Calcolo (2018) 55:46
https://doi.org/10.1007/s10092-018-0289-9

# A note on the condition number of the scaled total least squares problem

**Shaoxin Wang**[1] · **Hanyu Li**[2] · **Hu Yang**[2]

Received: 1 April 2018 / Accepted: 23 October 2018 / Published online: 25 October 2018


## Abstract

In this paper, we show that the normwise condition number of the scaled total least squares problem can be transformed into a new and compact form. Considering the relationship between the scaled total least squares problem and the total least squares problem, we obtain something new on the normwise condition number of the total least squares problem. The new forms of the normwise condition number are of particular interest in the following two aspects. Firstly, it is easy to use for the practitioners from applied disciplines. Secondly, the new forms enjoy great computational efficiency and require very little storage space compared with its original forms. Numerical examples are given to illustrate the results.

**Keywords** The scaled total least squares problem · Normwise condition number · Compact form · Condition number estimation

**Mathematics Subject Classification** 65F35 · 15A63 · 15A06

The work is supported by a project of Shandong Province Higher Educational Science and Technology Program (Grant No. J17KA160), and the National Natural Science Foundation of China (Grant Nos. 11671059, 11671060).

✉ Shaoxin Wang
shwangmy@163.com; shxwang@qfnu.edu.cn

Hanyu Li
hyli@cqu.edu.cn

Hu Yang
hy@cqu.edu.cn

[1] School of Statistics, Qufu Normal University, Qufu 273165, People's Republic of China

[2] College of Mathematics and Statistics, Chongqing University, Chongqing 401331, People's Republic of China

## 1 Introduction

The scaled total least squares (STLS) problem (or technique) was first proposed in [20] to give a unified treatment of the ordinary least squares (OLS) problem, the total least squares (TLS) problem and the data least squares (DLS) problem. Paige and Strakoš [19] reformulated the STLS problem and presented a detailed analysis of conditions that guarantee the STLS problem has a unique solution. With their formulation, the STLS problem is given as follows

$$\min \left\| \left[ E,\ f \right] \right\|_F, \quad \text{subject to } \lambda b - f \in \mathscr{R}(A + E), \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\lambda$ is a positive real number, $\|\cdot\|_F$ denotes the Frobenius norm and $\mathscr{R}(\cdot)$ is the range space. Let $[E_S,\ f_S]$ be the solution to (1), then the solution to the linear system $(A + E_S)\lambda x = \lambda b - f_S$ is called the STLS solution and denoted by $x_S$. As shown in [19], when $\lambda = 1$, $\lambda \to 0$ and $\lambda \to \infty$, $x_S$ becomes the TLS solution $x_T$, the OLS solution $x_O$ and the DLS solution $x_D$, respectively.

The condition number gives a quantitative measurement of the maximum amplification of the resulting change in solution with respect to a perturbation in the data, and has been extensively studied. The interested reader is referred to the comprehensive survey [5]. For the STLS problem, Zhou et al. [28] considered its perturbation analysis and presented the normwise, mixed and componentwise condition numbers. With the perturbation theory of singular value decomposition (SVD) given in [24], Li and Jia [16] gave a different approach to derive the normwise and componentwise condition numbers of the STLS problem, and the corresponding structured condition numbers were also discussed. It should be noted that the normwise condition number in [28] contains Kronecker products which make it impractical to compute for large-scale problems. Based on the fact that $\|A\|_2 = \|A^T A\|_2^{1/2}$, where $\|\cdot\|_2$ denotes the spectral norm of matrix or Euclidean norm of vector, some closed formulas and upper (or lower) bounds of the normwise condition number for the TLS problem were given in [1,14], and these results are easy to compute and do not contain Kronecker product any more. Xie et al. [27] showed that the expressions of the condition numbers given in [1,14] are mathematically equivalent. However, computing exact value of the condition number with the formula given in [1,27] needs to calculate the matrix cross product, which is a source of rounding error and potentially numerical unstable [11, p. 386]. By designing iterative procedure or exploiting the SVDs in solving the TLS problem, some progress to avoid computing matrix cross product was made in [1,14,27]. In this paper, we present a new explicit expression of the normwise condition number of the STLS problem. The new expression is easy to compute and does not need to compute Kronecker product or matrix cross product. Meanwhile, the new expression should be of particular interest to the practitioners from applied disciplines, who are more likely to directly compute the normwise condition number of the STLS problem.

The rest of the paper is organized as follows: Sect. 2 contains the main results of the paper. Numerical experiments are presented in Sect. 3. Concluding remarks are given in Sect. 4. Before proceeding to the following sections, we introduce some notation first. For any matrix $B$, $A \otimes B = [a_{ij}B]$ denotes the Kronecker product of $A$ and $B$.

$A \circ B = [a_{ij}b_{ij}]$ denotes the Hadamard product of $A$ and $B$. vec$(\cdot)$ is a linear map defined by $\text{vec}(A) = [a_{1,1}, \ldots, a_{m,1}, \ldots, a_{1,n}, \ldots, a_{m,n}]^T$.

## 2 Main results

As stated in the Introduction, the TLS problem can be treated as a special case of the STLS problem. An interesting result is that we can solve the STLS problem by finding the solution to a special TLS problem. When $\lambda = 1$, we get the following TLS problem

$$\min \left\| \left[ E,\ f \right] \right\|_F, \quad \text{subject to } \ b - f \in \mathscr{R}(A + E). \tag{2}$$

It is easy to check that when $x_S$ is the solution of (1), $\lambda x_S$ is the TLS solution to the following TLS problem

$$\min \left\| \left[ E,\ f \right] \right\|_F, \quad \text{subject to } \ \lambda b - f \in \mathscr{R}(A + E). \tag{3}$$

Let the SVDs of the matrices $[A,\ \lambda b]$ and $A$ be

$$U^T \left[ A,\ \lambda b \right] V = \Sigma, \quad \hat{U}^T A \hat{V} = \hat{\Sigma},$$

where $U = [u_1, \ldots, u_{n+1}] \in \mathbb{R}^{m\times(n+1)}$, $V = [v_1, \ldots, v_{n+1}] \in \mathbb{R}^{(n+1)\times(n+1)}$, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{n+1})$ with $\sigma_1 \geq, \ldots, \geq \sigma_{n+1} \geq 0$, $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_n] \in \mathbb{R}^{m\times n}$, $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_n] \in \mathbb{R}^{n\times n}$, and $\hat{\Sigma} = \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$ with $\hat{\sigma}_1 \geq, \ldots, \hat{\sigma}_n \geq 0$. Analogous to the Golub and Van Loan condition [9] for the TLS problem to guarantee the existence and uniqueness of a solution, Zhou et al. [28] presented the following sufficient condition to ensure the STLS problem has a unique solution

$$\hat{\sigma}_n > \sigma_{n+1} > 0, \tag{4}$$

which implies that both $A$ and $[A, b]$ are of full column rank. Therefore, by (4) the TLS solution to (3) is

$$\lambda x_S = (A^T A - \sigma_{n+1} I_n)^{-1} A^T (\lambda b),$$

which gives

$$x_S = (A^T A - \sigma_{n+1} I_n)^{-1} A^T b. \tag{5}$$

From (5), we get the normal equation of the STLS problem

$$(A^T A - \sigma_{n+1} I_n) x_S = A^T b. \tag{6}$$

When $\lambda = 1$, the Rayleigh quotient iteration (RQI) and preconditioned conjugate gradient method (PCG) were combined to solve the TLS problem with the normal

equation (6) in [4]. For a given $\lambda$, by substituting the block matrix $[A, b]$ with $[A, \lambda b]$, the RQIPCG method can be directly applied to solve the STLS problem.

Let $\Delta A$ and $\Delta b$ be the corresponding perturbations to $A$ and $b$, then we have the following perturbed STLS problem

$$\min \left\| \left[ E,\ f \right] \right\|_F,\ \text{subject to } \lambda(b + \Delta b) - f \in \mathscr{R}\left((A + \Delta A) + E\right). \tag{7}$$

For the perturbed STLS problem, Zhou et al. [28] and Li and Jia [16] presented two different approaches to show that when the perturbation $[\Delta A,\ \Delta b]$ is sufficiently small, the perturbed STLS problem admits a unique solution. We adapt the result given in [16] as the following theorem.

**Theorem 1** *Under the assumption (4), if $\|[\Delta A,\ \Delta b]\|_F$ is small enough, then the perturbed STLS problem (7) has unique solution. Moreover, if we denote the solution by $x_{PS}$, then*

$$\Delta x = x_{PS} - x_S = K \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \Delta b \end{bmatrix} + \mathscr{O}\left( \|[\Delta A, \Delta b]\|_F^2 \right), \tag{8}$$

*where*

$$K = M^{-1} \left( \left( \frac{2}{\|r\|_2^2} A^T r r^T - A^T \right) \left( \left[ x_S^T, -1 \right] \otimes I_m \right) - \left[ I_n \otimes r^T,\ 0_{n \times m} \right] \right) \tag{9}$$

*with $M = A^T A - \sigma_{n+1}^2 I_n$ and $r = A x_S - b$.*

Li and Jia [16] also presented a very detailed comparison of the above results with those given in [28], and showed that their perturbation estimate is the same as that given in [28]. According to Theorem 1, it can be easily deduced that if we set

$$\begin{aligned} F : \mathbb{R}^{m \times n} \times \mathbb{R}^m &\to \mathbb{R}^n, \\ \left[ A,\ \lambda b \right] &\to x_S = M^{-1} A^T b, \end{aligned}$$

then the map $F$ is Fréchet differentiable at $[A,\ \lambda b]$ under the Assumption (4) and the Fréchet derivative of $F$ at $[A,\ \lambda b]$ is given by

$$DF(A, \lambda b) := K.$$

According to the definition of condition number given in [8,21], the relative normwise condition number of the STLS problem is given by

$$\kappa_{rF}(A, \lambda b) = \lim_{\delta \to 0} \sup_{\|[\Delta A,\ \lambda \Delta b]\|_F < \delta} \frac{\frac{\|F(A+\Delta A, \lambda(b+\Delta b)) - F(A, \lambda b)\|_2}{\|F(A, \lambda b)\|_2}}{\frac{\|[\Delta A,\ \lambda \Delta b]\|_F}{\|[A,\ \lambda b]\|_F}}. \tag{10}$$

When $F$ is Fréchet differentiable, $\kappa_{rF}(A, \lambda b)$ reduces to

$$\kappa_{rF}(A, \lambda b) = \frac{\|DF\,(A, \lambda b)\|_2\, \|[A,\ \lambda b]\|_F}{\|F(A, \lambda b)\|_2},$$

and $\kappa_F(A, \lambda b) = \|DF\,(A, \lambda b)\|_2$ is the absolute condition number. For the convenience of presentation, we summarize the above discussion as the following theorem.

**Theorem 2** *Under the assumption* (4)*, the relative normwise condition number of the STLS problem defined by* (10) *is*

$$\kappa_{rF}(A, \lambda b) = \frac{\|K\|_2\, \|[A,\ \lambda b]\|_F}{\|F(A, \lambda b)\|_2},$$

*and its absolute condition number is*

$$\kappa_F(A, \lambda b) = \|K\|_2\,, \tag{11}$$

*where $K$ is given by* (9)*.*

It should be noted that the Kronecker product enlarges the size of matrix and may make it impractical to explicitly form $K$ when $m$ and $n$ are large. To eliminate the influence of the Kronecker product, we present a new *compact form* of the normwise condition number for the STLS problem in the following theorem. Considering the relationship between relative and absolute condition numbers, we only focus on $\kappa_F(A, \lambda b)$ in the following parts.

**Theorem 3** *The absolute condition number $\kappa_F(A, \lambda b)$ for the STLS problem has the following two equivalent forms*

$$\kappa_{F1}(A, \lambda b) = \left\| M^{-1} \left( (1 + \|x_S\|_2^2) A^T A - A^T r x_S^T - x_S r^T A + \|r\|_2^2 I_n \right) M^{-1} \right\|_2^{\frac{1}{2}}, \tag{12}$$

*and*

$$\kappa_{F2}(A, \lambda b) = \left\| M^{-1} \left[ A^T,\ \|x_S\|_2 A^T \left( I_m - \tfrac{1}{\|r\|_2^2} r r^T \right),\ \|r\|_2 \left( I_n - \tfrac{1}{\|r\|_2^2} A^T r x_S^T \right) \right] \right\|_2, \tag{13}$$

*where $M$ and $r$ are given in* (9)*.*

***Proof*** For a real matrix $X$, $\|X\|_2 = \|X^T X\|_2^{1/2} = \|X X^T\|_2^{1/2}$ holds. Thus with (9) and (11) we have

$$\kappa_F(A, \lambda b) = \|K\|_2 = \left\| K K^T \right\|_2^{\frac{1}{2}}.$$

Since $M$ is symmetric, by the equality $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$ [13, Ch. 4] we can get

$$\begin{aligned} KK^T &= M^{-1}\left(\left(\frac{2}{\|r\|_2^2}A^T rr^T - A^T\right)\left([x_S^T, -1]\otimes I_m\right) - \left[I_n \otimes r^T, 0_{n\times m}\right]\right) \\ &\quad \times \left(\left(\begin{bmatrix} x_S \\ -1 \end{bmatrix}\otimes I_m\right)\left(\frac{2}{\|r\|_2^2}rr^T A - A\right) - \begin{bmatrix} I_n \otimes r \\ 0_{m\times n} \end{bmatrix}\right) M^{-1} \\ &= M^{-1}\left(\left(1+\|x_S\|_2^2\right)A^T A - A^T r x_S^T - x_S r^T A + \|r\|_2^2 I_n\right) M^{-1} \qquad (14) \\ &= M^{-1}\left(\left[A^T, I_n\right]\begin{bmatrix} \left(1+\|x_S\|_2^2\right)I_m, & -rx_S^T \\ -x_S r^T, & \|r\|_2^2 I_n \end{bmatrix}\begin{bmatrix} A \\ I_n \end{bmatrix}\right) M^{-1}. \qquad (15) \end{aligned}$$

Since

$$\begin{aligned} \begin{bmatrix} \left(1+\|x_S\|_2^2\right)I_m, & -rx_S^T \\ -x_S r^T, & \|r\|_2^2 I_n \end{bmatrix} &= \begin{bmatrix} I_m, & -\frac{1}{\|r\|_2^2}rx_S^T \\ 0_{n\times m}, & I_n \end{bmatrix}\begin{bmatrix} I_m + \|x_S\|_2^2\left(I_m - \frac{1}{\|r\|_2^2}rr^T\right), & 0_{m\times n} \\ 0_{n\times m}, & \|r\|_2^2 I_n \end{bmatrix} \\ &\quad \times \begin{bmatrix} I_m, & 0_{m\times n} \\ -\frac{1}{\|r\|_2^2}x_S r^T, & I_n \end{bmatrix} \qquad (16) \end{aligned}$$

and

$$\begin{aligned} &I_m + \|x_S\|_2^2\left(I_m - \frac{1}{\|r\|_2^2}rr^T\right) \\ &= \left[I_m, \|x_S\|_2\left(I_m - \tfrac{1}{\|r\|_2^2}rr^T\right)\right]\begin{bmatrix} I_m \\ \|x_S\|_2\left(I_m - \frac{1}{\|r\|_2^2}rr^T\right) \end{bmatrix}, \qquad (17) \end{aligned}$$

we substitute (16) and (17) into (15) and get

$$KK^T = M^{-1}WW^T M^{-1}, \qquad (18)$$

where

$$W = \left[A^T, \|x_S\|_2 A^T\left(I_m - \tfrac{1}{\|r\|_2^2}rr^T\right), \|r\|_2\left(I_n - \tfrac{1}{\|r\|_2^2}A^T r x_S^T\right)\right].$$

Using the formula $\|X\|_2 = \|XX^T\|_2^{1/2}$ again, (12) and (13), the two equivalent forms of $\kappa_F(A, \lambda b)$, follow from (14) and (18), respectively. □

***Remark 1*** Theorem 3 shows that the two equivalent forms do not contain Kronecker product any more, and the sizes of the matrices in (11), (12) and (13) are $n \times m(n+1)$, $n \times n$ and $n \times (2m+n)$, respectively. When $m$ and $n$ are comparable and large, if

we compute the normwise condition number of the STLS problem with its explicit expressions, which is always preferred by the practitioners from applied disciplines, we note that the computation of (12) and (13) requires much less storage space compared with (11). However, as pointed out in [1,11], computing the matrix cross product may lead to large rounding errors. The explicit formulation of (12) and (13) needs to calculate $A^TA$ and $M^{-1}$, which is not desired. But, it should be claimed that $M^{-1}$ is often an intermediate result when the STLS problem is solved with its normal equation (6). For example, in the RQIPCG method, finding $M^{-1}$ can be transformed into solving triangular linear systems which can be efficiently computed and preserves better numerical stability [11, Ch. 8].

The TLS problem is a special case of the STLS problem, with Theorems 2 and 3 we get something new on the normwise condition number of the TLS problem.

**Corollary 1** *When $\lambda = 1$, the STLS problem degenerates into the TLS problem. From Theorems 2 and 3, the absolute condition number of the TLS problem has the following equivalent expressions*

$$\kappa_{TLSF}(A,b) = \left\| M^{-1}\left(\left(\frac{2}{\|r\|_2^2}A^Trr^T - A^T\right)([x_T^T, -1]\otimes I_m) - [I_n\otimes r^T, 0_{n\times m}]\right)\right\|_2,$$

$$\kappa_{TLSF1}(A,b) = \left\| M^{-1}\left((1+\|x_T\|_2^2)A^TA - A^Trx_T^T - x_Tr^TA + \|r\|_2^2 I_n\right)M^{-1}\right\|_2^{\frac{1}{2}}, \tag{19}$$

*and*

$$\kappa_{TLSF2}(A,b) = \left\| M^{-1}\left[A^T, \|x_T\|_2A^T\left(I_m - \tfrac{1}{\|r\|_2^2}rr^T\right), \|r\|_2\left(I_n - \tfrac{1}{\|r\|_2^2}A^Trx_T^T\right)\right]\right\|_2, \tag{20}$$

*where $M = A^TA - \sigma_{n+1}^2 I_n$, $\sigma_{n+1}$ is the smallest singular value of $[A,b]$, and $r = Ax_T - b$.*

***Remark 2*** We note that $\kappa_{TLSF1}(A,b)$ was an intermediate result in the proof of Theorem 1 in [1, Equation 3.8], and $\kappa_{TLSF}(A,b)$ was given by Jia and Li [14, Theorem 2]. Based on the normal equation $Mx_T = A^Tb$ and its variants, Baboulin and Gratton [1] showed that

$$\kappa_{TLSF1}(A,b) = \left\|(1+\|x_T\|_2^2)M^{-1}\left(A^TA + \sigma_{n+1}\left(I_n - \tfrac{2}{1+\|x_T\|_2^2}x_Tx_T^T\right)\right)M^{-1}\right\|_2^{\frac{1}{2}}. \tag{21}$$

To avoid computing the matrix cross product in (21), a power method [10, Ch. 7] based iterative procedure was proposed to compute the normwise condition number. Baboulin and Gratton [1] also suggested that when the TLS problem is solved by the SVD method, the computation of (21) can be further simplified. But their simplified expression needs the SVDs of both $A$ and $[A,b]$, which may be expensive. Jia and Li [14] further showed that only the SVD of $[A,b]$ will be enough. Based on the SVDs of $A$ and/or $[A,b]$, some computable upper and lower bounds of the condition number were also presented in [1,14]. In addition, it can be easily checked that

$$A^T A + \sigma_{n+1}\left(I_n - \frac{2}{1+\|x_T\|_2^2} x_T x_T^T\right)$$

is positive definite. Xie et al. [27, Remark 2] suggested to use Cholesky decomposition to further simplify the expression of normwise condition number, but no explicit expression was given there. According to Remark 1, our new compact form $\kappa_{TLSF2}(A, b)$ requires less storage space, and does not need to calculate Cholesky decomposition and the SVD. So we may say that the $\kappa_{TLSF2}(A, b)$ is a new and more efficient result on the explicit computation of the normwise condition number of the TLS problem.

As in [16,28], when $\lambda \to 0$, we get $\sigma_{n+1} \to 0$. Therefore, $M^{-1} = (A^T A - \sigma_{n+1} I_n)^{-1} \to (A^T A)^{-1}$ and $x_S$ converges to $x_O$. By the equality $A^T r = 0$ with $r = Ax_O - b$, from Theorems 2 and 3 we get the following three equivalent expressions of the normwise condition number for the OLS problem

$$\kappa_{OLSF}(A, b) = \left\|(A^T A)^{-1}\left(-A^T\left(\left[x_O^T, -1\right] \otimes I_m\right) - \left[I_n \otimes r^T, 0_{n\times m}\right]\right)\right\|_2,$$

$$\kappa_{OLSF1}(A, \lambda b) = \left\|(A^T A)^{-1}\left((1+\|x_O\|_2^2)A^T A + \|r\|_2^2 I_n\right)(A^T A)^{-1}\right\|_2^{\frac{1}{2}}, \quad (22)$$

and

$$\kappa_{OLSF2}(A, \lambda b) = \left\|(A^T A)^{-1}\left[A^T, \|x_O\|_2 A^T, \|r\|_2 I_n\right]\right\|_2. \quad (23)$$

With a little algebra, we can check that $\kappa_{OLSF}(A, b)$ can be rewritten as follows

$$\kappa_{OLSF}(A, b) = \left\|\left[-(x_O^T \otimes A^\dagger) - (A^T A)^{-1} \otimes r^T, A^\dagger\right]\right\|_2, \quad (24)$$

where $A^\dagger = (A^T A)^{-1} A^T$ is the Moore-Penrose inverse of matrix $A$ (see [2,26]). It should be noted that (24), (22) and (23) have been given by Li and Wang [18] in investigating the condition numbers for the indefinite least squares problem.

## 3 Numerical experiment

The computation and/or estimation of the condition numbers for the TLS problem has been extensively studied [1,6,14,27], which can be easily adapted to compute the normwise condition number of the STLS problem. From computational aspect, direct calculation of the condition number is not desired due to the heavy computation burden. When we solve the STLS problem, some intermediate results will be produced. Using these intermediate results to compute the condition number will largely reduce the computational burden. For example, in the implementation of RQIPCG method for solving the TLS problem, the approximate singular value of $\sigma_{n+1}$ and the Cholesky factor $R$ of $A^T A$ are available. So with these intermediate results the formulation of the condition number becomes much easier. Diao and Sun [6] proposed a power

method based on the RQIPCG procedure to estimate the upper bounds of the mixed and componentwise condition numbers of the TLS problem, which can also be modified to estimate the normwise condition number of the STLS problem. Here, we will not focus on devising algorithms to compute or estimate the normwise condition numbers. But, for completeness and to show the condition number estimation should be solver based, we present a RQIPCG procedure based power method to compute the normwise condition number of the STLS problem in the Appendix part.

As we have shown, our new compact forms require less storage space and are easy to calculate. This should be convenient for the practitioners to directly compute the normwise condition number of the STLS problem via software, like Matlab. Here, we only focus on the "naive" method to compute the normwise condition number of the STLS problem, which means that we first formulate the explicit expression of the matrix and then compute its spectral norm as the value of condition number. We use two built-in commands `norm(·,2)` and `normest(·,tol)` in Matlab R2010b. All the computations are performed on a PC with Intel i5-6600M CPU 3.30 GHz and 4.00 GB RAM.

***Example 1*** Investigating the influence of different forms on the computation of the normwise condition number for the STLS problem is our main purpose. Similar to [1], we construct the following random STLS problem. Let $[A,\ \lambda b]$ be defined by

$$[A,\ \lambda b] = Y \begin{bmatrix} D \\ 0 \end{bmatrix} Z^T \in \mathbb{R}^{m\times(n+1)}, \quad Y = I_m - 2yy^T, \quad Z = I_{n+1} - 2zz^T,$$

where $y \in \mathbb{R}^m, z \in \mathbb{R}^{n+1}$ are random unit vectors, and $D = \mathrm{diag}(n, n-1, \ldots, 1, 1-e_p)$ for given parameter $e_p$. Due to the interlacing property [3, p. 178], we get

$$\hat{\sigma}_n - \sigma_{n+1} \le \sigma_n - \sigma_{n+1} = e_p.$$

Thus $e_p$ gives a measure of the distance of the problem to nongenericity, and the solution $x_S$ is given by (5). By varying $\lambda$, $e_p$ and the size of the matrix, we report the CPU time in seconds of computing the condition number of the STLS problem with its different forms.

We repeat the computation 200 times for one group of settings. Since these two commands give same values of the condition numbers, we only report the mean values of CPU time in Table 1. From Table 1, we can see that when $m = 500, n = 300$, computing (11) becomes very time consuming due to the large size of the matrix, but (13) still works well. When we increase $(m, n)$ to (1000, 700), the computation of $\kappa_F(A, \lambda b)$ with `norm(·,2)` breaks down due to the lack of memory. The numerical results also show that the `normest(·,tol)` can largely reduce the CPU time in computing the spectral norm of large and sparse matrix. Of course, the matrix $K$ in (11) is sparse due to the Kronecker product. Moreover, we can also find that the CPU time for computing (13) is always the smallest in our numerical experiment. Therefore, we may say that the new compact form can greatly improve the efficiency of computing the exact value of the normwise condition number for the STLS problem with its explicit expression, and should be of much interest to the practical applications especially from applied disciplines.

**Table 1** Average CPU time in seconds of two "naive" methods

| Method | $\lambda$ | $e_p$ | $m = 100, n = 70$ $\kappa_F(A, \lambda b)\|\kappa_{F2}(A, \lambda b)$ | $m = 200, n = 150$ $\kappa_F(A, \lambda b)\|\kappa_{F2}(A, \lambda b)$ | $m = 500, n = 300$ $\kappa_F(A, \lambda b)\|\kappa_{F2}(A, \lambda b)$ |
|---|---|---|---|---|---|
| `norm(·, 2)` | 0.05 | 0.1 | 0.0424\|0.0022 | 0.6898\|0.0072 | 9.0942\|0.0380 |
| | | 0.001 | 0.0414\|0.0032 | 0.9154\|0.0156 | 9.3054\|0.0495 |
| | 5 | 0.1 | 0.0378\|0.0019 | 0.7182\|0.0075 | 9.6923\|0.0531 |
| | | 0.001 | 0.0365\|0.0019 | 0.6982\|0.0071 | 9.4639\|0.0520 |
| `normest(·, 10^-4)` | 0.05 | 0.1 | 0.0248\|0.0001 | 0.2263\|0.0021 | 2.4444\|0.0189 |
| | | 0.001 | 0.0260\|0.0001 | 0.2768\|0.0038 | 2.4759\|0.0190 |
| | 5 | 0.1 | 0.0239\|0.0001 | 0.2362\|0.0055 | 2.6477\|0.0227 |
| | | 0.001 | 0.0226\|0.0001 | 0.2269\|0.0020 | 2.4926\|0.0214 |

***Example 2*** From the definition of condition number (10), the relative forward error is bounded by the relative condition number multiplied by the relative backward error. We consider the following example adopted from [15] and arising from the application in signal restoration. Let $\alpha = 1.25$, $n = 500$ and $\omega = 80$. The convolution matrix $\bar{A}$ is an $n \times (n - 2\omega)$ Toeplitz matrix, and its first column is given by

$$a_{i,1} = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left[\frac{-(\omega - i + 1)^2}{2\alpha^2}\right], \quad i = 1, 2, \ldots, 2\omega + 1$$

and $t_{i,1} = 0$ otherwise. The elements in the first row are all zeros except $a_{11} = 1$. A Toeplitz matrix $A$ and a right-hand side vector $b$ are constructed as $A = \bar{A} + E$ and $b = \bar{g} + e$, where $\bar{g} = [1, \ldots, 1]^T$, $E$ is a random Toeplitz matrix with the same structure as $\bar{A}$ and $e$ is a random vector. The entries of $E$ and $e$ are generated from the standard normal distribution $\mathcal{N}(0, 1)$, and scaled such that

$$\|E\|_F = \gamma\|\bar{A}\|_F, \quad \|e\|_2 = \gamma\|\bar{g}\|_2, \quad \gamma = 10^{-3}.$$

The elements of the perturbations $\Delta A$ and $\Delta b$ to $A$ and $b$ are randomly generated from the open interval $(-1, 1)$, and scaled so that

$$\|\Delta A\|_F = \epsilon\|A\|_F, \quad \|\Delta b\|_2 = \epsilon\|b\|_2, \quad \epsilon = 10^{-8}.$$

Under the above settings, the perturbations to the coefficient matrices are not structured perturbation and can be interpreted as the additive white noise to the model [25].

Let $x + \Delta x$ be the solution to the perturbed STLS problem. The relative forward error is defined as $\|\Delta x\|_2/\|x\|_2$, and its approximate upper bounds with respect to different expressions of the normwise condition number are given by $\epsilon\kappa_{rF}(A, \lambda b)$ and $\epsilon\kappa_{rF2}(A, \lambda b)$. To compare the running time for computing the approximate upper forward error bound, we use $\text{CPUi}_{\kappa_*}$ with $i = 1, 2$ to denote the CPU time of computing the corresponding approximate upper forward error bounds with respect to `norm(·, 2)` and `normest(·, 10^-4)`, respectively.

**Table 2** Comparison of performances of different expressions in estimating the forward error

| | $\frac{\|\Delta x\|_2}{\|x\|_2}$ | $\epsilon\kappa_{rF}(A, \lambda b)$ | $\text{CPU1}_{\kappa_{rF}}$ | $\text{CPU2}_{\kappa_{rF}}$ | $\epsilon\kappa_{rF2}(A, \lambda b)$ | $\text{CPU1}_{\kappa_{rF2}}$ | $\text{CPU2}_{\kappa_{rF2}}$ |
|---|---|---|---|---|---|---|---|
| $\lambda = 10$ | 1.5247e−07 | 3.5895e−04 | 12.4836 | 3.1906 | 3.5895e−04 | 0.0445 | 0.0238 |
| $\lambda = 0.1$ | 2.3560e−07 | 3.6260e−05 | 12.5292 | 5.3745 | 3.6260e−05 | 0.0450 | 0.0301 |
| $\lambda = 0.0001$ | 2.0759e−08 | 7.4255e−07 | 12.6668 | 7.7938 | 7.4255e−07 | 0.0469 | 0.0333 |

The numerical results are reported in Table 2, from which we can find that although $\epsilon\kappa_{rF}(A, \lambda b)$ and $\epsilon\kappa_{rF2}(A, \lambda b)$ give same upper bounds, $\epsilon\kappa_{rF2}(A, \lambda b)$ requires much less CPU time compared with $\epsilon\kappa_{rF}(A, \lambda b)$. This shows the great efficiency of the new compact form of the normwise condition number in estimating the forward error of the STLS problem.

## 4 Concluding remark

In this paper, new and compact forms of the normwise condition number for the STLS problem are presented. Compared with the original expression, the new forms enjoy great computational efficiency in calculating the exact value of the normwise condition number, which may extend the applications of the condition number theory of the STLS problem to other areas. Since the estimation of the condition numbers for the TLS problem has been extensively studied and these methods can be directly applied to estimate the normwise condition number of the STLS problem, to avoid repetitive work we only outline a RQIPCG procedure based power method to estimate the normwise condition number of the STLS problem in the Appendix part. This is mainly to show that using the intermediate results produced in solving the STLS problem can largely reduce the computational burden in the condition number estimation.

In addition, Li and Jia [16] also considered the linear structured condition number for the STLS problem. Here, we need to point out that the structured normwise condition number of the STLS problem can also be transformed into some compact form with the same method given in [17, Remark 3.3]. But as Li and Wang [17] claimed, the compact form becomes very complicated due to the structure matrices. Thus, in this paper we will not consider the simplification of the structured normwise condition number of the STLS problem, and more researches on the structured condition number theory should be referred to [12,22,23].

**Acknowledgements** The authors are grateful to the anonymous referees and the Editor for their detailed and helpful comments that led to a substantial improvement to the paper.

## Appendix

When the large and sparse STLS problem is solved by the RQIPCG method [4], the Cholesky factor $R$ of $A^T A$, the approximate singular value of $\sigma_{n+1}$ and the residual

vector $r$[1] will be produced. With these intermediate results, we give the following algorithm to compute the normwise condition number of the STLS problem.

**Algorithm 1** RQIPCG based Power method

**Input:** $A$, termination criterion $\epsilon$, and intermediate results $-r$, Cholesky factor $R$, approximate singular value $\bar{\sigma}_{n+1}$.
**Output:** The normwise condition number $\kappa_F(A, \lambda b)$.

1. Given the initial vector $y_0$, and set $v_0 = 0$. $y_0$ is chosen as in [10, p. 366].
2. **for i** $= \mathbf{1}, \mathbf{2}, \cdots,$ **i**$_{\max}$ **do**

   (a) With $R$ and $\bar{\sigma}_{n+1}$, if we set $(A^TA - \bar{\sigma}_{n+1}^2 I_n)^{-1} y_{i-1} = z$, then we get $z$ by solving the following equivalent linear system
   $$\left[R,\ -\bar{\sigma}_{n+1} I_n\right] \begin{bmatrix} R^T \\ \bar{\sigma}_{n+1} I_n \end{bmatrix} z = y.$$
   The triangular linear system is much more easier to solve [11, Ch. 8]. Then
   $$K^T y = \mathrm{vec}\left(\left[wx^T - rz^T,\ -w\right]\right)$$
   with $w = \left(\frac{2}{\|r\|_2^2} rr^T A - A\right) z$.

   (b) $\left[A_i,\ b_i\right] \leftarrow \left[wx^T - rz^T,\ -w\right]$

   (c) $v_i \leftarrow \left\|\left[A_i,\ b_i\right]\right\|_F$
   **if** $|v_i - v_{i-1}| < \epsilon$ **then**
   $v = v_i$
   break
   **end if**

   (d) $\left[A_i,\ b_i\right] \leftarrow \frac{1}{v}\left[A_i,\ b_i\right]$

   (e) $y_i \leftarrow M^{-1}\left(\left(\frac{2}{\|r\|_2^2} A^T rr^T - A^T\right)(A_i x - b_i) - A_i^T r\right)$

   **end for**
3. The normwise condition number $\kappa_F(A, \lambda b)$ is given by
   $$\kappa_F(A, \lambda b) = \sqrt{v}.$$

Algorithm 1 computes the maximum eigenvalue of $KK^T$, so we use its square root as the condition number. Similar treatment of the normwise condition number of the TLS problem has been proposed in [1, Algorithm 1]. The main difference is that the power method in [1] gave direct manipulation to the original matrix and did not use the intermediate results in solving the TLS problem.

## References

1. Baboulin, M., Gratton, S.: A contribution to the conditioning of the total least-squares problem. SIAM J. Matrix Anal. Appl. **32**(3), 685–699 (2011)

[1] In [4], the residual vector is defined as $r = b - Ax_S$ which is different from the one used in our paper $r = Ax_S - b$, and $r$ appears in the termination criterion of RQI.

2. Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications, 2nd edn. Springer, New York (2003)
3. Björck, A.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
4. Björck, Å., Heggernes, P., Matstoms, P.: Methods for large scale total least squares problems. SIAM J. Matrix Anal. Appl. **22**(2), 413–429 (2000)
5. Bürgisser, P., Cucker, F.: Condition: The Geometry of Numerical Algorithms. Grundlehren der mathematischen Wissenschaften, vol. 349. Springer, Heidelberg (2013)
6. Diao, H.A., Sun, Y.: Mixed and componentwise condition numbers for a linear function of the solution of the total least squares problem. Linear Algebra Appl. **544**, 1–29 (2018)
7. Diao, H.A., Wei, Y., Xie, P.: Small sample statistical condition estimation for the total least squares problem. Numer. Algorithms **75**(2), 1–21 (2017)
8. Geurts, A.J.: A contribution to the theory of condition. Numer. Math. **39**(1), 85–96 (1982)
9. Golub, G.H., Van Loan, C.F.: An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**(6), 883–893 (1980)
10. Golub, G.H., Van Loan, C.F.: Matrix Computation, 4th edn. Johns Hopkins University Press, Baltimore (2013)
11. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)
12. Higham, D.J., Higham, N.J.: Backward error and condition of structured linear systems. SIAM J. Matrix Anal. Appl. **13**(1), 162–175 (1992)
13. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, New York (1991)
14. Jia, Z., Li, B.: On the condition number of the total least squares problem. Numer. Math. **125**(1), 61–87 (2013)
15. Kamm, J., Nagy, J.: A total least squares method for Toeplitz systems of equations. BIT **38**, 560–582 (1998)
16. Li, B., Jia, Z.: Some results on condition numbers of the scaled total least squares problem. Linear Algebra Appl. **435**(3), 674–686 (2011)
17. Li, H., Wang, S.: Partial condition number for the equality constrained linear least squares problem. Calcolo **54**(4), 1121–1146 (2017)
18. Li, H., Wang, S.: On the partial condition numbers for the indefinite least squares problem. Appl. Numer. Math. **123**, 200–220 (2018)
19. Paige, C.C., Strakoš, Z.: Scaled total least squares fundamentals. Numer. Math. **91**(1), 117–146 (2002)
20. 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 Modelling, pp. 11–20. SIAM, Philadelphia (1997)
21. Rice, J.R.: A theory of condition. SIAM J. Numer. Anal. **3**(2), 287–310 (1966)
22. Rump, S.M.: Structured perturbations part I: Normwise distances. SIAM J. Matrix Anal. Appl. **25**(1), 1–30 (2003)
23. Rump, S.M.: Structured perturbations part II: componentwise distances. SIAM J. Matrix Anal. Appl. **25**(1), 31–56 (2003)
24. Sun, J.G.: A note on simple non-zero singular values. J. Comput. Math. **6**, 258–266 (1988)
25. Tuzlukov, V.: Signal Processing Noise. CRC Press, Boca Raton (2002)
26. Wang, G., Wei, Y., Qiao, S.: Generalized Inverses: Theory and Computations. Science Press, Beijing (2004)
27. Xie, P., Wei, Y., Xiang, H.: Perturbation analysis and randomized algorithms for large-scale total least squares problems. arXiv preprint arXiv:1401.6832 (2014)
28. Zhou, L., Lin, L., Wei, Y., Qiao, S.: Perturbation analysis and condition numbers of scaled total least squares problems. Numer. Algorithms **51**(3), 381–399 (2009)