AN ANALYSIS OF THE TOTAL LEAST SQUARES PROBLEM

by

Gene H. Golub[+]

and

Charles F. Van Loan[++]

TR 80-411

Department of Computer Science
Cornell University
Ithaca, New York 14853

AN ANALYSIS OF THE TOTAL LEAST SQUARES PROBLEM

by

Gene H. Golub[+]
and
Charles F. Van Loan[++]

Abstract

Total Least Squares (TLS) is a method of fitting that is appropriate when there are errors in both the observation vector b (mxl) and in the data matrix A (mxn) . The technique has been discussed by several authors and amounts to fitting a "best" subspace to the points (a_i^T, b_i) , i=1,...,m , where a_i^T is the i-th row of A . In this paper a singular value decomposition analysis of the TLS problem is presented. The sensitivity of the TLS problem as well as its relationship to ordinary least squares regression is explored. An algorithm for solving the TLS problem is proposed that utilizes the singular value decomposition and which provides a measure of the underlying problem's sensitivity.

[+]Department of Computer Science, Stanford University, Stanford California, 94305

[++]Department of Computer Science, Cornell University, Ithaca, New York, 14853

The first author wishes to acknowledge the support of DOE contract DE-AS03-76SF0032(and United States Army Research Grant DAAG 29-78-G-0179.

1. Introduction

In the least squares (LS) problem we are given an mxn "data matrix" A, a "vector of observations" b having m components, a nonsingular diagonal matrix $D = \mathrm{diag}(d_1,\ldots,d_m)$, and are asked to find a vector x such that

$$\| D(b - Ax) \|_2 = \min . \tag{1.1}$$

Here, $\| \cdot \|_2$ denotes Euclidean length. It is well-known that any solution to the LS problem satisfies the following system of "normal equations":

$$A^T D^2 A x = A^T D^2 b . \tag{1.2}$$

The solution is unique if $\mathrm{rank}(A) = n$. However, regardless of the rank of A there is always a unique minimal 2-norm solution to the LS problem given by

$$x_{LS} = (DA)^+ D b \tag{1.3}$$

where $(DA)^+$ denotes the Moore-Penrose Pseudo-inverse of DA .

In the (classical) LS problem there is an underlying assumption that all the errors are confined to the observation vector b . Unfortunately, this assumption is frequently unrealistic; sampling errors, human errors, modelling errors, and instrument errors may preclude the possibility of knowing the data matrix A exactly. Methods for estimating the effect of such errors on x_{LS} are given in Hodges and Moore [11] and Stewart [19]. The representation of data errors in a statistically meaningful way is a difficult task that can be appreciated by reading the survey article by Cochrane [2].

In this paper we analyze the method of total least squares (TLS), which is one of several fitting techniques that has been devised to compensate for data errors. A good way to motivate the method is to recast the ordinary LS problem as follows:

$$\underset{r}{\text{minimize}} \ \|Dr\|_2$$

$$\text{subject to} \quad b+r \in \text{Range}(A)$$

If $\|Dr\|_2 = \min$ and $b + r = Ax$, then x solves the LS problem (1.1). Thus the LS problem amounts to perturbing the observation b by a minimum amount r so that $b+r$ can be "predicted" by the columns of A.

Now simply put, the idea behind total least squares is to consider perturbations of both b and A. More precisely, given the nonsingular weighting matrices

$$D = \text{diag}(d_1, \ldots, d_m) \qquad d_i > 0 \ , \ i = 1, \ldots, m$$

$$T = \text{diag}(t_1, \ldots, t_{n+1}) \qquad t_i > 0 \ , \ i = 1, \ldots, n+1$$

we seek to

$$\underset{E,r}{\text{minimize}} \ \| D[E|r]T \|_F \tag{1.4}$$

$$\text{subject to} \quad b+r \in \text{Range}(A + E)$$

Here, $\| \cdot \|_F$ denotes the Frobenius norm, viz $\| B \|_F^2 = \sum_i \sum_j |b_{ij}|^2$. Once a minimizing $[\tilde{E}|\tilde{r}]$ is found, then any x satisfying

$$(A+\tilde{E})\,x = b+\tilde{r}$$

is said to solve the TLS problem (1.4). Thus, the TLS problem is equivalent to the problem of solving a nearest compatible LS problem $\min \| (A+\tilde{E})x - (b+\tilde{r}) \|_2$ where "nearness" is measured by the weighted Frobenius norm above.

Total least squares is not a new method of fitting; the $n=1$ case has been scrutinized since the turn of the century. More recently, the method has been discussed in the context of the subset selection problem, see [9], [10], and [20]. In Deming [3] and Gerhold [4] the following more general problem is analyzed:

$$\text{(1.5)} \qquad \begin{aligned} &\underset{E,r}{\text{minimize}} \ \sum_{i=1}^{m} \{ \Delta_i r_i^2 + \sum_{j=1}^{n} \omega_{ij} e_{ij}^2 \} \\ &\text{subject to } b + r \in \text{Range}(A+E) \end{aligned}$$

where $E = (e_{ij})$, $r^T = (r_1,\ldots,r_m)$, and the Δ_i and ω_{ij} are given positive weights.

The TLS approach to fitting has also attracted interest outside of statistics. For example, many algorithms for nonlinearly constrained minimization require estimates of the vector of Lagrange multipliers. This typically involves the solution of an LS problem where the matrix is the Jacobian of the "active contraints." Because of uncertainties in this matrix, Gill and Murray[5] have suggested using total least squares. Similar in spirit is the work of Barrera and Dennis [1] who have developed a "fuzzy Broyden" method for systems of nonlinear equations.

In the present paper we analyze the TLS problem by making heavy use of the singular value decomposition (SVD). As is pointed out in Golub and Reinsch [7] and more fully in Golub [6], this decomposition can be used to solve the TLS problem. We indicate how this can be accomplished in §2. An interesting aspect of the TLS problem is that it may fail to have a solution. For example, if

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad , \quad D = T = I_2$$

then for every $\varepsilon > 0$, $b \in \text{Range}(A + E_\varepsilon)$ where $E_\varepsilon = \text{diag}(0,\varepsilon)$. Thus, there is no "smallest" $\| [\, E \mid r \,] \|_F$ for which $b + r \in \text{Range}(A+E)$ since $b \notin \text{Range}(A)$. This kind of pathological situation raises several important questions. Under what set of circumstances does the TLS problem lack a solution? More generally, what constitutes an ill-conditioned TLS problem? Answers to these and other related theoretical questions of practical importance are offered in §3 and §4. In §5 some algorithmic considerations are briefly mentioned.

2. The TLS Problem and the Singular Value Decomposition

If $b+r$ is in the range of $A+E$, then there is a vector $x \in R^n$ such that

$$(A + E)x = b + r$$

i.e.,

$$\{ D[A|b]T + D[E|r]T \} T^{-1} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0 \quad . \tag{2.1}$$

This equation shows that the TLS problem involves finding a perturbation matrix $\Delta \in R^{mx(n+1)}$ having minimal norm such that $C + \Delta$ is rank deficient where

$$C = D[A|b]T \, . \tag{2.2}$$

The singular value decomposition can be used for this purpose. Let

$$U^T C V = \text{diag}(\sigma_1 , \ldots , \sigma_{n+1})$$

$$U = [u_1 , \ldots , u_m] \qquad V = [v_1 , \ldots , v_{n+1}] \qquad u_i \in R^m , \; v_i \in R^{n+1} \tag{2.3}$$

$$\sigma_1 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_{n+1}$$

be the SVD of C with $U^TU = I_m$ and $V^TV = I_n$. A discussion of this decomposition and its elementary properties may be found in Stewart[17]. In particular, it can be shown that

$$\sigma_{n+1} = \min_{\text{rank}(C+\Delta) < n+1} \| \Delta \|_F \quad . \tag{2.4}$$

Moreover, the minimum is attained by setting $\Delta = -Cvv^T$ where v is any unit vector in the subspace S_C defined by

$$S_C = \text{span} \{ v_{k+1} , \ldots , v_{n+1} \} \quad . \tag{2.5}$$

Suppose we can find a vector v in S_C having the following form:

$$v = \begin{bmatrix} y \\ \alpha \end{bmatrix} \qquad y \in R^n \ , \ \alpha \neq 0 \ ,$$

If

$$x = \frac{-1}{\alpha\, t_{n+1}} T_1 y \qquad T_1 = \mathrm{diag}(t_1, \ldots, t_n) \tag{2.6}$$

and we define $\tilde{E}$ and $\tilde{r}$ by

$$D[\tilde{E}|\tilde{r}]T = -Cvv^T \ ,$$

then

$$\{ D[A|b]T + D[\tilde{E}|\tilde{r}]T \}\, T^{-1} \begin{bmatrix} x \\ -1 \end{bmatrix} = C\,(I - vv^T)\left(\frac{-v}{\alpha t_{n+1}}\right) = 0$$

In light of the remarks made after (2.1), it follows that x solves the TLS problem

If $e_{n+1} = (0,\ldots,0,1)^T$ is orthogonal to S_C, then the TLS problem has no solution. On the other hand, if σ_{n+1} is a repeated singular value of C, then the TLS problem may lack a unique solution. However, whenever this is the case it is possible to single out a unique "minimum norm" TLS solution which we denote by x_{TLS} In particular, let Q be an orthogonal matrix of order $n-k+1$ with the property tha

$$[\, v_{k+1}, \ldots, v_{n+1}\,]\, Q = \begin{bmatrix} W & y \\ 0 & \alpha \end{bmatrix} \begin{matrix} n-k \\ 1 \end{matrix} \qquad \begin{matrix} \\ n-k \end{matrix} \quad \begin{matrix} \\ 1 \end{matrix} \ . \tag{2.7}$$

If we set $x_{TLS} = -T_1 y/(\alpha t_{n+1})$ and if we define the τ-norm by

$$\| w \|_\tau = \| T_1^{-1} w \| \qquad w \in R^n \tag{2.8}$$

then it is easy to show that $\| x_{TLS} \|_\tau < \| x \|_\tau$ for all other solutions x to the TLS problem (1.4).

3. A Geometric Interpretation of the TLS Problem

If the SVD of $C = D[A|b]T$ is given by (2.3), then it is easy to verify that

$$\frac{\| D[A|b]T\, v \|_2}{\| v \|_2} \geq \sigma_{n+1} \qquad v \neq 0$$

and that equality holds for nonzero v if and only if v is in the subspace S_C defined by (2.5). Combining this fact with (2.6), we see that the TLS problem amounts to finding an $x \in R^n$ (if possible) such that

$$\frac{\| D[A|b]T\, T^{-1}\begin{bmatrix} x \\ -1 \end{bmatrix} \|_2}{\| T^{-1}\begin{bmatrix} x \\ -1 \end{bmatrix} \|_2} = \sigma_{n+1} .$$

The geometry of the TLS problem comes to light when we write

$$(3.1) \qquad \frac{\| D[A|b]T\, T^{-1}\begin{bmatrix} x \\ -1 \end{bmatrix} \|_2^2}{\| T^{-1}\begin{bmatrix} x \\ -1 \end{bmatrix} \|_2^2} = \sum_{i=1}^{m} d_i^2 \, \frac{| a_i^T x - b_i |^2}{x^T T_1^{-2} x + t_{n+1}^{-2}}$$

where $a_i^T = (a_{i1}, \ldots, a_{in})$, the i-th row of A. The quantity

$$\frac{| a_i^T x - b_i |^2}{x^T T_1^{-2} x + t_{n+1}^{-2}}$$

is the square of the distance from $\begin{bmatrix} a_i \\ b_i \end{bmatrix} \in R^{n+1}$ to the nearest point in the subspace P_x defined by

$$P_x = \{ \begin{bmatrix} a \\ b \end{bmatrix} \mid a \in R^n , \; b \in R , \; b = x^T a \} .$$

Here, the "distance" between two points u and v in R^{n+1} is given by $\| T(u - v) \|_2$.

Thus, the TLS problem is tantamount to finding a "closest" subspace P_x to the $(n+1)$-tuples $\begin{bmatrix} a_i \\ b_i \end{bmatrix}$, $i=1,\ldots,m$. The simple case when $n=1$ and D and T are both identities is worth illustrating. In Figure 1 the LS and the TLS measures of goodness-of-fit are depicted. In the LS problem it is the vertical distances that are important while in the TLS problem it is the perpendicular distances that are critical. (When $T \neq I$, these perpendiculars are "skewed".) To say that the

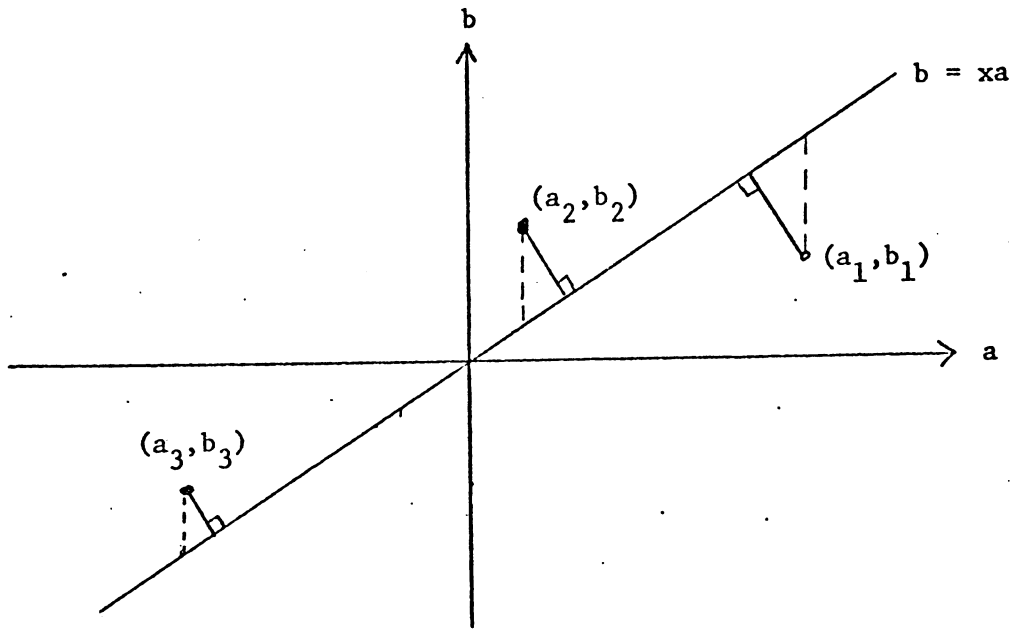

Figure 1. Least Squares vs. Total Least Squares

TLS problem has no solution in the $n=1$ case is to say the the TLS fitting line is vertical. This would be the case, for example, if the three data points in Figur 1 are (1,8), (2,-2), and (4,-1) for then the line $a = 7/3$ is closest to the data in the sense of minimizing the sum of the squared perpendicular distances.

The fitting of straight lines when both variables are subject to error has received a lot of attention in the statistics literature. We refer the interested reader to the papers by Pearson [15], Madansky [14], Riggs et al [16], and York [22], as well as Chapter 13 of Linnik [13].

4. The Sensitivity of the TLS Problem

In this section we establish some inequalities that shed light on the sensitivity of the TLS problem as well as on the relationship between x_{LS} and x_{TLS}. The starting point in the analysis is to formulate the TLS problem as an eigenvalue problem. Recall the definitions of the matrix C and the subspace S_C in §2. It is easy to show that the "singular vectors" v_i in (2.3) are eigenvectors of C^TC and that in particular, S_C is the invariant subspace associated σ_{n+1}^2, the smallest eigenvalue of this matrix. Thus, if $x \in R^n$ is such that

$$C^TC \begin{bmatrix} T^{-1}x \\ -1 \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} T^{-1}x \\ -1 \end{bmatrix} \tag{4.1}$$

then x solves the TLS problem. With the definitions

$$\hat{A} = D A T_1^{-1} \quad , \quad \hat{b} = Db \quad , \quad \lambda = t_{n+1} \tag{4.2}$$

equation (4.1) is readily seen to have the following block structure:

$$\begin{bmatrix} \hat{A}^T\hat{A} & \lambda\hat{A}^T\hat{b} \\ \lambda\hat{b}^T\hat{A} & \lambda^2\hat{b}^T\hat{b} \end{bmatrix} \begin{bmatrix} T_1^{-1}x \\ -\lambda^{-1} \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} T_1^{-1}x \\ -\lambda^{-1} \end{bmatrix} . \tag{4.3}$$

Moreover, if

$$\hat{U}^T \hat{A} \hat{V} = \hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n) \qquad \begin{matrix} \hat{U}^T\hat{U} = I_m \, , \, \hat{V}^T\hat{V} = I_n \\ \hat{\sigma}_1 \geqslant \hat{\sigma}_2 \geqslant \ldots \geqslant \hat{\sigma}_n \geqslant 0 \end{matrix} \tag{4.4}$$

is the SVD of $\hat{A}$, and if we define

$$K = \hat{\Sigma}^T\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2) \quad , \quad g = \hat{\Sigma}^T \hat{U}^T b \quad , \quad h^2 = \hat{b}^T\hat{b} \quad , \quad z = \hat{V}^T T_1^{-1} x \quad , \tag{4.5}$$

then (4.3) transforms to

$$\begin{bmatrix} K & \lambda g \\ \lambda g^T & \lambda^2 h^2 \end{bmatrix} \begin{bmatrix} z \\ -\lambda^{-1} \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} z \\ -\lambda^{-1} \end{bmatrix} \tag{4.6}$$

From this equation we see that

$$(K - \sigma_{n+1}^2 I)z = g \tag{4.7}$$

and

$$\frac{\sigma_{n+1}^2}{\lambda^2} + g^T z = h^2 \ . \tag{4.8}$$

With these reductions, we now obtain some useful characterizations of both x_{TLS} and σ_{n+1}. In order for the subsequent analysis to be uncluttered, we freely make use of the notation established in (2.2)-(2.8) and (4.2)-(4.5).

Theorem 4.1

If $\hat{\sigma}_n > \sigma_{n+1}$ then x_{TLS} exists and is the only solution to the TLS problem. Moreover,

$$x_{TLS} = T_1(\hat{A}^T\hat{A} - \sigma_{n+1}^2 I)^{-1} \hat{A}^T\hat{b} \tag{4.9}$$

and

$$\sigma_{n+1}^2 \left[\frac{1}{\lambda^2} + \sum_{i=1}^{n} \frac{c_i^2}{\hat{\sigma}_i^2 - \sigma_{n+1}^2} \right] = \rho_{LS}^2 \tag{4.10}$$

where

$$c = (c_1, \ldots, c_m)^T = \hat{U}^T \hat{b} \tag{4.11}$$

$$\rho_{LS}^2 = \min_x \| D(b - Ax) \|_2^2 = \| D(b - Ax_{LS}) \|_2^2 \ . \tag{4.12}$$

Proof

The separation theorem [21,p.103] for eigenvalues of symmetric matrices implies that

$$\sigma_1 \geq \hat{\sigma}_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq \hat{\sigma}_n \geq \sigma_{n+1} \quad . \tag{4.13}$$

The assumption $\hat{\sigma}_n > \sigma_{n+1}$ thus insures that σ_{n+1} is not a repeated singular value of C. If $C^TC \begin{bmatrix} y \\ 0 \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} y \\ 0 \end{bmatrix}$ and $0 \neq y \in R^n$, then it clearly follows that $\hat{A}^T\hat{A}y = \sigma_{n+1}^2 y$, a contradiction since $\hat{\sigma}_n^2$ is the smallest eigenvalue of $\hat{A}^T\hat{A}$. Thus, S_C must contain a vector whose (n+1)st component is nonzero. This implies that TLS problem has a solution. Since S_C has dimension 1, this solution is unique. The formula (4.9) follows directly from the "top half" of (4.3).

To establish (4.10) we observe from (4.7) and (4.8) that

$$\frac{\sigma_{n+1}^2}{\lambda^2} + g^T(K - \sigma_{n+1}^2 I)^{-1} g = h^2 \quad .$$

By using the definitions (4.5) and (4.11) this can be rewritten as

$$\frac{\sigma_{n+1}^2}{\lambda^2} + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2 c_i^2}{\hat{\sigma}_i^2 - \sigma_{n+1}^2} = \sum_{i=1}^{m} c_i^2$$

or

$$\sigma_{n+1}^2 \left[\frac{1}{\lambda^2} + \sum_{i=1}^{n} \frac{c_i^2}{\hat{\sigma}_i^2 - \sigma_{n+1}^2} \right] = \sum_{i=n+1}^{m} c_i^2 \quad .$$

Inequality (4.10) now follows since

$$\min_x \| D(b - Ax) \|_2^2 = \min_y \| \hat{b} - \hat{A}y \|_2^2 = \min_w \| c - \hat{S} w \|_2^2 = \sum_{i=n+1}^{m} c_i^2 \quad \blacksquare$$

We shall make use of (4.10) in the next section. The characterization (4.9) points out an interesting connection between total least squares and ridge regression. Ridge regression is a way of "regularizing" the solution to an ill-conditioned LS problem. (See [12,p.190ff.].) Consider, for example, the minimization of

$$\phi(x;\mu) = \| D(b - Ax) \|_2^2 + \mu \| T_1^{-1} x \|_2^2$$

where μ is a positive scalar. It is easy to show that

$$x_{LS}(\mu) = T_1 (\hat{A}^T \hat{A} + \mu I)^{-1} \hat{A}^T \hat{b}$$

solves this problem and that $\| T_1^{-1} x_{LS}(\mu) \|_2 = \| x_{LS}(\mu) \|_\tau$ becomes small as μ becomes large. This is the key to ridge regression; by controlling μ we can control the τ - norm of $x_{LS}(\mu)$.

What is particularly interesting, however, is that $x_{TLS} = x_{LS}(-\sigma_{n+1}^2)$. That is, total least squares is a deregularizing procedure, a kind of "reverse" ridge regression. As we shall see, this implies that the condition of the TLS problem is always worse than the condition of the corresponding LS problem. For this reason it is interesting to compare the LS and TLS fits with one another.

Corollary 4.2

Let $\rho_{LS} = \| D(b - Ax_{LS}) \|_2$. If $\hat{\sigma}_n > \sigma_{n+1}$ then

$$\| x_{TLS} - x_{LS} \|_\tau \leq \frac{\lambda \| \hat{b} \|_2 \rho_{LS}}{\hat{\sigma}_n^2 - \sigma_{n+1}^2} \tag{4.14}$$

and

$$\| D(b - x_{TLS}) \|_2 \leq \rho_{LS} \left[1 + \frac{\lambda \| \hat{b} \|_2}{\hat{\sigma}_n - \sigma_{n+1}} \right] \tag{4.15}$$

Proof

Using (1.2) it is clear that $x_{LS} = T_1(\hat{A}^T\hat{A})^{-1}\hat{A}^T\hat{b}$ and so from (4.9) we have

$$x_{TLS} - x_{LS} = T_1[(\hat{A}^T\hat{A} - \sigma_{n+1}^2 I)^{-1} - (\hat{A}^T\hat{A})^{-1}]\hat{A}^T\hat{b}$$

(4.16)

$$= \sigma_{n+1}^2 T_1(\hat{A}^T\hat{A} - \sigma_{n+1}^2 I)^{-1} T_1^{-1} x_{LS}$$

Applying T_1^{-1} to both sides of this equation and taking norms gives

$$\| x_{TLS} - x_{LS} \|_\tau \leq \frac{\sigma_{n+1}^2 \| x_{LS} \|_\tau}{\hat{\sigma}_n^2 - \sigma_{n+1}^2}$$

This result coupled with the inequalities

(4.17) $$\rho_{LS} = \| D(Ax_{LS} - b) \|_2 = \| D[A|b]T\, T^{-1} \begin{bmatrix} x_{LS} \\ -1 \end{bmatrix} \|_2 \geq \sigma_{n+1} \| x_{LS} \|_\tau$$

(4.18) $$\lambda \| \hat{b} \|_2 = \| D[A|b]T\, e_{n+1} \| \geq \sigma_{n+1} \qquad (e_{n+1}^T = (0,\ldots,0,1))$$

establish (4.14).

To prove (4.15), note that

(4.19) $$\| D(b - Ax_{TLS}) \|_2 \leq \rho_{LS} + \| DA(x_{TLS} - x_{LS}) \|_2 .$$

Now by (4.16),

$$DA(x_{TLS} - x_{LS}) = \sigma_{n+1}^2 \hat{A}(\hat{A}^T\hat{A} - \sigma_{n+1}^2 I)^{-1} T_1^{-1} x_{LS}$$

and so by invoking (4.17) and (4.18) we find

$$\| DA(x_{TLS} - x_{LS}) \|_2 \leq \rho_{LS} \lambda \| \hat{b} \|_2 \| \hat{A}(\hat{A}^T\hat{A} - \sigma_{n+1}^2)^{-1} \|_2$$

$$= \rho_{LS} \lambda \| \hat{b} \|_2 \max_{1 \leq k \leq n} \frac{\hat{\sigma}_k}{\hat{\sigma}_k + \sigma_{n+1}} \cdot \frac{1}{\hat{\sigma}_k - \sigma_{n+1}}$$

$$\leq \rho_{LS} \lambda \| \hat{b} \|_2 / (\hat{\sigma}_n - \sigma_{n+1})$$

Inequality (4.15) follows by substituting this result into (4.19). □

The corollary shows that $x_{TLS} \to x_{LS}$ as $\lambda \to 0$. Thus, by reducing the "observation weight" $\lambda = t_{n+1}$, the TLS problem "converges" to the LS problem. Of course, if $\rho_{LS} = 0$ and A has full rank, then $x_{TLS} = x_{LS}$ regardless of λ.

The bounds in (4.14) and (4.15) are large whenever σ_{n+1} is close to $\hat{\sigma}_n$. (This occurs, for example, whenever σ_{n+1} is a nearly repeated singular value.) Our next results indicate the extent to which $(\hat{\sigma}_n - \sigma_{n+1})^{-1}$ measures the sensitivity of the TLS problem.

Lemma 4.3

If $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_m]$ is a column partitioning of the matrix $\hat{U}$ in the SVD (4.4) and if $\hat{\sigma}_n > \sigma_{n+1}$, then

$$\frac{|\hat{u}_n^T \hat{b}|}{2(\hat{\sigma}_n - \sigma_{n+1})} \leq \| x_{TLS} \|_\tau \leq \frac{\| \hat{b} \|_2}{\hat{\sigma}_n - \sigma_{n+1}}$$

Proof

Substituting the SVD (4.4) into (4.9) and taking the τ-norm of both sides gives

$$\| x_{TLS} \|_\tau^2 = \sum_{i=1}^{n} \left[\frac{\hat{\sigma}_i}{(\hat{\sigma}_i + \sigma_{n+1})(\hat{\sigma}_i - \sigma_{n+1})} \frac{\hat{u}_i^T \hat{b}}{} \right]^2$$

The lemma follows from the inequalities $\frac{1}{2} \leq \frac{\hat{\sigma}_i}{(\hat{\sigma}_i + \sigma_{n+1})} \leq 1$ □

Theorem 4.4

If $A' \in R^{mxn}$ and $b' \in R^m$ are such that

$$\eta = \| D[\, A' - A \,|\, b' - b\,]\, T \|_F \leq \varepsilon / 6$$

where

$$\varepsilon = \hat{\sigma}_n - \sigma_{n+1} > 0$$

then the perturbed TLS problem

$$\min_{E,r} \| D[E|r]T \|_F \tag{4.20}$$

$$\text{subject to } b' + r \in \text{Range}(A' + E)$$

has a unique solution x'_{TLS}. Moreover, if $x_{TLS} \neq 0$, then

$$\frac{\| x_{TLS} - x'_{TLS} \|_\tau}{\| x_{TLS} \|_\tau} \leq \frac{9\,\eta\,\sigma_1}{\sigma_n - \sigma_{n+1}} \left\{ 1 + \frac{\lambda \| \hat{b} \|}{\hat{\sigma}_n - \sigma_{n+1}} \right\} \frac{1}{\| \lambda \hat{b} \|_2 - \sigma_{n+1}} \,. \tag{4.21}$$

Proof

Denote the singular values of the matrices $\hat{A}' = DA'T$ and $C' = D[A'|b']T$ by $\hat{\sigma}'_1 \geq \ldots \geq \hat{\sigma}'_n$ and $\sigma'_1 \geq \ldots \geq \sigma'_{n+1}$ respectively. Well-known perturbation results for singular values insure that

$$|\hat{\sigma}'_n - \sigma'_{n+1}| \geq |\hat{\sigma}_n - \sigma_{n+1}| - |\hat{\sigma}'_n - \hat{\sigma}_n| - |\sigma'_{n+1} - \sigma_{n+1}| \geq \varepsilon - \frac{\varepsilon}{6} - \frac{\varepsilon}{6} = \frac{2}{3}\varepsilon \tag{4.22}$$

In view of Theorem 4.1, this implies that the perturbed TLS problem above has a unique solution x'_{TLS}.

Let $\begin{bmatrix} y \\ \alpha \end{bmatrix}$ ($y \in R^n$, $\alpha \in R$) be a unit right singular vector of C associated with σ_{n+1}. Using the SVD perturbation theory of Stewart[18], it is possible to bound the difference between $\begin{bmatrix} y \\ \alpha \end{bmatrix}$ and $\begin{bmatrix} z \\ \beta \end{bmatrix}$, a corresponding singular vector of C' associated σ'_{n+1}. Not surprisingly, the bound involves the separation of σ_n and σ_{n+1} :

$$\left\| \begin{bmatrix} y \\ \alpha \end{bmatrix} - \begin{bmatrix} z \\ \beta \end{bmatrix} \right\|_2 \leq \frac{3\eta}{\sigma_n - \sigma_{n+1}}$$

Now from §2 we have $x_{TLS} = -T_1 y/(\lambda\alpha)$ and $x'_{TLS} = -T_1 z/(\lambda\beta)$ where $\lambda = t_{n+1}$ and $T_1 = \mathrm{diag}(t_1,\ldots,t_n)$. Thus,

$$\| x_{TLS} - x'_{TLS} \|_\tau = \frac{1}{\lambda} \| (y/\alpha) - (z/\beta) \|_2 \leq \frac{1}{|\lambda\alpha|} \left\{ \| y - z \|_2 + \frac{\| z \|_2}{|\beta|} |\alpha - \beta| \right\}$$

and so

$$\| x_{TLS} - x'_{TLS} \|_\tau \leq \frac{3\eta}{\sigma_n - \sigma_{n+1}} \frac{\| x_{TLS} \|_\tau}{\| y \|_2} [\, 1 + \lambda \| x'_{TLS} \|_\tau \,] .$$

Set $\hat{b}' = Db'$. From Lemma 4.3, equation (4.22), and the fact that $\| \lambda(\hat{b} - \hat{b}') \|_2 \leq$ [illegible] we have

$$\| x'_{TLS} \|_\tau \leq \frac{\| \hat{b}' \|_2}{\hat{\sigma}_n - \sigma_{n+1}} \leq \frac{3}{2} \left\{ \frac{\| \hat{b} \|_2}{\hat{\sigma}_n - \sigma_{n+1}} + \frac{1}{6\lambda} \right\}$$

and so

$$\frac{\| x_{TLS} - x'_{TLS} \|_\tau}{\| x_{TLS} \|_\tau} \leq \frac{3\eta}{\sigma_n - \sigma_{n+1}} \left\{ \frac{5}{4} + \frac{3}{2} \cdot \frac{\lambda \| \hat{b} \|}{\hat{\sigma}_n - \sigma_{n+1}} \right\} \frac{1}{\| y \|_2} . \tag{4.23}$$

In order to get a lower bound on $\| y \|_2$, observe that

$$\lambda |\alpha| \| \hat{b} \|_2 \leq \| [\hat{A} | \lambda \hat{b}] \begin{bmatrix} y \\ \alpha \end{bmatrix} \|_2 + \| \hat{A} y \|_2 \leq \sigma_{n+1} + \| \hat{A} \|_2 \| y \|_2$$

i.e.,

$$\begin{aligned} \lambda \| \hat{b} \|_2 - \sigma_{n+1} &\leq (1 - |\alpha|) \lambda \| \hat{b} \|_2 + \| \hat{A} \|_2 \| y \|_2 \\ &\leq \| y \|_2 [\, \lambda \| \hat{b} \|_2 + \| \hat{A} \|_2 \,] \leq 2 \| y \|_2 \| C \|_2 . \end{aligned}$$

The assumption that $x_{TLS} \neq 0$ implies that $\lambda \| \hat{b} \|_2 > \sigma_{n+1}$ for otherwise $\begin{bmatrix} y \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is a singular vector of C. The theorem now follows because

$$\frac{1}{\| y \|_2} \leq \frac{2\,\sigma_1}{\| \lambda \hat{b} \|_2 - \sigma_{n+1}} \quad \square$$

Both the lemma and the theorem suggest that the TLS problem is unstable whenever $\hat{\sigma}_n$ is close to σ_{n+1}. This is borne out by some results established in [23] where it is shown that a change of order $|\alpha|$ in C can result in an insoluble TLS problem. Using Lemma 4.3, this translates into the assertion that $\hat{\sigma}_n - \sigma_{n+1}$ is a measure of how close (1.4) is to the class of insoluble TLS problems.

Finally, we remark that if the LS problem is ill-conditioned, i.e., $\hat{\sigma}_n$ is small, then the TLS problem is likewise sensitive.

5. Algorithmic Considerations

Although a stable and efficient algorithm for computing the SVD exists [7], there are numerical difficulties associated with the determination of the dimension of S_c, i.e., the multiplicity of σ_{n+1}. One approach is to regard all computed singular values in the interval $[\sigma_{n+1}, \sigma_{n+1} + \varepsilon]$ as being identical where $\varepsilon > 0$ is some small machine dependent parameter. This leads to the following overall procedure for computing the solution to the TLS problem:

1. Compute the SVD $U^T(D[A|b]T)V = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n+1})$. Accumulate V.

2. Define the index p by $\sigma_p > \sigma_{n+1} + \varepsilon \geq \sigma_{p+1} \geq \cdots \geq \sigma_{n+1}$.

3. Let $V = [v_1, \ldots, v_n]$ be a column partition of V and compute a Householder matrix Q such that

$$[v_{p+1}, \ldots, v_{n+1}]Q = \left[\begin{array}{c|c} \times & y \\ \hline 0 \ \ldots \ 0 & \alpha \end{array}\right]$$

4. If $\alpha = 0$, then the TLS problem has no solution. Otherwise, $x_{TLS} = -T_1 y/(\alpha t_{n+1})$.

A shortcoming of this scheme is that it does not compute $\hat{\sigma}_n - \sigma_{n+1}$, which as we have seen, is a measure of TLS sensitivity. To rectify this it may be more desirable to compute the SVD $\hat{U}^T \hat{A} \hat{V} = \text{diag}(\hat{\sigma}_1,\ldots,\hat{\sigma}_n) = \hat{\Sigma}$ and then make use of the TLS "secular equation":

$$\Psi(\sigma) = \sigma^2 \left[\frac{1}{\lambda^2} + \sum_{i=1}^{n} \frac{c_i^2}{\hat{\sigma}_i^2 - \sigma^2} \right] = \rho_{LS}^2$$

In view of (4.9) and (4.11), if a σ can be found that satisfies this equation and is less than $\hat{\sigma}_n$, then

$$x_{TLS} = T_1 \hat{V} (\hat{\Sigma}^T \hat{\Sigma} - \sigma^2 I)^{-1} \hat{\Sigma}^T \hat{U}^T \hat{b} .$$

Standard root-finding techniques can be used for this purpose. (The function Ψ has monotonicity properties in the bracketing interval $[0,\hat{\sigma}_n]$.) Notice how easy it is to compute the TLS solution for different values of the weight $\lambda = t_{n+1}$. A detailed discussion of these and other algorithmic aspects of the TLS problem, such as the choosing of the weights, will appear elsewhere.

Acknowledgements

We are grateful to the following people for calling our attention to various aspects of the TLS problem: A.Bjork, R.Byers, P.Diaconis, C.Moler, C.Paige, C.Reinsch B.Rust, P.Velleman, and J.H.Wilkinson.

REFERENCES

[1] P.Barrera and J.E.Dennis, "Fuzzy Broyden Methods", private communication, 1979.

[2] W.G.Cochrane, "Errors of Measurement in Statistics," Technometrics, 10(1968), pp. 637-666.

[3] W.E.Deming, Statistical Adjustment of Data, J.Wiley & Sons Inc., New York, 1946.

[4] G.A.Gerhold, "Least Squares Adjustment of Weighted Data to a General Linear Equation," Amer.J.Physics, 37(1969), pp. 156-161 .

[5] P.E.Gill and W.Murray, "Computation of Lagrange Multiplier Estimates for Constrained Minimization," Math.Prog., 17(1979), pp. 32-60.

[6] G.H.Golub, "Some Modified Eigenvalue Problems," SIAM Review, 15(1973), pp. 318-344.

[7] G.H.Golub and C.Reinsch, "Singular Value Decomposition and Least Squares Solutions," Numer.Math., 14(1970), pp. 403-420.

[8] G.H.Golub and C.Van Loan, "Total Least Squares," in Smoothing Techniques for Curve Estimation, T.Gasser and M.Rosenblatt(eds), Springer-Verlag, New York,1979, pp. 69-76.

[9] R.F.Gunst, J.T.Webster, and R.L.Mason, "A Comparison of Least Squares and Latent Root Regression Estimators," Technometrics, 18(1976), pp.75-83.

[10] D.M.Hawkins, "On the Investigation of Alternative Regressions by Principle Component Analysis," Applied Statistics, 22(1973), pp.275-286.

[11] S.D.Hodges and P.G.Moore, "Data Uncertainties and Least Squares Regression," Applied Statistics, 21(1972), pp. 185-195.

[12] C.Lawson and R.Hanson, Solving Least Squares Problems, Prentice Hall, Englewood Cliffs, 1974.

[13] I.Linnik, Method of Least Squares and Principles of the Theory of Observations, Permagon Press, New York, 1961.

[14] A.Madansky, "The Fitting of Straight Lines When Both Variables are Subject to Error, JASA, 54(1959), pp. 173-205.

[15] K.Pearson, "On Lines and Planes of Closest Fit to Points in Space," Phil.Mag., 2(1901), pp. 559-572.

[16] D.Riggs, J.Guarnieri, and S.Adelman, "Fitting Straight Lines when Both Variables are Subject to Error," Life Sciences, 22(1978), pp. 1305-1360.

[17] G.W.Stewart, Introduction to Matrix Computations, Academic Press, New York, 197

[18] G.W.Stewart, "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," SIAM Review, 15(1973), pp.727-764.

[19] G.W.Stewart, "Sensitivity Coefficients for the Effects of Errors in the Independent Variables in a Linear Regression," Technical Report TR-571, Department of Computer Science, University of Maryland, 1977.

[20] J.T.Webster, R.F.Gunst, and R.L.Mason, "Latent Root Regression Analysis," Technometrics, 16(1974), pp. 513-522.

[21] J.H.Wilkinson, The Algebraic Eigenvalue Problem, Oxford University Press, London, 1965.

[22] D.York, "Least Squares Fitting of a Straight Line," Canadian J.Physics, 44(1966), 1079-1086.

[23] C.Van Loan, "On Stewart's Singular Value Decomposition for Partitioned Orthogonal Matrices," Department of Computer Science Report STAN-CS-79-767, Stanford University, 1979.