A CONTRIBUTION TO THE CONDITIONING OF THE TOTAL LEAST SQUARES PROBLEM

MARC BABOULIN* AND SERGE GRATTON‡

LAPACK Working Note 236

Abstract. We derive closed formulas for the condition number of a linear function of the total least squares solution. Given an over determined linear system $Ax = b$, we show that this condition number can be computed using the singular values and the right singular vectors of $[A, b]$ and A. We also provide an upper bound that requires the computation of the largest and the smallest singular value of $[A, b]$ and the smallest singular value of A. In numerical examples, we compare these values and the resulting forward error bounds with the error estimates given in [17].
Keywords: total least squares, condition number, normwise perturbations, errors-in-variables model
AMS Subject Classification (2000): 65F35

1. Introduction. Given a matrix $A \in \mathbb{R}^{m \times n}$ $(m > n)$ and an observation vector $b \in \mathbb{R}^m$, the standard over determined linear least squares (LS) problem consists in finding a vector $x \in \mathbb{R}^n$ such that Ax is the best approximation of b. Such a problem can be formulated using what is referred to as the linear statistical model

$$b = Ax + \epsilon, \ A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m, \ \text{rank}(A) = n,$$

where ϵ is a vector of random errors having expected value $E(\epsilon) = 0$ and variance-covariance $V(\epsilon) = \sigma^2 I$.

In the linear statistical model, random errors affect exclusively the observation vector b while A is considered as known exactly. However it is often more realistic to consider that measurement errors might also affect A. This case is treated by the statistical model referred to as Errors-In-Variables model (see e.g [17, p. 230] and [5, p. 176]), where we have the relation

$$(A + E)x = b + \epsilon.$$

In general it is assumed in this model that the rows of $[E, \epsilon]$ are independently and identically distributed with common zero mean vector and common covariance matrix. The corresponding linear algebra problem, discussed originally in [12], is called the Total Least Squares (TLS) problem and can be expressed as:

$$\min_{E, \epsilon} \|(E, \epsilon)\|_F, \ \ (A + E)x = b + \epsilon, \tag{1.1}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. As mentioned in [17, p. 238], the TLS method enables us to obtain a more accurate solution when entries of A are perturbed under certain conditions.

In error analysis, condition numbers are considered as fundamental tools since they measure the effect on the solution of small changes in the data. In particular the conditioning of the least squares problem was extensively studied in the numerical linear algebra literature (see e.g [5, 7, 8, 9, 10, 15, 16, 18, 19, 22]). The more general case of the conditioning of a linear function of an LS solution was studied in [2] and [4]

---

*Université Paris-Sud and INRIA, France (`marc.baboulin@inria.fr`).
†ENSEEIHT and CERFACS, France (`serge.gratton@enseeiht.fr`).

when perturbations on data are measured respectively normwise and componentwise (note that the componentwise and normwise condition numbers for LS problems were also treated in [9] but without the generalization to a linear function of the solution). Moreover we can find in [3] algorithms using the software libraries LAPACK [1] and ScaLAPACK [6] as well as physical applications.

The notion of **Total Least Squares** was initially defined in the seminal paper [12] that was the first to propose a numerically stable algorithm. Then various aspects of the TLS problem were developed in the comprehensive book [17] including a large survey of theoretical bases, computational methods and applications but also sensitivity analysis with for instance upper bounds for the TLS perturbation. The so-called **Scaled Total Least Squares** (STLS) problem $(\min_{E,\epsilon} \|(E,\epsilon)\|_F, \quad (A + E)x\gamma = \gamma b + \epsilon$, for a given scaling parameter γ) was formulated in [20] in which were addressed the difficulties coming from non existence of TLS solution. In a recent paper [23], we can find sharp estimates of the normwise, mixed and componentwise condition numbers of the Scaled Total Least Squares (STLS) problem.

Here we are concerned with the TLS problem, which is a special case of the STLS problem, and we will consider perturbations on data (A, b) that are measured normwise using a product norm. Contrary to [23], we will consider the general case of the conditioning of $L^T x$, linear function of the TLS solution for which we will derive an exact formula. The common situations correspond to the special cases where L is the identity matrix (condition number of the TLS solution) or a canonical vector (condition number of one solution component). The conditioning of a nonlinear function of a TLS solution can also be obtained by replacing, in the condition number expression, the quantity L^T by the Jacobian matrix at the solution.

We notice that the normwise condition number expression proposed in [23] is based on the evaluation of the norm of a matrix expressed as a Kronecker product resulting in large matrices which may be, as pointed out by the authors, impractical to compute, especially for large size problems. We propose here a computable expression for the resulting condition number (exact formula and upper bound) using data that could be already available from the TLS solution process, namely by-products of the SVD decomposition of A and $[A, b]$. We also make use of the adjoint operator which enables us to work on a space of lower dimension and we propose a practical algorithm based on the power method.

2. Definitions and notations.

2.1. The total least squares problem. Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, with $m > n$. Following [17], we consider the two singular value decompositions of A, and $[A, b]$: $A = U'\Sigma'V'^T$ and $[A, b] = U\Sigma V^T$. We also set $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n+1})$, $\Sigma' = \mathrm{diag}(\sigma'_1, \ldots, \sigma'_n)$, where the singular values are in nonincreasing order, and define $\lambda_i = \sigma_i^2$, and $\lambda'_i = \sigma_i'^2$. From [5, p. 178], we have the interlacing property

$$\sigma_1 \geq \sigma'_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq \sigma'_n \geq \sigma_{n+1}. \tag{2.1}$$

We consider the total least squares problem expressed in Equation (1.1) and we assume in this text that the *genericity* condition $\sigma'_n > \sigma_{n+1}$ holds (for more information about the "nongeneric" problem see e.g [17, 20]). From [17, Theorems 2.6 and 2.7], it follows that the TLS solution x exists, is unique, and satisfies

$$x = \left(A^T A - \lambda_{n+1} I_n\right)^{-1} A^T b. \tag{2.2}$$

In addition, $\begin{bmatrix} x \\ -1 \end{bmatrix}$ is an eigenvector of $[A, b]^T[A, b]$ associated with the simple eigenvalue λ_{n+1}, i.e $\sigma'_n > \sigma_{n+1}$ guarantees that λ_{n+1} is not a semi-simple eigenvalue of $[A, b]^T[A, b]$. As for linear least squares problems, we define the total least squares residual $r = b - Ax$, which enables us to write

$$\lambda_{n+1} = \frac{1}{1 + x^T x} \begin{bmatrix} x^T, & -1 \end{bmatrix} \begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = \frac{r^T r}{1 + x^T x}. \tag{2.3}$$

As mentioned in [17, p. 35], the TLS solution is obtained by scaling the last right singular vector v_{n+1} of $[A, b]$ until its last. component is -1 and, if $v_{i,n+1}$ denotes the ith component of v_{n+1}, we have

$$x = -\frac{1}{v_{n+1,n+1}}[v_{1,n+1}, \ldots, v_{n,n+1}]^T. \tag{2.4}$$

The TLS method involves an SVD computation and the computational cost is higher than that of a classical LS problem (about $2mn^2 + 12n^3$ as mentioned in [13, p. 598], to be compared with the approximately $2mn^2$ flops required for LS solved via Householder QR factorization). However, there exist faster methods referred to as "partial SVD" (PSVD) that calculate only the last right singular vector or a basis of the right singular subspace associated with the smallest singular values of $[A, b]$ (see [17, p. 97]).

2.2. Condition number of the TLS problem. To measure the perturbations on data A and b, we consider the product norm defined on $\mathbb{R}^{m \times n} \times \mathbb{R}^m$ by $\|(A, b)\|_F = \sqrt{\|A\|_F^2 + \|b\|_2^2}$ and we take the Euclidean norm $\|x\|_2$ for the solution space $\mathbb{R}^n$. In the following, the $n \times n$ identity matrix is denoted by I_n.

Let L be a given $n \times k$ matrix, with $k \le n$. We suppose here that L is not perturbed numerically and we consider the mapping

$$g : \begin{array}{ccc} \mathbb{R}^{m \times n} \times \mathbb{R}^m & \longrightarrow & \mathbb{R}^k \\ (A, b) & \longmapsto & g(A, b) = L^T x = L^T (A^T A - \lambda_{n+1} I_n)^{-1} A^T b, \end{array}$$

Since λ_{n+1} is simple, g is a Fréchet-differentiable function of A and b, and the genericity assumption ensures that the matrix $(A^T A - \lambda_{n+1} I_n)^{-1}$ is also Fréchet-differentiable in a neighborhood of (A, b). As a result, g is Fréchet-differentiable in a neighborhood of (A, b).

The approach that we follow here is based on the work by [11, 21] where the mathematical difficulty of a problem is measured by the norm of the Fréchet derivative of the problem solution expressed as a function of data. This measure is an attainable bound at first order, and may therefore be approximate when large perturbations are considered.

Using the definition given in [11, 21], we can express the condition number of $L^T x$, linear function of the TLS solution as

$$K(L, A, b) = \max_{(\Delta A, \Delta b) \ne 0} \frac{\|g'(A, b).(\Delta A, \Delta b)\|_2}{\|(\Delta A, \Delta b)\|_F}. \tag{2.5}$$

$K(L, A, b)$ is sometimes called the *absolute* condition number of $L^T x$ as opposed to the *relative* condition number of $L^T x$ and defined, when $L^T x$ is nonzero by

$$K^{(rel)}(L, A, b) = K(L, A, b)\|(A, b)\|_F / \|L^T x\|_2. \tag{2.6}$$

4

In the remainder, the quantity $K(L, A, b)$ will be simply referred to as the TLS condition number, even though the proper conditioning of the TLS solution corresponds to the special case when L is the identity matrix.

In the expression $g'(A, b).(\Delta A, \Delta b)$, the "." operator denotes that we apply the linear function $g'(A, b)$ to the variable $(\Delta A, \Delta b)$. We will use this notation throughout this paper to designate the image of a vector or a matrix by a linear function.

REMARK 1. The case where $g(A, b) = h(x)$, with h being a differentiable nonlinear function mapping $\mathbb{R}^n$ to $\mathbb{R}^k$ is also covered because we have

$$g'(A, b).(\Delta A, \Delta b) = h'(x).(x'(A, b).(\Delta A, \Delta b)),$$

and L^T would correspond to the Jacobian matrix $h'(x)$. The nonlinear function h can be for instance the Euclidean norm of part of the solution (e.g in the computation of Fourier coefficients when we are interested in the quantity of signal in a given frequency band).

3. Explicit formula for the TLS condition number.

3.1. Fréchet derivative. In this section, we compute the Fréchet dérivative of g under the genericity assumption, which enables us to obtain an explicit formula for the TLS condition number in Proposition 2.

PROPOSITION 1. *Under the genericity assumption, g is Fréchet differentiable in a neighborhood of (A, b). Setting $B_\lambda = A^T A - \lambda_{n+1} I_n$, the Fréchet derivative of g at (A, b) is expressed by*

$$g'(A, b) : \mathbb{R}^{m \times n} \times \mathbb{R}^m \longrightarrow \mathbb{R}^k$$
$$(\Delta A, \Delta b) \longmapsto L^T B_\lambda^{-1} \left(A^T + \frac{2xr^T}{1 + x^T x} \right) (\Delta b - \Delta A x) + \qquad (3.1)$$
$$L^T B_\lambda^{-1} \Delta A^T r.$$

<u>Proof:</u> The result is obtained from the chain rule. Since λ_{n+1}, expressed in Equation (2.3), is a simple eigenvalue of $[A, b]^T [A, b]$ with corresponding unit eigenvector $\frac{1}{\sqrt{1 + x^T x}} \begin{bmatrix} x^T & -1 \end{bmatrix}^T$, λ_{n+1} is differentiable in a neighborhood of (A, b) and then we have

$$\lambda'_{n+1}(A, b).(\Delta A, \Delta b) = \frac{1}{1 + x^T x} \begin{bmatrix} x^T & -1 \end{bmatrix} \begin{bmatrix} \Delta A^T A + A^T \Delta A & \Delta A^T b + A^T \Delta b \\ b^T \Delta A + \Delta b^T A & \Delta b^T b + b^T \Delta b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix}$$

$$= \frac{2}{1 + x^T x} \left(x^T \Delta A^T A x - x^T \Delta A^T b - x^T A^T \Delta b + b^T \Delta b \right)$$

$$= \frac{2}{1 + x^T x} \left(-x^T \Delta A^T r + (b^T - x^T A^T) \Delta b \right)$$

$$= \frac{2}{1 + x^T x} \left(-r^T \Delta A x + r^T \Delta b \right),$$

yielding

$$\lambda'_{n+1}(A, b).(\Delta A, \Delta b) = \frac{2r^T (\Delta b - \Delta A x)}{1 + x^T x}. \qquad (3.2)$$

Applying the chain rule to B_λ^{-1}, we obtain

$$(B_\lambda^{-1})'(A, b).(\Delta A, \Delta b) = -B_\lambda^{-1} \left(\Delta A^T A + A^T \Delta A - \lambda'_{n+1}(A, b).(\Delta A, \Delta b) I_n \right) B_\lambda^{-1}$$

$$= -B_\lambda^{-1} \left(\Delta A^T A + A^T \Delta A - \frac{2r^T (\Delta b - \Delta A x)}{1 + x^T x} I_n \right) B_\lambda^{-1}.$$

The chain rule now applied to $g(A, b)$ leads to

$$g'(A,b).(\Delta A, \Delta b) = -L^T B_\lambda^{-1}\left(\Delta A^T A + A^T \Delta A - \lambda'_{n+1}(A,b).(\Delta A, \Delta b)I_n\right) B_\lambda^{-1}A^T b + L^T B_\lambda^{-1}\Delta A^T b + L^T B_\lambda^{-1} A^T \Delta b$$

$$= -L^T B_\lambda^{-1}\left(\Delta A^T A + A^T \Delta A - \lambda'_{n+1}(A,b).(\Delta A, \Delta b)I_n\right) x + L^T B_\lambda^{-1}\left(\Delta A^T b + A^T \Delta b\right)$$

$$= L^T B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^T x}\right)(\Delta b - \Delta A x) + L^T B_\lambda^{-1}\Delta A^T r,$$

which gives the result.

□

We now introduce the vec operation that stacks all the columns of a matrix into a long vector: for $A = [a_1, \ldots, a_n] \in \mathbb{R}^{m \times n}$, $\mathrm{vec}(A) = [a_1^T, \ldots, a_n^T]^T \in \mathbb{R}^{mn \times 1}$. Let $P \in \mathbb{R}^{mn \times mn}$ denote the permutation matrix that represents the matrix transpose by $vec(B^T) = Pvec(B)$. We remind also that $vec(AXB) = (B^T \otimes A)vec(X)$, where $\otimes$ denotes the Kronecker product of two matrices [14, p. 21].

Let us now express the matrix representing $g'(A, b)$, denoted by $\mathcal{M}_{g'}$. Since $g'(A,b).(\Delta A, \Delta b) \in \mathbb{R}^k$, we have $g'(A,b).(\Delta A, \Delta b) = vec(g'(A,b).(\Delta A, \Delta b))$ and setting in addition $D_\lambda = L^T B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^T x}\right) \in \mathbb{R}^{k \times m}$, we obtain from (3.1)

$$g'(A,b).(\Delta A, \Delta b) = \mathrm{vec}\left(D_\lambda(\Delta b - \Delta A x) + L^T B_\lambda^{-1}\Delta A^T r\right)$$

$$= \left(-x^T \otimes D_\lambda\right)\mathrm{vec}(\Delta A) + \left(r^T \otimes (L^T B_\lambda^{-1})\right)\mathrm{vec}(\Delta A^T) + D_\lambda \Delta b$$

$$= \left[-x^T \otimes D_\lambda + \left(r^T \otimes (L^T B_\lambda^{-1})\right)P, \quad D_\lambda\right]\begin{bmatrix}\mathrm{vec}(\Delta A) \\ \Delta b\end{bmatrix}.$$

Then we get

$$\mathcal{M}_{g'} = \left[-x^T \otimes D_\lambda + \left(r^T \otimes (L^T B_\lambda^{-1})\right)P, \quad D_\lambda\right] \in \mathbb{R}^{k \times (nm+m)}.$$

But we have $\|(\Delta A, \Delta b)\|_F = \left\|\begin{bmatrix}\mathrm{vec}(\Delta A) \\ \Delta b\end{bmatrix}\right\|_2$ and then, from Proposition 1 and using the definition of $K(L, A, b)$ given in Expression (2.5), we get the following proposition that expresses the TLS condition number in terms of the norm of a matrix.

PROPOSITION 2. *The condition number of $g(A, b)$ is given by*

$$K(L, A, b) = \|\mathcal{M}_{g'}\|_2,$$

where

$$\mathcal{M}_{g'} = \left[-x^T \otimes D_\lambda + \left(r^T \otimes (L^T B_\lambda^{-1})\right)P, \quad D_\lambda\right] \in \mathbb{R}^{k \times (nm+m)}.$$

3.2. Adjoint operator and algorithm. Computing $K(L, A, b)$ reduces to computing the spectral norm of the $k \times (nm + m)$ matrix $\mathcal{M}_{g'}$. For large values of n or m, it is not possible to build explicitly the generally dense matrix $\mathcal{M}_{g'}$. Iterative techniques based on the power method [16, p. 289] or on the Lanczos method [13] are better suited. These algorithms involve however the computation of the product of $\mathcal{M}_{g'}^T$ by a vector $y \in \mathbb{R}^k$. We describe now how to perform this operation.

Using successively the fact that $B_\lambda^{-T} = B_\lambda^{-1}$, $(A \otimes B)^T = A^T \otimes B^T$, $vec(AXB) = (B^T \otimes A)vec(X)$ and $P^T = P^{-1}$ we have

$$
\begin{aligned}
\mathcal{M}_{g'}^T y &= \begin{bmatrix} -x \otimes D_\lambda^T + P^T \left(r \otimes (B_\lambda^{-T} L) \right) \\ D_\lambda^T \end{bmatrix} y \\
&= \begin{bmatrix} -(x \otimes D_\lambda^T)\mathrm{vec}(y) + P^T \left(r \otimes (B_\lambda^{-1} L) \right) \mathrm{vec}(y) \\ D_\lambda^T y \end{bmatrix} \\
&= \begin{bmatrix} P^{-1} \left(P\mathrm{vec} \left(-D_\lambda^T y x^T \right) + \mathrm{vec} \left(B_\lambda^{-1} L y r^T \right) \right) \\ D_\lambda^T y \end{bmatrix} \\
&= \begin{bmatrix} P^{-1} \left(\mathrm{vec} \left((-D_\lambda^T y x^T)^T \right) + \mathrm{vec} \left(B_\lambda^{-1} L y r^T \right) \right) \\ D_\lambda^T y \end{bmatrix} \\
&= \begin{bmatrix} P^{-1}\mathrm{vec} \left(-x y^T D_\lambda + B_\lambda^{-1} L y r^T \right) \\ D_\lambda^T y \end{bmatrix},
\end{aligned}
$$

and since for any matrix B we have $P^{-1}vec(B) = vec(B^T)$, we get

$$
\mathcal{M}_{g'}^T y = \begin{bmatrix} \mathrm{vec} \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1} \right) \\ D_\lambda^T y \end{bmatrix}. \tag{3.3}
$$

This leads us to the following proposition.

PROPOSITION 3. *The adjoint operator of $g'(A, b)$ using the scalar products* $\mathrm{trace}(A_1^T A_2) + b_1^T b_2$ *and $y_1^T y_2$ respectively on $\mathbb{R}^{m \times n} \times \mathbb{R}^m$ and $\mathbb{R}^k$ is*

$$
\begin{aligned}
g'^*(A, b) \ : \mathbb{R}^k &\longrightarrow & \mathbb{R}^{m \times n} \times \mathbb{R}^m \\
y &\longmapsto & \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}, D_\lambda^T y \right)
\end{aligned} \tag{3.4}
$$

In addition, if $k = 1$ we have

$$
K(L, A, b) = \sqrt{\| -D_\lambda^T x^T + r L^T B_\lambda^{-1} \|_F^2 + \| D_\lambda \|_2^2} \tag{3.5}
$$

Proof: Let us denote by $< (A_1, b_1), (A_2, b_2) >$ the scalar product $\mathrm{trace}(A_1^T A_2) + b_1^T b_2$ on $\mathbb{R}^{m \times n} \times \mathbb{R}^m$. We have for any $y \in \mathbb{R}^k$,

$$
\begin{aligned}
y^T(g'(A, b).(\Delta A, \Delta b)) &= y^T \mathcal{M}_{g'} \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \Delta b \end{bmatrix} \\
&= (\mathcal{M}_{g'}^T y)^T \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \Delta b \end{bmatrix} \\
&= \mathrm{vec} \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1} \right)^T \mathrm{vec}(\Delta A) + (D_\lambda^T y)^T \Delta b.
\end{aligned}
$$

Using now the fact that, for matrices A_1 and A_2 of identical sizes, $\mathrm{vec}(A_1)^T \mathrm{vec}(A_2) = \mathrm{trace}(A_1^T A_2)$, we get

$$
\begin{aligned}
y^T(g'(A, b).(\Delta A, \Delta b)) &= \mathrm{trace} \left((-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1})^T \Delta A \right) + (D_\lambda^T y)^T \Delta b \\
&= \ < \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}, D_\lambda^T y \right), (\Delta A, \Delta b) > \\
&= \ < g'^*(A, b).y, (\Delta A, \Delta b) >,
\end{aligned}
$$

which concludes the first part of the proof.

For the second part, we use

$$K(L, A, b) = \|\mathcal{M}_{g'}\|_2 = \|\mathcal{M}_{g'}^T\|_2 = \max_{y \neq 0} \frac{\left\| \begin{bmatrix} \text{vec}\left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}\right) \\ D_\lambda^T y \end{bmatrix} \right\|_2}{\|y\|_2}$$

Since $k = 1$, we have $y \in \mathbb{R}$, and $K(L, A, b) = \left\| \begin{bmatrix} \text{vec}\left(-D_\lambda^T x^T + r L^T B_\lambda^{-1}\right) \\ \text{vec}(D_\lambda^T), \end{bmatrix} \right\|_2$ and the result follows from the relation $\text{vec}(A_1)^T \text{vec}(A_1) = \text{trace} A_1^T A_1 = \|A_1\|_F^2$. $\qquad \square$

REMARK 2. The special case $k = 1$ recovers the situation where we compute the conditioning of the ith solution component. In that case L is the ith canonical vector of $\mathbb{R}^n$ and, in Equation (3.5), $L^T B_\lambda^{-1}$ is the ith row of B_λ^{-1} and D_λ is the ith row of $B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^T x}\right)$.

Using (3.1) and (3.4), we can now write in Algorithm 1 the iteration of the power method ([16, p. 289]) to compute the TLS condition number $K(L, A, b)$.

Algorithm 1 : Condition number of TLS problem
Select initial vector $y \in \mathbb{R}^k$
for p=1,2,...
$\quad (A_p, b_p) = \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}, D_\lambda^T y\right)$
$\quad \nu = \|(A_p, b_p)\|_F$
$\quad (A_p, b_p) \leftarrow (\frac{1}{\nu} \cdot A_p, \frac{1}{\nu} \cdot b_p)$
$\quad y = L^T B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^T x}\right)(b_p - A_p x) + L^T B_\lambda^{-1} A_p^T r$
end
$K(L, A, b) = \sqrt{\nu}$

The quantity ν computed by Algorithm 1 is the largest eigenvalue of $\mathcal{M}_{g'}\mathcal{M}_{g'}^T$. Since $K(L, A, b) = \|\mathcal{M}_{g'}\|_2$ then the condition number $K(L, A, b)$ is also the largest singular value of $\mathcal{M}_{g'}$ i.e $\sqrt{\nu}$. As mentioned in [13, p. 331], the algorithm will converge if the initial y has a component in the direction of the corresponding dominant eigenvector of $\mathcal{M}_{g'}\mathcal{M}_{g'}^T$. When there is an estimate of this dominant eigenvector, the initial y can be set to this estimate but in many implementations, y is initialized as a random vector. The algorithm is terminated by a "sufficiently" large number of iterations or by evaluating the difference between two successive values of ν and comparing it to a tolerance given by the user.

3.3. Closed formula. Using the adjoint formulas obtained in Section 3.2, we now get a closed formula for the total least squares conditioning.

THEOREM 1. *We consider the total least squares problem and assume that the genericity assumption holds. Setting $B_\lambda = A^T A - \lambda_{n+1} I_n$, then the condition number of $L^T x$, linear function of the TLS solution, is expressed by*

$$K(L, A, b) = \|C\|_2^{\frac{1}{2}},$$

where C is the $k \times k$ symmetric matrix

$$C = (1 + \|x\|_2^2) L^T B_\lambda^{-1}\left(A^T A + \lambda_{n+1}(I_n - \frac{2xx^T}{1 + \|x\|_2^2})\right) B_\lambda^{-1} L.$$

<u>Proof:</u> We have $K(L, A, b)^2 = \left\|\mathcal{M}_{g'}^T\right\|_2^2 = \max_{\|y\|_2=1} \left\|\mathcal{M}_{g'}^T y\right\|_2^2$. If y is a unit vector in $\mathbb{R}^k$, then using Equation (3.3) we obtain

$$\left\|\mathcal{M}_{g'}^T y\right\|_2^2 = \left\|\text{vec}\left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}\right)\right\|_2^2 + \left\|D_\lambda^T y\right\|_2^2$$
$$= \|-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}\|_F^2 + \left\|D_\lambda^T y\right\|_2^2$$
$$= \|D_\lambda^T y x^T\|_F^2 + \|r y^T L^T B_\lambda^{-1}\|_F^2 - 2\,\text{trace}(x y^T D_\lambda r y^T L^T B_\lambda^{-1}) + \left\|D_\lambda^T y\right\|_2^2.$$

For all vectors u and v, we have $\|uv^T\|_F = \|u\|_2 \|v\|_2$. Moreover we have

$$\text{trace}\left((x y^T D_\lambda r)(y^T L^T B_\lambda^{-1})\right) = \text{trace}\left((y^T L^T B_\lambda^{-1})(x y^T D_\lambda r)\right) = y^T L^T B_\lambda^{-1} x r^T D_\lambda^T y.$$

Thus

$$\left\|\mathcal{M}_{g'}^T y\right\|_2^2 = \|x\|_2^2 \|D_\lambda^T y\|_2^2 + \|r\|_2^2 \|B_\lambda^{-1} L y\|_2^2 - 2\, y^T L^T B_\lambda^{-1} x r^T D_\lambda^T y + \|D_\lambda^T y\|_2^2$$
$$= (1 + x^T x) y^T D_\lambda D_\lambda^T y + \|r\|_2^2 y^T L^T B_\lambda^{-2} L y - 2\, y^T L^T B_\lambda^{-1} x r^T D_\lambda^T y$$
$$= y^T \left((1 + x^T x) D_\lambda D_\lambda^T + \|r\|_2^2 L^T B_\lambda^{-2} L - 2\, L^T B_\lambda^{-1} x r^T D_\lambda^T\right) y,$$

i.e $\left\|\mathcal{M}_{g'}^T\right\|_2^2 = \|C\|_2$ with

$$C = (1 + x^T x) D_\lambda D_\lambda^T + \|r\|_2^2 L^T B_\lambda^{-2} L - 2\, L^T B_\lambda^{-1} x r^T D_\lambda^T. \tag{3.6}$$

Replacing D_λ by $L^T B_\lambda^{-1}\left(A^T + \frac{2 x r^T}{1 + x^T x}\right)$, Equation (3.6) simplifies to

$$C = L^T B_\lambda^{-1}\left((1 + x^T x) A^T A + \|r\|_2^2 I_n + 2 A^T r x^T\right) B_\lambda^{-1} L. \tag{3.7}$$

But $A^T r x^T = A^T (b - Ax) x^T = A^T b x^T - A^T A x x^T$ and, since from Equation (2.2) we have $A^T b = B_\lambda x$, we get $A^T r x^T = B_\lambda x x^T - A^T A x x^T = (A^T A - \lambda_{n+1} I_n) x x^T - A^T A x x^T = -\lambda_{n+1} x x^T$. From Equation (2.3) we also have $\|r\|_2^2 = \lambda_{n+1}(1 + x^T x)$ and thus Equation (3.7) becomes

$$C = L^T B_\lambda^{-1}\left((1 + x^T x) A^T A + \lambda_{n+1}(1 + x^T x) I_n - 2\lambda_{n+1} x x^T\right) B_\lambda^{-1} L$$

$$= (1 + \|x\|_2^2) L^T B_\lambda^{-1}\left(A^T A + \lambda_{n+1}(I_n - \frac{2 x x^T}{1 + \|x\|_2^2})\right) B_\lambda^{-1} L.$$

$\square$

4. TLS condition number and SVD.

4.1. Closed formula and upper bound. Computing $K(L, A, b)$ using Theorem 1 requires the explicit formation of the normal equations matrix $A^T A$ which is a source of rounding errors and also generates an extra computational cost of about mn^2 flops. In practice the TLS solution is obtained by Equation (2.4) and involves an SVD computation. In the following theorem, we propose a formula for $K(L, A, b)$ that can be computed with quantities that may be already available from the solution process. In the following $0_{n,1}$ (resp. $0_{1,n}$) denotes the zero column (resp. row) vector of length n.

THEOREM 2. *Let V and V' be the matrices whose columns are the right singular vectors of respectively $[A, b]$ and A associated with the singular values $(\sigma_1, \ldots, \sigma_{n+1})$*

and $(\sigma'_1, \ldots, \sigma'_n)$. *Then the condition number of $L^T x$, linear function of the TLS solution is expressed by*

$$K(L, A, b) = (1 + \|x\|_2^2)^{\frac{1}{2}} \left\| L^T V' D' \begin{bmatrix} V'^T, & 0_{n,1} \end{bmatrix} V \begin{bmatrix} D, & 0_{n,1} \end{bmatrix}^T \right\|_2, \text{ where}$$

$D' = diag\left((\sigma_1'^2 - \sigma_{n+1}^2)^{-1}, \ldots, (\sigma_n'^2 - \sigma_{n+1}^2)^{-1}\right)$ *and* $D = diag\left((\sigma_1^2 + \sigma_{n+1}^2)^{\frac{1}{2}}, \ldots, (\sigma_n^2 + \sigma_{n+1}^2)^{\frac{1}{2}}\right)$.
When L is the identity matrix, then the condition number reduces to

$$K(L, A, b) = (1 + \|x\|_2^2)^{\frac{1}{2}} \left\| D' \begin{bmatrix} V'^T, & 0_{n,1} \end{bmatrix} V \begin{bmatrix} D, & 0_{n,1} \end{bmatrix}^T \right\|_2.$$

<u>Proof:</u> From $[A, b] = U\Sigma V^T$, we have $[A, b]^T[A, b] = V\Sigma^2 V^T = \sum_{i=1}^{n+1} \sigma_i^2 v_i v_i^T$ and

$$\begin{aligned}
[A, b]^T[A, b] + \lambda_{n+1} I_{n+1} &= \sum_{i=1}^{n+1} \sigma_i^2 v_i v_i^T + \lambda_{n+1} \sum_{i=1}^{n+1} v_i v_i^T \\
&= \sum_{i=1}^{n+1} (\sigma_i^2 + \lambda_{n+1}) v_i v_i^T \\
&= \sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T + 2\lambda_{n+1} v_{n+1} v_{n+1}^T,
\end{aligned}$$

leading to

$$[A, b]^T[A, b] + \lambda_{n+1} I_{n+1} - 2\lambda_{n+1} v_{n+1} v_{n+1}^T = \sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \qquad (4.1)$$

From Equation (2.4), we have $v_{n+1} = -v_{n+1,n+1} \begin{bmatrix} x \\ -1 \end{bmatrix}$ and, since v_{n+1} is a unit vector, $v_{n+1,n+1}^2 = \frac{1}{1+\|x\|_2^2}$. Then Equation (4.1) can be expressed in matrix notation as

$$\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} + \lambda_{n+1} \begin{bmatrix} I_n & 0_{n,1} \\ 0_{1,n} & 1 \end{bmatrix} - \frac{2\lambda_{n+1}}{1+\|x\|_2^2} \begin{bmatrix} xx^T & -x \\ -x^T & 1 \end{bmatrix} = \sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \qquad (4.2)$$

The quantity $A^T A + \lambda_{n+1}(I_n - \frac{2xx^T}{1+\|x\|_2^2})$ corresponds to the left-hand side of Equation (4.2) in which the last row and the last column have been removed. Thus it can also be written

$$A^T A + \lambda_{n+1}(I_n - \frac{2xx^T}{1+\|x\|_2^2}) = \begin{bmatrix} I_n, & 0_{n,1} \end{bmatrix} \left(\sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \right) \begin{bmatrix} I_n \\ 0_{1,n} \end{bmatrix},$$

and the matrix C from Theorem 1 can be expressed

$$C = (1 + \|x\|_2^2) L^T \begin{bmatrix} B_\lambda^{-1}, & 0_{n,1} \end{bmatrix} \left(\sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \right) \begin{bmatrix} B_\lambda^{-1} \\ 0_{1,n} \end{bmatrix} L. \qquad (4.3)$$

Moreover from $A = U'\Sigma'V'^{T}$, we have $A^{T}A = V'\Sigma'^{2}V'^{T} = \sum_{i=1}^{n} \sigma_i'^{2} v_i' v_i'^{T}$ and

$$B_\lambda = A^{T}A - \lambda_{n+1}I_n$$

$$= \sum_{i=1}^{n} \sigma_i'^{2} v_i' v_i'^{T} - \sigma_{n+1}^2 \sum_{i=1}^{n} v_i' v_i'^{T}$$

$$= \sum_{i=1}^{n} (\sigma_i'^{2} - \sigma_{n+1}^2) v_i' v_i'^{T}$$

$$= V'D'^{-1}V'^{T}.$$

Hence $B_\lambda^{-1} = V'^{-T}D'V'^{-1} = V'D'V'^{T}$ and $\left[B_\lambda^{-1}, \quad 0_{n,1} \right] = V'D' \left[V'^{T}, \quad 0_{n,1} \right]$.
We also have $\sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^{T} = V \begin{bmatrix} D \\ 0_{1,n} \end{bmatrix} \left[D, \quad 0_{n,1} \right] V^{T}$.

Then, by replacing in Equation (4.3), we obtain $C = (1 + \|x\|_2^2)\widetilde{V}\widetilde{V}^{T}$ with $\widetilde{V} = L^{T}V'D' \left[V'^{T}, \quad 0_{n,1} \right] V \left[D, \quad 0_{n,1} \right]^{T}$. As a result, using Theorem 1,

$$K(L, A, b)^2 = \|C\|_2 = (1 + \|x\|_2^2) \left\| \widetilde{V}\widetilde{V}^{T} \right\|_2 = (1 + \|x\|_2^2) \left\| \widetilde{V} \right\|_2^2.$$

When $L = I_n$, we use the fact that V' is an orthogonal matrix and can be removed from the expression of $\left\| \widetilde{V} \right\|_2^2$.

$\square$

In many applications, an upper bound would be sufficient to give an estimate of the conditioning of the TLS solution. The following corollary gives an upper bound for $K(L, A, b)$.

COROLLARY 1. *The condition number of $L^{T}x$, linear function of the TLS solution is bounded by*

$$\bar{K}(L, A, b) = (1 + \|x\|_2^2)^{\frac{1}{2}} \|L\|_2 \frac{(\sigma_1^2 + \sigma_{n+1}^2)^{\frac{1}{2}}}{(\sigma_n'^{2} - \sigma_{n+1}^2)}.$$

Proof: This result comes from the inequality $\|AB\|_2 \le \|A\|_2 \|B\|_2$, followed by $\|D'\|_2 = \max_i(\sigma_i'^{2} - \sigma_{n+1}^2)^{-1} = (\sigma_n'^{2} - \sigma_{n+1}^2)^{-1}$ and $\|D\|_2^2 = \max_i(\sigma_i^2 + \sigma_{n+1}^2) = (\sigma_1^2 + \sigma_{n+1}^2)$.

$\square$

4.2. Numerical examples. In the following examples we study the condition number of x i.e L is here the identity matrix I_n. Then, to simplify the notations, we removed the variable L from the expressions and the condition number of x will be denoted by $K(A, b)$ and its upper bound by $\bar{K}(A, b)$. All the experiments were performed with MATLAB 7.6.0 using a machine precision $2.22 \cdot 10^{-16}$.

4.2.1. First example. In the first example we consider the TLS problem $Ax \approx b$ where $[A, b]$ is defined by

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

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n+1}$ are random unit vectors, $D = diag(n, n-1, \cdots, 1, 1-e_p)$ for a given parameter e_p. The quantity $\sigma_n' - \sigma_{n+1}$ measures the distance of our problem to nongenericity and, due to Equation (2.1), we have in exact arithmetic

$$\sigma_n' - \sigma_{n+1} \le \sigma_n - \sigma_{n+1} = e_p.$$

Then by varying e_p, we can generate different TLS problems and by considering small values of e_p, it is possible to study the behavior of the TLS condition number in the context of close-to-nongeneric problems. The TLS solution x is computed using an SVD of $[A, b]$ and Equation (2.4).

In Table 4.1, we compare the exact condition number $K(A, b)$ given in Theorem 2, the upper bound $\bar{K}(A, b)$ given in Corollary 1, and the upper bound obtained from [17, p. 212] and expressed by

$$\kappa(A, b) = \frac{9\sigma_1 \|x\|_2}{\sigma_n - \sigma_{n+1}} \left(1 + \frac{\|b\|_2}{\sigma_n' - \sigma_{n+1}}\right) \frac{1}{\|b\|_2 - \sigma_{n+1}}.$$

We also report the condition number computed by Algorithm 1, denoted by $K_p(A, b)$, and the corresponding number of power iterations (the algorithm terminates when the difference between two successive values is lower than 10^{-8}). When $\sigma_n' - \sigma_{n+1}$ decreases, the TLS problem becomes worse conditioned and there is a factor $\mathcal{O}(10)$ between the exact condition number $K(A, b)$ and its upper bound $\bar{K}(A, b)$. We also observe that $\bar{K}(A, b)$ is an estimate of better order of magnitude than $\kappa(A, b)$ and that, for small values of $\sigma_n' - \sigma_{n+1}$, $\kappa(A, b)$ is much less reliable. $K_p(A, b)$ is always equal or very close to $K(A, b)$.

TABLE 4.1
TLS conditioning for several values of $\sigma_n' - \sigma_{n+1}$.

$\sigma_n' - \sigma_{n+1}$	$K(A, b)$	$\bar{K}(A, b)$	$\kappa(A, b)$	$K_p(A, b)$	#iter
$9.99976032 \cdot 10^{-1}$	$1.18 \cdot 10^0$	$2.36 \cdot 10^1$	$1.29 \cdot 10^2$	$1.18 \cdot 10^0$	11
$9.99952397 \cdot 10^{-5}$	$8.36 \cdot 10^3$	$1.18 \cdot 10^5$	$1.31 \cdot 10^{10}$	$8.36 \cdot 10^3$	6
$9.99952365 \cdot 10^{-9}$	$8.36 \cdot 10^7$	$1.18 \cdot 10^9$	$1.31 \cdot 10^{18}$	$8.36 \cdot 10^7$	4
$9.99644811 \cdot 10^{-13}$	$8.36 \cdot 10^{11}$	$1.18 \cdot 10^{13}$	$1.31 \cdot 10^{26}$	$8.32 \cdot 10^{11}$	5

4.2.2. Second example. Let us now consider the following example from [17, p. 42] also used in [23] where

$$A = \begin{pmatrix} m-1 & -1 & \cdots & -1 \\ -1 & m-1 & \cdots & -1 \\ . & & & \\ . & & & \\ . & & & \\ -1 & -1 & \cdots & m-1 \\ -1 & -1 & \cdots & -1 \\ -1 & -1 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{m \times (m-2)}, \ b = \begin{pmatrix} -1 \\ -1 \\ . \\ . \\ . \\ -1 \\ m-1 \\ -1 \end{pmatrix} \in \mathbb{R}^m.$$

The exact solution of the TLS problem $Ax \approx b$ can be computed analytically [17, p. 42] and is equal to $x = -(1, \cdots, 1)^T$. We consider a random perturbation $(\Delta A, \Delta b)$ of small norm $\|(\Delta A, \Delta b)\|_F = 10^{-10}$ and we denote by $\tilde{x}$ the computed solution of the perturbed system $(A + \Delta A)x \approx b + \Delta b$.

In Table 4.2, we report for several values of m the *relative* condition number as defined in (2.6) and we compare the computed relative forward error $\frac{\|\tilde{x} - x\|_2}{\|x\|_2}$ with

the forward error bounds that can be expected from the computation of $K^{(rel)}(A, b)$ and its upper bounds $\bar{K}^{(rel)}(A, b)$ and $\kappa^{(rel)}(A, b)$. Since the condition number corresponds to the worst case in error amplification at first order, these quantities are, as observed in Table 4.2, always larger than the computed forward error (there is approximately a factor 10^2 between those quantities). We also observe that, in this example, $\bar{K}^{(rel)}(A, b)$ and $\kappa^{(rel)}(A, b)$ produce forward error estimates that are of same order of magnitude.

TABLE 4.2
Forward error and upper bounds for a perturbed TLS problem.

m	$K^{(rel)}(A, b)$	$\frac{\|\tilde{x}-x\|_2}{\|x\|_2}$	$K^{(rel)}(A, b)\frac{\|(\Delta A, \Delta b)\|_F}{\|(A, b)\|_F}$	$\bar{K}^{(rel)}(A, b)\frac{\|(\Delta A, \Delta b)\|_F}{\|(A, b)\|_F}$	$\kappa^{(rel)}(A, b)\frac{\|(\Delta A, \Delta b)\|_F}{\|(A, b)\|_F}$
50	$5.05 \cdot 10^1$	$2.45 \cdot 10^{-13}$	$2.21 \cdot 10^{-11}$	$1.55 \cdot 10^{-10}$	$6.72 \cdot 10^{-10}$
100	$1.01 \cdot 10^2$	$1.08 \cdot 10^{-13}$	$1.55 \cdot 10^{-11}$	$1.54 \cdot 10^{-10}$	$4.26 \cdot 10^{-10}$
500	$5.01 \cdot 10^2$	$8.79 \cdot 10^{-14}$	$6.85 \cdot 10^{-12}$	$1.53 \cdot 10^{-10}$	$1.66 \cdot 10^{-10}$
1000	$1.00 \cdot 10^3$	$4.33 \cdot 10^{-14}$	$4.84 \cdot 10^{-12}$	$1.53 \cdot 10^{-10}$	$1.13 \cdot 10^{-10}$

5. Conclusion. We proposed sensitivity analysis tools for the total least squares problem when the genericity condition is satisfied. We provided closed formulas for the condition number of a linear function of the TLS solution when the perturbations of data are measured normwise. We also described an algorithm based on an adjoint formula and we expressed this condition number and an upper bound of it in terms of the SVDs of $[A, b]$ and A. We illustrated the use for these quantities in two numerical examples.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide.* Society for Industrial and Applied Mathematics, 3 edition, 1999.
[2] M. Arioli, M. Baboulin, and S. Gratton. A partial condition number for linear least-squares problems. *SIAM J. Matrix Anal. and Appl.*, 29(2):413–433, 2007.
[3] M. Baboulin, J. Dongarra, S. Gratton, and J. Langou. Computing the conditioning of the components of a linear least squares solution. *Numerical Linear Algebra with Applications*, 16(7):517–533, 2009.
[4] M. Baboulin and S. Gratton. Using dual techniques to derive componentwise and mixed condition numbers for a linear function of a linear least squares solution. *BIT Numerical Mathematics*, 49(1):3–19, 2009.
[5] Å. Björck. *Numerical Methods for Least Squares Problems.* Society for Industrial and Applied Mathematics, 1996.
[6] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK Users' Guide.* Society for Industrial and Applied Mathematics, 1997.
[7] S. Chandrasekaran and I. C. F. Ipsen. On the sensitivity of solution components in linear systems of equations. *SIAM J. Matrix Anal. and Appl.*, 16(1):93–112, 1995.
[8] A. J. Cox and N. J. Higham. Accuracy and stability of the null space method for solving the equality constrained least squares problem. *BIT*, 39(1):34–50, 1999.
[9] F. Cucker, H. Diao, and Y. Wei. On mixed and componentwise condition numbers for moore-penrose inverse and linear least squares problems. *Mathematics of Computation*, 76(258):947–963, 2007.

[10] L. Eldén. Perturbation Theory for the Least Squares Problem with Linear Equality Constraints. *SIAM J. Numerical Analysis*, 17:338–350, 1980.

[11] A. J. Geurts. A contribution to the theory of condition. *Numerische Mathematik*, 39:85–96, 1982.

[12] G. H. Golub and C. F. van Loan. An analysis of the Total Least Squares problem. *SIAM J. Numerical Analysis*, 17:883–893, 1980.

[13] G. H. Golub and C. F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996. Third edition.

[14] A. Graham. *Kronecker products and matrix calculus with application*. Wiley, New York, 1981.

[15] S. Gratton. On the condition number of linear least squares problems in a weighted Frobenius norm. *BIT Numerical Mathematics*, 36(3):523–530, 1996.

[16] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2 edition, 2002.

[17] S. Van Huffel and J. Vandewalle. *The total least squares problem. Computational aspects and analysis*. SIAM, 1991.

[18] I. C. F. Ipsen. *Numerical matrix analysis: Linear systems and least squares*. SIAM, 2009.

[19] C. S. Kenney, A. J. Laub, and M. S. Reese. Statistical condition estimation for linear least squares. *SIAM J. Matrix Anal. and Appl.*, 19(4):906–923, 1998.

[20] C. Paige and Z. Strakoš. Core problems in linear algebraic systems. *SIAM J. Matrix Anal. and Appl.*, 27(3):861–875, 2006.

[21] J. Rice. A theory of condition. *SIAM J. Numerical Analysis*, 3:287–310, 1966.

[22] G. W. Stewart and Jiguang Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1991.

[23] Liangmin Zhou, Lijing Lin, Yimin Wei, and Sanzheng Qiao. Perturbation analysis and condition numbers of scaled total least squares problems. *Numerical Algorithms*, 51:381–399, 2009.