Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 09–44

A CONTRIBUTION TO THE CONDITIONING OF THE TOTAL LEAST SQUARES PROBLEM

MARC BABOULIN AND SERGE GRATTON

Abstract: We derive closed formulas for the condition number of a linear function of the total least squares solution. Given an over determined linear systems $Ax = b$, 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 a numerical example, we compare these values with the condition estimate 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

Received November 16, 2009.

originally in [12], is called the Total Least Squares (TLS) problem and can be expressed as:

$$\min_{E,\epsilon} \|(E,\epsilon)\|_F, \ (A+E)x = b+\epsilon, \tag{1}$$

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, 10, 15, 16, 18, 19, 23]). Recently, the more general case of the conditioning of a linear function of an LS solution was studied in [2, 4, 9]. Also one can find in [3] algorithms using the software libraries LAPACK [1] and ScaLAPACK [6] as well as physical applications.

As far as we are aware, there is no closed formula for the conditioning of the TLS problem. In this paper, we propose to derive an exact formula for the condition number of the TLS problem when perturbations of (A, b) are measured using a product norm. To be as general as possible, we consider again here the condition number of $L^T x$, linear function of the TLS solution. 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 L^T by the Jacobian matrix at the solution.

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 $\lambda'_i = \sigma_i'^2$.

We consider the total least squares problem expressed in Equation (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

$$x = \left(A^T A - \lambda_{n+1} I_n\right)^{-1} A^T b. \tag{2}$$

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

$$\lambda_{n+1} = \frac{1}{1+x^T x}\begin{bmatrix} x^T, & -1 \end{bmatrix}\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix}\begin{bmatrix} x \\ -1 \end{bmatrix} = \frac{r^T r}{1+x^T x}. \tag{3}$$

As mentioned [17, p. 35], the TLS solution is obtained by scaling the last 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

$$x = -\frac{1}{v_{n+1,n+1}}[v_{1,n+1}, \ldots, v_{n,n+1}]^T. \tag{4}$$

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 LLS solved via Householder QR factorization).

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 $\|(A,b)\|_F = \sqrt{\|A\|_{\mathrm{F}}^2 + \|b\|_2^2}$ 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 $k \leq n$. We suppose here that L is not perturbed numerically and we consider the mapping

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

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).

Then the condition number as defined in [11, 21] of $L^T x$, linear function of the TLS solution can be expressed as

$$K(L, A, b) = \max_{(\Delta A, \Delta b) \neq 0} \frac{\|g'(A, b).(\Delta A, \Delta b)\|_2}{\|(\Delta A, \Delta b)\|_F}. \tag{5}$$

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 a special where L is the identity matrix.

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 $h'(x)$.

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 dérivative of g at (A, b) is expressed by*

$$\begin{array}{rcl} g'(A, b) \ : \mathbb{R}^{m \times n} \times \mathbb{R}^m & \longrightarrow & \mathbb{R}^k \\ (\Delta A, \Delta b) & \longmapsto & L^T B_\lambda^{-1} \left(A^T + \frac{2 x r^T}{1 + x^T x} \right) (\Delta b - \Delta A x) + \\ & & L^T B_\lambda^{-1} \Delta A^T r. \end{array} \tag{6}$$

Proof: The result is obtained from the chain rule. Since λ_{n+1}, expressed in Equation (3), is a simple eigenvalue of $[A, b]^T [A, b]$ with corresponding unit eigenvector $\frac{1}{\sqrt{1+x^T x}} \begin{bmatrix} x^T & -1 \end{bmatrix}^T$ we know that, up to first order in ΔA and Δb,

λ_{n+1} can be written [22] $\lambda_{n+1} + \Delta\lambda$, where

$$\begin{aligned}
\Delta\lambda &= \frac{1}{1+x^Tx}\begin{bmatrix}x^T & -1\end{bmatrix}\begin{bmatrix}\Delta A^TA + A^T\Delta A & \Delta A^Tb + A^T\Delta b\\ b^T\Delta A + \Delta b^TA & \Delta b^Tb + b^T\Delta b\end{bmatrix}\begin{bmatrix}x\\-1\end{bmatrix}\\
&= \frac{2}{1+x^Tx}\left(x^T\Delta A^TAx - x^T\Delta A^Tb - x^TA^T\Delta b + b^T\Delta b\right)\\
&= \frac{2}{1+x^Tx}\left(-x^T\Delta A^Tr + (b^T - x^TA^T)\Delta b\right)\\
&= \frac{2}{1+x^Tx}\left(-r^T\Delta Ax + r^T\Delta b\right),
\end{aligned}$$

yielding

$$\Delta\lambda = \frac{2r^T(\Delta b - \Delta Ax)}{1+x^Tx}. \tag{7}$$

Considering B_λ^{-1}, using (7) and applying the chain rule, we obtain a first order expansion $B_\lambda^{-1} + \Delta B^{-1}$ with

$$\begin{aligned}
\Delta B^{-1} &= -B_\lambda^{-1}\left(\Delta A^TA + A^T\Delta A - \Delta\lambda I_n\right)B_\lambda^{-1}\\
&= -B_\lambda^{-1}\left(\Delta A^TA + A^T\Delta A - \frac{2r^T(\Delta b - \Delta Ax)}{1+x^Tx}I_n\right)B_\lambda^{-1}.
\end{aligned}$$

The chain rule now applied to $g(A,b)$ leads to $g(A,b) + L^T\delta$, with

$$\begin{aligned}
\delta &= -B_\lambda^{-1}\left(\Delta A^TA + A^T\Delta A - \Delta\lambda I_n\right)B_\lambda^{-1}A^Tb + B_\lambda^{-1}\Delta A^Tb + B_\lambda^{-1}A^T\Delta b\\
&= -B_\lambda^{-1}\left(\Delta A^TA + A^T\Delta A - \Delta\lambda I_n\right)x + B_\lambda^{-1}\left(\Delta A^Tb + A^T\Delta b\right)\\
&= B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^Tx}\right)(\Delta b - \Delta Ax) + B_\lambda^{-1}\Delta A^Tr,
\end{aligned}$$

and left multiplying δ by L^T 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 $D_\lambda = L^T B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^Tx}\right) \in \mathbb{R}^{k\times m}$, we obtain from (6)

$$\begin{aligned} g'(A,b).(\Delta A, \Delta b) &= \operatorname{vec}\left(D_\lambda\left(\Delta b - \Delta A x\right) + L^T B_\lambda^{-1}\Delta A^T r\right) \\ &= \left(-x^T \otimes D_\lambda\right)\operatorname{vec}(\Delta A) + \left(r^T \otimes (L^T B_\lambda^{-1})\right)\operatorname{vec}(\Delta A^T) + D_\lambda \Delta b \\ &= \left[-x^T \otimes D_\lambda + \left(r^T \otimes (L^T B_\lambda^{-1})\right) P, \;\; D_\lambda\right] \begin{bmatrix} \operatorname{vec}(\Delta A) \\ \Delta b \end{bmatrix}. \end{aligned}$$

Then we get

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

But we have $\|(\Delta A, \Delta b)\|_F = \left\| \begin{bmatrix} \operatorname{vec}(\Delta A) \\ \Delta b \end{bmatrix} \right\|_2$ and then, from Proposition 1 and using the definition of $K(L, A, b)$ given in Expression (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

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

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

$$\mathcal{M}_{g'}^T y = \begin{bmatrix} \mathrm{vec}\left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}\right) \\ D_\lambda^T y \end{bmatrix}. \tag{8}$$

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 $y^T y$ respectively on $\mathbb{R}^{m\times n} \times \mathbb{R}^m$ and $\mathbb{R}^k$ is*

$$\begin{aligned}
g'^*(A,b) \ : \mathbb{R}^k &\longrightarrow \mathbb{R}^{m\times n} \times \mathbb{R}^m \\
y &\longmapsto \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}, D_\lambda^T y\right)
\end{aligned} \tag{9}$$

In addition, if $k = 1$ we have

$$K(L,A,b) = \sqrt{\|-D_\lambda^T x^T + r L^T B_\lambda^{-1}\|_F^2 + \|D_\lambda\|_2^2} \tag{10}$$

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

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

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

□

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 (10), $L^T B_\lambda^{-1}$ 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 (6) and (9), 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
: $y = (1,\ldots,1)^T$
: **repeat**
: $(A_n, b_n) = \left(-D_\lambda^T y x^T + r y^T L^T B_\lambda^{-1}, D_\lambda^T y\right)$
: $\nu = \|(A_n, b_n)\|_F$
: $(A_n, b_n) \leftarrow (\frac{1}{\nu}\cdot A_n, \frac{1}{\nu}\cdot b_n)$
: $y = L^T B_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^T x}\right)(b_n - A_n x) + L^T B_\lambda^{-1} A_n^T r$
: **end**
: $K(L,A,b) = \sqrt{\nu}$

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* $B_\lambda = \left(A^TA - \lambda_{n+1}I_n\right)$*, then the condition number of* L^Tx*, 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*

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

Proof: We have $K(L, A, b)^2 = \left\|\mathcal{M}_{g'}^T\right\|_2^2 = \max_{\|y\|_2=1}\left\|\mathcal{M}_{g'}^Ty\right\|_2^2$. If y is a unit vector in $\mathbb{R}^k$, then using Equation (8) we obtain

$$\begin{aligned}\left\|\mathcal{M}_{g'}^Ty\right\|_2^2 &= \left\|\mathrm{vec}\left(-D_\lambda^Tyx^T + ry^TL^TB_\lambda^{-1}\right)\right\|_2^2 + + \left\|D_\lambda^Ty\right\|_2^2\\ &= \|-D_\lambda^Tyx^T + ry^TL^TB_\lambda^{-1}\|_F^2 + \left\|D_\lambda^Ty\right\|_2^2\\ &= \|D_\lambda^Tyx^T\|_F^2 + \|ry^TL^TB_\lambda^{-1}\|_F^2 - 2\,\mathrm{trace}(xy^TD_\lambda ry^TL^TB_\lambda^{-1}) + \left\|D_\lambda^Ty\right\|_2^2.\end{aligned}$$

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

$$\mathrm{trace}\left((xy^TD_\lambda r)(y^TL^TB_\lambda^{-1})\right) = \mathrm{trace}\left((y^TL^TB_\lambda^{-1})(xy^TD_\lambda r)\right) = y^TL^TB_\lambda^{-1}xr^TD_\lambda^Ty.$$

Thus

$$\begin{aligned}\left\|\mathcal{M}_{g'}^Ty\right\|_2^2 &= \|x\|_2^2\left\|D_\lambda^Ty\right\|_2^2 + \|r\|_2^2\left\|B_\lambda^{-1}Ly\right\|_2^2 - 2\,y^TL^TB_\lambda^{-1}xr^TD_\lambda^Ty + \left\|D_\lambda^Ty\right\|_2^2\\ &= (1+x^Tx)y^TD_\lambda D_\lambda^Ty + \|r\|_2^2\,y^TL^TB_\lambda^{-2}Ly - 2\,y^TL^TB_\lambda^{-1}xr^TD_\lambda^Ty\\ &= y^T\left((1+x^Tx)D_\lambda D_\lambda^T + \|r\|_2^2\,L^TB_\lambda^{-2}L - 2\,L^TB_\lambda^{-1}xr^TD_\lambda^T\right)y,\end{aligned}$$

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

$$C = (1+x^Tx)D_\lambda D_\lambda^T + \|r\|_2^2\,L^TB_\lambda^{-2}L - 2\,L^TB_\lambda^{-1}xr^TD_\lambda^T. \tag{11}$$

Replacing D_λ by $L^TB_\lambda^{-1}\left(A^T + \frac{2xr^T}{1+x^Tx}\right)$, Equation (11) simplifies to

$$C = L^TB_\lambda^{-1}\left((1+x^Tx)A^TA + \|r\|_2^2\,I_n + 2A^Trx^T\right)B_\lambda^{-1}L. \tag{12}$$

But $A^Trx^T = A^T(b-Ax)x^T = A^Tbx^T - A^TAxx^T$ and, since from Equation (2) we have $A^Tb = B_\lambda x$, we get $A^Trx^T = B_\lambda xx^T - A^TAxx^T = (A^TA -$

$\lambda_{n+1}I_n)xx^T - A^TAxx^T = -\lambda_{n+1}xx^T$. From Equation (3) we also have $\|r\|_2^2 = \lambda_{n+1}(1+x^Tx)$ and thus Equation (12) becomes

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

□

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^TA 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 (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^Tx, linear function of the TLS solution is expressed by*

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

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

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

Proof: 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

$$[A,b]^T[A,b] + \lambda_{n+1} I_{n+1} - 2\lambda_{n+1} v_{n+1} v_{n+1}^T = \sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \tag{13}$$

From Equation (4), we have $v_{n+1} = -v_{n+1,n+1} \begin{bmatrix} x \\ -1 \end{bmatrix}$ and, since v is a unit vector, $v_{n+1,n+1}^2 = \frac{1}{1+\|x\|_2^2}$. Then Equation (13) can be expressed in matrix notation as

$$\begin{bmatrix} A^TA & A^Tb \\ b^TA & b^Tb \end{bmatrix} + \lambda_{n+1} \begin{bmatrix} I_n & 0_{n,1} \\ 0_{1,n} & 1 \end{bmatrix} - \frac{2\lambda_{n+1}}{1+\|x\|_2^2} \begin{bmatrix} xx^T & -x \\ -x^T & 1 \end{bmatrix} = \sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \tag{14}$$

The quantity $A^TA + \lambda_{n+1}(I_n - \frac{2xx^T}{1+\|x\|_2^2})$ corresponds to the left-hand side of Equation (14) in which the last row and the last column have been removed. Thus it can also be written

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

and the matrix C from Theorem 1 can be expressed

$$C = (1+\|x\|_2^2) L^T \begin{bmatrix} B_\lambda^{-1}, & 0_{n,1} \end{bmatrix} \left(\sum_{i=1}^{n} (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T \right) \begin{bmatrix} B_\lambda^{-1} \\ 0_{1,n} \end{bmatrix} L. \tag{15}$$

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

$$\begin{aligned} B_\lambda &= A^T A - \lambda_{n+1} I_n \\ &= \sum_{i=1}^n \sigma_i'^2 v_i' v_i'^T - \sigma_{n+1}^2 \sum_{i=1}^n v_i' v_i'^T \\ &= \sum_{i=1}^n (\sigma_i'^2 - \sigma_{n+1}^2) v_i' v_i'^T \\ &= V' D'^{-1} V'^T. \end{aligned}$$

Hence $B_\lambda^{-1} = V'^{-T} D' V'^{-1} = V' D' V'^T$ and $\left[B_\lambda^{-1},\ 0_{n,1}\right] = V' D' \left[V'^T,\ 0_{n,1}\right]$. We also have $\sum_{i=1}^n (\sigma_i^2 + \sigma_{n+1}^2) v_i v_i^T = V \begin{bmatrix} D \\ 0_{1,n} \end{bmatrix} \left[D,\ 0_{n,1}\right] V^T$.
Then, by replacing in Equation (15), we obtain $C = (1 + \|x\|_2^2) \widetilde{V} \widetilde{V}^T$ with $\widetilde{V} = L^T V' D' \left[V'^T,\ 0_{n,1}\right] V \left[D,\ 0_{n,1}\right]^T$. 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 $\left\| \widetilde{V} \right\|_2^2$.

□

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 $\|AB\|_2 \leq \|A\|_2 \|B\|_2$, followed by $\|D'\|_2 = \max_i (\sigma_i'^2 - \sigma_{n+1}^2)^{-1} = (\sigma_n'^2 - \sigma_{n+1}^2)^{-1}$ and $\|D\|_2^2 = \max_i (\sigma_i^2 + \sigma_{n+1}^2) = (\sigma_1^2 + \sigma_{n+1}^2)$.

□

4.2. Numerical example. We consider $A \in \mathbb{R}^{30\times 10}$ and $b \in \mathbb{R}^{30}$ whose values are random numbers (uniform distribution). x is the TLS solution computed with Matlab (machine precision $2.22 \cdot 10^{-16}$) using an SVD of $[A, b]$ and Equation (4). We study here the condition number of x (i.e L is the identity matrix).

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

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

As observed in Table 1, $\bar{K}(L, A, b)$ is an estimate of better order of magnitude than $\kappa(A, b)$.

TABLE 1. Exact value and estimates for the condition number of the TLS solution.

Computed quantity	$K(L, A, b)$	$\bar{K}(L, A, b)$	$\kappa(A, b)$
Obtained value	$6.22 \cdot 10^0$	$5.97 \cdot 10^1$	$4.21 \cdot 10^3$

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.

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. *Eigensystems, Matrix Algorithms, vol. 2*. SIAM, 2001.
[23] G. W. Stewart and Jiguang Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1991.

Marc Baboulin
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
E-mail address: baboulin@mat.uc.pt

Serge Gratton
CERFACS, Toulouse, France
E-mail address: gratton@cerfacs.fr