SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 5, pp. 2841–2859

RIGOROUS PERTURBATION BOUNDS OF SOME MATRIX FACTORIZATIONS*

X.-W. CHANG† AND D. STEHLÉ‡

Abstract. This article presents rigorous normwise perturbation bounds for the Cholesky, LU, and QR factorizations with normwise or componentwise perturbations in the given matrix. The considered componentwise perturbations have the form of backward rounding errors for the standard factorization algorithms. The used approach is a combination of the classic and refined matrix equation approaches. Each of the new rigorous perturbation bounds is a small constant multiple of the corresponding first-order perturbation bound obtained by the refined matrix equation approach in the literature and can be estimated efficiently. These new bounds can be much tighter than the existing rigorous bounds obtained by the classic matrix equation approach, while the conditions for the former to hold are almost as moderate as the conditions for the latter to hold.

Key words. perturbation analysis, normwise perturbation, componentwise perturbation, Cholesky factorization, LU factorization, QR factorization, column and row scaling, first-order bounds, rigorous bounds

AMS subject classifications. 15A23, 65F35

DOI. 10.1137/090778535

1. Introduction.

Let A be a given matrix and have a factorization

$$A = BC. \tag{1.1}$$

Suppose that A is perturbed to $A + \Delta A$, where a normwise or componentwise bound on ΔA is known. Let the same factorization for $A + \Delta A$ be

$$A + \Delta A = (B + \Delta B)(C + \Delta C). \tag{1.2}$$

The aim of a perturbation analysis is to assess the effects of ΔA on ΔB and ΔC. In the analysis, normwise or componentwise bounds on ΔB and ΔC are derived.

The perturbation theory of matrix factorizations has been extensively studied. The following table summarizes the relevant works on perturbation bounds of Cholesky, LU, and QR factorizations which are known to the authors.

P	B	Cholesky	LU	QR
N	FN	[2], [8], [18], [19], [20]	[2], [6], [12], [18], [19]	[2], [9], [18], [20], [23]
N	RN	[8], [12], [13], [17], [20]	[1], [12]	[17], [20], [23]
C	FN	[4]	[5]	[7], [25]
C	RN	[3], [13]		[7], [10]
C	FC	[3]	[5]	[7]
C	RC	[12], [21], [22]	[12], [22]	[22]

In the first column, "P" stands for the type of *perturbation* in the matrix to be factorized, and "N" and "C" stand for *normwise* perturbation and *componentwise*

*Received by the editors November 30, 2009; accepted for publication (in revised form) by R.-C. Li September 9, 2010; published electronically November 4, 2010.
http://www.siam.org/journals/simax/31-5/77853.html

†School of Computer Science, McGill University, Montreal, QC, H3A 2A7, Canada (chang@cs.mcgill.ca). The work of this author was supported by NSERC of Canada grant RGPIN217191-07.

‡CNRS, Laboratoire LIP (U. Lyon, CNRS, ENS de Lyon, INRIA, UCBL), École Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France (damien.stehle@ens-lyon.fr).

perturbation, respectively; in the second column, "B" stands for perturbation bound of the factor, "FN", "RN","FC" and "RC" stand for *first-order normwise* perturbation bound, *rigorous normwise* perturbation bound, *first-order componentwise* perturbation bound, and *rigorous componentwise* perturbation bound, respectively. In the present article, we call a bound rigorous if it does not neglect any higher-order terms as the first-order bound does: Under appropriate assumptions, it always holds true.

Two types of approaches are often used to derive normwise perturbation bounds. One is the matrix-vector equation approach, and the other is the matrix equation approach; see [3]. Here we give a brief explanation about these two approaches in the context of first-order analysis. From (1.1) and (1.2) we have by dropping the second-order term that

$$\Delta A \approx B\Delta C + \Delta BC. \tag{1.3}$$

The basic idea of the matrix-vector equation approach is to write this approximate matrix equation (1.3) as a matrix-vector equation by using the special structures and properties of the involved matrices, then obtain the vector-type expressions for ΔB and ΔC, from which normwise bounds on ΔB and ΔC can be derived. The approach can be extended to obtain rigorous bounds. This approach usually leads to sharp bounds, but the bounds (first-order bounds or rigorous bounds) are expensive to estimate, and the conditions for the rigorous bounds to hold are often too restrictive and complicated. The matrix equation approach comes in two flavors. The classic matrix equation approach keeps (1.3) in the matrix-matrix form and drives bounds on ΔB and ΔC. The approach can be extended to obtain rigorous bounds. The bounds (first-order bounds or rigorous bounds) can be efficiently estimated, and the conditions for the rigorous bounds to hold are less restrictive and simpler. But the bounds are usually not tight. The refined matrix equation approach additionally uses row or column scaling techniques. It has been mainly used to derive first-order bounds, which numerical experiments showed are often good approximations to the sharp first-order bounds derived by the matrix-vector equation approach.

It is often unclear whether a first-order bound is a good approximate bound, as the ignored higher-order terms may dominate the true perturbation (see, e.g., Remark 5.1). Furthermore, in some applications rigorous bounds are needed in order to certify the accuracy of computations; see, e.g., [10, 15] for an application with the QR factorization and [16] for an application with the Cholesky factorization.

The present article aims at providing tight rigorous perturbation bounds for the Cholesky, LU, and QR factorizations, which can be efficiently estimated in $O(n^2)$ flops, where n is the number of columns of the matrix to be factorized. Additionally, the conditions for the bounds to hold are simple and moderate. We consider both normwise and componentwise perturbations in the matrix to be factorized. The componentwise perturbations have the form of backward errors resulting from standard factorization algorithms. In [10] we obtained such a rigorous bound for the R-factor of the QR factorization under a componentwise perturbation which has the form of backward rounding errors of standard QR factorization algorithms. The approach used in the latter work is actually a combination of the classic and refined matrix equation approaches. We will use a similar approach in this article.

The rest of this article is organized as follows. In section 2, we introduce notation and give some basics that will be necessary for the following three sections. Sections 3, 4, and 5 are devoted to Cholesky, LU, and QR factorizations, respectively. Finally a summary is given in section 6.

2. Notation and basics. For a matrix $X \in \mathbb{R}^{n\times n}$, we use $X(i,:)$ and $X(:,j)$ to denote its ith row and jth column, respectively, and use X_k to denote its $k\times k$ leading principal submatrix. We define a lower triangular matrix and two upper triangular matrices associated with $X \in \mathbb{R}^{n\times n}$ as follows:

$$\mathrm{slt}(X) = (s_{ij}), \qquad s_{ij} = \begin{cases} x_{ij} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

$$\mathrm{ut}(X) = X - \mathrm{slt}(X), \tag{2.2}$$

$$\mathrm{up}(X) = (s_{ij}), \qquad s_{ij} = \begin{cases} x_{ij} & \text{if } i < j, \\ \frac{1}{2}x_{ij} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

For any absolute matrix norm $\|\cdot\|$ (i.e., $\|A\| = \|\,|A|\,\|$ for any A), we have

$$\|\mathrm{slt}(X)\| \le \|X\|, \quad \|\mathrm{ut}(X)\| \le \|X\|, \quad \|\mathrm{up}(X)\| \le \|X\|. \tag{2.4}$$

Let $\mathcal{D}_n$ denote the set of all real $n\times n$ positive definite diagonal matrices. We will use the following properties, which hold for any $D \in \mathcal{D}_n$:

$$\mathrm{slt}(DX) = D\,\mathrm{slt}(X), \quad \mathrm{ut}(XD) = \mathrm{ut}(X)D, \quad \mathrm{up}(XD) = \mathrm{up}(X)D. \tag{2.5}$$

It can be verified that if $X^T = X$, then

$$\|\mathrm{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F. \tag{2.6}$$

It is proved in [9, Lemma 5.1] that, for any $D = \mathrm{diag}(\delta_1,\dots,\delta_n) \in \mathcal{D}_n$,

$$\|\mathrm{up}(X) + D^{-1}\mathrm{up}(X^T)D\|_F \le \rho_D\|X\|_F, \quad \rho_D = \left[1 + \max_{1\le i<j\le n}(\delta_j/\delta_i)^2\right]^{1/2}. \tag{2.7}$$

For any matrix $X \in \mathbb{R}^{m\times n}$ and any consistent matrix norm $\|\cdot\|_\nu$, we define

$$\kappa_\nu(X) = \|X^\dagger\|_\nu\|X\|_\nu, \qquad \mathrm{cond}_\nu(X) = \|\,|X^\dagger|\cdot|X|\,\|_\nu,$$

where $X^\dagger$ is the Moore-Penrose pseudo-inverse of X.

The following well-known results are due to van der Sluis [24].

LEMMA 2.1. *Let $S, T \in \mathbb{R}^{n\times n}$ with S nonsingular and define*

$$D_{rp} = \mathrm{diag}(\|S(i,:)\|_p), \quad D_{cp} = \mathrm{diag}(\|S(:,j)\|_p), \qquad p = 1, 2.$$

Then

$$\big\|\,|T||S|\,\big\|_\infty = \|TD_{r1}\|_\infty\|D_{r1}^{-1}S\|_\infty = \min_{D\in\mathcal{D}_n}\|TD\|_\infty\|D^{-1}S\|_\infty, \tag{2.8}$$

$$\|\,|S||T|\,\|_1 = \|SD_{c1}^{-1}\|_1\|D_{c1}T\|_1 = \min_{D\in\mathcal{D}_n}\|SD^{-1}\|_1\|DT\|_1, \tag{2.9}$$

$$\|TD_{r2}\|_2\|D_{r2}^{-1}S\|_2 \le \sqrt{n}\inf_{D\in\mathcal{D}_n}\|TD\|_2\|D^{-1}S\|_2, \tag{2.10}$$

$$\|SD_{c2}^{-1}\|_2\|D_{c2}T\|_2 \le \sqrt{n}\inf_{D\in\mathcal{D}_n}\|SD^{-1}\|_2\|DT\|_2. \tag{2.11}$$

This lemma indicates that if one wants to estimate the rightmost sides of (2.8)–(2.11), one can select appropriate scaling matrices and estimate the norms of scaled matrices T and S. If both S and T are available, or if only S is available but S is triangular and $T = \bar{D}S^{-1}$ in (2.8) and (2.10), or $T = S^{-1}\bar{D}$ in (2.9), and (2.11) for a known $\bar{D} \in \mathcal{D}_n$, then the above estimations can be done by norm estimators in $O(n^2)$ flops; see, e.g., [14, Chap. 15]. These results can be used to estimate the perturbation bounds to be presented.

Finally, we give the following basic result, which will be used in later sections many times.

LEMMA 2.2. *Let $a, b > 0$. Let $c(\cdot)$ be a continuous function of a parameter $t \in [0,1]$ such that $b^2 - 4ac(t) > 0$ holds for all t. Suppose that a continuous function $x(t)$ satisfies the quadratic inequality $ax(t)^2 - bx(t) + c(t) \geq 0$. If $c(0) = x(0) = 0$, then $x(1) \leq \frac{1}{2a}\big(b - \sqrt{b^2 - 4ac(1)}\big)$.*

Proof. The two roots of $ax(t)^2 - bx(t) + c(t) = 0$ are

$$x_1(t) = \frac{1}{2a}\big(b - \sqrt{b^2 - 4ac(t)}\big), \quad x_2(t) = \frac{1}{2a}\big(b + \sqrt{b^2 - 4ac(t)}\big).$$

Notice that $x_1(t) < x_2(t)$ and both are continuous. Since $ax(t)^2 - bx(t) + c(t) \geq 0$, we have either $x(t) \leq x_1(t)$ or $x(t) \geq x_2(t)$. But $x(t)$ is continuous and $x(0) = c(0) = 0$ so that $x(0) = x_1(0) < x_2(0)$, and therefore we must have $x(t) \leq x_1(t)$ for all t. ∎

3. Cholesky factorization. We first present rigorous perturbation bounds for the Cholesky factor when the given symmetric positive definite matrix has a general normwise perturbation.

THEOREM 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $A = R^T R$, where $R \in \mathbb{R}^{n\times n}$ is upper triangular with positive diagonal entries and let $\Delta A \in \mathbb{R}^{n\times n}$ be symmetric. If*

$$\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2} < 1/2, \tag{3.1}$$

then $A + \Delta A$ has the unique Cholesky factorization

$$A + \Delta A = (R + \Delta R)^T (R + \Delta R), \tag{3.2}$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}\kappa_2(R)\left[\inf_{D\in\mathcal{D}_n}\kappa_2(D^{-1}R)\right]\frac{\|\Delta A\|_F}{\|A\|_2}}{\sqrt{2} - 1 + \sqrt{1 - 2\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}} \tag{3.3}$$

$$\leq (2 + \sqrt{2})\kappa_2(R)\left[\inf_{D\in\mathcal{D}_n}\kappa_2(D^{-1}R)\right]\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{3.4}$$

Proof. From the condition (3.1),

$$\|A^{-1}\Delta A\|_2 \leq \kappa_2(A)\|\Delta A\|_2/\|A\|_2 < 1.$$

Thus, the matrix $A + t\Delta A$ for $t \in [0,1]$ is symmetric positive definite and has the unique Cholesky factorization

$$A + t\Delta A = (R + \Delta R(t))^T (R + \Delta R(t)), \tag{3.5}$$

which, with $\Delta R(1) = \Delta R$, leads to (3.2). Notice that $\Delta R(t)$ is a continuous function of t.

From (3.5) we obtain

$$R^{-T}\Delta R(t)^T + \Delta R(t)R^{-1} = tR^{-T}\Delta AR^{-1} - R^{-T}\Delta R(t)^T\Delta R(t)R^{-1}. \tag{3.6}$$

As $\Delta R(t)R^{-1}$ is upper triangular, it follows from (2.3) that

$$\Delta R(t)R^{-1} = \mathrm{up}(tR^{-T}\Delta AR^{-1} - R^{-T}\Delta R(t)^T\Delta R(t)R^{-1}). \tag{3.7}$$

Taking the Frobenius norm on both sides of (3.7) and using the inequality (2.6) and the fact that $\|A^{-1}\|_2 = \|R^{-1}\|_2^2$, we obtain

$$\|\Delta R(t)R^{-1}\|_F \le \frac{1}{\sqrt{2}}\|tR^{-T}\Delta AR^{-1} - R^{-T}\Delta R(t)^T\Delta R(t)R^{-1}\|_F \tag{3.8}$$

$$\le \frac{1}{\sqrt{2}}\left(t\|A^{-1}\|_2\|\Delta A\|_F + \|\Delta R(t)R^{-1}\|_F^2\right). \tag{3.9}$$

Therefore, as the assumption (3.1) guarantees that the condition of Lemma 2.2 holds, we have by Lemma 2.2 that

$$\|\Delta RR^{-1}\|_F \le \frac{1}{\sqrt{2}}\left(1 - \sqrt{1 - 2\|A^{-1}\|_2\|\Delta A\|_F}\right). \tag{3.10}$$

Taking $t = 1$ in (3.7), multiplying both sides by a diagonal $D \in \mathcal{D}_n$ from the right, and using the fact that $\mathrm{up}(X)D = \mathrm{up}(XD)$ (see (2.5)), we have

$$\Delta RR^{-1}D = \mathrm{up}(R^{-T}\Delta AR^{-1}D - R^{-T}\Delta R^T\Delta RR^{-1}D). \tag{3.11}$$

Taking the Frobenius norm on both sides of (3.11) and using $\|\mathrm{up}(X)\|_F \le \|X\|_F$ (see (2.4)), we obtain

$$\|\Delta RR^{-1}D\|_F \le \|R^{-1}\|_2\|R^{-1}D\|_2\|\Delta A\|_F + \|\Delta RR^{-1}\|_F\|\Delta RR^{-1}D\|_F.$$

Then, it follows by using (3.10) that

$$\|\Delta RR^{-1}D\|_F \le \frac{\sqrt{2}\|R^{-1}\|_2\|R^{-1}D\|_2\|\Delta A\|_F}{\sqrt{2} - 1 + \sqrt{1 - 2\|A^{-1}\|_2\|\Delta A\|_F}}.$$

Therefore,

$$\|\Delta R\|_F \le \|\Delta RR^{-1}D\|_F\|D^{-1}R\|_2 \le \frac{\sqrt{2}\|R^{-1}\|_2\|R^{-1}D\|_2\|D^{-1}R\|_2\|\Delta A\|_F}{\sqrt{2} - 1 + \sqrt{1 - 2\|A^{-1}\|_2\|\Delta A\|_F}}.$$

Since $D \in \mathcal{D}_n$ is arbitrary and $\|A\|_2 = \|R\|_2^2$, we have (3.3) and then (3.4). ∎

Now we make some remarks to show the relations between the new results and existing results in the literature.

Remark 3.1. In [8], the following first-order perturbation bound, which can be estimated in $O(n^2)$ flops, was derived:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_2(R)\left[\inf_{D\in\mathcal{D}_n}\kappa_2(D^{-1}R)\right]\frac{\|\Delta A\|_F}{\|A\|_2} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_2^2}\right). \tag{3.12}$$

Note that the difference between this first-order bound and the rigorous bound (3.4) is a factor of $2+\sqrt{2}$. Numerical experiments indicated that (3.12) is a good approximation to the optimal first-order bound derived by the matrix-vector equation approach in [8]:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_C(A)\frac{\|\Delta A\|_F}{\|A\|_2} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_2^2}\right),$$

where

$$\frac{1}{2}\kappa_2(R) \leq \kappa_C(A) \leq \kappa_2(R)\left[\inf_{D\in\mathcal{D}_n} \kappa_2(D^{-1}R)\right]. \tag{3.13}$$

The expression of $\kappa_C(A)$ involves an $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ lower triangular matrix defined by the entries of R. The best known method to estimate it requires $O(n^3)$ flops; see [8, Remark 6].

If the standard symmetric pivoting strategy is used in computing the Cholesky factorization, the quantity $\inf_{D\in\mathcal{D}_n} \kappa_2(D^{-1}R)$ is bounded by a function of n; see [8, sections 4 and 5].

Remark 3.2. One of the rigorous bounds derived by the classic matrix equation approach presented in [20] is as follows:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}{1+\sqrt{1-2\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}} \tag{3.14}$$

under the same condition as (3.1). If we take $D = I$ in (3.3), we obtain

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}{\sqrt{2}-1+\sqrt{1-2\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}}. \tag{3.15}$$

Comparing (3.15) with (3.14), we observe that the new rigorous bound (3.3) is at most $\sqrt{2}+1$ times as large as (3.14). But $\kappa_2(R)\inf_{D\in\mathcal{D}_n} \kappa_2(D^{-1}R)$ can be much smaller than $\kappa_2(A)$ when R has bad row scaling. For example, for $R = \mathrm{diag}(1,\gamma)$ with large $\gamma > 0$, $\kappa_2(R)\kappa_2(D^{-1}R) = \Theta(\gamma)$ with $D = \mathrm{diag}(1,\gamma)$, and $\kappa_2(A) = \Theta(\gamma^2)$. Thus the bound (3.3) can be much tighter than (3.14).

Remark 3.3. In [8, Theorem 9], the following rigorous perturbation bound was derived by the matrix-vector equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq 2\kappa_C(A)\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{3.16}$$

By (3.13), the new bound (3.4) is not as tight as this bound, but no numerical experiment has indicated that the former can be significantly larger than the latter; see Remark 3.1. As we mentioned in Remark 3.1, it is more expensive to estimate the latter than the former. A more serious problem with (3.16) is that the condition for it to hold given in [8, Theorem 9] can be as bad as $\kappa_C^2(A)\|\Delta A\|_F/\|A\|_2 < 1/4$. This is much more constraining than the condition (3.1) if $\inf_{D\in\mathcal{D}_n} \kappa_2(D^{-1}R)$ is not bounded by a constant; see (3.13). For example, for $R = \left[\begin{smallmatrix} 1 & \gamma \\ 0 & 1 \end{smallmatrix}\right]$ with large $\gamma > 0$, $\kappa_C^2(A) = \Theta(\gamma^4)$ and $\kappa_2(A) = \Theta(\gamma^2)$.

Remark 3.4. In [3, Theorem 2.2.8], the following rigorous bound was derived by the refined matrix equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le 2\kappa_2(R)\kappa_2(D^{-1}R)\frac{\|\Delta A\|_F}{\|A\|_2} \tag{3.17}$$

under the condition

$$\kappa_2(R)\|R\|_2\|R^{-1}D\|_2\|D^{-1}\|_2\frac{\|\Delta A\|_F}{\|A\|_2} < 1/4 \tag{3.18}$$

for any $D \in \mathcal{D}_n$. Notice that $\kappa_2(R)\|R\|_2\|R^{-1}D\|_2\|D^{-1}\|_2 \ge \kappa_2^2(R) = \kappa_2(A)$. Thus the condition (3.18) is not only more complicated but also more constraining than the condition (3.1). If we want to make the bound (3.17) similar to the new bound (3.4), then we may minimize $\kappa_2(D^{-1}R)$ over the set $\mathcal{D}_n$. But the optimal choice of D may make the condition (3.18) much more constraining than the condition (3.1). Here is an example. Let

$$R = \begin{bmatrix} 1 & \gamma & \gamma^2 \\ 0 & \gamma & \gamma^2 \\ 0 & 0 & \gamma \end{bmatrix}$$

with large $\gamma > 0$. By (2.10), $D_{r2} = \operatorname{diag}(\sqrt{1+\gamma^2+\gamma^4}, \sqrt{\gamma^2+\gamma^4}, \gamma)$ is an approximate optimal D. It is easy to verify that $\kappa_2(R)\|R\|_2\|R^{-1}D_{r2}\|_2\|D_{r2}^{-1}\|_2 = \Theta(\gamma^5)$ and $\kappa_2(A) = \Theta(\gamma^4)$. Thus the former can be arbitrarily larger than the latter.

In the following we present rigorous perturbation bounds for the Cholesky factor when the perturbation ΔA has the form we could expect from the backward error in A resulting from a standard Cholesky factorization algorithm (see [11] and [14, section 10.1]).

THEOREM 3.2. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $A = R^TR$ and let $A = D_cHD_c$ with $D_c = \operatorname{diag}(a_{11}^{1/2}, \ldots, a_{nn}^{1/2})$. Let $\Delta A \in \mathbb{R}^{n\times n}$ be symmetric such that $|\Delta A| \le \varepsilon dd^T$ for some constant ε and $d = [a_{11}^{1/2}, \ldots, a_{nn}^{1/2}]^T$. If*

$$n\|H^{-1}\|_2\,\varepsilon < 1/2, \tag{3.19}$$

then $A + \Delta A$ has the unique Cholesky factorization

$$A + \Delta A = (R + \Delta R)^T(R + \Delta R), \tag{3.20}$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\frac{\sqrt{2}n\|D_cR^{-1}\|_2\left(\inf_{D\in\mathcal{D}_n}\|D_cR^{-1}D\|_2\|D^{-1}R\|_2\right)}{\|R\|_2}\,\varepsilon}{\sqrt{2}-1+\sqrt{1-2n\|H^{-1}\|_2\,\varepsilon}} \tag{3.21}$$

$$\le \frac{(2+\sqrt{2})n\|D_cR^{-1}\|_2\left(\inf_{D\in\mathcal{D}_n}\|D_cR^{-1}D\|_2\|D^{-1}R\|_2\right)}{\|R\|_2}\,\varepsilon. \tag{3.22}$$

Proof. In the proof, we will use the following fact:

$$\|D_c^{-1}\Delta AD_c^{-1}\|_F \le \varepsilon\|D_c^{-1}dd^TD_c^{-1}\|_F = \varepsilon\|ee^T\|_F = n\varepsilon,$$

where $e = [1, \ldots, 1]^T$. Note that the spectral radius of $A^{-1}\Delta A$ satisfies

$$\begin{aligned}\rho(A^{-1}\Delta A) &= \rho(D_c^{-1}H^{-1}D_c^{-1}\Delta A) = \rho(H^{-1}D_c^{-1}\Delta AD_c^{-1}) \\ &\le \|H^{-1}\|_2\|D_c^{-1}\Delta AD_c^{-1}\|_2 \le n\|H^{-1}\|_2\,\varepsilon < 1.\end{aligned}$$

Thus, the matrix $A + t\Delta A$ for $t \in [0,1]$ is symmetric positive definite and has the unique Cholesky factorization (3.5), which, with $\Delta R(1) = \Delta R$, leads to (3.20).

From (3.7) we obtain

$$\Delta R(t)R^{-1} = \mathrm{up}\big(tR^{-T}D_cD_c^{-1}\Delta AD_c^{-1}D_cR^{-1} - R^{-T}\Delta R(t)^T\Delta R(t)R^{-1}\big). \tag{3.23}$$

Then, using (2.6) and the fact that $\|H^{-1}\|_2 = \|D_cR^{-1}\|_2^2$, we obtain

$$\|\Delta R(t)R^{-1}\|_F \le \frac{1}{\sqrt{2}}\left(tn\|H^{-1}\|_2\,\varepsilon + \|\Delta R(t)R^{-1}\|_F^2\right).$$

Therefore, as the assumption (3.19) guarantees that the condition of Lemma 2.2 holds, we have by Lemma 2.2 that

$$\|\Delta RR^{-1}\|_F \le \frac{1}{\sqrt{2}}\left(1 - \sqrt{1 - 2n\|H^{-1}\|_2\,\varepsilon}\right). \tag{3.24}$$

Taking $t = 1$ in (3.23), multiplying both sides by a diagonal $D \in \mathcal{D}_n$ from the right and then taking the Frobenius norm, we obtain

$$\|\Delta RR^{-1}D\|_F \le n\|D_cR^{-1}\|_2\|D_cR^{-1}D\|_2\,\varepsilon + \|\Delta RR^{-1}\|_F\|\Delta RR^{-1}D\|_F.$$

Then, using (3.24), we obtain

$$\|\Delta RR^{-1}D\|_F \le \frac{\sqrt{2}n\|D_cR^{-1}\|_2\|D_cR^{-1}D\|_2\,\varepsilon}{\sqrt{2} - 1 + \sqrt{1 - 2n\|H^{-1}\|_2\,\varepsilon}}.$$

This, combined with the inequality $\|\Delta R\|_F \le \|\Delta RR^{-1}D\|_F\|D^{-1}R\|_2$, leads to (3.21) and then (3.22). ∎

In the following we make some remarks, which are analogous to Remarks 3.1–3.4.

Remark 3.5. In [4] the following first-order perturbation bound, which can be estimated in $O(n^2)$ flops, was presented:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{n\|D_cR^{-1}\|_2\left(\inf_{D\in\mathcal{D}_n}\|D_cR^{-1}D\|_2\|D^{-1}R\|_2\right)}{\|R\|_2}\,\varepsilon + O(\varepsilon^2).$$

Note that the difference between the above first-order bound and the rigorous bound (3.22) is a factor of $2 + \sqrt{2}$ (cf. Remark 3.2). Numerical experiments indicated that the above first-order bound is often a reasonable approximation to the nearly optimal first-order bound derived by the matrix-vector equation approach in [4]:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \chi_C(A)\varepsilon + O(\varepsilon^2),$$

where (the first inequality below was proved in [3, Remark 2.3.5])

$$\frac{na_{nn}^{1/2}}{2\|A\|_2^{1/2}}\|H^{-1}\|_2^{1/2} \le \chi_C(A) \le \frac{n\|D_cR^{-1}\|_2\left(\inf_{D\in\mathcal{D}_n}\|D_cR^{-1}D\|_2\|D^{-1}R\|_2\right)}{\|R\|_2}. \tag{3.25}$$

The expression of $\chi_C(A)$ involves an $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ lower triangular matrix defined by the entries of RD_c^{-1} and the best known estimator of $\chi_C(A)$ requires $O(n^3)$ flops. Here we would like to point out that an example given in [3, Remark 2.3.9] shows that

in the second inequality in (3.25) the right-hand side can be arbitrarily larger than the left-hand side, although numerical tests have shown that usually the former is a reasonable approximation to the latter.

If the standard symmetric pivoting strategy is used in computing the Cholesky factorization, the quantity $\inf_{D\in\mathcal{D}_n} \|D_c R^{-1} D\|_2 \|D^{-1} R\|_2 / \|R\|_2$ is bounded by a function of n; see [3, Theorem 2.3.8 and section 2.3.4].

Remark 3.6. In [13] rigorous bounds on $\|\Delta R R^{-1}\|_{F,2}$ were derived. The bound on $\|\Delta R R^{-1}\|_F$, which was credited to Ji-guang Sun, is identical to (3.24) under the identical condition (3.19). As mentioned in [13], the bound on $\|\Delta R\|_F$ can be obtained by using $\|\Delta R\|_F \le \|\Delta R R^{-1}\|_F \|R\|_2$, leading to

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{1}{\sqrt{2}} \left(1 - \sqrt{1 - 2n\|H^{-1}\|_2\,\varepsilon}\right) = \frac{\sqrt{2}n\|H^{-1}\|_2\,\varepsilon}{1 + \sqrt{1 - 2n\|H^{-1}\|_2\,\varepsilon}}. \tag{3.26}$$

If we take $D = I$ in the bound in (3.21), we obtain

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\sqrt{2}n\|H^{-1}\|_2\,\varepsilon}{\sqrt{2} - 1 + \sqrt{1 - 2n\|H^{-1}\|_2\,\varepsilon}}.$$

Thus the bound in (3.21) is at most $\sqrt{2}+1$ times as large as the bound in (3.26). But $\|D_c R^{-1}\|_2 \left(\inf_{D\in\mathcal{D}_n} \|D_c R^{-1} D\|_2 \|D^{-1} R\|_2\right) / \|R\|_2$ in (3.21) can be much smaller than $\|H^{-1}\|_2$. For example, for $R = \begin{bmatrix} \gamma & \gamma \\ 0 & 1 \end{bmatrix}$ with large $\gamma > 0$, $\frac{\|D_c R^{-1}\|_2 \|D_c R^{-1} D\|_2 \|D^{-1} R\|_2}{\|R\|_2} = \Theta(\gamma)$ with $D = \mathrm{diag}(\gamma, 1)$, and $\|H^{-1}\|_2 = \Theta(\gamma^2)$. Thus the bound (3.21) can be much tighter than the bound (3.26).

Remark 3.7. In [3, Theorem 2.3.9], the following rigorous perturbation bound was derived by the matrix-vector equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le 2\chi_C(A)\varepsilon \tag{3.27}$$

under the condition (see [3, Theorem 2.3.9, Remark 2.3.4])

$$\frac{\|A\|_2}{n \min_i a_{ii}} \chi_C^2(A)\varepsilon < \frac{1}{4}. \tag{3.28}$$

By the second inequality in (3.25), the new bound (3.22) is not as tight as (3.27). But, as we mentioned in Remark 3.5, estimating the latter is more expensive than estimating the former. A more serious problem is that the condition (3.28) can be much more constraining than the condition (3.19). In fact, by the first inequality in (3.25), we have

$$\frac{\|A\|_2}{n \min_i a_{ii}} \chi_C^2(A) \ge \frac{\|A\|_2}{n \min_i a_{ii}} \cdot \frac{n^2 a_{nn}}{4\|A\|_2} \|H^{-1}\|_2 \ge \frac{1}{4} n \|H^{-1}\|_2.$$

Thus, if $a_{nn}$is much larger than $\min_i a_{ii}$, then (3.28) is much more constraining than (3.19).

Remark 3.8. In [3, Theorem 2.3.10], the following rigorous bound was derived by the refined matrix equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{2n\|D_c R^{-1}\|_2 \|D_c R^{-1} D\|_2 \|D^{-1} R\|_2}{\|R\|_2}\,\varepsilon \tag{3.29}$$

under the condition

$$n\|D_cR^{-1}\|_2\|D_cR^{-1}D\|_2\|D^{-1}\|_2\varepsilon < 1/4 \tag{3.30}$$

for any $D \in \mathcal{D}_n$. Notice that $\|D_cR^{-1}\|_2\|D_cR^{-1}D\|_2\|D^{-1}\|_2 \geq \|D_cR^{-1}\|_2^2 = \|H^{-1}\|_2$. Thus the condition (3.30) is not only more complicated but also more constraining than the condition (3.19). If we want to make the bound (3.29) similar to the new bound (3.22), then we may minimize $\|D_cR^{-1}D\|_2\|D^{-1}R\|_2$ over the set $\mathcal{D}_n$. But the optimal choice of D may make the condition (3.30) much more constraining than the condition (3.19). Here is an example. Let $R = \left[\begin{smallmatrix} 1 & 1 \\ 0 & \gamma \end{smallmatrix}\right]$ with a large $\gamma > 0$. By (2.10), $D_{r2} = \mathrm{diag}(\sqrt{2}, \gamma)$ is an approximate optimal D. It is easy to verify that $\|D_cR^{-1}\|_2\|D_cR^{-1}D_{r2}\|_2\|D_{r2}^{-1}\|_2 = \Theta(\gamma)$ and $\|H^{-1}\|_2 = \Theta(1)$. Thus the former can be arbitrarily larger than the latter.

4. LU factorization. We first present rigorous perturbation bounds for the LU factors when the given matrix has a general normwise perturbation.

THEOREM 4.1. *Let $A \in \mathbb{R}^{n\times n}$ have nonsingular leading principal submatrices with the LU factorization $A = LU$, where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular, and let $\Delta A \in \mathbb{R}^{n\times n}$ be a small perturbation in A. If*

$$\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F < 1/4, \tag{4.1}$$

then $A + \Delta A$ has the unique LU factorization

$$A + \Delta A = (L + \Delta L)(U + \Delta U), \tag{4.2}$$

where

$$\frac{\|\Delta L\|_F}{\|L\|_F} \leq \frac{2\left(\inf_{D_L\in\mathcal{D}_n}\kappa_2(LD_L^{-1})\right)\frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}}{1+\sqrt{1-4\|L^{-1}\|_2\|U^{-1}\|_2\|A\|_F\frac{\|\Delta A\|_F}{\|A\|_F}}} \tag{4.3}$$

$$\leq 2\left(\inf_{D_L\in\mathcal{D}_n}\kappa_2(LD_L^{-1})\right)\frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}, \tag{4.4}$$

$$\frac{\|\Delta U\|_F}{\|U\|_F} \leq \frac{2\left(\inf_{D_U\in\mathcal{D}_n}\kappa_2(D_U^{-1}U)\right)\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}}{1+\sqrt{1-4\|L^{-1}\|_2\|U^{-1}\|_2\|A\|_F\frac{\|\Delta A\|_F}{\|A\|_F}}} \tag{4.5}$$

$$\leq 2\left(\inf_{D_U\in\mathcal{D}_n}\kappa_2(D_U^{-1}U)\right)\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}. \tag{4.6}$$

Proof. With the condition (4.1), we have for $1 \leq k \leq n$,

$$\|A_k^{-1}\Delta A_k\|_2 \leq \|L_k^{-1}\|_2\|U_k^{-1}\|_2\|\Delta A_k\|_F \leq \|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F < 1.$$

Thus $A_k + t\Delta A_k$ for $t \in [0, 1]$ is nonsingular. In other words, all the leading principal submatrices of $A + t\Delta A$ are nonsingular. Thus, the matrix $A + t\Delta A$ has a unique LU factorization

$$A + t\Delta A = (L + \Delta L(t))(U + \Delta U(t)), \tag{4.7}$$

which, with $\Delta L(1) = \Delta L$ and $\Delta U(1) = \Delta U$, leads to (4.2).

From (4.7), we obtain

$$L^{-1}\Delta L(t) + \Delta U(t)U^{-1} = tL^{-1}\Delta AU^{-1} - L^{-1}\Delta L(t)\Delta U(t)U^{-1}. \tag{4.8}$$

Notice that $L^{-1}\Delta L(t)$ is strictly lower triangular and $\Delta U(t)U^{-1}$ is upper triangular. Taking the Frobenius norm on both sides of (4.8), we obtain

(4.9)
$$\|L^{-1}\Delta L(t)+\Delta U(t)U^{-1}\|_F \le t\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F + \|L^{-1}\Delta L(t)\|_F\|\Delta U(t)U^{-1}\|_F.$$

Let $x(t)=\max(\|L^{-1}\Delta L(t)\|_F, \|\Delta U(t)U^{-1}\|_F)$. Then we have

$$x(t) \le \|L^{-1}\Delta L(t)+\Delta U(t)U^{-1}\|_F, \quad \|L^{-1}\Delta L(t)\|_F\|\Delta U(t)U^{-1}\|_F \le x(t)^2.$$

Thus, from (4.9) it follows that $x(t)^2 - x(t) + t\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F \ge 0$. The assumption (4.1) ensures that the condition of Lemma 2.2 is satisfied. Therefore, by Lemma 2.2 we obtain

$$\max(\|L^{-1}\Delta L\|_F, \|\Delta U U^{-1}\|_F) \le \frac{1}{2}(1-\sqrt{1-4\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F}). \tag{4.10}$$

We now derive perturbation bounds for the L-factor. Let $\left[\begin{smallmatrix} U_{n-1} & u \\ 0 & u_{nn}\end{smallmatrix}\right]$. From (4.8) with $t=1$ it follows that

$$\begin{aligned} L^{-1}\Delta L &= \mathrm{slt}\left(L^{-1}\Delta A \begin{bmatrix} U_{n-1}^{-1} & -U_{n-1}^{-1}u/u_{nn} \\ 0 & 1/u_{nn}\end{bmatrix}\right) - \mathrm{slt}(L^{-1}\Delta L \Delta U U^{-1}) \\ &= \mathrm{slt}\left(L^{-1}\Delta A \begin{bmatrix} U_{n-1}^{-1} & 0 \\ 0 & 0\end{bmatrix}\right) - \mathrm{slt}(L^{-1}\Delta L \Delta U U^{-1}). \end{aligned} \tag{4.11}$$

Multiplying both sides of (4.11) from the left by a diagonal $D_L \in \mathcal{D}_n$ and taking the Frobenius norm, we obtain

$$\|D_L L^{-1}\Delta L\|_F \le \|D_L L^{-1}\|_2\|U_{n-1}^{-1}\|_2\|\Delta A\|_F + \|D_L L^{-1}\Delta L\|_F\|\Delta U U^{-1}\|_F.$$

Using (4.10), we have

$$\|D_L L^{-1}\Delta L\|_F \le \frac{2\|D_L L^{-1}\|_2\|U_{n-1}^{-1}\|_2\|\Delta A\|_F}{1+\sqrt{1-4\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F}}.$$

Combining the inequality $\|\Delta L\|_F \le \|L D_L^{-1}\|_2\|D_L L^{-1}\Delta L\|_F$ and the above inequality leads to (4.3) and then (4.4).

Now we derive perturbation bounds for the U-factor. From (4.8) with $t=1$,

$$\Delta U U^{-1} = \mathrm{ut}(L^{-1}\Delta A U^{-1}) - \mathrm{ut}(L^{-1}\Delta L \Delta U U^{-1}). \tag{4.12}$$

Multiplying both sides of (4.12) from the right by a diagonal $D_U \in \mathcal{D}_n$ and taking the Frobenius norm, we obtain

$$\|\Delta U U^{-1} D_U\|_F \le \|L^{-1}\|_2\|U^{-1}D_U\|_2\|\Delta A\|_F + \|L^{-1}\Delta L\|_F\|\Delta U U^{-1} D_U\|_F.$$

Using (4.10), we have

$$\|\Delta U U^{-1} D_U\|_F \le \frac{2\|L^{-1}\|_2\|U^{-1}D_U\|_2\|\Delta A\|_F}{1+\sqrt{1-4\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_F}}.$$

Therefore, with the inequality $\|\Delta U\|_F \le \|\Delta U U^{-1} D_U\|_F \|D_U^{-1}U\|_2$, we can obtain (4.5) and (4.6). ∎

Remark 4.1. In [6] the following first-order perturbation bounds, which can be estimated in $O(n^2)$ flops, were presented:

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \left(\inf_{D_L\in\mathcal{D}_n} \kappa_2(LD_L^{-1})\right) \frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F}\frac{\|\Delta A\|_F}{\|A\|_F} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_F^2}\right),$$
$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \left(\inf_{D_U\in\mathcal{D}_n} \kappa_2(D_U^{-1}U)\right) \frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F}\frac{\|\Delta A\|_F}{\|A\|_F} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_F^2}\right).$$

Note that the difference between the above first-order bound for the L-factor and the rigorous bound (4.4) is a factor of 2. The same holds for the U-factor as well. Numerical experiments have indicated that the above first-order bounds are good approximations to the corresponding optimal first-order bounds derived by the matrix-vector equation approach in [6]:

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \kappa_L(A)\frac{\|\Delta A\|_F}{\|A\|_F} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_F^2}\right),$$
$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \kappa_U(A)\frac{\|\Delta A\|_F}{\|A\|_F} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_F^2}\right),$$

where

$$\frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F} \le \kappa_L(A) \le \left(\inf_{D_L\in\mathcal{D}_n} \kappa_2(LD_L^{-1})\right)\frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F}, \tag{4.13}$$

$$\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F} \le \kappa_U(A) \le \left(\inf_{D_U\in\mathcal{D}_n} \kappa_2(D_U^{-1}U)\right)\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F}. \tag{4.14}$$

The expressions of $\kappa_L(A)$ and $\kappa_U(A)$ involve an $n^2\times n^2$ matrix defined by the entries of L and U and are expensive to estimate.

To see how partial pivoting and complete pivoting affect the bounds in (4.13) and (4.14), we refer to [6, sections 4 and 5].

Remark 4.2. In [1] the following rigorous bounds were presented:

$$\|\Delta L\|_F \le \frac{\|L\|_2\|L^{-1}\Delta AU^{-1}\|_F}{1-\|L^{-1}\Delta AU^{-1}\|_2}, \quad \|\Delta U\|_F \le \frac{\|U\|_2\|L^{-1}\Delta AU^{-1}\|_F}{1-\|L^{-1}\Delta AU^{-1}\|_2} \tag{4.15}$$

under the condition that $\|L^{-1}\Delta AU^{-1}\|_2 < 1$. If we know only $\|\Delta A\|_F$ or $\|\Delta A\|_2$ rather than $\|L^{-1}\Delta AU^{-1}\|_2$ (this is often the case), then the tightest bounds we can derive from (4.15) are as follows:

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \frac{\kappa_2(L)\frac{\|U^{-1}\|_2\|A\|_F}{\|L\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}}{1-\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_2}, \quad \frac{\|\Delta U\|_F}{\|U\|_F} \le \frac{\kappa_2(U)\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F}\frac{\|\Delta A\|_F}{\|A\|_F}}{1-\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_2},$$

where we assume $\|L^{-1}\|_2\|U^{-1}\|_2\|\Delta A\|_2 < 1$, which is a little less restrictive than (4.1). A comparison between these two bounds with (4.3) and (4.5) shows that the formers can be much larger than the latters when L has bad column scaling or $\|U^{-1}\|_2$ is much larger than $\|U_{n-1}^{-1}\|_2$ (for the L-factor) and when U has bad row scaling (for the U-factor).

If the Gaussian elimination is used for computing the LU factorization of A and runs to completion, then the computed LU factors $\widetilde{L}$ and $\widetilde{U}$ satisfy

$$A+\Delta A = \widetilde{L}\widetilde{U}, \qquad |\Delta A| \le \varepsilon|\widetilde{L}||\widetilde{U}|, \tag{4.16}$$

where $\varepsilon = nu/(1-nu)$ with u being the unit roundoff; see, for example, [14, Theorem 9.3]. In the following theorem we will consider the perturbation ΔA, which has the same form as in (4.16). The perturbation bounds will involve the LU factors of $A+\Delta A$, unlike other perturbation bounds given in this paper, which involve the factors of A. The reason is that the bound on $|\Delta A|$ in (4.16) involves the LU factors of $A+\Delta A$. The perturbation bounds will use a consistent absolute matrix norm (e.g., the 1-norm, ∞-norm, and F-norm), unlike other bounds given in this paper, which use the F-norm or 2-norm.

THEOREM 4.2. *Suppose that $\Delta A \in \mathbb{R}^{n\times n}$ is a perturbation in $A \in \mathbb{R}^{n\times n}$ and $A+\Delta A$ has nonsingular leading principal submatrices with the LU factorization satisfying* (4.16). *Let $\|\cdot\|$ denote a consistent absolute matrix norm. If*

$$\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon < 1/4, \tag{4.17}$$

then A has the unique LU factorization $A = LU$. Let $\Delta L = \widetilde{L} - L$ and $\Delta U = \widetilde{U} - U$. Then

$$\frac{\|\Delta L\|}{\|\widetilde{L}\|} \le \frac{2\frac{\inf_{D_L\in\mathcal{D}_n}\|\widetilde{L}D_L^{-1}\|\cdot\|D_L|\widetilde{L}^{-1}||\widetilde{L}|\,\|}{\|\widetilde{L}\|}\mathrm{cond}(\widetilde{U}_{n-1}^{-1})\varepsilon}{1+\sqrt{1-4\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon}} \tag{4.18}$$

$$\le 2\frac{\inf_{D_L\in\mathcal{D}_n}\|\widetilde{L}D_L^{-1}\|\cdot\|D_L|\widetilde{L}^{-1}||\widetilde{L}|\,\|}{\|\widetilde{L}\|}\mathrm{cond}(\widetilde{U}_{n-1}^{-1})\varepsilon, \tag{4.19}$$

$$\frac{\|\Delta U\|}{\|\widetilde{U}\|} \le \frac{2\frac{\inf_{D_U\in\mathcal{D}_n}\|\,|\widetilde{U}||\widetilde{U}^{-1}|D_U\|\cdot\|D_U^{-1}\widetilde{U}\|}{\|\widetilde{U}\|}\mathrm{cond}(\widetilde{L})\varepsilon}{1+\sqrt{1-4\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon}} \tag{4.20}$$

$$\le 2\frac{\inf_{D_U\in\mathcal{D}_n}\|\,|\widetilde{U}||\widetilde{U}^{-1}|D_U\|\cdot\|D_U^{-1}\widetilde{U}\|}{\|\widetilde{U}\|}\mathrm{cond}(\widetilde{L})\varepsilon. \tag{4.21}$$

Proof. The proof is similar to the proof of Theorem 4.1 and we mainly reverse the roles of A and $A+\Delta A$. Using the bound on $|\Delta A|$ in (4.16) and (4.17), we have for $1 \le k \le n$,

$$\|\widetilde{L}_k^{-1}\Delta A_k\widetilde{U}_k^{-1}\| = \|\,|\widetilde{L}_k^{-1}|\cdot|\widetilde{L}_k|\cdot|\widetilde{U}_k|\cdot|\widetilde{U}_k^{-1}|\varepsilon\,\| \le \mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon < 1. \tag{4.22}$$

For $t \in [0,1]$,

$$(A_k+\Delta A_k) - t\Delta A_k = \widetilde{L}_k\widetilde{U}_k - t\Delta A_k = \widetilde{L}_k[I - t\widetilde{L}_k^{-1}\Delta A_k\widetilde{U}_k^{-1}]\widetilde{U}_k.$$

Thus, by (4.22), the matrix $(A_k+\Delta A_k)-t\Delta A_k$ is nonsingular. Therefore $(A+\Delta A)-t\Delta A$ has the unique LU factorization

$$(A+\Delta A) - t\Delta A = \big(\widetilde{L} - \Delta L(t)\big)\big(\widetilde{U} - \Delta U(t)\big), \tag{4.23}$$

which, with $\Delta L(1) = \Delta L$ and $\Delta U(1) = \Delta U$, gives the LU factorization $A = LU$.

From (4.23), we obtain

$$\widetilde{L}^{-1}\Delta L(t) + \Delta U(t)\widetilde{U}^{-1} = t\widetilde{L}^{-1}\Delta A\widetilde{U}^{-1} + \widetilde{L}^{-1}\Delta L(t)\Delta U(t)\widetilde{U}^{-1}, \tag{4.24}$$

where $\widetilde{L}^{-1}\Delta L(t)$ is strictly lower triangular and $\Delta U(t)\widetilde{U}^{-1}$ is upper triangular. Taking the consistent absolute matrix norm $\|\cdot\|$ on both sides of (4.24) and using the bound on $|\Delta A|$ in (4.16), we obtain

$$\|\widetilde{L}^{-1}\Delta L(t)+\Delta U(t)\widetilde{U}^{-1}\| \le t\,\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon + \|\widetilde{L}^{-1}\Delta L(t)\|\|\Delta U(t)\widetilde{U}^{-1}\|. \tag{4.25}$$

Let $x(t) = \max(\|\widetilde{L}^{-1}\Delta L(t)\|, \|\Delta U(t)\widetilde{U}^{-1}\|)$. Then we have

$$x(t) \le \|\widetilde{L}^{-1}\Delta L(t) + \Delta U(t)\widetilde{U}^{-1}\|, \quad \|\widetilde{L}^{-1}\Delta L(t)\|\|\Delta U(t)\widetilde{U}^{-1}\| \le x(t)^2.$$

Thus, from (4.25) it follows that $x(t)^2 - x(t) + t\,\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon \ge 0$. The assumption (4.17) ensures that the condition of Lemma 2.2 is satisfied. Therefore, by Lemma 2.2,

$$\max(\|\widetilde{L}^{-1}\Delta L\|, \|\Delta U\widetilde{U}^{-1}\|) \le \frac{1}{2}\left(1 - \sqrt{1 - 4\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon}\right). \tag{4.26}$$

We now derive perturbation bounds for the L-factor. Let $\left[\begin{smallmatrix}\widetilde{U}_{n-1} & \tilde{u}\\ 0 & \tilde{u}_{nn}\end{smallmatrix}\right]$. Similarly to (4.11), from (4.24) with $t = 1$ we have

$$\widetilde{L}^{-1}\Delta L = \mathrm{slt}\left(\widetilde{L}^{-1}\Delta A\begin{bmatrix}\widetilde{U}_{n-1}^{-1} & 0\\ 0 & 0\end{bmatrix}\right) + \mathrm{slt}(\widetilde{L}^{-1}\Delta L\Delta U\widetilde{U}^{-1}). \tag{4.27}$$

Then, with the bound on $|\Delta A|$ in (4.16), from (4.27) we obtain that for any $D_L \in \mathcal{D}_n$,

$$\|D_L\widetilde{L}^{-1}\Delta L\| \le \|D_L|\widetilde{L}^{-1}||\widetilde{L}|\,\|\,\mathrm{cond}(\widetilde{U}_{n-1}^{-1})\varepsilon + \|D_L\widetilde{L}^{-1}\Delta L\|\cdot\|\Delta U\widetilde{U}^{-1}\|.$$

Therefore, using (4.26), we have

$$\|D_L\widetilde{L}^{-1}\Delta L\| \le \frac{2\|D_L|\widetilde{L}^{-1}||\widetilde{L}|\,\|\,\mathrm{cond}(\widetilde{U}_{n-1}^{-1})\varepsilon}{1 + \sqrt{1 - 4\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon}}.$$

Combining the inequality $\|\Delta L\| \le \|\widetilde{L}D_L^{-1}\|\cdot\|D_L\widetilde{L}^{-1}\Delta L\|$ and the above inequality leads to (4.18) and then (4.19).

Now we derive perturbation bounds for the U-factor. From (4.24) with $t = 1$, it follows that

$$\Delta U\widetilde{U}^{-1} = \mathrm{ut}(\widetilde{L}^{-1}\Delta A\widetilde{U}^{-1}) + \mathrm{ut}(\widetilde{L}^{-1}\Delta L\Delta U\widetilde{U}^{-1}).$$

Then, for any $D_U \in \mathcal{D}_n$, with (4.16) we obtain

$$\|\Delta U\widetilde{U}^{-1}D_U\| \le \mathrm{cond}(\widetilde{L})\|\,|\widetilde{U}||\widetilde{U}^{-1}|D_U\|\varepsilon + \|\widetilde{L}^{-1}\Delta L\|\cdot\|\Delta U\widetilde{U}^{-1}D_U\|.$$

Therefore, using (4.26), we have

$$\|\Delta U\widetilde{U}^{-1}D_U\| \le \frac{2\|\,|\widetilde{U}||\widetilde{U}^{-1}|D_U\|\mathrm{cond}(\widetilde{L})\varepsilon}{1 + \sqrt{1 - 4\mathrm{cond}(\widetilde{L})\mathrm{cond}(\widetilde{U}^{-1})\varepsilon}}.$$

Combining the inequality $\|\Delta U\| \le \|\Delta U\widetilde{U}^{-1}D_U\|\cdot\|D_U^{-1}\widetilde{U}\|$ and the above inequality, we obtain (4.20) and then (4.21). ∎

Remark 4.3. For some choices of norms, we can remove the scaling matrices in Theorem 4.2. From Lemma 2.1 we observe that if we take the 1-norm or ∞-norm, the bounds (4.18), (4.19), (4.20), and (4.21) can be written as ($p = 1, \infty$)

$$\frac{\|\Delta L\|_p}{\|\widetilde{L}\|_p} \le \frac{2\frac{\||\widetilde{L}||\widetilde{L}^{-1}||\widetilde{L}|\|_p}{\|\widetilde{L}\|_p}\mathrm{cond}_p(\widetilde{U}_{n-1}^{-1})\varepsilon}{1 + \sqrt{1 - 4\mathrm{cond}_p(\widetilde{L})\mathrm{cond}_p(\widetilde{U}^{-1})\varepsilon}} \le 2\frac{\||\widetilde{L}||\widetilde{L}^{-1}||\widetilde{L}|\|_p}{\|\widetilde{L}\|_p}\mathrm{cond}_p(\widetilde{U}_{n-1}^{-1})\varepsilon,$$

$$\frac{\|\Delta U\|_p}{\|\widetilde{U}\|_p} \le \frac{2\frac{\||\widetilde{U}||\widetilde{U}^{-1}||\widetilde{U}|\|_p}{\|\widetilde{U}\|_p}\mathrm{cond}_p(\widetilde{L})\varepsilon}{1 + \sqrt{1 - 4\mathrm{cond}_p(\widetilde{L})\mathrm{cond}_p(\widetilde{U}^{-1})\varepsilon}} \le 2\frac{\||\widetilde{U}||\widetilde{U}^{-1}||\widetilde{U}|\|_p}{\|\widetilde{U}\|_p}\mathrm{cond}_p(\widetilde{L})\varepsilon$$

under the condition $\mathrm{cond}_p(\widetilde{L})\mathrm{cond}_p(\widetilde{U}^{-1})\varepsilon < 1/4$. In [5] the following first-order bounds were derived (with $\|\cdot\|$ being a consistent absolute matrix norm):

$$\frac{\|\Delta L\|}{\|\widetilde{L}\|} \le \frac{\| \, |\widetilde{L}||\widetilde{L}^{-1}||\widetilde{L}| \, \|}{\|\widetilde{L}\|}\mathrm{cond}(\widetilde{U}_{n-1}^{-1})\varepsilon + O(\varepsilon^2), \tag{4.28}$$

$$\frac{\|\Delta U\|}{\|\widetilde{U}\|} \le \frac{\| \, |\widetilde{U}||\widetilde{U}^{-1}||\widetilde{U}| \, \|}{\|\widetilde{U}\|}\mathrm{cond}(\widetilde{L})\varepsilon + O(\varepsilon^2). \tag{4.29}$$

We can see the obvious relation between these first-order bounds and the rigorous bounds derived above when the 1-norm and ∞-norm are used. For estimation of the perturbation bounds, we refer to [5]. We would like to point out that to our knowledge there are no optimal or nearly optimal first-order bounds or rigorous bounds derived by the matrix-vector equation approach in the literature.

To see how partial pivoting, rook pivoting, and complete pivoting affect the first-order bounds in (4.28) and (4.29), we refer to [5, section 4.2].

5. QR factorization. In this section we consider the perturbation of the R-factor of the QR factorization of A. As we do not have any new result concerning the Q-factor, we will not consider it. First we present rigorous perturbation bounds when the given matrix A has a general normwise perturbation.

THEOREM 5.1. *Let $A \in \mathbb{R}^{m\times n}$ be of full column rank with QR factorization $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns and $R \in \mathbb{R}^{n\times n}$ is upper triangular with positive diagonal entries. If the perturbation matrix $\Delta A \in \mathbb{R}^{m\times n}$ satisfies*

$$\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2} < \sqrt{3/2} - 1, \tag{5.1}$$

then $A + \Delta A$ has a unique QR factorization

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R), \tag{5.2}$$

where, with ρ_D defined in (2.7),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\sqrt{2}\left(\inf_{D\in\mathcal{D}_n}\rho_D\kappa_2(D^{-1}R)\right)\left(\frac{\|Q^T\Delta A\|_F}{\|A\|_2} + \kappa_2(A)\frac{\|\Delta A\|_F^2}{\|A\|_2^2}\right)}{\sqrt{2} - 1 + \sqrt{1 - 4\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2} - 2\kappa_2^2(A)\frac{\|\Delta A\|_F^2}{\|A\|_2^2}}} \tag{5.3}$$

$$\le \frac{\sqrt{3}\left(\inf_{D\in\mathcal{D}_n}\rho_D\kappa_2(D^{-1}R)\right)\frac{\|\Delta A\|_F}{\|A\|_2}}{\sqrt{2} - 1 + \sqrt{1 - 4\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2} - 2\kappa_2^2(A)\frac{\|\Delta A\|_F^2}{\|A\|_2^2}}} \tag{5.4}$$

$$\le (\sqrt{6} + \sqrt{3})\left(\inf_{D\in\mathcal{D}_n}\rho_D\kappa_2(D^{-1}R)\right)\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{5.5}$$

Proof. Notice that for any $t \in [0,1]$, $Q^T(A + t\Delta A) = R(I + tR^{-1}Q^T\Delta A) = R(I + tA^\dagger\Delta A)$ and $\|A^\dagger\Delta A\|_2 < 1$ by (5.1). Thus $Q^T(A + t\Delta A)$ is nonsingular, and then $A + t\Delta A$ has full column rank and has the unique QR factorization

$$A + t\Delta A = (Q + \Delta Q(t))(R + \Delta R(t)), \tag{5.6}$$

which, with $\Delta Q(1) = \Delta Q$ and $\Delta R(1) = \Delta R$, gives (5.2).

From (5.6), we obtain

$$R^T\Delta R(t) + \Delta R(t)^T R = tR^TQ^T\Delta A + t\Delta A^TQR + t^2\Delta A^T\Delta A - \Delta R(t)^T\Delta R(t).$$

Multiplying the above by R^{-T} from left and R^{-1} from right, we obtain

$$\begin{aligned} &R^{-T}\Delta R(t)^T + \Delta R(t)R^{-1} \\ &= tQ^T\Delta AR^{-1} + tR^{-T}\Delta A^TQ + R^{-T}\left(t^2\Delta A^T\Delta A - \Delta R(t)^T\Delta R(t)\right)R^{-1}. \end{aligned}$$

Since ΔRR^{-1} is upper triangular, it follows that

$$\begin{aligned} \Delta R(t)R^{-1} = \operatorname{up}\big[&tQ^T\Delta AR^{-1} + tR^{-T}\Delta A^TQ \\ &+ R^{-T}\big(t^2\Delta A^T\Delta A - \Delta R(t)^T\Delta R(t)\big)R^{-1}\big]. \end{aligned} \tag{5.7}$$

Thus, by (2.6), the quantity $\|\Delta R(t)R^{-1}\|_F$ verifies

$$\|\Delta R(t)R^{-1}\|_F \le \frac{1}{\sqrt{2}}\left(2t\|R^{-1}\|_2\|Q^T\Delta A\|_F + t^2\|R^{-1}\|_2^2\|\Delta A\|_F^2 + \|\Delta R(t)R^{-1}\|_F^2\right).$$

It can easily be verified that $1 - 4t\|R^{-1}\|_2\|Q^T\Delta A\|_F - 2t^2\|R^{-1}\|_2^2\|\Delta A\|_F^2 > 0$ when $\|R^{-1}\|_2\|\Delta A\|_F < \sqrt{3/2} - 1$, which is equivalent to the condition (5.1). The condition of Lemma 2.2 is thus satisfied and we can apply it, with $x(t) = \|\Delta R(t)R^{-1}\|_F$, to get

$$\|\Delta RR^{-1}\|_F \le \frac{1}{\sqrt{2}}\left(1 - \sqrt{1 - 4\|R^{-1}\|_2\|Q^T\Delta A\|_F - 2\|R^{-1}\|_2^2\|\Delta A\|_F^2}\right). \tag{5.8}$$

For any $D \in \mathcal{D}_n$, we have from (5.7) with $t = 1$ that

$$\begin{aligned} \Delta RR^{-1}D = \operatorname{up}&\left[(Q^T\Delta AR^{-1}D) + D^{-1}(DR^{-T}\Delta A^TQ)D\right] \\ &+ \operatorname{up}\left[R^{-T}\left(\Delta A^T\Delta A - \Delta R^T\Delta R\right)R^{-1}D\right]. \end{aligned} \tag{5.9}$$

Then, by (2.7), it follows that

$$\begin{aligned} \|\Delta RR^{-1}D\|_F \le{}& \rho_D\|Q^T\Delta A\|_F\|R^{-1}D\|_2 + \|R^{-1}\|_2\|\Delta A\|_F^2\|R^{-1}D\|_2 \\ &+ \|\Delta RR^{-1}\|_F\|\Delta RR^{-1}D\|_F. \end{aligned}$$

Therefore, using (5.8) and the fact that $\rho_D \ge 1$ (see (2.7)), we obtain

$$\|\Delta RR^{-1}D\|_F \le \frac{\sqrt{2}\rho_D\|R^{-1}D\|_2(\|Q^T\Delta A\|_F + \|R^{-1}\|_2\|\Delta A\|_F^2)}{\sqrt{2} - 1 + \sqrt{1 - 4\|R^{-1}\|_2\|\Delta A\|_F - 2\|R^{-1}\|_2^2\|\Delta A\|_F^2}}. \tag{5.10}$$

Combining the inequality $\|\Delta R\|_F \le \|\Delta RR^{-1}D\|_F\|D^{-1}R\|_2$ and the above inequality we obtain (5.3). Since $\|Q^T\Delta A\|_F \le \|\Delta A\|_F$ and (5.1) holds, (5.4) follows from (5.3). Then (5.5) is obtained. ∎

Remark 5.1. In [9] the following first-order bound was derived by the refined matrix equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \left(\inf_{D\in\mathcal{D}_n}\rho_D\kappa_2(D^{-1}R)\right)\frac{\|Q^T\Delta A\|_F}{\|A\|_2} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_2^2}\right). \tag{5.11}$$

Some practice choices of D were given in [9] to estimate the above bound. This first-order bound (5.11) has some similarity to (5.3). But if $Q^T\Delta A = 0$ (i.e., ΔA lies in the orthogonal complement of the range of A), then this first-order bound becomes useless, but the rigorous bound (5.3) clearly shows how R is sensitive to the perturbation ΔA. Numerical experiments have indicated that this first-order bound is

a good approximation to the optimal first-order bound derived by the matrix-vector equation approach in [9]:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_R(A)\frac{\|Q^T\Delta A\|_F}{\|A\|_2} + O\left(\frac{\|\Delta A\|_F^2}{\|A\|_2^2}\right),$$

where

$$1 \le \kappa_R(A) \le \inf_{D\in\mathcal{D}_n} \rho_D\kappa_2(D^{-1}R).$$

The expression of $\kappa_R(A)$ involves an $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ lower triangular matrix and an $\frac{n(n+1)}{2} \times n^2$ matrix defined by the entries of R and is expensive to estimate.

If the standard column pivoting strategy is used in computing the QR factorization, the quantity $\inf_{D\in\mathcal{D}_n} \rho_D\kappa_2(D^{-1}R)$ can be bounded by a function of n; see [9, sections 5 and 6].

Remark 5.2. The following rigorous bound was derived in [20] by the classic matrix equation approach:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\sqrt{2}\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}{1-\kappa_2(A)\frac{\|\Delta A\|_2}{\|A\|_2}} \tag{5.12}$$

under the condition $\kappa_2(A)\|\Delta A\|_2/\|A\|_2 < 1$, which is a little less restrictive than (5.1). Note that if $D = I$, then $\rho_D\kappa_2(D^{-1}R) = \sqrt{2}\kappa_2(A)$. If R has bad row scaling and the 2-norm of its rows decreases from the top to bottom, then $\inf_{D\in\mathcal{D}_n} \rho_D\kappa_2(D^{-1}R)$ can be much smaller than $\kappa_2(A)$. For example, for $R = \mathrm{diag}(\gamma, 1)$ with large γ, $\rho_D\kappa_2(D^{-1}R) = \Theta(1)$ with $D = \mathrm{diag}(\gamma, 1)$, $\kappa_2(A) = \kappa_2(R) = \Theta(\gamma)$. Thus the new rigorous bounds can be much tighter than (5.12). Here we would like to point out that to our knowledge there are no rigorous bounds derived by the matrix-vector equation approach in the literature.

For the componentwise perturbation ΔA which has the form of backward error we could expect from a standard QR factorization algorithm, the analysis has been done in [10]. For completeness, we give the result here, without a proof.

THEOREM 5.2. *Let $A \in \mathbb{R}^{m\times n}$ be of full column rank with QR factorization $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns and $R \in \mathbb{R}^{n\times n}$ is upper triangular with positive diagonal entries. Let $\Delta A \in \mathbb{R}^{m\times n}$ be a perturbation matrix in A such that*

$$|\Delta A| \le \varepsilon C|A|, \quad C \in \mathbb{R}^{m\times m}, \quad 0 \le c_{ij} \le 1, \quad \varepsilon \text{ a small constant.} \tag{5.13}$$

If

$$\mathrm{cond}_2(R^{-1})\varepsilon < \frac{\sqrt{3/2}-1}{m\sqrt{n}}, \tag{5.14}$$

then $A + \Delta A$ has a unique QR factorization

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R), \tag{5.15}$$

where, with ρ_D defined in (2.7),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \sqrt{6}mn^{1/2}\frac{\inf_{D\in\mathcal{D}_n}\rho_D\||R||R^{-1}|D\|_2\|D^{-1}R\|_2}{\|R\|_2}\varepsilon. \tag{5.16}$$

The assumption (5.13) on the perturbation ΔA can (essentially) handle two special cases. First, there is a small relative componentwise perturbation in A; i.e., $|\Delta A| \leq \varepsilon |A|$. Second, there is a small relative columnwise perturbation in A; i.e., $\|\Delta A(:,j)\|_2 \leq \varepsilon \|A(:,j)\|_2$ for $1 \leq j \leq n$; see, e.g., [7, section 2]. The second case may arise when ΔA is the backward error of the QR factorization by a standard algorithm; see [14, Chap. 19]. For practical choices of D to estimate the bound (5.16), we refer to [7, 10].

6. Summary. We have presented new rigorous normwise perturbation bounds for the Cholesky, LU, and QR factorizations with normwise and componentwise perturbations in the given matrix by using a hybrid approach of the classic and refined matrix equation approaches. Each of the new rigorous perturbation bounds is a small constant multiple of the corresponding first-order perturbation bound obtained by the refined matrix equation approach in the literature and can be estimated efficiently. These new bounds can be much tighter than the existing rigorous bounds obtained by the classic matrix equation approach, while the conditions for the former to hold are almost as moderate as the conditions for the latter to hold.

Acknowledgments. The authors thank Gilles Villard for early discussions on this work. They are grateful to the referees' very helpful suggestions, which improved the presentation of the paper.

REFERENCES

[1] A. Barrlund, *Perturbation bounds for the LDL^H and the LU factorizations*, BIT, 31 (1991), pp. 358–363.
[2] R. Bhatia, *Matrix factorizations and their perturbations*, Linear Algebra Appl., 197–198 (1994), pp. 245–276.
[3] X.-W. Chang, *Perturbation Analysis of Some Matrix Factorizations*, Ph.D. thesis, Computer Science, McGill University, Montreal, Canada, February 1997.
[4] X.-W. Chang, *Perturbation analyses for the Cholesky factorization with backward rounding errors*, in Proceedings of the Workshop on Scientific Computing, Hong Kong, 1997, pp. 180–187.
[5] X.-W. Chang, *Some features of Gaussian elimination with rook pivoting*, BIT, 42 (2002), pp. 66–83.
[6] X.-W. Chang and C. C. Paige, *On the sensitivity of the LU factorization*, BIT, 38 (1998), pp. 486–501.
[7] X.-W. Chang and C. C. Paige, *Componentwise perturbation analyses for the QR factorization*, Numer. Math., 88 (2001), pp. 319–345.
[8] X.-W. Chang, C. Paige, and G. Stewart, *New perturbation analyses for the Cholesky factorization*, IMA J. Numer. Anal., 16 (1996), pp. 457–484.
[9] X.-W. Chang, C. Paige, and G. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 775–791.
[10] X.-W. Chang, D. Stehlé, and G. Villard, *Perturbation analysis of the QR factor R in the context of LLL lattice basis reduction*, 25 pages, submitted. Available at http://perso.ens-lyon.fr/damien.stehle/QRPERTURB.html.
[11] J. W. Demmel, *On floating point errors in Cholesky*, Technical report CS 89-87, Department of Computer Science, University of Tennessee, Knoxville, TN, 1989, 6 pages. LAPACK Working Note 14.
[12] F. M. Dopico and J. M. Molera, *Perturbation theory for factorizations of LU type through series expansions*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 561–581.
[13] Z. Drmač, M. Omladič, K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.
[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002.
[15] I. Morel, G. Villard, D. Stehlé, *H-LLL: Using Householder inside LLL*, in Proceedings of ISSAC, Seoul, Korea, 2009, pp. 271–278.

[16] P. Q. Nguyen, D. Stehlé, *An LLL algorithm with quadratic complexity*, SIAM J. Comput., 39 (2009), pp. 874–903.
[17] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.
[18] G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1141–1146.
[19] G. W. Stewart, *On the perturbation of LU and Cholesky factors*, IMA J. Numer. Anal., 17 (1997), pp. 1–6.
[20] J.-G. Sun, *Perturbation bounds for the Cholesky and QR factorizations*, BIT, 31 (1991), pp. 341–352.
[21] J.-G. Sun, *Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations*, Linear Algebra Appl., 173 (1992), pp. 77–97.
[22] J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.
[23] J.-G. Sun, *On the perturbation bounds for the QR factorization*, Linear Algebra Appl., 215 (1995), pp. 95–112.
[24] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[25] H. Zha, *A componentwise perturbation analysis of the QR decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1124–1131.