ELSEVIER Linear Algebra and its Applications 280 (1998) 267–287

LINEAR ALGEBRA AND ITS APPLICATIONS

Perturbation theory for the Eckart-Young-Mirsky theorem and the constrained total least squares problem [1]

Musheng Wei

Department of Mathematics, East China Normal University, Shanghai 200062, People's Republic of China

Received 2 July 1991; received in revised form 4 February 1998; accepted 12 February 1998

Submitted by G.H. Golub

Abstract

Golub et al. (Linear Algebra Appl. 88/89 (1987) 317–327), J.Demmel (SIAM J. Numer. Anal. 24 (1987) 199–206), generalized the Eckart-Young-Mirsky (EYM) theorem, which solves the problem of approximating a matrix by one of lower rank with only a specific rectangular subset of the matrix allowed to be changed. Based on their results, this paper presents perturbation analysis for the EYM theorem and the constrained total least squares problem (CTLS). © 1998 Elsevier Science Inc. All rights reserved.

AMS classification: 65F20; 15A18; 15A12

Keywords: Singular value; Least squares; Rank deficient; Smallest perturbation

1. Introduction

Let

$$G_0 = D, \quad G_1 = (C, D), \quad G_2 = \begin{pmatrix} B \\ D \end{pmatrix}, \quad G_3 = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

[1] This work was supported by the National Natural Sciences Foundation, People's Republic of China.

PII: S0024-3795(98)10018-6

be the partitioned rectangular matrices. Eckart, Young and Mirsky (see the Ref. in [1]) solved the problem of finding the perturbation δD of D with smallest Frobenius or two norm which reduces the rank of $G_0 - \delta D$ to a smaller number. Golub et al. [1] and Demmel [2] generalized the analysis to matrices G_1, G_2 or G_3, according to the orthogonal decompositions of G_j. Demmel also obtained some elegant results of determining the range of rank of G_3, and the condition under which one can obtain a smallest perturbation δD of D which reduces the rank of G_3 to a specific integer.

In this paper we will restate Demmel's result in [2] according to the submatrices A, B, C and D in G_3, and then derive a perturbation analysis for the smallest perturbation δD for G_j, $j = 1, 2, 3$ for general case in which the resulting matrices may be rank deficient.

We will use the following notation. For any matrix R, we will denote by $\mathrm{rank}(R)$ the rank of R, $\mathrm{R}(R)$ the range of R, R^{H} the conjugate transpose of R, $R^{-\mathrm{H}} = (R^{\mathrm{H}})^{-1}$, R^{+} the Moore–Penrose pseudoinverse of R. $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\| = \|\cdot\|_2$ the 2-norm, $\|\cdot\|_u$ denotes any (normalized) unitarily invariant norm.

The analysis heavily relies on the singular value decomposition (SVD)[3]. For any matrices $D_1, D_1' \in \mathbb{C}^{m\times n}$, there exist unitary matrices Z, W, Z', W' and diagonal matrices

$$T = \mathrm{diag}(t_1, \ldots, t_l), \qquad T' = \mathrm{diag}(t_1', \ldots, t_l'),$$

with $l = \min\{m, n\}$, $t_1 \geqslant \cdots \geqslant t_l \geqslant 0$ and $t_1' \geqslant \cdots \geqslant t_l' \geqslant 0$ the singular values of D_1, D_1', respectively, such that

$$D_1 = ZTW^{\mathrm{H}}, \qquad D_1' = Z'T'W'^{\mathrm{H}}, \tag{1.1}$$

and the difference $t_j - t_j'$ satisfies

$$|t_j - t_j'| \leqslant \|D_1 - D_1'\|,\ j = 1, \ldots, l, \qquad \sum_{j=1}^{l} (t_j - t_j')^2 \leqslant \|D_1 - D_1'\|_F^2. \tag{1.2}$$

The perturbation bounds in Eq. (1.2) can be used to analyze the least squares (LS) and the total least squares (TLS) problems [3–8].

The paper is arranged as follows. In Section 2 we give an alternative statement of the main result of Demmel ([2] Theorem 3); in Section 3 we present the perturbation analysis for G_j related to the Eckart-Young-Mirsky (EYM) theorem; in Section 4 we derive the perturbation bounds for the constrained total least squares problem (CTLS); finally, in Section 5 we conclude the paper with several remarks. We mention the following result for our further discussion.

Lemma 1.1. *Suppose that* A, $A' \in \mathbb{C}^{m\times n}$ *with* $\mathrm{rank}(A) = \mathrm{rank}(A') = r$. *Then*

$$\begin{aligned} \|AA^{+} - A'A'^{+}\|_u &\leqslant a(u)\|AA^{+}(I - A'A'^{+})\|_u \leqslant a(u)\|\|A' - A\|_u\|A^{+}\|, \\ \|A^{+}A - A'^{+}A'\|_u &\leqslant a(u)\|A^{+}A(I - A'^{+}A')\|_u \leqslant a(u)\|\|A' - A\|_u\|A^{+}\|, \end{aligned} \tag{1.3}$$

in which $a(u) = 1$ for the 2-norm, $a(u) = \sqrt{2}$ for the F-norm and $a(u) = 2$ for any other unitarily invariant norm.

Proof. Let the unitary matrices $U = (U_1, U_2)$, $U' = (U_1', U_2') \in \mathbb{C}^{m\times m}$ be such that $\mathrm{R}(U_1) = \mathrm{R}(A)$ and $\mathrm{R}(U_1') = \mathrm{R}(A')$. Then like the proof of Theorem 2.6.2 in [3].

$$\begin{aligned}\|AA^+ - A'A'^+\|_u &= \|U_1U_1^{\mathrm{H}} - U_1'U_1'^{\mathrm{H}}\|_u = \|U^{\mathrm{H}}(U_1U_1^{\mathrm{H}} - U_1'U_1'^{\mathrm{H}})U'\|_u \\ &= \left\|\begin{pmatrix} 0 & U_1^{\mathrm{H}}U_2' \\ -U_2^{\mathrm{H}}U_1' & 0\end{pmatrix}\right\|_u \leqslant a(u)\|U_1^{\mathrm{H}}U_2'\|_u = a(u)\|U_1U_1^{\mathrm{H}}U_2'U_2'^{\mathrm{H}}\|_u \\ &= a(u)\|AA^+(I - A'A'^+)\|_u = a(u)\|(A^+)^{\mathrm{H}}(A - A')^{\mathrm{H}}(I - A'A'^+)\|_u,\end{aligned}$$

proving the first formula of Eq. (1.3). The second one can be obtained in a similar manner. □

2. Restatement of Theorem 3 in [2]

In this paper we intend to present a perturbation theory for the problems related to the EYM theory. For this purpose, in this section we will restate Theorem 3 of [2], according to the submatrices A, B, C and D in G_3.

Let

$$G_3 = \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{matrix} m_1 \\ m_2 \end{matrix} \qquad n_1,\ n_2 \tag{2.1}$$

and let $P_{N(A)} = I - AA^+$ and $P_{N(A^{\mathrm{H}})} = I - A^+A$ be the orthogonal projections onto the orthogonal complements of the range of A and A^{H}, respectively, and

$$M = P_{N(A)}B, \qquad N = CP_{N(A^{\mathrm{H}})}. \tag{2.2}$$

Now we restate Theorem 3 of [2].

Theorem 2.1. *Let G_3 be defined in Eq.* (2.1). *Then* $\mathrm{rank}(G_3)$ *must satisfy*

$$\begin{aligned}&\mathrm{rank}(A) + \mathrm{rank}(M) + \mathrm{rank}(N) \leqslant \mathrm{rank}(G_3) \\ &\quad = \mathrm{rank}(A) + \mathrm{rank}(M) + \mathrm{rank}(N) + \mathrm{rank}(D_1)\end{aligned} \tag{2.3}$$

where M and N are defined in (2.2) *and*

$$D_1 = (I - NN^+)(D - CA^+B)(I - M^+M). \tag{2.4}$$

If r satisfies $\mathrm{rank}(A) + \mathrm{rank}(M) + \mathrm{rank}(N) \leqslant r < \mathrm{rank}(G_3)$, *then a smallest perturbation δD of D which reduces the rank of*

$$\begin{pmatrix} A & B \\ C & D-\delta D \end{pmatrix}$$

to r is given as follows. Let $p = r - \operatorname{rank}(A) - \operatorname{rank}(M) - \operatorname{rank}(N)$ and let $D_1 = Z\operatorname{diag}(t_1,\ldots,t_l)W^{\mathrm{H}}$ be the SVD of D_1 where $t_1 \geqslant \cdots \geqslant t_l \geqslant 0$ with $l = \min\{m_2, n_2\}$. Then $\delta D = Z\operatorname{diag}(0,\ldots,0,t_{p+1},\ldots,t_l)W^{\mathrm{H}}$. This smallest perturbation has Frobenius norm $\|\delta D\|_F = \sqrt{\sum_{j=p+1}^{l} t_j^2}$ and 2-norm $\|\delta D\| = t_{p+1}$. If $t_p > t_{p+1}$, then δD is unique.

Proof. It can be shown ([2], Lemma 2) that there exist unitary matrices U_1, U_2, V_1 and V_2, such that

$$G_3 = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} A_{11} & 0 & 0 & B_{11} & B_{12} \\ 0 & 0 & 0 & B_{21} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ C_{11} & C_{12} & 0 & D_{11} & D_{12} \\ C_{21} & 0 & 0 & D_{21} & D_{22} \end{pmatrix} \begin{pmatrix} V_1^{\mathrm{H}} & 0 \\ 0 & V_2^{\mathrm{H}} \end{pmatrix}, \tag{2.5}$$

where each of A_{11}, B_{21} and C_{12} is either square and nonsingular, or null. Let

$$p_1 = \operatorname{rank}(A_{11}), \quad p_2 = \operatorname{rank}(B_{21}) \quad \text{and} \quad p_3 = \operatorname{rank}(C_{12}). \tag{2.6}$$

Partition U_i, V_i as follows for $i = 1, 2$:

$$\begin{aligned} U_1 &= (\underset{p_1}{U_{11}}, \underset{p_2}{U_{12}}, \underset{m_1-p_1-p_2}{U_{13}}), & U_2 &= (\underset{p_3}{U_{21}}, \underset{m_2-p_3}{U_{22}}), \\ V_1 &= (\underset{p_1}{V_{11}}, \underset{p_3}{V_{12}}, \underset{n_1-p_1-p_3}{V_{13}}), & V_2 &= (\underset{p_2}{V_{21}}, \underset{n_2-p_2}{V_{22}}), \end{aligned} \tag{2.7}$$

then from Eqs. (2.5) and (2.7), one can show that

$$\begin{aligned} &A = U_{11}A_{11}V_{11}^{\mathrm{H}}, \quad P_{N(A)} = I - U_{11}U_{11}^{\mathrm{H}}, \quad P_{N(A^{\mathrm{H}})} = I - V_{11}V_{11}^{\mathrm{H}}, \\ &M = P_{N(A)}B = U_{12}B_{21}V_{21}^{\mathrm{H}}, \qquad I - M^+M = I - V_{21}V_{21}^{\mathrm{H}} = V_{22}V_{22}^{\mathrm{H}}, \\ &N = CP_{N(A^{\mathrm{H}})} = U_{21}C_{12}V_{22}^{\mathrm{H}}, \qquad I - NN^+ = I - U_{21}U_{21}^{\mathrm{H}} = U_{22}U_{22}^{\mathrm{H}}. \end{aligned} \tag{2.8}$$

So one carries out from Eqs. (2.5)–(2.8) that

$$\begin{aligned} &B(I - M^+M) = U_{11}B_{12}V_{22}^{\mathrm{H}}, \qquad B_{12} = U_{11}^{\mathrm{H}}B(I - M^+M)V_{22}, \\ &(I - NN^+)C = U_{22}C_{21}V_{11}^{\mathrm{H}}, \qquad C_{21} = U_{22}^{\mathrm{H}}(I - NN^+)CV_{11}, \\ &(I - NN^+)D(I - M^+M) = U_{22}D_{22}V_{22}^{\mathrm{H}}, \\ &D_{22} = U_{22}^{\mathrm{H}}(I - NN^+)D(I - M^+M)V_{22}. \end{aligned} \tag{2.9}$$

From Eqs. (2.5)–(2.9),

$$D_1 = (I - NN^+)(D - CA^+B)(I - M^+M) = U_{22}(D_{22} - C_{21}A_{11}^{-1}B_{12})V_{22}^{\mathrm{H}},$$
$$D_{22} - C_{21}A_{11}^{-1}B_{12} = U_{22}^{\mathrm{H}}D_1V_{22}. \tag{2.10}$$

Then it follows from Eqs. (2.8)–(2.10) that

$$\begin{aligned}&\operatorname{rank}(A) = \operatorname{rank}(A_{11}),\ \operatorname{rank}(M) = \operatorname{rank}(B_{21}),\\ &\operatorname{rank}(N) = \operatorname{rank}(C_{12}),\ \operatorname{rank}(D_1) = \operatorname{rank}(D_{22} - C_{21}A_{11}^{-1}B_{12}).\end{aligned} \tag{2.11}$$

So by applying Theorem 3 of [2] and the EYM theorem, one obtains the assertions of the theorem. □

3. Perturbation theory for the EYM theorem

In this section we will present the perturbation theory for the EYM theorem. When considering the LS, the TLS and the equality constrained least squares (LSE) problems, one usually assumes that the coefficient matrices have full rank to simplify the discussion. However, in the practical computations of extracting poles from some transient data using the LS, TLS and LSE techniques, the author found that the results for the rank deficient problems are always better than their full rank counterparts (see the numerical examples in [6,7,9]). Thus one needs to analyze the general cases, including both full rank and rank deficient cases, with a special care.

We first consider the perturbation theory for G_3. Then the perturbation bounds for G_1 and G_2 are just special cases of G_3 with some submatrices set to be zero matrices.

3.1. The perturbation theory for G_3

In this subsection we will present a perturbation theory for G_3. We have the following theorem.

Theorem 3.1. *Let G_3 be defined in Eq. (2.2) and $G_3' = G_3 + \Delta G_3$ its counterpart, with $A' = A + \Delta A$, $B' = B + \Delta B$, $C' = C + \Delta C$, $D' = D + \Delta D$. Let*

$$\begin{aligned}&M = (I - AA^+)B,\ M' = (I - A'A'^+)B',\\ &N = C(I - A^+A), N' = C'(I - A'^+A'),\end{aligned} \tag{3.1}$$

and

$$D_1 = (I - NN^+)(D - CA^+B)(I - M^+M),$$
$$D_1' = (I - N'N'^+)(D' - C'A'^+B')(I - M'^+M'). \tag{3.2}$$

If

$$\operatorname{rank}(A) = \operatorname{rank}(A'), \quad \operatorname{rank}(M) = \operatorname{rank}(M'), \\ \operatorname{rank}(N) = \operatorname{rank}(N'), \tag{3.3}$$

then

$$\begin{aligned} \|D_1 - D_1'\|_u \leqslant {} & \|\Delta D\|_u + \|\Delta C\|_u \|A^+B(I - M^+M)\| \\ & + \|\Delta B\|_u \|(I - N'N'^+)C'A'^+\| \\ & + \|\Delta A\|_u \|(I - N'N'^+)C'A'^+\| \|A^+B(I - M^+M)\| \\ & + a(u)\|M^+\|(\|\Delta B\|_u + a(u)\|\Delta A\|_u \|B\| \|A^+\|)\|(I - N'N'^+)(D' - C'A'^+B')\| \\ & + a(u)\|N^+\|(\|\Delta C\|_u + a(u)\|\Delta A\|_u \|C\| \|A^+\|)\|(D - CA^+B)(I - M^+M)\|, \end{aligned} \tag{3.4}$$

where $a(u)$ *is defined in Lemma* 1.1. *If* $\|\Delta A\|_u \leqslant \xi$, $\|\Delta B\|_u \leqslant \xi$ *and* $\|\Delta C\|_u \leqslant \xi$, *then to the first order*,

$$\begin{aligned} \|D_1 - D_1'\|_u \leqslant {} & \xi(1 + \|A^+B(I - M^+M)\|)(1 + \|(I - NN^+)CA^+\|) \\ & + \xi a(u)(1 + a(u)\|B\| \|A^+\|)\|M^+\| \|(I - NN^+)(D - CA^+B)\| \\ & + \xi a(u)(1 + a(u)\|C\| \|A^+\|)\|N^+\| \|(D - CA^+B)(I - M^+M)\| + O(\xi^2). \end{aligned} \tag{3.5}$$

Remarks 3.1. (1) We enforce the conditions in Eq. (3.3) in order to make A'^+, M'^+ and N'^+ change continuously with respect to the small perturbations in G_3. In the case that some of A, M and N are not of full ranks, the conditions in Eq. (3.3) are too restrictive. But if one has known the ranks of A, M and N, then one can use efficient algorithms such as column pivoting QR factorization (CPQR) [3], rank revealing QR factorization (RRQR) [10] or SVD [3], to keep the computed A', M' and N' (which we also denote, resp., by A', M' and N') having the same ranks as their original counterparts.

(2) The first four terms of the right-hand side in Eq. (3.4) are due to the perturbations ΔD, ΔC, ΔB and ΔA, respectively, while the fifth and sixth terms are due to the perturbations in the orthogonal projections $I - M^+M$ and $I - NN^+$ with $M = (I - AA^+)B$ and $N = C(I - A^+A)$, respectively, as can be shown in Eq. (3.6).

Proof. From Eq. (3.2),

$$\begin{aligned} \|D_1 - D_1'\|_u \leqslant {} & \|(I - N'N'^+)[D' - D - C'A'^+(B' - B) - C'(A'^+ - A^+)B \\ & - (C' - C)A^+B](I - M^+M)\|_u \\ & + \|(I - N'N'^+)(D' - C'A'^+B')[(I - M'^+M') - (I - M^+M)]\|_u \\ & + \|[(I - N'N'^+) - (I - NN^+)](D - CA^+B)(I - M^+M)\|_u. \end{aligned} \tag{3.6}$$

Notice that ([11], Theorem 4.1)

$$A'^{+} - A^{+} = -A'^{+}\Delta AA^{+} + A'^{+}(I - AA^{+}) - (I - A'^{+}A')A^{+},$$

so

$$\begin{aligned}(I - N'N'^{+})&C'(A'^{+} - A^{+})B(I - M^{+}M)\\ &= (I - N'N'^{+})C'[-A'^{+}\Delta AA^{+} + A'^{+}(I - AA^{+})\\ &\quad - (I - A'^{+}A')A^{+}]B(I - M^{+}M)\\ &= -(I - N'N'^{+})C'A'^{+}\Delta AA^{+}B(I - M^{+}M),\end{aligned} \tag{3.7}$$

because $(I - AA^{+})B = M$ and $C'(I - A'^{+}A') = N'$. On the other hand, one has from Eq. (3.1) and Lemma 1.1 that

$$\begin{aligned}\|M - M'\|_u &= \|(I - AA^{+})B - (I - A'A'^{+})B'\|_u\\ &\leqslant \|(I - A'A'^{+})\Delta B\|_u + \|AA^{+} - A'A'^{+}\|_u\|B\|\\ &\leqslant \|\Delta B\|_u + a(u)\|B\|\|\Delta A\|_u\|A^{+}\|.\end{aligned}$$

Also one obtains from Lemma 1.1 that

$$\begin{aligned}\|M'^{+}M' - M^{+}M\|_u &\leqslant a(u)\|M^{+}\|\|M - M'\|_u\\ &\leqslant a(u)\|M^{+}\|(\|\Delta B\|_u + a(u)\|\Delta A\|_u\|B\|\|A^{+}\|).\end{aligned} \tag{3.8a}$$

Similarly, one has

$$\|N'N'^{+} - NN^{+}\|_u \leqslant a(u)\|N^{+}\|(\|\Delta C\|_u + a(u)\|\Delta A\|_u\|C\|\|A^{+}\|). \tag{3.8b}$$

By substituting Eqs. (3.7), (3.8a) and (3.8b) into Eq. (3.6) we obtain the desired estimates in Eqs. (3.4) and (3.5). □

Before making remarks on Theorem 3.1, we first provide an example.

Example 3.1. Let $0 < \xi \ll a^2 < a < 1$ and

$$A = \begin{pmatrix} a & 0\\ 0 & 0\end{pmatrix}, \quad B = C = \begin{pmatrix} 0 & 1\\ 1 & 0\end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0\\ 0 & d\end{pmatrix},$$

$$A' = \begin{pmatrix} a - \xi & -\xi\\ -\xi & \frac{\xi^2}{a-\xi}\end{pmatrix}, \quad B' = C' = \begin{pmatrix} 0 & 1\\ 1 & \xi\end{pmatrix}, \quad D' = \begin{pmatrix} 0 & 0\\ 0 & d - \xi\end{pmatrix}.$$

Then

$$A' = \begin{pmatrix} a - \xi\\ -\xi\end{pmatrix}(a - \xi)^{-1}(a - \xi, -\xi),$$

$$A'^{+} = (a - \xi, -\xi)^{+}(a - \xi)\begin{pmatrix} a - \xi\\ -\xi\end{pmatrix}^{+},$$

so

$$M = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad I - M^+M = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad I - NN^+ = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 0 & 0 \\ 0 & d - a^{-1} \end{pmatrix},$$

$$M' = \frac{1}{(a-\xi)^2 + \xi^2} \begin{pmatrix} \xi \\ a - \xi \end{pmatrix} (a - \xi, \xi(a + 1 - \xi)), I - M'^+M'$$

$$= \frac{1}{(a-\xi)^2 + \xi^2(a+1-\xi)^2} \begin{pmatrix} -\xi(a+1-\xi) \\ a - \xi \end{pmatrix} (-\xi(a+1-\xi), a - \xi),$$

$$N' = \frac{1}{(a-\xi)^2 + \xi^2} \begin{pmatrix} a - \xi \\ \xi(a+1-\xi) \end{pmatrix} (\xi, a - \xi) = (M')^{\mathrm{H}},$$

$$I - N'N'^+ = I - M'^+M'.$$

Notice that all rank conditions in Eq. (3.3) hold. After some calculation we obtain

$$D_1' = \frac{(a-\xi)^2(d-\xi) - (a-\xi)}{((a-\xi)^2 + \xi^2(a+1-\xi)^2)^2}$$
$$\times \begin{pmatrix} \xi^2(a+1-\xi) - \xi(a-\xi)(a+1-\xi) \\ -\xi(a-\xi)(a+1-\xi)(a-\xi)^2 \end{pmatrix}$$
$$= D_1 + \begin{pmatrix} 0 & -\xi(1+a^{-1})(d-a^{-1}) \\ -\xi(1+a^{-1})(d-a^{-1}) & -\xi(1+a^{-2}) \end{pmatrix} + \mathrm{O}(\xi^2).$$

Remarks 3.2. (1) The perturbation bound drawn in Eq. (3.5) is a generalization of (∗) in p. 206 of [2] where Demmel just considered the simplest case that both G_3 and G_3' can be transformed into the standard forms as in Eq. (2.5) by the same pairs of unitary matrices

$$\begin{pmatrix} U_1 & \\ & U_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} V_1 & \\ & V_2 \end{pmatrix}.$$

In this case $M^+M = M'^+M'$ and $NN^+ = N'N'^+$, see Eq. (2.8), and the estimate in Eq. (3.5) reduces to that obtained in p. 206 of [2]. In general case, the inequality (∗) of p. 206 in [2] is not true. For example, if in Example 3.1 we take $d = -a$, then $\|G_3\| = \sqrt{1 + a^2}$ and

$$\|D_1 - D_1'\| \approx \xi(1+a^{-2})\left\|\begin{pmatrix} 0 & 1+a \\ 1+a & -1 \end{pmatrix}\right\|$$

$$= \xi(1+a^{-2})\frac{1+\sqrt{4(1+a)^2+1}}{2},$$

while the upper bound in (*) of p. 206 in [2] is

$$\|D_1 - D_1'\| \leqslant \xi(1+\|G_3\|a^{-1})^2 = \xi(1+\sqrt{1+a^{-2}})^2,$$

which is not true in Example 3.1 for $0<\xi \ll a^2 \ll 1$.

(2) In [12], Zha proposed the following problem: Given matrices $G \in \mathbb{C}^{m\times n}$, $W_1 \in \mathbb{C}^{m\times q}$ and $W_2 \in \mathbb{C}^{s\times n}$, and an integer $r < \text{rank}(G)$, find a matrix $\hat{E} \in \mathbb{C}^{q\times s}$, such that

$$\text{rank}(G - W_1\hat{E}W_2) = r, \qquad \|\hat{E}\|_F = \min_{\substack{\text{rank}(G-W_1EW_2)=r \\ E\in\mathbb{C}^{q\times s}}} \|E\|_F, \tag{3.9}$$

he then obtained the restricted singular value decomposition (RSVD). In [13], Van Huffel and Zha then proposed the restricted total least squares problem (RTLS). The problem proposed by Demmel [2] is a special case of the RSVD problem with

$$G = G_3, \quad W_1 = \begin{pmatrix} 0_{m_1} & \\ & I_{m_2} \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0_{n_1} & \\ & I_{n_2} \end{pmatrix}.$$

For general matrices W_1 and W_2, the RSVD problem is more complicated.

3.2. The perturbation theory for G_1 and G_2

We now consider the perturbation theory for $G_2 = (B^{\text{H}}, D^{\text{H}})^{\text{H}}$ and $G_2' = (B'^{\text{H}}, D'^{\text{H}})^{\text{H}}$, where B, $B' = B + \Delta B \in \mathbb{C}^{m_1\times n_2}$ and D, $D' = D + \Delta D \in \mathbb{C}^{m_2\times n_2}$, with $\text{rank}(B) = \text{rank}(B') = s$. Define

$$D_1 = D(I - B^+B) \quad \text{and} \quad D_1' = D'(I - B'^+B'). \tag{3.10}$$

Let the SVD for D_1 and D_1' be

$$D_1 = ZTW^{\text{H}} \quad \text{and} \quad D_1' = Z'T'W'^{\text{H}}, \tag{3.11}$$

where Z, Z', W, W' are unitary matrices, T and T' are diagonal matrices with the diagonal elements the singular values t_j and t_j' of D_1, D_1', respectively, for $j = 1, \ldots, l = \min\{m_2, n_2\}$, and both t_j and t_j' are arranged in decreasing orders. Golub et al. [1] found that for any positive integer p with $0 \leqslant p < \text{rank}(D_1)$ and $0 \leqslant p < \text{rank}(D_1')$, the matrices

$$\delta D = Z_2T_2W_2^{\text{H}}, \qquad \delta D' = Z_2'T_2'W_2'^{\text{H}} \tag{3.12}$$

satisfy

$$\|\delta D\|_u = \min_{\mathrm{rank}(E) \leqslant p} \|D_1 - E\|_u, \|\delta D'\|_u = \min_{\mathrm{rank}(E) \leqslant p} \|D'_1 - E\|_u, \tag{3.13}$$

where Z_2, Z'_2 are, respectively, the last $m_2 - p$ columns of Z and Z', W_2, W'_2 are, respectively, the last $n_2 - p$ columns of W and W', $T_2 = \mathrm{diag}(t_{p+1}, \dots, t_l)$ and $T'_2 = \mathrm{diag}(t'_{p+1}, \dots, t'_l)$. We have

Corollary 3.1. *Suppose that* B, $B' = B + \Delta B \in C^{m_1 \times n_2}$, $D, D' = D + \Delta D \in C^{m_2 \times n_2}$ *and* $\mathrm{rank}(B) = \mathrm{rank}(B') = s$. *Let* D_1 *and* D'_1 *be defined in* (3.10) *and the SVD of them be in* (3.11). *Then*

$$\|T_2 - T'_2\|_u \leqslant \|D_1 - D'_1\|_u \leqslant \|\Delta D\|_u + a(u)\|\Delta B\|_u \|B^+\| \|D\|. \tag{3.14}$$

Proof. In the matrices G_3 and G'_3 considered in Theorem 3.1, set $A = A' = 0_{m_1 \times n_1}$, $C = C' = 0_{m_2 \times n_1}$. Then

$$A^+ = A'^+ = 0_{n_1 \times m_1}, \quad \Delta A = 0_{m_1 \times n_1}, \quad \Delta C = 0_{m_2 \times n_1}, \quad M = B,$$
$$M' = B', \quad N = N' = 0_{m_2 \times n_1}.$$

Then the estimates in Eq. (3.14) are direct consequences of Eqs. (3.4) and (1.2). □

Notice that if $\mathrm{rank}(D_1) = p$, then from Corollary 3.1, for $j \geqslant p + 1$,

$$t'_j \leqslant \|\delta D'\| \leqslant \|\Delta D\| + \|\Delta B\| \|B^+\| \|D\|.$$

One can also derive the perturbation bound for t'_j according to the modified CS decomposition [14]. Let $\mathrm{rank}(G_2) = k$ and the SVD for G_2 be

$$G_2 = YFP^{\mathrm{H}}, \tag{3.15}$$

with $F = \mathrm{diag}(F_1, \emptyset)$, $F_1 = \mathrm{diag}(f_1, \dots, f_k)$, $f_1 \geqslant f_2 \geqslant \cdots \geqslant f_k > 0$ and Y, P unitary matrices. Partition Y as

$$Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} \begin{matrix} m_1 \\ m_2 \end{matrix} \tag{3.16}$$
$$\quad k, \quad m_1 + m_2 - k.$$

Then $\mathrm{rank}(Y_{11}) = \mathrm{rank}(B) = s$ with $d_1 = \cdots = d_q = 1 > d_{q+1} \geqslant \cdots \geqslant d_s > 0$ the nonzero singular values of Y_{11}. Let $C_1 = \mathrm{diag}(d_{q+1}, \dots, d_s)$ and $S_1 = \mathrm{diag}\left(\sqrt{1 - d_{q+1}^2}, \dots, \sqrt{1 - d_s^2}\right)$. Then it is well known [14] that there exist unitary matrices U_1, U_2, V_1 and V_2 with appropriate sizes, such that Y has a modified CS decomposition

$$Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} = \begin{pmatrix} U_1 & \\ & U_2 \end{pmatrix} \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \begin{pmatrix} V_1^{\mathrm{H}} & \\ & V_2^{\mathrm{H}} \end{pmatrix}, \tag{3.17}$$

where

$$\begin{aligned} D_{11} &= \mathrm{diag}(I_q, C_1, \emptyset_{(m_1-s)\times(k-s)}),\\ D_{12} &= \mathrm{diag}(\emptyset_{q\times(m_2+q-k)}, S_1, I_{m_1-s}),\\ D_{21} &= \mathrm{diag}(\emptyset_{(m_2+q-k)\times q}, S_1, I_{k-s}),\\ D_{22} &= \mathrm{diag}(I_{m_2+q-k}, -C_1, \emptyset_{(k-s)\times(m_1-s)}). \end{aligned} \tag{3.18}$$

Now we have the following theorem.

Theorem 3.2. *Suppose that* $\mathrm{rank}(B) = \mathrm{rank}(B') = s$, $\|G_2 - G_2'\|\|B^+\| < \frac{1}{2}$ *and* $\|G_2 - G_2'\|\|G_2^+\| < d_s/2\sqrt{2}$. *Then* $\mathrm{rank}(G_2') = k_1 \geqslant \mathrm{rank}(G_2) = k$. *Let* D_1 *and* D_1' *be defined in Eq.* (3.10) *and suppose that* $\mathrm{rank}(D_1) = p = k - s$. *Let* δD, $\delta D'$ *be defined in Eq.* (3.12), *then* $\delta D = 0$, *and*

$$\|\delta D'\| \leqslant \frac{1}{d_s}\|G_2' - G_2\|(1+\epsilon), \tag{3.19}$$

with $\epsilon = \mathrm{O}(\|G_2 - G_2'\|\|G_2^+\|)$.

Proof. In Lemma 3.2 of [9], set $L = B$, $K = D$, $n = n_2$. □

For $G_1 = (C, D)$, noting that $G_1^{\mathrm{H}} = (C, D)^{\mathrm{H}}$, one can apply the results in Corollary 3.1 and Theorem 3.2 to derive the perturbation bounds. We omit the detail.

Remarks 3.3. The upper bounds derived in Corollary 3.1 and Theorem 3.2 are realistic in the sense that one can find an example such that the true error in $|t_j - t_j'|$ is close to the bound in Eq. (3.14) or Eq. (3.19).

Example 3.2. Let

$$B = (a, 0), \quad B' = (a(1+e), ae), \quad D = \begin{pmatrix} a & 0 \\ -a & 0 \end{pmatrix},$$
$$D' = \begin{pmatrix} a(1+e) & ae \\ -a(1+e) & ae \end{pmatrix},$$

in which $a > 0$ and $0 < e \ll 1$. Then $m_1 = 1$, $m_2 = 2$, $\mathrm{rank}(B) = \mathrm{rank}(B') = 1$. Further more, one can easily derive that

$$I - B^+B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$
$$I - B'^+B' = \frac{1}{(1+e)^2 + e^2}\begin{pmatrix} e^2 & -e(1+e) \\ -e(1+e) & (1+e)^2 \end{pmatrix},$$

$$D_1 = D(I - B^+B) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$D_1' = D'(I - B'^+B') = \frac{2e(1+e)a}{(1+e)^2 + e^2} \begin{pmatrix} 0 & 0 \\ -e & 1+e \end{pmatrix}.$$

Then one has $\operatorname{rank}(D_1) = 0$, $\operatorname{rank}(D_1') = 1, t_1 = t_2 = 0, t_1' = 2e(1+e)a/((1+e)^2 + e^2)^{1/2}$ and $t_2' = 0$. That is,

$$|t_1 - t_1'| = t_1' = \frac{2e(1+e)a}{((1+e)^2 + e^2)^{1/2}} \approx 2ea, \quad |t_2 - t_2'| = 0. \tag{3.20}$$

On the other hand, it turns out that

$$\|\Delta D\| = \left\| \begin{pmatrix} ae & ae \\ -ae & ae \end{pmatrix} \right\| = \sqrt{2}ae, \quad \|\Delta B\| = \|(ae, ae)\| = \sqrt{2}ae,$$

$$\|D\| = \sqrt{2}a, \quad \|B^+\| = a^{-1},$$

so

$$\|\Delta D\| + \|\Delta B\| \|B^+\| \|D\| = (2 + \sqrt{2})ae. \tag{3.21}$$

Also note that

$$G_2 = \begin{pmatrix} B \\ D \end{pmatrix} = \begin{pmatrix} a & 0 \\ a & 0 \\ -a & 0 \end{pmatrix}, \quad G_2' = \begin{pmatrix} B' \\ D' \end{pmatrix} = \begin{pmatrix} a(1+e) & ae \\ a(1+e) & ae \\ -a(1+e) & ae \end{pmatrix},$$

$$\|G_2 - G_2'\| = \left\| \begin{pmatrix} ae & ae \\ ae & ae \\ -ae & ae \end{pmatrix} \right\| = 2ae,$$

the SVD for G_2 is $G_2 = YFP^{\mathrm{H}}$, where

$$Y = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ -1/\sqrt{3} & 0 & 2/\sqrt{6} \end{pmatrix}, \quad F = \begin{pmatrix} \sqrt{3}a & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Notice that $k = m_1 = 1$, $m_2 = 2$ and $m_1 + m_2 - k = 2$. So according to Eq. (3.16),

$$Y_{22} = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{6} \\ 0 & 2/\sqrt{6} \end{pmatrix}, \qquad \sigma_1(Y_{22}) = 1, \quad \sigma_2(Y_{22}) = \frac{1}{\sqrt{3}}.$$

Then

$$\|Y_{22}^+\| \|G_2 - G_2'\| = 2\sqrt{3}ae. \tag{3.22}$$

Comparing Eqs. (3.20)–(3.22), one observes that the perturbation bounds in Eqs. (3.14) and (3.19) are realistic.

(2) The perturbation bounds obtained in Theorem 3.2 and Corollary 3.1 can be used to analyze the generalized TLS (GTLS or LS-TLS) problems [5,15,16,19] and the (LSE) problems [17,9].

3.3. Perturbation bound for $\|\delta D - \delta D'\|_u$

In this subsection we will provide a bound for $\|\delta D - \delta D'\|_u$. Let $\eta = \|D_1 - D_1'\|$, $\eta_F = \|D_1 - D_1'\|_F$ and $\eta_u = \|D_1 - D_1'\|_u$. It has been observed that when $\eta \sim t_p - t_{p+1}$, then $\|\delta D - \delta D'\|_u$ could be large even when η is small. For example, [2] if $D_1 = \mathrm{diag}(1, 1-\epsilon)$, $D_1' = \mathrm{diag}(1-\epsilon, 1)$ with $0 < \epsilon \ll 1$, then $\|D_1 - D_1'\|_u = a(u)\epsilon$. However, for $p = 1$, $\delta D = \mathrm{diag}(0, 1-\epsilon)$ and $\delta D' = \mathrm{diag}\ (1-\epsilon, 0)$. Therefore,

$$\|\delta D - \delta D'\|_u = a(u)(1-\epsilon).$$

In the following theorem we will show that if $\eta < (t_p - t_{p+1})/2$, then the quantity $\|\delta D - \delta D'\|_u$ should be of order $\mathrm{O}(\eta_u)$.

Theorem 3.3. *Suppose that D_1, $D_1' \in \mathbb{C}^{m\times n}$, and the SVD for D_1 and D_1' be given in (1.1). For some integer p with $0 \leqslant p < l = \min\{m, n\}$, let*

$$\begin{aligned} &Z = (\underset{p}{Z_1}, \underset{m-p}{Z_2}), \quad Z' = (\underset{p}{Z_1'}, \underset{m-p}{Z_2'}), \quad W = (\underset{p}{W_1}, \underset{n-p}{W_2}), \\ &W' = (\underset{p}{W_1'}, \underset{n-p}{W_2'}), \end{aligned} \tag{3.23}$$

$T_1 = \mathrm{diag}(t_1, \ldots, t_p)$, $T_2 = \mathrm{diag}(t_{p+1}, \ldots, t_l)$, $T_1' = \mathrm{diag}(t_1', \ldots, t_p')$ *and* $T_2' = \mathrm{diag}\ (t_{p+1}', \ldots, t_l')$. *If* $\delta D = Z_2 T_2 W_2^{\mathrm{H}}$, $\delta D' = Z_2' T_2' W_2'^{\mathrm{H}}$, *then*

$$\|\delta D - \delta D'\|_u \leqslant \eta_u + a(u)\max\{\|T_2\|_u, \ \|T_2'\|_u\}. \tag{3.24}$$

Furthermore, if $t_p > t_{p+1}$ and $\eta < (t_p - t_{p+1})/2$, then

$$\|\delta D - \delta D'\|_u \leqslant \eta_u \left(1 + a(u)\frac{t_{p+1} + \eta}{t_p - t_{p+1} - \eta}\right), \tag{3.25}$$

in which we define $t_p = \infty$ for $p = 0$.

Proof. If $p = 0$, then $\delta D = D_1$ and $\delta D' = D_1'$ so Eq. (3.25) holds. For $p > 0$, Notice that

[2] This example was provided by one referee.

$$\eta_u = \|D_1 - D_1'\|_u = \|Z^{\mathrm{H}}(D_1 - D_1')W'\|_u$$
$$= \left\| \begin{pmatrix} T_1 W_1^{\mathrm{H}} W_1' - Z_1^{\mathrm{H}} Z_1' T_1' & T_1 W_1^{\mathrm{H}} W_2' - Z_1^{\mathrm{H}} Z_2' T_2' \\ T_2 W_2^{\mathrm{H}} W_1' - Z_2^{\mathrm{H}} Z_1' T_1' & T_2 W_2^{\mathrm{H}} W_2' - Z_2^{\mathrm{H}} Z_2' T_2' \end{pmatrix} \right\|_u \tag{3.26}$$

and

$$\|\delta D - \delta D'\|_u = \|Z^{\mathrm{H}}(\delta D - \delta D')W'\|_u$$
$$= \left\| \begin{pmatrix} 0 & -Z_1^{\mathrm{H}} Z_2' T_2' \\ T_2 W_2^{\mathrm{H}} W_1' & T_2 W_2^{\mathrm{H}} W_2' - Z_2^{\mathrm{H}} Z_2' T_2' \end{pmatrix} \right\|_u. \tag{3.27}$$

It is obvious that,

$$\bar{\eta}_u \equiv \max_{i,j=1,2}\{\|T_i W_i^{\mathrm{H}} W_j' - Z_i^{\mathrm{H}} Z_j' T_j'\|_u, \|T_i' W_i'^{\mathrm{H}} W_j - Z_i'^{\mathrm{H}} Z_j T_j\|_u\} \leqslant \eta_u. \tag{3.28}$$

One then has from Eqs. (3.27) and (3.28),

$$\|\delta D - \delta D'\|_u \leqslant \|T_2 W_2^{\mathrm{H}} W_2' - Z_2^{\mathrm{H}} Z_2' T_2'\|_u + \left\| \begin{pmatrix} 0 & -Z_1^{\mathrm{H}} Z_2' T_2' \\ T_2 W_2^{\mathrm{H}} W_1' & 0 \end{pmatrix} \right\|_u$$
$$\leqslant \eta_u + a(u)\max\{\|T_2 W_2^{\mathrm{H}} W_1'\|_u, \|Z_1^{\mathrm{H}} Z_2' T_2'\|_u\}$$
$$\leqslant \eta_u + a(u)\max\{\|T_2\|_u, \|T_2'\|_u\}.$$

This is the inequality in Eq. (3.24).

Furthermore, if $\eta < (t_p - t_{p+1})/2$, then from Eq. (1.2), $t_p' - t_{p+1}' \geqslant t_p - t_{p+1} - 2\eta > 0$, $t_p' - t_{p+1} \geqslant t_p - t_{p+1} - \eta > 0$, and from Eq. (3.28),

$$\|T_2 W_2^{\mathrm{H}} W_1' T_1'^{-1} - Z_2^{\mathrm{H}} Z_1'\|_u \leqslant \|T_2 W_2^{\mathrm{H}} W_1' - Z_2^{\mathrm{H}} Z_1' T_1'\|_u \|T_1'^{-1}\| \leqslant \bar{\eta}_u / t_p',$$

and so

$$\|Z_2^{\mathrm{H}} Z_1'\|_u \leqslant \bar{\eta}_u / t_p' + \|T_2\| \|W_2^{\mathrm{H}} W_1'\|_u \|T_1'^{-1}\| = (\bar{\eta}_u + t_{p+1}\|W_2^{\mathrm{H}} W_1'\|_u)/t_p'. \tag{3.29a}$$

Similarly, one can derive

$$\|W_2^{\mathrm{H}} W_1'\|_u \leqslant (\bar{\eta}_u + t_{p+1}\|Z_2^{\mathrm{H}} Z_1'\|_u)/t_p'. \tag{3.29b}$$

Substituting Eq. (3.29a) into Eq. (3.29b) one obtains

$$\frac{t_p'^2 - t_{p+1}^2}{t_p'^2}\|W_2^{\mathrm{H}} W_1'\|_u \leqslant \frac{\bar{\eta}_u(t_p' + t_{p+1})}{t_p'^2}.$$

Because $t_p' > t_{p+1} \geqslant 0$, so

$$\|W_2^{\mathrm{H}} W_1'\|_u \leqslant \frac{\bar{\eta}_u}{t_p' - t_{p+1}} \leqslant \frac{\bar{\eta}_u}{t_p - t_{p+1} - \eta} \leqslant \frac{\eta_u}{t_p - t_{p+1} - \eta}. \tag{3.30a}$$

Similarly, one has

$$\|Z_1^{\mathrm{H}} Z_2'\|_u \leqslant \frac{\bar{\eta}_u}{t_p - t_{p+1} - \eta} \leqslant \frac{\eta_u}{t_p - t_{p+1} - \eta}. \tag{3.30b}$$

Then from (3.27)–(3.30), one obtains

$$\begin{aligned}
\|\delta D - \delta D'\|_u &\leqslant \|T_2 W_2^{\mathrm{H}} W_2' - Z_2^{\mathrm{H}} Z_2' T_2'\|_u + \left\| \begin{pmatrix} 0 & -Z_1^{\mathrm{H}} Z_2' T_2' \\ T_2 W_2^{\mathrm{H}} W_1' & 0 \end{pmatrix} \right\|_u \\
&\leqslant \bar{\eta}_u + a(u) \max\{\|T_2 W_2^{\mathrm{H}} W_1'\|_u, \|Z_1^{\mathrm{H}} Z_2' T_2'\|_u\} \\
&\leqslant \bar{\eta}_u + a(u) \frac{\bar{\eta}_u}{t_p - t_{p+1} - \eta} \max\{\|T_2\|, \|T_2'\|\} \\
&\leqslant \bar{\eta}_u \left(1.+a(u) \frac{t_{p+1} + \eta}{t_p - t_{p+1} - \eta} \right) \\
&\leqslant \eta_u \left(1 + a(u) \frac{t_{p+1} + \eta}{t_p - t_{p+1} - \eta} \right).
\end{aligned}$$

One then obtains the desired estimate in Eq. (3.25). □

Remark. From the discussion of this section, we see that when $\|D_1 - D_1'\|_u = \eta_u$ is small, then the quantity $|t_j - t_j'| \leqslant \eta$ for $j = 1, \ldots, l$, that is, the perturbations of the singular values of D_1 are also small. Notice that the quantity $\|\delta D - \delta D'\|_u$ could be large. However, if $t_p - t_{p+1}$, the gap between the singular values, is large enough such that $\eta < (t_p - t_{p+1})/2$, then $\|\delta D - \delta D'\|_u$ is also small. The above observations can be used to study perturbation analysis of the TLS, LSE and GTLS problems.

Wedin [18] first obtained the estimates (3.30a) and (3.30b). Derivation in this paper is simpler.

4. Perturbation analysis for the CTLS problem

In this section we will derive perturbation analysis for the constrained TLS problem (CTLS). For given matrices

$$L = \overset{}{\underset{n_1 \quad n_2}{\begin{pmatrix} A & B_{01} \\ C & D_{01} \end{pmatrix}}} \begin{matrix} m_1 \\ m_2 \end{matrix}, \qquad F = \underset{d}{\begin{pmatrix} B_{02} \\ D_{02} \end{pmatrix}} \begin{matrix} m_1 \\ m_2 \end{matrix}, \tag{4.1}$$

with $m = m_1 + m_2$ and $n = n_1 + n_2$, consider a system of linear equations

$$LX \approx F, \tag{4.2}$$

in which L and F are approximations of the unobservable data matrices L_0 and F_0, respectively, which satisfy the exact relation

$$L_0 X_0 = F_0, \tag{4.3}$$

with

$$L_0 = \begin{pmatrix} A & B_{01} \\ C & D_{01}^{(0)} \end{pmatrix} \begin{matrix} m_1 \\ m_2 \end{matrix}, \qquad F_0 = \begin{pmatrix} B_{02} \\ D_{02}^{(0)} \end{pmatrix} \begin{matrix} m_1 \\ m_2 \end{matrix}, \tag{4.4}$$
$$\qquad n_1 \quad n_2 \qquad\qquad\qquad d$$

that is, all errors in L and F are contained in D_{01} and D_{02}. Denote $B = (B_{01}, B_{02})$ $D = (D_{01}, D_{02})$ and

$$G_3 = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Then the CTLS problem is: Find integer r which satisfies

$$u = \operatorname{rank}(A) + \operatorname{rank}(M) + \operatorname{rank}(N) \leqslant r \leqslant \operatorname{rank}(G_3), \tag{4.5}$$

and an estimate $\hat{D} = (\hat{D}_{01}, \hat{D}_{02}) = D - \delta D$, such that

$$\|\delta D\|_F = \min \left\{ \|E\|_F\colon\ E \in \mathbb{C}^{m_2 \times (n_2+d)},\ \operatorname{rank}\begin{pmatrix} A & B \\ C & D-E \end{pmatrix} = r \right\},$$
$$\text{s.t.}\quad \hat{F} \in R(\hat{L}), \tag{4.6}$$

where M and N are defined in Eq. (3.1), and

$$\hat{L} = \begin{pmatrix} A & B_{01} \\ C & \hat{D}_{01} \end{pmatrix}, \quad \hat{F} = \begin{pmatrix} B_{02} \\ \hat{D}_{02} \end{pmatrix}, \quad \hat{G}_3 = (\hat{L}, \hat{F}).$$

Here we would like to point out that the CTLS problem defined in this way is always solvable for $r = u$. In fact, because Eq. (4.3) is exact, $B_{02} = (A, B_{01})(A, B_{01})^+ B_{02}$ and so

$$\operatorname{rank}(M) = \operatorname{rank}(P_{N(A)} B) = \operatorname{rank}(P_{N(A)} B_{01}).$$

Notice that for $r = u$, $\operatorname{rank}(\hat{G}_3) = u$. Then from Theorem 3 of [2] and Theorem 2.1,

$$u = \operatorname{rank}(A) + \operatorname{rank}(P_{N(A)} B_{01}) + \operatorname{rank}(N) \leqslant \operatorname{rank}\left(\begin{pmatrix} A & B_{01} \\ C & \hat{D}_{01} \end{pmatrix}\right)$$
$$= \operatorname{rank}(\hat{L}) \leqslant \operatorname{rank}(\hat{G}_3) = u,$$

so for $r = u$,

$$\hat{L} X = \hat{F} \tag{4.7}$$

is consistent, and according to Theorem 2.1, δD satisfies the first constraint of Eq. (4.6).

In general a given problem will have solutions with different r. The solution with the maximum such r will often be the most useful, but not always, as this may for example have unacceptably large $\|X\|$, and a solution corresponding to a smaller r may be physically more meaningful.

Now we turn to study the perturbation theory of the CTLS problem.

When formulating the CTLS problem, one assumes that the matrices A, B and C are known exactly, and so therefore are their structures and ranks. We are then interested in how perturbation in D affects the solutions. In the practical computations, because of finite precision computation, even using a numerically stable algorithm in the computation will produce computed errors corresponding to slightly different initial data [3]. Notice that in general this effective error in the initial matrices due to round off is much smaller than the error caused by uncertainty in the data. To simplify the analysis, we therefore make the following assumptions, which allow any perturbations in D, but only relatively small perturbations in A, B and C. When considering the perturbed CTLS problem, we suppose that our perturbed data $A' = A + \Delta A$, $B' = B + \Delta B$, $C' = C + \Delta C$ and $D' = D + \Delta D$ satisfy

$$\|\Delta A\| \leqslant \alpha_{11}\epsilon,\ \ \|\Delta B\| \leqslant \alpha_{12}\epsilon,\ \ \|\Delta C\| \leqslant \alpha_{21}\epsilon,\ \ \|\Delta D\| \leqslant \epsilon_2, \tag{4.8}$$

and

$$\begin{aligned} &\operatorname{rank}(A') = \operatorname{rank}(A),\ \ \operatorname{rank}(M') = \operatorname{rank}(M),\\ &\operatorname{rank}(N') = \operatorname{rank}(N),\ \ \operatorname{rank}(A', B'_{01}) = \operatorname{rank}(A', B'), \end{aligned} \tag{4.9}$$

where M, M', N, N' are defined in Eq. (3.1), α_{ij} are constants depending on the dimensions and the submatrices of G_3, ϵ is the machine precision unit and ϵ_2 can be large. Then the perturbed CTLS problem is: Find an integer r with

$$u = \operatorname{rank}(A') + \operatorname{rank}(M') + \operatorname{rank}(N') \leqslant r \leqslant \operatorname{rank}(G'_3), \tag{4.5$'$}$$

and an estimate $\hat{D}' = (\hat{D}'_{01}, \hat{D}'_{02}) = D' - \delta D'$, such that

$$\begin{aligned} &\|\delta D'\|_F = \min\left\{\|E\|_F\colon\ E \in \mathbb{C}^{m_2\times(n_2+d)},\ \operatorname{rank}\begin{pmatrix} A' & B' \\ C' & D' - E\end{pmatrix} = r\right\},\\ &\text{s.t.}\quad \hat{F}' \in R(\hat{L}'), \end{aligned} \tag{4.6$'$}$$

where

$$\hat{L}' = \begin{pmatrix} A' & B'_{01} \\ C' & \hat{D}'_{01}\end{pmatrix},\quad \hat{F}' = \begin{pmatrix} B'_{02} \\ \hat{D}'_{02}\end{pmatrix},\quad \hat{G}'_3 = (\hat{L}', \hat{F}').$$

If for an r, Eq. (4.6′) is solvable, then a CTLS solution X is a solution of the consistent system

$$\hat{L}'X = \hat{F}'. \tag{4.7$'$}$$

We now have the following theorem.

Theorem 4.1. *Let the matrices L, F be given in Eq.* (4.1) *and L', F' be their perturbed versions, respectively, and perturbations satisfy Eqs.* (4.8) *and* (4.9). *Let D_1 and D'_1 be defined in Eq.* (3.2) *and the* SVD *of them be as in Eq.* (1.1). *Let*

$$\epsilon_L = \|L - L'\|, \quad \epsilon_G = \|G_3 - G_3'\|, \quad \eta = \|D_1 - D_1'\|,$$

and

$$\begin{aligned}\eta_T = {} & \epsilon_2 + \epsilon(\alpha_{21}\|A^+B(I - M^+M)\| + \alpha_{12}\|(I - N'N'^+)C'A'^+\|) \\ & + \epsilon\alpha_{11}\|(I - N'N'^+)C'A'^+\|\|A^+B(I - M^+M)\| \\ & + \epsilon\|M^+\|(\alpha_{12} \quad + \alpha_{11}\|B\|\|A^+\|)\|(I - N'N'^+)(D' - C'A'^+B')\| \\ & + \epsilon\|N^+\|(\alpha_{21} \quad + \alpha_{11}\|C\|\|A^+\|)\|(D - CA^+B)(I - M^+M)\|. \end{aligned} \tag{4.10}$$

If for some integer r satisfying $u \leqslant r \leqslant \operatorname{rank}(G_3)$ and for $p = r - u$,

$$\sigma_r(L) > t_{p+1} + \epsilon_L + \eta_T \quad \text{and} \quad t_p > t_{p+1} + 2\eta_T, \tag{4.11}$$

where $\sigma_r(L)$ is the rth largest singular value of L, then for this r both the original and the perturbed CTLS *problems are solvable. Furthermore, in this case, for the original and the perturbed minimum F-norm (and so 2-norm)* CTLS *solutions X_{CTLS} and X'_{CTLS} we have the following estimates:*

(1) *When $r = n$, then*

$$\|X_{\mathrm{CTLS}} - X'_{\mathrm{CTLS}}\|_u \leqslant \frac{\epsilon_G + \eta(1 + \frac{t_{p+1}+\eta}{t_p - t_{p-1} - \eta})}{\sigma_r(L) - t_{p+1} - \epsilon_L - \eta}\sqrt{\|X_{\mathrm{CTLS}}\|_u^2 + b(u)}. \tag{4.12}$$

(2) *When $r<n$, then*

$$\|X_{\mathrm{CTLS}} - X'_{\mathrm{CTLS}}\|_u \leqslant \left[\left(\frac{\epsilon_G + \eta\left(1 + \frac{t_{p-1}+\eta}{t_p - t_{p+1} - \eta}\right)}{\sigma_r(L) - t_{p+1} - \epsilon_L - \eta}\right)^2 (\|X_{\mathrm{CTLS}}\|_u^2 + b(u)) + \left(\frac{\epsilon_L + \eta\left(1 + \frac{t_{p+1}+\eta}{t_p - t_{p+1} - \eta}\right)}{\sigma_r(L) - t_{p+1}}\|X_{\mathrm{CTLS}}\|_u\right)^2\right]^{1/2}; \tag{4.13}$$

in which $b(u) = d$ for the F-norm and $b(u) = 1$ for the 2-norm. Furthermore, when $r < n$, for any solution X of the original CTLS *problem, there exists a solution X' of the perturbed* CTLS *problem, such that*

$$\|X - X'\|_u \leqslant \sqrt{2}\frac{\epsilon_G + \eta\left(1 + \frac{t_{p+1}+\eta}{t_p - t_{p+1} - \eta}\right)}{\sigma_r(L) - t_{p+1} - \epsilon_L - \eta}\sqrt{\|X\|_u^2 + b(u)}, \tag{4.14}$$

and vice versa.

Proof. First we have from Eqs. (3.4) and (4.10) that $\eta \leqslant \eta_T$. we then obtain from

$$\hat{G}_3 = G_3 - \begin{pmatrix} 0 & 0 \\ 0 & \delta D \end{pmatrix}, \qquad \hat{G}_3' = G_3' - \begin{pmatrix} 0 & 0 \\ 0 & \delta D' \end{pmatrix}$$

and Eq. (1.2) that

$$\sigma_r(\hat{L}) \geqslant \sigma_r(L) - \|\delta D\| = \sigma_r(L) - t_{p+1} > 0 \tag{4.15a}$$

and

$$\sigma_r(\hat{L}') \geqslant \sigma_r(L') - \|\delta D'\| \geqslant \sigma_r(L) - \epsilon_L - t_{p+1} - \eta > 0. \tag{4.15b}$$

So $r \geqslant \operatorname{rank}(\hat{G}_3) \geqslant \operatorname{rank}(\hat{L}) \geqslant r$ and Eq. (4.7) is consistent. With the same argument Eq. (4.7′) is also consistent. We now have

$$\begin{aligned} X_{\text{CTLS}} - X'_{\text{CTLS}} &= \hat{L}^+\hat{F} - \hat{L}'^+\hat{F}' = \hat{L}'^+(\hat{F} - \hat{F}') \\ &\quad + (\hat{L}'^+(\hat{L}' - \hat{L})\hat{L}^+ + (I - \hat{L}'^+\hat{L}')\hat{L}^+)\hat{F} \\ &= \hat{L}'^+(\hat{G}'_3 - \hat{G}_3)\begin{pmatrix} X_{\text{CTLS}} \\ -I \end{pmatrix} + (I - \hat{L}'^+\hat{L}')\hat{L}^+\hat{L}X_{\text{CTLS}} \end{aligned}$$

and so

$$\begin{aligned} \|X_{\text{CTLS}} - X'_{\text{CTLS}}\|_u^2 &\leqslant (\|\hat{L}'^+\|\|\hat{G}'_3 - \hat{G}_3\|)^2(\|X_{\text{CTLS}}\|_u^2 + b(u)) \\ &\quad + (\|(I - \hat{L}'^+\hat{L}')\|\|\hat{L}^+\|\|\hat{L} - \hat{L}'\|\|X_{\text{CTLS}}\|_u)^2. \end{aligned} \tag{4.16}$$

Also we have

$$\|\hat{G}_3 - \hat{G}'_3\| \leqslant \|G_3 - G'_3\| + \|\delta D - \delta D'\| \leqslant \epsilon_G + \eta\left(1 + \frac{t_{p+1} + \eta}{t_p - t_{p+1} - \eta}\right) \tag{4.17}$$

by applying Theorem 3.3.

1. When $r = n$, $I - \hat{L}'^+\hat{L}' = 0$. By substituting Eqs. (4.15a) and (4.17) into Eq. (4.16) we obtain the desired estimate in Eq. (4.12).
2. When $r < n$, we also have

$$\|\hat{L} - \hat{L}'\| \leqslant \|L - L'\| + \|\delta D - \delta D'\| \leqslant \epsilon_L + \eta\left(1 + \frac{t_{p+1} + \eta}{t_p - t_{p+1} - \eta}\right). \tag{4.18}$$

By substituting Eqs. (4.15a), (4.17) and (4.18) into Eq. (4.16) we obtain the desired estimate in Eq. (4.13).

Furthermore, for $r < n$, any CTLS solution X of Eq. (4.7) is of the form

$$X = \hat{L}^+\hat{F} + (I - \hat{L}^+\hat{L})Z, \tag{4.19}$$

where Z is an arbitrary $n \times d$ matrix. Define X' as

$$X' = \hat{L}'^+\hat{F}' + (I - \hat{L}'^+\hat{L}')(\hat{L}^+\hat{F} + (I - \hat{L}^+\hat{L})Z), \tag{4.20}$$

then X' is a CTLS solution of Eq. (4.7′), and we have

$$X' - X = -\hat{L}'^+(\hat{G}'_3 - \hat{G}_3)\begin{pmatrix} X_{\text{CTLS}} \\ -I \end{pmatrix} - \hat{L}'^+(\hat{L}' - \hat{L})(I - \hat{L}^+\hat{L})Z.$$

Then by applying (4.15)–(4.18) and the Cauchy–Schwartz inequality we obtain the estimate in Eq. (4.14). □

Remark 4.1. (1) We use the conditions in Eq. (4.11) with the following consideration. Suppose that Eq. (4.3) is consistent and $\mathrm{rank}(L_0) = \mathrm{rank}(L_0, F_0) = r$, and that Eq. (4.2) is slightly inconsistent, then we can expect $t_p > 0$ and $t_{p+1} \approx 0$ and $\sigma_r(L) - t_{p+1} > 0$. Furthermore, if $\|G_3' - G_3\|$ is very small such that conditions in Eq. (4.11) hold, then both the original and the perturbed CTLS problems are solvable for this number r, as shown in Eqs. (4.15a) and (4.15b).

(2) If $\epsilon \ll \epsilon_2$, we can assume that

$$\epsilon_G \approx \epsilon_2, \epsilon_L \approx \epsilon_2 \text{ and } \eta \leqslant \eta_T \approx \epsilon_2,$$

and we can further simplify the estimates in Eqs. (4.12)–(4.14).

5. Concluding remarks

Golub et al. [1], Demmel [2] generalized the EYM theorem, which solves the problem of approximating a matrix by one of lower rank with only a specific rectangular subset of the matrix allowed to be changed. Based on an alternative statement of a main result of Demmel ([2], Theorem 3), in this paper the perturbation bounds for the EYM theorem for G_j, $j = 1, 2, 3$ and $\|\delta D - \delta D'\|_u$ have been deduced which generalizes the result of Demmel in [2]. Based on these perturbation bounds a perturbation analysis for the CTLS problem has also been presented.

Acknowledgements

The author is grateful to Professor Gene H. Golub and the referees for their useful comments and suggestions.

References

[1] G.H. Golub, A. Hoffman, G.W. Stewart, A generalization of the Eckart–Young–Mirsky matrix approximation theorem, Linear Algebra Appl. 88/89 (1987) 317–327.

[2] J. Demmel, The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems, SIAM J. Numer. Anal. 24 (1987) 199–206.

[3] G.H. Golub, C.F. Van Loan, Matrix Computations, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[4] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[5] S. Van Huffel, J. Vandewalle, The total least squares problem: Computational aspects and analysis, SIAM 1991.
[6] M. Wei, Perturbation of the least squares problem, Linear Algebra Appl. 141 (1990) 177–182.
[7] M. Wei, The analysis for the total least squares problem with more than one solution, SIAM J. Matrix Anal. Appl. 13 (1992) 746–763.
[8] M. Wei, Algebraic relations between the total least squares and the least squares problems with more than one solution, Numer. Math. 62 (1992) 123–148.
[9] M. Wei, Perturbation theory for the rank-deficient equality constrained least squares problem, SIAM J. Numer. Anal. 29 (1992) 1462–1481.
[10] T.F. Chan, Rank revealing QR factorization, Linear Algebra Appl. 88/89 (1987) 67–82.
[11] P.-Å. Wedin, Perturbation theory for pseudoinverse, BIT 13 (1973) 217–232.
[12] H. Zha, Restricted singular value decomposition of matrix triplets, SIAM J. Matrix Anal. Appl. 12 (1991) 172–194.
[13] S. Van Huffel, H. Zha, Restricted total least squares problem: Formulation, algorithms, and properties, SIAM J. Matrix Anal. Appl. 12 (1991) 292–309.
[14] C.C. Paige, M.A. Saunders, Towards a generalized singular value decomposition, SIAM J. Numer. Anal. 18 (1981) 398–405.
[15] S. Van Huffel, J. Vandewalle, Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns in A are subject to error, SIAM J. Matrix Anal. Appl. 10 (1989) 294–315.
[16] C.C. Paige, M. Wei, Analysis of the generalized total least squares problem $AX \approx B$ when some columns of A are free of error, Numer. Math. 65 (1993) 177–202.
[17] L. Eldén, Perturbation theory for the least squares problem with equality constraints, SIAM J. Numer. Anal. 17 (1980) 338–350.
[18] P.-Å. Wedin, Perturbation bounds inconnection with singular value decomposition, BIT 12 (1972) 99–111.
[19] M. Wei, G. Majda, On the accuracy of the least squares and the total least squares methods, Numer. Math. J. Chinese Univ. 3 (1994) 135–153.