Available online at www.sciencedirect.com
ScienceDirect
Linear Algebra and its Applications 428 (2008) 469–478
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

A note on the scaled total least squares problem ☆

Wei Xu [a], Sanzheng Qiao [a,1], Yimin Wei [b,*]

[a] *Department of Computing and Software, McMaster University, Hamilton, Ont., Canada L8S 4K1*
[b] *School of Mathematical Sciences, Fudan University, Shanghai 200433, PR China*

Received 25 September 2006; accepted 8 March 2007
Available online 6 April 2007
Submitted by M.K. Ng

Abstract

In this note, we present two results on the scaled total least squares problem. First, we discuss the relation between the scaled total least squares and the least squares problems. We derive an upper bound for the difference between the scaled total least squares solution and the least squares solution and establish a quantitative relation between the scaled total least squares residual and the least squares residual. Second, we give a perturbation analysis of the scaled total least squares problem. Numerical experiments in comparing our results with existing results are demonstrated.

AMS classification: 15A18; 65F20; 65F25; 65F50

Keywords: Scaled total least squares; Total least squares; Least squares; Perturbation analysis

1. Introduction

The scaled total least squares (STLS) problem is a generalization of the total least squares (TLS) problem. For given $A \in \mathbb{R}^{m\times n}$ $(m > n)$ and $b \in \mathbb{R}^m$, the TLS problem is to find $E \in \mathbb{R}^{m\times n}$ and $r \in \mathbb{R}^m$ solving the problem

☆ The first and second authors are partially supported by the Natural Sciences and Engineering Research Council of Canada. The third author is supported by the National Natural Science Foundation of China under grant 10471027 and Shanghai Education Committee.

* Corresponding author.
E-mail addresses: xuw5@mcmaster.ca (W. Xu), qiao@mcmaster.ca (S. Qiao), ymwei@fudan.edu.cn (Y. Wei).

[1] This work is partially supported by the Shanghai Key Laboratory of Contemporary Applied Mathematics of Fudan University during Sanzheng Qiao's visit.

0024-3795/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2007.03.015

$$\min_{(b-r)\in\text{range}(A+E)} \|[E \quad r]\|_{\text{F}}. \tag{1.1}$$

The STLS generalizes the TLS by introducing a scaling factor. Rao [6] proposed the STLS problem as

$$\min_{(b-r)\in\text{range}(A+E)} \|[E \quad \lambda r]\|_{\text{F}} \quad \text{for } E \in \mathbb{R}^{m\times n} \text{ and } r \in \mathbb{R}^m,$$

where $\lambda > 0$ is a given scalar. Obviously, the TLS is a special case of the STLS when $\lambda = 1$. Alternatively, Paige and Strakoš [5] suggested the formulation:

$$\min_{(\lambda b-r)\in\text{range}(A+E)} \|[E \quad r]\|_{\text{F}}. \tag{1.2}$$

If $[E_{\text{STLS}}\ r_{\text{STLS}}]$ solves the above problem, then the solution x_{STLS} for x in the equation $(A + E_{\text{STLS}})\lambda x = \lambda b - r_{\text{STLS}}$ is called the STLS solution.

As described above, the relation between the STLS and the TLS is obvious. What is the relation between the STLS and the least squares (LS) problem $\min \|b - Ax\|_2$?

In this note, after giving explicit expressions for the STLS solution x_{STLS} in Section 2, we derive an upper bound for $\|x_{\text{STLS}} - x_{\text{LS}}\|_2$, where x_{LS} denotes the LS solution, and establish a relation between the residuals $\bar{r}_{\text{STLS}} = b - Ax_{\text{STLS}}$ and $r_{\text{LS}} = b - Ax_{\text{LS}}$ in Section 3. Then in Section 4 we present a perturbation analysis of the STLS problem. Finally, in Section 5, we demonstrate our numerical experiments in comparing our bounds with existing ones.

2. Solving STLS problem

In this section, we give existence conditions and explicit expressions for the STLS solution. From the formulation (1.2), if x_{STLS} is the solution of (1.2), then λx_{STLS} is the solution of the TLS problem with A and λb.

The following theorem by Wei [10] gives existence conditions and explicit expressions for the TLS solution.

Theorem 2.1 (Theorem 2.2, [10]). *Let*

$$\breve{C} := [A \quad b] = \breve{U}\breve{\Sigma}\breve{V}^{\text{T}} \tag{2.1}$$

be the SVD of $[A \quad b]$, *where* $\breve{\Sigma} = \text{diag}(\sigma_1(\breve{C}), \ldots, \sigma_{n+1}(\breve{C}))$, $\sigma_1(\breve{C}) \geqslant \cdots \geqslant \sigma_{n+1}(\breve{C}) \geqslant 0$ *and* $\breve{U} \in \mathbb{R}^{m\times(n+1)}$ *and* $\breve{V} \in \mathbb{R}^{(n+1)\times(n+1)}$ *have orthonormal columns. Let* $k = \text{rank}(A)$, *then* $\text{rank}(\breve{C}) = k + 1$ *assuming* $b \notin \text{range}(A)$. *Partitioning* $\breve{\Sigma}$, $\breve{U}$, *and* $\breve{V}$ *in* (2.1):

$$\breve{\Sigma} = \begin{bmatrix} \breve{\Sigma}_1 & 0 \\ 0 & \breve{\Sigma}_2 \end{bmatrix} \begin{matrix} k \\ n-k+1 \end{matrix}, \quad \breve{U} = \begin{matrix} [\breve{U}_1 & \breve{U}_2] \\ k & n-k+1 \end{matrix}, \tag{2.2}$$

and

$$\breve{V} = \begin{bmatrix} \breve{V}_{11} & \breve{V}_{12} \\ \breve{v}_{21}^{\text{T}} & \breve{v}_{22}^{\text{T}} \end{bmatrix} \begin{matrix} n \\ 1. \end{matrix} \tag{2.3}$$

If the conditions

(i) $\sigma_k(\breve{C}) > \sigma_{k+1}(\breve{C}) > \sigma_{k+2}(\breve{C}) = \cdots = \sigma_{n+1}(\breve{C}) = 0$,
(ii) $\breve{V}_{11}$ *is of full column rank, or equivalently,* $\breve{v}_{22} \neq 0$,

are satisfied, then

$$\begin{aligned} x_{\mathrm{TLS}} &= (\breve{V}_{11}^{\mathrm{T}})^{\dagger}\breve{v}_{21} = \breve{V}_{11}\breve{v}_{21}/(1-\breve{v}_{21}^{\mathrm{T}}\breve{v}_{21}) \\ &= -\breve{V}_{12}(\breve{v}_{22}^{\mathrm{T}})^{\dagger} = -\breve{V}_{12}\breve{v}_{22}/(1-\breve{v}_{21}^{\mathrm{T}}\breve{v}_{21}) \\ &= (A^{\mathrm{T}}A - \breve{V}_{12}\breve{\Sigma}_2^2\breve{V}_{12}^{\mathrm{T}})^{\dagger}(A^{\mathrm{T}}b - \breve{V}_{12}\breve{\Sigma}_2^2\breve{v}_{22}) \end{aligned}$$

is the minimal norm TLS solution. Moreover, let $q = \breve{v}_{22}/\|\breve{v}_{22}\|_2$, *then*

$$[E_{\mathrm{TLS}} \;\; r_{\mathrm{TLS}}] = \breve{U}_2\breve{\Sigma}_2 q q^{\mathrm{T}}[\breve{V}_{12}^{\mathrm{T}} \;\; \breve{v}_{22}]$$

solves (1.1) *and*

$$\|[E_{\mathrm{TLS}} \;\; r_{\mathrm{TLS}}]\|_{\mathrm{F}} = \sigma_{k+1}(\breve{C}).$$

For the STLS problem, following the formulation (1.2), we consider the SVD

$$C := [A \;\; \lambda b] = U\Sigma V^{\mathrm{T}}, \tag{2.4}$$

where $U \in \mathbb{R}^{m\times(n+1)}$ has orthonormal columns, $V \in \mathbb{R}^{(n+1)\times(n+1)}$ is orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1(C), \ldots, \sigma_{n+1}(C))$, $\sigma_1(C) \geqslant \cdots \geqslant \sigma_{n+1}(C) \geqslant 0$. Applying Theorem 2.1, substituting b in (1.1) with λb, and partitioning Σ, U, and V in the SVD (2.4) of C as $\breve{\Sigma}$, $\breve{U}$, and $\breve{V}$ in (2.2) and (2.3), we can express the STLS solution as

$$\lambda x_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{\dagger}v_{21} = -V_{12}(v_{22}^{\mathrm{T}})^{\dagger} = (A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}(\lambda A^{\mathrm{T}}b - V_{12}\Sigma_2^2 v_{22}), \tag{2.5}$$

provided that $\sigma_k(C) > \sigma_{k+1}(C)$ and $v_{22} \neq 0$.

Thus, the STLS problem can be solved by the SVD using, for example, $\lambda x_{\mathrm{STLS}} = -V_{12}(v_{22}^{\mathrm{T}})^{\dagger}$. In this case, since only V_{12} and v_{22} in V are required, a complete SVD is unnecessary. The SVD can be replaced by any of its approximations as long as a good approximation of the last $n-k+1$ columns of the V matrix can be obtained. For example, the complete orthogonal decomposition (COD) [2] can be used in place of the SVD. In 1993, Van Huffel and Zha proposed a rank revealing ULV decomposition (RRULVD) method [3]. Although such method is more efficient than the SVD method, its accuracy depends on the estimator for the smallest singular value and its corresponding singular vector. In Section 4, we use the RRULVD method for our perturbation analysis.

3. Relating STLS to LS

While the TLS is a special case of the STLS when $\lambda = 1$, the relation between the STLS and LS is not so obvious. In [4], it is shown that $x_{\mathrm{STLS}} = x_{\mathrm{LS}}$ and $\sigma_{k+1}(C)/\lambda = \|r_{\mathrm{LS}}\|_2$ as $\lambda \to 0$. In this section, we present quantitative comparisons between the solutions and residuals of the STLS and the LS. Specifically, we derive upper bounds for $\|x_{\mathrm{STLS}} - x_{\mathrm{LS}}\|_2$ and $\|\bar{r}_{\mathrm{STLS}}\|_2$ in terms of $\|r_{\mathrm{LS}}\|_2$.

Theorem 3.1. *If the existence conditions* (i) *and* (ii) *in Theorem* 2.1 *are satisfied, then*

$$\begin{aligned} \|x_{\mathrm{STLS}} - x_{\mathrm{LS}}\|_2 &\leqslant \rho^2\|V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22}\|_2 + \beta\|x_{\mathrm{STLS}}\|_2 \\ &\leqslant \frac{\rho^2+\beta}{\lambda\|v_{22}\|_2}, \end{aligned}$$

where

$$\rho = \frac{\sigma_{k+1}(C)}{\sigma_k(A)} \quad \textit{and} \quad \beta = \min\left(1, \frac{\rho^2}{1-\rho^2}\right). \tag{3.1}$$

Also, the residual norm

$$\|\bar{r}_{\mathrm{STLS}}\|_2 \leqslant \|r_{\mathrm{LS}}\|_2 + \frac{\rho^2\sigma_k(A)}{\lambda\|v_{22}\|_2}.$$

Proof. First, we show some equalities used in our derivation. Using the partitions of Σ, U, and V in the SVD (2.4) of C, we can verify

$$A^{\mathrm{T}}A = V_{11}\Sigma_1^2V_{11}^{\mathrm{T}} + V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}, \quad \lambda A^{\mathrm{T}}b = V_{11}\Sigma_1^2v_{21} + V_{12}\Sigma_2^2v_{22}, \tag{3.2}$$

and

$$V_{12}^{\mathrm{T}}V_{12} + v_{22}v_{22}^{\mathrm{T}} = I. \tag{3.3}$$

From the generalized inverse theory [8], we have

$$(A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}} = A^{\dagger} \quad \text{and} \quad (I - A^{\dagger}A)A^{\mathrm{T}} = 0. \tag{3.4}$$

Then, using the first equation in (3.4), $x_{\mathrm{LS}} = A^{\dagger}b = (A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}b$, and the second equation in (3.4), we get

$$\begin{aligned}
&x_{\mathrm{STLS}} - x_{\mathrm{LS}} \\
&\quad = (I - A^{\dagger}A)x_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}(A^{\mathrm{T}}A)x_{\mathrm{STLS}} \\
&\qquad - (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - (A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}b \\
&\quad = (A^{\mathrm{T}}A)^{\dagger}[(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})x_{\mathrm{STLS}} - \lambda^{-1}V_{11}\Sigma_1^2v_{21}] - \lambda^{-1}(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2v_{22} \\
&\qquad + (I - A^{\dagger}A)x_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}x_{\mathrm{STLS}}.
\end{aligned}$$

From the first equation in (3.2) and $\lambda x_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{\dagger}v_{21}$ in (2.5), the expression in the square bracket in the above equation:

$$\begin{aligned}
&(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})x_{\mathrm{STLS}} - \lambda^{-1}V_{11}\Sigma_1^2v_{21} \\
&\quad = \lambda^{-1}V_{11}\Sigma_1^2V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{\dagger}v_{21} - \lambda^{-1}V_{11}\Sigma_1^2v_{21} \\
&\quad = 0,
\end{aligned}$$

since, by Theorem 2.1, V_{11} is of full column rank and $V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{\dagger} = I$. Thus

$$x_{\mathrm{STLS}} - x_{\mathrm{LS}} = (I - A^{\dagger}A)x_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22}). \tag{3.5}$$

In the following, we show that the first term in the right side of (3.5) satisfies $\|(I - A^{\dagger}A)x_{\mathrm{STLS}}\|_2 \leqslant \beta\|x_{\mathrm{STLS}}\|_2$, where β is defined in (3.1).

On the one hand, $\|(I - A^{\dagger}A)x_{\mathrm{STLS}}\|_2 \leqslant \|x_{\mathrm{STLS}}\|_2$ since $I - A^{\dagger}A$ is an orthogonal projection. On the other hand, (2.5) and the symmetry of $A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}$ imply that

$$\begin{aligned}
x_{\mathrm{STLS}} &= (A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})x_{\mathrm{STLS}} \\
&= (A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}x_{\mathrm{STLS}}.
\end{aligned}$$

Hence, from the second equation in (3.4),

$$\begin{aligned}
&\|(I - A^{\dagger}A)x_{\mathrm{STLS}}\|_2 \\
&\quad = \|(I - A^{\dagger}A)(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}x_{\mathrm{STLS}}\|_2 \\
&\quad = \|(I - A^{\dagger}A)V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}x_{\mathrm{STLS}}\|_2 \\
&\quad \leqslant \|V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}\|_2 \|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 \|x_{\mathrm{STLS}}\|_2 \\
&\quad \leqslant \sigma_{k+1}^2(C)\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 \|x_{\mathrm{STLS}}\|_2.
\end{aligned}$$

Now, we claim that

$$\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 \leqslant \frac{1}{\sigma_k^2(A) - \sigma_{k+1}^2(C)},$$

then we have $\|(I - A^{\dagger}A)x_{\mathrm{STLS}}\|_2 \leqslant \beta\|x_{\mathrm{STLS}}\|_2$. Indeed, from the first equation in (3.2), $A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}$ is of rank k, so

$$\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 = \frac{1}{\sigma_k(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})}.$$

From Mirsky theorem [7, p. 204], we have

$$\sigma_k(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}) - \sigma_k(A^{\mathrm{T}}A) \geqslant -\|V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}\|_2 \geqslant -\sigma_{k+1}^2(C)$$

and consequently

$$\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 = \frac{1}{\sigma_k(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})} \leqslant \frac{1}{\sigma_k^2(A) - \sigma_{k+1}^2(C)}.$$

For the second term in the right side of (3.5), from $\lambda x_{\mathrm{STLS}} = -V_{12}(v_{22}^{\mathrm{T}})^{\dagger}$ in (2.5) and (3.3), we have

$$\begin{aligned}
&V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22} \\
&\quad = -\lambda^{-1}(V_{12}^{\mathrm{T}}V_{12}(v_{22}^{\mathrm{T}})^{\dagger} + v_{22}) \\
&\quad = -\lambda^{-1}(V_{12}^{\mathrm{T}}V_{12} + v_{22}v_{22}^{\mathrm{T}})v_{22}/(v_{22}^{\mathrm{T}}v_{22}) \\
&\quad = -\lambda^{-1}(v_{22}^{\mathrm{T}})^{\dagger},
\end{aligned}$$

which implies

$$\begin{aligned}
&\|(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22})\|_2 \\
&\quad \leqslant \rho^2\|V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22}\|_2 \\
&\quad = \frac{\rho^2}{\lambda}\|v_{22}^{\dagger}\|_2 = \frac{\rho^2}{\lambda\|v_{22}\|_2}.
\end{aligned} \tag{3.6}$$

Putting things together, we get

$$\|x_{\mathrm{STLS}} - x_{\mathrm{LS}}\|_2 \leqslant \frac{\rho^2}{\lambda\|v_{22}\|_2} + \beta\|x_{\mathrm{STLS}}\|_2 \leqslant \frac{\rho^2 + \beta}{\lambda\|v_{22}\|_2},$$

since $\|\lambda x_{\mathrm{STLS}}\|_2 = \|V_{12}(v_{22}^{\mathrm{T}})^{\dagger}\|_2 \leqslant \|v_{22}^{\dagger}\|_2 = \|v_{22}\|_2^{-1}$.

Finally, using (3.5) and (3.6), we get the residual norm

$$\begin{aligned}&\|b - Ax_{\mathrm{STLS}}\|_2\\ &\leqslant \|b - Ax_{\mathrm{LS}}\|_2 + \|A(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}x_{\mathrm{STLS}} - \lambda^{-1}v_{22})\|_2\\ &\leqslant \|r_{\mathrm{LS}}\|_2 + \frac{\sigma_{k+1}^2(C)}{\lambda\sigma_k(A)}\|v_{22}^{\dagger}\|_2 = \|r_{\mathrm{LS}}\|_2 + \frac{\rho^2\sigma_k(A)}{\lambda\|v_{22}\|_2}.\end{aligned}$$

This completes the proof. □

Since $\sigma_{k+1}(C)/\lambda = \|r_{\mathrm{LS}}\|_2$ as $\lambda \to 0$ [4], we have $(\rho^2 + \beta)/\lambda \to 0$ as $\lambda \to 0$. The above theorem then implies that $x_{\mathrm{STLS}} = x_{\mathrm{LS}}$ and $\|r_{\mathrm{STLS}}\|_2 = \|r_{\mathrm{LS}}\|_2$ when $\lambda \to 0$, which is consistent with the results in [4].

4. Perturbation analysis

Consider the RRULVD method for solving the STLS problem [3]. Let

$$C := [A\ \lambda b] = P_C \begin{bmatrix} L_C & 0 \\ H_C & F_C \end{bmatrix} Q_C^{\mathrm{T}},$$

be an RRULVD of C, where L_C is lower triangular and of order $k+1$ and H_C and F_C are small blocks introduced by rounding errors and approximations. This can also be viewed as a perturbed COD of C. In the RRULVD method, we actually compute the STLS solution of the truncated decomposition:

$$P_C \begin{bmatrix} L_C & 0 \\ 0 & 0 \end{bmatrix} Q_C^{\mathrm{T}} =: [\widehat{A}\ \lambda\hat{b}] = \widehat{C}.$$

In this section, we derive an upper bound for $\|x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}\|_2$, where x_{STLS} and $\hat{x}_{\mathrm{STLS}}$ are the STLS solutions of C and $\widehat{C}$ respectively. Our analysis can be readily applied to the SVD method, where L_C is diagonal, $H_C = 0$, and F_C a small diagonal matrix.

Since H_C and F_C are introduced by rounding errors, we assume that

$$\Delta C := C - \widehat{C} = -P_C \begin{bmatrix} 0 & 0 \\ H_C & F_C \end{bmatrix} Q_C^{\mathrm{T}}$$

is small, specifically,

$$\|H_C\|_2 + \|F_C\|_2 = c\mathbf{u}\|C\|_2 =: \eta, \tag{4.1}$$

where c is a moderate constant and $\mathbf{u}$ is the unit of roundoff.

Before deriving the error bound, it is necessary to verify the existence condition (i) in Theorem 2.1. From (4.1), it follows that

$$\begin{aligned}&\sigma_k(\widehat{A}) - \sigma_{k+1}(\widehat{C})\\ &\quad= \sigma_k(A) - \sigma_{k+1}(C) + \sigma_k(\widehat{A}) - \sigma_k(A) + \sigma_{k+1}(C) - \sigma_{k+1}(\widehat{C})\\ &\quad\geqslant \sigma_k(A) - \sigma_{k+1}(C) - 2\eta.\end{aligned}$$

Thus, if $\sigma_k(A) - \sigma_{k+1}(C) > 2\eta$, then the existence condition $\sigma_k(\widehat{A}) > \sigma_{k+1}(\widehat{C})$ for the perturbed STLS problem is satisfied.

Now, we derive the error bound. Using the SVD (2.4) of C and the partitions of U, Σ, and V, we define

$$E_A := A - U_2\Sigma_2V_{12}^{\mathrm{T}} = U_1\Sigma_1V_{11}^{\mathrm{T}} \quad \text{and} \quad \lambda e_b := \lambda b - U_2\Sigma_2v_{22} = U_1\Sigma_1v_{21}.$$

Then, from (2.5), it can be verified that

$$\lambda x_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{\dagger} v_{21} = \lambda E_A^{\dagger} e_b. \tag{4.2}$$

Note that when $\sigma_k(A) > \sigma_{k+1}(C)$, V_{11} is of full column rank [10], implying that $I = V_{11}^{\dagger} V_{11} = V_{11}^{\mathrm{T}} (V_{11}^{\mathrm{T}})^{\dagger}$. Consequently,

$$E_A x_{\mathrm{STLS}} = U_1 \Sigma_1 V_{11}^{\mathrm{T}} x_{\mathrm{STLS}} = \lambda^{-1} U_1 \Sigma_1 V_{11}^{\mathrm{T}} (V_{11}^{\mathrm{T}})^{\dagger} v_{21} = \lambda^{-1} U_1 \Sigma_1 v_{21} = e_b.$$

Similarly, letting $\widehat{C} = \widehat{U}\widehat{\Sigma}\widehat{V}^{\mathrm{T}}$ be the SVD of $\widehat{C}$, partitioning $\widehat{U}$, $\widehat{\Sigma}$, and $\widehat{V}$ according to (2.2) and (2.3), and defining

$$E_{\hat{A}} := \widehat{A} - \widehat{U}_2 \widehat{\Sigma}_2 \widehat{V}_{12}^{\mathrm{T}} = \widehat{U}_1 \widehat{\Sigma}_1 \widehat{V}_{11}^{\mathrm{T}} \quad \text{and} \quad \lambda e_{\hat{b}} := \lambda \hat{b} - \widehat{U}_2 \widehat{\Sigma}_2 \hat{v}_{22} = \widehat{U}_1 \widehat{\Sigma}_1 \hat{v}_{21},$$

we have the solution

$$\hat{x}_{\mathrm{STLS}} = E_{\hat{A}}^{\dagger} e_{\hat{b}}. \tag{4.3}$$

Comparing the two solutions (4.2) and (4.3), we get

$$\begin{aligned} \widehat{U}_1^{\mathrm{T}} E_{\hat{A}} (x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}) &= \widehat{U}_1^{\mathrm{T}} (E_{\hat{A}} x_{\mathrm{STLS}} - E_{\hat{A}} \hat{x}_{\mathrm{STLS}} + E_A x_{\mathrm{STLS}} - E_A x_{\mathrm{STLS}}) \\ &= \widehat{U}_1^{\mathrm{T}} (E_{\hat{A}} - E_A) x_{\mathrm{STLS}} - \widehat{U}_1^{\mathrm{T}} (e_{\hat{b}} - e_b). \end{aligned}$$

Taking the norm on the both sides, we obtain

$$\begin{aligned} \sigma_k(E_{\hat{A}}) \|x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}\|_2 &\leqslant \|\widehat{U}_1^{\mathrm{T}} E_{\hat{A}} (x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}})\|_2 \\ &\leqslant \|\widehat{U}_1^{\mathrm{T}} (E_{\hat{A}} - E_A)\|_2 \|x_{\mathrm{STLS}}\|_2 + \|\widehat{U}_1^{\mathrm{T}} (e_{\hat{b}} - e_b)\|_2. \end{aligned} \tag{4.4}$$

Obviously, from $E_{\hat{A}} = \widehat{A} - \widehat{U}_2 \widehat{\Sigma}_2 \widehat{V}_{12}^{\mathrm{T}}$, we get

$$\sigma_k(E_{\hat{A}}) \geqslant \sigma_k(\widehat{A}) - \|\widehat{U}_2 \widehat{\Sigma}_2 \widehat{V}_{12}^{\mathrm{T}}\|_2 \geqslant \sigma_k(\widehat{A}) - \sigma_{k+1}(\widehat{C}) \geqslant \sigma_k(A) - \sigma_{k+1}(C) - 2\eta, \tag{4.5}$$

since $\mathrm{rank}(E_{\hat{A}}) = k$. Furthermore, we have

$$\begin{aligned} \|\widehat{U}_1^{\mathrm{T}} (E_{\hat{A}} - E_A)\|_2 &= \|\widehat{U}_1^{\mathrm{T}} (\widehat{A} - A - \widehat{U}_2 \widehat{\Sigma}_2 \widehat{V}_{12}^{\mathrm{T}} + U_2 \Sigma_2 V_{12}^{\mathrm{T}})\|_2 \\ &\leqslant \|\widehat{A} - A\|_2 + \|\widehat{U}_1^{\mathrm{T}} (\widehat{U}_2 \widehat{\Sigma}_2 \widehat{V}_{12}^{\mathrm{T}} + U_2 \Sigma_2 V_{12}^{\mathrm{T}})\|_2 \\ &\leqslant \|\widehat{C} - C\|_2 + \|\widehat{U}_1^{\mathrm{T}} U_2 \Sigma_2 V_{12}^{\mathrm{T}}\|_2 \\ &\leqslant \eta + \sigma_{k+1}(C) \frac{\sigma_{k+1}(\widehat{C})\eta}{\sigma_{k+1}^2(\widehat{C}) - \eta^2} \end{aligned} \tag{4.6}$$

and, similarly,

$$\begin{aligned} \|\widehat{U}_1^{\mathrm{T}} (e_{\hat{b}} - e_b)\|_2 &= \|\widehat{U}_1^{\mathrm{T}} (\hat{b} - b - \lambda^{-1} \widehat{U}_2 \widehat{\Sigma}_2 \hat{v}_{22} + \lambda^{-1} U_2 \Sigma_2 v_{22})\|_2 \\ &\leqslant \|\hat{b} - b\|_2 + \lambda^{-1} \|\widehat{U}_1^{\mathrm{T}} U_2 \Sigma_2 v_{22}\|_2 \\ &\leqslant \eta + \lambda^{-1} \sigma_{k+1}(C) \frac{\sigma_{k+1}(\widehat{C})\eta}{\sigma_{k+1}^2(\widehat{C}) - \eta^2}, \end{aligned} \tag{4.7}$$

since $\|\widehat{U}_1^{\mathrm{T}} U_2\|_2 \leqslant \sigma_{k+1}(\widehat{C}) \|H_C\|_2 / (\sigma_{k+1}^2(\widehat{C}) - \|F_C\|_2^2)$ [1, Corollary 2.5] and $\|H_C\|_2, \|F_C\|_2 \leqslant \eta$. Applying the above three inequalities (4.5), (4.6), and (4.7), from (4.4), we obtain

$$\begin{aligned}
&\|x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}\|_2 \\
&\leqslant \frac{1}{\sigma_k(E_{\hat{A}})}(\|\widehat{U}_1^T(E_{\hat{A}} - E_A)\|_2\|x_{\mathrm{STLS}}\|_2 + \|\widehat{U}_1^T(e_{\hat{b}} - e_b)\|_2) \\
&\leqslant \frac{\eta((1+\|x_{\mathrm{STLS}}\|_2)(\sigma_{k+1}^2(\widehat{C}) - \eta^2) + \sigma_{k+1}(\widehat{C})\sigma_{k+1}(C)(\|x_{\mathrm{STLS}}\|_2 + \lambda^{-1}))}{(\sigma_k(A) - \sigma_{k+1}(C) - 2\eta)(\sigma_{k+1}^2(\widehat{C}) - \eta^2)}.
\end{aligned}$$

The above argument is valid for any sufficiently small perturbation ΔC. Ignoring η^2, we have the following theorem.

Theorem 4.1. *Let $C = [A\ \lambda b]$. Suppose that $\widehat{C} = C + \Delta C =: [\widehat{A}\ \lambda\hat{b}]$ and $\|\Delta C\|_2 \approx c\mathbf{u}\|C\|_2 =: \eta \ll 1$, where c is a moderate constant and $\mathbf{u}$ is the unit of roundoff. Let x_{STLS} and $\hat{x}_{\mathrm{STLS}}$ be the STLS solutions corresponding to C and $\widehat{C}$ respectively, then*

$$\|x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}\|_2 \leqslant \eta\frac{(1+\|x_{\mathrm{STLS}}\|_2)\sigma_{k+1}(\widehat{C}) + (\|x_{\mathrm{STLS}}\|_2 + \lambda^{-1})\sigma_{k+1}(C)}{(\sigma_k(A) - \sigma_{k+1}(C) - 2\eta)\sigma_{k+1}(\widehat{C})},$$

provided that $\sigma_k(A) - \sigma_{k+1}(C) > 2\eta$.

This theorem shows that if the perturbation $\eta = \|\Delta C\|_2$ is small, we can expect a small error $\|x_{\mathrm{STLS}} - \hat{x}_{\mathrm{STLS}}\|_2$ as long as $\sigma_k(A)$ and $\sigma_{k+1}(C)$ are not very close.

5. Numerical experiments

In this section, we compare our bounds given by Theorems 3.1 and 4.1 with their existing counterparts given in [4] and [9]. The experiments were carried out in MATLAB. As we shall see, our numerical experiments have shown that our bounds are very close to the existing ones. It is probably very difficult, if possible, to prove which ones are superior. Moreover, the bounds in [9] are valid for the differences between the TLS and LS, whereas our bounds are valid for more general STLS and LS. Thus we compare the bounds in [9] with our bounds for the special case when $\lambda = 1$.

Our test matrix is the matrix of rank k that is closest in the matrix 2-norm to a randomly generated matrix with entries uniformly distributed on [0, 1]. Our test right-hand side vector is a vector with entries uniformly distributed on [0, 1]. The STLS solution is given by $x_{\mathrm{STLS}} = -\lambda^{-1}V_{12}(v_{22}^{\mathrm{T}})^{\dagger}$ and the LS solution by $x_{\mathrm{LS}} = A^{\dagger}b$.

Example 1. Theorem 3.1 gives two bounds for $\|x_{\mathrm{STLS}} - x_{\mathrm{LS}}\|_2$. As shown in the results, the second one is slightly larger than the first, but much simpler. In [9, (3.4)], Wei gives the bound, using our notations,

$$\|x_{\mathrm{TLS}} - x_{\mathrm{LS}}\|_2 \leqslant \rho^2\sqrt{\|x_{\mathrm{TLS}}\|_2^2 + 1} + \rho\|x_{\mathrm{TLS}}\|_2.$$

To compare, we set $\lambda = 1$ in our bounds. Table 1 shows that our bounds are only slightly larger than Wei's bound. However, Wei's bound is implicit in that it involves the solution norm $\|x_{\mathrm{TLS}}\|_2$, whereas our second bound is simpler and can be obtained without $\|x_{\mathrm{STLS}}\|_2$. Moreover, our bounds are more general in that they are for STLS with TLS as a special case when $\lambda = 1$.

Example 2. Paige and Strakoš [4] give an expression for $\|\bar{r}_{\mathrm{STLS}}\|_2$ [4, (4.13)] and an approximation for $\|r_{\mathrm{LS}}\|_2$ [4, (4.14)], from which, using our notations, we have

Table 1
Comparison of the bound in [9] and the first and second bounds in Theorem 3.1

Size	Rank	$\|x_{\mathrm{TLS}} - x_{\mathrm{LS}}\|_2$	Bound in [9]	First bound	Second bound
200×100	80	1.4094	3.6853	3.7912	4.0397
400×300	100	0.2455	1.0248	1.2000	1.7626
800×700	600	1.3091	3.8888	4.1931	4.4009

Table 2
Comparison of the estimate for $\|\bar{r}_{\mathrm{STLS}}\|_2 - \|r_{\mathrm{LS}}\|_2$ from [4] and the bound in Theorem 3.1

λ	$\|\bar{r}_{\mathrm{STLS}}\|_2 - \|r_{\mathrm{LS}}\|_2$	Estimate from [4]	Bound in Theorem 3.1
0.005	1.111E−9	2.221E−9	2.302E−3
0.5	1.627E−1	2.816E−1	2.405E+0
1	1.605E+0	2.298E+0	4.703E+0
5	3.012E+0	4.169E+0	6.363E+0
10	3.196E+0	4.250E+0	6.454E+0

Table 3
Comparison of the bounds for perturbation in the TLS solution

Rank	$\|x_{\mathrm{TLS}} - \hat{x}_{\mathrm{TLS}}\|_2$	Bound in [9, (7.9)]	Bound in Theorem 4.1
500	4.787E−4	4.436E+1	4.254E−1
600	9.253E−5	7.083E+1	8.840E−2
800	2.569E−5	1.539E+2	2.770E−2

$$\|\bar{r}_{\mathrm{STLS}}\|_2 - \|r_{\mathrm{LS}}\|_2 \approx \lambda^{-1} \left| \sigma_{k+1}(C)\sqrt{1 + \|\lambda z\|_2^2} - \sigma_{k+1}(C)\sqrt{1 + \|\lambda x_{\mathrm{LS}}\|_2^2} \right|,$$

where $z = (A^{\mathrm{T}}A - \sigma_{k+1}^2(C)I)^{-1}(A^{\mathrm{T}}b)$. To compare the above estimate with our bound in Theorem 3.1, we generated random A and b and used various values of λ. Table 2 shows the results for a 500×400 random matrix A with rank 350.

Table 2 shows that our bound for $\|\bar{r}_{\mathrm{STLS}}\|_2 - \|r_{\mathrm{LS}}\|_2$ given in Theorem 3.1 is about 1.5 times as large as the estimate from [4] when λ is larger than 1, which indicates that our bound is quite close for $\lambda > 1$. When λ is less than 1, our bound is larger, which means that our bound does not converge to zero as fast as the one from [4]. However, the evaluation of the expression for $\|\bar{r}_{\mathrm{STLS}}\|_2$ in [4] involves solving for z in the system $(A^{\mathrm{T}}A - \sigma_{k+1}^2(C)I)z = A^{\mathrm{T}}b$. Whereas our bound for $\|\bar{r}_{\mathrm{STLS}}\|_2 - \|r_{\mathrm{LS}}\|_2$ given in Theorem 3.1 can be readily obtained.

Example 3. In this example, we compare our STLS solution perturbation bounds given in Theorem 4.1 for $\lambda = 1$ with the TLS solution perturbation bound [9, (7.9)]

$$\|x_{\mathrm{TLS}} - \hat{x}_{\mathrm{TLS}}\|_2 \leqslant \frac{6(\eta + \sigma_{k+1}(C))}{\sigma_k(A) - \sigma_{k+1}(C)} \sqrt{\|x_{\mathrm{TLS}}\|_2^2 + 1}.$$

We generated random matrices A of size 1000×800 with various ranks k and random vectors b. Then we constructed random perturbation matrices ΔC such that $\|\Delta C\|_2 = \eta$, where η was set to $(\sigma_k(A) - \sigma_{k+1}([A, b]))/6$. A typical value of η was 0.0867.

The results in Table 3 show that our bound is tighter than [9, (7.9)]. This can be explained by the fact that our bound has η as a factor. For the same reason, our experiments also show that for small $\eta \approx c\mathbf{u}\|C\|_2$, our bound gives close estimate for the perturbation in the solution.

6. Conclusion

In this paper, we first present quantitative relations between the scaled total least squares and least squares solutions and residuals. Theorem 3.1 shows that the two solutions and two residuals equal when $\lambda \to 0$. They can be very different when λ is not small and $\|v_{22}\|_2$ is small. Second, we give a perturbation analysis of the scaled total least squares problem. Theorem 4.1 shows that the solution of the perturbed problem is close to the original problem if the perturbation is small and $\sigma_{k+1}(C)$ and $\sigma_k(A)$ are not close to each other. Finally, our numerical experiments demonstrate that our bounds are competitive with existing bounds.

References

[1] R.D. Fierro, J.R. Bunch, Bounding the subspaces from rank revealing two-sided orthogonal decompositions, SIAM J. Matrix Anal. Appl. 16 (1995) 743–759.
[2] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[3] S. Van Huffel, H. Zha, An efficient total least squares algorithm based on a rank-revealing two-sided orthogonal decomposition, Numer. Algorithm 4 (1993) 101–133.
[4] C.C. Paige, Z. Strakoš, Bounds for the least squares distance using scaled total least squares, Numer. Math. 91 (2002) 93–115.
[5] C.C. Paige, Z. Strakoš, Scaled total least squares fundamentals, Numer. Math. 91 (2002) 117–146.
[6] B.D. Rao, Unified treatment of ls, tls and truncated svd methods using a weighted tls framework, in: S. Van Huffel (Ed.), Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling, SIAM, Philadelphia, PA, 1997, pp. 11–20.
[7] G.W. Stewart, J. Sun, Matrix Perturbation Theory, Academic Press, San Diego, CA, 1990.
[8] G. Wang, Y. Wei, S. Qiao, Generalized Inverses: Theory and Computations, Science Press, Beijing, New York, 2004.
[9] M. Wei, Algebraic relations between the total least squares and lest squares problems with more than one solution, Numer. Math. 62 (1992) 123–148.
[10] M. Wei, The analysis for the total least square problem with more than one solution, SIAM J. Matrix Anal. Appl. 13 (1992) 746–763.