Numerische Mathematik manuscript No.
(will be inserted by the editor)

Perturbation theory for the LDU factorization and accurate computations for diagonally dominant matrices

Froilán M. Dopico · **Plamen Koev**

Received: date / Accepted: date; Version of April 19, 2011

Abstract We present a structured perturbation theory for the LDU factorization of (row) diagonally dominant matrices and we use this theory to prove that a recent algorithm of Q. Ye [Math. Comp. 77, 2195-2230 (2008)] computes the L, D and U factors of these matrices with relative errors less than $14n^3\mathbf{u}$, where $\mathbf{u}$ is the unit roundoff and $n \times n$ is the size of the matrix. The relative errors for D are componentwise and for L and U are normwise with respect the "max norm" $\|A\|_M = \max_{ij} |a_{ij}|$. These error bounds guarantee that for any diagonally dominant matrix A we can compute accurately its singular value decomposition and the solution of the linear system $Ax = b$ for most vectors b, independently of the magnitude of the traditional condition number of A and in $O(n^3)$ flops.

Keywords diagonally dominant matrices · high relative accuracy · LDU factorization · linear systems · singular value decomposition

Mathematics Subject Classification (2000) 65F05 · 65F15 · 15A18 · 15A23 · 15B99

1 Introduction

Diagonally dominant matrices are very important in applications. For instance, they arise out of finite difference and finite element discretizations of partial differential

F. M. Dopico was partially supported by the Ministerio de Ciencia e Innovación of Spain through grant MTM-2009-09281.

Froilán M. Dopico
Instituto de Ciencias Matemáticas CSIC-UAM-UC3M-UCM and Departamento de Matemáticas, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain
E-mail: dopico@math.uc3m.es

Plamen Koev
Department of Mathematics, San Jose State University,
One Washington Square, San Jose, CA 95192, USA
E-mail: koev@math.sjsu.edu

equations and in the solution of Markov modeling problems [1,5,8,26]. Diagonally dominant matrices enjoy excellent theoretical and numerical properties that are explained in classical references [7,29–32]. However, if the condition number of an $n \times n$ diagonally dominant matrix A is very large, then conventional algorithms compute its singular value decomposition (SVD) and solve linear systems $Ax = b$ with very large relative errors. The main goal of this paper is to prove rigorously that it is possible to compute with high accuracy SVDs and solutions of linear systems (for most vectors b) for any diagonally dominant matrix independently of its condition number and at $O(n^3)$ cost, i.e., roughly the same cost as that of conventional algorithms for dense matrices. These results are direct consequences of the rigorous error analysis that we will develop for an algorithm recently presented by Q. Ye [42] for computing the LDU factorization of diagonally dominant matrices in $2n^3 + O(n^2)$ operations.

This work is part of the intensive research effort that has been made in the last twenty years to derive algorithms for computing the SVD of important classes of structured $n \times n$ matrices *to high relative accuracy* at $O(n^3)$ cost. Some selected references in this area are [10–16,22–24,27,42]. By *high relative accuracy* of singular values we mean that the exact singular values σ_i and their computed counterparts $\hat{\sigma}_i$ satisfy

$$|\hat{\sigma}_i - \sigma_i| \leq \mathbf{u}\, p(n)\, \sigma_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $\mathbf{u}$ is the unit roundoff of the computer and $p(n)$ is a polynomial of low degree in n. These error bounds guarantee that *all* singular values, including the tiniest ones, are computed with correct leading digits. We refer the reader to [12] for the appropriate meaning of "high relative accuracy" in singular vectors.

Introduced by Demmel et al. [12], and motivated by multiplicative perturbation results developed in [25,33], the key unifying idea in high accuracy computations of SVDs is to first compute an accurate *rank revealing decomposition* (RRD), i.e., a decomposition $A = XDY^T$, where X and Y are well conditioned and $D = \mathrm{diag}(d_1, \ldots, d_n)$ is diagonal, and then recover the singular values and vectors of A from the factors of the RRD through algorithms of Jacobi type. More precisely, it is shown in [12] that if a certain algorithm computes in floating point arithmetic factors $\widehat{X}$, $\widehat{D}$ and $\widehat{Y}$ with errors

$$|\widehat{d}_i - d_i| \leq \mathbf{u}\, q(n)\, |d_i|, \quad \text{for} \quad i = 1, \ldots, n, \tag{1}$$

$$\|\widehat{X} - X\|_2 \leq \mathbf{u}\, q(n) \|X\|_2, \qquad \|\widehat{Y} - Y\|_2 \leq \mathbf{u}\, q(n) \|Y\|_2, \tag{2}$$

where $q(n)$ is a polynomial of low degree in n and $\|\cdot\|_2$ is the matrix spectral norm[1] [30, Ch. 6], then the Jacobi type algorithms proposed in [12] compute the singular values of A at $O(n^3)$ cost and with errors

$$|\hat{\sigma}_i - \sigma_i| \leq \mathbf{u}\, p(n)\, \max\{\kappa(X), \kappa(Y)\}\, |\sigma_i| \quad \text{for} \quad i = 1, \ldots, n, \tag{3}$$

with $\kappa(X) := \|X\|_2 \|X^{-1}\|_2$ the spectral condition number of X. The fundamental point in (3) is that the error bound is governed by the condition numbers of the well

[1] Obviously any other matrix norm like the 1-norm, ∞-norm, Frobenius norm or "max norm" can be used at the cost of modifying the degree of $q(n)$. We use the spectral norm in this Introduction simply to follow reference [12].

conditioned factors X and Y, and not by $\kappa(A)$ which may be extremely large. Symmetric RRDs have been also used to compute accurate eigenvalues and eigenvectors of symmetric matrices [21,18,19].

Very recently [20], it has been shown that computing an accurate RRD, in the sense of (1)-(2), of an $n \times n$ matrix A also leads to very important benefits in the accuracy of the numerical solution of the system $Ax = b$. More precisely, if the solution is computed in $O(n^3)$ flops as $Xz = b$, $Dy = z$, and $Y^T x = y$, then the computed solution $\widehat{x}$ satisfies

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \leq \mathbf{u}\, p(n)\, \max\{\kappa(X),\kappa(Y)\}\, \frac{\|A^{-1}\|_2\, \|b\|_2}{\|x\|_2}. \tag{4}$$

This is a very satisfactory bound because if u_n is the left singular vector of A associated with its smallest singular value, then $\frac{\|A^{-1}\|_2\,\|b\|_2}{\|x\|_2} \leq \frac{\|b\|_2}{|u_n^T b|}$. This ratio is much smaller than $1/\mathbf{u}$ for all vectors b that are not close to be orthogonal to u_n, i.e., for most vectors b [9].

The previous discussion illustrates that the computation of RRDs with errors given by (1)-(2) is a task from which important benefits can be obtained for the accuracy of basic problems in Numerical Linear Algebra. In principle, this task may be undertaken by using Gaussian Elimination with Complete Pivoting (GECP) because it computes, in general, LDU factorizations with very well conditioned factors $L\,(=X)$ and $U\ (=Y^T)$. However, it is well known that standard GECP produces very large forward errors for ill conditioned matrices, and so it does not achieve (1)-(2). Error bounds as those in (1)-(2) can be obtained only for certain classes of relevant structured matrices through highly structured and nontrivial implementations of GECP or variations of it [6,10–12,14,15,18,37]. These classes of matrices include: Cauchy, Vandermonde, acyclic, diagonally dominant M-matrices, γ-scaled diagonally dominant, polynomial Vandermonde, and many others.

In this context a novel algorithm has been recently developed by Q. Ye [42, Algorithm 1] for computing in $2n^3$ flops the LDU factorization with complete pivoting of a *(row*[2]*) diagonally dominant matrix* A for which its off-diagonal entries and diagonally dominant parts are accurately known (see Section 2.1 for definitions). Numerical experiments presented in [42] show that *in practice* this algorithm behaves perfectly well and achieves errors of type (1)-(2). However, if $\widehat{L},\widehat{D} = \mathrm{diag}(\widehat{d}_1,\ldots,\widehat{d}_n),\widehat{U}$ are the computed factors and $L,D = \mathrm{diag}(d_1,\ldots,d_n),U$ are the exact ones, the best error bounds that have been proved so far are [42, Theorem 3]

$$|\widehat{d}_i - d_i| \leq (\mathbf{u}\, 5\cdot 8^{n-1} + O(\mathbf{u}^2))\, |d_i|, \quad \text{for} \quad i = 1,\ldots,n,$$
$$\|\widehat{L}-L\|_\infty \leq (\mathbf{u}\, n\, 6\cdot 8^{n-1} + O(\mathbf{u}^2))\, \|L\|_\infty, \quad \|\widehat{U}-U\|_\infty \leq (\mathbf{u}\, 6\cdot 8^{n-1} + O(\mathbf{u}^2))\, \|U\|_\infty.$$

Observe that these bounds do not guarantee any single correct digit even for very small matrices, since in double precision IEEE arithmetic $\mathbf{u} = 2^{-53}$ and, then, $\mathbf{u}\, 8^{n-1} \geq 2^4$ if $n \geq 20$. Despite of being pessimistic, these bounds have a remarkable feature:

[2] Note that results for the LDU factorization of *column* diagonally dominant matrices follow easily from the row case: if $A = LDU$ is column diagonally dominant, then $A^T = U^T D L^T$ is row diagonally dominant.

they show that the forward errors are independent of any condition number of the matrix A. This motivates a search of sharper error bounds for Algorithm 1 in [42].

The first main result in this work is to prove rigorously that Algorithm 1 in [42] with complete pivoting achieves high accuracy, because it computes LDU factors with relative errors bounded by $14n^3\mathbf{u}$. This is presented in Theorem 4, which relies on a *new perturbation theory for the LDU factorization of diagonally dominant matrices*. This is presented in Theorem 3 and *is the second main result in this work*.

Algorithm 1 in [42] uses as inputs the diagonally dominant parts of A, that are defined as $v_i := |a_{ii}| - \sum_{j\neq i} |a_{ij}|$ for $i = 1,\ldots,n$. We discuss in this paragraph that this is not an obstacle that spoils the accuracy of this algorithm. First, diagonally dominant parts are often known in applications, because they may have physical significance or they may be directly provided by the discretization scheme of a differential equation [4]. In addition, even when only the entries of A are known, v_i can be accurately computed either by standard recursive summation or, if $|a_{ii}| \approx \sum_{j\neq i} |a_{ij}|$, by *doubly compensated summation* at $10n$ cost [38] (see also [30, Section 4.3]). Of course, these accuracy issues refer to a matrix A stored in the computer. Errors coming from the storage process are intrinsic in numerical computations and can only be avoided with the use of extended precision.

The rest of the paper is organized as follows. Section 2 establishes basic concepts. In Section 3, we develop a structured perturbation theory for the LDU factorization of diagonally dominant matrices. This theory is summarized in Theorem 3, which is used in Section 4 to perform a detailed error analysis of Algorithm 1 in [42]. The main results of this analysis are included in Theorem 4. Conclusions and discussion of future work are presented in Section 5.

2 Notation and preliminaries

Notation: We consider only real matrices and denote the set of $m \times n$ real matrices by $\mathbb{R}^{m\times n}$. The entries of a matrix A are a_{ij} and $|A|$ is the matrix with entries $|a_{ij}|$. Inequalities $A \geq B$ for matrices mean $a_{ij} \geq b_{ij}$ for all i, j. We use MATLAB [34] notation for submatrices, e.g., $A(i:j,k:l)$ indicates the submatrix of A consisting of rows i through j and columns k through l, and $A(:,k:l)$ indicates the submatrix of A consisting of columns k through l. $A(i',j')$ denotes the submatrix of A with row i and column j deleted. Upper (resp. lower) triangular matrices whose diagonal entries are equal to one are called unit upper (resp. lower) triangular matrices. I_s is the $s \times s$ identity matrix and 0_s the $s \times s$ zero matrix. The sign of $x \in \mathbb{R}$ is $\mathrm{sign}(x)$, with the convention that $\mathrm{sign}(0) = 1$. We present componentwise perturbation theory and error analysis, but we use occasionally three matrix norms in the rest of the paper to translate componentwise into normwise bounds: the "max norm" $\|A\|_M := \max_{ij} |a_{ij}|$, $\|A\|_1$ and $\|A\|_\infty$ as they are defined in [30, Ch. 6].

In this section we summarize some basic properties of row diagonally dominant matrices that *may be singular*. The potential singularity of the matrix complicates the results, but this complication is necessary to develop perturbation theory and error analysis valid for singular matrices in Sections 3 and 4. Similar results to the ones in this section hold for column diagonally dominant matrices with the obvious mod-

ifications. In addition, we introduce the model of floating point arithmetic used in Section 4 and some notation for error analysis.

Let us recall that $A \in \mathbb{R}^{n\times n}$ is row diagonally dominant if

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|, \quad i = 1,\ldots,n.$$

The results in Theorem 1 are often used in this paper. The reader may refresh the relation between Schur complements and Gaussian elimination in [29, p. 103].

Theorem 1 *If $A \in \mathbb{R}^{n\times n}$ is row diagonally dominant, then*

(a) *Every principal submatrix of A is row diagonally dominant;*
(b) *PAP^T is row diagonally dominant for every permutation matrix $P \in \mathbb{R}^{n\times n}$;*
(c) *If $a_{11} \neq 0$ then the Schur complement of a_{11} in A is row diagonally dominant;*
(d) *If $\det A \neq 0$ then $\det A$ has the same sign as the product $a_{11}a_{22}\ldots a_{nn}$; and*
(e) *$|\det A(i',i')| \geq |\det A(i',j')|$, for all $i = 1,\ldots,n$ and all $j \neq i$.*

Proof Parts (a) and (b) are immediate. For nonsingular matrices, the proofs of parts (c), (d) and (e) can be found in: part (c) in [30, Theorem 13.7], part (d) in [32, p. 125] and part (e) is proven inside the proof of [32, Theorem 2.5.12]. The proofs in [32] are done for strictly row diagonally dominant matrices. The reader is invited to check that these proofs remain essentially the same for row diagonally dominant matrices that may be singular. □

The entry with largest absolute value of a row diagonally dominant matrix is in the main diagonal. This property and parts (b) and (c) of Theorem 1 allow us to perform Gaussian elimination with complete pivoting on these matrices as follows: in each stage make the same row and column exchanges to place in the pivot position the diagonal entry with largest absolute value of the corresponding Schur complement. This strategy will be called *complete-diagonal pivoting* and is of paramount importance in this paper. However, many of the perturbation and rounding error bounds we have obtained remain valid for any *diagonal pivoting* strategy, i.e., a strategy that chooses as pivot any nonzero entry in the diagonal of each Schur complement.

The main object in this paper is the LDU factorization of a matrix A, $A = LDU$, where D is diagonal and L (resp. U) is a lower (resp. upper) unit triangular matrix. It is well known that nonsingular row diagonally dominant matrices always have LDU factorization without pivoting [29,30]. This is no longer true in the singular case. In general, we only have Theorem 2.

Theorem 2 *Let $A \in \mathbb{R}^{n\times n}$ be a row diagonally dominant matrix with rank r. Then there exist a permutation matrix $P \in \mathbb{R}^{n\times n}$, a unit lower triangular matrix $L_{11} \in \mathbb{R}^{r\times r}$, a unit upper triangular matrix $U_{11} \in \mathbb{R}^{r\times r}$, and a nonsingular diagonal matrix $D_{11} = \operatorname{diag}(d_1,\ldots,d_r) \in \mathbb{R}^{r\times r}$ such that*

$$PAP^T = LDU \tag{5}$$

where

$$L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & I_{n-r} \end{bmatrix}, \qquad D = \begin{bmatrix} D_{11} & 0 \\ 0 & 0_{n-r} \end{bmatrix}, \qquad U = \begin{bmatrix} U_{11} & U_{12} \\ 0 & I_{n-r} \end{bmatrix}.$$

If $\operatorname{rank}(A) = n$ then P may be taken equal to I_n.

Proof The result follows easily from performing Gaussian elimination with any diagonal pivoting strategy on A to compute the LDU factorization of A. □

For simplicity, we will often assume that the matrix A is arranged in such a way that it satisfies (5) with $P = I$. This motivates Definition 1.

Definition 1 A row diagonally dominant matrix $A \in \mathbb{R}^{n\times n}$ with rank r is said to have LDU factorization if A can be factorized as in (5) with $P = I$. In this case, equation (5) is $A = LDU$ and this is called the LDU factorization of A.

The LDU factorization of A in Definition 1 is unique and the nontrivial entries of L, D and U are given by [28, p. 35]

$$\ell_{ij} = \frac{\det A([1:j-1,i],1:j)}{\det A(1:j,1:j)}, \quad i > j \quad \text{and} \quad j = 1,\ldots,r, \tag{6}$$

$$d_i = \frac{\det A(1:i,1:i)}{\det A(1:i-1,1:i-1)}, \quad i = 1,\ldots,r, \quad (\det A(1:0,1:0) := 1) \tag{7}$$

$$u_{ij} = \frac{\det A(1:i,[1:i-1,j])}{\det A(1:i,1:i)}, \quad i < j \quad \text{and} \quad i = 1,\ldots,r. \tag{8}$$

These determinantal formulas are essential in the perturbation theory of Section 3. Recall also that the factor U of a row diagonally dominant matrix is also row diagonally dominant, so $|u_{ij}| \leq 1$ for $i < j$.

Determinantal formulas for the entries of the Schur complements of A are also of interest to us. Let us recall them. Assume that $A \in \mathbb{R}^{n\times n}$ with rank r has LDU factorization as in Definition 1. Then Gaussian elimination without pivoting applied to A finishes after $\min\{r,n-1\}$ stages[3]. Define $A^{(1)} := A$ and let $A^{(k+1)} = [a_{ij}^{(k+1)}] \in \mathbb{R}^{n\times n}$ be the matrix obtained after k stages of Gaussian elimination have been performed. The matrix $A^{(k+1)}(k+1:n,k+1:n)$ is the Schur complement of $A(1:k,1:k)$ in A [29, p. 103]. We have that $A^{(r+1)}(r+1:n\,,\,r+1:n) = 0$ and also (see [28, p. 26])

$$a_{ij}^{(k+1)} = \frac{\det A([1:k,i],[1:k,j])}{\det A(1:k,1:k)}, \quad i,j = k+1,\ldots,n, \quad k = 1,\ldots,\min\{r,n-1\}. \tag{9}$$

Since complete-diagonal pivoting will play an essential role in this paper, it is convenient to define those matrices that are not permuted by Gaussian elimination with complete-diagonal pivoting. This is done in Definition 2.

Definition 2 A row diagonally dominant matrix $A \in \mathbb{R}^{n\times n}$ with rank r and having LDU factorization is said to be arranged for complete-diagonal pivoting if

$$|a_{kk}^{(k)}| = \max_{k\leq i\leq n} |a_{ii}^{(k)}|, \quad k = 1,\ldots,\min\{r,n-1\}.$$

We will also say that the matrix A is *almost* arranged for complete-diagonal pivoting if $|a_{kk}^{(k)}|$ is not much smaller than $\max_{k\leq i\leq n}|a_{ii}^{(k)}|$ for all k, i.e.,

$$|a_{kk}^{(k)}| \geq \frac{\max_{k\leq i\leq n}|a_{ii}^{(k)}|}{c}, \quad k = 1,\ldots,\min\{r,n-1\},$$

[3] As usual [30], in this paper the kth stage is the one that makes zero the entries in the kth column below the main diagonal. Therefore Gaussian elimination finishes after at most $n-1$ stages.

for some modest constant $c \geq 1$ independent of k.

2.1 Parametrization in terms of diagonally dominant parts and determinants

In the rest of the paper we consider matrices $A \in \mathbb{R}^{n\times n}$ such that

$$a_{ii} \geq 0, \quad i = 1,\ldots n, \tag{10}$$

as it was done in [42]. This does not impose any theoretical or numerical restriction for solving linear systems whose matrix is A or for computing the singular value decomposition of A, since we can multiply A by a diagonal matrix $D = \mathrm{diag}(\pm 1,\ldots,\pm 1)$ to turn the diagonal entries of A into nonnegative numbers. This process does not produce any rounding errors and does not change the singular values of A. In addition, the modifications in the *LDU* decomposition of A or in its singular vectors are trivial.

We follow [42] and parameterize every row diagonally dominant matrix $A \in \mathbb{R}^{n\times n}$ in terms of its diagonally dominant parts and off-diagonal entries. The *diagonally dominant parts* of A are defined as

$$v_i := a_{ii} - \sum_{j\neq i} |a_{ij}|, \quad i = 1,\ldots n, \tag{11}$$

and they will be frequently stored in a vector $v := [v_1,\ldots,v_n]^T$. Note that a matrix satisfying (10) is row diagonally dominant if and only if $v_i \geq 0$ for $i = 1,\ldots,n$. The off-diagonal entries of A are stored in the matrix A_D whose entries are

$$(A_D)_{ij} := \begin{cases} 0 & \text{for } i = j \\ a_{ij} & \text{for } i \neq j \end{cases}.$$

The pair (A_D, v) allows us to recover the matrix A and, therefore, it provides a parametrization of A. A matrix A parameterized is this way will be denoted as

$$A = \mathscr{D}(A_D, v). \tag{12}$$

Q. Ye has introduced very recently the use of the parametrization $A = \mathscr{D}(A_D, v)$ of row diagonally dominant matrices in [42,43]. This author used previously this parametrization in the particular case of diagonally dominant matrices that are also M-matrices [3,4] and it has been also applied by other authors for these matrices [14, 37]. It should be noted that the concept of *diagonally dominant part* is not new: it can be traced back to early references, where it is applied to bound the condition number of diagonally dominant matrices [2,40,41].

The parametrization (12) allows us to express in Lemma 1 the determinant of row diagonally dominant matrices as a summation of nonnegative terms and this is the main reason why a satisfactory perturbation theory can be developed in Section3. Here and in the rest of the paper $v \geq 0$ denotes $v_i \geq 0$ for $i = 1,\ldots,n$.

Lemma 1 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be such that $v \geq 0$, i.e., A is row diagonally dominant with nonnegative diagonal entries. Denote the algebraic cofactors of A by*

$$C_{ij} := (-1)^{i+j} \det A(i', j'), \quad i, j = 1, \ldots, n.$$

Then

$$\det A = v_i C_{ii} + \sum_{j\neq i} \left(|a_{ij}|C_{ii} + a_{ij}C_{ij}\right), \quad i = 1, \ldots, n,$$

with $v_i C_{ii} \geq 0$ and $(|a_{ij}|C_{ii} + a_{ij}C_{ij}) \geq 0$ for $j \neq i$.

Proof Use $a_{ii} = v_i + \sum_{j\neq i} |a_{ij}|$ in the cofactor expansion $\det A = a_{ii}C_{ii} + \sum_{j\neq i} a_{ij}C_{ij}$. For the signs: recall that $A(i', i')$ is also row diagonally dominant with nonnegative diagonal entries, so $\det A(i', i') \geq 0$ by Theorem 1, part (d). Finally, apply part (e) of Theorem 1 to prove $(|a_{ij}|C_{ii} + a_{ij}C_{ij}) \geq 0$. □

2.2 Model of floating point arithmetic and notation for error analysis

In the rounding error analysis of Section 4 we use the conventional error model for floating point arithmetic [30, section 2.2]: $fl(a \odot b) = (a \odot b)(1 + \psi)$, where a and b are real floating point numbers, $\odot \in \{+, -, \times, /\}$, and $|\psi| \leq \mathbf{u}$, with $\mathbf{u}$ the unit roundoff. We assume that neither overflow nor underflow occurs. We will also use the following result [30, Lemma 3.1]: if q is a positive integer, $|\psi_i| \leq \mathbf{u}$ and $\rho_i = \pm 1$ for $i = 1, \ldots, q$, and $q\mathbf{u} < 1$, then

$$\prod_{i=1}^{q} (1 + \psi_i)^{\rho_i} = 1 + \theta_q, \quad \text{where} \quad |\theta_q| \leq \frac{q\mathbf{u}}{1 - q\mathbf{u}} =: \gamma_q. \tag{13}$$

In proofs and auxiliary lemmas, we will use for simplicity Stewart's notation for keeping track of products of $(1 + \psi_i)^{\rho_i}$ factors [39] (see also [30, p. 68]):

$$<q> := \prod_{i=1}^{q} (1 + \psi_i)^{\rho_i}, \tag{14}$$

which satisfies the rules $<j><k> = <j + k>$ and $<j> \; / \; <k> = <j + k>$. In general, we use the same symbol $<k>$ for different products of k factors $(1 + \psi_i)^{\rho_i}$, so $<k> \; / \; <k> = <2k>$. Of course, $<0> = 1$. Observe that for all q, $<q> > 0$, because $\mathbf{u} < 1$. In addition, if $s > q$, then $<q>$ can be replaced in any expression by $<s>$, because $<q> = <q> \prod_{i=1}^{s-q} (1 + \psi_i')$, with $0 = |\psi_i'| < \mathbf{u}$.

3 Structured perturbation theory for the LDU factorization

This section is organized in two parts. In Subsection 3.1, we state without proofs the main results we have obtained on perturbation bounds for the LDU factorization of a row diagonally dominant matrix with nonnegative diagonal entries. The proofs are presented in Subsections 3.2 and 3.3, together with some auxiliary lemmas that we

consider interesting on their own. We hope that this organization will allow the reader to find quickly the most relevant information.

The main perturbation result is that small entrywise relative perturbations of the diagonally dominant parts, v, and the off-diagonal entries, A_D, of a row diagonally dominant matrix A with nonnegative diagonal entries produce small relative variations of the entries of D and small absolute variations of the entries of U. However, the absolute variations of the entries of L are not small in general and we can only prove that they are small if the matrix A is *almost* arranged for complete-diagonal pivoting. We stress that complete-diagonal pivoting is essential, since, for other pivoting strategies, we present an example where the variations in the factor L are very large. Note that small *absolute* variations of the entries of L and U imply small *relative* normwise variations of L and U as a consequence of the lower bounds $1 \leq \|L\|_\infty$ and $1 \leq \|U\|_\infty$. The satisfactory perturbation bounds that we present do not depend on any condition number.

3.1 Main perturbation results and comments

We gather the perturbation bounds we have proved for the LDU factorization in Theorem 3. This theorem only pays attention to the variation of those entries of the *LDU* factors that are not trivial according to Definition 1 and the rank of the matrix.

Theorem 3 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be such that $v \geq 0$, i.e., A is row diagonally dominant with nonnegative diagonal entries. Suppose* $\operatorname{rank}(A) = r$ *and that A has* LDU *factorization $A = LDU$ as in Definition 1. Let $\widetilde{A} = \mathscr{D}(\widetilde{A}_D, \widetilde{v}) \in \mathbb{R}^{n\times n}$ be such that*

$$|\widetilde{v} - v| \leq \delta\, v \quad \textit{and} \quad |\widetilde{A}_D - A_D| \leq \delta\, |A_D|, \quad \textit{for some } 0 \leq \delta < 1. \tag{15}$$

Then

(a) *$\widetilde{A}$ is also row diagonally dominant with nonnegative diagonal entries,* $\operatorname{rank}(\widetilde{A}) = r$*, and it has LDU factorization $\widetilde{A} = \widetilde{L}\widetilde{D}\widetilde{U}$;*

(b) *For $i = 1, \ldots, r$,*

$$\widetilde{d}_i = d_i \frac{(1+\eta_{i1})\cdots(1+\eta_{ii})}{(1+\eta_{i-1,1})\cdots(1+\eta_{i-1,i-1})}, \quad \textit{where} \quad |\eta_{ik}| \leq \delta,\ |\eta_{i-1,p}| \leq \delta,$$

for $k = 1, \ldots, i$ and $p = 1, \ldots, i-1$;

(c)

$$|\widetilde{u}_{ij} - u_{ij}| \leq 3\, i\, \delta, \qquad i = 1, \ldots, \min\{r, n-1\} \quad \textit{and} \quad i < j;$$

(d) *For $j = 1, \ldots, \min\{r, n-1\}$ and $i > j$*

$$|\widetilde{\ell}_{ij} - \ell_{ij}| \leq |\ell_{ij}| \left(\frac{1}{(1-\delta)^j} - 1 \right) + 2\, \frac{(1+\delta)^j - 1}{(1-\delta)^j} \left| \frac{a_{ii}^{(j)}}{a_{jj}^{(j)}} \right|,$$

where $A^{(j)}$ is the matrix obtained after $(j-1)$ stages of Gaussian elimination;

(e) *In addition, let* $r' = \min\{r, n-1\}$ *and* $\beta_1, \beta_2, \ldots, \beta_{r'}$ *be numbers such that* $0 \leq \beta_j$, $(1+\beta_j)\,|a_{jj}^{(j)}| \geq |a_{ii}^{(j)}|$, *for* $j = 1, \ldots, r'$ *and* $j < i \leq n$, *and* $j\,\delta < 1$, *then*

$$|\widetilde{\ell}_{ij} - \ell_{ij}| \leq (1+\beta_j)\,\frac{j\,\delta}{1-j\,\delta}\left(3 + \frac{2\,j\,\delta}{1-j\,\delta}\right). \tag{16}$$

Note that if A is arranged for complete-diagonal pivoting then (16) *holds with* $\beta_j = 0$ *for* $j = 1, \ldots, r'$.

Theorem 3 is our main perturbation result and it deserves some comments. Note first that *if the matrix A does not have LDU factorization* but, according to Theorem 2, there exists a permutation matrix P such that PAP^T has LDU factorization, then part (a) of Theorem 3 guarantees that, *for the same* P, $P\widetilde{A}P^T$ has also LDU factorization. Second, recall that $1 \leq \|U\|_\infty$, hence the *absolute entrywise* bounds in part (c) of Theorem 3 provide the following very satisfactory *relative normwise* bound for the factor U:

$$\frac{\|\widetilde{U} - U\|_\infty}{\|U\|_\infty} \leq \frac{3}{4}\,n^2\,\delta.$$

However, the bounds for the entries of the L factors in part (d) of Theorem 3 are large if $\delta\,|a_{ii}^{(j)}| \gtrsim |a_{jj}^{(j)}|$ for some $i > j$. Fortunately, as a direct consequence of part (d), part (e) of Theorem 3 establishes good absolute entrywise bounds when the matrix A is *almost* arranged for complete-diagonal pivoting, i.e., if the pivot used in each stage j of Gaussian elimination is, in absolute value, larger than the maximum possible pivot divided by a moderate number larger than one. We consider matrices that are almost arranged for complete-diagonal pivoting because they naturally appear in the error analysis of Section 4, since the permutation matrix determined by Algorithm 1 in [42] with complete-diagonal pivoting in *floating point arithmetic* may be different from the permutation matrix corresponding to complete-diagonal pivoting in *exact aritmetic*. We will see that the perturbation behaviors of the L and U factors are different as a consequence of the different properties of L and U factors of row diagonally dominant matrices: U is also row diagonally dominant while L does not inherit, in general, any particular property.

Diagonal pivoting strategies in Gaussian elimination that compute column diagonally dominant L factors of $n \times n$ row diagonally dominant matrices were presented in [42, Algorithm 1] (see also [42, pp. 2198-2199]). Strategies of this type were previously introduced in [37] for row diagonally dominant M-matrices and have the advantage over complete-diagonal pivoting that the L and U factors satisfy

$$\kappa_1(L) \leq 2n \quad \text{and} \quad \kappa_\infty(U) \leq 2n.$$

Therefore, these strategies would guarantee rigorously small error bounds in (3) for the singular values, if they can be implemented to compute LDU factorizations with errors (1)-(2). The drawback of these strategies is that part (d) of Theorem 3 does not provide good perturbation bounds for the L factor, and this is not an artifact of our proof as Example 1 below shows.

Example 1 This example illustrates that complete-diagonal pivoting is essential to guarantee a good behavior of the factor L under structured small perturbations of type (15). Consider the LDU factorization of the following row diagonally dominant matrix A

$$A = \begin{bmatrix} 1000 & 100 & 500 \\ 0 & 0.1 & 0.05 \\ 100 & 10 & 120 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0.1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1000 & & \\ & 0.1 & \\ & & 70 \end{bmatrix} \begin{bmatrix} 1 & 0.1 & 0.5 \\ & 1 & 0.5 \\ & & 1 \end{bmatrix}, \qquad (17)$$

and note that the vector of the diagonally dominant parts of A is $v(A) = [400, 0.05, 10]$. The factor L is column diagonally dominant. Moreover, the reader is invited to check that A is arranged in such a way that the pivots used without permutations satisfy

$$\frac{\sum_{i=k+1}^{n} |a_{ik}^{(k)}|}{|a_{kk}^{(k)}|} = \min_{k \le j \le n} \frac{\sum_{\substack{i=k \\ i \neq j}}^{n} |a_{ij}^{(k)}|}{|a_{jj}^{(k)}|}, \qquad k = 1, \ldots, n-1, \qquad (18)$$

and, therefore, the kth stage of Gaussian elimination minimizes over all possible choices of diagonal pivots the sum of the absolute values of the off diagonal entries of $L(:,k)$ for $k = 1, \ldots, n-1$. In our example, it is easy to check that there is no permutation PAP^T of A that produces a column diagonally dominant L factor with a smaller sum of the absolute values of all its off-diagonal entries. Pivoting strategies that satisfy (18) are called in [36] *column maximal relative diagonal dominance* pivoting. Despite of these properties the factor L is very sensitive to tiny structured perturbations of type (15). To this purpose, consider now the LDU factorization of the row diagonally dominant matrix $\widetilde{A} = \widetilde{L}\widetilde{D}\widetilde{U}$.

$$\widetilde{A} = \begin{bmatrix} 1000 & 101 & 500 \\ 0 & 0.1 & 0.05 \\ 100 & 10 & 120 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0.1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1000 & & \\ & 0.1 & \\ & & 70.05 \end{bmatrix} \begin{bmatrix} 1 & 0.101 & 0.5 \\ & 1 & 0.5 \\ & & 1 \end{bmatrix},$$

whose diagonally dominant parts are $v(\widetilde{A}) = [399, 0.05, 10]$. Note that A and $\widetilde{A}$ satisfy (15) with $\delta = 10^{-2}$ but that their L factors are very different, since $|\widetilde{\ell}_{32} - \ell_{32}| = 1$. Theorem 3 (d) explains this large variation because $|a_{33}^{(2)}|/|a_{22}^{(2)}| = 700 > \delta^{-1}$. The reader is invited to check that if P is the permutation matrix such that PAP^T is arranged for complete-diagonal pivoting, then the LDU factorizations of PAP^T and $P\widetilde{A}P^T$ are very close each other, as it is predicted by part (e) of Theorem 3. Observe also that L and $\widetilde{L}$ are both very well conditioned.

In the error analysis of Section 4, we will use how the diagonal entries of Schur complements vary under structured perturbations of type (15). We establish these variations in Lemma 2.

Lemma 2 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n \times n}$ be such that $v \geq 0$. Suppose* $\operatorname{rank}(A) = r$ *and that A has an* LDU *factorization as in Definition 1. Let $\widetilde{A} = \mathscr{D}(\widetilde{A}_D, \widetilde{v}) \in \mathbb{R}^{n \times n}$ be such that*

$$|\widetilde{v} - v| \leq \delta\, v \quad \textit{and} \quad |\widetilde{A}_D - A_D| \leq \delta\, |A_D|, \quad \textit{for some } 0 \leq \delta < 1.$$

Then, for $k=1,\ldots,\min\{r+1,n\}$ *and* $i=k,\ldots,n$,

$$\widetilde{a}_{ii}^{(k)}=a_{ii}^{(k)}\frac{(1+\eta_1^{(ki)})\cdots(1+\eta_k^{(ki)})}{(1+\eta_{k-1,1})\cdots(1+\eta_{k-1,k-1})},\quad \textit{where}\quad |\eta_j^{(ki)}|\leq\delta,\ |\eta_{k-1,p}|\leq\delta,$$

for $j=1,\ldots,k$ *and* $p=1,\ldots,k-1$.

3.2 Perturbation of principal minors and proofs of parts (a) and (b) of Theorem 3 and Lemma 2

We will often use that equations (15) for the perturbed matrix $\widetilde{A}=\mathscr{D}(\widetilde{A}_D,\widetilde{v})$ are equivalent to

$$\widetilde{v}_i=v_i(1+\phi_i),\quad \text{where}\quad |\phi_i|\leq\delta<1\quad \text{for}\quad i=1,\ldots,n, \tag{19}$$

$$\widetilde{a}_{ij}=a_{ij}(1+\varphi_{ij}),\quad \text{where}\quad |\varphi_{ij}|\leq\delta<1\quad \text{for}\quad i\neq j,\quad i,j=1,\ldots,n. \tag{20}$$

Observe that $1+\phi_i\geq 1-\delta>0$, for all i, and that $1+\varphi_{ij}>0$, for all $i\neq j$. So $\widetilde{v}_i\geq 0$ if and only if $v_i\geq 0$ or, equivalently, $\widetilde{A}$ is row diagonally dominant with nonnegative diagonal entries if and only if A is row diagonally dominant with nonnegative diagonal entries. As a consequence, perturbations of type (15) (equivalently, of type (19)-(20)) can be properly termed as structured-preserving perturbations in the set of row diagonally dominant matrices with nonnegative diagonal entries. This preservation property is essential in subsequent developments. We will also use that

$$|\widetilde{a}_{ij}|=|a_{ij}|(1+\varphi_{ij}). \tag{21}$$

All the perturbation results that we present are based on Lemma 3, that studies the variation of the determinant under structured perturbations of type (15).

Lemma 3 *Let* $A=\mathscr{D}(A_D,v)\in\mathbb{R}^{n\times n}$ *be such that* $v\geq 0$.

(a) *If* $B^{[i]}=\mathscr{D}(B_D^{[i]},v^{[i]})\in\mathbb{R}^{n\times n}$ *is a matrix that differs from A only in the* ith *row, i.e.,* $B^{[i]}(p,:)=A(p,:)$ *for* $p\neq i$, *and whose* ith *row parameters satisfy*

$$|v_i^{[i]}-v_i|\leq\delta\,v_i\quad \textit{and}\quad |b_{ij}^{[i]}-a_{ij}|\leq\delta\,|a_{ij}|,\quad \textit{for}\ \ j\neq i\ \textit{and}\ 0\leq\delta<1, \tag{22}$$

then

$$\det B^{[i]}=(\det A)(1+\eta_i),\quad \textit{where}\quad |\eta_i|\leq\delta.$$

(b) *If* $\widetilde{A}=\mathscr{D}(\widetilde{A}_D,\widetilde{v})\in\mathbb{R}^{n\times n}$ *is a matrix that satisfies* (15)*, then*

$$\det\widetilde{A}=(\det A)(1+\eta_1)\cdots(1+\eta_n),\quad \textit{where}\quad |\eta_k|\leq\delta\quad \textit{for}\quad k=1,\ldots,n.$$

Proof Note that $v^{[i]}\geq 0$ and $\widetilde{v}\geq 0$. The algebraic cofactors for the ith row of A and $B^{[i]}$ are equal. So, use Lemma 1 to get

$$\det B^{[i]}=v_i^{[i]}C_{ii}+\sum_{j\neq i}\left(|b_{ij}^{[i]}|C_{ii}+b_{ij}^{[i]}C_{ij}\right), \tag{23}$$

where C_{ij}, $j = 1,\ldots,n$, are cofactors of A. Observe, as in (19) and (20), that $v_i^{[i]} = v_i(1+\phi_i)$ and $b_{ij}^{[i]} = a_{ij}(1+\varphi_{ij})$, where $|\phi_i| \leq \delta$ and $|\varphi_{ij}| \leq \delta$. Recall also that $(1+\phi_i) > 0$ and $(1+\varphi_{ij}) > 0$. Then, (23) and Lemma 1 imply

$$\begin{aligned}\det B^{[i]} &= v_i(1+\phi_i)C_{ii} + \sum_{j\neq i}\left(|a_{ij}|(1+\varphi_{ij})C_{ii} + a_{ij}(1+\varphi_{ij})C_{ij}\right)\\ &= \det A + v_i\phi_i C_{ii} + \sum_{j\neq i}\varphi_{ij}\left(|a_{ij}|C_{ii} + a_{ij}C_{ij}\right),\end{aligned}$$

and

$$|\det B^{[i]} - \det A| \leq \delta\left(v_i C_{ii} + \sum_{j\neq i}\left(|a_{ij}|C_{ii} + a_{ij}C_{ij}\right)\right) = \delta\det A.$$

This is equivalent to the result in part (a).

Part (b) is a direct consequence of part (a). Consider that the perturbed matrix $\widetilde{A}$ is obtained from A through a sequence of n "only-one-row" perturbations of type (22): first only the parameters of the 1st row are modified, then the parameters of the 2nd row are modified, and so on until the parameters of the nth row are modified and we obtain $\widetilde{A}$. According to the result in part (a), in each of these "only-one-row" perturbation steps the determinant of the matrix obtained after the perturbation is equal to the determinant before the perturbation times a factor of type $1+\eta$, with $|\eta| \leq \delta$. □

Lemma 4 considers structured perturbations of principal minors of row diagonally dominant matrices with nonnegative diagonal entries. It is a corollary of Lemma 3.

Lemma 4 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be such that $v \geq 0$, let $B^{[i]} = \mathscr{D}(B_D^{[i]}, v^{[i]}) \in \mathbb{R}^{n\times n}$ be a matrix that differs from A only in the i*th *row and that satisfies* (22)*, and let $\widetilde{A} = \mathscr{D}(\widetilde{A}_D, \widetilde{v}) \in \mathbb{R}^{n\times n}$ be a matrix that satisfies* (15)*. Let $1 \leq i_1 < i_2 < \cdots < i_q \leq n$ and $\alpha = [i_1, i_2, \ldots, i_q]$, and denote the principal submatrix of A that lies in rows and columns indexed by α as $A(\alpha,\alpha)$. Then*

(a)

$$\det B^{[i]}(\alpha,\alpha) = \begin{cases} \det A(\alpha,\alpha) & \text{if } i \notin \alpha \\ (\det A(\alpha,\alpha))\left(1+\eta_i^{(\alpha)}\right) & \text{if } i \in \alpha \end{cases},$$

where $|\eta_i^{(\alpha)}| \leq \delta$;

(b)

$$\det \widetilde{A}(\alpha,\alpha) = (\det A(\alpha,\alpha))\,(1+\eta_1^{(\alpha)})\cdots(1+\eta_q^{(\alpha)}),$$

where $|\eta_k^{(\alpha)}| \leq \delta$ for $k = 1,\ldots,q$.

Proof For $B^{[i]}$, we assume that $i \in \alpha$, because otherwise the result is trivial. The discussion after (19)-(20) implies that $B^{[i]}$ and $\widetilde{A}$ are row diagonally dominant with nonnegative diagonal entries, since A has these properties. Then part (a) of Theorem 1 guarantees that $A(\alpha,\alpha)$, $B^{[i]}(\alpha,\alpha)$ and $\widetilde{A}(\alpha,\alpha)$ are also row diagonally dominant with nonnegative diagonal entries. They can be parameterized in terms of their diagonally dominant parts and off diagonal entries. Denote by w, $w^{[i]}$ and $\widetilde{w}$, respectively,

the vectors of the diagonally dominant parts of $A(\alpha,\alpha)$, $B^{[i]}(\alpha,\alpha)$ and $\widetilde{A}(\alpha,\alpha)$. Then, the parameterizations of these matrices are $A(\alpha,\alpha)=\mathscr{D}(A_D(\alpha,\alpha),w)$, $B^{[i]}(\alpha,\alpha)=\mathscr{D}(B_D^{[i]}(\alpha,\alpha),w^{[i]})$, and $\widetilde{A}(\alpha,\alpha)=\mathscr{D}(\widetilde{A}_D(\alpha,\alpha),\widetilde{w})$, whose off-diagonal entries obviously satisfy

$$|\widetilde{a}_{kj}-a_{kj}|\le\delta\,|a_{kj}| \quad\text{and}\quad |b_{is}^{[i]}-a_{is}|\le\delta\,|a_{is}|, \quad\text{for}\quad k\ne j, s\ne i,\ k,j,s\in\alpha. \tag{24}$$

For the diagonally dominant parts, observe that if the entries of w are indexed as $w=[w_{i_1},w_{i_2},\ldots,w_{i_q}]^T$, then

$$w_p=v_p+\sum_{\substack{j\ne p\\ j\notin\alpha}}|a_{pj}|, \quad\text{for}\quad p\in\alpha.$$

Use (19)-(20) to show that, for $p\in\alpha$,

$$\begin{aligned}\widetilde{w}_p&=\widetilde{v}_p+\sum_{\substack{j\ne p\\ j\notin\alpha}}|\widetilde{a}_{pj}|=v_p\,(1+\phi_p)+\sum_{\substack{j\ne p\\ j\notin\alpha}}|a_{pj}|\,(1+\varphi_{pj})\\&=w_p+v_p\,\phi_p+\sum_{\substack{j\ne p\\ j\notin\alpha}}|a_{pj}|\,\varphi_{pj}.\end{aligned}$$

Therefore, $|\widetilde{w}-w|\le\delta\,w$, and the same argument shows that $|w_i^{[i]}-w_i|\le\delta\,w_i$. These results together with (24) allow us to apply Lemma 3 to $A(\alpha,\alpha)=\mathscr{D}(A_D(\alpha,\alpha),w)$, $B^{[i]}(\alpha,\alpha)=\mathscr{D}(B_D^{[i]}(\alpha,\alpha),w^{[i]})$, and $\widetilde{A}(\alpha,\alpha)=\mathscr{D}(\widetilde{A}_D(\alpha,\alpha),\widetilde{w})$ to prove Lemma 4. □

Lemma 5 is a slightly stronger version of part (a) of Theorem 3 that may be useful in some situations.

Lemma 5 *Let $A=\mathscr{D}(A_D,v)\in\mathbb{R}^{n\times n}$ and $\widetilde{A}=\mathscr{D}(\widetilde{A}_D,\widetilde{v})\in\mathbb{R}^{n\times n}$ be such that*

$$|\widetilde{v}-v|\le\delta\,|v| \quad\textit{and}\quad |\widetilde{A}_D-A_D|\le\delta\,|A_D|, \quad\textit{for some}\ \ 0\le\delta<1.$$

(a) *Then $v\ge 0$ if and only if $\widetilde{v}\ge 0$, i.e., A is row diagonally dominant with nonnegative diagonal entries if and only if $\widetilde{A}$ is row diagonally dominant with nonnegative diagonal entries.*

(b) *Assume that $v\ge 0$. Then:* (i) $\operatorname{rank}(A)=\operatorname{rank}(\widetilde{A})$*; and* (ii) *$A$ has LDU factorization if and only if $\widetilde{A}$ has LDU factorization.*

Proof We have already seen that the discussion after (19)-(20) implies that $v\ge 0$ if and only if $\widetilde{v}\ge 0$. Assume $v\ge 0$ in the rest of the proof. Let $\widetilde{r}=\operatorname{rank}(\widetilde{A})$. According to Theorem 2, there exists a permutation matrix $\widetilde{P}$ such that $\widetilde{C}=\widetilde{P}\widetilde{A}\widetilde{P}^T$ has factorization LDU as in (5). So, by (7), $\det\widetilde{C}(1:\widetilde{r},1:\widetilde{r})\ne 0$. Define $C=\widetilde{P}A\widetilde{P}^T$, then part (b) of Lemma 4 implies

$$\det\widetilde{C}(1:\widetilde{r},1:\widetilde{r})=(\det C(1:\widetilde{r},1:\widetilde{r}))\,(1+\eta_1)\cdots(1+\eta_{\widetilde{r}}),$$

where $|\eta_j|\le\delta<1$ and, therefore, $(1+\eta_j)>0$, for $j=1,\ldots,\widetilde{r}$. In conclusion, $\det C(1:\widetilde{r},1:\widetilde{r})\ne 0$ and $\operatorname{rank}(A)\ge\operatorname{rank}(\widetilde{A})$. The same argument applied to the LDU

factorization of a certain permutation PAP^T of A leads to $\text{rank}(A) \leq \text{rank}(\widetilde{A})$. Therefore, $\text{rank}(A) = \text{rank}(\widetilde{A})$.

Now, we suppose that A has LDU factorization and prove that $\widetilde{A}$ has also LDU factorization. Let $r = \text{rank}(A)$ and note that, for $p = 1,\ldots,r$, $\det A(1:p,1:p) \neq 0$ because A has LDU factorization and equations (7) hold. Then, $\det \widetilde{A}(1:p,1:p) \neq 0$ again by Lemma 4. So, $\widetilde{A}(1:r,1:r)$ has a unique LDU factorization, that we write as $\widetilde{A}(1:r,1:r) = \widetilde{L}_{11}\widetilde{D}_{11}\widetilde{U}_{11}$ (see [30, Theorem 9.1]). Denote $\widetilde{A}_{11} := \widetilde{A}(1:r,1:r)$, partition $\widetilde{A}$ accordingly, and write the identity

$$\widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix} = \begin{bmatrix} \widetilde{L}_{11} & 0 \\ \widetilde{A}_{21}(\widetilde{D}_{11}\widetilde{U}_{11})^{-1} & I_{n-r} \end{bmatrix} \begin{bmatrix} \widetilde{D}_{11} & 0 \\ 0 & \widetilde{S}_{22} \end{bmatrix} \begin{bmatrix} \widetilde{U}_{11} & (\widetilde{L}_{11}\widetilde{D}_{11})^{-1}\widetilde{A}_{12} \\ 0 & I_{n-r} \end{bmatrix}, \quad (25)$$

with $\widetilde{S}_{22} = \widetilde{A}_{22} - \widetilde{A}_{21}\widetilde{A}_{11}^{-1}\widetilde{A}_{12}$. The fact that $\text{rank}(A) = \text{rank}(\widetilde{A})$ implies $\widetilde{S}_{22} = 0$, and so (25) provides the LDU factorization of $\widetilde{A}$. A similar argument proves that if $\widetilde{A}$ has LDU factorization then A has also LDU factorization. □

We can prove now parts (a) and (b) of Theorem 3 and Lemma 2.

Proof of part (a) *of Theorem 3* It is a corollary of Lemma 5. □

Proof of part (b) *of Theorem 3* Combine (7) and part (b) of Lemma 4. □

Proof of Lemma 2 Combine (9) and part (b) of Lemma 4. □

3.3 Perturbation of nonprincipal minors and proofs of parts (c), (d) and (e) of Theorem 3

We need Lemma 6 below to prove parts (c), (d) and (e) of Theorem 3. Note that Lemma 6 is not a perturbation result. It establishes certain technical relationships between the minors appearing as numerators in the entries of Schur complements (see (9)) and principal minors for row diagonally dominant matrices. The proof of Lemma 6 is the most complicated one in this section. We will use for simplicity the following notation:

$$g_{pq}^{(k+1)} := \det A([1:k,p],[1:k,q]), \quad (26)$$

for $k = 1,\ldots,n-1$ and $p,q = k+1,\ldots,n$. We denote as $\left(g^{[i]}\right)_{pq}^{(k+1)}$ and $\widetilde{g}_{pq}^{(k+1)}$ the corresponding minors of the perturbed matrices $B^{[i]}$ and $\widetilde{A}$ that satisfy (22) and (15), respectively.

Lemma 6 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be such that $v \geq 0$. For $k = 1,\ldots,n-2$, $p \neq q$, and $p,q = k+1,\ldots,n$, let G_{ij} be the algebraic cofactor of $A([1:k,p],[1:k,q])$ for the entry a_{ij}. Then the following statements hold for the minors defined in* (26).

(a)

$$g_{pq}^{(k+1)} = a_{p1}G_{p1} + \cdots + a_{pk}G_{pk} + a_{pq}G_{pq}, \quad \textit{and} \quad (27)$$

$$2\,g_{pp}^{(k+1)} \geq |a_{p1}G_{p1}| + \cdots + |a_{pk}G_{pk}| + |a_{pq}G_{pq}|. \quad (28)$$

(b) *For* $1 \le i \le k$,

$$g_{pq}^{(k+1)} = \left(v_i + \sum_{j \notin [1:k,q]} |a_{ij}| \right) G_{ii} + \sum_{\substack{j \neq i \\ j \in [1:k,q]}} \left(a_{ij} G_{ij} + |a_{ij}| G_{ii} \right), \quad \textit{and} \tag{29}$$

$$2\, g_{pp}^{(k+1)} \ge \left(v_i + \sum_{j \notin [1:k,q]} |a_{ij}| \right) |G_{ii}| + \sum_{\substack{j \neq i \\ j \in [1:k,q]}} \left| a_{ij} G_{ij} + |a_{ij}| G_{ii} \right|. \tag{30}$$

Proof Note first that for every row diagonally dominant matrix $C \in \mathbb{R}^{n \times n}$ with nonnegative diagonal entries, the principal submatrix $C([1:k,p,q],[1:k,p,q])$ is also row diagonally dominant with nonnegative diagonal entries. Therefore, according to parts (d) and (e) of Theorem 1,

$$\det C([1:k,p],[1:k,p]) \ge |\det C([1:k,p],[1:k,q])|. \tag{31}$$

Part (a). Equation (27) is obvious. Define a matrix $A' = \mathscr{D}(A'_D, v') \in \mathbb{R}^{n \times n}$ through its diagonally dominant parts and off-diagonal entries as follows: $v' = v$, $A'_D(l,:) = A_D(l,:)$ for $l \neq p$, $a'_{pj} = a_{pj}$ for $j \notin [1:k,q]$ and $j \neq p$, and

$$a'_{pj} = a_{pj}\,(1 + \delta\, s\, s_{pj}), \quad \text{for} \quad j \in [1:k,q],$$

where $s = \operatorname{sign}(g_{pq}^{(k+1)})$, $s_{pj} = \operatorname{sign}(a_{pj} G_{pj})$, and $0 \le \delta < 1$ is an arbitrary parameter. Observe that A' is row diagonally dominant with nonnegative diagonal entries because $v' \ge 0$ and it differs from A only in the pth row. Therefore, by (31) applied to A' and part (a) of Lemma 4, we have that

$$(1+\delta)\, g_{pp}^{(k+1)} \ge \left(g'\right)_{pp}^{(k+1)} \ge \left| \left(g'\right)_{pq}^{(k+1)} \right|. \tag{32}$$

The cofactors G_{pj} in equation (27) are equal in A that in A', so

$$\begin{aligned} \left(g'\right)_{pq}^{(k+1)} &= a'_{p1} G_{p1} + \cdots + a'_{pk} G_{pk} + a'_{pq} G_{pq} \\ &= g_{pq}^{(k+1)} + \delta\, s \left(|a_{p1} G_{p1}| + \cdots + |a_{pk} G_{pk}| + |a_{pq} G_{pq}| \right). \end{aligned}$$

Both terms in the last equation have the same sign. Combine this property with (32) to get

$$(1+\delta)\, g_{pp}^{(k+1)} \ge \delta \left(|a_{p1} G_{p1}| + \cdots + |a_{pk} G_{pk}| + |a_{pq} G_{pq}| \right), \tag{33}$$

that is valid for *any* $0 \le \delta < 1$. Note that $|a_{p1} G_{p1}| + \cdots + |a_{pk} G_{pk}| + |a_{pq} G_{pq}|$ and $g_{pp}^{(k+1)}$ do not depend on δ. Therefore the inequality (33) also holds for $\delta = 1$, by continuity, which gives (28).

Part (b). This proof follows the pattern of the one of part (a), but it is more complicated. The cofactor expansion of $g_{pq}^{(k+1)}$ along row i and the definition (11) give (29). Let $s = \operatorname{sign}(g_{pq}^{(k+1)})$ and $0 \le \delta < 1$ be an arbitrary parameter. Define an auxiliary

matrix $A' = \mathscr{D}(A'_D, v') \in \mathbb{R}^{n\times n}$ such that $A'(l,:) = A(l,:)$, for $l \neq i$, and the parameters of its ith row are:

$$\begin{aligned}
v'_i &= v_i\,(1+\delta\, s\, s_{ii}), \quad \text{where } s_{ii} = \operatorname{sign}(G_{ii}),\\
a'_{ij} &= a_{ij}\,(1+\delta\, s\, s_{ii}), \quad \text{for } j \notin [1:k,q],\\
a'_{ij} &= a_{ij}\,(1+\delta\, s\, s_{ij}), \quad \text{for } j \in [1:k,q],\ j \neq i, \text{ where } s_{ij} = \operatorname{sign}\left(a_{ij}G_{ij} + |a_{ij}|G_{ii}\right).
\end{aligned}$$

Observe that this matrix A' also satisfies (32) as the one in the proof of part (a). In addition, the cofactors G_{ii} and G_{ij} in equation (29) are equal in A that in A' and, for $j \in [1:k,q]$, $j \neq i$, $\left(a'_{ij}G_{ij} + |a'_{ij}|G_{ii}\right) = (1+\delta\, s\, s_{ij})\left(a_{ij}G_{ij} + |a_{ij}|G_{ii}\right)$ by (21). So

$$\begin{aligned}
\left(g'\right)^{(k+1)}_{pq} &= \left(v'_i + \sum_{j\notin[1:k,q]} |a'_{ij}|\right) G_{ii} + \sum_{\substack{j\neq i\\ j\in[1:k,q]}} \left(a'_{ij}G_{ij} + |a'_{ij}|G_{ii}\right),\\
&= g^{(k+1)}_{pq} + \delta\, s \left[\left(v_i + \sum_{j\notin[1:k,q]} |a_{ij}|\right)|G_{ii}| + \sum_{\substack{j\neq i\\ j\in[1:k,q]}} \left|a_{ij}G_{ij} + |a_{ij}|G_{ii}\right|\right].
\end{aligned}$$

Again, both terms in the last equation have the same sign. Combine this property with (32) to get

$$(1+\delta)\, g^{(k+1)}_{pp} \geq \delta \left[\left(v_i + \sum_{j\notin[1:k,q]} |a_{ij}|\right)|G_{ii}| + \sum_{\substack{j\neq i\\ j\in[1:k,q]}} \left|a_{ij}G_{ij} + |a_{ij}|G_{ii}\right|\right].$$

An argument similar to that in part (a) shows that this inequality holds for $\delta = 1$, which gives (30). □

Lemma 7 studies the variations of the nonprincipal minors appearing in the entries (9) of the Schur complements under structured perturbations of type (15).

Lemma 7 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be such that $v \geq 0$, let $B^{[i]} = \mathscr{D}(B^{[i]}_D, v^{[i]}) \in \mathbb{R}^{n\times n}$ be a matrix that differs from A only in the i*th *row and that satisfies* (22)*, and let $\widetilde{A} = \mathscr{D}(\widetilde{A}_D, \widetilde{v}) \in \mathbb{R}^{n\times n}$ be a matrix that satisfies* (15)*. Then, for $k = 1,\ldots,n-2$, $p \neq q$, and $p,q = k+1,\ldots,n$, the following statements hold for the minors defined in* (26)*.*

(a)

$$\left|\left(g^{[i]}\right)^{(k+1)}_{pq} - g^{(k+1)}_{pq}\right| \leq \begin{cases} 0 & \text{if } i \notin [1:k,p]\\ 2\,\delta\, g^{(k+1)}_{pp} & \text{if } i \in [1:k,p]\end{cases}.$$

(b)

$$|\widetilde{g}^{(k+1)}_{pq} - g^{(k+1)}_{pq}| \leq 2\left((1+\delta)^{k+1} - 1\right) g^{(k+1)}_{pp}.$$

Proof Let us start with *Part* (a). Assume that $i \in [1:k,p]$, because otherwise the result is trivial. First we consider $1 \leq i \leq k$ and use (29) to get

$$\left(g^{[i]}\right)_{pq}^{(k+1)} = \left(v_i^{[i]} + \sum_{j \notin [1:k,q]} |b_{ij}^{[i]}|\right) G_{ii} + \sum_{\substack{j \neq i \\ j \in [1:k,q]}} \left(b_{ij}^{[i]} G_{ij} + |b_{ij}^{[i]}| G_{ii}\right),$$

since the cofactors G_{ii} and G_{ij} are equal in A and in $B^{[i]}$. Observe, as in (19) and (20), that $v_i^{[i]} = v_i(1+\phi_i)$ and $b_{ij}^{[i]} = a_{ij}(1+\varphi_{ij})$, where $|\phi_i| \leq \delta$ and $|\varphi_{ij}| \leq \delta$. So, by (21), $\left(b_{ij}^{[i]} G_{ij} + |b_{ij}^{[i]}| G_{ii}\right) = (1+\varphi_{ij})\left(a_{ij} G_{ij} + |a_{ij}| G_{ii}\right)$, and

$$\left(g^{[i]}\right)_{pq}^{(k+1)} = g_{pq}^{(k+1)} + \left(\phi_i v_i + \sum_{j \notin [1:k,q]} \varphi_{ij} |a_{ij}|\right) G_{ii} + \sum_{\substack{j \neq i \\ j \in [1:k,q]}} \varphi_{ij} \left(a_{ij} G_{ij} + |a_{ij}| G_{ii}\right),$$

and

$$\begin{aligned} \left|\left(g^{[i]}\right)_{pq}^{(k+1)} - g_{pq}^{(k+1)}\right| &\leq \delta \left[\left(v_i + \sum_{j \notin [1:k,q]} |a_{ij}|\right) |G_{ii}| + \sum_{\substack{j \neq i \\ j \in [1:k,q]}} \left|a_{ij} G_{ij} + |a_{ij}| G_{ii}\right|\right] \\ &\leq \delta\, 2\, g_{pp}^{(k+1)}, \end{aligned}$$

by (30). This proves part (a) if $1 \leq i \leq k$. For $i = p$, use (27) and (28) and follow a similar argument.

Now, we prove *Part* (b). As in the proof of part (b) of Lemma 3, we consider again that $\widetilde{A}$ is obtained from A as a sequence of n "only-one-row" perturbations of type (22). Note that all matrices in this sequence are row diagonally dominant with nonnegative diagonal entries. Obviously, the variation of $g_{pq}^{(k+1)}$ is consequence only of the perturbations of rows with indices in $[1:k,p]$. Let α be a subset of $[1:k,p]$ and denote by $(g^{\alpha})_{pq}^{(k+1)}$ the minor (26) corresponding to a matrix obtained from A through perturbations of the rows with indices in α, while the remaining rows remain unchanged. Observe that $\widetilde{g}_{pq}^{(k+1)} = \left(g^{[1:k,p]}\right)_{pq}^{(k+1)}$. So,

$$\begin{aligned} |\widetilde{g}_{pq}^{(k+1)} - g_{pq}^{(k+1)}| &\leq \left|\widetilde{g}_{pq}^{(k+1)} - \left(g^{[1:k]}\right)_{pq}^{(k+1)}\right| + \left|\left(g^{[1:k]}\right)_{pq}^{(k+1)} - \left(g^{[1:k-1]}\right)_{pq}^{(k+1)}\right| + \cdots \\ &\quad + \left|\left(g^{[1]}\right)_{pq}^{(k+1)} - g_{pq}^{(k+1)}\right| \\ &\leq 2\,\delta \left(\left(g^{[1:k]}\right)_{pp}^{(k+1)} + \left(g^{[1:k-1]}\right)_{pp}^{(k+1)} + \cdots + \left(g^{[1]}\right)_{pp}^{(k+1)} + g_{pp}^{(k+1)}\right), \end{aligned}$$

where the last inequality follows from applying part (a) to each matrix in the sequence of "only-one-row" perturbations. Apply part (a) of Lemma 4 iteratively to obtain

$$|\widetilde{g}_{pq}^{(k+1)} - g_{pq}^{(k+1)}| \leq 2\,\delta \left((1+\delta)^k + (1+\delta)^{k-1} + \cdots + (1+\delta) + 1\right) g_{pp}^{(k+1)},$$

which proves the result. □

We can prove now parts (c), (d) and (e) of Theorem 3.

Proof of part (c) *of Theorem 3* By using $u_{1j} = a_{1j}/a_{11}$, $j > 1$, the reader may easily check that the off-diagonal entries of the first rows of U and $\widetilde{U}$ satisfy $\widetilde{u}_{1j} = u_{1j}(1+\varphi_{1j})/(1+\varphi_{11})$, where $|\varphi_{1s}| \leq \delta$, for $s = 1,\ldots,n$. This implies $|\widetilde{u}_{1j} - u_{1j}| \leq 2\delta$, because A and $\widetilde{A}$ are both row diagonally dominant and so $|\widetilde{u}_{1j}| \leq 1$ and $|u_{1j}| \leq 1$.

To study the variation of the entries u_{ij} with $j > i > 1$, consider again that $\widetilde{A}$ is obtained from A as a sequence of n "only-one-row" perturbations of type (22). Observe, by (8), that the variation of u_{ij} depends only on the perturbations of rows $1,\ldots,i$ of A. Let $B^{[s]}$, with $s \in [1:i]$, be a matrix that differs from A only in the sth row and that satisfies (22) for this row. Let $u_{ij}^{[s]}$ be the entries of the U factor of the LDU factorization of $B^{[s]}$. This factorization exists by part (a) of Theorem 3 applied to $B^{[s]}$. Use (8), (26), part (a) of Lemma 4 and part (a) of Lemma 7 to show

$$u_{ij}^{[s]} = \frac{\left(g^{[s]}\right)_{ij}^{(i)}}{\left(g^{[s]}\right)_{ii}^{(i)}} = \frac{g_{ij}^{(i)} + 2\,\xi_1\, g_{ii}^{(i)}}{g_{ii}^{(i)}\,(1+\xi_2)} = \frac{u_{ij} + 2\,\xi_1}{(1+\xi_2)},$$

where $|\xi_1| \leq \delta$ and $|\xi_2| \leq \delta$. Therefore $u_{ij}^{[s]} = u_{ij} + 2\,\xi_1 - \xi_2\, u_{ij}^{[s]}$. The matrix $B^{[s]}$ is row diagonally dominant and, so, $u_{ij}^{[s]} \leq 1$. As a consequence

$$|u_{ij}^{[s]} - u_{ij}| \leq 3\,\delta, \tag{34}$$

i.e., an "only-one-row" perturbation causes in u_{ij} an absolute variation of at most 3δ. Let α be a subset of $[1:i]$ and denote by u_{ij}^{α} the entries of the U factor of a matrix obtained from A through perturbations of the rows with indices in α, while the remaining rows remain unchanged. Note that $\widetilde{u}_{ij} = u_{ij}^{[1:i]}$ and that

$$|\widetilde{u}_{ij} - u_{ij}| \leq |\widetilde{u}_{ij} - u_{ij}^{[1:i-1]}| + |u_{ij}^{[1:i-1]} - u_{ij}^{[1:i-2]}| + \cdots + |u_{ij}^{[1]} - u_{ij}| \leq 3\,i\,\delta,$$

where we have used (34) and the fact that all the matrices in the sequence of "only-one-row" perturbations are row diagonally dominant. □

Proof of part (d) *of Theorem 3* This proof does not follow the pattern of the one of part (c) because the entries $|\ell_{ij}|$ are not bounded by 1. Instead, use (6), (26), part (b) of Lemma 4 and part (b) of Lemma 7 to show

$$\widetilde{\ell}_{ij} = \frac{\widetilde{g}_{ij}^{(j)}}{\widetilde{g}_{jj}^{(j)}} = \frac{g_{ij}^{(j)} + 2\,\chi\, g_{ii}^{(j)}}{g_{jj}^{(j)}\,(1+\xi_1)\cdots(1+\xi_j)}, \tag{35}$$

where $|\xi_1| \leq \delta, \ldots, |\xi_j| \leq \delta$ and $|\chi| \leq \left((1+\delta)^j - 1\right)$. Define

$$\zeta := \frac{1}{(1+\xi_1)\cdots(1+\xi_j)} - 1, \quad \text{and note} \quad |\zeta| \leq \frac{1}{(1-\delta)^j} - 1.$$

So, from (35) and (9), we get

$$\widetilde{\ell}_{ij} = \left(\ell_{ij} + 2\chi \, \frac{a_{ii}^{(j)}}{a_{jj}^{(j)}} \right) (1+\zeta) \quad \text{and} \quad \widetilde{\ell}_{ij} - \ell_{ij} = \zeta \, \ell_{ij} + 2\chi \, (1+\zeta) \, \frac{a_{ii}^{(j)}}{a_{jj}^{(j)}}.$$

Now, take absolute values, triangular inequalities and get part (d) of Theorem 3. □

Proof of part (e) *of Theorem 3* It follows directly from the use of (13), with δ instead of $\mathbf{u}$, in the bound of part (d). The only point to remark is that $|\ell_{ij}| = \left| a_{ij}^{(j)} / a_{jj}^{(j)} \right| \leq 1 + \beta_j$, because the Schur complement $A^{(j)}(j:n, j:n)$ is row diagonally dominant, so its largest entry in absolute value is on the diagonal, i.e., $|a_{ij}^{(j)}| \leq |a_{ii}^{(j)}|$. □

4 Error analysis of Q. Ye's algorithm for the LDU factorization

This section is organized as follows. In Subsection 4.1 we state without proofs the main results we have obtained for the errors committed by Algorithm 1 in Q. Ye's paper [42]. The proofs are presented in Subsections 4.2 and 4.3. In the first order error analysis included in [42], all summations of nonnegative numbers are performed with the method of compensated summation [30, Sec. 4.3]. In contrast, we use only standard summation here and the error bounds we present are rigorous, i.e., they do not neglect any high-order terms. The improvements that may be obtained from the use of compensated summation are briefly explained in Subsection 4.4.

Recall that Algorithm 1 in [42] computes the LDU factorization of a row diagonally dominant matrix A with nonnegative diagonal entries, *for which its diagonally dominant parts v and off-diagonal entries A_D are known*. The most important feature of this algorithm is that in the kth stage of Gaussian elimination the parameters $(A_D^{(k+1)}, v^{(k+1)})$ of $A^{(k+1)}$ are obtained from the parameters $(A_D^{(k)}, v^{(k)})$ of $A^{(k)}$ in such a way that each entry of $v^{(k+1)}$ is a sum of nonnegative terms.

Our ability to obtain tiny error bounds relies on the perturbation results in Theorem 3. As a consequence, the relative error bounds we get for the entries of D and the absolute error bounds we get for the entries of U hold *for any diagonal pivoting strategy*, while the absolute error bounds we get for the entries of L hold *only for complete-diagonal pivoting*. These strategies were introduced in Section 2.

4.1 Main rounding error results and comments

Let P be the permutation matrix constructed in floating point arithmetic by Algorithm 1 in [42] when it is applied to a matrix A with a certain diagonal pivoting strategy. *For simplicity, we assume in the error analysis that the matrix A has been permuted by P in advance, so that no permutations are needed in the process.* The reader can find a detailed description of the first stage of Algorithm 1 in [42] in Section 4.2 (just before Lemma 8). The remaining stages consist in applying exactly the same procedure on the corresponding Schur complements.

Theorem 4 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be a row diagonally dominant matrix with nonnegative diagonal entries and assume that $v \geq 0$ and A_D are known. Let $\widehat{L}$, $\widehat{D} = diag(\widehat{d}_1, \ldots, \widehat{d}_n)$ and $\widehat{U}$ be the LDU factors of A computed by Algorithm 1 in [42] in a computer with unit roundoff $\mathbf{u}$ and let L, $D = diag(d_1, \ldots, d_n)$ and U be the exact factors of A. Suppose that $36n^3\mathbf{u} < 1$. Then*

(a) *For any diagonal pivoting strategy*

$$|\widehat{d}_i - d_i| \leq |d_i| \frac{6ni^2\mathbf{u}}{1-6ni^2\mathbf{u}} \leq |d_i| \frac{6n^3\mathbf{u}}{1-6n^3\mathbf{u}}, \qquad i = 1, \ldots, n;$$

(b) *For any diagonal pivoting strategy*

$$|\widehat{u}_{ij} - u_{ij}| \leq 8ni^2\mathbf{u} < 8n^3\mathbf{u}, \qquad 1 \leq i < j \leq n;$$

(c) *For the complete-diagonal pivoting strategy*

$$|\widehat{\ell}_{ij} - \ell_{ij}| \leq 14nj^2\mathbf{u} < 14n^3\mathbf{u}, \qquad 1 \leq j < i \leq n.$$

Remark 1 Observe that part (a) implies that $\widehat{d}_i = 0$ if and only if $d_i = 0$, so Algorithm 1 in [42] determines exactly the rank of A. This was also shown in [42]. Therefore, if $\mathrm{rank}(A) = r < n$, then Algorithm 1 in [42] stops in floating point arithmetic after r stages of Gaussian elimination have been performed. As a consequence, $\widehat{u}_{ij} = u_{ij} = 0$ for $r+1 \leq i < j \leq n$, and $\widehat{\ell}_{ij} = \ell_{ij} = 0$ for $r+1 \leq j < i \leq n$.

From Theorem 4 and the fact that L and U are unit triangular, we can get the following normwise relative error bounds for L and U:

$$\frac{\|\widehat{L}-L\|_M}{\|L\|_M} \leq 14n^3\mathbf{u}, \qquad \frac{\|\widehat{U}-U\|_M}{\|U\|_M} \leq 8n^3\mathbf{u}, \tag{36}$$

$$\frac{\|\widehat{L}-L\|_1}{\|L\|_1} \leq \frac{56}{27}n^4\mathbf{u}, \qquad \frac{\|\widehat{U}-U\|_\infty}{\|U\|_\infty} \leq \frac{32}{27}n^4\mathbf{u}. \tag{37}$$

The n^4 factor in (37) can be replaced by n^3 if compensated summation is used in Algorithm 1 in [42] (see Subsection 4.4), at the cost of modifying somewhat the numerical constants and getting only first order error bounds in $\mathbf{u}$.

4.2 Rounding error analysis of the first stage of Q. Ye's algorithm

The proof of Theorem 4 follows an inductive argument on the size of the matrix. For the error bound on the D factor this argument has the same flavor as the one in [35], but for U and L the proofs are different. Before performing the induction, we need to carefully analyze the errors committed in the first stage of Algorithm 1 in [42]. The perturbation theory developed in Section 3 plays a stellar role in this process, in particular in Lemma 9 below.

In order to simplify the presentation we introduce one more bit of notation. In Section 2, $A^{(2)} = [a_{ij}^{(2)}] \in \mathbb{R}^{n\times n}$ denoted the matrix obtained after the first stage of Gaussian elimination is applied to $A^{(1)} := A \in \mathbb{R}^{n\times n}$. We will denote in this section

$$\mathscr{A}^{(2)} := A^{(2)}(2:n,2:n) \in \mathbb{R}^{(n-1)\times(n-1)}, \tag{38}$$

i.e., the Schur complement of a_{11} in A. This is the submatrix of $A^{(2)}$ that is active in the second stage of Gaussian elimination. The entries of the matrix $\mathscr{A}^{(2)} = [a_{ij}^{(2)}]_{i,j=2}^{n}$ will be indexed from 2 to n. In addition, it is well known [17] that $\mathscr{A}^{(2)}$ satisfies the following two properties that are fundamental in the sequel:

1. If $\mathscr{A}^{(2)} = L_{22}D_{22}U_{22}$ is the LDU factorization of $\mathscr{A}^{(2)}$ and $A = LDU$ is the LDU factorization of A, then $L_{22} = L(2:n,2:n)$, $D_{22} = D(2:n,2:n)$ and $U_{22} = U(2:n,2:n)$.
2. If $k \geq 2$, then the kth Schur complement $[a_{ij}^{(k)}]_{i,j=k}^{n}$ of A (recall (9)) is the $(k-1)$th Schur complement of $\mathscr{A}^{(2)}$.

To begin with, we recall how Algorithm 1 in [42] performs the first stage of Gaussian elimination. The inputs are the parameters $(A_D^{(1)}, v^{(1)}) := (A_D, v)$, $v \geq 0$, corresponding to $A^{(1)} := A$. The outputs are: (a) the first entry d_1 of D; (b) the first column of L; (c) the first row of U; and, (d) the parameters $(\mathscr{A}_D^{(2)}, v^{(2)}(2:n))$ corresponding to $\mathscr{A}^{(2)}$. These outputs are computed as follows[4]:

1. $a_{ii}^{(1)} = v_i^{(1)} + \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}^{(1)}|$, for $i = 1,\ldots,n$, and $d_1 = a_{11}^{(1)}$;
2. $\ell_{i1} = a_{i1}^{(1)}/a_{11}^{(1)}$ and $u_{1i} = a_{1i}^{(1)}/a_{11}^{(1)}$, for $i = 2,\ldots,n$;
3. $a_{ij}^{(2)} = a_{ij}^{(1)} - \ell_{i1}a_{1j}^{(1)}$ for $i,j = 2,\ldots,n$, $i \neq j$;
4. For $i = 2,\ldots,n$ (see also [42, Theorem 1])

$$v_i^{(2)} = v_i^{(1)} + \sum_{\substack{j=2 \\ j\neq i}}^{n} (1 - s_{ij}^{(1)})\,|a_{ij}^{(1)}| + |\ell_{i1}|\,v_1^{(1)} + \sum_{j=2}^{n} (1 - t_{ij}^{(1)})\,|\ell_{i1}|\,|a_{1j}^{(1)}|, \tag{39}$$

where $\quad s_{ij}^{(1)} = \operatorname{sign}\left(a_{ij}^{(2)}\right)\operatorname{sign}\left(a_{ij}^{(1)}\right)$, $\quad$ and

$$t_{ij}^{(1)} = \begin{cases} -\operatorname{sign}\left(a_{ij}^{(2)}\right)\operatorname{sign}\left(a_{i1}^{(1)}\right)\operatorname{sign}\left(a_{1j}^{(1)}\right), & \text{if } i \neq j \\ \operatorname{sign}\left(a_{i1}^{(1)}\right)\operatorname{sign}\left(a_{1i}^{(1)}\right), & \text{if } i = j. \end{cases}$$

Observe that $v_i^{(2)}$ is a a sum of nonnegative terms because $v^{(1)} \geq 0$.

The second stage of Algorithm 1 in [42] applies the same to $\mathscr{D}(\mathscr{A}_D^{(2)}, v^{(2)}(2:n))$ and so on.

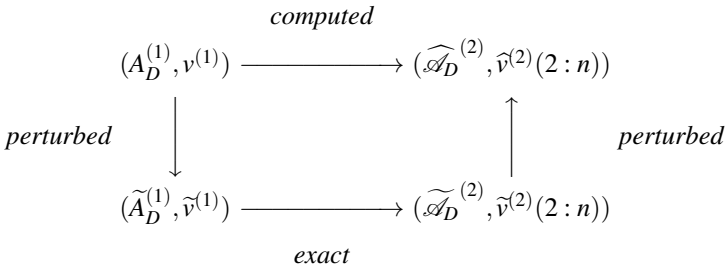
Lemma 8 establishes the rounding errors for the first stage of Algorithm 1 in [42]. The notation introduced in (14) is used in this lemma and in the rest of this section.

[4] Note that Algorithm 1 in [42] needs to compute all the diagonal entries to determine the pivot, but the only entry that is an output is $d_1 = a_{11}^{(1)}$ because we assume that A has been permuted in advance.

Lemma 8 *Let $\widehat{a}_{ii}^{(1)}$ for $i=1,\dots,n$, $\widehat{\ell}_{i1}$, $\widehat{u}_{1i}$, for $i=2,\dots,n$, and $(\widehat{\mathscr{A}_D}^{(2)},\widehat{v}^{(2)}(2:n))$ be the quantities computed in the first stage of Algorithm 1 in [42] applied to $(A_D^{(1)},v^{(1)})$, where $A=\mathscr{D}(A_D^{(1)},v^{(1)})\in\mathbb{R}^{n\times n}$. Let $a_{ii}^{(1)}$, ℓ_{i1}, and u_{1i} be corresponding exact quantities. Then the following statements hold.*

(a) $\widehat{a}_{ii}^{(1)}=a_{ii}^{(1)}\ <n-1>$. *In particular,* $\widehat{d}_1=d_1\ <n-1>$.
(b) $\widehat{\ell}_{i1}=\ell_{i1}\ <n>$ *and* $\widehat{u}_{1i}=u_{1i}\ <n>$.
(c) *The computation of $(\widehat{\mathscr{A}_D}^{(2)},\widehat{v}^{(2)}(2:n))$ is equivalent to the following sequence of operations:*
   1. *Multiply each input parameter in $(A_D^{(1)},v^{(1)})$ by a factor $<n+1>$, getting $(\widetilde{A}_D^{(1)},\widetilde{v}^{(1)})$. More precisely this means that $\widetilde{v}_i^{(1)}=v_i^{(1)}\ <n+1>$ for $i=1,\dots,n$, and $\widetilde{a}_{ij}^{(1)}=a_{ij}^{(1)}\ <n+1>$ for $i\neq j$, $i,j=1,\dots,n$, where the factors $<n+1>$ may be different for each parameter.*
   2. *Obtain in exact arithmetic the parameters $(\widetilde{\mathscr{A}_D}^{(2)},\widetilde{v}^{(2)}(2:n))$ corresponding to $(\widetilde{A}_D^{(1)},\widetilde{v}^{(1)})$.*
   3. *Multiply each parameter of $(\widetilde{\mathscr{A}_D}^{(2)},\widetilde{v}^{(2)}(2:n))$ by a factor $<4n>$, getting $(\widehat{\mathscr{A}_D}^{(2)},\widehat{v}^{(2)}(2:n))$.*

   This sequence of operations is represented graphically as follows:



Proof Parts (a) and (b) follows from standard accumulation of rounding errors and the fact that $a_{ii}^{(1)}$ is obtained as a summation of nonnegative terms. Simply follow [30, Section 3.1] and note $a_{ii}^{(1)}(1-\mathbf{u})^{n-1}\le\widehat{a}_{ii}^{(1)}\le a_{ii}^{(1)}(1+\mathbf{u})^{n-1}$.

For part (c), we will show that the parameters $(\widetilde{A}_D^{(1)},\widetilde{v}^{(1)})$ are as follows:

$$\widetilde{v}^{(1)}=v^{(1)}\quad\text{and, for } i\neq j,\quad \widetilde{a}_{ij}^{(1)}=\begin{cases} a_{ij}^{(1)} & \text{if } i=1 \text{ or } j=1\\ a_{ij}^{(1)}\ <n+1> & \text{otherwise.}\end{cases}\tag{40}$$

To this purpose, let $1+\xi_{ij} = <n+1>$, and note that for $i \neq j$, with $2 \leq i,j \leq n$,

$$\begin{aligned}
\widehat{a}_{ij}^{(2)} &= \left(a_{ij}^{(1)} - \frac{a_{i1}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}} (1+\xi_{ij}) \right) <1> \\
&= \left(\frac{a_{ij}^{(1)}}{1+\xi_{ij}} - \frac{a_{i1}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}} \right) (1+\xi_{ij}) <1> = \left(\widetilde{a}_{ij}^{(1)} - \frac{a_{i1}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}} \right) <n+2> \\
&= \widetilde{a}_{ij}^{(2)} <n+2>,
\end{aligned}$$

where $\widetilde{a}_{ij}^{(1)} := a_{ij}^{(1)}/(1+\xi_{ij})$. Therefore $\widehat{\mathscr{A}_D}^{(2)}$ satisfies part (c) for $(\widetilde{A}_D^{(1)}, \widetilde{v}^{(1)})$ defined in (40). Next, we prove that $\widehat{v}^{(2)}(2:n)$ also satisfies part (c) for the same $(\widetilde{A}_D^{(1)}, \widetilde{v}^{(1)})$.

Consider the computed signs $\widehat{s}_{ij}^{(1)}$ and $\widehat{t}_{ij}^{(1)}$ appearing in (39), the corresponding exact signs $\widetilde{s}_{ij}^{(1)}$ and $\widetilde{t}_{ij}^{(1)}$ for $(\widetilde{A}_D^{(1)}, \widetilde{v}^{(1)})$, and observe

$$\widehat{s}_{ij}^{(1)} = \widetilde{s}_{ij}^{(1)} \quad \text{and} \quad \widehat{t}_{ij}^{(1)} = \widetilde{t}_{ij}^{(1)}.$$

Now, for $i = 2, \ldots, n$, define the auxiliary variables

$$w_i^{(2)} = v_i^{(1)} + \sum_{\substack{j=2 \\ j \neq i}}^{n} (1 - \widetilde{s}_{ij}^{(1)}) \, |a_{ij}^{(1)}| + |\ell_{i1}| \, v_1^{(1)} + \sum_{j=2}^{n} (1 - \widetilde{t}_{ij}^{(1)}) \, |\ell_{i1}| \, |a_{1j}^{(1)}|,$$

and proceed as in [30, Section 3.1] to show[5]

$$\widehat{v}_i^{(2)} = w_i^{(2)} <3n-1>. \tag{41}$$

On the other hand, by (40), the exact entries of $\widetilde{v}^{(2)}(2:n)$ for $(\widetilde{A}_D^{(1)}, \widetilde{v}^{(1)})$ are

$$\widetilde{v}_i^{(2)} = v_i^{(1)} + \sum_{\substack{j=2 \\ j \neq i}}^{n} (1 - \widetilde{s}_{ij}^{(1)}) \, |\widetilde{a}_{ij}^{(1)}| + |\ell_{i1}| \, v_1^{(1)} + \sum_{j=2}^{n} (1 - \widetilde{t}_{ij}^{(1)}) \, |\ell_{i1}| \, |a_{1j}^{(1)}|.$$

It is straightforward to prove that $\widetilde{v}_i^{(2)} = w_i^{(2)} <n+1>$. Combine this with (41) to get $\widehat{v}_i^{(2)} = \widetilde{v}_i^{(2)} <4n>$, which proves the result. □

A trivial but key observation in subsequent developments is that the *LDU* factors computed by Algorithm 1 in [42] applied to $(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n))$ are precisely

$$\widehat{L}(2:n,2:n) =: \widehat{L}_{22}, \; \widehat{D}(2:n,2:n) =: \widehat{D}_{22}, \; \widehat{U}(2:n,2:n) =: \widehat{U}_{22}, \tag{42}$$

where recall that $\widehat{L}$, $\widehat{D}$ and $\widehat{U}$ are the computed factors of $A = \mathscr{D}(A_D^{(1)}, v^{(1)}) \in \mathbb{R}^{n \times n}$. Related with this fact, we will also need the *exact LDU factors* of the matrix $\mathscr{A}_2' := \mathscr{D}(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)) \in \mathbb{R}^{(n-1)\times(n-1)}$, that are denoted by

$$\mathscr{A}_2' = \mathscr{D}(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)) = L_{22}' D_{22}' U_{22}', \tag{43}$$

[5] As in [42], we start the computation of $v_i^{(2)}$ with the summation $v_i^{(1)} + |\ell_{i1}| v_1^{(1)}$.

and their entries are indexed from 2 to n. These exact factors are needed because the essential induction hypothesis in Section 4.3 is that Algorithm 1 in [42] is accurate for matrices of size $(n-1)\times(n-1)$, and so that $\widehat{L}_{22}$, $\widehat{D}_{22}$ and $\widehat{U}_{22}$ are, respectively, close to L'_{22}, D'_{22}, and U'_{22} in a sense that will be made precise later. Our goal now is to use the perturbation theory in Section 3 for relating in Lemma 9, L'_{22}, D'_{22}, and U'_{22} with $L(2:n,2:n)$, $D(2:n,2:n)$ and $U(2:n,2:n)$, where L, D and U are the exact factors of $A=\mathscr{D}(A_D^{(1)},v^{(1)})\in\mathbb{R}^{n\times n}$. For technical reasons that will be clear in Section 4.3, we also relate in Lemma 9 the exact diagonal entries of the Schur complements of A with the exact diagonal entries of the Schur complements of $\mathscr{A}'_2$.

Lemma 9 *Use the same notation and assumptions as in Lemma 8, and denote* $r=\operatorname{rank}\left(\mathscr{D}(A_D^{(1)},v^{(1)})\right)>0$. *Let L, D and U be the exact LDU factors of* $A=\mathscr{D}(A_D^{(1)},v^{(1)})$, *let* $L'_{22}=[\ell'_{ij}]_{i,j=2}^{n}$, $D'_{22}=\operatorname{diag}(d'_2,\ldots,d'_n)$ *and* $U'_{22}=[u'_{ij}]_{i,j=2}^{n}$ *be the exact LDU factors of* $\mathscr{A}'_2$ *in* (43), *and let* $\widetilde{A}:=\mathscr{D}(\widetilde{A}_D^{(1)},\widetilde{v}^{(1)})$. *For* $2\le k\le\min\{r+1,n\}$, *let* $[a_{ii}^{(k)}]_{i=k}^{n}$, $[\widetilde{a}_{ii}^{(k)}]_{i=k}^{n}$, *and* $[a_{ii}^{\prime(k)}]_{i=k}^{n}$ *be, respectively, the exact diagonal entries of the Schur complements of A,* $\widetilde{A}$, *and* $\mathscr{A}'_2$. *Then the following statements hold.*

(a) $\operatorname{rank}(A)=\operatorname{rank}(\widetilde{A})=1+\operatorname{rank}(\mathscr{A}'_2)$.
(b) *For* $2\le k\le\min\{r+1,n\}$ *and* $k\le i\le n$

$$a_{ii}^{\prime(k)}=a_{ii}^{(k)}\ <10nk-13n+2k-1>\ .$$

In particular, $d'_p=d_p\ <10np-13n+2p-1>$ *for* $2\le p\le n$, *since* $d_p=a_{pp}^{(p)}$.
(c) *For* $2\le i\le n-1$ *and* $j>i$

$$|u'_{ij}-u_{ij}|\le\mathbf{u}\,(15ni-12n+3i).$$

(d) *In addition, let* $r'=\min\{r,n-1\}$ *and* $\beta_2,\beta_3,\ldots,\beta_{r'}$ *be numbers such that* $0\le\beta_k$, $(1+\beta_k)\,|a_{kk}^{(k)}|\ge|a_{ii}^{(k)}|$ *and* $(1+\beta_k)\,|\widetilde{a}_{kk}^{(k)}|\ge|\widetilde{a}_{ii}^{(k)}|$, *for* $k=2,\ldots,r'$ *and* $k<i\le n$. *Let* $36n^3\mathbf{u}<1$. *Then, for* $2\le j\le n-1$ *and* $i>j$,

$$|\ell'_{ij}-\ell_{ij}|\le\mathbf{u}\,(1+\beta_j)\,\frac{2}{3}\,(27nj-22n+5j).$$

Proof Parts (a), (b) and (c) are direct consequences of Theorem 3 and Lemma 2. We simply sketch the proofs. Throughout this proof $\widetilde{L}$, $\widetilde{D}$ and $\widetilde{U}$ denote the exact *LDU* factors of $\widetilde{A}:=\mathscr{D}(\widetilde{A}_D^{(1)},\widetilde{v}^{(1)})$.

Note first that according to Lemma 8, one can consider that $\widetilde{A}$ is obtained from A by applying a sequence of $n+1$ perturbations that can be: (1) of type (15) with $\delta=\mathbf{u}$ (recall (19) and (20)); or, (2) perturbations that are reversals of type (15) (i.e., with the roles of the matrices A and $\widetilde{A}$ in (15) exchanged) with $\delta=\mathbf{u}$ and that take into account the factors $(1+\psi_i)^{-1}$ in $<n+1>$. Analogously, $\mathscr{A}'_2$ is obtained from $\mathscr{D}(\widetilde{\mathscr{A}_D}^{(2)},\widetilde{v}^{(2)}(2:n))$ as a sequence of $4n$ perturbations of type (15) with $\delta=\mathbf{u}$ or reversals of them.

Note that part (a) of Theorem 3 and the equality $\mathrm{rank}(\widetilde{A}) = 1 + \mathrm{rank}(\mathscr{D}(\widetilde{\mathscr{A}_D}^{(2)}, \widetilde{v}^{(2)}(2:n)))$ imply part (a). Lemma 2 implies

$$\widetilde{a}_{ii}^{(k)} = a_{ii}^{(k)} \; <(2k-1)(n+1)> \quad \text{and} \quad a_{ii}^{\prime(k)} = \widetilde{a}_{ii}^{(k)} \; <(2k-3)4n>, \tag{44}$$

where in the last equality we have used that $[\widetilde{a}_{ij}^{(k)}]_{i,j=k}^{n}$ is the $(k-1)$th Schur complement of $\mathscr{D}(\widetilde{\mathscr{A}_D}^{(2)}, \widetilde{v}^{(2)}(2:n))$. Therefore

$$a_{ii}^{\prime(k)} = a_{ii}^{(k)} \; <(2k-1)(n+1) + (2k-3)4n>,$$

which is the result in part (b). Part (c) of Theorem 3 implies

$$|\widetilde{u}_{ij} - u_{ij}| \leq 3\,(n+1)\,i\,\mathbf{u} \quad \text{and} \quad |\widetilde{u}_{ij} - u'_{ij}| \leq 12\,n\,(i-1)\,\mathbf{u},$$

where in the last inequality we have used that $\widetilde{u}_{ij}$ corresponds to the $(i-1)$th row of the U factor of $\mathscr{D}(\widetilde{\mathscr{A}_D}^{(2)}, \widetilde{v}^{(2)}(2:n))$. Therefore

$$|u'_{ij} - u_{ij}| \leq (3\,(n+1)\,i + 12\,n\,(i-1))\,\mathbf{u},$$

which is the result in part (c) for the entries of U that are not trivial according to the rank of A. The entries of U that are identically equal to 0 or 1 (see Definition 1) satisfy part (c) trivially.

The proof of part (d) requires more work. Lemma 8 and (13) imply

$$|\widetilde{v}^{(1)} - v^{(1)}| \leq \gamma_{n+1}\, v^{(1)} \quad \text{and} \quad |\widetilde{A}_D^{(1)} - A_D^{(1)}| \leq \gamma_{n+1}\, |A_D^{(1)}|.$$

Observe that $\gamma_{n+1} < 1$ because $36n^3\mathbf{u} < 1$. From part (e) of Theorem 3, we get

$$|\widetilde{\ell}_{ij} - \ell_{ij}| \leq (1+\beta_j)\,\frac{j\,\gamma_{n+1}}{1 - j\,\gamma_{n+1}}\left(3 + \frac{2\,j\,\gamma_{n+1}}{1 - j\,\gamma_{n+1}}\right).$$

The condition $36n^3\mathbf{u} < 1$ together with some algebraic manipulations gives

$$|\widetilde{\ell}_{ij} - \ell_{ij}| \leq (1+\beta_j)\,\frac{2}{3}\,5\,j(n+1)\,\mathbf{u}. \tag{45}$$

Lemma 8 and (13) also imply

$$|\widehat{v}^{(2)}(2:n) - \widetilde{v}^{(2)}(2:n)| \leq \gamma_{4n}\,\widetilde{v}^{(2)}(2:n) \quad \text{and} \quad \left|\widehat{\mathscr{A}_D}^{(2)} - \widetilde{\mathscr{A}_D}^{(2)}\right| \leq \gamma_{4n}\left|\widetilde{\mathscr{A}_D}^{(2)}\right|.$$

Observe that $\gamma_{4n} < 1$ because $36n^3\mathbf{u} < 1$. From part (e) of Theorem 3, we get

$$|\ell'_{ij} - \widetilde{\ell}_{ij}| \leq (1+\beta_j)\,\frac{(j-1)\,\gamma_{4n}}{1 - (j-1)\,\gamma_{4n}}\left(3 + \frac{2\,(j-1)\,\gamma_{4n}}{1 - (j-1)\,\gamma_{4n}}\right),$$

where we have used that $\widetilde{\ell}_{ij}$ corresponds to the $(j-1)$th column of the L factor of $\mathscr{D}(\widetilde{\mathscr{A}_D}^{(2)}, \widetilde{v}^{(2)}(2:n))$. The condition $36n^3\mathbf{u} < 1$ together with some algebraic manipulations gives

$$|\ell'_{ij} - \widetilde{\ell}_{ij}| \leq (1+\beta_j)\,\frac{2}{3}\,22\,(j-1)n\,\mathbf{u}. \tag{46}$$

Combine (45) and (46) to get the result in part (d). □

4.3 Inductive proof of Theorem 4

We can now prove Theorem 4. The proof of part (a) follows directly from Lemma 10 below by using $d_k = a_{kk}^{(k)}$. The errors in the diagonal entries of the Schur complements in Lemma 10 will be used in the proof of part (c) of Theorem 4, where it is important to show that complete-diagonal pivoting in *floating point arithmetic* implies that the initial matrix $A = \mathscr{D}(A_D, v)$ is almost arranged for complete-diagonal pivoting in *exact arithmetic*. Recall that we assume that A has been permuted in advance.

Lemma 10 *Let $A = \mathscr{D}(A_D, v) \in \mathbb{R}^{n\times n}$ be a row diagonally dominant matrix with nonnegative diagonal entries and assume that $v \geq 0$ and A_D are known. Let $r =$ rank$(\mathscr{D}(A_D, v)) > 0$. Let us apply Algorithm 1 in [42] to (A_D, v) with any diagonal pivoting strategy in a computer with unit roundoff* $\mathbf{u}$*. Then the following statements hold.*

(a) *The off-diagonal entries and the diagonally dominant parts computed after r stages of Gaussian elimination satisfy $(\widehat{A}_D^{(r+1)}(r+1:n, r+1:n), \widehat{v}^{(r+1)}(r+1:n)) = (0,0)$, i.e., Algorithm 1 in [42] computes exactly the rank of A and stops after r stages.*

(b) *For $1 \leq k \leq \min\{r+1, n\}$, let $[a_{ii}^{(k)}]_{i=k}^{n}$ and $[\widehat{a}_{ii}^{(k)}]_{i=k}^{n}$ be, respectively, the exact diagonal entries of the Schur complements of A and the corresponding computed entries. Then,*

$$\widehat{a}_{ii}^{(k)} = a_{ii}^{(k)} \ <6nk^2> \qquad \text{for} \quad i = k, \ldots, n.$$

Proof Before we get into the details of the proof, note that $a_{11} = a_{11}^{(1)} > 0$ and therefore $\widehat{a}_{11} = \widehat{a}_{11}^{(1)} > 0$ by Lemma 8. This follows from the fact that $(A_D, v) \neq 0$, because the rank of A is not zero. So, taking into account our definition in Section 2, any diagonal pivoting strategy will select as first pivot a nonzero diagonal entry. Recall also that we are assuming that A has been permuted in advance according to the diagonal pivoting strategy that we are using.

Part (a). The result is obviously true for size $n = 1$. Assume that it holds for $(n-1)\times(n-1)$ matrices. With the notation of Lemma 8: stages $2, 3, \ldots$ of Algorithm 1 in [42] on $(A_D, v) = (A_D^{(1)}, v^{(1)})$ are precisely stages $1, 2, \ldots$ of Algorithm 1 in [42] on $(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n))$. By the induction hypothesis the rank of $\mathscr{A}_2' = \mathscr{D}(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)) \in \mathbb{R}^{(n-1)\times(n-1)}$ is exactly computed by Algorithm 1 in [42] and, by Lemma 9 (a), also the rank of A.

Part (b). We follow again an inductive argument. More precisely, assume that

$$\widehat{a}_{ii}^{(k)} = a_{ii}^{(k)} \ <\Phi(n,k)>, \tag{47}$$

where $\Phi(n,k)$ is a constant that depends on n and k but not on A. Obviously (47) holds for $n = k = 1$ with $\Phi(1,1) = 0$, since with $n = 1$ there is nothing to compute, and, more general, (47) also holds for $k = 1$ and any $n \geq 1$ with $\Phi(n,1) = n-1$, as a consequence of Lemma 8 (a). Our aim is to verify (47) inductively and at the same time to derive a recurrence for the unknown constant $\Phi(n,k)$.

Our induction hypothesis is that for any matrix $B = \mathscr{D}(B_D, v_B) \in \mathbb{R}^{(n-1)\times(n-1)}$, $v_B \geq 0$, Algorithm 1 in [42] with any diagonal pivoting strategy computes the diagonal entries of the Schur complements of B with errors,

$$\widehat{b}_{ii}^{(k)} = b_{ii}^{(k)} \ <\Phi(n-1,k)> \tag{48}$$

for $k = 1,\ldots,\min\{\mathrm{rank}(B)+1, n-1\}$ and $i = k,\ldots,n-1$. Therefore, this happens in particular for the matrix $\mathscr{A}_2' = \mathscr{D}\left(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)\right) \in \mathbb{R}^{(n-1)\times(n-1)}$ appearing in Lemmas 9 and 8. Recall that the computed $(k-1)$th Schur complement of $\mathscr{A}_2'$ is precisely $[\widehat{a}_{ij}^{(k)}]_{i,j=k}^{n}$ and that with the notation of Lemma 9 the corresponding exact Schur complement of $\mathscr{A}_2'$ is $[a_{ij}^{\prime(k)}]_{i,j=k}^{n}$. So (48) implies,

$$\widehat{a}_{ii}^{(k)} = a_{ii}^{\prime(k)} \ <\Phi(n-1,k-1)> \tag{49}$$

for $k = 2,\ldots,\min\{r+1,n\}$ and $i = k,\ldots,n$. Combine this equation with Lemma 9 (b) to get

$$\widehat{a}_{ii}^{(k)} = a_{ii}^{(k)} \ <\Phi(n-1,k-1) + 10nk - 13n + 2k - 1>,$$

and from (47)

$$\Phi(n,k) = \Phi(n-1,k-1) + 10nk - 13n + 2k - 1, \quad k = 2,3,\ldots \tag{50}$$

This recurrence relation, together with $\Phi(n,1) = n-1$ for all $n \geq 1$, determines completely $\Phi(n,k)$ for all $n \geq k \geq 1$. It is easy to check inductively that $\Phi(n,k) \leq 6nk^2$. □

Proof of part (a) *of Theorem 4* If $\mathrm{rank}(A) = r$, then $d_{r+1} = \cdots = d_n = 0$. Lemma 10 (a) implies that $\widehat{d}_{r+1} = \cdots = \widehat{d}_n = 0$. So $|\widehat{d}_i - d_i| = 0$, for $i = r+1,\ldots,n$. For $d_1,\ldots,d_r$, the result follows from Lemma 10 (b), $a_{ii}^{(i)} = d_i$, and (13). □

Proof of part (b) *of Theorem 4* We follow an inductive argument and assume that

$$|\widehat{u}_{ij} - u_{ij}| \leq \mathbf{u}\,\Psi(n,i) \qquad \text{for} \quad 1 \leq i < j \leq n, \tag{51}$$

where $\Psi(n,i)$ is a constant that depends on n and i but not on A. Observe that Lemma 8 (b), (13), the fact that U is row diagonally dominant, and $36n^3\,\mathbf{u} < 1$ imply

$$|\widehat{u}_{1j} - u_{1j}| \leq |u_{1j}|\,\frac{n\mathbf{u}}{1-n\mathbf{u}} < \mathbf{u}\,2n \qquad \text{for} \quad 1 < j \leq n.$$

Therefore, (51) holds for $i = 1$ and any $n \geq 1$ with $\Psi(n,1) = 2n$.

To derive a recurrence for the unknown $\Psi(n,i)$, we assume that for any matrix $B = \mathscr{D}(B_D, v_B) \in \mathbb{R}^{(n-1)\times(n-1)}$, $v_B \geq 0$, Algorithm 1 in [42] with any diagonal pivoting strategy computes the entries (i,j) of the U-factor of B with errors as in (51) but with $n-1$ instead of n, i.e., with errors $\mathbf{u}\,\Psi(n-1,i)$. In particular, this happens for the matrix $\mathscr{A}_2' = \mathscr{D}\left(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)\right) \in \mathbb{R}^{(n-1)\times(n-1)}$ appearing in Lemmas 9 and 8.

The computed U-factor of $\mathscr{A}_2'$ is precisely $[\widehat{u}_{ij}]_{i,j=2}^n$ and, with the notation of Lemma 9, the exact U-factor of $\mathscr{A}_2'$ is $[u'_{ij}]_{i,j=2}^n$. So,

$$|\widehat{u}_{ij} - u'_{ij}| \leq \mathbf{u}\,\Psi(n-1,i-1) \qquad \text{for} \quad 2 \leq i < j \leq n, \tag{52}$$

where we have used that the entries u'_{ij} are in the $(i-1)$th row of the U-factor of $\mathscr{A}_2'$. Combine (52) with Lemma 9 (c) and get

$$|\widehat{u}_{ij} - u_{ij}| \leq \mathbf{u}\,(\,\Psi(n-1,i-1) + 15ni - 12n + 3i\,) \qquad \text{for} \quad 2 \leq i < j \leq n.$$

This and (51) give

$$\Psi(n,i) = \Psi(n-1,i-1) + 15ni - 12n + 3i \qquad \text{for} \quad 2 \leq i \leq n.$$

This recurrence relation, together with $\Psi(n,1) = 2n$ for all $n \geq 1$, determines completely $\Psi(n,i)$ for all $n \geq i \geq 1$. It is easy to check inductively that $\Psi(n,i) \leq 8ni^2$. $\square$

Proof of part (c) *of Theorem 4* The proof of this part needs more work than the proofs of parts (a) and (b). The main difficulty comes from the fact that an effective application of Lemma 9 (d) requires to determine the numbers $\beta_2, \ldots, \beta_r$ in that lemma. For this purpose, recall that in this part we assume that complete-diagonal pivoting has been used in Algorithm 1 in [42]. This means that the computed diagonal entries of the Schur complements satisfy

$$|\widehat{a}_{kk}^{(k)}| \geq |\widehat{a}_{ii}^{(k)}|, \quad k \leq i \leq n, \quad 1 \leq k \leq \min\{r, n-1\}. \tag{53}$$

But from Lemma 10 (b), we can only guarantee for the exact entries that

$$<12nk^2>\, |a_{kk}^{(k)}| \geq |a_{ii}^{(k)}|, \quad k \leq i \leq n, \quad 1 \leq k \leq \min\{r, n-1\}.$$

This inequality together with (13) and (14) imply

$$\left(1 + \frac{12nk^2\mathbf{u}}{1 - 12nk^2\mathbf{u}}\right) |a_{kk}^{(k)}| \geq |a_{ii}^{(k)}|, \quad k \leq i \leq n, \quad 1 \leq k \leq \min\{r, n-1\}. \tag{54}$$

Now, use the notation in Lemma 9, equation (44) in the proof of Lemma 9 and equation (49) in the proof of Lemma 10 to get

$$\widehat{a}_{ii}^{(k)} = \widetilde{a}_{ii}^{(k)} \,<\Phi(n-1,k-1) + 8nk - 12n>, \quad k \leq i \leq n, \quad 2 \leq k \leq \min\{r, n-1\}.$$

From (50)

$$\Phi(n-1,k-1) + 8nk - 12n \leq \Phi(n,k) \leq 6nk^2.$$

So, $\widehat{a}_{ii}^{(k)} = \widetilde{a}_{ii}^{(k)} \,<6\,nk^2>$ and from (53)

$$\left(1 + \frac{12nk^2\mathbf{u}}{1 - 12nk^2\mathbf{u}}\right) |\widetilde{a}_{kk}^{(k)}| \geq |\widetilde{a}_{ii}^{(k)}|, \quad k \leq i \leq n, \quad 2 \leq k \leq \min\{r, n-1\}. \tag{55}$$

Equations (54) and (55) imply that we can take as numbers β_k, $k = 2, \ldots, \min\{r, n-1\}$, in Lemma 9 (d),

$$\beta_k = \frac{12nk^2\mathbf{u}}{1 - 12nk^2\mathbf{u}} \leq \frac{1/3}{1 - (1/3)} = \frac{1}{2}, \tag{56}$$

where we have used that $36n^3\mathbf{u} < 1$.

The rest of the proof follows the same pattern as the one for Theorem 4 (b). We simply sketch it. Assume that

$$|\widehat{\ell}_{ij} - \ell_{ij}| \leq \mathbf{u}\,\Omega(n,j) \qquad \text{for} \quad 1 \leq j < i \leq n. \tag{57}$$

Lemma 8 (b), the fact that complete-diagonal pivoting is used in floating point arithmetic (so $|\widehat{\ell}_{ij}| \leq 1$) and $36n^3\,\mathbf{u} < 1$ imply

$$|\widehat{\ell}_{i1} - \ell_{i1}| \leq |\widehat{\ell}_{i1}|\,\frac{n\mathbf{u}}{1-n\mathbf{u}} < \mathbf{u}\,2n \qquad \text{for} \quad 1 < i \leq n,$$

so (57) holds for $j = 1$ and any $n \geq 1$ with $\Omega(n,1) = 2n$.

Assume that for any matrix $B = \mathscr{D}(B_D, v_B) \in \mathbb{R}^{(n-1)\times(n-1)}$, $v_B \geq 0$, Algorithm 1 in [42] with complete-diagonal pivoting computes the entries (i,j) of the L-factor of B with errors $\mathbf{u}\,\Omega(n-1,j)$. So, this happens for $\mathscr{A}_2' = \mathscr{D}(\widehat{\mathscr{A}_D}^{(2)}, \widehat{v}^{(2)}(2:n)) \in \mathbb{R}^{(n-1)\times(n-1)}$ in Lemma 9. The computed L-factor of $\mathscr{A}_2'$ is $[\widehat{\ell}_{ij}]_{i,j=2}^{n}$ and, with the notation of Lemma 9, the exact L-factor of $\mathscr{A}_2'$ is $[\ell'_{ij}]_{i,j=2}^{n}$. Therefore, by the induction assumption,

$$|\widehat{\ell}_{ij} - \ell'_{ij}| \leq \mathbf{u}\,\Omega(n-1,j-1) \qquad \text{for} \quad 2 \leq j < i \leq n, \tag{58}$$

because the entries ℓ'_{ij} are in the $(j-1)$th column of the L-factor of $\mathscr{A}_2'$. Combine (58) with Lemma 9 (d) and (56) to get

$$|\widehat{\ell}_{ij} - \ell_{ij}| \leq \mathbf{u}\,(\,\Omega(n-1,j-1) + 27nj - 22n + 5j\,) \qquad \text{for} \quad 2 \leq j < i \leq n.$$

This and (57) give

$$\Omega(n,j) = \Omega(n-1,j-1) + 27nj - 22n + 5j \qquad \text{for} \quad 2 \leq j \leq n.$$

This recurrence relation is completed with $\Omega(n,1) = 2n$ for all $n \geq 1$. By induction, it can be checked that $\Omega(n,j) \leq 14nj^2$. □

4.4 Improving the bounds through compensated summation

Algorithm 1 in [42] requires to compute summations of nonnegative numbers in several instances. We have shown in Theorem 4 excellent error bounds for the computed *LDU* factors when these summations are computed through *standard recursive summation* [30, Sec. 4.1]. In fact, the componentwise errors in Theorem 4 are nearly optimal because the cost of Algorithm 1 in [42] is $O(n^3)$. These error bounds can be further improved if summations of nonnegative numbers are computed with the method of *compensated summation* [30, Sec. 4.3] and we discuss briefly in this section this improvement. Nevertheless, *we do not recommend at all to use compensated summation* since we have not observed any significative improvement in the accuracy in practice, and the total cost of the algorithm is increased by almost a factor 4.

If compensated summation is used to sum n nonnegative numbers, then the relative error is bounded by $2\mathbf{u} + O(n\mathbf{u}^2)$, i.e., the number of summands only affects

the error in second order terms. Then the relative errors and relative perturbations established in Lemma 8 are of the type $c\,\mathbf{u} + O(n\,\mathbf{u}^2)$, where c is a small integer constant. The reader may check that factors n also disappear from the first order terms in the rest of the error analysis presented in subsections 4.2 and 4.3. Finally the error bounds committed by Algorithm 1 in [42] when compensated summation is used are

(a) For any diagonal pivoting strategy

$$|\widehat{d_i} - d_i| \leq |d_i|\,(c_1\, i^2\,\mathbf{u} + O(n\, i^2\,\mathbf{u}^2)), \qquad i = 1, \ldots, n;$$

(b) For any diagonal pivoting strategy

$$|\widehat{u}_{ij} - u_{ij}| \leq (c_2\, i^2\,\mathbf{u} + O(n\, i^2\,\mathbf{u}^2)), \qquad 1 \leq i < j \leq n;$$

(c) For the complete-diagonal pivoting strategy

$$|\widehat{\ell}_{ij} - \ell_{ij}| \leq (c_3\, j^2\,\mathbf{u} + O(n\, j^2\,\mathbf{u}^2)), \qquad 1 \leq j < i \leq n,$$

where c_1, c_2 and c_3 are small integer constants. These bounds improve those of Theorem 4 up to first order in $\mathbf{u}$, but the higher order terms remain unknown. These bounds allow us to reduce by one the exponent of n in the normwise bounds (36) and (37).

5 Conclusions

We have proved that tiny relative componentwise perturbations of the diagonally dominant parts and the off-diagonal entries of a (row) diagonally dominant matrix with nonnegative diagonal entries produce tiny relative variations in its L, D, and U factors obtained through complete-diagonal pivoting. These tiny relative variations are componentwise for D and normwise for L and U. This perturbation result has been the key to prove that Algorithm 1 in [42] for the LDU factorization with complete diagonal pivoting computes accurately rank revealing decompositions of diagonally dominant matrices, and so that accurate computations of SVDs and solutions of linear systems are possible at cost $O(n^3)$ for any diagonally dominant matrix stored in a computer, independently of its traditional condition number.

The perturbation theory in this paper can be combined with the perturbation theory for the SVD in [12] to prove that the diagonally dominant parts and off diagonal entries determine to high relative accuracy the SVD of any diagonally dominant matrix. However, from [43], it seems possible to develop a sharper perturbation theory for the SVD, independent of the intermediate LDU factorization and of the condition numbers of L and U. This remains as an open problem.

Acknowledgements The authors thank Juan M. Molera for fruitful discussions and constructive comments. We are also grateful to two anonymous Referees for several useful suggestions that have helped us to improve the paper.

References

1. Abate, J., Choudhury, G.L., Whitt, W.: Asymptotics for steady-state tail probabilities in structured Markov queueing models. Comm. Statist. Stochastic Models **10**(1), 99–143 (1994)
2. Ahlberg, J.H., Nilson, E.N.: Convergence properties of the spline fit. J. Soc. Indust. Appl. Math. **11**, 95–104 (1963)
3. Alfa, A.S., Xue, J., Ye, Q.: Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comp. **71**(237), 217–236 (2002)
4. Alfa, A.S., Xue, J., Ye, Q.: Entrywise perturbation theory for diagonally dominant M-matrices with applications. Numer. Math. **90**(3), 401–414 (2002)
5. Avron, H., Chen, D., Shklarski, G., Toledo, S.: Combinatorial preconditioners for scalar elliptic finite-element problems. SIAM J. Matrix Anal. Appl. **31**(2), 694–720 (2009)
6. Barlow, J., Demmel, J.: Computing accurate eigensystems of scaled diagonally dominant matrices. SIAM J. Num. Anal. **27**(3), 762–791 (1990)
7. Berman, A., Plemmons, R.J.: Nonnegative matrices in the mathematical sciences, *Classics in Applied Mathematics*, vol. 9. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1994). Revised reprint of the 1979 original
8. Boman, E.G., Hendrickson, B., Vavasis, S.: Solving elliptic finite element systems in near-linear time with support preconditioners. SIAM J. Numer. Anal. **46**(6), 3264–3284 (2008)
9. Chan, T.F., Foulser, D.E.: Effectively well-conditioned linear systems. SIAM J. Sci. Statist. Comput. **9**(6), 963–969 (1988)
10. Demmel, J.: Accurate singular value decompositions of structured matrices. SIAM J. Matrix Anal. Appl. **21**(2), 562–580 (1999)
11. Demmel, J., Gragg, W.: On computing accurate singular values and eigenvalues of acyclic matrices. Linear Algebra Appl. **185**, 203–218 (1993)
12. Demmel, J., Gu, M., Eisenstat, S., Slapničar, I., Veselić, K., Drmač, Z.: Computing the singular value decomposition with high relative accuracy. Linear Algebra Appl. **299**(1–3), 21–80 (1999)
13. Demmel, J., Kahan, W.: Accurate singular values of bidiagonal matrices. SIAM J. Sci. Statist. Comput. **11**(5), 873–912 (1990)
14. Demmel, J., Koev, P.: Accurate SVDs of weakly diagonally dominant M-matrices. Numer. Math. **98**(1), 99–104 (2004)
15. Demmel, J., Koev, P.: Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials. Linear Algebra Appl. **417**(2-3), 382–396 (2006)
16. Demmel, J., Veselić, K.: Jacobi's method is more accurate than QR. SIAM J. Matrix Anal. Appl. **13**(4), 1204–1246 (1992)
17. Demmel, J.W.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
18. Dopico, F.M., Koev, P.: Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices. SIAM J. Matrix Anal. Appl. **28**(4), 1126–1156 (2006)
19. Dopico, F.M., Koev, P., Molera, J.M.: Implicit standard Jacobi gives high relative accuracy. Numer. Math. **113**(2), 519–553 (2009)
20. Dopico, F.M., Molera, J.M.: Accurate solution of structured linear systems via rank-revealing decompositions. Submitted. (2010)
21. Dopico, F.M., Molera, J.M., Moro, J.: An orthogonal high relative accuracy algorithm for the symmetric eigenproblem. SIAM J. Matrix Anal. Appl. **25**(2), 301–351 (2003)
22. Drmač, Z.: Accurate computation of the product induced singular value decomposition with applications. SIAM J. Num. Anal. **35**(5), 1969–1994 (1998)
23. Drmač, Z., Veselić, K.: New fast and accurate Jacobi SVD algorithm. I. SIAM Journal on Matrix Analysis and Applications **29**(4), 1322–1342 (2008)
24. Drmač, Z., Veselić, K.: New fast and accurate Jacobi SVD algorithm. II. SIAM Journal on Matrix Analysis and Applications **29**(4), 1343–1362 (2008)
25. Eisenstat, S., Ipsen, I.: Relative perturbation techniques for singular value problems. SIAM J. Numer. Anal. **32**(6), 1972–1988 (1995)
26. Falkenberg, E.: On the asymptotic behaviour of the stationary distribution of Markov chains of $M/G/1$-type. Comm. Statist. Stochastic Models **10**(1), 75–97 (1994)
27. Fernando, K., Parlett, B.: Accurate singular values and differential qd algorithms. Numer. Math. **67**, 191–229 (1994)
28. Gantmacher, F.: The Theory of Matrices. AMS Chelsea, Providence, RI (1998)

29. Golub, G., Van Loan, C.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore, MD (1996)
30. Higham, N.J.: Accuracy and stability of numerical algorithms, second edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2002)
31. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
32. Horn, R.A., Johnson, C.R.: Topics in matrix analysis. Cambridge University Press, Cambridge (1994). Corrected reprint of the 1991 original
33. Li, R.C.: Relative perturbation theory. II. Eigenspace and singular subspace variations. SIAM J. Matrix Anal. Appl. **20**(2), 471–492 (1999)
34. The MathWorks, Inc., Natick, MA: MATLAB Reference Guide (1992)
35. O'Cinneide, C.A.: Relative-error bounds for the *LU* decomposition via the GTH algorithm. Numer. Math. **73**(4), 507–519 (1996)
36. Peña, J.M.: Pivoting strategies leading to diagonal dominance by rows. Numer. Math. **81**(2), 293–304 (1998)
37. Peña, J.M.: LDU decompositions with L and U well conditioned. Electron. Trans. Numer. Anal. **18**, 198–208 (electronic) (2004)
38. Priest, D.: On properties of floating point arithmetics: Numerical stability and the cost of accurate computations. Ph.D. thesis, Mathematics Department, University of California, Berkeley, CA, USA (1992)
39. Stewart, G.W.: Introduction to Matrix Computations. Academic Press, New York (1973)
40. Varah, J.M.: A lower bound for the smallest singular value of a matrix. Linear Algebra and Appl. **11**, 3–5 (1975)
41. Varga, R.S.: On diagonal dominance arguments for bounding $\| A^{-1} \|_\infty$. Linear Algebra and Appl. **14**(3), 211–217 (1976)
42. Ye, Q.: Computing singular values of diagonally dominant matrices to high relative accuracy. Math. Comp. **77**(264), 2195–2230 (2008)
43. Ye, Q.: Relative perturbation bounds for eigenvalues of symmetric positive definite diagonally dominant matrices. SIAM J. Matrix Anal. Appl. **31**(1), 11–17 (2009)