

THE ANALYTIC SOLUTIONS OF A CLASS OF CONSTRAINED MATRIX MINIMIZATION AND MAXIMIZATION PROBLEMS WITH APPLICATIONS*

WEIWEI XU[†], WEN LI[‡], LEI ZHU[§], AND XUEPING HUANG[†]

Abstract. In this paper we present the analytic solutions of the following constrained matrix determinant and trace minimization and maximization problems: $\min_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(cI_n \pm \prod_{j=1}^m A_j U_j)|$, $\max_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(cI_n \pm \prod_{j=1}^m A_j U_j)|$ and $\min_{U_1, \dots, U_m \in \mathbb{U}_n} |\operatorname{tr}(cI_n \pm \prod_{j=1}^m A_j U_j)|$, $\max_{U_1, \dots, U_m \in \mathbb{U}_n} |\operatorname{tr}(cI_n \pm \prod_{j=1}^m A_j U_j)|$, where $c \in \mathbb{R}$, A_1, \dots, A_m are $n \times n$ complex matrices, I_n is the $n \times n$ identity matrix, \mathbb{U}_n is the set of $n \times n$ unitary matrices, and $\det(\cdot)$ and $\operatorname{tr}(\cdot)$ denote the matrix determinant function and the trace function, respectively. The given results improve on the corresponding ones in Marshall, Olkin, and Arnold [*Inequalities: Theory of Majorization and Its Applications*, Springer, New York, 2009], Lu [*Acta Math. Sinica*, 13 (1963), pp. 49–62], and Sun [*SIAM J. Matrix Anal. Appl.*, 20 (1983), pp. 611–625]. Some theoretical and practical applications are presented. In particular, some examples of applications to test signals of mechanical systems and aero engine fault diagnosis are given to show the efficiency of the proposed theoretical results.

Key words. nonconvex nonlinear matrix optimization problem, analytic solutions, compact real analytic manifold, Riemannian metric

AMS subject classifications. 90C06, 90C25, 93E12

DOI. 10.1137/17M1140777

1. Introduction. Matrix optimization problems play a vital role in scientific and engineering computing and in data science, for example, in control theory, model reduction, classification and comparative analysis of gene expression data, electronic structure calculations, image content authentication, aero engine fault diagnosis, and test signals of mechanical systems; see [6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38].

In this paper we consider the following special constrained matrix optimization problems:

$$(1.1a) \quad \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|,$$

$$(1.1b) \quad \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|$$

*Received by the editors July 26, 2017; accepted for publication (in revised form) April 10, 2019; published electronically June 20, 2019.

<http://www.siam.org/journals/siopt/29-2/M114077.html>

Funding: The work was supported by the Natural Science Foundation of Jiangsu Province under grant BK20181405, by the Natural Science Foundation of China under grants 11671158, U1811464, U1733201, and U1533202, and by the Major Project and Team Project of Guangdong Provincial Universities under grants 2016K2DXM025 and 2015KCXTD007.

[†]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, People's Republic of China (www19840904@sina.com, hxp_innocence@126.com).

[‡]School of Mathematical Sciences, South China Normal University, Guangzhou 510631, People's Republic of China (liwen@scnu.edu.cn).

[§]College of Engineering, Nanjing Agricultural University, Nanjing 210031, People's Republic of China (zhulei@njau.edu.cn).

and

$$(1.2a) \quad \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|,$$

$$(1.2b) \quad \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|,$$

where $c \in \mathbb{R}$ and A_1, \dots, A_m are $n \times n$ complex matrices, and we discuss their analytic solutions. The major contributions of this paper are the following.

- The analytic solutions of (1.1) and (1.2) are given explicitly by using the matrix analysis technique involving Jacobi's formula, matrix exponential, matrix function, majorization, etc.
- A new sharp estimation for an arbitrary generalized singular value of a matrix pair is presented.
- Some new practical numerical examples for the constrained matrix minimization and maximization problem are given to show the efficiency of our theoretical results.

1.1. Literature review. The existing work related to the constrained matrix minimization and maximization problems (1.1) and (1.2) is summarized as follows.

- A special case of (1.1b) was studied by Lu [4], where $c = 1$, $m = 2$, and both A_1 and A_2 are positive diagonal matrices with the main diagonal elements between 0 and 1 descending simultaneously or in ascending order.
- For a slightly extended case, similar results were presented by Sun [5], where $c = 1$, $m = 2$, and both A_1 and A_2 are nonnegative diagonal matrices with the main diagonal elements descending simultaneously or in ascending order.
- Similar results for (1.2b) with $c = 0$, $m = 2$ were given by von Neumann (1937) and Fan (1951), respectively (see, e.g., [2]).

Moreover, Sun [5] provided the Hoffman–Wielandt-type theorem for generalized singular values of Grassman matrix pairs. Other related work, such as Bai and Zha [23], Bai and Demmel [24], Chen and Li [25], Li [26], and Xu et al. [12] also consider Grassman matrix pairs. Compared with the above special cases, problems (1.1) and (1.2) are more complicated because they involve more constraints U_i and an arbitrary matrix A_i . This motivated us to use a new technique for giving the analytic solutions of (1.1) and (1.2). On the other hand, the given theoretical results can be applied to test signals for mechanical systems and aero engine fault diagnosis.

1.2. Organization. The rest of this paper is organized as follows. In section 2 we give some lemmas that will be useful for deducing the main results. In section 3 we provide analytic solutions of the constrained matrix determinant and trace optimization problems in (1.1) and (1.2), respectively. In section 4 we present theoretical applications. By using the proposed theoretical results we give a refined two-sided estimation of any generalized singular value of a matrix pair. In section 5 we give some practical numerical examples to show the efficiency of the theoretical results. Concluding remarks are given in section 6.

1.3. Notation. Throughout this paper we always use the following notation and definitions. Let \mathbb{R} , \mathbb{C} , \mathbb{R}^n , $\mathbb{C}^{m \times n}$, and \mathbb{U}_n be the sets of real numbers, complex numbers, n -dimensional real vectors, $m \times n$ complex matrices, and $n \times n$ unitary matrices, respectively. $|\cdot|$, $\Im(\cdot)$, and $\Re(\cdot)$ stand for absolute value, imaginary part, and real part of a complex number, respectively. The symbols I_n and $O_{m \times n}$

stand for the identity matrix of order n and the $m \times n$ zero matrix, respectively. For a square matrix $A \in \mathbb{C}^{n \times n}$, \bar{A} , A^T , A^H , A^{-1} , $\det(A)$, and $\operatorname{tr}(A)$ denote the conjugate, transpose, conjugate transpose, inverse, determinant, and trace of a matrix A , respectively. By $\|\cdot\|_2$ we denote the spectral norm of a matrix. The conjugate of a number $c \in \mathbb{C}$ is denoted by c^* . We denote the Cartesian product of two sets X and Y by $X \times Y := \{(x, y) : x \in X, y \in Y\}$. If X and Y have some algebraic or topological structure, $X \times Y$ is understood with the natural product structure. For a matrix $A \in \mathbb{C}^{n \times n}$, we denote by $\sigma(A)$ the set of its singular values; usually, we assume that its singular values are arranged in decreasing order, i.e., $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A) \geq 0$. We review some basic notions from topology; see [1] for details. Recall that a metric space (X, d) is a nonempty set X equipped with a distance function $d(\cdot, \cdot) : X \rightarrow [0, \infty)$ that satisfies the following conditions:

1. $\forall x, y \in X, d(x, y) = d(y, x)$;
2. $d(x, y) = 0 \Leftrightarrow x = y$;
3. $\forall x, y, z \in X, d(x, z) \leq d(x, y) + d(y, z)$.

For each $x \in X$ and $r > 0$, we denote the open ball with radius r and center x by $B(x, r) = \{y \in X : d(x, y) < r\}$.

From two metric spaces (X, d_X) and (Y, d_Y) , we can define a product metric space $X \times Y$ with a distance function

$$d((x_1, y_1), (x_2, y_2)) := \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}.$$

For a vector $\mathbf{a} = (a_1, \dots, a_n)^T \in \mathbb{R}^n$, we denote its decreasing rearrangement by $\mathbf{a}_\downarrow = (a_{[1]}, \dots, a_{[n]})^T$ with

$$a_{[1]} \geq \cdots \geq a_{[n]}$$

being the components of \mathbf{a} arranged in decreasing order, and where $[1], \dots, [n]$ denote subscripts.

DEFINITION 1.1 (Marshall, Olkin, and Arnold [2]). *Let $\mathbf{a} = (a_1, \dots, a_n)$, $\mathbf{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$ be two vectors. We say that \mathbf{a} is majorized by \mathbf{b} (\mathbf{b} majorizes \mathbf{a}), denoted by $\mathbf{a} \prec \mathbf{b}$, if*

$$\sum_{i=1}^k a_{[i]} \leq \sum_{i=1}^k b_{[i]} \quad \forall 1 \leq k < n, \quad \sum_{i=1}^n a_{[i]} = \sum_{i=1}^n b_{[i]}.$$

2. Preliminaries. The following results include many classical inequalities as their special cases.

LEMMA 2.1 (Marshall, Olkin, and Arnold [2, Theorem A.4]). *Let $I \subseteq \mathbb{R}$ be an open interval. Let $\phi : I^n \rightarrow \mathbb{R}$ be a continuously differentiable function. Assume that ϕ is symmetric, that is,*

$$\phi(a_1, \dots, a_n) = \phi(a_{\eta(1)}, \dots, a_{\eta(n)})$$

for any $\eta \in S_n$, the permutation group of $\{1, \dots, n\}$. If, for each $(x_1, \dots, x_n) \in I^n$, $1 \leq i, j \leq n$,

$$(2.1) \quad (x_i - x_j) \left(\frac{\partial \phi}{\partial x_i} - \frac{\partial \phi}{\partial x_j} \right) \geq 0,$$

then $\phi(\mathbf{a}) \leq \phi(\mathbf{b})$ whenever $\mathbf{a} \prec \mathbf{b}$. The converse holds as well.

Remark 2.1. A continuously differentiable function ϕ is said to be Schur-convex if ϕ is symmetric and satisfies (2.1). It is easy to check that functions ϕ_1 and ϕ_2 in the following example are Schur-convex, which can be used later in the paper.

Example 2.1. Consider two symmetric functions

$$\phi_1(x_1, \dots, x_n) = \sum_{i=1}^n \log(1 + e^{x_i})$$

on \mathbb{R}^n and

$$\phi_2(x_1, \dots, x_n) = - \sum_{i=1}^n \log(1 - e^{x_i})$$

on I^n , where $I = (-\infty, 0)$, satisfying (2.1).

The next lemma is about logarithmic majorization of singular values.

LEMMA 2.2 (Marshall, Olkin, and Arnold [2, Theorem H.1.b]). *Let $A_1, \dots, A_m \in \mathbb{C}^{n \times n}$. Then*

$$\prod_{i=1}^k \sigma_i(A_1 \cdots A_m) \leq \prod_{i=1}^k \sigma_i(A_1) \cdots \sigma_i(A_m), \quad 1 \leq k < n,$$

and $\prod_{i=1}^n \sigma_i(A_1 \cdots A_m) = \prod_{i=1}^n \sigma_i(A_1) \cdots \sigma_i(A_m)$. In particular, if $\sigma_n(A_1 \cdots A_m) > 0$, then

$$\begin{aligned} & (\log(\sigma_1(A_1 \cdots A_m)), \dots, \log(\sigma_n(A_1 \cdots A_m))) \\ & \prec (\log(\sigma_1(A_1) \cdots \sigma_1(A_m)), \dots, \log(\sigma_n(A_1) \cdots \sigma_n(A_m))). \end{aligned}$$

We will next provide a series of elementary facts from the perspective of linear algebra and topology. The proofs of the following three lemmas are easy and can be found in [12].

LEMMA 2.3. *If a square matrix A satisfies that $\operatorname{tr}(AH) \in \mathbb{R}$ for arbitrary Hermitian matrix H , then A is Hermitian as well.*

LEMMA 2.4. *Let $A \in \mathbb{C}^{n \times n}$ be a matrix such that $I_n + A$ is invertible. Then $I_n - (I_n + A)^{-1}A$ is invertible and $A = (I_n - (I_n + A)^{-1}A)^{-1} - I_n$. If $(I_n + A)^{-1}A$ is furthermore Hermitian, then A is Hermitian as well.*

LEMMA 2.5. *Let D be a diagonal matrix with pairwise different diagonal elements. If A satisfies $AD = DA$, then A is diagonal.*

LEMMA 2.6 (Berge [21]). *Let X and Y be metric spaces and Y be compact. Let F be a continuous function on $X \times Y$. Define $G(\cdot) = \max_{y \in Y} F(\cdot, y)$ as a function on X . Then G is continuous.*

LEMMA 2.7 (Xu et al. [12]). *Let*

$$\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_n), \quad \gamma_1 \geq \dots \geq \gamma_n \geq 0.$$

Then

$$\max_{U \in \mathbb{U}_n} |\det(I_n + \Gamma U)| = \prod_{i=1}^n (1 + \gamma_i).$$

LEMMA 2.8. *Let*

$$\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n), \quad \gamma_1 \geq \dots \geq \gamma_n \geq 0.$$

Then

$$\min_{U \in \mathbb{U}_n} |\det(I_n + \Gamma U)| = \begin{cases} \prod_{i=1}^n (1 - \gamma_i), & \gamma_1 \leq 1, \\ \prod_{i=1}^n (\gamma_i - 1), & \gamma_n \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. First, we make some observations for simplifying the proof. We note that $F(\Gamma, U) := |\det(I_n + \Gamma U)|^2$ can be understood as an analytic function on the manifold $\mathbb{R}^n \times \mathbb{U}_n$ (here \mathbb{U}_n is understood to be a compact analytic manifold). Fixing Γ , we see that the infimum of $F(\Gamma, \cdot)$ is always achieved at some $U_1 \in \mathbb{U}_n$ (depending on Γ).

Second, in the following arguments, we can focus on the case when $\gamma_1 \leq 1$. If $\gamma_n \geq 1$, the problem can be reduced to the case when $\gamma_1 \leq 1$. In fact, noting that

$$|\det(I_n + \Gamma U)| = |\det(I_n + U^H \Gamma^{-1}) \det(\Gamma) \det(U)| = |\det(I_n + \Gamma^{-1} U^H) \det(\Gamma)|,$$

we then obtain that

$$\min_{U \in \mathbb{U}_n} |\det(I_n + \Gamma U)| = \prod_{i=1}^n \gamma_i \cdot \prod_{i=1}^n (1 - \gamma_i^{-1}) = \prod_{i=1}^n (\gamma_i - 1).$$

For the remaining case, that is, if $\gamma_1 > 1$, $1 > \gamma_n \geq 0$, we claim that

$$\min_{U \in \mathbb{U}_n} |\det(I_n + \Gamma U)| = 0.$$

Indeed, since $\gamma_1 > 1$, $1 > \gamma_n \geq 0$, we can find $\mathbf{v} = (v_1, \dots, v_n)^T \in \mathbb{R}^n$ such that $\|\mathbf{v}^T\| = \|\mathbf{v}^T \Gamma\| = 1$. The geometric picture is clear: the unit sphere in \mathbb{R}^n and the ellipse (or cylinder, if for some j , $\gamma_j = \dots = \gamma_n = 0$) $\sum_{i=1}^n \gamma_i^2 |v_i|^2 = 1$ must intersect. By elementary geometry we see that there is $V \in O(n) \subset \mathbb{U}_n$ such that $\mathbf{v}^T \Gamma V = -\mathbf{v}^T$. Thus, $\det(I_n + \Gamma V) = 0$ and hence $\min_{U \in \mathbb{U}_n} |\det(I_n + \Gamma U)| = 0$.

The last observation is a continuity argument. Set $X = \mathbb{R}^n$ and $Y = \mathbb{U}_n$. Then they have metric space structures induced from their natural Riemannian metrics and F can be viewed as a continuous function on the product space $X \times Y$. Define

$$G(\Gamma) := \min_{\mathbb{U}_n} F(\Gamma, \cdot).$$

It follows from Lemma 2.6 that G is continuous on X . It is sufficient to check that

$$G(\Gamma) = \prod_{i=1}^n (1 - \gamma_i)^2$$

on

$$\{(\gamma_1, \dots, \gamma_n) : 1 \geq \gamma_1 \geq \dots \geq \gamma_n \geq 0\}.$$

It is known that the set $\{(\gamma_1, \dots, \gamma_n) : 1 > \gamma_1 > \dots > \gamma_n > 0\}$ is a dense subset of the set $\{(\gamma_1, \dots, \gamma_n) : 1 \geq \gamma_1 \geq \dots \geq \gamma_n \geq 0\}$. Then without loss of generality we assume that

$$1 > \gamma_1 > \dots > \gamma_n > 0$$

in the following discussion.

Let $U_1 \in \mathbb{U}_n$ satisfy $F(\Gamma, U_1) = \min_{\mathbb{U}_n} F(\Gamma, \cdot)$. We claim that $I_n + \Gamma U_1$ is invertible. In fact, suppose that there exists a nonzero vector $\mathbf{c} = (c_1, \dots, c_n)^T \in \mathbb{C}^n$ such that

$$(I_n + \Gamma U_1)\mathbf{c} = 0.$$

We have

$$\|\mathbf{c}\| = \|\Gamma U_1 \mathbf{c}\| \leq \gamma_1 \|\mathbf{c}\|,$$

from which one can get $\|\mathbf{c}\| = 0$ since $\gamma_1 < 1$.

Let H be an arbitrary Hermitian matrix. For $t \in \mathbb{R}$, we have $e^{itH} \in \mathbb{U}_n$. Define

$$F_1(t) = |\det(I_n + \Gamma U_1 e^{itH})|^2.$$

We have $F_1(t) \geq F_1(0) \forall t$. Hence, $F_1'(0) = 0$.

As we have shown, $I_n + \Gamma U_1$ is invertible. Recall Jacobi's formula [3] for the derivative of the determinant of a matrix function:

$$\frac{d}{dt} \det(A(t))|_{t=0} = \det(A(0)) \operatorname{tr} \left[A(0)^{-1} \frac{d}{dt} (A(t))|_{t=0} \right],$$

where $A(t)$ is a differentiable curve in $\mathbb{C}^{n \times n}$ with $A(0)$ invertible. Hence,

$$\begin{aligned} & \frac{d}{dt} |\det(A(t))|^2|_{t=0} \\ &= \frac{d}{dt} \det(A(t))|_{t=0} (\det(A(0)))^H + \det(A(0)) \left(\frac{d}{dt} \det(A(t)) \right)^H|_{t=0} \\ &= |\det(A(0))|^2 \operatorname{tr} \left[A(0)^{-1} \frac{d}{dt} (A(t))|_{t=0} \right] \\ &\quad + |\det(A(0))|^2 \left\{ \operatorname{tr} \left[A(0)^{-1} \frac{d}{dt} (A(t))|_{t=0} \right] \right\}^H. \end{aligned}$$

Applying the above to F_1 , we obtain that

$$\begin{aligned} 0 = F_1'(0) &= i |\det(I_n + \Gamma U_1)|^2 \operatorname{tr}((I_n + \Gamma U_1)^{-1} \Gamma U_1 H) \\ &\quad - i |\det(I_n + \Gamma U_1)|^2 [\operatorname{tr}((I_n + \Gamma U_1)^{-1} \Gamma U_1 H)]^H. \end{aligned}$$

It follows that, for each Hermitian matrix H ,

$$\operatorname{tr}((I_n + \Gamma U_1)^{-1} \Gamma U_1 H) \in \mathbb{R}.$$

By Lemma 2.3, $(I_n + \Gamma U_1)^{-1} \Gamma U_1$ is Hermitian. Thus, by Lemma 2.4, ΓU_1 is Hermitian. Note that

$$(\Gamma U_1)^2 = (\Gamma U_1)(\Gamma U_1)^H = (\Gamma U_1)^H(\Gamma U_1).$$

A simple calculation shows that $U_1 \Gamma^2 = \Gamma^2 U_1$. By Lemma 2.5 and the assumption that $1 > \gamma_1 > \dots > \gamma_n > 0$, U_1 is diagonal. But U_1 is unitary and ΓU_1 is Hermitian, so we have $U_1 = \operatorname{diag}(a_1, \dots, a_n)$, where $a_i \in \{\pm 1\} \forall i$.

Set $D(n) = \{D \in M_{n \times n}(\mathbb{C}) : U = \operatorname{diag}(d_1, \dots, d_n), d_i \in \{\pm 1\} \forall i\}$. We have shown that

$$\min_{\mathbb{U}_n} F(\Gamma, \cdot) = F(\Gamma, U_1) = \min_{D(n)} F(\Gamma, \cdot).$$

Since $1 > \gamma_1 > \cdots > \gamma_n > 0$, we have

$$\min_{D(n)} F(\Gamma, \cdot) = \min_{d_i \in \{\pm 1\} \forall i} |1 + \gamma_i d_i|^2 = \prod_{i=1}^n |1 - \gamma_i|^2,$$

which proves the lemma. \square

COROLLARY 2.9. *Let $A \in \mathbb{C}^{n \times n}$ have the singular value decomposition (SVD)*

$$A = W_1 \Gamma W_2^H,$$

where $W_1, W_2 \in \mathbb{U}_n$ and $\Gamma = \text{diag}(\sigma_1(A), \dots, \sigma_n(A))$ arranged in decreasing order. Then

$$\max_{U \in \mathbb{U}_n} |\det(I_n + AU)| = \prod_{i=1}^n (1 + \sigma_i(A))$$

and

$$\min_{U \in \mathbb{U}_n} |\det(I_n + AU)| = \begin{cases} \prod_{i=1}^n (1 - \sigma_i(A)), & \sigma_1(A) \leq 1, \\ \prod_{i=1}^n (\sigma_i(A) - 1), & \sigma_n(A) \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. Note that

$$|\det(I_n + AU)| = |\det(I_n + W_1 \Gamma W_2^H U)| = |\det(I_n + \Gamma W_2^H U W_1)|.$$

When U runs over \mathbb{U}_n , $W_2^H U W_1$ runs over \mathbb{U}_n as well. The assertions then follow from Lemmas 2.7 and 2.8. \square

3. Solutions of problems (1.1) and (1.2). In this section we consider the analytic solutions of the constrained matrix determinant and trace optimization problems in (1.1) and (1.2), respectively.

3.1. Solutions of problem (1.1). We first provide the analytic solutions of problem (1.1) as follows.

THEOREM 3.1. *Let $A_1, \dots, A_m \in \mathbb{C}^{n \times n}$ and $c \in \mathbb{R}$. Then we have*

$$\begin{aligned} (1) \quad & \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right| = \prod_{i=1}^n \left(|c| + \prod_{j=1}^m \sigma_i(A_j) \right), \\ (2) \quad & \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right| \\ &= \begin{cases} \prod_{i=1}^n (|c| - \prod_{j=1}^m \sigma_i(A_j)), & \prod_{j=1}^m \sigma_1(A_j) \leq |c|, \\ \prod_{i=1}^n (\prod_{j=1}^m \sigma_i(A_j) - |c|), & \prod_{j=1}^m \sigma_n(A_j) \geq |c|, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Proof. Note that $-U_1$ runs over \mathbb{U}_n when U_1 runs over \mathbb{U}_n . The extreme values of $|\det(cI_n - \prod_{j=1}^m A_j U_j)|$ are the same as those of $|\det(cI_n + \prod_{j=1}^m A_j U_j)|$. We shall prove the assertion for $|\det(cI_n + \prod_{j=1}^m A_j U_j)|$ with three cases.

Case 1. If $c > 0$, then by replacing A_1 with $\frac{A_1}{c}$ we only need to consider the case when $c = 1$.

First, we will assume that $\prod_{j=1}^m \sigma_1(A_j) \leq 1$ for the minimum part of the theorem.

By a continuity argument similar to that in the proof of Lemma 2.7, we can assume that, $\forall 1 \leq j \leq m$,

$$\sigma_1(A_j) > \sigma_2(A_j) > \cdots > \sigma_n(A_j) > 0.$$

Hence, we may further assume $\prod_{j=1}^m \sigma_1(A_j) < 1$. Fix $U_1, \dots, U_{m-1} \in \mathbb{U}_n$ and consider the supremum over $U(n) \in \mathbb{U}_n$. Set $A = A_1 U_1 \cdots A_{m-1} U_{m-1} A_m$. By Corollary 2.9 we have

$$\max_{U_m \in \mathbb{U}_n} |\det(I_n + AU_m)| = \prod_{i=1}^n (1 + \sigma_i(A)).$$

Similarly,

$$\min_{U_m \in \mathbb{U}_n} |\det(I_n + AU_m)| = \prod_{i=1}^n (1 - \sigma_i(A))$$

if $\sigma_1(A) \leq 1$. Now, by Lemma 2.2 we obtain that

$$\prod_{i=1}^k \sigma_i(A) \leq \prod_{i=1}^k \sigma_i(A_1 U_1) \cdots \sigma_i(A_{m-1} U_{m-1}) \sigma_i(A_m), \quad 1 \leq k < n,$$

and $\prod_{i=1}^n \sigma_i(A) = \prod_{i=1}^n \sigma_i(A_1 U_1) \cdots \sigma_i(A_{m-1} U_{m-1}) \sigma_i(A_m)$. Note that $\sigma_i(A_j U_j) = \sigma_i(A_j)$. We see that $\sigma_n(A) > 0$ and

$$\begin{aligned} & (\log(\sigma_1(A)), \dots, \log(\sigma_n(A))) \\ & \prec (\log(\sigma_1(A_1) \cdots \sigma_1(A_m)), \dots, \log(\sigma_n(A_1) \cdots \sigma_n(A_m))). \end{aligned}$$

It follows from Lemmas 2.1 and 2.2 that

$$\begin{aligned} & \phi(\log(\sigma_1(A)), \dots, \log(\sigma_n(A))) \\ & \leq \phi(\log(\sigma_1(A_1) \cdots \sigma_1(A_m)), \dots, \log(\sigma_n(A_1) \cdots \sigma_n(A_m))) \end{aligned}$$

for each Schur-convex function ϕ . Taking $\phi = \phi_1$ or ϕ_2 , where ϕ_1 and ϕ_2 are defined by Example 2.1, gives

$$\sum_{i=1}^n \log(1 + \sigma_i(A)) \leq \sum_{i=1}^n \log \left(1 + \prod_{j=1}^m \sigma_i(A_j) \right)$$

and

$$-\sum_{i=1}^n \log(1 - \sigma_i(A)) \leq -\sum_{i=1}^n \log \left(1 - \prod_{j=1}^m \sigma_i(A_j) \right).$$

It directly follows that

$$\prod_{i=1}^n (1 + \sigma_i(A)) \leq \prod_{i=1}^n \left(1 + \prod_{j=1}^m \sigma_i(A_j) \right)$$

and

$$\prod_{i=1}^n (1 - \sigma_i(A)) \geq \prod_{i=1}^n \left(1 - \prod_{j=1}^m \sigma_i(A_j) \right).$$

Hence, we have

$$\max_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(I_n + A_1 U_1 \cdots A_m U_m)| \leq \prod_{i=1}^n \left(1 + \prod_{j=1}^m \sigma_i(A_j) \right)$$

and

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(I_n + A_1 U_1 \cdots A_m U_m)| \geq \prod_{i=1}^n \left(1 - \prod_{j=1}^m \sigma_i(A_j) \right).$$

In order to show the desired assertions we need only to prove the equalities can be achieved in the above inequalities. In fact, let A_j have the SVD

$$A_j = W_j \operatorname{diag}(\sigma_1(A_j), \dots, \sigma_n(A_j)) V_j^H \quad \forall j.$$

Then, taking $U_j = V_j W_{j+1}^H$, where $W_{m+1} = W_1$, gives the first desired equation. The second one can be obtained by taking $U_j = V_j W_{j+1}^H \forall j < m$ and $U_m = -V_m W_1^H$.

Next, we consider the remaining two cases for the minimum part of this theorem. If $\prod_{j=1}^m \sigma_1(A_j) \geq 1$, then $\prod_{j=1}^m \sigma_1(A_j^{-1}) \leq 1$.

Since

$$\begin{aligned} |\det(I_n + A_1 U_1 \cdots A_m U_m)| &= |\det(I_n + U_m^H A_m^{-1} \cdots U_1^H A_1^{-1})| |\det(A_1 U_1 \cdots A_m U_m)| \\ &= |\det(I_n + A_m^{-1} U_{m-1}^H \cdots U_1^H A_1^{-1} U_m^H)| |\det(A_1 \cdots A_m)|. \end{aligned}$$

By the above proof we have

$$\begin{aligned} \min_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(I_n + A_1 U_1 \cdots A_m U_m)| &= \prod_{i=1}^n \left(1 - \prod_{j=1}^m \sigma_i^{-1}(A_j) \right) \cdot \prod_{j=1}^m |\det(A_j)| \\ &= \prod_{i=1}^n \left(\prod_{j=1}^m \sigma_i(A_j) - 1 \right), \end{aligned}$$

which proves statement (2).

If $\prod_{j=1}^m \sigma_1(A_j) > 1$ and $1 > \prod_{j=1}^m \sigma_n(A_j) \geq 0$, then in the following we shall prove that there exist $U_1, \dots, U_m \in \mathbb{U}_n$ such that

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(I_n + A_1 U_1 \cdots A_m U_m)| = 0.$$

First, we set $U_j = V_j W_{j+1}^H$ for $1 \leq j \leq m-1$ and $U_m = V_m U W_1^H$, where $U \in \mathbb{U}_n$ is undetermined. Then we have

$$\begin{aligned} &|\det(I_n + A_1 U_1 \cdots A_m U_m)| \\ &= \left| \det \left(I_n + W_1 \operatorname{diag} \left(\prod_{j=1}^m \sigma_1(A_j), \dots, \prod_{j=1}^m \sigma_n(A_j) \right) U W_1^H \right) \right| \\ &= \left| \det \left(I_n + \operatorname{diag} \left(\prod_{j=1}^m \sigma_1(A_j), \dots, \prod_{j=1}^m \sigma_n(A_j) \right) U \right) \right|. \end{aligned}$$

Let

$$\Gamma = \operatorname{diag} \left(\prod_{j=1}^m \sigma_1(A_j), \dots, \prod_{j=1}^m \sigma_n(A_j) \right).$$

By the same argument as in the proof of Lemma 2.8, we can choose some $U \in O(n) \subset \mathbb{U}_n$ and some nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}^T \Gamma U = -\mathbf{v}^T$. It follows that $\det(I_n + \Gamma U) = 0$. This proves

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} |\det(I_n + A_1 U_1 \cdots A_m U_m)| = 0.$$

By the above proof, we have shown that statements (1) and (2) of this theorem hold for the case when $c = 1$.

Case 2. If $c = 0$, it is easy to see that (1) and (2) are true.

Case 3. If $c < 0$, one may easily check that the following equations hold:

$$\begin{aligned} \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right| &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left((-1) \left(|c|I_n \mp \prod_{j=1}^m A_j U_j \right) \right) \right| \\ &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(|c|I_n \pm \prod_{j=1}^m A_j U_j \right) \right| \end{aligned}$$

and

$$\begin{aligned} \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right| &= \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left((-1) |c|I_n \mp \prod_{j=1}^m A_j U_j \right) \right| \\ &= \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(|c|I_n \pm \prod_{j=1}^m A_j U_j \right) \right|. \end{aligned}$$

Hence, statements (1) and (2) follow from case 1. This completes the proof of the theorem. \square

The following two results are special cases of Theorem 3.1.

COROLLARY 3.2. *Let $c \in \mathbb{R}$ and*

$$\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n), \quad \Delta = \text{diag}(\delta_1, \dots, \delta_n)$$

with

$$\gamma_1 \geq \dots \geq \gamma_n \geq 0, \quad \delta_1 \geq \dots \geq \delta_n \geq 0.$$

Then

$$\max_{U, V \in \mathbb{U}_n} |\det(cI_n \pm \Gamma U \Delta V)| = \prod_{i=1}^n (|c| + \gamma_i \delta_i).$$

COROLLARY 3.3. *Let $c \in \mathbb{R}$ and*

$$\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n), \quad \Delta = \text{diag}(\delta_1, \dots, \delta_n)$$

with

$$\gamma_1 \geq \dots \geq \gamma_n \geq 0, \quad \delta_1 \geq \dots \geq \delta_n \geq 0,$$

and $\gamma_1 \delta_1 \leq |c|$. Then

$$\min_{U, V \in \mathbb{U}_n} |\det(cI_n \pm \Gamma U \Delta V)| = \prod_{i=1}^n (|c| - \gamma_i \delta_i).$$

Remark 3.1. To the best of our knowledge, Corollary 3.3 with $c = 1$ was first shown by Lu in [4]. Corollary 3.2 with $1 \geq \gamma_1$ and $1 \geq \delta_1$ was given in [5] as a consequence of Lu's approach. However, Lu's result in [4] holds under the additional assumption that $1 \geq \gamma_1 \delta_1$.

3.2. Solutions of problem (1.2). We now give the analytic solutions of problem (1.2) as follows.

LEMMA 3.4 (Marshall, Olkin, and Arnold [2]). *Let $A_1, \dots, A_m \in \mathbb{C}^{n \times n}$. We have the following equations:¹*

$$\begin{aligned} \max_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left[\operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right] &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| = \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j), \\ \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| &= 0. \end{aligned}$$

THEOREM 3.5. *Let $c \in \mathbb{R}$ and $A_1, \dots, A_m \in \mathbb{C}^{n \times n}$ with $\sigma_1(A_i) \geq \sigma_2(A_i) \geq \dots \geq \sigma_n(A_i)$, $i = 1, \dots, m$. We have the following:*

$$\begin{aligned} (1) \quad & \max_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left[\operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right] = nc + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j), \\ & \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| = n|c| + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j); \\ (2) \quad & \min_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left[\operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right] = nc - \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j), \\ & \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| \\ &= \begin{cases} n|c| - \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j), & \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j) \leq |c|, \\ 0, & \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j) \geq |c|. \end{cases} \end{aligned}$$

Proof. (1) We first consider the case when $c \geq 0$. It follows from Lemma 3.4 that

$$\begin{aligned} \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| &\leq \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left(\left| \operatorname{tr}(cI_n) \right| + \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| \right) \\ &\leq nc + \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| \\ (3.1) \quad &\leq nc + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j). \end{aligned}$$

Let $A_j = S_j^H \Sigma_j T_j$, $j = 1, \dots, m$, be SVDs of A_j , where S_j and T_j are unitary matrices. Then we set $V_1 = T_m^H S_1$, $V_j = T_{j-1}^H S_j$, $j = 2, \dots, m$, and $H_1 = -T_m^H S_1$, $H_j = T_{j-1}^H S_j$, $j = 2, \dots, m$. Therefore,

$$\begin{aligned} \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n + \prod_{j=1}^m U_j A_j \right) \right| &\geq \left| \operatorname{tr} \left(cI_n + \prod_{j=1}^m V_j A_j \right) \right| \\ (3.2) \quad &\geq nc + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j) \end{aligned}$$

¹The first equation can be derived by using the results of von Neumann (1937) and Fan (1951) given in [2, Chapter 20, Theorem B.1].

and

$$\begin{aligned}
 \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n - \prod_{j=1}^m U_j A_j \right) \right| &\geq \left| \operatorname{tr} \left(cI_n - \prod_{j=1}^m H_j A_j \right) \right| \\
 (3.3) \qquad \qquad \qquad &\geq nc + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j).
 \end{aligned}$$

For $c < 0$, we have

$$\begin{aligned}
 \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left((-1) \left(|c|I_n \mp \prod_{j=1}^m U_j A_j \right) \right) \right| \\
 (3.4) \qquad \qquad \qquad &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(|c|I_n \pm \prod_{j=1}^m U_j A_j \right) \right|.
 \end{aligned}$$

From Lemma 3.4 we have

$$\begin{aligned}
 \max_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left(\operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right) \\
 &= \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left(\Re(\operatorname{tr}(cI_n)) \pm \Re \left(\operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right) \right) \\
 &= nc + \max_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left(\operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right) \\
 (3.5) \qquad \qquad \qquad &= nc + \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j).
 \end{aligned}$$

Combining (3.1)–(3.5) yields the desired assertion (1).

(2) Since

$$\begin{aligned}
 \min_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left[\operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right] \\
 &= \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left(\Re(\operatorname{tr}(cI_n)) \pm \Re \left(\operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right) \right) \\
 &= nc - \max_{U_1, \dots, U_m \in \mathbb{U}_n} \Re \left(\operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right) \\
 (3.6) \qquad \qquad \qquad &= nc - \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j),
 \end{aligned}$$

it is easy to see that

$$\begin{aligned}
 \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| &\geq \left| \operatorname{tr}(cI_n) - \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| \\
 (3.7) \qquad \qquad \qquad &= \left| n|c| - \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right|.
 \end{aligned}$$

If $\max_{U_1, \dots, U_m \in \mathbb{U}_n} |\operatorname{tr}(\pm \prod_{j=1}^m U_j A_j)| = \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j) \leq n|c|$, then

$$\begin{aligned} \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| &\geq \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| |\operatorname{tr}(cI_n)| - \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| \right| \\ &= n|c| - \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(\prod_{j=1}^m U_j A_j \right) \right| \\ (3.8) \qquad \qquad \qquad &= n|c| - \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j). \end{aligned}$$

Since there exist $V_1, \dots, V_m, H_1, \dots, H_m \in \mathbb{U}_n$ such that

$$\operatorname{tr} \left(\prod_{j=1}^m V_j A_j \right) = \operatorname{tr} \left(- \prod_{j=1}^m H_j A_j \right) = \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j)$$

and we set $\mathbf{V}_1 = V_1$ or $\mathbf{V}_1 = H_1$ and $\mathbf{V}_j = V_j$, $j = 2, \dots, m$, we have

$$\begin{aligned} \min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| &\leq \left| \operatorname{tr}(cI_n) - \operatorname{tr} \left(\prod_{j=1}^m \mathbf{V}_j A_j \right) \right| \\ (3.9) \qquad \qquad \qquad &= n|c| - \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j). \end{aligned}$$

If $\max_{U_1, \dots, U_m \in \mathbb{U}_n} |\operatorname{tr}(\prod_{j=1}^m U_j A_j)| = \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j) > n|c|$, then by Lemma 3.4 we know that $0 \leq |\operatorname{tr}(\prod_{j=1}^m U_j A_j)| \leq \sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j)$. When $n = 2k$, $k = 1, 2, \dots$, we set

$$\begin{aligned} 0 \leq \sin(\alpha) &= \frac{n|c|}{\sum_{i=1}^n \prod_{j=1}^m \sigma_i(A_j)} < 1, \\ C_{k \times k} &= \sin(\alpha)I_k, \quad S_{k \times k} = \cos(\alpha)I_k, \quad L = \begin{pmatrix} C_{k \times k} & S_{k \times k} \\ -S_{k \times k} & C_{k \times k} \end{pmatrix}. \end{aligned}$$

When $n = 2k + 1$, $k = 0, 1, \dots$, we set

$$\begin{aligned} -1 < \sin(\beta) &= \frac{n|c| - \prod_{j=1}^m \sigma_n(A_j)}{\sum_{i=1}^{n-1} \prod_{j=1}^m \sigma_i(A_j)} < 1, \\ C_{k \times k} &= \sin(\beta)I_k, \quad S_{k \times k} = \cos(\beta)I_k, \quad L = \begin{pmatrix} C_{k \times k} & S_{k \times k} & O_{k \times 1} \\ -S_{k \times k} & C_{k \times k} & O_{k \times 1} \\ O_{1 \times k} & O_{1 \times k} & 1 \end{pmatrix}. \end{aligned}$$

Then there exist $\mathbf{T}_1 = T_m^H L S_1$ or $\mathbf{T}_1 = -T_m^H L S_1$, $\mathbf{T}_j = T_{j-1}^H S_j$, $j = 2, \dots, m \in U(n)$, such that $\operatorname{tr}(\prod_{j=1}^m \mathbf{T}_j A_j) = n|c|$. It follows that

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m U_j A_j \right) \right| \leq \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m \mathbf{T}_j A_j \right) \right| = 0,$$

which, together with (3.6)–(3.9), gives assertion (2). \square

Remark 3.2. The analytic solutions of the constrained matrix determinant and trace minimization and maximization problems in (1.1) and (1.2) are given by Theorems 3.1 and 3.5, respectively. When $c = 0$ Theorem 3.5 reduces to the result given by von Neumann (1937) and Fan (1951) in [2, Chapter 20, Theorem B.1]. Hence, Theorem 3.5 generalizes the existing results in [2].

4. Sharp upper and lower bounds of any generalized singular value of a matrix pair. In this section we will employ the analytic solutions of the constrained matrix determinant and trace minimization and maximization problems in (1.1) and (1.2) to deduce sharp upper and lower bounds of any generalized singular value of a matrix pair.

Let $\{E_1, E_2\}$ be an (m, p, s) -Grassmann matrix pair (assume $s \leq m, s \leq p$) and denote the generalized singular values of $\{E_1, E_2\}$ by $\sigma\{E_1, E_2\} = \{(\alpha_i, \beta_i), i = 1, 2, \dots, n\}$. The generalized SVD of a matrix pair $\{E_1, E_2\}$ is given as follows (see, e.g., [12]): there exist unitary matrices $U \in \mathbb{C}^{m \times m}$, $V \in \mathbb{C}^{p \times p}$, and a nonsingular matrix $R \in \mathbb{C}^{s \times s}$ such that

$$(4.1) \quad \begin{aligned} U^H E_1 R^{-1} &= \Sigma_{E_1}, & V^H E_2 R^{-1} &= \Sigma_{E_2}, \\ \Sigma_{E_1} &= \begin{pmatrix} \Lambda & \\ & O_{E_1(m-k-t) \times (n-k-t)} \end{pmatrix}, & \Sigma_{E_2} &= \begin{pmatrix} O_{E_2(p+k-s) \times k} & \\ & \Omega \end{pmatrix}, \end{aligned}$$

where O_{E_1} and O_{E_2} are zero matrices, and

$$\Lambda = \text{diag}(\alpha_1, \dots, \alpha_{k+t}), \quad \Omega = \text{diag}(\beta_{k+1}, \dots, \beta_n)$$

with

$$\begin{aligned} 1 &= \alpha_1 = \dots = \alpha_k > \alpha_{k+1} \geq \dots \geq \alpha_{k+t} > \alpha_{k+t+1} = \dots = \alpha_s = 0, \\ 0 &= \beta_1 = \dots = \beta_k < \beta_{k+1} \leq \dots \leq \beta_{k+t} < \beta_{k+t+1} = \dots = \beta_s = 1, \end{aligned}$$

and $\alpha_i^2 + \beta_i^2 = 1, 1 \leq i \leq s$.

Next, we may use the results in section 3 to deduce the upper and lower bounds of α_i and β_i .

Since

$$\begin{aligned} 1 &\geq \alpha_i^2 \geq \alpha_{i+1}^2 \geq \dots \geq \alpha_s^2 \geq 0, \quad i = 1, \dots, s, \\ 0 &\leq \beta_i^2 \leq \beta_{i+1}^2 \leq \dots \leq \beta_s^2 \leq 1, \quad i = 1, \dots, s, \end{aligned}$$

we have

$$(4.2) \quad 1 \geq \frac{\alpha_i^2}{2 - \alpha_i^2} \geq \frac{\alpha_{i+1}^2}{2 - \alpha_{i+1}^2} \geq \dots \geq \frac{\alpha_s^2}{2 - \alpha_s^2} \geq 0,$$

$$(4.3) \quad 0 \leq \frac{\beta_i^2}{2 - \beta_i^2} \leq \frac{\beta_{i+1}^2}{2 - \beta_{i+1}^2} \leq \dots \leq \frac{\beta_s^2}{2 - \beta_s^2} \leq 1.$$

Let $Q_1 = \text{diag}(1, 1, \dots, 1_s, 0_{s+1}, 0, \dots, 0_m) \in \mathbb{C}^{m \times m}$ and $\Sigma_{E_1} = \begin{pmatrix} \hat{\Sigma}_{E_1} \\ O_{(m-s) \times s} \end{pmatrix}$. Then by Theorem 3.1 we have

$$\begin{aligned} \max_{\Phi_1 \in \mathbb{U}_m} \det(E_1^H \Phi_1^H Q_1 \Phi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1}) &= \max_{\Phi_1 \in \mathbb{U}_m} \det(\Sigma_{E_1}^H U^H \Phi_1^H Q_1 \Phi_1 U \Sigma_{E_1}) \\ &= \max_{\phi_{11} \phi_{11}^H \leq I_s} \det(\hat{\Sigma}_{E_1}^H \phi_{11}^H I_s \phi_{11} \hat{\Sigma}_{E_1}) \\ &= \max_{\phi_{11} \phi_{11}^H = I_s} \det(\hat{\Sigma}_{E_2}^H \phi_{11}^H I_s \phi_{11} \hat{\Sigma}_{E_2}) \\ &= \alpha_1^2 \dots \alpha_s^2, \end{aligned}$$

where $\Phi_1 U = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \in \mathbb{U}_m$ with $\phi_{11} \in \mathbb{U}_s$. Let

$$K_1 = \text{diag}(0, 0, \dots, 0_{p-s}, 1_{p-s+1}, \dots, 1_p) \in \mathbb{C}^{p \times p},$$

$$K_2 = \text{diag}(0, 0, \dots, 0_{p-j}, 1_{p-j+1}, \dots, 1_p) \in \mathbb{C}^{p \times p},$$

$1 \leq j \leq s$, and

$$\Sigma_{E_2} = \begin{pmatrix} O_{(p-s) \times s} \\ \hat{\Sigma}_{E_2} \end{pmatrix}.$$

Hence, by Theorem 3.1 and (4.1) we have

$$\begin{aligned} & \min_{\Psi_2 \in \mathbb{U}_p} \det(2I_s - E_2^H \Psi_2^H K_1 \Psi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1}) \\ &= \min_{\Psi_2 \in \mathbb{U}_p} \det(2I_s - \Sigma_{E_2}^H V^H \Psi_2^H K_1 \Psi_2 V \Sigma_{E_2}) \\ &= \min_{\tilde{\phi}_{22} \tilde{\phi}_{22}^H \leq I_s} \det(2I_s - \hat{\Sigma}_{E_2}^H \tilde{\phi}_{22}^H I_s \tilde{\phi}_{22} \hat{\Sigma}_{E_2}) \\ &= \min_{\tilde{\phi}_{22} \tilde{\phi}_{22}^H = I_s} \det(2I_s - \hat{\Sigma}_{E_2}^H \tilde{\phi}_{22}^H I_s \tilde{\phi}_{22} \hat{\Sigma}_{E_2}) \\ &= (2 - \beta_1^2)(2 - \beta_2^2) \dots (2 - \beta_s^2), \end{aligned}$$

where $\Psi_2 V = \begin{pmatrix} \tilde{\phi}_{11} & \tilde{\phi}_{12} \\ \tilde{\phi}_{21} & \tilde{\phi}_{22} \end{pmatrix} \in \mathbb{U}_p$ with $\tilde{\phi}_{22} \in \mathbb{U}_s$. Similarly, we get

$$\begin{aligned} & \min_{\Psi_1 \in \mathbb{U}_m} \det(2I_s - E_1^H \Psi_1^H Q_1 \Psi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1}) = (2 - \alpha_1^2) \dots (2 - \alpha_s^2), \\ & \max_{\Phi_2 \in \mathbb{U}_p} \det(E_2^H \Phi_2^H K_1 \Phi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1}) = \beta_1^2 \beta_2^2 \dots \beta_s^2. \end{aligned}$$

Then by (4.2) we have

$$\begin{aligned} & \frac{\max_{\Phi_1 \in \mathbb{U}_m} \det(E_1^H \Phi_1^H Q_1 \Phi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1})}{\min_{\Psi_1 \in \mathbb{U}_m} \det(2I_s - E_1^H \Psi_1^H Q_1 \Psi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1})} \\ &= \frac{\alpha_1^2}{(2 - \alpha_1^2)} \frac{\alpha_2^2}{(2 - \alpha_2^2)} \dots \frac{\alpha_s^2}{(2 - \alpha_s^2)} \\ &\leq \alpha_i^{2(s-i+1)} \end{aligned}$$

for $i = 1, \dots, s$. It follows that

$$\begin{aligned} \alpha_i^2 &\geq \left(\frac{\max_{\Phi_1 \in \mathbb{U}_m} \det(E_1^H \Phi_1^H Q_1 \Phi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1})}{\min_{\Psi_1 \in \mathbb{U}_m} \det(2I_s - E_1^H \Psi_1^H Q_1 \Psi_1 E_1 (E_1^H E_1 + E_2^H E_2)^{-1})} \right)^{\frac{1}{s-i+1}} \\ (4.4) \quad &:= a_0 \end{aligned}$$

with $0 \leq a_0 \leq 1$ and

$$(4.5) \quad \beta_i^2 = 1 - \alpha_i^2 \leq 1 - a_0.$$

By (4.3) we get

$$\begin{aligned} & \frac{\max_{\Phi_2 \in \mathbb{U}_p} \det(E_2^H \Phi_2^H K_1 \Phi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1})}{\min_{\Psi_2 \in \mathbb{U}_p} \det(2I_s - E_2^H \Psi_2^H K_1 \Psi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1})} \\ &= \frac{\beta_1^2}{(2 - \beta_1^2)} \frac{\beta_2^2}{(2 - \beta_2^2)} \dots \frac{\beta_s^2}{(2 - \beta_s^2)} \\ &\leq \beta_i^{2i}, \end{aligned}$$

then

$$(4.6) \quad \beta_i^2 \geq \left(\frac{\max_{\Phi_2 \in \mathbb{U}_p} \det(E_2^H \Phi_2^H K_1 \Phi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1})}{\min_{\Psi_2 \in \mathbb{U}_p} \det(2I_s - E_2^H \Psi_2^H K_1 \Psi_2 E_2 (E_1^H E_1 + E_2^H E_2)^{-1})} \right)^{\frac{1}{i}} := b_0$$

with $0 \leq b_0 \leq 1$ and

$$(4.7) \quad \alpha_i^2 = 1 - \beta_i^2 \leq 1 - b_0.$$

By partitioning the matrices and using Strassen's matrix multiplication, computing a_0, b_0 costs roughly less than $O(s^2 \max\{m, p\})$. Therefore, (4.4)–(4.7) lead to

$$(4.8) \quad a_0 \leq \alpha_i^2 \leq 1 - b_0, \quad b_0 \leq \beta_i^2 \leq 1 - a_0, \quad i = 1, 2, \dots, s.$$

From (4.8) it follows that $\alpha_i \in [0, 1]$, $\beta_i \in [0, 1]$, which implies the upper and lower bounds of α_i and β_i are sharp.

5. Practical applications. In this section we give practical applications to report the performance of the proposed theoretical results, which are carried out in MATLAB (R2011b). We will apply the proposed results to the test signal of a mechanical system and aero engine fault diagnosis, respectively.

5.1. Implementation details. A static collision developing in an aero engine is a serious form of damage or failure. In order to analyze the fault, the data analysis needs to find out not only whether there is rubbing but also the time the rubbing occurs. For a rotor rubbing fault, the most basic features are the collision and friction. Hence, effective analysis for monitoring the cyclic impulse response of the rotor system is needed. For more details we refer the reader to [29, 30, 31, 32, 33, 34, 35, 36, 37]. In recent years, state-of-the-art methods for assessing the impact on components caused by rubbing include the wavelet transform method [29, 32, 36, 37], the generalized S transform method [27, 30, 35], and the noise reduction method based on singular values [28, 31, 33, 34]. In the following, we briefly introduce these methods.

5.1.1. The wavelet transform method. In recent years, with the development of wavelet theory, wavelet analysis has been widely used as a new type of signal processing method in many fields, such as fault diagnosis and information detection. The basic idea of multiresolution analysis in the wavelet transform method is to project the signal onto a subspace composed of a set of mutually orthogonal wavelet functions. It forms the expansion of the signal at different scales because its scale changes by binary dilation. The low-frequency signal is decomposed by the upper layer into two parts, of low frequency and high frequency. The high-frequency part is not decomposed further, so both the frequency resolution of the high-frequency band and the time resolution of the low-frequency band are poor. By multilevel division of the frequency band, the high-frequency part without subdivision is further decomposed and based on the analyzed signal; see [29, 32, 36, 37].

5.1.2. The generalized S transform method. The generalized S transform method is a reversible time-frequency localization analysis method and has been applied in processing seismic signals, power quality disturbance signals, and the collision diagnostics of aero engines. The generalized S transform is used to change the radial component of the rub-impact signal in the phase space. The rub-impact characteristics are observed and extracted in the phase space.

5.1.3. The noise reduction method based on singular values. For signal reconstruction we obtain signals with different band energies in different directions in the reconstruction space. The mutation information has different effects on singular values. Phase space reconstruction based on singular values is performed to obtain the signal component reflecting the mutation information. This method uses some improvements in the aero engine processing field. The signal is more complicated, which is unfavorable for extracting information on the local mutation. The signal reconstructed by the singular value feature can be completed. To get some information about the fault signal, it is separated from the background signal. For details we refer the reader to [28, 31, 33, 34, 37, 38].

5.2. Test signal of the mechanical system. For fault diagnosis, assume that the test signal of the mechanical system is the following numerical sequence: x_i , $i = 1, 2, \dots, 2n - 1$. Then the available intersegment attractor reconstruction matrix D is

$$D = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \vdots & \vdots & & \vdots \\ x_n & x_{n+1} & \cdots & x_{2n-1} \end{pmatrix};$$

see [33, 34]. Assuming that the device is operating normally, we denote the sensor's test signal by $x(t)$, which involves a deterministic signal $s(t)$; Gaussian white noise is denoted by $n(t)$. Then $x(t) = s(t) + n(t)$, where signal $s(t)$ is not correlated with noise $n(t)$. Here the reconstruction matrix D_x of test signal $x(t)$ is the superposition of the reconstruction matrix D_s of signal $s(t)$ and the reconstruction matrix D_n of noise $n(t)$, i.e., $D_x = D_s + D_n$. Further, suppose after some time the new test signal $y(t)$ adds a new fault signal, such as impact component $d(t)$; then, $y(t) = s(t) + n(t) + d(t)$. The corresponding reconstruction matrix D_y of the new test signal $y(t)$ is $D_y = D_s + D_d + D_n$, where D_d is the reconstruction matrix of $d(t)$ and $D_d = D_y - D_x$ with D_x , D_y being tested; for details see [30, 31]. We enumerate the process of fault friction mutation information as shown in Figure 5.1. The energy of the fault signal $d(t)$ can be estimated by using the singular value sum of matrix D_d . We denote by N and by $Thr = \sum_{i=1}^n f^2(m_i)$ the number of samples and the threshold, respectively, where f is the signal time-frequency window function before the fault and m_i is the band.

We reconstruct matrix

$$C = D_x + D_d + \sum_{k=1}^n \chi_k e_k e_k^T + \sum_{k=1}^n \tilde{\chi}_k e_k e_k^T,$$

where e_k is the k th column of the identity matrix, and χ_k and $\tilde{\chi}_k$ are the optimal upper bounds of $\sum_{i=1}^k \sigma_i(D_d)$ and $\sum_{i=1}^k \sigma_i(D_x)$. The process of reconstructing the decomposition signal x' is

$$x'(i) = \begin{cases} \sum_{j=1}^i C(j, i-j+1)/i & \text{if } 1 \leq i \leq N, \\ \sum_{j=1}^i C(N-j+1, i-N+j)/(2N-i) & \text{if } N < i \leq 2N-1. \end{cases}$$

The visible oscillation mainly appears in the end portion. During data processing the endpoint part can be extended appropriately, and the extended part is discarded once the processing is completed. The characteristic frequency of the rub-impact signal can be reproduced from the reconstructed decomposition signals.

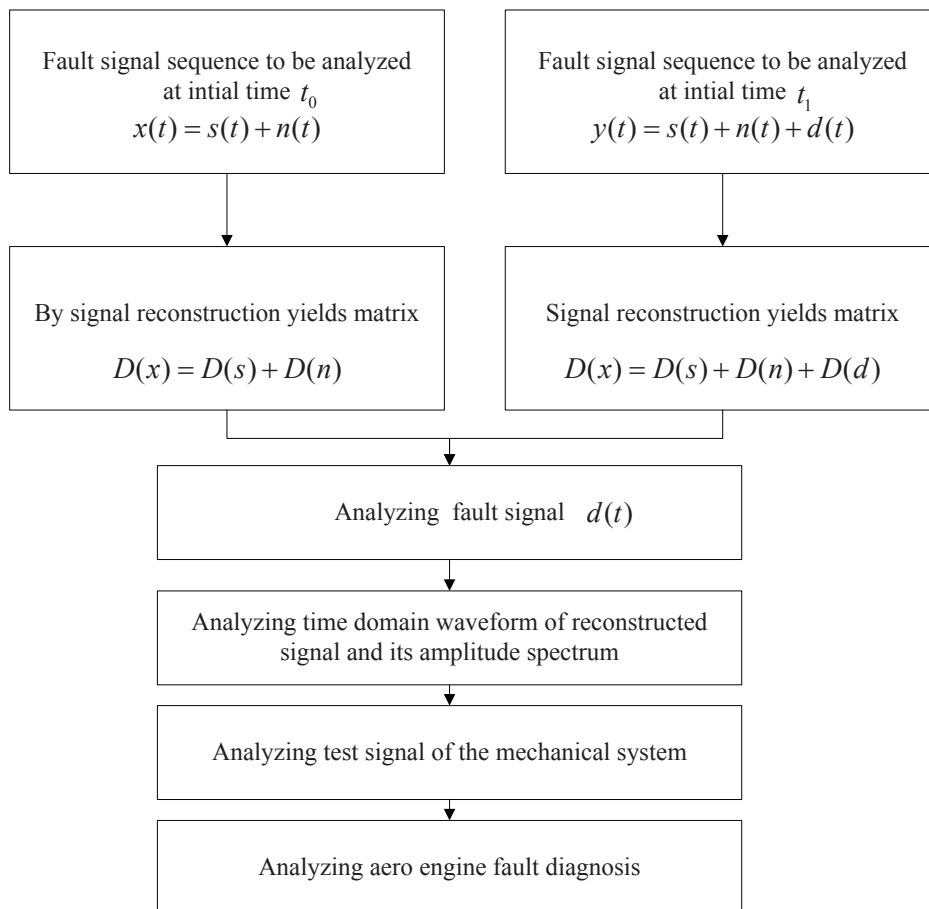


FIG. 5.1. Outline of extraction process of abrupt signal and aero engine fault diagnosis.

5.2.1. The *NRSVE* and *NRSVF* methods. Assume that $B \in \mathbb{C}^{n \times n}$ is Hermitian positive semidefinite and

$$A_j = \text{diag}(1, \dots, 1_{j-1}, 1_j, 0, \dots, 0) \in \mathbb{C}^{n \times n}, \quad j = 1, \dots, n.$$

We set $\text{tr}(BA_j) = c(B, A_j)$. It is easy to check that each diagonal element of BA_j is nonnegative and $\text{tr}(BA_j) \geq 0$. By Theorem 3.5 we give the following analysis.

(i) It follows from the first equality of Theorem 3.5(1) that

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) \\
 &= \max_{U_1, U_2 \in \mathbb{U}_n} \Re \left[\text{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right] - nc \\
 &\geq \Re[\text{tr}(cI_n + BA_j)] - nc \\
 (5.1) \quad &= nc + \text{tr}(BA_j) - nc = \text{tr}(BA_j)
 \end{aligned}$$

and

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) \\
 &= \max_{U_1, U_2 \in \mathbb{U}_n} \Re \left[\operatorname{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right] - nc \\
 &\geq \Re[\operatorname{tr}(cI_n - BA_j)] - nc \\
 (5.2) \quad &= nc - \operatorname{tr}(BA_j) - nc = -\operatorname{tr}(BA_j).
 \end{aligned}$$

Since $\operatorname{tr}(BA_j) \geq 0$, by (5.1) and (5.2) we have

$$\sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) = \sum_{i=1}^j \sigma_i(B) \geq \operatorname{tr}(BA_j).$$

(ii) From the second equality of Theorem 3.5(1) we have that if $c \geq 0$, then

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - nc \\
 &\geq \Re[\operatorname{tr}(cI_n + BA_j)] - nc \\
 (5.3) \quad &= nc + \operatorname{tr}(BA_j) - nc = \operatorname{tr}(BA_j).
 \end{aligned}$$

If $c < 0$ and $\operatorname{tr}(BA_j) \geq n|c|$, then

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - n|c| \\
 &\geq |\operatorname{tr}(cI_n + BA_j)| - n|c| \\
 (5.4) \quad &= \operatorname{tr}(BA_j) - 2n|c|.
 \end{aligned}$$

If $c < 0$ and $\operatorname{tr}(BA_j) \leq n|c|$, then

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - n|c| \\
 &\geq |\operatorname{tr}(cI_n + BA_j)| - n|c| \\
 (5.5) \quad &= n|c| - n|c| - \operatorname{tr}(BA_j) = -\operatorname{tr}(BA_j).
 \end{aligned}$$

Similarly, if $c \leq 0$, then

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - n|c| \\
 &\geq |\operatorname{tr}(cI_n - BA_j)| - n|c| \\
 (5.6) \quad &= n|c| + \operatorname{tr}(BA_j) - n|c| = \operatorname{tr}(BA_j).
 \end{aligned}$$

If $c > 0$ and $\operatorname{tr}(BA_j) \geq n|c|$, then

$$\begin{aligned}
 \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - n|c| \\
 &\geq |\operatorname{tr}(cI_n - BA_j)| - n|c| \\
 (5.7) \quad &= \operatorname{tr}(BA_j) - 2n|c|.
 \end{aligned}$$

If $c > 0$ and $\text{tr}(BA_j) \leq n|c|$, then

$$\begin{aligned} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = \max_{U_1, U_2 \in \mathbb{U}_n} \left| \text{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right| - n|c| \\ &\geq |\text{tr}(cI_n - BA_j)| - n|c| \\ &= n|c| - n|c| - \text{tr}(BA_j) = -\text{tr}(BA_j). \end{aligned} \quad (5.8)$$

By (5.3)–(5.8) we have $\sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) = \sum_{i=1}^j \sigma_i(B) \geq \text{tr}(BA_j)$.

(iii) It follows from the first equality of Theorem 3.5(2) that

$$\begin{aligned} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = nc - \min_{U_1, U_2 \in \mathbb{U}_n} \Re \left[\text{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right] \\ &\geq nc - \Re[\text{tr}(cI_n + BA_j)] \\ &= nc - \text{tr}(BA_j) - nc = -\text{tr}(BA_j), \end{aligned} \quad (5.9)$$

and similarly,

$$\begin{aligned} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = nc - \min_{U_1, U_2 \in \mathbb{U}_n} \Re \left[\text{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right] \\ &\geq nc - \Re[\text{tr}(cI_n - BA_j)] \\ &= nc + \text{tr}(BA_j) - nc = \text{tr}(BA_j). \end{aligned} \quad (5.10)$$

By (5.9) and (5.10) we have $\sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) = \sum_{i=1}^j \sigma_i(B) \geq \text{tr}(BA_j)$.

(iv) By Theorem 3.5(2) we first set $c = \text{tr}(B) \geq 0$ or $c = \|B\|_2$ (or a larger number). Then we have $\frac{1}{n} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) \leq |c|$. It follows that

$$\begin{aligned} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = n|c| - \min_{U_1, U_2 \in \mathbb{U}_n} \left| \text{tr} \left(cI_n + \prod_{j=1}^m U_1 B U_2 A_j \right) \right| \\ &\geq n|c| - |\text{tr}(cI_n + BA_j)| \\ &= n|c| - \text{tr}(BA_j) - n|c| = -\text{tr}(BA_j), \end{aligned} \quad (5.11)$$

and similarly,

$$\begin{aligned} \sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) &= \sum_{i=1}^j \sigma_i(B) = n|c| - \min_{U_1, U_2 \in \mathbb{U}_n} \left| \text{tr} \left(cI_n - \prod_{j=1}^m U_1 B U_2 A_j \right) \right| \\ &\geq n|c| - \left| \text{tr} \left(cI_n - \prod_{j=1}^m BA_j \right) \right| \\ &= n|c| + \text{tr}(BA_j) - n|c| = \text{tr}(BA_j). \end{aligned} \quad (5.12)$$

By (5.11) and (5.12) we also have $\sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) = \sum_{i=1}^j \sigma_i(B) \geq \text{tr}(BA_j)$.

Therefore, by (i)–(iv) we can conclude $\sum_{i=1}^n \prod_{j=1}^m \sigma_i(B) \sigma_i(A_j) = \sum_{i=1}^j \sigma_i(B) \geq \text{tr}(BA_j) = c(B, A_j)$, which implies that

$$\sigma_j(B) \leq \text{tr}(B) - c(B, A_{j-1}). \quad (5.13)$$

We set A_0 to be a zero matrix here. In (5.13), by partitioning the matrices and using Strassen's matrix multiplication, computing $c(B, A_{j-1})$ costs roughly less than $O((j-1)n^2)$. We also note that, for arbitrary matrix $\mathcal{D} \in \mathbb{C}^{n \times n}$, we could first get $B = \mathcal{D}^H \mathcal{D}$ by \mathcal{D} . Then, by the above analysis and the fact that $\sigma_j(\mathcal{D}) = \sqrt{\sigma_j(B)}$, we could estimate the bounds of $\sigma_j(\mathcal{D})$.

The following bound can be found in [31]: for $k = 1, \dots, n$,

$$\sum_{j=1}^k (\sigma(D_d)_j + \sigma(D_x)_j) \leq k[(\text{tr}(D_d^H D_d))^{\frac{1}{2}} + (\text{tr}(D_x^H D_x))^{\frac{1}{2}}] := \mathbf{E}(k) = \mathbf{E}.$$

Let N be determined by $\mathbf{E}(N) \leq \sqrt{n}Thr$ (see [34, 35]). By (5.13) it is easy to see that, for $k = 1, \dots, n$,

$$\sum_{j=1}^k (\sigma(D_d)_j + \sigma(D_x)_j) \leq \sum_{j=1}^k [(\text{tr}(D_d^H D_d) + \varphi_j)^{\frac{1}{2}} + (\text{tr}(D_x^H D_x) + \tilde{\varphi}_j)^{\frac{1}{2}}] := \mathbf{F}(k) = \mathbf{F},$$

where $\varphi_j = -c(D_d^H D_d, A_{j-1}) \leq 0$ and $\tilde{\varphi}_j = -c(D_x^H D_x, A_{j-1}) \leq 0$. N can be determined by $\mathbf{F}(N) \leq \sqrt{n}Thr$. It is easy to check that, for any unitary matrices U, V , it holds that

$$\sum_{j=1}^k [(\text{tr}(D_d^H D_d) + \varphi_j)^{\frac{1}{2}} + (\text{tr}(D_x^H D_x) + \tilde{\varphi}_j)^{\frac{1}{2}}] \leq k[(\text{tr}(D_d^H D_d))^{\frac{1}{2}} + (\text{tr}(D_x^H D_x))^{\frac{1}{2}}],$$

i.e., $\mathbf{F} \leq \mathbf{E}$. Hence, in our numerical performance we set $U = V = I_n$ for simple computations. In an intermediate step of the noise reduction method, which is called the *NRSVE* (where "NRSV" stands for "noise reduction by singular values") method for convenience, \mathbf{E} can usually be used as an estimate bound.

Here we use \mathbf{F} as an estimate bound in the intermediate step in the noise reduction method called the *NRSVF* method. In the following we apply our analysis as simulation analysis in aero engine fault diagnosis.

5.3. Testing problems. We consider the following simulation signals frequently used in aero engine fault diagnosis; see [37, 38]. At initial time t_0 the test signal is described by

$$s_0(t) = s_1 + \sigma e(t).$$

After t time, the test signal with engine fault is described by

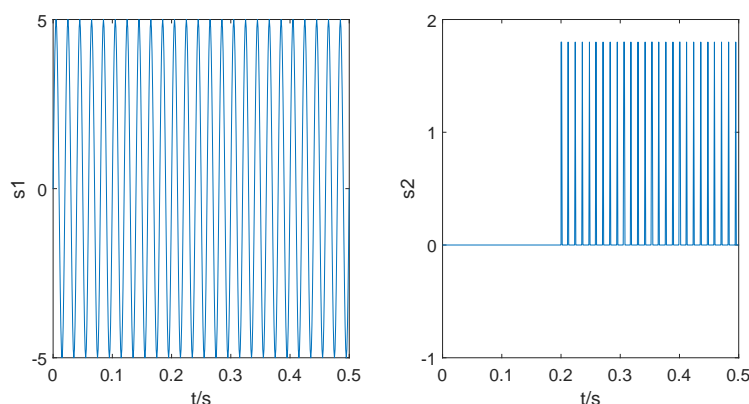
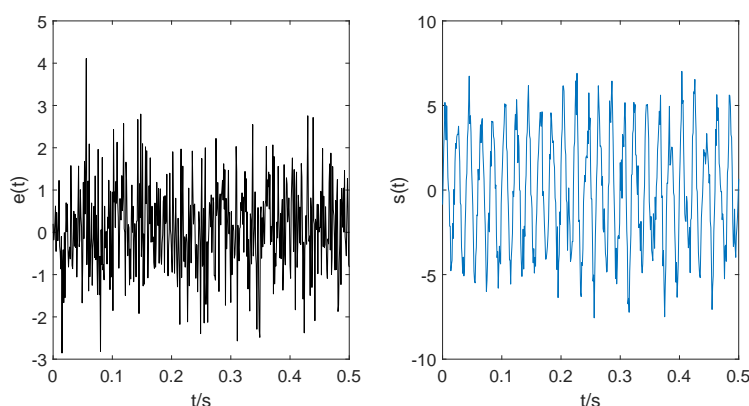
$$s(t) = s_1 + s_2 + \sigma e(t),$$

where

$$s_1 = 5 \sin(2\pi\omega_1 t), \quad \omega_1 = 50 \text{ Hz},$$

and s_2 is the half-sinusoidal pulse sequence that begins to occur after 0.2 seconds with amplitude 1.8 and frequency 85 Hz, as shown in Figure 5.2. $\sigma e(t)$ is Gaussian white noise with a mean of 0 and variance of 1 with $\sigma = 1$ as shown on the left in Figure 5.3. The three superimposed signals are shown on the right in Figure 5.3.

5.3.1. Simulation analysis in aero engine fault diagnosis. We will give simulation analysis in aero engine fault diagnosis by the above numerical methods.

FIG. 5.2. Signals of s_1 (left) and s_2 (right).FIG. 5.3. Signals of $e(t)$ (left) and $s(t)$ (right).

The wavelet transform method. Using the wavelet transform method we perform $J = 3$ layers of db5 wavelet packet decomposition on the original signal, and perform threshold denoising on the signal. Then we select the group with the largest energy, perform the Fourier transform on the decomposition coefficient and decompose the spectrum after wavelet packet decomposition. This involves the following parameters: the mean $\varrho_1 = 3$ and the variance $\varrho_2 = 5$, the threshold $Thr = \varrho_1 + \varrho_2 = 8$, multiplier $X = 100.2 \text{ Hz}$, and the spectrum of the best coefficient vector of the signal is shown in Figure 5.4. The corresponding reconstructed signal and amplitude spectra are shown in Figures 5.9 and 5.11.

The generalized S transform method. By the generalized S transform method we have the energy decay rate $\alpha = 8$, the energy delay time $\beta = 10$, the time shifting parameter $\tau = 5$, and the signal-to-noise ratio is 13 dB. The radial component of the vibration response is shown in Figure 5.5. The generalized S transform of (5.2) is used to decompose the rubbing signal, and a representation of the signal in phase space can be obtained. The result of decomposing the impact-rub signal is shown in Figure 5.6, which reflects its time-frequency local characteristics. The corresponding reconstructed signal and amplitude spectrum are shown in Figures 5.9 and 5.11.

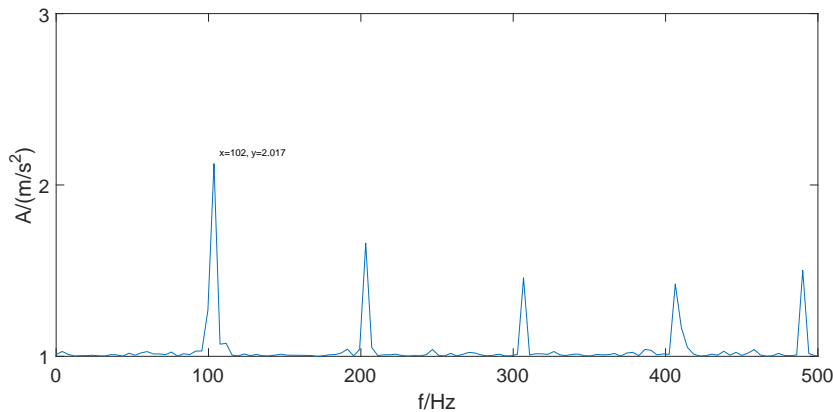


FIG. 5.4. Spectrum of the best coefficient vector of the signal.

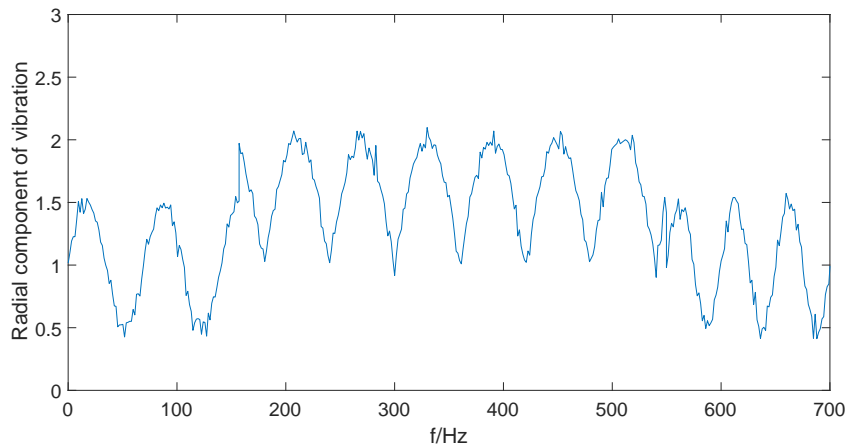
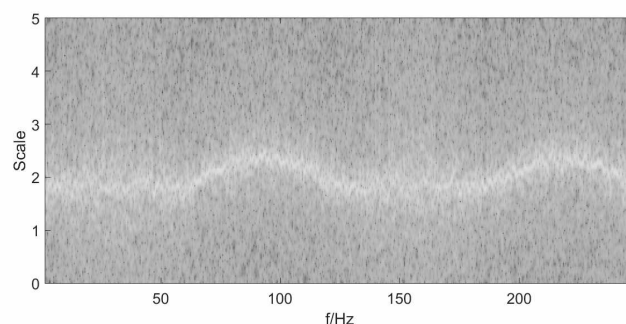


FIG. 5.5. Radial component of vibration.

The NRSVE and NRSVF methods. We set $\chi_k + \tilde{\chi}_k = \mathbf{F}$ in the *NRSVF* method and $\chi_k + \tilde{\chi}_k = \mathbf{E}$ in the *NRSVE* method. The reconstructed decomposition signals and amplitude spectrum obtained by the *NRSVF* and *NRSVE* methods, respectively, are shown in Figures 5.7 and 5.9.

The different mutation information abilities of the *NRSVE* method, the *NRSVF* method, the wavelet transform method, and the generalized *S* transform method are shown in Figures 5.7–5.10, respectively. In the following subsection we compare their performance numerically in terms of the time domain waveform and the amplitude spectrum of the mutation signal.

5.4. Performance comparisons. In this subsection, we compare the numerical efficiency of the *NRSVF*, *NRSVE*, wavelet transform, and generalized *S* transform methods. As we know, in aero engine fault diagnosis, the first mutation time and the amplitude spectrum are important mutation information extraction factors, because

FIG. 5.6. *Decomposition result of the impact-rub signal.*TABLE 5.1
Comparisons of the CPU time, the first mutation time, and the amplitude spectrum.

N	Method	CPU time	First mutation time	Amplitude spectrum
160	$NRSV\mathbf{F}$	81.1 s	0.208 s	85.4 Hz
160	$NRSV\mathbf{E}$	81.5 s	0.269 s	100.3 Hz
160	Wavelet transform	80.5 s	0.272 s	100.2 Hz
160	Generalized S transform	83.7 s	0.241 s	101.0 Hz
280	$NRSV\mathbf{F}$	83.8 s	0.205 s	86.1 Hz
280	$NRSV\mathbf{E}$	83.9 s	0.273 s	101.1 Hz
280	Wavelet transform	83.2 s	0.281 s	106.8 Hz
280	Generalized S transform	86.4 s	0.255 s	104.9 Hz
420	$NRSV\mathbf{F}$	92.6 s	0.212 s	85.5 Hz
420	$NRSV\mathbf{E}$	92.2 s	0.269 s	102.3 Hz
420	Wavelet transform	90.5 s	0.277 s	104.7 Hz
420	Generalized S transform	95.2 s	0.259 s	107.8 Hz

if the time of fault diagnosis can be advanced, it may play a major role in preventing aviation accidents. The amplitude spectrum is also helpful for understanding the fault signal. Therefore, the first mutation time and the amplitude spectrum of fault diagnosis and mutation information are very important. In addition, calculation time is related to the complexity of the computation methods, which is also considered. In the following we compare the first mutation time, the amplitude spectrum, and the CPU time of these methods. Here the CPU time denotes the time from deriving test signals $s_0(t)$ and $s(t)$ to the time the fault signal is reconstructed. Numerical performance comparisons are given in Table 5.1.

We now give an error formula for the relative error analysis of the first mutation time and the amplitude spectrum in order to compare the accuracy. Define

$$Error_1 = \frac{|M_v - E_v|}{|M_v|}, \quad Error_2 = \frac{|M_v - E_v|}{|E_v|},$$

where $Error_1$ and $Error_2$ are two kinds of relative errors and M_v and E_v denote the measured value and reference exact value, respectively. As mentioned above, the

TABLE 5.2

Comparisons of relative error analysis for the first mutation time.

N	Method	R_{error_1}	R_{error_2}
160	$NRSV\mathbf{F}$	0.0244	0.025
160	$NRSV\mathbf{E}$	0.2453	0.3250
160	Wavelet transform	0.2647	0.3600
160	Generalized S transform	0.1701	0.2050
280	$NRSV\mathbf{F}$	0.0244	0.025
280	$NRSV\mathbf{E}$	0.2674	0.3650
280	Wavelet transform	0.2883	0.4050
280	Generalized S transform	0.2157	0.2750
420	$NRSV\mathbf{F}$	0.0566	0.0600
420	$NRSV\mathbf{E}$	0.2565	0.3450
420	Wavelet transform	0.2780	0.3850
420	Generalized S transform	0.2278	0.2950

TABLE 5.3

Comparisons of relative error analysis for the amplitude spectrum.

N	Method	R_{error_1}	R_{error_2}
160	$NRSV\mathbf{F}$	0.0023	0.0024
160	$NRSV\mathbf{E}$	0.1508	0.1776
160	Wavelet transform	0.1517	0.1788
160	Generalized S transform	0.1584	0.1882
280	$NRSV\mathbf{F}$	0.0128	0.0129
280	$NRSV\mathbf{E}$	0.1592	0.1894
280	Wavelet transform	0.2041	0.2565
280	Generalized S transform	0.1897	0.2341
420	$NRSV\mathbf{F}$	0.0058	0.0059
420	$NRSV\mathbf{E}$	0.1691	0.2035
420	Wavelet transform	0.1882	0.2318
420	Generalized S transform	0.2115	0.2682

exact values of the first mutation time and the amplitude spectrum are 0.2s and 85 Hz.

Let N be the number of samples. Then the decision coefficient R^2 and the predicting root mean square error $x_{RMSE,p}$ are defined by (see [31, 32, 38])

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{\vartheta}_i - \vartheta_i)^2}{\sum_{i=1}^N (\vartheta_i - \bar{\vartheta})^2}, \quad x_{RMSE,p} = \sqrt{\frac{\sum_{i=1}^N (\hat{\vartheta}_i - \vartheta_i)^2}{N-1}},$$

respectively, where $\hat{\vartheta}_i$, $i = 1, 2, \dots, N$, is the predicted value of the sample, $\bar{\vartheta}$ is the mean of ϑ_i , and ϑ_i is the corresponding accurate value. The scope of the decision coefficient R^2 is from 0 to 1. If R^2 is closer to 1, the method or model fit is higher. Predicted root mean square error $x_{RMSE,p}$ is an evaluation of all standard deviations of the difference between the predicted value of the sample and the reference exact value. It represents the overall residual error. In general, it will be better when its value is smaller.

From Figures 5.7–5.10 and Tables 5.1–5.3 we conclude the following results.

- The first mutation time of mutation information at “0.2s” and the amplitude spectrum at “85 Hz” are not only detected effectively, but also show that the

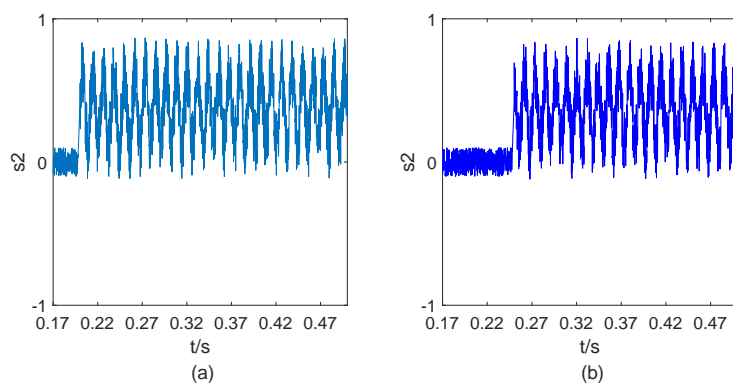


FIG. 5.7. Signal reconstructed by (a) the $NRSVF$ method and (b) the $NRSVE$ method.

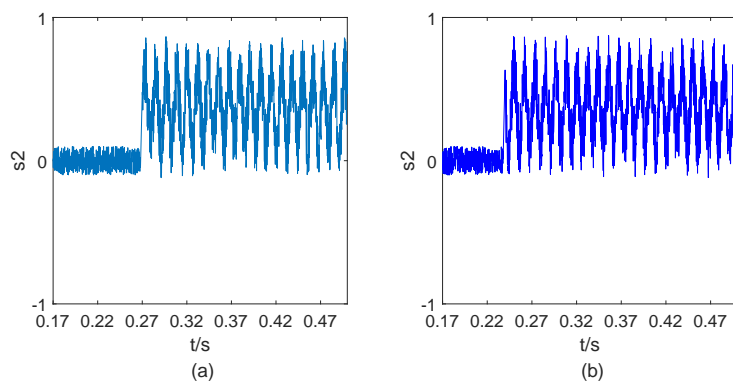


FIG. 5.8. Signal reconstructed by (a) the wavelet transform method and (b) the generalized S transform method.

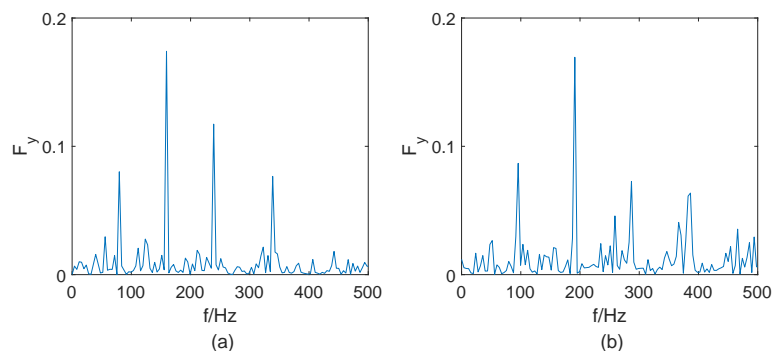


FIG. 5.9. Amplitude spectrum reconstructed by (a) the $NRSVF$ method and (b) the $NRSVE$ method.

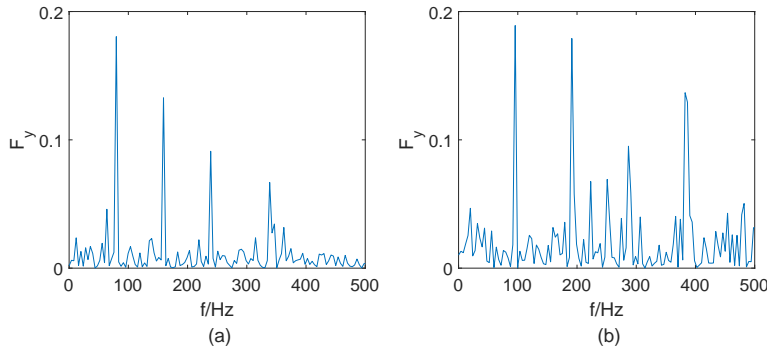


FIG. 5.10. Amplitude spectrum reconstructed by (a) the wavelet transform method and (b) the generalized S transform method.

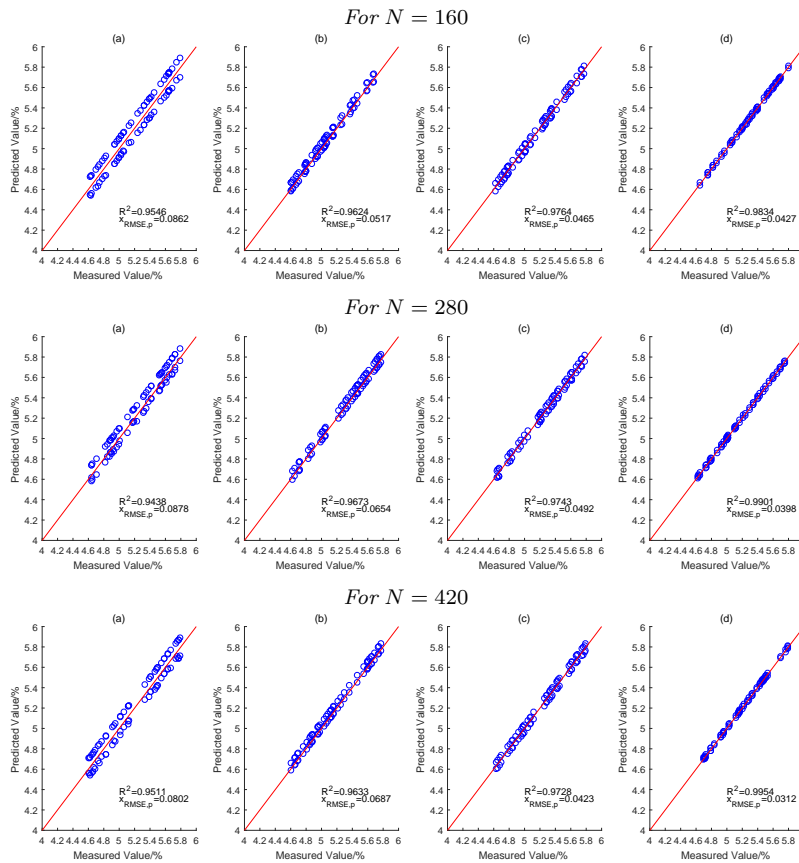


FIG. 5.11. Comparisons of the first failure time and amplitude frequency prediction results given by our four methods with the actual analysis results: (a) the wavelet transform method; (b) the generalized S transform method; (c) the NRSVE method; (d) the NRSVF method.

shape of the variable information curve is very close to that of the original half sine pulse sequence.

- The time domain waveform and amplitude spectrum of the reconstructed signal show that the smooth background signal is basically eliminated. Half the effect of a sinusoidal pulse is small.
- For the first mutation time and the amplitude spectrum, the *NRSVF* method is more efficient than the other three methods. For the CPU time, the methods give almost the same results but the wavelet transform method is a little more efficient. However, the difference in computational cost for the four methods is not very big, and computational cost of the methods takes almost the same CPU time. As mentioned above, the most important issue in aero engine fault diagnosis is to perform fault diagnosis faster. The important factors in aero engine fault diagnosis are the first mutation time and the amplitude spectrum. Hence, by using the *NRSVF* method we could test aero engine faults faster.

The *NRSVF* method is better at mutation information extraction than some previous methods; it is more efficient for aero engine fault diagnosis and could prevent faults earlier. It can be seen from Figure 5.11, the performance of the proposed *NRSVF* method is more efficient. R^2 increases from 0.9546 to 0.9834 for $N = 160$, 0.9438 to 0.9901 for $N = 280$, and 0.9511 to 0.9954 for $N = 420$. $x_{RMSE,p}$ decreases from 0.0862 to 0.0427 for $N = 160$, 0.0878 to 0.0398 for $N = 280$, and 0.0802 to 0.0312 for $N = 420$. This effectively improves predictive accuracy and method validity.

6. Concluding remarks. In this paper we give the analytic solutions of the following constrained matrix determinant and trace minimization and maximization problems:

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|, \quad \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \det \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|$$

and

$$\min_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|, \quad \max_{U_1, \dots, U_m \in \mathbb{U}_n} \left| \operatorname{tr} \left(cI_n \pm \prod_{j=1}^m A_j U_j \right) \right|,$$

respectively, where $c \in \mathbb{R}$ is a real number, A_1, \dots, A_m are $n \times n$ complex matrices, I_n is an $n \times n$ identity matrix, and U_1, \dots, U_m are $n \times n$ unitary matrices. The main results improve on the corresponding existing results in [2, 4, 5]. In particular, our results can be applied to some practical applications, such as the test signal of mechanical systems and aero engine fault diagnosis.

Acknowledgments. The authors would like to thank the associate editor, anonymous referees, and Professor Weiwei Sun for their valuable comments, which improved the presentation of the paper greatly.

REFERENCES

- [1] J. MUNKRES, *Topology: Pearson New International Edition*, 2nd ed., Pearson, London, 2014.
- [2] A. W. MARSHALL, I. OLKIN, AND B. C. ARNOLD, *Inequalities: Theory of Majorization and Its Applications*, 2nd ed., Springer Ser. Statist., Springer, New York, 2009.
- [3] R. BHATIA, *Matrix Analysis*, Springer, New York, 1997.
- [4] Q. K. LU, *The elliptic geometry of extended spaces*, Acta Math. Sinica, 13 (1963), pp. 49–62.

- [5] J. G. SUN, *Perturbation analysis for the generalized singular value problem*, SIAM J. Matrix Anal. Appl., 20 (1983), pp. 611–625.
- [6] O. ALTER, P. O. BROWN, AND D. BOTSTEIN, *Processing and modeling genome-wide expression data using singular value decomposition*, in Proceedings of BIOS 2001, Microarrays: Optical Technologies and Informatics, Proc. SPIE 4266, SPIE Press, Bellingham, WA, 2001, pp. 171–186.
- [7] O. ALTER, P. O. BROWN, AND D. BOTSTEIN, *Singular value decomposition for genome-wide expression data processing and modeling*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 10101–10106.
- [8] O. ALTER, P. O. BROWN, AND D. BOTSTEIN, *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 3351–3356.
- [9] O. ALTER, G. H. GOLUB, P. O. BROWN, AND D. BOTSTEIN, *Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models*, in The Cell Cycle, Chromosomes and Cancer, Proc. Miami Nature Biotech. Winter Symp. 15, M. P. Deutscher et al., eds., Miller School of Medicine, University of Miami, Miami, FL, 2004.
- [10] Z. BAI AND J. W. DEMMEL, *Computing the generalized singular value decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 1464–1486.
- [11] H. KIM, G. H. GOLUB, AND H. PARK, *Missing value estimation for DNA microarray gene expression data: Local least squares imputation*, Bioinformatics, 21 (2005), pp. 187–198.
- [12] W. W. XU, H. K. PANG, W. LI, X. P. HUANG, AND W. J. SUN, *On the explicit expression of chordal metric between generalized singular values of Grassmann matrix pairs with applications*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1547–1563.
- [13] W. M. MIAO, S. H. PAN, AND D. F. SUN, *A rank-corrected procedure for matrix completion with fixed basis coefficients*, Math. Program., 159 (2016), pp. 289–338.
- [14] C. DING, D. F. SUN, AND K. C. TOH, *An introduction to a class of matrix cone programming*, Math. Program., 144 (2014), pp. 141–179.
- [15] M. FAZEL, T. K. PONG, D. F. SUN, AND P. TSENG, *Hankel matrix rank minimization with applications to system identification and realization*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 946–977.
- [16] D. F. SUN AND J. SUN, *Löwner’s operator and spectral functions in Euclidean Jordan algebras*, Math. Oper. Res., 33 (2008), pp. 421–445.
- [17] H. QI AND X. M. YUAN, *Computing the nearest Euclidean distance matrix with low embedding dimensions*, Math. Program., 147 (2014), pp. 351–389.
- [18] B. S. HE, M. H. XU, AND X. M. YUAN, *Solving large-scale least squares semidefinite programming by alternating direction methods*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 136–152.
- [19] X. J. CHEN AND S. H. XIANG, *Perturbation bounds of P-matrix linear complementarity problems*, SIAM J. Optim., 18 (2008), pp. 1250–1265.
- [20] X. J. CHEN, Z. S. LU, AND T. K. PONG, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492.
- [21] C. BERGE, *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*, Macmillan, London, 1963.
- [22] Y. B. HUA AND T. K. SARKAR, *On SVD for estimating generalized eigenvalues of singular matrix pencil in noise*, IEEE Trans. Signal Process., 39 (1991), pp. 892–900.
- [23] Z. J. BAI AND H. ZHA, *A new preprocessing algorithm for the computation of the generalized singular value decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 1007–1012.
- [24] Z. J. BAI AND J. W. DEMMEL, *Computing the generalized singular value decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 1464–1486.
- [25] X. S. CHEN AND W. LI, *A note on backward error analysis of the generalized singular value decomposition*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1358–1370.
- [26] R. C. LI, *Bounds on perturbations of generalized singular values and of associated subspaces*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 195–234.
- [27] Y. I. PORTNYAGIN, E. G. MERZLYAKOV, C. H. JACOBI, N. J. MITCHELL, H. G. MULLER, A. H. MANSION, W. SINGER, P. HOFFMANN, AND A. N. FACHRUTDINOVA., *Some results of S-transform analysis of the transient planetary-scale wind oscillations in the lower thermosphere*, Earth Planets Space, 51 (1999), pp. 711–717.
- [28] W. X. YANG AND P. W. TSE, *Development of an advanced noise reduction method for vibration analysis based on singular value decomposition*, NDT&E Internat., 36 (2003) pp. 419–432.
- [29] J. LIN AND L. S. QU, *Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis*, J. Sound Vib., 234 (2000), pp. 135–147.
- [30] L. T. YAN AND D. Y. WANG, *Vibration features from rubbing between rotor and casing for a*

- dual-shaft aeroengine*, J. Aerosp. Power, 13 (2) (1998), pp. 173–176.
- [31] P. P. KANJILAL, S. PALIT, AND G. SAHA, *Fetal ECG extraction from single-channel maternal ECG using singular value decomposition*, IEEE Trans. Biomed. Eng., 44 (1997), pp. 51–59.
- [32] S. BHUNIA AND K. ROY, *A novel wavelet transform-based transient current analysis for fault detection and localization*, IEEE Trans. Very Large Scale Integr. (VLSI) Syst., 13 (2005), pp. 503–507.
- [33] H. TIAN, X. D. LIU, AND Q. H. LI, *Improved fault diagnosis method for aero engine rotor-stator [sic] rubs*, J. Aerosp. Power, 23 (2008), pp. 1093–1097.
- [34] H. TIAN, X. D. LIU, Y. N. CHEN, AND Q. H. LI, *Method for diagnosing rub fault of rotor-stator based on differences of singularly values*, J. Aerosp. Power, 24 (2009), pp. 2296–2301.
- [35] J. Z. SUN, H. F. ZUO, P. P. LIU, AND L. ZHU, *A method of condition monitoring and on-wing life prediction for civil aviation aircraft engine based on dynamic linear model*, Syst. Eng. Theory Practice, 33 (2013), pp. 3243–3250.
- [36] L. ZHU, H. F. ZUO, AND J. CAI, *Performance reliability prediction for civil aviation aircraft engine based on Wiener process*, J. Aerosp. Power, 28 (2013), pp. 1006–1012.
- [37] L. HAI, J. HONG, AND D. WANG, *Fault diagnosis of aero-engine bearings based on wavelet package analysis*, J. Propulsion Technol., 3 (2009), pp. 328–341.
- [38] L. ZHU AND H. F. ZUO, *Predicting compressor of gas turbine power plant on-line washing interval using proportional hazards model*, Adv. Mater. Res., 6 (2012), pp. 195–199.