Statistical Condition Estimates and Randomized Algorithms for Large-Scale Total Least Squares Problems

Pengpeng Xie\textsuperscript{a} \* Yimin Wei\textsuperscript{a,b}† Hua Xiang\textsuperscript{c}‡

\textsuperscript{a} School of Mathematical Sciences, Fudan University, Shanghai, 200433, P.R. China
\textsuperscript{b} and Key Laboratory of Mathematics for Nonlinear Sciences
\textsuperscript{c} School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, P.R. China

January 28, 2014

Abstract

In this paper, we present new perturbation analysis and randomized algorithms for the total least squares (Tls) problem. Our new perturbation results are sharper than the earlier ones. We prove that the condition numbers obtained by Zhou et al. [Numer. Algorithm, 51 (2009), pp. 381-399], Barboulin and Gratton [SIAM J. Matrix Anal. Appl., 32 (2011), pp. 685-699], Li and Jia [Linear Algebra Appl., 435 (2011), pp. 674-686] are mathematically equivalent. Statistical condition estimates (Sce) are applied to the Tls problem. Motivated by the recently popular probabilistic algorithms for low-rank approximations, we develop randomized algorithms for the Tls and the truncated total least squares (Ttls) solutions of large-scale discrete ill-posed problems, which can greatly reduce the computational time and still keep good accuracy. Its efficiency is well demonstrated by numerical examples.

Keywords: Condition number; Singular value decomposition; Total least squares; Truncated total least squares; Statistical condition estimation; Randomized algorithms.

AMS Classification: 15A09, 65F35

\*E-mail: 09110180005@fudan.edu.cn. P. Xie is supported by 973 Program Project (No. 2010CB327900).
\†E-mail: ymwei@fudan.edu.cn. Y. Wei is supported by the National Natural Science Foundation of China under grant 11271084.
\‡Corresponding author (H. Xiang). E-mail: hxiang@whu.edu.cn. H. Xiang is supported by the National Natural Science Foundation of China under grants 10901125 and 91130022. Partial work was completed when he visited Fudan University and The Chinese University of Hong Kong in 2014.
1 Introduction

Given an overdetermined set of \( m \) linear equations \( Ax \approx b \) in \( n \) unknowns \( x \), the total least squares (Tls) problem can be formulated as [26]

\[
\min \| [E \ f] \|_F \quad \text{subject to} \quad b + f \in \mathcal{R}(A + E),
\]

(1.1)

where \( \| \cdot \|_F \) denotes the Frobenius matrix norm and \( \mathcal{R}(\cdot) \) denotes the range space. When the sampling or modeling or measurement errors also affect the coefficient matrix \( A \), the Tls method is more realistic, while the underlying assumption in the least squares (Ls) problem is that errors only occur in the right-hand side vector \( b \).

The term “total least squares” was coined in [7]. It has been also known as errors-in-variables model, orthogonal regression, or measurement errors in the statistical literature, and blind deconvolution in image deblurring. In the monograph based on the doctoral thesis of Sabine Van Huffel, the authors show the readers how to use Tls for solving a variety of problems, especially those arising in signal processing, medical imaging, and geophysics, etc. The applications and theory associated with the Tls are still being studied, for example [12, 18]. In recent years, perturbation analysis for the Tls problem has been studied extensively in the numerical linear algebra (see e.g. [2, 4, 5, 9, 13, 14, 22, 23, 24, 28, 27, 30, 31]).

It is well known that the condition number indicates the sensitivity of the problem itself, and that an approximate bound for the forward error can be given by the multiplication of the condition number and the backward error. For the perturbation in the solution of the scaled total least squares problem, Zhou et al. [31] presented a first order estimation. But as pointed out by the authors, it is impractical to compute since the condition number formula is a Kronecker product-based one. Baboulin and Gratton [2] derived a computable expression for the condition number. At almost the same time, Li and Jia [14] made a first order perturbation analysis. Recently, Jia and Li [15] proposed a formula which only used the singular values and the right singular vectors of \( [A, b] \), and presented the lower and upper bounds for the condition number. In this paper, we will present a relative perturbation bound without considering the condition number only. We first give a new perturbation result in this paper. And its significant improvements will be demonstrated by numerical examples. As a by-product, we also show that these three condition numbers in [2, 14, 31] mentioned above are mathematically equivalent. To estimate the sensitivity and conditioning of the total least squares problem, we apply the statistical condition estimation (Sce), which is a statistically based method for estimating the condition of general matrix functions firstly proposed by Kenney and Laub in [16], and then applied to the linear least squares problem [17].

For the numerical solution of the Tls problem, a simple and elegant solver based on the Svd of the augmented matrix \( [A \ b] \) can be used. When \( A \) is large, a complete Svd will be very costly. One improvement is to compute a partial Svd based on Lanczos bi-diagonalization [6]. But the partial Svd is still prohibitive for large-scale sparse or structured matrices, since the initial reduction of \( [A \ b] \) to bi-diagonal form will destroy the sparsity or structure of the matrix. For the Tls problem with very ill-conditioned coefficient matrix whose singular values decay gradually, the task is even more challenging. Without regularization, the ordinary least squares or total least squares solvers yield physically meaningless solutions. For such discrete ill-posed problems, there already exist several regularization strategies of the Tls solution. For example, the solution can be stabilized by truncating small singular
values of \([A, b]\) via an iterative algorithm based on Lanczos bi-diagonalization [6]. Tikhonov regularization strategy is used in [35, 34, 33, 18, 36], where a Cholesky decomposition is computed in each step in [35], and the linear systems are projected onto Krylov subspace of much smaller dimensions to reduce the problem size in [33]. Regularization by an additional quadratic constraint is another choice [37, 38, 39, 32]. It is the regularized Tls based on quadratic eigenvalue problems (QEP): adding a quadratic constraint to the Tls, and then iteratively solving the QEP. For the large-scale discrete ill-conditioned problem, a complete Svd is prohibitive, and the choice of regularization parameter is also time consuming. The classical Svd of a matrix can be well approximated by the randomized Svd [10], and the regularization parameter can also be located by randomized algorithms [29]. Such randomized algorithms can greatly reduce the computational time, and still keep good accuracy with very high probability. Motivated by these randomized matrix algorithms, we present two randomized algorithms for the solution of total least squares (Tls). One is designed for the well-conditioned cases; the other one is for the ill-conditioned cases.

Throughout this paper, \(R^{m \times n}\) denotes the set of \(m \times n\) matrices with real entries. \(I_n\) stands for the identity matrix with order \(n\) and \(e_j\) is the \(j\)th canonical vector. Single vertical bars around a matrix or vector indicate the componentwise absolute value of a matrix or vector. For a matrix \(A \in R^{m \times n}\), \(A^T\) is the transpose of \(A\); \(\|A\|_2\) and \(\|A\|_F\) denote the spectral norm and the Frobenius norm of \(A\), respectively. For any matrix \(A = [a_1, a_2, \ldots, a_n] = (a_{ij}) \in R^{m \times n}\) and \(B = (b_{ij}) \in R^{p \times q}\), the Kronecker product \(A \otimes B\) is defined as \(A \otimes B = (a_{ij}B) \in R^{mp \times nq}\). We define \(\text{vec}(A) = [a_1^T, a_2^T, \ldots, a_n^T]^T \in R^{mn}\) and the ‘unvec’ operator undoes the operation. For a vector \(a\), \(\text{diag}(a)\) is a diagonal matrix whose diagonals are given as components of \(a\). The normal distribution with mean \(\mu\) and variance \(\sigma^2\) is written as \(N(\mu, \sigma^2)\) and the uniform continuous distribution between \(a\) and \(b\) is abbreviated \(U(a, b)\). We denote the Fréchet derivative of a matrix function \(g : R^n \to R^m\) with respect to the variable \(X \in R^p\) by \(D_g(X) = (\partial g_i/\partial x_j) \in R^{m \times p}\). The Fréchet derivative of \(g\) with respect to \(X\) evaluated in the direction \(Y \in R^p\) is denoted \(D_g(X; Y) = D_g(X)Y\). When \(X\) and \(Y\) are matrices, we replace them by their Kronecker vectors in the derivative.

## 2 Preliminaries

Let \(A \in R^{m \times n}\) and \(b \in R^n\) with \(m > n\). Let \([A, b]\) and \(A\) have singular value decompositions, respectively

\[
U^T[A, b]V = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{n+1}) = \Sigma,
\]

\[
\bar{U}^TAV = \text{diag}(\bar{\sigma}_1, \bar{\sigma}_2, \ldots, \bar{\sigma}_n),
\]

where \(U, V\), and \(\Sigma\) are partitioned as follows:

\[
U = [U_1, u_{n+1}]_{m \times (n+1)}, \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}_{(n+1) \times (n+1)}, \quad \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \sigma_{n+1} \end{bmatrix}.
\]

Throughout this paper, we assume the genericity condition:

\[
\bar{\sigma}_n > \sigma_{n+1}, \tag{2.1}
\]
to ensure the existence and uniqueness of the TLS solution \(x\) (see [7]). From best rank-1 approximation [8] of matrix \([A, b]\), we know that

\[
[E, f] = -U \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{n+1} \end{bmatrix} V^T
= -\sigma_{n+1} u_{n+1} \begin{bmatrix} v_{12}^T, v_{22} \end{bmatrix}
= -\sigma_{n+1} u_{n+1} v_{n+1}^T,
\]

where \(v_{n+1} = \begin{bmatrix} v_{12}^T, v_{22} \end{bmatrix}^T\). Therefore, \(x = -v_{12}/v_{22}\). From [26 Theorem 2.7], it follows that the solution \(x\) can also be expressed as a function of \([A, b]\), i.e.,

\[
x = g ([A, b]) = (A^T A - \sigma_{n+1}^2 I)^{-1} A^T b, \tag{2.2}
\]

and it holds that

\[
\begin{bmatrix} x^T, -1 \end{bmatrix}^T = -\frac{1}{v_{22}} v_{n+1}.
\tag{2.3}
\]

### 3 New perturbation results

First, we give a lemma which will be very useful in our analysis.

**Lemma 3.1** Consider the total least squares problem (1.1) and assume that the genericity condition (2.1) holds. If \([A, b]\) is perturbed to \([A + \delta A, b + \delta b]\), then we have

\[
\sigma_{n+1} u_{n+1}^T [\delta A, \delta b] v_{n+1} = \frac{r^T [\delta b - (\delta A)x]}{1 + x^T x},
\]

where \(r = b - Ax\).

**Proof.** From (2.3) and the singular value decomposition of \([A, b]\), we know that

\[
r = b - Ax = -[A, b] \begin{bmatrix} x \\ -1 \end{bmatrix} = \frac{1}{v_{22}} [A, b] v_{n+1} = \frac{1}{v_{22}} \sigma_{n+1} u_{n+1}.
\]

Therefore we have

\[
r^T [\delta b - (\delta A)x] = \frac{\sigma_{n+1} u_{n+1}^T [\delta b - (\delta A)x]}{1 + x^T x}
= -\sigma_{n+1} v_{22} u_{n+1}^T [\delta A, \delta b] \begin{bmatrix} x \\ -1 \end{bmatrix}
= \sigma_{n+1} u_{n+1}^T [\delta A, \delta b] v_{n+1},
\]

where we use \(v_{22}^2 = \frac{1}{1 + x^T x}\), which is a direct result of (2.3).

The following lemma [23] is also needed for deriving our new perturbation result.
Lemma 3.2. Let $\sigma_{\text{min}}$ be the smallest nonzero and simple singular value of a matrix $X$ with $u_{\text{min}}$ and $v_{\text{min}}$ being its corresponding left and right singular vectors, respectively. If $\delta X$ is sufficiently small, then the smallest nonzero singular value $\tilde{\sigma}_{\text{min}}$ of the perturbed matrix $\tilde{X} = X + \delta X$ is simple and

$$\tilde{\sigma}_{\text{min}} = \sigma_{\text{min}} + u_{\text{min}}^T(\delta X)v_{\text{min}} + O(||\delta X||_F^2).$$

In the following, we present our perturbation bound under the genericity condition \eqref{eq:2.1}.

Theorem 3.1. Consider the total least squares problem \eqref{eq:1.1} and assume that the genericity condition \eqref{eq:2.1} holds. If

$$||[\delta A, \delta b]||_2 < \frac{\tilde{\sigma}_n^2 - \sigma_{n+1}^2}{2\sigma_{n+1}},$$

then

$$\frac{||\delta x||_2}{||x||_2} \leq \frac{1}{1 - c} \left( \frac{||A||_2||r||_2}{||x||_2} + 2\sigma_{n+1}||A||_2 \right) \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||A||_2}{||A||_2} \left( \frac{||\delta A||_2}{||A||_2} \right) + \frac{1}{1 - c} \left( \frac{||b||_2}{||x||_2} \right) \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||b||_2}{||b||_2},$$

where

$$c = 2 \left( \frac{r^T(\delta b - (\delta A)x)}{1 + x^T x} \right) \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||b||_2}{||b||_2}.$$

Proof. From $A^TA - \sigma_{n+1}^2 x = A^T b$, perturbing $[A, b]$ yields

$$(A + \delta A)^T(A + \delta A) - \tilde{\sigma}_{n+1}^2 f_n(x + \delta x) = (A + \delta A)^T(b + \delta b),$$

where $\tilde{\sigma}_{n+1}$ is the smallest singular value of $[A + \delta A, b + \delta b]$. Subtracting two equations \eqref{eq:1.1} and \eqref{eq:3.2}, and omitting the higher order terms, we have

$$(A^TA - \sigma_{n+1}^2 f_n)\delta x = (\delta A)^T \delta r + A^T [\delta b - (\delta A)x] + (\tilde{\sigma}_{n+1}^2 - \sigma_{n+1}^2)(x + \delta x).$$

Therefore we have

$$\frac{||\delta x||_2}{||x||_2} \leq \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||\delta A||_2||r||_2}{||x||_2} + \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||A||_2}{||A||_2} \left( \frac{||\delta A||_2}{||A||_2} \right) + \left( \frac{||b||_2}{||x||_2} \right) \left( (A^TA - \sigma_{n+1}^2)^{-1} \right) \frac{||b||_2}{||b||_2} \left( 1 + \frac{||\delta x||_2}{||x||_2} \right).$$

Furthermore, combining Lemma \ref{lem:3.1} and Lemma \ref{lem:3.2} and ignoring higher order terms, we have

$$\tilde{\sigma}_{n+1}^2 - \sigma_{n+1}^2 = (\sigma_{n+1} - \sigma_{n+1}) (\sigma_{n+1} + \sigma_{n+1})$$

$$= \left( \frac{u_{n+1}^T[A, \delta b]v_{n+1} + O(||[A, \delta b]||_F^2)}{2\sigma_{n+1} + u_{n+1}^T[A, \delta b]v_{n+1} + O(||[A, \delta b]||_F^2)} \right) \frac{2\sigma_{n+1} + u_{n+1}^T[A, \delta b]v_{n+1} + O(||[A, \delta b]||_F^2}{2\sigma_{n+1} + u_{n+1}^T[A, \delta b]v_{n+1} + O(||[A, \delta b]||_F^2)}$$

$$\approx \frac{r^T(\delta b - (\delta A)x)}{1 + x^T x}. \quad \quad \text{(3.3)}$$
When the assumption holds, \( c \) is smaller than 1 and therefore we obtain
\[
\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{1}{1 - c} \left( \frac{\|A\|_2}{\|x\|_2} \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} \right\|_2 \right) + \frac{1}{\|A\|_2} \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} A^T \right\|_2 \frac{\|\delta A\|_2}{\|A\|_2} \left\| b \right\|_2
\]
\[
+ \frac{1}{1 - c} \left( \frac{\|b\|_2}{\|x\|_2} \right) \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} A^T \right\|_2 \|\delta b\|_2 + \frac{1}{1 - c} \|\sigma_{n+1}^2 - \sigma_{n+1}^2\|_2 \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} \right\|_2.
\]

Since \( \|\sigma_{n+1}^2 - \sigma_{n+1}^2\|_2 \leq 2\sigma_{n+1} ||\delta A, \delta b||_2 \leq 2\sigma_{n+1} (||\delta A||_2 + ||\delta b||_2) \) from (3.5), we get the final result. \( \square \)

**Corollary 3.1** For the total least squares problem \((1.1)\) with genericity condition \((2.1)\) holding, we have
\[
\frac{\|\delta x\|_2}{\|x\|_2} \leq \left( \frac{\|\delta A\|_2}{\|A\|_2} \right) \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} A^T \right\|_2 \frac{\|\delta b\|_2}{\|b\|_2} \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} \right\|_2 + \frac{\|A\|_2}{\|\delta A\|_2} \left\| (A^T A - \sigma_{n+1}^2 I)^{-1} \right\|_2 \frac{\|b\|_2}{\|x\|_2} + \frac{\|\delta b\|_2}{\|A\|_2} \left\| A^T + 2 \frac{x^T}{1 + x^T x} \right\|_2 \left\| A^T + 2 \frac{\delta b}{1 + x^T x} \right\|_2.
\]

**Proof.** Under the conditions \((1.1), (3.2)\) and \((3.3)\), dropping the second and higher order terms, we have
\[
\delta x \approx (A^T A - \sigma_{n+1}^2 I)^{-1} \left[ A^T (\delta b) - A^T (\delta A) x + (\delta A)^T r \right]
\]
\[
+ 2 (A^T A - \sigma_{n+1}^2 I)^{-1} \frac{r^T (\delta b - (\delta A) x)}{1 + x^T x} (x + \delta x)
\]
\[
\approx (A^T A - \sigma_{n+1}^2 I)^{-1} \left[ A^T (\delta b) - A^T (\delta A) x + (\delta A)^T r \right]
\]
\[
+ 2 (A^T A - \sigma_{n+1}^2 I)^{-1} \frac{r^T (\delta b - (\delta A) x)}{1 + x^T x} x,
\]
and then
\[
\delta x \approx (A^T A - \sigma_{n+1}^2 I)^{-1} \left[ A^T + 2 \frac{x^T}{1 + x^T x} \right] \delta b
\]
\[
+ (A^T A - \sigma_{n+1}^2 I)^{-1} \left( (\delta A)^T r - \left[ A^T + 2 \frac{x^T}{1 + x^T x} \right] (\delta A) x \right).
\]

Since \( r = b - A x = \frac{\sigma_{n+1}^2}{v_{22}} u_{n+1} \), using the MATLAB notation we have
\[
\frac{x^T}{1 + x^T x} = -v_{n+1}(1 : n)\sigma_{n+1} u_{n+1}^T = -[V_{11}, v_{12}] \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{n+1} \end{bmatrix} U^T.
\]

Moreover, it follows that the SVD of \([A, b]\) and
\[
A = [A, b] \begin{bmatrix} I_n \\ 0_{1 \times n} \end{bmatrix} = U \Sigma \begin{bmatrix} V_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}^T \begin{bmatrix} I_n \\ 0_{1 \times n} \end{bmatrix} = U \Sigma [V_{11}, v_{12}]^T, \quad \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \sigma_{n+1} \end{bmatrix}.
\]
Therefore, we obtain that

\[ A^T + 2 \frac{xx^T}{1 + x^T x} = [V_{11}, \; v_{12}] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & -\sigma_{n+1} \end{bmatrix} U^T. \]  

(3.5)

Furthermore,

\[
\left[A^T + 2 \frac{xx^T}{1 + x^T x}\right] \cdot \left[A^T + 2 \frac{xx^T}{1 + x^T x}\right]^T = [V_{11}, \; v_{12}] \Sigma^2 [V_{11}, \; v_{12}]^T = A^T A,
\]

and

\[
\left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} \left[A^T + 2 \frac{xx^T}{1 + x^T x}\right] \right\|_2 = \left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2.
\]

Then we have the following estimation by (3.4),

\[
\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2 \|b\|_2}{\|x\|_2} \|\delta b\|_2
\]

\[
+ \frac{\left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} \left(I_n \otimes x^T - x^T \otimes \left[A^T + 2 \frac{xx^T}{1 + x^T x}\right] \right) \right\|_2 \|\delta A\|_F}{\|A\|_F} \|\delta b\|_2.
\]

Then we can draw the conclusion.

**Remark 1** The assumption 

\[
\|\|\delta A, \delta b\|\|_2 \leq \frac{\sigma_n^2 - \sigma_{n+1}^2}{2\sigma_{n+1}}
\]

can be easily satisfied, when the perturbations are sufficiently small. In practice, to estimate it, we can omit the factor $1/(1 - c)$ and just compute two terms

\[
\left( \frac{\|A\|_2 \|r\|_2}{\|x\|_2} + 2\sigma_{n+1} \|A\|_2 \right) \left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2
\]

and

\[
\frac{\|b\|_2}{\|x\|_2} \left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2 + 2\sigma_{n+1} \|b\|_2 \left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2,
\]

since $[\delta A, \delta b]$ is unknown and $1/(1 - c)$ is close to 1, if the perturbations are small enough. Moreover, it is easy to check that

\[
\left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2 = \frac{\sigma_n}{\sigma_n^2 - \sigma_{n+1}^2}.
\]

The term $\|\|b\|\|_2 \left\| \left( A^T A - \sigma_{n+1}^2 I_n \right)^{-1} A^T \right\|_2$ is reminiscent of the “effective condition number” of the total least squares problem. The effective condition number is defined as [19, 20, 21]

\[
\text{Cond}_{\text{eff}} = \frac{\|b\|_2}{\|x\|_2} \frac{\|A\|_2 \|r\|_2}{\|x\|_2}
\]

for the linear system $Ax = b$ with $\sigma_r$ being the smallest positive singular value of $A$, and $A^\dagger$ is the Moore-Penrose inverse of $A$ [8].
Denote
\[ k_{\text{TLS}}^Z = \frac{\|M + N\|_2\|[A, b]\|_F}{\|x\|_2}, \]
\[ M = \begin{bmatrix} K \otimes b^T - x^T \otimes (KA^T) - K \otimes (Ax)^T & KA^T \end{bmatrix}, \]
\[ N = 2\sigma_{n+1}y \left( v_{n+1}^T \otimes u_{n+1}^T \right) \]
with \( K = (A^T A - \sigma_{n+1}^2 I)^{-1} \) and \( y = Kx \). Omitting the complicated higher order term \( R(\delta A, \delta b) \) in Eqn.(3.5), then the upper bound derived in [31] becomes
\[ k_{\text{TLS}}^Z \frac{\|[\delta A, \delta b]\|_F}{\|[A, b]\|_F}. \]

The upper bound obtained from [2] is expressed by
\[ K^{(\text{rel})}(A, b) \frac{\|[\delta A, \delta b]\|_F}{\|[A, b]\|_F}, \]
where \( K^{(\text{rel})}(A, b) \) is the relative condition number
\[ K^{(\text{rel})}(A, b) = \left( 1 + \|x\|^2 \right)^{\frac{1}{2}} \frac{\|D \begin{bmatrix} V^T, 0_{n \times 1} \end{bmatrix} V [D, 0_{n \times 1}]^T \|_2 \|[A, b]\|_F}{\|x\|_2}. \]

Later, Li and Jia [14] established the following relative condition number
\[ K_{\text{LJ}} = \frac{\|K_1\|_2\|[A, b]\|_F}{\|x\|_2} \]
in which
\[ K_1 = (A^T A - \sigma_{n+1}^2 I_n)^{-1} \left( 2A^T \frac{r^T}{\|r\|_2} \frac{r^T}{\|r\|_2} G(x) - A^T G(x) + [I_n \otimes r^T, 0_{n \times m}] \right) \]
with \( G(x) = [x^T, -1] \otimes I_n. \)

**Remark 2** Note that \( K^{(\text{rel})}(A, b) \) has another closed formula [2]

\[ K^{(\text{rel})}(A, b) = \left( 1 + \|x\|^2 \right)^{\frac{1}{2}} \frac{\|A^T A - \sigma_{n+1}^2 I_n \|_2 \|[A, b]\|_F}{\|x\|_2}. \]

Since \( A^T A + \sigma_{n+1}^2 \left( I_n - \frac{2x^T}{1 + \|x\|^2} \right) = A^T A - \sigma_{n+1}^2 I_n + 2\sigma_{n+1}^2 \left( I_n - \frac{x^T}{1 + \|x\|^2} \right) \) is symmetric positive definite, We can define \( LL^T \) as its Cholesky factorization. Then we have
\[ K^{(\text{rel})}(A, b) = \left( 1 + \|x\|^2 \right)^{\frac{1}{2}} \frac{\|A^T A - \sigma_{n+1}^2 I_n \|_2 \|[A, b]\|_F}{\|x\|_2}, \]
which is another expression of \( K_{\text{LJ}} \) [15].
Moreover, using Lemma 3.1 and the proof of [21, Lemma 3.2], we can get the following equation

$$M + N = \left[ -x^T \otimes D_{\sigma_{n+1}^2} + \left( A^T A - \sigma_{n+1}^2 I \right)^{-1} \otimes r^T, \ D_{\sigma_{n+1}^2} \right],$$

where $D_{\sigma_{n+1}^2} = \left( A^T A - \sigma_{n+1}^2 I \right)^{-1} \left( A^T + 2 - \frac{\nu^T}{1+\nu^T} \right)$. Denote $P \in \mathbb{R}^{mn \times mn}$ the permutation matrix that represents the matrix transpose by vec$(B^T) = P \text{vec}(B)$. Note that $K^{(rel)}(A, b)$ can also be expressed by [21]

$$K^{(rel)}(A, b) = \frac{\|M_{g'}\|_2\|\{A, b\}\|_F}{\|x\|_2}$$

with

$$M_{g'} = \left[ -x^T \otimes D_{\sigma_{n+1}^2} + \left( r^T \otimes \left( A^T A - \sigma_{n+1}^2 I \right)^{-1} \right) P, \ D_{\sigma_{n+1}^2} \right].$$

Easily we find that $M + N = M_{g'}$, which means that $K^{(rel)}(A, b) = \kappa_{TLS}^Z$.

Therefore, we see that the condition numbers derived respectively in [21, 14, 31] are mathematically equivalent. But as pointed by the authors, the normwise condition number proposed in [31] is impractical to compute.

4 Sce for the total least squares

Statistical condition estimation (Sce) is a statistically based method for estimating the condition of general matrix functions by Kenney and Laub in [16]. It is then applied to the linear least squares problem [17]. In what follows, we consider the Sce for the total least squares problem.

4.1 Review of statistical condition estimation

Suppose $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is at least twice continuously differentiable. Denote the gradient of $f$ at $x \in \mathbb{R}^p$ by $\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_p} \right)^T$. Then using the first order Taylor expansion of $f$ at $x$ along $d$, we have

$$f(x + \delta d) = f(x) + \delta (\nabla f(x))^T d + O(\delta^2),$$

where $\delta \in \mathbb{R}$ is small and $d \in \mathbb{R}^p$ has unit 2-norm. We can see that the norm of the gradient can measure the local sensitivity of $f$ appropriately. If $d$ is selected uniformly and randomly from the unit sphere $S_{p-1}$ in $\mathbb{R}^p$, which is denoted by $d \in U(S_{p-1})$. Then from [16], the expected value of the condition estimator $\nu = \frac{\|\nabla f(x)\|_2}{\omega_p}$ satisfies that

$$E(\nu) = \|\nabla f(x)\|_2,$$

where $\omega_p$ is the Wallis factor [16], and for $\gamma > 1$ we have

$$\text{Prob}\left( \frac{\|\nabla f(x)\|_2}{\gamma} \leq \nu \leq \gamma\|\nabla f(x)\|_2 \right) \geq 1 - \frac{2}{\pi\gamma} + O\left( \frac{1}{\gamma^2} \right).$$

Therefore, we can use the absolute value

$$\left| \frac{f(x + \delta d) - f(x)}{\delta \omega_p} \right|$$
as a first order condition estimator, which can estimate $\|\nabla f(x)\|_2$ with high probability for the function $f$ at $x$. In practice, the Wallis factor can be approximated accurately \cite{16} by

$$
\omega_p \approx \sqrt{\frac{2}{\pi(p - \frac{1}{2})}}.
$$

(4.1)

In situations where we need more reliability, we use more function evaluations to get different values $\nu^{(1)}, \nu^{(2)}, \ldots, \nu^{(m)}$ corresponding to independently randomly generated vectors $d^{(1)}, d^{(2)}, \ldots, d^{(m)}$ in $S_{p-1}$ and then take the average

$$
\nu(m) = \frac{\nu^{(1)} + \nu^{(2)} + \cdots + \nu^{(m)}}{m}.
$$

(Here we use superscripts in parentheses to distinguish between different vectors)

This is the so called “averaged small-sample statistical method” \cite{16} and it can be shown that for $\gamma > 1$,

$$
\text{Prob}\left(\frac{\|\nabla f(x)\|_2}{\gamma} \leq \nu(m) \leq \gamma\|\nabla f(x)\|_2\right) \geq 1 - \frac{1}{m!} \left(\frac{2m}{\pi \gamma}\right)^m + O\left(\frac{1}{\gamma^{m+1}}\right).
$$

Thus $\nu(m)$ is a $m$th-order condition estimator.

Compared with this method, the subspace statistical method can give sharper estimates. Firstly, select $k$ vectors from $S_{p-1}$ and find an orthonormal basis $d_1, d_2, \ldots, d_k$ for their span by using a Gram-Schmidt procedure or a QR decomposition \cite{8}. Thus the norm of the projection of $\nabla f(x)$ onto the span of $d_1, d_2, \ldots, d_k$ is $(|\langle \nabla f(x) \rangle^T d_1|^2 + |\langle \nabla f(x) \rangle^T d_2|^2 + \cdots + |\langle \nabla f(x) \rangle^T d_k|^2)^{1/2}$. From \cite{16}, we know that

$$
\mathbb{E} \left( \frac{\omega_k}{\omega_p} \sqrt{|\langle \nabla f(x) \rangle^T d_1|^2 + |\langle \nabla f(x) \rangle^T d_2|^2 + \cdots + |\langle \nabla f(x) \rangle^T d_k|^2} \right) = \|\nabla f(x)\|_2.
$$

Therefore, we can define the subspace condition estimator as

$$
\xi(k) = \frac{\omega_k}{\omega_p} \sqrt{|\langle \nabla f(x) \rangle^T d_1|^2 + |\langle \nabla f(x) \rangle^T d_2|^2 + \cdots + |\langle \nabla f(x) \rangle^T d_k|^2}.
$$

As shown in \cite{16}, these condition estimators give better results than the averaged statistical estimators and are analytically very tractable. From Theorem 3.3 in \cite{16}, we find that

$$
\text{Prob}\left(\frac{\|\nabla f(x)\|_2}{\gamma} \leq \xi(2) \leq \gamma\|\nabla f(x)\|_2\right) \approx 1 - \frac{\pi}{4\gamma^2},
$$

and the similar expressions for $\xi(3), \xi(4), \text{etc.}$ These estimates are generally very accurate for $\gamma \geq 10$.

The function $f$ may be scalar valued, but we can easily extend $\text{ScE}$ to vector and matrix-valued functions by viewing $f$ as a map from $\mathbb{R}^p$ to $\mathbb{R}^q$ (using the operations ‘vec’ and ‘unvec’ to convert between matrices and vectors).
4.2 The Fréchet derivative and total least squares

Perturbing $[A, b]$ to $[A + e\Omega, b + e\beta]$ in the equation $(A^TA - \sigma^2_{n+1}I_n)x = A^Tb$, where $e \in \mathbb{R}$ and $[\Omega, \beta]$ has Frobenius norm equal to one, we get the corresponding equation similar to (3.2)

$$\left[(A + e\Omega)^T(A + e\Omega) - \sigma^2_{n+1}I_n\right](x + ey) = (A + e\Omega)^T(b + e\beta).$$

Let $y = D_g([A, b]; [\Omega, \beta])$ be the Fréchet derivative of (2.2) with respect to $[A, b]$ evaluated in the direction $[\Omega, \beta]$. Then we know that

$$y = \lim_{\epsilon \to 0} \frac{\left[(A + e\Omega)^T(A + e\Omega) - \sigma^2_{n+1}I_n\right]^{-1}(A + e\Omega)^T(b + e\beta) - (A^TA - \sigma^2_{n+1}I_n)^{-1}A^Tb}{\epsilon}$$

$$= \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}A^T(\beta - \Omega x) + \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}\Omega^T \lim_{\epsilon \to 0} \frac{(\sigma^2_{n+1} - \sigma^2_{n+1})\left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}x}{\epsilon}$$

$$= \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}\left[A^T + 2\frac{\sigma^T}{1 + \sigma^T}x\right](\beta - \Omega x) + \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}\Omega^T r,$$

(4.2)

where we use (3.3) in the third equality.

**Remark 3** Baboulin and Gratton in [2] exploited the chain rule to derive the Fréchet derivative of $g$ at $[A, b]$ as follows

$$g'(A, b) : \mathbb{R}^{m \times n} \times \mathbb{R}^m \to \mathbb{R}^k$$

$$([\Delta A, \Delta b]) \mapsto \left[LT \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}A^T + 2\frac{\sigma^T}{1 + \sigma^T}x\right][\Delta b - (\Delta A)x]$$

$$+ LT \left(A^TA - \sigma^2_{n+1}I_n\right)^{-1}(\Delta A)^T r.$$

The conditioning of the TLS problem corresponds to the case where $L$ is the identity matrix. We can see that the result is the same as the one we derived by the definition of the Fréchet derivative.

With the SVD of $[A, b]$, and $A = U\Sigma[V_{11}, V_{12}]^T$, we have

$$A^TA - \sigma^2_{n+1}I_n = \begin{bmatrix} \sigma^2_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{n+1} \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \end{bmatrix}^T \begin{bmatrix} V_{11} & V_{12} \end{bmatrix}^T - \sigma^2_{n+1}I_n \begin{bmatrix} V_{11} & V_{12} \end{bmatrix}^T$$

$$= \begin{bmatrix} \sigma^2_1 - \sigma^2_{n+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{n+1} - \sigma^2_{n+1} \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \end{bmatrix}^T$$

$$= \begin{bmatrix} \sigma^2_1 - \sigma^2_{n+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{n+1} - \sigma^2_{n+1} \end{bmatrix} \begin{bmatrix} V_{11}^T, & V_{12}^T \end{bmatrix}$$
and again by (3.5)

\[
A^T + 2 \frac{xx^T}{1 + x^T x} = [V_{11}, v_{12}] \begin{pmatrix}
\sigma_1 \\
& \ddots \\
& & \sigma_n \\
& & & -\sigma_{n+1}
\end{pmatrix} U^T.
\]

So we can compute the Fréchet derivative.

Suppose now that \(\Omega\) and \(\beta\) are random variables whose entries are i.i.d. \(\mathbb{N}(0, 1)\), but \([\Omega, \beta]\) does not necessarily have Frobenius norm equal to one. We can rewrite (4.2) as

\[
y = \left(\frac{A^T A - \sigma_{n+1}^2}{\|\Omega, \beta\|_F}\right)^{-1} \left( A^T + 2 \frac{xx^T}{1 + x^T x} \right) (\beta - \Omega x) + \Omega^T r \right].
\]

(4.3)

4.3 Sce for the total least squares under the genericity condition

Denote \(\tilde{D} = \text{diag} \left( \left(\tilde{\sigma}_1^2 - \sigma_{n+1}^2\right)^{-1}, \ldots, \left(\tilde{\sigma}_n^2 - \sigma_{n+1}^2\right)^{-1} \right)\), \(D = \text{diag} \left( \sqrt{\sigma_1^2 + \sigma_{n+1}^2}, \ldots, \sqrt{\sigma_n^2 + \sigma_{n+1}^2} \right)\).

For the condition number of Tls solution, we use the formula derived in [2, Theorem 2]

\[
K(A, b) = \left( 1 + \|x\|_2^2 \right)^{\frac{1}{2}} \|D [\tilde{F} V^T, 0_{n \times 1}] V [D, 0_{n \times 1}]^T \|_2,
\]

(4.4)

which is the absolute condition number of \(x\). Denote \(K_j\) the condition number of the function \(z_j^T x\), where \(z_j\)'s are random orthogonal vectors selected uniformly and randomly from the unit sphere in \(n\) dimensions. From [2], we know that \(K_j\) can be computed by

\[
K_j = \left( 1 + \|x\|_2^2 \right)^{\frac{1}{2}} \|z_j^T \tilde{D} [\tilde{V}^T, 0_{n \times 1}] V [D, 0_{n \times 1}]^T \|_2.
\]

(4.5)

Taking the technique and notations adopted in [3], we see that

\[
\tilde{k}_{\text{Tls}} = \frac{\omega_q}{\omega_n} \sqrt{\sum_{j=1}^q K_j^2}
\]

is an estimate for \(K(A, b)\).

We use the results above to give the Sce-based method for estimating the condition of the solution to the total least squares problem under the genericity condition (2.1). Inputs of the method are the matrix \(A \in \mathbb{R}^{m \times n}\) and the vector \(b \in \mathbb{R}^m\), and the output is the statistical estimate for the absolute condition number, a scalar quantity that must be distinguished from the estimate given in [17] where it is a vector. In Algorithm Sce0 the integer \(q \geq 1\) refers to the number of Sce samples.
1. Generate $q$ vectors $z_1, z_2, \ldots, z_q \in \mathbb{R}^n$ with entries in the uniform continuous distribution $\mathcal{U}(0, 1)$. Orthonormalize the vectors using a QR factorization.

2. For $j = 1, 2, \ldots, q$, compute $\kappa_j$ by (4.5).

3. Approximate $\omega_q$ and $\omega_n$ using (4.1) and compute $\bar{\kappa}_{Tls} = \frac{\omega_q}{\omega_n} \sqrt{\sum_{j=1}^{q} \kappa_j^2}$.

### 4.4 Componentwise condition estimates

As shown in [17], we consider the SCE for the condition numbers of the components of $x$. The exact value of the condition number for the $i$th component of $x$ is computed by

$$\kappa_i^e = \left(1 + ||x||_2^2\right)^{\frac{1}{2}} \left|e_i^T \bar{V} \tilde{D} [\bar{V}^T, 0_{n \times 1}] V [D, 0_{n \times 1}]^T\right|_2, \quad (i = 1, 2, \ldots, n),$$

which is simply obtained by taking $L = e_i$ in the condition number for $L^T x$ in [2, Theorem 2]. The following Algorithm SCE1 computes a vector $\bar{K}_{abs}$ containing the statistical estimate for the $\kappa_i^e$'s. We should point out that for the componentwise case, the techniques taken in this algorithm are different from the ones taken in the subsection above. Note that $y_j$'s in Algorithm SCE1 are just the Fréchet derivative estimates for the components of the Tls solution. This algorithm is based on the original idea of SCE [16], which means that we estimate the Fréchet derivative.

To see the process more clearly, the readers need to notice that $y_j$'s are just the Fréchet derivative estimates for the components of the Tls solution. The manipulation here is somewhat similar to that in [40, Section 3.2]. After generating and normalizing the random elements in the first step, these random elements are overwritten by the componentwise product of $[A, b]$ and these elements. In the main step we need to calculate $y_j$ as

$$y_j = (A^T A - \sigma_{n+1}^2 I)^{-1} \left[ (A^T + 2 x r^T 1 + x^T x) (\beta_j - \Omega_j x) + \Omega_j^T r \right].$$

At last, the condition vector containing $\kappa_i^e$’s are computed by

$$\bar{K}_{abs} = \frac{\omega_q}{\omega_p} \sqrt{|y_1|^2 + |y_2|^2 + \cdots + |y_q|^2},$$

where the square root and power operation are performed componentwise.

### 5 Randomized algorithms for Tls problems

The randomized algorithms open the possibility of dealing with truly massive data sets, and have become more and more popular in the matrix approximation in the last decade [10]. Numerical experiments and detailed error analysis show that these random sampling techniques can beat the classical
Algorithm Sce1: Subspace condition estimate for relative perturbations

1. Generate matrices $[\Omega_1, \beta_1], [\Omega_2, \beta_2], \ldots, [\Omega_q, \beta_q]$ with entries in $\mathbb{N}(0, 1)$. Apply a QR factorization on the matrix

$$\begin{bmatrix}
\text{vec}([\Omega_1, \beta_1]) & \text{vec}([\Omega_2, \beta_2]) & \cdots & \text{vec}([\Omega_q, \beta_q])
\end{bmatrix}$$

and form an orthonormal matrix $[\xi_1, \xi_2, \ldots, \xi_q]$, where $\xi_i$ can be converted into the desired matrices $[\Omega_i, \beta_i]$ with the ‘unvec’ operation.

2. For $j = 1, 2, \ldots, q$, set $[\Omega_j, \beta_j]$ equal to the componentwise product of $[A, b]$ with $[\Omega_j, \beta_j]$.

3. Let $p = m(n + 1)$. Approximate $\omega_p$ and $\omega_q$ by using (4.1).

4. For $j = 1, 2, \ldots, q$, calculate $y_j$ by (4.7). Then compute the absolute condition vector $\bar{K}_{abs}$ by (4.8).

competitors in many aspects. Avron et al. in [1] derived a randomized least-squares solver which outperforms LAPACK by large factors for dense highly overdetermined systems. Recently, Xiang and Zou [29] used the randomized strategy for solving large-scale discrete inverse problems. In this section, we propose two algorithms using the similar randomized strategies. One is the randomize algorithm for total least squares (Rtls for short), and the other is the randomized algorithm for truncated total least squares (Rttls for short). These randomized algorithms can greatly reduce the computational time, and still give good approximate solutions.

5.1 Randomized algorithms for well-conditioned cases

Algorithm Rtls: Randomized algorithm for TLS

1. Generate an $(n + 1) \times l$ Gaussian random matrix $\Omega$.

2. Solve $(C^T C)X = \Omega$, where $C = [A, b] \in \mathbb{R}^{m \times (n + 1)}$.

3. Compute the $(n + 1) \times l$ orthonormal matrix $Q$ via QR factorization $X = QR$.

4. Solve $(C^T C)Y = Q$.

5. Form the $l \times l$ matrix $Z = Q^T Y$.

6. Compute the SVD of the smaller matrix, $Z = W \Sigma W^T$, where $W$ is orthogonal.

7. Form the $(n + 1) \times l$ matrix $V = QW$, and define $v = V(:, 1)$.

8. Form the solution $x_{Rtls} = -v(1 : n)/v(n + 1)$. 

14
For the total least squares problem, it is very crucial to obtain the right singular vector $v_{n+1}$ associated with the smallest singular values of $[A,b]$. Then the total least squares solution can be expressed by (2.3). The essential step of this traditional algorithm is the SVD of $[A,b]$. But when the size of $A$ is large, SVD can be very costly, or even prohibitive. How to reduce the computational cost and still ensure the accuracy of the approximate solution is our main concern. Our new randomized algorithm Rtls is presented in Algorithm Rtls.

Note that $l$ is a pre-specified parameter. In [10] the index $l$ is usually selected in the form $l = k + p$, where $p$ is an oversampling parameter, and $k$ corresponds to the rank $k$ specified in advance for the best rank-$k$ approximation of $A$. To understand Algorithm Rtls more, we make some remarks about each step of the algorithm. In Step 2 we obtain $X = (C^T C)^{-1} \Omega$ to extract the column information, which is further represented by an orthogonal matrix $Q$ in Step 3. The linear system involving $C^T C$ in Step 2 and 4 can be solved by direct methods or Krylov subspace iterative methods. After Step 5 the problem is reduced to a smaller matrix $Z = Q^T (C^T C)^{-1} Q$, and SVD is applied to this small matrix in Step 6. This leads to an SVD approximation, $(C^T C)^{-1} \approx V \Sigma V^T$, where $V = QW$ and $W \Sigma W^T = Z$. We then use this approximated SVD to seek the approximate total least squares solution in Step 8.

5.2 Randomized algorithms for ill-conditioned cases

Algorithm Rtls works well for the well-conditioned case. For the total least squares problem with very ill-conditioned coefficient matrices, the condition number of $C^T C$ can be very large. We need to use regularization techniques to avoid noise contaminations and obtain a meaningful approximate solution. Fierro et al. in [6] focused on the truncated Tls for solving discrete ill-posed problems, where the singular values of the coefficient matrix decay gradually. The technique of truncated Tls is similar in spirit to truncated SVD (TSVD), where the small singular values of $[A,b]$ are treated as zeros, and the problem is reduced to an exactly rank-deficient one [6]. The traditional truncated total least squares solution is given by Algorithm Ttls.

(ALGORITHM Ttls: TRADITIONAL TRUNCATED Tls)

1. Compute the SVD: $[A,b] = U \Sigma V^T = \Sigma_{i=1}^{n+1} \sigma_i u_i v_i^T$.

2. Partition the matrix, $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, where $V_{12} \in \mathbb{R}^{n \times (n+1-k)}$, and the other matrix blocks have compatible dimensions.

3. Form the minimum-norm Tls solution: $x_{Ttls} = -V_{12} V_{22}^\dagger$.

In Algorithm Ttls, the truncation parameter $k$ is user-specified or determined adaptively [6]. It is chosen such that the first $k$ large singular values dominate and $\|V_{22}\| \neq 0$. Here the Moore-Penrose inverse $V_{22}^\dagger = V_{22}^T \|V_{22}\|^{-2}$.

When the discrete ill-posed problems is of medium size, we can compute the complete SVD of $[A,b]$ directly like Step 1 in Algorithm Ttls. When the size of $A$ is large, the SVD in Step 1 is very costly since the SVD needs about $6mn^2 + 20n^3$ flops [8]. For large-scale discrete ill-posed problems, Lanczos bi-diagonalization in [6] is used to achieve a good approximation to the singular triplets associated
with several largest singular values. This approach will lose the sparsity or structure of the coefficient matrix in the first step of bi-diagonal reduction. What’s more, Lanczos procedure needs to access the coefficient matrix many times and use the BLAS-2 operations, i.e., the matrix-vector multiplications. Here we propose an alternative technique based on randomized strategies, that is, a randomized version of truncated total least squares (Rttls). This is new randomized algorithm, most flops spent on the matrix-matrix multiplications, which are the so-called nice BLAS-3 operations, and the algorithm can be realized by accessing the original large-scale matrix \( A \) only once.

\[ \text{(Algorithm Rttls: Randomized algorithm for truncated TLS)} \]

1. Generate an \((n + 1) \times l\) Gaussian random matrix \( \Omega \).
2. Form the \( m \times l \) matrix \( Y = C\Omega \), where \( C = [A, b] \).
3. Apply QR decomposition to \( Y \), i.e., \( Y = QR \), where \( Q \in \mathbb{R}^{m \times l} \).
4. Form the \( l \times (n + 1) \) matrix \( Z \) such that \( Z = Q^T C \).
5. Apply Svd to the smaller matrix \( Z \), i.e., \( Z = W\Sigma V^T \), where \( V \in \mathbb{R}^{(n+1) \times l} \).
6. Let \( V_{11} = V(1:n, 1:k) \), \( V_{21} = V(\text{end}, 1:k) \), and form the solution \( x_{\text{Rttls}} = (V_{11}^T)^+ V_{21}^T \).

Since in Algorithm Rttls we obtain a good approximation of the right singular vectors associated with largest singular values, we use an alternative expression for the TLS solution [26, Theorem 3.10] in Step 6. In Algorithm Rttls the parameter \( l \) stands for the number of sampling, and the number \( k \) is the parameter for truncating \( k \leq l \). A larger \( l \) will improve the reliability of the algorithm [10], but also increase the computational complexity. In practice, we choose \( l \ll n \), and make a balance between the reliability and the computational complexity. The truncation parameter \( k \) can be user-specified or determined by some regularization technique if no a priori estimate. Here we use randomized regularization techniques in [29] to obtain an estimation for this parameter. We first perform randomized algorithms to obtain an approximate Svd of \( A \), then a Gcv function based on this approximation is used to determine the truncation parameter \( k \) for the Tsvd solution of \( Ax \approx b \). This procedure can be performed very fast [29]. This parameter cannot be the optimal for the total least squares based on the Svd on the augmented matrix \([A, b]\), but should be a reasonable estimate for the truncation parameter in Tls. Other rules such as the L-curve, quasi-optimality, and discrepancy principle can be also used for regularization parameter choice. Our randomized Tts is constructed in the spirit of the truncated Svd (Tsvd). Tikhonov regularization for Tts [35, 34, 38, 18, 36] can be also combined with randomized algorithms, together with the existing rules for regularization parameter choice, such as L-curves, Gcv, quasi-optimality, and discrepancy principle, etc. The detailed discussion about some important issues, such as the regularization parameter choice, the scaling of \( A \) and \( b \) [26 Section 3.6.2], is beyond the scope of this paper.
| Step | \textbf{RTLS} | \textbf{RtLS} |
|------|--------------|--------------|
| 1    | \(O(nl)\)   | \(O(nl)\)   |
| 2    | \(2mn^2 + \frac{8}{3}n^3\) | \(2mnl\) |
| 3    | \(4nl^2 - \frac{4}{3}l^3\) | \(4ml^2 - \frac{4}{3}l^3\) |
| 4    | \(\frac{8}{3}n^3\) | \(2mnl\) |
| 5    | \(2nl^2\) | \(6ml^3 + 20l^3\) |
| 6    | \(26l^3\) | \(O(nk^2)\) |
| 7    | \(2nl^2\) | - |
| 8    | \(O(n)\) | - |

### 5.3 Computational complexity

We shall say a few words about the computational complexity of the randomized algorithms. The cost of each step of algorithms is listed in Table 1. For the matrix \(C = [A, b] \in \mathbb{R}^{m \times (n+1)}\), the flops count of the classical SVD based on R-bidiagonalization is about \(6mn^2 + 20n^3\) [8], while the cost of Algorithm \text{	extbf{RTLS}} is about \(2mn^2 + \frac{8}{3}n^3 + 8nl^2 + O(l^3)\); the cost of Algorithm \text{	extbf{RtLS}} is about \(4mnl^2 + (4m + 6n)l^2 + O(l^3)\). The cost of Algorithm \text{	extbf{RTLS}} is several times cheaper than the classical one. Note that the most flops are performed in Step 2 by very efficient BLAS-3 operations, and that fast Krylov subspace iterative solvers can be used in Step 2 and 4 instead. The advantage of our \text{	extbf{RTLS}} can be more obvious than just what the flops account tells.

For the cases where singular values decay rapidly, we can choose a small parameter \(l\). For most cases, \(m \geq n \gg l\). According to the flops, the ratio of the cost for Algorithm \text{	extbf{RtLS}} over that for the classical SVD is of the order \(O(l/n)\). Hence, the randomized algorithms can be essentially faster than the traditional counterpart.

### 6 Numerical examples

In this section we give numerical examples to check the perturbation bounds, the statistical condition estimates and our randomized total least squares algorithms (\text{	extbf{RTLS}} and \text{	extbf{RtLS}}). The following numerical tests are performed via MATLAB R2010a in a laptop with Intel Core i5 by using double precision.

#### 6.1 Perturbation bounds

We compare our new upper bounds (3.1) with those derived in [2, 14, 31]. We will see that these three are equal, and ours are sharper.

##### 6.1.1 Example I

In this example [2, Example 1] we consider the TLS problem \(Ax \approx b\), where \([A, b]\) is defined by

\[
[A, b] = Y \left[ \begin{array}{c}
D \\
0
\end{array} \right] Z^T \in \mathbb{R}^{m \times (n+1)}, Y = I_m - 2yy^T, Z = I_{n+1} - 2zz^T,
\]
The exact solution of the Tikhonov (Tikh) problem of computing, we also find that the bound in [31] is quite impractical for computing, since the results are listed in Table 3. From the tests, Example I in Section 6.1.1 is adopted here.

Throughout this section, we take two samples and the average of the ratios are obtained by 1000 random tests. Example I from [26, p. 42], where the condition numbers in [2, 14, 31] are the same.

We consider a random perturbation to the interlacing property, we have in exact arithmetic

\[ y \in \mathbb{R}^m \text{ and } z \in \mathbb{R}^{n+1} \text{ are random unit vectors, } D = \text{diag}(n, n - 1, \ldots, 1, 1 - \epsilon_p) \text{ for a given parameter } \epsilon_p. \]

The quantity \( \bar{\sigma}_n - \sigma_{n+1} \) measures the distance of our problem to nongenericity and, due to the interlacing property, we have in exact arithmetic

\[ \bar{\sigma}_n - \sigma_{n+1} \leq \sigma_n - \sigma_{n+1} = \epsilon_p. \]

We consider a random perturbation \( \|\delta A, \delta b\|_F = 10^{-10} \). We take \( m = 100, n = 40 \) in this example.

In Table 2 we compare the exact relative error with the upper bounds (3.1) and the above bounds derived in [2, 14, 31]. We observe that our new bounds are sharp and smaller than the bounds derived in the literature. It is also important to point out that the bound in [31] is impractical to compute for large size problems while our bound is cheaply computable.

Without considering the computational cost, we can see that the numerical results of the three condition numbers in [2, 14, 31] are the same.

### 6.1.2 Example II

Consider the second example from [26, p. 42], where

\[
A = \begin{bmatrix}
m - 1 & -1 & \cdots & -1 \\
-1 & m - 1 & \cdots & -1 \\
\vdots & \ddots & \ddots & \vdots \\
-1 & -1 & \cdots & m - 1 \\
-1 & -1 & \cdots & -1
\end{bmatrix} \in \mathbb{R}^{m \times (m - 2)}, \quad b = \begin{bmatrix}
-1 \\
-1 \\
\vdots \\
-1 \\
-1
\end{bmatrix} \in \mathbb{R}^m.
\]

The exact solution of the TLS problem \( Ax \approx b \) is \( x = -[1, 1, \ldots, 1]^\top \) and \( \sigma_{n+1} = \sqrt{m}, \bar{\sigma}_n = \sqrt{2m}. \) We consider the same random perturbations as the Example I. The results are listed in Table 3. From the experience of computing, we also find that the bound in [31] is quite impractical for computing, since MATLAB will be out of memory on our Microsoft Windows operating system.

### 6.2 Accuracy of the statistical estimates

Throughout this section, we take two samples and the average of the ratios are obtained by 1000 random tests. Example I in Section 6.1.1 is adopted here.

### Table 2: Comparisons of forward error and upper bounds for a perturbed TLS problem.

| \( \epsilon_p \) | \( \|x - x_h\|_F \) | \( K_{TLS}^Z \) | \( K^{(rel)}_1(A, b) \) | \( K^{(rel)}_1(A, b) \) | \( K_{UL} \) |
|----------------|-----------------|-------------|-----------------|-----------------|-------------|
| 9.99976032e-1 | 2.6475e-10      | 5.2973e-8   | 5.2973e-8       | 5.2973e-8       | 4.1458e-9   |
| 9.99952397e-5 | 4.3907e-6       | 2.0401e-3   | 2.0401e-3       | 2.0401e-3       | 1.2773e-3   |

### Table 3: Comparisons of forward error and upper bounds for a perturbed TLS problem.

| \( m \) | \( \|x - x_h\|_F \) | \( K_{TLS}^Z \) | \( K^{(rel)}_1(A, b) \) | \( K^{(rel)}_1(A, b) \) | \( K_{UL} \) |
|--------|-----------------|-------------|-----------------|-----------------|-------------|
| 100    | 8.8065e-11      | 7.2548e-10  | 7.2548e-10      | 7.2548e-10      | 6.7737e-10  |
| 250    | 8.7539e-11      | 1.1220e-9   | 1.1220e-9       | 1.1220e-9       | 6.8075e-10  |
Figure 1: Scé results $\tilde{\kappa}_{\text{Tls}}$ compared with the exact condition numbers $K(A, b)$ of Tls. Ratio = $\tilde{\kappa}_{\text{Tls}}/K(A, b)$. The tested matrices are of size $100 \times 40$ with $\epsilon_p=9.99976032\times 10^{-1}$ (top plot) and $\epsilon_p=9.99952397\times 10^{-5}$ (bottom plot). The horizontal dotted lines stand for the average ratios.

Table 4: Average ratio between exact condition numbers $K(A, b)$ and statistical ones $\tilde{\kappa}_{\text{Tls}}$ ($q = 2$).

| $\epsilon_p$   | $m=100, n=40$ | $m=200, n=100$ | $m=500, n=200$ |
|---------------|---------------|----------------|----------------|
| 9.99976032\times 10^{-1} | 1.4554        | 1.9209         | 1.5702         |
| 9.99952397\times 10^{-5} | 1.0336        | 1.7334         | 1.2144         |

6.2.1 Conditioning of Tls solution

We compare the statistical result obtained via Algorithm Scé0 with the exact condition number given in (4.4). Figure 1 shows the performance of Scé on the case where $m = 100, n = 40$. The plots of statistical behaviors of Scé on the matrix of size $200 \times 100$ or $500 \times 200$ are similar. We just report the average ratios in Table 4 where the values of parameter $\epsilon_p$ are taken from [2]. It shows the accuracy of our estimates, and we can find that the ratios are close to 1, which confirms the accuracy of the statistical results.

6.2.2 Componentwise condition estimation

We take $m = 200, n = 100$ in this part. In Figure 2 we plot the ratio for $\epsilon_p = 9.99976032\times 10^{-1}$ and $\epsilon_p = 9.99952397\times 10^{-5}$ between the statistical condition estimate via Algorithm Scé1 and the exact value computed by (4.6). It is observed that the ratio is close to 1 for every component of Tls solution. Therefore the statistical estimate is accurate with high probability in practical computing.
Figure 2: Componentwise comparison between the ScE results $\bar{K}_{abs}$ and the exact condition numbers $\kappa^c = [\kappa^c_1, \kappa^c_2, \cdots, \kappa^c_n]$ of TLS by using 1000 samples. Ratio = $\bar{K}_{abs} / \kappa^c$ in MATLAB notation. The tested matrices are of size $200 \times 100$ with $\epsilon_p = 9.99976032e-1$ (top plot) and $\epsilon_p = 9.99952397e-5$ (bottom plot).

6.3 Numerical experiments for randomized algorithms

In this subsection, we apply Algorithm Rtls and Algorithm Rtls to Example I, Example II and some cases in Hansen’s Regularization Tool [11]. We will compare the computational time and solution accuracy of our new randomized TLS algorithms with the traditional algorithms.

6.3.1 Algorithm Rtls on well-conditioned cases

For the case Example I in Table 5, we choose $\epsilon_p = 9.9976032e-1$, and set $n = \frac{2}{5}m$. The solution $x_{\text{Tls}}$ is computed by (2.3), while $x_{\text{Rtls}}$ is obtained by Algorithm Rtls. The corresponding execution time $\text{Time}_{\text{Tls}}$ or $\text{Time}_{\text{Rtls}}$ are measured by the MATLAB tic-toc pairs in seconds. From Table 5, we

| Matrix size | Cond(A) | Cond([A, b]) | Time_{Tls} | Time_{Rtls} | $\|x_{\text{Tls}} - x_{\text{Rtls}}\|_{\infty}$ |
|------------|---------|-------------|------------|-------------|-----------------|
| Example I  | m = 500 | 2.00E+2     | 8.34E+6    | 0.0916      | 0.0044          | 7.15E-12        |
|           | m = 1000 | 4.00E+2    | 1.67E+7    | 0.6865      | 0.0215          | 3.45E-12        |
| Example II | m = 500 | 15.8        | 22.4       | 0.1512      | 0.0261          | 3.71E-02        |
|           | m = 1000 | 22.4       | 31.6       | 1.5025      | 0.1656          | 3.73E-02        |
| Deriv2     | m = 500 | 3.04E+5     | 3.33E+5    | 0.8707      | 0.0285          | 2.64E-06        |
|           | m = 1000 | 1.22E+6    | 1.33E+6    | 11.984      | 0.3284          | 3.17E-05        |
Table 6: Tests on Shaw with different relative noise levels.

| $\delta$ | $k$ | Time$_{Ttls}$ | Time$_{Rtls}$ | $\|x_{Ttls} - x_{Rtls}\|_\infty$ |
|----------|-----|---------------|---------------|----------------------------------|
| 1E-1     | 4   | 0.01273       | 0.00349       | 5.98E-2                          |
| 1E-2     | 5   | 0.01174       | 0.00282       | 4.96E-2                          |
| 1E-3     | 7   | 0.01159       | 0.00270       | 4.10E-2                          |
| 1E-4     | 7   | 0.01135       | 0.00310       | 1.37E-4                          |

can see that our Rtls algorithm on large matrices outperforms the traditional counterpart according to computational time, while the accuracy of solutions of two methods is comparable. For the small matrices, the advantage of Rtls will not be so obvious. For an ill-conditioned matrix, MATLAB reports inaccuracy warning due to the ill-conditioned linear system in Step 2 and 4 of Algorithm Rtls. Even for the ill-conditioned case Deriv2, Algorithm Rtls can still give approximate solution with good accuracy. But for the very ill-conditioned cases, we need Algorithm Rttls.

6.3.2 Algorithm Rttls on ill-conditioned cases

Our ill-conditioned cases are taken from Hansen’s Regularizaton Tools [11]. For example, the case Shaw is generated by the command $[\vec{A}, \vec{b}, x_{true}] = \text{Shaw}(m)$. Then noises are added to $\vec{A}$ and $\vec{b}$. Suppose that $\delta$ is the relative noise level. We define

$$b = \vec{b} + \delta \|\vec{b}\| \frac{\vec{\zeta}}{\|\vec{\zeta}\|}, \quad A = \vec{A} + \delta \|\vec{A}\|_F \frac{Z}{\|Z\|_F},$$

where $\vec{\zeta}$ is a random vector, $\vec{\zeta} = 2 \ast \text{rand}(m, 1) - 1$; $Z$ is a random matrix, $Z = 2 \ast \text{rand}(m, n) - 1$. It is easy to verify that

$$\frac{\|b - \vec{b}\|}{\|b\|} = \frac{\|A - \vec{A}\|_F}{\|\vec{A}\|_F} = \delta.$$

Then we seek the total least squares solution of $Ax \approx b$.

We first test Algorithm Rttls on the $100 \times 100$ matrix Shaw with different relative noise levels $\delta$. The results are given in Table [5]. From this table we can see that the results of Algorithm Rttls are very close to those of traditional Ttls even for the relative noise level as large as 10%. According to the computational time, Algorithm Rttls is substantially faster than the traditional Ttls. The compute solutions for the case where the relative noise level $\delta=1E-1$ are presented in Figure [3].

We then test Algorithm Rttls on larger matrices. We set the parameter for sampling size $l = 10$ and the relative noise level $\delta=1E-3$ for all cases. The truncation parameter $k$ is estimated by the randomized algorithm with Gcv and Tsvd [29]. After the determination of parameter $k$, the computational time for implementing Algorithm Ttls and Algorithm Rttls is recorded in Time$_{Ttls}$ and Time$_{Rttls}$ respectively. The results are given in Table [7]. The randomized strategy can greatly speed up the traditional Algorithm Ttls. The advantage of our Algorithm Rttls is more obvious when we test the larger matrices. The plots of the computed solutions are given in Figure [4].
Table 7: Algorithm Rttls on ill-conditioned cases.

| Matrix   | Size | $k$ | $\text{Time}_{TLS}$ | $\text{Time}_{Rttls}$ | $\|x_{TLS} - x_{Rttls}\|_\infty$ |
|----------|------|-----|----------------------|------------------------|----------------------------------|
| Baart    | $100$ | $4$ | $0.0114$            | $0.0031$               | $1.27E-2$                        |
|          | $1000$ | $4$ | $14.75$             | $0.0100$               | $1.94E-2$                        |
| Deriv2   | $100$ | $6$ | $0.0109$            | $0.0022$               | $5.33E-3$                        |
|          | $1000$ | $8$ | $13.04$             | $0.0127$               | $5.41E-3$                        |
| Foxgood  | $100$ | $2$ | $0.0119$            | $0.0041$               | $3.31E-6$                        |
|          | $1000$ | $3$ | $14.93$             | $0.0114$               | $2.23E-3$                        |
| Gravity  | $100$ | $7$ | $0.0115$            | $0.0038$               | $5.67E-3$                        |
|          | $1000$ | $9$ | $14.65$             | $0.0183$               | $2.54E-3$                        |
| Heat     | $100$ | $9$ | $0.0118$            | $0.0020$               | $7.27E-2$                        |
|          | $1000$ | $8$ | $13.78$             | $0.0152$               | $9.68E-2$                        |
| Laplace  | $100$ | $7$ | $0.0121$            | $0.0037$               | $1.35E-3$                        |
|          | $1000$ | $9$ | $14.93$             | $0.0138$               | $5.69E-3$                        |
| Phillips | $100$ | $7$ | $0.0107$            | $0.0035$               | $1.03E-3$                        |
|          | $1000$ | $7$ | $13.81$             | $0.0169$               | $8.31E-4$                        |
Figure 4: Rtls for ill-conditioned cases of size $n = 1000$ with relative noise level $\delta = 1E-3$. 

(a) BAART 
(b) DERIV2 
(c) FOXGOOD 
(d) GRAVITY 
(e) HEAT 
(f) LAPLACE 
(g) PHILLIPS 
(h) SHAW
7 Conclusion

In this paper, we derive a new perturbation bound for the total least squares problem. This sharper and numerically computable perturbation bound is well illustrated by the numerical examples. Also we show that three kinds of condition numbers in \[2, 14, 31\] obtained through different ways are mathematically equivalent. To estimate the conditioning of the total least squares problem, Scé techniques are applied, and the Scé results are compared with the exact values in numerical experiments. We propose randomized algorithms \(\text{R tls}\) and \(\text{R ttls}\) for the numerical solutions of well-conditioned and ill-conditioned total least squares problems, respectively. These randomized algorithms can greatly reduce the computational time, and still give solutions with good accuracy. The regularization parameter in \(\text{R ttls}\) is estimated by the truncated parameter of the Tsvd solution of \(Ax \approx b\) based on a fast randomized Svd of \(A\) \[29\]. Then a randomized Svd of \([A, b]\) together with this truncation parameter yields a good approximate Ttls solutions to the large-scale ill-conditioned total least squares problem. The detailed investigation on other regularization parameter choices, and other techniques such as Tikhonov regularization, will be our future research.

References

[1] Blendenpik: Supercharging LAPACK’s least-squares solver, Avron, Haim and Maymounkov, Petar and Toledo, Sivan, SIAM J. Sci. Comput., Vol. 32, No. 3, pp. 1217–1236.
[2] Baboulin, Marc and Gratton, Serge, A contribution to the conditioning of the total least-squares problem, SIAM J. Matrix Anal. Appl., Vol. 32, 2011, No. 3, pp. 685–699.
[3] Baboulin, Marc and Gratton, Serge and Lacroix, Rémi and Laub, Alan, Efficient computation of condition estimates for linear least squares problems. To appear in the Proceedings of the 10th International Conference on Parallel Processing and Applied Mathematics, PPAM 2013(09/2013).
[4] Chang, X.-W. and Titley-Peloquin, D., Backward perturbation analysis for scaled total least-squares problems, Numer. Linear Algebra Appl., Vol. 16, 2009, No. 8, pp. 627–648.
[5] De Moor, Bart and David, Johan, Total linear least squares and the algebraic Riccati equation, Systems Control Lett., Vol. 18, 1992, No. 5, pp. 329–337.
[6] Fierro, R. D. and Golub, G. H. and Hansen, P. C. and O’Leary, D. P., Regularization by truncated total least squares, SIAM J. Sci. Comput., Vol. 18, 1997, No. 4, pp. 1223–1241.
[7] Golub, Gene H. and Van Loan, Charles F., An analysis of the total least squares problem, SIAM J. Numer. Anal., Vol. 17, 1980, No. 6, pp. 883–893.
[8] Golub, Gene H. and Van Loan, Charles F., Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences, 4th, Johns Hopkins University Press, Baltimore, MD, 2013, pp. xxx+698.
[9] Gratton, Serge and Titley-Peloquin, David and Ilunga, Jean Tshimanga, Sensitivity and conditioning of the truncated total least squares solution, SIAM J. Matrix Anal. Appl., Vol. 34, 2013, No. 3, pp. 1257–1276.
[10] Halko, N. and Martinsson, P. G. and Tropp, J. A., Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev., Vol. 53, 2011, No. 2, pp. 217–288.
[11] Hansen, Per Christian, Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems, Numer. Algorithms, Vol. 6, 1994, No. 1-2, pp. 1–35.
[12] Hansen, Per Christian and Nagy, James G and O’leary, Dianne P, Deblurring Images: Matrices, Spectra, and Filtering, Vol. 3, 2006, Society for Industrial and Applied Mathematics, Philadelphia.

[13] Hnětynková, Iveta and Plešinger, Martin and Sima, Diana Maria and Strakoš, Zdeněk and Van Huffel, Sabine, The total least squares problem in $AX \approx B$: a new classification with the relationship to the classical works, SIAM J. Matrix Anal. Appl., Vol. 32, 2011, No. 3, pp. 748–770.

[14] Li, Bingyu and Jia, Zhongxiao, Some results on condition numbers of the scaled total least squares problem, Linear Algebra Appl., Vol. 435, 2011, No. 3, pp. 674–686.

[15] Jia, Zhongxiao and Li, Bingyu, On the condition number of the total least squares problem, Numer. Math., Vol. 125, 2013, No. 1, pp. 61–87.

[16] Kenney, C. S. and Laub, A. J., Small-sample statistical condition estimates for general matrix functions, SIAM J. Sci. Comput., Vol. 15, 1994, No. 1, pp. 36–61.

[17] Kenney, C. S. and Laub, A. J. and Reese, M. S., Statistical condition estimation for linear least squares, SIAM J. Matrix Anal. Appl., Vol. 19, 1998, No. 4, pp. 906–923.

[18] Lee, Geunseop and Fu, Haoying and Barlow, Jesse L, Fast High-Resolution Image Reconstruction Using Tikhonov Regularization Based Total Least Squares, SIAM J. Sci. Comput., Vol. 35, No. 1, pp. B275–B290, 2013.

[19] Li, Zi-Cai and Chien, Cheng-Sheng and Huang, Hung-Tsai, Effective condition number for finite difference method, J. Comput. Appl. Math., Vol. 198, 2007, No. 1, pp. 208–235.

[20] Li, Zi-Cai and Huang, Hung-Tsai and Chen, Jeng-Tzong and Wei, Yimin, Effective condition number and its applications, Computing, Vol. 89, 2010, No. 1-2, pp. 87–112.

[21] Li, Zi-Cai and Huang, Hung-Tsai and Wei, Yimin, Ill-conditioning of the truncated singular value decomposition, Tikhonov regularization and their applications to numerical partial differential equations, Numer. Linear Algebra Appl., Vol. 18, 2011, No. 2, pp. 205–221.

[22] Liu, Xin-Guo, Solvability and perturbation analysis of the total least squares problem, Acta Math. Appl. Sinica, Vol. 19, 1996, No. 2, pp. 254–262.

[23] Markovsky, Ivan and Rastello, Maria Luisa and Premoli, Amedeo and Kukush, Alexander and Van Huffel, Sabine, The element-wise weighted total least-squares problem, Comput. Statist. Data Anal., Vol. 50, 2006, No. 1, pp. 181–209.

[24] Paige, Christopher C. and Strakoš, Zdeněk, Scaled total least squares fundamentals, Numer. Math., Vol. 91, 2002, No. 1, pp. 117–146.

[25] Stewart, G. W., A second order perturbation expansion for small singular values, Linear Algebra Appl., Vol. 56, 1984, pp. 231–235.

[26] Van Huffel, Sabine and Vandewalle, Joos, The Total Least Squares Problem, Frontiers in Applied Mathematics, Vol. 9, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1991, pp. xiv+300.

[27] Wei, Musheng, The analysis for the total least squares problem with more than one solution, SIAM J. Matrix Anal. Appl., Vol. 13, 1992, No. 3, pp. 746–763.

[28] Wei, Musheng, Algebraic relations between the total least squares and least squares problems with more than one solution, Numer. Math., Vol. 62, 1992, No. 1, pp. 123–148.

[29] Xiang, Hua and Zou, Jun, Regularization with randomized SVD for large-scale discrete inverse problems, Inverse Problems, Vol. 29, 2013, No. 8, pp. 085008, 23.
[30] Xu, Wei and Qiao, Sanzheng and Wei, Yimin, A note on the scaled total least squares problem, Linear Algebra Appl., Vol. 428, 2008, No. 2-3, pp. 469–478.

[31] Zhou, Liangmin and Lin, Lijing and Wei, Yimin and Qiao, Sanzheng, Perturbation analysis and condition numbers of scaled total least squares problems, Numer. Algorithms, Vol. 51, 2009, No. 3, pp. 381–399.

[32] Sima, Diana M. and Van Huffel, Sabine and Golub, Gene H., Regularized total least squares based on quadratic eigenvalue problem solvers, BIT, Vol. 44, 2004, No. 4, pp. 793–812.

[33] Lampe, Jörg and Voss, Heinrich, Large-scale Tikhonov regularization of total least squares, J. Comput. Appl. Math., Vol. 238, 2013, pp. 95–108.

[34] Golub, Gene H. and Hansen, Per Christian and O’Leary, Dianne P., Tikhonov regularization and total least squares, SIAM J. Matrix Anal. Appl., Vol. 21, 1999, No. 1, pp. 185–194.

[35] Beck, Amir and Ben-Tal, Aharon, On the solution of the Tikhonov regularization of the total least squares problem, SIAM J. Optim., Vol. 17, 2006, No. 1, pp. 98–118.

[36] Lu, Shuai and Pereverzev, Sergei V. and Tautenhahn, Ulrich, Regularized total least squares: computational aspects and error bounds, SIAM J. Matrix Anal. Appl., Vol. 31, 2009, No. 3, pp. 918–941.

[37] Beck, Amir and Ben-Tal, Aharon and Teboulle, Marc, Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares, SIAM J. Matrix Anal. Appl., Vol. 28, 2006, No. 2, pp. 425–445.

[38] Lampe, J. and Voss, H., Global convergence of RTLSQEP: a solver of regularized total least squares problems via quadratic eigenproblems, Math. Model. Anal., Vol. 13, 2008, No. 1, pp. 55–66.

[39] Renaut, Rosemary A. and Guo, Hongbin, Efficient algorithms for solution of regularized total least squares, SIAM J. Matrix Anal. Appl., Vol. 26, 2004, No. 2, pp. 457–476.

[40] Diao, Huaian and Shi, Xinghua and Wei, Yimin, Effective condition numbers and small sample statistical condition estimation for the generalized Sylvester equation, Sci. China Math., Vol. 56, 2013, No. 5, pp. 967–982.