SIAM J. MATRIX ANAL. APPL.
Vol. 00, No. 0, pp. 000-000

CHARACTERIZING MATRICES THAT ARE CONSISTENT WITH GIVEN SOLUTIONS*

X.-W. CHANG†, C. C. PAIGE‡, AND D. TITLEY-PELOQUIN§

Abstract. For given vectors $b \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$ we describe a unitary transformation approach to deriving the set $\mathcal{F}$ of all matrices $F \in \mathbb{C}^{m\times n}$ such that y is an exact solution to the compatible system $Fy = b$. This is used for deriving minimal backward errors E and f such that $(A+E)y = b+f$ when possibly noisy data $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$ are given, and the aim is to decide if y is a satisfactory approximate solution to $Ax = b$. The approach might be different, but the above results are not new. However we also prove the apparently new result that two well known approaches to making this decision are theoretically equivalent, and discuss how such knowledge can be used in designing effective stopping criteria for iterative solution techniques. All these ideas generalize to the following formulations. We extend our constructive approach to derive a superset $\mathcal{F}_{STLS+}$ of the set $\mathcal{F}_{STLS}$ of all matrices $F \in \mathbb{C}^{m\times n}$ such that y is a scaled total least squares solution to $Fy \approx b$. This is a new general result that specializes in two important ways. The ordinary least squares problem is an extreme case of the scaled total least squares problem, and we use our result to obtain the set $\mathcal{F}_{LS}$ of all matrices $F \in \mathbb{C}^{m\times n}$ such that y is an exact least squares solution to $Fy \approx b$. This complements the original less-constructive derivation of Waldén, Karlson and Sun [Numerical Linear Algebra with Applications, 2:271–286 (1995)]. We do the equivalent for the data least squares problem—the other extreme case of the scaled total least squares problem. Not only can the results be used as indicated above for the compatible case, but the constructive technique we use could also be applicable to other backward problems—such as those for under-determined systems, the singular value decomposition, and the eigenproblem.

Key words. matrix characterization, approximate solutions, iterative methods, linear algebraic equations, least squares, data least squares, total least squares, scaled total least squares, backward errors, stopping criteria.

AMS subject classifications. 15A06, 15A29, 65F05, 65F10, 65F20, 65F25, 65G99.

DOI. (Digital Object Identifier)

1. Introduction.

We will study a class of 'backward' problems for linear systems $Fy \approx b$. Specifically, given two vectors y and b we want to find the sets of all matrices F such that y is the exact solution (i.e., $Fy = b$), the least squares (LS) solution, the data least squares (DLS) solution, and the scaled total least square (STLS) solution. We will propose a unified unitary transformation approach to handling these problems.

Some of these problems have been investigated before. The result for the compatible case is well-known and the result for the least squares case was obtained elegantly by Waldén, Karlson and Sun in [31]. But while [31] presents, then proves, the least squares result, our approach shows how to derive a more general result in a fairly simple way, and we suspect that this constructive approach is not only easier to comprehend for non-mathematicians, but perhaps easier to apply to other problems.

Thus the technique we use is widely applicable and an important part of this paper: it is to transform the unknown matrices F from the left and right by certain theoretical unitary matrices related to the given vectors b and y. These are designed so that the constraints, such as $Fy = b$ in the compatible case, reveal the structure of the set of possible matrices F.

* Submitted ...
†‡§ School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7
† (chang@cs.mcgill.ca) Research supported by NSERC of Canada Grant RGPIN217191-07.
‡ (paige@cs.mcgill.ca) Research supported by NSERC of Canada Grant RGPIN9236.
§ (dtitle@cs.mcgill.ca) Research supported by NSERC of Canada PGS-M Fellowship.

One of the main uses for finding sets of matrices consistent with given approximate solutions is to find minimal backward errors, see for example section 2.2 here. If the normwise relative backward error is of the order of the unit round-off then we say that the approximate solution is a (normwise) backward stable solution. This is useful in practice, for sometimes we do not know if an algorithm for solving a problem is numerically stable—but if we know that a computed solution of a specific problem is a backward stable solution, we are usually satisfied with this computed solution. Also when we solve a problem by an iterative algorithm, the minimal backward error can often be used to design effective stopping criteria. There has been a lot of work on backward error problems, especially in recent years. For example for consistent linear systems (including structured problems), see [10], [12], [19], [24], [25], [30] and [33]; for unconstrained least squares problems, see [6], [8], [11]–[14], [20]–[23] and [31]; for constrained least squares problems, see [3], [13] and [14]; and for data least squares problems, see [2].

To illustrate the basic ideas and techniques of our approach and the uses of the results, we start with the simplest compatible case in section 2. In section 3 we find a useful superset $\mathcal{F}_{STLS+}$ of the set $\mathcal{F}_{STLS}$ of matrices consistent with given STLS solutions. The STLS problem is a generalization of the ordinary least squares (LS) and data least squares (DLS) problems. From the results for the STLS problem we obtain the set $\mathcal{F}_{LS}$ of consistent matrices for the LS problem, and a superset $\mathcal{F}_{DLS+}$ of the set $\mathcal{F}_{DLS}$ of consistent matrices for DLS problem. The results are given in sections 4 and 5. We have not been able to find simple and practical representations of $\mathcal{F}_{DLS}$ and $\mathcal{F}_{STLS}$, but we discuss in section 6 how the sets $\mathcal{F}_{STLS+}$ and $\mathcal{F}_{DLS+}$ can be just as useful.

In problems which have known structure we will sometimes only be interested in those matrices with that structure, see for example several papers in [27] and [28] for total least squares problems. We have not looked at such problems, but "structure" can take many forms, and it might be that some structures can be described as subsets of the sets we derive here.

We will use $I = [e_1, \ldots, e_n]$ to denote the unit matrix; $\|x\|^2 \equiv x^H x$; $\|B\|_2 \equiv \sigma_{max}(B)$, the maximum singular value of B; $\|B\|_F^2 \equiv \mathrm{trace}(B^H B)$. We will use $B^\dagger$ to represent the Moore-Penrose generalized inverse of B. For any complex vector v,

$$v^\dagger \equiv \begin{cases} 0 & \text{if } v = 0, \\ v^H/\|v\|^2 & \text{if } v \neq 0, \end{cases}$$

and $P_{v^\perp} = I - vv^\dagger$ is always the projector onto the orthogonal complement of range(v). We will regularly use the following: if $v \in \mathbb{C}^n$ and $V_2 \in \mathbb{C}^{n\times(n-1)}$, then

$$V_2 V_2^H = I - vv^\dagger \quad \Leftrightarrow \quad V \equiv [v/\|v\|, V_2] \in \mathbb{C}^{n\times n} \quad \text{is unitary.} \tag{1.1}$$

2. The compatible case. Our analysis for known results for compatible linear systems $Ax = b$ will provide the basic ideas and techniques used throughout this paper. The useful Lemma 2.1 and an apparently new result Corollary 2.3 will be given.

2.1. The set of consistent matrices F for $Fy = b$. The backward problem is the following: given $b \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$ we wish to characterize all $F \in \mathbb{C}^{m\times n}$ such that y is the exact solution to $Fy = b$. We can write

$$\mathcal{F} = \mathcal{F}(b, y) \equiv \{F \in \mathbb{C}^{m\times n} : \ Fy = b\}. \tag{2.1}$$

An explicit expression for $\mathcal{F}$ can be used in various problems such as finding optimal such F, structured such F, *etc.*. We now obtain such an explicit characterization of

all $F \in \mathcal{F}$. Note that if $y = 0$ then every $F \in \mathbb{C}^{m\times n}$ will do if $b = 0$, but none will do if $b \neq 0$, and we now only consider $y \neq 0$.

LEMMA 2.1. *For given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$ define*

$$\begin{aligned} \mathcal{F} &\equiv \{F \in \mathbb{C}^{m\times n} : \ Fy = b\}, \\ \mathcal{N} &\equiv \{by^\dagger + Z(I - yy^\dagger) : \ Z \in \mathbb{C}^{m\times n}\}. \end{aligned} \tag{2.2}$$

Then $\mathcal{F} = \mathcal{N}$, and any $F \in \mathcal{F}$ can also be written as

$$F = by^\dagger + G_2 Y_2^H \quad \textit{where} \quad G_2 \in \mathbb{C}^{m\times(n-1)}, \quad Y_2 Y_2^H = I - yy^\dagger. \tag{2.3}$$

Proof. In the theory of generalized inverses any solution X of $XA = B$ can be written as $X = BA^\dagger + Z(I - AA^\dagger)$, so $\mathcal{F} = \mathcal{N}$ follows immediately. However the following constructive derivation provides the useful representation (2.3), and can be extended to solve other backward problems such as those in sections 3, 4, and 5.

Note in (2.3) that Y_2 has $n-1$ columns, so from (1.1) we see that $Y = [y/\|y\|, Y_2]$ is unitary. Such unitary matrices are the tools we use in our constructive derivations of sets such as those above. For any $F \in \mathcal{F}$, let Y be any unitary matrix of the form $Y = [y/\|y\|, Y_2]$, so that $Y^H y = e_1\|y\|$. In order to describe our sets we introduce an unknown matrix G. Specifically we define $G \equiv FY \in \mathbb{C}^{m\times n}$ so that the constraint $Fy = b$ can be rewritten as

$$Fy = FYY^H y = Ge_1\|y\| = b. \tag{2.4}$$

We will show how (2.4) limits the possible G. Express G as $G \equiv [g_1, G_2]$ for some vector g_1. Then (2.4) gives $g_1\|y\| = b$, so that $g_1 = b/\|y\|$ and $F = GY^H = by^\dagger + G_2 Y_2^H$, proving (2.3). Here G_2, and only G_2, is independent of the constraint $Fy = b$.

We can replace $G_2 Y_2^H$ by $ZY_2 Y_2^H$ as follows. For any $Z \in \mathbb{C}^{m\times n}$ define $G_2 \equiv ZY_2$, so that $G_2 Y_2^H = ZY_2 Y_2^H$. Conversely, for any G_2 we can define $Z \equiv G_2 Y_2^H$ so that $G_2 Y_2^H = G_2 Y_2^H Y_2 Y_2^H = ZY_2 Y_2^H$. Thus we can use (1.1) to rewrite any F in (2.3) as

$$F = by^\dagger + ZY_2 Y_2^H = by^\dagger + Z(I - yy^\dagger)$$

for some totally unknown $Z \in \mathbb{C}^{m\times n}$. Thus $F \in \mathcal{F} \Rightarrow F \in \mathcal{N}$, and so $\mathcal{F} \subseteq \mathcal{N}$. But if $F \in \mathcal{N}$, then clearly $Fy = b$, so $F \in \mathcal{F}$, proving $\mathcal{N} \subseteq \mathcal{F}$, and thus $\mathcal{F} = \mathcal{N}$. ∎

If $y = 0$ and $b = 0$ it is easy to see that $\mathcal{F} = \mathcal{N}$ still holds, but (2.3) does not since no such Y_2 can exist.

Notice that although (2.2) is a compact explicit representation of all possible matrices F such that $Fy = b$, the equivalent (2.3) shows there are other representations. The most useful representation will depend on the problem being solved.

Compatible linear systems are a distinct and important special case of each of the later problems we examine. It is helpful to continue this introduction by applying Lemma 2.1 to give some well-known results and an interesting corollary. These illustrate how these set representations might be used in general.

2.2. Minimal Backward Errors and Acceptable Solutions. In this section we will only consider the matrix 2- and F-norms, and use one description for both. Thus $\eta_{2,F}$ *etc.* will indicate that one can either use the matrix 2-norm throughout, or the F-norm throughout.

Given $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, suppose we wish to find the smallest (in some sense) perturbations E and f in A and b such that $(A+E)y = b+f$. One approach proposed by Rigal and Gaches [19, section 3.1] is to essentially solve

$$\eta_{2,F} \equiv \min_{\eta,E,f}\{\eta :\ \|E\|_{2,F} \le \eta\alpha\|A\|_{2,F},\ \|f\| \le \eta\beta\|b\|,\ (A+E)y = b+f\} \tag{2.5}$$

for given $\alpha \ge 0$ and $\beta \ge 0$ (not both zero). See Remark 2.1 for comments on $\eta_{2,F}$. Another well-known approach, see for example [9, Problem 7.8], is to solve

$$\zeta \equiv \min_{E,f}\{\|[E, f\theta]\|_F :\ (A+E)y = b+f\} \tag{2.6}$$

for some given real scalar $\theta \ge 0$. This approach was originally used in [31] for least squares probelms.

Although the solutions are known for the above two approaches, these are apparently not compared in the literature. We prove in Corollary 2.3 that for the 2- and F-norms the two approaches are in fact theoretically equivalent.

THEOREM 2.2. *Given $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$, nonzero $y \in \mathbb{C}^n$, along with nonnegative real scalars α and β (not both zero), and $\theta \ge 0$, then with the definitions*

$$r \equiv b - Ay, \quad \mu_{2,F} \equiv \frac{\beta\|b\|}{\alpha\|A\|_{2,F}\|y\| + \beta\|b\|}, \quad \nu \equiv \frac{1}{1+\theta^2\|y\|^2}, \tag{2.7}$$

the minimum in (2.5) is

$$\eta_{2,F} = \frac{\|r\|}{\alpha\|A\|_{2,F}\|y\| + \beta\|b\|} \tag{2.8}$$

which is reached with the optimal

$$\hat{E} = r(1-\mu_{2,F})y^\dagger, \quad \hat{f} = -r\mu_{2,F}, \tag{2.9}$$

while the minimum in (2.6) is

$$\zeta = \left(\frac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2}\right)^{1/2} \tag{2.10}$$

which is reached with the optimal

$$\hat{E} = r(1-\nu)y^\dagger, \quad \hat{f} = -r\nu. \tag{2.11}$$

Proof. The quickest proof is to follow the approach of Higham [9, Thm. 7.1]: for each of (2.8) and (2.10) it is straightforward to show that the righthand side is a lower bound on the minimand, and that the stated optimal values give the lower bound.

But for possible future work it is useful to see how the actual solutions can be *found* via Lemma 2.1. Using the notation of Lemma 2.1 we see from (2.1) and (2.3) that for any given f, any E satisfying the constraint $(A+E)y = b+f$ has the form

$$E = (b+f)y^\dagger + G_2Y_2^H - A, \qquad Y_2Y_2^H = I - yy^\dagger,$$

for some $G_2 \in \mathbb{C}^{m\times(n-1)}$. Therefore with unitary Y of the form $Y = [y/\|y\|, Y_2]$,

$$\begin{aligned}\|E\|_{2,F}^2 = \|EY\|_{2,F}^2 &= \|[(b+f)y^\dagger + G_2Y_2^H - A][y/\|y\|, Y_2]\|_{2,F}^2 \\ &= \|[(b+f-Ay)/\|y\|, G_2 - AY_2]\|_{2,F}^2 \\ &\geq \|r+f\|^2/\|y\|^2.\end{aligned}$$

The last inequality becomes an equality if $G_2 = AY_2$, which is independent of f. Thus $G_2 = AY_2$ is optimal for both (2.5) and (2.6), and so with this G_2 we have $E = (b+f)y^\dagger - Ayy^\dagger = (r+f)y^\dagger$ for both problems.

Note that we can always write $f = -r\mu + u$ for some (possibly complex) scalar μ and some $u \in \mathbb{C}^m$ such that $u^H r = 0$, so

$$\|f\|^2 = \|r\|^2|\mu|^2 + \|u\|^2, \qquad \|r+f\|^2 = \|r\|^2|1-\mu|^2 + \|u\|^2.$$

But $\|E\|_{2,F} = \|r+f\|/\|y\|$, from which we can see that the minima in both (2.5) and (2.6) require $u = 0$ and μ real, since for a given real part $\mu_{\mathcal{R}}$, both $|\mu|$ and $|1-\mu|$ are minimized by taking $\mu = \mu_{\mathcal{R}}$. This gives $f = -r\mu$, $E = r(1-\mu)y^\dagger$.

The theorem is obvious when $r = 0$, so assume $r \neq 0$. For (2.5) we solve

$$\min_\mu\{\eta : |1-\mu|\cdot\|r\| \leq \eta\alpha\|A\|_{2,F}\|y\|,\ |\mu|\cdot\|r\| \leq \eta\beta\|b\|\},$$

from which we can see that if $\alpha = 0$ then $\mu = 1$; if $\beta = 0$ then $\mu = 0$; otherwise the minimum occurs when

$$\eta = \frac{|1-\mu|\cdot\|r\|}{\alpha\|A\|_{2,F}\|y\|} = \frac{|\mu|\cdot\|r\|}{\beta\|b\|}, \quad 0 < \mu < 1,$$

giving $\mu = \beta\|b\|/(\alpha\|A\|_{2,F}\|y\|+\beta\|b\|) = \mu_{2,F}$ in all cases, so that the optimal $\eta = \eta_{2,F}$, proving (2.8) with its minimizers (2.9).

In (2.10)

$$\|[E, f\theta]\|_F^2 = [(1-\mu)^2 + \mu^2\theta^2\|y\|^2]\|r\|^2/\|y\|^2,$$

which is minimized by $\mu = (1+\theta^2\|y\|^2)^{-1} = \nu$, $1-\mu = \nu\theta^2\|y\|^2$, proving (2.10) with its minimizers (2.11), and completing this longer but constructive proof. ∎

Rigal and Gaches [19, section 3.1] essentially proved (2.8), while the result (2.10) is well known, see for example [9, Problem 7.8]. Here we relate these two results.

COROLLARY 2.3. *With the notation in Theorem 2.2, taking θ in (2.6) to be*

$$\theta_{2,F} \equiv \begin{cases} \left(\dfrac{\alpha\|A\|_{2,F}}{\beta\|b\|\cdot\|y\|}\right)^{1/2}, & \text{if } \beta > 0 \\ \infty, & \text{if } \beta = 0 \end{cases} \tag{2.12}$$

makes the optimal $\hat{E}$ and $\hat{f}$ for (2.10) identical to the optimal $\hat{E}$ and $\hat{f}$ for (2.8).

Proof. From Theorem 2.2 we see that the optimal $\hat{E}$ and $\hat{f}$ have the same forms $E = r(1-\mu)y^\dagger$ and $f = -r\mu$, where the only differences are in the values of μ. The values of μ become the same by choosing θ so that $\nu = \mu_{2,F}$, that is

$$\nu^{-1} = 1 + \theta^2\|y\|^2 = \mu_{2,F}^{-1} = (\alpha\|A\|_{2,F}\|y\| + \beta\|b\|)/\beta\|b\|,$$

giving $\theta = \theta_{2,F}$ in (2.12) when $\beta > 0$. If $\beta = 0$ then taking $\theta = \infty$ results in $\nu = \mu_{2,F} = 0$, which forces $f = 0$, *c.f.* the DLS case in section 5. □

Thus in order to define optimal backward perturbations in these cases, it does not matter which of the theoretical approaches (2.8) or (2.10) we take, as long as we choose α and β, or θ, according to (2.12).

The quantity $\eta_{2,F}$ can be used to check if an approximate solution to $Ax = b$ is an *acceptable solution*. Most practical problems contain uncertainties in the data, and instead of solving $Ax = b$ with ideal data A and b, we solve some system

$$(A + \delta A)\tilde{x} = b + \delta b, \quad \text{where} \quad \|\delta A\|_{2,F} \le \alpha \|A\|_{2,F}, \ \|\delta b\| \le \beta \|b\| \tag{2.13}$$

for some hopefully approximately known $\alpha \ge 0$ and $\beta \ge 0$. Notice that the given y solves a problem within the range of uncertainty in the data (2.13) if and only if $\eta_{2,F} \le 1$, so that if $\eta_{2,F} \le 1$ we can conclude that the given y is an acceptable solution to the compatible system $Ax = b$, and this can be used as a stopping criterion for iterative methods such as MGS-GMRES (see [15]).

REMARK 2.1. *If $\alpha = \beta = 1$ in (2.5), then from (2.8) $\eta_{2,F}$ becomes the normwise relative backward error (NRBE) $\|r\|/(\|A\|_{2,F}\|y\|+\|b\|)$ in [9, p. 120]. This is excellent for plotting the performance of an iterative solution of equations algorithm, and can be used in the stopping criterion $\eta_{2,F} \le O(\epsilon)$ for a numerically stable algorithm. To handle $\alpha \ne \beta$ in (2.13) we chose $\eta_{2,F}$ as in (2.5). This is then neither a NRBE nor a direct measure of backward error, and it has to be used with the very different stopping criterion $\eta_{2,F} \le 1$. But in this case it is easy to define and compute the* two distinct *NRBEs—that in A: $\eta_{2,F}\alpha = \|\hat{E}\|_{2,F}/\|A\|_{2,F}$, and that in b: $\eta_{2,F}\beta = \|\hat{f}\|/\|b\|$.*

A knowledge of the uncertainties will usually suggest rough values for α and β. If we do not know such values, or want maximum accuracy in a normwise backward sense, we can use a backward stable algorithm and take $\alpha = \beta = O(\epsilon)$, where ϵ is the floating point arithmetic unit roundoff and $O(\epsilon)$ depends on the algorithm. For example it was shown in [15, sections 1 and 8.2] that for sufficiently nonsingular $n \times n$ A in the real problem $Ax = b$, for MGS-GMRES we would use the F-norm, and might take $\alpha = \beta = 100kn\epsilon$ at step k if we wanted to be unrealistically careful, or more sensibly $\alpha = \beta = 10n\epsilon$, where experience suggests we can usually obtain even better accuracy than this.

Sometimes we will only have an estimate of the α/β ratio, or of the equivalent θ satisfying (2.12). For example we might only know that the relative error in b can be about ten times that in A. If we have no idea of the individual α, β values, we do not have a clear acceptance criterion. For certainty we could assume that α and β were very small, and in the case $\beta/\alpha \ge 1$ we could set $\alpha = O(\epsilon)$ and $\beta = (\beta/\alpha)O(\epsilon)$. For example, when using MGS-GMRES, if we know that the relative error in b is about ten times that in A, we might set $\alpha = 10n\epsilon$, $\beta = 100n\epsilon$.

If only θ, or the ratio α/β, is available, the quantity ζ in (2.6) is sometimes referred to as a normwise backward error, see, for example, [9, Problem 7.8]. But it is important to be aware that this quantity can be a poor measure of backward error for small θ. This is because $\zeta \to 0$ as $\theta \to 0$, see (2.10), while (2.6) shows that the optimal $E \to 0$, $f \to Ay - b = -r$ as $\theta \to 0$, so that if $r \ne 0$ then ζ will be an inappropriate measure when θ is small. A generally more appropriate measure of backward error for the $\|[E, f\theta]\|_F$ approach might be $\|[\hat{E}, \hat{f}]\|_F^2$, where with (2.11) and (2.7)

$$\|[\hat{E}, \hat{f}]\|_F^2 = \frac{1 + \theta^4\|y\|^2}{(1 + \theta^2\|y\|^2)^2}\|r\|^2. \tag{2.14}$$

Note that this quantity is equal to ζ when $\theta = 1$ and in the limit as $\theta \to \infty$, but tends to the desired $\|r\|^2$ as $\theta \to 0$. Thus although for a given θ we *minimize* $\|[E, f\theta]\|_F$, a more meaningful *indicator* of the backward error might be $\|[\hat{E}, \hat{f}]\|_F$.

Finally for fixed $\|y\|$ and $\|r\|$ the minimum of (2.14) is given by $\theta^2 = 1$. This is one argument for taking $\theta = 1$ if we have no reasonable *a priori* idea of α/β or θ.

3. The scaled total least squares problem. Given $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$, and $\gamma \in (0, \infty)$, the scaled total least squares (STLS) problem was formulated in [18] as finding $\hat{E}$, $\hat{f}$, and $\hat{x}$ which solve

$$\sigma_S \equiv \min_{E,f,x} \|[E, f\gamma]\|_F \quad \text{subject to} \quad (A + E)x = b + f. \tag{3.1}$$

By taking $g = f\gamma$, (3.1) was reformulated in [16, (5.1)] as

$$\sigma_S \equiv \min_{E,g,x} \|[E, g]\|_F \quad \text{subject to} \quad (A + E)x\gamma = b\gamma + g. \tag{3.2}$$

The scalar σ_S is called the STLS distance, and $\hat{x} = \hat{x}(\gamma)$ the STLS solution. The formulation (3.1) is closely related to the minimal backward error problem for compatible systems (2.6), while (3.2) has the advantage of being an unscaled total least squares problem, for which codes are easily available.

Let $\mathcal{U}_{min}$ be the left singular vector subspace of A corresponding to its minimum singular value $\sigma_{min}(A)$. In [16] it was shown that a satisfactory condition for building the theory for the STLS problem is the condition that we will now assume holds:

$$\text{the } m \times n \text{ matrix } A \text{ has rank } n, \text{ and } b \not\perp \mathcal{U}_{min}. \tag{3.3}$$

Under this condition the solution to (3.2) must exist and be unique.

The STLS solution reduces to the ordinary least squares solution in the limit as $\gamma \to 0$ (so $E = 0$), to the unscaled total least squares solution when $\gamma = 1$, and to the data least squares solution in the limit as $\gamma \to \infty$ (so $f = 0$), see, for example, [16].

It was shown in [4] for the real case, and in [16, (5.9)] for the complex case, that the STLS solution $\hat{x}$ solves

$$\sigma_S^2 = \min_x \left\{ \sigma_S^2(x) \equiv \frac{\|b - Ax\|^2}{\gamma^{-2} + \|x\|^2} \right\}. \tag{3.4}$$

If we differentiate the real version of $\sigma_S^2(x)$ in (3.4) with respect to x and equate the result to zero, we see that $\hat{x}$ satisfies the real version of

$$A^H(b - A\hat{x}) = -\hat{x}\frac{\|b - A\hat{x}\|^2}{\gamma^{-2} + \|\hat{x}\|^2} = -\hat{x}\sigma_S^2(\hat{x}).$$

This is a necessary optimality condition, but it is not sufficient since the function $\sigma_S^2(x)$ is not convex. In fact it can be proven (see [29, Theorem 2.7], [16, Sec 6]) that when (3.3) holds, $\hat{x}$ solves (3.2) if and only if

$$A^H(b - A\hat{x}) = -\hat{x}\sigma_S^2, \quad \sigma_S^2 \equiv \frac{\|b - A\hat{x}\|^2}{\gamma^{-2} + \|\hat{x}\|^2} < \sigma_{min}^2(A). \tag{3.5}$$

Given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, the backward STLS problem is then to characterize the set $\mathcal{F}_{STLS}$ of all $F \in \mathbb{C}^{m\times n}$ such that y is the exact STLS solution to

$Fy \approx b$, see (3.1). From (3.4) and (3.5), the sets $\mathcal{F}_{STLS}$ and $\mathcal{F}_{STLS+}$ can be defined as follows:

$$\mathcal{F}_{STLS} \equiv \left\{ F \in \mathbb{C}^{m\times n} : \ \frac{\|b-Fy\|^2}{\gamma^{-2}+\|y\|^2} = \min_{x\in\mathbb{C}^n} \frac{\|b-Fx\|^2}{\gamma^{-2}+\|x\|^2} \right\} \tag{3.6}$$

$$\equiv \left\{ F \in \mathbb{C}^{m\times n} : F^H(b-Fy) = -y\frac{\|b-Fy\|^2}{\gamma^{-2}+\|y\|^2}, \ \frac{\|b-Fy\|^2}{\gamma^{-2}+\|y\|^2} < \sigma^2_{\min}(F) \right\} \tag{3.7}$$

$$\subseteq \mathcal{F}_{STLS+} \equiv \left\{ F \in \mathbb{C}^{m\times n} : \ F^H(b-Fy) = -y\frac{\|b-Fy\|^2}{\gamma^{-2}+\|y\|^2} \right\}. \tag{3.8}$$

There might be elements of $\mathcal{F}_{STLS+}$ which do not satisfy the inequality in (3.7). In other words there might be $F \in \mathcal{F}_{STLS+}$ for which the given y is not the minimizer, but merely a stationary point, of the right hand side in (3.6). In the use of such sets optimizing over $\mathcal{F}_{STLS}$ would be difficult, so in practice we would usually choose to use the more amenable $\mathcal{F}_{STLS+}$. One reason for this is that any particular element found in $\mathcal{F}_{STLS+}$ could be tested to see if it also satisfied (3.7). A more important reason is that in all the problems we can imagine we would be given a y that is a reasonable approximation to $\hat{x}$, and use this to find an $F \in \mathcal{F}_{STLS+}$ which is as close as possible to A. It can be seen from (3.5) that if $y = \hat{x}$ then the closest F in any measure would be A itself, so that F *would* satisfy (3.7). If we only had $y \approx \hat{x}$ then finding F as close as possible to A would tend to force (3.7) to hold. A good example of this is in [2] which deals with the minimum backward error for an approximate solution y to the DLS problem, see section 5. Using the notation in (5.4), in [2] we used

$$\mu_F(y) \equiv \min_{A+\Delta A\in\mathcal{F}_{DLS+}} \|\Delta A\|_F \quad \text{in order to find} \quad \hat{\mu}_F(y) \equiv \min_{A+\Delta A\in\mathcal{F}_{DLS}} \|\Delta A\|_F,$$

and proved in [2, Theorem 2.8] that if $\hat{x}$ is the DLS solution to $Ax \approx b$ then there exists an $\epsilon > 0$ such that if $\|y-\hat{x}\|_2 < \epsilon$ then $\mu_F(y) = \hat{\mu}_F(y)$. So that for a good approximation y nothing is lost by using $\mathcal{F}_{DLS+}$ instead of $\mathcal{F}_{DLS}$. In fact in the thousands of numerical tests in [2, Section 5] no example was found where using $\mathcal{F}_{DLS+}$ gave the wrong answer, where the y were chosen to have relative errors up to 10^{-1}. Since $\mathcal{F}_{DLS}$ is a limiting case of $\mathcal{F}_{STLS}$, we suspect that in many practical cases $\mathcal{F}_{STLS+}$ will also be a useful and usable replacement for $\mathcal{F}_{STLS}$.

To develop an explicit expression for all $F \in \mathcal{F}_{STLS+}$ we will use the following lemma as a guide.

LEMMA 3.1. *If $F \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$, nonzero $y \in \mathbb{C}^n$ and $\gamma \in (0,\infty)$, then*

$$F^H(b-Fy) = -y\|b-Fy\|^2/(\gamma^{-2}+\|y\|^2) \tag{3.9}$$

$$\iff w = b-Fy, \quad (I-yy^\dagger)F^Hw = 0, \quad b^Hw = \frac{\|w\|^2}{1+\gamma^2\|y\|^2} \tag{3.10}$$

$$\iff w = b-Fy, \quad (F^H+y\gamma^2b^H)w = 0. \tag{3.11}$$

Proof. Define $w \equiv b-Fy$, then

$$\|w\|^2 = b^Hw - y^HF^Hw. \tag{3.12}$$

Suppose that (3.9) holds. Multiplying (3.9) on the left by y^H gives

$$y^HF^Hw = -\|y\|^2\|w\|^2/(\gamma^{-2}+\|y\|^2),$$

which with (3.12) leads to the last equality in (3.10):

$$b^H w = \|w\|^2/(1+\gamma^2\|y\|^2). \tag{3.13}$$

The second equality in (3.10) can be obtained immediately by multiplying (3.9) on the left by $I - yy^\dagger$. Thus (3.10) holds. From (3.9) and (3.13) we obtain

$$F^H w = -y\|w\|^2/(\gamma^{-2}+\|y\|^2) = -y\gamma^2 b^H w,$$

leading to (3.11).

Conversely if (3.10) holds, then using its second, first and third equalities we have

$$F^H w = \frac{y(y^H F^H w)}{\|y\|^2} = \frac{y(b^H w - w^H w)}{\|y\|^2} = \frac{y(\frac{\|w\|^2}{1+\gamma^2\|y\|^2} - \|w\|^2)}{\|y\|^2} = -y\frac{\|w\|^2}{\gamma^{-2}+\|y\|^2},$$

so that (3.9) holds. Finally if (3.11) holds, then from its two equalities we obtain

$$\|w\|^2 = b^H w - y^H F^H w = b^H w - y^H(-y\gamma^2 b^H w) = (1+\gamma^2\|y\|^2) b^H w.$$

Therefore the second equality in (3.11) can be rewritten as

$$F^H w = -y\gamma^2 b^H w = -y\gamma^2\|w\|^2/(1+\gamma^2\|y\|^2),$$

so that (3.9) holds. ∎

We now obtain two new characterizations of all matrices $F \in \mathcal{F}_{STLS+}$ in (3.8).

THEOREM 3.2. *For given $b \in \mathbb{C}^m$, nonzero $y \in \mathbb{C}^n$ and $\gamma \in (0,\infty)$ write*

$$\mathcal{F}_{STLS+} \equiv \Big\{ F \in \mathbb{C}^{m\times n} :\ F^H(b-Fy) = -y\frac{\|b-Fy\|^2}{\gamma^{-2}+\|y\|^2} \Big\}, \tag{3.14}$$

$$\mathcal{N}_{STLS+} \equiv \Big\{ (b-w)y^\dagger + (I-ww^\dagger)Z(I-yy^\dagger) : \tag{3.15}$$
$$w \in \mathbb{C}^m,\ b^H w = \frac{\|w\|^2}{1+\gamma^2\|y\|^2},\ Z \in \mathbb{C}^{m\times n} \Big\},$$

$$\widetilde{\mathcal{N}}_{STLS+} \equiv \{ -\tilde{w}\tilde{w}^\dagger b\gamma^2 y^H + (I-\tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I-yy^\dagger)] : \tilde{w} \in \mathbb{C}^m, Z \in \mathbb{C}^{m\times n} \}. \tag{3.16}$$

Then $\mathcal{F}_{STLS+} = \mathcal{N}_{STLS+} = \widetilde{\mathcal{N}}_{STLS+}$, and for any matrix F in these identical sets, the corresponding w in the representation (3.15) satisfies $w = b - Fy$ and is a scalar multiple of the corresponding $\tilde{w}$ in the representation (3.16).

Proof. In order to show that $\mathcal{F}_{STLS+} \subseteq \mathcal{N}_{STLS+}$ and $\mathcal{F}_{STLS+} \subseteq \widetilde{\mathcal{N}}_{STLS+}$, consider any $F \in \mathcal{F}_{STLS+}$ so that by Lemma 3.1

$$w \equiv b - Fy, \quad (F^H + y\gamma^2 b^H)w = 0. \tag{3.17}$$

If $w = 0$ then $Fy = b$, so from Lemma 2.1 we see that $F \in \mathcal{F} = \mathcal{N}$. But obviously $\mathcal{N} \subseteq \mathcal{N}_{STLS+}$ and $\mathcal{N} \subseteq \widetilde{\mathcal{N}}_{STLS+}$, thus $F \in \mathcal{N}_{STLS+}$ and $F \in \widetilde{\mathcal{N}}_{STLS+}$. Now assume $w \neq 0$. Write $\hat{w} \equiv w/\|w\|$, $\hat{y} \equiv y/\|y\|$, and let $W = [\hat{w}, W_2] \in \mathbb{C}^{m\times m}$ and $Y = [\hat{y}, Y_2] \in \mathbb{C}^{n\times n}$ be unitary matrices, so that $Y^H y = e_1\|y\|$ and $W^H w = e_1\|w\|$. Define $G \equiv W^H F Y \in \mathbb{C}^{m\times n}$, so $F = WGY^H$. Then multiplying the first and second equalities in (3.17) by W^H and Y^H from the left, respectively, leads to

$$Ge_1\|y\| = W^H(b-w), \quad G^H e_1\|w\| = -e_1(\|y\|\gamma^2 b^H w). \tag{3.18}$$

We will now show how (3.18) limits the possible G. Write

$$m \times n \;\; G = \begin{bmatrix} g_{11} & g_1^H \\ g_2 & G_{22} \end{bmatrix}, \;\; \text{with} \;\; (m-1) \times (n-1) \;\; G_{22}.$$

Then from (3.18) we obtain

$$\begin{bmatrix} g_{11} \\ g_2 \end{bmatrix} \|y\| = \begin{bmatrix} \hat{w}^H(b-w) \\ W_2^H b \end{bmatrix}, \quad \begin{bmatrix} g_{11}^H \\ g_1 \end{bmatrix} \|w\| = \begin{bmatrix} -\|y\|\gamma^2 b^H w \\ 0 \end{bmatrix},$$

leading to

$$g_1 = 0, \quad g_2 = W_2^H b/\|y\|, \quad g_{11} = \hat{w}^H(b-w)/\|y\| = -\|y\|\gamma^2 \hat{w}^H b. \tag{3.19}$$

From these it follows that

$$\begin{aligned} F = WGY^H &= (\hat{w}g_{11} + W_2 g_2)\hat{y}^H + W_2 G_{22} Y_2^H \\ &= ww^\dagger(b-w)y^\dagger + W_2 W_2^H b y^\dagger + W_2 G_{22} Y_2^H \qquad (3.20) \\ &= -ww^\dagger b\gamma^2 y^H + W_2 W_2^H b y^\dagger + W_2 G_{22} Y_2^H. \qquad (3.21) \end{aligned}$$

Similarly to what we did in the proof of Lemma 2.1, we can replace $W_2 G_{22} Y_2^H$ in (3.20) and (3.21) by $W_2 W_2^H Z Y_2 Y_2^H$ for some totally unknown $Z \in \mathbb{C}^{m\times n}$. Then using (1.1) we have from (3.20) and (3.21) that

$$\begin{aligned} F &= ww^\dagger(b-w)y^\dagger + (I-ww^\dagger)by^\dagger + (I-ww^\dagger)Z(I-yy^\dagger) \\ &= (b-w)y^\dagger + (I-ww^\dagger)Z(I-yy^\dagger) \qquad (3.22) \\ &= -ww^\dagger b\gamma^2 y^H + (I-ww^\dagger)[by^\dagger + Z(I-yy^\dagger)]. \qquad (3.23) \end{aligned}$$

From (3.22) and (3.10) it follows that $F \in \mathcal{N}_{STLS+}$, so $\mathcal{F}_{STLS+} \subseteq \mathcal{N}_{STLS+}$; and from (3.23) it follows that $F \in \widetilde{\mathcal{N}}_{STLS+}$, and therefore $\mathcal{F}_{STLS+} \subseteq \widetilde{\mathcal{N}}_{STLS+}$.

Conversely suppose that $F \in \mathcal{N}_{STLS+}$, so that

$$F = (b-w)y^\dagger + (I-ww^\dagger)Z(I-yy^\dagger)$$

for some Z and some w satisfying $b^H w = \|w\|^2/(1+\gamma^2\|y\|^2)$. Then it follows that

$$Fy = b - w, \quad (I-yy^\dagger)F^H w = 0. \tag{3.24}$$

Therefore by Lemma 3.1 and (3.14) $F \in \mathcal{F}_{STLS+}$, thus $\mathcal{N}_{STLS+} \subseteq \mathcal{F}_{STLS+}$, proving that $\mathcal{N}_{STLS+} = \mathcal{F}_{STLS+}$. Finally suppose that $F \in \widetilde{\mathcal{N}}_{STLS+}$, so that

$$F = -\tilde{w}\tilde{w}^\dagger b\gamma^2 y^H + (I-\tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I-yy^\dagger)]$$

for some Z and $\tilde{w}$. Then $\tilde{w}^H F = -\tilde{w}^H b\gamma^2 y^H$, so $(F^H + y\gamma^2 b^H)\tilde{w} = 0$, and

$$Fy = -\tilde{w}\tilde{w}^\dagger b\gamma^2\|y\|^2 + (I-\tilde{w}\tilde{w}^\dagger)b = b - \tilde{w}\tilde{w}^\dagger b(1+\gamma^2\|y\|^2).$$

This gives an expression for w defined by

$$w \equiv b - Fy = \tilde{w}[\tilde{w}^\dagger b(1+\gamma^2\|y\|^2)]. \tag{3.25}$$

We see by Lemma 3.1 and (3.14) that $F \in \mathcal{F}_{STLS+}$, and thus $\widetilde{\mathcal{N}}_{STLS+} \subseteq \mathcal{F}_{STLS+}$, proving that $\mathcal{F}_{STLS+} = \widetilde{\mathcal{N}}_{STLS+}$. The first equality in (3.24) and (3.25) indicates that w in (3.15) is a scalar multiple of $\tilde{w}$ in (3.16) for the same matrix F. ∎

Here we make two remarks. Unlike the expression for $\mathcal{N}_{STLS+}$ in (3.15), the expression for $\widetilde{\mathcal{N}}_{STLS+}$ in (3.16) does not involve any constraint and so is easier to use. But if we want to consider $\gamma \to \infty$, it is easier to use $\mathcal{N}_{STLS+}$, see section 5.

The condition (3.3) does not necessarily hold for every $F \in \mathcal{F}_{STLS+}$— an example is the rank-1 matrix $by^\dagger \in \mathcal{F}_{STLS+}$ in (3.8) which does not have full column rank if $n > 1$. Nor need it hold for every $F \in \mathcal{F}_{STLS}$. This knowledge needs to be taken into account in the use of these sets, but at least we know that for every $F \in \mathcal{F}_{STLS+}$, y gives a stationary point of $\|b - Fx\|^2/(\gamma^{-2} + \|x\|^2)$, see (3.4).

This completes our theory for the general STLS formulation. We will now use Theorem 3.2 to characterize matrices consistent with given approximate solutions for its two extreme cases: the least squares and data least squares problems.

4. The Least Squares Problem. Given $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$, the ordinary least squares (LS) problem is defined as

$$\sigma_{LS} \equiv \min_{f,x} \|f\| \quad \text{subject to} \quad Ax = b + f.$$

It is well known that $\hat{x}$ is the LS solution if and only if it satisfies the normal equations:

$$A^H(b - A\hat{x}) = 0.$$

See for example [1] or [5] for useful background.

Given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, the backward LS problem is then to characterize the set $\mathcal{F}_{LS}$ of all $F \in \mathbb{C}^{m\times n}$ such that y is the exact LS solution to $Fy \approx b$. Obviously we have

$$\mathcal{F}_{LS} \equiv \{F \in \mathbb{C}^{m\times n} : \|b - Fy\|^2 = \min_{x\in\mathbb{C}^n} \|b - Fx\|^2\} = \{F \in \mathbb{C}^{m\times n} : F^H(b - Fy) = 0\}.$$

We now give an alternative derivation to that in [31] of an explicit representation for all $F \in \mathcal{F}_{LS}$.

THEOREM 4.1. *Given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, write*

$$\begin{aligned}\mathcal{F}_{LS} &\equiv \{F \in \mathbb{C}^{m\times n} :\ \|b - Fy\|^2 = \min_{x\in\mathbb{C}^n} \|b - Fx\|^2\} \\ &= \{F \in \mathbb{C}^{m\times n} :\ F^H(b - Fy) = 0\}, \qquad (4.1)\\ \widetilde{\mathcal{N}}_{LS} &\equiv \{(I - \tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I - yy^\dagger)] :\ \tilde{w} \in \mathbb{C}^m, Z \in \mathbb{C}^{m\times n}\}. \qquad (4.2)\end{aligned}$$

Then $\mathcal{F}_{LS} = \widetilde{\mathcal{N}}_{LS}$, and for any matrix F in these two identical sets, $w \equiv b - Fy$ is a scalar multiple of the corresponding $\tilde{w}$ in the representation (4.2).

Proof. This theorem could be proved by the same approach as that used in proving Theorem 3.2. But we can obtain the results directly from Theorem 3.2. Notice from (3.14) and (4.1) that

$$\mathcal{F}_{LS} = \lim_{\gamma\to 0} \mathcal{F}_{STLS+}.$$

Now since $\mathcal{F}_{STLS+} = \widetilde{\mathcal{N}}_{STLS+}$ from Theorem 3.2, it follows that

$$\begin{aligned}\mathcal{F}_{LS} &= \lim_{\gamma\to 0} \mathcal{F}_{STLS+} = \lim_{\gamma\to 0} \widetilde{\mathcal{N}}_{STLS+} \\ &= \left\{(I - \tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I - yy^\dagger)] : \tilde{w} \in \mathbb{C}^m, Z \in \mathbb{C}^{m\times n}\right\} = \widetilde{\mathcal{N}}_{LS}.\end{aligned}$$

The conclusion that $w = b - Fy$ is a scalar multiple of $\tilde{w}$ still holds. In fact this can be seen from (3.25) by taking $\gamma \to 0$. ∎

In [31] Waldén, Karlson, and Sun gave the original and elegant proof that $\mathcal{F}_{LS} = \widetilde{\mathcal{N}}_{LS}$, and used the result to find the minimal backward error for the LS problem. It is not the intent of this paper to find minimal backward errors, and we will now specialize the general result of Theorem 3.2 to DLS problems.

5. The Data Least Squares Problem. Given $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$, the data least squares (DLS) problem is defined as (see [7] and for example [16, 17]):

$$\sigma_{_D} \equiv \min_{E,x} \|E\|_F \quad \text{subject to} \quad (A+E)x = b. \tag{5.1}$$

When $\gamma \to \infty$, the STLS problem (3.1) becomes the DLS problem (5.1), see [16]. The condition (3.3) is still needed for building the theory for the DLS problem.

It is easy to show that (5.1) is equivalent to (see, e.g., [16])

$$\sigma_{_D}^2 = \min_x \frac{\|b - Ax\|^2}{\|x\|^2}. \tag{5.2}$$

From [16, (5.14)–(5.17)], when (3.3) holds, $\hat{x}$ solves (5.1) if and only if

$$A^H(b - A\hat{x}) = -\hat{x}\sigma_{_D}^2, \quad \sigma_{_D}^2 \equiv \frac{\|b - A\hat{x}\|^2}{\|\hat{x}\|^2} < \sigma_{min}^2(A). \tag{5.3}$$

Both (5.2) and (5.3) can also be obtained by taking $\gamma \to \infty$ in (3.4) and (3.5).

Given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, the backward DLS problem is then to characterize the set $\mathcal{F}_{DLS}$ of all $F \in \mathbb{C}^{m\times n}$ such that y is the exact DLS solution to $Fy \approx b$. As in the STLS problem, the sets $\mathcal{F}_{DLS}$ and $\mathcal{F}_{DLS+}$ can be defined as follows:

$$\begin{aligned}\mathcal{F}_{DLS} &\equiv \left\{F \in \mathbb{C}^{m\times n} : \ \frac{\|b - Fy\|^2}{\|y\|^2} = \min_{x\in\mathbb{C}^n} \frac{\|b - Fx\|^2}{\|x\|^2}\right\} \\ &\subseteq \mathcal{F}_{DLS+} \equiv \left\{F \in \mathbb{C}^{m\times n} : \ F^H(b - Fy) = -y\frac{\|b - Fy\|^2}{\|y\|^2}\right\}.\end{aligned} \tag{5.4}$$

Comments paralleling those given after (3.8) and Theorem 3.2 apply here as well.

We now obtain an explicit characterization for all $F \in \mathcal{F}_{DLS+}$.

THEOREM 5.1. *For given $b \in \mathbb{C}^m$ and nonzero $y \in \mathbb{C}^n$, write*

$$\begin{aligned}\mathcal{F}_{DLS+} &\equiv \left\{F \in \mathbb{C}^{m\times n} : \ F^H(b - Fy) = -y\frac{\|b - Fy\|^2}{\|y\|^2}\right\}, \\ \mathcal{N}_{DLS+} &\equiv \{(b - w)y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger) : w \in \mathbb{C}^m,\ b^Hw = 0,\ Z \in \mathbb{C}^{m\times n}\}.\end{aligned} \tag{5.5}$$

Then $\mathcal{F}_{DLS+} = \mathcal{N}_{DLS+}$, and for any $F \in \mathcal{N}_{DLS+}$, the w in the representation (5.5) satisfies $w = b - Fy$.

Proof. We could prove this theorem by using a constructive derivation similar to that used in proving Theorem 3.2. Instead we obtain the results by taking the limit $\gamma \to \infty$ for the results in Theorem 3.2. In fact we have

$$\begin{aligned}\mathcal{F}_{DLS+} &= \lim_{\gamma\to\infty} \mathcal{F}_{STLS+} = \lim_{\gamma\to\infty} \mathcal{N}_{STLS+} \\ &= \{(b - w)y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger) : \ b^Hw = 0,\ Z \in \mathbb{C}^{m\times n}\} \\ &= \mathcal{N}_{DLS+}.\end{aligned}$$

The conclusion that w in (5.5) satisfies $w = b - Fy$ still holds, and can also be verified by forming Fy for any $F \in \mathcal{N}_{DLS+}$. ∎

The result of Theorem 5.1 is used in [2] for the backward perturbation analysis for the DLS problem.

6. Summary and Comments. Given $b \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$ we have presented a unitary transformation approach to finding sets, or supersets, of all matrices $F \in \mathbb{C}^{m\times n}$ such that y is the solution to $Fy \approx b$ for some common classes of approximation problems.

Our approach is constructive and easy to follow. We have used the well-known compatible case $Fy = b$ to illustrate this approach in its simplest setting, as well as illustrating one of the uses of such sets—finding minimal backward errors. In doing so we have shown the equivalence of two often used problem formulations for such errors—an apparently new result.

We then applied this approach to finding new and useful supersets of matrices consistent with the STLS solution to $Fy \approx b$. From (3.1) or (3.5) the STLS solution becomes the LS solution as $\gamma \to 0$, and becomes the DLS solution as $\gamma \to \infty$, see, for example, [16, §6]. Based on these facts, we derived the results for the LS and DLS problems using the results of Theorem 3.2 directly, although we could have separately given a full constructive derivation for these two problems, similar to that in Theorem 3.2.

We summarize the different problems and sets we have obtained as follows:

- The STLS problem, see (3.8) and Theorem 3.2:

$$\begin{aligned}
\mathcal{F}_{STLS} &\equiv \left\{ F \in \mathbb{C}^{m\times n} : \frac{\|b - Fy\|^2}{\gamma^{-2} + \|y\|^2} = \min_{x\in\mathbb{C}^n} \frac{\|b - Fx\|^2}{\gamma^{-2} + \|x\|^2} \right\} \\
\subseteq \mathcal{F}_{STLS+} &\equiv \left\{ F \in \mathbb{C}^{m\times n} : F^H(b - Fy) = -y\frac{\|b - Fy\|^2}{\gamma^{-2} + \|y\|^2} \right\} \\
= \mathcal{N}_{STLS+} &\equiv \Big\{ (b - w)y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger) : \\
&\qquad w \in \mathbb{C}^m,\, w^H b = \frac{\|w\|^2}{1 + \gamma^2\|y\|^2},\, Z \in \mathbb{C}^{m\times n} \Big\} \\
= \widetilde{\mathcal{N}}_{STLS+} &\equiv \{ -\tilde{w}\tilde{w}^\dagger b\gamma^2 y^H + (I - \tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I - yy^\dagger)] : \tilde{w} \in \mathbb{C}^m, Z \in \mathbb{C}^{m\times n} \}.
\end{aligned}$$

- Compatible systems, see Lemma 2.1:

$$\begin{aligned}
\mathcal{F} &\equiv \{ F \in \mathbb{C}^{m\times n} : Fy = b \} \\
= \mathcal{N} &\equiv \{ by^\dagger + Z(I - yy^\dagger) :\ Z \in \mathbb{C}^{m\times n} \}.
\end{aligned}$$

- The LS problem, see Theorem 4.1:

$$\begin{aligned}
\mathcal{F}_{LS} &\equiv \{ F \in \mathbb{C}^{m\times n} : \|b - Fy\| = \min_{x\in\mathbb{C}^n} \|b - Fx\| \} \\
&= \{ F \in \mathbb{C}^{m\times n} :\ F^H(b - Fy) = 0 \} \\
= \widetilde{\mathcal{N}}_{LS} &\equiv \{ (I - \tilde{w}\tilde{w}^\dagger)[by^\dagger + Z(I - yy^\dagger)] :\ \tilde{w} \in \mathbb{C}^m,\ Z \in \mathbb{C}^{m\times n} \}.
\end{aligned}$$

- The DLS problem, see Theorem 5.1:

$$\begin{aligned}\mathcal{F}_{DLS} &\equiv \left\{ F \in \mathbb{C}^{m\times n} : \frac{\|b-Fy\|^2}{\|y\|^2} = \min_{x\in\mathbb{C}^n} \frac{\|b-Fx\|^2}{\|x\|^2} \right\} \\ \subseteq \mathcal{F}_{DLS+} &\equiv \left\{ F \in \mathbb{C}^{m\times n} : F^H(b-Fy) = -y\frac{\|b-Fy\|^2}{\|y\|^2} \right\} \\ = \mathcal{N}_{DLS+} &\equiv \{(b-w)y^\dagger + (I-ww^\dagger)Z(I-yy^\dagger) : b^Hw = 0, w\in\mathbb{C}^m, Z\in\mathbb{C}^{m\times n}\}.\end{aligned}$$

The sets $\mathcal{F}_{STLS+}$ and $\mathcal{F}_{DLS+}$ are supersets of $\mathcal{F}_{STLS}$ and $\mathcal{F}_{DLS}$, respectively. But theoretical arguments and numerical experiments given in [2] have shown that when y is a reasonable approximation to the solution of the DLS problem for $Ax \approx b$, the set $\mathcal{N}_{DLS+}$ can usually be used with no further constraints to obtain the minimal backward errors for the DLS problem. This is probably true for the STLS problem as well. However since such behavior is problem dependent we will not discuss it further here, except to state that for many practical uses $\mathcal{N}_{STLS+}$ or $\widetilde{\mathcal{N}}_{STLS+}$ can be used in place of $\mathcal{F}_{STLS}$, and $\mathcal{N}_{DLS+}$ can be used in place of $\mathcal{F}_{DLS}$.

The constructive technique we use could also be applicable to other backward problems, e.g., finding a matrix whose partial eigenvalues and eigenvectors are known.

Acknowledgment. We would like to thank the referees for their helpful comments.

REFERENCES

[1] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[2] X.-W. Chang, G. H. Golub and C. C. Paige, *Towards a backward perturbation analysis for data least squares problems*, to appear in SIAM J. Matrix Anal. Appl..
[3] A. J. Cox and N. J. Higham, *Backward error bounds for constrained least squares problems*, BIT, 39 (1999), pp. 210–227.
[4] 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.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore MD, 3rd Edn., 1996.
[6] J. F. Grcar, *Optimal sensitivity analysis of linear least squares*, Technical Report LBNL-52434. Lawrence Berkeley National Laboratory, 2003.
[7] R. D. D. Groat and E. M. Dowling, *The data least squares problem and channel equalization*, IEEE Trans. Signal Processing, 42 (1993), pp. 407–411.
[8] M. Gu, *Backward perturbation bounds for linear least squares problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 363–372.
[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms* SIAM Publications, Philadelphia PA, 2nd Edn., 2002.
[10] D. J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.
[11] R. Karlson and B. Waldén, *Estimation backward perturbation bounds for the linear least squares problem*, BIT, 37 (1997), pp. 862-869.
[12] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured Mapping Problems for Matrices Associated with Scalar Products Part I: Lie and Jordan Algebras*, Technical Report. Manchester Institute for Mathematical Sciences, 2006.
[13] A. N. Malyshev, *Optimal backward perturbation bounds for the LSS problems*, BIT, 41 (2001), pp. 430–432.
[14] A. N. Malyshev and M. Sadkane, *Computation of optimal backward perturbation bounds for large sparse linear least squares problems*, BIT, 41 (2002), pp. 739–747.
[15] C. C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram-Schmidt (MGS), Least Squares, and Backward Stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284.
[16] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numerische Mathematik, 91 (2002), pp. 117–146.

[17] 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.

[18] 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 Publications, Philadelphia PA, 1997, pp. 11–20.

[19] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, JACM, 14 (1967), pp. 543–548.

[20] G. W. Stewart, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Review, 19 (1977), pp. 634–662.

[21] Z. Su, *Computational Methods for Least Squares Problems and Clinical Trials*, Ph.D Thesis, Scientific Computing & Computational Mathematics, Stanford University, 2005.

[22] J.-G. Sun, *Optimal backward perturbation bounds for the linear least squares problem with multiple right-hand sides*, IMA J. Numer. Anal., 16 (1996), pp. 1–11.

[23] J.-G. Sun, *On optimal backward perturbation bounds for the linear least-squares problem*, BIT, 37 (1997), pp. 179–188.

[24] J.-G. Sun, *Bounds for the structured backward errors of Vandermonde systems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 45–59.

[25] J.-G. Sun, *A note on backward errors for structured linear systems*, Numerical Linear Algebra with Applications, 12(7), 2005, pp 585-603.

[26] J.-G. Sun and Z. Sun, *Optimal backward perturbation bounds for underdetermined systems*. SIAM J. Matrix Anal. Appl., 18 (1997), pp. 393–402.

[27] S. Van Huffel, ed., *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, SIAM Publications, Philadelphia PA, 1997.

[28] S. Van Huffel and P. Lemmerling, eds., *Total Least Squares and Errors-in-Variables Modeling*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[29] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM Publications, Philadelphia PA, 1991.

[30] J. M. Varah, *Backward error estimates for Toeplitz systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 408–417.

[31] B. Waldén, R. Karlson and J. Sun, *Optimal backward perturbation bounds for the linear least squares problem*, Numerical Linear Algebra with Applications, 2 (1995), pp. 271–286.

[32] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, 1965.

[33] H. Xiang and Y. Wei *On normwise structured backward errors for saddle point systems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 838-849.