MULTIPLICATIVE PERTURBATION THEORY AND ACCURATE SOLUTION OF LEAST SQUARES PROBLEMS [*]

NIEVES CASTRO-GONZÁLEZ[†], JOHAN CEBALLOS[‡], FROILÁN M. DOPICO[§], AND JUAN M. MOLERA[‡]

Abstract. Least squares problems $\min_x \|b-Ax\|_2$ where the matrix $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ has some particular structure arise frequently in applications. Polynomial data fitting is a well-known instance of problems that yield highly structured matrices A, but many other examples exist. Very often, structured matrices have huge condition numbers $\kappa_2(A) = \|A\|_2 \, \|A^\dagger\|_2$ (here $A^\dagger$ denotes the Moore-Penrose pseudo-inverse of A) and, therefore, standard algorithms fail to compute accurate minimum 2-norm solutions of least squares problems. In this work, we introduce a framework that allows us to compute minimum 2-norm solutions of many classes of structured least squares problems accurately, i.e., with errors $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2 = O(\mathtt{u})$, where $\mathtt{u}$ is the unit roundoff, independently of the magnitude of $\kappa_2(A)$ for most vectors b. The cost of these accurate computations is $O(n^2m)$ flops, i.e., roughly the same cost as standard algorithms for least squares problems. The approach in this work relies in computing first an accurate rank-revealing decomposition of A, an idea that has been widely used in the last decades to compute, for structured ill-conditioned matrices, singular value decompositions, eigenvalues and eigenvectors in the Hermitian case, and solutions of linear systems with high relative accuracy. In order to prove that accurate solutions are computed, it is needed to develop a multiplicative perturbation theory of least squares problems and Moore-Penrose pseudo-inverses. The results and algorithms presented in this paper are valid in the case of both full rank and rank deficient problems and also in the case of underdetermined linear systems $(m < n)$. Among other types of matrices, the new method applies to rectangular Cauchy, Vandermonde, and graded matrices and detailed numerical tests for these matrices are presented.

Key words. accurate solutions, Cauchy matrices, diagonally dominant matrices, graded matrices, least squares problems, Moore-Penrose pseudo-inverse, multiplicative perturbation theory, rank revealing decompositions, structured matrices, Vandermonde matrices

AMS subject classifications. 65F20, 65F35, 15A09, 15A12, 15A23, 15B05.

1. Introduction.

Matrices with particular structures arise frequently in theory and applications [38, 39]. As a consequence, the design and analysis of special algorithms for performing structured matrix computations is a classical area of Numerical Linear Algebra that attracts the attention of many researchers. Special algorithms for solving structured linear systems of equations or structured eigenvalue problems are included in many standard references [14, 24, 28, 32, 45], but special algorithms for solving structured least squares problems do not appear so often in the literature. The goal of special algorithms is to exploit the structure of the problem to increase the speed of computations, and/or to decrease storage requirements, and/or to improve the accuracy of the solutions in comparison with standard algorithms. On this latter goal, let us mention that special algorithms for solving structured linear systems of equations more accurately than standard methods have been developed from the early days of Numerical Linear Algebra [4] and many papers have been published on this topic since then (see the references in [18, 28]). The development of accurate algorithms for structured eigenvalue problems is much more recent, since it started in early 90's with the celebrated paper [10] and has also received considerable attention (see, for instance, [2, 9, 13, 17, 19, 20, 21, 23, 31, 43, 47] among many other references). The present paper focuses on a part of *"Accurate Numerical Linear Algebra"* for which there are

[*] This work was partially supported by the Ministerio de Economía y Competitividad of Spain through grants MTM-2009-09281 (J. Ceballos, F. M. Dopico, and J. M. Molera) and MTM2010-18057 (N. Castro-González).

[†] Facultad de Informática, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain (nieves@fi.upm.es).

[‡] Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain (jceballo@math.uc3m.es, molera@math.uc3m.es).

[§] Instituto de Ciencias Matemáticas CSIC-UAM-UC3M-UCM and Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain (dopico@math.uc3m.es).

not many references available in the literature: algorithms for solving structured least squares problems $\min_x \|b - Ax\|_2$, where $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^{m\times 1}$, with much more accuracy than the one provided by standard algorithms and roughly with the same computational cost, that is, $O(n^2 m)$ flops. We only know reference [35] on this topic, which focuses on a very particular class of least squares problems.

The standard method for solving full column-rank least squares (LS) problems $\min_x \|b - Ax\|_2$ is via the QR factorization computed with the Householder algorithm [28, Chapters 19 and 20]. This method is backward stable, that is, the computed solution $\widehat{x}_0$ is the exact solution of a LS problem $\min_x \|(b + \Delta b) - (A + \Delta A)x\|_2$, where $\|\Delta b\|_2 \leq c\,\mathrm{u}\,mn\,\|b\|_2$, $\|\Delta A\|_2 \leq c\,\mathrm{u}\,m\,n^{3/2}\,\|A\|_2$, u is the unit roundoff of the computer, and c denotes a small integer constant [28, Theorem 20.3]. Backward error results of the same type hold for other methods of solution of LS problems based on orthogonal decompositions as, for instance, the singular value decomposition (SVD)*. This strong backward error result, together with classical normwise perturbation theory of LS problems [46, Theorem 5.1] (see also [3, Theorem 1.4.6, p. 30]), implies the following forward error bound in the computed solution $\widehat{x}_0$ with respect the exact solution x_0

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq (c\,\mathrm{u}\,m\,n^{3/2}) \left(\kappa_2(A) + \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} + \kappa_2(A)^2 \, \frac{\|b - Ax_0\|_2}{\|A\|_2\,\|x_0\|_2} \right), \tag{1.1}$$

where $A^\dagger$ is the Moore-Penrose pseudo-inverse of A, $\|A\|_2$ denotes the spectral norm of A, and $\kappa_2(A) = \|A\|_2\,\|A^\dagger\|_2$ is the spectral condition number of A. The bound in (1.1) is larger than $\mathrm{u}\,\kappa_2(A)$ (in fact, it can be much larger under certain conditions) and, so, (1.1) does not guarantee any digit of accuracy in the computed solution if $\kappa_2(A) \gtrsim 1/\mathrm{u}$, that is, if A is ill-conditioned with respect to the inverse of the unit roundoff. Unfortunately, many types of structured matrices arising in applications are extremely ill-conditioned and standard algorithms for LS problems may compute solutions with huge relative errors. Two famous examples are Vandermonde matrices, which arise in polynomial data fitting, and Cauchy matrices [28, Chapters 22 and 28].

Our goal in this work is to present a numerical framework for the solution of LS problems and to prove rigorously that it allows us to compute for many classes of structured matrices solutions with guaranteed error bounds much smaller than the one in (1.1). The framework we introduce relies on the concept of *rank-revealing decomposition* (RRD), originally introduced in [9] for computing the SVD with high relative accuracy -see also [28, Sec 9.12]. An RRD of $A \in \mathbb{C}^{m\times n}$ is a factorization $A = XDY$, where $X \in \mathbb{C}^{m\times r}$, $D = \mathrm{diag}(d_1, d_2, \ldots, d_r) \in \mathbb{C}^{r\times r}$ is diagonal and nonsingular, and $Y \in \mathbb{C}^{r\times n}$, $\mathrm{rank}\,(X) = \mathrm{rank}\,(Y) = r$, and X and Y are well conditioned. Note that this means that the rank of A is r, and that if A is ill conditioned, then the diagonal factor D is also ill conditioned. We propose to compute the minimum 2-norm solution of $\min_x \|b - Ax\|_2$ in two main stages:

1. First stage. Compute an RRD of $A = XDY$, accurately in the sense of [9] (we revise the precise meaning of "accuracy" in this context in Definition 2.3).
2. Second stage. It has three steps: (1) compute the unique solution x_1 of $\min_x \|b - Xx\|_2$ via Householder QR factorization; (2) compute the solution x_2 of the linear system $Dx_2 = x_1$ as $x_2(i) = x_1(i)/d_i$, $i = 1 : r$; and (3) compute the minimum 2-norm solution x_0 of the underdetermined linear system $Yx = x_2$ using

*It should be noted that the backward error in A committed by solving LS problems via the Householder QR factorization is columnwise, i.e., $\|\Delta A(:,j)\|_2 \leq c\,\mathrm{u}\,mn\,\|A(:,j)\|_2$ for $j = 1 : n$ (MATLAB notation), and, therefore, it is stronger than the one mentioned above. However, this columwise bound does not hold for the solution via the SVD, since orthogonal transformations are applied to A from both sides.

the Q method [28, Chapter 21]. The vector x_0 is the minimum 2-norm solution of $\min_x \|b - Ax\|_2$.

The intuition behind why this procedure computes accurate solutions, even for extremely ill conditioned matrices A, is that each entry of x_2 is computed with a relative error less than $\mathtt{u}$, that is, the ill conditioned linear system $Dx_2 = x_1$ is solved very accurately, together with the fact that $\min_x \|b - Xx\|_2$ and $Yx = x_2$ are also solved accurately because X and Y are well conditioned. We will prove in Section 6 that the relative error for the minimum 2-norm solution $\widehat{x}_0$ computed by the proposed procedure is

$$\text{(1.2)} \qquad \frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, f(m,n) \left(\kappa_2(Y) + \kappa_2(X) \, \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} \right),$$

where $f(m,n)$ is a modestly growing function of m and n. Note first that (1.2) improves (1.1), because X and Y are well conditioned and, so, the only potentially large factor in (1.2) is $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$, which also appears in (1.1). But the really important point on the bound (1.2) is that if A is fixed, then $\|A^\dagger\|_2 \, \|b\|_2 / \|x_0\|_2$ is small for most right-hand sides b, even for very ill conditioned matrices A. This fact is well-known if A is square and nonsingular (see [1, 7] and [18, Section 3.2]) and, as we will explain in Subsection 4.1, it also holds for general matrices in two senses: for most vectors b that are everywhere in the space, and for most vectors b with a fixed value of the relative residual $\|Ax_0 - b\|_2 / \|b\|_2$ not too close to 1. In this paper the sentence "for most vectors b" may be understood in any of these two senses.

The framework and the results discussed above resemble those presented in [18] for computing accurate solutions of structured linear systems $Ax = b$ with A nonsingular. However, not surprisingly, the analysis for LS problems is much more complicated and requires completely different techniques for developing the new multiplicative perturbation theory that is needed to prove the error bound in (1.2). In addition, the results and algorithms we present are fully general, since they remain valid both for full rank and rank defective matrices A, and, although we focus mainly on LS problems, they can be also applied to solve accurately underdetermined linear systems.

The computation of an *accurate* RRD $A = XDY$ is the difficult part in the framework proposed above. For almost any matrix $A \in \mathbb{C}^{m \times n}$ an RRD (potentially inaccurate) can be computed by applying standard Gaussian elimination with complete pivoting (GECP) to get, barring permutations, an LDU factorization, where $X = L \in \mathbb{C}^{m \times r}$ is unit lower trapezoidal (notation from [28, p. 355]), $D \in \mathbb{C}^{r \times r}$ is diagonal and nonsingular, and $Y = U \in \mathbb{C}^{r \times n}$ is unit upper trapezoidal [9, 28]. Other option is to use the Householder QR algorithm with column-pivoting and take $X = Q$, $D = \operatorname{diag}(R)$, and $Y = D^{-1}R$, barring permutations. Very rarely, GECP or QR with column-pivoting fail to produce well conditioned X and Y factors, but then other pivoting strategies that guarantee well conditioned factors are available in [25, 37, 40]. However neither standard GECP nor QR with column-pivoting are accurate for ill conditioned matrices and, nowadays, RRDs with guaranteed accuracy can be computed only for particular classes of structured matrices through special implementations of GECP that exploit carefully the structure to obtain accurate factors and, in the case of graded matrices, also through Householder QR factorization with *complete* pivoting [27].

Fortunately, as a by-product of the intense research performed in the last two decades on computing SVDs with high relative accuracy, there are algorithms to compute accurate RRDs of many classes of $m \times n$ structured matrices in $O(m\,n^2)$ operations. These classes include: Cauchy matrices, diagonally scaled Cauchy matrices, Vandermonde matrices, and some "related unit-displacement-rank" matrices [8]; graded matrices (that is, matrices of the form $S_1 B S_2$ with B well conditioned and S_1 and S_2 diagonal) [9, 27]; acyclic matrices (which include bidiagonal matrices), total signed compound matrices, diagonally scaled totally unimodular matrices [9]; diagonally dominant M-matrices [11, 41]; polynomial Vandermonde matrices involving orthonormal polynomials [12]; and diagonally dominant matrices [16, 47]. In addition, for certain real symmetric structured matrices, it is possible to compute accurate RRDs that preserve the symmetry. These symmetric matrices include: symmetric positive definite matrices SHS, with H well conditioned and S diagonal [13, 36]; and symmetric Cauchy, symmetric diagonally scaled Cauchy, and symmetric Vandermonde matrices [15]. Most of the algorithms cited in this paragraph determine exactly the rank of rank-deficient matrices and for all classes of matrices listed in this paragraph, the framework introduced in this paper solves LS problems with relative errors bounded as in (1.2). This error bound is $O(\mathtt{u}\, f(m,n))$ for most right-hand sides independently of the traditional condition number of the matrices and so guarantees accurate solutions.

The paper is organized as follows. We introduce in Section 2 the basic notations, concepts, and results that will be used throughout the paper. Section 3 studies the variation of the Moore-Penrose pseudo-inverse under multiplicative perturbations and, based on these results, Section 4 presents multiplicative perturbation bounds for LS problems. As a consequence, we get in Section 5 perturbation bounds for LS problems whose coefficient matrix is given as an RRD under perturbations of the factors. Section 6 presents a new algorithm for solving accurately LS problems via RRDs and the corresponding backward and forward error analyses. The accuracy of this algorithm is checked in practice via extensive numerical tests in Section 7. Finally, conclusions and lines of future work are discussed in Section 8.

2. Preliminaries and basic concepts. Since we consider LS problems, we will use the most natural norms for these problems: the Euclidean vector norm, i.e., given $x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n$, $\|x\|_2 := \left(|x_1|^2 + \cdots + |x_n|^2\right)^{1/2}$, and for matrices $A \in \mathbb{C}^{m\times n}$ the corresponding subordinate matrix norm $\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$, that is called the spectral or 2-norm of A. In Subsection 3.1, we will also use arbitrary (normalized) unitarily invariant matrix norms [44, Chapter II. Sec. 3], that will be denoted by $\|\cdot\|$. The symbol I_n stands for the $n \times n$ identity matrix, but we will use simply I if the size is clear from the context, and A^* denotes the conjugate-transpose of A. We will use MATLAB notation for submatrices: $A(i : j, :)$ indicates the submatrix of A consisting of rows i through j and $A(:, k : l)$ indicates the submatrix of A consisting of columns k through l. Given $A \in \mathbb{C}^{m\times n}$, with $m \geq n$, its singular values are denoted as $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$.

Lemma 2.1 will be needed to derive some perturbation bounds.

LEMMA 2.1. *Let $B, C \in \mathbb{C}^{m\times n}$, let $\mathcal{S} \subseteq \mathbb{C}^m$ and $\mathcal{W} \subseteq \mathbb{C}^n$ be vector subspaces, and let $P_{\mathcal{S}} \in \mathbb{C}^{m\times m}$ and $P_{\mathcal{W}} \in \mathbb{C}^{n\times n}$ be the orthogonal projectors onto $\mathcal{S}$ and $\mathcal{W}$, respectively. Then the following statements hold:*

(a) $\|P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}$.

(b) $\|BP_{\mathcal{W}} + C(I - P_{\mathcal{W}})\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}$.

Proof. Part (a). Let $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$. Since the vectors $P_{\mathcal{S}}Bx$ and $(I - P_{\mathcal{S}})Cx$ are orthogonal, then $\|(P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C)x\|_2^2 = \|P_{\mathcal{S}}Bx\|_2^2 + \|(I - P_{\mathcal{S}})Cx\|_2^2 \leq \|Bx\|_2^2 + \|Cx\|_2^2 \leq \|B\|_2^2 + \|C\|_2^2$ and

$$\|P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C\|_2 = \max_{\|x\|_2=1} \|(P_{\mathcal{S}}B + (I - P_{\mathcal{S}})C)x\|_2 \leq \sqrt{\|B\|_2^2 + \|C\|_2^2}\,.$$

Part (b) follows from part (a) applied to the conjugate-transpose matrix and the fact that for any matrix $\|A\|_2 = \|A^*\|_2$. □

In Sections 5 and 6, we will need the entrywise absolute value of a matrix. Given a matrix $G \in \mathbb{C}^{m\times n}$ with entries g_{ij}, we denote by $|G|$ the matrix with entries $|g_{ij}|$. Expressions like $|G| \leq |B|$, where $B \in \mathbb{C}^{m\times n}$, mean $|g_{ij}| \leq |b_{ij}|$ for $1 \leq i \leq m,\ 1 \leq j \leq n$.

The Moore-Penrose pseudo-inverse of $A \in \mathbb{C}^{m\times n}$ plays a key role in this work. It is defined to be the unique matrix $Z \in \mathbb{C}^{n\times m}$ such that

$$\text{(i) } AZA = A, \quad \text{(ii) } ZAZ = Z, \quad \text{(iii) } (AZ)^* = AZ, \quad \text{(iv) } (ZA)^* = ZA, \tag{2.1}$$

or, equivalently, such that

$$AZ = P_A \quad \text{and} \quad ZA = P_Z, \tag{2.2}$$

where P_A and P_Z stand for the orthogonal projectors onto the column spaces of A and Z, respectively. The equivalence of the four conditions in (2.1) and the two conditions in (2.2) can be easily established and can be found at [6, Theorem 1.1.1]. We will denote by $A^\dagger \in \mathbb{C}^{n\times m}$ the Moore-Penrose pseudo-inverse of $A \in \mathbb{C}^{m\times n}$. Recall that if $A \in \mathbb{C}^{n\times n}$ is nonsingular, then $A^\dagger = A^{-1}$. Recall also that the SVD of A allows us to get an expression for $A^\dagger$ and to prove many of its properties [44, Chapter 3]. $\mathcal{R}(A)$ will denote the column space of A and $\mathcal{N}(A)$ its null space. It is easy to see that $\mathcal{R}(A^*) = \mathcal{R}(A^\dagger)$, so, according to (2.2), $P_A = AA^\dagger$ and $P_{A^*} = P_{A^\dagger} = A^\dagger A$ are the orthogonal projectors onto $\mathcal{R}(A)$ and $\mathcal{R}(A^*)$, respectively.

We state without proof in Lemma 2.2 some well-known properties of the Moore-Penrose pseudo-inverse that will be needed throughout the paper. The proofs can be found in [6].

LEMMA 2.2.

(a) *If A has full row rank, then $A^\dagger = A^*(AA^*)^{-1}$ and $AA^\dagger = I$.*

(b) *If A has full column rank, then $A^\dagger = (A^*A)^{-1}A^*$ and $A^\dagger A = I$.*

(c) *Let $F \in \mathbb{C}^{m\times r}$ and $G \in \mathbb{C}^{r\times n}$. If $\operatorname{rank}(F) = \operatorname{rank}(G) = r$, then $(FG)^\dagger = G^\dagger F^\dagger$.*

The minimum 2-norm solution of the LS problem $\min_{x\in\mathbb{C}^n} \|b - Ax\|_2$ is $x_0 = A^\dagger b$ and the minimum 2-norm solution of an underdetermined linear system $Ax = b$ is also given by $x_0 = A^\dagger b$. If $A = XDY \in \mathbb{C}^{m\times n}$ is an RRD of A, then two successive applications of Lemma 2.2-(c) imply that $A^\dagger = Y^\dagger D^{-1} X^\dagger$ and the minimum 2-norm solution of LS problems or underdetermined linear systems is $x_0 = Y^\dagger D^{-1} X^\dagger b$.

Following [9], next we define the precise meaning of an *accurate* computed RRD of a matrix A. We add, with respect [9], the condition (2.5) that guarantees that the computed and exact "well conditioned" factors X and Y have condition numbers of similar magnitude.

DEFINITION 2.3. *Let $A = XDY$ be an RRD of $A \in \mathbb{C}^{m\times n}$, where $X \in \mathbb{C}^{m\times r}$, $D = \operatorname{diag}(d_1, \ldots, d_r) \in \mathbb{C}^{r\times r}$, and $Y \in \mathbb{C}^{r\times n}$, and let $\widehat{X} \in \mathbb{C}^{m\times r}$, $\widehat{D} = \operatorname{diag}(\widehat{d}_1, \ldots, \widehat{d}_r) \in \mathbb{C}^{r\times r}$, and $\widehat{Y} \in \mathbb{C}^{r\times n}$ be the factors computed by a certain algorithm in a computer with unit roundoff $\mathtt{u}$. We say that the factorization $\widehat{X}\widehat{D}\widehat{Y}$ has been accurately computed, or is an accurate RRD, if*

$$\frac{\|\widehat{X} - X\|_2}{\|X\|_2} \leq \mathtt{u}\, p(m,n), \quad \frac{\|\widehat{Y} - Y\|_2}{\|Y\|_2} \leq \mathtt{u}\, p(m,n), \quad \textit{and} \tag{2.3}$$

$$\frac{|\widehat{d}_i - d_i|}{|d_i|} \leq \mathtt{u}\, p(m,n), \quad i = 1:r, \tag{2.4}$$

where $p(m,n)$ is a modestly growing function of m and n, i.e., a function bounded by a low degree polynomial in m and n, such that

$$\max\{\kappa_2(X)\,,\, \kappa_2(Y)\}\, \mathtt{u}\, p(m,n) < 1/2\,. \tag{2.5}$$

For example, the algorithm to compute an RRD of an $m \times n$ $(m \geq n)$ real Cauchy matrix presented[†] in [8, Section 4] computes the factors with an entrywise relative error bounded by $9n\mathtt{u}/(1-9n\mathtt{u})$.

Let us discuss briefly, the role of condition (2.5). According to Weyl perturbation theorem [44], the differences between the ordered singular values of X and $\widehat{X}$ are bounded as follows $|\sigma_i(\widehat{X}) - \sigma_i(X)| \leq \|\widehat{X} - X\|_2 \leq \mathtt{u}\, p(m,n)\, \|X\|_2$, for $i = 1 : r$. Therefore, $|\sigma_i(\widehat{X}) - \sigma_i(X)|/\sigma_i(X) \leq \mathtt{u}\, p(m,n)\, \kappa_2(X)$, for $i = 1 : r$. A similar discussion holds for Y and $\widehat{Y}$. As a consequence, condition (2.5) implies $\operatorname{rank}(X) = \operatorname{rank}(\widehat{X}) = r$, $\operatorname{rank}(Y) = \operatorname{rank}(\widehat{Y}) = r$, and

$$\text{(2.6)} \qquad \frac{\kappa_2(X)}{3} \leq \kappa_2(\widehat{X}) \leq 3\,\kappa_2(X) \quad \text{and} \quad \frac{\kappa_2(Y)}{3} \leq \kappa_2(\widehat{Y}) \leq 3\,\kappa_2(Y)\,.$$

Equation (2.6) will allow us to use either $\kappa_2(X)$ and $\kappa_2(Y)$, or $\kappa_2(\widehat{X})$ and $\kappa_2(\widehat{Y})$ in the rounding error bounds obtained in Section 6 at the cost of modifying somewhat the constants involved in the bounds.

In the rounding error analysis of Section 6 we will use the conventional error model for floating point arithmetic [28, Section 2.2]

$$fl(a \odot b) = (a \odot b)(1 + \delta),$$

where a and b are real floating point numbers, $\odot \in \{+, -, \times, /\}$, and $|\delta| \leq \mathtt{u}$. Recall that this model also holds for complex floating point numbers if $\mathtt{u}$ is replaced by a slightly larger constant, see [28, Section 3.6]. In addition, we will assume that neither overflow nor underflow occurs.

3. Multiplicative perturbation results for the Moore-Penrose pseudo-inverse.

In this section and in Section 4, we consider a multiplicative perturbation of a general matrix $A \in \mathbb{C}^{m\times n}$, that is, a matrix $\widetilde{A} = (I+E)A(I+F)$, where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices. The final goal is to bound, in Section 4, $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$, where x_0 and $\widetilde{x}_0$ are the minimum 2-norm solutions of the LS problems $\min_{x\in\mathbb{C}^n} \|Ax - b\|_2$ and $\min_{x\in\mathbb{C}^n} \|\widetilde{A}x - \widetilde{b}\|_2$, respectively. This goal is achieved via the main theorem in this section, Theorem 3.2, where we obtain two expressions for $\widetilde{A}^\dagger$ in terms of $A^\dagger$, $(I+E)^{-1}$, and $(I+F)^{-1}$. We use the expressions of $\widetilde{A}^\dagger$ to develop in Subsection 3.1 bounds for $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|$ in any unitarily invariant norm and in the 2-norm. Although these bounds are not needed for our final purpose, we emphasize that the development of perturbation bounds for the Moore-Penrose pseudo-inverse is a classical topic in Matrix Analysis (see [46] and [44, Chapter 3, Sec. 3]) that has attracted the attention of many researchers. We show that the results in Subsection 3.1 are superior than those presented in [5], which are of a different nature and are obtained through a different approach. Multiplicative perturbation theory of matrices has received considerable attention in the literature in the context of accurate computations of eigenvalues and singular values [22, 29, 30, 33, 34] and also in the context of accurate solution of linear systems of equations [18, Lemma 3.1] but, as far as we know, it has not been studied yet in the context of accurate solution of LS problems.

Lemma 3.1 is a technical result that is used in the proof of Theorem 3.2.

LEMMA 3.1. *Let $A \in \mathbb{C}^{m\times n}$ and $\widetilde{A} = (I+E)A(I+F) \in \mathbb{C}^{m\times n}$, where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices. Then the following equalities hold:*

(a) $P_A(I+E^*)(I - P_{\widetilde{A}}) = 0$.

[†] The algorithm presented in [8, Section 4] covers only the square case $m = n$, but it is immediate to modify it for rectangular Cauchy matrices. This point will be further discussed in Section 7.

(b) $(I - P_{\widetilde{A}^*})(I + F^*)P_{A^*} = 0$.

Proof. (a) Since $\mathcal{R}(\widetilde{A}) = \mathcal{R}((I+E)A)$ then $(I-P_{\widetilde{A}})(I+E)A = 0$. Thus, $(I-P_{\widetilde{A}})(I+E)AA^\dagger = (I - P_{\widetilde{A}})(I + E)P_A = 0$, which is equivalent to $P_A(I + E^*)(I - P_{\widetilde{A}}) = 0$.

(b) Apply (a) to $\widetilde{A}^* = (I + F^*)A^*(I + E^*)$ and conjugate and transpose the equality. □

Next, we state the main result in this section, which is valid both for full rank and rank deficient matrices and for perturbations of any magnitude.

THEOREM 3.2. *Let $A \in \mathbb{C}^{m\times n}$ and $\widetilde{A} = (I + E)A(I + F) \in \mathbb{C}^{m\times n}$, where $(I + E) \in \mathbb{C}^{m\times m}$ and $(I + F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices. Then*

$$\widetilde{A}^\dagger = P_{\widetilde{A}^*}(I + F)^{-1}A^\dagger(I + E)^{-1}P_{\widetilde{A}} \tag{3.1}$$

and

$$\widetilde{A}^\dagger = \left(I + (I - P_{\widetilde{A}^*})F^* - P_{\widetilde{A}^*}\widehat{F}\right) A^\dagger \left(I + E^*(I - P_{\widetilde{A}}) - \widehat{E}P_{\widetilde{A}}\right), \tag{3.2}$$

where $\widehat{E} = (I + E)^{-1}E$ and $\widehat{F} = (I + F)^{-1}F$.

Proof. We prove first (3.1). To this purpose, we define $Z := P_{\widetilde{A}^*}(I + F)^{-1}A^\dagger(I + E)^{-1}P_{\widetilde{A}}$ to be the right hand side of (3.1). We will prove that Z satisfies the conditions (2.2) with A replaced by $\widetilde{A}$. Recall that $P_{\widetilde{A}^*} = \widetilde{A}^\dagger\widetilde{A}$ and $P_{\widetilde{A}} = \widetilde{A}\widetilde{A}^\dagger$. Then

$$\begin{aligned}\widetilde{A}Z &= \widetilde{A}(I + F)^{-1}A^\dagger(I + E)^{-1}P_{\widetilde{A}} = (I + E)AA^\dagger(I + E)^{-1}P_{\widetilde{A}} \\ &= (I + E)AA^\dagger(I + E)^{-1}\widetilde{A}\widetilde{A}^\dagger = (I + E)AA^\dagger A(I + F)\widetilde{A}^\dagger = \widetilde{A}\widetilde{A}^\dagger = P_{\widetilde{A}}.\end{aligned}$$

In a similar way,

$$\begin{aligned}Z\widetilde{A} &= P_{\widetilde{A}^*}(I + F)^{-1}A^\dagger(I + E)^{-1}\widetilde{A} = P_{\widetilde{A}^*}(I + F)^{-1}A^\dagger A(I + F) \\ &= \widetilde{A}^\dagger\widetilde{A}(I + F)^{-1}A^\dagger A(I + F) = \widetilde{A}^\dagger(I + E)AA^\dagger A(I + F) = \widetilde{A}^\dagger\widetilde{A} = P_{\widetilde{A}^*}.\end{aligned} \tag{3.3}$$

The equality (3.3) implies $\mathcal{R}(\widetilde{A}^*) \subseteq \mathcal{R}(Z)$ and the definition of Z implies $\mathcal{R}(Z) \subseteq \mathcal{R}(\widetilde{A}^*)$. Thus, $\mathcal{R}(Z) = \mathcal{R}(\widetilde{A}^*)$ and equation (3.3) implies $Z\widetilde{A} = P_Z$. Therefore, conditions (2.2) for $\widetilde{A}$ hold, $Z = \widetilde{A}^\dagger$ and (3.1) is proved.

Next, we use (3.1) to prove (3.2). First, we write $(I+E)^{-1} = I-(I+E)^{-1}E = I-\widehat{E}$ and $(I + F)^{-1} = I - (I + F)^{-1}F = I - \widehat{F}$. Substituting these expressions in (3.1), we get

$$\widetilde{A}^\dagger = P_{\widetilde{A}^*}(I - \widehat{F})A^\dagger(I - \widehat{E})P_{\widetilde{A}} = P_{\widetilde{A}^*}(P_{A^*} - \widehat{F})A^\dagger(P_A - \widehat{E})P_{\widetilde{A}}. \tag{3.4}$$

From Lemma 3.1-(a) it follows that $P_A(I + E^*(I - P_{\widetilde{A}})) = P_AP_{\widetilde{A}}$. Analogously, from Lemma 3.1-(b), $((I - P_{\widetilde{A}^*})F^* + I)P_{A^*} = P_{\widetilde{A}^*}P_{A^*}$. Finally, substitute these relations in (3.4), use $A^\dagger P_A = A^\dagger$ and $P_{A^*}A^\dagger = A^\dagger$, and get (3.2). □

If $m = n$ and A is nonsingular, then $P_{\widetilde{A}} = P_{\widetilde{A}^*} = I_n$, and (3.1) and (3.2) just become $\widetilde{A}^{-1} = (I + F)^{-1}A^{-1}(I + E)^{-1}$. If A has full column rank, then $\widetilde{A}$ has also full column rank, $P_{\widetilde{A}^*} = I_n$, (3.1) simplifies to $\widetilde{A}^\dagger = (I + F)^{-1}A^\dagger(I + E)^{-1}P_{\widetilde{A}}$, and (3.2) to $\widetilde{A}^\dagger = \left(I - \widehat{F}\right) A^\dagger \left(I + E^*(I - P_{\widetilde{A}}) - \widehat{E}P_{\widetilde{A}}\right)$. Finally, if A has full row rank, then $\widetilde{A}$ has also full row rank, $P_{\widetilde{A}} = I_m$, (3.1) simplifies to $\widetilde{A}^\dagger = P_{\widetilde{A}^*}(I + F)^{-1}A^\dagger(I + E)^{-1}$, and (3.2) to $\widetilde{A}^\dagger = \left(I + (I - P_{\widetilde{A}^*})F^* - P_{\widetilde{A}^*}\widehat{F}\right) A^\dagger \left(I - \widehat{E}\right)$.

We emphasize that expression (3.2) ensures that under “small" multiplicative perturbations of A, i.e., small E and F, we obtain “small" multiplicative perturbations of $A^\dagger$.

The assumptions of Theorem 3.2 guarantee that $\operatorname{rank}(A) = \operatorname{rank}(\widetilde{A})$. This has simplified considerably the analysis of the variation of the Moore-Penrose pseudo-inverse with respect general "additive" perturbations $\widetilde{A} = A + \Delta A$ [44, 46]. In addition, Theorem 3.3 implies that if the mild condition $\max\{\|E\|_2, \|F\|_2\} < 1$ holds, then $\widetilde{A} = (I+E)A(I+F)$ is an *acute perturbation* of A (see the original definition in [46, Definition 7.2] and also in [44, Ch. III, Definition 3.2]). It is well known that acute perturbations introduce simplifications even for additive perturbations $\widetilde{A} = A + \Delta A$. The bounds in Theorem 3.3 will be used in Section 4.

THEOREM 3.3. *Let $A \in \mathbb{C}^{m\times n}$ and $\widetilde{A} = (I+E)A(I+F) \in \mathbb{C}^{m\times n}$, where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices, and let $P_{\mathcal{N}(A)}$ and $P_{\mathcal{N}(\widetilde{A})}$ be the orthogonal projectors onto the null spaces of A and $\widetilde{A}$, respectively. Then:*

(a) $\|P_{\widetilde{A}} - P_A\|_2 = \|P_{\widetilde{A}}(I - P_A)\|_2 = \|P_A(I - P_{\widetilde{A}})\|_2 \leq \|E\|_2$.

(b) $\|P_{\widetilde{A}^*} - P_{A^*}\|_2 = \|P_{\widetilde{A}^*}(I - P_{A^*})\|_2 = \|P_{A^*}(I - P_{\widetilde{A}^*})\|_2 \leq \|F\|_2$.

(c) $\|P_{\mathcal{N}(\widetilde{A})} - P_{\mathcal{N}(A)}\|_2 \leq \|F\|_2$.

Proof. Part (a). The subspaces $\mathcal{R}(A)$ and $\mathcal{R}(\widetilde{A})$ have the same dimension. Thus, from [44, Ch. I, Theorem 5.5], $\|P_{\widetilde{A}} - P_A\|_2 = \|P_{\widetilde{A}}(I - P_A)\|_2 = \|P_A(I - P_{\widetilde{A}})\|_2$. Moreover, by Lemma 3.1-(a), $\|P_A(I - P_{\widetilde{A}})\|_2 = \| - P_A E^*(I - P_{\widetilde{A}})\|_2 \leq \|E^*\|_2 = \|E\|_2$.

Part (b) follows from applying part (a) to $\widetilde{A}^* = (I + F^*)A^*(I + E^*)$. Finally, part (c) follows from part (b), $P_{\mathcal{N}(A)} = I - P_{A^*}$, and $P_{\mathcal{N}(\widetilde{A})} = I - P_{\widetilde{A}^*}$. □

Corollary 3.4 presents an expression for $\widetilde{A}^\dagger - A^\dagger$ that follows directly from (3.2). This will be used in Subsection 3.1 and, more important, in Theorem 4.1.

COROLLARY 3.4. *Let $A \in \mathbb{C}^{m\times n}$ and $\widetilde{A} = (I+E)A(I+F) \in \mathbb{C}^{m\times n}$, where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices, and let $\widehat{E} = (I+E)^{-1}E$ and $\widehat{F} = (I+F)^{-1}F$. Then*

$$\widetilde{A}^\dagger - A^\dagger = A^\dagger \Theta_E + \Theta_F A^\dagger + \Theta_F A^\dagger \Theta_E, \tag{3.5}$$

where

$$\Theta_E = E^*(I - P_{\widetilde{A}}) - \widehat{E} P_{\widetilde{A}} \quad \textit{and} \quad \Theta_F = (I - P_{\widetilde{A}^*})F^* - P_{\widetilde{A}^*}\widehat{F}\,. \tag{3.6}$$

3.1. Mutiplicative perturbation bounds for the Moore-Penrose pseudo-inverse.

As explained in the first paragraph of Section 3, the results in this subsection are not used elsewhere in the rest of the paper. The main goal in this subsection is to present bounds for $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|$. We assume $A \neq 0$, since otherwise the problem is trivial. The main result is Theorem 3.5.

THEOREM 3.5. *Let $A \in \mathbb{C}^{m\times n}$ and $\widetilde{A} = (I+E)A(I+F) \in \mathbb{C}^{m\times n}$, where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices, and let $\widehat{E} = (I+E)^{-1}E$ and $\widehat{F} = (I+F)^{-1}F$. Let us denote by $\|\cdot\|$ a normalized unitarily invariant norm and by $\|\cdot\|_2$ the spectral norm. Then the following bounds hold:*

(a) $$\frac{\|\widetilde{A}^\dagger - A^\dagger\|}{\min\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \|E\| + \|\widehat{E}\| + \|F\| + \|\widehat{F}\| + \left(\|E\| + \|\widehat{E}\|\right)\left(\|F\| + \|\widehat{F}\|\right).$$

(b) $$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|A^\dagger\|_2} \leq \sqrt{\|E\|_2^2 + \|F\|_2^2 + \left(\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2\right)^2}, \quad \textit{and}$$

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|\widetilde{A}^\dagger\|_2} \leq \sqrt{\|\widehat{E}\|_2^2 + \|\widehat{F}\|_2^2 + (\|E\|_2 + \|F\|_2 + \|E\|_2\,\|F\|_2)^2}\,.$$

Proof. Part (a). The bound for $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$ follows directly from (3.5), just by taking into account that for any matrices B and C, $\|BC\| \leq \|B\|_2\|C\|$ and $\|BC\| \leq \|B\|\|C\|_2$ [44, p. 80]. The bound for $\|\widetilde{A}^\dagger - A^\dagger\|/\|\widetilde{A}^\dagger\|_2$ follows from the one for $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$ by exchanging the roles of A and $\widetilde{A}$, i.e., by considering A a multiplicative perturbation of $\widetilde{A}$ as $A = (I+E)^{-1}\widetilde{A}(I+F)^{-1} = (I-\widehat{E})\widetilde{A}(I-\widehat{F})$, which amounts to interchanging in the bounds $\|E\|, \|F\|$ by $\|\widehat{E}\|, \|\widehat{F}\|$, respectively, and vice versa.

Part (b). First note that from (3.6), Lemma 3.1-(b), and $P_{A^*} = A^\dagger A$, we get

$$
\begin{aligned}
(I+\Theta_F)A^\dagger A &= (I + (I - P_{\widetilde{A}^*})F^* - P_{\widetilde{A}^*}\widehat{F})P_{A^*} \\
&= P_{A^*} + (I - P_{\widetilde{A}^*})F^* P_{A^*} - P_{\widetilde{A}^*}\widehat{F}P_{A^*} \\
&= P_{A^*} - (I - P_{\widetilde{A}^*})P_{A^*} - P_{\widetilde{A}^*}\widehat{F}P_{A^*} \\
&= P_{\widetilde{A}^*}(I - \widehat{F})P_{A^*}\,,
\end{aligned}
$$

and, if we multiply by $A^\dagger$ on the right the previous equation, then we obtain

$$
(I+\Theta_F)A^\dagger = P_{\widetilde{A}^*}(I-\widehat{F})A^\dagger\,. \tag{3.7}
$$

In a similar way, but using Lemma 3.1-(a), we get

$$
A^\dagger\,(I+\Theta_E) = A^\dagger(I-\widehat{E})P_{\widetilde{A}}\,. \tag{3.8}
$$

Next, we use (3.5), (3.7)-(3.8), and (3.6) to prove

$$
\begin{aligned}
\widetilde{A}^\dagger - A^\dagger &= (I+\Theta_F)A^\dagger\Theta_E + \Theta_F A^\dagger \\
&= P_{\widetilde{A}^*}(I-\widehat{F})A^\dagger\Theta_E + \Theta_F A^\dagger \\
&= P_{\widetilde{A}^*}\left[(I-\widehat{F})A^\dagger\Theta_E - \widehat{F}A^\dagger\right] + (I - P_{\widetilde{A}^*})F^*A^\dagger \\
&= P_{\widetilde{A}^*}\left[A^\dagger\Theta_E - \widehat{F}A^\dagger(I+\Theta_E)\right] + (I - P_{\widetilde{A}^*})F^*A^\dagger \\
&= P_{\widetilde{A}^*}\left[A^\dagger E^*(I - P_{\widetilde{A}}) - A^\dagger\widehat{E}P_{\widetilde{A}} - \widehat{F}A^\dagger(I-\widehat{E})P_{\widetilde{A}}\right] + (I - P_{\widetilde{A}^*})F^*A^\dagger\,. \qquad (3.9)
\end{aligned}
$$

It remains to apply Lemma 2.1 to (3.9) and get

$$
\begin{aligned}
\|\widetilde{A}^\dagger - A^\dagger\|_2^2 &\leq \|A^\dagger E^*(I-P_{\widetilde{A}}) - [A^\dagger\widehat{E} + \widehat{F}A^\dagger(I-\widehat{E})]\,P_{\widetilde{A}}\|_2^2 + \|F^*A^\dagger\|_2^2 \\
&\leq \|A^\dagger E^*\|_2^2 + \|A^\dagger\widehat{E} + \widehat{F}A^\dagger(I-\widehat{E})\|_2^2 + \|F\|_2^2\,\|A^\dagger\|_2^2 \\
&\leq\ \|A^\dagger\|_2^2\left(\,\|E\|_2^2 + \|F\|_2^2 + (\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2)^2\,\right)\,,
\end{aligned}
$$

which gives the bound for $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|A^\dagger\|_2$ in part (b). From here, the bound for $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|\widetilde{A}^\dagger\|_2$ follows by exchanging the roles of A and $\widetilde{A}$ as we did in the proof of part (a). □

REMARK 3.6. We highlight the following points on Theorem 3.5.

(a) The bounds in Theorem 3.5 improve significatively the classical bounds for the relative variation of the Moore-Penrose pseudo-inverse under general additive perturbations $\widetilde{A} = A + \Delta A$ (see [46, Theorem 4.1] or the rearrangement in [44, Ch. III, Corollary 3.10]). The crucial point is that the bounds in Theorem 3.5 do not depend on $\kappa_2(A)$, while the classical bounds do.

(b) The bound in part (a) of Theorem 3.5 has the advantage that is valid for any normalized unitarily invariant norm, but when it is particularized to $\|\cdot\|_2$, then the bound in part (b) is always sharper than the one in part (a), since $\sqrt{x^2+y^2} \leq x+y$, for $x \geq 0$, $y \geq 0$ real numbers, and

$$\sqrt{\|E\|_2^2 + \|F\|_2^2 + \left(\|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2\right)^2} \leq$$

$$\begin{aligned} &\leq \sqrt{\|E\|_2^2 + \|F\|_2^2} \, + \|\widehat{E}\|_2 + \|\widehat{F}\|_2 + \|\widehat{E}\|_2\,\|\widehat{F}\|_2 \\ &\leq \|E\|_2 + \|\widehat{E}\|_2 + \|F\|_2 + \|\widehat{F}\|_2 + (\|E\|_2 + \|\widehat{E}\|_2)\,(\|F\|_2 + \|\widehat{F}\|_2)\,. \end{aligned}$$

(c) If A has full row rank, then $\widetilde{A}$ has also full row rank and $P_{\widetilde{A}} = I_m$. Thus, Θ_E in (3.6) simplifies to $\Theta_E = -\widehat{E}$ and all the terms containing $\|E\|$ or $\|E\|_2$ in the bounds of Theorem 3.5 vanish (but one should keep $\|\widehat{E}\|$ and $\|\widehat{E}\|_2$).

(d) If A has full column rank, then $\widetilde{A}$ has also full column rank and $P_{\widetilde{A}^*} = I_n$. Thus, Θ_F in (3.6) simplifies to $\Theta_F = -\widehat{F}$ and all the terms containing $\|F\|$ or $\|F\|_2$ in the bounds of Theorem 3.5 vanish (but one should keep $\|\widehat{F}\|$ and $\|\widehat{F}\|_2$).

(e) If we restrict in Theorem 3.5 the magnitude of the perturbations to be $\max\{\|E\|_2, \|F\|_2\} < 1$, a condition that in fact guarantees that $(I+E)$ and $(I+F)$ are nonsingular, then standard matrix norm inequalities [24] imply

$$\|\widehat{E}\|_2 \leq \frac{\|E\|_2}{1-\|E\|_2} \quad \text{and} \quad \|\widehat{F}\|_2 \leq \frac{\|F\|_2}{1-\|F\|_2}\,. \tag{3.10}$$

These inequalities can be used in part (b) of Theorem 3.5 to obtain bounds that are easily computable in terms of $\|E\|_2$ and $\|F\|_2$.

(g) Finally, again with the additional restriction $\max\{\|E\|_2, \|F\|_2\} < 1$, Theorem 3.5 can be completed with

$$\frac{\|A^\dagger\|_2}{(1+\|E\|_2)(1+\|F\|_2)} \leq \|\widetilde{A}^\dagger\|_2 \leq \frac{\|A^\dagger\|_2}{(1-\|E\|_2)(1-\|F\|_2)}.$$

The rightmost inequality follows from (3.1), which implies $\|\widetilde{A}^\dagger\|_2 \leq \|(I+F)^{-1}\|_2 \, \|A^\dagger\|_2 \, \|(I+E)^{-1}\|_2$. For the leftmost inequality: consider A as a multiplicative perturbation of $\widetilde{A}$, i.e., $A = (I+E)^{-1}\widetilde{A}(I+F)^{-1}$, and apply (3.1) with the roles of A and $\widetilde{A}$ exchanged to get $A^\dagger = P_{A^*}(I+F)\widetilde{A}^\dagger(I+E)P_A$. This implies $\|A^\dagger\|_2 \leq (1+\|F\|_2)\|\widetilde{A}^\dagger\|_2(1+\|E\|_2)$.

Recently multiplicative perturbation bounds for the Moore-Penrose pseudo-inverse have been presented in [5, Section 4]. The bounds in [5] are not based on expressions for $\widetilde{A}^\dagger$ as those in Theorem 3.2, they are obtained following a different approach. In the rest of this section we compare the bounds in Theorem 3.5 with those in [5]. In the notation of Theorem 3.5, the following two multiplicative bounds[‡] are presented in [5, Theorems 4.1 and 4.2]:

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|}{\max\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \|E\| + \|\widehat{E}\| + \|F\| + \|\widehat{F}\|\,, \tag{3.11}$$

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\max\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}} \leq \sqrt{\frac{3}{2}}\,\sqrt{\|E\|_2^2 + \|\widehat{E}\|_2^2 + \|F\|_2^2 + \|\widehat{F}\|_2^2}\,. \tag{3.12}$$

[‡] The bound (3.12) is presented in [5] for the class of unitarily invariant norms called *Q-norms*, which includes, among others, the spectral and the Frobenius norms.

To begin with note that the presence of $\max\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ makes difficult to compare in general (3.11)-(3.12) with the bounds in Theorem 3.5. For instance, if $\|A^\dagger\|_2 = \|\widetilde{A}^\dagger\|_2$, then the bound in (3.11) is obviously sharper than the one in Theorem 3.5-(a), but if $\|A^\dagger\|_2 \ll \|\widetilde{A}^\dagger\|_2$, then (3.11) does not give any information on $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2$, while Theorem 3.5-(a) does. *However, as we discuss next, the bounds in Theorem 3.5 are superior than* (3.11)-(3.12) *both to first order and in terms of wider applicability.*

If we consider tiny perturbations and neglect second order terms, then we can replace both $\max\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ and $\min\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$, simply by $\|A^\dagger\|_2$, which allows us to make comparisons easier. Theorem 3.5-(a) and (3.11) give both the same bound to first order, that is, $\|\widetilde{A}^\dagger - A^\dagger\|/\|A^\dagger\|_2 \le 2(\|E\| + \|F\|)$. However, to first order, the right-hand side of (3.12) is $\Xi = \sqrt{3}\,\sqrt{\|E\|_2^2 + \|F\|_2^2}$ and the bound in Theorem 3.5-(b) is $\Gamma_b = \sqrt{\|E\|_2^2 + \|F\|_2^2 + (\|E\|_2 + \|F\|_2)^2}$. To compare Γ_b and Ξ, use that $(x+y)^2 \le 2(x^2+y^2)$, for $x \ge 0$, $y \ge 0$ real numbers. Thus

$$\Gamma_b \le \sqrt{3\,\|E\|_2^2 + 3\,\|F\|_2^2} = \Xi,$$

which implies that, to first order, the bound in Theorem 3.5-(b) is always sharper than (3.12).

For sufficiently large perturbations, the presence of $\max\{\|A^\dagger\|_2, \|\widetilde{A}^\dagger\|_2\}$ makes (3.11)-(3.12) unapplicable in certain situations, since one of the standard goals of perturbation theory is to bound $\|\widetilde{A}^\dagger - A^\dagger\|$ *without knowing* $\widetilde{A}^\dagger$ *and having only some bounds on the norms of the perturbations E and F*. Let us illustrate this point with an example. Let $A = [1\ 0; 0\ 1; 0\ 0] \in \mathbb{C}^{3\times 2}$ (we have used MATLAB notation for matrices), $E = \operatorname{diag}(-4/5, -4/5, -4/5)$, and $F = \operatorname{diag}(-4/5, -4/5)$. An easy computation shows that

$$\frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|A^\dagger\|_2} = 24, \quad \frac{\|\widetilde{A}^\dagger - A^\dagger\|_2}{\|\widetilde{A}^\dagger\|_2} = 0.96$$
$$\|E\|_2 = 0.8, \quad \|F\|_2 = 0.8, \quad \|\widehat{E}\|_2 = 4, \quad \|\widehat{F}\|_2 = 4,$$

which give: 9.6 for the bound in (3.11); 32.64 for the bound in Theorem 3.5-(a); 7.07 for the bound in (3.12); 24.03 for the 1st bound in Theorem 3.5-(b); and 6.084 for the 2nd bound in Theorem 3.5-(b). So, in this example, (3.11)-(3.12) fail in getting a bound on $\|\widetilde{A}^\dagger - A^\dagger\|_2/\|A^\dagger\|_2$, while the bounds in Theorem 3.5 give sharp estimates (in particular those in Theorem 3.5-(b)).

4. Multiplicative perturbation results for least squares problems.

In this section we consider the LS problem

$$\min_{x\in\mathbb{C}^n} \|Ax - b\|_2, \quad A \in \mathbb{C}^{m\times n}, \quad b \in \mathbb{C}^m, \tag{4.1}$$

and the multiplicatively perturbed LS problem

$$\min_{x\in\mathbb{C}^n} \|\widetilde{A}x - \widetilde{b}\|_2, \quad \widetilde{A} = (I+E)A(I+F) \in \mathbb{C}^{m\times n}, \quad \widetilde{b} = b + h \in \mathbb{C}^m\,, \tag{4.2}$$

where $(I+E) \in \mathbb{C}^{m\times m}$ and $(I+F) \in \mathbb{C}^{n\times n}$ are nonsingular matrices. We are interested in finding an upper bound for the relative variation $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$, where $x_0 = A^\dagger b$ and $\widetilde{x}_0 = \widetilde{A}^\dagger\widetilde{b}$ are the minimum 2-norm solutions of (4.1) and (4.2), respectively. We will also examine the variation of the associated residuals $r = b - Ax_0$ and $\widetilde{r} = \widetilde{b} - \widetilde{A}\widetilde{x}_0$. Theorem 4.1 is the main result in this section.

THEOREM 4.1. *Let x_0 and $\widetilde{x}_0$ be the minimum 2-norm solutions of (4.1) and (4.2), respectively, and let $r = b - Ax_0$ and $\widetilde{r} = \widetilde{b} - \widetilde{A}\widetilde{x}_0$ be the corresponding residuals. Let*

$\widehat{E} = (I+E)^{-1}E$ and $\widehat{F} = (I+F)^{-1}F$, define $\alpha_E := \sqrt{\|E\|_2^2 + \|\widehat{E}\|_2^2}$ and $\alpha_F := \sqrt{\|F\|_2^2 + \|\widehat{F}\|_2^2}$, and assume that $\|h\|_2 \leq \epsilon \|b\|_2$. Then the following two bounds hold

$$\|\widetilde{x}_0 - x_0\|_2 \leq \alpha_F \, \|x_0\|_2 + \left[\alpha_E \, (1+\alpha_F)\,(1+\epsilon) + \epsilon\,(1+\alpha_F)\,\right] \, \|A^\dagger\|_2 \, \|b\|_2 \, , \tag{4.3}$$

$$\|\widetilde{r} - r\|_2 \leq \|b\|_2 \, \sqrt{(\epsilon + \|E\|_2)^2 + \|E\|_2^2} \, . \tag{4.4}$$

Proof. Let us prove first (4.3). The proof is based on Corollary 3.4 that implies:

$$\begin{aligned}
\widetilde{x}_0 - x_0 &= \widetilde{A}^\dagger (b+h) - A^\dagger b \\
&= \left(\widetilde{A}^\dagger - A^\dagger\right)(b+h) + A^\dagger h \\
&= \left(A^\dagger \, \Theta_E + \Theta_F \, A^\dagger + \Theta_F \, A^\dagger \, \Theta_E\right)(b+h) + A^\dagger h \\
&= \left(A^\dagger \, \Theta_E + \Theta_F \, A^\dagger \, \Theta_E\right)(b+h) + \Theta_F \, x_0 + \Theta_F \, A^\dagger h + A^\dagger h \, .
\end{aligned}$$

Apply norm inequalities and get

$$\|\widetilde{x}_0 - x_0\|_2 \leq \|\Theta_F\|_2 \|x_0\|_2 + \left[\|\Theta_E\|_2 \, (1 + \|\Theta_F\|_2)\,(1+\epsilon) + \epsilon\,(1 + \|\Theta_F\|_2)\right] \, \|A^\dagger\|_2 \, \|b\|_2 \, .$$

Now, (4.3) follows from Lemma 2.1 that implies $\|\Theta_E\|_2 \leq \alpha_E$ and $\|\Theta_F\|_2 \leq \alpha_F$.

Next, we prove (4.4). First, observe that

$$\begin{aligned}
\widetilde{r} - r &= h - \widetilde{A}\,\widetilde{x}_0 + Ax_0 \\
&= h - \widetilde{A}\widetilde{A}^\dagger \, \widetilde{b} + Ax_0 \\
&= (I - \widetilde{A}\widetilde{A}^\dagger)\, h + Ax_0 - \widetilde{A}\widetilde{A}^\dagger \, b \\
&= (I - \widetilde{A}\widetilde{A}^\dagger)\, h + Ax_0 - \widetilde{A}\widetilde{A}^\dagger \, (r + Ax_0) \\
&= (I - \widetilde{A}\widetilde{A}^\dagger)\, (h + Ax_0) - \widetilde{A}\widetilde{A}^\dagger \, r \, . \qquad (4.5)
\end{aligned}$$

Note that the summands in (4.5) are orthogonal vectors, since $P_{\widetilde{A}} = \widetilde{A}\widetilde{A}^\dagger$, use $Ax_0 = P_A b$ and $r = (I - P_A)b$, recall Theorem 3.3, and get (4.4) as follows

$$\begin{aligned}
\|\widetilde{r} - r\|_2^2 &= \|(I - P_{\widetilde{A}})\,(h + P_A b)\|_2^2 + \|P_{\widetilde{A}}\,(I - P_A)b\|_2^2 \\
&\leq \left(\|h\|_2 + \|(I - P_{\widetilde{A}})P_A b\|_2\right)^2 + \|P_{\widetilde{A}}\,(I - P_A)b\|_2^2 \\
&\leq \left(\epsilon\|b\|_2 + \|E\|_2\|b\|_2\right)^2 + \|E\|_2^2 \, \|b\|_2^2 \, .
\end{aligned}$$

□

The bound (4.3) simplifies if A has full row or full column rank in the way explained in parts (d) and (e) of Remark 3.6. If A has full row rank, then $\widetilde{r} = r = 0$ and $\|\widetilde{r} - r\|_2 = 0$.

The bounds in Theorem 4.1 improve significatively the classical bounds for the relative variation of minimum 2-norm solutions and residuals of LS problems under general additive perturbations $\widetilde{A} = A + \Delta A$ [46, Theorem 5.1]. For the purpose of comparison, let us state these classical perturbation bounds as they are stated in [3, Theorem 1.4.6] in the case $\operatorname{rank}(A) = \operatorname{rank}(\widetilde{A})$ and $\eta := \kappa_2(A)\,\|\Delta A\|_2/\|A\|_2 < 1$. Let $\widetilde{x}_0$ be the minimum 2-norm solution of the LS problem $\min_{x\in\mathbb{C}^n} \|(b + \Delta b) - (A + \Delta A)x\|_2$ and x_0 be the minimum 2-norm solution of $\min_{x\in\mathbb{C}^n} \|b - Ax\|_2$, and let $\widetilde{r} := (b + \Delta b) - (A + \Delta A)\widetilde{x}_0$ and $r := b - Ax_0$.

Then, assuming $x_0 \neq 0$ and defining $\epsilon_A := \|\Delta A\|_2 / \|A\|_2$ and $\epsilon_b := \|\Delta b\|_2 / \|b\|_2$, a minor variation of [3, Theorem 1.4.6] states that

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \frac{1}{1-\eta} \left(2\,\kappa_2(A)\,\epsilon_A + \frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2}\,\epsilon_b + \kappa_2(A)^2\,\frac{\|r\|_2}{\|A\|_2\,\|x_0\|_2}\,\epsilon_A \right), \tag{4.6}$$

$$\frac{\|\widetilde{r} - r\|_2}{\|b\|_2} \leq \left(\frac{\|A\|_2 \|x_0\|_2}{\|b\|_2}\,\epsilon_A + \epsilon_b + \kappa_2(A)\,\frac{\|r\|_2}{\|b\|_2}\,\epsilon_A \right). \tag{4.7}$$

In (4.7), it is convenient to bear in mind that $(\|A\|_2 \|x_0\|_2)/\|b\|_2 \leq \kappa_2(A)$ and $\|r\|_2 \leq \|b\|_2$. Next, observe that the bound for $\|\widetilde{r} - r\|_2 / \|b\|_2$ in (4.7) includes terms that can be very large even if ϵ_A and ϵ_b are very tiny. This happens if $\kappa_2(A)$ is large and $\|r\|_2 \neq 0$ is not very small. In contrast, if $\|E\|_2$ and ϵ are tiny, then the bound for $\|\widetilde{r} - r\|_2 / \|b\|_2$ that follows from (4.4) is always tiny. With respect to the bounds for $\|\widetilde{x}_0 - x_0\|_2 / \|x_0\|_2$: the bound in (4.6) amplifies the perturbations in the data at least by a factor $\kappa_2(A)$ and the amplification can be much larger under certain conditions. In addition, (4.6) includes the amplification factor $\|A^\dagger\|_2\,\|b\|_2 / \|x_0\|_2$, which is the only potentially large factor in the bound that follows from (4.3). We will show in Subsection 4.1 that $\|A^\dagger\|_2\,\|b\|_2 / \|x_0\|_2$ is a moderate number except for very particular choices of b. Therefore, (4.3) always improves (4.6) and, if $\|E\|_2$, $\|F\|_2$, and ϵ are tiny, then (4.3) produces tiny bounds for $\|\widetilde{x}_0 - x_0\|_2 / \|x_0\|_2$ for almost all b.

The bounds in Theorem 4.1 cannot be directly applied due to the presence of $\widehat{E}$ and $\widehat{F}$. Corollary 4.2 overcomes this shortcoming by restricting the magnitude of the perturbations and by using (3.10). Corollary 4.2 follows directly from Theorem 4.1 and is stated in a way that is convenient for its use in Section 5.

COROLLARY 4.2. *With the same notation and hypotheses that in Theorem 4.1, assume in addition that* $\|E\|_2 \leq \mu < 1$ *and* $\|F\|_2 \leq \nu < 1$, $x_0 \neq 0$, *and* $b \neq 0$. *Define*

$$\theta_\mu := \mu \sqrt{1 + \frac{1}{(1-\mu)^2}} \quad \textit{and} \quad \theta_\nu := \nu \sqrt{1 + \frac{1}{(1-\nu)^2}}\,. \tag{4.8}$$

Then the following bounds hold:

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \theta_\nu + [\theta_\mu (1+\theta_\nu)(1+\epsilon) + \epsilon(1+\theta_\nu)]\,\frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2}\,, \tag{4.9}$$

$$\frac{\|\widetilde{r} - r\|_2}{\|b\|_2} \leq \sqrt{(\epsilon+\mu)^2 + \mu^2}\,. \tag{4.10}$$

The bound (4.9) *yields to first order in* ϵ, μ, ν

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \sqrt{2}\,\nu + \left(\epsilon + \sqrt{2}\,\mu\right) \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} + h.o.t\,, \tag{4.11}$$

where $h.o.t$ *stands for "higher order terms" in* ϵ, μ, ν.

We will prove in Subsection 4.2 that, to first order, the perturbation bound for $\|\widetilde{x}_0 - x_0\|_2 / \|x_0\|_2$ that follows from Theorem 4.1 is optimal, i.e., it can be attained modulo a moderate constant. In this context, it is interesting to observe that another approach to get multiplicative perturbation bounds for LS problems is via Theorem 3.5 as follows: we know that $\widetilde{x}_0 - x_0 = \left(\widetilde{A}^\dagger - A^\dagger\right)(b+h) + A^\dagger h$, thus $\|\widetilde{x}_0 - x_0\|_2 \leq \|\widetilde{A}^\dagger - A^\dagger\|_2 (\|b\|_2 + \|h\|_2) + \|A^\dagger\|_2 \|h\|_2$, and this can be combined with Theorem 3.5-(b) and $\|h\|_2 \leq \epsilon \|b\|_2$ to get a bound for $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2$. The bound so obtained (let us call it Γ_{LP}) includes in all its terms the factor $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$. This is in contrast with the bound coming from (4.3) (let us call it Γ_{LS}) which has a term that is simply α_F (see also the terms θ_ν in (4.9) or $\sqrt{2}\,\nu$ in (4.11)). Thus, Γ_{LP} is much larger than Γ_{LS} if $\max\{\|E\|_2, \epsilon\} \ll \|F\|_2$ and $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ is large (this is easily checked to first order). In addition, it may be proved to first order that $\Gamma_{LS} \leq 2\,\Gamma_{LP}$ always, and $\Gamma_{LS} \leq \Gamma_{LP}$ if $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \geq 2$. So, we can conclude that the approach in Theorem 4.1 is better than the use of Theorem 3.5.

Finally, observe that all the results in this section, as well as those in Section 3, are valid for any values of m and n, that is, both if $m \geq n$ or if $m < n$. Thus, they are valid also for multiplicative perturbations of solutions of underdetermined linear systems.

4.1. Why is the factor $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ usually small?. This section is related to [18, Section 3.2], which considered the same problem for a nonsingular matrix A. Although the fact that $A \in \mathbb{C}^{m\times n}$ is rectangular forces nontrivial modifications, the main conclusions remain the same. Theorem 4.1 and Corollary 4.2 prove that the sensitivity of the minimum 2-norm solution $x_0 = A^\dagger b$ of a LS problem under multiplicative perturbations is governed by $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$. This quantity is well known because it is the condition number for LS problems when only the right-hand side b of $\min_{x\in\mathbb{C}^n} \|Ax - b\|_2$ is perturbed. More precisely, it is easy to prove that if $x_0 = A^\dagger b$, then

$$\frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} = \lim_{\epsilon\to 0} \sup \left\{ \frac{\|\widetilde{x}_0 - x_0\|_2}{\epsilon \|x_0\|_2} \, : \, \widetilde{x}_0 = A^\dagger \, (b+h), \quad \|h\|_2 \leq \epsilon \, \|b\|_2 \right\} .$$

Thus Theorem 4.1 essentially proves that multiplicative perturbations have an effect on the minimum 2-norm solution of LS problems similar to perturbing only the right-hand side b.

Note that $1 \leq \|A^\dagger\|_2\|b\|_2/\|x_0\|_2$, but, in general, $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \not\lesssim \kappa_2(A)$, in contrast to the case when A is nonsingular[§] [18, Section 3.2]. Nevertheless, (4.6) shows that

$$\kappa_{LS}(A,b) := \left(2\,\kappa_2(A) + \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} + \kappa_2(A)^2 \, \frac{\|r\|_2}{\|A\|_2 \, \|x_0\|_2} \right)$$

can be considered as a condition number for LS problems under additive tiny normwise perturbations of A and b (in fact, it is proved in [46, Section 6] that the bound (4.6) is approximately attained to first order in the perturbations), and $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \leq \kappa_{LS}(A,b)$. But the first key point in this section is to show that *if A is fixed*, then $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ is a moderate number *for most vectors* b, even if $\kappa_2(A) \gg 1$, and so $\kappa_{LS}(A,b) \gg 1$, which implies that $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \ll \kappa_{LS}(A,b)$ for most ill-conditioned LS problems whose coefficient matrix is A. However, this is not enough for our purposes, because if $\operatorname{rank}(A) < m$, then for most vectors b the acute angle $\theta(b, \mathcal{R}(A))$ between b and the column space of A is not small, which is equivalent to say that the relative residual $\|Ax_0 - b\|_2/\|b\|_2 = \sin\theta(b, \mathcal{R}(A))$ is not small. But, very often in practice LS problems have small relative residuals, since the problems correspond to inconsistent linear systems $Ax \approx b$ that are close to be consistent. Therefore, the second key point in this section is *if A is fixed* to consider all vectors b such that $\Upsilon = \theta(b, \mathcal{R}(A)) < \pi/2$ *is also fixed*, and then to show that for most *of these vectors* b the factor $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ is a moderate number much smaller than $\kappa_{LS}(A,b)$ whenever A is very ill-conditioned.

To explain the properties mentioned above, assume $\operatorname{rank}(A) = r$ and let $A = U\Sigma V^*$ be the SVD of A, where $U \in \mathbb{C}^{m\times r}$ and $V \in \mathbb{C}^{n\times r}$ have orthonormal columns, $\Sigma =$

[§] If A is nonsingular or, more general, if $Ax_0 = b$, then $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \leq \kappa_2(A)$. However, consider $A = [1 \ 0; 0 \ 1; 0 \ 0] \in \mathbb{C}^{3\times 2}$ and $b = [\eta; 0; 1] \in \mathbb{C}^{3\times 1}$. In this case $x_0 = [\eta; 0] \in \mathbb{C}^{2\times 1}$ and $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 = \sqrt{|\eta|^2 + 1}/|\eta|$ which tends to ∞ if $\eta \to 0$ while $\kappa_2(A) = 1$.

$\mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{C}^{r\times r}$, and $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Observe that $\|x_0\|_2 = \|A^\dagger b\|_2 = \|\Sigma^{-1}U^* b\|_2 \geq |u_r^* b|/\sigma_r$ and

$$\text{(4.12)} \qquad \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} = \frac{\|b\|_2}{\sigma_r \|x_0\|_2} \leq \frac{\|b\|_2}{|u_r^* b|} = \frac{1}{\cos\,\theta(u_r, b)},$$

where u_r is the last column of U and $\theta(u_r, b)$ is the acute angle between u_r and b. Note that the bound on $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ in (4.12) may be large only if b is "almost" orthogonal to u_r. For example, if A is an extremely ill conditioned fixed matrix (think that $\kappa_2(A) = 10^{1000}$ to be concrete) and b is considered as a random vector whose direction is uniformly distributed in the whole space, then the probability that $0 \leq \theta(u_r, b) \leq \pi/2 - 10^{-6}$ is approximately $1 - 10^{-6}$. Note that the condition $0 \leq \theta(u_r, b) \leq \pi/2 - 10^{-6}$ implies $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 \lesssim 10^6$, which is a moderate number compared to 10^{1000}. In particular, if the perturbation parameters μ, ν, and ϵ in Corollary 4.2 are 10^{-16}, then $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2 \lesssim 10^6$ provides a very good bound for the variation of the minimum 2-norm solution of the LS problem. Even more, it is possible that $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2$ is moderate although $\cos\,\theta(u_r, b) \approx 0$. This can be seen by extending from nonsingular to general matrices the original result by Chan and Foulser in [7, Theorem 1]. We do not present this easy generalization here and refer the reader to the discussion in [18, Section 3.2].

In the argument above, the random vector b may be everywhere in the space. Next, we consider vectors b such that $\Upsilon = \theta(b, \mathcal{R}(A)) < \pi/2$ is kept constant. Let us describe all these vectors as follows: let $y \in \mathbb{C}^r$ be any vector and let $U_\perp \in \mathbb{C}^{m\times(m-r)}$ be such that $[\,U \;\, U_\perp\,] \in \mathbb{C}^{m\times m}$ is unitary. Then chose any $z \in \mathbb{C}^{m-r}$ such that $\|z\|_2 = \|y\|_2 \tan \Upsilon$, and define $b = Uy + U_\perp z$. It is obvious that $\Upsilon = \theta(b, \mathcal{R}(A))$, because $\mathcal{R}(U) = \mathcal{R}(A)$. In addition, from (4.12), it can be easily proved that these vectors b satisfy

$$\text{(4.13)} \qquad \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2} = \frac{\|b\|_2}{\sigma_r \|x_0\|_2} \leq \frac{\|b\|_2}{|u_r^* b|} = \frac{\sqrt{1+\tan^2\Upsilon}}{\cos\,\theta(e_r, y)} = \frac{1}{(\cos\,\Upsilon)\cdot(\cos\,\theta(e_r, y)\,)},$$

where e_r is the rth column of I_r. The bound in (4.13) is a "geometrical" quantity that does not depend on $\kappa_2(A)$ and that, assuming that Υ is not very close to $\pi/2$, is a moderate number for most vectors y, i.e., for most vectors[¶] b such that $\Upsilon = \theta(b, \mathcal{R}(A))$.

Finally, we discuss an interesting relationship of the factor $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2$ with the term of $\kappa_{LS}(A, b)$ that depends on $\kappa_2(A)^2$. Note that this term can be upper bounded as follows

$$\text{(4.14)} \qquad \Phi := \kappa_2(A)^2 \, \frac{\|r\|_2}{\|A\|_2 \, \|x_0\|_2} \; = \kappa_2(A) \, \frac{\|A^\dagger\|_2 \, \|r\|_2}{\|x_0\|_2} \; \leq \kappa_2(A) \, \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2}\,.$$

According to our discussion in this subsection $\|A^\dagger\|_2 \|b\|_2/\|x_0\|_2$ is a moderate number for most vectors b. Therefore, Φ is upper bounded by a moderate number times $\kappa_2(A)$ for most vectors b and, as a consequence, $\kappa_2(A)^2$ only affects the sensitivity of LS problems in very particular situations. In addition, Φ can be written as follows

$$\text{(4.15)} \qquad \left(\kappa_2(A) \, \frac{\|A^\dagger\|_2 \, \|b\|_2}{\|x_0\|_2}\right) \frac{\|r\|_2}{\|b\|_2} \; = \Phi\,,$$

[¶] It seems possible to give a rigorous probabilistic meaning to the loose sentences *"for most vectors b"* that we have used in this section and throughout the paper. A possible strategy would be to consider random vectors whose entries follow uniform distributions in symmetric intervals and to develop results in the spirit of those in [42, Section 3]. This would likely lead to some extra $\sqrt{m}$ factor in the bounds. Due to the length of the paper, we postpone the investigation of these topics.

which implies that for large enough relative residuals (think, for instance, in $\|r\|_2/\|b\|_2 \geq 10^{-3}$) and very ill conditioned matrices A, we have $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \ll \Phi \leq \kappa_{LS}(A,b)$, even if $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ is large.

4.2. The condition number under multiplicative perturbations of LS problems. In this subsection we prove that $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ is essentially, i.e., up to a moderate constant, the condition number under multiplicative perturbations of LS problems. The reader should notice that, for simplicity, we consider in our definition of condition number that the left and right mulplicative perturbations and the relative variation of b all have the same order.

THEOREM 4.3. *Let us use the same notation and assumptions as in Corollary 4.2 with the parameters μ, ν, and ϵ set equal to η, and let us define the condition number*

$$\kappa_{LS}^{(M)}(A,b) := \lim_{\eta\to 0}\sup\left\{ \frac{\|\widetilde{x}_0 - x_0\|_2}{\eta\,\|x_0\|_2} \;:\; \widetilde{x}_0 = [(I+E)A(I+F)]^\dagger\,(b+h), \right.$$
$$\left. \|E\|_2 \leq \eta,\ \|F\|_2 \leq \eta,\ \|h\|_2 \leq \eta\|b\|_2 \right\}.$$

Then

$$\frac{1}{1+2\sqrt{2}}\,\kappa_{LS}^{(M)}(A,b) \leq \frac{\|A^\dagger\|_2\,\|b\|_2}{\|x_0\|_2} \leq \kappa_{LS}^{(M)}(A,b)\ . \tag{4.16}$$

Proof. From (4.11) and $1 \leq \|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$, we get

$$\frac{\|\widetilde{x}_0 - x_0\|_2}{\eta\,\|x_0\|_2} \leq \sqrt{2} + \left(1+\sqrt{2}\right)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2} + O(\eta) \leq \left(1+2\sqrt{2}\right)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2} + O(\eta)\,,$$

which implies the left inequality in (4.16). To prove the right inequality choose a perturbation such that $E=0$, $F=0$, and $h=\eta w$, where $\|w\|_2 = \|b\|_2$ and $\|A^\dagger w\|_2 = \|A^\dagger\|_2\|w\|_2$. For this perturbation $\|\widetilde{x}_0 - x_0\|_2/\|x_0\|_2 = \|A^\dagger h\|_2/\|x_0\|_2 = \eta\,\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$. So, the "sup" appearing in the definition of $\kappa_{LS}^{(M)}(A,b)$ implies $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2 \leq \kappa_{LS}^{(M)}(A,b)$. □

4.3. Multiplicative perturbation bounds for other solutions of LS problems. Bounds for the variation of solutions different from the minimum 2-norm solution are easily obtained from Theorem 4.1 and Theorem 3.3-(c) and are a minor modification of (4.3). Since the residual of a LS problem is the same for all its solutions, it is not needed to consider again perturbation bounds for the residuals.

THEOREM 4.4. *If $y \in \mathbb{C}^n$ is a solution of the LS problem* (4.1)*, then there exists a solution $\widetilde{y} \in \mathbb{C}^n$ of the LS problem* (4.2) *such that*

$$\|\widetilde{y} - y\|_2 \leq\ (\alpha_F + \|F\|_2)\,\|y\|_2 + \left[\alpha_E\,(1+\alpha_F)\,(1+\epsilon) + \epsilon\,(1+\alpha_F)\right]\ \|A^\dagger\|_2\,\|b\|_2\,,$$

where α_E, α_F, and ϵ are defined as in the statement of Theorem 4.1.

Proof. Given y, there exists a vector $z \in \mathbb{C}^n$ such that $y = x_0 + P_{\mathcal{N}(A)}z$, where x_0 is the minimum 2-norm solution of (4.1). Recall also that $\|y\|_2^2 = \|x_0\|_2^2 + \|P_{\mathcal{N}(A)}z\|_2^2$ and, so, $\|P_{\mathcal{N}(A)}z\|_2 \leq \|y\|_2$. Let us choose the following solution of (4.2), $\widetilde{y} = \widetilde{x}_0 + P_{\mathcal{N}(\widetilde{A})}P_{\mathcal{N}(A)}z$, where $\widetilde{x}_0$ is the minimum 2-norm solution of (4.2). Therefore

$$\|\widetilde{y} - y\|_2 \leq \|\widetilde{x}_0 - x_0\|_2 + \|(P_{\mathcal{N}(\widetilde{A})} - P_{\mathcal{N}(A)})P_{\mathcal{N}(A)}z\|_2 \leq \|\widetilde{x}_0 - x_0\|_2 + \|F\|_2\|y\|_2\,,$$

where we have used Theorem 3.3-(c). Now, use (4.3) and $\|x_0\|_2 \leq \|y\|_2$ and get the result. □

Note that the relative variation $\|\widetilde{y}-y\|_2/\|y\|_2$ is governed by $\max\{1, \|A^\dagger\|_2\|b\|_2/\|y\|_2\}$, which is smaller than or equal to $\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$. Therefore, the minimum 2-norm solution is the most sensitive of the solutions under multiplicative perturbations.

5. Perturbation of least squares problems through factors. As explained in the Introduction, we present in Section 6 an accurate algorithm, Algorithm 6.1, for the solution of LS problems $\min_{x\in\mathbb{C}^n}\|b-Ax\|_2$ that makes use of an accurate RRD XDY of A. The error analysis of Algorithm 6.1 is presented in Theorem 6.2 and it shows that the computed solution is the exact solution of a LS problem corresponding to an RRD with nearby factors $(X+\Delta X)(D+\Delta D)(Y+\Delta Y)$, where the perturbations are normwise for the well conditioned factors X and Y, and componentwise for the diagonal and potentially ill-conditioned factor D. Therefore, we need to develop perturbation bounds for the solution of LS problems whose coefficient matrix is given as an RRD under perturbations of the factors. This is done in Theorem 5.1, whose proof relies on the key idea of writing the perturbations in the factors as a multiplicative perturbation of the whole matrix.

Let us recall that, according to Lemma 2.2-(c), if $A = XDY$ is an RRD of A, then $A^\dagger = Y^\dagger D^{-1}X^\dagger$. Consequently, the minimum 2-norm solution of the LS problem $\min_{x\in\mathbb{C}^n}\|b-XDYx\|_2$ is $x_0 = Y^\dagger D^{-1}X^\dagger b$.

THEOREM 5.1. *Let $X\in\mathbb{C}^{m\times r}$, $D\in\mathbb{C}^{r\times r}$, and $Y\in\mathbb{C}^{r\times n}$ be such that $\operatorname{rank}(X) = \operatorname{rank}(Y) = r$ and D is diagonal and nonsingular, and let $b\in\mathbb{C}^m$. Let x_0 be the minimum 2-norm solution of $\min_{x\in\mathbb{C}^n}\|b-XDY\,x\|_2$, and $\widetilde{x}_0$ be the minimum 2-norm solution of $\min_{x\in\mathbb{C}^n}\|(b+h)-(X+\delta X)(D+\delta D)(Y+\delta Y)\,x\|_2$, where $\|\delta X\|_2\le\alpha\|X\|_2$, $\|\delta Y\|_2\le\beta\|Y\|_2$, $|\delta D|\le\rho|D|$, and $\|h\|_2\le\epsilon\|b\|_2$. Let $r = b - XDY\,x_0$ and $\widetilde{r} = (b+h)-(X+\delta X)(D+\delta D)(Y+\delta Y)\,\widetilde{x}_0$. Assume that*

$$\mu := \alpha\,\kappa_2(X) < 1 \quad \textit{and} \quad \nu := [\beta+\rho(1+\beta)]\kappa_2(Y) < 1, \tag{5.1}$$

and define for these parameters θ_μ and θ_ν as in (4.8)*. Then, the bound* (4.9) *holds with $A^\dagger$ replaced by $Y^\dagger D^{-1}X^\dagger$, the bound* (4.10) *holds, and to first order in α, β, ρ, and ϵ*

$$\frac{\|\widetilde{x}_0-x_0\|_2}{\|x_0\|_2}\le\sqrt{2}\,(\beta+\rho)\,\kappa_2(Y)+\left(\epsilon+\sqrt{2}\,\alpha\,\kappa_2(X)\right)\frac{\|Y^\dagger D^{-1}X^\dagger\|_2\|b\|_2}{\|x_0\|_2}+h.o.t. \tag{5.2}$$

Proof. Let us call $A = XDY$ and $\widetilde{A} = (X+\delta X)(D+\delta D)(Y+\delta Y)$. Let us write $\widetilde{A}$ as a multiplicative perturbation of A as follows

$$\begin{aligned}\widetilde{A} &= (I+\delta X X^\dagger)\,XD\,(I+D^{-1}\,\delta D)Y(I+Y^\dagger\delta Y)\\ &= (I+\delta X X^\dagger)\,XDY\,(I+Y^\dagger\,D^{-1}\,\delta D\,Y)\,(I+Y^\dagger\delta Y)\\ &=: (I+E)A(I+F),\end{aligned}$$

where $E = \delta X X^\dagger$ and $F = Y^\dagger\delta Y + Y^\dagger D^{-1}\,\delta D\,Y + Y^\dagger D^{-1}\,\delta D\,\delta Y$. Next, taking into account that $\|\delta D\,D^{-1}\|_2\le\rho$, we get

$$\|E\|_2\le\alpha\,\kappa_2(X)=\mu<1, \quad \|F\|_2\le[\beta+\rho(1+\beta)]\,\kappa_2(Y)=\nu<1,$$

and Theorem 5.1 follows immediately from Corollary 4.2. □

Since the factors X and Y of an RRD are well conditioned, we see from (5.2) that the sensitivity with respect perturbations of the factors of the minimum 2-norm solution of the LS problem $\min_{x\in\mathbb{C}^n}\|b-XDYx\|_2$ is again controlled by $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$, where $A = XDY$, which is a moderate number for most b (see Subsection 4.1). Note that Theorem 5.1 is valid both if $m\ge n$ or if $m<n$, i.e., for LS problems or for undertermined linear systems, since Corollary 4.2 is valid in both cases.

6. Algorithm and error analysis. We present in this section Algorithm 6.1 for solving a LS problem $\min_{x\in\mathbb{C}^n} \|b - A\,x\|_2$ and we prove that it computes the minimum 2-norm solution with relative error bounded by $O(\mathtt{u})\,\|A^\dagger\|_2\,\|b\|_2/\|x_0\|_2$, which is simply $O(\mathtt{u})$ for most vectors b according to the discussion in Subsection 4.1. The first step of the algorithm computes an accurate RRD of $A = XDY \in \mathbb{C}^{m\times n}$ in the sense of Definition 2.3, something that is possible for many classes of structured matrices as we have discussed in the Introduction. Next steps of Algorithm 6.1 are based on the fact that according to Lemma 2.2-(c) the minimum 2-norm solution is $x_0 = Y^\dagger(D^{-1}(X^\dagger b))$ and the following observations: (1) $x_1 = X^\dagger b$ is the unique solution of the *full column rank* LS problem $\min_{x\in\mathbb{C}^r} \|b - X\,x\|_2$; (2) $x_2 = D^{-1}(X^\dagger b)$ is the unique solution of the linear system $Dx = x_1$; and (3) $Y^\dagger(D^{-1}(X^\dagger b))$ is the minimum 2-norm solution of the *full row rank* underdetermined linear system $Yx = x_2$. Observe that this procedure is valid both if $m \geq n$ and if $m < n$. Therefore, in the latter case and if $\operatorname{rank}(A) = m$, the procedure solves accurately the underdetermined linear system $Ax = b$.

The minimum 2-norm solution x_0 of the underdetermined system $Yx = x_2$ is computed via the Q-method described in [28, Sec. 21.1] and that we recall here briefly. The idea is to compute first via the Householder QR algorithm a *thin* QR factorization of $Y^* = WR_Y \in \mathbb{C}^{n\times r}$, where $W \in \mathbb{C}^{n\times r}$ satisfies $W^*W = I_r$ and $R_Y \in \mathbb{C}^{r\times r}$ is upper triangular and nonsingular. Thus $Y = R_Y^* W^* \in \mathbb{C}^{r\times n}$, and Lemma 2.2-(c) implies $Y^\dagger = (W^*)^\dagger\,(R_Y^*)^\dagger = W\,R_Y^{-*}$, where R_Y^{-*} denotes the inverse of R_Y^*. Finally, $x_0 = W\,(R_Y^{-*}x_2)$ and $R_Y^{-*}x_2$ is computed by solving the triangular system $R_Y^* x = x_2$ by forward substitution. In practice, it is important to note that the factor W is not required explicitly, we just need the ability of multiplying W times a vector, and this can be done by multiplying the sequence of $n \times n$ Householder reflectors involved in the QR factorization of Y^* times $[(R_Y^{-*}x_2)^*, 0]^* \in \mathbb{C}^n$.

We are now in position of stating Algorithm 6.1.

ALGORITHM 6.1. (Accurate solution of LS problems via RRD)

```
Input:   $A \in \mathbb{C}^{m\times n}, b \in \mathbb{C}^m$
Output:  $x_0$ minimum 2-norm solution of $\min_{x\in\mathbb{C}^n} \|b - A\,x\|_2$

Step 1:  Compute an accurate RRD of $A = XDY$ in the sense of Definition 2.3,
         where $X \in \mathbb{C}^{m\times r}$, $D \in \mathbb{C}^{r\times r}$ is diagonal, $Y \in \mathbb{C}^{r\times n}$, and
         $\operatorname{rank}(A) = \operatorname{rank}(X) = \operatorname{rank}(Y) = \operatorname{rank}(D) = r$.

Step 2:  Compute the unique solution $x_1$ of $\min_{x\in\mathbb{C}^r} \|b - X\,x\|_2$ using the
         Householder $QR$ factorization of $X$.

Step 3:  Compute the unique solution $x_2$ of the diagonal linear system $D\,x = x_1$ as
         $x_2(i) = x_1(i)/d_{ii}$, $i = 1, \ldots, r$.

Step 4:  Compute the minimum 2-norm solution $x_0$ of $Yx = x_2$ using the $Q$ method,
         i.e., via Householder $QR$ factorization of $Y^*$.
```

The computational cost of `Step 1` of Algorithm 6.1 depends on the specific type of matrices and on the specific algorithm used among those mentioned in the Introduction. Anyway all these algorithms cost $O(mn^2)$ flops if $m \geq n$ and $O(m^2 n)$ flops if $m < n$. The leading terms of the costs of `Steps 2, 3,` and `4` are $2r^2(m - r/3)$, r, and $2r^2(n - r/3)$ flops, respectively. Since $r \leq \min\{m, n\}$, the total cost of Algorithm 6.1 is $O(mn^2)$ flops if $m \geq n$ and $O(m^2 n)$ flops if $m < n$.

The backward rounding errors committed by Algorithm 6.1 are analyzed in Theorem 6.2.

We will use the following notation introduced in [28, Secs. 3.1 and 3.4]

$$\gamma_n := \frac{n\mathtt{u}}{1-n\mathtt{u}} \quad \text{and} \quad \widetilde{\gamma}_n := \frac{cn\mathtt{u}}{1-cn\mathtt{u}}, \tag{6.1}$$

where c denotes a small integer constant whose exact value is not essential in the analysis. Before stating and proving Theorem 6.2, let us comment the need and the meaning of the assumptions used in the theorem. First, we assume that the factors $\widehat{X}$, $\widehat{D}$, $\widehat{Y}$ computed in `Step 1` in floating point arithmetic satisfy (2.3), (2.4), and (2.5), which imply $\operatorname{rank}(X) = \operatorname{rank}(\widehat{X}) = r$, $\operatorname{rank}(D) = \operatorname{rank}(\widehat{D}) = r$, $\operatorname{rank}(Y) = \operatorname{rank}(\widehat{Y}) = r$, and (2.6). Therefore, we can use $\kappa_2(X)$ and $\kappa_2(Y)$ in the errors of `Steps 2` and `4` instead of $\kappa_2(\widehat{X})$ and $\kappa_2(\widehat{Y})$ at the cost of not paying attention to the exact values of the numerical constants in the error bounds. The assumption $\max\{\kappa_2(X), \kappa_2(Y)\}\,\sqrt{r}\,\widetilde{\gamma}_{r\max\{m,n\}} < 1$ guarantees that the backward errors $\Delta\widehat{X}$ on $\widehat{X}$ in `Step 2` preserve the full rank, i.e., $\operatorname{rank}(\widehat{X}) = \operatorname{rank}(\widehat{X} + \Delta\widehat{X}) = r$, and the same for the backward errors on $\widehat{Y}$ in `Step 4`. Finally, the technical assumption $\kappa_2(Y)\, n\, r^2\, \widetilde{\gamma}_n < 1$ is needed for applying [28, Theorem 21.4] in the error analysis of `Step 4`.

We present in Theorem 6.2 two statements for the backward errors of Algorithm 6.1, one with respect the computed factors $\widehat{X}$, $\widehat{D}$, and $\widehat{Y}$ of A and another with respect the exact ones, which is the result to be used in practice. The reason for presenting these two statements is that the former gives stronger column-wise and row-wise backward errors in $\widehat{X}$ and $\widehat{Y}$, respectively, than the latter. This may be used to give stronger final backward errors for some particular classes of matrices, as Cauchy matrices. We do not follow this line here.

THEOREM 6.2. *Let $\widehat{X} \in \mathbb{C}^{m\times r}$, $\widehat{D} \in \mathbb{C}^{r\times r}$, and $\widehat{Y} \in \mathbb{C}^{r\times n}$ be the factors of A computed in* `Step 1` *of Algorithm 6.1 and assume that they satisfy the error bounds* (2.3) *and* (2.4) *with respect to the exact factors X, D, and Y of A. Assume also that* (2.5)*,*

$$\max\{\kappa_2(X), \kappa_2(Y)\}\,\sqrt{r}\,\widetilde{\gamma}_{r\max\{m,n\}} < 1, \qquad \textit{and} \tag{6.2}$$

$$\kappa_2(Y)\, n\, r^2\, \widetilde{\gamma}_n < 1 \tag{6.3}$$

hold. Let $\widehat{x}_0$ be the computed minimum 2-norm solution of $\min_{x\in\mathbb{C}^n} \|b - A\,x\|_2$ using Algorithm 6.1 in finite precision with unit roundoff $\mathtt{u}$. Then the following statements hold.

(a) *$\widehat{x}_0$ is the exact minimum 2-norm solution of*

$$\min_{x\in\mathbb{C}^n} \|(b+\Delta b) - (\widehat{X}+\Delta\widehat{X})(\widehat{D}+\Delta\widehat{D})(\widehat{Y}+\Delta\widehat{Y})\,x\|_2, \tag{6.4}$$

where

$$\begin{aligned} &\|\Delta\widehat{X}(:,j)\|_2 \le \widetilde{\gamma}_{mr}\,\|\widehat{X}(:,j)\|_2, \quad && \|\Delta\widehat{Y}(j,:)\|_2 \le \widetilde{\gamma}_{nr}\,\|\widehat{Y}(j,:)\|_2, \ \textit{for } j = 1,\ldots,r \\ &|\Delta\widehat{D}| \le \widetilde{\gamma}_1\,|\widehat{D}|, && \|\Delta b\|_2 \le \widetilde{\gamma}_{mr}\,\|b\|_2\,. \end{aligned}$$

(b) *$\widehat{x}_0$ is the exact minimum 2-norm solution of*

$$\min_{x\in\mathbb{C}^n} \|(b+\Delta b) - (X+\Delta X)(D+\Delta D)(Y+\Delta Y)\,x\|_2, \tag{6.5}$$

where

$$\begin{aligned} \|\Delta X\|_2 &\le\ (\mathtt{u}\,p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{mr} + \sqrt{r}\,\widetilde{\gamma}_{mr}\,\mathtt{u}\,p(m,n))\,\|X\|_2, \\ \|\Delta Y\|_2 &\le\ (\mathtt{u}\,p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{nr} + \sqrt{r}\,\widetilde{\gamma}_{nr}\,\mathtt{u}\,p(m,n))\,\|Y\|_2, \\ |\Delta D| &\le\ (\mathtt{u}\,p(m,n) + \widetilde{\gamma}_1 + \widetilde{\gamma}_1\,\mathtt{u}\,p(m,n))\,|D|, \qquad \|\Delta b\|_2 \le \widetilde{\gamma}_{mr}\,\|b\|_2\,. \end{aligned}$$

(c) *If* x_0 *is the exact minimum 2-norm solution of* $\min_{x\in\mathbb{C}^n}\|b - A\,x\|_2$*, then* $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ *can be bounded as in Theorem 5.1 with* $\alpha = (\mathtt{u}\,p(m,n) + \sqrt{r}\,\widetilde{\gamma}_{mr} + \sqrt{r}\,\widetilde{\gamma}_{mr}\,\mathtt{u}\,p(m,n))$*,* $\beta = (\mathtt{u}\,p(m,n)+\sqrt{r}\,\widetilde{\gamma}_{nr}+\sqrt{r}\,\widetilde{\gamma}_{nr}\,\mathtt{u}\,p(m,n))$*,* $\rho = (\mathtt{u}\,p(m,n)+ \widetilde{\gamma}_1 + \widetilde{\gamma}_1\,\mathtt{u}\,p(m,n))$*, and* $\epsilon = \widetilde{\gamma}_{mr}$*. In particular, to first order in* $\mathtt{u}$*, and if* c *is an small integer constant, then*

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\,\mathtt{u}\left[p_y(m,n)\,\kappa_2(Y) + p_x(m,n)\,\kappa_2(X)\,\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right] + O(\mathtt{u}^2)\,,$$

where $p_y(m,n) := (p(m,n) + nr^{3/2})$ *and* $p_x(m,n) := (p(m,n) + mr^{3/2})$*.*

Proof. In order to prove part (a) let us write the backward errors in steps 2, 3, and 4 of Algorithm 6.1.

1. The backward errors of `Step 2` are given in [28, Theorem 20.3]: the solution computed in `Step 2`, $\widehat{x}_1$, is the exact solution of the LS problem

   $$\min_{x\in\mathbb{C}^r}\|(b+\Delta b) - (\widehat{X} + \Delta\widehat{X})\,x\|_2\,, \tag{6.6}$$

   where $\|\Delta\widehat{X}(:,j)\|_2 \leq \widetilde{\gamma}_{mr}\|\widehat{X}(:,j)\|_2$, for $j = 1,\ldots,r$, and $\|\Delta b\|_2 \leq \widetilde{\gamma}_{mr}\,\|b\|_2$. Therefore, $\|\Delta\widehat{X}\|_2 \leq \|\Delta\widehat{X}\|_F \leq \widetilde{\gamma}_{mr}\|\widehat{X}\|_F \leq \sqrt{r}\widetilde{\gamma}_{mr}\|\widehat{X}\|_2$. Note also that, as we have commented before, (2.3) and (2.5) imply $\operatorname{rank}(X) = \operatorname{rank}(\widehat{X}) = r$, so Weyl perturbation theorem [44] for singular values and (6.2) imply $|\sigma_r(\widehat{X} + \Delta\widehat{X}) - \sigma_r(\widehat{X})|/\sigma_r(\widehat{X}) \leq \|\Delta\widehat{X}\|_2/\sigma_r(\widehat{X}) \leq \sqrt{r}\widetilde{\gamma}_{mr}\kappa_2(\widehat{X}) < 1$, and, finally, $\operatorname{rank}(\widehat{X}) = \operatorname{rank}(\widehat{X} + \Delta\widehat{X}) = r$. As a consequence, $\widehat{x}_1$ satisfies

   $$\widehat{x}_1 = (\widehat{X} + \Delta\widehat{X})^\dagger(b + \Delta b), \tag{6.7}$$

   with $\widehat{X} + \Delta\widehat{X} \in \mathbb{C}^{m\times r}$ and $\operatorname{rank}(\widehat{X} + \Delta\widehat{X}) = r$.
2. As a consequence of [28, Lemma 3.5], the solution, $\widehat{x}_2$, computed in `Step 3` obeys

   $$(\widehat{D} + \Delta\widehat{D})\,\widehat{x}_2 = \widehat{x}_1 \quad \text{with} \quad |\Delta\widehat{D}| \leq \widetilde{\gamma}_1|\widehat{D}|, \tag{6.8}$$

   with $\widehat{D}+\Delta\widehat{D} \in \mathbb{C}^{r\times r}$ diagonal and nonsingular, since (2.4) and (2.5) imply $\operatorname{rank}(D) = \operatorname{rank}(\widehat{D}) = r$ and $\widetilde{\gamma}_1 < 1$ by (6.2).
3. The backward errors of `Step 4` are given in [28, Theorem 21.4]. The application of [28, Theorem 21.4] requires $\operatorname{rank}(\widehat{Y}) = r$, which follows from (2.3) and (2.5), and the assumption

   $$\|\,|\widehat{Y}^\dagger|\,|\widehat{Y}|\,\|_2\,r\,n\,\gamma_n < 1,$$

   which is guaranteed by (6.3), since $\|\,|\widehat{Y}^\dagger|\,|\widehat{Y}|\,\|_2\,r\,n\,\gamma_n \leq \kappa_2(\widehat{Y})\,r^2\,n\,\gamma_n < 1$. With this condition, the minimum 2-norm solution computed in `Step 4`, $\widehat{x}_0$, is the exact minimum 2-norm solution of the underdetermined system

   $$(\widehat{Y} + \Delta\widehat{Y})x = \widehat{x}_2\,,$$

   with $\|\Delta\widehat{Y}(j,:)\|_2 \leq \widetilde{\gamma}_{nr}\|\widehat{Y}(j,:)\|_2$, for $j = 1,\ldots,r$. In addition, we can prove $\operatorname{rank}(\widehat{Y}) = \operatorname{rank}(\widehat{Y} + \Delta\widehat{Y}) = r$ via an argument similar to the one we used to prove the same for $\widehat{X} + \Delta\widehat{X}$. Therefore, $\widehat{x}_0$ obeys

   $$\widehat{x}_0 = (\widehat{Y} + \Delta\widehat{Y})^\dagger\widehat{x}_2, \tag{6.9}$$

   with $\widehat{Y} + \Delta\widehat{Y} \in \mathbb{C}^{r\times n}$ and $\operatorname{rank}(\widehat{Y} + \Delta\widehat{Y}) = r$.

From (6.7), (6.8), and (6.9) we have that

$$\widehat{x}_0 = (\widehat{Y} + \Delta\widehat{Y})^\dagger (\widehat{D} + \Delta\widehat{D})^{-1} (\widehat{X} + \Delta\widehat{X})^\dagger (b + \Delta b) \tag{6.10}$$

$$= \left[(\widehat{X} + \Delta\widehat{X})\, (\widehat{D} + \Delta\widehat{D})\, (\widehat{Y} + \Delta\widehat{Y}) \right]^\dagger (b + \Delta b)\,, \tag{6.11}$$

where the second equality follows from Lemma 2.2-(c). This and the bounds we have developed for $\widehat{X}$, $\widehat{D}$, and $\widehat{Y}$ prove part (a) of Theorem 6.2.

The proof of Theorem 6.2-(b) follows easily from part-(a). Equations (2.3) and (2.4) allow us to write $\widehat{X} = X + E_X$, $\widehat{D} = D + E_D$, and $\widehat{Y} = Y + E_Y$, where $\|E_X\|_2 \leq \mathtt{u}\, p(m,n)\, \|X\|_2$, $|E_D| \leq \mathtt{u}\, p(m,n) |D|$, and $\|E_Y\|_2 \leq \mathtt{u}\, p(m,n) \|Y\|_2$. Therefore, we can write

$$\widehat{X} + \Delta\widehat{X} = X + E_X + \Delta\widehat{X} =: X + \Delta X\,, \tag{6.12}$$

where

$$\begin{aligned}
\|\Delta X\|_2 &\leq \|E_X\|_2 + \|\Delta\widehat{X}\|_2 \\
&\leq \mathtt{u}\, p(m,n)\, \|X\|_2 + \sqrt{r}\, \widetilde{\gamma}_{mr} \|\widehat{X}\|_2 \\
&\leq \mathtt{u}\, p(m,n)\, \|X\|_2 + \sqrt{r}\, \widetilde{\gamma}_{mr} \left(\|X\|_2 + \|E_X\|_2 \right) \\
&\leq \left(\mathtt{u}\, p(m,n) + \sqrt{r}\, \widetilde{\gamma}_{mr} + \sqrt{r}\, \widetilde{\gamma}_{mr}\, \mathtt{u}\, p(m,n) \right) \|X\|_2 \,.
\end{aligned} \tag{6.13}$$

Analogously, we can write

$$\widehat{D} + \Delta\widehat{D} =: D + \Delta D\,, \quad \text{with } |\Delta D| \leq \left(\mathtt{u}\, p(m,n) + \widetilde{\gamma}_1 + \widetilde{\gamma}_1\, \mathtt{u}\, p(m,n) \right) |D|\,,$$

$$\widehat{Y} + \Delta\widehat{Y} =: Y + \Delta Y\,, \quad \text{with } \|\Delta Y\|_2 \leq \left(\mathtt{u}\, p(m,n) + \sqrt{r}\, \widetilde{\gamma}_{nr} + \sqrt{r}\, \widetilde{\gamma}_{nr}\, \mathtt{u}\, p(m,n) \right) \|Y\|_2 \,.$$

If these equations and (6.12)-(6.13) are inserted into (6.4), then (6.5) is obtained and part (b) is proved. Finally, part (c) is an immediate consequence of part (b) and Theorem 5.1. □

Observe that, since in an RRD the factors X and Y are well conditioned, Theorem 6.2-(c) guarantees that the forward error in the solution computed by Algorithm 6.1 is bounded by $O(\mathtt{u}) \|A^\dagger\|_2 \|b\|_2 / \|x_0\|_2$.

7. Numerical experiments. In this section we will show numerical tests done using MATLAB™ that illustrate how well the errors committed by Algorithm 6.1 compare with the theoretical predictions and with the errors committed by the usual method to solve LS problems using the QR factorization computed with the traditional Householder algorithm as implemented in MATLAB™ [28, Section 20.2]. For that, we will use three important classes of rectangular structured matrices that may have huge condition numbers: Cauchy, Vandermonde, and Graded matrices. For matrices in these classes, accurate RRDs in the sense of Definition 2.3 can be computed using the algorithms in [8] and [27]. We will present tests only for matrices $A \in \mathbb{R}^{m \times n}$ with real entries, $m \geq n$, and such that $\operatorname{rank}(A) = n$, which means that we consider only LS problems with unique solutions.

We know from (1.1) and (4.14) that if $\widehat{x}_0$ is the unique solution of $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$ computed by the QR algorithm in MATLAB™ and x_0 is the exact solution, then

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq c\, m\, n^{3/2}\, \mathtt{u}\, \kappa_2(A) \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2}, \tag{7.1}$$

which is a bound larger than (1.1) but reliable in most situations. In contrast, Algorithm 6.1 satisfies (see (1.2) and Theorem 6.2-(c))

$$\frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq \mathtt{u}\, f(m,n) \, \left(\kappa_2(Y) + \kappa_2(X) \frac{\|A^\dagger\|_2 \|b\|_2}{\|x_0\|_2} \right). \tag{7.2}$$

In our tests we have computed the relative error in the solution for both algorithms, and also the quantities

$$(7.3)\quad \Theta_{QR} := \mathtt{u}\left(\kappa_2(A)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right), \quad \Theta_{RRD} := \mathtt{u}\left(\kappa_2(Y)+\kappa_2(X)\frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right),$$

to check, erasing pessimistic dimensional constants, how sharp the bounds (7.1) and (7.2) are. We will see that Algorithm 6.1 is by far the most accurate of the two: for random right-hand sides b, it achieves relative normwise errors of order unit roundoff, which means that Θ_{RRD} is $O(\mathtt{u})$ almost always even for extremely ill-conditioned matrices A.

7.1. Cauchy matrices. The entries of a Cauchy matrix, $C \in \mathbb{R}^{m\times n}$, $m \geq n$, are defined in terms of two vectors $z = [z_1, \ldots, z_m]^T \in \mathbb{R}^m$, $y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ as

$$(7.4)\qquad c_{ij} = \frac{1}{z_i + y_j}, \qquad i = 1, \ldots, m,\ j = 1, \ldots, n.$$

Matrices of the form $G = S_1 C S_2$, where C is Cauchy and S_1, S_2 are diagonal and nonsingular, are called in [8] quasi-Cauchy matrices, which include, as a particular case, Cauchy matrices for $S_1 = I_m,\ S_2 = I_n$. Quasi-Cauchy matrices have full column rank if $z_i \neq z_j$ for any $i \neq j$, $y_k \neq y_l$ for any $k \neq l$, and $z_i \neq -y_j$ for all i, j. Algorithm 3 in [8] uses a structured version of GECP to compute an accurate RRD of any *square* quasi-Cauchy matrix. This algorithm can be very easily extended to deal with rectangular matrices, and this version is the one used in the tests of this section to compute the RRD in `Step 1` of Algorithm 6.1. The overall cost of this step is $2mn^2 - 2n^3/3 + O(n^2 + mn)$ operations plus $mn^2/2 - n^3/6 + O(n^2 + mn)$ comparisons.

In order to make easy references, let us summarize and give names to the two algorithms that are used in this section for solving $\min_{x\in\mathbb{R}^n} \|Cx - b\|_2$, with C a Cauchy matrix:

- `LS-QR`: given the vectors z and y, the entries of C are computed as in (7.4) and the LS problem is solved using the Householder QR factorization implemented in `MATLAB`™.
- `LS-RRD`: the LS problem is solved using Algorithm 6.1 with the RRD in `Step 1` computed with the rectangular version discussed above of Algorithm 3 in [8].

The QR factorizations needed in `Steps 2` and `4` of Algorithm 6.1 are computed with the routine in `MATLAB`™. Note that in this case the linear system $Yx = x_2$ in `Step 4` has the matrix Y nonsingular and GE with partial pivoting can be used in its solution.

In our tests, we have generated Cauchy matrices with random z and y vectors, we have generated also random right-hand side vectors b, and we have computed the solution of $\min_{x\in\mathbb{R}^n} \|Cx - b\|_2$ using the algorithms `LS-QR` and `LS-RRD`. To compute the relative errors $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, we take as "exact" solution x_0 the one computed via the `svd` command of `MATLAB`™run in variable precision arithmetic. In each test we have set the precision to $2\log_{10}|D_1/D_n| + 30$ decimal digits, where D_1 and D_n are, respectively, the largest and the smallest (in absolute value) diagonal entries of the diagonal matrix D in the RRD of C computed in `Step 1` of Algorithm 6.1. The motivation of taking $2\log_{10}|D_1/D_n| + 30$ decimal digits comes from the facts that $|D_1/D_n|$ has a magnitude similar to $\kappa_2(C)$, because X and Y are well-conditioned, and that, according to (7.1) and to the discussion in Subsection 4.1, the error in traditional algorithms for LS problems is almost always much smaller than $\mathtt{u}\,\kappa_2(C)^2$. The random vectors z, y, and b have been chosen either from the uniform distribution in the interval $[0, 1]$ (command `rand` in `MATLAB`™), or from the standard normal distribution (command `randn` in `MATLAB`™). In all experiments we have tested the eight resulting possibilities in the choice of the random distributions for z, y, and b.

Two kind of experiments have been done. In the first group, we have fixed the size of the matrix: $m \times n = 100 \times 50, 50 \times 30$, or 25×10. For each size we have generated 50×8 different sets of the random vectors z, y, and b, therefore, generating a total of 400 different LS problems for each size. Figure 7.1 shows the results for the size 100×50 when the vectors z and b are selected from the standard normal distribution and the vector y from the uniform distribution in $[0, 1]$. We have plotted in a log-log scale the relative error $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ of the solution against the condition number (computed from the "exact" SVD in variable precision arithmetic used to compute the "exact solution" x_0) of the matrices for the algorithms LS−QR and LS−RRD. Besides this, we have displayed also the quantities Θ_{QR} and Θ_{RRD} appearing in (7.3). We observe that the relative error in the LS−RRD algorithm is of order $\mathbf{u}$ times a small constant, as predicted, while the error for LS−QR scales almost linearly with $\kappa_2(C)$ until it saturates. The linear dependence on $\kappa_2(C)$ of the relative error in LS−QR is the predicted by (7.1) since $\|C^\dagger\|_2\|b\|_2/\|x_0\|_2$ has been always moderate in these tests. It can be also observed that the bound Θ_{RRD} is rather sharp and does not overestimate the actual errors. For other sizes and other ways to generate z, y, and b the results have been similar.

In our second group of tests we have fixed the number of rows of the matrix and varied the number of the columns. We have tested matrices of sizes $m = 100,\ n = 10 : 10 : 90$ (5×8 sets of random vectors z, y, and b for each size), $m = 50,\ n = 10 : 2 : 40$ (10×8 sets of random vectors z, y, and b for each size), and $m = 25,\ n = 5 : 5 : 20$ (20×8 sets of random vectors z, y, and b for each size). This makes a total of 2280 matrices. Figure 7.2 shows the results for $m = 50,\ n = 10 : 2 : 40$ for four different combinations of the random distributions for z, y, and b. For each size we plot the maximum relative error out of the 10 samples. Again the relative errors of the solution for the LS-RRD algorithm are of order $\mathbf{u}$ times a moderate constant, while for the LS-QR algorithm are huge. For other sizes and other ways to generate z, y, and b the results have been similar.

For all our experiments with Cauchy matrices, the range of the condition numbers has been $10^0 \lesssim \kappa_2(C) \lesssim 10^{100}$, the maximum value of the term $\|C^\dagger\|_2\|b\|_2/\|x_0\|_2$ has been 1376, $8 \leq \kappa_2(X) \leq 72$, and $13 \leq \kappa_2(Y) \leq 58$.

7.2. Vandermonde matrices. We have performed numerical tests of Algorithm 6.1 similar to those in Subsection 7.1 with Vandermonde matrices. Vandermonde matrices appear naturally in polynomial data fitting. Given a vector of points $z = [z_1, \ldots, z_m]^T \in \mathbb{R}^m$, such that $z_i \neq z_j$ if $i \neq j$, and a set of "function values" $b = [b_1, \ldots, b_m]^T \in \mathbb{R}^m$, the problem of finding the polynomial $P_{n-1}(z)$, of degree less than or equal to $n-1$, $n \leq m$, that best fits the data z, b in the LS sense is equivalent as solving the LS problem $\min_{c \in \mathbb{R}^n} \|Vc - b\|_2$ where $V \in \mathbb{R}^{m \times n}$ is a Vandermonde matrix, whose entries are given by

$$v_{ij} = z_i^{j-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n, \tag{7.5}$$

and the sought solution $c \in \mathbb{R}^n$ is the vector that contains the coefficients of the polynomial $P_{n-1}(z) = c_1 + c_2 z + \cdots + c_n z^{n-1}$. A method to compute an accurate RRD of any Vandermonde matrix was presented in [8, Section 5]. It is based on the fact that if $F \in \mathbb{C}^{n \times n}$ is the $n \times n$ discrete Fourier transform, then VF is a quasi-Cauchy matrix whose parameters can be accurately computed in $O(mn)$ operations, as well as the sums and subtractions of any pair of these parameters. Then, an accurate RRD of $VF = XDY$ can be computed with the Algorithm 3 in [8] adapted to deal with rectangular matrices as in Subsection 7.1. Finally, $V = X\,D\,(YF^*)$ is an accurate RRD of V. The overall cost is the same as for Cauchy matrices: $2mn^2 - 2n^3/3 + O(n^2 + mn)$ operations plus $mn^2/2 - n^3/6 + O(n^2 + mn)$

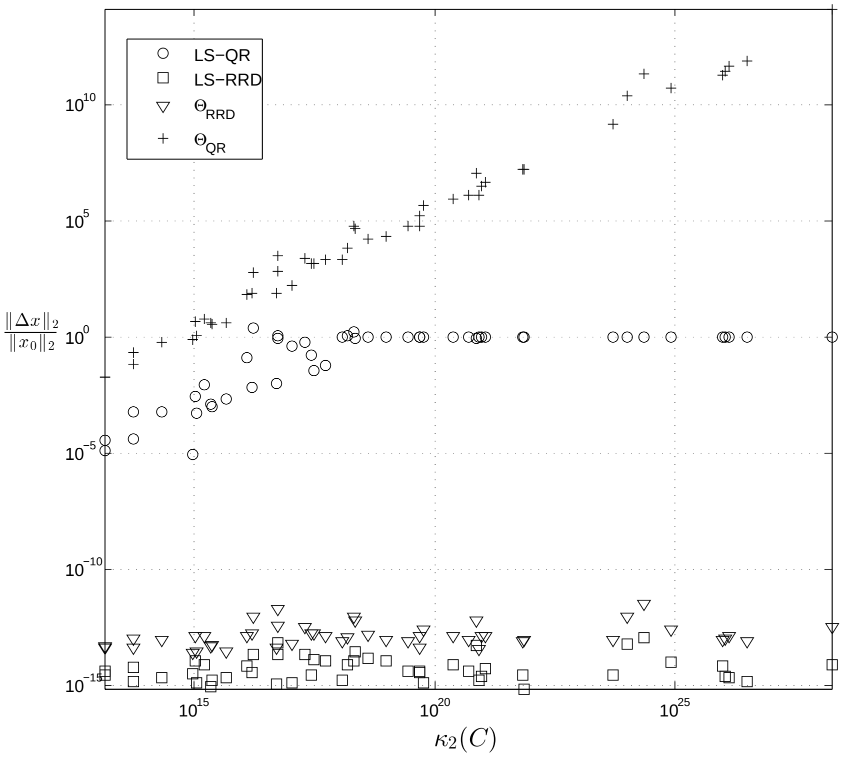

FIG. 7.1. *Forward relative error* $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ *against* $\kappa_2(C)$. *C are random* 100×50 *Cauchy matrices. The vectors z and b are selected from the standard normal distribution and the vector y from the uniform distribution in* $[0, 1]$.

comparisons. This algorithm will be used to compute the RRD in `Step 1` of Algorithm 6.1. The two algorithms used in this section for solving $\min_{x\in\mathbb{R}^n} \|Vx - b\|_2$ are:

- `LS-QR`: given the vector z, the entries of V are computed as in (7.5) and the LS problem is solved using the Householder QR factorization implemented in `MATLAB`™.
- `LS-RRD`: the LS problem is solved using Algorithm 6.1 with the RRD in `Step 1` computed with the rectangular version discussed above of the method in [8, Sec. 5].

In our tests, we have generated Vandermonde matrices with random z vectors, we have generated also random right-hand side vectors b, and we have computed the solution of $\min_{x\in\mathbb{R}^n} \|Vx - b\|_2$ using the algorithms `LS-QR` and `LS-RRD`. To compute the relative error $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, we follow the procedure presented in Subsection 7.1 to obtain the “exact” solution x_0. The random vectors z and b have been chosen either from the uniform distribution in $[0, 1]$ or from the standard normal distribution. In all experiments we have tested the four resulting combinations in the choice of the random distributions for z and b.

We have tested $m \times n$ matrices of sizes $m = 50, n = 5\!:\!5\!:\!30$; $m = 100, n = 10\!:\!5\!:\!60$; and $m = 500, n = 100 : 50 : 250$. For each size we have generated different sets of the random vectors z and b (for $m = 50, 100$, 25×4 different sets, and for $m = 500$, 10×4 different sets), generating a total of 1860 different LS problems. Figure 7.3 shows the results for the sizes $m = 100,\ n = 10\!:\!5\!:\!60$, when the vectors z and b are chosen from the standard normal distribution. We have plotted in a log-log scale the relative error $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ of the solution against the condition number (computed from the “exact" SVD as in Subsection 7.1) of the matrices for the algorithms `LS-QR` and `LS-RRD`. Besides this, we have displayed

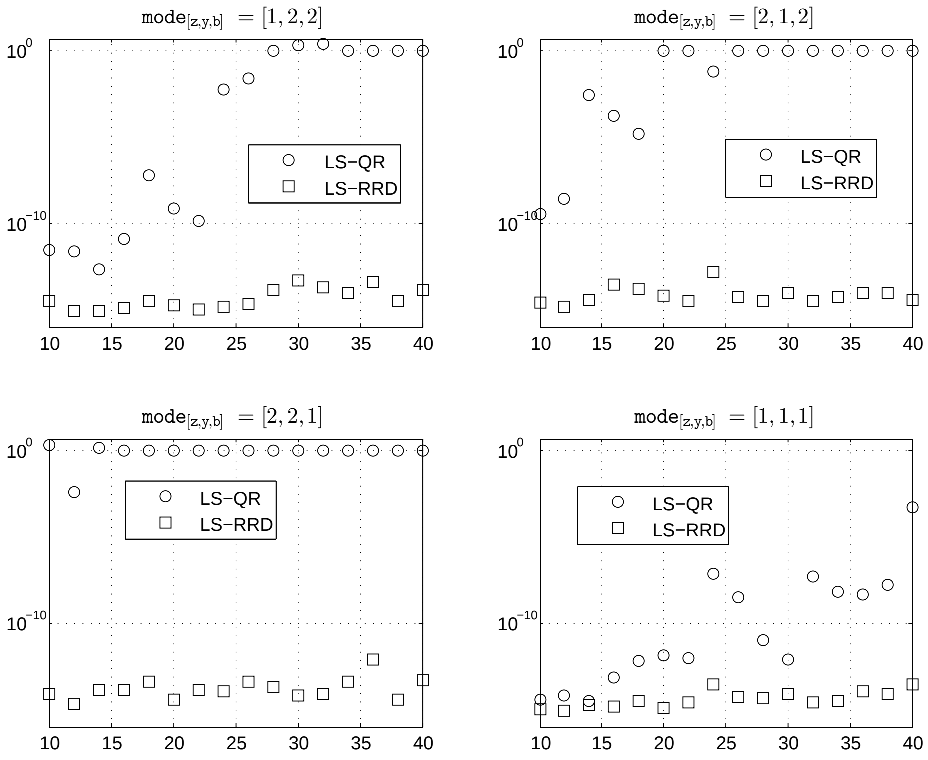

FIG. 7.2. *Forward relative error* $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ *against* n, *for* $m \times n$ *Cauchy matrices of sizes* $m = 50$, $n = 10:2:40$ *(*10 *matrices for each size) for four different combinations of the random distributions (mode=1 denotes the standard normal distribution and mode=2 denotes the uniform distribution in* $[0,1]$*) for* z, y*, and* b*.*

also the quantities[‖] Θ_{QR} and Θ_{RRD} appearing in (7.3). The results are similar to those obtained in Figure 7.1 for Cauchy matrices and we refer the reader to the comments made in Subsection 7.1. For other sizes and other ways to generate z and b the results have been similar, producing the algorithm LS−RRD relative errors that are always of order $\mathbf{u}$ times a moderate constant and the algorithm LS−QR relative errors that scale linearly with $\kappa_2(V)$ and that are very large for very large $\kappa_2(V)$. For all our experiments with Vandermonde matrices, the range of the condition numbers has been $10^0 \lesssim \kappa_2(V) \lesssim 10^{70}$, the maximum value of the term $\|V^\dagger\|_2\|b\|_2/\|x_0\|_2$ has been 1076, $4 \leq \kappa_2(X) \leq 65$, and $3 \leq \kappa_2(Y) \leq 87$.

7.3. Graded matrices. Another class of matrices for which it is possible to compute an accurate RRD under certain conditions are graded matrices: matrices of the form $A = S_1 B S_2 \in \mathbb{R}^{m\times n}$, with $S_1 \in \mathbb{R}^{m\times m}$ and $S_2 \in \mathbb{R}^{n\times n}$ nonsingular diagonal matrices that may be arbitrarily ill-conditioned, $B \in \mathbb{R}^{m\times n}$ a well-conditioned matrix, and $\operatorname{rank}(A) = n$. Therefore, the matrix A can have a huge condition number. Higham in [27] determined conditions such that if the QR factorization with *complete pivoting* (column pivoting together with row sorting or row pivoting, see [27] for details) of a graded matrix A is computed and the SVD of the permuted R factor is computed via the one-sided Jacobi algorithm [24], then the SVD of A is obtained with high relative accuracy. In this section we show that under the

[‖] To keep the scale of the plot we have plotted $\min(\Theta_{QR}, 10)$ instead of Θ_{QR}, since values of Θ_{QR} much larger than in Figure 7.1 have appeared in these tests.

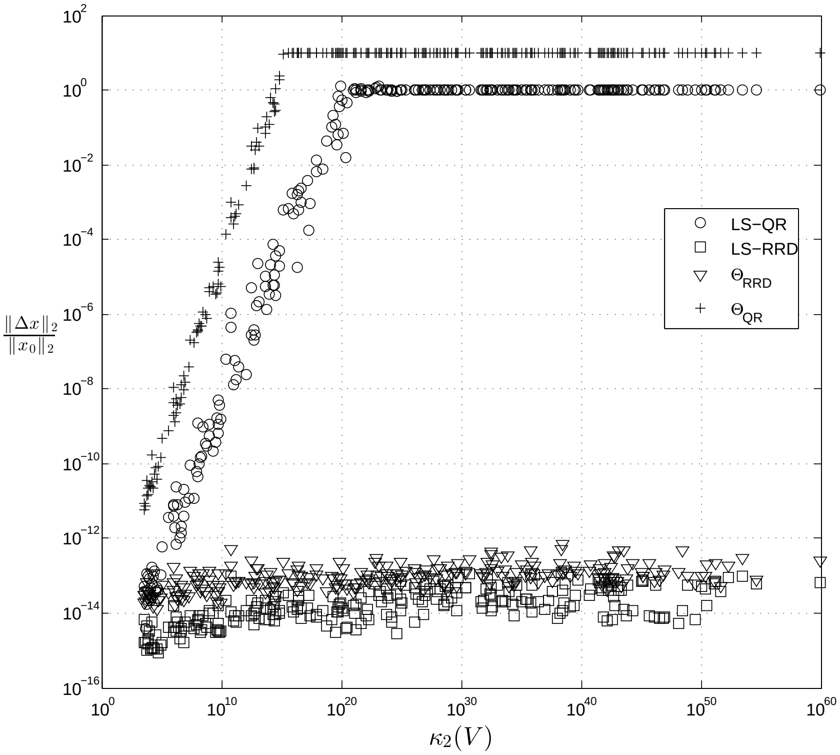

FIG. 7.3. *Forward relative error* $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$ *against* $\kappa_2(V)$ *for random Vandermonde matrices* V *of sizes* $100 \times 10\!:\!5\!:\!60$*. The random vectors* z *and* b *are selected from the standard normal distribution.*

same conditions, the QR factorization with *complete pivoting* can be used to solve accurately and very efficiently LS problems whose coefficient matrix is graded. To this purpose, note that if $P_R A P_C = QR$ is the thin QR factorization with complete pivoting, where P_R and P_C are permutation matrices, $Q \in \mathbb{R}^{m\times n}$, and $R \in \mathbb{R}^{n\times n}$, then an RRD $A = XDY$ can be obtained as follows:

$$A = P_R^T Q R P_C^T = (P_R^T\, Q)\, D\, (D^{-1}\, R\, P_C^T),$$

where $D := \operatorname{diag}(R)$, $X := P_R^T Q$, and $Y := D^{-1}\, R\, P_C^T$. In this case, `Step 2` of Algorithm 6.1 reduces to $x_1 = Q^T P_R b$ and `Steps 3-4` can be merged in only one step without affecting the rounding errors, since the errors are componentwise and D diagonal. This merged step is simply to solve the linear system $(R P_C^T) x = x_1$. Therefore, D is not necessary and Algorithm 6.1 simplifies to Algorithm 7.1, which is almost the usual algorithm to solve LS problems using the QR factorization.

```
ALGORITHM 7.1. (Accurate solution of graded LS problems via QR-complete pivoting)
Input:  $A \in \mathbb{R}^{m\times n}$ graded matrix, $b \in \mathbb{R}^m$, $m \geq n = \operatorname{rank}(A)$.
Output: $x_0$ unique solution of $\min_{x\in\mathbb{R}^n} \|b - A\,x\|_2$

Step 1: Compute the thin QR decomposition with complete pivoting (see [27]) of $A$,
        $A = P_R^T Q R P_C^T$.
```

`Step 2`: Compute $x_1 = Q^T P_R b$.

`Step 3`: Solve the consistent linear system $R P_C^T x = x_1$ to get x_0.

The cost of Algorithm 7.1 is essentially the same as the standard Householder QR method, i.e., $2mn^2 - \frac{2}{3}n^3$ flops, since the cost of pivoting is $O(mn)$. As usual it is not necessary to form explicitly the matrix Q. Algorithm 7.1 runs also for matrices A that are not graded, but then the accuracy is not guaranteed.

It is important to note that GECP can be also used to compute under the same conditions an accurate RRD of a graded matrix [9, Section 4] and then Algorithm 6.1 can be used to solve accurately graded LS problems. However, this procedure needs to compute a QR factorization in `Step 2` of Algorithm 6.1 and, so, it is more expensive than Algorithm 7.1.

Next, we explain briefly and in a simplified non-rigorous way which are the errors committed by Algorithm 7.1. These errors are essentially equal to those of Algorithm 6.1 taking as RRD $A = (P_R^T Q)\, D\, (D^{-1} R P_C^T)$, and therefore they are determined by the errors committed in the computation of the QR factorization with complete pivoting. In order to make notation simpler, we assume that the matrix A has been pre-pivoted, i.e., that `Step 1` of Algorithm 7.1 produces $P_R = I_m$ and $P_C = I_n$. Observe that this induces corresponding pre-permutations in S_1, B, and S_2. First, it was proved in [27, Theorem 2.5] that if $A = S_1 B S_2 \in \mathbb{R}^{m \times n} (m \geq n)$, for arbitrary nonsingular diagonal matrices S_1 and S_2, then the factor $\widehat{R}$ computed in `Step 1` of Algorithm 7.1, obeys

$$S_1(B + \Delta B)S_2 = Q\widehat{R}, \qquad \text{with} \quad \|\Delta B\|_2 = O(\mathtt{u})\|B\|_2, \tag{7.6}$$

where $Q \in \mathbb{R}^{m \times n}$ is a matrix having exactly orthonormal columns. The big-O notation in (7.6) hides some factors: growth factors and low degree polynomials in m and n, that might be important in some rare cases. See [27] for the details. However our experiments (as those in [27]) show that $O(\mathtt{u})$ is in practice a small constant times $\mathtt{u}$. If we would compute the factor Q explicitly, then we would obtain a matrix $\widehat{Q}$ such that $\|Q - \widehat{Q}\|_2 = O(\mathtt{u})$ [28, eq. (19.13)], where Q is the matrix in (7.6). Therefore, we will not distinguish between the exact Q and the computed one in the following discussion.

Now we need to get, from the backward error in (7.6), forward errors on the RRD of the type appearing in Definition 2.3. To this purpose, recall that if $B = LU$ has an LU factorization (without pivoting), $L \in \mathbb{R}^{m \times n}$ and $U \in \mathbb{R}^{n \times n}$, whose factors are well-conditioned and such that $\|B^\dagger\|_2 \approx \|L^\dagger\|_2 \|U^{-1}\|_2$, then it was proven in the proof of [9, Theorem 4.1] that $S_1(B + \Delta B)S_2$ can be written as[**]

$$S_1(B + \Delta B)S_2 = (I + E)A(I + F), \tag{7.7}$$

$$\text{with} \quad \max(\|E\|_2, \|F\|_2) = O(\tau\, \|\Delta B\|_2 \|B^\dagger\|_2) = O(\mathtt{u})\, \tau\, \kappa_2(B), \tag{7.8}$$

where the factor τ controls the grading (after the permutations in `Step 1` of Algorithm 7.1) and it is given by

$$\tau = \max(1, \tau_1, \tau_2), \quad \text{with} \quad \tau_1 = \max_{\substack{1 \leq j \leq n \\ j \leq k \leq m}} \frac{|(S_1)_{kk}|}{|(S_1)_{jj}|}, \quad \tau_2 = \max_{1 \leq j \leq k \leq n} \frac{|(S_2)_{kk}|}{|(S_2)_{jj}|}. \tag{7.9}$$

Combining (7.6) and (7.7), neglecting second order terms, and defining $\widehat{D} := \mathrm{diag}(\widehat{R})$, we get $A = (I - E)Q\widehat{D}(\widehat{D}^{-1}\widehat{R})(I - F)$. So, $X = (I - E)Q$, the matrix $\widehat{D}$, and $Y =$

[**] The actual bounds are more complicated and include $\kappa_2(L)$ and $\kappa_2(U)$, that may be much larger than $\kappa_2(B)$ since the pivoting is made on A and not on B (see [9, Section 4] and [27] for details). In our discussion, we pretend to emphasize the main factors that control the errors in practice and not to present a rigorous analysis.

$(\widehat{D}^{-1}\widehat{R})(I-F)$ can be considered as the factors of an exact RRD of A, whose computed factors would be $\widehat{X} = Q$, $\widehat{D}$, and $\widehat{Y} = (\widehat{D}^{-1}\widehat{R})$. Observe that here the exact and the computed diagonal factors are the same. Therefore, we can use (7.8) to get $\max\{\|\widehat{X} - X\|_2/\|\widehat{X}\|_2\,,\ \|\widehat{Y}-Y\|_2/\|\widehat{Y}\|_2\} = O(\mathtt{u})\,\tau\,\kappa_2(B)$, which allows us to apply Theorem 6.2-(c) with $p(m,n)$ replaced by $\tau\,\kappa_2(B)$ and to get to first order the following forward error on the solution computed by Algorithm 7.1

$$\text{(7.10)} \qquad \frac{\|\widehat{x}_0 - x_0\|_2}{\|x_0\|_2} \leq O(\mathtt{u})\,\tau\,\kappa_2(B)\left(\kappa_2(\widehat{D}^{-1}\widehat{R}) + \frac{\|A^\dagger\|_2\|b\|_2}{\|x_0\|_2}\right) + O(\mathtt{u}^2)\,,$$

where we have used that $\kappa_2(Q) = 1$. A key point in the bound (7.10) is the parameter τ, which penalizes $\kappa_2(B)$. It might be very large, because the diagonal matrices S_1 and S_2 can be arbitrarily ill-conditioned. However, the permutations coming from QR with complete pivoting almost always reorder S_1 and S_2 in such a way that τ is of order one.

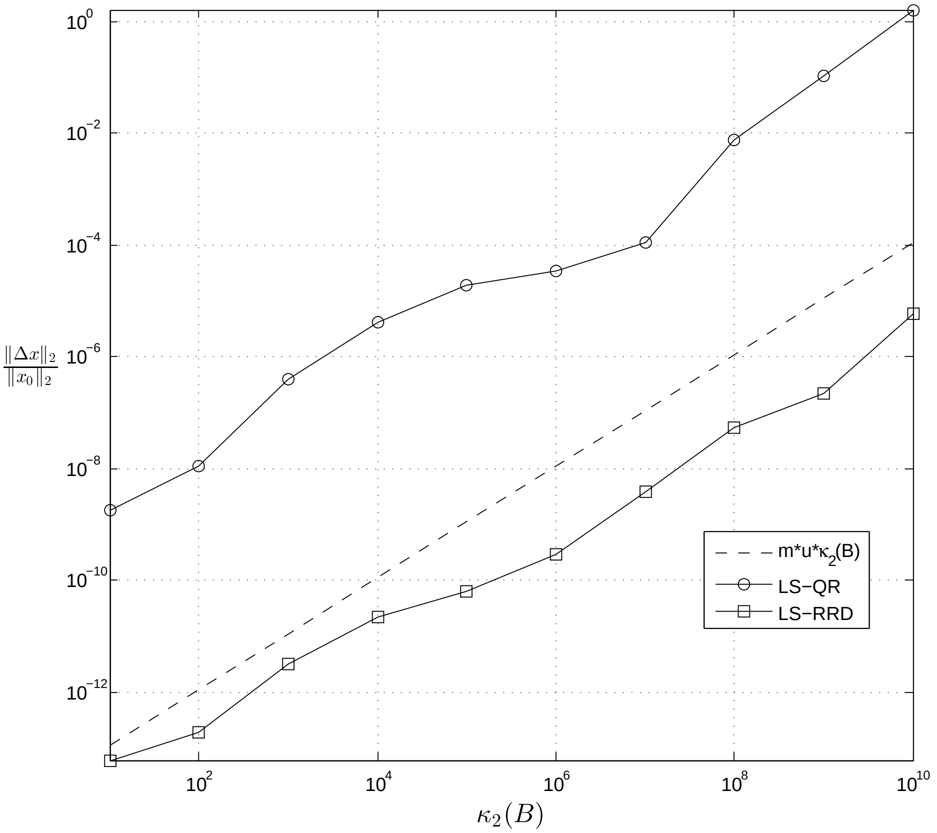

FIG. 7.4. *Forward relative error* $\|\widehat{x}_0-x_0\|_2/\|x_0\|_2$ *against* $\kappa_2(B)$ *for random graded matrices* $A = S_1BS_2$ *of size* 100×40*, with* B*,* S_1*, and* S_2 *generated with the option (b) explained in text.*

To test the accuracy of Algorithm 7.1 for graded matrices A we have performed several numerical experiments similar to those in [27] and we have used the following two algorithms for solving $\min_{x\in\mathbb{R}^n}\|Ax-b\|_2$:

- LS-QR: computes the solution using the Householder QR factorization with *column* pivoting as implemented in MATLAB™.

- LS-RRD: uses Algorithm 7.1 with the QR factorization in Step 1 computed with *row sorting and column pivoting* [27].

Observe that the algorithms LS-QR in Subsections 7.1 and 7.2 used Householder QR *without* any pivoting. Here, we use column pivoting to illustrate that the additional row sorting in LS-RRD is fundamental to get accurate solutions. To compute the relative error $\|\widehat{x}_0 - x_0\|_2/\|x_0\|_2$, we follow the procedure presented in Subsection 7.1 with $D = \operatorname{diag}(R)$.

In our tests, we have generated random matrices of the form $A = S_1 B S_2$, where B is constructed always using the mode 3 in the routine randsvd from the Test Matrix Toolbox developed by Higham [26], i.e., B is a random dense matrix with given condition number and with its singular values distributed geometrically. The diagonal matrices S_1 and S_2 are also generated with randsvd by using a variety of distributions for its singular values, i.e., for its diagonal entries. We have run experiments where the sizes $m \times n$ of the matrices A and B have been $m = 50,\ n = 10:10:30$ and $m = 100,\ n = 20:20:60$. The matrices A have been built in the following way: we have chosen matrices B with condition numbers $\kappa_2(B) = 10^i$, for $i = 1:10$. We have generated diagonal matrices S_1 and S_2 with positive diagonal entries chosen from one of the three pairs of distributions: (a) the entries of both S_1 and S_2 having uniformly distributed logarithm (mode 5 in randsvd), but with decreasing order for the entries for S_1 and increasing order for S_2; (b) the entries of both S_1 and S_2 being geometrically distributed (mode 3 in randsvd) but with increasing order for S_1 and decreasing for S_2; (c) geometrically distributed entries in decreasing order for S_1 and entries with uniformly distributed logarithm in increasing order for S_2. We took $\kappa_2(S_1) = \kappa_2(S_2) = 10^k$ with $k = 2:2:16$. For each size, for each option (a), (b), or (c), and for each triplet $(\kappa_2(B), \kappa_2(S_1), \kappa_2(S_2))$, ten matrices were generated and ten right-hand sides b, which follow the standard normal distribution. This makes, for each size and each $\kappa_2(B)$, a total of $10 \times 8 \times 3 = 240$ matrices.

In Figure 7.4 we show the results for the size 100×40 and with option (b) for matrices S_1 and S_2. We have plotted in a log-log scale the maximum relative error for all the 80 matrices with that particular $\kappa_2(B)$. It is observed that the error for the LS-RRD algorithm behaves as $O(\mathrm{u})\,\kappa_2(B)$, which according to (7.10) implies that the numbers τ, $\kappa_2(\widehat{D}^{-1}\widehat{R})$, and $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ have been of order one in all these tests. The error of the algorithm LS-QR loses as many as six more digits of precision, behaving notably worse than LS-RRD. We have plotted also a dashed line showing the quantity $m\,\mathrm{u}\,\kappa_2(B)$. The behaviors for all the other sizes of the matrices and all the other modes of randsvd are similar.

For all our experiments with graded matrices, the range of the condition numbers has been $10 \lesssim \kappa_2(A) \lesssim 10^{40}$, the maximum value of the term $\|A^\dagger\|_2\|b\|_2/\|x_0\|_2$ has been 108, and $2 \leq \kappa_2(\widehat{D}^{-1}\widehat{R}) \leq 18$.

7.4. Numerical tests controlling the residual. In the numerical tests done previously the right-hand sides b where chosen randomly. Therefore, the relative residual $\rho_r := \|b - Ax_0\|_2/\|b\|_2$ has been really small very rarely, because if $\operatorname{rank}(A) = n < m$, then it is very unlikely that $\theta(b, \mathcal{R}(A))$ is very small. In this subsection, we consider a different type of tests in which we generate random vectors b with a fixed value of ρ_r. For this purpose, we have performed experiments with $m \times n$ Cauchy and Vandermonde matrices in the following way. We first get the RRD of the matrix $A = XDY$. We compute the full SVD of the matrix X (something that can be done accurately via the command svd of MATLAB™ since X is well-conditioned): $X = U_X \Sigma_X V_X^T$, where $U_X \in \mathbb{R}^{m\times m}$, $\Sigma_X \in \mathbb{R}^{m\times n}$, and $V_X \in \mathbb{R}^{n\times n}$. Then, we partition $U_X = [U_1\ U_2]$, where $U_1 \in \mathbb{R}^{m\times n}$ and $U_2 \in \mathbb{R}^{m\times(m-n)}$, and generate random vectors $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{m-n}$ such that $\|\alpha\|_2 = \|\beta\|_2 = 1$. With this we define

$$b_0 := U_1\alpha \in \mathcal{R}(A), \quad \Delta b := t\,U_2\beta \in \mathcal{R}(A)^\perp, \quad \text{and} \quad b := b_0 + \Delta b\,, \tag{7.11}$$

where we have used that $\mathcal{R}(A) = \mathcal{R}(X)$ and $t \geq 0$ is a parameter. Observe that in this way

$$\rho_r = \frac{\|b - Ax_0\|_2}{\|b\|_2} = \frac{t}{\sqrt{1+t^2}}, \tag{7.12}$$

because the solution x_0 obeys $Ax_0 = b_0$. We have used Cauchy matrices of sizes $m = 100 \times n = 20 : 20 : 60$ and Vandermonde matrices of sizes $m = 50 \times n = 5 : 5 : 25$. For each size we have changed the value of t to get relative residuals $\rho_r = 10^{[-16:2:-2]}$. For each size and for each value of ρ_r we have generated 10 matrices and 10 right-hand sides b, using the standard normal distribution for the parameters of the matrices and for the vectors α and β. Finally, for each value of ρ_r, we get the maximum value of all relative forward errors, for all sizes and all random tests. The results are displayed in Table 7.1 for Vandermonde matrices. Similar results are obtained for Cauchy matrices. It can be observed that the analysis in Section 4.1 holds independently of the size of the relative residual, even for very small ones. We stress again that the really important point on the bound (1.2) is that $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ is small for most right-hand sides b for any fixed size of the relative residual not too close to one, independently of the ill-conditioning of the matrix A.

$\log_{10}(\|b - Ax_0\|_2/\|b\|_2)$	-16	-14	-12	-10	-8	-6	-4	-2
QR : $\log_{10}(\|\Delta x\|_2/\|x_0\|_2)$	-2.7	-2.8	-2.1	-2.5	-3.8	-3.5	-3.3	-2.6
RRD : $\log_{10}(\|\Delta x\|_2/\|x_0\|_2)$	-14.1	-14.0	-13.8	-13.9	-14.1	-14.0	-13.8	-13.8

TABLE 7.1

Experiments controlling the residual. The forward relative error $\|\Delta x\|_2/\|x_0\|_2$ is displayed for both algorithms LS-QR *and* LS-RRD *described in Subsection 7.2 for different values of the relative residual. This experiment was done with Vandermonde matrices of sizes $m = 50 \times n = 5 : 5 : 25$. All the necessary random vectors were chosen from the standard normal distribution.*

7.5. Tests with $\|A^\dagger\|_2 \, \|b\|_2 \, / \, \|x_0\|_2$ not small. In the all random tests presented in Subsections 7.1, 7.2, 7.3, and 7.4 the factor $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ has been moderate and therefore Algorithm 6.1 has solved accurately all tested LS problems. So, these tests have confirmed the discussion in Subsection 4.1. Of course, it is possible to prepare tests where $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ is not small and Algorithm 6.1 is not accurate, but this requires to select very carefully the vectors b. We have proceed as follows: the right-hand side has been prepared to be $b = u_1 + b_\perp$, where u_1 is the left singular vector of A corresponding to its largest singular value and $b_\perp$ is any random vector orthogonal to $\mathcal{R}(A)$. Note that for vectors of this type $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2 = \kappa_2(A) \sqrt{1 + \|b_\perp\|_2^2}$ and, as a consequence, the forward relative errors in the solutions committed by Algorithm 6.1 have been big and proportional to the unit roundoff times the condition numbers of the matrices. However, despite of this fact, the errors of Algorithm 6.1 have been much smaller than those committed by the standard Householder-QR algorithm. The reason is that the error (1.1) for Householder-QR includes the term Φ defined in (4.14) and, for vectors $b = u_1 + b_\perp$, $\Phi = \kappa_2(A)^2 \|b_\perp\|_2$, which is really huge if $\kappa_2(A)$ is large and $\|b_\perp\|_2 = \|r\|_2$ is not too small. Recall in this context our discussion of (4.15). As it was pointed out in [18], notice that for ill-conditioned matrices the vectors $b = u_1 + b_\perp$ have to be generated using highly accurate algorithms for getting u_1 and $b_\perp$ (see [9] and Subsection 7.4). If the vector b is prepared using usual floating point arithmetic and the `svd` command in MATLAB™, the rounding errors make it impossible for b to have exactly the required structure and Algorithm 6.1 computes solutions with high relative accuracy.

8. Conclusions and future work. In this paper we have introduced, and carefully analyzed, a new algorithm to compute accurate solutions of those least squares problems $\min_{x\in\mathbb{C}^n} \|Ax - b\|_2$ such that an accurate rank-revealing decomposition of the coefficient matrix A can be computed. This is nowadays possible for many classes of structured matrices that may have extremely large traditional condition numbers, as it was explained in the Introduction, and, probably, it will be possible for more classes in the future. In addition, the new algorithm can be also applied to compute accurate minimum 2-norm solutions of underdetermined linear systems. This work together with the previous papers [9, 17, 18] show that, for those matrices for which accurate rank-revealing decompositions can be computed, we can perform accurately and efficiently almost all basic tasks of Numerical Linear Algebra, i.e., solution of linear systems, solution of least squares problems, computation of eigenvalues and eigenvectors of symmetric matrices, and computation of the singular value decomposition, and to obtain relative errors of order $\mathbf{u}$ for very ill-conditioned problems where standard algorithms fail to provide even a single correct digit of accuracy. The only basic problem that is excluded from this framework is the nonsymmetric eigenvalue problem. To investigate at which extent rank-revealing decompositions allow us to solve accurately nonsymmetric eigenvalue problems will be the subject of our future research.

Acknowledgements. The authors thank Nick Higham for providing the codes of the algorithms and numerical tests presented in [27]. The authors also thank Siegfried Rump for illuminating discussions on why $\|A^\dagger\|_2 \, \|b\|_2/\|x_0\|_2$ is usually small for all vectors b which form the same acute angle with the column space of A. Froilán M. Dopico wishes to especially thank Rodolfo Cuerno and Rocío Ortega for all the encouragement provided during the completion of this paper. Froilán finishes this paper *"with a big help from his Friends"*.

REFERENCES

[1] J. M. Banoczi, N.-C. Chiu, G. E. Cho, and I. C. F. Ipsen, *The lack of influence of the right-hand side on the accuracy of linear system solution*, SIAM J. Sci. Comput., 20 (1998), pp. 203–227.

[2] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Num. Anal., 27 (1990), pp. 762–791.

[3] Å. Björck, *Numerical methods for least squares problems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[4] Å. Björck and V. Pereyra, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[5] L.-X. Cai, W.-W. Xu, and W. Li, *Additive and multiplicative perturbation bounds for the Moore-Penrose inverse*, Linear Algebra Appl., 434 (2011), pp. 480–489.

[6] S. L. Campbell and C. D. Meyer, Jr., *Generalized inverses of linear transformations*, Dover Publications Inc., New York, 1991. Corrected reprint of the 1979 original.

[7] T. F. Chan and D. E. Foulser, *Effectively well-conditioned linear systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 963–969.

[8] J. Demmel, *Accurate singular value decompositions of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–580.

[9] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[10] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[11] J. Demmel and P. Koev, *Accurate SVDs of weakly diagonally dominant M-matrices*, Numer. Math., 98 (2004), pp. 99–104.

[12] ———, *Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials*, Linear Algebra Appl., 417 (2006), pp. 382–396.

[13] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1246.

[14] J. W. Demmel, *Applied Numerical Linear Algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[15] F. M. Dopico and P. Koev, *Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1126–1156.

[16] ———, *Perturbation theory for the LDU factorization and accurate computations for diagonally dominant matrices*, Numer. Math., 119 (2011), pp. 337–371.

[17] F. M. Dopico, P. Koev, and J. M. Molera, *Implicit standard Jacobi gives high relative accuracy*, Numer. Math., 113 (2009), pp. 519–553.

[18] F. M. Dopico and J. M. Molera, *Accurate solution of structured linear systems via rank-revealing decompositions*, IMA J. Numer. Anal., 32 (2012), pp. 1096–1116.

[19] F. M. Dopico, J. M. Molera, and J. Moro, *An orthogonal high relative accuracy algorithm for the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 301–351.

[20] Z. Drmač and K. Veselić, *New fast and accurate Jacobi SVD algorithm. I*, SIAM Journal on Matrix Analysis and Applications, 29 (2008), pp. 1322–1342.

[21] ———, *New fast and accurate Jacobi SVD algorithm. II*, SIAM Journal on Matrix Analysis and Applications, 29 (2008), pp. 1343–1362.

[22] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[23] K. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

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

[25] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.

[26] N. J. Higham, *The Test Matrix Toolbox for Matlab (version 3.0)*, Numerical Analysis Report No. 276, Manchester Center for Computational Mathematics, Manchester, England, (1995).

[27] ———, *QR factorization with complete pivoting and accurate computation of the SVD*, Linear Algebra Appl., 309 (2000), pp. 153–174.

[28] ———, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002.

[29] I. C. F. Ipsen, *Relative perturbation results for matrix eigenvalues and singular values*, in Acta numerica, 1998, vol. 7 of Acta Numer., Cambridge Univ. Press, Cambridge, 1998, pp. 151–201.

[30] ———, *An overview of relative* $\sin\Theta$ *theorems for invariant subspaces of complex matrices*, J. Comput. Appl. Math., 123 (2000), pp. 131–153. Numerical analysis 2000, Vol. III. Linear algebra.

[31] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 1–23.

[32] D. Kressner, *Numerical methods for general and structured eigenvalue problems*, vol. 46 of Lecture Notes in Computational Science and Engineering, Springer-Verlag, Berlin, 2005.

[33] R.-C. Li, *Relative perturbation theory. I. Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.

[34] ———, *Relative perturbation theory. II. Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 471–492.

[35] A. Marco and J.-J. Martínez, *Polynomial least squares fitting in the Bernstein basis*, Linear Algebra Appl., 433 (2010), pp. 1254–1264.

[36] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.

[37] L. Miranian and M. Gu, *Strong rank revealing LU factorizations*, Linear Algebra Appl., 367 (2003), pp. 1–16.

[38] V. Olshevsky, ed., *Structured matrices in mathematics, computer science, and engineering. I*, vol. 280 of Contemporary Mathematics, Providence, RI, 2001, American Mathematical Society.

[39] ———, ed., *Structured matrices in mathematics, computer science, and engineering. II*, vol. 281 of Contemporary Mathematics, Providence, RI, 2001, American Mathematical Society.

[40] C.-T. Pan, *On the existence and computation of rank-revealing LU factorizations*, Linear Algebra Appl., 316 (2000), pp. 199–222.

[41] J. M. Peña, *LDU decompositions with L and U well conditioned*, Electron. Trans. Numer. Anal., 18 (2004), pp. 198–208 (electronic).

[42] S. M. Rump, *Inversion of extremely ill-conditioned matrices in floating-point*, Japan J. Indust. Appl. Math., 26 (2009), pp. 249–277.

[43] I. Slapničar, *Highly accurate symmetric eigenvalue decomposition and hyperbolic SVD*, Linear Algebra Appl., 358 (2003), pp. 387–424.

[44] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[45] D. S. Watkins, *The matrix eigenvalue problem: GR and Krylov subspace methods*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[46] P.-Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT, 13 (1973), pp. 217–232.

[47] Q. Ye, *Computing singular values of diagonally dominant matrices to high relative accuracy*, Math. Comp., 77 (2008), pp. 2195–2230.