An Algorithm for Solving Scaled Total Least Squares Problems

Sanzheng Qiao and Wei Xu
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada L8S 4K1

Yimin Wei
Institute of Mathematics, School of Mathematical Science
Fudan University, Shanghai, P.R. China 200433 and
Key Laboratory of Nonlinear Science (Fudan University), Education of Ministry *

Abstract - *In this paper, we present a rank-revealing two-sided orthogonal decomposition method for solving the STLS problem. An error analysis of the algorithm is given. Our numerical experiments show that this algorithm computes the STLS solution as accurate as the SVD method with less computation.*

Keywords: Scaled total least squares, total least squares, least squares, rank revealing decompositions.

1 Introduction

Given an m-by-n, $m \geq n$, matrix A and an m-vector $\mathbf{b}$, the problem of the least squares (LS) is to find a minimizer $\mathbf{x}$ for

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2.$$

Equivalently, it is to find an m-vector $\mathbf{r}$ for the following minimization problem:

$$\min_{\mathbf{b}-\mathbf{r}\in \text{range}(A)} \|\mathbf{r}\|_2.$$

The problem of the total least squares (TLS) is to find an m-by-n matrix E and an m-vector $\mathbf{r}$ for the following minimization problem:

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

Rao [6] unified the LS and the TLS problems by introducing the scaled total least square (STLS) problem:

$$\min_{(\mathbf{b}-\mathbf{r})\in \text{range}(A+E)} \|[E \;\; \lambda\mathbf{r}]\|_{\text{F}},$$

where λ is a scaling factor. Paige and Strakoš [5] suggested a slightly different but equivalent formulation:

$$\min_{(\lambda\mathbf{b}-\mathbf{r})\in \text{range}(A+E)} \|[E \;\; \text{r}]\|_{\text{F}}. \tag{1}$$

If the pair E_{STLS} and $\mathbf{r}_{\text{STLS}}$ solves the above problem (1), then the solution $\mathbf{x}_{\text{STLS}}$ for $\mathbf{x}$ in the consistent system $(A + E_{\text{STLS}})\lambda\mathbf{x} = \lambda\mathbf{b} - \mathbf{r}_{\text{STLS}}$ is called the STLS solution.

The scaled total least squares formulation unifies LS and TLS in that the STLS reduces to the TLS when $\lambda = 1$ and the STLS solution approaches the LS solution as $\lambda \to 0$ [4].

In the STLS literatures [4, 5, 6], A is assumed to be of full rank. In this paper, we consider the general case when $\text{rank}(A) = k$, $k \leq n$. Let

$$C := [A \;\; \lambda\mathbf{b}] = U\Sigma V^{\text{T}}, \tag{2}$$

be the singular value decomposition (SVD) of C, where U is m-by-$(n+1)$ and has orthonormal columns, V is of order $n+1$ and orthogonal, and $\Sigma = \text{diag}(\sigma_1(C), ..., \sigma_{n+1}(C))$, $\sigma_1(C) \geq \sigma_2(C) \geq \cdots \geq \sigma_{k+1}(C) > \sigma_{k+2}(C) = \cdots = \sigma_{n+1}(C) = 0$. Then we partition U, Σ and V in (2) as:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix},$$

*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 and Shanghai Education Committee.

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \quad V = \begin{bmatrix} V_{11} & V_{12} \\ \mathbf{v}_{21}^{\mathrm{T}} & \mathbf{v}_{22}^{\mathrm{T}} \end{bmatrix}, \tag{3}$$

such that $\Sigma_1 = \mathrm{diag}(\sigma_1(C), \ldots, \sigma_k(C))$, $\Sigma_2 = \mathrm{diag}(\sigma_{k+1}(C), 0, \cdots, 0)$, U_1 and U_2 are respectively the first k columns and last $n+1-k$ columns of U, $V_{11} \in \mathrm{R}^{n\times k}$, $V_{12} \in \mathrm{R}^{n\times(n+1-k)}$, $\mathbf{v}_{21} \in \mathrm{R}^{k\times 1}$, and $\mathbf{v}_{22} \in \mathrm{R}^{(n+1-k)\times 1}$. Accordingly, we denote the SVD of A as

$$A = U_A \begin{bmatrix} \Sigma_A & 0 \\ 0 & 0 \end{bmatrix} V_A^{\mathrm{T}}, \quad U_A = \begin{bmatrix} U_{A1} & U_{A2} \end{bmatrix},$$

where $U_A \in \mathrm{R}^{m\times m}$ and $V_A \in \mathrm{R}^{n\times n}$ are orthogonal, $\Sigma_A = \mathrm{diag}(\sigma_1(A), ..., \sigma_k(A))$, $\sigma_1(A) \geq \cdots \geq \sigma_k(A) > 0$, and U_{A1} and U_{A2} are respectively the first k columns and the last $m-k$ columns of U_A.

The STLS problem can be solved by using the SVD [10]. Specifically, the solution

$$\begin{aligned} & \lambda \mathbf{x}_{\mathrm{STLS}} \\ = \; & -V_{12}(\mathbf{v}_{22}^{\mathrm{T}})^{+} \\ = \; & (V_{11}^{\mathrm{T}})^{+}\mathbf{v}_{21} \\ = \; & (A^{\mathrm{T}}A - V_{12}\Sigma_2^2 V_{12}^{\mathrm{T}})^{+}(\lambda A^{\mathrm{T}}\mathbf{b} - V_{12}\Sigma_2^2\mathbf{v}_{22}), \end{aligned} \tag{4}$$

where $(\mathbf{v}_{22}^{\mathrm{T}})^{+}$ denotes the pseudoinverse of $\mathbf{v}_{22}^{\mathrm{T}}$. As we know, computing the SVD is expensive. In this paper, we present an algorithm for solving the STLS problem using a rank revealing decomposition. This algorithm is more efficient than the SVD method and it is particularly efficient for the STLS problems with same coefficient matrix A but multiple right hand side vectors $\mathbf{b}$. In Section 2, we first describe a complete orthogonal decomposition (COD) [2] to illustrate the ideas behind our algorithm. Then we present a practical algorithm for solving the STLS problem using the rank revealing ULV decomposition (RRULVD) [7]. The the computation of the RRULVD is given in Section 3. A perturbation analysis of our STLS algorithm is given in Section 4 and numerical experiments are presented in Section 5.

2 Main Idea

The STLS solution expression (4) shows that to compute the solution, we need only V_{12} and $\mathbf{v}_{22}$, which, from the partition of V in (3), form the null space and the right singular vector corresponding to the smallest nonzero singular value of the augmented matrix C defined in (2). It is unnecessary to compute all the individual singular values and singular vectors. Now, we consider the complete orthogonal decomposition (COD):

$$C = \bar{P} \begin{bmatrix} \bar{L} & 0 \\ 0 & 0 \end{bmatrix} \bar{Q}^{\mathrm{T}},$$

where $\bar{P} \in \mathrm{R}^{m\times(n+1)}$ has orthonormal columns, $\bar{Q} \in R^{(n+1)\times(n+1)}$ is orthogonal, and $\bar{L}$ is a $(k+1)$-by-$(k+1)$ nonsingular lower triangular matrix. Let $\mathbf{w}$ be the right singular vector corresponding to the smallest nonzero singular value $\sigma_{k+1}(\bar{L})$ of $\bar{L}$ and $\bar{Q} = [\bar{Q}_1\ \bar{Q}_2]$, where $\bar{Q}_1$ and $\bar{Q}_2$ are respectively the first $k+1$ and the last $n-k$ columns of $\bar{Q}$. It is shown in [8] that if $U_{A1}^{\mathrm{T}}\mathbf{b} \neq 0$, V_{11} is of full rank and $\mathbf{v}_{22}$ is a nonzero vector, then we can find a Householder matrix H of order $n-k+1$ such that $\widetilde{Q} = [\bar{Q}_1\mathbf{w}\ \ \bar{Q}_2]H$ and $\widetilde{Q}(n+1,\ 2:n-k+1) = 0$, that is, the last row of $\widetilde{Q}$ has the structure $[\times, 0, ..., 0]$, specifically, $\mathbf{v}_{22}^T H = \|\mathbf{v}_{22}\|_2[1, 0, ..., 0]$, and, from (4), the STLS solution can be explicitly expressed as

$$\lambda \mathbf{x}_{\mathrm{STLS}} = -\widetilde{Q}(1:n, 1)/\widetilde{Q}(n+1, 1).$$

Note that $\widetilde{Q}(n+1, 1) = \|\mathbf{v}_{22}\|_2 \neq 0$, since $\mathbf{v}_{22}$ is a nonzero vector.

Now that we have described a COD method for computing the STLS solution. This method has the following issues to be dealt with. First, the COD is sensitive to perturbations and rounding errors when the matrix is rank deficient. Second, we still need to compute the right singular vector corresponding to the smallest nonzero singular value of C. Third, we may want to check the solution existence condition $\sigma_k(A) > \sigma_{k+1}(C)$, recalling that $\sigma_k(A)$ and $\sigma_{k+1}(C)$ are the smallest nonzero singular values of A and C respectively. To alleviate these problems, we propose a rank revealing ULV decomposition [7] (RRULVD) algorithm, which is an approximation of the COD. The RRULVD of $A \in \mathrm{R}^{m\times n}$ is defined as

$$A = P_A \begin{bmatrix} L_A & \\ H_A & F_A \end{bmatrix} Q_A^{\mathrm{T}}, \tag{5}$$

where L_A and F_A are lower triangular, L_A is of order $k = \mathrm{rank}(A)$, the numerical rank of A, $\|F_A\|_2 \approx \sigma_{k+1}(A) = 0$ and $\|H_A\|_2$ is sufficiently small so that $\|F_A\|_2 + \|H_A\|_2 \approx \sigma_{k+1}(A) = 0$. Thus RRULVD reveals the numerical rank of A. When both $\|H_A\|_2$ and $\|F_A\|_2$ are small, the RRULVD can be viewed as an approximation of the COD of a rank-deficient matrix. In addition, in the next section, we will show that in the computation of the RRULVD of A, we get an estimate for $\sigma_k(A)$. Moreover, the RRULVD can be efficiently updated when a column $\lambda\mathbf{b}$ is appended to A. Also, in updating the decomposition,

we can get an estimate for $\sigma_{k+1}(C)$ and its corresponding right singular vector. Thus, all the information needed for computing the STLS solution and checking the condition $\sigma_k(A) > \sigma_{k+1}(C)$ can be obtained during the computation of the RRULVDs of A and C. Letting

$$C := [A\ \lambda\mathbf{b}] = P_C \begin{bmatrix} L_C & \\ H_C & F_C \end{bmatrix} Q_C^{\mathrm{T}} \qquad (6)$$

be the updated RRULVD after $\lambda\mathbf{b}$ is appended to A, we present the following algorithm. The computation of the RRULVD, the crucial part of the algorithm, is described in the next section.

Algorithm 1 (RRULVD based) *Given a pair A and $\mathbf{b}$, and λ, this algorithm computes the STLS solution $\mathbf{x}_{STLS}$ using the RRULVD.*

1. Compute the RRULVD (5) of A and an estimate of $\sigma_k(A)$;
2. Append $\lambda\mathbf{b}$ to A, update the RRULVD, as in (6), and compute an estimate for $\sigma_{k+1}(C)$ and its corresponding right singular vector $\mathbf{w}$;
3. **if** $(\sigma_k(A) = \sigma_{k+1}(C))$ quit **end**;
4. Partition $Q_C = [Q_{C1}\ Q_{C2}]$ such that Q_{C1} and Q_{C2} contain the first $k+1$ and the last $n-k$ columns of Q_C respectively;
5. Find a Householder matrix H such that $\widetilde{Q} = [Q_{C1}\mathbf{w}\ \ Q_{C2}]H$ and $\widetilde{Q}(n+1,\ 2:n-k+1) = 0$, that is, the last row of $\widetilde{Q}$ has the structure $[\times, 0, ..., 0]$;
6. $\mathbf{x}_{\mathrm{STLS}} = -\lambda^{-1}\widetilde{Q}(1:n,\ 1)/\widetilde{Q}(n+1,\ 1)$.

3 Computing RRULVD

The RRULVD algorithm presented in this section is based on Stewart's method [7]. It is a column updating scheme.

Let us consider one step of the RRULVD algorithm: Update the RRULVD of a matrix when a column is appended to the matrix. Assume that the RRULVD (5) of A is available and a column $\mathbf{a} = \lambda\mathbf{b}$ is appended to A. Let

$$\mathbf{y} = P_A^{\mathrm{T}}\mathbf{a} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix},$$

then, from (5), we have

$$[A\ \mathbf{a}] = P_A \begin{bmatrix} L_A & 0 & \mathbf{y}_1 \\ H_A & F_A & \mathbf{y}_2 \end{bmatrix} \begin{bmatrix} Q_A^{\mathrm{T}} & 0 \\ 0 & 1 \end{bmatrix}. \qquad (7)$$

What we need to do next is to restore the rank revealing triangular structure of the middle matrix, denoted by $\widehat{L}$, on the right side of the above equation (7). It consists of two steps described by the following algorithm. See [9] for details.

Algorithm 2 (Triangularization) *Denote*

$$\widehat{L} = \begin{bmatrix} L_A & 0 & \mathbf{y}_1 \\ H_A & F_A & \mathbf{y}_2 \end{bmatrix}$$

as the middle matrix on the right side of (7), this algorithm triangularizes $\widehat{L}$ using two-side orthogonal transformations.

1. Two sequences of rotations are applied to the both sides of the bottom part of $\widehat{L}$ to eliminate $\mathbf{y}_2$ except its first entry, while keeping the lower triangular structure of the F_A blcok;
2. A sequence of rotations is applied to the columns of $\widehat{L}$ resulted from the previous step to eliminate $\mathbf{y}_1$ and the first entry of $\mathbf{y}_2$ using the diagonal of $\widehat{L}$.

After the above triangularization, we obtain the decomposition (6), where L_C is lower triangular of order $k+1$. However, the rank of C can be either k or $k+1$. To reveal the numerical rank of C, we apply the deflation procedure presented in [7], using the Van Loan's 2-norm condition number estimator [3] to estimate the smallest singular value $\sigma_{k+1}(L_C)$ of L_C in (6) and its corresponding right singular vector $\mathbf{w}$. We refer the details of the deflation to [7] or [9].

It is shown in [1] that the quality of the subspaces obtained by the RRULVD algorithm, which determines the accuracy of the computed STLS solution, depends on the quality of the condition estimator of the lower triangular matrix L_C. We propose the following techniques of improving the approximations of $\sigma_{k+1}(L_C)$ and its corresponding right singular vector $\mathbf{w}$.

We first apply the Van Loan's method [3] to get an approximation $\mathbf{y}$ of the right singular vector of L_C^{T}. Then we solve the linear system $L_C\mathbf{x} = \mathbf{y}$. Now, $\mathbf{w} = \mathbf{x}/\|\mathbf{x}\|_2$ is an improved right singular vector, and an improved singular value $\sigma_{k+1}(L_C)$ can be obtained from $\mathbf{w}$. Since L_C is lower triangular, the overhead introduced by this technique is insignificant comparing with the total cost, while the accuracy is significantly improved.

The accuracy of the computed STLS solution depends on the quality of not only the estimates of the

smallest singular value and its corresponding singular vector, but also the null space of C, measured by the norm of the block H_C in (6). The refinement technique in [7] can be used to improve the accuracy of the null space by reducing the norm of H_C.

In summary, to compute the RRULVD of A, starting with the RRULVD of the first column of A, we append one column of A at a time and update the RRULVD using Algorithm 2 followed by deflation. Then, we append $\lambda\,\mathbf{b}$ to A and update the RRULVD. Refinement may be applied in updating to improve the quality of the null space. Since only one right singular vector and the null space of C are required for computing the STLS solution, updating P_A in (5) is unnecessary when we compute the RRULVD of C. This saves the computational cost significantly when m is much larger than n.

4 Perturbation Analysis

Algorithm 1 first computes the RRULVD:

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

where the blocks H_C and F_C are introduced by rounding errors and approximations. Then the algorithm computes the STLS solution using the truncated RRULVD as an approximation of the COD of C:

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

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

$$E := C - \widehat{C} = -P \begin{bmatrix} 0 & 0 \\ H_C & F_C \end{bmatrix} Q^{\mathrm{T}},$$

is small, specifically,

$$\|H_C\|_2 + \|F_C\|_2 \le c\,u\,\|C\|_2 =: \eta, \tag{8}$$

where c is a moderate constant and u is the unit of roundoff. What is the difference between the solution corresponding to $C = [A\ \lambda\mathbf{b}]$ and that of $\widehat{C} = [\widehat{A}\ \lambda\widehat{\mathbf{b}}]$? In this section, we give an upper bound for the error $\|\mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}}\|_2$, where $\mathbf{x}_{\mathrm{STLS}}$ and $\widehat{\mathbf{x}}_{\mathrm{STLS}}$ denote the solutions corresponding to C and $\widehat{C}$ respectively.

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

$$\begin{aligned} &\|\mathbf{x}_{STLS} - \widehat{\mathbf{x}}_{STLS}\|_2 \\ \le\ & \frac{(\sigma_k(A) + \sigma_{k+1}(C))(\|\mathbf{x}_{STLS}\|_2 + \lambda^{-1}) + \eta}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta}, \end{aligned}$$

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

Proof. Before deriving a bound for $\|\mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}}\|_2$, it is necessary to verify the existence condition. From (8), it follows that

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

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) of C and the partitions (3), we define

$$E_A := A - U_2\Sigma_2 V_{12}^{\mathrm{T}} = U_1\Sigma_1 V_{11}^{\mathrm{T}}$$

and

$$\lambda\,\mathbf{e}_b := \lambda\mathbf{b} - U_2\Sigma_2\mathbf{v}_{22} = U_1\Sigma_1\mathbf{v}_{21}.$$

Then, from (4), it can be verified that

$$\lambda\,\mathbf{x}_{\mathrm{STLS}} = (V_{11}^{\mathrm{T}})^{+}\mathbf{v}_{21} = \lambda E_A^{+}\mathbf{e}_b. \tag{9}$$

Note that when $\sigma_k(A) > \sigma_{k+1}(C)$ V_{11} is of full column rank [8], implying that $I = V_{11}^{+}V_{11} = V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{+}$. Consequently,

$$\begin{aligned} E_A\mathbf{x}_{\mathrm{STLS}} &= U_1\Sigma_1 V_{11}^{\mathrm{T}}\mathbf{x}_{\mathrm{STLS}} \\ &= \lambda^{-1}U_1\Sigma_1 V_{11}^{\mathrm{T}}(V_{11}^{\mathrm{T}})^{+}\mathbf{v}_{21} \\ &= \lambda^{-1}U_1\Sigma_1\mathbf{v}_{21} \\ &= \mathbf{e}_b. \end{aligned}$$

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 (3), and defining

$$E_{\widehat{A}} := \widehat{A} - \widehat{U}_2\widehat{\Sigma}_2\widehat{V}_{12}^{\mathrm{T}} = \widehat{U}_1\widehat{\Sigma}_1\widehat{V}_{11}^{\mathrm{T}}$$

and

$$\lambda\mathbf{e}_{\hat{b}} := \lambda\widehat{\mathbf{b}} - \widehat{U}_2\widehat{\Sigma}_2\widehat{\mathbf{v}}_{22} = \widehat{U}_1\widehat{\Sigma}_1\widehat{\mathbf{v}}_{21},$$

we have the solution

$$\widehat{\mathbf{x}}_{\mathrm{STLS}} = E_{\widehat{A}}^{+}\mathbf{e}_{\hat{b}}. \tag{10}$$

Comparing the two solutions (9) and (10), we get

$$\begin{aligned}
& \mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}} \\
= \;& \mathbf{x}_{\mathrm{STLS}} - E_{\widehat{A}}^{+}\mathbf{e}_{\hat{b}} \\
= \;& \mathbf{x}_{\mathrm{STLS}} - E_{\widehat{A}}^{+}E_{\widehat{A}}\mathbf{x}_{\mathrm{STLS}} + E_{\widehat{A}}^{+}E_{\widehat{A}}\mathbf{x}_{\mathrm{STLS}} \\
& - E_{\widehat{A}}^{+}\mathbf{e}_b - E_{\widehat{A}}^{+}(\mathbf{e}_{\hat{b}} - \mathbf{e}_b) \\
= \;& \mathbf{x}_{\mathrm{STLS}} - E_{\widehat{A}}^{+}E_{\widehat{A}}\mathbf{x}_{\mathrm{STLS}} + E_{\widehat{A}}^{+}E_{\widehat{A}}\mathbf{x}_{\mathrm{STLS}} \\
& - E_{\widehat{A}}^{+}E_A\mathbf{x}_{\mathrm{STLS}} - E_{\widehat{A}}^{+}(\mathbf{e}_{\hat{\mathbf{b}}} - \mathbf{e}_b) \\
= \;& (I - E_{\widehat{A}}^{+}E_{\widehat{A}})\mathbf{x}_{\mathrm{STLS}} + E_{\widehat{A}}^{+}(E_{\widehat{A}} - E_A)\mathbf{x}_{\mathrm{STLS}} \\
& - E_{\widehat{A}}^{+}(\mathbf{e}_{\hat{b}} - \mathbf{e}_b).
\end{aligned}$$

Obviously, $\|(I - E_{\widehat{A}}^{+}E_{\widehat{A}})\mathbf{x}_{\mathrm{STLS}}\|_2 \le \|\mathbf{x}_{\mathrm{STLS}}\|_2$. From $E_{\widehat{A}} = \widehat{A} - \widehat{U}_2\widehat{\Sigma}_2\widehat{V}_{12}^{\mathrm{T}}$, we have

$$\begin{aligned}
\sigma_k(E_{\widehat{A}}) &\ge \sigma_k(\widehat{A}) - \|\widehat{U}_2\widehat{\Sigma}_2\widehat{V}_{12}^{\mathrm{T}}\|_2 \\
&\ge \sigma_k(\widehat{A}) - \sigma_{k+1}(\widehat{C}) \\
&\ge \sigma_k(A) - \sigma_{k+1}(C) - 2\eta,
\end{aligned}$$

which implies that

$$\begin{aligned}
\|E_{\widehat{A}}^{+}\|_2 &= (\sigma_k(E_{\widehat{A}}))^{-1} \\
&\le \frac{1}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta}, \qquad (11)
\end{aligned}$$

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

$$\begin{aligned}
& \|E_{\widehat{A}} - E_A\|_2 \\
= \;& \|\widehat{A} - A - \widehat{U}_2\widehat{\Sigma}_2\widehat{V}_{12}^{\mathrm{T}} + U_2\Sigma_2V_{12}^{\mathrm{T}}\|_2 \\
\le \;& \|\widehat{A} - A\|_2 + \|\widehat{U}_2\widehat{\Sigma}_2\widehat{V}_{12}^{\mathrm{T}}\|_2 + \|U_2\Sigma_2V_{12}^{\mathrm{T}}\|_2 \\
\le \;& \|\widehat{C} - C\|_2 + \sigma_{k+1}(\widehat{C}) + \sigma_{k+1}(C) \\
\le \;& \eta + \sigma_{k+1}(C) + \sigma_{k+1}(\widehat{C}) \\
\le \;& 2\eta + 2\sigma_{k+1}(C) \qquad (12)
\end{aligned}$$

and

$$\begin{aligned}
& \|\mathbf{e}_{\hat{b}} - \mathbf{e}_b\|_2 \\
= \;& \|\widehat{\mathbf{b}} - \mathbf{b} - \lambda^{-1}\widehat{U}_2\widehat{\Sigma}_2\widehat{\mathbf{v}}_{22} + \lambda^{-1}U_2\Sigma_2\mathbf{v}_{22}\|_2 \\
\le \;& \|\widehat{\mathbf{b}} - \mathbf{b}\|_2 + \lambda^{-1}(\sigma_{k+1}(\widehat{C}) + \sigma_{k+1}(C)) \\
= \;& \eta + \lambda^{-1}(2\sigma_{k+1}(C) + \eta). \qquad (13)
\end{aligned}$$

Putting the above three inequalities (11), (12), and (13) together, we get

$$\begin{aligned}
& \|\mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}}\|_2 \\
\le \;& \|\mathbf{x}_{\mathrm{STLS}}\|_2 + \|E_{\widehat{A}}^{+}\|_2\,\|E_{\widehat{A}} - E_A\|_2\,\|\mathbf{x}_{\mathrm{STLS}}\|_2 \\
& + \|E_{\widehat{A}}^{+}\|_2\,\|\mathbf{e}_{\hat{b}} - \mathbf{e}_b\|_2 \\
\le \;& \|\mathbf{x}_{\mathrm{STLS}}\|_2 + \frac{2\sigma_{k+1}(C) + 2\eta}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta}\,\|\mathbf{x}_{\mathrm{STLS}}\|_2 \\
& + \frac{\lambda^{-1}(2\sigma_{k+1}(C) + \eta) + \eta}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta} \\
= \;& \frac{\sigma_k(A) + \sigma_{k+1}(C)}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta}\|\mathbf{x}_{\mathrm{STLS}}\|_2 \\
& + \frac{\lambda^{-1}(2\sigma_{k+1}(C) + \eta) + \eta}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta} \\
< \;& \frac{(\sigma_k(A) + \sigma_{k+1}(C))(\|\mathbf{x}_{\mathrm{STLS}}\|_2 + \lambda^{-1}) + \eta}{\sigma_k(A) - \sigma_{k+1}(C) - 2\eta},
\end{aligned}$$

since $\eta < \sigma_k(A) - \sigma_{k+1}(C)$. □

This theorem says that if the perturbation $\eta = \|E\|_2$ is small, we can expect a small error $\|\mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}}\|_2$ as long as $\sigma_k(A)$ and $\sigma_{k+1}(C)$ are not closely clustered. If $\sigma_k(A)$ is very close to $\sigma_{k+1}(C)$, the computed solution $\widehat{\mathbf{x}}_{\mathrm{STLS}}$ may be very different from the exact solution $\mathbf{x}_{\mathrm{STLS}}$. Moreover, as λ approaches to zero, both $\sigma_{k+1}(C)$ and $\sigma_{k+1}(\widehat{C})$ approach to zero as fast as λ does. Specifically, $\lim_{\lambda\to 0}\sigma_{k+1}(C)/\lambda = \|\mathbf{r}\|_2$, where $\mathbf{r}$ is the residual of the least square problem $\min_{\mathbf{x}}\|A\mathbf{x} - \mathbf{b}\|_2$ [10]. Then the inequality in the theorem reduces to

$$\begin{aligned}
\|\mathbf{x}_{\mathrm{STLS}} - \widehat{\mathbf{x}}_{\mathrm{STLS}}\|_2 &\le (1 + \frac{\eta}{\sigma_k(A)})\|\mathbf{x}_{\mathrm{STLS}}\|_2 \\
&\quad + \frac{\eta}{\sigma_k(A)}(1 + \|\hat{\mathbf{r}}\|_2 + \|\mathbf{r}\|_2).
\end{aligned}$$

It shows that the difference between $\mathbf{x}_{\mathrm{STLS}}$ and $\widehat{\mathbf{x}}_{\mathrm{STLS}}$ is independent of the scalar λ, when λ approaches to zero.

5 Numerical Experiments

In the STLS formulation (1), a scalar λ is introduced to the right side vector $\mathbf{b}$. The residual to be minimized is $[E \;\; \mathbf{r}]$, same as the TLS problem. In this section, we compare STLS with TLS. The STLS problem is solved by the RRULVD method presented in the previous sections, whereas the TLS problem is solved by the SVD method.

All of our numerical experiments were performed in MATLAB on a Sun SPARC workstation Ultra 10 using double precision. The rank deficient matrices were generated as the product

$$A = U\begin{bmatrix} \Sigma & 0 \\ 0 & Z \end{bmatrix}V^{\mathrm{T}},$$

where $U \in \mathrm{R}^{m\times n}$ and $V \in \mathrm{R}^{n\times n}$, $m > n$, are random matrices with orthonormal columns, Σ diagonal of order k, whose diagonal elements are random variables uniformly distributed over $[0, 1]$, and Z a zero matrix of order $n - k$. The right-hand side vectors $\mathbf{b}$ were generated as vectors with entries uniformly

λ	$\cos\theta_S$	$\cos\theta_T$	res_S	res_T
0.01	0.9428	0.2610	0.0164	0.8050
0.1	0.9428	0.2610	0.1632	0.8050
1	0.6073	0.2610	0.9489	0.8050
5	0.8916	0.2610	1.0450	0.8050

Table 1: Comparison of the STLS solution with the TLS solution for a 64-by-48 matrix A of rank 43.

distributed over $[0, 1]$. The random perturbations E and $\mathbf{r}$ on A and $\mathbf{b}$ respectively were constructed by

$$E = \xi\,\mathrm{randn}(m, n), \qquad \mathbf{r} = \xi\,\mathrm{randn}(m, 1),$$

where ξ is a parameter controlling the magnitude of the perturbations, and the function randn generates random numbers normally distributed with zero mean and unit variance. In all examples, we set $\xi = 3 \times 10^{-8}$ and the numerical rank tolerance to 2×10^{-5}. Since the perturbations are smaller than the numerical rank tolerance, all matrices are numerically rank deficient.

To compare STLS and TLS, we denote θ_S and θ_T as the angles between $\mathbf{b}$ and $A\mathbf{x}_{\text{STLS}}$ and between $\mathbf{b}$ and $A\mathbf{x}_{\text{TLS}}$, respectively, that is,

$$\begin{aligned} \cos\theta_S &:= \|\mathbf{b}^{\mathrm{T}} A\mathbf{x}_{\text{STLS}}\|_2/(\|A\mathbf{x}_{\text{STLS}}\|_2\|\mathbf{b}\|_2) \quad \text{and} \\ \cos\theta_T &:= \|\mathbf{b}^{\mathrm{T}} A\mathbf{x}_{\text{TLS}}\|_2/(\|A\mathbf{x}_{\text{TLS}}\|_2\|\mathbf{b}\|_2). \end{aligned}$$

Also, we denote the residual

$$r_S := \|[E_{\text{STLS}} \quad \mathbf{r}_{\text{STLS}}]\|_{\mathrm{F}},$$

which is equal to $\sigma_{k+1}(C)$ [10], and $r_T := \|[E_{\text{TLS}} \quad \mathbf{r}_{\text{TLS}}]\|_{\mathrm{F}} = \sigma_{k+1}(C)$ [8]. Note that θ_T and r_T are independent of λ.

Table 1 shows:

- For small values of λ, $A\mathbf{x}_{\text{STLS}}$ is closer to $\mathbf{b}$ than $A\mathbf{x}_{\text{TLS}}$ is, and the STLS residual is much smaller than the TLS residual;
- When λ is small, θ_S is insensitive to the change of λ.

We note that

- In theory, when $\lambda = 1$, $\mathbf{x}_{\text{STLS}} = \mathbf{x}_{\text{TLS}}$. The differences in the table when $\lambda = 1$ are due to the different algorithms used to compute the STLS solution and the TLS solution. In the STLS algorithm, the RRULVD, which is an approximation of the COD, is computed, whereas in the TLS algorithm, the SVD is computed. However, we can see that the corresponding values are in the same magnitude order.
- For large values of λ, large vectors $\lambda\mathbf{b}$ are appended to A to form C. Consequently, the right singular vectors corresponding to $\sigma_{k+1}(C)$ of the resulting matrices C vary little with different $\mathbf{b}$. Recall that the STLS solution depends on the right singular vector and the null space. Thus, the STLS solutions vary little for large values of λ with different $\mathbf{b}$.

Conclusion: Choose $\lambda < 1$.

6 Conclusion

We presented an algorithm for solving the scaled total least squares problems using the rank revealing ULV decomposition, which is an approximation of the complete orthogonal decomposition. To improve accuracy, in addition to the refinement, we proposed a technique for improving the accuracy of the estimates for the smallest nonzero singular value and its corresponding right singular vector. Our perturbation analysis showed that if the smallest nonzero singular values $\sigma_k(A)$ and $\sigma_{k+1}(C)$ of the coefficient matrix A and the augmented matrix C respectively are not closely clustered, accurate solutions are expected from our method. Experiments demonstrated that our method produces solutions as accurate as the SVD method with less computation.

References

[1] Ricardo. D. Fierro and James 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 and C. F. Van Loan, *Matrix Computations*, 3rd Ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[3] Charles Van Loan. On estimating the condition of eigenvalues and eigenvectors. *Linear Algebra and its Applications*, 88/89(1987), 715–732.

[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 fundermentals. *Numer. Math.* 91(2002), 117–146.

[6] B.D. Rao. Unified treatment of LS, TLS and Truncated SVD methods using a weighted TLS framework. *Recent Advanced in Total Least Squares Techniques and Errors-in-Variables Modeling*, edited by S. Van Huffel. SIAM, Philadelphia PA, 1997, 11–20.

[7] G.W. Stewart. Matrix Algorithms, volume I: Basic Decompositions. SIAM, Philadelphia, 1998.

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

[9] Wei Xu, Yimin Wei and Sanzheng Qiao. An algorithm for solving rank-deficient scaled total least squares problems. *Technical Report No. CAS 04-04-SQ*, McMaster University, Hamilton, Ont. Canada.

[10] Wei Xu, Sanzheng Qiao and Yimin Wei. A note on the scaled total least squares problem. *Linear Algebra and Its Applications*, 428/2+3(2008), 469-478.