_{21} + \delta F\Delta A_{11} - \delta\Delta A_{22}F + \delta F\Delta A_{12}F,
\end{aligned} \tag{7.2}$$

and satisfy

$$\|E\|_F = \mathscr{O}(\delta), \quad \|F\|_F = \mathscr{O}(\delta), \text{ as } \delta \to 0.$$

So ΔP has the first order perturbation expansion

$$\Delta P \approx \delta S \begin{bmatrix} \mathbf{0} & -\widehat{E} \\ \widehat{F} & \mathbf{0} \end{bmatrix} S^{-1},$$

where $\widehat{E}$ and $\widehat{F}$ can be derived by solving the two Sylvester equations:

$$\begin{aligned} A_{11}\widehat{E} - \widehat{E}A_{22} &= -\Delta A_{12}, \\ A_{22}\widehat{F} - \widehat{F}A_{11} &= -\Delta A_{21}. \end{aligned} \tag{7.3}$$

In practice, let $a^{(1)}, a^{(2)}, \ldots, a^{(k)}$ be orthonormal vectors of length $p = n^2$ such that their span is uniformly and randomly generated from the space of all k-dimensional subspaces of $\mathbb{R}^n$. Form the matrices $\widehat{A}_i = \text{unvec}(a^{(i)})$ and we then define the subspace condition estimators for each entry of P by

$$\xi_{ij}(k) = \frac{\omega_k}{\omega_p}\sqrt{\left(K_{ij}^{(1)}\right)^2 + \left(K_{ij}^{(2)}\right)^2 + \cdots + \left(K_{ij}^{(k)}\right)^2},$$

where $K^{(l)}$ is given by

$$K^{(l)} = S \begin{bmatrix} \mathbf{0} & -\widehat{E}^{(l)} \\ \widehat{F}^{(l)} & \mathbf{0} \end{bmatrix} S^{-1}.$$

Based on the above analysis, we now present the SCE algorithm for the spectral projection.

Algorithm 7.1 (Subspace condition estimation for the spectral projection)**.**

1. *Generate matrices $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_k$ with entries in $\mathbb{N}(0,1)$. Use a QR factorization for the matrix*

$$\begin{bmatrix} \text{vec}(\mathscr{A}_1) & \text{vec}(\mathscr{A}_2) & \cdots & \text{vec}(\mathscr{A}_k) \end{bmatrix}$$

and form an orthonormal matrix $[q_1, q_2, \ldots, q_k]$. Each q_i can be converted into the desired matrices $\mathscr{A}_i$ with the unvec *operation.*

2. *Let $p = n^2$. Approximate ω_p and ω_k using* (7.1)*.*

3. *For $l = 1, 2, \ldots, k$, solve* (7.3) *to get $\widehat{E}^{(l)}$, $\widehat{F}^{(l)}$ and form $K^{(l)}$. Calculate the absolute condition matrix*

$$\xi(k) = \frac{\omega_k}{\omega_p}\sqrt{\left|K^{(1)}\right|^2 + \left|K^{(2)}\right|^2 + \cdots + \left|K^{(k)}\right|^2},$$

where the square root and power operation are performed componentwise.

In Algorithm 7.1, we generate matrices $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_k$ in $\mathbb{N}(0,1)$ instead of $U(S_{p-1})$, which is based on the fact that the vector $z = z/\|z\|_2 \in U(S_{p-1})$ if $z \in \mathbb{N}(0,1)$. Specifically, to get a basis for a subspace of dimension m that is selected uniformly and randomly from the set of all subspaces of dimension m in $\mathbb{R}^n$, it is sufficient to generate m independent vectors uniformly and randomly on S_{n-1} and then find an orthonormal basis for these vectors [17]. Now we can define the normwise, mixed and componentwise condition number as follows:

(7.4)
$$n_{\mathrm{sp}}^{\mathrm{SCE,(k)}} = \|\xi(k)\|_F / \|P\|_F, \quad m_{\mathrm{sp}}^{\mathrm{SCE,(k)}} = \|\xi(k)\|_{\max} / \|P\|_{\max}, \quad c_{\mathrm{sp}}^{\mathrm{SCE,(k)}} = \|\xi(k)/P\|_{\max}.$$

7.3. SCE for the generalized spectral projections

Take $\Delta A = \delta\widehat{A}$, $\Delta B = \delta\widehat{B}$, where $\left\|\left[\widehat{A}, \widehat{B}\right]\right\|_F = 1$. Denote

$$Q^{-1}\widehat{A}S = \begin{bmatrix} \Delta A_{11} & \Delta A_{12} \\ \Delta A_{21} & \Delta A_{22} \end{bmatrix}, \quad Q^{-1}\widehat{B}S = \begin{bmatrix} \Delta B_{11} & \Delta B_{12} \\ \Delta B_{21} & \Delta B_{22} \end{bmatrix}, \quad \Delta A_{11}, \Delta B_{11} \in \mathbb{C}^{m\times m}.$$

Using the similar technique, we obtain

$$\Delta P_r \approx \delta S \begin{bmatrix} \mathbf{0} & -\widehat{G} \\ \widehat{H} & \mathbf{0} \end{bmatrix} S^{-1}, \quad \Delta P_l \approx \delta Q \begin{bmatrix} \mathbf{0} & -\widehat{J} \\ \widehat{L} & \mathbf{0} \end{bmatrix} Q^{-1},$$

where $\widehat{G}$, $\widehat{H}$, $\widehat{J}$ and $\widehat{L}$ satisfy that

$$\begin{aligned} A_{11}\widehat{G} - \widehat{J}A_{22} &= -\Delta A_{12}, \qquad B_{11}\widehat{G} - \widehat{J}B_{22} = -\Delta B_{12}, \\ A_{22}\widehat{H} - \widehat{L}A_{11} &= -\Delta A_{21}, \qquad B_{22}\widehat{H} - \widehat{L}B_{11} = -\Delta B_{21}. \end{aligned} \tag{7.5}$$

The statistical condition estimation for the generalized spectral projections can be derived similarly. Hence we can obtain the condition numbers.

7.4. SCE for the matrix sign function

The analysis in this subsection is based on [2, Theorem 3.2]. Taking $\Delta A = \delta\widehat{A}$ with $\left\|\widehat{A}\right\|_F = 1$ and δ sufficiently small, we partition $\widehat{A}$ conformally as

$$U^H\widehat{A}U = \begin{bmatrix} \Delta A_{11} & \Delta A_{12} \\ \Delta A_{21} & \Delta A_{22} \end{bmatrix}.$$

Then $\Delta_{\mathrm{sign}}(A)$ has the first order expansion

$$\Delta_{\mathrm{sign}}(A) \approx \delta U \begin{bmatrix} Y\widetilde{E}_{21} & 2\widetilde{E}_{12} - \frac{Y\widetilde{E}_{21}Y}{2} \\ 2\widetilde{E}_{21} & -\widetilde{E}_{21}Y \end{bmatrix} U^H,$$

where $\widetilde{E}_{21}$ and $\widetilde{E}_{12}$ satisfy

$$\begin{aligned} &A_{22}\widetilde{E}_{21} - \widetilde{E}_{21}A_{11} = \Delta A_{21}, \\ &\widetilde{E}_{12}A_{22} - A_{11}\widetilde{E}_{12} = \Delta A_{12} - \frac{Y\Delta A_{22}}{2} + \frac{\Delta A_{11}Y}{2} - \frac{Y\Delta_{21}Y}{4}. \end{aligned} \tag{7.6}$$

The condition numbers of the matrix sign function can be derived similarly.

8. Numerical examples

All computations are performed in MATLAB 7.12, with the usual double precision and the floating point accuracy is 2.22×10^{-16}. We now use three numerical examples to illustrate our results of Sections 3–7.

Example 8.1. Consider the matrix

$$A = U \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix} U^T,$$

where the orthogonal matrix U is chosen to be the unitary factor of QR factorization of a matrix with entries chosen to randomly uniformly distributed in the interval $[0, 1]$. Let

$$A_{11} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} 10^l & 1 & 0 & 0 & 0 \\ -1 & 10^l & 1 & 0 & 0 \\ 0 & -1 & 10^l & 1 & 0 \end{bmatrix},$$

and

$$A_{22} = \begin{bmatrix} 1-10^{-k} & 1 & 0 & 0 & 0 \\ 0 & 1-10^{-k} & 1 & 0 & 0 \\ 0 & 0 & 1-10^{-k} & 1 & 0 \\ 0 & 0 & 0 & 1-10^{-k} & 1 \\ 0 & 0 & 0 & 0 & 1-10^{-k} \end{bmatrix}.$$

This example is originally arising from [30]. The target spectral projection P of A is corresponding to the triple eigenvalue 1. The computed condition numbers are listed in Tables 8.1–8.6.

$l\backslash k$	-1	0	1	2
0	1.2097e-001	1.9917e+002	1.0849e+014	2.0474e+027
1	3.1520e-001	1.7467e+004	1.9624e+014	7.7607e+026
2	2.0708e+001	1.7980e+006	9.7305e+016	2.0001e+027
3	2.0590e+003	1.8039e+008	1.1087e+019	9.6443e+029

Table 8.1: The absolute condition number $c_{\text{abs}}(P)$ in [30, Eqn. (2.23)].

$l\backslash k$	-1	0	1	2
0	1.4154e+000	1.7123e+002	1.0585e+008	1.9279e+015
1	3.4275e+000	7.6358e+003	1.7543e+009	2.9487e+015
2	2.0705e+002	7.0351e+005	6.7218e+011	1.7327e+017
3	2.0555e+004	6.9934e+007	6.4911e+013	6.6815e+019

Table 8.2: The relative condition number $c_{\rm rel}(P)$ in [30, Eqn. (2.23)].

$l\backslash k$	-1	0	1	2
0	9.9394e-001	1.2841e+002	8.0650e+007	9.1299e+014
1	3.1290e+000	5.5280e+003	1.0625e+009	1.5370e+015
2	2.8411e+002	4.4438e+005	4.0703e+011	1.2084e+017
3	3.3187e+004	3.4991e+007	3.4447e+013	3.7777e+019

Table 8.3: The mixed condition number $m_P(A)$ in (3.4).

$l\backslash k$	-1	0	1	2
0	3.0528e+001	3.2737e+003	2.2567e+008	1.1403e+015
1	6.1359e+001	3.9940e+005	1.2467e+009	1.8935e+015
2	8.1988e+003	1.4488e+007	4.5762e+011	1.2889e+017
3	6.4062e+005	3.8123e+010	4.0801e+013	3.8488e+019

Table 8.4: The upper bound of mixed condition number $m_P(A)$ in (3.6).

$l\backslash k$	-1	0	1	2
0	4.6966e+000	5.4408e+002	1.6475e+009	1.2134e+016
1	1.2599e+001	1.5111e+004	3.4706e+009	5.0122e+016
2	7.1813e+002	1.1169e+006	4.2969e+011	3.6637e+017
3	9.3672e+004	8.9718e+007	3.8798e+013	3.7950e+019

Table 8.5: The componentwise condition number $c_P(A)$ in (3.5).

$l\backslash k$	-1	0	1	2
0	1.5856e+002	4.8141e+004	2.2962e+009	1.2381e+016
1	4.7307e+002	3.0632e+006	5.5438e+009	5.9997e+016
2	2.6608e+004	1.0709e+008	5.0128e+011	3.8025e+017
3	2.4070e+006	1.5770e+011	4.4983e+013	3.8940e+019

Table 8.6: The upper bound of componentwise condition number $c_P(A)$ in (3.7).

For the cases of $k = 0, 1, 2$, the mixed condition numbers $m_P(A)$ are much smaller than the absolute condition numbers $c_{\rm abs}(P)$ in [30, Eqn. (2.23)]. But for the case of $k = -1$, the mixed condition numbers $m_P(A)$ are not smaller. The upper bounds of the mixed and

componentwise condition numbers are a little bit bigger with no surprise.

The corresponding SCE results for this example are listed in the next three tables.

$l\backslash k$	-1	0	1	2
0	1.2945e+000	4.1552e+001	6.2213e+007	7.2531e+014
1	7.9716e-001	9.6332e+002	3.5800e+008	5.6281e+014
2	1.2622e+001	2.7743e+004	5.0356e+009	1.4264e+015
3	9.0830e+001	7.5957e+004	6.4316e+010	2.4724e+017

Table 8.7: Normwise condition number from SCE by $n_{\rm sp}^{\rm SCE,(3)}$ in (7.4).

$l\backslash k$	-1	0	1	2
0	1.0067e+000	5.7192e+001	5.7191e+007	9.9939e+014
1	8.3759e-001	1.0148e+003	3.1131e+008	6.0435e+014
2	1.4625e+001	2.9274e+004	6.3020e+009	1.4904e+015
3	9.9826e+001	6.5180e+004	5.0960e+010	2.3166e+017

Table 8.8: Mixed condition number from SCE by $m_{\rm sp}^{\rm SCE,(3)}$ in (7.4).

$l\backslash k$	-1	0	1	2
0	4.4925e+000	1.3145e+002	4.7887e+008	1.2694e+015
1	2.6200e+000	7.1676e+003	5.0040e+009	5.6004e+015
2	3.0031e+001	1.5961e+005	4.0064e+010	4.0442e+015
3	2.3232e+002	2.5736e+005	2.5850e+011	1.3577e+018

Table 8.9: Componentwise condition number from SCE by $c_{\rm sp}^{\rm SCE,(3)}$ in (7.4).

It is easy to see that in most cases the condition numbers devised from SCE are sufficient. To see the performance of SCE more precisely, we plot the following three ratios in Figure 8.1, for the case of $l = 0$, $k = 2$,

$$\text{ratio}_{\rm n} := \frac{n_{\rm sp}^{\rm SCE,(3)}}{c_{\rm rel}(P)}, \quad \text{ratio}_{\rm m} := \frac{m_{\rm sp}^{\rm SCE,(3)}}{m_P(A)}, \quad \text{ratio}_{\rm c} := \frac{c_{\rm sp}^{\rm SCE,(3)}}{c_{\rm sign}(A)}.$$

We test 1000 examples and the means of the three ratios are 0.5538, 1.0313, 8.3572, respectively. So it is sufficient in practice.

Example 8.2. [30] Consider the real regular matrix pair (A, B) with

$$A = V \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix} U^T, \quad B = V \begin{bmatrix} B_{11} & B_{12} \\ \mathbf{0} & B_{22} \end{bmatrix} U^T,$$

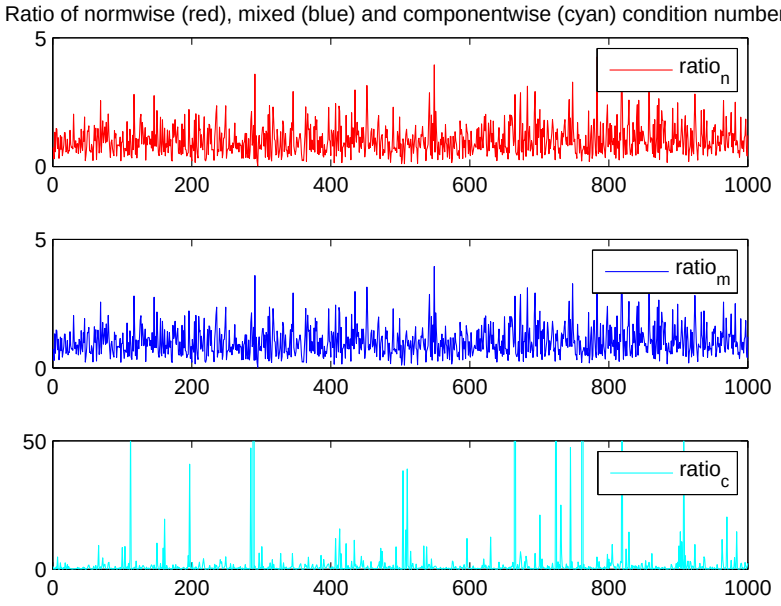

Figure 8.1: SCE results compared with the exact condition numbers of the spectral projection.

where $V = I_6 - \frac{1}{3}vv^T$, $v = [1,1,1,1,1,1]^T$ and $U = I_6 - \frac{1}{3}uu^T$, $u = [1,-1,1,-1,1,-1]^T$, with

$$A_{11} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 10^{-k} \end{bmatrix}, \quad B_{11} = \begin{bmatrix} -4\times 10^{-k} & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2\times 10^{-l} \end{bmatrix}$$

and

$$A_{22} = \begin{bmatrix} 10^k & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -10^l \end{bmatrix}, \quad B_{22} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -10^k \end{bmatrix},$$

$$A_{12} = \begin{bmatrix} 2\times 10^k & 1 & 0 \\ 0 & -1 & 3 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_{12} = \begin{bmatrix} 7 & 2 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & -3\times 10^k \end{bmatrix}.$$

Solving the generalized Sylvester equation (4.2), we obtain (M, N) and the spectral projections P_r and P_l of (A, B) corresponding to the eigenvalue $-\frac{1}{4}\times 10^k$, $-\frac{5}{3}$, and $-\frac{1}{2}\times$

10^{l-k}. The computational condition numbers are listed in Tables 8.10–8.19, respectively.

$l\backslash k$	0	1	2	3	4
0	2.0245e+000	9.0994e+001	9.1815e+003	9.1677e+005	9.1654e+007
1	7.4983e+001	8.8198e+002	4.7726e+004	4.7146e+006	4.7141e+008
2	7.4540e+003	7.4685e+004	8.8192e+005	4.7726e+007	4.7146e+009
3	7.4536e+005	7.4537e+006	7.4685e+007	8.8192e+008	4.7726e+010
4	7.4536e+007	7.4536e+008	7.4537e+009	7.4685e+010	8.8192e+011

Table 8.10: The absolute condition number $c_{\text{abs}}(P_r)$ in [30, Eqn. (3.37)].

$l\backslash k$	0	1	2	3	4
0	9.1146e+000	9.8595e+001	6.8553e+003	6.6838e+005	6.6684e+007
1	1.6910e+001	1.0481e+002	3.0731e+003	2.3800e+005	2.3918e+007
2	1.5209e+002	7.5144e+002	9.0821e+003	2.9075e+005	2.2907e+007
3	1.5520e+003	7.4193e+003	6.9310e+004	8.9087e+005	2.8908e+007
4	1.5552e+004	7.4217e+004	6.7502e+005	6.8919e+006	8.8909e+007

Table 8.11: The mixed condition number $m_{P_r}(A)$ in (4.13).

$l\backslash k$	0	1	2	3	4
0	1.9832e+001	1.4795e+002	7.2723e+003	6.7243e+005	6.6724e+007
1	2.6190e+001	1.7445e+002	3.6828e+003	2.5194e+005	2.4058e+007
2	1.8591e+002	1.0653e+003	1.1710e+004	3.1657e+005	2.3314e+007
3	1.7859e+003	9.7696e+003	9.3243e+004	1.1171e+006	3.1165e+007
4	1.7786e+004	9.6576e+004	9.0232e+005	9.1353e+006	1.1117e+008

Table 8.12: The upper bound of mixed condition number $m_{P_r}(A)$ in (4.15).

$l\backslash k$	0	1	2	3	4
0	1.9838e+002	2.6748e+003	3.2182e+005	2.9302e+008	2.9017e+011
1	9.8888e+001	2.2926e+002	6.8042e+003	6.6746e+005	6.6674e+007
2	1.8690e+002	8.3626e+002	9.4673e+003	2.9638e+005	2.4784e+007
3	1.5859e+003	7.5461e+003	7.0786e+004	8.9452e+005	2.8961e+007
4	1.5586e+004	7.4352e+004	6.7928e+005	6.9106e+006	8.8945e+007

Table 8.13: The componentwise condition number $c_{P_r}(A)$ in (4.14).

$l\backslash k$	0	1	2	3	4
0	9.0216e+002	5.4314e+003	3.5402e+005	2.9596e+008	2.9046e+011
1	2.0424e+002	5.1983e+002	1.0130e+004	6.9978e+005	6.6996e+007
2	2.2343e+002	1.2857e+003	1.3391e+004	3.5307e+005	2.6024e+007
3	1.8197e+003	1.0053e+004	9.7111e+004	1.1330e+006	3.1514e+007
4	1.7819e+004	9.6878e+004	9.1331e+005	9.1825e+006	1.1133e+008

Table 8.14: The upper bound of componentwise condition number $c_{P_r}(A)$ in (4.17).

$l\backslash k$	0	1	2	3	4
0	4.3943e+000	2.9272e+001	2.8947e+002	2.8767e+003	2.8747e+004
1	9.7310e+000	4.4111e+001	2.8947e+002	2.8767e+003	2.8747e+004
2	9.1954e+001	2.0183e+002	5.5980e+002	2.8767e+003	2.8747e+004
3	9.1927e+002	9.1929e+002	4.2778e+003	5.8048e+003	2.8747e+004
4	9.1927e+003	9.1927e+003	1.9852e+004	5.2423e+004	5.8271e+004

Table 8.15: The absolute condition number $c_{\text{abs}}(P_l)$ in [30, Eqn. (3.37)].

$l\backslash k$	0	1	2	3	4
0	5.0748e+000	6.1820e+001	3.3112e+003	2.9785e+005	2.9440e+007
1	1.1563e+001	1.2273e+002	3.5132e+003	2.9975e+005	2.9459e+007
2	6.4579e+001	5.5135e+002	8.6434e+003	3.2251e+005	2.9686e+007
3	6.0613e+002	2.9053e+003	6.3764e+004	8.1741e+005	3.1984e+007
4	6.0201e+003	3.0036e+004	3.7667e+005	7.5064e+006	8.1251e+007

Table 8.16: The mixed condition number $m_{P_l}(A)$ in (4.13).

$l\backslash k$	0	1	2	3	4
0	1.7665e+001	9.1293e+001	3.4443e+003	2.9915e+005	2.9453e+007
1	2.9913e+001	2.4133e+002	4.0597e+003	3.0243e+005	2.9486e+007
2	1.1401e+002	1.1092e+003	1.5053e+004	3.6619e+005	2.9823e+007
3	9.6862e+002	7.0000e+003	1.0663e+005	1.4091e+006	3.6214e+007
4	9.5128e+003	6.2640e+004	7.3348e+005	1.1095e+007	1.3994e+008

Table 8.17: The upper bound of mixed condition number $m_{P_l}(A)$ in (4.16).

$l\backslash k$	0	1	2	3	4
0	2.4394e+001	3.0123e+002	2.1942e+004	2.3950e+006	2.4394e+008
1	8.9995e+001	7.1714e+002	2.7404e+004	2.3598e+006	2.4358e+008
2	1.1597e+003	4.4183e+003	8.5667e+004	2.7769e+006	2.4007e+008
3	1.2250e+004	2.6681e+004	5.1276e+005	9.1441e+006	2.7822e+008
4	1.2323e+005	2.4874e+005	3.1769e+006	6.2703e+007	9.2143e+008

Table 8.18: The componentwise condition number $c_{P_l}(A)$ in (4.14).

$l\backslash k$	0	1	2	3	4
0	6.8742e+001	4.7580e+002	2.4061e+004	2.4362e+006	2.4436e+008
1	2.2992e+002	1.5524e+003	3.4133e+004	2.5473e+006	2.4548e+008
2	1.7387e+003	8.6260e+003	1.2997e+005	3.6548e+006	2.5680e+008
3	1.6964e+004	5.4188e+004	9.2589e+005	1.3991e+007	3.6877e+008
4	1.6940e+005	4.7888e+005	6.9002e+006	1.0430e+008	1.4119e+009

Table 8.19: The upper bound of componentwise condition number $c_{P_l}(A)$ in (4.18).

The mixed condition numbers $m_{P_r}(A)$ are smaller than the absolute condition numbers $c_{\text{abs}}(P_r)$ in [30, Eqn. (3.37)] except for $l = 0$, $k = 0$, and the mixed condition numbers $m_{P_l}(A)$ are larger than the absolute condition numbers $c_{\text{abs}}(P_l)$ in [30, Eqn. (3.37)]. But

the mixed condition numbers $m_{P_r}(A) \approx m_{P_l}(A)$. The upper bounds of the mixed and componentwise condition numbers increase a little bit.

The corresponding SCE results can be obtained similarly.

Example 8.3. [29] For a nonzero real scalar x, let

$$A = UX \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X^{-1}U^T, \quad X = \begin{bmatrix} 1 & -x/2 \\ 0 & 1 \end{bmatrix},$$

where the orthogonal matrix U is chosen to be the unitary factor of QR factorization of a matrix with entries chosen to be randomly uniformly distributed in the interval $[0, 1]$. It follows from [29] that

$$\operatorname{sign}(A) = UX \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X^{-1}U^T.$$

The mixed and componentwise condition numbers and the corresponding SCE (3 samples) for $\operatorname{sign}(A)$ from one random U are listed in Tables 8.20 and 8.21, respectively.

x	[29, Eqn. (3.27)]	$m_{\text{sign}}(A)$	upper bound (5.4)	$c_{\text{sign}}(A)$	upper bound (5.5)	$m_{\text{sign}}^{\text{SCE}}(A)$	$c_{\text{sign}}^{\text{SCE}}(A)$
−1.0	1.5000	1.2504	4.4971	1.4992	2.2477	2.8251	6.8673
−0.9	1.4050	1.1933	4.7630	1.6134	2.6415	1.1223	3.7065
−0.8	1.3200	1.4497	3.0961	1.4497	1.6706	0.8062	3.6463
−0.7	1.2450	1.2427	1.2427	2.1159	2.1159	0.2645	2.2803
−0.6	1.1800	1.1223	4.9538	1.7553	3.3869	0.1662	0.9946
−0.5	1.1250	1.3218	3.0684	1.3564	1.8737	3.6262	26.4495
−0.4	1.0800	1.3595	2.7345	1.3595	1.7089	0.4402	2.1016
−0.3	1.0450	1.5496	1.6949	1.6533	1.8083	0.3255	4.1436
−0.2	1.0200	1.0971	4.4427	1.8058	3.8006	0.7538	7.0907
−0.1	1.0050	1.0951	4.2536	1.8098	3.8391	1.3298	4.9779

Table 8.20: The mixed and componentwise condition numbers.

x	[29, Eqn. (3.27)]	$m_{\text{sign}}(A)$	upper bound (5.4)	$c_{\text{sign}}(A)$	upper bound (5.5)	$m_{\text{sign}}^{\text{SCE}}(A)$	$c_{\text{sign}}^{\text{SCE}}(A)$
0.1	1.0050	1.3371	2.3566	1.3371	1.8045	2.0661	4.6214
0.2	1.0200	0.9237	0.9523	1.8797	1.9379	1.8488	3.5279
0.3	1.0450	0.8373	0.8478	1.9504	1.9749	0.5134	7.0444
0.4	1.0800	1.0872	5.1590	1.8256	4.0040	0.9117	1.2996
0.5	1.1250	1.6501	1.9986	1.6501	1.8063	0.2159	0.8094
0.6	1.1800	1.5759	2.3573	1.5759	1.7589	2.0993	22.9109
0.7	1.2450	1.1485	1.1485	2.1633	2.1633	0.2634	0.2634
0.8	1.3200	1.0119	14.9809	1.9763	10.0598	0.5284	0.6544
0.9	1.4050	1.5740	1.5740	2.2307	2.2307	0.8489	6.8765
1.0	1.5000	1.8976	2.2481	1.8976	1.9474	0.5551	2.9394

Table 8.21: The mixed and componentwise condition numbers.

If $x = 0.1 \sim 1.0$, the mixed condition numbers are smaller than the normwise condition numbers in [29, Eqn. (3.27)]. But if $x = -1.0 \sim -0.1$, some mixed condition numbers are larger. The upper bounds of the mixed and componentwise condition numbers increase a little bit, if x is positive. Also we can see that the statistical condition estimate is adequate because of its efficiency in computing.

To more clearly show the overestimation of the statistical condition estimation, we generate random matrix U 1000 times, take $x = 0.1$ and 3 samples and plot the ratio in figure. In Figure 8.2, we denote

$$\text{ratio}_{\text{m}} := \frac{m_{\text{sign}}^{\text{SCE},(3)}(A)}{m_{\text{sign}}(A)}, \quad \text{ratio}_{\text{c}} := \frac{c_{\text{sign}}^{\text{SCE},(3)}(A)}{c_{\text{sign}}(A)}.$$

The averages of the two ratios are 1.6408 and 16.9633, respectively. As shown in Figure 8.2, the statistical condition estimations are within a factor of 100 to the true condition numbers except several exceptional cases. It is computationally efficient and adequate in practice. We suggest the statistical condition estimations for the applications.

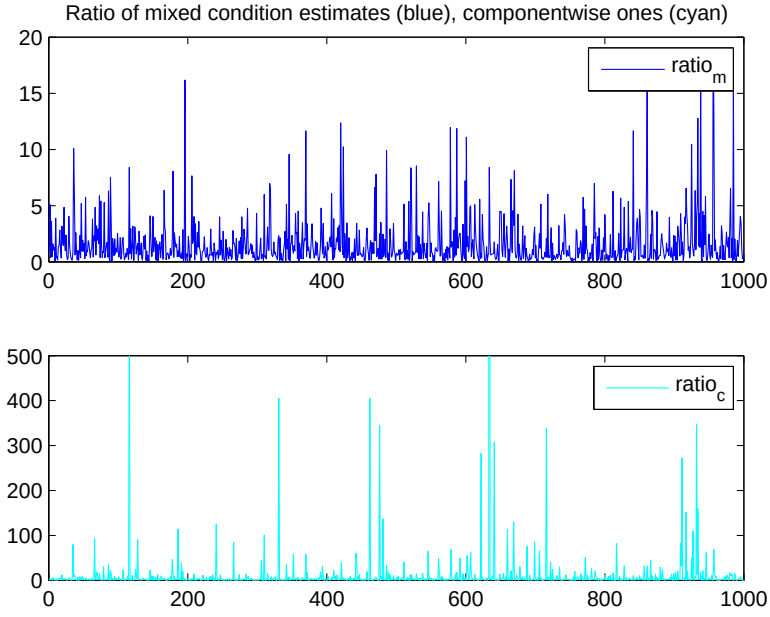

Figure 8.2: SCE results compared with the exact condition numbers of the matrix sign function

9. Conclusions

The perturbation analysis for the spectral projections, generalized spectral projections and the matrix sign functions are derived in this paper. The explicit expressions for the mixed and componentwise condition numbers are also presented. From the numerical examples,

we can see that the statistical condition estimation is enough in practice for estimating the actual conditioning.

Acknowledgments

We are very grateful to the editor and two anonymous referees for their detailed comments and suggestions.

References

[1] Z. Bai and J. Demmel, *Using the matrix sign function to compute invariant subspaces*, SIAM J. Matrix Anal. Appl. **19** (1998), no. 1, 205–225. http://dx.doi.org/10.1137/s0895479896297719

[2] R. Byers, C. He and V. Mehrmann, *The matrix sign function method and the computation of invariant subspaces*, SIAM J. Matrix Anal. Appl. **18** (1997), no. 3, 615–632. http://dx.doi.org/10.1137/s0895479894277454

[3] X. Chen, W. Li and W.-K. Ching, *Perturbation analysis for the sign functions of regular matrix pairs*, Numer. Linear Algebra Appl. **18** (2011), no. 2, 189–203. http://dx.doi.org/10.1002/nla.759

[4] D. Chu, L. Lin, R. C. E. Tan and Y. Wei, *Condition numbers and perturbation analysis for the Tikhonov regularization of discrete ill-posed problems*, Numer. Linear Algebra Appl. **18** (2011), no. 1, 87–103. http://dx.doi.org/10.1002/nla.702

[5] F. Cucker, H. Diao and Y. Wei, *On mixed and componentwise condition numbers for Moore-Penrose inverse and linear least squares problems*, Math. Comp. **76** (2007), no. 258, 947–963.

[6] J. W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math. **51** (1987), no. 3, 251–289. http://dx.doi.org/10.1007/bf01400115

[7] J. W. Demmel and B. Kågström, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl. **88/89** (1987), 139–186. http://dx.doi.org/10.1016/0024-3795(87)90108-x

[8] H. Diao, H. Xiang and Y. Wei, *Mixed, componentwise condition numbers and small sample statistical condition estimation of Sylvester equations*, Numer. Linear Algebra Appl. **19** (2012), no. 4, 639–654. http://dx.doi.org/10.1002/nla.790

[9] S. K. Godunov and M. Sadkane, *Computation of pseudospectra via spectral projectors*, Linear Algebra Appl. **279** (1998), no. 1-3, 163–175. http://dx.doi.org/10.1016/s0024-3795(98)00011-1

[10] I. Gohberg and I. Koltracht, *Mixed, componentwise, and structured condition numbers*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 3, 688–704. http://dx.doi.org/10.1137/0614049

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Fourth edition, Johns Hopkins University Press, Baltimore, MD, 2013.

[12] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl. **212/213** (1994), 3–20. http://dx.doi.org/10.1016/0024-3795(94)90393-x

[13] ———, *Accuracy and Stability of Numerical Algorithms*, Second edition, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. http://dx.doi.org/10.1137/1.9780898718027

[14] ———, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. http://dx.doi.org/10.1137/1.9780898717778

[15] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991. http://dx.doi.org/10.1017/cbo9780511840371

[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966. http://dx.doi.org/10.1007/978-3-662-12678-3

[17] C. S. Kenney and A. J. Laub, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput. **15** (1994), no. 1, 36–61. http://dx.doi.org/10.1137/0915003

[18] ———, *The matrix sign function*, IEEE Trans. Automat. Control **40** (1995), no. 8, 1330–1348. http://dx.doi.org/10.1109/9.402226

[19] C. S. Kenney, A. J. Laub and M. S. Reese, *Statistical condition estimation for linear systems*, SIAM J. Sci. Comput. **19** (1998), no. 2, 566–583. http://dx.doi.org/10.1137/s1064827595282519

[20] A. J. Laub and J. Xia, *Applications of statistical condition estimation to the solution of linear systems*, Numer. Linear Algebra Appl. **15** (2008), no. 6, 489–513. http://dx.doi.org/10.1002/nla.570

[21] ———, *Fast condition estimation for a class of structured eigenvalue problems*, SIAM J. Matrix Anal. Appl. **30** (2009), no. 4, 1658–1676. http://dx.doi.org/10.1137/070707713

[22] Y. Lin, L. Bao and Y. Wei, *Matrix sign function methods for solving projected generalized continuous-time Sylvester equations*, IEEE Trans. Automat. Control **55** (2010), no. 11, 2629–2634. http://dx.doi.org/10.1109/tac.2010.2064590

[23] Y. Lin and Y. Wei, *Condition numbers of the generalized Sylvester equation*, IEEE Trans. Automat. Control **52** (2007), no. 12, 2380–2385. http://dx.doi.org/10.1109/tac.2007.910727

[24] R. Mathias, *Condition estimation for matrix functions via the Schur decomposition*, SIAM J. Matrix Anal. Appl. **16** (1995), no. 2, 565–578. http://dx.doi.org/10.1137/s0895479893244389

[25] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal. **3** (1966), 287–310.

[26] S. M. Rump, *Structured perturbations Part I: Normwise distances*, SIAM J. Matrix Anal. Appl. **25** (2003), no. 1, 1–30. http://dx.doi.org/10.1137/s0895479802405732

[27] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, Inc., Boston, MA, 1990.

[28] T. Stykel, *On criteria for asymptotic stability of differential-algebraic equations*, ZAMM Z. Angew. Math. Mech. **82** (2002), no. 3, 147–158. http://dx.doi.org/10.1002/1521-4001(200203)82:3<147::aid-zamm147>3.0.co;2-h

[29] J.-G. Sun, *Perturbation analysis of the matrix sign function*, Linear Algebra Appl. **250** (1997), 177–206. http://dx.doi.org/10.1016/0024-3795(95)00522-6

[30] ———, *Condition numbers of the spectral projections*, Report UMINF 02.18, December 2, 2002. Umeå University.

[31] ———, *On the sensitivity of the spectral projection*, Linear Algebra Appl. **395** (2005), 83–94. http://dx.doi.org/10.1016/j.laa.2004.06.005

[32] X. Sun and E. S. Quintana-Ortí, *The generalized Newton iteration for the matrix sign function*, SIAM J. Sci. Comput. **24** (2002), no. 2, 669–683. http://dx.doi.org/10.1137/s1064827598348696

[33] ———, *Spectral division methods for block generalized Schur decompositions*, Math. Comp. **73** (2004), no. 248, 1827–1847. http://dx.doi.org/10.1090/s0025-5718-04-01667-9

[34] J. H. Wilkinson, *Sensitivity of eigenvalues*, Utilitas Math. **25** (1984), 5–76.

[35] L. Zhou, L. Lin, Y. Wei and S. Qiao, *Perturbation analysis and condition numbers of scaled total least squares problems*, Numer Algorithms **51** (2009), no. 3, 381–399. http://dx.doi.org/10.1007/s11075-009-9269-0

[36] L. Zhou, Y. Lin, Y. Wei and S. Qiao, *Perturbation analysis and condition numbers of symmetric algebraic Riccati equations*, Automatica J. IFAC **45** (2009), no. 4, 1005–1011. http://dx.doi.org/10.1016/j.automatica.2008.11.010

Wei-Guo Wang
School of Mathematical Sciences, Ocean University of China, Qingdao, 266100, P. R. China
E-mail address: wgwang@ouc.edu.cn

Chern-Shuh Wang
Department of Mathematics, National Cheng Kung University, Tainan City, Taiwan
E-mail address: chenshu@mail.ncku.edu.tw

Yi-Min Wei
School of Mathematical Sciences, Fudan University, Shanghai, 200433, P. R. China
and
Shanghai Key Laboratory of Contemporary Applied Mathematics
E-mail address: ymwei@fudan.edu.cn

Peng-Peng Xie
School of Mathematical Sciences, Ocean University of China, Qingdao, 266100, P. R. China
and
School of Mathematical Sciences, Fudan University, Shanghai, 200433, P. R. China
E-mail address: xie@ouc.edu.cn