Numer Algor (2009) 51:381–399
DOI 10.1007/s11075-009-9269-0

ORIGINAL PAPER

Perturbation analysis and condition numbers of scaled total least squares problems

Liangmin Zhou · Lijing Lin · Yimin Wei · Sanzheng Qiao

Received: 23 April 2008 / Accepted: 1 February 2009 /
Published online: 19 February 2009

Abstract The standard approaches to solving an overdetermined linear system $Ax \approx b$ find minimal corrections to the vector b and/or the matrix A such that the corrected system is consistent, such as the least squares (LS), the data least squares (DLS) and the total least squares (TLS). The scaled total least squares (STLS) method unifies the LS, DLS and TLS methods. The classical normwise condition numbers for the LS problem have been widely studied.

In memory of Prof. Gene H. Golub.

Yimin Wei is supported by the National Natural Science Foundation of China under grant 10871051, Shanghai Science & Technology Committee under grant 08DZ2271900 and Shanghai Education Committee under grant 08SG01. Sanzheng Qiao is partially supported by Shanghai Key Laboratory of Contemporary Applied Mathematics of Fudan University during his visiting.

L. Zhou · L. Lin · Y. Wei (✉)
Institute of Mathematics, School of Mathematical Science,
Fudan University, Shanghai, 200433, People's Republic of China
e-mail: ymwei@fudan.edu.cn

L. Zhou
e-mail: 062018027@fudan.edu.cn

L. Lin
e-mail: 042018028@fudan.edu.cn

L. Zhou · L. Lin · Y. Wei
Key Laboratory of Nonlinear Science (Fudan University),
Ministry of Education, Shanghai, 200433, People's Republic of China

S. Qiao
Department of Computing and Software, McMaster University,
Hamilton, Ontario, Canada, L8S 4K1
e-mail: qiao@mcmaster.ca

However, there are no such similar results for the TLS and the STLS problems. In this paper, we first present a perturbation analysis of the STLS problem, which is a generalization of the TLS problem, and give a normwise condition number for the STLS problem. Different from normwise condition numbers, which measure the sizes of both input perturbations and output errors using some norms, componentwise condition numbers take into account the relation of each data component, and possible data sparsity. Then in this paper we give explicit expressions for the estimates of the mixed and componentwise condition numbers for the STLS problem. Since the TLS problem is a special case of the STLS problem, the condition numbers for the TLS problem follow immediately from our STLS results. All the discussions in this paper are under the Golub-Van Loan condition for the existence and uniqueness of the STLS solution.

Keywords Scaled total least squares · Least squares · Condition number · Mixed and componentwise condition number · Perturbation · Error bounds

Mathematics Subject Classifications (2000) 65F20 · 65F35

1 Introduction

The standard approaches to solving an overdetermined linear system $Ax \approx b$ find minimal corrections to the vector b and/or the matrix A such that the corrected system is consistent, such as the least squares (LS) method, the data least squares (DLS) method and the total least squares (TLS) method.

The scaled total least squares (STLS) method unifies the LS, DLS, and TLS methods. For given $A \in \mathbb{R}^{m\times n}$ $(m > n)$ and $b \in \mathbb{R}^m$, based on the work of Rao [18], Paige and Strakoš [16] formulated the STLS problem:

$$\min \|[r \quad E]\|_F \quad \text{subject to} \quad \lambda b - r \in \mathcal{R}(A + E), \tag{1.1}$$

where λ is a real positive parameter, $\|\cdot\|_F$ denotes the Frobenius norm and $\mathcal{R}(\cdot)$ represents the range space. Suppose that $[r_{_{STLS}} \quad E_{_{STLS}}]$ solves the above problem, then the solution $x_{_{STLS}}$ for x in the equation $(A + E_{_{STLS}})\lambda x = \lambda b - r_{_{STLS}}$ is called the STLS solution. Obviously, the STLS problem becomes the TLS problem when $\lambda = 1$. Also, Paige and Strakoš [15] showed that the STLS solution approaches the LS solution as $\lambda \to 0$.

Condition numbers play an important role in sensitivity analysis, a study of the worst case sensitivity of the solution of a problem to small perturbations in the input data [7]. The product of a condition number and backward error provides an approximate local linear upper bound on the forward error in a computed solution.

There have been many results about the classical normwise condition numbers for the LS problem. However, there are no such results for the TLS

and STLS problems. In this paper, using the terminologies introduced by Gohberg and Koltracht [4], we derive the relative normwise, mixed and componentwise condition numbers for the STLS problem. To ensure the existence and uniqueness of the STLS solution, we assume that the Golub-Van Loan condition is satisfied. That is, the smallest singular value of A is strictly greater than the smallest singular value of $[A \quad \lambda b]$. Since the TLS problem is a special case of the STLS problem where $\lambda = 1$, our results are readily applied to the TLS problem.

Throughout this paper, the following notations are used: $\|X\|_2$ denotes the spectral norm of a matrix, given by the square root of the largest eigenvalue of $X^T X$; $\|X\|_\infty$ the infinity norm of a matrix, given by $\|X\|_\infty = \max_i \sum_j |X_{ij}|$; X^T the transpose of X; $|X|$ is the absolute value of the matrix X, whose entries are $|X_{ij}|$; the matrix inequality $|X| \le |Y|$ means the componentwise $|X_{ij}| \le |Y_{ij}|$, for all i and j; $X^\dagger$ the Moore-Penrose inverse [1, 30] of X; diag(a) the diagonal matrix whose diagonal is given by a vector a; $\|a\|_2$ the Euclidean norm of a vector, given by $\|a\|_2 = \sqrt{\sum_i |a_i|^2}$; $\|a\|_\infty$ the infinity norm of a vector, given by $\|a\|_\infty = \max_i |a_i|$; $|a|$ the vector whose elements are $|a_i|$; I the $n \times n$ identity matrix. For any $a, b \in \mathbb{R}^n$, we define $\frac{a}{b} = [c_1, c_2, \cdots, c_n]^T$ by

$$c_i = \begin{cases} a_i/b_i, & \text{if } b_i \neq 0, \\ 0, & \text{if } a_i = b_i = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

For matrices $X = [x_1, x_2, \cdots, x_n] = [X_{ij}]$ and Y, $X \otimes Y = [X_{ij} Y]$ is the Kronecker product of X and Y and for X the linear operator vec : $\mathbb{R}^{m \times n} \to \mathbb{R}^{mn}$ is defined by $\text{vec}(X) = \left[x_1^T, x_2^T, \cdots, x_n^T\right]^T$.

This paper is organized as follows. In Section 2, we give explicit expressions for the STLS solution and some preliminaries necessary for this paper. In Section 3, we derive a perturbation estimate and a relative normwise condition number for the STLS problem. The mixed and componentwise condition numbers for the STLS problem are given in Section 4. In Section 5, we present our numerical experiments to confirm our results. Finally, in Section 6, we make some conclusions.

2 Preliminaries

In this section, we begin with a brief review of some properties of the TLS problem, which play an essential role throughout our study on the condition numbers for the STLS problem. Then using the solution for the TLS problem, we give a sufficient condition for the existence of the STLS solution and an explicit expression for the solution. Finally, we give a useful identity necessary for our study of the condition numbers for the STLS problem.

As we know, the TLS problem can be formulated as

$$\min \|[r \quad E]\|_F \quad \text{subject to} \quad b - r \in \mathcal{R}(A + E). \tag{2.1}$$

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

The existence of the TLS solution has been studied by a number of authors [5, 9–11, 13, 14, 17, 23–26, 29, 31]. Golub and Van Loan [5] gave the following condition for the existence and uniqueness of the TLS solution. Let the singular value decompositions of A and $[A \quad b]$ be

$$\hat{U}^T A \hat{V} = \hat{\Sigma} \quad \text{and} \quad \check{U}^T [A \quad b] \check{V} = \check{\Sigma}$$

respectively, where

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1 & & \\ & \ddots & \\ & & \hat{\sigma}_n \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad \hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_n \geq 0,$$

and

$$\check{\Sigma} = \begin{bmatrix} \check{\sigma}_1 & & \\ & \ddots & \\ & & \check{\sigma}_{n+1} \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad \check{\sigma}_1 \geq \check{\sigma}_2 \geq \cdots \geq \check{\sigma}_{n+1} \geq 0,$$

with $\hat{U}, \check{U} \in \mathbb{R}^{m \times m}$, $\hat{V} \in \mathbb{R}^{n \times n}$, $\check{V} \in \mathbb{R}^{(n+1) \times (n+1)}$ being real orthogonal matrices. If $\hat{\sigma}_n > \check{\sigma}_{n+1}$, then x_{TLS} exists uniquely and

$$x_{TLS} = \left(A^T A - \check{\sigma}_{n+1}^2 I\right)^{-1} A^T b. \tag{2.2}$$

In the following part, we make use of the TLS solution to give a sufficient condition for the existence of the STLS solution and an explicit expression of the solution under the condition. Comparing (1.1) and (2.1), we can see that if x_{STLS} is the solution of (1.1) then λx_{STLS} is the solution of the TLS problem with A and λb. For a sufficient condition for the existence of the STLS solution, let the singular value decomposition of $[A \quad \lambda b]$ be

$$U^T [A \quad \lambda b] V = \Sigma,$$

where

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{n+1} \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n+1} \geq 0,$$

with $U \in \mathbb{R}^{m\times(n+1)}$, $V \in \mathbb{R}^{(n+1)\times(n+1)}$ being real orthogonal matrices. If $\hat{\sigma}_n > \sigma_{n+1}$, by (2.2), there exists a unique solution to the STLS problem, which can be expressed as

$$\lambda x_{STLS} = \lambda \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b,$$

that is,

$$x_{STLS} = \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b. \tag{2.3}$$

Throughout this paper we assume $\hat{\sigma}_n > \sigma_{n+1}$ to ensure the existence and uniqueness of the STLS solution. Note that this condition implies that $\hat{\sigma}_n > 0$, then it follows that the LS problem

$$\min \|s\|_2 \quad \text{subject to} \quad b + s \in \mathcal{R}(A), \tag{2.4}$$

has a unique solution given by

$$x_{LS} = \left(A^T A\right)^{-1} A^T b.$$

From the expression (2.3), we can see that if $\sigma_{n+1} = 0$, the STLS solution becomes the LS solution. Thus in the following discussion, we further assume that $\sigma_{n+1} > 0$.

Finally, we give an identity about the Kronecker product in the following lemma. See [6] for a proof.

Lemma 2.1 *Let $B \in \mathbb{R}^{m\times p}$, $C \in \mathbb{R}^{q\times n}$, $X \in \mathbb{R}^{p\times q}$, then*

$$\operatorname{vec}(BXC) = \left(C^T \otimes B\right) \operatorname{vec}(X). \tag{2.5}$$

The above identity plays an important role in this paper.

3 Normwise condition number

In this section we first derive a perturbation estimate and then present a relative normwise condition number for the STLS problem. Let $\tilde{A} = A + \Delta_A$ and $\tilde{b} = b + \Delta_b$, where Δ_A and Δ_b are the perturbations of the input data A and b respectively. Consider the perturbed STLS problem

$$\min \|[r \;\; E]\|_F \quad \text{subject to} \quad \lambda\tilde{b} - r \in \mathcal{R}\left(\tilde{A} + E\right).$$

Denote the singular values of the matrix $[\tilde{A} \;\; \lambda\tilde{b}]$ by $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_{n+1} \geq 0$. If the norm $\|[\Delta_A \;\; \lambda\Delta_b]\|_F$ of the perturbations is sufficiently small, then the well-known perturbation analysis of singular values ensures that the perturbed STLS problem above has a unique solution $\tilde{x}_{STLS}$ and it can be expressed as

$$\tilde{x}_{STLS} = \left(\tilde{A}^T \tilde{A} - \tilde{\sigma}_{n+1}^2 I\right)^{-1} \tilde{A}^T \tilde{b}. \tag{3.1}$$

Let the change in the solution be

$$\Delta_x := \tilde{x}_{STLS} - x_{STLS}. \tag{3.2}$$

Now we introduce the following definition of the relative normwise condition number for the STLS problem.

Definition 3.1 Using the above notations, the relative normwise condition number for the STLS problem is defined by

$$\kappa_{STLS} := \lim_{\epsilon \to 0} \sup_{\|[\Delta_A \quad \lambda\Delta_b]\|_F \le \epsilon \|[A \quad \lambda b]\|_F} \frac{\|\Delta_x\|_2}{\epsilon \|x_{STLS}\|_2}. \tag{3.3}$$

Before deriving the explicit expression for this normwise condition number, we give a useful expansion of the smallest singular value of the perturbed matrix in terms of the smallest singular value of the original matrix. See [21] or [22] for a proof.

Lemma 3.1 *Let $\sigma_{\min}$ be the smallest nonzero and simple singular value of a matrix X with $u_{\min}$ and $v_{\min}$ being its corresponding left and right singular vectors respectively. If $\|\Delta_X\|_F$ is sufficiently small, then $\tilde{\sigma}_{\min}$, the smallest nonzero singular value of the perturbed matrix $\tilde{X} = X + \Delta_X$, is simple and*

$$\tilde{\sigma}_{\min} = \sigma_{\min} + u_{\min}^T \Delta_X v_{\min} + \mathcal{O}\left(\|\Delta_X\|_F^2\right).$$

In our case, from the interlacing property [33, p.103] of the eigenvalues of a symmetric matrix we obtain that

$$\sigma_1 \ge \hat{\sigma}_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge \hat{\sigma}_n \ge \sigma_{n+1}.$$

By the above inequality and our assumption $\hat{\sigma}_n > \sigma_{n+1}$, we can further conclude that σ_{n+1} is a simple singular value of $[A \quad \lambda b]$. Then by Lemma 3.1, we have

$$\Delta_{\sigma_{n+1}} := \tilde{\sigma}_{n+1} - \sigma_{n+1} = u_{n+1}^T [\Delta_A \quad \lambda\Delta_b] v_{n+1} + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right), \tag{3.4}$$

where u_{n+1} and v_{n+1} are respectively the left and right singular vectors corresponding to the smallest singular value σ_{n+1} of $[A \quad \lambda b]$.

Now we derive an explicit expression of an estimate for the perturbation Δ_x defined in (3.2) in the following lemma.

Lemma 3.2 *For Δ_x defined in (3.2), we have an estimate*

$$\Delta_x = (M + N)\mathrm{vec}\left([\Delta_A \quad \lambda\Delta_b]\right) + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right), \tag{3.5}$$

where

$$M := \left[K \otimes b^T - x_{STLS}^T \otimes \left(KA^T\right) - K \otimes \left(Ax_{STLS}\right)^T \qquad \lambda^{-1} KA^T\right], \tag{3.6}$$

$$N := 2\sigma_{n+1} y \left(v_{n+1}^T \otimes u_{n+1}^T\right), \tag{3.7}$$

and

$$\begin{aligned} R(\Delta_A, \Delta_b) := {} & K\Delta_A^T\Delta_b - K\left(\Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I\right) K\left(A^Tb + A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right) \\ & - K\left(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I\right) K\left(A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right) \\ & - K\sum_{i=2}^{\infty}\left(\left(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I + \Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I\right) K\right)^i \\ & \cdot \left(A^Tb + A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right). \end{aligned}$$

with $K := \left(A^TA - \sigma_{n+1}^2 I\right)^{-1}$ *and* $y = Kx_{STLS}$.

Proof From (3.1), (3.2) and (3.4), we have

$$\begin{aligned} \Delta_x &= \tilde{x}_{STLS} - x_{STLS} \\ &= \left(\tilde{A}^T\tilde{A} - \tilde{\sigma}_{n+1}^2 I\right)^{-1}\tilde{A}^T\tilde{b} - x_{STLS} \\ &= \left((A+\Delta_A)^T(A+\Delta_A) - \left(\sigma_{n+1} + \Delta_{\sigma_{n+1}}\right)^2 I\right)^{-1}(A+\Delta_A)^T(b+\Delta_b) - x_{STLS}. \end{aligned}$$

Expanding the last equality of the above formula, we have

$$\begin{aligned} \Delta_x = {} & \left(\left(A^TA - \sigma_{n+1}^2 I\right) + \left(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I + \Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I\right)\right)^{-1} \\ & \cdot \left(A^Tb + A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right) - x_{STLS} \\ = {} & \left(A^TA - \sigma_{n+1}^2 I\right)^{-1} \\ & \cdot \left(I + \left(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I + \Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I\right)\left(A^TA - \sigma_{n+1}^2 I\right)^{-1}\right)^{-1} \\ & \cdot \left(A^Tb + A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right) - x_{STLS}. \end{aligned}$$

When $\|[\Delta_A \lambda \Delta_b]\|_F$ is sufficiently small, we may assume that the norm of $(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I + \Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I)(A^TA - \sigma_{n+1}^2 I)^{-1}$ is less than 1. Then we have

$$\begin{aligned} \Delta_x = {} & \left(A^TA - \sigma_{n+1}^2 I\right)^{-1} \\ & \cdot \left(I - \sum_{i=1}^{\infty}\left(\left(A^T\Delta_A + \Delta_A^TA - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I + \Delta_A^T\Delta_A - \Delta_{\sigma_{n+1}}^2 I\right)\right.\right. \\ & \qquad \left.\left. \times \left(A^TA - \sigma_{n+1}^2 I\right)^{-1}\right)^i\right) \\ & \cdot \left(A^Tb + A^T\Delta_b + \Delta_A^Tb + \Delta_A^T\Delta_b\right) - x_{STLS}. \end{aligned}$$

Further expanding the above formula, replacing $(A^TA - \sigma_{n+1}^2 I)^{-1}$ with K, and using the definition of $R(\Delta_A, \Delta_b)$, we obtain

$$\begin{aligned}\Delta_x = &\left(A^TA - \sigma_{n+1}^2 I\right)^{-1} A^T b + K\left(\Delta_A^T b + A^T\Delta_b\right)\\ &- K\left(A^T\Delta_A + \Delta_A^T A - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I\right) KA^T b + R\left(\Delta_A, \Delta_b\right) - x_{STLS}.\end{aligned}$$

Then by (2.3), we get

$$\Delta_x = K\left(\Delta_A^T b + A^T\Delta_b\right) - K\left(A^T\Delta_A + \Delta_A^T A - 2\sigma_{n+1}\Delta_{\sigma_{n+1}} I\right) x_{STLS} + R\left(\Delta_A, \Delta_b\right).$$

Substituting $\Delta_{\sigma_{n+1}}$ in the above formula with (3.4), we have

$$\begin{aligned}\Delta_x = &K\left(\Delta_A^T b + A^T\Delta_b\right) - K\left(A^T\Delta_A + \Delta_A^T A - 2\sigma_{n+1}u_{n+1}^T[\Delta_A \quad \lambda\Delta_b]v_{n+1} I\right) x_{STLS}\\ &+ R\left(\Delta_A, \Delta_b\right) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

Further expanding the above formula and replacing Kx_{STLS} with y, we get

$$\begin{aligned}\Delta_x &= K\Delta_A^T b + KA^T\Delta_b - KA^T\Delta_A x_{STLS} - K\Delta_A^T A x_{STLS}\\ &\quad + 2\sigma_{n+1}u_{n+1}^T[\Delta_A \quad \lambda\Delta_b]v_{n+1} Kx_{STLS} + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right)\\ &= K\Delta_A^T b + KA^T\Delta_b - KA^T\Delta_A x_{STLS} - K\Delta_A^T A x_{STLS}\\ &\quad + 2\sigma_{n+1}u_{n+1}^T[\Delta_A \quad \lambda\Delta_b]v_{n+1} y + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

Applying the operator vec to both sides of the last equation above, we have

$$\begin{aligned}\Delta_x &= KA^T\Delta_b + \operatorname{vec}(K\Delta_A^T b) - \operatorname{vec}\left(KA^T\Delta_A x_{STLS}\right) - \operatorname{vec}\left(K\Delta_A^T A x_{STLS}\right)\\ &\quad + 2\sigma_{n+1} y \operatorname{vec}\left(u_{n+1}^T[\Delta_A \quad \lambda\Delta_b]v_{n+1}\right) + R\left(\Delta_A, \Delta_b\right) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right)\\ &= KA^T\Delta_b + \operatorname{vec}\left(b^T\Delta_A K^T\right) - \operatorname{vec}\left(KA^T\Delta_A x_{STLS}\right) - \operatorname{vec}\left((Ax_{STLS})^T\Delta_A K^T\right)\\ &\quad + 2\sigma_{n+1} y \operatorname{vec}\left(u_{n+1}^T[\Delta_A \quad \lambda\Delta_b]v_{n+1}\right) + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

Applying (2.5) to the above formula we obtain

$$\begin{aligned}\Delta_x &= KA^T\Delta_b + \left(K\otimes b^T\right)\operatorname{vec}(\Delta_A) - \left(x_{STLS}^T\otimes(KA^T)\right)\operatorname{vec}(\Delta_A)\\ &\quad - \left(K\otimes(Ax_{STLS})^T\right)\operatorname{vec}(\Delta_A) + 2\sigma_{n+1} y\left(v_{n+1}^T\otimes u_{n+1}^T\right)\operatorname{vec}([\Delta_A \quad \lambda\Delta_b])\\ &\quad + R\left(\Delta_A, \Delta_b\right) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right)\\ &= \left[K\otimes b^T - x_{STLS}^T\otimes\left(KA^T\right) - K\otimes(Ax_{STLS})^T \quad \lambda^{-1}KA^T\right]\begin{bmatrix}\operatorname{vec}(\Delta_A)\\ \lambda\Delta_b\end{bmatrix}\\ &\quad + 2\sigma_{n+1} y(v_{n+1}^T\otimes u_{n+1}^T)\operatorname{vec}([\Delta_A \ \ \lambda\Delta_b]) + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \ \ \lambda\Delta_b]\|_F^2\right)\\ &= (M+N)\operatorname{vec}([\Delta_A \ \ \lambda\Delta_b]) + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

This completes the proof. □

Now that we have obtained an expression of Δ_x, a key result in this paper, we give an explicit expression for the estimate of the relative normwise condition number defined in (3.3) in the following theorem.

Theorem 3.1 *Using the notations above, we have the condition number*

$$\kappa_{_{STLS}} \leq \frac{\|M+N\|_2\|\begin{bmatrix} A & \lambda b\end{bmatrix}\|_F}{\|x_{_{STLS}}\|_2} := \kappa^M_{_{STLS}} \tag{3.8}$$

defined in Definition 3.1.

Proof From (3.5), we have

$$\begin{aligned}\|\Delta_x\|_2 &\leq \|M+N\|_2\|\mathrm{vec}([\Delta_A \quad \lambda\Delta_b])\|_2 + \|R(\Delta_A,\Delta_b)\|_2 + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right)\\ &= \|M+N\|_2\|[\Delta_A \quad \lambda\Delta_b]\|_F + \|R(\Delta_A,\Delta_b)\|_2 + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

For the term $\|R(\Delta_A,\Delta_b)\|_2$, the definition of $R(\Delta_A,\Delta_b)$ shows that every term in the expansion of $\|R(\Delta_A,\Delta_b)\|_2$ contains a factor either $\|\Delta_A\|_F^2$ or $\|\Delta_b\|_2^2$ or $\|\Delta_A\|_F\|\Delta_b\|_2$. The condition $\|[\Delta_A \quad \lambda\Delta_b]\|_F \leq \epsilon\|\begin{bmatrix} A & \lambda b\end{bmatrix}\|_F$ in (3.3) implies that $\|\Delta_A\|_F \leq \epsilon\|[A\lambda b]\|_F$ and $\|\Delta_b\|_2 \leq \epsilon\|[A\lambda b]\|_F$. Therefore, $\|R(\Delta_A,\Delta_b)\|_2 \leq \mathcal{O}(\epsilon^2)$. Consequently, a perturbation bound for the STLS solution x is obtained:

$$\frac{\|\Delta_x\|_2}{\|x_{_{STLS}}\|_2} \leq \epsilon\frac{\|M+N\|_2}{\|x_{_{STLS}}\|_2}\|[A \quad \lambda b]\|_F + \mathcal{O}\left(\epsilon^2\right),$$

which, from (3.3), leads to (3.8). □

We will demonstrate in Section 5 that the estimate (3.5) is accurate and the condition number $\kappa_{_{STLS}}$ (3.8) is very sharp.

Note that $\kappa^M_{_{STLS}}$ involves the matrices M and N, which may be impractical to compute. However, (3.6) and (3.7) indicate that when $\hat{\sigma}_n$, the smallest singular value of A, is close to σ_{n+1}, the smallest singular value of $[A \quad \lambda b]$, then K is ill-conditioned, implying that the norm of $y = Kx_{_{STLS}}$ thus the norm of N and the norm of M can be large. In other words, when the gap $\hat{\sigma}_n - \sigma_{n+1}$ is small, then the STLS problem is ill-conditioned. Our experiments have shown that when the gap is small, $\kappa^M_{_{STLS}}$ is about the reciprocal of the gap.

4 Componentwise condition numbers

A drawback of the relative normwise condition numbers as defined in (3.3) is that they ignore the structure of both input and output data with respect to scaling and/or sparsity. When the data are badly scaled or contain many zeros, measuring the size of a perturbation in terms of its norm leaves us in the dark concerning the relative size of the perturbation on its small (or zero) entries.

To tackle this drawback, another approach in the perturbation theory, known as componentwise analysis, has been increasingly considered [8, 19, 20].

Specifically, two kinds of condition numbers are studied. The first kind, called mixed condition numbers by Gohberg and Koltracht [4], measures the output errors in norms and the input perturbations componentwise. The second kind, called componentwise condition numbers by Gohberg and Koltracht [4], measures both the output errors and the input perturbations componentwise. We adopt their terminology and define the mixed and componentwise condition numbers for the STLS problem as follows.

Definition 4.1 The mixed condition number for the STLS problem is defined by

$$\mu_{STLS} := \lim_{\epsilon \to 0} \sup_{\substack{|\Delta_A| \le \epsilon |A| \\ |\Delta_b| \le \epsilon |b|}} \frac{\|\Delta_x\|_\infty}{\epsilon \|x_{STLS}\|_\infty} \tag{4.1}$$

and the componentwise condition number is defined by

$$\nu_{STLS} := \lim_{\epsilon \to 0} \sup_{\substack{|\Delta_A| \le \epsilon |A| \\ |\Delta_b| \le \epsilon |b|}} \frac{1}{\epsilon} \left\| \frac{\Delta_x}{x_{STLS}} \right\|_\infty, \tag{4.2}$$

recalling that $|\Delta_A| \le \epsilon |A|$ and $|\Delta_b| \le \epsilon |b|$ means $|\Delta_{A_{ij}}| \le \epsilon |A_{ij}|$ and $|\Delta_{b_i}| \le \epsilon |b_i|$.

The following theorem gives explicit expressions for the estimates of the above mixed and componentwise condition numbers for the STLS problem.

Theorem 4.1 *Using the notations above, we have the condition numbers*

$$\mu_{STLS} \le \frac{\| \, |M + N| \mathrm{vec}([|A| \quad \lambda |b|]) \|_\infty}{\|x_{STLS}\|_\infty} := \mu_{STLS}^M$$

and

$$\nu_{STLS} \le \left\| \frac{|M + N| \mathrm{vec}([|A| \quad \lambda |b|])}{x_{STLS}} \right\|_\infty := \nu_{STLS}^M$$

defined in Definition 4.1.

Proof If the perturbation Δ_A is structured as A, then the zero elements of A are not perturbed, that is, if $A_{ij} = 0$ then $\Delta_{A_{ij}} = 0$. Therefore,

$$\mathrm{vec}(\Delta_A) = D_A D_A^\dagger \mathrm{vec}(\Delta_A),$$

where $D_A = \mathrm{diag}(\mathrm{vec}(A))$ and $D_A^\dagger$ is the Moore-Penrose inverse of D_A [1, 30].

Denoting $D_b = \mathrm{diag}(b)$, rewriting (3.5) as

$$\begin{aligned}\Delta_x &= (M+N)\mathrm{vec}([\Delta_A \quad \lambda\Delta_b]) + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right) \\ &= (M+N)\begin{bmatrix}\mathrm{vec}(\Delta_A) \\ \lambda\Delta_b\end{bmatrix} + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right) \\ &= (M+N)\begin{bmatrix}D_A & \\ & \lambda D_b\end{bmatrix}\begin{bmatrix}D_A^\dagger \mathrm{vec}(\Delta_A) \\ D_b^\dagger \Delta_b\end{bmatrix} + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right),\end{aligned} \tag{4.3}$$

taking norms we have

$$\begin{aligned}\|\Delta_x\|_\infty \le &\left\|(M+N)\begin{bmatrix}D_A & \\ & \lambda D_b\end{bmatrix}\right\|_\infty \left\|\begin{bmatrix}D_A^\dagger \mathrm{vec}(\Delta_A) \\ D_b^\dagger \Delta_b\end{bmatrix}\right\|_\infty \\ &+ \|R(\Delta_A, \Delta_b)\|_\infty + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right).\end{aligned}$$

Similar to the proof of (3.8), we can finally obtain $\|R(\Delta_A, \Delta_b)\|_\infty \le \mathcal{O}(\epsilon^2)$ given that $|\Delta_A| \le \epsilon|A|$ and $|\Delta_b| \le \epsilon|b|$ in (4.1). Besides, we can also get $\mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right) \le \mathcal{O}(\epsilon^2)$. Therefore, we have

$$\|\Delta_x\|_\infty \le \epsilon \left\|(M+N)\begin{bmatrix}D_A & \\ & \lambda D_b\end{bmatrix}\right\|_\infty + \mathcal{O}\left(\epsilon^2\right).$$

According to (4.1), we have the mixed condition number

$$\begin{aligned}\mu_{STLS} &\le \frac{\left\|(M+N)\begin{bmatrix}D_A & \\ & \lambda D_b\end{bmatrix}\right\|_\infty}{\|x_{STLS}\|_\infty} \\ &= \frac{\left\||M+N|\begin{bmatrix}|D_A| & \\ & \lambda|D_b|\end{bmatrix}\right\|_\infty}{\|x_{STLS}\|_\infty} \\ &= \frac{\left\||M+N|\begin{bmatrix}|D_A| & \\ & \lambda|D_b|\end{bmatrix}e\right\|_\infty}{\|x_{STLS}\|_\infty} \\ &= \frac{\left\||M+N|\begin{bmatrix}\mathrm{vec}(|A|) \\ \lambda|b|\end{bmatrix}\right\|_\infty}{\|x_{STLS}\|_\infty} \\ &= \frac{\|\,|M+N|\mathrm{vec}([|A| \quad \lambda|b|])\|_\infty}{\|x_{STLS}\|_\infty},\end{aligned}$$

where e is an $(n+1)m$ dimensional vector with all entries equal to one.

According to the definition of ν in (4.2), if the i-th entry of x is 0 but the i-th entry of Δ_x is not 0, then $\nu = \infty$. Otherwise, from (4.3) we get

$$D^{\dagger}_{x_{STLS}}\Delta_x = D^{\dagger}_{x_{STLS}}(M+N)\begin{bmatrix} D_A & \\ & \lambda D_b \end{bmatrix}\begin{bmatrix} D_A^{\dagger}\mathrm{vec}(\Delta_A) \\ D_b^{\dagger}\Delta_b \end{bmatrix} + R(\Delta_A, \Delta_b) + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right),$$

taking norms and using the conditions $|\Delta_A| \le \epsilon|A|$ and $|\Delta_b| \le \epsilon|b|$ in (4.2), we have

$$\begin{aligned}\left\|\frac{\Delta_x}{x_{STLS}}\right\|_\infty &= \left\|D^{\dagger}_{x_{STLS}}\Delta_x\right\|_\infty \\ &\le \left\|D^{\dagger}_{x_{STLS}}(M+N)\begin{bmatrix} D_A & \\ & \lambda D_b \end{bmatrix}\right\|_\infty \left\|\begin{bmatrix} D_A^{\dagger}\mathrm{vec}(\Delta_A) \\ D_b^{\dagger}\Delta_b \end{bmatrix}\right\|_\infty \\ &\quad + \|R(\Delta_A, \Delta_b)\|_\infty + \mathcal{O}\left(\|[\Delta_A \quad \lambda\Delta_b]\|_F^2\right) \\ &\le \epsilon\left\|D^{\dagger}_{x_{STLS}}(M+N)\begin{bmatrix} D_A & \\ & \lambda D_b \end{bmatrix}\right\|_\infty + \mathcal{O}\left(\epsilon^2\right).\end{aligned}$$

According to (4.2), we have the componentwise condition number

$$\begin{aligned}\nu_{STLS} &\le \left\|D^{\dagger}_{x_{STLS}}(M+N)\begin{bmatrix} D_A & \\ & \lambda D_b \end{bmatrix}\right\|_\infty \\ &= \left\|\left|D^{\dagger}_{x_{STLS}}\right||M+N|\begin{bmatrix} |D_A| & \\ & \lambda|D_b| \end{bmatrix}\right\|_\infty \\ &= \left\|\left|D^{\dagger}_{x_{STLS}}\right||M+N|\begin{bmatrix} |D_A| & \\ & \lambda|D_b| \end{bmatrix}e\right\|_\infty \\ &= \left\|\left|D^{\dagger}_{x_{STLS}}\right||M+N|\begin{bmatrix} \mathrm{vec}(|A|) \\ \lambda|b| \end{bmatrix}\right\|_\infty \\ &= \left\|\frac{|M+N|\begin{bmatrix} \mathrm{vec}(|A|) \\ \lambda|b| \end{bmatrix}}{x_{STLS}}\right\|_\infty \\ &= \left\|\frac{|M+N|\mathrm{vec}([|A| \quad \lambda|b|])}{x_{STLS}}\right\|_\infty,\end{aligned}$$

where e is a $(n+1)m$ dimensional vector with all entries equal to one. □

In Section 5 we will show that the estimates of the mixed and componentwise condition numbers we obtained in this section are tight and μ^M_{STLS} and ν^M_{STLS} are nearly the reciprocal of the gap between the smallest singular value of A and the smallest singular value of $[A\ \ \lambda b]$ as long as the gap is sufficiently small.

As we know, the STLS solution approaches the LS solution when $\lambda \to 0$. In the following part, we show that these mixed and componentwise condition numbers for the STLS problem respectively approach some existing mixed and componentwise condition numbers for the LS problem, i.e., $\mu^M_{STLS} \to \mu_{LS}$ and $\nu^M_{STLS} \to \nu_{LS}$, when $\lambda \to 0$. Under the condition that A is of full column rank, Cucker et al. [2] derived the following mixed and componentwise condition numbers for the LS problem (2.4):

$$\mu_{LS} = \frac{\left\| \left| -\left(x_{LS}^T \otimes A^\dagger\right) + \left(A^TA\right)^{-1} \otimes \left(b - Ax_{LS}\right)^T \right| \mathrm{vec}(|A|) + \left|A^\dagger\right| |b| \right\|_\infty}{\|x_{LS}\|_\infty},$$

$$\nu_{LS} = \left\| \frac{\left| -\left(x_{LS}^T \otimes A^\dagger\right) + \left(A^TA\right)^{-1} \otimes \left(b - Ax_{LS}\right)^T \right| \mathrm{vec}(|A|) + \left|A^\dagger\right| |b|}{x_{LS}} \right\|_\infty.$$

By Theorem 4.1, we get

$$\mu^M_{STLS} = \frac{\| \, |M+N| \mathrm{vec}([|A| \ \ \lambda|b|]) \|_\infty}{\|x_{STLS}\|_\infty} = \frac{\left\| |M+N| \begin{bmatrix} \mathrm{vec}(|A|) \\ \lambda|b| \end{bmatrix} \right\|_\infty}{\|x_{STLS}\|_\infty}.$$

As $\lambda \to 0$, we have $\sigma_{n+1} \to 0$, $x_{STLS} \to x_{LS}$, $K \to (A^TA)^{-1}$, and $KA^T \to (A^TA)^{-1}A^T = A^\dagger$. From (3.6) and (3.7), we get

$$M \to \left[\left(A^TA\right)^{-1} \otimes b^T - x_{LS}^T \otimes A^\dagger - \left(A^TA\right)^{-1} \otimes \left(Ax_{LS}\right)^T \ \ \middle| \ \ \lambda^{-1}A^\dagger \right]$$

and $N \to 0$ as $\lambda \to 0$. Therefore, we have

$$\begin{aligned} \mu^M_{STLS} &\to \frac{\left\| \left| \left(A^TA\right)^{-1} \otimes b^T - x_{LS}^T \otimes A^\dagger - \left(A^TA\right)^{-1} \otimes \left(Ax_{LS}\right)^T \right| \mathrm{vec}(|A|) + \left|A^\dagger\right| |b| \right\|_\infty}{\|x_{LS}\|_\infty} \\ &= \frac{\left\| \left| -x_{LS}^T \otimes A^\dagger + \left(A^TA\right)^{-1} \otimes \left(b - Ax_{LS}\right)^T \right| \mathrm{vec}(|A|) + \left|A^\dagger\right| |b| \right\|_\infty}{\|x_{LS}\|_\infty} \\ &= \mu_{LS}. \end{aligned}$$

Similarly, we can show that $\nu_{STLS} \to \nu_{LS}$ as $\lambda \to 0$. Thus, the condition numbers for the STLS problem presented in this paper are a generalization of the

condition numbers for the LS problem given in [2]. Moreover, since the TLS problem is a special case of the STLS problem when $\lambda = 1$, the condition numbers we derived for the STLS problem become the condition numbers for the TLS problem when $\lambda = 1$.

5 Numerical examples

Since there are no existing condition numbers for the STLS problem, we first compare our result (3.5) when $\lambda = 1$ with some existing results of the perturbation analysis of the TLS problem. Then we demonstrate the accuracy of our mixed and componentwise condition numbers for various values of λ. All the experiments were carried out using MATLAB 7.0.

First of all, we review some existing results of the perturbation analysis of the TLS problem. Golub and Van Loan [5] provided a numerical analysis of the TLS problem under the condition $\hat{\sigma}_n > \sigma_{n+1}$. Van Huffel and Vandewalle [27], generalized the results in [5] to multidimensional problems. Then Van Huffel and Vandewalle [25] and Wei [31] considered the case when $\hat{\sigma}_p > \sigma_{p+1} = \cdots = \sigma_{n+1}$ for some $p < n$. Later, Van Huffel and Vandewalle [28] showed that even if the former condition is not satisfied, a TLS solution (unique or not unique) may still exist. For rank deficient problems $k \le n$, where k denotes numerical rank of A, Fierro and Bunch [3] derived perturbation bounds for TLS solutions in terms of subspace angle between approximate nullspaces. Further analysis can also be found in Liu [12] and Wei [32]. We summarize some existing results as follows.

Recall the SVDs:

$$A = \hat{U}\hat{\Sigma}\hat{V}^T, \quad \hat{\Sigma} = \operatorname{diag}(\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_n),$$

$$[A \;\; \lambda b] = U\Sigma V^T, \quad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, ..., \sigma_{n+1})$$

and partition $U = [U_1 \quad u_{n+1}]$ and $V = [V_1 \quad v_{n+1}]$, where u_{n+1} and v_{n+1} are the last columns of U and V respectively. Denote

$$\left[\tilde{A} \;\; \tilde{b}\right] := \left[A + \Delta_A \;\; b + \Delta_b\right] = \tilde{U}\tilde{\Sigma}\tilde{V}^T, \quad \tilde{\Sigma} = \operatorname{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, ..., \tilde{\sigma}_{n+1}),$$

and partition $\tilde{U} = [\tilde{U}_1\tilde{u}_{n+1}]$ and $\tilde{V} = [\tilde{V}_1\tilde{v}_{n+1}]$, where $\tilde{u}_{n+1}$ and $\tilde{v}_{n+1}$ are the last columns of $\tilde{U}$ and $\tilde{V}$ respectively. Let x_{TLS} be the TLS solution corresponding to $[A \;\; b]$ and $\tilde{x}_{TLS}$ the TLS solution corresponding to $\left[\tilde{A} \;\; \tilde{b}\right]$.

[5, Thm. 4.4] Condition: $\|[\Delta_A \;\; \Delta_b]\|_F \le \hat{\sigma}_n - \sigma_{n+1}$.
Result:

$$\|x_{TLS} - \tilde{x}_{TLS}\|_2 \le \frac{9\sigma_1\|[\Delta_A \;\; \Delta_b]\|_F}{\sigma_n - \sigma_{n+1}}\left(1 + \frac{\|b\|_2}{\hat{\sigma}_n - \sigma_{n+1}}\right) \times \frac{1}{\|b\|_2 - \sigma_{n+1}}\|x_{TLS}\|_2.$$

[31, Thm. 4.1] Condition: $\|[\Delta_A \;\; \Delta_b]\|_2 \leq (\hat{\sigma}_n - \sigma_{n+1})/6$.
Result:

$$\begin{aligned}\|x_{TLS} - \tilde{x}_{TLS}\|_2 &\leq \frac{3\big(\|[\Delta_A \Delta_b]\|_2 + \sigma_{n+1}\big)}{\hat{\sigma}_n - \sigma_{n+1}}\big(1 + \|x_{TLS}\|_2\big) \\ &\leq \frac{9\big(\|[\Delta_A \Delta_b]\|_2 + \sigma_{n+1}\big)}{2(\hat{\sigma}_n - \sigma_{n+1})}\sqrt{1 + \|x_{TLS}\|_2^2} \\ &\leq \frac{9\big(\|[\Delta_A \Delta_b]\|_2 + \sigma_{n+1}\big)}{\hat{\sigma}_n - \sigma_{n+1}}\frac{\sigma_1}{\|b\|_2 - \sigma_{n+1}}\|x_{TLS}\|_2.\end{aligned}$$

[12, Thm. 3] Condition: $\|[\Delta_A \;\; \Delta_b]\|_2 \leq (\hat{\sigma}_n - \sigma_{n+1})/2$.
Result:

$$\|x_{TLS} - \tilde{x}_{TLS}\|_2 \leq 2\sqrt{2}\sqrt{\alpha^{-2} + 2\mu^2}\,\frac{\|[\Delta_A \quad \Delta_b]\|_F}{\sigma_n - \sigma_{n+1}},$$

where $\alpha = e_{n+1}^T v_{n+1}$, $\beta = e_{n+1}^T \tilde{v}_{n+1}$, and $\mu = \max\{\alpha^{-2}, \beta^{-2}\}$.

[32, Thm. 3.2] Condition: $\|[\Delta_A \;\; \Delta_b]\|_2 \leq \tilde{\sigma}_n - \sigma_{n+1}$.
Result:

$$\begin{aligned}\sin\phi_{TLS} &\leq \|x_{TLS} - \tilde{x}_{TLS}\|_2 \\ &\leq \sin\phi_{TLS}\sqrt{\big(\|x_{TLS}\|_2^2 + 1\big)\big(\|\tilde{x}_{TLS}\|_2^2 + 1\big)},\end{aligned}$$

where $\sin\phi_{TLS}$ denotes the sine of the angle between the nullspaces of $[A \;\; b]$ and $[\tilde{A} \;\; \tilde{b}]$ and under reasonable assumptions, $\sin\phi_{TLS} \approx \|[\Delta_A \quad \Delta_b]\|_2/\sigma_n$.

[32, Theorem 3.2] Conditions: $\|[\Delta_A \quad \Delta_b]\|_2 < (\sigma_n - \sigma_{n+1})/2$ and $e_{n+1}^T v_{n+1}$, $e_{n+1}^T \tilde{v}_{n+1} > 0$.
Result:

$$\begin{aligned}\left\|\tilde{V}_1^T v_{n+1}\right\|_2 &\leq \|x_{TLS} - \tilde{x}_{TLS}\|_2 \\ &\leq \left\|\tilde{V}_1^T v_{n+1}\right\|_2\sqrt{\big(\|x_{TLS}\|_2^2 + 1\big)\big(\|\tilde{x}_{TLS}\|_2^2 + 1\big)}.\end{aligned}$$

Table 1 Comparison of the perturbation estimate $\|x_{TLS} - \tilde{x}_{TLS}\|_2$

	$N = 5$	$N = 60$	$N = 200$
(3.5)	$6.8055e-9$	$6.1083e-8$	$9.1965e-8$
[5, Thm. 4.4]	$3.4064e-6$	$8.3802e-5$	$4.1706e-4$
[31, Thm.4.1]	$8.4152e+1$	$1.9185e+2$	$3.2989e+2$
[12, Thm. 3]	$2.8454e-7$	$1.2788e-5$	$6.9388e-5$
[3, Thm. 3.2]	$3.2072e-8$	$5.8260e-7$	$1.9715e-6$
[32, Thm. 3.2]	$8.7926e-9$	$1.8890e-7$	$5.6663e-7$
Exact perturbation	$4.6959e-9$	$2.6764e-8$	$5.8839e-8$

Table 2 Comparison of the computed perturbation $\|\tilde{x}_{STLS} - x_{STLS}\|_2$ with our estimate (3.5)

m	n	λ	$\|\Delta_x\|_2$	Estimate (3.5)
75	20	0.005	$2.7511e-009$	$2.7511e-009$
750	50	0.005	$1.1130e-009$	$1.1130e-009$
75	20	1	$3.7443e-005$	$3.7471e-005$
750	50	1	$5.2471e-006$	$5.2502e-006$
75	20	10	$4.2346e-006$	$4.2356e-006$
750	50	10	$4.1477e-005$	$4.1528e-005$

Example 1 The data $A \in \mathbb{R}^{N\times(N-2)}$ and $b \in \mathbb{R}^N$ were adopted from [23, page 20]:

$$A = \begin{bmatrix} N-1 & -1 & \cdots & -1 \\ -1 & N-1 & \ldots & -1 \\ \vdots & \vdots & \cdots & \vdots \\ -1 & -1 & \cdots & N-1 \\ -1 & -1 & \cdots & -1 \\ -1 & -1 & \cdots & -1 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ N-1 \\ -1 \end{bmatrix}.$$

So the exact

$$x_{TLS} = -\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \sigma_{n+1} = \sqrt{N}, \quad \text{and} \quad \hat{\sigma}_n = \sqrt{2N}.$$

The entries of the perturbations Δ_A and Δ_b were generated by $\eta_{ij}A_{ij}$ and $\xi_i b_i$ respectively, where η_{ij} and ξ_i are random variables uniformly distributed on $(0, 10^{-9})$. Along with the exact perturbation, Table 1 compares our perturbation estimate (3.5) ($\lambda = 1$) with the above existing results, with various values of N.

In the following two examples, the entries of the data A and b were generated as random variables normally distributed with mean zero and variance one. The entries of the perturbations Δ_A and Δ_b were generated by $\eta_{ij}A_{ij}$ and $\xi_i b_i$ respectively, where η_{ij} and ξ_i are random variables uniformly distributed on the interval $(0, 10^{-8})$. Thus the sizes of the relative perturbations $|\Delta_A|/|A|$

Table 3 Comparisons of our mixed and componentwise condition numbers with the computed relative perturbations $\|\Delta_x\|_\infty/\|x_{STLS}\|_\infty$ and $\left\|\frac{\Delta_x}{x_{STLS}}\right\|_\infty$

m	n	λ	$\frac{\|\Delta_x\|_\infty}{\|x_{STLS}\|_\infty}$	$\epsilon\,\mu^M_{STLS}$	$\left\|\frac{\Delta_x}{x_{STLS}}\right\|_\infty$	$\epsilon\,\nu^M_{STLS}$
75	20	0.005	$4.8368e-009$	$1.8065e-007$	$4.2757e-007$	$2.9083e-005$
750	50	0.005	$4.3023e-009$	$3.4165e-007$	$1.9138e-006$	$5.1544e-004$
75	20	1	$4.3625e-008$	$4.7337e-006$	$1.9570e-006$	$1.0771e-004$
750	50	1	$5.0974e-008$	$2.6190e-005$	$1.3555e-006$	$6.5079e-004$
75	20	10	$9.0560e-008$	$7.9714e-006$	$5.4786e-007$	$5.3745e-005$
750	50	10	$4.4682e-008$	$1.7573e-005$	$1.8704e-006$	$8.4454e-004$

Table 4 Comparisons of the difference $\hat{\sigma}_n - \sigma_{n+1}$ with our condition numbers

m	n	λ	$\hat{\sigma}_n - \sigma_{n+1}$	κ^M_{STLS}	μ^M_{STLS}	ν^M_{STLS}
6	4	0.005	$1.5343e-013$	$6.0531e+012$	$5.0590e+012$	$5.5046e+012$
100	15	0.005	$9.0949e-013$	$1.9718e+013$	$1.0324e+013$	$8.3093e+013$
6	4	1	$8.2057e-013$	$2.3134e+012$	$1.3575e+012$	$1.7667e+012$
100	15	1	$2.5846e-013$	$3.2349e+013$	$2.2703e+013$	$8.3680e+013$
6	4	10	$2.7338e-012$	$4.3990e+011$	$2.3608e+011$	$8.5680e+011$
100	15	10	$7.1498e-014$	$1.8827e+014$	$1.2201e+014$	$3.9294e+014$

and $|\Delta_b|/|b|$ were about $\epsilon = 10^{-8}$. The STLS solutions were computed using (2.3). For each size of A, given by m and n, we carried out 500 random experiments. Each figure in the following tables is the averages of 500 experiments.

Example 2 In this example, we investigate the accuracy of our first order perturbation estimate (3.5) for the STLS solution. We experimented on three values of λ, namely 0.005 (a value close to 0), 1, and 10 (a large value). Table 2 lists the results. Our experiments show that the estimate (3.5) is very accurate as expected.

Example 3 In this example, we compare the relative perturbations $\|\Delta_x\|_\infty / \|x_{STLS}\|_\infty$ and $\left\| \frac{\Delta_x}{x_{STLS}} \right\|_\infty$ with our mixed and componentwise condition numbers respectively. From Definition 4.1, for small ϵ,

$$\frac{\|\Delta_x\|_\infty}{\|x_{STLS}\|_\infty} \le \epsilon\, \mu^M_{STLS} + \mathcal{O}\left(\epsilon^2\right) \quad \text{and} \quad \left\| \frac{\Delta_x}{x_{STLS}} \right\|_\infty \le \epsilon\, \nu^M_{STLS} + \mathcal{O}\left(\epsilon^2\right).$$

As shown in Table 3, our experimental results demonstrate that our condition numbers are tight.

Example 4 In this example, we constructed examples of $[A\ b]$ such that $\hat{\sigma}_n$, the smallest singular value of A, is close to σ_{n+1} and compared the difference $\hat{\sigma}_n - \sigma_{n+1}$ with the condition numbers we derived in our paper. As shown in Table 4, our experimental results demonstrate that our condition numbers are approximately the reciprocal of the difference $\hat{\sigma}_n - \sigma_{n+1}$ when the difference is sufficiently small.

6 Concluding remarks

In this paper, we derive a first order estimate for the perturbation in the STLS solution (3.5). Based on the estimate, we present the normwise, mixed and componentwise condition numbers for the STLS problem under the condition $\hat{\sigma}_n > \sigma_{n+1}$. Our condition numbers include the previous results on the condition numbers for the LS problems as special cases and give the condition numbers for the TLS problem by setting $\lambda = 1$. Our numerical experiments demonstrate

that our perturbation estimate is very accurate and condition numbers are very sharp. In the case when $\hat{\sigma}_p > \sigma_{p+1}$ for some $p < n$, the STLS problem may have more than one solution [31]. Hence how to derive the explicit expressions of these three condition numbers for STLS problems under this more general condition is worthy of further study.

Acknowledgements The authors would like to thank Professors Sabine van Huffel and Z. Strakoš with two referees for their useful suggestions on this topic.

References

1. Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications, 2nd edn. Springer, New York (2003)
2. Cucker, F., Diao, H., Wei, Y.: On mixed and componentwise condition numbers for Moore-Penrose inverse and linear least squares problems. Math. Comput. **76**, 947–963 (2007)
3. Fierro, R.D., Bunch, J.R.: Perturbation theory for orthogonal projection methods with applications to least squares and total least squares. Linear Algebra Appl. **234**, 71–96 (1996)
4. Gohberg, I., Koltracht, I.: Mixed, componentwise, and structured condition numbers. SIAM J. Matrix Anal. Appl. **14**, 688–704 (1993)
5. Golub, G.H., Van Loan, C.F.: An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**, 883–893 (1980)
6. Graham, A.: Kronecker Products and Matrix Calculus with Application. Wiley, New York (1981)
7. Higham, N.J.: Accuracy, Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)
8. Higham, N.J.: A survey of componentwise perturbation theory in numerical linear algebra. Proc. Symp. Appl. Math. **48**, 49–77 (1994)
9. Hnětynková, I., Strakoš, Z.: Lanczos tridiagonalization and core problems. Linear Algebra Appl. **421**, 243–251 (2007)
10. Kukush, A., Markovsky, I., Van Huffel, S.: Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. J. Stat. Plan. Inference **133**, 315–358 (2005)
11. Kukush, A., Van Huffel, S.: Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model $AX = B$. Metrika **59**, 75–97 (2004)
12. Liu, X.: On the solvability and perturbation analysis for tatal least squares problem. Acta Math. Appl. Sin. **19**, 254–262 (1996) (in Chinese)
13. Markovsky, I., Rastello, M.L., Premoli, A., Kukush, A., Van Huffel, S.: The element-wise weighted total least-squares problem. Comput. Stat. Data Anal. **50**, 181–209 (2006)
14. Markovsky, I., Van Huffel, S.: Overview of total least-squares methods. Signal Process. **87**, 2283–2302 (2007)
15. Paige, C.C., Strakoš, Z.: Bounds for the least squares distance using scaled total least squares. Numer. Math. **91**, 93–115 (2002)
16. Paige, C.C., Strakoš, Z.: Scaled total least squares fundamentals. Numer. Math. **91**, 117–146 (2002)
17. Paige, C.C., Strakoš, Z.: Core problems in linear algebraic systems. SIAM J. Matrix Anal. Appl. **27**, 861–875 (2006)
18. Rao, B.D.: Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework. In: Van Huffel, S. (ed.) Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling, pp. 11–20. SIAM, Philadelphia (1997)
19. Rohn, J.: New condition numbers for matrices and linear systems. Computing **41**, 167–169 (1989)
20. Skeel, R.D.: Scaling for numerical stability in Gaussian elimination. J. Assoc. Comput. Math. **26**, 167–169 (1979)

21. Stewart, G.W.: A second order perturbation expansion for small singular values. Linear Algebra Appl. **56**, 231–235 (1984)
22. Sun, J.-G.: A note on simple non-zero singular values. J. Comput. Math. **6**, 258–266 (1988)
23. Van Huffel, S.: Analysis of the total least squares problem and its use in parameter estimation. Dissertation, ESAT Lab., Dept. Electr. Eng., K.U. Leuven (1987)
24. Van Huffel, S.: On the significance of nongeneric total least squares problems. SIAM J. Matrix Anal. Appl. **13**, 20–35 (1992)
25. Van Huffel, S., Vandewalle, J.: Analysis and solution of the nongeneric total least squares problems. SIAM J. Matrix Anal. Appl. **9**, 360–372 (1988)
26. Van Huffel, S., Vandewalle, J.: Analysis and properties of the generalized total least squares problems $AX \approx B$ when some or all columns in A are subject to error. SIAM J. Matrix Anal. Appl. **10**, 294–315 (1989)
27. Van Huffel, S., Vandewalle, J.: Algebraic connections between the least squares and total least squares problems. Numer. Math. **55**, 431–449 (1989)
28. Van Huffel, S., Vandewalle, J.: The Total Least Squares Problem: Computational Aspects and Analysis. SIAM, Philadelphia (1991)
29. Van Huffel, S., Zha, H.: The restricted total least squares problem: formulation, algorithm, and properties. SIAM J. Matrix Anal. Appl. **12**, 292–309 (1991)
30. Wang, G., Wei, Y., Qiao, S.: Generalized Inverses: Theory and Computations. Science, Beijing (2004)
31. Wei, M.: The analysis for the total least squares problem with more than one solution. SIAM J. Matrix Anal. Appl. **13**, 746–763 (1992)
32. Wei, M.: On the perturbation of the LS and TLS problems. Math. Numer. Sinica **20**(3), 267–278 (1998) (in Chinese)
33. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford University Press, London (1965)