Springer

Dear Author:

Please find attached the final pdf file of your contribution, which can be viewed using the Acrobat Reader, version 3.0 or higher. We would kindly like to draw your attention to the fact that copyright law is also valid for electronic products. This means especially that:

- You may not alter the pdf file, as changes to the published contribution are prohibited by copyright law.

- You may print the file and distribute it amongst your colleagues in the scientific community for scientific and/or personal use.

- You may make an article published by Springer-Verlag available on your personal home page provided the source of the published article is cited and Springer-Verlag is mentioned as copyright holder. You are requested to create a link to the published article in LINK, Springer's internet service. The link must be accompanied by the following text: The original publication is available on LINK **http://link.springer.de.** Please use the appropriate URL and/or DOI for the article in LINK. Articles disseminated via LINK are indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks and consortia.

- You are not allowed to make the pdf file accessible to the general public, e.g. your institute/your company is not allowed to place this file on its homepage.

- Please address any queries to the production editor of the journal in question, giving your name, the journal title, volume and first page number.

Yours sincerely,

Springer-Verlag Berlin Heidelberg

Numer. Math. (2002) 91: 117–146
Digital Object Identifier (DOI) 10.1007/s002110100314

Numerische Mathematik

Scaled total least squares fundamentals

Christopher C. Paige[1,⋆], **Zdeněk Strakoš**[2,⋆⋆]

[1] School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2A7; e-mail: paige@cs.mcgill.ca
[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic; e-mail: strakos@cs.cas.cz

Received June 2, 1999 / Revised version received July 3, 2000 /
Published online July 25, 2001 – © Springer-Verlag 2001

Summary. The standard approaches to solving overdetermined linear systems $Bx \approx c$ construct minimal corrections to the vector c and/or the matrix B such that the corrected system is compatible. In ordinary least squares (LS) the correction is restricted to c, while in data least squares (DLS) it is restricted to B. In scaled total least squares (STLS) [22], corrections to both c and B are allowed, and their relative sizes depend on a real positive parameter γ. STLS unifies several formulations since it becomes total least squares (TLS) when $\gamma = 1$, and in the limit corresponds to LS when $\gamma \to 0$, and DLS when $\gamma \to \infty$. This paper analyzes a particularly useful formulation of the STLS problem. The analysis is based on a new assumption that guarantees existence and uniqueness of meaningful STLS solutions for all parameters $\gamma > 0$. It makes the whole STLS theory consistent. Our theory reveals the necessary and sufficient condition for preserving the smallest singular value of a matrix while appending (or deleting) a column. This condition represents a basic matrix theory result for updating the singular value decomposition, as well as the rank-one modification of the Hermitian eigenproblem. The paper allows complex data, and the equivalences in the limit of STLS with DLS and LS are proven for such data. It is shown how any linear system $Bx \approx c$ can be reduced to a minimally dimensioned core system satisfying our assumption. Consequently, our theory and algorithms can be applied to fully general systems. The basics of practical algorithms for both the STLS and DLS problems are indicated for either dense or large

⋆ Supported by NSERC of Canada Grant OGP0009236.
⋆⋆ Part of this work was performed during the academic years 1998/1999 and 1999/2000 while visiting Emory University, Atlanta, GA, USA
Correspondence to: Z. Strakoš

sparse systems. Our assumption and its consequences are compared with earlier approaches.

Mathematics Subject Classification (1991): 15A18, 65F20, 65F25, 65F50

1 Introduction

We will use $\mathcal{R}(B)$ to denote the range (column space) of a matrix B. Two useful approaches to solving the overdetermined approximate linear system

$$Bx \approx c, \quad B \text{ an } n \text{ by } k \text{ matrix, } c \text{ an } n\text{-vector}, \quad c \notin \mathcal{R}(B), \tag{1.1}$$

are ordinary least squares (LS, or OLS, see for example [1], [12, §5.3]) and total least squares (TLS, see [10, 11], and for example [1, §4.6], [12, §12.3], [16]). In LS we seek (we use $\|\cdot\|$ to denote the vector 2-norm)

$$\text{LS distance} \equiv \min_{r,y} \|r\| \quad \text{subject to} \quad By = c - r. \tag{1.2}$$

In TLS, E and s are sought to minimize the Frobenius (F) norm in

$$\text{TLS distance} \equiv \min_{s,E,z} \|[s, E]\|_F \quad \text{s. t.} \quad (B + E)z = c - s. \tag{1.3}$$

In both LS and TLS we look for a minimal correction such that the corrected problem is compatible. While in LS the correction is restricted to the vector c (which corresponds to the assumption that all errors are confined to the vector of observations), in TLS the correction is allowed to compensate for errors in the data matrix B as well as in the vector of observations c. The LS and TLS problems have statistical relevance for different situations, see Van Huffel and Vandewalle [16] for an excellent discussion and history. They also carefully delineated the TLS theory and how it is related to LS.

The opposite case to LS is the data least squares problem (DLS), see [13]. In DLS the correction is allowed only in B (errors are assumed to affect only the data matrix)

$$\text{DLS distance} \equiv \min_{G,w} \|G\|_F \quad \text{subject to} \quad (B + G)w = c. \tag{1.4}$$

All these approaches can be unified by considering the following very general scaled TLS problem (STLS), see the paper [22] by Rao, who called it "weighted TLS": for a given $\gamma > 0$,

$$\text{STLS distance} \equiv \min_{\tilde{s},\tilde{E},\tilde{z}} \|[\tilde{s}\gamma, \tilde{E}]\|_F \quad \text{s. t.} \quad (B + \tilde{E})\tilde{z} = c - \tilde{s}. \tag{1.5}$$

Here the relative sizes of the corrections in B and c are determined by the real parameter $\gamma > 0$. When $\gamma \to 0$ the STLS solution approaches the LS solution, when $\gamma = 1$ (1.5) coincides with the TLS formulation, and when

$\gamma \to \infty$ it approaches DLS. The case $\gamma \to 0$ is not completely obvious since setting $\gamma = 0$ in (1.5) leads to $\tilde{E} = 0$, but allows *arbitrary* $\tilde{s}$, which does not necessarily mean the LS solution. The case $\gamma = 1$ is obvious, and we see that $\gamma \to \infty$ requires $\tilde{s} \to 0$, leading to DLS. For more on STLS and DLS see also [3], [4], [5]. Scaling by a diagonal matrix was considered in [11], and this motivated later researchers, leading eventually to the STLS formulation in [22]. The paper [8] considered the case where only some of the columns of the data matrix are contaminated, and this also suggested a way of treating LS as well as TLS in the one formulation.

The formulation of the STLS problem that we use is slightly different from that in (1.5). For any positive bounded γ, substitute in (1.5) $s \equiv \tilde{s}\gamma$, $z \equiv \tilde{z}$ and $E \equiv \tilde{E}$ to obtain the new formulation of the STLS problem: for a given $\gamma > 0$,

$$\text{(1.6) STLS distance} \quad \equiv \quad \min_{s,E,z} \|[s, E]\|_F \quad \text{s. t.} \quad (B + E)z\gamma = c\gamma - s.$$

We call the $z = z(\gamma)$ that minimizes this the *STLS solution* of (1.6). In analogy with (1.3), we call $z(\gamma)\gamma$ the *TLS solution* of (1.6). In (1.6) we could have written z instead of $z\gamma$. We chose the present form so that for positive bounded γ, the STLS solution $z = z(\gamma)$ of (1.6) is identical to the solution $\tilde{z}$ of (1.5). Thus (1.5) and (1.6) have identical distances and solutions for positive bounded γ. Therefore our results and discussions based on (1.6) apply fully to the scaled TLS problem (1.5).

We show for (1.6) that as $\gamma \to 0$, $z(\gamma)$ becomes the LS solution y of (1.2), and $(\text{STLS distance})/\gamma$ becomes the LS distance. As $\gamma \to \infty$, $z(\gamma)$ becomes the DLS solution w of (1.4), and the STLS distance becomes the DLS distance. The convergence of the STLS problem to the LS problem has been described in [22], and essentially in [11], for the real case. Here the convergence is proven for complex data by explicitly taking the limits for both solutions and distances.

We found that the development of our results was more simple and intuitive using the formulation (1.6) rather than (1.5). In particular, all the known TLS theory and algorithms can be applied directly to (1.6). The equivalence of (1.6) and (1.5) is extremely useful. This equivalence was pointed out to us by Sabine Van Huffel [15] after she read an earlier version of our work based on (1.6). We have not seen it stated in the literature, but it is implicit in the paper by Rao [22].

In (1.6), γ simply scales the right-hand side vector c (and the STLS solution $z = z(\gamma)$). Therefore it is appropriate to call the formulation (1.6) the *scaled* TLS problem, rather than the "weighted" TLS problem as was done in [22]. This also avoids the possibility of confusing the meaning of "weighted" here with its different meaning in "weighted least squares".

Using γ can have a statistical significance. Suppose that the elements of B are known to have independent zero-mean random errors of equal standard deviation δ_B. Suppose also that the elements of c have been observed with independent zero-mean random errors of equal standard deviation δ_c, and that the errors in c and B are independent. Then taking $\gamma = \delta_B/\delta_c$ in (1.6) will ensure that all the errors in that model have equal standard deviation (and so variance), and (1.6) is the ideal formulation for providing estimates. This agrees with the limiting behaviour described above, for clearly if $\delta_B = 0$ and $\delta_c \neq 0$, then LS is the correct choice, while if $\delta_B \neq 0$ and $\delta_c = 0$, then DLS is the correct choice. However (1.6) can also be useful outside any statistical context, see for example [20], and then γ does not have the above interpretation.

In all these formulations, if $c \in \mathcal{R}(B)$, then zero distance can be obtained via a direct solution. Otherwise TLS, and so STLS solutions can be found via the singular value decomposition (SVD). Let $\sigma_{min}(\cdot)$ denote the smallest singular value of a given matrix. To be precise, $\sigma_{min}(G)$ will denote the j-th largest singular value of an n by j matrix G, and will be zero if $n < j$. The interlacing property for the eigenvalues of $[B, c]^H[B, c]$ and of $B^H B$ [23, Ch2, §47, pp. 103–4] tells us that $\sigma_{min}([B, c]) \leq \sigma_{min}(B)$. When

$$\sigma_{min}([B, c]) < \sigma_{min}(B) \tag{1.7}$$

the n by k matrix B must have rank k, the unique solution of the TLS problem (1.3) is obtained from scaling the right singular vector of $[B, c]$ corresponding to $\sigma_{min}([B, c])$, and the norm of the TLS correction satisfies $\min_{s,E,z} \|[s, E]\|_F = \sigma_{min}([B, c])$, (see for example [12, §12.3]). When

$$\sigma_{min}([B, c\gamma]) < \sigma_{min}(B) \quad \text{for a given } \gamma > 0, \tag{1.8}$$

it follows that

$$\text{STLS distance in (1.6)} \;=\; \sigma_{min}([B, c\gamma]). \tag{1.9}$$

In the general case, let $\mathcal{U}_{min}$ be the left singular vector subspace of B corresponding to $\sigma_{min}(B)$. We explain in full later why (1.8) should not be used as a basis for the STLS theory. Very briefly, if $c \perp \mathcal{U}_{min}$, then $[B, c\gamma]$ has a singular value equal to $\sigma_{min}(B)$ for all $\gamma > 0$. But for a *particular* value of γ, $[B, c\gamma]$ might have a *smaller* singular value than $\sigma_{min}(B)$. Thus we can have (1.8) *and* $c \perp \mathcal{U}_{min}$. But $c \perp \mathcal{U}_{min}$ means $\sigma_{min}(B)$ plays no role in solving the LS problem, so the comparison with $\sigma_{min}(B)$ in (1.8) should not form the basis for deciding if there is a solution to the STLS, or even TLS, problem.

We argue that a satisfactory condition for building the theory of the TLS, DLS and STLS formulations for solving (1.1) is the γ-independent criterion:

$$\text{the } n \times k \text{ matrix } B \text{ has rank } k, \text{ and } \; c \not\perp \mathcal{U}_{min}. \tag{1.10}$$

We will show in Theorem 3.1 (see (3.7)) that this implies

$$\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]) < \sigma_{min}(B) \quad \text{for all } \gamma \geq 0, \tag{1.11}$$

which of course implies (1.7) and (1.8). The condition (1.10) is the simplest one. It only requires the SVD of B, while the others each require two SVDs. Note also that (1.10) is purely geometric.

A trivial example of B with rank k, but c not satisfying (1.10), is

$$\left[\begin{array}{c|cc} c & B \end{array}\right] = \left[\begin{array}{c|cc} 2 & 4 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{array}\right]. \tag{1.12}$$

Note that for sufficiently large γ (even $\gamma = 1$), this example gives

$$\sigma_{min}([B, c\gamma]) = \sigma_{min}(B). \tag{1.13}$$

But dropping the last row and column of $[c, B]$, to give $[c_1, B_{11}]$, results in a "core" problem $B_{11}x_1 \approx c_1$ satisfying (1.10) and (1.11).

Almost all practical problems will satisfy (1.10), and so (1.11), but to complete the theoretical foundations of the STLS problem, Theorem 3.1 analyzes when it is possible to have (1.13). This case seems never to have been fully analyzed before. Clearly (1.13) corresponds to the smallest singular value of a matrix being preserved when appending (or deleting) a column. This is useful in the theory of updating the singular value decomposition, and the rank-one modification of the Hermitian eigenproblem.

A crucial property of the criterion (1.10) is that *any* linear system $Bx \approx c$ can in theory be reduced to a "core" problem satisfying (1.10). In practice this can be done by direct computations that can be usefully applied to all small and dense STLS problems. We also suggest an algorithm for the large sparse matrix case.

Thus in this paper we present a new and thorough analysis of the theoretical foundations of the STLS problem, and of its relationships to the LS and DLS problems. But we also develop some more generally applicable matrix theory, and suggest the basics for useful approaches to solving STLS and DLS problems.

The rest of the paper is organized as follows. We start in Sect. 2 by reviewing some mathematical tools that we will need. In Sect. 3 we prove just when (1.13) can hold, since this is poorly understood, but is needed for a full understanding of TLS, DLS and STLS problems, and for our choice of criterion (1.10). It represents a general matrix theory result that might be useful outside the context of this paper. In Sect. 4 we give the valuable secular equation which $\sigma_{min}([B, c\gamma])$ in (1.9) must satisfy. Section 5 derives alternative forms of the STLS formulation (1.6) and the DLS formulation (1.4), as well as the DLS solution. We allow complex data, and the functional

to be optimized is not an analytic function, so the derivative cannot be taken in the usual way. Thus we give new algebraic proofs of optimality, instead of using derivatives. In Sect. 6 we show how the STLS problem (1.6) corresponds to the LS problem (1.2) when $\gamma \to 0$, and to the DLS problem (1.4) when $\gamma \to \infty$. Section 7 shows why the formulations (1.3)–(1.6) are incomplete without the criterion (1.10), since when this does not hold, they at best contain computationally dangerous irrelevant data, and at worst do not lead to meaningful solutions. Section 8 shows how to handle the completely general STLS (or even $Bx \approx c$) problem by reducing it to a core problem that satisfies an even stronger criterion than (1.10) — one which ensures the core problem is irreducible in the sense of containing no information that is irrelevant to the solution. This suggests practical approaches to solving STLS problems, whether the data matrix B is small and dense, or large and sparse. Section 9 discusses how STLS problems can be solved computationally, and describes a simple solution to the DLS problem. Section 10 compares our chosen assumptions for ensuring unique STLS solutions with the criteria for "generic" TLS problems given in [16]. We will always use these quotes here because we show that some of the problems labelled "generic" in [16] are not generic in the more usual sense of the word. This is in no way a criticism of [16] — the authors were probably using the terminology to indicate that all such problems could be solved by the standard algorithm of Golub and Van Loan [11].

This paper deals with equalities, and is the first in a sequence. The next one [20] will deal with bounds and the LS–STLS relationship when $\gamma > 0$.

We use $[c\gamma, B]$ for some purposes, and $[B, c\gamma]$ for others. Their SVDs are trivially related. Of course all the ideas given here for general γ apply to the TLS problem (1.3) by taking $\gamma = 1$.

The philosophy behind this paper is radically different from that of previous TLS, STLS or DLS work known to us. The STLS formulation (1.6) makes it easy to analyze and solve the STLS problem (it shows the STLS problem is just the TLS problem with its right-hand side c scaled by γ, so all the TLS artillery is available). But more importantly than that, the approach of reducing a problem $Bx \approx c$ to its "core" problem (Sect. 8) and solving that core problem simplifies our understanding of the area. It also simplifies the development of algorithms, while unifying the theoretical problems in the area. Crucial to all this is the new (γ-independent) criterion (1.10) for STLS (also TLS, DLS and even possibly LS) problems. This is based on c and the SVD of B, whereas the previous main TLS criterion (1.7) involved the SVDs of both B and $[B, c]$ (and so would be dependent on γ for STLS problems). The key here is that *any* STLS (or LS or TLS or DLS) problem can in theory be transformed by *direct* unitary transformations into two independent problems: a (possibly nonexistent) trivial problem, and a core

problem, where the core problem automatically satisfies (1.10). Solving the core problem then solves the original problem. Thus no complicated conditions such as (1.7) or (1.10) need be tested, and no special cases need be treated. All the decisions can be made by examining the sizes of elements in the unitarily transformed data. Both theory and computations can thus be unified, simplified and clarified.

2 Mathematical preliminaries

Here we introduce some notation and general theory that we will need in the paper. We use $\bar{\alpha}$ to denote the complex conjugate of the scalar α, and a^H to denote the complex conjugate transpose of the vector a. The $k \times k$ unit matrix is $I = [e_1, \ldots, e_k]$. We will need the following lemma. It is a generalization of the familiar result obtained by taking $m = -1$ in the lemma.

Lemma 2.1 *For any integer m and matrix Z,*

$$Z(Z^HZ - \lambda I)^m Z^H - \lambda(ZZ^H - \lambda I)^m = (ZZ^H - \lambda I)^{m+1}, \tag{2.1}$$

where if $m < 0$, the scalar λ must be such that the inverses exist.

Proof. Clearly (2.1) is true for $m = 0$. Multiply each side of (2.1) by $ZZ^H - \lambda I$, giving

$$Z(Z^HZ - \lambda I)^{m+1} Z^H - \lambda(ZZ^H - \lambda I)^{m+1} = (ZZ^H - \lambda I)^{m+2}, \tag{2.2}$$

which is (2.1) with m increased by unity. Thus since (2.1) holds for $m = 0$, it holds for all integers $m \geq 0$. Now if $m = -1$, (2.2) is true, so (2.1) is also true if $ZZ^H - \lambda I$ is nonsingular. Similarly we can show (2.1) is true for $m = -2, -3, \ldots$. □

We can avoid the restriction on λ as follows.

Corollary 2.1 *For $m < 0$, (2.1) also holds for* any *scalar λ if we replace each inverse by the Moore–Penrose pseudo-inverse (the G_{1234} generalized inverse).*

Proof. Replace Z by its singular value decomposition, and use (2.1) for the nonsingular part. □

In order to analyze or solve (1.2)–(1.6) we usually transform the data $[B, c]$. The transformations that we use here take the form

$$[\tilde{B}, \tilde{c}] = P^H[BQ, c], \qquad P \text{ and } Q \text{ unitary}, \tag{2.3}$$

(for obtaining LS solutions, Q need only be nonsingular). These do not alter the distances defined in Sect. 1, and the solution vector for the original data $[B, c]$ is Q times that for the transformed data.

We regularly use the following for r, c, B and y in (1.2). Let $n \geq k$ in (1.1). Let the $n \times k$ matrix B have rank k and singular values σ_i with singular value decomposition (SVD)

$$B = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^H, \quad \Sigma \equiv \text{diag}(\sigma_1, \ldots, \sigma_k), \quad \sigma_1 \geq \ldots \geq \sigma_k > 0. \tag{2.4}$$

Here U is an $n \times n$ unitary matrix, Σ is $k \times k$, and $k \times k$ V is a unitary matrix. If $n > k$, in $U = [U_B | \hat{U}_B] = [u_1, \ldots, u_k | u_{k+1}, \ldots, u_n]$, $\hat{U}_B$ is arbitrary up to multiplication on the right by a unitary matrix, so assume it is chosen to give $\hat{U}_B^H c = e_1 \rho$, $\rho \geq 0$. If $n = k$, this ρ will not exist. For this study, an important part of the SVD of B is

$$\mathcal{U}_{min} \equiv \text{the left singular vector subspace of } B \text{ for } \sigma_{min}(B). \tag{2.5}$$

A useful allowable transformation of the data is

$$U^H [B, c] \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{cc} \Sigma & a \\ 0 & \rho \\ \hline 0 & 0 \end{array}\right],$$

$$a \equiv (\alpha_1, \ldots, \alpha_k)^T \equiv [u_1, \ldots, u_k]^H c = U_B^H c. \tag{2.6}$$

We also denote, for $\gamma \geq 0$,

$$N \equiv U^H [B, c\gamma] \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{cc} \Sigma & a\gamma \\ 0 & \rho\gamma \\ \hline 0 & 0 \end{array}\right]. \tag{2.7}$$

N has the same singular values as $[B, c\gamma]$. The elements of a are the components of the vector of observations c in the directions of the left singular vectors of the data matrix B. With (1.2) we see that

$$U^H r = U^H (c - By) = \begin{bmatrix} a - \Sigma V^H y \\ \rho \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \rho \\ 0 \end{bmatrix}$$

gives the minimum for $\|r\|$. Then for the LS solution and residual

$$y = V \Sigma^{-1} a, \qquad \|y\|^2 = \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2}, \tag{2.8}$$

$$\|r\| = \rho. \tag{2.9}$$

For analysis of the STLS problem (1.6), we will be interested in the singular values σ of $[B, c\gamma]$, see (1.9), and so the eigenvalues σ^2 of $N^H N$

for N in (2.7). We use the following classical results (see for example [14]) to analyze these. Consider a matrix G with partitioning $G = \begin{bmatrix} C & D \\ E & F \end{bmatrix}$. When C is square and nonsingular, the Schur complement (G/C) of C in G is defined as (we will also need $C = C^H$, $D = E^H$ and $F = F^H$ for (2.12) below)

$$(G/C) \equiv F - EC^{-1}D, \text{ so } G = \begin{bmatrix} I & 0 \\ EC^{-1} & I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & (G/C) \end{bmatrix} \begin{bmatrix} I & C^{-1}D \\ 0 & I \end{bmatrix}. \tag{2.10}$$

Define the "inertia" $\mathrm{In}(M)$ of a Hermitian matrix M to be the ordered triple $\{i_+, i_-, i_0\}$, where i_+ denotes the number of positive eigenvalues of M, i_- the number of negative eigenvalues, and i_0 the number of zero eigenvalues. We will use results that follow from (2.10):

$$\det(G) = \det(C) \cdot \det(G/C), \tag{2.11}$$

$$\mathrm{In}(G) = \mathrm{In}(C) + \mathrm{In}(G/C). \tag{2.12}$$

3 The minimum singular values of $[B, c\gamma]$ and B

When the minimum singular values of $[B, c\gamma]$ and B are distinct, the STLS problem (1.6) has a unique solution. The other possibility,

$$\sigma_{min}([B, c\gamma]) = \sigma_{min}(B), \tag{3.1}$$

is important but subtle, so here we show just when this can happen. This will lead us to a full understanding of the different possible meanings of STLS problems.

The condition for (3.1) to hold is also the condition for the smallest singular value of a full column rank matrix (here B) to remain unchanged when we append a column (here $c\gamma$) to the matrix B (or delete the column $c\gamma$ from the matrix $[B, c\gamma]$). The singular value 0 is clearly unchanged if B does not have full column rank. Clearly, a similar condition can be formulated with respect to the rows of a matrix. This represents a general result independent of the context of our paper. The proof is an extension of the following result.

Lemma 3.1 *If the vector a has at least one element, the Hermitian arrow matrix $A \equiv \begin{bmatrix} 0 & a \\ a^H & \alpha \end{bmatrix}$ is positive semi-definite (singular with no negative eigenvalues) if and only if $a = 0$ and the scalar $\alpha \geq 0$.*

Proof. Suppose $a \neq 0$, then without loss of generality we can assume its last element is nonzero. By considering the determinant, we see the last 2×2 principal submatrix of A has a negative eigenvalue, and by the interlacing

property (see [23, Ch2, §47, pp. 103–4]) the whole of A must have a negative eigenvalue. If $a = 0$, A is singular but has no negative eigenvalue if and only if $\alpha \geq 0$. $\square$

Theorem 3.1 *Let $\gamma > 0$ be a scalar, and for $n \geq k \geq 1$ let $[B, c\gamma]$ be an n by $k+1$ matrix with n by k submatrix B. Let B have singular values $\sigma_1 \geq \ldots \geq \sigma_j > \sigma_{j+1} = \ldots = \sigma_k \equiv \sigma_{min}(B) > 0$ with the corresponding left singular vectors $u_1, \ldots, u_k$. Let $\rho = \|r\|$ be the minimum in (1.2), see (2.6), (2.9). Then*

$$\sigma_{min}([B, c\gamma]) = \sigma_{min}(B) \tag{3.2}$$

if and only if (see (2.6))

$$\alpha_i \equiv u_i^H c = 0, \qquad i = j+1, \ldots, k, \tag{3.3}$$

and

$$\psi_j(\sigma_k, \gamma) \geq 0, \quad \psi_j(\sigma, \gamma) \equiv \gamma^2\|r\|^2 - \sigma^2 - \gamma^2\sigma^2 \sum_{i=1}^{j} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma^2}. \tag{3.4}$$

The summation term is ignored if all singular values of B are equal.

Proof. Write $\Sigma_1 \equiv \text{diag}(\sigma_1, \ldots, \sigma_j)$, $\Sigma_2 \equiv \text{diag}(\sigma_{j+1}, \ldots, \sigma_k) = \sigma_k I$, $a_1 \equiv (\alpha_1, \ldots, \alpha_j)^T$, $a_2 \equiv (\alpha_{j+1}, \ldots, \alpha_k)^T$, $a \equiv (a_1^T, a_2^T)^T$. Σ_1 and a_1 need not exist (j can be zero), but Σ_2 and a_2 do ($k - j > 0$). To prove the theorem, we need to show for any given $\gamma > 0$,

$$\sigma_{min}([B, c\gamma]) = \sigma_{min}(B) \Leftrightarrow \{a_2 = 0 \;\&\; \psi_j(\sigma_k, \gamma) \geq 0\}. \tag{3.5}$$

Clearly (3.2) holds if and only if σ_k is the minimum singular value of N in (2.7), that is, if and only if

$$N^H N - \sigma_k^2 I = \begin{bmatrix} \Sigma_1^2 - \sigma_k^2 I_j & 0 & \Sigma_1 a_1 \gamma \\ 0 & 0 \cdot I_{k-j} & a_2 \sigma_k \gamma \\ \gamma a_1^H \Sigma_1 & \gamma \sigma_k a_2^H & \gamma^2(a^H a + \rho^2) - \sigma_k^2 \end{bmatrix} \tag{3.6}$$

is positive semi-definite. If $j > 0$ the Schur complement M of positive definite $\Sigma_1^2 - \sigma_k^2 I$ in $N^H N - \sigma_k^2 I$ is, see (2.10),

$$\begin{aligned} M &= \begin{bmatrix} 0 & a_2 \sigma_k \gamma \\ \gamma \sigma_k a_2^H & \psi \end{bmatrix}, \\ \psi &\equiv \gamma^2(a^H a + \rho^2) - \sigma_k^2 - \gamma^2 a_1^H \Sigma_1 (\Sigma_1^2 - \sigma_k^2 I)^{-1} \Sigma_1 a_1 \\ &= \gamma^2(a_2^H a_2 + \rho^2) - \sigma_k^2 - \gamma^2 \sigma_k^2 a_1^H (\Sigma_1^2 - \sigma_k^2 I)^{-1} a_1 \\ &= \psi_j(\sigma_k, \gamma) + \gamma^2 a_2^H a_2, \end{aligned}$$

using Lemma 2.1 with $m = -1$, and $\psi_j(\sigma_k, \gamma)$ in (3.4). From (2.12)

$$\mathrm{In}(N^H N - \sigma_k^2 I) = \mathrm{In}(\Sigma_1^2 - \sigma_k^2 I) + \mathrm{In}(M),$$

so (3.2) holds if and only if M is positive semi-definite. But a_2 has at least one element, so from Lemma 3.1 this is true if and only if $a_2 = 0$ and $\psi = \psi_j(\sigma_k, \gamma) + \gamma^2 a_2^H a_2 = \psi_j(\sigma_k, \gamma) \geq 0$. Thus (3.3) and (3.4) are necessary and sufficient for (3.2). If $j = 0$ the same result follows by applying Lemma 3.1 directly to (3.6). □

Since $\sigma_{min}([B, c \cdot 0]) = 0$, this has proven for $\gamma \geq 0$ (the left hand side is (1.10))

(3.7) $\{B$ full column rank & $a_2 \neq 0\} \Rightarrow \sigma_{min}([B, c\gamma]) < \sigma_{min}(B).$

The theorem also tells us that for B and c representing data from some real world application, having $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ exactly is a rare event. It requires all left singular vectors of B corresponding to its smallest singular value σ_k to be orthogonal to c, as well as (3.4). The first condition ($a_2 = 0$ in the theorem) is highly unlikely to be satisfied. Moreover, even when it is true, we cannot necessarily find γ satisfying (3.4). For a particular B and c it is possible to have

$$\|r\|^2 - \sigma_k^2 \sum_{i=1}^{j} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma_k^2} \leq 0, \tag{3.8}$$

giving $\psi_j(\sigma_k, \gamma) < 0$ for all $\gamma > 0$, see (3.4). In fact we have:

Corollary 3.1 *Using the notation of Theorem 3.1, where $a_2 \neq 0$ corresponds to $c \not\perp \mathcal{U}_{min}$ in (1.10), if B has rank k then*

$$\begin{aligned}
\{a_2 \neq 0\} \quad & \textit{or} \quad \{a_2 = 0 \ \& \ (3.8)\} \\
& \Leftrightarrow \quad \{\sigma_{min}([B, c\gamma]) < \sigma_{min}(B) \ \forall\, \gamma \geq 0\}, \\
\{a_2 = 0\} \quad & \& \quad \{\exists\, \gamma_0 > 0 \ \textit{such that} \ \psi_j(\sigma_k, \gamma_0) = 0 \ \textit{in} \ (3.4)\} \\
& \Leftrightarrow \begin{cases} \sigma_{min}([B, c\gamma]) < \sigma_{min}(B) & \forall\, 0 < \gamma < \gamma_0, \\ \sigma_{min}([B, c\gamma]) = \sigma_{min}(B) & \forall\, \gamma \geq \gamma_0. \end{cases}
\end{aligned}$$

Proof. These follow from (3.5) and the form of $\psi_j(\sigma_k, \gamma)$ in (3.4). □

Remark 3.1 From this we can see that for an *arbitrary* B and c with $\sigma_{min}([B, c\gamma_1]) < \sigma_{min}(B)$ for some $\gamma_1 > 0$, one cannot always get $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ by increasing γ. But sometimes for a B, c and γ_1 with $\sigma_{min}([B, c\gamma_1]) < \sigma_{min}(B)$ there exists $\gamma_0 > \gamma_1$ such that $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ for all $\gamma \geq \gamma_0$.

4 The secular equation for singular values of $[B, c\gamma]$

When (1.10) holds, the smallest singular value of $[B, c\gamma]$ is the STLS distance in (1.6). We now derive several forms of the secular equation for this STLS distance. These forms will be useful for examining the limiting behaviour in Sect. 6, and for obtaining bounds in [20].

Lemma 4.1 *For any $n\times k$ matrix B and n-vector c let $\sigma(\gamma)\equiv\sigma_{min}([B, c\gamma])$. If (1.10) holds, then for all $\gamma \geq 0$ the STLS distance in (1.6) is $\sigma(\gamma)$, which is the smallest nonnegative scalar σ satisfying*

$$
\begin{aligned}
(4.1)\quad 0 = \psi_k(\sigma,\gamma) &\equiv \det([B, c\gamma]^H[B, c\gamma] - \sigma^2 I)/\det(B^H B - \sigma^2 I)\\
(4.2)\quad &= \gamma^2 c^H c - \sigma^2 - \gamma^2 c^H B(B^H B - \sigma^2 I)^{-1} B^H c\\
(4.3)\quad &= \gamma^2 c^H [I - B(B^H B - \sigma^2 I)^{-1} B^H]c - \sigma^2\\
(4.4)\quad &= -\gamma^2\sigma^2 c^H (BB^H - \sigma^2 I)^{-1} c - \sigma^2\\
(4.5)\quad &= \gamma^2\rho^2 - \sigma^2 - \gamma^2\sigma^2 a^H(\Sigma\Sigma^H - \sigma^2 I)^{-1} a\\
(4.6)\quad &= \gamma^2\|r\|^2 - \sigma^2 - \gamma^2\sigma^2 \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma^2},
\end{aligned}
$$

where these last two lines use the notation of (2.4), (2.6), (2.9).

Proof. When (1.10) holds, we proved in Theorem 3.1 that (1.11) holds, so $\sigma(\gamma)$ is the STLS distance in (1.6) for all $\gamma \geq 0$, see (1.9). But (1.11) shows $B^H B - \sigma^2(\gamma) I$ is positive definite, so $\sigma(\gamma)$ is the smallest nonnegative σ satisfying (4.1). Since

$$
[B, c\gamma]^H[B, c\gamma] - \sigma^2 I = \begin{bmatrix} B^H B - \sigma^2 I & B^H c\gamma \\ \gamma c^H B & \gamma^2 c^H c - \sigma^2 \end{bmatrix},
$$

(4.1) and (2.11) show that

$$
\psi_k(\sigma,\gamma) = \det(([B, c\gamma]^H[B, c\gamma] - \sigma^2 I)/(B^H B - \sigma^2 I)),
$$

which is the Schur complement of $B^H B - \sigma^2 I$ in $[B, c\gamma]^H[B, c\gamma] - \sigma^2 I$, since the Schur complement is a scalar here. This with the definition in (2.10) proves (4.2). But (4.3) is just (4.2) rearranged, and (4.4) follows from (4.3) by using Lemma 2.1 with $m = -1$. Finally (4.5) and (4.6) follow from (4.4) by using the SVD of B in (2.4), and the notation of (2.6) and (2.9). □

When the elements α_i of a are nonzero and the σ_i are distinct in (4.6) (see [23, §39, pp. 94–6], which also handles the case when this last is not so) *all* the singular values of $[B, c\gamma]$ are given by the $k+1$ solutions $\sigma \geq 0$ of the secular equation $0 = \psi_k(\sigma,\gamma)$. When some $\alpha_i = 0$, (2.6) and (2.7) show both B and $[B, c\gamma]$ have the singular value σ_i. However we are only

interested in $\sigma(\gamma)$, the smallest singular value, thus when (1.10) holds, we see from (4.6) $\sigma(\gamma)$ satisfies

$$0 = \psi_k(\sigma(\gamma), \gamma) = \gamma^2\|r\|^2 - \sigma(\gamma)^2 - \gamma^2\sigma(\gamma)^2 \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma(\gamma)^2}. \tag{4.7}$$

With $\gamma = 1$, (4.7) was derived in [11], see also [16, Thm. 2.7, & (6.36)], where [16, (6.36)] corresponds to a more general case. These derivations assumed the weaker condition (1.7), and so cannot be generalized to STLS for all $\gamma \geq 0$, see Remark 3.1.

It is of interest to examine how $\sigma(\gamma)$ changes with γ.

Corollary 4.1 *If (1.10) holds, $\gamma > 0$, and c is not in the range of B, then $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$ increases as γ increases, and decreases as γ decreases, strictly monotonically.*

Proof. (1.10) implies $n \times k$ B has rank k. If finite $\gamma > 0$ and c is not in the range of B, then $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]) > 0$. Differentiating ((4.7) divided by $\gamma^2\sigma(\gamma)$) with respect to γ gives

$$\dot{\sigma}(\gamma) \left[\sigma(\gamma) \sum_{i=1}^{k} \frac{|\alpha_i|^2}{(\sigma_i^2 - \sigma(\gamma)^2)^2} + \frac{\|r\|^2}{\sigma(\gamma)^3} \right] = \frac{1}{\gamma^3} .$$

But when (1.10) holds, (3.7) shows $\sigma(\gamma) < \sigma_k$ for all $\gamma > 0$, so the factor $[\cdot]$ here represents a positive finite number, and thus $\dot{\sigma}(\gamma) > 0$ for all $\gamma > 0$. □

It is revealing to put the result of Theorem 3.1 in the context of work on updating the SVD, or on rank-one modification of the Hermitian eigenproblem (see for example [2], which is based on the ideas of Wilkinson described in [23, Ch.2, §39, pp. 94–96]). Assume that the condition (3.3) is satisfied. Then $0 = \psi_j(\sigma, \gamma)$ represents the secular equation of the corresponding deflated problem (where the $k - j$ deflation steps correspond to $\alpha_{j+1} = \ldots = \alpha_k = 0$). Then $[B, c\gamma]$ has $k - j$ singular values equal to $\sigma_{min}(B)$. The condition (3.4) guarantees that the deflated secular equation does not have a solution σ less than $\sigma_{min}(B)$ (when $\psi_j(\sigma_k, \gamma) < 0$, it does have such a solution). Conversely, if $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$, then (3.3) must hold and $\sigma_{min}(B)$ must be deflated, otherwise the function $\psi_k(\sigma, \gamma)$ will have a pole at $\sigma_{min}(B)$ and a positive solution $\sigma(\gamma)$ strictly less than $\sigma_{min}(B)$ (which gives a contradiction). Moreover, the deflated secular equation $0 = \psi_j(\sigma, \gamma)$ must not have a positive solution less than $\sigma_{min}(B)$, which gives (3.4). We see that we could have proved our Theorem 3.1 directly using the ideas of Wilkinson, but we prefer our way above, because it is logically simpler, and it also provides some algebraic relations that we use later in the paper. Some related questions were also studied in [6], but a statement similar to our Theorem 3.1 was not considered there.

5 Alternative STLS and DLS formulations

When the minimum singular values of $[B, c]$ and B are distinct, the SVD approximation theory used to provide the TLS solution (see for example [11, 12, 1, 16]) is so powerful that two intermediate formulations which we need are usually not mentioned. These hold even when the minimum singular values are equal. The STLS and DLS versions are needed here to prove a theoretical weakness of the formulations (1.3)–(1.6), and may even be useful otherwise.

For the generality of this paper we allow the data to be complex. This leads to nontrivial proofs, since (1.6) is a constrained optimization problem, but for example $\|E\|_F^2$ is not an analytic function of the elements of the complex matrix E. Thus we avoid differentiation in our proofs. We learnt a technique for doing this by listening to J. H. Wilkinson. The idea is to start with the answer — perhaps found by differentiating the Lagrangian for the real case and generalizing — and show that any change to the answer increases the functional. This also allows us to give a rigorous proof of the form of the DLS solution when the data can be complex.

For clarity in this analysis we define $x \equiv z\gamma$, so (1.6) becomes:

$$\sigma_S^2 \equiv \min_{s,E,x} \|[s, E]\|_F^2, \quad \text{subject to} \quad (B + E)x = c\gamma - s. \tag{5.1}$$

Suppose for a given $\gamma > 0$ that $\hat{x}$ and $\hat{s}$ are the *vectors* in the solution of the STLS problem (5.1). We now show that the matrix part of the solution is

$$\hat{E} = \hat{d}\hat{x}^H/\hat{x}^H\hat{x}, \qquad \hat{d} \equiv c\gamma - B\hat{x} - \hat{s}. \tag{5.2}$$

For $\hat{E}$ and the solution vectors $\hat{x}$ and $\hat{s}$, (5.1) simplifies to

$$\sigma_S^2 = \min_F\{\|\hat{s}\|^2 + \|\hat{E} + F\|_F^2\} \quad \text{s. t.} \quad (B + \hat{E} + F)\hat{x} = c\gamma - \hat{s}. \tag{5.3}$$

For any F satisfying these constraints,

$$\hat{d} \equiv c\gamma - B\hat{x} - \hat{s} = (\hat{E} + F)\hat{x} = \hat{d} + F\hat{x},$$

so $F\hat{x} = 0$. Thus $F\hat{E}^H = F\hat{x}\hat{d}^H/\hat{x}^H\hat{x} = 0$, and

$$\|\hat{E} + F\|_F^2 = \text{trace}[(\hat{E} + F)(\hat{E} + F)^H] = \|\hat{E}\|_F^2 + \|F\|_F^2.$$

This shows that the unique minimum in (5.3) is at $F = 0$, and (5.2) is the matrix part of the solution to (5.1). It follows that we can substitute

$$E = (c\gamma - Bx - s)x^H/x^Hx \tag{5.4}$$

in (5.1) to give the first alternative formulation of STLS :

$$\sigma_S^2 = \min_{s,x}\{\|s\|^2 + \|c\gamma - Bx - s\|^2/\|x\|^2\}, \tag{5.5}$$

since the constraints in (5.1) are automatically satisfied by E in (5.4).

Suppose $\hat{x}$ and $\hat{s}$ solve (5.5). We will show that

$$\hat{s} = \tilde{s} \equiv (c\gamma - B\hat{x})/(1 + \hat{x}^H\hat{x}), \quad \text{so that } \hat{E} = \hat{s}\hat{x}^H, \tag{5.6}$$

where the expression for $\hat{E}$ follows by substituting $\hat{s}$ in (5.2). Define

$$\tilde{d} \equiv c\gamma - B\hat{x} - \tilde{s} = \tilde{s}(1 + \|\hat{x}\|^2) - \tilde{s} = \tilde{s}\|\hat{x}\|^2. \tag{5.7}$$

Our proof that $\tilde{s}$ solves (5.5) will also show that $\tilde{d} = \hat{d}$ in (5.2). If $\hat{x}$ is known, we can replace x and s in (5.5) by $\hat{x}$ and $\tilde{s} + t$ to give

$$\begin{aligned} \sigma_S^2 &= \min_t \zeta(t), \quad \zeta(t) \equiv \{\|\tilde{s} + t\|^2 + \|\tilde{d} - t\|^2/\|\hat{x}\|^2\}, \\ \zeta(t) &= \|\tilde{s}\|^2 + \tilde{s}^H t + t^H\tilde{s} + \|t\|^2 + (\|\tilde{d}\|^2 - \tilde{d}^H t - t^H\tilde{d} + \|t\|^2)/\|\hat{x}\|^2. \end{aligned} \tag{5.8}$$

But from (5.7) $t^H\tilde{s} - t^H\tilde{d}/\|\hat{x}\|^2 = 0$, so the unique minimum of (5.8) is given by $t = 0$. Thus if $\hat{x}$ and $\hat{s}$ solve (5.5), (5.6) holds, giving with (5.5) our second alternative formulation of STLS (5.1):

$$\sigma_S^2 = \min_x \|c\gamma - Bx\|^2/(1 + \|x\|^2). \tag{5.9}$$

For the real case, this was derived in [11], see also [3, (3.21), p.57].

This is the result we need for our analysis of STLS, so we go no further with solving STLS here, but we will continue with the solution of the DLS formulation (1.4). Suppose w_D is the *vector* in the solution of the DLS problem

$$\sigma_D^2 \equiv \min_{G,w} \|G\|_F^2, \quad \text{subject to} \quad (B + G)w = c. \tag{5.10}$$

Doing the analysis (5.1)–(5.5) while insisting $s = 0$ proves the matrix part of the solution of this is

$$G = dw_D^H/w_D^H w_D, \qquad d \equiv c - Bw_D, \tag{5.11}$$

so that (5.10) simplifies to the unconstrained DLS formulation

$$\sigma_D^2 = \min_w \|Bw - c\|^2/\|w\|^2. \tag{5.12}$$

For the real case, this was stated in [3, (4.47), p.120], with a proof in Appendix B of that Thesis.

Now we derive a closed form DLS solution. We assume that (1.10) holds, and that $\rho > 0$ in (2.9). Using (2.4), (2.6), and remembering that $\sigma_1 \geq \cdots \geq \sigma_k \equiv \sigma_{min}(B) > 0$, consider the equation

$$\begin{aligned} 0 = \psi(\sigma^2) &\equiv c^H(BB^H - \sigma^2 I)^{-1}c \\ &= c^H U(U^H U_B \Sigma^2 U_B^H U - \sigma^2 I)^{-1} U^H c = \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma^2} - \frac{\rho^2}{\sigma^2}, \end{aligned} \tag{5.13}$$

where (1.10) ensures at least one of the α_i corresponding to $\sigma_{min}(B)$ will be nonzero. Clearly $\psi(\sigma^2)$ is unbounded below as $\sigma^2 \searrow 0$, and unbounded above as $\sigma^2 \nearrow \sigma_k^2$. Thus (5.13) has its minimal solution $\sigma_{_M}^2$ satisfying

$$0 < \sigma_{_M}^2 < \sigma_k^2 \quad \text{when (1.10) holds and } \rho > 0. \tag{5.14}$$

In this case we will show $\sigma_{_D}^2 = \sigma_{_M}^2$ and the solution of (5.12) is

$$w_{_D} \equiv (B^H B - \sigma_{_M}^2 I)^{-1} B^H c. \tag{5.15}$$

This $w_{_D}$ with $m = -1$ in (2.1) gives

$$Bw_{_D} - c = [B(B^H B - \sigma_{_M}^2 I)^{-1} B^H - I]c = \sigma_{_M}^2 (BB^H - \sigma_{_M}^2 I)^{-1} c. \tag{5.16}$$

So using (2.1) with $m = -2$, and (5.13),

$$\begin{aligned} &\sigma_{_M}^2 \|w_{_D}\|^2 - \|Bw_{_D} - c\|^2 \\ &= \sigma_{_M}^2 c^H B(B^H B - \sigma_{_M}^2 I)^{-2} B^H c - \sigma_{_M}^4 c^H (BB^H - \sigma_{_M}^2 I)^{-2} c \\ &= \sigma_{_M}^2 c^H (BB^H - \sigma_{_M}^2 I)^{-1} c = 0. \end{aligned}$$

This shows that $\sigma_{_M}$ and $w_{_D}$ are candidates for solving (5.12), since

$$\sigma_{_M}^2 = \|Bw_{_D} - c\|^2 / \|w_{_D}\|^2. \tag{5.17}$$

It remains for us to show that any nonzero change v to $w_{_D}$ *increases* this functional. Define

$$\begin{aligned} \phi(v) &\equiv \|B(w_{_D} + v) - c\|^2 - \sigma_{_M}^2 \|w_{_D} + v\|^2 \\ &= \|Bw_{_D} - c\|^2 + (Bw_{_D} - c)^H Bv + v^H B^H (Bw_{_D} - c) + \|Bv\|^2 \\ &\quad -\sigma_{_M}^2 (\|w_{_D}\|^2 + w_{_D}^H v + v^H w_{_D} + \|v\|^2). \end{aligned}$$

But if we use (5.15), we see that

$$v^H B^H (Bw_{_D} - c) - \sigma_{_M}^2 v^H w_{_D} = v^H [(B^H B - \sigma_{_M}^2 I) w_{_D} - B^H c] = 0.$$

This with (5.17) and (5.14) gives for nonzero v

$$\phi(v) = \|Bv\|^2 - \sigma_{_M}^2 \|v\|^2 > 0,$$

so $\sigma_{_M}^2 < \|B(w_{_D} + v) - c\|^2 / \|w_{_D} + v\|^2$ if $v \neq 0$. But this shows (5.17) *is* the optimum. When (1.10) holds and $\rho > 0$, $w_{_D}$ in (5.15) and G in (5.11) uniquely solve (5.10), and $\sigma_{_D}^2 = \sigma_{_M}^2$ is the minimum σ^2 in (5.13).

Relations (5.9) and (5.12) represent a formulation of the STLS and DLS problems (1.5) ((1.6)) and (1.4) analogous to the classical formulation of the LS problem

$$\|r\|^2 = \min_y \|c - By\|^2.$$

These were known before, but we proved them for the complex case assuming (1.10). The generalized total least squares approach used in [3–5] can be extended to complex data in a similar way.

6 Equivalence in the limit of STLS with LS, DLS

We need to prove that when $\gamma \to 0$ the STLS solution of the STLS formulation (1.6) becomes the LS solution, and when $\gamma \to \infty$ the STLS formulation corresponds to DLS (1.4). For DLS this seems reasonable, since for any positive bounded γ, (1.5) and (1.6) are equivalent with the substitutions $s \equiv \tilde{s}\gamma$, $z \equiv \tilde{z}$ and $E \equiv \tilde{E}$. Clearly (1.5) becomes DLS as $\gamma \to \infty$, so it appears that (1.6) becomes DLS too. Alternatively for any positive bounded γ we can rewrite (1.6) as

$$\text{STLS distance} \;\equiv\; \min_{s,E,z} \|[s,E]\|_F \quad \text{s. t.} \quad (B+E)z = c - s/\gamma.$$

As $\gamma \to \infty$ it appears that we can take $s = 0$, corresponding to DLS. But neither of these arguments is rigorous, so we resort to the closed form solution of (1.6) to prove these equivalences. For the case of real data, the basic ideas for $\gamma \to 0$ were given in [11, Corollary 4.2], and more precisely in [22, Theorem 3.1]. This section is thus an extension of these works.

We will assume (1.10) holds, so in particular (1.11) holds, and B has full column rank. For the case of (1.11), [16, Thm. 2.7] showed (for the real case with $\gamma = 1$) that the closed form TLS solution of (1.6) is, with $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$,

$$z(\gamma)\gamma = [B^H B - \sigma^2(\gamma) I]^{-1} B^H c\gamma. \tag{6.1}$$

If $v \equiv (\tilde{v}^T, \nu)^T$ with $\nu \neq 0$ is a right singular vector of $[B, c\gamma]$ corresponding to $\sigma(\gamma)$, then we know $z(\gamma)\gamma = -\tilde{v}/\nu$. But $\{[B, c\gamma]^H [B, c\gamma] - \sigma^2(\gamma) I\} v = 0$, and the first k elements of this give (6.1). This could also have been proven from the formulation (5.9) (remembering $x \equiv z\gamma$), see the proof of (5.15) from (5.12).

The definition $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$ shows $\lim_{\gamma \to 0} \sigma(\gamma) = 0$, so

$$\lim_{\gamma \to 0} z(\gamma) = (B^H B)^{-1} B^H c = \text{ the LS solution } y \text{ for (1.2).} \tag{6.2}$$

Next we relate the *distances* when $\gamma \to 0$. The STLS distance is the smallest singular value $\sigma(\gamma)$ of $[B, c\gamma]$, see (1.9), and so is the smallest solution $\sigma \geq 0$ of (4.1). If we define $M \equiv I - B(B^H B)^{-1} B^H = M^H = M^2$, (4.3) shows that for the LS residual $r = c - By = Mc$ in (1.2),

$$\begin{aligned} \lim_{\gamma \to 0} \frac{\text{STLS distance in (1.6)}}{\gamma} &= \lim_{\gamma \to 0} \frac{\sigma(\gamma)}{\gamma} = \sqrt{c^H M c} = \sqrt{r^H r} \\ &= \text{ LS distance in (1.2).} \end{aligned} \tag{6.3}$$

This completes our proof that as $\gamma \to 0$, the STLS solution of the STLS formulation (1.6) becomes the LS solution, and the STLS distance divided by γ becomes the LS distance.

For the DLS equivalence we have the added difficulty of unknown $\sigma(\infty) \equiv \lim_{\gamma\to\infty} \sigma(\gamma)$. Taking the limit $\gamma \to \infty$ in (4.4) shows that the STLS distance $\sigma(\infty)$ must be the smallest positive solution $\sigma_M < \sigma_{min}(B)$ of (5.13), see (5.14). But this means that $\sigma(\infty)$ is also the DLS distance σ_D. Also from (6.1) and (5.15) we see in the limit the STLS solution $z(\gamma)$ of STLS (1.6) becomes the solution vector w_D of (1.4). Summarizing:

$$\lim_{\gamma\to\infty} \text{STLS distance} = \text{DLS distance}, \qquad \lim_{\gamma\to\infty} z(\gamma) = w_D. \tag{6.4}$$

This completes the proof that when $\gamma \to \infty$, the STLS formulation (1.6) corresponds exactly to the DLS formulation (1.4).

7 Conditions for meaningful solutions

Here we show when the problem formulations (1.3)–(1.6) are not good for solving $Bx \approx c$ in (1.1). Because (1.3) is a special case of (1.6), and (1.5) is equivalent to (1.6) for bounded $\gamma > 0$, we need only consider the DLS (1.4) and STLS (1.6) formulations. Of course the LS formulation (1.2) always has a meaningful solution.

We first show that (1.3)–(1.6) are not good when n by k B does not have rank k. The functional in each case is nonnegative. Suppose c does not lie in the range of B, so the functional is positive. For the STLS problem an alternative formulation is (5.9) with $x \equiv z\gamma$. But taking *any* x and adding to it a large enough component in the null space of B will make the functional in (5.9) arbitrarily close to zero. A similar argument holds for DLS via (5.12). Thus the formulations should at least demand the solution vectors be orthogonal to the null space. It is preferable to eliminate the null space.

We argue that (1.3)–(1.6) are best restricted to problems of the form (1.1) satisfying (1.10), that is,

$$\text{the } n \times k \text{ matrix } B \text{ has rank } k, \text{ and } \; c \not\perp \mathcal{U}_{min},$$

where $\mathcal{U}_{min}$ is the left singular vector subspace of B corresponding to $\sigma_{min}(B)$. If this holds, then Theorem 3.1 shows (1.11) holds, see (3.7), and we have the standard, meaningful solutions. But if it does not hold, we will show these four formulations either have solutions that do not make sense as solutions to (1.1), or contain data which is irrelevant to the solution and could cause unnecessary inaccuracies with finite precision computation. It is rarely possible to tell ahead of time which is the case, and we recommend that the formulations (1.3)–(1.6) each come with the proviso that B and c must obey (1.10).

Suppose the data can be unitarily transformed, see (2.3), so that

$$\left[\tilde{c}\middle\|\tilde{B}\right] = P^H \left[c\middle\|BQ\right] = \left[\begin{array}{c|c||c|c} c_1 & B_{11} & 0 \\ \hline 0 & 0 & B_{22} \end{array}\right]. \tag{7.1}$$

Note that in this case the SVD problems of $[\tilde{c}, \tilde{B}]$ and of $\tilde{B}$ each split into two independent SVD problems. The approximation problem $Bx \approx c$ then represents two *independent* approximation problems:

$$B_{11}x_1 \approx c_1, \qquad B_{22}x_2 \approx 0, \qquad x \equiv Q \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{7.2}$$

in that the solution to each of these has no effect upon, and can be found independently of, the other. Each of (1.2)–(1.6) applied to $B_{22}x_2 \approx 0$ gives zero distance and $x_2 = 0$, an eminently meaningful solution.

If (1.10) does not hold, a transformation (2.3) clearly exists giving (7.1) where B_{22} contains all the singular values of B equal to $\sigma_{min}(B)$. In the worst case we will show, see (7.3)–(7.6), that (1.3)–(1.6) applied directly to the combined problem $Bx \approx c$ can give meaningless solutions. But even in the best case these minimum singular values are irrelevant, and should be removed from the problem, lest rounding errors effectively introduce a nonzero vector below c_1 in (7.1), and so cause these irrelevant singular values to contaminate the solution. This is more likely the smaller $\sigma_{min}(B)$ is. Although (1.2) in theory gives $x_2 = 0$, this last comment suggests we might gain by insisting on (1.10) for LS too. The rest of this section will further develop our argument justifying the fundamental role of (7.1).

The practical reader, who agrees that problems $Bx \approx c$ with data that can be transformed to (7.1) should be solved as two independent problems, can ignore the rest of this section and go to Sect. 8. That shows how transformations may be applied to produce $[c_1, B_{11}]$ in (7.1) that cannot be reduced any further.

We examine TLS. From (5.9) with $\gamma = 1$ we see that (1.3) corresponds to

$$(\text{TLS distance})^2 = \min_x \|Bx - c\|^2/(1 + \|x\|^2). \tag{7.3}$$

Suppose x_1 solves

$$\sigma_{11}^2 \equiv \min_{s,E,x} \|[s, E]\|_F^2 \quad \text{s. t.} \quad (B_{11} + E)x = c_1 - s,$$

then from (7.3)

$$\sigma_{11}^2 = \|B_{11}x_1 - c_1\|^2/(1 + \|x_1\|^2). \tag{7.4}$$

Suppose (to give the worst case mentioned above, see (1.12) as a numerical example of this),

$$\sigma_k \equiv \sigma_{min}(B_{22}) < \sigma_{11}, \tag{7.5}$$
$$B_{22}v = u\sigma_k, \qquad u^H B_{22} = \sigma_k v^H, \qquad v^H v = u^H u = 1.$$

We now show taking $x = Q \begin{pmatrix} x_1 \\ v \end{pmatrix}$ in (7.3) with (7.1) gives a functional value less than σ_{11}^2, so $Q \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$ does not minimize (7.3) when (7.1) and (7.5) hold. Using (7.4), the functional in (7.3) becomes

$$\frac{\|B_{11}x_1 - c_1\|^2 + \sigma_k^2}{1 + \|x_1\|^2 + 1} = \frac{\sigma_{11}^2 + \sigma_k^2/(1 + \|x_1\|^2)}{1 + 1/(1 + \|x_1\|^2)} < \sigma_{11}^2, \tag{7.6}$$

since $\sigma_k^2 < \sigma_{11}^2$. Thus the meaningful solution $x = Q \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$ does not solve the combined problem correctly using the formulation (1.3). Using (5.12) instead of (7.3) shows the DLS formulation (1.4) has exactly the same weakness.

The case for STLS is more dangerous still, since Theorem 3.1 showed that when (1.10) does not hold (giving (7.1)), we could have $\sigma_{min}([B, c]) < \sigma_{min}(B)$, but $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ for some γ, see Corollary 3.1. It can be shown that this also allows the possibility that (7.5) holds — the worst case above.

The fundamental difficulty revealed here in a clear way by the form (7.1) has been noticed and described in various different ways before. Van Huffel and Vandewalle [16] developed a rigorous and fascinating, but quite complicated theory allowing them to construct a meaningful solution to the approximation problem $Bx \approx c$. Later workers assumed (1.11), and applied this theory in [16] directly to the STLS problem.

We argue for the criterion (1.10) for all the formulations (1.3)–(1.6), since unlike (1.11) this criterion is independent of γ, but it ensures (1.11) holds; it is simpler than (1.11), it leads to a clear and consistent theory, and it ensures that the minimum singular value of B is relevant to the solution. This argument is easy to accept when we realize there is an elegant transformation which produces the minimally dimensioned core problem obeying (1.10) from any given $[c, B]$.

8 The core problem within $Bx \approx c$

Here we answer the following important question. Given a general n by k matrix B and n-vector c, how can the data be transformed so that the problem $Bx \approx c$ splits into two independent problems as in (7.1) and (7.2), giving a trivial problem $B_{22}u_2 \approx 0$ of maximal dimensions, and the minimally dimensioned core problem $B_{11}u_1 \approx c_1$ satisfying (1.10). This last condition ensures each of the formulations (1.2)–(1.6) has a unique meaningful solution, which can be expressed via a simple closed form.

The answer we give immediately suggests a very useful direct (that is, not iterative) practical computation, but we only give the theoretical (exact precision) version here.

Remember the STLS solution requires some knowledge of the SVD of $[c\gamma, B]$. Our reduction leads to a core problem from which this SVD information can be computed easily and efficiently. Choose unitary matrices P and Q to produce the following real bidiagonal matrix, see for example [12, §5.4.3–5, pp. 251–254]. In the usual case of $n \times k$ B with $n > k$ we obtain, where a blank means a zero element, and the bottom 0 could represent a zero vector or be nonexistent:

$$[\tilde{c}, \tilde{B}] \equiv P^H \left[\, c \middle| B \,\right] \begin{bmatrix} 1 & \\ & Q \end{bmatrix} = \left[\begin{array}{c|cccc} \gamma_1 & \beta_1 & & & \\ & \gamma_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \gamma_k & \beta_k \\ & & & & \gamma_{k+1} \\ & & & & 0 \end{array}\right]. \tag{8.1}$$

Notice how the SVD of $[c\gamma, B]$ can quickly be computed from this bidiagonal form for any choice of γ, see for example [12, §8.6.2, pp. 452–456].

There are two ways this algorithm can terminate prematurely, so we describe the relevant partial reductions. Initially we design unitary P_1 so that $P_1^H c = e_1\gamma_1$, then unitary Q_1 so that $(e_1^T P_1^H B)Q_1 = \beta_1 e_1^T$, *etc.*. After the first half of the j-th step, $j \le k+1$, we have

$$P_j^H \cdots P_2^H P_1^H \left[\, c \middle| BQ_1Q_2\cdots Q_{j-1} \,\right] = \left[\begin{array}{c|cc|c} \gamma_1 & \beta_1 & & \\ & \ddots & \ddots & \\ \hline & & \gamma_j & \times\ \times \\ & & 0 & \times\ \times \\ & & 0 & \times\ \times \end{array}\right]. \tag{8.2}$$

Stop if $\gamma_j = 0$, since then the exact solution (zero STLS distance) can be found by discarding columns $j+1, \ldots, k$, and rows $j, \ldots, n$ of the transformed $[c, B]$. Otherwise if $j \le k$, choose unitary Q_j so that

$$P_j^H \cdots P_2^H P_1^H \left[\, c \middle| BQ_1Q_2\cdots Q_j \,\right] = \left[\begin{array}{c|cc|c} \gamma_1 & \beta_1 & & \\ & \ddots & \ddots & \\ \hline & & \gamma_j & \beta_j\ 0 \\ \hline & & 0 & \times\ \times \\ & & 0 & \times\ \times \end{array}\right]. \tag{8.3}$$

Stop if $\beta_j = 0$, discarding columns $j+1, \ldots, k$ and rows $j+1, \ldots, n$ of the transformed $[c, B]$, leaving a STLS problem with a j by j upper bidiagonal matrix $[\tilde{c}, \tilde{B}]$. In both these terminations we assume

$$\gamma_i\beta_i \neq 0, \quad i = 1, \ldots, j-1. \tag{8.4}$$

Notice that in each of these early terminations, direct transformations have split the SVD of $[c\gamma, B]$ (and of B) into two independent SVDs.

The computations described in [12, §5.4.3–5, pp. 251–254] are designed for dense matrices. If we have large sparse $[c, B]$, then we could consider the iterative bidiagonalization suggested by Golub and Kahan in [9], see also [17]. This iterative bidiagonalization is the basis for the valuable LSQR algorithm in [18,19] which solves large sparse LS (as well as consistent) problems. The bidiagonalization "Bidiag 1" of [18, p.47] is used for the LSQR algorithm (and code) in [19]. In theory after j and a half steps, "Bidiag 1" applied to $[c, B]$ ($[b, A]$ in [18]) produces the first $j+1$ columns of P, the first j columns of Q, and the leading $j+1$ by $j+1$ block of the right-hand side in (8.1). Åke Björck [1, §7.6.5, pp.310-311] suggested applying the iterative bidiagonalization (as in LSQR) to the TLS problem, see also [7, Section 4.1]. Now we see this approach is also applicable to solving the STLS problem, as well as (at least in theory) delivering the core problem, for any large sparse linear system $Bx \approx c$. The adaptation of LSQR for solving large sparse STLS or DLS problems using finite precision computations will be further investigated. See Sect. 9 for the DLS solution using (8.1).

The main theoretical importance of the reduction (8.1) here is that if (8.4) holds, then our main criterion (1.10) holds for the reduced bidiagonal matrix. If $\gamma_j = 0$ this is the bidiagonal matrix in the top left corner of the transformed $[c, B]$ in (8.2); or if $\gamma_j \neq 0$, it is the bidiagonal matrix in the top left corner of the transformed $[c, B]$ in (8.3) if $\beta_j = 0$. Also (1.10) holds for $[c, B]$ in (8.1) if the algorithm is not stopped prematurely. We now prove this.

Theorem 8.1 *Suppose n by k B has SVD $B = \sum_{i=1}^{k} u_i \sigma_i v_i^H$, and there exist unitary matrices P and Q giving $[\tilde{c}, \tilde{B}] \equiv P^H [c, BQ]$ where*

$$(8.5)\quad \left[\, \tilde{c} \middle| \tilde{B} \,\right] \equiv \left[\begin{array}{c|cccc} \gamma_1 & \beta_1 & & & \\ & \gamma_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \gamma_k & \beta_k \\ & & & & \gamma_{k+1} \\ & & & & 0 \end{array}\right], \quad \gamma_j \beta_j \neq 0, \quad j = 1, \ldots, k.$$

Then we have a stronger condition than (1.10) for this c and B:

$$(8.6)\qquad \textit{rank}(B) = k; \qquad c^H u_i \neq 0, \quad i = 1, \ldots, k.$$

The k singular values of B are distinct and nonzero; the $k+1$ singular values of $[c, B]$ are distinct, and all nonzero if and only if $\gamma_{k+1} \neq 0$.

Proof. Clearly $\tilde{B}$ and B have the same singular values, as do $[\tilde{c}, \tilde{B}]$ and $[c, B]$, and $\tilde{B} = P^H B Q$ has the SVD $\tilde{B} = \sum_{i=1}^{k} \tilde{u}_i \sigma_i \tilde{v}_i^H \equiv$

$\sum_{i=1}^{k} P^H u_i \sigma_i v_i^H Q$, so

$$c^H u_i = c^H P P^H u_i = \tilde{c}^H \tilde{u}_i, \quad i = 1, \ldots, k.$$

Write $\tilde{B} \equiv [b_1, B_1]$, then $\tilde{B}^H \tilde{B}$ is $k \times k$ tridiagonal with nonzero next to diagonal elements, and $B_1^H B_1$ remains when the first row and column are deleted. Thus the eigenvalues of $B_1^H B_1$ strictly separate those of $\tilde{B}^H \tilde{B}$, see [23, Ch.5, §37, p.300], and the singular values of B_1 strictly separate those of $\tilde{B}$. Thus $\tilde{B}$, and so B, has distinct singular values (see also [21, Lemma 7.7.1, p.134]). A similar argument holds for $[c, B]$. B clearly has rank k, and $[c, B]$ has rank $k+1$ if and only if $\gamma_{k+1} \neq 0$. Suppose σ is a singular value of $\tilde{B}$ with singular vectors u and v such that

$$\tilde{c}^H u = \bar{\gamma}_1 e_1^T u = 0, \quad u\sigma = \tilde{B} v, \quad \sigma v^H = u^H \tilde{B}, \quad \|u\| = \|v\| = 1,$$

then $0 = e_1^T u \sigma = e_1^T \tilde{B} v = \beta_1 e_1^T v$, and $e_1^T v = 0$. Writing $v = \begin{pmatrix} 0 \\ q \end{pmatrix}$ shows

$$\tilde{B} v = B_1 q = u\sigma, \quad u^H B_1 = \sigma q^H, \quad \|u\| = \|q\| = 1,$$

so σ is also a singular value of B_1. This is a contradiction since the singular values of B_1 strictly separate those of $\tilde{B}$, so (8.6) holds. □

Thus we need not derive results for the most general possible $[c\gamma, B]$. We can instead assume (1.10). Any more general $Bx \approx c$ problem can be reduced to a core problem that satisfies (8.6) (and so (1.10)) by applying the reduction (8.1) and stopping at the first zero γ_j or β_j. Suppose the resulting core data is $[c_1, B_{11}]$, see (7.1). Then the theorem also showed that B_{11} has no multiple singular values, so any singular value repeats must appear in B_{22}.

We do not insist on (8.6), because a problem only satisfying (1.10) will in theory give the same solution and distance as it would if it were reduced to one satisfying (8.6). This can be seen for example by using the transformations of (2.6) in (6.1) to give

$$z(\gamma) = V[\Sigma^2 - \sigma^2(\gamma) I]^{-1} \Sigma U_B^H c.$$

Clearly when (1.10) holds and $\alpha_i \equiv u_i^H c = 0$ for some i, $1 \leq i \leq k$, the corresponding σ_i in Σ does not contribute to the solution, and need not, at least in theory, be eliminated. In practice it is preferable to carry out the reduction (8.1) leading to (8.6), see Sect. 9.

9 Computing STLS and DLS solutions

In order to compute either STLS solutions or the DLS solution for given data $[c, B]$, we recommend first carrying out a reduction of the form (8.1) to the core problem in Sect. 8 — unless there are clear reasons for not doing so. The reasons for doing so are hard to reject. For general data we will not know if the formulations (1.2)–(1.6) have unique meaningful solutions, but the reduction will give us a subproblem for which this is so. Even if we know the original data satisfies (1.10), it is (from the computational point of view) highly preferable to remove all the irrelevant information from our data as early in the solution process as possible, and this is exactly what the transformation (8.1) does. In any case we still need some sort of SVD of the data, and this will usually first perform a reduction as costly as that in (8.1). But (8.1) allows us to find the SVD of $[c\gamma, B]$ easily for different choices of γ and so is the obvious choice. There are excellent fast and accurate algorithms for finding all or part of the SVD of (8.1) with γ_1 replaced by $\gamma_1\gamma$. We can find just the smallest singular value and its singular vectors, from which the solution vector $z(\gamma)$ can be simply attained, see (6.1) and the two sentences following it. If we have some idea of the accuracy of our data, then when we use numerically reliable unitary transformations in (8.1), we will have a good idea of what element of (8.1) (if any) we can set to zero to obtain one of the stopping criteria as soon as possible in (8.1)–(8.4). Thus the crucial decisions can be made *before* any SVD computations are carried out. This is more efficient, but it is almost certainly more reliable to make such decisions from unitary transformations of the original data than from the elements of singular vectors, (see for example [16, p.23] or (10.1) later). The remaining computations for STLS are fairly obvious. Finally (8.1) leads to a solution to the DLS problem (1.4), which we now describe.

We saw from (5.13) and (5.15) that when (1.10) holds, the solution $w_{_D}$ and distance $\sigma_{_D}$ of the DLS problem (1.4) are

$$w_{_D} \equiv (B^HB - \sigma_{_M}^2 I)^{-1}B^Hc, \qquad \sigma_{_D} = \sigma_{_M} \geq 0, \tag{9.1}$$

where $\sigma_{_M}^2$ is the minimal solution σ^2 of

$$0 = \psi(\sigma^2) \equiv c^H(BB^H - \sigma^2 I)^{-1}c. \tag{9.2}$$

Now suppose that the core part $[\tilde{c}, \tilde{B}]$ of the transformed $[c, B]$ has the form in (8.5). This obviously applies to the usual case where the reduction does not stop prematurely, but it also applies to the core problem in (8.2) or (8.3) by replacing k here by j. We will solve the DLS problem for this reduced, or core data. Now Theorem 8.1 proved (1.10) holds. If $\gamma_{k+1} = 0$ the DLS distance is zero, and the solution is obvious. Otherwise, writing

$$[\tilde{c}|\tilde{B}] \equiv \left[\begin{array}{c|c} \gamma_1 & \beta_1 e_1^T \\ \hline 0 & B_2 \end{array}\right],$$

we see for this reduced problem that σ_M^2 must be the minimal solution σ^2 of

$$\begin{aligned} 0 &= \tilde{c}^H(\tilde{B}\tilde{B}^H - \sigma^2 I)^{-1}\tilde{c} = |\gamma_1|^2 e_1^T(\tilde{B}\tilde{B}^H - \sigma^2 I)^{-1}e_1 \\ &= |\gamma_1|^2 \det(B_2 B_2^H - \sigma^2 I)/\det(\tilde{B}\tilde{B}^H - \sigma^2 I), \end{aligned}$$

since for nonsingular A, $A^{-1} = \text{adjugate}(A)/\det(A)$, see for example [23, (36.3), p.39]. But because the γ_i and β_i in (8.5) are nonzero for $i = 1, \ldots, k$, no singular value of B_2 is a singular value of $\tilde{B}$ (by strict separation, see the proof of Theorem 8.1), so σ_M must be the smallest singular value of the nonsingular bidiagonal matrix B_2. This is relatively easy to find, see for example [12, §8.6.2, pp. 452–456].

Now let v be the right singular vector of B_2 corresponding to σ_M, then $e_1^T v \neq 0$ (otherwise σ_M would also be a singular value of $\tilde{B}$) and

$$w_D = v\gamma_1/(\beta_1 e_1^T v), \qquad \sigma_D = \sigma_M = \sigma_{min}(B_2), \tag{9.3}$$

are the DLS solution and distance in (1.4) for the reduced data $[\tilde{c}, \tilde{B}]$. We see w_D satisfies the equivalent of (9.1) for this reduced data, since

$$\begin{aligned} &\tilde{B}^H\tilde{B} = |\beta_1|^2 e_1 e_1^T + B_2^H B_2, \quad \tilde{B}^H\tilde{c} = e_1\bar{\beta}_1\gamma_1, \quad B_2^H B_2 v = v\sigma_M^2, \\ &(\tilde{B}^H\tilde{B} - \sigma_M^2 I)w_D = |\beta_1|^2 e_1 e_1^T v\gamma_1/(\beta_1 e_1^T v) = e_1\bar{\beta}_1\gamma_1 = \tilde{B}^H\tilde{c}. \end{aligned}$$

10 "Generic" TLS problems

It is useful in the light of our new knowledge to compare (1.10) with the criterion for "generic" TLS [16] as applied to STLS (1.6), and we do this now. We simplify the results of [16] to the case of a single right hand side c in (1.6), but allow $\gamma \neq 1$ in order to extend their results to the STLS problem.

Van Huffel and Vandewalle used the following definition of the "generic" (S)TLS problem in [16, p.23]. Consider the singular value decomposition of the extended matrix $[B, c\gamma]$ for some $\gamma > 0$

$$[B, c\gamma] = U'\Sigma' V'^H, \text{ with } \Sigma' \equiv [\text{diag}(\sigma'_1, \ldots, \sigma'_{k+1}), 0]^T,$$

for $n \times n$ unitary U', $(k+1) \times (k+1)$ unitary $V' \equiv [v'_1, \ldots, v'_{k+1}]$ with elements ν'_{ij}, and $n \times (k+1)$ Σ', with $\sigma'_1 \geq \ldots \geq \sigma'_{k+1} \geq 0$. The STLS problem (1.6) is "generic" if for $j \leq k$ defined so that

$$\sigma'_j > \sigma'_{j+1} = \cdots = \sigma'_{k+1}, \quad \text{we have } [\nu'_{k+1,j+1}, \ldots, \nu'_{k+1,k+1}] \neq 0. \tag{10.1}$$

This includes the $\text{rank}(B) < k$ case. The TLS solution of a "generic" problem is called the "generic" TLS solution, and can be computed by the algorithm of Golub and Van Loan [11].

Note (10.1) used the SVD of $[B, c\gamma]$, whereas (1.10) used that of B. Let $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$ be the singular values of B, see (2.4). The interlacing

property for the eigenvalues of $[B, c\gamma]^H[B, c\gamma]$ and of $B^H B$ [23, Ch2, §47, pp. 103–4] tells us that

$$\sigma'_1 \geq \sigma_1 \geq \cdots \geq \sigma'_j \geq \sigma_j \geq \sigma'_{j+1} \geq \sigma_{j+1} \geq \cdots \geq \sigma'_k \geq \sigma_k \geq \sigma'_{k+1}. \tag{10.2}$$

In their Lemma 3.1 and Corollary 3.4 in [16, pp. 64-5], Van Huffel and Vandewalle proved another necessary and sufficient condition for (1.6) to be "generic". For a given $\gamma > 0$ this condition can be stated in the following way: if $j \leq k$ is defined so that $\sigma'_j > \sigma'_{j+1} = \cdots = \sigma'_{k+1}$ then the STLS problem (1.6) is "generic" if and only if

$$\sigma_j > \sigma'_{j+1} = \cdots = \sigma'_{k+1}. \tag{10.3}$$

(In fact they considered $\gamma = 1$, and proved that (10.3) is equivalent to $[\nu'_{k+1,j+1}, \ldots, \nu'_{k+1,k+1}] \neq 0$). With (10.2), (10.3) becomes

$$\sigma'_j \geq \sigma_j > \sigma'_{j+1} = \sigma_{j+1} = \cdots = \sigma'_k = \sigma_k = \sigma'_{k+1}, \tag{10.4}$$

meaning the STLS problem (1.6) is "generic" if and only if

$$\sigma_{min}(B) > \sigma_{min}([B, c\gamma]) \quad \text{when } \sigma_{min}([B, c\gamma]) \text{ is simple,} \tag{10.5}$$

$$\text{or, when } \sigma_{min}([B, c\gamma]) \text{ is multiple:} \tag{10.6}$$

$$\text{multiplicity}(\sigma_{min}([B, c\gamma])) > \text{multiplicity}(\sigma_{min}(B)). \tag{10.7}$$

Since (10.5) is just (1.8), this new formulation (for a single right-hand side c) emphasizes that the purpose for using the "generic" TLS criterion [16] is to provide solutions where possible in the subtle case where $[B, c]$ has a multiple minimum singular value.

Our criterion (1.10) is far more brutal than (10.5)–(10.7) — it rejects some cases where (10.5) holds, see Corollary 3.1, and all cases where (10.6) holds. This last because (10.6) implies $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$, so that $a_2 = 0$ in Corollary 3.1, and (1.10) does not hold. Yet the intentions and outcomes of the criteria in [16] and our criterion (1.10) are not very different. In particular note that in nearly all practical problems our restrictive criterion (1.10) *will* hold, and so (10.5) will also hold, and any differences in the criteria apply to a small number of problems at best.

If our data $[c, B]$ does not meet the criterion (1.10), we do not wish to reject it — we want to transform it to obtain a reduced problem that satisfies (1.10). In fact, we go even further. We suggest that the data $[c, B]$ should *always* be transformed to a reduced system (7.1) with B_{11} of minimal dimensions. If we do this via the approach in Sect. 8, we discard *all* the components of the SVD of B that are irrelevant to the main approximation problem in $Bx \approx c$. But this is partly what is done in [16]. In the case of isolated $\sigma_{min}([B, c\gamma])$, (10.5) shows if $\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ the "generic" STLS solution does not exist. Moreover, in Theorem 3.1

$\sigma_{min}([B, c\gamma]) = \sigma_{min}(B)$ implies $a_2 = 0$ and therefore $\sigma_{min}(B) \equiv \sigma_k$ makes no contribution to the main approximation problem in $Bx \approx c$ (which is why we eliminate it). For this case Van Huffel and Vandewalle [16, §3.4.1] proposed a nongeneric TLS solution that effectively also discards $\sigma_{min}(B)$. So in this case of isolated $\sigma_{min}([B, c\gamma])$ the intentions are the same and the outcomes are similar.

When (10.6) and (10.7) hold, the "generic" TLS solution exists but it is not unique, and Van Huffel and Vandewalle construct the minimum norm TLS solution [16, Thm. 3.7]. The "generic" solution exists because the minimum eigenvalue of

$$\begin{bmatrix} \Sigma_1^2 & \Sigma_1 a_1 \gamma \\ \gamma a_1^H \Sigma_1 & \gamma^2 (a^H a + \rho^2) \end{bmatrix}$$

is the minimum eigenvalue of $N^H N$ in (3.6) (see (10.7)). It is not unique because this minimum eigenvalue is also equal to the unwanted minimum eigenvalues $\sigma_{j+1}^2 = \cdots = \sigma_k^2$ of $B^H B$ in Theorem 3.1. This shows this form of "generic" problem is extremely unlikely — and in STLS problems any minute change in γ will upset this equality, see Corollary 4.1. Thus such problems (even TLS problems) are not generic in the usual sense of the word. One definition of generic is 'general, not specific or special', so we would expect a generic problem to satisfy (8.6) and so (10.5), but certainly not (10.6) and (10.7).

For (10.6)–(10.7) our approach would first get rid of *all* the unwanted singular values of B (not only those equal to $\sigma_{min}(B)$), leading to a unique solution of a reduced problem. This will provide unique solutions in all cases. So again the intentions are the same, though the outcomes may differ. The philosophy here is to reduce the problem to one of minimal dimensions with a unique meaningful solution. The tendency in [16] was more to seek such solutions without such a reduction — but by applying orthogonality conditions to the solution instead.

In summary, the stronger but simpler criterion (1.10) together with the concept of the core problem in Sect. 8 has allowed us to achieve simply, clearly, thoroughly, and with one uniform approach, what [16] sought to do, and partially achieved through the ingenious use of several techniques.

11 Summary and conclusion

The total least squares (TLS) problem for the matrix B and the right-hand side $c\gamma$, $\gamma > 0$, represents a formulation (1.6) of the scaled TLS (STLS) problem. For positive bounded γ it is equivalent to the usual formulation (1.5) of the STLS problem for B and c, where the relative sizes of the corrections in B and c are determined by γ. Our results bring, we believe,

a new view to the theoretical foundations of STLS problems, and a new understanding of these, including TLS problems, as well as of data least squares (DLS) problems.

In Theorem 3.1 we proved the necessary and sufficient condition for $\sigma_{min}(B) = \sigma_{min}([c\gamma, B])$, which reveals that this undesirable event must be rare in practical STLS problems. This is a general matrix theory result — it gives a necessary and sufficient condition for preserving the smallest singular value of a matrix while appending or deleting a column. This led us to a new criterion for scaled total least squares (STLS) problems, and we showed that when this criterion is not obeyed, the standard formulations can lead to computationally risky, or even meaningless solutions. We have given algebraic proofs of alternative formulations of the STLS and DLS problems, and proven the form of the DLS solution for the case of possibly complex data. We proved how our formulation (1.6) of the STLS problem corresponds to LS as γ goes to zero, and to DLS as γ goes to infinity. We showed how to reduce any general LS, STLS or DLS problem to the core and transparent problem where the system matrix B has full column rank and distinct singular values, and the right-hand side c is not orthogonal to any left singular vector of B. This removes any irrelevant information from the data and it more than obeys our criterion (1.10). We briefly indicated new algorithms for solving STLS and DLS problems, when the data $[c, B]$ is small and dense, and when it is large and sparse.

Van Huffel and Vandewalle [16, p.19] call the TLS problem "basic" when it has only one right-hand side vector and a unique solution. If $[B, c]$ satisfies our criterion (1.10) then the STLS formulation (1.6) yields a unique solution for any $\gamma > 0$. The LS and DLS formulations then also yield unique solutions. The reduction in Sect. 8 yields a core problem that has minimal dimensions and satisfies (8.6). This last criterion is even stronger than (1.10). So perhaps we could call such problems, or the general approximation problem $Bx \approx c$, "basic" when $[B, c]$ satisfies (1.10), and "core" when it satisfies (8.6).

If $[B, c]$ satisfies (1.10), then in theory there is no need to perform the reduction to the minimally dimensioned core problem satisfying (8.6). Both the original problem satisfying (1.10) and the reduced minimally dimensioned core problem have identical solutions and distances. Computationally however, it seems always desirable to perform the proposed reduction.

Throughout this paper we have only dealt with problems $Bx \approx c$ with one right-hand side vector c. For this case Sect. 10 developed a new formulation (10.5)–(10.7) of the existence condition for the "generic" TLS solution in [16]. We used this to show that reducing the problem to one which satisfied the simpler but stronger criterion (1.10) (or preferably the even stronger (8.6)), then solving this problem, achieved everything that this

difficult "generic" concept and its related solution methods did, and more. In Sect. 9 we argued that the reduction in Sect. 8 to the core problem be applied to any STLS problem unless there is a good reason not to do so. Thus for problems with one right-hand side, if we use this reduction there is no need for the subtle and sophisticated concept of "generic TLS" and the related solution methods for special cases introduced in [16]. Perhaps this reduction and the criterion (1.10) can be developed to apply to problems with more than one right-hand side?

As we mentioned earlier, this paper deals with exact relationships. Our next paper [20] follows on from this, and will deal with bounds and the LS–STLS relationship when $\gamma > 0$. A crucial element in that is the amount by which $\sigma_{min}([c\gamma, B])$ is less than $\sigma_{min}(B)$, and many of the results will depend on $\delta(\gamma) \equiv \sigma_{min}([c\gamma, B])/\sigma_{min}(B)$.

Acknowledgements. The authors are indebted to Sabine Van Huffel and Åke Björck for their gracious help and their many valuable suggestions on this work. The insights we gained from Sabine greatly improved our understanding, and led us to present the content here as a separate paper. We can but marvel at the thoroughness, depth and rigour of the treatment of the topic by Van Huffel and Vandewalle in [16]. Discussions with Xiao-Wen Chang were very helpful for the analysis in Sections 5 and 6. Other suggestions by Xiao-Wen Chang, Gene Golub, Volker Mehrmann and Mike Saunders were also very helpful.

References

1. Å. Björck: Numerical Methods for Least Squares Problems. SIAM Publications, Philadelphia PA 1996
2. J. R. Bunch, C. P. Nielsen: Updating the singular value decomposition. Numerische Mathematik **31**, 111–129 (1978)
3. G. Cirrincione: A Neural Approach to the Structure of Motion Problem. PhD thesis, LIS INPG Grenoble 1998
4. G. Cirrincione, G. Ganesan, K. V. S. Hari, S. V. Huffel: Direct and Neural techniques for the data least squares problem. Proceedings of the 14th International symposium on Mathematical theory of networks and systems (MTNS 2000), Perpignan, France, June 19–23, 2000 (To appear)
5. G. Cirrincione, A. Premoli, M. L. Rastello: Generalization and scheduling of TLS problems. (1999) (In preparation)
6. L. Elsner, C. He, V. Mehrmann: Minimization of the norm, the norm of the inverse and the condition number of a matrix by completion. Numerical Linear Algebra Appl. **2**, 155–171 (1995)
7. R. D. Fierro, G. H. Golub, P. C. Hansen, D. P. O'Leary: Regularization by truncated total least squares. SIAM J. Sci. Comput. **18**, 1223–1241 (1997)
8. G. H. Golub, A. Hoffman, G. W. Stewart: A generalization of the Eckart-Young-Mirsky matrix approximation theorem. Linear Algebra Appl. **88/89**, 317–327 (1987)
9. G. H. Golub, W. Kahan: Calculating the singular values and pseudo-inverse of a matrix. J. SIAM, Series B, Numer. Anal. **2**, 205–224 (1965)

10. G. H. Golub, C. Reinsch: Singular value decomposition and least squares solutions. Numerische Mathematik **14**, 403–420 (1970). Also in "Handbook for Automatic Computation Vol. 2: Linear Algebra", by J. H. Wilkinson and C. Reinsch, (eds.) pp. 134–151. New York: Springer 1971
11. G. H. Golub, C. F. van Loan: An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**, 883–893 (1980)
12. G. H. Golub, C. F. Van Loan: Matrix Computations. Baltimore MD: The Johns Hopkins University Press, third ed. 1996
13. R. D. D. Groat, E. M. Dowling: The data least squares problem and channel equalization. IEEE Transactions on Signal Processing **42:1**, 407–411 (1993)
14. E. V. Haynsworth: Determination of the inertia of a partitioned Hermitian matrix. Linear Algebra Appl. **1**, 73–81 (1968)
15. S. Van Huffel: Personal communication, June 1999
16. S. Van Huffel, J. Vandewalle: The Total Least Squares Problem: Computational Aspects and Analysis. Philadelphia PA: SIAM Publications 1991
17. C. C. Paige: Bidiagonalization of matrices and solution of linear equations. SIAM J. Numer. Anal. **11**, 197–209 (1974)
18. C. C. Paige, M.A. Saunders: LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Software **8**, 43–71 (1982)
19. C. C. Paige, M. A. Saunders: Algorithm 583 LSQR: Sparse linear equations and least squares problems. ACM Trans. Math. Software **8**, 195–209 (1982)
20. C. C. Paige, Z. Strakoš: Bounds for the least squares distance using scaled total least squares. Accepted for publication in Numerische Mathematik, February 2001
21. B. N. Parlett: The Symmetric Eigenvalue Problem. Philadelphia PA: SIAM Publications 1998
22. B. D. Rao: Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework. In: "Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling", S. Van Huffel (ed.), pp. 11–20. Philadelphia PA: SIAM Publications 1997
23. J. Wilkinson: The Algebraic Eigenvalue Problem. Oxford England: Clarendon Press 1965