Perturbation Analyses for the Cholesky Factorization with Backward Rounding Errors

Xiao-Wen Chang

Technical Report No. SOCS-96.3
July 1996

PERTURBATION ANALYSES FOR THE CHOLESKY FACTORIZATION WITH BACKWARD ROUNDING ERRORS

XIAO-WEN CHANG*

Abstract. This paper gives perturbation analyses for the Cholesky factorization $A = R^T R$ of a real symmetric positive definite matrix A, with the form of perturbations we could expect from the equivalent backward error in A resulting from numerically stable computations on finite precision floating point computers. First-order and rigorous perturbation bounds are are obtained. The analyses more accurately reflect the sensitivity of the problem than previous such results. Both numerical results and an analysis show the standard method of pivoting usually improves the condition of the problem. A useful upper bound on the condition of the problem is derived.

Key Words. Cholesky factorization, perturbation analysis, backward rounding error, condition estimation, pivoting

AMS Subject Classifications: 15A23, 65F35

1. Introduction. Let $A \in \mathbf{R}^{n\times n}$ be a symmetric positive definite matrix. Then there exists a unique upper triangular matrix $R \in \mathbf{R}^{n\times n}$ with positive diagonal elements such that

$$A = R^T R.$$

This factorization is called the Cholesky factorization, and R is called the Cholesky factor of A.

There are different algorithmic forms of Cholesky factorization. The following algorithm is the 'bordered' form.

Algorithm CHOL: Given a symmetric positive definite $A \in \mathbf{R}^{n\times n}$ this algorithm computes the Cholesky factorization $A = R^T R$.

```
for j = 1 : n
    for i = 1 : j - 1
        r_ij = (a_ij - sum_{k=1}^{i-1} r_ki r_kj) / r_ii
    end
    r_jj = (a_jj - sum_{k=1}^{j-1} r_kj^2)^{1/2}
end
```

Suppose there is a perturbation ΔA in A, and $A + \Delta A$ is also symmetric positive definite, then $A + \Delta A$ has a unique Cholesky factorization

$$A + \Delta A = (R + \Delta R)^T (R + \Delta R).$$

The goal of the perturbation analysis is to determine a bound on $\|\Delta R\|$ (or $|\Delta R|$) in terms of $\|\Delta A\|$ (or $|\Delta A|$). There have been several papers dealing with the perturbation analysis for the Cholesky factor for a general perturbation ΔA. The first norm-based result was that of Stewart [6]. It was further modified and improved by Sun [9, 1991], who included a *first-order* perturbation result. Using a different approach, Stewart [7] obtained the same first-order perturbation result. Component

* School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7, (chang@cs.mcgill.ca). This research was supported by NSERC of Canada Grant OGP0009236.

perturbation analyses have been given by Sun [10, 11]. Recently Drmač et al [3] presented perturbation results of a different flavor. They made a perturbation analysis for $H = D_c^{-1} A D_c$ with $D_c = \mathrm{diag}(a_{ii}^{1/2})$ instead of A. Most recently, Chang et al [1] presented new perturbation results using two different approaches, which gave significant improvements on previous corresponding results. For ΔA corresponding to the backward rounding error in A resulting from numerical stable computations, e.g. Algorithm CHOL, on finite precision floating point computers, Drmač et al [3] presented a nice norm-based perturbation result using their $H = D_c^{-1} A D_c^{-1}$ approach. Sun [10, 11] also included component perturbation bounds for a different form of the backward rounding error in A.

The main purpose of this paper is to establish first-order perturbation bounds for ΔA corresponding to the backward rounding error, and present a new condition number which more closely reflect the true sensitivity of the problem. Also we obtain rigorous perturbation bounds. Our bounds are potentially much smaller than the equivalent bound in [3].

In Section 2 we define some notation and cite a well-known result of van der Sluis, which will be used in estimating condition numbers. In Section 3 we survey important first-order results for a general perturbation ΔA. In Section 4 we give perturbation analyses for backward rounding errors ΔA, first by the so called matrix-vector equation approach, then by the so called matrix-equation approach, and obtain an upper bound on the condition number when standard pivoting is used in computing the Cholesky factorization. In Section 5 we give numerical results and suggest a practical condition estimator.

2. Notation, and van der Sluis's results. To simplify the presentation, for any $n \times n$ matrix X, we define the upper triangular matrix

$$\mathrm{up}(X) \equiv \begin{bmatrix} \frac{1}{2}x_{11} & x_{12} & \cdots & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{2}x_{nn} \end{bmatrix}, \tag{2.1}$$

and the diagonal matrix

$$\mathrm{diag}(X) \equiv \mathrm{diag}(x_{11}, x_{22}, \ldots, x_{nn}) \equiv \mathrm{diag}(x_{ii}).$$

For any matrix $C \equiv (c_{ij}) \equiv [c_1, \ldots, c_n] \in \mathcal{R}^{n \times n}$, denote by $c_j^{(i)}$ the vector of the leading i elements of c_j. Using 'u' to denote 'upper', we define

$$\mathrm{uvec}(C) \equiv \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ \cdot \\ c_n^{(n)} \end{bmatrix}, \tag{2.2}$$

the vector formed by stacking the columns of the upper triangular part of C into one long vector.

Note that for any upper triangular X, $\|\mathrm{uvec}(X)\|_2 = \|X\|_F$. To help our norm analysis, for any matrix $C \in \mathcal{R}^{n \times n}$ we define

$$\mathrm{duvec}(C) \equiv \mathcal{D}\, \mathrm{uvec}(C), \tag{2.3}$$

where

$$\mathcal{D} = \operatorname{diag}\,(1, \underbrace{\sqrt{2}, 1}_{2}, \ldots, \underbrace{\sqrt{2}, \sqrt{2}, \ldots, \sqrt{2}, 1}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}.$$

Thus for any symmetric matrix X we have $\|\mathrm{duvec}(X)\|_2 = \|X\|_F$.

Let $\mathbf{D}_n$ be the set of all $n \times n$ real positive definite diagonal matrices, and let $e = (1, 1, \ldots, 1)^T \in \mathbf{R}^n$.

Now we introduce the following results due to van der Sluis [12], which will be used later.

THEOREM 2.1. [12] *Let S, $T \in \mathbf{R}^{n\times n}$ and let S be nonsingular, and define*

$$D_r = \operatorname{diag}(\|S(i,:)\|_2).$$

Then

$$\|TD_r\|_2\,\|D_r^{-1}S\|_2 \le \sqrt{n} \inf_{D\in\mathbf{D}_n} \|TD\|_2\,\|D^{-1}S\|_2. \tag{2.4}$$

Thus if $T = S^{-1}$, then

$$\kappa_2(D_r^{-1}S) \le \sqrt{n} \inf_{D\in\mathbf{D}_n} \kappa_2(D^{-1}S). \tag{2.5}$$

In particular, if S is symmetric positive definite, define $D_* = \operatorname{diag}(S)^{1/2}$, then

$$\kappa_2(D_*^{-1}SD_*^{-1}) \le n \inf_{D\in\mathbf{D}_n} \kappa_2(D^{-1}SD^{-1}). \tag{2.6}$$

3. Previous perturbation bounds for a general perturbation ΔA. In this section we summarize the strongest norm-based first-order results for a general perturbation ΔA presented in Chang et al [1], since those results will be compared with the new results for ΔA having the form of backward rounding error.

THEOREM 3.1. [1] *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^TR$. Let ΔA be symmetric. If $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ satisfies*

$$\kappa_2(A)\epsilon < 1,$$

where $\kappa_2(A) \equiv \|A\|_2\,\|A^{-1}\|_2$, then $A + \Delta A$ has the Cholesky factorization

$$A + \Delta A = (R + \Delta R)^T(R + \Delta R),$$

such that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_C(A)\epsilon + O(\epsilon^2), \tag{3.1}$$

$$\frac{1}{2}\kappa_2^{1/2}(A) \le \kappa_C(A) \le \kappa_{C1}(A) \le \kappa_2(A), \tag{3.2}$$

where

$$\kappa_C(A) \equiv \|\widehat{W}_R^{-1}\|_2\,\|A\|_2^{1/2} \equiv \|(\widehat{\mathcal{D}}W_R)^{-1}\|_2\,\|A\|_2^{1/2}, \tag{3.3}$$

with

$$\widehat{\mathcal{D}} \equiv \operatorname{diag}(2, \underbrace{\sqrt{2}, 2}_{2}, \ldots, \underbrace{\sqrt{2}, \sqrt{2}, \ldots, \sqrt{2}, 2}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$$

and $W_R \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ *as*

$$\tag{3.4} \left[\begin{array}{cccccccccccc}
r_{11} & & & & & & & & & & & \\
r_{12} & r_{11} & & & & & & & & & & \\
 & r_{12} & r_{22} & & & & & & & & & \\
r_{13} & & & r_{11} & & & & & & & & \\
 & r_{13} & r_{23} & r_{12} & r_{22} & & & & & & & \\
 & & & r_{13} & r_{23} & r_{33} & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & \\
r_{1n} & & & & & & & r_{11} & & & & \\
 & r_{1n} & r_{2n} & & & & & r_{12} & r_{22} & & & \\
 & & & r_{1n} & r_{2n} & r_{3n} & & r_{13} & r_{23} & r_{33} & & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \\
 & & & & & & & r_{1n} & r_{2n} & r_{3n} & \cdot & r_{nn}
\end{array}\right],$$

and

$$\kappa_{C1}(A) \equiv \inf_{D \in \mathbf{D}_n} \kappa_{C1}(A, D),$$
$$\kappa_{C1}(A, D) \equiv \kappa_2(R)\kappa_2(D^{-1}R).$$

If standard pivoting is used in computing the Cholesky factorization $PAP^T = R^TR$, *we have*

$$\tag{3.5} \frac{1}{2}\kappa_2^{1/2}(A) \le \kappa_C(PAP^T) \le \kappa_{C1}(PAP^T) \le \kappa_2^{1/2}(A)\sqrt{2n(n+1)(4^n + 6n - 1)}/6.$$

[1] showed the bound (3.1) is attainable to first order in ϵ. So $\kappa_C(A)$ is the true condition number for the Cholesky factor R. It is expensive to evaluate $\kappa_C(A)$ directly by using the usual methods. But we can use $\kappa_{C1}(A)$ as an approximation of $\kappa_C(A)$. According to van de Sluis's Theorem 2.1, if D is chosen to equilibrate the rows of R, then the resulting $\kappa_{C1}(A, D)$ will be nearly minimal. This results in a practical estimator — $\kappa_{C1}(A, D)$, which can be estimated using a standard condition estimator in $O(n^2)$.

4. Perturbation bounds for backward rounding errors ΔA. In this section we first use the so called matrix-vector equation approach to derive a tight perturbation bound, leading to the condition number $\chi_C(A)$ for perturbation having the form of equivalent backward rounding error, then use the so called matrix equation approach to derive a practical perturbation bound, leading to a practical estimate $\chi_{C1}(A)$ of $\chi_C(A)$. We also compare $\chi_C(A)$ with $\kappa_C(A)$, and $\chi_{C1}(A)$ with $\kappa_{C1}(A)$. Finally we show how standard pivoting improves the condition number $\chi_C(A)$.

Before proceeding we introduce the following result due to Demmel [2], also see Higham [4, Theorems 10.5 and 10.7].

LEMMA 4.1. [2] *Let* $A \equiv D_c H D_c \in \mathbf{R}^{n\times n}$ *be a symmetric positive definite floating point matrix, where* $D_c = \mathrm{diag}(A)^{1/2}$. *If*

$$n\epsilon\,\|H^{-1}\|_2 < 1, \tag{4.1}$$

where $\epsilon \equiv (n+1)u/(1-2(n+1)u)$ *with* u *being the unit round-off, then the Cholesky factorization applied to* A *succeeds (barring underflow and overflow) and produces a nonsingular* $\tilde{R}$, *which satisfies*

$$A + \Delta A = \tilde{R}^T\tilde{R}, \qquad |\Delta A| \le \epsilon\, dd^T, \tag{4.2}$$

where $d_i = a_{ii}^{1/2}$.

This lemma is applicable not only to Algorithm CHOL but also to its standard mathematically equivalent variants.

4.1. Matrix-vector equation approach. Based on Lemma 4.1, we can establish the following bound for the computed Cholesky factor $\tilde{R}$.

THEOREM 4.2. *With* $A = R^T R$ *and the same assumptions as in Lemma* 4.1, *for the perturbation* ΔA *and result* $\tilde{R}$ *in* (4.2) *we have*

$$|\mathrm{uvec}(\Delta R)| \le \epsilon\,|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}|\,\mathrm{duvec}(ee^T) + O(\epsilon^2), \tag{4.3}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \chi_C(A)\,\epsilon + O(\epsilon^2), \tag{4.4}$$

where $\Delta R \equiv \tilde{R} - R$, $\hat{R} \equiv R D_c^{-1}$ *and*

$$\chi_C(A) \equiv n\,\|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2/\|A\|_2^{1/2}, \tag{4.5}$$

with

$$\mathcal{D}_c \equiv \mathrm{diag}\,(a_{11}^{1/2}, \underbrace{a_{22}^{1/2}, a_{22}^{1/2}}_{2}, \ldots, \underbrace{a_{nn}^{1/2}, a_{nn}^{1/2}, \ldots, a_{nn}^{1/2}}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}, \tag{4.6}$$

and $\widehat{W}_{\hat{R}}$ *being just* $\widehat{W}_R$ *in* (3.3) *with each entry* r_{ij} *replaced by* $\hat{r}_{ij}$.

Proof. Let $G \equiv \Delta A/\epsilon$. By (4.2) we have

$$\|D_c^{-1}\Delta A D_c^{-1}\|_2 \le \epsilon\,\|D_c^{-1} dd^T D_c^{-1}\|_F = \epsilon\,\|ee^T\|_F = n\epsilon,$$

which, combined with (4.1), implies

$$\|H^{-1}\|_2 \|D_c^{-1}\Delta A D_c^{-1}\|_2 < 1.$$

Thus $A + tG = D_c(H + D_c^{-1}\Delta A D_c^{-1} t/\epsilon)D_c$ is symmetric positive definite for all $t \in [0, \epsilon]$, and has the Cholesky factorization

$$A + tG = R^T(t)R(t), \qquad 0 \le t \le \epsilon, \tag{4.7}$$

with $R(0) = R$ and $R(\epsilon) = \tilde{R} \equiv R + \Delta R$.

Differentiating (4.7) and setting $t = 0$ in the result gives

$$G = R^T\dot{R}(0) + \dot{R}(0)^T R, \tag{4.8}$$

which, combined with $R = \hat{R}D_c$, gives

$$\hat{R}^T \dot{R}(0) D_c^{-1} + D_c^{-1} \dot{R}^T(0) \hat{R} = D_c^{-1} G D_c^{-1}. \tag{4.9}$$

The upper and lower triangular parts of (4.9) contain identical information, and it is straightforward to show that by using (2.2) and (2.3) the upper triangular part of (4.9) can be rewritten as the following matrix-vector equation form

$$\widehat{W}_{\hat{R}}\, \mathrm{uvec}(\dot{R}(0) D_c^{-1}) = \mathrm{duvec}(D_c^{-1} G D_c^{-1}). \tag{4.10}$$

Notice from (2.2) that $\mathrm{uvec}(\dot{R}(0) D_c^{-1}) = \mathcal{D}_c^{-1}\, \mathrm{uvec}(\dot{R}(0))$ with $\mathcal{D}_c$ as in (4.6), then from (4.10) we have

$$\mathrm{uvec}(\dot{R}(0)) = \mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\, \mathrm{duvec}(D_c^{-1} G D_c^{-1}), \tag{4.11}$$

which with $G = \Delta A/\epsilon$ and (4.2) gives both

$$|\mathrm{uvec}(\dot{R}(0))| \le |\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}|\, \mathrm{duvec}(D_c^{-1} d d^T D_c^{-1}) = |\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}|\, \mathrm{duvec}(e e^T), \tag{4.12}$$

and

$$\|\dot{R}(0)\|_F \le \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2\, \|D_c^{-1} d d^T D_c^{-1}\|_F = n \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2. \tag{4.13}$$

The Taylor expansion of $R(t)$ about $t = 0$ gives at $t = \epsilon$

$$R + \Delta R \equiv \tilde{R} = R(\epsilon) = R(0) + \epsilon \dot{R}(0) + O(\epsilon^2), \tag{4.14}$$

which, combined with (4.12) and (4.13), gives (4.3) and (4.4), respectively. □

REMARK 4.1. From (4.4) we see the quantity $\chi_C(A)$ can be thought of as a condition number for perturbation errors satisfying (4.2). It is easy to verify that $D_c \equiv \mathrm{diag}(a_{ii}^{1/2}) = \mathrm{diag}(\|R(:,j)\|_2)$. Thus $\hat{R}$, the Cholesky factor of H ($H \equiv D_c^{-1} A D_c^{-1} = \hat{R}^T \hat{R}$), has columns of unit Euclidean length. That is the reason that we use the notation D_c, where 'c' denote 'column'. We might want to simplify $\chi_C(A)$. Since

$$\|\mathcal{D}_c\|_2 = \|D_c\|_2 \le \|A\|_2^{1/2},$$

we have

$$\chi_C(A) \le n \|\widehat{W}_{\hat{R}}^{-1}\|_2.$$

It is easy to see that if the diagonal elements of A have almost the same magnitude and n is not very large, $n\|\widehat{W}_{\hat{R}}^{-1}\|_2$ will be close to $\chi_C(A)$. But in some other cases $n\|\widehat{W}_{\hat{R}}^{-1}\|_2$ can be arbitrarily larger than $\chi_C(A)$. For example, if with small ε

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1+\varepsilon^2 & \varepsilon^2 \\ 0 & \varepsilon^2 & \varepsilon^2+\varepsilon^4 \end{bmatrix}, \quad \text{then} \quad R = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \varepsilon & \varepsilon \\ 0 & 0 & \varepsilon^2 \end{bmatrix}.$$

Some simple computations give

$$\chi_C(A) = O(1/\varepsilon^2), \qquad \|\widehat{W}_{\hat{R}}^{-1}\|_2 = O(1/\varepsilon^3).$$

REMARK 4.2. We now give a lower bound on $\chi_C(A)$. Since $\widehat{W}_{\hat{R}}$ (cf. (3.3) and (3.4)) has the form

$$\widehat{W}_{\hat{R}} = \begin{bmatrix} \times & 0 \\ \times & \hat{D}\hat{R}^T \end{bmatrix},$$

where

$$\hat{D} \equiv \mathrm{diag}(\sqrt{2}, \sqrt{2}, \ldots, \sqrt{2}, 2),$$

we have

$$\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1} = \begin{bmatrix} \times & 0 \\ \times & a_{nn}^{1/2} \hat{R}^{-T} \hat{D}^{-1} \end{bmatrix},$$

$$a_{nn}^{1/2} \|\hat{R}^{-T} \hat{D}^{-1}\|_2 \le \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2. \tag{4.15}$$

Hence we get the following lower bound on $\chi_C(A)$

$$\frac{n a_{nn}^{1/2} \|\hat{R}^{-1}\hat{D}^{-1}\|_2}{\|A\|_2^{1/2}} \le \chi_C(A). \tag{4.16}$$

This bound is tight for any n, since equality will hold by taking $R = \mathrm{diag}(r_{ii})$, with $0 < r_{ii} \le r_{nn}$, $i \ne n$. But it is a little complicated. In fact we can get a slightly weaker but simpler lower bound. Since

$$\|\hat{R}^{-1}\hat{D}^{-1}\|_2 \ge \frac{1}{2}\|\hat{R}^{-1}\|_2 = \frac{1}{2}\|H^{-1}\|^{1/2}, \tag{4.17}$$

we have from (4.16) that

$$\frac{1}{2}\sqrt{n}\,\frac{a_{nn}^{1/2}}{\|A\|_2^{1/2}}\,\|H^{-1}\|_2^{1/2} \le \chi_C(A).$$

But $\|H\|_2^{1/2} \le \sqrt{n}$ since H is positive definite with $h_{ii} = 1$, and $\|A\|_2^{1/2} \le \sqrt{n}\max_i a_{ii}^{1/2}$, hence we have

$$\frac{1}{2}\frac{a_{nn}^{1/2}}{\max_i a_{ii}^{1/2}}\,\kappa_2^{1/2}(H) \le \chi_C(A).$$

Numerical tests suggest that usually $\chi_C(A)$ is smaller or much smaller than $\kappa_C(A)$, the condition number for a general perturbation ΔA, defined by (3.3). This is not surprising since (4.2) provides more information about the perturbation in the data. We now relate these two condition numbers mathematically for the general case. For the case where standard pivoting is used in computing the Cholesky factorization, see Remark 4.8.

THEOREM 4.3.

$$\frac{1}{n}\chi_C(A) \le \kappa_C(A) \le \frac{\max_i a_{ii}}{\min_i a_{ii}}\chi_C(A). \tag{4.18}$$

Proof. Since $R = \hat{R}D_c$, it is easy to verify by the the structure of $\widehat{W}_R$ that

$$\widehat{W}_R = \widehat{\mathcal{D}}_c \widehat{W}_{\hat{R}} \mathcal{D}_c^{-1}, \tag{4.19}$$

where $\mathcal{D}_c$ is defined by (4.6) and

$$\widehat{\mathcal{D}}_c \equiv \operatorname{diag}(\sqrt{a_{11}a_{11}}, \underbrace{\sqrt{a_{11}a_{22}}, \sqrt{a_{22}a_{22}}}_{2}, \ldots, \underbrace{\sqrt{a_{11}a_{nn}}, \sqrt{a_{22}a_{nn}}, \ldots, \sqrt{a_{nn}a_{nn}}}_{n}).$$

Thus using (4.19) and $\max_i a_{ii} \leq \|A\|_2$, we have

$$\begin{aligned}
\chi_C(A) &\equiv n\|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2 / \|A\|_2^{1/2} = n\|\widehat{W}_R^{-1}\widehat{\mathcal{D}}_c\|_2/\|A\|_2^{1/2} \\
&\leq n\|\widehat{W}_R^{-1}\|_2 \, \|A\|_2^{1/2} \, \|\widehat{\mathcal{D}}_c\|_2/\|A\|_2 = n\kappa_C(A) \max_i a_{ii}/\|A\|_2 \\
&\leq n\kappa_C(A).
\end{aligned}$$

We now prove the second inequality. Using (4.19) and $\|A\|_2 \leq n \max_i a_{ii}$, we obtain

$$\begin{aligned}
\kappa_C(A) &\equiv \|\widehat{W}_R^{-1}\|_2 \, \|A\|_2^{1/2} = \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1} \hat{\mathcal{D}}_c^{-1}\|_2 \, \|A\|_2^{1/2} \\
&\leq \|A\|_2 \, \|\hat{\mathcal{D}}_c^{-1}\|_2 \, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2/\|A\|_2^{1/2} \leq \frac{\max_i a_{ii}}{\min_i a_{ii}} \chi_C(A).
\end{aligned}$$

The proof is complete. $\square$

REMARK 4.3. A reader might want to know why the first inequality in (4.18) is not $\chi_C(A) \leq \kappa_C(A)$. This can be easily explained. For a general ΔA, we have by Theorem 3.1 that $\|\Delta R\|_F/\|R\|_2 \lesssim \kappa_C(A)\epsilon$, where ϵ satisfies $\|\Delta A\|_F = \epsilon \, \|A\|_2$. For the backward rounding error ΔA, we have by Theorem 4.2 that $\|\Delta R\|_F/\|R\|_2 \lesssim \chi_C(A)\epsilon$, where ϵ satisfies $|\Delta A| \leq \epsilon dd^T$ (see (4.2)). It follows that $\|\Delta A\|_F \leq \epsilon \, \|dd^T\|_F = \epsilon \, \|R\|_F^2$, where $\|R\|_F$ satisfies attainable inequalities $\|A\|_2 \leq \|R\|_F^2 \leq n\|A\|_2$.

REMARK 4.4. The first inequality in (4.18) is attainable, since equality will hold by taking $A = cI$ with $c > 0$. The second inequality is at least nearly attainable. In fact by taking $A = R^T R$, with $R = \operatorname{diag}(\varepsilon^{n-1}, \varepsilon^{n-2}, \ldots, \varepsilon, 1) + e_1 e_n^T$ with small $\varepsilon > 0$, we easily obtain

$$\kappa_2(A) = O(\frac{1}{\varepsilon^{2n-2}}), \qquad \frac{\max_i a_{ii}}{\min_i a_{ii}} \chi_C(A) = \frac{2}{\varepsilon^{2n-2}} O(1) = O(\frac{1}{\varepsilon^{2n-2}}).$$

This example also suggests that possibly $\kappa_C(A)$ is much larger than $\chi_C(A)$ if the maximum element is much larger then the minimum element on the diagonal of A.

Like $\kappa_C(A)$, $\chi_C(A)$ is difficult to estimate directly. For a practical approach, see Theorem 4.6.

Usually a first-order bound is satisfactory, but sometimes more careful work is needed. Now we would like to present rigorous bounds for the Cholesky factorization with back rounding errors. We need the following lemma.

LEMMA 4.4. (*A trivial variant of Theorem* 2.11 *in Stewart and Sun* [8]) *Let* $\mathbf{T}$ *be a bounded linear operator on a Banach space* $\mathcal{B}$. *Assume that* $\mathbf{T}$ *has a bounded inverse, and set* $\delta = \|\mathbf{T}^{-1}\|^{-1}$. *Let* $\varphi : \mathcal{B} \to \mathcal{B}$ *be a function that satisfies*

$$\|\varphi(x)\| \leq \eta \|x\|^2$$

and

$$\|\varphi(x) - \varphi(y)\| \leq 2\eta \max\{\|x\|, \|y\|\}\|x - y\|$$

for some $\eta \geq 0$*. For any* $g \in \mathcal{B}$*, let* $\tau = \|\mathbf{T}^{-1}g\|$*. If* $\rho = 4\tau\eta/\delta < 1$*, then the equation*

$$\mathbf{T}x = g + \varphi(x)$$

has a unique solution x *that satisfies*

$$\|x\| \leq \frac{2\tau}{1 + \sqrt{1-\rho}} \leq 2\tau.$$

THEOREM 4.5. *Let* $A \in \mathbf{R}^{n\times n}$ *be a symmetric positive definite floating point matrix, with the exact Cholesky factorization* $A = R^T R$*. Define* $D_c \equiv \mathrm{diag}(A)^{1/2}$ *and* $\hat{R} \equiv RD_c^{-1}$*. If*

$$n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2\, \|D_c^{-1}\|_2^2 < 1/4, \tag{4.20}$$

where $\epsilon \equiv (n+1)u/(1-2(n+1)u)$ *with* u *being the unit round-off and* $\mathcal{D}_c$ *is defined by* (4.6)*, then the Cholesky factorization applied to* A *succeeds (barring underflow and overflow) and produces a nonsingular* $\tilde{R}$*, which satisfies*

$$\begin{aligned}\|\Delta R\|_2 &\leq \frac{2n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2}{1 + \sqrt{1 - 4n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2\, \|D_c^{-1}\|_2^2}} \\ &\leq 2n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2,\end{aligned} \tag{4.21}$$

where $\Delta R \equiv \tilde{R} - R$*. Obviously the weaker bound above can be rewritten in the following form:*

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq 2\chi_C(A)\epsilon. \tag{4.22}$$

Proof. Let $H \equiv D_c^{-1}AD_c^{-1}$ as before. Then $H = \hat{R}^T\hat{R}$. From (4.17) and (4.15), we get

$$\|H^{-1}\|_2 = \|\hat{R}^{-1}\|_2^2 \leq 4\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2/a_{nn} \leq 4\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2\, \|D_c^{-1}\|_2^2,$$

which with (4.20) implies that (4.1) holds. Thus by Lemma 4.1 the Cholesky factorization applied to A succeeds and the computed Cholesky factor $\tilde{R}$ satisfies

$$A + \Delta A = \tilde{R}^T\tilde{R}, \qquad |\Delta A| \leq \epsilon\, dd^T,$$

where $d_i = a_{ii}^{1/2}$. Combining this with $A = R^T R$ and $\Delta R = \tilde{R} - R$, we have

$$R^T\Delta R + \Delta R^T R = \Delta A - \Delta R^T \Delta R,$$

or

$$\hat{R}^T \Delta R D_c^{-1} + D_c^{-1}\Delta R^T \hat{R} = D_c^{-1}\Delta A D_c^{-1} - D_c^{-1}\Delta R^T \Delta R D_c^{-1}. \tag{4.23}$$

Define the linear operator $\mathbf{T}$ that maps the space of upper triangular matrices X into the space of symmetric matrices by $\mathbf{T}X \equiv \hat{R}^T X D_c^{-1} + D_c^{-1} X^T \hat{R}$, and define $\phi(X) = D_c^{-1} X^T X D_c^{-1}$. Then (4.23) can be rewritten as

$$\mathbf{T}X = D_c^{-1} \Delta A D_c^{-1} - \phi(X) \tag{4.24}$$

with $X = \Delta R$. Clearly the function $\phi(X)$ satisfies

$$\|\phi(X)\|_F \le \|D_c^{-1}\|_2^2 \, \|X\|_F^2,$$

and for an upper triangular matrices X and Y,

$$\|\phi(X) - \phi(Y)\|_F \le 2\|D_c^{-1}\|_2^2 \max\{\|X\|_F,\, \|Y\|_F\}\|X - Y\|_F,$$

so we take $\eta = \|D_c^{-1}\|_2^2$. From the proof of Theorem 4.2, we know for any symmetric $S \in \mathbf{R}^{n\times n}$, $\mathbf{T}X = S$ can be rewritten as $\widehat{W}_{\hat{R}} \mathcal{D}_c^{-1} \operatorname{uvec}(X) = \operatorname{duvec}(S)$. Using this we can easily verify

$$\|\mathbf{T}^{-1}\|_F \equiv \sup_{S:\text{symmetric}\neq 0} \frac{\|\mathbf{T}^{-1} S\|_F}{\|S\|_F} = \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2,$$

so we take $\delta = \|\mathbf{T}^{-1}\|_F^{-1} = \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^{-1}$. Notice that

$$\tau = \|\mathbf{T}^{-1}(D_c^{-1} \Delta A D_c^{-1})\|_F = \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1} \operatorname{duvec}(D_c^{-1} \Delta A D_c^{-1})\|_2 \le n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2,$$

then by the assumption (4.20) we have $\rho = 4\tau\eta/\delta < 1$. Thus, by Lemma 4.4, ΔR is the unique upper triangular solution of (4.24) and satisfies

$$\begin{aligned}\|\Delta R\|_F &\le \frac{2\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1} \operatorname{duvec}(D_c^{-1} \Delta A D_c^{-1})\|_2}{1 + \sqrt{1 - 4\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2 \, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1} \operatorname{duvec}(D_c^{-1} \Delta A D_c^{-1})\|_2 \, \|D_c^{-1}\|_2^2}} \\ &\le \frac{2n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2}{1 + \sqrt{1 - 4n\epsilon\, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2 \, \|D_c^{-1}\|_2^2}}.\end{aligned}$$

The proof is complete. $\square$

For a somewhat weaker but practical rigorous bound, see Theorem 4.9.

4.2. Matrix equation approach. We now derive practical perturbation bounds for errors of the form (4.2) using the matrix equation approach.

THEOREM 4.6. *With $A = R^T R$ and the same assumptions as in Lemma* 4.1, *for the perturbation ΔA and result $\tilde{R}$ in* (4.2) *we have*

$$|\Delta R| \le \epsilon \operatorname{up}(|\hat{R}^{-T}| e e^T |\hat{R}^{-1}|) |R| + O(\epsilon^2), \tag{4.25}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \chi_{C1}(A)\epsilon + O(\epsilon^2), \tag{4.26}$$

$$\chi_C(A) \le \chi_{C1}(A), \tag{4.27}$$

where $\Delta R \equiv \tilde{R} - R$, $\hat{R} \equiv R D_c^{-1}$, and

$$\chi_{C1}(A) \equiv \inf_{D \in \mathbf{D}_n} \chi_{C1}(A, D), \tag{4.28}$$

$$\chi_{C1}(A, D) \equiv n\|\hat{R}^{-1}\|_2 \, \|\hat{R}^{-1} D\|_2 \, \|D^{-1} R\|_2 / \|R\|_2. \tag{4.29}$$

Proof. Let $G \equiv \Delta A/\epsilon$. From (4.8) with $R = \hat{R}D_c$, it follows that

$$\dot{R}(0)R^{-1} + R^{-T}\dot{R}^T(0) = \hat{R}^{-T}D_c^{-1}GD_c^{-1}\hat{R}^{-1}.$$

Then we have with (2.1)

$$\dot{R}(0) = \mathrm{up}(\hat{R}^{-T}D_c^{-1}GD_c^{-1}\hat{R}^{-1})R, \tag{4.30}$$

which with $|G| \le dd^T$ gives

$$|\dot{R}(0)| \le \mathrm{up}(|\hat{R}^{-T}|ee^T|\hat{R}^{-1}|)|R|.$$

Then by the Taylor expansion (4.14), (4.25) follows immediately.

Also from (4.30) we have for any $D \in \mathbf{D}_n$ that

$$\dot{R}(0) = \mathrm{up}(\hat{R}^{-T}D_c^{-1}GD_c^{-1}\hat{R}^{-1}D)D^{-1}R.$$

Thus taking the Frobenius norm we have

$$\|\dot{R}(0)\|_F \le \|\hat{R}^{-1}\|_2\,\|D_c^{-1}GD_c^{-1}\|_F\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2, \tag{4.31}$$

which with $|G| \le dd^T$ gives

$$\|\dot{R}(0)\|_F \le n\|\hat{R}^{-1}\|_2\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2.$$

Since this holds for any $D \in \mathbf{D}_n$, (4.26) follows by using Taylor expansion (4.14).

It remains to prove (4.27). From (4.11) and (4.31) we have

$$\|\mathcal{D}_c\widehat{W}_R^{-1}\mathrm{duvec}(D_c^{-1}GD_c^{-1})\|_2 \le \|\hat{R}^{-1}\|_2\,\|D_c^{-1}GD_c^{-1}\|_F\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2. \tag{4.32}$$

Actually this holds for any symmetric $G \in \mathbf{R}^{n\times n}$ since it was essentially obtained from the matrix equation $R^TX + X^TR = G$ with X triangular by the two different approaches. Notice $\|\mathrm{duvec}(D_c^{-1}GD_c^{-1})\|_2 = \|D_c^{-1}GD_c^{-1}\|_F$, thus from (4.32) we must have

$$\|\mathcal{D}_c\widehat{W}_R^{-1}\|_2 \le \|\hat{R}^{-1}\|_2\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2,$$

which implies (4.27). $\square$

REMARK 4.5. Notice that $\|\mathrm{up}(X)\|_F \le \frac{\sqrt{2}}{2}\|X\|_F$ for any *symmetric* X, and $\|\hat{R}\|_2^2 = \|H\|_2$, then from (4.30) it is easy to obtain

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\sqrt{2}}{2}n\|H^{-1}\|_2\epsilon + O(\epsilon^2), \tag{4.33}$$

which is essentially the first-order bound of Drmač et al [3, Theorem 3.1], except that their bound is rigorous and only the 2-norm is used. But this bound can severely overestimate the true relative errors of the computed Cholesky factor. In fact

$$\chi_{C1}(A) \le \chi_{C1}(A, I) = n\|H^{-1}\|_2, \tag{4.34}$$

and $n\|H^{-1}\|_2$ can be arbitrarily larger than $\chi_{C1}(A)$. For example, if $A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon^2 \end{bmatrix}$ with small ε, then $R = \begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}$. Take $D = \mathrm{diag}(\sqrt{2}, \epsilon)$, then simple computations give

$$\chi_{C1}(A) \le \chi_{C1}(A, D) = O(1/\varepsilon), \qquad \|H^{-1}\|_2 = O(1/\varepsilon^2).$$

Thus the new approximation $\chi_{C1}(A)$ to the condition number $\chi_C(A)$ is a significant improvement on that of Drmač et al. Furthermore, it is easy to see $\|H^{-1}\|_2$ is invariant if pivoting is used in computing the Cholesky factorization of A, whereas $\chi_C(A)$ and $\chi_{C1}(A)$ are dependent on any pivoting. Thus the new bounds (4.4) and (4.26) more closely reflect the true sensitivity of the Cholesky factorization than (4.33). However, if the ill-conditioning of R is mostly due to bad scaling of its columns, then $n\|H^{-1}\|_2$ is small and is as good as $\chi_{C1}(A)$.

(4.34) implies $n\|H^{-1}\|_2$ is also an upper bound on $\chi_C(A)$. In fact we have the following stronger result

$$\chi_C(A) \le \frac{\sqrt{2}}{2} n\|H^{-1}\|_2, \tag{4.35}$$

which can be easily proved by using (4.11), (4.30) and the fact that for any *symmetric* X, $\|\mathrm{up}(X)\|_F \le \frac{\sqrt{2}}{2}\|X\|_F$. That is to say the first-order bound in (4.4) is smaller than that in (4.33). The example above suggest the former can be much smaller than the latter.

REMARK 4.6. The practical outcome of Theorem 4.6 is that $\chi_{C1}(A)$ is quite easy to estimate — all we need to do is choose a suitable D in $\chi_{C1}(A, D)$ in (4.29). We consider how to do this in the next section.

Numerical experiments suggest usually $\chi_{C1}(A)$ is a reasonable approximation of $\chi_C(A)$. But the following example shows $\chi_{C1}(A)$ can still be very larger than $\chi_C(A)$, even though it can be much smaller than $n\|H^{-1}\|_2$.

$$A = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & \varepsilon^2+\varepsilon^4 \end{bmatrix}, \qquad R = \begin{bmatrix} 1 & \varepsilon \\ 0 & \varepsilon^2 \end{bmatrix}.$$

It is easy to show

$$\chi_C(A) = O(1), \qquad \chi_{C1}(A) = O(1/\varepsilon), \qquad \|H^{-1}\|_2 = O(1/\varepsilon^2).$$

REMARK 4.7. If we use a monotone and consistent matrix norm $\|\cdot\|$, i.e., $|A| \le |B|$ implies $\|A\| \le \|B\|$ and $\|AB\| \le \|A\|\,\|B\|$, then from (4.30) it is straightforward to show we have the following perturbation bound instead of (4.26),

$$\frac{\|\Delta R\|}{\|R\|} \le \frac{\|ee^T\|\,\|\hat{R}^{-1}\|\,\|\,|\hat{R}^{-1}||R|\,\|}{\|R\|}\epsilon + O(\epsilon^2).$$

The advantage of this bound is that we do not need to choose the scaling matrix D.

In Section 4.1 we established a relationship (4.18) between $\chi_C(A)$ and $\kappa_C(A)$ in Theorem 3.1. Is there a similar relationship between $\chi_{C1}(A, D)$ (or $\chi_{C1}(A)$) and $\kappa_{C1}(A, D)$ (or $\kappa_{C1}(A)$) in Theorem 3.1? The answer is yes.

THEOREM 4.7.

$$\frac{1}{n}\chi_{C1}(A,D) \le \kappa_{C1}(A,D) \le \frac{\max_i a_{ii}}{\min_i a_{ii}}\chi_{C1}(A,D), \tag{4.36}$$

and from this,

$$\frac{1}{n}\chi_{C1}(A) \le \kappa_{C1}(A) \le \frac{\max_i a_{ii}}{\min_i a_{ii}}\,\chi_{C1}(A). \tag{4.37}$$

Proof.

$$\begin{aligned}
\frac{1}{n}\chi_{C1}(A,D) &\equiv \|\hat{R}^{-1}\|_2\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2/\|R\|_2 \\
&= \|D_cR^{-1}\|_2\,\|D_cR^{-1}D\|_2\,\|D^{-1}R\|_2/\|R\|_2 \\
&\le \|D_c\|_2^2\,\|R^{-1}\|_2\,\|R^{-1}D\|_2\,\|D^{-1}R\|_2/\|R\|_2 \\
&\le \|R\|_2\,\|R^{-1}\|_2\,\|R^{-1}D\|_2\,\|D^{-1}R\|_2 \quad (\text{using } \|D_c\| \le \|R\|_2) \\
&\equiv \kappa_{C1}(A,D).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\kappa_{C1}(A,D) &\equiv \|R\|_2\,\|R^{-1}\|_2\,\|R^{-1}D\|_2\,\|D^{-1}R\|_2 \\
&= \|R\|_2\,\|D_c^{-1}\hat{R}^{-1}\|_2\,\|D_c^{-1}\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2 \\
&\le \|R\|_2^2\,\|D_c^{-1}\|_2^2\,\|\hat{R}^{-1}\|_2\,\|\hat{R}^{-1}D\|_2\,\|D^{-1}R\|_2/\|R\|_2 \\
&= \frac{1}{n}\|A\|_2\,\|D_c^{-1}\|_2^2\,\chi_{C1}(A,D) \\
&\le \frac{\max_i a_{ii}}{\min_i a_{ii}}\,\chi_{C1}(A,D), \quad (\text{using } \|A\|_2 \le n\max_i a_{ii}).
\end{aligned}$$

The proof is complete. □

The standard pivoting strategy can usually improve $\chi_C(A)$, just as it can usually improve $\kappa_{C1}(A)$. In fact we have the following theorem.

THEOREM 4.8. *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $PAP^T = R^TR$ when the standard pivoting strategy is used. Then*

$$\chi_C(PAP^T) \le \chi_{C1}(PAP^T) \le \|H^{-1}\|_2^{1/2} n\sqrt{2n(n+1)(4^n+6n-1)}/6, \tag{4.38}$$

where $PAP^T \equiv D_cHD_c$ with $D_c \equiv \mathrm{diag}(PAP^T)^{1/2}$.

Proof. Since $R \equiv \hat{R}D_c$ and $\|D_c\|_2 \le \|R\|_2$, we have

$$\|\hat{R}^{-1}D\|_2 = \|D_cR^{-1}D\|_2 \le \|R\|_2\,\|R^{-1}D\|_2,$$

and so from (4.29) we obtain

$$\chi_{C1}(PAP^T) \le n \inf_{D\in\mathbf{D}_n} \|\hat{R}^{-1}\|_2\,\kappa_2(D^{-1}R). \tag{4.39}$$

Standard pivoting ensures $|r_{ii}| \ge |r_{ij}|$ for all $j \ge i$. If we take $D = \mathrm{diag}(R)$, then the elements of $\bar{R} \equiv D^{-1}R$ satisfy $1 = |\bar{r}_{ii}| \ge |\bar{r}_{ij}|$. It is easy to to show that

$$|(\bar{R}^{-1})_{ij}| \le 2^{j-i-1} \quad \text{for } j > i.$$

Thus

$$\|\bar{R}\|_2^2 \le \|\bar{R}\|_F^2 \le n(n+1)/2,$$
$$\|\bar{R}^{-1}\|_2^2 \le \|\bar{R}^{-1}\|_F^2 \le \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(2^{j-i-1})^2 + \sum_{i=1}^{n} 1 = (4^n + 6n - 1)/9.$$

Then from (4.39) and $\|\hat{R}^{-1}\|_2 = \|H^{-1}\|_2^{1/2}$, (4.38) follows immediately. □

REMARK 4.8. In Theorem 3.1 we have

$$\kappa_C(PAP^T) \le \kappa_{C1}(PAP^T) \le \kappa_2^{1/2}(A)\sqrt{2n(n+1)(4^n+6n-1)}/6.$$

By van de Sluis's result (2.6),

$$\kappa_2^{1/2}(H) \le \sqrt{n}\,\kappa_2^{1/2}(A),$$

where it is possible that

$$\kappa_2^{1/2}(H) \ll \kappa_2^{1/2}(A)$$

if A is badly scaled—the columns of R are badly scaled. But $1 \le \|H\|_2 < n$ since H is positive definite with $h_{ii} = 1$, hence

$$\|H^{-1}\|_2^{1/2} \le \sqrt{n}\,\kappa_2^{1/2}(A),$$

and it is possible that

$$\|H^{-1}\|_2^{1/2} \ll \kappa_2^{1/2}(A)$$

if R has badly scaled columns. Thus it is expected that $\chi_C(PAP^T)$ can be arbitrarily smaller than $\kappa_C(PAP^T)$.

REMARK 4.9. Suppose the Cholesky factorization of A be approached by using the standard symmetric pivoting strategy: $PAP^T = R^TR$. If the permutation matrix P is known beforehand and the Cholesky factorization is applied to PAP^T directly, it is easy to observe that Lemma 4.1 still holds except that now D_c and H should be redefined as

$$D_c \equiv \mathrm{diag}(PAP^T)^{1/2}, \qquad H \equiv D_c^{-1}PAP^TD_c^{-1},$$

and $A + \Delta A = \tilde{R}^T\tilde{R}$ in (4.2) should be replaced by

$$P(A+\Delta A)P^T = \tilde{R}^T\tilde{R}.$$

Then by Theorem 4.2 we have

$$\|\tilde{R} - R\|_F/\|R\|_2 \lesssim \chi_C(PAP^T)\epsilon.$$

(4.38) suggests that the computed Cholesky factor $\tilde{R}$ has high accuracy. However, usually the permutation matrix P for A is not known beforehand. The permutation matrix for $A + \Delta A$, which is produced in the computing process, is possibly different from that for A. Fortunately, Higham [4, Lemma 10.11] showed that if there are no ties in the pivoting strategy for $PAP^T = R^TR$, then for sufficiently small ΔA, the two permutation matrices are the same. Thus the Cholesky factorization with standard symmetric pivoting will most likely gives an $\tilde{R}$ which is about as accuracy as possible.

Finally we use Lemma 4.4 to derive a practical rigorous bound by the matrix equation approach.

THEOREM 4.9. *With the same assumption as in Theorem* 4.5, *except that requirement* (4.20) *is replaced by*

$$n\epsilon \,\|\hat{R}^{-1}\|_2 \,\|\hat{R}^{-1}D\|_2 \,\|D^{-1}\|_2 < 1/4, \tag{4.40}$$

where $D \in \mathbf{D}_n$, Cholesky factorization applied to A succeeds (*barring underflow and overflow*) *and produces a nonsingular $\tilde{R}$, which satisfies*

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{2n\epsilon \,\|\hat{R}^{-1}\|_2 \,\|\hat{R}^{-1}D\|_2 \,\|D^{-1}R\|_2/\|R\|_2}{1+\sqrt{1-4n\epsilon \,\|\hat{R}^{-1}\|_2 \,\|\hat{R}^{-1}D\|_2 \,\|D^{-1}\|_2}} \tag{4.41}$$

$$\leq 2\chi_{C1}(A, D)\epsilon, \tag{4.42}$$

where $\Delta R \equiv \tilde{R} - R$.

Proof. Let $H \equiv D_c^{-1}AD_c^{-1}$. Then $H = \hat{R}^T\hat{R}$. Since

$$\|H^{-1}\|_2 = \|\hat{R}^{-1}\|_2^2 \leq \|\hat{R}^{-1}\|_2 \,\|\hat{R}^{-1}D\|_2 \,\|D^{-1}\|_2,$$

which with (4.40) implies that (4.1) holds. Thus by Lemma 4.1 the Cholesky factorization applied to A succeeds and the computed Cholesky factor $\tilde{R}$ satisfies

$$A + \Delta A = \tilde{R}^T\tilde{R}, \qquad |\Delta A| \leq \epsilon\, dd^T,$$

where $d_i = a_{ii}^{1/2}$. Combining this with $A = R^TR$ and $\Delta R = \tilde{R} - R$, we obtain

$$R^T\Delta R + \Delta R^T R = \Delta A - \Delta R^T\Delta R,$$

or

$$(\Delta RR^{-1})^T + \Delta RR^{-1} = R^{-T}(\Delta A - \Delta R^T\Delta R)R^{-1},$$

which gives

$$\Delta RR^{-1} = \mathrm{up}[R^{-T}(\Delta A - \Delta R^T\Delta R)R^{-1}]. \tag{4.43}$$

Let $R \equiv D\bar{R}$. Then from (4.43) we have

$$\Delta R\bar{R}^{-1} = \mathrm{up}(R^{-T}\Delta A\bar{R}^{-1}) - \mathrm{up}(R^{-T}\Delta R^T\Delta R\bar{R}^{-1}).$$

Let $X = \Delta R\bar{R}^{-1}$, and define $\phi(X) = \mathrm{up}(D^{-1}X^TX)$, then the equation above can be rewritten as

$$X = \mathrm{up}(R^{-T}\Delta A\bar{R}^{-1}) - \phi(X). \tag{4.44}$$

It is easy to see that the function $\phi(X)$ satisfies

$$\|\phi(X)\|_F \leq \|D^{-1}\|_2 \,\|X\|_F^2,$$

and for any upper triangular matrices X and Y,

$$\|\phi(X)-\phi(X)\|_F \leq 2\|D^{-1}\|_2 \max\{\|X\|_F, \|Y\|_F\}\|X-Y\|_F,$$

so we take $\eta = \|D^{-1}\|_2$. Notice

$$\|\mathrm{up}(R^{-T}\Delta A\bar{R}^{-1})\|_F \leq \|\hat{R}^{-1}\|_2\, \|D_c^{-1}\Delta A D_c^{-1}\|_F\, \|\hat{R}^{-1}D\|_F \leq n\epsilon\, \|\hat{R}^{-1}\|_2\, \|\hat{R}^{-1}D\|_2,$$

then by the assumption (4.40) we have $\rho = 4\tau\eta/\delta < 1$, with $\tau = \|\mathrm{up}(R^{-T}\Delta A\bar{R}^{-1})\|_F$, and $\delta = 1$ since here the linear operator $\mathbf{T}$ is an identity operator. Thus, by Lemma 4.4, $\Delta R\bar{R}^{-1}$ is the unique upper triangular solution of (4.44), and satisfies

$$\begin{aligned}\|\Delta R\bar{R}^{-1}\|_F &\leq \frac{2\|\mathrm{up}(R^{-T}\Delta A\bar{R}^{-1})\|_F}{1+\sqrt{1-4\|\mathrm{up}(R^{-T}\Delta A\bar{R}^{-1})\|_F\, \|D^{-1}\|_2}} \\ &\leq \frac{2n\epsilon\, \|\hat{R}^{-1}\|_2\, \|\hat{R}^{-1}D\|_2}{1+\sqrt{1-4n\epsilon\, \|\hat{R}^{-1}\|_2\, \|\hat{R}^{-1}D\|_2\, \|D^{-1}\|_2}}.\end{aligned}$$

Then the bound (4.41) follows using $\|\Delta R\|_F \leq \|\Delta R\bar{R}^{-1}\|_F\, \|\bar{R}\|_2$. □

REMARK 4.10. If we take $D = I$, it is to show by using the fact that $\|\mathrm{up}(X)\|_F \leq \frac{1}{\sqrt{2}}\|X\|_F$ for any symmetric $X \in \mathbf{R}^{n\times n}$ that the assumption (4.40) can be weakened to

$$n\|H^{-1}\|_2\, \epsilon \leq 1/2, \tag{4.45}$$

and we have the following bound:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}n\|H^{-1}\|_2\epsilon}{1+\sqrt{1-2n\|H^{-1}\|_2\epsilon}} \leq \sqrt{2}n\|H^{-1}\|_2\epsilon, \tag{4.46}$$

which is a slightly stronger than (4.41) where $D = I$. An equivalent rigorous bound where only the 2-norm is used was obtained by Drmač et al [3]

As we pointed out in Remark 4.5, with a correct choice of D, e.g., $D = D_r \equiv \mathrm{diag}(\|(R(i,:)\|_2)$, possibly $\chi_{C1}(A, D)$ can be much smaller than $n\|H^{-1}\|_2$. So the bound (4.46) is potentially weak, although the condition (4.45) is not as constraining as (4.40).

5. Numerical experiments and condition estimator. In Section 4 we presented new first-order and rigorous perturbation bounds for the change caused by backward rounding errors for the Cholesky factorization using two different approaches, defined $\chi_C(A) \equiv n\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2/\|A\|_2^{1/2}$ as the condition number, and suggested that usually $\chi_C(A)$ could be estimated in practice by

$$\chi_{C1}(A, D) \equiv n\|\hat{R}^{-1}\|_2\, \|\hat{R}^{-1}D\|_2\, \|D^{-1}R\|_2/\|R\|_2.$$

Our bounds are potentially much smaller than the equivalent bound in [3]. Also we compare $\chi_C(A)$ with $\kappa_C(A)$, the condition number for general perturbation ΔA, and compare the corresponding estimates $\chi_{C1}(A, D)$ with $\kappa_{C1}(A, D)$ as well.

We would like to choose D such that $\chi_{C1}(A, D)$ is a good approximation to the infimum $\chi_{C1}(A)$ in (4.28). By van der Sluis's result Theorem 2.1, $\|\hat{R}^{-1}D\|_2\, \|D^{-1}R\|_2$ will be nearly minimum when the rows of $D^{-1}R$ are equilibrated. Then a procedure for obtaining an $O(n^2)$ condition estimator for the Cholesky factorization with backward rounding errors is to choose $D = D_r \equiv \mathrm{diag}(\|R(i,:)\|_2)$, then use a standard norm estimator to estimate all factors in $\chi_{C1}(A, D)$.

Now we give a set of examples to show our findings. The matrices are $n \times n$ Pascal matrices, (with elements $a_{1j} = a_{i1} = 1$, $a_{ij} = a_{i,j-1} + a_{i-1,j}$), $n = 1, 2, \ldots, n$. The results are shown in Table 5.1 without pivoting and in Table 5.2 with pivoting, where D is chosen by the method presented in the last paragraph.

Table 5.1
Results for Pascal matrices without pivoting, $A = R^T R$

n	$\chi_C(A)$	$\chi_{C1}(A, D)$	$\kappa_C(A)$	$\kappa_{C1}(A, D)$	$\frac{\sqrt{2}}{2} n\|H^{-1}\|_2$
1	5.0e–01	1.0e+00	5.0e–01	1.0e+00	7.1e–01
2	2.2e+00	6.0e+00	2.1e+00	6.3e+00	4.8e+00
3	8.9e+00	3.6e+01	9.7e+00	5.0e+01	3.6e+01
4	3.9e+01	2.3e+02	5.5e+01	4.8e+02	3.0e+02
5	1.7e+02	1.4e+03	3.5e+02	4.9e+03	2.5e+03
6	7.8e+02	9.2e+03	2.5e+03	5.2e+04	2.2e+04
7	3.5e+03	5.8e+04	1.9e+04	5.7e+05	2.0e+05
8	1.6e+04	3.7e+05	1.5e+05	6.3e+06	1.7e+06
9	7.7e+04	2.3e+06	1.3e+06	7.0e+07	1.5e+07
10	3.6e+05	1.5e+07	1.1e+07	7.9e+08	1.4e+08
11	1.7e+06	9.1e+07	9.8e+07	9.0e+09	1.2e+09
12	8.4e+06	5.6e+08	8.7e+08	1.0e+11	1.1e+10
13	4.1e+07	3.5e+09	7.8e+09	1.2e+12	9.8e+10
14	2.0e+08	2.2e+10	7.1e+10	1.4e+13	8.8e+11
15	9.7e+08	1.3e+11	6.5e+11	1.6e+14	7.9e+12

Table 5.2
Results for Pascal matrices with pivoting, $\tilde{A} \equiv PAP^T = R^T R$

n	$\chi_C(\tilde{A})$	$\chi_{C1}(\tilde{A}, D)$	$\kappa_C(\tilde{A})$	$\kappa_{C1}(\tilde{A}, D)$	$\frac{\sqrt{2}}{2} n\|H^{-1}\|_2$
1	5.0e–01	1.0e+00	5.0e–01	1.0e+00	7.1e–01
2	1.6e+00	4.9e+00	1.5e+00	4.2e+00	4.8e+00
3	4.1e+00	1.8e+01	5.1e+00	1.6e+01	3.6e+01
4	1.3e+01	6.1e+01	2.2e+01	8.0e+01	3.0e+02
5	3.6e+01	2.1e+02	8.3e+01	3.3e+02	2.5e+03
6	7.7e+01	6.7e+02	2.5e+02	1.3e+03	2.2e+04
7	1.8e+02	2.2e+03	9.4e+02	5.1e+03	2.0e+05
8	4.8e+02	7.8e+03	4.0e+03	2.4e+04	1.7e+06
9	1.2e+03	2.7e+04	1.6e+04	1.0e+05	1.5e+07
10	3.6e+03	9.0e+04	7.6e+04	4.7e+05	1.4e+08
11	7.5e+03	2.7e+05	2.4e+05	1.8e+06	1.2e+09
12	1.8e+04	9.2e+05	8.3e+05	8.2e+06	1.1e+10
13	3.9e+04	2.9e+06	3.2e+06	3.1e+07	9.8e+10
14	9.4e+04	8.5e+06	1.3e+07	1.2e+08	8.8e+11
15	2.2e+05	2.8e+07	5.4e+07	4.9e+08	7.9e+12

Note in Table 5.1 and Table 5.2 how $\frac{\sqrt{2}}{2} n\|H^{-1}\|_2$ can be worse than $\chi_C(A)$. In

Table 5.2 pivoting is seen to give a significant improvement on $\chi_C(A)$. Also we observe from both Table 5.1 and Table 5.2 that $\chi_{C1}(A)$ is a reasonable approximation of $\chi_C(A)$. We see $\chi_C(A)$ is smaller than $\kappa_C(A)$ for $n > 2$.

Acknowledgement. I would like to extend my sincere thanks to my supervisor, Professor Chris Paige, for his invaluable supervision. This paper has been considerably improved by his comments, suggestions and corrections.

REFERENCES

[1] X.-W. Chang, C. C. Paige and G. W. Stewart. New perturbation analyses for the Cholesky factorization. *IMA J. Numer. Anal.*, to appear.

[2] J. W. Demmel. On floating point errors in Cholesky. Technical Report CS-89-87, Department of Computer Science, University of Tennessee, October 1989. 6pp. LAPACK Working Note 14.

[3] Z. Drmač, M. Omladič and K. Veselić. On the perturbation of the Cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 15(4):1319–1332, 1994.

[4] N. J. Higham. Accuracy and Stability of Numerical Algorithms. *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA, 1996.

[5] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15(4):727–764, 1973.

[6] G. W. Stewart. Perturbation bounds for the QR factorization of a matrix. *SIAM J. Numer. Anal.*, 14(3):509–518, 1977.

[7] G. W. Stewart. On the perturbation of LU, Cholesky, and QR factorizations. *SIAM J. Matrix Anal. Appl.*, 4(4):1141–1145, 1993.

[8] G. W. Stewart and J. G. Sun. *Matrix perturbation theory*, Academic Press, London, 1990.

[9] J.-G. Sun. Perturbation bounds for the Cholesky and QR factorization. *BIT*, 31:341–352, 1991.

[10] J.-G. Sun. Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations. *Linear Algebra and Appl.*,173:77–97, 1992.

[11] J.-G. Sun. Componentwise perturbation bounds for some matrix decompositions. *BIT*, 32:702–714, 1992.

[12] A. van der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14:14–23, 1969.