SIAM J. MATRIX ANAL. APPL.
Vol. 27, No. 3, pp. 861–875


# CORE PROBLEMS IN LINEAR ALGEBRAIC SYSTEMS*

CHRISTOPHER C. PAIGE† AND ZDENĚK STRAKOŠ‡

**Abstract.** For any linear system $Ax \approx b$ we define a set of core problems and show that the orthogonal upper bidiagonalization of $[b, A]$ gives such a core problem. In particular we show that these core problems have desirable properties such as minimal dimensions. When a total least squares problem is solved by first finding a core problem, we show the resulting theory is consistent with earlier generalizations, but much simpler and clearer. The approach is important for other related solutions and leads, for example, to an elegant solution to the data least squares problem. The ideas could be useful for solving ill-posed problems.

**Key words.** scaled total least squares, least squares, data least squares, orthogonal regression, core problem, orthogonal reduction, minimum 2-norm solutions, bidiagonalization, singular value decomposition, ill-posed problems

**AMS subject classifications.** 15A06, 15A18, 15A21, 65F20, 65F25, 65G50

**DOI.** 10.1137/040616991

## 1. Introduction.

We will use uppercase Roman letters to denote matrices, lowercase Roman to denote vectors and indices, and lowercase Greek to denote scalars. The $i$th column of the unit matrix $I$ is $e_i$, $\|\cdot\|$ denotes the 2-norm, $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathcal{R}(M)$ denotes the range (column space) of a matrix $M$.

Consider estimating $\tilde{x}$ from the (possibly compatible) real linear system

$$\tilde{A}\tilde{x} \approx \tilde{b}, \quad \tilde{A} \text{ a nonzero } n \text{ by } k \text{ matrix}, \quad \tilde{b} \text{ a nonzero } n\text{-vector}. \tag{1.1}$$

Suppose (1.1) can be transformed to $Ax \equiv (P^T\tilde{A}Q)(Q^T\tilde{x}) \approx P^T\tilde{b} \equiv b$, where

$$P^T\left[\ \tilde{b} \,\|\, \tilde{A}Q\ \right] = \left[\ b \,\|\, A\ \right] = \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right]; \ P^{-1}=P^T,\ Q^{-1}=Q^T. \tag{1.2}$$

We say this is a nontrivial decomposition if $A_{22}$ has at least one row and one column, even if $A_{22} = 0$. In this nontrivial case the singular value decompositions (SVDs) of $[b, A]$ and $A$ can each be split into two independent SVDs, with the SVD of $A_{22}$ being common to both. More importantly for this exposition, the approximation problem $\tilde{A}\tilde{x} \approx \tilde{b}$ can then be transformed to two *independent* approximation problems as follows:

$$A_{11}x_1 \approx b_1, \qquad A_{22}x_2 \approx 0, \qquad \tilde{x} \equiv Qx, \qquad x \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{1.3}$$

It can be seen from (1.2) that the solution to each of the approximation problems in (1.3) does not affect, and can be found independently of, the other. The problem $A_{22}x_2 \approx 0$ says only that $x_2$ lies approximately in the nullspace of $A_{22}$. Thus unless

*Received by the editors October 4, 2004; accepted for publication (in revised form) by P. C. Hansen July 7, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/simax/27-3/61699.html

†School of Computer Science, McGill University, Montreal, Quebec, H3A 2A7, Canada (paige@cs.mcgill.ca). The research of this author was supported by NSERC of Canada grant OGP0009236.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic (strakos@cs.cas.cz). The research of this author was supported by the National Program of Research "Information Society" under project 1ET400300415.

there is some reason not to (see, for example, Remark 1.1), we can take $x_2 = 0$, and only $A_{11}x_1 \approx b_1$ need be solved. For orthogonally transformed problems we suggest the following definition.

DEFINITION 1.1. *We say* $A_{11}x_1 \approx b_1$ *is a* core problem *in* $\tilde{A}\tilde{x} \approx \tilde{b}$ *if* $[b_1, A_{11}]$ *is minimally dimensioned (or* $A_{22}$ *is maximally dimensioned) subject to* (1.2).

The ideas presented here may be applicable to other than orthogonal transformations of (1.1), but we will concentrate on orthogonal transformations because of their practicality (their relationship with scaled total least squares (scaled TLS)) and the elegance of the results (their relationship with the SVD). Even when we restrict ourselves to orthogonal transformations, the concept of core problems is meaningful outside optimal approximation methods, but these give the best motivation for introducing core problems. For (1.1), the unifying example of optimal approximation problems whose optima are invariant under orthogonal transformations is then scaled TLS: for given $\gamma > 0$:

$$\text{scaled TLS distance} \equiv \min_{\tilde{g},\tilde{E},\tilde{x}} \|[\tilde{g}, \tilde{E}]\|_F \quad \text{subject to} \quad (\tilde{A} + \tilde{E})\tilde{x}\gamma = \tilde{b}\gamma + \tilde{g}\,. \tag{1.4}$$

This is a reformulation of the unification in [22]; see [20, 21]. When $\gamma = 1$ scaled TLS becomes total least squares (TLS, see in particular [10], [7, section 6], [11], [9, pp. 324–326], [25, 23]). TLS is also known in the statistical literature as orthogonal regression. In the limit scaled TLS corresponds to ordinary least squares (LS) when $\gamma \to 0$, and to data least squares (DLS, see [13]) when $\gamma \to \infty$; see [11, 22, 20, 21].

It is not theoretically necessary, but for the remainder of this paper we will assume

$$\tilde{b} \not\perp \mathcal{R}(\tilde{A}) \quad (\text{that is, } \tilde{A}^T\tilde{b} \neq 0). \tag{1.5}$$

This eliminates the annoying trivial case where $A_{11}$ in (1.2) has no columns.

It is often important, even essential, to find a *core* problem $A_{11}x_1 \approx b_1$. It was shown in [20, section 7] that the existence of a nontrivial decomposition of the form (1.2) can prevent the TLS, scaled TLS, and DLS formulations for solving (1.1) from having meaningful solutions when applied directly to $[\tilde{b}, \tilde{A}]$. Here we give a simple example to make this obvious.

For our analysis it is sufficient and convenient to consider the TLS problem ((1.4) with $\gamma = 1$) applied to $\tilde{A}\tilde{x} \approx \tilde{b}$ with nontrivial decomposition (1.2). If $A_{11}x_1 \approx b_1$ is a core problem, then its TLS solution exists and is unique (see Remark 2.1), and its TLS distance is $\sigma_{min}([b_1, A_{11}])$, where $\sigma_{min}(\cdot)$ denotes the minimum singular value; see, for example, [12, section 12.3]. Using temporary notation, suppose $\sigma_k \equiv \sigma_{min}(A_{22})$ and

$$\sigma_k < \sigma_{min}([b_1, A_{11}]), \quad A_{22}v_k = u_k\sigma_k, \quad u_k^T A_{22} = \sigma_k v_k^T, \quad \|u_k\| = \|v_k\| = 1.$$

For *any* real vector $z$ define $r_1 \equiv b_1 - A_{11}z$, then for any real scalar $\theta > 0$

$$\left[\begin{array}{c|c} A_{11} & r_1\theta^{-1}v_k^T \\ \hline 0 & A_{22} - u_k\sigma_k v_k^T \end{array}\right] \left[\begin{array}{c} z \\ \hline v_k\theta \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline 0 \end{array}\right],$$

so the square of the Frobenius norm of the corresponding correction to $A$ is $\|r_1\|^2\theta^{-2} + \sigma_k^2 \to \sigma_k^2$ as $\theta \to \infty$.

Thus by applying (1.4) directly to $[\tilde{b}, \tilde{A}]$ we can get a nonoptimal "TLS distance" *less than* $\sigma_{min}([b_1, A_{11}])$, the TLS distance for the meaningful core problem $A_{11}x_1 \approx b_1$. The above "solution vector" need have nothing to do with the TLS solution vector for $A_{11}x_1 \approx b_1$, but is *essentially determined by* $v_k$ *from the noncore part of the problem* $A_{22}x_2 \approx 0$. This does not reflect any useful information contained in the data. Moreover, in this case the optimal solution to (1.4) does not even exist. Instead, in general we recommend finding a core problem via (1.2); then

$$\text{solve the core problem } A_{11}x_1 \approx b_1 \text{ and set } x_2 = 0 \text{ in (1.3).} \tag{1.6}$$

Section 2 introduces an SVD based decomposition of the form (1.2) with minimally dimensioned $[b_1, A_{11}]$. It also suggests a possible application to ill-posed problems with uncertain data. Section 3 shows how an orthogonal bidiagonalization will give an optimally partitioned decomposition of the form (1.2) directly. This bidiagonalization will also show whether the original problem (1.1) is compatible or not, and is an ideal first step in obtaining the TLS, scaled TLS, or DLS solutions to (1.1); see [20, 21]. For completeness this is briefly reviewed in section 4, with an emphasize on an elegant solution of the following DLS formulation applied to a core problem:

$$\text{DLS distance} \quad \equiv \quad \min_{E_{11}, x_1} \|E_{11}\|_F \quad \text{subject to} \quad (A_{11} + E_{11})x_1 = b_1. \tag{1.7}$$

Section 5 summarizes our ideas and compares them with other approaches. In all cases the extension to the complex case is straightforward; see, for example, [20]. Throughout the paper we will use the following easily proven result.

LEMMA 1.2. *For* (1.1)*, and* (1.3) *satisfying* (1.2)*,* $A_{11}x_1 \approx b_1$ *is a compatible (incompatible) system if and only if* $\tilde{A}\tilde{x} \approx \tilde{b}$ *is a compatible (incompatible) system.*

*Remark* 1.1. In some applications such as solving noisy ill-posed problems, one may be tempted to use a nonzero $x_2$ in order to enforce some regularization constraints on the solution $\tilde{x}$. Such constraints are typically formulated in terms of the generalized norm $\|L\tilde{x}\|$, which for a given matrix $L$ can combine $x_1$ and $x_2$; see [16]. In [6, *section* 5] the problem involving $\tilde{A}$, $L$, and $\tilde{b}$ is transformed into a standard form TLS problem with the matrix $A_{sf}$ and the right-hand side $b_{sf}$. Then standard algorithms for finding the regularized TLS solution of the transformed problem are applied, followed by the transformation of the regularized solution $\tilde{x}_{sf}$ back to the general setting. Within such a framework the core problem theory and computations can be applied to solving the transformed problem $A_{sf}x_{sf} \approx b_{sf}$.

Here we focus on the basic theory of core problems in (1.1). We are aware of the possible implementation difficulties, and of promising applications to regularization of ill-posed problems; see Remarks 2.2 and 3.2 below. These issues are, however, not in the scope of this paper, and we leave them for further investigation. Therefore, within the scope of our paper, the obvious choice for the solution of $A_{22}x_2 \approx 0$ is $x_2 = 0$.

**2. Understanding core problems in $\tilde{A}\tilde{x} \approx \tilde{b}$.** In Definition 1.1 we said $A_{11}x_1 \approx b_1$ is a core problem in $\tilde{A}\tilde{x} \approx \tilde{b}$ if an orthogonal decomposition of the form of (1.2) exists where $[b_1, A_{11}]$, and so $A_{11}$, has minimal dimensions. An understanding of such minimal dimensions can be gained by the following construction, which shows how to concentrate the relevant information into $A_{11}$ and $b_1$ while moving the irrelevant and redundant information into $A_{22}$. Let $\tilde{A}$ have rank $r$ and SVD

$$\tilde{A} = U \left[\begin{array}{c|c} S & 0 \\ \hline 0 & 0 \end{array}\right] V^T,$$

where $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, the $n \times n$ matrix $U^{-1} = U^T$, and the $k \times k$ matrix $V^{-1} = V^T$. Then

$$U^T \left[\tilde{b} \,\|\, \tilde{A}V\right] = \left[\begin{array}{c||c|c} \tilde{c} & S & 0 \\ \hline d & 0 & 0 \end{array}\right]. \tag{2.1}$$

Although singular values are unique, in any SVD representation their ordering, and sometimes some singular vectors, are not. In order to obtain the core problem we will seek to transform $U^T[\tilde{b}, \tilde{A}V]$ further, while maintaining the SVD of $\tilde{A}$. Consider the partitioning $U \equiv [U_1, U_2]$, $V \equiv [V_1, V_2]$, where $U_1$ and $V_1$ have $r$ columns. The vectors $\tilde{c}$ and $d$ might already have some zero elements, and our aim is to introduce as many more as possible. For $d$, choose orthogonal $P_{22}$ so that $P_{22}^T d = e_1\delta$, $\delta \equiv \|d\|$, and replace $U_2$ by $U_2P_{22}$. In this way, $d$ is transformed into a vector having at most one nonzero entry. Now consider $\tilde{c} \equiv [\tilde{\gamma}_1, \ldots, \tilde{\gamma}_r]^T$. Suppose that some of the singular values of $A$ have multiplicities greater than one, for example $\sigma_i = \sigma_{i+1} = \cdots = \sigma_j$, $j > i$. Choose orthogonal $P_{(i,j)}$ so that $P_{(i,j)}^T[\tilde{\gamma}_i, \ldots, \tilde{\gamma}_j]^T = [\gamma_{i,j}, 0, \ldots, 0]^T$. Then transform $U_1$ and $V_1$ by application of $P_{(i,j)}$ from the right to the block of columns numbered $i, \ldots, j$. This transformation will leave $S$ unchanged and therefore preserves the SVD. Do this for each block of multiple singular values, and so obtain $P_{11}$ where $P_{11}^T\tilde{c}$ has at most one nonzero element corresponding to each block of equal singular values of $S$, and replace $U_1$ by $U_1P_{11}$, $V_1$ by $V_1P_{11}$. Next permute the columns of $U_1P_{11}$ and $V_1P_{11}$ identically, to move the zero elements of $P_{11}^T\tilde{c}$ to the bottom of this vector, leaving $c$ at the top with nonzero elements while keeping $S$ diagonal. Finally, if $\delta > 0$, move its row so $\delta$ is immediately below $c$ by a further permutation from the left to give, with obvious new notation and indexing,

$$U^T \left[\ \tilde{b} \,\|\, \tilde{A}V\ \right] = \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right] = \left[\begin{array}{c||c|c} c & S_1 & 0 \\ \delta & 0 & 0 \\ \hline 0 & 0 & S_2 \end{array}\right], \tag{2.2}$$

where, assuming $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$,

$$\begin{aligned} &c \equiv [\gamma_1, \ldots, \gamma_p]^T; \quad \gamma_i \neq 0,\ i = 1, \ldots, p; \quad U^{-1} = U^T, \quad V^{-1} = V^T; \\ &S_1 \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_p), \quad \sigma_1 > \cdots > \sigma_p > 0; \quad \text{the row with} \\ &\text{the scalar } \delta \text{ is nonexistent if and only if } \tilde{A}\tilde{x} \approx \tilde{b} \text{ is compatible.} \end{aligned} \tag{2.3}$$

The final partitioning corresponds to that in (1.2). Here $S_2$ has the remaining $r - p$ singular values of $\tilde{A}$, and the comment in (2.3) follows from Lemma 1.2. We emphasize that the diagonal elements of $S_1$ are different from each other and that all entries in $c$ are nonzero. In this way, the redundant information (multiplicities of singular values) and irrelevant data are removed to $S_2$.

We now show that $A_{11}x_1 \approx b_1$ obtained by the transformation process described above has the desired minimality property. For the SVD of $\tilde{A}$ write $U \equiv [u_1, \ldots, u_n]$. If $u_i, u_{i+1}, \ldots, u_j$, $i \leq j$, are *all* the left singular vectors corresponding to a given singular value $\sigma$, we say $\mathcal{R}([u_i, u_{i+1}, \ldots, u_j])$ is the *left singular subspace* of $\tilde{A}$ corresponding to $\sigma$. But the left and right singular subspaces corresponding to a given (possibly multiple) nonzero singular value $\sigma$ of $\tilde{A}$ are unique; see, for example, [17, Thm. 3.1.1', p. 147]. Thus if $\tilde{b}$ is orthogonal (or not orthogonal) to the left singular subspace of $\tilde{A}$ corresponding to a given singular value $\sigma > 0$, this will be obvious in *all* SVD representations of $\tilde{A}$. In particular for the SVD of $\tilde{A}$ in (2.2)–(2.3), for $i = 1, \ldots, p$

we see that $u_i\gamma_i = u_i(u_i^T\tilde{b})$ is the projection of $\tilde{b}$ onto the left singular subspace of $\tilde{A}$ corresponding to $\sigma_i > 0$ (since the construction has ensured that the other singular vectors of any multiple $\sigma_i$ are orthogonal to $\tilde{b}$). This construction of the decomposition (1.2) in the special form (2.2)–(2.3) leads to Lemma 2.1. The subsequent minimality theorem proves the minimal dimensions of the resulting core problem $A_{11}x_1 \approx b_1$.

LEMMA 2.1. $[\tilde{b}, \tilde{A}]$ *has a decomposition of the form in* (2.2)–(2.3) *if and only if* $\tilde{b}$ *has nonzero projections on exactly* $p$ *left singular subspaces of* $\tilde{A}$ *corresponding to distinct nonzero singular values. The projections correspond to those in* $[b_1, A_{11}]$.

THEOREM 2.2. *Suppose* $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$ *has nonzero projections on exactly* $p$ *left singular subspaces of* $\tilde{A}$ *corresponding to distinct nonzero singular values. Then among all decompositions of the form* (1.2)*, the minimally dimensioned* $A_{11}$ *is* $p\times p$ *if* $\tilde{A}\tilde{x} \approx \tilde{b}$ *is compatible, and* $(p+1)\times p$ *if* $\tilde{A}\tilde{x} \approx \tilde{b}$ *is incompatible.*

*Proof.* From Lemmas 1.2 and 2.1, $[\tilde{b}, \tilde{A}]$ has a decomposition of the form in (2.2)–(2.3), where $A_{11}$ is $p \times p$ if $\tilde{A}\tilde{x} \approx \tilde{b}$ is compatible and $(p+1) \times p$ if $\tilde{A}\tilde{x} \approx \tilde{b}$ is incompatible. Suppose that there exists another decomposition of the form

$$\bar{P}^T \left[\ \tilde{b} \parallel \tilde{A}\bar{Q}\ \right] = \left[\begin{array}{c||c|c} \bar{b}_1 & \bar{A}_{11} & 0 \\ \hline 0 & 0 & \bar{A}_{22} \end{array}\right],$$

where $\bar{P}$ and $\bar{Q}$ are orthogonal matrices and $\bar{A}_{11}$ has $q$ columns. Here we can assume that $\bar{A}_{11}$ has full column rank $q$, and $[\bar{b}_1, \bar{A}_{11}]$ has full row rank $\tilde{q}$, otherwise $\bar{Q}$ and $\bar{P}$ could be chosen to give $\bar{A}_{22}$ more columns or rows. From Lemma 1.2 we see that $\tilde{q} = q$ if $\tilde{A}\tilde{x} \approx \tilde{b}$ is compatible and $\tilde{q} = q + 1$ if it is not.

Suppose $q < p$; then $\bar{A}_{11}$ must have fewer columns *and* rows than $A_{11}$. Obtain an SVD of $\tilde{A}$ by obtaining the individual SVDs of $\bar{A}_{11}$ (and transforming $\bar{b}_1$ accordingly) and $\bar{A}_{22}$, leading to the form of (2.2)–(2.3) with $p$ replaced by $q$. But this would mean $\tilde{b}$ has nonzero projections on at most $q < p$ left singular subspaces of $\tilde{A}$ (see Lemma 2.1), which by assumption is false; so $q \geq p$, and (2.3) provides a minimally dimensioned, or *core*, problem within $\tilde{A}\tilde{x} \approx \tilde{b}$. ∎

*Remark* 2.1. It follows from Definition 1.1, Lemma 2.1, and Theorem 2.2 that $[b_1, A_{11}]$ in (2.2)–(2.3) represents a core problem. From the form of this it can be shown that the TLS, scaled TLS, and DLS formulations have unique and meaningful solutions for any core problem $A_{11}x_1 \approx b_1$; see [20, (1.10) et seq.] and [21, (9) et al.]. This theory, and the method of solution of such problems, is discussed further in sections 3 (following Theorem 3.2), 4, and 5. In fact (1.10) in [20], and (9) in [21], is just (5.11) here.

The SVD is costly to compute, and the computation is necessarily iterative; see, for example, [12, sections 5.4.3–5, pp. 251–254]. In order to find a core problem, we do not need to follow the costly procedure described above. In section 3 we show how to find a core problem directly and cheaply. This will also give us the ideal first step towards computing $c$, $\delta$, and $S_1$ in (2.2)–(2.3), should we want them.

*Remark* 2.2. Because of data and rounding errors, few practical problems will decompose computationally as in (1.2). However, if we have a good idea of the accuracy of our data and computer arithmetic, this analysis will allow us to go from (2.1) to (2.2)–(2.3) within this accuracy (cf. [11, *section* 5]), and this could be particularly useful for ill-posed problems. Suppose $\tilde{b}$ is only accurate to within $\beta\|\tilde{b}\|$, and $\tilde{A}$ to within $\alpha\|\tilde{A}\|$. Then here is the outline of an approach to get from (2.1) to the form in (2.2)–(2.3):

- Any elements of $|\tilde{c}|$ less than say $\beta\|\tilde{b}\|$ can be set to zero.
- Any diagonal elements of $S$ less than say $\alpha\|\tilde{A}\|$ can be set to zero.

- Any block of diagonal elements of $S$ which are equal to within say $2\alpha\|\tilde{A}\|$ can be set equal to their midvalue.
- The resulting new (2.1) can be transformed as above to (2.2) with (2.3).

**3. Computing a core problem within $\tilde{A}\tilde{x} \approx \tilde{b}$.** We can compute a decomposition of the form (1.2) directly by choosing orthogonal matrices $P$ and $Q$ to reduce $[\tilde{b}, \tilde{A}]$ to a real upper-bidiagonal matrix; see, for example, [12, section 5.4.3, p. 251] (and also [20, section 8]). Partitions $P = [P_1, P_2]$ and $Q = [Q_1, Q_2]$ are obtained by stopping at the first zero element, giving (1.2) where $A_{22}$ has not been bidiagonalized, while upper-bidiagonal $[b_1, A_{11}] = P_1^T[\tilde{b}, \tilde{A}Q_1]$ has nonzero bidiagonal elements and is either

$$[b_1|A_{11}] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \cdot & \cdot & \\ & & & \beta_p & \alpha_p \end{array}\right], \quad \beta_i\alpha_i \neq 0, \quad i = 1, \ldots, p \tag{3.1}$$

if $\beta_{p+1} = 0$ or $p = n$; or

$$[b_1|A_{11}] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \cdot & \cdot & \\ & & & \beta_p & \alpha_p \\ & & & & \beta_{p+1} \end{array}\right], \quad \beta_i\alpha_i \neq 0, \; i = 1, \ldots, p; \; \beta_{p+1} \neq 0 \tag{3.2}$$

if $\alpha_{p+1} = 0$ or $p = k$. In the first case $A_{11}x_1 = b_1$ in (1.3) is a compatible system since $A_{11}$ is $p \times p$ and nonsingular. In the second case $A_{11}x_1 \approx b_1$ is an incompatible system since $[b_1, A_{11}]$ has rank $p + 1$. Note that under the assumption $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$ we have $\alpha_1 \neq 0$ (see (1.5)).

*Remark* 3.1. Whether (3.1) or (3.2) results, this bidiagonalization has two important alternative interpretations, and these help us to understand its effectiveness. With the above partitioning of $P$ and $Q$ we see that

$$P^T \left[\begin{array}{c|c} \tilde{b} & \tilde{A}Q_1 \end{array}\right] = \left[\begin{array}{c|c} b_1 & A_{11} \\ 0 & 0 \end{array}\right],$$

so that the bidiagonalization gives the QR factorization of $[\tilde{b}, \tilde{A}Q_1]$, ensuring that $[b_1, A_{11}]$ has full row rank. Next we see that

$$\left[\; P_1^T\tilde{A} \;\right] Q = \left[\; A_{11}, \quad 0 \;\right],$$

so that the bidiagonalization gives the LQ factorization of $P_1^T\tilde{A}$, ensuring that $A_{11}$ has full column rank.

The proof that (3.1) and (3.2) correspond to core problems will be given in Theorems 3.2 and 3.3. But first we need a lemma.

LEMMA 3.1. *Let $J = [b_1, A_{11}]$ be bidiagonal as in* (3.1) *or* (3.2)*. Then all its left and right singular vectors (for its $p$ or $p + 1$ nonzero singular values) have nonzero first and last elements. The nonzero null-vector of* (3.1) *has no zero elements.*

*Proof.* With $u \equiv (\mu_1, \ldots, \mu_{p+1})^T$, $v \equiv (\nu_1, \ldots, \nu_{p+1})^T$, we see that

$$u^TJ = \sigma v^T \Rightarrow \mu_1\beta_1 = \sigma\nu_1; \quad \mu_i\alpha_i + \mu_{i+1}\beta_{i+1} = \sigma\nu_{i+1}, \; i = 1, \ldots, p; \tag{3.3}$$

$$Jv = u\sigma \Rightarrow \beta_i\nu_i + \alpha_i\nu_{i+1} = \mu_i\sigma, i = 1, \ldots, p; \; \beta_{p+1}\nu_{p+1} = \mu_{p+1}\sigma; \tag{3.4}$$

where $\mu_{p+1}$ and $\beta_{p+1}$ are nonexistent in (3.1). For $\sigma > 0$, (3.3) shows that if either $\mu_1$ or $\nu_1$ is zero, then so is the other, and then (3.4) and (3.3) show all the remaining elements are zero. Similar arguments give the rest of the proof. ∎

THEOREM 3.2. *For* $[b_1, A_{11}]$, $b_1 \neq 0$, $\alpha_1 \neq 0$, *in* (3.1) *or* (3.2) *with SVD* $A_{11} = \sum_{i=1}^{p} u_i \sigma_i v_i^T$, *the* $p$ *singular values* $\sigma_i$ *of* $A_{11}$ *are distinct and nonzero, and they strictly separate the* $p+1$ *distinct and nonzero singular values of* $[b_1, A_{11}]$ *in* (3.2) *(or the* $p$ *distinct and nonzero singular values of* $[b_1, A_{11}]$ together with 0 *in* (3.1)*). In both cases*

$$\text{rank}\,(A_{11}) = p; \qquad b_1^T u_i \neq 0, \quad i = 1, \ldots, p. \tag{3.5}$$

*Proof.* From their obvious ranks, $A_{11}$ has exactly $p$ nonzero singular values, and $[b_1, A_{11}]$ has exactly $p$ nonzero singular values in (3.1) and exactly $p+1$ in (3.2). But $T \equiv [b_1, A_{11}]^T [b_1, A_{11}]$ is $(p+1) \times (p+1)$ symmetric tridiagonal with nonzero next to diagonal elements, and $A_{11}^T A_{11}$ remains when the first row and column are deleted. Thus the eigenvalues of $A_{11}^T A_{11}$ strictly separate those of $T$; see, for example, [26, Ch. 5, sect. 37, p. 300]. This proves the first part of the theorem. For the second part of (3.5) $b_1^T u_i = \beta_1 e_1^T u_i \neq 0$ from Lemma 3.1. ∎

The condition (3.5) also directly ensures that the TLS, scaled TLS, and DLS formulations have unique and meaningful solutions for $A_{11} x_1 \approx b_1$ in the incompatible case (3.2); see [20, (1.10)], [21, (9)], and (5.11) with the discussion in section 5 ((3.5) implies (5.11) for $[b_1, A_{11}]$).

In Theorem 3.3 we will show that the orthogonal bidiagonalization leading to (3.1) or (3.2) gives a core problem $A_{11} x_1 \approx b_1$ in $\tilde{A}\tilde{x} \approx \tilde{b}$. It will help if we first briefly restate the relevant parts of Theorems 2.2 and 3.2.

THEOREM 2.2. *Suppose* $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$ *has nonzero projections on exactly* $p$ *left singular subspaces of* $\tilde{A}$ *corresponding to distinct nonzero singular values. Then among all decompositions of the form* (1.2)*, the minimally dimensioned* $A_{11}$ *is* $p \times p$ *if* $\tilde{A}\tilde{x} \approx \tilde{b}$ *is compatible, and* $(p+1) \times p$ *if* $\tilde{A}\tilde{x} \approx \tilde{b}$ *is incompatible.*

THEOREM 3.2. *For* $[b_1, A_{11}]$, $b_1 \neq 0$, $\alpha_1 \neq 0$, *in* (3.1) *or* (3.2) *with SVD* $A_{11} = \sum_{i=1}^{p} u_i \sigma_i v_i^T$, *we have that* $\text{rank}(A_{11}) = p$ *and* $b_1^T u_i \neq 0$ *for* $i = 1, \ldots, p$. *Also the* $p$ *singular values of* $A_{11}$ *are distinct and nonzero. (The rest is omitted here.)*

THEOREM 3.3. *If* $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$ *and* $n \times (k+1)$ $[\tilde{b}, \tilde{A}]$ *has an orthogonal decomposition of the form* (1.2) *with* $[b_1, A_{11}]$ *as in* (3.1) *or* (3.2)*, then*

(a) $A_{11}$ *has no zero or multiple singular values, and thus any zero singular values or repeats that* $\tilde{A}$ *has must appear in* $A_{22}$*;*

(b) $[b_1, A_{11}]$ *(and thus* $A_{11}$*) has minimal dimensions, and* $A_{22}$ *maximal dimensions, over all orthogonal transformations of the form shown in* (1.2)*;*

(c) *orthogonal* $\hat{U}_{11}$ *and* $\hat{V}_{11}$ *in the transformation* $\hat{U}_{11}^T [b_1, A_{11} \hat{V}_{11}]$ *can be designed to produce the form of* $[b_1, A_{11}]$ *in* (2.2)–(2.3).

*Proof.* (a) This follows immediately from Theorem 3.2. (b) Theorem 3.2 shows that $\tilde{b}$ has nonzero projections on exactly $p$ left singular subspaces of $\tilde{A}$ corresponding to distinct nonzero singular values. Thus, following Theorem 2.2, $[b_1, A_{11}]$ is minimally dimensioned, so that $A_{22}$ is maximally dimensioned. (c) Using the SVDs $A_{11} = U_{11} S_1 V_{11}^T$, $A_{22} = U_{22} S_2 V_{22}^T$, $U_{11}^T U_{11} = V_{11}^T V_{11} = I_p$, etc.,

$$\left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array}\right] = \left[\begin{array}{c||c|c} b_1 & U_{11} S_1 V_{11}^T & 0 \\ \hline 0 & 0 & U_{22} S_2 V_{22}^T \end{array}\right]$$

$$= \left[\begin{array}{cc|c} U_{11} & r_1 & 0 \\ \hline 0 & 0 & U_{22} \end{array}\right] \left[\begin{array}{c||c|c} c & S_1 & 0 \\ \delta & 0 & 0 \\ \hline 0 & 0 & S_2 \end{array}\right] \left[\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 0 & V_{11}^T & 0 \\ \hline 0 & 0 & V_{22}^T \end{array}\right],$$

where $c \equiv U_{11}^T b_1$, $w \equiv b_1 - U_{11}c$, and if $w \neq 0$, $\delta \equiv \|w\|$, $r_1 \equiv w/\delta$. If $\tilde{A}\tilde{x} = \tilde{b}$ is compatible, then $U_{11}$ is square, $w = 0$, and in this case $\delta$ and its row, and $r_1$ and its column, are nonexistent. In the incompatible case $w \neq 0$, $\|r_1\| = 1$, $U_{11}^T r_1 = 0$, and denoting $\hat{U}_{11} \equiv [U_{11}, r_1]$, $\hat{V}_{11} \equiv V_{11}$ gives the structure in (2.2), while Theorem 3.2 shows (2.3) holds for this structure. □

Theorems 2.2 and 3.3 are new. The result (b) of Theorem 3.3 was mentioned in [20, 21], but not proven, because we had not then obtained a sufficiently readable proof of what seemed a fairly obvious result. Theorem 2.2 allowed this, but we hope that others can provide an even simpler proof. The essence of Theorem 3.2 was given in [21, Thm. 1], which is an extended version of [20, Thm. 8.1].

*Remark* 3.2. In practical computations which involve rounding errors or noise in the data, one must consider threshold criteria to decide which elements of the bidiagonal matrix are small in magnitude and should be set to zero. The criteria will be problem-dependent, but unlike the case in Remark 2.2, their choice is less obvious and needs further investigation.

It will be interesting to relate the core problem formulation to the work on truncated TLS [4, 5]; see also [15, section 6.6] and [24]. In particular, the core problem formulation can be considered as a theoretical basis for the Lanczos truncated TLS proposed in [6, section 4.1] as well as for the partial least squares (PLS) method of Wold, et al. [27] which is equivalent to the Lanczos bidiagonalization-based truncated least squares; see [3]. Various related regularization aspects are described, for example, in [2, 18, 14, 8]; see also [28].

**4. Solving the LS, scaled TLS, and DLS problems using bidiagonalization and the core problem.** Consider the upper bidiagonalization

$$[b, A] = P^T[\tilde{b}, \tilde{A}Q] \quad \text{of the form in (1.2)},$$

with the core problem part $[b_1, A_{11}] = P_1^T[\tilde{b}, \tilde{A}Q_1]$ given by (3.1) or (3.2). In (3.1) $A_{11}x_1 = b_1$ is a compatible system, therefore $\tilde{A}\tilde{x} = \tilde{b}$ is a compatible system (see Lemma 1.2). Then the LS residual, the scaled TLS distance (for any positive finite $\gamma$), and the DLS distance are zero, and the solutions are obvious. We will now consider only the incompatible case (3.2) and take $x_2 = 0$, $\tilde{x} = Q_1 x_1$.

The LS solution of $A_{11}x_1 \approx b_1$ with $[b_1, A_{11}]$ in (3.2) is obtained by orthogonal reduction of the matrix $[A_{11}, b_1]$ to upper triangular form (note the reversal of $A_{11}$ and $b_1$), followed by solution of a triangular system to give $x_1$.

For any given finite positive scaling $\gamma$, the scaled TLS solution (see (1.4)) of $A_{11}x_1 \approx b_1$ with $[b_1, A_{11}]$ in (3.2) is uniquely determined by the unique SVD component $\bar{u}\bar{\sigma}\bar{v}^T$ of $[\gamma b_1, A_{11}]$ corresponding to $\bar{\sigma} \equiv \sigma_{min}([\gamma b_1, A_{11}])$, its minimal singular value. From Lemma 3.1, $\nu = e_1^T \bar{v} \neq 0$, $\bar{v}^T \equiv [\nu, \bar{w}^T]$ and

$$x_1 = -\bar{w}/\nu \quad \text{with the scaled TLS correction} \quad -\bar{u}\bar{\sigma}\bar{v}^T \quad \text{to} \quad [\gamma b_1, A_{11}].$$

In the original variables the scaled TLS solution of $\tilde{A}\tilde{x} \approx \tilde{b}$ is $\tilde{x} = Q_1 x_1$. The scaled TLS correction to $[\gamma\tilde{b}, \tilde{A}]$ is obtained from the following exercise:

$$[\tilde{g}, \tilde{E}] = P \begin{bmatrix} -\bar{u}\bar{\sigma}\bar{v}^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q^T \end{bmatrix} = [P_1|P_2] \left[\begin{array}{c|c} -\bar{u}\bar{\sigma}(\nu, \bar{w}^T) & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 0 & Q_1^T \\ \hline 0 & Q_2^T \end{array}\right]$$
$$= -(P_1\bar{u})\,\bar{\sigma}\,(\nu, \bar{w}^T Q_1^T) = (P_1\bar{u})\,\bar{\sigma}\nu\,[-1, \tilde{x}^T]\,.$$

The DLS problem was proposed and used in [13]. Suppose that the core part $[b_1, A_{11}]$ of the transformed $[\tilde{b}, \tilde{A}]$ has the form in (3.2). We will show how to solve the DLS problem (1.7) for this core data; see [20, 21]. Write

$$[b_1|A_{11}] \equiv \left[\begin{array}{c|c} \beta_1 & \alpha_1 e_1^T \\ \hline 0 & A_2 \end{array}\right] \equiv P_1^T[\tilde{b}|\tilde{A}Q_1], \quad E_{11} \equiv \left[\begin{array}{c} e^T \\ \hline E_2 \end{array}\right] \equiv P_1^T\tilde{E}Q_1, \quad \tilde{x} \equiv Q_1 x_1,$$

where $A_2$ is square with all its singular values distinct and nonzero. Let $\sigma$, $u$, and $v$ be the minimum singular value $\sigma \equiv \sigma_{min}(A_2)$ and its left and right vectors for $A_2$. For the above reduced data the DLS problem (1.7) becomes

$$\min_{e, E_2, x_1} \{\|e\|^2 + \|E_2\|_F^2\} \quad \text{subject to} \quad \left[\begin{array}{c|c} \beta_1 & \alpha_1 e_1^T + e^T \\ \hline 0 & A_2 + E_2 \end{array}\right] \left[\begin{array}{c} -1 \\ x_1 \end{array}\right] = 0.$$

Since $\beta_1$ is nonzero, $x_1 \neq 0$, and the minimum $\|E_2\|_F$ in $(A_2+E_2)x_1 = 0$ is $\sigma$, given by $E_2 = -u\sigma v^T$ and $x_1 = v\xi$ for some $\xi \neq 0$. To make this $x_1$ satisfy the full constraints we need $\beta_1 = \alpha_1 e_1^T v\xi + e^T v\xi$. But $e_1^T v \neq 0$ from Lemma 3.1, so $e = 0$ gives the overall minimum with $\xi = \beta_1/(\alpha_1 e_1^T v)$, and

$$x_D \equiv x_1 = v\beta_1/(\alpha_1 e_1^T v), \qquad \sigma_D \equiv \sigma = \sigma_{min}(A_2)$$

are the DLS solution and distance in (1.7) for the reduced data $[b_1, A_{11}]$. In the original variables $\tilde{x} = Q_1 x_1 = Q_1 v\xi$ and

$$\tilde{E} = P \left[\begin{array}{cc} E_{11} & 0 \\ 0 & 0 \end{array}\right] Q^T = P_1 E_{11} Q_1^T = P_1 \left[\begin{array}{c} 0 \\ -u\sigma v^T \end{array}\right] Q_1^T = P_1 \left[\begin{array}{c} 0 \\ -u\,(\sigma/\xi)\,\tilde{x}^T \end{array}\right].$$

From these we see that the solutions of the scaled TLS and DLS problems are reduced to computing the smallest singular value and its right singular vector for the nonsingular bidiagonal matrices $[b_1, A_{11}]$ and $A_2$, respectively, and these are relatively easy to find; see, for example, [12, section 8.6.2, pp. 452–456].

**5. Review and comparisons with earlier work.** The main topics we have dealt with so far are (in order of presentation) as follows:

T1. Core problems $A_{11}x_1 \approx b_1$ in $\tilde{A}\tilde{x} \approx \tilde{b}$; see Definition 1.1 and Theorem 2.2.

T2. Some criteria for solving $\tilde{A}\tilde{x} \approx \tilde{b}$; see (1.4), (1.7).

T3. A special form of the SVD of $\tilde{A}$ related to $\tilde{b}$, which gives the core problem; see (2.2)–(2.3).

T4. The orthogonal upper bidiagonalization of $[\tilde{b}, \tilde{A}]$ as the way of computing the core problem efficiently; see (3.1)–(3.2).

T5. Implementing topic T2 for topic T1; see section 4 (also [1, 11, 12, 25, 20, 21]).

Our main purpose has been to introduce T3, show its relationship to T4, and prove that the core problem in T1 could be found immediately from either T3 or T4. Using T4 to obtain T1 is a direct computation, and is also an ideal first step in either computing T3 (it gives an orthogonal transformation of $\tilde{A}$ to lower bidiagonal form) or in solving (1.4) or (1.7). Using T3 is an iterative process (but very fast after the bidiagonalization in T4) and gives an excellent guide as to what elements can be considered zero or equal in different cases; see Remark 2.2. An analogous guide which would apply directly to T4 needs further work; see Remark 3.2.

For applications, our suggested approach to finding a scaled TLS solution to (1.1) is to determine the core problem (3.1)–(3.2) and continue according to (1.6). Then (1.4) with finite positive $\gamma$ for $A_{11}x_1 \approx b_1$ can be solved in the manner reviewed briefly in section 4 and described in more detail from here to (5.5) (based on the algorithm of Golub and Van Loan in [11]; see [12, section 12.3]), while the DLS problem can be solved as in section 4. These will always give meaningful solutions (see Remark 2.1), and because of minimal dimensions of $[b_1, A_{11}]$, they will also minimize the cost.

On reading this, Sabine van Huffel suggested that we also compare the core problem approach here to the approaches in [25] for solving the TLS problem (with a single right-hand side $\tilde{b}$, since the extension of the theory here to multiple right-hand side problems remains a subject for further investigation). This will require some basic analysis. We need only consider the incompatible case and $\tilde{b} \not\perp \mathcal{R}(\tilde{A})$.

One reason for developing our formulation (1.4) was that its solution is clearly equivalent to the TLS solution of $\tilde{A}(\tilde{x}\gamma) \approx \tilde{b}\gamma$. Thus by just considering

$$\text{TLS distance} \equiv \min_{g,E,x} \|[g, E]\|_F \quad \text{subject to} \quad (A + E)x = b + g, \tag{5.1}$$

the results we obtain will also cover the extension of the results in [25] to {(1.4) with any finite positive $\gamma$}. We will also assume that

$$[b, A] \equiv P^T[\tilde{b}, \tilde{A}Q] \quad \text{has the form in (1.2) with } [b_1, A_{11}] \text{ given by (3.2),} \tag{5.2}$$

and in this case we will show that the minimum 2-norm solution approaches that were given in [25] theoretically give the same answers to (5.1) as (1.6) does. Since the optimum in (5.1) is independent of such orthogonal transformations, this will mean that the answers for any $[\tilde{b}, \tilde{A}]$ will theoretically be the same here and for the minimum 2-norm solution approaches in (the extended versions of) [25].

The approaches to solving (5.1)–(5.2) make use of the SVD of $n \times (k+1)$ $[b, A]$, which we write as:

$$[b, A] = \tilde{U}\tilde{S}\tilde{V}^T = \sum_{i=1}^{k+1} \tilde{u}_i\tilde{\sigma}_i\tilde{v}_i^T, \qquad \tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_{k+1} \geq 0. \tag{5.3}$$

Here we have assumed $n > k$, which can be attained by adding zero rows to $[b, A]$ if necessary. Note that the SVD used in [25] (see, for example, [25, equation (1.22), p. 22]) does not take account of the possible structure in (1.2), while here we only use the SVD of $[b_1, A_{11}]$; the SVD of $A_{22}$ need not be computed. Thus in our case the SVD of $[b, A]$ in (5.3) is the direct sum of the SVDs of $[b_1, A_{11}]$ and $A_{22}$. We can still use the ordering of vectors and singular values in (5.3), but for equal singular values we will always assume the first in order comes from the $[b_1, A_{11}]$ block. Clearly for our version of (5.1)–(5.3), any SVD component $\tilde{u}_i\tilde{\sigma}_i\tilde{v}_i^T$ coming from the $[b_1, A_{11}]$ block can be nonzero only in its leading principle $(p+1) \times (p+1)$ block. Similarly for our version, any SVD component coming from the $A_{22}$ block can be nonzero only in the block corresponding to $A_{22}$.

The SVD used in [25] is not precisely defined, but since we want to compare the solution $\left[\begin{smallmatrix} x_1 \\ 0 \end{smallmatrix}\right]$ here based on $A_{11}x_1 \approx b_1$ to the solution $x$ in [25] based on $Ax \approx b$ in (1.2), it is easiest to assume [25] uses the same SVD as here and comment on the effects of using a more general SVD.

First we derive our unique solution (1.6) to (1.1) with core problem $A_{11}x_1 \approx b_1$. As in section 4, let $[b_1, A_{11}]$ have unique "minimum" SVD component $\bar{u}\bar{\sigma}\bar{v}^T$, where

$$\bar{\sigma} \equiv \sigma_{min}([b_1, A_{11}]) = \tilde{\sigma}_m \quad \text{say, in (5.3).} \tag{5.4}$$

$\bar{\sigma}$ will be unique in the SVD of $[b_1, A_{11}]$ (see Theorem 3.2), but we choose $\tilde{\sigma}_m$ to be the first of any equals in the SVD (5.3). From Lemma 3.1, $\nu \equiv e_1^T \bar{v} \neq 0$. Define

$$[g, E] \equiv -\tilde{u}_m \tilde{\sigma}_m \tilde{v}_m^T = -\begin{bmatrix} \bar{u}\bar{\sigma}\bar{v}^T & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{u}_m = \begin{bmatrix} \bar{u} \\ 0 \end{bmatrix}, \quad \tilde{v}_m = \begin{bmatrix} \bar{v} \\ 0 \end{bmatrix}.$$

Denote

$$\tilde{v}_m \equiv \begin{bmatrix} \nu \\ w \end{bmatrix}, \quad w \equiv \begin{bmatrix} \bar{w} \\ 0 \end{bmatrix}, \quad \bar{v} = \begin{bmatrix} \nu \\ \bar{w} \end{bmatrix}.$$

The solution to the TLS problem (5.1) for $A_{11}x_1 \approx b_1$ is then $x_1 = -\bar{w}/\nu$, with correction $-\bar{u}\bar{\sigma}\bar{v}^T$ to $[b_1, A_{11}]$. With (1.6) this means our solution to $Ax \approx b$ is

$$x = -w/\nu, \text{where } \tilde{v}_m \equiv \begin{bmatrix} \nu \\ w \end{bmatrix}, \text{ with correction } [g, E] \equiv -\tilde{u}_m \tilde{\sigma}_m \tilde{v}_m^T \text{ to } [b, A]. \tag{5.5}$$

We will show that the minimum 2-norm solutions in [25] are identical to this. To help in this we will use the following lemma.

LEMMA 5.1. *The singular values of $A$ interlace those of* $[b, A]$ ([12, Cor.8.6.3]).

There are exactly three cases, denoted 1, 2(a), and 2(b) below.

*Case* 1. If the most well-known criterion for a unique solution to (5.1) holds, i.e.

$$\tilde{\sigma}_{k+1} \equiv \sigma_{min}([b, A]) < \sigma_{min}(A), \tag{5.6}$$

(see [11, Thm.4.1], [12, Thm.12.3.1], [25, Thm.2.6, p.35]), then in (5.4), $\bar{\sigma} = \tilde{\sigma}_m = \tilde{\sigma}_{k+1}$, which from Lemma 5.1 is the unique minimum singular value of $[b, A]$, and the solution in [25, Thm. 2.6, p. 35] to (5.1) is unique and given by (5.5), and so is identical to the solution here.

*Case* 2. Next suppose for a general SVD in (5.3) that for some $j \leq k$

$$\tilde{\sigma}_j > \tilde{\sigma}_{j+1} = \cdots = \tilde{\sigma}_{k+1}, \quad V' \equiv [\tilde{v}_{j+1}, \ldots, \tilde{v}_{k+1}], \quad U' \equiv [\tilde{u}_{j+1}, \ldots, \tilde{u}_{k+1}]. \tag{5.7}$$

If $j < k$, then from Lemma 5.1, (5.6) does not hold. If $e_1^T V' = 0$, then it can be seen from Lemma 3.1 that $\tilde{\sigma}_{j+1} = \cdots = \tilde{\sigma}_{k+1}$ must correspond wholly to $A_{22}$, and (5.6) does not hold. These lead to the special cases in [25].

*Case* 2(a). If $j < k$ and $e_1^T V' \neq 0$, the TLS solution is not unique. Golub and Van Loan [11, pp. 885–886], and later Van Huffel and Vandewalle, effectively design orthogonal $Q'$ so that $e_1^T V' Q' = \nu e_1^T$ and set $\tilde{u} \equiv U'Q'e_1$ and $\tilde{v} \equiv V'Q'e_1 = (\nu, w^T)^T$. Then it was proven in [25, Thm. 3.7, p. 58] that $[g, E] \equiv -\tilde{u}\tilde{\sigma}_{j+1}\tilde{v}^T$ is an optimal correction to $[b, A]$ in (5.1) with minimum 2-norm solution $x = -w/\nu$ and distance $\tilde{\sigma}_{j+1}$. For our structured form of SVD in (5.3), we see from Theorem 3.2 that the $p+1$ singular values of $[b_1, A_{11}]$ in (3.2) are distinct and nonzero, so at most one can come from among $\tilde{\sigma}_{j+1} = \cdots = \tilde{\sigma}_{k+1}$ above, and the other $k - j$ (or $k - j + 1$) in this set of equal values must come from $A_{22}$ in (1.2). Now if $\tilde{v}_{j+1}, \ldots, \tilde{v}_{k+1}$ *all* came from $A_{22}$, then $e_1^T V' = 0$, a contradiction. So exactly one such vector (with our chosen ordering for (5.3) it is $\tilde{v}_{j+1}$) comes from $[b_1, A_{11}]$, and by Lemma 3.1 it has nonzero first element. What has happened is that the splitting (1.2) caused by our bidiagonalization ensures that $V'$ *already* satisfies $e_1^T V' = \pm\nu e_1^T$. Thus $m = j + 1$ in (5.4), and our solution (5.5) is again obtained in [25, Thm. 3.7, p. 58].

*Case* 2(b). The remaining case is where $e_1^T V' = 0$ in (5.7) (which may happen when $j = k$ or $j < k$). This is called the "nongeneric" case in [25, section 3.4] and corresponds to the unpleasant example that we gave near the end of section 1. In linear regression the columns of such $V'$ correspond to nonpredictive multicolinearities; they are of no value in predicting the response $b$ [25, p. 71]. To handle this, Van Huffel and Vandewalle [25, Def. 3.2, p. 68] define a new problem, a more constrained version of (5.1). The strategy is to eliminate those directions in $A$, corresponding to the smallest singular value or to the several smallest singular values of $[b, A]$, that are not at all correlated with the observation vector $b$. For any general SVD (5.3), let $q$ be the maximum value of $i$ such that $e_1^T \tilde{v}_i \neq 0$. Note that we might have $q < j$ in (5.7). The "nongeneric" problem of [25, Def. 3.2, p. 68] is then just (5.1) with the added restriction that $[g, E][\tilde{v}_{q+1}, \ldots, \tilde{v}_{k+1}] = 0$, and any solution $x$ for this is called a "nongeneric TLS solution." In [25, Thm. 3.12, p. 72] it is shown how to obtain such a solution, and the comments following that indicate how to compute the minimum 2-norm solution.

For our structured form of SVD in (5.3), define $q$ as above, set $\tilde{V}_1 = [\tilde{v}_1, \ldots, \tilde{v}_q]$, and with conformable partitioning in (5.3) write

$$\tilde{U} = [\tilde{U}_1, \tilde{U}_2], \quad \tilde{S} = \mathrm{diag}(\tilde{S}_1, \tilde{S}_2), \quad \tilde{V} = [\tilde{V}_1, \tilde{V}_2] \quad \text{so that } [g, E]\,\tilde{V}_2 = 0 \tag{5.8}$$

is the added constraint. Lemma 3.1 with $e_1^T \tilde{V}_2 = 0$ shows that the singular vectors $\tilde{v}_{q+1}, \ldots, \tilde{v}_{k+1}$ come from $A_{22}$; $e_1^T \tilde{v}_q \neq 0$ shows that $\tilde{v}_q$ comes from $[b_1, A_{11}]$, which with the ordering in (5.3) implies $m = q$ in (5.4); and (5.7) shows $\tilde{\sigma}_q \equiv \sigma_{min}([b_1, A_{11}]) > \sigma_{min}(A_{22})$, so (5.6) does not hold. Clearly our solution (5.5) satisfies the constraints in (5.1) and (5.8), but we still need to show it is optimal and minimal in length.

We can eliminate the explicit constraint $[g, E]\,\tilde{V}_2 = 0$ by taking

$$[g, E] \equiv \tilde{U} H \tilde{V}_1^T \equiv (\tilde{U}_1 H_1 + \tilde{U}_2 H_2)\tilde{V}_1^T$$

and now minimize $\|H\|_F$. By defining $y^T \equiv (y_1^T, y_2^T) \equiv (-1, x^T)[\tilde{V}_1, \tilde{V}_2]$ and transforming the constraints in (5.1) to

$$0 = \tilde{U}^T\{[b, A] + [g, E]\}\tilde{V}\tilde{V}^T \begin{bmatrix} -1 \\ x \end{bmatrix} = \begin{bmatrix} \tilde{S}_1 + H_1 & 0 \\ H_2 & \tilde{S}_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

we can reformulate the "nongeneric problem" in [25, Def. 3.2, p. 68] to

$$\begin{aligned} &\min_{H_1, H_2, y_1, y_2, x} (\|H_1\|_F^2 + \|H_2\|_F^2) \\ &\text{subject to} \quad \begin{bmatrix} \tilde{S}_1 + H_1 & 0 \\ H_2 & \tilde{S}_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0, \quad \begin{bmatrix} -1 \\ x \end{bmatrix} = \tilde{V}_1 y_1 + \tilde{V}_2 y_2. \end{aligned} \tag{5.9}$$

Now whatever $H_1$ and $y_1$ are, we can take $H_2 = 0$ and $y_2 = 0$ (since $e_1^T \tilde{V}_2 = 0$, only $y_1$ contributes to the $-1$ in (5.9), and thus the last constraint gives no restriction on $y_2$). Therefore, the problem simplifies to

$$\min_{H_1, y_1, x} \|H_1\|_F^2 \ \text{ subject to } \ (\tilde{S}_1 + H_1) y_1 = 0, \quad \begin{bmatrix} -1 \\ x \end{bmatrix} = \tilde{V}_1 y_1. \tag{5.10}$$

Since $\tilde{S}_1 = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_q)$ with $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_q = \sigma_{min}([b_1, A_{11}]) > 0$ (see Theorem 3.2), a solution is given by $H_1 = -e_q \tilde{\sigma}_q e_q^T$, $y_1 = e_q/(-\nu)$, giving our solution (5.5) with $m = q$. But this is the minimum norm solution, since $\tilde{v}_q$ is the only right singular vector with nonzero first element that corresponds to $\tilde{\sigma}_q$, using our version of (5.3). Thus our approach (1.6) immediately gives the minimum 2-norm solution *satisfying* (5.1) *together with the added restriction in* (5.8), so again the minimum 2-norm solution in [25] is the same as the solution here. This concludes our discussion of equivalent solutions.

Now let $\mathcal{U}_{min}$ be the left singular vector subspace of $A$ corresponding to $\sigma_{min}(A)$. We argued in [20, section 7] that (5.6) was not ideal and that a satisfactory condition for ensuring unique solutions to, and for building the theory of, the TLS, DLS, and scaled TLS formulations for solving (1.1) is the following $\gamma$-independent criterion:

$$\text{the } n \times k \text{ matrix } A \text{ has rank } k \text{ and } \; b \not\perp \mathcal{U}_{min}. \tag{5.11}$$

We showed in [20, Thm. 3.1] that this implies $\sigma_{min}([b\gamma, A]) < \sigma_{min}(A)$ for *all* $\gamma \geq 0$, which is important for scaled TLS, and of course implies (5.6).

A crucial property of any core problem $A_{11}x_1 \approx b_1$ is that $A_{11}$ and $b_1$ *always* satisfy (5.11), see (3.5), or (2.2)–(2.3), and so its scaled TLS solution exists, is unique, and can be computed from the SVD of $[b_1\gamma, A_{11}]$. This can be computed efficiently from either of the forms of $[b_1, A_{11}]$ in (3.2) or (2.2). The bidiagonalization leads to $[b_1, A_{11}]$ satisfying even more than just (5.11), since it removes all redundancies and irrelevant parts of the problem corresponding to all singular values, and in theory it does this implicitly before any singular value computation. Minimizing the dimensions of $[b_1, A_{11}]$ also maximizes the computational efficiency.

The solution of the original problem (compatible or not) obtained via (1.6) is then the minimum 2-norm solution of (1.1) such that the core problem is solved, and any corrections correspond to corrections only in $[b_1, A_{11}]$ in (1.2).

**6. Conclusion.** If $\tilde{A}$ has full column rank, these results show us that we can only find a decomposition of the form (1.2) with nontrivial $A_{22}$ if either $\tilde{b}$ is orthogonal to a left singular vector subspace of $\tilde{A}$ or $\tilde{A}$ has at least one repeated singular value or both. For any $[\tilde{b}, \tilde{A}]$, the bidiagonalization (3.1) or (3.2) (a direct computation, that is, noniterative) will provide that decomposition with minimally dimensioned $A_{11}$ in (1.2). This bidiagonalization will also show whether the original problem (1.1) is compatible or not, and is an ideal first step in solving the TLS, scaled TLS, or DLS formulations for finding the optimal solution to (1.1); see [20, 21]. In some cases (for example, using LSQR in [19]) it is also an excellent first step for ordinary linear LS problems. We showed how this bidiagonalization and the core problem formulation can be used to solve the LS, scaled TLS, and DLS problems.

The TLS solutions obtained via (1.6) were shown to be theoretically identical to the minimum 2-norm solutions of all formulations of TLS in [25]. The one simple and efficient approach given here can be applied to all such problems with a single right-hand side, while different, and more complicated, approaches were needed to solve different classes of problems in [25].

We feel that in addition to its more general contributions, this study simplifies and extends the body of work which was essentially started in [10, 7, 11], but which has been so extensively developed in [25] and elsewhere; see, for example, the collections containing [22, 21].

These results revise our understanding of both the theory and computations in all forms of linear LS problems with a single right-hand side. So far we have presented essentially theoretical and algorithmic ideas, while many implementation details still need to be worked out; see, for example, Remarks 2.2 and 3.2.

**Acknowledgments.** We thank Sabine van Huffel for her valuable suggestion that we show in all cases the mathematical equivalence of the TLS solutions in [25] with the corresponding solutions obtained by the theoretically more direct approach here. It was difficult to find a rigorous way to do this that was still reasonably readable, but we agree it was important to complete this link. We feel the end result of section 5 rounds out the work nicely, even though the proofs could be more elegant.

For improvements to the revised version we thank Per Christian Hansen both for pointing out a useful reference and offering valuable suggestions which led to Remarks 1.1 and 3.2; Gene Golub for supplying three historical references; and an anonymous referee for carefully reading the manuscript and providing comments which improved the presentation nicely.

## REFERENCES

[1] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[2] Å. Björck, E. Grimme, and P. Van Dooren, *An implicit shift bidiagonalization algorithm for ill-posed problems*, BIT, 34 (1994), pp. 510–534.

[3] L. Eldén, *Partial least-squares vs. Lanczos bidiagonalization*–I: *Analysis of a projection method for multiple regression*, Comput. Statist. Data Anal., 46 (2004), pp. 11–31.

[4] R. D. Fierro and J. R. Bunch, *Collinearity and total least squares*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1167–1181.

[5] R. D. Fierro and J. R. Bunch, *Perturbation theory for orthogonal projection methods with applications to least squares and total least squares*, Linear Algebra Appl., 234 (1996), pp. 71–96.

[6] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Regularization by truncated total least squares*, SIAM J. Sci. Comput., 18 (1997), pp. 1223–1241.

[7] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[8] G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Tikhonov regularization and total least squares*, SIAM J. Matrix. Anal. Appl., 21 (1999), pp. 185–194.

[9] G. H. Golub, A. Hoffman, and G. W. Stewart, *A generalization of the Eckart-Young-Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.

[10] G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.

[11] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[13] R. D. D. Groat and E. M. Dowling, *The data least squares problem and channel equalization*, IEEE Trans. Signal Process., 42 (1993), pp. 407–411.

[14] M. Hanke, *On Lanczos based methods for the regularization of discrete ill-posed problems*, BIT, 41 (2001), pp. 1008–1018.

[15] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed problems: Numerical Aspects of Linear Inversion*, SIAM Monogr. Math. Model. Comput. 4, SIAM, Philadelphia, 1997.

[16] P. C. Hansen, T. Sekii, and H. Shibahashi, *The modified truncated SVD method for regularization in general form*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1142–1150.

[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[18] D. P. O'Leary and J. A. Simmons, *A bidiagonalization-regularization procedure for large-scale discretizations of ill-posed problems*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 474–489.

[19] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.

[20] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., 91 (2002), pp. 117–146.

[21] C. C. Paige and Z. Strakoš, *Unifying least squares, total least squares and data least squares*, in Total Least Squares and Errors-in-Variables Modeling, S. van Huffel and P. Lemmerling, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 25–34.

[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., SIAM, Philadelphia, 1997, pp. 11–20.

[23] A. Van der Sluis, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.

[24] A. Van der Sluis and G. W. Veltkamp, *Restoring rank and consistency by orthogonal projection*, Linear Algebra Appl., 28 (1979), pp. 257–278.
[25] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers Appl. Math. 9, SIAM, Philadelphia, 1991.
[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.
[27] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 735–743.
[28] A. I. Zhdanov and O. A. Katsyuba, *Strong consistency of estimates made by the method of orthogonal projections*, Internat. J. Systems Sci., 21 (1990), pp. 1463–1471.