IMA Journal of Numerical Analysis (2011) Page 1 of 21
doi:10.1093/imanum/drnxxx

Accurate solution of structured linear systems via rank-revealing decompositions

FROILÁN M. DOPICO†
Instituto de Ciencias Matemáticas CSIC-UAM-UC3M-UCM and Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Spain

AND

JUAN M. MOLERA‡
Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Spain

[Received on xxxxxxxx; revised on xxxxxxx]

Linear systems of equations $Ax = b$, where the matrix A has some particular structure, arise frequently in applications. Very often structured matrices have huge condition numbers $\kappa(A) = \|A^{-1}\|\,\|A\|$ and, therefore, standard algorithms fail to compute accurate solutions of $Ax = b$. We say in this paper that a computed solution $\widehat{x}$ is accurate if $\|\widehat{x}-x\|/\|x\| = O(\mathtt{u})$, being $\mathtt{u}$ the unit roundoff. In this work, we introduce a framework that allows to solve accurately many classes of structured linear systems, independently of the condition number of A and efficiently, that is, with cost $O(n^3)$. For most of these classes no algorithms are known that are both accurate and efficient. The approach in this work relies on computing first an accurate rank-revealing decomposition of A, an idea that has been widely used in the last decades to compute singular value and eigenvalue decompositions of structured matrices with high relative accuracy. In particular, we illustrate the new method solving accurately Cauchy and Vandermonde linear systems with any distribution of nodes, i.e., without requiring A to be totally positive, for most right-hand sides b.

Keywords: accurate solutions, acyclic matrices, Cauchy matrices, diagonally dominant matrices, graded matrices, linear systems, polynomial Vandermonde matrices, rank-revealing decompositions, structured matrices, Vandermonde matrices

1. Introduction

Structured matrices arise very frequently in applications -see for example Olshevsky (2001a,b). As a consequence, the design and analysis of special algorithms for solving linear systems $Ax = b$, $A \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$, whose matrix coefficient A has some particular structure is a classical area of research in Numerical Linear Algebra that has attracted the attention of many researchers -see (Golub & Van Loan, 1996, Chapter 4) and (Higham, 2002, Chapters 8,10,11,22) and the references therein. The goal of these

†Email: dopico@math.uc3m.es
‡Corresponding author. Email: molera@math.uc3m.es

special algorithms is to exploit the structure to increase the speed of solution, and/or to decrease storage requirements, and/or to improve the accuracy in comparison with standard algorithms for general matrices, as for instance Gaussian elimination with partial pivoting (GEPP) or the use of the QR factorization (see in (Higham, 2002) the accuracy properties of these methods). In this work, we establish a framework that allows us to solve many classes of structured linear systems with much more accuracy than the one provided by standard algorithms and roughly with the same computational cost, that is, with cost $O(n^3) \leqslant c\,n^3$, where c denotes a small real constant. This framework enlarges significantly the number of classes of structured matrices for which it is possible to compute accurately and efficiently solutions of linear systems.

Given a vector norm $\|\cdot\|$ in $\mathbb{C}^n$ and its subordinate matrix norm in the set $\mathbb{C}^{n\times n}$ of complex $n\times n$ matrices, it is well known that if $\widehat{x}$ is the solution of $Ax = b$ computed by GEPP or QR in a computer with unit roundoff $\mathtt{u}$ ($\mathtt{u} \approx 10^{-16}$ in double precision IEEE arithmetic), then the relative error satisfies

$$\frac{\|x-\widehat{x}\|}{\|x\|} \leqslant \mathtt{u}\, g(n)\, \kappa(A), \tag{1.1}$$

where $\kappa(A) = \|A^{-1}\|\,\|A\|$ is the condition number* of A and $g(n)$ is a modestly growing function of n, that is, a function bounded by a low degree polynomial in n. The bound (1.1) does not guarantee a single digit of accuracy if $\kappa(A) \gtrsim 1/\mathtt{u}$, 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. Two very famous examples are Cauchy and Vandermonde matrices (Higham, 2002, Chapters 22, 28). Our goal is to prove that the numerical framework we propose provides for many classes of structured matrices error bounds where $\kappa(A)$ is replaced by a much smaller quantity.

The framework we introduce relies on the concept of *rank-revealing decomposition* (RRD), originally introduced by Demmel *et al.* (1999) for computing the singular value decomposition (SVD) with high relative accuracy -see also (Higham, 2002, Sec 9.12). An RRD of $A \in \mathbb{C}^{n\times n}$ is a factorization $A = XDY$, where $X \in \mathbb{C}^{n\times n}$, $D = \mathrm{diag}(d_1, d_2, \ldots, d_n) \in \mathbb{C}^{n\times n}$ is diagonal, and $Y \in \mathbb{C}^{n\times n}$, and X and Y are well conditioned. Note that this means that if A is ill conditioned, then the diagonal factor D is also ill conditioned. We propose to compute the solution of $Ax = b$ in two steps:

1. First, compute an RRD of $A = XDY$ accurately in the sense of Demmel *et al.* (1999) (we revise the precise meaning of "accuracy" in this context in Definition 2.1).

2. Second, solve the three linear systems $Xs = b$, $Dw = s$, and $Yx = w$, where $Xs = b$ and $Yx = w$ are solved by standard backward-stable methods as GEPP or QR, while $Dw = s$ is solved as $w_i = s_i/d_i$, $i = 1:n$.

The intuition behind why this procedure computes accurate solutions, even for extremely ill conditioned matrices, is that each entry of w is computed with a relative error less than $\mathtt{u}$, that is, the ill conditioned linear system $Dw = s$ is solved very accurately, together with the fact that $Xs = b$ and $Yx = w$ are also solved accurately because X and Y are well conditioned. We will prove in Section 4 that the relative error in the solution $\widehat{x}$ computed by this two step procedure is

$$\frac{\|x-\widehat{x}\|}{\|x\|} \leqslant \mathtt{u}\, f(n)\, \max\{\kappa(X), \kappa(Y)\}\, \frac{\|A^{-1}\|\,\|b\|}{\|x\|}, \tag{1.2}$$

*We assume throughout this work that A is nonsingular, i.e., that the solution of the linear system $Ax = b$ is unique for any b.

where $f(n)$ is a modestly growing function of n. Note that the only potentially big factor in (1.2) is $\|A^{-1}\|\,\|b\|/\|x\|$, because X and Y are well conditioned. It is well known that this factor is small for most right-hand sides b, even for very ill conditioned matrices A (Chan & Foulser, 1988; Banoczi *et al.*, 1998). We will explain this fact carefully in Subsection 3.2, where we revise some properties of $\|A^{-1}\|\,\|b\|/\|x\|$.

The computation of an *accurate* RRD of A is the difficult part in this framework. For almost any matrix $A \in \mathbb{C}^{n\times n}$ an RRD (potentially inaccurate) can be computed by applying standard Gaussian elimination with complete pivoting (GECP) to get an LDU factorization, where L is unit lower triangular, D is diagonal, and U is unit upper triangular (Demmel *et al.*, 1999; Higham, 2002). Very rarely GECP fails to produce well conditioned L and U factors, but then other pivoting strategies that guarantee well conditioned factors are available, for instance in Miranian & Gu (2003) and Pan (2000). However standard GECP is not 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.

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 $n \times n$ structured matrices in $O(n^3)$ operations. These classes include: Cauchy matrices, diagonally scaled Cauchy matrices, Vandermonde matrices, and some "related unit-displacement-rank" matrices (Demmel, 1999); graded matrices (that is, matrices of the form D_1BD_2 with B well conditioned and D_1 and D_2 diagonal), acyclic matrices (which include bidiagonal matrices), total signed compound matrices, diagonally scaled totally unimodular matrices (Demmel *et al.*, 1999); diagonally dominant M-matrices (Demmel & Koev, 2004; Peña, 2004); polynomial Vandermonde matrices involving orthonormal polynomials (Demmel & Koev, 2006); and diagonally dominant matrices (Ye, 2008; Dopico & Koev, 2010). 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 DHD, with H well conditioned and D diagonal (Demmel & Veselić, 1992; Mathias, 1995); symmetric Cauchy, symmetric diagonally scaled Cauchy, symmetric Vandermonde (Dopico & Koev, 2006); and symmetric diagonally scaled totally unimodular and total signed compound matrices (Peláez & Moro, 2006). These symmetric RRDs have been used to compute accurate eigenvalues and eigenvectors of symmetric matrices (Dopico *et al.*, 2003, 2009). For all classes of matrices listed in this paragraph, the framework introduced in this paper solves linear systems with relative errors bounded as in (1.2). This error bound is $O(\mathrm{u}\,f(n))$ for most right-hand sides independently of the traditional condition number of the matrices and so guarantees accurate solutions.

We want to remark that for Cauchy and Vandermonde matrices A, the use of the RRDs computed by the algorithms in (Demmel, 1999) allows us to solve accurately $Ax = b$ *for any distribution of the nodes* defining Cauchy and Vandermonde matrices and for most right-hand sides b. The situation is different for currently available fast algorithms with $O(n^2)$ cost for Vandermonde and Cauchy linear systems. The analyses (Boros *et al.*, 1999; Higham, 2002) show that they deliver componentwise bounds for errors (forward, backward and the residual), though these error bounds are small, for most right-hand sides, only for those distributions of nodes corresponding to totally positive (TP) matrices. In conclusion, fast methods compute solutions with guaranteed accuracy only in the TP case. This is further shown in the numerical experiments in this paper. It remains as an open problem to find fast algorithms for Cauchy and Vandermonde matrices that guarantee accurate solutions for any distribution of nodes.

The paper is organized as follows. We introduce in Section 2 basic notations and concepts that will be used throughout the paper. Section 3 develops a structured perturbation theory for linear systems through the factors of RRDs. These perturbation results are used in Section 4 to perform a general

error analysis of the framework proposed in this work. Numerical experiments that show the significant improvements obtained in accuracy from the use of the new method are presented in Section 5. For brevity, we restrict only to Cauchy and Vandermonde matrices in the numerical tests. Finally, Section 6 contains conclusions and establishes some lines of future research.

2. Preliminaries

We present in this section basic facts on vector and matrix norms, the definition of the entrywise absolute value of a matrix, the exact meaning of accurate rank-revealing decomposition in this work, and the model of floating point arithmetic that we use in the error analysis of Section 4.

We denote throughout the paper by $\|\cdot\|$ any vector norm in $\mathbb{C}^n$ and also the corresponding subordinate matrix norm, defined as

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

for any $A \in \mathbb{C}^{n\times n}$. We will use the *dual norm* of the vector norm $\|\cdot\|$, which is defined as

$$\|x\|_D = \max_{z\neq 0} \frac{|z^* x|}{\|z\|} \tag{2.1}$$

for any $x \in \mathbb{C}^n$. As usual z^* denotes the conjugate-transpose of $z \in \mathbb{C}^n$. Given a vector $y \in \mathbb{C}^n$, we will need the vector z dual to y, which is defined by the property

$$z^* y = \|z\|_D \, \|y\| = 1. \tag{2.2}$$

See (Higham, 2002, Chapter 6) or (Horn & Johnson, 1985, Chapter 5) for more information on these concepts.

In Sections 3.1 and 4, we will need the entrywise absolute value of a matrix. Given a matrix $G \in \mathbb{C}^{n\times n}$ with entries g_{ij}, we denote by $|G|$ the matrix with entries $|g_{ij}|$. Expressions like $|G| \leqslant |B|$, where $B \in \mathbb{C}^{n\times n}$, mean $|g_{ij}| \leqslant |b_{ij}|$ for $1 \leqslant i, j \leqslant n$.

Following Demmel *et al.* (1999), next we define the precise meaning of an *accurate* computed RRD of a matrix A. We consider only nonsingular matrices A, since these are the ones of interest in this work.

DEFINITION 2.1 Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix, let $A = XDY$ with $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ be an RRD of A, and let $\widehat{X}$, $\widehat{D} = \mathrm{diag}(\widehat{d_1}, \widehat{d_2}, \ldots, \widehat{d_n})$, and $\widehat{Y}$ 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 if the factors satisfy

$$\frac{\|\widehat{X} - X\|}{\|X\|} \leqslant \mathtt{u}\, p(n), \quad \frac{\|\widehat{Y} - Y\|}{\|Y\|} \leqslant \mathtt{u}\, p(n), \quad \text{and} \quad \frac{|\widehat{d_i} - d_i|}{|d_i|} \leqslant \mathtt{u}\, p(n), \quad i = 1:n, \tag{2.3}$$

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

For example, the algorithm to compute an RRD of a Cauchy matrix presented in (Demmel, 1999, Section 4), satisfies (2.3) with $p(n) \leqslant 9n$, in norm $\|\,.\,\|_1$.

In the rounding error analysis of Section 4 we will use the conventional error model for floating point arithmetic (Higham, 2002, 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 $|\delta| \leqslant \mathtt{u}$. Recall that this model also holds for complex floating point numbers if $\mathtt{u}$ is replaced by a slightly larger constant, see (Higham, 2002, Section 3.6). In addition, we will assume that neither overflow nor underflow occurs.

3. Structured perturbation theory for linear systems

In this section we present structured perturbation results for the solution of linear systems that will be used in the error analysis of Section 4. We deal with perturbations of the coefficient matrix coming from perturbing the factors of an RRD according to the errors in (2.3). A crucial point here is the technical Lemma 3.1. This Lemma shows that the results of this section are closely connected to multiplicative perturbations of the coefficient matrix of the system. Multiplicative perturbation theory of matrices has received considerable attention in the literature in the context of accurate computations of eigenvalues and singular values (Eisenstat & Ipsen, 1995; Ipsen, 1998, 2000; Li, 1998, 1999), but, as far as we know, it has not been studied yet for linear systems.

We stress that the most relevant factor in all perturbation bounds in this section is $\|A^{-1}\|\,\|b\|/\|x\|$. We will revise briefly some results existing in the literature for this factor in Subsection 3.2 which justify that this factor has a moderate magnitude for most vectors b even for very ill conditioned matrices A.

We are concerned in this section with a linear system $Ax = b$, where $A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$, and A *is nonsingular*. For brevity, we will not repeat these assumptions in the statements of the results. As usual, I denotes the $n \times n$ identity matrix.

3.1 *Perturbation theory through factors*

We start with Lemma 3.1 that is the base of our analysis.

LEMMA 3.1 Let $Ax = b$ and $(I+E)A(I+F)\widetilde{x} = b+h$, where $I+E \in \mathbb{C}^{n\times n}$ and $I+F \in \mathbb{C}^{n\times n}$ are nonsingular matrices. Then

$$\widetilde{x} - x = (I+F)^{-1}\left(A^{-1}(I+E)^{-1}(h-Eb) - Fx\right). \tag{3.1}$$

In addition, if $\|h\| \leqslant \varepsilon\|b\|$ and $\max\{\|E\|,\|F\|\} \leqslant \varepsilon < 1$, then

$$\widetilde{x} - x = A^{-1}h - A^{-1}Eb - Fx + O(\varepsilon^2). \tag{3.2}$$

Proof. Define $f \in \mathbb{C}^n$ as $x+f := (I+F)\widetilde{x}$. Then $(I+E)A(x+f) = b+h$ and $(I+E)(b+Af) = b+h$. Therefore, $(I+E)Af = h-Eb$ and

$$f = A^{-1}(I+E)^{-1}(h-Eb). \tag{3.3}$$

Use the definition of f to get

$$\widetilde{x} - x = (I+F)^{-1}(x+f) - x = (I+F)^{-1}(f-Fx),$$

and introduce (3.3) in this expression to prove (3.1). To prove (3.2) note that if $\varepsilon < 1$, then $(I+E)^{-1} = \sum_{k=0}^{\infty}(-E)^k = I + O(\varepsilon)$ (Horn & Johnson, 1985, Corollary 5.6.16) and, analogously, $(I+F)^{-1} = I + O(\varepsilon)$. This allows us to obtain (3.2) from (3.1). □

Our next result measures the sensitivity of the system for perturbations of the factors of an RRD of the coefficient matrix.

THEOREM 3.1 Let $A = XDY \in \mathbb{C}^{n\times n}$, where $D \in \mathbb{C}^{n\times n}$ is a diagonal matrix, and let $\|\cdot\|$ be a vector norm in $\mathbb{C}^n$ whose subordinate matrix norm satisfies $\|\Lambda\| = \max_{i=1:n}|\lambda_i|$ for all diagonal matrices $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Let $XDY\,x = b$ and $(X+\Delta X)(D+\Delta D)(Y+\Delta Y)\widetilde{x} = b+h$, where $\|\Delta X\| \leqslant \varepsilon\,\|X\|$, $\|\Delta Y\| \leqslant \varepsilon\,\|Y\|$, $|\Delta D| \leqslant \varepsilon\,|D|$ and $\|h\| \leqslant \varepsilon\,\|b\|$, and assume that $\varepsilon\kappa(Y) < 1$ and $\varepsilon(2+\varepsilon)\kappa(X) < 1$. Then,

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leqslant \frac{\varepsilon}{1-\varepsilon\kappa(Y)}\left(\kappa(Y) + \frac{1+(2+\varepsilon)\|X\|\frac{\|X^{-1}b\|}{\|b\|}}{1-\varepsilon(2+\varepsilon)\kappa(X)}\;\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) \tag{3.4}$$

and, to first order in ε,

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leqslant \varepsilon\left(\kappa(Y) + \left(1+2\,\|X\|\frac{\|X^{-1}b\|}{\|b\|}\right)\;\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) + O(\varepsilon^2). \tag{3.5}$$

Proof. The proof consists in transforming the perturbation of the factors into a multiplicative perturbation and then to use Lemma 3.1. The matrix $(X+\Delta X)(D+\Delta D)(Y+\Delta Y)$ can be written as

$$\begin{aligned}(X+\Delta X)(D+\Delta D)(Y+\Delta Y) &= (I+\Delta X\,X^{-1})X(I+\Delta D\,D^{-1})DY(I+Y^{-1}\Delta Y)\\ &= (I+\Delta X\,X^{-1})(I+X\Delta D\,D^{-1}X^{-1})XDY(I+Y^{-1}\Delta Y),\\ &=: (I+E)XDY(I+F),\end{aligned}$$

where

$$E = \Delta X\,X^{-1} + X\Delta D\,D^{-1}X^{-1} + \Delta X\,\Delta D\,D^{-1}X^{-1}, \tag{3.6}$$

$$F = Y^{-1}\Delta Y. \tag{3.7}$$

Observe also that

$$\|E\| \leqslant \varepsilon(2+\varepsilon)\kappa(X) < 1, \quad \|F\| \leqslant \varepsilon\kappa(Y) < 1, \quad \text{and,} \quad \|Eb\| \leqslant \varepsilon(2+\varepsilon)\,\|X\|\,\|X^{-1}b\|, \tag{3.8}$$

that guarantee in particular that $I+E$ and $I+F$ are nonsingular. Now use (3.1) for the systems $XDYx = b$ and $(I+E)XDY(I+F)\widetilde{x} = b+h$, apply standard norm inequalities to get

$$\|\widetilde{x}-x\| \leqslant \|(I+F)^{-1}\|\,\left(\|A^{-1}\|\,\|(I+E)^{-1}\|\,(\|h\|+\|Eb\|) + \|F\|\|x\|\right),$$

use

$$\|(I+F)^{-1}\| \leqslant \frac{1}{1-\|F\|} \leqslant \frac{1}{1-\varepsilon\kappa(Y)}, \quad \text{and} \quad \|(I+E)^{-1}\| \leqslant \frac{1}{1-\|E\|} \leqslant \frac{1}{1-\varepsilon(2+\varepsilon)\kappa(X)},$$

and (3.8) to obtain (3.4). Finally, the bound (3.5) to first order follows immediately from (3.4). □

REMARK 3.1 We will see in Theorem 3.2 that the bounds (3.4) and (3.5) are "essentially" the best that can be obtained under the assumptions on Theorem 3.1. However the following weaker bounds may be more useful in practice

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leqslant \frac{\varepsilon}{1-\varepsilon\kappa(Y)}\left(\kappa(Y) + \frac{1+(2+\varepsilon)\,\kappa(X)}{1-\varepsilon(2+\varepsilon)\kappa(X)}\;\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right), \tag{3.9}$$

$$\frac{\|\widetilde{x}-x\|}{\|x\|} \leqslant \varepsilon\left(\kappa(Y) + (1+2\,\kappa(X))\;\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) + O(\varepsilon^2), \tag{3.10}$$

because they do not require to compute $\|X^{-1}b\|$ and do not overestimate significantly the variation of the solution if $\kappa(X)$ is small.

We have not been able to prove that the bound (3.5) can be attained to first order in ε for some specific perturbations h, ΔX, ΔD, and ΔY and, as a consequence, we have not found an expression of the condition number of linear equation solving for perturbations of the factors of an RRD of the coefficient matrix. In fact, we think that (3.5) cannot be attained since it is independent of D. However, we show in Theorem 3.2 that what we get, after dividing by ε the right-hand side of (3.5) and taking the limit $\varepsilon \to 0$, is smaller than three times the condition number, and, so, (3.5) cannot overestimate significantly the best possible first order perturbation bound. The condition number for perturbation of the factors is denoted by $\kappa_f(X,D,Y,b)$, where the subindex f stands for "factors".

THEOREM 3.2 Let us use the same notation and assumptions as in Theorem 3.1 and define the condition number

$$\kappa_f(X,D,Y,b) := \lim_{\varepsilon\to 0} \sup \left\{ \frac{\|\widetilde{x}-x\|}{\varepsilon\|x\|} \;:\; (X+\Delta X)\,(D+\Delta D)\,(Y+\Delta Y)\,\widetilde{x} = b+h, \right.$$
$$\left. \|h\| \leqslant \varepsilon\|b\|,\ \|\Delta X\| \leqslant \varepsilon\,\|X\|,\ |\Delta D| \leqslant \varepsilon\,|D|,\ \|\Delta Y\| \leqslant \varepsilon\,\|Y\| \right\}.$$

Then

$$\kappa_f(X,D,Y,b) \leqslant \left(\kappa(Y) + \left(1+2\,\|X\|\frac{\|X^{-1}b\|}{\|b\|}\right)\,\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) \leqslant 3\,\kappa_f(X,D,Y,b). \tag{3.11}$$

Proof. The lower bound in (3.11), that is, the part "$\kappa_f(X,D,Y,b) \leqslant \cdots$", follows immediately from (3.5) and the definition of the condition number. Therefore, we focus on proving the upper bound in (3.11).

We prove first that $\kappa(Y) \leqslant \kappa_f(X,D,Y,b)$. For this purpose choose a perturbation such that $h = 0$, $\Delta D = 0$, $\Delta X = 0$, and

$$\Delta Y \;=\; \varepsilon\,\|x\|\,\|Y\|\,zy^*,$$

where y is a vector dual to x, i.e., $y^*x = \|y\|_D\|x\| = 1$, $\|z\| = 1$, and $\|Y^{-1}z\| = \|Y^{-1}\|$. Note that $\|\Delta Y\| = \varepsilon\|Y\|$, and that the matrices defined in (3.6) and (3.7) are in this case $E = 0$ and $F = Y^{-1}\,\Delta Y$. From (3.2), we obtain that for this perturbation

$$\widetilde{x} - x = -Y^{-1}\,\Delta Y\,x + O(\varepsilon^2) = -\,\varepsilon\,\|x\|\,\|Y\|\,(Y^{-1}z)\,(y^*x) + O(\varepsilon^2) = -\,\varepsilon\,\|x\|\,\|Y\|\,(Y^{-1}z) + O(\varepsilon^2).$$

So, for this perturbation,

$$\lim_{\varepsilon\to 0}\frac{\|\widetilde{x}-x\|}{\varepsilon\|x\|} = \kappa(Y) \leqslant \kappa_f(X,D,Y,b), \tag{3.12}$$

where the last inequality follows from the "sup" appearing in the definition of $\kappa_f(X,D,Y,b)$.

Next we prove that

$$\|X\|\frac{\|X^{-1}b\|}{\|b\|}\,\frac{\|A^{-1}\|\,\|b\|}{\|x\|} \leqslant \left(1+\|X\|\frac{\|X^{-1}b\|}{\|b\|}\right)\,\frac{\|A^{-1}\|\,\|b\|}{\|x\|} \leqslant \kappa_f(X,D,Y,b). \tag{3.13}$$

For this purpose choose a perturbation such that $\Delta D = 0$, $\Delta Y = 0$, $h = \varepsilon\,\|b\|\,w$, where $\|w\| = 1$ and $\|A^{-1}w\| = \|A^{-1}\|$, and, finally,

$$\Delta X \;=\; -\varepsilon\,\|X^{-1}b\|\,\|X\|\,ws^*,$$

where s is a vector dual to $X^{-1}b$. Note that $\|h\| = \varepsilon\|b\|$, $\|\Delta X\| = \varepsilon\|X\|$, and that the matrices defined in (3.6) and (3.7) are in this case $E = \Delta X\, X^{-1}$ and $F = 0$. From (3.2), we obtain that for this perturbation

$$\begin{aligned}\widetilde{x} - x &= \varepsilon \left(\|b\| A^{-1}w + \|X^{-1}b\|\,\|X\|\,(A^{-1}w)\,(s^* X^{-1}b) \right) + O(\varepsilon^2) \\ &= \varepsilon \left(\|b\| + \|X^{-1}b\|\,\|X\| \right) A^{-1}w + O(\varepsilon^2).\end{aligned}$$

So, for this perturbation,

$$\lim_{\varepsilon\to 0} \frac{\|\widetilde{x}-x\|}{\varepsilon\|x\|} = \left(1 + \|X\|\frac{\|X^{-1}b\|}{\|b\|}\right) \frac{\|A^{-1}\|\,\|b\|}{\|x\|} \leqslant \kappa_f(X,D,Y,b),$$

where the last inequality follows again from the "sup" appearing in the definition of $\kappa_f(X,D,Y,b)$.

It only remains to combine (3.12) and (3.13) to obtain

$$\left(\kappa(Y) + \left(1 + 2\,\|X\|\frac{\|X^{-1}b\|}{\|b\|}\right) \frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right) \leqslant 3\,\kappa_f(X,D,Y,b).$$

□

3.2 *Why is* $\dfrac{\|A^{-1}\|\,\|b\|}{\|x\|}$ *usually small?*

Theorem 3.1, together with Theorem 3.2, prove that the sensitivity of the system $Ax = b$ to perturbations through factors is mainly governed by $\|A^{-1}\|\,\|b\|/\|x\|$, assuming that the factors X and Y of the RRD of A are well conditioned. The quantity $\|A^{-1}\|\,\|b\|\,/\,\|x\|$ is well known in Numerical Linear Algebra because it is the usual condition number for normwise perturbations when only the right-hand side of the linear system $Ax = b$ is perturbed. For instance, it is proved in (Higham, 2002, p. 121) that

$$\kappa(A,b) := \lim_{\varepsilon\to 0} \sup\left\{ \frac{\|\widetilde{x}-x\|}{\varepsilon\|x\|} : A\widetilde{x} = b + h,\ \ \|h\| \leqslant \varepsilon\,\|b\| \right\} = \frac{\|A^{-1}\|\,\|b\|}{\|x\|}. \tag{3.14}$$

Therefore we have proved in Theorem 3.1 that, loosely speaking, perturbations trough the factors of an RRD of $A = XDY$ have an effect on the solution similar to perturbing only the right-hand side b.

It is obvious that

$$1 \leqslant \kappa(A,b) \leqslant \kappa(A),$$

but the key point in this section is to show that if A is fixed and $\kappa(A) \gg 1$, then $\kappa(A,b) \ll \kappa(A)$ for most vectors b, that is, $\kappa(A,b)$ is usually a moderate number. This was put to light by Chan & Foulser (1988) -see also (Higham, 2002, Problem 7.5). The explanation of this fact is particularly simple for the Euclidean vector norm $\|\cdot\|_2$ and its subordinate matrix norm (Higham, 2002, p. 108), with the use of the SVD of A.

Let $A = U\Sigma V^*$ be the SVD of A, where $U, V \in \mathbb{C}^{n\times n}$ are unitary, $\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_n)$, and $\sigma_1 \geqslant \cdots \geqslant \sigma_n > 0$. Observe that the 2-norm of the solution of $Ax = b$ can be written as

$$\|x\|_2 = \|A^{-1}b\|_2 = \|\Sigma^{-1}U^*b\|_2 \geqslant \frac{|u_n^*b|}{\sigma_n},$$

and bound (3.14) simply as

$$\kappa_2(A,b) = \frac{\|b\|_2}{\sigma_n \|x\|_2} \leqslant \frac{\|b\|_2}{|u_n^* b|} = \frac{1}{\cos\theta(u_n,b)}, \tag{3.15}$$

being u_n the last column of U and $\theta(u_n,b)$ the acute angle between u_n and b. Observe that the bound on $\kappa_2(A,b)$ in (3.15) may be large only if b is "almost" orthogonal to u_n. For example, if A is an extremely ill conditioned fixed matrix (think that $\kappa(A) = 10^{1000}$ to be concrete) and the right-hand side b is considered as a random vector whose direction is uniformly distributed in the space, then the probability that $0 \leqslant \theta(u_n,b) \leqslant \pi/2 - 10^{-6}$ is approximately $1 - 10^{-6}$. Note that the condition $0 \leqslant \theta(u_n,b) \leqslant \pi/2 - 10^{-6}$ implies $\kappa_2(A,b) \lesssim 10^6$, which is a moderate number compared to 10^{1000}. In particular, if the perturbation parameter in Theorem 3.1 is $\varepsilon = 10^{-16}$, then $\|A^{-1}\|_2\|b\|_2/\|x\|_2 \lesssim 10^6$ provides a very good bound for the variation of the solution. Even more, it is possible that $\kappa_2(A,b)$ is moderate although $\cos\theta(u_n,b) \approx 0$. This is consequence of the following theorem.

THEOREM 3.3 (Chan & Foulser (1988)) Let $A = U\Sigma V^*$ be the SVD of $A \in \mathbb{C}^{n\times n}$ and P_k be the orthogonal projector onto the subspace spanned by the last k columns of U. Then

$$\kappa_2(A,b) \leqslant \frac{\sigma_{n+1-k}}{\sigma_n} \frac{\|b\|_2}{\|P_k b\|_2}, \quad \text{for} \quad k = 1:n. \tag{3.16}$$

If $\cos\theta(u_n,b) \approx 0$, then $\|P_2 b\|_2 \approx |u_{n-1}^* b|$, being u_{n-1} the next to last column of U,

$$\kappa_2(A,b) \lesssim \frac{\sigma_{n-1}}{\sigma_n} \frac{1}{\cos\theta(u_{n-1},b)}.$$

Whenever $\sigma_{n-1} \approx \sigma_n$, this bound will be moderate with high probability on random right-hand sides b such that $\cos\theta(u_n,b) \approx 0$. Observe that (3.16) can be interpreted as saying that the effective condition number for the problem $A\widetilde{x} = b + h$ is σ_{n+1-k}/σ_n, being k the smallest natural number for which $\|P_k b\|_2/\|b\|_2$ is not too small.

4. Algorithm and error analysis

We have seen in Remark 3.1 that the sensitivity of the solution of the system $Ax = b$, where $A = XDY$, for perturbations of the factors of A depends on the condition numbers of the factors X and Y, which are small if $A = XDY$ is an RRD, and also on $\|A^{-1}\|\,\|b\|/\|x\|$, which is moderate for most vectors b according to the discussion in Subsection 3.2. The purpose of this section is to prove that these quantities also determine the rounding errors of the solution of $Ax = b$ computed with the framework presented in the Introduction. Therefore, if an RRD of A can be computed accurately in the sense of Definition 2.1, then this framework computes accurate solutions of the system $Ax = b$ for most vectors b. We start by specifying carefully the method of solution, and then we present the error analysis.

Algorithm 4.1 **(Accurate solution of linear systems with rank-revealing decompositions)**
`Input:` $A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$
`Output:` x, solution of $Ax = b$

`Step 1:` Compute an accurate RRD of $A = XDY$, with $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, in the sense of Definition 2.1.

`Step 2`: Solve the three linear systems,

$$\begin{array}{rcl} X\,s = b & \longrightarrow & s, \\ D\,w = s & \longrightarrow & w, \\ Y\,x = w & \longrightarrow & x, \end{array}$$

where $X\,s = b$ and $Y\,x = w$ are solved by any backward stable (in norm) method as GEPP or the QR factorization, and $D\,w = s$ is solved as $w_i = s_i/d_i$, $i = 1:n$.

As mentioned in the Introduction, in most cases RRDs are computed trough variants of GECP, so X and Y are permutations of lower and upper triangular matrices respectively and $X\,s = b$ and $Y\,x = w$ are solved simply with forward and backward substitution in n^2 operations (Higham, 2002, Chapter 8). However, there are classes of matrices for which this does not happen, see for instance the algorithm for computing RRDs of Vandermonde matrices presented in (Demmel, 1999, Section 5), and we have decided to present Algorithm 4.1 and its error analysis in a more general context. Theorem 4.2 provides backward and forward errors for Algorithm 4.1. To simplify future references to Theorem 4.2, we include all necessary assumptions in the statement.

THEOREM 4.2 Let $\|\cdot\|$ be a vector norm in $\mathbb{C}^n$ whose subordinate matrix norm satisfies $\|\Lambda\| = \max_{i=1:n}|\lambda_i|$ for all diagonal matrices $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Let $\widehat{X}$, $\widehat{D}$, and $\widehat{Y}$ be the factors of A computed in `Step 1` of Algorithm 4.1 and assume that they satisfy

$$\frac{\|\widehat{X}-X\|}{\|X\|} \leqslant \mathtt{u}\,p(n), \quad \frac{\|\widehat{Y}-Y\|}{\|Y\|} \leqslant \mathtt{u}\,p(n), \quad \text{and} \quad |\widehat{D}-D| \leqslant \mathtt{u}\,p(n)\,|D|, \tag{4.1}$$

where $p(n)$ is a modestly growing function of n, X, D, and Y are the corresponding exact factors of A, and $\mathtt{u}$ is the unit roundoff. Assume also that the systems $X\,s = b$ and $Y\,x = w$ are solved with a backward stable algorithm that when applied to any linear system $Bz = c$, $B \in \mathbb{C}^{n\times n}$ and $c \in \mathbb{C}^n$, computes a solution $\widehat{z}$ that satisfies[†]

$$(B+\Delta B)\widehat{z} = c, \quad \text{with } \|\Delta B\| \leqslant \mathtt{u}\,q(n)\,\|B\|, \tag{4.2}$$

where $q(n)$ is a modestly growing function of n such that $q(n) \geqslant 4\sqrt{2}/(1-12\mathtt{u})$. Let

$$g(n) := p(n) + q(n) + \mathtt{u}\,p(n)q(n).$$

1. If $\widehat{x}$ is the computed solution of $A\,x = b$ using Algorithm 4.1, then

$$(X+\Delta X)(D+\Delta D)(Y+\Delta Y)\widehat{x} = b,$$

where $\|\Delta X\| \leqslant \mathtt{u}\,g(n)\,\|X\|$, $|\Delta D| \leqslant \mathtt{u}\,g(n)\,|D|$, and $\|\Delta Y\| \leqslant \mathtt{u}\,g(n)\,\|Y\|$.

2. In addition, if x is the exact solution of $Ax = b$, $(\,\mathtt{u}\,g(n)\,\kappa(Y)\,) < 1$ and $(\,\mathtt{u}\,g(n)\,(2+\mathtt{u}\,g(n))\,\kappa(X)\,) < 1$, then

$$\frac{\|\widehat{x}-x\|}{\|x\|} \leqslant \frac{\mathtt{u}\,g(n)}{1-\mathtt{u}\,g(n)\kappa(Y)}\left(\kappa(Y) + \frac{1+(2+\mathtt{u}\,g(n))\,\kappa(X)}{1-\mathtt{u}\,g(n)(2+\mathtt{u}\,g(n))\kappa(X)}\;\frac{\|A^{-1}\|\,\|b\|}{\|x\|}\right). \tag{4.3}$$

[†]We are assuming here that there are no backward errors in the right-hand side c. This happens for Gaussian elimination (Higham, 2002, p. 165) and also if the QR factorization is used for solving the system (Higham, 2002, eq. (19.14)). It is possible to consider backward errors in c at the cost of complicating somewhat the analysis.

REMARK 4.1 The error bound (4.3) simplifies considerably if we only pay attention to the first order term

$$\frac{\|\widehat{x}-x\|}{\|x\|} \leqslant \mathtt{u}\, g(n) \left(\kappa(Y) + (1+2\,\kappa(X)) \, \frac{\|A^{-1}\|\,\|b\|}{\|x\|} \right) + O\left((\mathtt{u}\, g(n))^2 \right), \tag{4.4}$$

$$\leqslant 4\mathtt{u}\, g(n) \max\{\kappa(X), \kappa(Y)\} \, \frac{\|A^{-1}\|\,\|b\|}{\|x\|} + O\left((\mathtt{u}\, g(n))^2 \right). \tag{4.5}$$

These bounds show clearly that the relative error in the solution is $O(\mathtt{u})$ if $\kappa(X)$, $\kappa(Y)$ and $\|A^{-1}\|\,\|b\|/\|x\|$ are moderate numbers.

Proof of Theorem 4.2. Part 2 follows directly from Part 1 and Remark 3.1. Therefore, we only need to prove Part 1. Observe that (4.1) implies

$$\widehat{X} = X + \Delta X_1, \quad \widehat{D} = D + \Delta D_1, \quad \widehat{Y} = Y + \Delta Y_1, \tag{4.6}$$

where $\|\Delta X_1\| \leqslant \mathtt{u}\, p(n)\, \|X\|$, $|\Delta D_1| \leqslant \mathtt{u}\, p(n)\, |D|$, and $\|\Delta Y_1\| \leqslant \mathtt{u}\, p(n)\, \|Y\|$. From (4.6), we obtain also

$$\|\widehat{X}\| \leqslant (1+\mathtt{u}\, p(n))\, \|X\|, \quad |\widehat{D}| \leqslant (1+\mathtt{u}\, p(n))\, |D|, \quad \|\widehat{Y}\| \leqslant (1+\mathtt{u}\, p(n))\, \|Y\|, \tag{4.7}$$

that will be used in the sequel.

Next, we use (4.2), Section 3.6 in (Higham, 2002) for the rounding error committed in a floating point complex division, and (4.7) to bound the backward errors when solving the linear systems in `Step 2` of Algorithm 4.1: (i) the computed solution, $\widehat{s}$, of $\widehat{X}s = b$ satisfies

$$(\widehat{X} + \Delta X_2)\, \widehat{s} = b, \quad \text{with} \quad \|\Delta X_2\| \leqslant \mathtt{u}\, q(n)\, \|\widehat{X}\| \leqslant \mathtt{u}\, q(n)\, (1+\mathtt{u}\, p(n))\, \|X\|; \tag{4.8}$$

(ii) the computed solution, $\widehat{w}$, of $\widehat{D}w = \widehat{s}$ satisfies

$$(\widehat{D} + \Delta D_2)\, \widehat{w} = \widehat{s}, \quad \text{with} \quad |\Delta D_2| \leqslant \mathtt{u} \frac{4\sqrt{2}}{1-12\,\mathtt{u}}\, |\widehat{D}| \leqslant \mathtt{u}\, q(n)\, (1+\mathtt{u}\, p(n))\, |D|; \tag{4.9}$$

(iii) the computed solution, $\widehat{x}$, of $\widehat{Y}x = \widehat{w}$ satisfies

$$(\widehat{Y} + \Delta Y_2)\, \widehat{x} = \widehat{w}, \quad \text{with} \quad \|\Delta Y_2\| \leqslant \mathtt{u}\, q(n)\, \|\widehat{Y}\| \leqslant \mathtt{u}\, q(n)\, (1+\mathtt{u}\, p(n))\, \|Y\|. \tag{4.10}$$

Putting together (4.8), (4.9) and (4.10), we found that the computed solution $\widehat{x}$ satisfies

$$(\widehat{X} + \Delta X_2)\, (\widehat{D} + \Delta D_2)(\widehat{Y} + \Delta Y_2)\, \widehat{x} = b,$$

or with (4.6)

$$(X + \Delta X_1 + \Delta X_2)(D + \Delta D_1 + \Delta D_2)(Y + \Delta Y_1 + \Delta Y_2)\, \widehat{x} = b,$$

that is

$$(X + \Delta X)\, (D + \Delta D)\, (Y + \Delta Y)\widehat{x} = b$$

with

$$\|\Delta X\| \leqslant \mathtt{u}\, g(n)\, \|X\|, \;\; |\Delta D| \leqslant \mathtt{u}\; g(n)\, |D|, \;\; \|\Delta Y\| \leqslant \mathtt{u}\, g(n)\, \|Y\|.$$

□

If we ignore in (4.4) second order terms and the pessimistic dimensional factor $g(n)$, then we obtain

$$\frac{\|\widehat{x}-x\|}{\|x\|} \lesssim \mathtt{u}\left(\kappa(\widehat{Y})+\left(1+2\,\kappa(\widehat{X})\right)\,\frac{\|A^{-1}\|\,\|b\|}{\|\widehat{x}\|}\right) =: \Theta_1, \tag{4.11}$$

that can be used to estimate the forward error in the solution. We have already mentioned that the RRD $\widehat{X}\widehat{D}\widehat{Y}$ is in most cases a (permuted) *LDU* factorization, so $\kappa(\widehat{X})$, $\kappa(\widehat{Y})$, and $\|A^{-1}\|$ can be estimated in $O(n^2)$ operations for p-norms (Higham, 2002, Chapter 15). However it is also possible to compute, from calculated quantities, a realistic bound without the use of A^{-1}. From $\|A^{-1}\| = \|Y^{-1}D^{-1}X^{-1}\| \leqslant \frac{\|Y^{-1}\|\,\|X^{-1}\|}{\min_i |d_i|}$, the following simpler estimation of the error follows

$$\frac{\|\widehat{x}-x\|}{\|x\|} \lesssim \mathtt{u}\left(\kappa(\widehat{Y})+\left(1+2\,\kappa(\widehat{X})\right)\,\frac{\|\widehat{Y}^{-1}\|\,\|\widehat{X}^{-1}\|}{\min_i |\widehat{d}_i|}\,\frac{\|b\|}{\|\widehat{x}\|}\right) =: \Theta_2. \tag{4.12}$$

Both quantities Θ_1 and Θ_2 will be computed in the numerical tests of Section 5 and we will see that they are reliable estimators of the true error in the solution.

5. Numerical Experiments

In this section we will show the results obtained from testing Algorithm 4.1 with two important classes of structured matrices: Cauchy and Vandermonde matrices. For matrices in these classes, accurate RRDs in the sense of Definition 2.1 can be computed using the algorithms developed by Demmel (1999) with a cost of $4n^3/3+O(n^2)$ operations plus $n^3/3+O(n^2)$ comparisons. Note that this cost is double than the cost of usual GECP. We will compare the relative errors committed by Algorithm 4.1 with those committed by other algorithms available in the literature and we will see that Algorithm 4.1 is by far the most accurate one[‡] and that, for random right-hand sides, it achieves relative normwise errors approximately equal to the unit roundoff $\mathtt{u}$ *for any distribution of the nodes defining Cauchy and Vandermonde matrices*. We will also check how well the quantities Θ_1 and Θ_2 defined in (4.11) and (4.12) estimate the actual errors. We will use in all the experiments the Euclidean or two norm and we will present tests only for matrices with real entries.

5.1 *Cauchy matrices*

The entries of a Cauchy matrix, $C \in \mathbb{R}^{n\times n}$, are defined in terms of two vectors $x = [x_1,\ldots,x_n]^T, y = [y_1,\ldots,y_n]^T \in \mathbb{R}^n$ as

$$c_{ij} = \frac{1}{x_i+y_j}. \tag{5.1}$$

It is well known that C is *totally positive* if $x_1+y_1>0$, $x_1<x_2<\cdots<x_n$, and $y_1<y_2<\cdots<y_n$ (Gantmacher & Krein, 2002, p. 78). Matrices of the form $G = D_1CD_2$, where C is Cauchy and D_1, D_2 are diagonal, are called by Demmel (1999) quasi-Cauchy matrices, which include as a particular case Cauchy matrices for $D_1 = D_2 = I$. Algorithm 3 in (Demmel, 1999) uses a structured version of GECP to compute accurate RRDs of any quasi-Cauchy matrix with a cost of $4n^3/3+O(n^2)$ operations plus $n^3/3+O(n^2)$ comparisons, and so this algorithm can be used in `Step 1` of Algorithm 4.1.

[‡]See however our comments below on the method BjP-PPP in Figure 4.

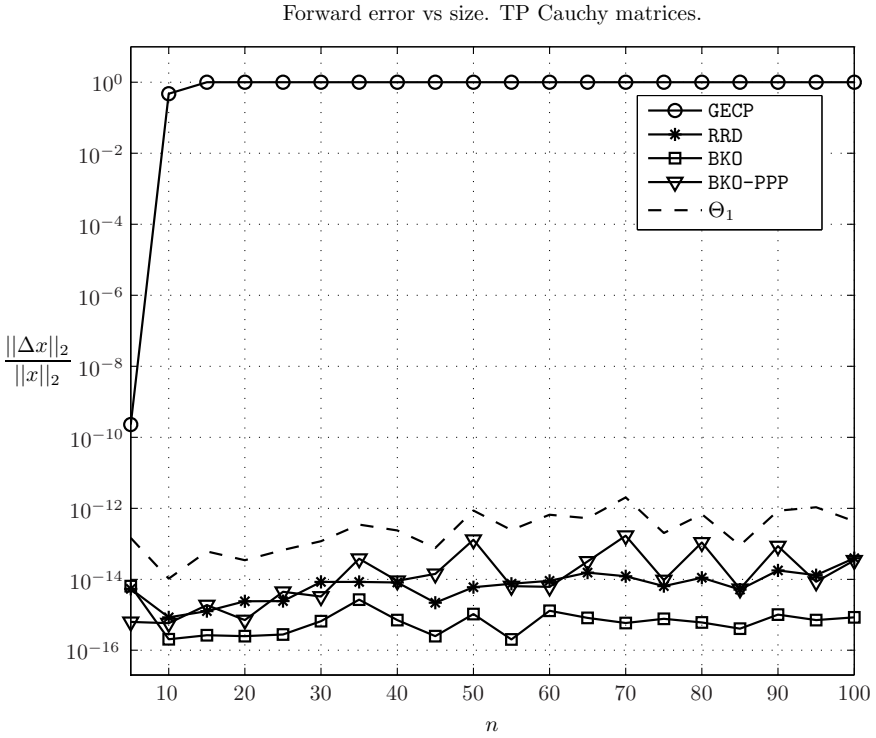

FIG. 1. Forward relative error $\frac{||\widehat{x}-x||_2}{||x||_2}$ against size for **Totally Positive Cauchy** matrices.

In the numerical experiments to solve linear systems $Ax = b$, where A is a Cauchy matrix, we have compared the following four algorithms implemented in MATLAB™:

- `GECP`: standard GECP, that is, we compute first the entries of the matrix and then apply GECP. Cost:[§] $2n^3/3$ operations + $n^3/3$ comparisons.
- `RRD`: Algorithm 4.1 implemented using Algorithm 3 in (Demmel, 1999) for `Step 1`, while for `Step 2` we solve $Xs = b$ with forward substitution (recall that X is a permutation of a lower triangular matrix), $Dw = s$ as $w_i = s_i/d_i$ for $i = 1 : n$, and $Yx = w$ with backward substitution (recall that Y is a permutation of an upper triangular matrix). Cost: $4n^3/3$ operations + $n^3/3$ comparisons.
- `BKO`: We use also Algorithm 3.2 by Boros *et al.* (1999), which is a fast method that solves Cauchy linear systems and that delivers satisfactory componentwise bounds for the relative forward error of the solution, for most right-hand sides (in the sense explained in Section 3.2) *only for totally positive Cauchy matrices* (Boros *et al.*, 1999). However, when the matrix is not totally positive, current analysis does not guarantee good bounds for the errors, even in normwise sense. Cost: $7n^2$ operations.

[§] When the cost is of order n^3 we omit the terms $O(n^2)$.

- `BKO-PPP`: Finally we use BKO algorithm above, but ordering the nodes *x in advance* by using Algorithm 5.2 by Boros *et al.* (2002). This ordering costs $O(n^2)$ operations and corresponds to the order of the rows of C that would be obtained by applying GEPP. This is why this strategy is called Predictive Partial Pivoting by Boros *et al.* (1999, 2002). This method has not guaranteed accuracy even when the matrix is totally positive (Boros *et al.*, 1999). Cost: $9n^2$ operations + $n^2/2$ comparisons.

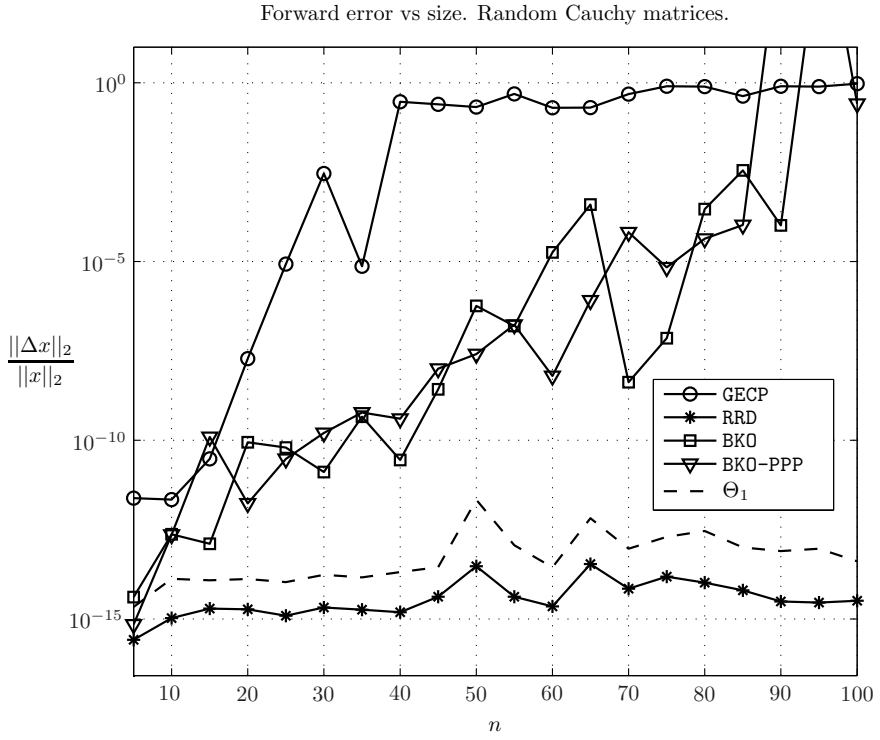

FIG. 2. Forward relative error $\frac{\|\widehat{x}-x\|_2}{\|x\|_2}$ against size for **non-Totally Positive Cauchy** matrices.

We have generated random Cauchy matrices both totally positive (TP) (uniformly distributed random, using `MATLAB`$^{\text{TM}}$ command `rand`, x and y vectors satisfying the TP requirements: $x_1 + y_1 > 0$, $x_1 < x_2 < \cdots < x_n$, and $y_1 < y_2 < \cdots < y_n$) and non totally positive (nTP) (normally distributed random, using `MATLAB`$^{\text{TM}}$ command `randn`, x and y vectors). We have used in all tests normally distributed random right-hand side vectors b. The sizes of the matrices have ranged from $n = 5$ to $n = 100$. The range of the condition numbers has been $10^6 \lesssim \kappa(C) \lesssim 10^{200}$. For each linear system we take as "exact" solution the one computed with the variable precision arithmetic of `MATLAB`$^{\text{TM}}$ set to $\log_{10}\kappa(C) + 30$ decimal digits, where $\kappa(C)$ has been estimated from the D factor of the RRD as $\kappa(C) \approx \max_i |d_i| / \min_i |d_i|$.

The results are shown in Figure 1 for TP matrices and in Figure 2 for non TP matrices. We have plotted in a log scale the relative error $\|\widehat{x}-x\|_2/\|x\|_2$ of the solution against the size of the matrices. We have run five different cases for every size and we have plotted the maximum error among the five cases. Besides the relative error for the four methods mentioned above, we have displayed (dashed line) also the quantity Θ_1 appearing in (4.11). It can be observed in both figures that the bound (4.11) is very sharp. We have also computed the quantity Θ_2 in (4.12). For clarity it is not shown in the figures, but it behaves as Θ_1, being approximately a factor 10 bigger, for the range of our experiments.

It can be seen in Figure 1 for TP matrices that, except `GECP` that produces huge errors due to the ill conditioning of Cauchy matrices, the other three methods get full accuracy ($\approx \mathtt{u}$) when the matrices are totally positive. This is the expected accuracy for the methods `BKO` and `RRD`, however the method with predictive partial pivoting, `BKO-PPP`, gets also full accuracy without having a theoretical result that guarantees this. We have reproduced the experiment in (Boros *et al.*, 1999, Figure 2), obtaining the same lost of accuracy for the `BKO-PPP` method for the TP matrices shown there, while our method (`RRD`) behaves perfectly well. More important, Figure 2 also shows that the algorithms introduced by Boros, Kailath and Olshevsky in (Boros *et al.*, 1999, 2002) do not deliver high relative accuracy when the Cauchy matrices are non totally positive, while the algorithm proposed in this paper, `RRD`, (continuous $*$-line) does.

An important remark about the sources of errors committed by `GECP` applied to Cauchy (and also to Vandermonde) matrices is in order here. A key difference between `GECP` and the other methods presented in this section is that the input data for the former are the entries of the matrix c_{ij}, and these have to be computed from the data vectors x and y, that are the input data for the other methods. Therefore, `GECP` is applied to a system $\widehat{C}x = b$, where $\widehat{C}$ is not the exact Cauchy matrix C. Due to the usually huge condition number of C, even the exact solution of $\widehat{C}x = b$ would differ enormously from the exact solution of $Cx = b$. In addition, `GECP` computes a solution $\widehat{x}$ of $\widehat{C}x = b$ that differs also greatly from the exact solution as a consequence again of the huge condition number of $\widehat{C}$, which is the second source of error for `GECP`. Both sources of error contribute essentially the same to the final relative error, i.e., ($\sim \mathtt{u}\kappa(C)$). A way to check that the first source of error in `GECP`, the one coming from forming the elements of the matrix, is not the only cause for the bad results of the forward error, is to create a matrix for which there are not rounding errors in forming it. This is not easy to do for Cauchy matrices, but it can be done easily for Vandermonde matrices, using floating point numbers (sums of power of two) for the components of the vector x. We have done those experiments, seeing that the forward errors behave in the same way as in Figures 3 and 4 (see below). In particular, those of `GECP` are still huge despite the fact that the matrix is formed without rounding errors.

5.2 *Vandermonde matrices*

We have performed similar numerical tests of Algorithm 4.1 on Vandermonde matrices. Given $x = [x_1,\ldots,x_n]^T \in \mathbb{R}^n$, a Vandermonde matrix, $V \in \mathbb{R}^{n\times n}$, is a matrix whose entries are given by

$$v_{ij} = x_i^{j-1}. \tag{5.2}$$

It is well known that V is totally positive if $0 < x_1 < \cdots < x_n$ (Gantmacher & Krein, 2002, p. 76). A method to compute an accurate RRD of any Vandermonde matrix was presented in (Demmel, 1999, 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 $O(n^2)$ 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 Algorithm 3 in (Demmel, 1999). Finally, $V = XD(YF^*)$ is an accurate RRD of V. For

most applications, in particular for linear solving, (YF^*) can be stored in factored form, then the method costs $4n^3/3 + O(n^2)$ operations plus $n^3/3 + O(n^2)$ comparisons. If one insists in computing explicitly YF^*, then this can be done accurately in $O(n^2 \log n)$ operations through the fast Fourier Transform (Higham, 2002, Chapter 24). Let us notice that the method above further requires, to guarantee the desired accuracy, that the primitive nth roots of unity be computed in advance with high precision. They can be computed and stored once for all. This is if *the nodes x_i are real* (our case in the experiments), if the nodes x_i are complex, further demands on the precision of the computations of the elements of the matrix F are necessary (Demmel, 1999).

In the numerical experiments to solve linear systems $Ax = b$, where A is a Vandermonde matrix, we have compared the following four algorithms implemented in MATLAB™:

- `GECP`: standard GECP, that is, we compute first the entries of the matrix and then apply GECP. Cost: $2n^3/3$ operations + $n^3/3$ comparisons.
- `RRD`: Algorithm 4.1 implemented using the method in Section 5 of (Demmel, 1999) for computing in `Step 1` an RRD $A = XD(YF^*)$, while for `Step 2` we solve $Xs = b$ with forward substitution (recall that X is a permutation of a lower triangular matrix), $Dw = s$ as $w_i = s_i/d_i$ for $i = 1 : n$, $Yx' = w$ with backward substitution (recall that Y is a permutation of an upper triangular matrix), and finally $x = Fx'$. Cost: $4n^3/3$ operations + $n^3/3$ comparisons.
- `BjP`: We use also the fast algorithm introduced by Björck & Pereyra (1970) -see also Algorithm 4.6.1 in Golub & Van Loan (1996). This algorithm guarantees excellent componentwise bounds for the relative forward error of the solution, for most right-hand sides (in the sense explained in Section 3.2) *only for totally positive Vandermonde matrices* (Higham, 1987, 2002). However, when the matrix is not totally positive, current analysis does not guarantee good bounds for the errors, even in normwise sense. Cost: $9n^2/2$.
- `BjP-PPP`: Finally we use BjP algorithm above, but ordering the nodes x in advance by using essentially the Leja ordering (Reichel, 1990). This ordering can be implemented in $n^2/2$ operations and $n^2/2$ comparisons (Higham, 1990, Algorithm 5.1) and corresponds to the order of the rows of V that would be obtained by applying GEPP. It is known that the accuracy of this last algorithm is not guaranteed even when the matrix is TP (Higham, 1990). Cost: $5n^2 + n^2/2$ comparisons.

For these algorithms we have proceed as in Section 5.1. We have generated random Vandermonde matrices both totally positive (TP) (uniformly distributed random, using MATLAB™ command `rand`, x vectors satisfying the TP requirements: $0 < x_1 < x_2 < \cdots < x_n$) and non totally positive (nTP) (normally distributed random, using MATLAB™ command `randn`, x vectors). We have used in all tests normally distributed random right-hand side vectors b. The sizes of the matrices have ranged from $n = 5$ to $n = 100$. The range of the condition numbers has been $10^2 \lesssim \kappa(V) \lesssim 10^{100}$. The results are shown in Figure 3 for TP matrices and in Figure 4 for non TP matrices. For TP matrices all algorithms, except `GECP`, get the full accuracy ($\approx \mathtt{u}$). This is the expected accuracy for the methods `BjP` and `RRD`; however the method with predictive partial pivoting, `BjP-PPP`, gets also full accuracy without having a theoretical result that guarantees this. There are experiments with TP Vandermonde matrices in (Boros *et al.*, 1999) that show that, for special distribution of nodes and right-hand sides, with small

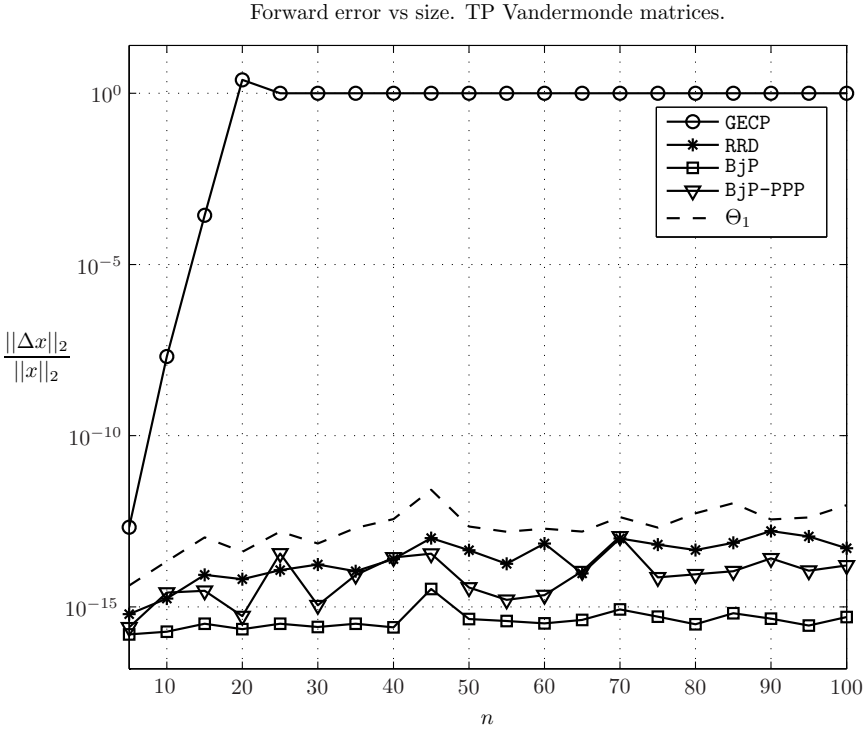

FIG. 3. Forward relative error $\frac{||\widehat{x}-x||_2}{||x||_2}$ against size for **Totally Positive Vandermonde** matrices.

$\|A^{-1}\|\|b\|/\|x\|$, the accuracy is lost by the `BjP-PPP` method. However, Figure 4 shows that, when Vandermonde matrices are non totally positive, Algorithm 4.1 delivers full accuracy while `BjP` does not. Therefore, as predicted by the error analysis, the behavior of the `RRD` algorithm is excellent both in the TP and the non-TP cases. It is remarkable also that the quantity Θ_1 (dashed line) is very close to the error of the method `RRD` (continuous $*$-line), being sometimes smaller. This shows how sharp the bound (4.5) is, and that the factor $g(n)$ is $O(1)$ for these experiments. There is a fact in Figure 4 for which we do not have an explanation: the `BjP-PPP` algorithm also shows full accuracy. As far as we know, there is no error analysis that guarantees this good behavior.

5.3 *Experiments with* $\dfrac{\|A^{-1}\|\,\|b\|}{\|x\|}$ *not small*

Besides the experiments with random matrices presented in subsections 5.1 and 5.2 we have performed experiments in which $\|A^{-1}\|\,\|b\|/\|x\|$ is not small. To do that, the right-hand side has been prepared to be $b = u_1$, the first left singular vector of A (see Theorem 3.3). In this case, both for Cauchy and Vandermonde matrices, the forward error of the solution has been, as expected, big and proportional to the unit roundoff times the condition number of the matrices. The results for six Cauchy matrices, three TP and three non-TP, are presented in Tables 1 and 2; the results for Vandermonde matrices are similar

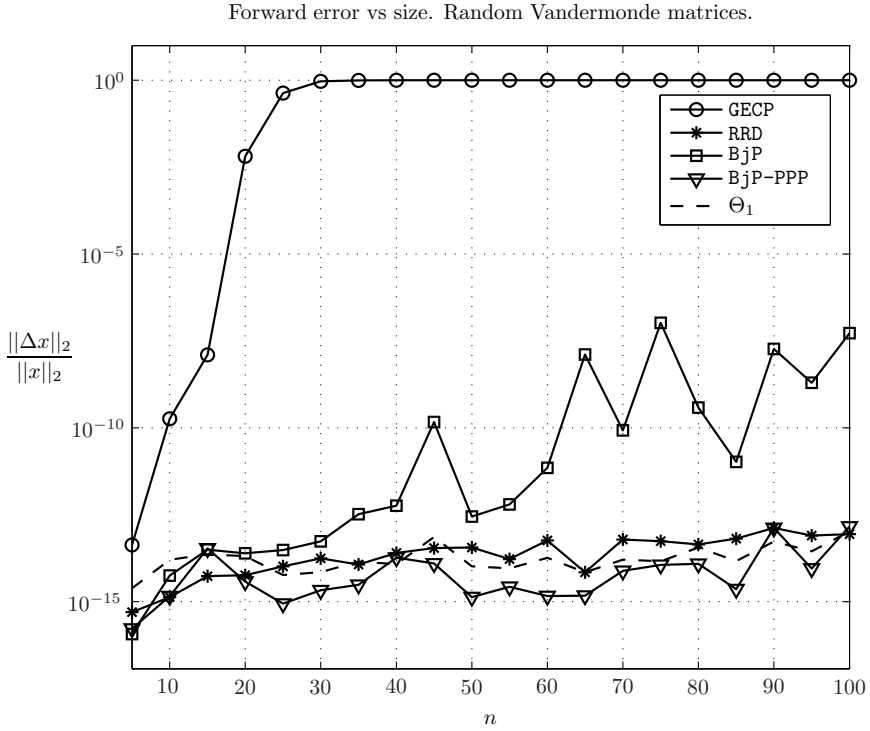

FIG. 4. Forward relative error $\frac{\|\widehat{x}-x\|_2}{\|x\|_2}$ against size for **non-Totally Positive Vandermonde** matrices.

and not presented here for brevity. Let us notice however that to get these results, the vector b has to be prepared using a high accurate algorithm for the SVD (Demmel *et al.*, 1999). 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 lay exactly in the direction of u_1, developing small components in the direction of other left singular vectors, and making the relative forward error be of order u.

6. Conclusions

We have introduced a numerical method that employs accurate rank-revealing decompositions (RRD) to solve accurately and with a cost of $O(n^3)$ operations many classes of structured linear systems $Ax = b$, even if the matrix $A \in \mathbb{C}^{n\times n}$ has a huge condition number. The precise cost of this procedure is given by the cost of computing an accurate RRD of the particular class of structured matrices we consider. We have presented a complete error analysis of this method that is based on a new structured perturbation theory for linear systems when A is perturbed through the factors of an RRD. We have performed very satisfactory numerical tests on Cauchy and Vandermonde matrices with arbitrary distributions of the nodes. These tests show that the accuracy of the proposed algorithm is essentially perfect in extreme situations where any other algorithm fails to compute accurate solutions. The accuracy of our method is governed, essentially, by the unit roundoff times the factor $\|A^{-1}\|\|b\|/\|x\|$, while for other methods, that

	$n = 10$	$n = 30$	$n = 50$
GECP	0.33	1.00	1.00
RRD	3.49	4.00	3.71
BKO	1.41	0.49	0.41

Table 1. Experiments with $\|A^{-1}\|\,\|b\|/\|x\|$ not small. The forward relative error $\|\Delta x\|_2/\|x\|_2$ is displayed for three different methods for **Totally Positive Cauchy** matrices, for sizes $n = 10$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 1.4 \cdot 10^{17}$, $n = 30$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 2.1 \cdot 10^{17}$, and $n = 50$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 2.0 \cdot 10^{17}$.

	$n = 10$	$n = 30$	$n = 50$
GECP	$1.46 \cdot 10^{-7}$	$4.30 \cdot 10^{-9}$	$2.04 \cdot 10^{-5}$
RRD	$9.13 \cdot 10^{-9}$	$4.44 \cdot 10^{-8}$	$7.85 \cdot 10^{-5}$
BKO	$1.81 \cdot 10^{-3}$	$5.33 \cdot 10^{-4}$	$2.49 \cdot 10^{3}$

Table 2. Experiments with $\|A^{-1}\|\,\|b\|/\|x\|$ not small. The forward relative error $\|\Delta x\|_2/\|x\|_2$ is displayed for three different methods for **non-Totally Positive Cauchy** matrices, for sizes $n = 10$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 2.5 \cdot 10^{10}$, $n = 30$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 1.4 \cdot 10^{10}$, and $n = 50$, $\|A^{-1}\|_2\,\|b\|_2/\|x\|_2 = 1.6 \cdot 10^{15}$.

get, for some experiments, the same accuracy (as BjP-PPP in Figure 4), this is not always guaranteed by theoretical bounds.

It remains an open problem to find fast algorithms, i.e., with $O(n^2)$ cost, for solving linear systems with guaranteed accuracy for Cauchy and Vandermonde matrices. It should be noted that the method we propose will compute accurate solutions of linear systems $Ax = b$ for any class of matrices A for which accurate RRDs can be computed, *now or in the future*, and for most vectors b.

Acknowledgements

We thank the referees and the Associate Editor of this paper, Nick Higham, for very useful comments and remarks that helped us to improve this paper. This research was partially supported by the Ministerio de Ciencia e Innovación of Spain through grant MTM2009-09281.

REFERENCES

BANOCZI, J. M., CHIU, N.-C., CHO, G. E. & IPSEN, I. C. F. (1998) The lack of influence of the right-hand side on the accuracy of linear system solution. *SIAM J. Sci. Comput.*, **20**, 203–227.

BJÖRCK, Å. & ELFVING, T. (1973/74) Algorithms for confluent Vandermonde systems. *Numer. Math.*, **21**, 130–137.

BJÖRCK, Å. & PEREYRA, V. (1970) Solution of Vandermonde systems of equations. *Math. Comp.*, **24**, 893–903.

BOROS, T., KAILATH, T. & OLSHEVSKY, V. (1999) A fast parallel Björck-Pereyra-type algorithm for solving Cauchy linear equations. *Linear Algebra Appl.*, **302/303**, 265–293.

BOROS, T., KAILATH, T. & OLSHEVSKY, V. (2002) Pivoting and backward stability of fast algorithms for solving Cauchy linear equations. *Linear Algebra Appl.*, **343/344**, 63–99.

CHAN, T. F. & FOULSER, D. E. (1988) Effectively well-conditioned linear systems. *SIAM J. Sci. Statist. Comput.*, **9**, 963–969.

DEMMEL, J. (1999) Accurate singular value decompositions of structured matrices. *SIAM J. Matrix Anal. Appl.*, **21**, 562–580.

DEMMEL, J., GU, M., EISENSTAT, S., SLAPNIČAR, I., VESELIĆ, K. & DRMAČ, Z. (1999) Computing the singular value decomposition with high relative accuracy. *Linear Algebra Appl.*, **299**, 21–80.

DEMMEL, J. & KOEV, P. (2004) Accurate SVDs of weakly diagonally dominant M-matrices. *Numer. Math.*, **98**, 99–104.

DEMMEL, J. & KOEV, P. (2006) Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials. *Linear Algebra Appl.*, **417**, 382–396.

DEMMEL, J. & VESELIĆ, K. (1992) Jacobi's method is more accurate than QR. *SIAM J. Matrix Anal. Appl.*, **13**, 1204–1246.

DOPICO, F. M., MOLERA, J. M. & MORO, J. (2003) An orthogonal high relative accuracy algorithm for the symmetric eigenproblem. *SIAM J. Matrix Anal. Appl.*, **25**, 301–351.

DOPICO, F. M., KOEV, P. & MOLERA, J. M. (2009) Implicit standard Jacobi gives high relative accuracy. *Numer. Math.*, **113**, 519–553.

DOPICO, F. M. & KOEV, P. (2006) Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices. *SIAM J. Matrix Anal. Appl.*, **28**, 1126–1156.

DOPICO, F. M. & KOEV, P. (2010). Perturbation theory for the LDU factorization and accurate computations for diagonally dominant matrices. Submitted.

EISENSTAT, S. & IPSEN, I. (1995) Relative perturbation techniques for singular value problems. *SIAM J. Numer. Anal.*, **32**, 1972–1988.

GANTMACHER, F. & KREIN, M. (2002) *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, revised edn. AMS Chelsea, Providence, RI, pp. viii+310.

GOLUB, G. & VAN LOAN, C. (1996) *Matrix Computations*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.

HIGHAM, N. J. (1987) Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems. *Numer. Math.*, **50**, 613–632.

HIGHAM, N. J. (1988) Fast solution of Vandermonde-like systems involving orthogonal polynomials. *IMA J. Numer. Anal.*, **8**, 473–486.

HIGHAM, N. J. (1990) Stability analysis of algorithms for solving confluent Vandermonde-like systems. *SIAM J. Matrix Anal. Appl.*, **11**, 23–41.

HIGHAM, N. J. (2002) *Accuracy and stability of numerical algorithms*, second edn. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xxx+680.

HORN, R. A. & JOHNSON, C. R. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

IPSEN, I. C. F. (1998) Relative perturbation results for matrix eigenvalues and singular values. *Acta numerica, 1998*. Acta Numer., vol. 7. Cambridge: Cambridge Univ. Press, pp. 151–201.

IPSEN, I. C. F. (2000) An overview of relative $\sin\Theta$ theorems for invariant subspaces of complex matrices. *J. Comput. Appl. Math.*, **123**, 131–153. Numerical analysis 2000, Vol. III. Linear algebra.

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

LI, R.-C. (1999) Relative perturbation theory. II. Eigenspace and singular subspace variations. *SIAM J. Matrix Anal. Appl.*, **20**, 471–492.

MATHIAS, R. (1995) Accurate eigensystem computations by Jacobi methods. *SIAM J. Matrix Anal. Appl.*, **16**, 977–1003.

MIRANIAN, L. & GU, M. (2003) Strong rank revealing *LU* factorizations. *Linear Algebra Appl.*, **367**, 1–16.

OLSHEVSKY, V. (ed.) (2001a). *Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Structured Matrices in Operator Theory, Numerical Analysis, Control, Signal and Image Processing held at the*

University of Colorado, Boulder, CO, June 27–July 1, 1999. Contemporary Mathematics, vol. 280. Providence, RI: American Mathematical Society, pp. xiv+327.

OLSHEVSKY, V. (ed.) (2001b). *Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference held at the University of Colorado, Boulder, CO, June 27–July 1, 1999*. Contemporary Mathematics, vol. 281. Providence, RI: American Mathematical Society, pp. xiv+344.

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

PELÁEZ, M. J. & MORO, J. (2006) Accurate factorization and eigenvalue algorithms for symmetric DSTU and TSC matrices. *SIAM J. Matrix Anal. Appl.*, **28**, 1173–1198.

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

REICHEL, L. (1990) Newton interpolation at Leja points. *BIT*, **30**, 332–346.

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