IMA Journal of Numerical Analysis (1995) **14**, 1–28

New perturbation analyses for the Cholesky factorization

XIAO-WEN CHANG†, CHRISTOPHER C. PAIGE†
School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2A7
AND
G. W. STEWART‡
Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A.

[Received 13 June 1995 and in revised form 7 November 1995]

We present new perturbation analyses for the Cholesky factorization $A = R^TR$ of a symmetric positive definite matrix A. The analyses more accurately reflect the sensitivity of the problem than previous normwise results. The condition numbers here are altered by any symmetric pivoting used in $PAP^T = R^TR$, and both numerical results and an analysis show that the standard method of pivoting is optimal in that it usually leads to a condition number very close to its lower limit for any given A. It follows that the computed R will probably have greatest accuracy when we use the standard symmetric pivoting strategy.

Initially we give a thorough analysis to obtain both first order and strict normwise perturbation bounds which are as tight as possible, leading to a definition of an optimal condition number for the problem. Then we use this approach to obtain reasonably clear first order and strict componentwise perturbation bounds.

We complete the work by giving a much simpler normwise analysis which provides a somewhat weaker bound, but which allows us to estimate the condition of the problem quite well with an efficient computation. This simpler analysis also shows why the factorization is often less sensitive than we previously thought, and adds further insight into why pivoting usually gives such good results. We derive a useful upper bound on the condition of the problem when we use pivoting.

1. Introduction

The Cholesky factorization is a fundamental tool in matrix computations: given an $n \times n$ real symmetric positive definite matrix A, there exists a unique upper triangular matrix R with positive diagonal entries such that

$$A = R^TR.$$

R is called the Cholesky factor.

There have been several papers dealing with the perturbation analysis for the Cholesky factor. The first norm-based result was that of Stewart (1977). It was

† This research was supported by Natural Sciences and Engineering Research Council of Canada Grant OGP0009236

‡ This research was supported in part by the US National Science Foundation under grant CCR 95503126.

further modified and improved by Sun (1991), who included a *first order* perturbation result. Using a different approach, Stewart (1993) obtained the same first order perturbation result. Componentwise perturbation analyses have been given by Sun (1992a, 1992b).

The main purpose of this paper is to establish new first order bounds on the norm of the perturbation in the Cholesky factor, sharper than that of Sun (1991) and Stewart (1993). Also, we obtain a new first order bound for the components of the perturbation, and give *strict* bounds on the norm and components of the perturbation.

In the remainder of this section we review some useful tools and results by showing one way of obtaining the first order normwise perturbation bound given by Sun (1991) and Stewart (1993) for the Cholesky factor.

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

$$\kappa_2(A)\epsilon < 1, \tag{1}$$

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

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

where

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

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$, the theorem is trivial). If (1) holds, $A+tG$ is symmetric positive definite for all $t \in [0, \epsilon]$, and $A + tG$ has the Cholesky factorization

$$A + tG = R^T(t)R(t), \quad 0 \leq t \leq \epsilon. \tag{3}$$

In particular, $R(0) = R$ and $R(\epsilon) = R + \Delta R$.

¿From the algorithm for the Cholesky factorization, it is easy to verify that $R(t)$ is a twice continuously differentiable function of $t \in [0, \epsilon]$. Differentiating (3) and setting $t = 0$ in the result gives

$$G = R^T\dot{R}(0) + \dot{R}^T(0)R, \tag{4}$$

$$R^{-T}GR^{-1} = \dot{R}(0)R^{-1} + (\dot{R}(0)R^{-1})^T = \text{Upper Triangular} + \text{Lower Triangular}. \tag{5}$$

If we follow Stewart (1993, 1995b) and define the upper triangular matrix

$$\text{up}(X) = \begin{pmatrix} \frac{1}{2}x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}x_{nn} \end{pmatrix}$$

for *any* $n \times n$ matrix X, then clearly

$$\dot{R}(0)R^{-1} = \text{up}(R^{-T}GR^{-1}). \tag{6}$$

But for any $n \times n$ *symmetric* matrix X,

$$2\|\mathrm{up}(X)\|_F^2 = \|X\|_F^2 - \frac{1}{2}(x_{11}^2 + x_{22}^2 + \cdots + x_{nn}^2) \le \|X\|_F^2. \tag{7}$$

Thus from (6) we have

$$\sqrt{2}\|R\|_2^{-1}\|\dot{R}(0)\|_F \le \sqrt{2}\|\dot{R}(0)R^{-1}\|_F \le \|R^{-T}GR^{-1}\|_F \le \|R^{-1}\|_2^2\|G\|_F, \tag{8}$$

which, together with $\|G\|_F = \|A\|_2 = \|R\|_2^2$, gives

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \frac{1}{\sqrt{2}}\kappa_2(A). \tag{9}$$

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

$$R + \Delta R = R(\epsilon) = R(0) + \epsilon\dot{R}(0) + O(\epsilon^2), \tag{10}$$

so that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\|\dot{R}(0)\|_F}{\|R\|_2}\epsilon + O(\epsilon^2),$$

which, combined with (9), gives (2). □

¿From the proof, it can be seen that the key point in this derivation of a first order perturbation bound is the use of (4) to give a good bound on the sensitivity $\|\dot{R}(0)\|_F/\|R\|_2$. We use the same idea in obtaining some other bounds.

The rest of the paper is organized as follows. In Section 2 we derive tight new first order and strict bounds on the norm of the perturbation, and use this approach to give first order and strict componentwise perturbation bounds. In Section 3 we give numerical results for these. Section 4 gives a quick summary of the key norm results of Section 2, followed by the derivation of somewhat weaker but more insightful and computationally applicable bounds than those in Section 2, together with numerical comparisons. In Section 5 we summarize our results and derive a useful upper bound on the condition of the problem when we use pivoting.

Since the derivations in Section 2 require a lot of detail, a reader who is only interested in the more immediately practical applications could start by reading Section 1, Section 3.1, Section 4 and Section 5.

2. New perturbation bounds

In the first part of this section we derive and analyse an important new first order perturbation bound for how R varies when A is altered. In the second part we derive new strict perturbation bounds.

2.1 *First order perturbation bound*

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

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

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

In order to estimate $\dot{R}(0)$ using (4), as in Stewart (1977) we define the linear operator $\mathbf{T}_R$ that maps the space of upper triangular matrices X into the space of symmetric matrices ($\mathbf{T}_R$ is in fact 1 to 1 and onto, as we shall see) by

$$\mathbf{T}_R X = R^T X + X^T R,$$

where $R \in \mathcal{R}^{n\times n}$ is nonsingular upper triangular. Let

$$\mathbf{T}_R X = R^T X + X^T R = G, \tag{11}$$

where G is an $n \times n$ symmetric matrix. The upper and lower triangular parts of (11) contain identical information, and the key to this analysis is writing the upper triangular part of (11) as

$$W_R \mathrm{uvec}(X) = \mathrm{uvec}(G), \tag{12}$$

where from Appendix A $W_R \in \mathcal{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is the lower triangular matrix

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

In retrospect this seems to be a sensible approach. The reason it was not used before is probably that it appeared it would lead to a rather complicated and obscure analysis, as we had very little understanding of W_R. It turns out that by using the correct tools, the relevant analysis of W_R and its inverse is quite tractable, and the results are extremely rewarding. To help the reader who wants to go further into this, we study W_R in more detail in Appendix A. However a first reading should be possible without recourse to the appendices.

Since R is nonsingular, W_R is also, and from (12),

$$\mathrm{uvec}(X) = W_R^{-1}\mathrm{uvec}(G), \tag{13}$$

so that for any symmetric matrix G, $\mathbf{T}_R X = G$ has a unique (upper triangular) solution, and the operator $\mathbf{T}_R$ is nonsingular.

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

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

where

$$D = \mathrm{diag}(1, \underbrace{\sqrt{2}, 1}_{2}, ..., \underbrace{\sqrt{2}, \sqrt{2}, ..., \sqrt{2}, 1}_{j}, ..., \underbrace{\sqrt{2}, \sqrt{2}, ..., \sqrt{2}, 1}_{n}) \in \mathcal{R}^{\frac{n(n+1)}{2}}. \tag{14}$$

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

Defining $\hat{W}_R \equiv DW_R$ for our norm-based analysis, we rewrite (12) as

$$\hat{W}_R\mathrm{uvec}(X) = \mathrm{duvec}(G). \tag{15}$$

Comparing (4) with (11) and its equivalent (12) or (15), we see that W_R or $\hat{W}_R$ is crucial to our analysis. Thus for our norm-based perturbation analysis we want to know something about $\|\hat{W}_R^{-1}\|_2$. We first define the operator norm subordinate to the matrix F-norm

$$\begin{aligned}
\|\mathbf{T}_R^{-1}\|_F &\equiv \sup_{G:\mathrm{symmetric}\neq 0} \frac{\|\mathbf{T}_R^{-1}G\|_F}{\|G\|_F} \\
&= \sup_{X:\mathrm{upper\ triangular}\neq 0} \frac{\|X\|_F}{\|R^TX + X^TR\|_F} \\
&= \sup_{X:\mathrm{upper\ triangular}\neq 0} \frac{\|\mathrm{uvec}(X)\|_2}{\|\mathrm{duvec}(R^TX + X^TR)\|_2} \\
&= \sup_{G:\mathrm{symmetric}\neq 0} \frac{\|\hat{W}_R^{-1}\mathrm{duvec}(G)\|_2}{\|\mathrm{duvec}(G)\|_2} \\
&= \sup_{g\neq 0} \frac{\|\hat{W}_R^{-1}g\|_2}{\|g\|_2} = \|\hat{W}_R^{-1}\|_2.
\end{aligned} \tag{16}$$

For any symmetric matrix G, $\mathbf{T}_R X = R^TX + X^TR = G$ gives $XR^{-1} + R^{-T}X^T = R^{-T}GR^{-1}$, so that $XR^{-1} = \mathrm{up}(R^{-T}GR^{-1})$ and $\sqrt{2}\|XR^{-1}\|_F \leq \|R^{-1}\|_2^2\|G\|_F$, see (6) and (7), giving

$$\|\mathbf{T}_R^{-1}G\|_F = \|X\|_F = \|XR^{-1}R\|_F \leq \|XR^{-1}\|_F\|R\|_2 \leq \frac{1}{\sqrt{2}}\|R\|_2\|R^{-1}\|_2^2\|G\|_F.$$

Combining this with (16), we get

$$\|\hat{W}_R^{-1}\|_2 = \sup_{G:\text{symmetric}\neq 0} \frac{\|\mathbf{T}_R^{-1}G\|_F}{\|G\|_F} \leq \frac{1}{\sqrt{2}}\|R\|_2\|R^{-1}\|_2^2, \tag{17}$$

so that

$$\|\hat{W}_R^{-1}\|_2\|R\|_2 \leq \frac{1}{\sqrt{2}}\|R\|_2^2\|R^{-1}\|_2^2 = \frac{1}{\sqrt{2}}\kappa_2^2(R). \tag{18}$$

The upper bound is achieved if R is an $n \times n$ identity matrix with $n \geq 2$, and so is tight. To motivate this bound we point out that $\|\hat{W}_R^{-1}\|_2\|R\|_2$ will later be shown to be our measure of the sensitivity of the Cholesky factor to changes in A.

Now we obtain a lower bound on $\|\hat{W}_R^{-1}\|_2\|R\|_2$. Observe that the $n \times n$ bottom right hand corner of W_R is just $\mathrm{diag}(1, 1,..., 1, 2)R^T$, so that $\hat{W}_R$ has the form

$$\hat{W}_R = \begin{pmatrix} \times & 0 \\ \times & \hat{R}^T \end{pmatrix},$$

where

$$\hat{R}^T = \mathrm{diag}(\sqrt{2}, \sqrt{2},..., \sqrt{2}, 2)R^T. \tag{19}$$

Therefore we have

$$\hat{W}_R^{-1} = \begin{pmatrix} \times & 0 \\ \times & \hat{R}^{-T} \end{pmatrix}.$$

It follows that

$$\|\hat{R}^{-T}\|_2 \leq \|\hat{W}_R^{-1}\|_2. \tag{20}$$

Note from (19) that $\|R^{-1}\|_2 \leq 2\|\hat{R}^{-1}\|_2$. Combining this with (20), we get

$$\|\hat{W}_R^{-1}\|_2 \geq \frac{1}{2}\|R^{-1}\|_2, \tag{21}$$

so that

$$\|\hat{W}_R^{-1}\|_2\|R\|_2 \geq \frac{1}{2}\|R^{-1}\|_2\|R\|_2 = \frac{1}{2}\kappa_2(R). \tag{22}$$

This bound is tight for any n, since equality will hold by taking $R = \mathrm{diag}(r_{ii})$, with $0 < \sqrt{2}r_{nn} \leq r_{ii}$, $i \neq n$.

This analysis leads to our first order, tight normwise and componentwise perturbation bound results.

Theorem 2 Let A, ΔA, R and ϵ be as in Theorem 1. Then $A + \Delta A$ has the Cholesky factorization

$$A + \Delta A = (R + \Delta R)^T(R + \Delta R), \tag{23}$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_C(A)\epsilon + O(\epsilon^2), \tag{24}$$

$$\frac{1}{2}\kappa_2^{1/2}(A) \le \kappa_C(A) \le \frac{1}{\sqrt{2}}\kappa_2(A), \tag{25}$$

with

$$\kappa_C(A) \equiv \|\hat{W}_R^{-1}\|_2 \|A\|_2^{1/2}. \tag{26}$$

Proof. By the same argument as in the proof of Theorem 1, $A + \Delta A$ has the Cholesky factorization (23). ¿From (4) and the expression for G we have

$$\mathrm{uvec}(\dot{R}(0)) = \hat{W}_R^{-1}\mathrm{duvec}(G) = \hat{W}_R^{-1}\mathrm{duvec}(\Delta A/\epsilon), \tag{27}$$

so taking the 2-norm gives

$$\|\dot{R}(0)\|_F \le \|\hat{W}_R^{-1}\|_2 \|\Delta A/\epsilon\|_F = \|\hat{W}_R^{-1}\|_2 \|A\|_2.$$

Combining this with $\|A\|_2 = \|R\|_2^2$, $\|A^{-1}\|_2 = \|R^{-1}\|_2^2$, (18) and (22), we can obtain (25) and

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \kappa_C(A). \tag{28}$$

Thus, from the Taylor series (10) for $R(t)$, (24) follows. □

REMARK 3 ¿From (25) we know the bound (24) is at least as sharp as (2), but it suggests it may be considerably sharper. By how much can the two bounds differ? The answer is by an arbitrary factor. Consider the following example.

Let $A = \mathrm{diag}\,(1, \varepsilon^2)$. Then $R = \mathrm{diag}\,(1, \varepsilon)$, $\hat{W}_R = \mathrm{diag}\,(2, \sqrt{2}, 2\varepsilon)$. When $0 < \varepsilon \le \frac{1}{\sqrt{2}}$, we obtain

$$\kappa_C(A) = \frac{1}{2\varepsilon}, \quad \kappa_2(A) = \frac{1}{\varepsilon^2}.$$

Here $\kappa_C(A)$ is at the lower bound in (25), and we see that the first order perturbation bound (2) can severely overestimate the effect of a perturbation in A.

But it is possible that $\kappa_C(A)$ has the same order as $\kappa_2(A)$. For example, if $A = \begin{pmatrix} \varepsilon^2 & \varepsilon \\ \varepsilon & 2 \end{pmatrix}$ with small ε, then $R = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 \end{pmatrix}$. Some simple computations give

$$\kappa_C(A) = O(\frac{1}{\varepsilon^2}), \quad \kappa_2(A) = O(\frac{1}{\varepsilon^2}).$$

Actually $\kappa_C(A)$ is the optimal measure of the sensitivity of the Cholesky factorization in the following sense. Suppose we are able to obtain a bound of the form

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le f(\|A\|_{2,F}, \|A^{-1}\|_{2,F}), \tag{29}$$

where f, a function of $\|A\|_2$ or $\|A\|_F$, and $\|A^{-1}\|_2$ or $\|A^{-1}\|_F$, is some other measure of the sensitivity of the Cholesky factorization than $\kappa_C(A)$ in (28). Solving (4) we have $\dot{R}(0) = \mathbf{T}_R^{-1}G$, so that

$$\frac{\|\mathbf{T}_R^{-1}G\|_F}{\|R\|_2} \le f(\|A\|_{2,F}, \|A^{-1}\|_{2,F}),$$

which, together with $\|G\|_F = \|A\|_2 = \|R\|_2^2$, gives

$$\frac{\|\mathbf{T}_R^{-1}G\|_F}{\|G\|_F} \le \frac{f(\|A\|_{2,F}, \|A^{-1}\|_{2,F})}{\|A\|_2^{1/2}}.$$

As a consequence, we have for our operator norm

$$\|\mathbf{T}_R^{-1}\|_F \le \frac{f(\|A\|_{2,F}, \|A^{-1}\|_{2,F})}{\|A\|_2^{1/2}},$$

so from (16),

$$\kappa_C(A) = \|\hat{W}_R^{-1}\|_2 \|A\|_2^{1/2} \le f(\|A\|_{2,F}, \|A^{-1}\|_{2,F}).$$

By the above analysis, we can consider $\kappa_C(A)$ to be the condition number for the Cholesky factorization.

REMARK 4 Suppose the Cholesky factorization of A is approached by using the standard symmetric pivoting strategy: $PAP^T = R^T R$, where P is an $n \times n$ permutation matrix designed so that rows and columns of A are interchanged, during the computation of the reduction, to make the leading diagonal elements of R as large as possible. Let the Cholesky factorization of $P(A+\Delta A)P^T$ be $P(A+\Delta A)P^T = (R+\Delta R)^T(R+\Delta R)$. Then by Theorem 2 we have

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

and

$$\frac{1}{2}\kappa_2^{1/2}(A) \le \kappa_C(PAP^T) \le \frac{1}{\sqrt{2}}\kappa_2(A).$$

Note that the first order bound (2) does not change when the Cholesky factorization of A is approached by using any pivoting strategy. Clearly the perturbation bound (24) more closely reflects the structure of the problem. Many numerical experiments with the standard pivoting strategy suggest that $\kappa_C(PAP^T)$ usually has the same order as $\kappa_2^{1/2}(A)$, and in fact $\kappa_C(PAP^T)$ can be bounded above by $\kappa_2^{1/2}(A)2^{n-1}\sqrt{n(n+1)/2}$, see (53). Using the standard symmetric pivoting strategy in the previous example gives $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $R = \begin{pmatrix} \sqrt{2} & \frac{1}{\sqrt{2}}\varepsilon \\ 0 & \frac{1}{\sqrt{2}}\varepsilon \end{pmatrix}$, and $\kappa_C(PAP^T) = O(\frac{1}{\varepsilon})$, showing how pivoting can improve the condition of the problem (as measured by our condition number) by an arbitrary amount.

REMARK 5 Cholesky factorization gives a computed $\tilde{R}$ for which $\tilde{R}^T\tilde{R} = P(A+\Delta A)P^T$, with the bound on $\|\Delta A\|_F$ independent of P (see for example Wilkinson & Reinsch (1971, p. 26)). ¿From our perturbation analysis we can see that $\|\tilde{R} - R\|_F/\|R\|_2 \lesssim \kappa_C(PAP^T)\|\Delta A\|_F/\|A\|_2$, and if as usually happens, the standard symmetric pivoting does give $\kappa_C(PAP^T) \approx \kappa_2^{1/2}(A)$, which is near optimal, then this pivoting will also most likely give an $\tilde{R}$ that is as accurate as possible. Meinguet

(1983) gave a careful rounding error analysis of the Cholesky factorization, and emphasized how well-behaved it is. The results here add even more evidence of its "good-hearted" nature.

REMARK 6 There have been several techniques for estimating the 2-norm of the inverse of a triangular matrix, e.g., Cline et al. (1979), Cline et al (1982) and Dixon (1983). A comprehensive, comparative survey on these techniques has been given by Higham (1987). Using these techniques, $\|\hat{W}_R^{-1}\|_2$ could be estimated at the cost of solving a few linear systems with matrices $\hat{W}_R$ and $\hat{W}_R^T$. To solve the former is equivalent to solving $R^TX + X^TR = G$ (G is symmetric) for upper triangular X, and the cost is $O(n^3)$; see Appendix A. Even though $\hat{W}_R^T y = c$ is not the transpose of the above matrix equation, it can also be solved in $O(n^3)$ by using the special structure of $\hat{W}_R^T$. However since the Cholesky factorization costs $O(n^3)$, such a computation would rarely be considered feasible. Of course if it is known that $\kappa_C(PAP^T) \approx \kappa_2^{1/2}(A)$, as usually happens when we use the standard symmetric pivoting, then we need only estimate $\kappa_2^{1/2}(A) = \kappa_2(R)$ for this case, and this can be done in $O(n^2)$. For a practical approach to the general case, see Section 4.

Similar ideas apply to bounding the individual components of ΔR, as we now show. Let $\|\cdot\|$ denote a family of monotone, consistent matrix norms; i.e., $|A| \leq |B|$ implies $\|A\| \leq \|B\|$, and $\|AB\| \leq \|A\|\|B\|$ (see for example Stewart & Sun (1990, p. 52)).

THEOREM 7 Let $A \in \mathcal{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^TR$. Let $\Delta A \in \mathcal{R}^{n\times n}$ be a symmetric matrix satisfying $|\Delta A| \leq \epsilon S$ for some nonnegative matrix S with nonnegative ϵ. If

$$\epsilon \| |A^{-1}|S\| < 1, \tag{30}$$

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

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

where

$$\mathrm{uvec}(|\Delta R|)| \leq |W_R^{-1}|\mathrm{uvec}(S)\epsilon + O(\epsilon^2). \tag{31}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). From (30), we have

$$\|A^{-1}\Delta A\| \leq \| |A^{-1}\Delta A| \| \leq \| |A^{-1}||\Delta A| \| \leq \epsilon \| |A^{-1}|S\| < 1,$$

so that all the eigenvalues of $A^{-1}\Delta A$ satisfy

$$|\lambda(A^{-1}\Delta A)| < 1.$$

Therefore $A + \Delta A$ is positive definite (see Stewart & Sun (1990, p.27)). It is easy to observe that the argument before (4) in the proof of Theorem 1 is applicable here. Thus $A + \Delta A$ has the Cholesky factorization, and from (4), (11) and (12),

$$\mathrm{uvec}(\dot{R}(0)) = W_R^{-1}\mathrm{uvec}(G) = W_R^{-1}\mathrm{uvec}(\Delta A/\epsilon), \tag{32}$$

so taking the absolute value gives

$$\mathrm{uvec}(|\dot{R}(0)|) \leq |W_R^{-1}|\mathrm{uvec}(|\Delta A/\epsilon|) \leq |W_R^{-1}|\mathrm{uvec}(S).$$

Thus (31) follows immediately from the Taylor series (10) for $R(t)$. □

2.2 *Strict perturbation bounds*

In this section, we will present strict perturbation bounds (with no higher order terms) for the Cholesky factorization. We give both norm based bounds, and component bounds.

Let $X = \Delta R$ and define $\phi(X) = X^T X$. Then (23) can be rewritten as

$$\mathbf{T}_R X = \Delta A - \phi(X). \tag{33}$$

Our purpose is to derive a condition for the existence of the upper triangular solution X and an upper bound on X. As in Stewart (1977), we can use the following lemma.

LEMMA 8 (A trivial variant of Theorem 3.1 in Stewart (1973) and Theorem 2.11 in Stewart & Sun (1990)) 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

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

Applying the above lemma to (33), we can obtain the next two theorems.

THEOREM 9 Let $A \in \mathcal{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^T R$. Let $\Delta A \in \mathcal{R}^{n\times n}$ be symmetric. If

$$\|\hat{W}_R^{-1}\|_2 \|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2 < 1/4, \tag{34}$$

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

$$A + \Delta A = (R + \Delta R)^T (R + \Delta R), \tag{35}$$

where

$$\begin{aligned} \|\Delta R\|_F &\leq \frac{2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}{1 + \sqrt{1 - 4\|\hat{W}_R^{-1}\|_2 \|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}} \\ &\leq 2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2 \leq 2\|\hat{W}_R^{-1}\|_2 \|\Delta A\|_F. \end{aligned} \tag{36}$$

Obviously, the weakest bound above can be rewritten in the following elegant form:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le 2\kappa_C(A)\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{37}$$

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

$$\|\phi(X)\|_F \le \|X\|_F^2,$$

and for any upper triangular matrices X and Y,

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

so we take $\eta = 1$. From (15), $\mathbf{T}_R X = \Delta A$ can be rewritten $\hat{W}_R \mathrm{uvec}(X) = \mathrm{duvec}(\Delta A)$, so

$$\|\mathbf{T}_R^{-1}\Delta A\|_F = \|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2, \quad \|\mathbf{T}_R^{-1}\|_F = \|\hat{W}_R^{-1}\|_2.$$

By the assumption (34), we have $\rho = 4\tau\eta/\delta < 1$ with $\tau = \|\mathbf{T}_R^{-1}\Delta A\|_F$, $\eta = 1$ and $\delta = \|\mathbf{T}_R^{-1}\|_F^{-1}$. Thus, by Lemma 8, (33) has a unique upper triangular solution, say ΔR, that satisfies

$$\|\Delta R\|_F \le \frac{2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}{1 + \sqrt{1 - 4\|\hat{W}_R^{-1}\|_2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}},$$

and (35) and (36) follow.

But $R + \Delta R$ in (35) must have positive diagonal to satisfy our definition of the Cholesky factor, where R was given and ΔR solved (33). We now prove this positivity. ¿From (34) and (36) it follows that

$$\|\Delta R\|_F < \frac{1}{2\|\hat{W}_R^{-1}\|_2}.$$

Combining this with (21), we have

$$\|\Delta R\|_F < \frac{1}{\|R^{-1}\|_2} = \sigma_n(R),$$

where $\sigma_n(R)$ is the smallest singular value of R. Thus $R + t\Delta R$ is nonsingular for any $t \in [0, 1]$, and by continuity of elements, $R + \Delta R$ has positive diagonal. □

REMARK 10 ¿From (17) and (21), it follows that

$$\frac{1}{4}\|A^{-1}\|_2\|\Delta A\|_F \le \|\hat{W}_R^{-1}\|_2^2\|\Delta A\|_F \le \frac{1}{2}\kappa_2(A)\|A^{-1}\|_2\|\Delta A\|_F.$$

In randomly perturbed problems we expect

$$\|\hat{W}_R^{-1}\|_2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2 \approx \|\hat{W}_R^{-1}\|_2^2\|\Delta A\|_F.$$

Thus the assumption (34) is generally stronger than (1), and may be greatly so. Following the same argument as Stewart (1993), however, (34) is needed to guarantee that the bound on $\|\Delta R\|_F$ will not explode. Furthermore, if the standard symmetric pivoting is used in computing the Cholesky factorization, then $\|\hat{W}_R^{-1}\|_2$ and $\frac{1}{2}\|R^{-1}\|$ will usually be of similar magnitude, see (54); that is, $\|\hat{W}_R^{-1}\|_2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2$ and $\frac{1}{4}\|A^{-1}\|_2\|\Delta A\|_F$ will usually have similar magnitude. So the condition (34) is not too constraining.

Numerical tests suggest that (36) is sharper than the equivalent bound in Sun (1991); see Section 3.2.

In order to analyze computed results, Theorem 9 may be rephrased as follows.

THEOREM 11 Let $A \in \mathcal{R}^{n\times n}$ be symmetric and $\tilde{A} \in \mathcal{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $\tilde{A} = \tilde{R}^T\tilde{R}$. Set $\Delta A = \tilde{A} - A$. If

$$\|\hat{W}_{\tilde{R}}^{-1}\|_2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2 < 1/4, \tag{38}$$

then A has the Cholesky factorization

$$A = R^T R,$$

and

$$\begin{aligned}\|\Delta R\|_F &\le \frac{2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2}{1+\sqrt{1-4\|\hat{W}_{\tilde{R}}^{-1}\|_2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2}} \\ &\le 2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2 \le 2\|\hat{W}_{\tilde{R}}^{-1}\|_2\|\Delta A\|_F, \end{aligned}\tag{39}$$

where $\Delta R \equiv \tilde{R} - R$. Clearly, the weakest bound can be rewritten in the following elegant form:

$$\frac{\|\Delta R\|_F}{\|\tilde{R}\|_2} \le 2\kappa_C(\tilde{A})\frac{\|\Delta A\|_F}{\|\tilde{A}\|_2}. \tag{40}$$

REMARK 12 In the above theorem, $\tilde{R}$ may be the computed Cholesky factor of A, in which case $\|\Delta A\|_F$ can be bounded from standard rounding error analysis results, so that (39) and (40) present computable *a posteriori* error bounds.

The componentwise perturbation bounds given by Sun (1992a, (2.2.3)) and Sun (1992b, (2.3)) reveal the structure characteristic of the problem. The former is rather complicated to state, while the latter is much simpler. But both are hard to interpret. Here, (40) not only respects the structure of the problem, but it is simple and can be easily interpreted. Its drawback is that assumption (38) on the size of ΔA is a bit restrictive.

Now we give a strict componentwise perturbation bound, and another strict normwise perturbation bound, neither of which requires strong restrictions on the size of ΔA.

THEOREM 13 Let both $A \in \mathcal{R}^{n\times n}$ and $\tilde{A} \in \mathcal{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^TR$ and $\tilde{A} = \tilde{R}^T\tilde{R}$, respectively. Set $\Delta A = \tilde{A} - A$ and $\Delta R = \tilde{R} - R$. Then

$$\mathrm{uvec}(|\Delta R|) \le |W_{R,\tilde{R}}^{-1}|\mathrm{uvec}(|\Delta A|), \tag{41}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} = \frac{\|\hat{W}_{R,\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2}{\|R\|_2} \le \|\hat{W}_{R,\tilde{R}}^{-1}\|_2\|A\|_2^{1/2}\frac{\|\Delta A\|_F}{\|A\|_2}, \tag{42}$$

where $W_{R,\tilde{R}} \in \mathcal{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is

$$\left[\begin{array}{cccccccc}
r_{11}+\tilde{r}_{11} & & & & & & & \\
r_{12} & \tilde{r}_{11} & & & & & & \\
 & r_{12}+\tilde{r}_{12} & r_{22}+\tilde{r}_{22} & & & & & \\
\cdot & \cdot & \cdot & \cdot & & & & \\
r_{1n} & & & & \tilde{r}_{11} & & & \\
 & r_{1n} & r_{2n} & & \tilde{r}_{12} & \tilde{r}_{22} & & \\
 & & & \cdot & \cdot & \cdot & \cdot & \\
 & & & & r_{1n}+\tilde{r}_{1n} & r_{2n}+\tilde{r}_{2n} & \cdot & r_{nn}+\tilde{r}_{nn}
\end{array}\right],$$

which has the same zero/nonzero structure as W_R, and $\hat{W}_{R,\tilde{R}} \equiv DW_{R,\tilde{R}}$ with D as in (14).

Proof. It is easy to verify that

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

Then its upper triangular part can be written as

$$W_{R,\tilde{R}}\mathrm{uvec}(\Delta R) = \mathrm{uvec}(\Delta A) \tag{43}$$

(for details see Appendix B), which is equivalent to

$$\hat{W}_{R,\tilde{R}}\mathrm{uvec}(\Delta R) = \mathrm{duvec}(\Delta A). \tag{44}$$

Note that $W_{R,\tilde{R}}$ and $\hat{W}_{R,\tilde{R}}$ are nonsingular; thus, (41) and (42) can be deduced from (43) and (44) respectively, the latter by using $\|A\|_2 = \|R\|_2^2$. $\square$

Remark 14 It is important to realize that knowing $W_{R,\tilde{R}}$ implies that both R and $\tilde{R}$ are known. But in this case $\Delta R = \tilde{R} - R$ is known exactly, so Theorem 13 has only theoretical, not practical value.

3. Numerical experiments

3.1 *Numerical examples for the first order perturbation bound*

In Section 2.1, we presented a first order perturbation analysis for the Cholesky factorization, defined $\kappa_C(A) \equiv \|\hat{W}_R^{-1}\|_2\|A\|_2^{1/2}$ as the condition number for the Cholesky factorization, gave a lower bound and upper bound for $\kappa_C(A)$, and showed that the first order result (24) is sharper than (2). Now we use some numerical tests to illustrate our results.

We computed three sets of examples by using Matlab. The matrices in the first set have the form

$$A = U\Lambda U^T,$$

Table 1
Results for matrix $A = U\Lambda U^T$ of order 10

ε	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(PAP^T)$	$\kappa_C(A)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$	$\frac{2\kappa_C(PAP^T)}{\kappa_2^{1/2}(A)}$
1.0	9.6348e-01	1.4864e+00	1.4966e+00	2.6256e+00	1.54
1.0e-01	2.4635e+00	4.0012e+00	6.1989e+00	1.7166e+01	1.62
1.0e-02	4.8870e+00	9.5844e+00	2.2304e+01	6.7554e+01	1.96
1.0e-03	1.3367e+01	2.7135e+01	9.5212e+01	5.0538e+02	2.03
1.0e-04	4.8424e+01	9.2252e+01	6.4081e+02	6.6319e+03	1.90
1.0e-05	1.4540e+02	3.9903e+02	1.3352e+03	5.9791e+04	2.74
1.0e-06	4.9581e+02	1.1100e+03	2.0200e+03	6.9530e+05	2.24
1.0e-07	1.5501e+03	3.6469e+03	2.9800e+05	6.7960e+06	2.35
1.0e-08	4.4806e+03	9.5121e+03	6.7845e+04	5.6786e+07	2.12
1.0e-09	1.5390e+04	3.2109e+04	6.6104e+04	6.6997e+08	2.09
1.0e-10	4.4247e+04	1.0561e+05	2.8532e+05	5.5377e+09	2.39

where $U \in \mathcal{R}^{n\times n}$ is an orthogonal matrix obtained from the QR factorization of a random $n \times n$ matrix, Λ=diag$(r_1,..., r_{n-2}, \varepsilon, \varepsilon)$ with $r_1, ..., r_{n-2}$ random positive numbers and $0 < \varepsilon \le 1$. We generated different matrices by taking all combinations of $n \in \{4, 6, 10, 15, 20\}$ and $\varepsilon \in \{1, 10^{-1},..., 10^{-10}\}$. The results for $n = 10$, $\varepsilon = 10^{-i}$, i=0, 1, ..., 10 are shown in Table 1. Results obtained by putting the two ε's at the front of Λ were mainly similar, but occasionally gave $\kappa_C(A)$ up to 100 times the value of $\kappa_C(PAP^T)$. Note in Table 1 how the $\kappa_2(A)/\sqrt{2}$ bound can be much worse (larger) than $\kappa_C(A)$ for ill-conditioned problems, and how symmetric pivoting always gave an improvement on $\kappa_C(A)$, bringing $\kappa_C(PAP^T)$ very close to its minimal value of $\kappa_2^{1/2}(A)/2$.

The second set of matrices are $n \times n$ Hilbert matrices, $n = 1,..., 10$. The results are shown in Table 2. Again $\kappa_C(A)$ is significantly smaller than $\kappa_2(A)/\sqrt{2}$, but pivoting gives little further improvement (for $n = 4$ it gives a marginally *worse* bound).

The third set of matrices are $n \times n$ Pascal matrices (with elements $a_{1j} = a_{i1} = 1$, $a_{ij} = a_{i,j-1} + a_{i-1,j}$), $n = 1,..., 10$. The results are shown in Table 3. Here pivoting is seen to make a significant difference for large n.

Remark 15 The numerical experiments confirm that the bound (2) can be very pessimistic. The bound (24) may be much sharper than (2). These tests suggest that if the Cholesky factorization of A is approached using the standard symmetric pivoting strategy, then the condition number of the Cholesky factorization $\kappa_C(PAP^T)$ will have the same order as its lower limit $\kappa_2^{1/2}(A)/2$ (the ratio in the last column was never larger than 3). However, though this appears to occur for most A, it is possible to give examples with n not small where it does not hold. We will show in (53) that $\kappa_2^{1/2}(A) \le 2\kappa_C(PAP^T) \le \kappa_2^{1/2}(A)2^n\sqrt{n(n+1)/2}$ when R is computed with standard symmetric pivoting, and an example was found where $A = R^TR$ is

TABLE 2
Results for Hilbert matrix of order n, $n = 1, \ldots, 10$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(PAP^T)$	$\kappa_C(A)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$	$\frac{2\kappa_C(PAP^T)}{\kappa_2^{1/2}(A)}$
1	5.0000e-01	5.0000e-01	5.0000e-01	7.0711e-01	1.00
2	2.1955e+00	2.4969e+00	2.4969e+00	1.3634e+01	1.14
3	1.1446e+01	1.6000e+01	1.6520e+01	3.7056e+02	1.40
4	6.2277e+01	1.2099e+02	1.1913e+02	1.0970e+04	1.94
5	3.4518e+02	4.5016e+02	8.9552e+02	3.3701e+05	1.30
6	1.9333e+03	2.8008e+03	6.9072e+03	1.0572e+07	1.45
7	1.0901e+04	1.8447e+04	5.4250e+04	3.3614e+08	1.69
8	6.1761e+04	8.9868e+04	4.3207e+05	1.0789e+10	1.46
9	3.5112e+05	6.1781e+05	3.4804e+06	3.4871e+11	1.76
10	2.0015e+06	3.2462e+06	2.8303e+07	1.1331e+13	1.62

a 15×15 matrix with

$$R = \mathrm{diag}(1, s, \cdots, s^{14}) \begin{bmatrix} 1 & -c & -c & \cdot & -c \\ & 1 & -c & \cdot & -c \\ & & 1 & \cdot & -c \\ & & & \cdot & \cdot \\ & & & & 1 \end{bmatrix}$$

with $c = \cos(\pi/4)$, $s = \sin(\pi/4)$; $2\kappa_C(PAP^T)/\kappa_2^{1/2}(A) = 1.0337\text{e+}03$. (The matrix R was introduced by Kahan (1966); see also Higham (1990).) However such examples are rare, and Table 3 shows how the sensitivity of the problem can be significantly improved by this symmetric pivoting, in that our new more accurate measure of sensitivity can be greatly decreased.

TABLE 3
Results for Pascal matrix of order n, $n = 1, \ldots, 10$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(PAP^T)$	$\kappa_C(A)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$	$\frac{2\kappa_C(PAP^T)}{\kappa_2^{1/2}(A)}$
1	5.0000e-01	5.0000e-01	5.0000e-01	7.0711e-01	1.00
2	1.3090e+00	1.5487e+00	2.1219e+00	4.8466e+00	1.18
3	3.9365e+00	5.0569e+00	9.7474e+00	4.3829e+01	1.28
4	1.3152e+01	2.2142e+01	5.4970e+01	4.8927e+02	1.68
5	4.6145e+01	8.3012e+01	3.5255e+02	6.0228e+03	1.80
6	1.6642e+02	2.5264e+02	2.5007e+03	7.8338e+04	1.52
7	6.1103e+02	9.4277e+02	1.9147e+04	1.0560e+06	1.54
8	2.2718e+03	4.0196e+03	1.5465e+05	1.4598e+07	1.77
9	8.5262e+03	1.6040e+04	1.2958e+06	2.0561e+08	1.88
10	3.2230e+04	7.5871e+04	1.1141e+07	2.9382e+09	2.35

3.2 *A numerical example for the strict perturbation bounds*

Numerical experiments suggest that the strict perturbation bounds presented in Sections 2.2 are sharper than those of Stewart (1977) and Sun (1991), and are comparable with those of Sun (1992a) (1992b). Now we give one simple numerical example.

$$A = \begin{pmatrix} 25 & 20 & 15 & 10 & 5 \\ 20 & 32 & 24 & 16 & 8 \\ 15 & 24 & 27 & 18 & 9 \\ 10 & 16 & 18 & 16 & 8 \\ 5 & 8 & 9 & 8 & 5 \end{pmatrix}$$

The matrix A is symmetric positive definite, and its Cholesky factor R is given by

$$R = \begin{pmatrix} 5 & 4 & 3 & 2 & 1 \\ 0 & 4 & 3 & 2 & 1 \\ 0 & 0 & 3 & 2 & 1 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Take a 5×5 random symmetric matrix

$$\Delta A = (1.0e-03) \begin{pmatrix} 0.2113 & 0.4649 & 0.5174 & 0.4857 & 0.6167 \\ 0.4649 & 0.4524 & 0.8441 & 0.7382 & 0.4374 \\ 0.5174 & 0.8441 & 0.6538 & 0.6630 & 0.5072 \\ 0.4857 & 0.7382 & 0.6630 & 0.7469 & 0.1891 \\ 0.6167 & 0.4374 & 0.5072 & 0.1891 & 0.1167 \end{pmatrix}$$

and let $\tilde{A} = A + \Delta A$. Let $\tilde{R}$ be the Cholesky factor of $\tilde{A}$, and $\Delta R = \tilde{R} - R$. Then we get

$$\|\Delta R\|_F = 2.7419\text{e-}04,$$

$$\|\hat{W}_R^{-1}\|_2 \|\hat{W}_R^{-1} \mathrm{duvec}(\Delta A)\|_2 = 1.8830\text{e-}04 < 1/4,$$

$$\|\hat{W}_{\tilde{R}}^{-1}\|_2 \|\hat{W}_{\tilde{R}}^{-1} \mathrm{duvec}(\Delta A)\|_2 = 1.8827\text{e-}04 < 1/4.$$

The bounds on ΔR given by Theorem 9, 11, and 13 are

$$\beta_1 \equiv \frac{2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}{1 + \sqrt{1 - 4\|\hat{W}_R^{-1}\|_2 \|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2}} = 2.7425\text{e-}04,$$

$$\beta_2 \equiv 2\|\hat{W}_R^{-1}\mathrm{duvec}(\Delta A)\|_2 = 5.4839\text{e-}04, \quad \beta_3 \equiv 2\|\hat{W}_R^{-1}\|_2 \|\Delta A\|_F = 3.8385\text{e-}03,$$

$$\beta_4 \equiv \frac{2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2}{1 + \sqrt{1 - 4\|\hat{W}_{\tilde{R}}^{-1}\|_2 \|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2}} = 2.7424\text{e-}04,$$

$$\beta_5 \equiv 2\|\hat{W}_{\tilde{R}}^{-1}\mathrm{duvec}(\Delta A)\|_2 = 5.4838\text{e-}04, \quad \beta_6 \equiv 2\|\hat{W}_{\tilde{R}}^{-1}\|_2\|\Delta A\|_F = 3.8381\text{e-}03,$$

$$\beta_7 \equiv \|\hat{W}_{R,\tilde{R}}^{-1}\|_2\|\Delta A\|_F = 1.9191\text{e-}03.$$

Under the same assumption as Theorem 9, the perturbation bound on ΔR given by Stewart (1977) is

$$\|\Delta R\|_F \leq 2n\|R^{-1}\|_F \left(1 + \frac{\|R\|_F\|R^{-1}\|_F}{\sqrt{2}}\|\Delta A\|_F\right) \equiv \beta_8.$$

Sun (1991) proved that if

$$\alpha \equiv \|A^{-1}\|_2\|\Delta A\|_2 < 1,$$

then

$$\|\Delta R\|_F \leq \frac{\|A\|_2^{1/2}\|A^{-1}\|_2\|\Delta A\|_F}{\sqrt{2}(1 - \|A^{-1}\|_2\|\Delta A\|_F)} \equiv \beta_9.$$

Through simple calculations, we get $\alpha = 3.5684\text{e-}03$ and

$$\beta_8 = 4.2907\text{e-}01, \quad \beta_9 = 2.3575\text{e-}02.$$

so certainly for this (well-conditioned) example our bounds are clearly stronger.

Componentwise perturbation analyses for the Cholesky factorization are given in Sun (1992a, 1992b). Sun (1992a) proved that

$$|\Delta R| \leq \mathrm{triu}(|\tilde{R}^{-T}||\Delta A||\tilde{R}^{-1}|)|\tilde{R}| \equiv B_{g,1},$$

and if

$$\tilde{\alpha} \equiv \min_{\substack{W = \mathrm{diag}(\omega_1,\dots,\omega_n)\\ \omega_j > 0 \forall j}} \left[(\|(WAW)^{-1}\|_2\|(W\tilde{A}W)^{-1}\|_2)^{1/2}\|W\Delta AW\|_F\right] < 1,$$

then

$$\begin{aligned} |\Delta R| &\leq \mathrm{triu}(|\tilde{R}^{-T}||\Delta A||\tilde{R}^{-1}|)|\tilde{R}| + \mathrm{triu}(|\tilde{R}^{-T}||\Delta A||\tilde{R}^{-1}|\Phi|\tilde{R}^{-T}||\Delta A||\tilde{R}^{-1}|)|\tilde{R}| \\ &\equiv B_{g,2}, \end{aligned}$$

where $\Phi = (\phi_{ij})$, $\phi_{ij} = 1/(1-\tilde{\alpha})^2 \ \forall i,j$. Sun (1992b) also proved that if

$$\rho(\Delta A_r) < 1,$$

where $\Delta A_r = |\tilde{R}^{-T}\Delta A\tilde{R}^{-1}|$, then

$$|\Delta R| \leq \mathrm{triu}(\Delta A_r(I - \Delta A_r)^{-1})|\tilde{R}| \equiv B_{ch}.$$

Calculations give

$$\tilde{\alpha} \leq (\|A^{-1}\|_2\|\tilde{A}^{-1}\|_2)^{1/2}\|\Delta A\|_F = 3.6782\text{e-}03 < 1, \quad \rho(\Delta A_r) = 2.7303\text{e-}04,$$

and

$$\beta_{10} \equiv \|B_{g,1}\|_F = 2.8236\text{e-}03, \quad \beta_{11} \equiv \|B_{g,2}\|_F \leq 2.8428\text{e-}03,$$

$$\beta_{12} \equiv \|B_{ch}\|_F \leq 4.1675\text{e-}04.$$

These are comparable with our β_1 to β_7.

4. Practical bounds and condition estimators

In Section 2 we looked at $A + tG = R^T(t)R(t)$, $0 \le t \le \epsilon$, for the perturbed Cholesky factorization $A + \Delta A = (R + \Delta R)^T(R + \Delta R)$, where $G \equiv \Delta A/\epsilon$ with $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ so $\|G\|_F = \|A\|_2$. We showed (28)

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \kappa_C(A) \equiv \|\hat{W}_R^{-1}\|_2\|A\|_2^{1/2},$$

from which we proved (24)

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

and saw that $\kappa_C(A)$ is the condition number (for bounding the above norm ratios) for the Cholesky factorization.

We showed via (29) that there can be no better such measure of sensitivity than $\kappa_C(A)$. To emphasize this, (15) may be written

$$\hat{W}_R \mathrm{uvec}(\dot{R}(0)) = \mathrm{duvec}(G),$$

where $\|\mathrm{uvec}(\dot{R}(0))\|_2 = \|\dot{R}(0)\|_F$ and $\|\mathrm{duvec}(G)\|_2 = \|G\|_F$. Now for any definite $A = R^TR$ it is clearly possible to select symmetric ΔA, and so G, giving

$$\|\dot{R}(0)\|_F = \|\mathrm{uvec}(\dot{R}(0))\|_2 = \|\hat{W}_R^{-1}\mathrm{duvec}(G)\|_2 = \|\hat{W}_R^{-1}\|_2\|G\|_F,$$

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} = \|\hat{W}_R^{-1}\|_2\|A\|_2^{1/2} = \kappa_C(A),$$

so the upper bound can always be obtained, and $\kappa_C(A)$ is (again) optimal.

The drawback with this approach is that $\kappa_C(A)$ is both difficult to understand and (so far) to compute or estimate, except when we use pivoting, in which case $\kappa_C(PAP^T)$ usually approaches its lower bound $\kappa_2^{1/2}(A)/2$, see (53) and Remark 15. Fortunately, the approach of Stewart (1993, 1995b) can be extended to obtain expressions which, though they are larger than $\kappa_C(A)$, are excellent estimates of this, and also give considerable insight into what is going on, and lead to efficient and practical condition estimators even when we do not use pivoting.

We want to bound $\|\dot{R}(0)\|_F/\|R\|_2$, where (6) tells us

$$\dot{R}(0)R^{-1} = \mathrm{up}(R^{-T}GR^{-1}).$$

A useful observation is that for any diagonal matrix $\mathcal{D}$,

$$\mathrm{up}(X)\mathcal{D} = \mathrm{up}(X\mathcal{D}).$$

For any nonsingular diagonal $\mathcal{D}$ we can take $R = \mathcal{D}\bar{R}$, giving

$$\dot{R}(0)R^{-1} = \mathrm{up}(R^{-T}G\bar{R}^{-1}\mathcal{D}^{-1}) = \mathrm{up}(R^{-T}G\bar{R}^{-1})\mathcal{D}^{-1},$$

which leads to cancellation of the $\mathcal{D}^{-1}$:

$$\dot{R}(0) = \mathrm{up}(R^{-T}G\bar{R}^{-1})\bar{R}, \tag{45}$$

and since for any matrix X, $\|\mathrm{up}(X)\|_F \le \|X\|_F$,

$$\|\dot{R}(0)\|_F \le \|\mathrm{up}(R^{-T}G\bar{R}^{-1})\|_F\|\bar{R}\|_2 \le \|R^{-1}\|_2\|G\|_F\kappa_2(\bar{R}).$$

Using $\|G\|_F = \|A\|_2 = \|R\|_2^2$ we get

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \kappa_2(R)\kappa_2(\bar{R}) \equiv \kappa_{C1}(A,\mathcal{D}), \quad \text{say}, \tag{46}$$

and $\kappa_{C1}(A,\mathcal{D})$ is a potential estimator for $\kappa_C(A)$. Since $\mathcal{D}$ is an arbitrary nonsingular diagonal matrix, it may be chosen to minimize $\kappa_{C1}(A,\mathcal{D})$,

$$\kappa_{C1}(A) \equiv \min_{\text{nonsingular diagonal } \mathcal{D}} \kappa_{C1}(A,\mathcal{D}), \tag{47}$$

with the minimum being no greater than $\kappa_2(A)$ by taking $\mathcal{D} = I$. This shows $\frac{1}{\sqrt{2}}$ times this estimate will be at least as good as (9), the first order normwise perturbation bound in Sun (1991) and Stewart (1993).

This approach is very simple yet gives considerable insight into why the Cholesky factorization can be far more well-conditioned than we previously thought. For example, if the ill-conditioning of R is mostly due to bad scaling of the rows, then correct choice of $\mathcal{D}$ in $R = \mathcal{D}\bar{R}$ can give $\kappa_2(\bar{R})$ very near one, so $\kappa_{C1}(A,\mathcal{D})$ will approach twice the lower bound $\kappa_2^{1/2}(A)/2$ on $\kappa_C(A)$. As an illustration suppose

$$R = \begin{pmatrix} 1 & \epsilon \\ 0 & \epsilon \end{pmatrix},$$

then for small ϵ, $\kappa_2(R) = O(\frac{1}{\epsilon})$. But if we set $\mathcal{D} = \mathrm{diag}(1,\epsilon)$, then $\kappa_2(\mathcal{D}) = 1/\epsilon$ and $\kappa_2(\bar{R}) = O(1)$, so that $\kappa_{C1}(A,\mathcal{D})$ is close to the lower bound on $\kappa_C(A)$. Note how almost all the ill-conditioning was revealed by the diagonal of R.

This also provides another explanation as to why the standard pivoting of A is so successful, making $\kappa_C(PAP^T)$ approach its lower bound in nearly all cases. If A is ill-conditioned (so there is a large distance between the lower and upper bounds on $\kappa_C(A)$) and the Cholesky factorization is computed with the standard pivoting, the ill-conditioning of A will usually reveal itself in the diagonal elements of R. Stewart (1995a) has shown that such upper triangular matrices are artificially ill-conditioned in the sense that they can be made well-conditioned by scaling the rows via $\mathcal{D}$. This implies that $\kappa_{C1}(PAP^T,\mathcal{D})$, and therefore (as we shall show) $\kappa_C(PAP^T)$, will approach its lower bound. We support this mathematically in Section 5.

Note the cancellation of $\mathcal{D}^{-1}$ with $\mathcal{D}$ leading to (45) is the crucial insight here — the up-notation just adds clarity to the derivation.

The practical outcome of this simple analysis is that we now have an $O(n^2)$ condition estimator for the Cholesky factor. By a well known result of van der Sluis

TABLE 4
Results for Hilbert matrix of order n, $n = 1,\ldots, 12$ without pivoting, $\mathcal{D} = \mathcal{D}_1$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(A)$	$\kappa_{C1}(A,\mathcal{D}_1)$	$\kappa_{C2}(A,\mathcal{D}_1)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$
1	5.0000e-01	5.0000e-01	1.0000e+00	7.0711e-01	7.0711e-01
2	2.1955e+00	2.4969e+00	7.1049e+00	6.0017e+00	1.3634e+01
3	1.1446e+01	1.6520e+01	6.5290e+01	6.4459e+01	3.7056e+02
4	6.2277e+01	1.1913e+02	6.6739e+02	8.0584e+02	1.0970e+04
5	3.4518e+02	8.9552e+02	7.1518e+03	1.0904e+04	3.3701e+05
6	1.9333e+03	6.9072e+03	3.3701e+05	1.5432e+05	1.0572e+07
7	1.0901e+04	5.4250e+04	8.7446e+05	2.2435e+06	3.3614e+08
8	6.1761e+04	4.3207e+05	9.8211e+06	3.3183e+07	1.0789e+10
9	3.5112e+05	3.4804e+06	1.1095e+08	4.9665e+08	3.4871e+11
10	2.0015e+06	2.8303e+07	1.2585e+09	7.4974e+09	1.1331e+13
11	1.1408e+07	2.3156e+08	1.4289e+10	1.1369e+11	3.6811e+14
12	6.3484e+07	1.8583e+09	1.5850e+11	1.6924e+12	1.1399e+16

(1969), $\kappa_2(\bar{R})$ will be nearly optimal when the rows of $\bar{R}$ are equilibrated. Thus the procedure is to choose $\mathcal{D}$ in $R = \mathcal{D}\bar{R}$ so that the rows of $\bar{R}$ are equilibrated, and use condition estimators to estimate $\kappa_2(\bar{R})$ and $\kappa_2(R)$ in

$$\frac{1}{2}\kappa_2^{1/2}(A) \leq \kappa_C(A) \leq \kappa_{C1}(A) \leq \kappa_{C1}(A,\mathcal{D}) = \kappa_2(R)\kappa_2(\bar{R}), \tag{48}$$

where we will prove the second inequality later. The tables of computations here illustrate the effectiveness of this estimate. In these, $\kappa_{C2}(A,\mathcal{D})$ will be described later, $\kappa_{C1}(A,\mathcal{D}) = \kappa_2(R)\kappa_2(\bar{R})$, and $\mathcal{D}_1 = \mathrm{diag}(d_{ii}^{(1)})$ where $d_{ii}^{(1)} = \sqrt{\sum_{j=i}^{n} r_{ij}^2}$.

Note in Table 4 for $n = 12$ the estimate $\kappa_{C1}(A,\mathcal{D}_1)$ is over 80 times $\kappa_C(A)$. A study of the $n = 2$ case shows $\kappa_{C1}(A)$ can never be much larger than $\kappa_C(A)$, so we suspect $\kappa_{C1}(A)$ is at worst $\kappa_C(A)$ times some function of n in general, and hopefully the same will be true for $\kappa_{C1}(A,\mathcal{D}_1)$.

Table 5 suggests for Hilbert matrices that the estimate will be significantly better when pivoting is used. However with pivoting we usually have $\frac{1}{2}\kappa_2^{1/2}(A) \approx \kappa_C(PAP^T)$, and for about half the cost we can estimate this lower bound, which, in this example at least, is superior. Of course estimating the upper bound $\kappa_{C1}(A,\mathcal{D}_1)$ also gives an estimate of the lower bound $\frac{1}{2}\kappa_2^{1/2}(A)$ at no extra cost, and this appears the most appealing strategy.

Similar comments can be made for Pascal matrices, as can be seen from Tables 6 & 7. For $n = 19$ without pivoting $\kappa_{C1}(A,\mathcal{D}_1)$ overestimates $\kappa_C(A)$ by a factor of about 600, but is much better for low n. This suggests the (assumed) function of n bounding the ratio $\kappa_{C1}(A,\mathcal{D}_1)/\kappa_C(A)$ may involve something like 2^n.

We know $\kappa_C(A)$ is the optimal condition number, so necessarily $\kappa_C(A) \leq \kappa_{C1}(A)$. But it is useful to prove this directly, and so obtain some indication of how weak $\kappa_{C1}(A)$ is as an approximation to $\kappa_C(A)$. ¿From (11) we see that for any symmetric

Table 5
Results for Hilbert matrix of order $n = 1, ..., 12$ with pivoting, $\tilde{A} = PAP^T$, $\mathcal{D} = \mathcal{D}_1$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(\tilde{A})$	$\kappa_{C1}(\tilde{A}, \mathcal{D}_1)$	$\kappa_{C2}(\tilde{A}, \mathcal{D}_1)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$
1	5.0000e-01	5.0000e-01	1.0000e+00	7.0711e-01	7.0711e-01
2	2.1955e+00	2.4969e+00	7.1049e+00	6.0017e+00	1.3634e+01
3	1.1446e+01	1.6000e+01	5.9448e+01	5.4321e+01	3.7056e+02
4	6.2277e+01	1.2099e+02	4.6857e+02	4.7277e+02	1.0970e+04
5	3.4518e+02	4.5016e+02	3.1085e+03	5.3306e+03	3.3701e+05
6	1.9333e+03	2.8008e+03	2.1668e+04	3.5214e+04	1.0572e+07
7	1.0901e+04	1.8447e+04	1.2176e+05	1.6765e+05	3.3614e+08
8	6.1761e+04	8.9868e+04	8.6676e+05	1.5050e+06	1.0789e+10
9	3.5112e+05	6.1781e+05	5.1312e+06	8.0254e+06	3.4871e+11
10	2.0015e+06	3.2462e+06	2.9591e+07	5.0379e+07	1.1331e+13
11	1.1408e+07	2.4775e+07	2.0836e+08	2.9586e+08	3.6811e+14
12	6.3484e+07	1.2052e+08	1.1566e+09	1.9383e+09	1.1399e+16

Table 6
Results for Pascal matrix of order n, $n = 1, ..., 19$ without pivoting, $\mathcal{D} = \mathcal{D}_1$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(A)$	$\kappa_{C1}(A, \mathcal{D}_1)$	$\kappa_{C2}(A, \mathcal{D}_1)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$
1	5.0000e-01	5.0000e-01	1.0000+00	7.0711e-01	7.0711e-01
2	1.3090e+00	2.1219e+00	6.3205e+00	5.1038e+00	4.8466e+00
3	3.9365e+00	9.7474e+00	5.0300e+01	5.1169e+01	4.3829e+01
4	1.3152e+01	5.4970e+01	4.7797e+02	6.5650e+02	4.8927e+02
5	4.6145e+01	3.5255e+02	4.8942e+03	9.3737e+03	6.0228e+03
6	1.6642e+02	2.5007e+03	5.2071e+04	1.4145e+05	7.8338e+04
7	6.1103e+02	1.9147e+04	5.6705e+05	2.2078e+06	1.0560e+06
8	2.2718e+03	1.5465e+05	6.2730e+06	3.5254e+07	1.4598e+07
9	8.5262e+03	1.2958e+06	7.0187e+07	5.7231e+08	2.0561e+08
10	3.2230e+04	1.1141e+07	7.9207e+08	9.4076e+09	2.9382e+09
11	1.2254e+05	9.7648e+07	8.9987e+09	1.5617e+11	4.2471e+10
12	4.6808e+05	8.6857e+08	1.0279e+11	2.6129e+12	6.1970e+11
13	1.7950e+06	7.8174e+09	1.1793e+12	4.4002e+13	9.1129e+12
14	6.9059e+06	7.1040e+10	1.3580e+13	7.4504e+14	1.3489e+14
15	2.6644e+07	6.5077e+11	1.5687e+14	1.2673e+16	2.0079e+15
16	1.0305e+08	6.0022e+12	1.8170e+15	2.1643e+17	3.0035e+16
17	3.9940e+08	5.5686e+13	2.1095e+16	3.7090e+18	4.5118e+17
18	1.5509e+09	5.1923e+14	2.4542e+17	6.3753e+19	6.8031e+18
19	6.0322e+09	4.8628e+15	2.8603e+18	1.0988e+21	1.0292e+20

TABLE 7
Results for Pascal matrix of order $n = 1, ..., 19$ with pivoting, $\tilde{A} = PAP^T$, $\mathcal{D} = \mathcal{D}_1$

n	$\frac{1}{2}\kappa_2^{1/2}(A)$	$\kappa_C(\tilde{A})$	$\kappa_{C1}(\tilde{A}, \mathcal{D}_1)$	$\kappa_{C2}(\tilde{A}, \mathcal{D}_1)$	$\frac{1}{\sqrt{2}}\kappa_2(A)$
1	5.0000e-01	5.0000e-01	1.0000e+00	7.0711e-01	7.0711e-01
2	1.3090e+00	1.5487e+00	4.2361e+00	3.5212e+00	4.8466e+00
3	3.9365e+00	5.0569e+00	1.5760e+01	1.2976e+01	4.3829e+01
4	1.3152e+01	2.2142e+01	8.0203e+01	7.5960e+01	4.8927e+02
5	4.6145e+01	8.3012e+01	3.3412e+02	3.4403e+02	6.0228e+03
6	1.6642e+02	2.5264e+02	1.2670e+03	1.5843e+03	7.8338e+04
7	6.1103e+02	9.4277e+02	5.0710e+03	6.6876e+03	1.0560e+06
8	2.2718e+03	4.0196e+03	2.3585e+04	3.0756e+04	1.4598e+07
9	8.5262e+03	1.6040e+04	1.0151e+05	1.2964e+05	2.0561e+08
10	3.2230e+04	7.5871e+04	4.6645e+05	5.6383e+05	2.9382e+09
11	1.2254e+05	2.3711e+05	1.7789e+06	2.7092e+06	4.2471e+10
12	4.6808e+05	8.3284e+05	8.2111e+06	1.5288e+07	6.1970e+11
13	1.7950e+06	3.2173e+06	3.1391e+07	5.6878e+07	9.1129e+12
14	6.9059e+06	1.3381e+07	1.2391e+08	2.2788e+08	1.3489e+14
15	2.6644e+07	5.3884e+07	4.9035e+08	8.2432e+08	2.0079e+15
16	1.0305e+08	1.9624e+08	2.1144e+09	4.3536e+09	3.0035e+16
17	3.9940e+08	8.2578e+08	7.8097e+09	1.3445e+10	4.5118e+17
18	1.5509e+09	3.0381e+09	3.4538e+10	6.7700e+10	6.8031e+18
19	6.0322e+09	1.1875e+10	1.6305e+11	3.6712e+11	1.0292e+20

ΔA and resulting G and $\dot{R}(0)$,

$$\mathbf{T}_R\dot{R}(0) = R^T\dot{R}(0) + \dot{R}(0)^T R = G,$$

so that for any nonsingular diagonal $\mathcal{D}$ in $R = \mathcal{D}\bar{R}$

$$\|\mathbf{T}_R^{-1}G\|_F = \|\dot{R}(0)\|_F = \|\dot{R}(0)\bar{R}^{-1}\bar{R}\|_F.$$

But from (45), $\dot{R}(0)\bar{R}^{-1} = \mathrm{up}(R^{-T}G\bar{R}^{-1})$, so

$$\begin{aligned} \|\mathbf{T}_R^{-1}G\|_F &\leq \|\mathrm{up}(R^{-T}G\bar{R}^{-1})\|_F\|\bar{R}\|_2 && (49)\\ &\leq \kappa_2(\bar{R})\|R^{-1}\|_2\|G\|_F, && (50) \end{aligned}$$

and thus from (16)

$$\|\hat{W}_R^{-1}\|_2 = \sup_{G:\mathrm{symmetric}\neq 0} \frac{\|\mathbf{T}_R^{-1}G\|_F}{\|G\|_F} \leq \kappa_2(\bar{R})\|R^{-1}\|_2,$$

which implies

$$\kappa_C(A) \equiv \|\hat{W}_R^{-1}\|_2\|R\|_2 \leq \kappa_2(\bar{R})\kappa_2(R) \equiv \kappa_{C1}(A, \mathcal{D}),$$

for all nonsingular diagonal $\mathcal{D}$. Note the two inequalities (49) and (50) in going from $\kappa_C(A)$ to $\kappa_{C1}(A, \mathcal{D})$.

Finally we note that we discarded some symmetry in deriving $\kappa_{C1}(A,\mathcal{D})$. We obtain a slightly different estimate by maintaining symmetry. For symmetric X and any diagonal matrix $\mathcal{D}$, $\mathcal{D}X\mathcal{D}$ is symmetric and

$$\mathcal{D}\mathrm{up}(X)\mathcal{D} = \mathrm{up}(\mathcal{D}X\mathcal{D}).$$

Again for any nonsingular diagonal $\mathcal{D}$ we can take $R = \mathcal{D}\bar{R}$, giving

$$\dot{R}(0)R^{-1} = \mathrm{up}(\mathcal{D}^{-1}\bar{R}^{-T}G\bar{R}^{-1}\mathcal{D}^{-1}) = \mathcal{D}^{-1}\mathrm{up}(\bar{R}^{-T}G\bar{R}^{-1})\mathcal{D}^{-1},$$

which leads to cancellation of the rightmost $\mathcal{D}^{-1}$:

$$\dot{R}(0) = \mathcal{D}^{-1}\mathrm{up}(\bar{R}^{-T}G\bar{R}^{-1})\bar{R},$$

$$\|\dot{R}(0)\|_F \le \|\mathcal{D}^{-1}\|_2\|\bar{R}\|_2\|\mathrm{up}(\bar{R}^{-T}G\bar{R}^{-1})\|_F.$$

Now using (7) for any $n\times n$ symmetric matrix X, and remembering $\|G\|_F = \|A\|_2 = \|R\|_2^2$, we see

$$\begin{aligned}\frac{\|\dot{R}(0)\|_F}{\|R\|_2} &\le \frac{1}{\sqrt{2}}\|\mathcal{D}^{-1}\|_2\|\bar{R}\|_2\|\bar{R}^{-1}\|_2^2\frac{\|G\|_F}{\|R\|_2^2}\|R\|_2 \\ &= \frac{1}{\sqrt{2}}\|\mathcal{D}^{-1}\|_2\|R\|_2\|\bar{R}^{-1}\|_2\kappa_2(\bar{R}) \equiv \kappa_{C2}(A,\mathcal{D}), \quad \text{say}, \qquad (51)\\ &\le \frac{1}{\sqrt{2}}\kappa_2(\mathcal{D})\kappa_2^2(\bar{R}) \equiv \kappa_{C3}(A,\mathcal{D}), \quad \text{say}, \qquad (52)\end{aligned}$$

and $\kappa_{C2}(A,\mathcal{D})$ is the quantity that appears in Tables 4 to 7. It can be estimated at somewhat less cost than $\kappa_{C1}(A,\mathcal{D})$, but the computations show it can sometimes be significantly larger than the latter for large n, and so should be used with care. In fact

$$\kappa_{C1}(A,\mathcal{D}) = \|\bar{R}^{-1}\mathcal{D}^{-1}\|_2\|R\|_2\kappa_2(\bar{R}) \le \sqrt{2}\kappa_{C2}(A,\mathcal{D}),$$

so little can be gained in accuracy using $\kappa_{C2}(A,\mathcal{D})$ instead of $\kappa_{C1}(A,\mathcal{D})$. Since $\kappa_{C3}(A,\mathcal{D})$ is even weaker still, there is nothing to be gained by using it.

5. Summary and conclusions

Although the Stewart and Sun first order perturbation bound (2) is relevant in the sense that some problems do attain close to the indicated condition, we have shown that it gives a large over-bound for most problems. The more refined bound (24) is usually significantly stronger, and is never weaker, and the resulting condition number $\kappa_C(A)$ more accurately reflects the true sensitivity of the problem. Further, the sizes of our condition numbers depend on any symmetric pivoting used, and numerical results and analyses show that the standard symmetric pivoting strategy leads to a near optimally conditioned factorization for a given A in $PAP^T = R^TR$.

In fact pivoting ensures $|r_{ii}| \ge |r_{ij}|$ for all $j \ge i$, and Higham (1995), Theorem 8.3, then shows

$$\frac{1}{\min_i |r_{ii}|} \le \|R^{-1}\|_2 \le \frac{2^{n-1}}{\min_i |r_{ii}|}.$$

(Higham (1987) also quotes the result with 2^{n-1} replaced by $\sqrt{4^n+6n-1}/3$.) But now if we take $\mathcal{D}_2 = \mathrm{diag}(r_{ii})$, $R = \mathcal{D}_2\bar{R}$, we have $1 = |\bar{r}_{ii}| \geq |\bar{r}_{ij}|$ and

$$\begin{aligned} 1 \leq \quad & \|\bar{R}^{-1}\|_2 && \leq 2^{n-1}, \\ \|\bar{R}\|_2^2 \leq \quad & \|\bar{R}\|_F^2 && \leq n(n+1)/2, \\ & \kappa_2(\bar{R}) && \leq 2^{n-1} \cdot \sqrt{n(n+1)/2}. \end{aligned}$$

It follows from (48) that with pivoting giving $PAP^T = R^TR$, with $R = \mathcal{D}_2\bar{R}$,

$$\begin{aligned} \frac{1}{2}\kappa_2^{1/2}(A) \leq \kappa_C(PAP^T) \leq \kappa_{C1}(PAP^T) \quad & \leq \quad \kappa_{C1}(PAP^T, \mathcal{D}_2) = \kappa_2(R)\kappa_2(\bar{R}) \\ & \leq \quad \kappa_2^{1/2}(A) \cdot 2^{n-1} \cdot \sqrt{n(n+1)/2}, \end{aligned} \tag{53}$$

which also shows

$$\frac{1}{2}\|A^{-1}\|_2^{1/2} \leq \|\hat{W}_R^{-1}\|_2 \leq \|A^{-1}\|_2^{1/2} \cdot 2^{n-1} \cdot \sqrt{n(n+1)/2}. \tag{54}$$

Computational experience shows that with pivoting $\kappa_C(PAP^T)$ is usually quite close to the lower bound, see Tables 1 to 3, but can be significantly larger, see Remark 15. Thus pivoting will probably give a computed R that is about as accurate as can be obtained using the standard Cholesky factorization algorithm.

Because of the difficulty in understanding and computing $\kappa_C(A)$, there was need for, and fortunately we have been able to give, a simpler analysis and bound. Although the new bound (46) is somewhat weaker, it provides a computationally practical and useful estimate $\kappa_{C1}(A, \mathcal{D}_1)$ of $\kappa_C(A)$, and at the same time gives us insight into why the Cholesky factorization is often less sensitive than we thought, and adds to our understanding as to why the standard pivoting usually gives a condition number approaching $\kappa_2^{1/2}(A)$.

Initial investigations suggest $\kappa_{C1}(A, \mathcal{D}_1)/\kappa_C(A)$ can be bounded above by a function of n alone (probably involving something like 2^n), but this needs to be studied further for the general case. When we use pivoting (53) shows $\kappa_{C1}(PAP^T)/\kappa_C(PAP^T)$ and $\kappa_{C1}(PAP^T, \mathcal{D}_2)/\kappa_C(PAP^T)$ have such a bound, and following van der Sluis (1969) we have

$$\kappa_{C1}(PAP^T) \leq \kappa_{C1}(PAP^T, \mathcal{D}_1) \leq n^{1/2}\kappa_{C1}(PAP^T)$$

so that $\kappa_{C1}(PAP^T, \mathcal{D}_1)/\kappa_C(PAP^T)$ has such a bound too. The analysis leading to (53) also suggests we could estimate $\kappa_{C1}(PAP^T, \mathcal{D}_2)$ instead of $\kappa_{C1}(PAP^T, \mathcal{D}_1)$ when we use pivoting, and this was found to be effective.

These ideas can be applied to other problems such as the QR factorization (Chang & Paige (1995)), LU factorization (Stewart (1995b)), and Cholesky downdating problem, the derived condition numbers generally giving tighter bounds than existing results.

Appendix A: Upper triangular solution of $R^TX + X^TR = G$

Given an upper triangular matrix $R \in \mathcal{R}^{n\times n}$ and a symmetric matrix $G \in \mathcal{R}^{n\times n}$, we derive the upper triangular solution X to the equation

$$R^TX + X^TR = G. \tag{1}$$

First let us introduce some notation. For $n \times n$ C, $C = [c_1, \cdots, c_n]$, C_{ij} is the leading $i \times j$ subblock of C, c_{ij} is the ith element of c_j, and $c_j^{(i)}$ is the the vector of the leading i elements of c_j. Also e_j is the jth column of the unit matrix. We can then define

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

With upper triangular X and R, the jth column of the upper triangular part of (1) is

$$R_{jj}^T x_j^{(j)} + X_{jj}^T r_j^{(j)} = g_j^{(j)},$$

so the system can be solved for each $x_j^{(j)}$ in the order j=1, 2,..., n via

$$\begin{aligned} (R_{jj}^T + e_j {r_j^{(j)}}^T) x_j^{(j)} &= g_j^{(j)} - \begin{bmatrix} X_{j,j-1}^T \\ 0 \end{bmatrix} r_j^{(j)} \\ &= g_j^{(j)} - \begin{bmatrix} {r_j^{(1)}}^T x_1^{(1)} \\ {r_j^{(2)}}^T x_2^{(2)} \\ \cdot \\ {r_j^{(j-1)}}^T x_{j-1}^{(j-1)} \\ 0 \end{bmatrix}, \end{aligned}$$

which clearly takes only $O(n^3)$ operations. It is easy to see that this is the solution of the lower triangular system

$$W_R \mathrm{uvec}(X) = \mathrm{uvec}(G), \tag{2}$$

where W_R is

$$\begin{bmatrix}
2r_{11} & & & & & & & & & & & \\
r_{12} & r_{11} & & & & & & & & & & \\
 & 2r_{12} & 2r_{22} & & & & & & & & & \\
r_{13} & & & r_{11} & & & & & & & & \\
 & r_{13} & r_{23} & r_{12} & r_{22} & & & & & & & \\
 & & & 2r_{13} & 2r_{23} & 2r_{33} & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & \\
r_{1n} & & & & & & & r_{11} & & & & \\
 & r_{1n} & r_{2n} & & & & & r_{12} & r_{22} & & & \\
 & & & r_{1n} & r_{2n} & r_{3n} & & r_{13} & r_{23} & r_{33} & & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \\
 & & & & & & & 2r_{1n} & 2r_{2n} & 2r_{3n} & \cdot & 2r_{nn}
\end{bmatrix}.$$

Through a careful examination of W_R and (2) we see the following important and interesting properties. W_R is lower triangular, and is nonsingular if and only if R

is. The column of W_R corresponding to x_{ij} has exactly $n-i+1$ nonzero elements, and these are multiples of the elements in the ith row of R. That is the first row of R appears n times, the second $n-1$ times, and the ith (with $n-i+1$ elements) appears $n-i+1$ times, so W_R has only

$$\sum_{i=1}^{n}(n-i+1)^2 = \sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$$

nonzero elements. A final crucial observation is that the $j\times j$ block on the diagonal of W_R, j=1, 2,..., n, is just

$$\mathrm{diag}(\underbrace{1, 1, \cdots, 1, 2}_{j})R_{jj}^T.$$

In particular, the $n\times n$ bottom right corner of W_R is

$$\mathrm{diag}(\underbrace{1, 1, \cdots, 1, 2}_{n})R^T.$$

Appendix B: Upper triangular solution of $\tilde{R}^T X + X^T R = G$

Let both $R\in\mathcal{R}^{n\times n}$ and $\tilde{R}\in\mathcal{R}^{n\times n}$ be upper triangular, and let $G\in\mathcal{R}^{n\times n}$ be symmetric. We derive the upper triangular solution X to

$$\tilde{R}^T X + X^T R = G. \qquad (1)$$

The approach here is similar to that in Appendix A.

In the notation of Appendix A, the jth column of the upper triangular part of (1) can be written as

$$\tilde{R}_{jj}^T x_j^{(j)} + X_{jj}^T r_j^{(j)} = g_j^{(j)},$$

so the system can be solved for each $x_j^{(j)}$ in the order j=1, 2,..., n via

$$\begin{aligned}(\tilde{R}_{jj}^T + e_j {r_j^{(j)}}^T)x_j^{(j)} &= g_j^{(j)} - \begin{bmatrix} X_{j,j-1}^T \\ 0 \end{bmatrix} r_j^{(j)} \\ &= g_j^{(j)} - \begin{bmatrix} {r_j^{(1)}}^T x_1^{(1)} \\ {r_j^{(2)}}^T x_2^{(2)} \\ \cdot \\ {r_j^{(j-1)}}^T x_{j-1}^{(j-1)} \\ 0 \end{bmatrix},\end{aligned}$$

which clearly takes only $O(n^3)$ operations. This is just the solution of the lower triangular system

$$W_{R,\tilde{R}}\mathrm{uvec}(X) = \mathrm{uvec}(G), \qquad (2)$$

where $W_{R,\tilde{R}}$ is

$$\begin{bmatrix}
r_{11}+\tilde{r}_{11} & & & & & & & \\
r_{12} & \tilde{r}_{11} & & & & & & \\
 & r_{12}+\tilde{r}_{12} & r_{22}+\tilde{r}_{22} & & & & & \\
\cdot & \cdot & \cdot & \cdot & & & & \\
r_{1n} & & & & \tilde{r}_{11} & & & \\
 & r_{1n} & r_{2n} & & \tilde{r}_{12} & \tilde{r}_{22} & & \\
 & & & \cdot & \cdot & \cdot & \cdot & \\
 & & & & r_{1n}+\tilde{r}_{1n} & r_{2n}+\tilde{r}_{2n} & \cdot & r_{nn}+\tilde{r}_{nn}
\end{bmatrix}.$$

$W_{R,\tilde{R}}$ has the same zero/nonzero structure as W_R, and the $j\times j$ block $R_{jj}^T+e_j{r_j^{(j)}}^T$ on the diagonal is replaced by $\tilde{R}_{jj}^T+e_j{r_j^{(j)}}^T$, j=1, 2,..., n.

Acknowledgement

We would like to thank Nick Higham and Ji-guang Sun for their interest and encouragement in this work.

References

CHANG, X. & PAIGE, C.C. 1995 A new perturbation analysis for the QR factorization. In preparation.

CLINE, A. K., CONN, A. R. & VAN LOAN, C. F. 1982 Generalizing the LINPACK condition estimator, in *Numerical Analysis*, ed. HENNART, J. P., Lecture Notes in Mathematics, no. 909, Springer-Verlag, NY.

CLINE, A. K., MOLER, C. B., STEWART, G. W. & WILKINSON, J. H. 1979 An estimate for the condition number of a matrix, *SIAM J. Numer. Anal.* **16**, 368–375.

DIXON, J. D. 1983 Estimating extremal eigenvalues and condition number of matrices, *SIAM J. Numer. Anal.* **20**, 812–814.

HIGHAM, N. J. 1987 A survey of condition number estimation for triangular matrices, *SIAM Rev.* **29**, 575–596

HIGHAM, N. J. 1990 Analysis of the Cholesky decomposition of a semi-definite matrix, in *Reliable Numerical Computation*, ed. COX, M. G. & HAMMARLING, S. J., Oxford University Press, 161–186.

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

KAHAN, W. 1966 Numerical linear algebra, *Can. Math. Bull.* **9**, 757–801.

MEINGUET, J. 1983 Refined Error Analyses of Cholesky Factorization, *SIAM J. Numer. Anal.* **20**, 1243–1250.

STEWART, G. W. 1973 Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Rev.* **15**, 727–764.

STEWART, G. W. 1977 Perturbation bounds for the QR factorization of a matrix, *SIAM J. Numer. Anal.* **14**, 509–518.

STEWART, G. W. 1993 On the perturbation of LU, Cholesky, and QR factorizations, *SIAM J. Matrix Anal. Appl.* **14**, 1141–1146.

STEWART, G. W. 1995a The triangular matrices of Gaussian elimination and related decompositions, Technical Report **CS-TR-3533 UMIACS-TR-95-91**, University of Maryland, Department of Computer Science.

STEWART, G. W. 1995b On the Perturbation of LU and Cholesky Factors, Technical Report **CS-TR-3535 UMIACS-TR-95-93**, University of Maryland, Department of Computer Science.

STEWART, G. W. & SUN, J.-G. 1990 *Matrix Perturbation Theory*, Academic Press, London.

SUN, J.-G. 1991 Perturbation bounds for the Cholesky and QR factorization, *BIT* **31**, 341–352.

SUN, J.-G. 1992a Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations, *Linear Algebra and Appl.* **173**, 77–97.

SUN, J.-G. 1992b Componentwise perturbation bounds for some matrix decompositions, *BIT* **32**, 702–714.

van der SLUIS, A. 1969 Condition numbers and equilibration of matrices, *Numerische Mathematik* **14**, 14–23.

WILKINSON, J. H. & REINSCH, C. eds. 1971 *Handbook for Automatic Computation*, Vol. 2, Linear Algebra, Springer, New York.