An Analysis of Rank-Deficient Scaled Total Least Squares Problem

Wei Xu[1] Yimin Wei[2] Sanzheng Qiao[3]

[1] Department of Computing and Software, McMaster University
Hamilton, Ont. L8S 4L7, Canada. xuw5@mcmaster.ca

[2] Department of Mathematics, Fudan University
Shanghai 200433, P. R. of China. ymwei@fudan.edu.cn

[3] Department of Computing and Software, McMaster University
Hamilton, Ont. L8S 4L7, Canada. qiao@mcmaster.ca

Abstract

In this paper, we study the scaled total least squares problems of rank-deficient linear systems. We present a solution for rank-deficient scaled total least squares and discuss the relation between scaled total least squares and least squares.

AMS Subject Classification(1991): 15A18, 65F20, 65F25, 65F50.
Keywords: Scaled total least squares, total least squares, least squares, rank-deficient.

1 Introduction

The least squares (LS) problem is to find $\mathbf{x}$ to minimize $\|\mathbf{b} - A\mathbf{x}\|_2$ for a given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ and a $\mathbf{b} \in \mathbb{R}^m$. Let the residual $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, the least squares problem can be recast as

$$\min_{(\mathbf{b}-\mathbf{r})\in \mathrm{range}(A)} \|\mathbf{r}\|_2 \qquad \text{for} \quad \mathbf{r} \in \mathbb{R}^m. \tag{1.1}$$

This formulation shows that the errors only occur on the vector $\mathbf{b}$. When $\mathbf{b} \in \mathrm{range}(A)$, (1.1) is solved by $\mathbf{r}_{\mathrm{LS}} = 0$. So, we assume $\mathbf{b} \notin \mathrm{range}(A)$ throughout this paper. The total least squares (TLS) problem allows errors to present in both $\mathbf{b}$ and A:

$$\min_{(\mathbf{b}-\mathbf{r})\in \mathrm{range}(A+E)} \|[E\ \mathbf{r}]\|_{\mathrm{F}} \qquad \text{for} \quad E \in \mathbb{R}^{m\times n} \quad \text{and} \quad \mathbf{r} \in \mathbb{R}^m. \tag{1.2}$$

When A is of full column rank, $\mathrm{rank}(A) = n$, the LS solution is unique and given by $\mathbf{x}_{\mathrm{LS}} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\mathbf{b}$. When A is rank-deficient, $\mathrm{rank}(A) = k < n$, the LS solution is not unique. The minimal 2-norm solution can be obtained by using the singular value decomposition (SVD) as follows. Suppose that

$$A = \widehat{U} \begin{bmatrix} \widehat{\Sigma} & 0 \\ 0 & 0 \end{bmatrix} \widehat{V}^{\mathrm{T}} \tag{1.3}$$

is the SVD of A, where $\widehat{U} \in \mathbb{R}^{m\times m}$ and $\widehat{V} \in \mathbb{R}^{n\times n}$ are orthogonal and $\widehat{\Sigma} = \mathrm{diag}(\widehat{\sigma}_1, ..., \widehat{\sigma}_k)$, $\widehat{\sigma}_1 \geq \cdots \geq \widehat{\sigma}_k > 0$. The minimal norm LS solution is given by

$$\mathbf{x}_{\mathrm{LS}} = A^{\dagger}\mathbf{b} \quad \text{where} \quad A^{\dagger} = \widehat{V} \begin{bmatrix} \widehat{\Sigma}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \widehat{U}^{\mathrm{T}}$$

and the vector

$$\mathbf{r}_{\mathrm{LS}} = \mathbf{b} - A\mathbf{x}_{\mathrm{LS}} = (I - AA^{\dagger})\mathbf{b}$$

solves (1.1).

For the TLS problem, we consider the SVD:

$$[A \;\; \mathbf{b}] = \breve{U}\breve{\Sigma}\breve{V}^{\mathrm{T}}, \tag{1.4}$$

where $\breve{\Sigma} = \mathrm{diag}(\breve{\sigma}_1, ..., \breve{\sigma}_{n+1})$, $\breve{\sigma}_1 \geq \cdots \geq \breve{\sigma}_{n+1} \geq 0$ and $\breve{U} \in \mathbb{R}^{m\times(n+1)}$ and $\breve{V} \in \mathbb{R}^{(n+1)\times(n+1)}$ have orthonormal columns, and partition

$$\breve{U} = \begin{array}{cc} [\breve{U}_1 & \breve{\mathbf{u}}_{n+1}] \\ n & 1 \end{array} \quad \text{and} \quad \breve{V} = \begin{array}{cc} \begin{bmatrix} \breve{V}_{11} & \breve{\mathbf{v}}_{12} \\ \breve{\mathbf{v}}_{21}^{\mathrm{T}} & \breve{\nu}_{22} \end{bmatrix} & \begin{array}{c} n \\ 1 \end{array} \\ \begin{array}{cc} n & \;\;1 \end{array} & \end{array}.$$

If $\widehat{\sigma}_n > \breve{\sigma}_{n+1}$, which implies that $\widehat{\sigma}_n > 0$ or A is of full column rank, then the matrix

$$[E_{\mathrm{TLS}} \;\; \mathbf{r}_{\mathrm{TLS}}] = -\breve{\sigma}_{n+1}\breve{\mathbf{u}}_{n+1}[\breve{\mathbf{v}}_{12}^{\mathrm{T}} \;\; \breve{\nu}_{22}]$$

solves (1.2) and

$$\mathbf{x}_{\mathrm{TLS}} = -(A^{\mathrm{T}}A - \breve{\sigma}_{n+1}^2 I_n)^{-1}A^{\mathrm{T}}\mathbf{b} = -\breve{\mathbf{v}}_{12}/\breve{\nu}_{22}$$

is the unique solution to $(A + E_{\mathrm{TLS}})\mathbf{x} = \mathbf{b} + \mathbf{r}_{\mathrm{TLS}}$ [2, Page 598].

If $\widehat{\sigma}_n = \breve{\sigma}_{n+1}$, then the solution to the TLS problem may still exist, although it may not be unique. Wei [10] considered the minimal norm TLS solution for a general case when

$$\breve{\sigma}_p > \breve{\sigma}_{p+1} = \cdots = \breve{\sigma}_q > \breve{\sigma}_{q+1} \geq \cdots \geq \breve{\sigma}_{n+1} \geq 0,$$

for some integers $1 \leq p \leq n$ and $q > p$. For convenience, we restate the theorem for $p = k$ and $q = k + 1$.

Theorem 1.1 *[10, Theorem 2.2] Partitioning $\breve{\Sigma}$, $\breve{U}$, and $\breve{V}$ in (1.4):*

$$\breve{\Sigma} = \begin{array}{cc} \begin{bmatrix} \breve{\Sigma}_1 & 0 \\ 0 & \breve{\Sigma}_2 \end{bmatrix} & \begin{array}{c} k \\ n-k+1 \end{array} \\ \begin{array}{cc} k & n-k+1 \end{array} & \end{array}, \quad \breve{U} = \begin{array}{cc} [\breve{U}_1 & \breve{U}_2 \;] \\ k & n-k+1 \end{array}, \tag{1.5}$$

and

$$\check{V} = \begin{array}{cc} \left[\begin{array}{cc} \check{V}_{11} & \check{V}_{12} \\ \check{\mathbf{v}}_{21}^{\mathrm{T}} & \check{\mathbf{v}}_{22}^{\mathrm{T}} \end{array}\right] & \begin{array}{c} n \\ 1 \end{array} \\ \begin{array}{cc} k & n-k+1 \end{array} & \end{array}, \tag{1.6}$$

if

$$\check{\sigma}_k > \check{\sigma}_{k+1} > \check{\sigma}_{k+2} \geq \cdots \geq \check{\sigma}_{n+1} \geq 0,$$

then $\check{V}_{11}$ *is of full column rank,* $\check{\mathbf{v}}_{22} \neq 0$*, and*

$$\begin{aligned} \mathbf{x}_{TLS} &= (\check{V}_{11}^{\mathrm{T}})^{\dagger}\check{\mathbf{v}}_{21} = \check{V}_{11}\check{\mathbf{v}}_{21}/(1 - \check{\mathbf{v}}_{21}^{\mathrm{T}}\check{\mathbf{v}}_{21}) \\ &= -\check{V}_{12}(\check{\mathbf{v}}_{22}^{\mathrm{T}})^{\dagger} = -\check{V}_{12}\check{\mathbf{v}}_{22}/(1 - \check{\mathbf{v}}_{21}^{\mathrm{T}}\check{\mathbf{v}}_{21}) \\ &= (A^{\mathrm{T}}A - \check{V}_{12}\check{\Sigma}_2^2\check{V}_{12}^{\mathrm{T}})^{\dagger}(A^{\mathrm{T}}\mathbf{b} - \check{V}_{12}\check{\Sigma}_2^2\check{\mathbf{v}}_{22}) \end{aligned} \tag{1.7}$$

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

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

solves (1.2) and

$$\|[E_{TLS} \;\; \mathbf{r}_{TLS}]\|_{\mathrm{F}} = \check{\sigma}_{k+1}.$$

We refer the details of LS to [3] and TLS to [8]. The problems of LS and TLS can be unified by introducing a scaling parameter into the TLS problem. Rao [6] proposed

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

where $\lambda > 0$ is a given scalar. Paige and Strakoš [5] suggested a slightly different but equivalent formulation:

$$\min_{(\lambda\mathbf{b}-\mathbf{r}) \in \mathrm{range}(A+E)} \|[E \;\; \mathbf{r}]\|_{\mathrm{F}}. \tag{1.8}$$

If $[E_{\mathrm{STLS}} \;\; \mathbf{r}_{\mathrm{STLS}}]$ solves the above problem (1.8), then the solution $\mathbf{x}_{\mathrm{STLS}}$ for $\mathbf{x}$ in $(A + E_{\mathrm{STLS}})\lambda\mathbf{x} = \lambda\mathbf{b} - \mathbf{r}_{\mathrm{STLS}}$ is called the scaled total least squares (STLS) solution. In this paper, we adopt the formulation (1.8) by Paige and Strakoš.

Obviously, when $\lambda = 1$, the STLS (1.8) reduces to TLS. It is shown in [5] that $\mathbf{x}_{\mathrm{STLS}}$ approaches $\mathbf{x}_{\mathrm{LS}}$ as $\lambda \to 0$. In the STLS literatures [4, 5, 6], it is assumed that A is of full column rank. This paper presents the STLS solution when A is rank-deficient. The rest of the paper is organized as follows. In Section 2, we analyze the STLS when A is rank-deficient. Then in section 3, we relate STLS to LS.

2 Solving Rank-Deficient STLS

Following the STLS formulation (1.8), we denote

$$C := [A \;\; \lambda\mathbf{b}] = U\Sigma V^{\mathrm{T}}, \tag{2.1}$$

where $U \in \mathbb{R}^{m\times(n+1)}$ has orthonormal columns, $V \in \mathbb{R}^{(n+1)\times(n+1)}$ orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_{n+1})$, $\sigma_1 \geq \cdots \geq \sigma_{k+1} > \sigma_{k+2} = \cdots = \sigma_{n+1} = 0$. From Theorem 1.1, if $\sigma_k > \sigma_{k+1} > 0$, then, substituting $\mathbf{b}$ in (1.2) with $\lambda\mathbf{b}$, we can obtain the minimal norm STLS solution by applying the minimal norm TLS solution (1.7). Specifically, $\lambda\mathbf{x}_{\text{STLS}} = \mathbf{x}_{\text{TLS}}$. When does condition $\sigma_k > \sigma_{k+1}$ hold? The following theorem gives a necessary and sufficient condition for $\sigma_{k+1} = \widehat{\sigma}_k$. The interlacing property says that $\sigma_k \geq \widehat{\sigma}_k \geq \sigma_{k+1}$. Thus, $\widehat{\sigma}_k \neq \sigma_{k+1}$ implies $\sigma_k > \sigma_{k+1}$, which is what we need for applying Theorem 1.1 to STLS.

Theorem 2.1 *Suppose that A has the singular values*

$$\widehat{\sigma}_1 \geq \cdots \geq \widehat{\sigma}_j > \widehat{\sigma}_{j+1} = \cdots = \widehat{\sigma}_k > \widehat{\sigma}_{k+1} = \cdots = \widehat{\sigma}_n = 0$$

for some $j < k$ and $\widehat{U} = [\widehat{\mathbf{u}}_1, ..., \widehat{\mathbf{u}}_m]$ is a column partition. Let

$$\widehat{U}_k = [\widehat{\mathbf{u}}_{j+1}, ..., \widehat{\mathbf{u}}_k], \quad \rho := \|\mathbf{r}_{LS}\|_2, \quad \alpha_i := \widehat{\mathbf{u}}_i^{\mathrm{T}}\mathbf{b}, \quad \textit{for } i = 1, ..., k,$$

and

$$\psi(\sigma) = \lambda^2\rho^2 - \sigma^2 - \lambda^2\sigma^2 \sum_{i=1}^{j} \frac{\alpha_i^2}{\widehat{\sigma}_j^2 - \sigma^2}, \tag{2.2}$$

then

$$\sigma_{k+1} = \widehat{\sigma}_k,$$

if and only if

$$\widehat{U}_k^{\mathrm{T}}\mathbf{b} = 0, \quad \textit{and} \quad \psi(\widehat{\sigma}_k) \geq 0.$$

Proof. We construct a matrix

$$N = \widehat{U}^{\mathrm{T}} C \begin{bmatrix} \widehat{V} & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.3}$$

which has the same singular values $\widehat{\sigma}_i$ as C in (2.1). Then, the $n+1$ eigenvalues of $N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I$ are:

$$\sigma_1^2 - \widehat{\sigma}_k^2, ..., \sigma_k^2 - \widehat{\sigma}_k^2,\ \sigma_{k+1}^2 - \widehat{\sigma}_k^2,\ -\widehat{\sigma}_k^2, ..., -\widehat{\sigma}_k^2.$$

From the interlacing property, the first k eigenvalues in the above list are nonnegative and there are exactly $n - k$ negative eigenvalues if and only if $\sigma_{k+1} = \widehat{\sigma}_k$. In the following, we transfrom $N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I$ while keeping the number of the negative eigenvalues. First, to simplify $N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I$, recall that in the LS problem,

$$\mathbf{r}_{\text{LS}} = (I - AA^{\dagger})\mathbf{b} = \widehat{U} \begin{bmatrix} 0 & 0 \\ 0 & I_{m-k} \end{bmatrix} \widehat{U}^{\mathrm{T}}\mathbf{b}.$$

It then follows that

$$\rho := \|\mathbf{r}_{\text{LS}}\|_2 = \|[\widehat{\mathbf{u}}_{k+1}, ..., \widehat{\mathbf{u}}_m]^{\mathrm{T}}\mathbf{b}\|_2.$$

Defining

$$\mathbf{a} = [\alpha_1, ..., \alpha_k]^{\mathrm{T}} := [\widehat{\mathbf{u}}_1, ..., \widehat{\mathbf{u}}_k]^{\mathrm{T}}\mathbf{b} \quad \text{and} \quad \widehat{\mathbf{b}} = [\widehat{\mathbf{u}}_{k+1}, ..., \widehat{\mathbf{u}}_m]^{\mathrm{T}}\mathbf{b},$$

from (2.3) and (2.1), we have

$$N = \widehat{U}^{\mathrm{T}}[A \ \lambda\mathbf{b}] \begin{bmatrix} \widehat{V} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \widehat{\Sigma} & 0 & \lambda\mathbf{a} \\ 0 & 0 & \lambda\widehat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} \widehat{\Sigma} & 0 & \lambda\mathbf{a} \\ 0 & 0 & \lambda\rho \\ 0 & 0 & 0 \end{bmatrix},$$

where H is a Householder matrix such that $H\widehat{\mathbf{b}} = \rho\mathbf{e}_1$. Thus

$$\begin{aligned} N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I &= \begin{bmatrix} \widehat{\Sigma} & 0 & \lambda\mathbf{a} \\ 0 & 0 & \lambda\rho \\ 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \widehat{\Sigma} & 0 & \lambda\mathbf{a} \\ 0 & 0 & \lambda\rho \\ 0 & 0 & 0 \end{bmatrix} - \widehat{\sigma}_k^2 I \\ &= \begin{bmatrix} \widehat{\Sigma}^2 - \widehat{\sigma}_k^2 I_k & 0 & \lambda\widehat{\Sigma}\mathbf{a} \\ 0 & -\widehat{\sigma}_k^2 I_{n-k} & 0 \\ \lambda\mathbf{a}^{\mathrm{T}}\widehat{\Sigma} & 0 & \lambda^2(\rho^2 + \mathbf{a}^{\mathrm{T}}\mathbf{a}) - \widehat{\sigma}_k^2 \end{bmatrix}. \end{aligned}$$

Partitioning

$$\widehat{\Sigma} = \mathrm{diag}(\widehat{\Sigma}_1, \widehat{\sigma}_k I_{k-j}), \quad \text{where } \widehat{\Sigma}_1 = \mathrm{diag}(\widehat{\sigma}_1, ..., \widehat{\sigma}_j),$$

and

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$$

accordingly, we get

$$N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I = \begin{bmatrix} \widehat{\Sigma}_1^2 - \widehat{\sigma}_k^2 I_j & 0 & 0 & \lambda\widehat{\Sigma}_1\mathbf{a}_1 \\ 0 & 0 & 0 & \lambda\widehat{\sigma}_k\mathbf{a}_2 \\ 0 & 0 & -\widehat{\sigma}_k^2 I_{n-k} & 0 \\ \lambda\mathbf{a}_1^{\mathrm{T}}\widehat{\Sigma}_1 & \lambda\widehat{\sigma}_k\mathbf{a}_2^{\mathrm{T}} & 0 & \lambda^2(\rho^2 + \mathbf{a}^{\mathrm{T}}\mathbf{a}) - \widehat{\sigma}_k^2 \end{bmatrix}.$$

Now, it can be verified that the Schur complement of $\widehat{\Sigma}_1^2 - \widehat{\sigma}_k^2 I_j$ is

$$M := \begin{bmatrix} 0 & 0 & \lambda\widehat{\sigma}_k\mathbf{a}_2 \\ 0 & -\widehat{\sigma}_k^2 I_{n-k} & 0 \\ \lambda\widehat{\sigma}_k\mathbf{a}_2^{\mathrm{T}} & 0 & \psi(\widehat{\sigma}_k) + \lambda^2\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 \end{bmatrix}, \tag{2.4}$$

since

$$\begin{aligned} & \lambda^2(\rho^2 + \mathbf{a}^{\mathrm{T}}\mathbf{a}) - \widehat{\sigma}_k^2 - \lambda^2\mathbf{a}_1^{\mathrm{T}}\widehat{\Sigma}_1(\widehat{\Sigma}_1^2 - \widehat{\sigma}_k^2 I_j)^{-1}\widehat{\Sigma}_1\mathbf{a}_1 \\ = \ & \lambda^2\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 + \lambda^2\rho^2 - \widehat{\sigma}_k^2 + \lambda^2\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - \lambda^2\mathbf{a}_1^{\mathrm{T}}\widehat{\Sigma}_1^2(\widehat{\Sigma}_1^2 - \widehat{\sigma}_k^2 I_j)^{-1}\mathbf{a}_1 \\ = \ & \lambda^2\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 + \psi(\widehat{\sigma}_k). \end{aligned}$$

Since $\widehat{\Sigma}_1^2 - \widehat{\sigma}_k^2 I_j$ is positive definite, from Sylvester law of inertia [2, Page 403], the number of the negative eigenvalues of $N^{\mathrm{T}}N - \widehat{\sigma}_k^2 I$ equals the number of the negative eigenvalues of M, which, from (2.4), has exactly $n-k$ negative eigenvalues if and only if

$$M_1 := \left[\begin{array}{cc} 0 & \lambda\widehat{\sigma}_k \mathbf{a}_2 \\ \lambda\widehat{\sigma}_k \mathbf{a}_2^{\mathrm{T}} & \psi(\widehat{\sigma}_k) + \lambda^2 \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 \end{array}\right]$$

is positive semi-definite. It follows from Lemma 3.1 in [5] that M_1 is positive semi-definite if and only if

$$0 = \mathbf{a}_2 = \widehat{U}_k^{\mathrm{T}}\mathbf{b} \quad \text{and} \quad \psi(\widehat{\sigma}_k) \geq 0.$$

This completes the proof. □

The condition $\widehat{\sigma}_k > \sigma_{k+1}$ for the existence of the minimal norm STLS solution requires the singular values of both A and C. This theorem provides the alternative conditions $\widehat{U}_k^{\mathrm{T}}\mathbf{b} \neq 0$ and $\psi(\widehat{\sigma}_k) < 0$ which require only the SVD of A. From this theorem, if $\widehat{U}_k^{\mathrm{T}}\mathbf{b} \neq 0$ or $\psi(\widehat{\sigma}_k) < 0$, then we can apply Theorem 1.1 to STLS. For example, if Σ, U, and V in (2.1) are partitioned as in (1.5) and (1.6), then, from (1.7), the minimal norm STLS solution $\mathbf{x}_{\mathrm{STLS}}$ can be given by

$$\lambda\mathbf{x}_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{\dagger}\mathbf{v}_{21} = -V_{12}(\mathbf{v}_{22}^{\mathrm{T}})^{\dagger} = (A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}(\lambda A^{\mathrm{T}}\mathbf{b} - V_{12}\Sigma_2^2\mathbf{v}_{22}). \tag{2.5}$$

Moreover, let $\mathbf{q} = \mathbf{v}_{22}/\|\mathbf{v}_{22}\|_2$, then

$$[E_{\mathrm{STLS}} \quad \mathbf{r}_{\mathrm{STLS}}] = U_2\Sigma_2\mathbf{q}\mathbf{q}^{\mathrm{T}}[V_{12}^{\mathrm{T}} \quad \mathbf{v}_{22}]$$

solves (1.8) and

$$\|[E_{\mathrm{STLS}} \quad \mathbf{r}_{\mathrm{STLS}}]\|_{\mathrm{F}} = \sigma_{k+1}.$$

Finally, we conclude this section by presenting two properties of σ_{k+1}.

Corollary 2.2 *If $\widehat{U}_k^{\mathrm{T}}\mathbf{b} \neq 0$ or $\psi(\widehat{\sigma}_k) < 0$, then σ_{k+1} is the smallest positive solution for σ in $\psi(\sigma) = 0$ defined in (2.2).*

Proof. On the one hand, the matrix N defined in (2.3) has the singular values $\sigma_1 \geq \cdots \geq \sigma_{k+1}$. Thus, σ_{k+1} is the smallest positive solution for σ in the equation $\det(N^{\mathrm{T}}N - \sigma^2 I) = 0$.

On the other hand, we consider

$$N^{\mathrm{T}}N - \sigma^2 I = \left[\begin{array}{ccc} \widehat{\Sigma}^2 - \sigma^2 I_k & 0 & \lambda\widehat{\Sigma}\mathbf{a} \\ 0 & -\sigma^2 I_{n-k} & 0 \\ \lambda\mathbf{a}^{\mathrm{T}}\widehat{\Sigma} & 0 & \lambda^2(\rho^2 + \mathbf{a}^{\mathrm{T}}\mathbf{a}) - \sigma^2 \end{array}\right].$$

Similar to (2.4), the Schur complement of $\widehat{\Sigma}^2 - \sigma^2 I_k$ is

$$\left[\begin{array}{cc} -\sigma^2 I_{n-k} & 0 \\ 0 & \psi(\sigma) \end{array}\right],$$

since

$$
\begin{aligned}
& \lambda^2(\rho^2 + \mathbf{a}^{\mathrm{T}}\mathbf{a}) - \sigma^2 - \lambda^2 \mathbf{a}^{\mathrm{T}}\widehat{\Sigma}(\widehat{\Sigma}^2 - \sigma^2 I_k)^{-1}\widehat{\Sigma}\mathbf{a} \\
= \; & \lambda^2\rho^2 - \sigma^2 + \lambda^2\mathbf{a}^{\mathrm{T}}\mathbf{a} - \lambda^2\mathbf{a}^{\mathrm{T}}\widehat{\Sigma}^2(\widehat{\Sigma}^2 - \sigma^2 I_k)^{-1}\mathbf{a} \\
= \; & \psi(\sigma).
\end{aligned}
$$

Thus, we have

$$
\det(N^{\mathrm{T}}N - \sigma^2 I) = (-1)^{n-k}\sigma^{2(n-k)}\psi(\sigma)\det(\widehat{\Sigma}^2 - \sigma^2 I_k).
$$

From Theorem 2.1, when $\widehat{U}_k^{\mathrm{T}}\mathbf{b} \neq 0$ or $\psi(\widehat{\sigma}_k) < 0$, we have $\widehat{\sigma}_k > \sigma_{k+1}$. Consequently, $\widehat{\Sigma}^2 - \sigma_{k+1}^2 I_k$ is positive definite. Therefore, σ_{k+1} is the smallest positive solution for σ in the equation $\psi(\sigma) = 0$, because it the smallest positive solution for σ in the equation $\det(N^{\mathrm{T}}N - \sigma^2 I) = 0$. $\square$

Corollary 2.3 *Under the condition that* $\widehat{U}_k^{\mathrm{T}}\mathbf{b} \neq 0$ *or* $\psi(\widehat{\sigma}_k) < 0$, σ_{k+1} *is a monotonically increasing function of* λ.

Proof. From Corollary 2.2, under the given condition, $\psi(\sigma_{k+1}) = 0$. Differentiating

$$
0 = \psi(\sigma_{k+1})/(\lambda^2\sigma_{k+1}^2),
$$

with respect to λ, we get

$$
\begin{aligned}
0 \; &= \; -\frac{2\rho^2\sigma'_{k+1}}{\sigma_{k+1}^3} + \frac{2}{\lambda^3} - \sum_{i=1}^{j} \frac{2\alpha_i^2\sigma_{k+1}\sigma'_{k+1}}{(\widehat{\sigma}_j^2 - \sigma_{k+1}^2)^2} \\
&= \; \frac{2}{\lambda^3} - 2\sigma'_{k+1}\left[\frac{\rho^2}{\sigma_{k+1}^3} + \sigma_{k+1}\sum_{i=1}^{j}\frac{\alpha_i^2}{(\widehat{\sigma}_j^2 - \sigma_{k+1}^2)^2}\right],
\end{aligned}
$$

which impies $\sigma'_{k+1} > 0$, since $\lambda > 0$ and the value of the expression in the square bracket is positive. $\square$

3 Relating STLS to LS

The relation between STLS and TLS is obvious. The TLS problem is a special case of STLS when $\lambda = 1$. In this section, we discuss the relation between STLS and LS. It is shown in [5] that $\mathbf{x}_{\mathrm{STLS}}$ approaches to $\mathbf{x}_{\mathrm{LS}}$ as λ tends to zero when A is of full column rank and $\widehat{U}_k^T\mathbf{b} \neq 0$. In this section, we extend their result to the case when A is rank-deficient.

Theorem 3.1 *If* $\widehat{U}_k\mathbf{b} \neq 0$ *or* $\psi(\widehat{\sigma}_k) < 0$, *then*

$$
\lim_{\lambda \to 0} \mathbf{x}_{STLS} = \mathbf{x}_{LS} \quad \textit{and} \quad \lim_{\lambda \to 0} \frac{\sigma_{k+1}}{\lambda} = \rho.
$$

Proof. We first show that

$$\lim_{\lambda\to 0}\frac{\sigma_{k+1}^2}{\lambda}=0.$$

Indeed, from Corollary 2.2, we have

$$\sigma_{k+1}^2=\lambda^2\left(\rho^2-\sigma_{k+1}^2\sum_{i=1}^{j}\frac{\alpha_i^2}{\widehat{\sigma}_j^2-\sigma_{k+1}^2}\right). \tag{3.1}$$

It follows that $\lim_{\lambda\to 0}(\sigma_{k+1}^2/\lambda)=0$, which implies $\lim_{\lambda\to 0}\sigma_{k+1}^2=0$. Then, noting that $\Sigma_2=\mathrm{diag}(\sigma_{k+1},0,...,0)$ and $(A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}=A^{\dagger}$, from (2.5), we have

$$\begin{aligned}\lim_{\lambda\to 0}\mathbf{x}_{\mathrm{STLS}} &= \lim_{\lambda\to 0}(A^{\mathrm{T}}A-V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}(A^{\mathrm{T}}\mathbf{b}-\lambda^{-1}V_{12}\Sigma_2^2\mathbf{v}_{22})\\ &= (A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}\mathbf{b}\\ &= \mathbf{x}_{\mathrm{LS}}.\end{aligned}$$

Also, from (3.1), we get

$$\lim_{\lambda\to 0}\frac{\sigma_{k+1}}{\lambda}=\lim_{\lambda\to 0}\sqrt{\rho^2-\sigma_{k+1}^2\sum_{i=1}^{j}\frac{\alpha_i^2}{\widehat{\sigma}_j^2-\sigma_{k+1}^2}}=\rho. \qquad \Box$$

In the following, we derive bounds for $\|\mathbf{x}_{\mathrm{STLS}}-\mathbf{x}_{\mathrm{LS}}\|_2$ and the residual norm $\|\mathbf{b}-A\mathbf{x}_{\mathrm{STLS}}\|_2$.

Theorem 3.2 *If $\widehat{U}_k\mathbf{b}\neq 0$ or $\psi(\widehat{\sigma}_k)<0$, then*

$$\begin{aligned}\|\mathbf{x}_{STLS}-\mathbf{x}_{LS}\|_2 &\leq \frac{\sigma_{k+1}^2}{\widehat{\sigma}_k^2}\|V_{12}^{\mathrm{T}}\mathbf{x}_{STLS}-\lambda^{-1}\mathbf{v}_{22}\|_2+\beta\|\mathbf{x}_{STLS}\|_2\\ &\leq \left(\frac{\sigma_{k+1}^2}{\widehat{\sigma}_k^2}+\beta\right)\frac{1}{\lambda\,\|\mathbf{v}_{22}\|_2},\end{aligned}$$

where

$$\beta=\min\left(1,\ \frac{\sigma_{k+1}^2}{\widehat{\sigma}_k^2-\sigma_{k+1}^2}\right). \tag{3.2}$$

Also, the residual norm

$$\|\mathbf{b}-A\mathbf{x}_{STLS}\|_2\leq\rho+\frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k\|\mathbf{v}_{22}\|_2}.$$

Proof. First, we show some equalities used in our derivation. Partitioning Σ, U, and V in the SVD (2.1) of C as $\breve{\Sigma}$, $\breve{U}$, and $\breve{V}$ in (1.5) and (1.6), we can verify

$$A^{\mathrm{T}}A=V_{11}\Sigma_1^2V_{11}^{\mathrm{T}}+V_{12}\Sigma_2^2V_{12}^{\mathrm{T}},\quad \lambda A^{\mathrm{T}}\mathbf{b}=V_{11}\Sigma_1^2\mathbf{v}_{21}+V_{12}\Sigma_2^2\mathbf{v}_{22}. \tag{3.3}$$

and

$$V_{12}^{\mathrm{T}}V_{12} + \mathbf{v}_{22}\mathbf{v}_{22}^{\mathrm{T}} = I. \tag{3.4}$$

From the generalized inverse theory [9], we have

$$(A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}} = A^{\dagger}, \qquad (I - A^{\dagger}A)A^{\mathrm{T}} = 0 \tag{3.5}$$

and

$$\mathbf{x}^{\dagger} = \mathbf{x}^{\mathrm{T}}/(\mathbf{x}^{\mathrm{T}}\mathbf{x}), \quad \mathbf{x} \neq 0. \tag{3.6}$$

Then, using the first equation in (3.5), $\mathbf{x}_{\mathrm{LS}} = A^{\dagger}\mathbf{b} = (A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}\mathbf{b}$ and the second equation in (3.3), we get

$$\begin{aligned}
& \mathbf{x}_{\mathrm{STLS}} - \mathbf{x}_{\mathrm{LS}} \\
= \; & (I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}(A^{\mathrm{T}}A)\mathbf{x}_{\mathrm{STLS}} \\
& -(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - (A^{\mathrm{T}}A)^{\dagger}A^{\mathrm{T}}\mathbf{b} \\
= \; & (A^{\mathrm{T}}A)^{\dagger}[(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}V_{11}\Sigma_1^2\mathbf{v}_{21}] - \lambda^{-1}(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2\mathbf{v}_{22} \\
& +(I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}}.
\end{aligned}$$

From the first equation in (3.3) and $\lambda\mathbf{x}_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{\dagger}\mathbf{v}_{21}$ in (2.5), the expression in the square bracket in the above equation:

$$\begin{aligned}
& (A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}V_{11}\Sigma_1^2\mathbf{v}_{21} \\
= \; & \lambda^{-1}V_{11}\Sigma_1^2V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{\dagger}\mathbf{v}_{21} - \lambda^{-1}V_{11}\Sigma_1^2\mathbf{v}_{21} \\
= \; & 0,
\end{aligned}$$

since $V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{\dagger} = I$, because, applying Theorem 1.1, V_{11} is of full column rank. Thus

$$\mathbf{x}_{\mathrm{STLS}} - \mathbf{x}_{\mathrm{LS}} = (I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}} + (A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}\mathbf{v}_{22}). \tag{3.7}$$

In the following, we show that the first term on the right side of (3.7) satisfies $\|(I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}}\|_2 \leq \beta\|\mathbf{x}_{\mathrm{STLS}}\|_2$, where β is defined in (3.2).

On the one hand, $\|(I-A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}}\|_2 \leq \|\mathbf{x}_{\mathrm{STLS}}\|_2$ 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

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

Hence, from the second equation in (3.5),

$$\begin{aligned}
& \|(I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}}\|_2 \\
= \; & \|(I - A^{\dagger}A)(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}\mathbf{x}_{\mathrm{STLS}}\|_2 \\
= \; & \|(I - A^{\dagger}A)V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}\mathbf{x}_{\mathrm{STLS}}\|_2 \\
\leq \; & \|V_{12}\Sigma_2^2V_{12}^{\mathrm{T}}\|_2\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}\|_2\|\mathbf{x}_{\mathrm{STLS}}\|_2 \\
\leq \; & \sigma_{k+1}^2\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2V_{12}^{\mathrm{T}})^{\dagger}\|_2\|\mathbf{x}_{\mathrm{STLS}}\|_2.
\end{aligned}$$

Now, we claim that

$$\|(A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{\dagger}\|_2 \leq \frac{1}{\widehat{\sigma}_k^2 - \sigma_{k+1}^2},$$

then we have $\|(I - A^{\dagger}A)\mathbf{x}_{\mathrm{STLS}}\|_2 \leq \beta\,\|\mathbf{x}_{\mathrm{STLS}}\|_2$. Indeed, from the first equation in (3.3), $A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}}$ 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, Page 204], we have

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

and consequently

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

For the second term on the right side of (3.7), from $\lambda\mathbf{x}_{\mathrm{STLS}} = -V_{12}(\mathbf{v}_{22}^{\mathrm{T}})^{\dagger}$ in (2.5), (3.4), and (3.6), we have

$$\begin{aligned} & V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}\mathbf{v}_{22} \\ = \; & -\lambda^{-1}(V_{12}^{\mathrm{T}}V_{12}(\mathbf{v}_{22}^{\mathrm{T}})^{\dagger} + \mathbf{v}_{22}) \\ = \; & -\lambda^{-1}(V_{12}^{\mathrm{T}}V_{12} + \mathbf{v}_{22}\mathbf{v}_{22}^{\mathrm{T}})\mathbf{v}_{22}/(\mathbf{v}_{22}^{\mathrm{T}}\mathbf{v}_{22}) \\ = \; & -\lambda^{-1}(\mathbf{v}_{22}^{\mathrm{T}})^{\dagger}, \end{aligned}$$

which implies

$$\begin{aligned} & \|(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}\mathbf{v}_{22})\|_2 \\ \leq \; & \frac{\sigma_{k+1}^2}{\widehat{\sigma}_k^2}\|V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}\mathbf{v}_{22}\|_2 \\ = \; & \frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k^2}\|\mathbf{v}_{22}^{\dagger}\|_2 = \frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k^2\|\mathbf{v}_{22}\|_2}. \end{aligned} \tag{3.8}$$

Putting things together, we get

$$\|\mathbf{x}_{\mathrm{STLS}} - \mathbf{x}_{\mathrm{LS}}\|_2 \leq \frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k^2\|\mathbf{v}_{22}\|_2} + \beta\,\|\mathbf{x}_{\mathrm{STLS}}\|_2 \leq \left(\frac{\sigma_{k+1}^2}{\widehat{\sigma}_k^2} + \beta\right)\frac{1}{\lambda\,\|\mathbf{v}_{22}\|_2}\lambda,$$

since $\|\lambda\mathbf{x}_{\mathrm{STLS}}\|_2 = \|V_{12}(\mathbf{v}_{22}^{\mathrm{T}})^{\dagger}\|_2 \leq \|\mathbf{v}_{22}^{\dagger}\|_2 = \|\mathbf{v}_{22}\|_2^{-1}$.

Finally, using (3.7) and (3.8), we get the residual norm

$$\begin{aligned} & \|\mathbf{b} - A\mathbf{x}_{\mathrm{STLS}}\|_2 \\ \leq \; & \|\mathbf{b} - A\mathbf{x}_{\mathrm{LS}}\|_2 + \|A(A^{\mathrm{T}}A)^{\dagger}V_{12}\Sigma_2^2(V_{12}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} - \lambda^{-1}\mathbf{v}_{22})\|_2 \\ \leq \; & \rho + \frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k}\|\mathbf{v}_{22}^{\dagger}\|_2 = \rho + \frac{\sigma_{k+1}^2}{\lambda\widehat{\sigma}_k\|\mathbf{v}_{22}\|_2}. \end{aligned}$$

□

Conclusion

In this paper, we showed the conditions for the existence of the minimal norm solution for rank-deficient STLS. Our conditions involve only the SVD of the coefficient matrix A. Also, we gave explicit forms of the minimal norm solution for rank-deficient STLS. In Section 3 we showed the difference norm $\|\mathbf{x}_{\text{STLS}} - \mathbf{x}_{\text{LS}}\|$ between an STLS solution and its corresponding LS solution and the STLS residual norm $\|\mathbf{b} - A\mathbf{x}_{\text{STLS}}\|$.

References

[1] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM J. Numer. Anal.* **17**(1980), 883–893.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd Ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[3] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.

[4] Christopher C. Paige and Zdeněk Strakoš. Bounds for the least squares distance using scaled total least squares. *Numer. Math.* **91**(2002) 93–115.

[5] Christopher C. Paige and Zdeněk Strakoš. Scaled total least squares fundamentals. *Numer. Math.* **91**(2002) 117–146.

[6] B. D. Rao. Unified treatment of LS, TLS and Treuncated SVD methoeds using a weihted TLS framework. *Recent Advances in Total Least Squaes Techeniques and Errors-in-Variables Modelling*, edited by S. Van Huffel. SIAM, Philadelpha PA, 1997, 11–20.

[7] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, Inc., 1990.

[8] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem*. SIAM, Philadelphia PA, 1991.

[9] G. Wang, Y. Wei and S. Qiao. *Generalized Inverses: Theory and Computations*. Science Press, Beijing, to appear.

[10] Musheng Wei. The analysis for the total least square problem with more than one solution. *SIAM J. Matrix Anal. Appl.* **13**(1992) 746–763.