Analysis of the Cholesky Decomposition of a Semi-definite Matrix

Higham, Nicholas J.

1990

MIMS EPrint: **2008.56**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Analysis of the Cholesky Decomposition of a Semi-Definite Matrix *

Nicholas J. Higham†

Abstract

Perturbation theory is developed for the Cholesky decomposition of an $n \times n$ symmetric positive semi-definite matrix A of rank r. The matrix $W = A_{11}^{-1}A_{12}$ is found to play a key role in the perturbation bounds, where A_{11} and A_{12} are $r \times r$ and $r \times (n-r)$ submatrices of A respectively.

A backward error analysis is given; it shows that the computed Cholesky factors are the exact ones of a matrix whose distance from A is bounded by $4r(r+1)\big(\|W\|_2+1\big)^2 u\|A\|_2 + O(u^2)$, where u is the unit roundoff. For the complete pivoting strategy it is shown that $\|W\|_2^2 \leq \frac{1}{3}(n-r)(4^r-1)$, and empirical evidence that $\|W\|_2$ is usually small is presented. The overall conclusion is that the Cholesky algorithm with complete pivoting is stable for semi-definite matrices.

Similar perturbation results are derived for the QR decomposition with column pivoting and for the LU decomposition with complete pivoting. The results give new insight into the reliability of these decompositions in rank estimation.

Key words. Cholesky decomposition, positive semi-definite matrix, perturbation theory, backward error analysis, QR decomposition, rank estimation, LINPACK.

AMS subject classifications. Primary 65F30, 65G05.

*This is a reprint of the paper: N. J. Higham. Analysis of the Cholesky decomposition of a semi-definite matrix. In M. G. Cox and S. J. Hammarling, editors, *Reliable Numerical Computation*, pages 161–185. Oxford University Press, 1990.

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov).

1 Introduction

The Cholesky decomposition $A = R^T R$ of a positive definite matrix A, in which R is upper triangular with positive diagonal elements, is a fundamental tool in matrix computations. The standard algorithm for its computation dates from the early part of this century (Dongarra *et al.* 1979, p. 3.16; Householder 1964, p. 208) and it is one of the most numerically stable of all matrix algorithms (Wilkinson 1968, Meinguet 1983, Kielbasinski 1987).

The Cholesky decomposition exists and is unique when A is positive definite (see, e.g., Golub and Van Loan (1983, p. 88)). The questions of existence and uniqueness of a Cholesky decomposition when A is positive *semi*-definite are answered by the following result (Dongarra *et al.* 1979, p. 8.3; Householder 1964, p. 13; Moler and Stewart 1978).

Lemma 1.1. *Let A be positive semi-definite, of rank r.*

(a) There exists at least one upper triangular R with nonnegative diagonal elements such that $A = R^T R$.

(b) There is a permutation Π such that $\Pi^T A \Pi$ has a unique Cholesky decomposition, which takes the form

$$\Pi^T A \Pi = R^T R, \qquad R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}, \tag{1.1}$$

where R_{11} is $r \times r$ upper triangular with positive diagonal elements.

Proof. (a): Let the symmetric positive semi-definite square root X of A have the QR decomposition $X = QR$ with $r_{ii} \geq 0$. Then $A = X^2 = X^T X = R^T Q^T Q R = R^T R$. (b): The algorithm with pivoting described below amounts to a constructive proof. ∎

Note that the factorisation in part (a) is not in general unique. For example,

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \equiv \begin{pmatrix} 0 & 0 \\ \cos\theta & \sin\theta \end{pmatrix} \begin{pmatrix} 0 & \cos\theta \\ 0 & \sin\theta \end{pmatrix}.$$

In several applications it is necessary to compute a decomposition of the form (1.1). One example is in the solution of rank-deficient least squares problems, where "$A = X^T X$" is the matrix of the normal equations (Björck 1987, Dongarra *et al.* 1979, Stewart 1984). Another example occurs in physics in the study of the spectra of molecules with high degrees of symmetry (Fox and Krohn 1977); in this application A is idempotent ($A^2 = A$) and of low rank. A further example is in optimisation problems with matrix semi-definiteness constraints (Fletcher 1985).

Software for computing a decomposition (1.1) is readily available, notably in LINPACK, in the routine SCHDC (Dongarra *et al.* 1979, Ch. 8). However, as pointed out in

Dongarra *et al.* (1979, p. 8.15) there is no published error analysis for the Cholesky decomposition of a semi-definite matrix, except for the case where A is idempotent (Moler and Stewart 1978). Our aim here is to fill this gap in the literature, by examining the mathematical and numerical properties of the Cholesky algorithm applied to semi-definite matrices, with particular reference to SCHDC.

This report is organised as follows. In section 2 we develop perturbation theory for the Schur complements associated with the Cholesky decomposition. A backward error analysis of the Cholesky algorithm is presented in section 3; unlike most backward error analyses this one draws on perturbation theory—that of section 2. The implications of the error analysis are discussed in section 4. In section 5 our perturbation theory is used to derive some new results concerning the reliability of the QR decomposition with column pivoting, and of Gaussian elimination with complete pivoting, for rank estimation. Numerical experiments are presented in section 6.

In the remainder of the introduction we describe the Cholesky algorithm as it is implemented in LINPACK's SCHDC. Note that this is one of several possible implementations of the Cholesky method, all of them equivalent both mathematically and numerically (for other versions see Dongarra *et al.* (1979, Ch. 3), Golub and Van Loan (1983, p. 89)).

The algorithm consists of $r = \mathrm{rank}(A)$ steps, in the kth of which a rank one matrix is subtracted from A so as to introduce zeros into positions $k, \ldots, n$ in the kth row and column. Ignoring pivoting for the moment, at the start of the kth stage we have

$$A^{(k)} = \left(a_{ij}^{(k)}\right) = A - \sum_{i=1}^{k-1} r_i r_i^T = \begin{array}{c} \\ k-1 \\ n-k+1 \end{array} \begin{array}{c} \begin{array}{cc} k-1 & n-k+1 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 0 & A_k \end{pmatrix} \end{array} \tag{1.2}$$

where $r_i^T = (0, \ldots, 0, r_{ii}, \ldots, r_{in})$.

The reduction is carried one stage further by computing

$$\begin{aligned} r_{kk} &= \sqrt{a_{kk}^{(k)}}, \\ r_{kj} &= a_{kj}^{(k)} / r_{kk}, \qquad j = k+1, \ldots, n, \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - r_{ki} r_{kj}, \quad i, j = k+1, \ldots, n. \end{aligned} \tag{1.3}$$

Overall we have,

$$A = \sum_{i=1}^{r} r_i r_i^T = R^T R, \qquad R^T = (r_1, \ldots, r_r).$$

To avoid breakdown when $a_{kk}^{(k)}$ vanishes (or is negative because of rounding errors), pivoting is incorporated into the algorithm as follows. At the start of the kth stage an element $a_{ss}^{(k)} > 0$ $(s \geq k)$ is selected as pivot, and rows and columns k and s of A_k, and

the kth and sth elements of r_i $(i = 1, \ldots, k-1)$, are interchanged. The overall effect is to compute the decomposition (1.1), where the permutation Π takes account of all the interchanges.

The strategy used by SCHDC in its pivoting option is defined by

$$s = \min\{j : a_{jj}^{(k)} = \max_{k \le i \le n} a_{ii}^{(k)}\}.$$

This is equivalent to complete pivoting in Gaussian elimination, since A_k is positive semi-definite so its largest element lies on the diagonal. We note for later reference that this pivoting strategy produces a matrix R that satisfies (Dongarra *et al.* 1979, p. 8.4)

$$r_{kk}^2 \ge \sum_{i=k}^{\min\{j,r\}} r_{ij}^2, \qquad j = k+1, \ldots, n, \quad k = 1, \ldots, r. \tag{1.4}$$

Finally, a word on notation. We will use two matrix norms, the spectral norm

$$\|A\|_2 = \sup_{x \ne 0} \frac{\|Ax\|_2}{\|x\|_2} \qquad \left(\|x\|_2 = (x^T x)^{\frac{1}{2}} \right),$$

and the Frobenius norm

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{trace}(A^T A)}.$$

It will be convenient to denote by $\mathrm{cp}(A) = \Pi^T A \Pi$ the permuted matrix obtained from the Cholesky algorithm with complete pivoting.

2 Perturbation Theory

We begin by analysing the effect on the Cholesky decomposition of perturbations in the data. This perturbation theory will be used in the error analysis of the next section, and in section 5, but it is also of intrinsic interest.

Throughout this section A is assumed to be a positive semi-definite matrix of rank r whose leading principal submatrix of order r is positive definite. For $1 \le k \le r$ we will write

$$A = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix} \end{array} \tag{2.1}$$

and other matrices will be partitioned conformally.

We have the identity

$$A = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} k \\ \begin{pmatrix} R_{11}^T \\ R_{12}^T \end{pmatrix} \end{array} \begin{pmatrix} R_{11}, & R_{12} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & S_k(A) \end{pmatrix}, \tag{2.2}$$

where R_{11} is the Cholesky factor of A_{11}, $R_{12} = R_{11}^{-T} A_{12}$, and

$$S_k(A) = A_{22} - A_{12}^T A_{11}^{-1} A_{12}$$

is the *Schur complement* of A_{11} in A. Note that $S_r(A) \equiv 0$ and so for $k = r$, (2.2) is the (unique) Cholesky decomposition of A. The following lemma shows how $S_k(A)$ changes when A is perturbed.

Lemma 2.1. *Let E be symmetric and assume $\|A_{11}^{-1} E_{11}\|_2 < 1$. Then*

$$S_k(A+E) = S_k(A) + E_{22} - (E_{12}^T W + W^T E_{12}) + W^T E_{11} W + O\big(\|E\|_2^2\big), \tag{2.3}$$

where $W = A_{11}^{-1} A_{12}$. The second order term (which will be required in section 5) takes the form

$$-E_{12}^T A_{11}^{-1} E_{12} + E_{12}^T A_{11}^{-1} E_{11} W + W^T E_{11} A_{11}^{-1} E_{12} - W^T E_{11} A_{11}^{-1} E_{11} W + O\big(\|E_{11}\|_2^3\big).$$

Proof. The condition $\|A_{11}^{-1} E_{11}\|_2 < 1$ ensures that $A_{11} + E_{11}$ is nonsingular and that we can expand

$$(A_{11} + E_{11})^{-1} = A_{11}^{-1} - A_{11}^{-1} E_{11} A_{11}^{-1} + A_{11}^{-1} E_{11} A_{11}^{-1} E_{11} A_{11}^{-1} + O\big(\|E_{11}\|_2^3\big).$$

The result is obtained by substituting this expansion into $S_k(A+E) = (A_{22} + E_{22}) - (A_{12} + E_{12})^T (A_{11} + E_{11})^{-1} (A_{12} + E_{12})$, and collecting terms. ∎

Lemma 2.1 shows that the sensitivity of $S_k(A)$ to perturbations in A is governed by the matrix $W = A_{11}^{-1} A_{12}$. The question arises of whether, for a given A, the potential $\|W\|_2^2$ magnification of E indicated by (2.3) is attainable. For the no-pivoting strategy, $\Pi = I$, the answer is trivially "yes", since we can take $E = \left(\begin{smallmatrix} \gamma I & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with $|\gamma|$ small, to obtain $\|S_k(A+E) - S_k(A)\|_2 = \|W\|_2^2 \|E\|_2 + O\big(\|E\|_2^2\big)$. For complete pivoting, however, the answer is complicated by the possibility that the sequence of pivots will be different for $A + E$ than for A, in which case Lemma 2.1 is not applicable. Fortunately, a mild assumption on A is enough to rule out this technical difficulty, for small $\|E\|_2$. In the next lemma we redefine $A := \mathrm{cp}(A)$ in order to simplify the notation.

Lemma 2.2. *Let $A := \mathrm{cp}(A)$. Suppose that*

$$\big(S_i(A)\big)_{11} > \big(S_i(A)\big)_{jj}, \qquad 2 \le j \le n-i, \quad 0 \le i \le r-1 \tag{2.4}$$

(where $S_0(A) := A$). Then, for sufficiently small $\|E\|_2$, $A + E = \mathrm{cp}(A + E)$. For $E = \left(\begin{smallmatrix} \gamma I & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with $|\gamma|$ sufficiently small,

$$\|S_k\big(\mathrm{cp}(A+E)\big) - S_k(A)\|_2 = \|W\|_2^2 \|E\|_2 + O\big(\|E\|_2^2\big).$$

Proof. Note that since $A = \mathrm{cp}(A)$, (2.4) simply states that there are no ties in the pivoting strategy (since $\big(S_i(A)\big)_{11} \equiv a_{i+1,i+1}^{(i+1)}$ in (1.2)). Applying Lemma 2.1 inductively, since

$$S_i(A+E) = S_i(A) + O\big(\|E\|_2\big),$$

then in view of (2.4), for sufficiently small $\|E\|_2$,

$$\big(S_i(A+E)\big)_{11} > \big(S_i(A+E)\big)_{jj}, \qquad 2 \le j \le n-i, \quad 0 \le i \le r-1.$$

This shows that $A + E = \mathrm{cp}(A + E)$. The last part then follows from Lemma 2.1. ∎

We now examine the quantity $\|W\|_2 = \|A_{11}^{-1} A_{12}\|_2$. In general, this can be arbitrarily large; for example, consider the positive semi-definite matrix

$$A = \begin{pmatrix} \alpha I_{k,k} & I_{k,n-k} \\ I_{n-k,k} & \alpha^{-1} I_{n-k,n-k} \end{pmatrix}$$

for small $\alpha > 0$, where $I_{p,q}$ is the $p \times q$ identity matrix. However, for $A := \mathrm{cp}(A)$, $\|W\|_2$ can be bounded solely in terms of n and k. The essence of the proof, in the next lemma, is that large elements in A_{11}^{-1} are countered by small elements in A_{12}. Hereafter we set $k = r$, the value of interest in the following sections.

Lemma 2.3. *Let $A := \mathrm{cp}(A)$ and set $k = r$. Then*

$$\|A_{11}^{-1} A_{12}\|_{2,\mathrm{F}} \le \sqrt{\frac{1}{3}(n-r)(4^r - 1)}. \tag{2.5}$$

There is a parametrised family of rank-r matrices $A(\theta) = \mathrm{cp}(A(\theta))$, $\theta \in (0, \frac{\pi}{2}]$, for which

$$\|A_{11}(\theta)^{-1} A_{12}(\theta)\|_{2,\mathrm{F}} \to \sqrt{\frac{1}{3}(n-r)(4^r - 1)} \qquad \text{as } \theta \to 0.$$

Proof. From (2.2) we have $W = A_{11}^{-1} A_{12} = R_{11}^{-1} R_{11}^{-T} A_{12} = R_{11}^{-1} R_{12}$, so we may work with R instead of A. Writing $D = \mathrm{diag}(r_{11}, \ldots, r_{rr})$ we have

$$W = R_{11}^{-1} D \cdot D^{-1} R_{12} \equiv T_{11}^{-1} T_{12}$$

where, in view of the inequalities (1.4), the elements of $T_{11} = (t_{ij})$ satisfy

$$t_{ii} = 1, \quad |t_{ij}| \le 1 \quad \text{for } j > i, \qquad i = 1, \ldots, r,$$

and each element of T_{12} is bounded in absolute value by 1. It is easy to show that if $|x_i| \le 1$ for all i, then

$$|T_{11}^{-1}x| \le (2^{r-1}, 2^{r-2}, \ldots, 1)^T =: y,$$

where absolute values and inequalities for vectors or matrices are defined elementwise. It follows that

$$|W| = |T_{11}^{-1}T_{12}| \le ye^T, \qquad e^T = (1, 1, \ldots, 1) \in \mathbb{R}^{n-r}.$$

Hence for the 2– or Frobenius norms, $\|W\| \le \|ye^T\|$. But $\|ye^T\|_2^2 = \|ye^T\|_{\mathrm{F}}^2 = \mathrm{trace}(ey^Tye^t) = (n-r)y^Ty = \frac{1}{3}(n-r)(4^r-1)$, which completes the proof of (2.5).

For the last part, let $A(\theta) = R(\theta)^TR(\theta)$, where

$$R(\theta) = \mathrm{diag}(1, s, \ldots, s^{r-1}) \begin{pmatrix} 1 & -c & -c & \ldots & -c & -c & \ldots & -c \\ & 1 & -c & \ldots & -c & -c & \ldots & -c \\ & & 1 & & \vdots & \vdots & & \vdots \\ & & & \ddots & \vdots & \vdots & & \vdots \\ & & & & 1 & -c & \ldots & -c \end{pmatrix} \in \mathbb{R}^{r\times n}, \qquad (2.6)$$

with $c = \cos\theta, s = \sin\theta$. (This is the $r \times n$ version of a matrix introduced by Kahan (1966); see also Lawson and Hanson (1974, p.31.)) R satisfies the inequalities (1.4) (as equalities) and so $A(\theta) = \mathrm{cp}(A(\theta))$. Some computations analogous to those in the first part show that

$$R_{11}(\theta)^{-1}R_{12}(\theta) = -cze^T, \quad \text{where} \quad z = \big((1+c)^{r-1}, (1+c)^{r-2}, \ldots, 1\big)^T.$$

Thus

$$\|R_{11}(\theta)^{-1}R_{12}(\theta)\|_{2,\mathrm{F}}^2 = c(n-r)\frac{(1+c)^{2r}-1}{(1+c)^2-1} \to \frac{1}{3}(n-r)(4^r-1) \qquad \text{as } \theta \to 0. \qquad (2.7)$$

∎

Example 2.1 We conclude this section with a "worst-case" example for the Cholesky decomposition with complete pivoting. Let $U(\theta) = \mathrm{diag}(r, r-1, \ldots, 1)R(\theta)$, where $R(\theta)$ is given by (2.6), and define the rank-r matrix $C(\theta) = U(\theta)^TU(\theta)$. Then $C(\theta)$ satisfies the conditions of Lemma 2.2. Also,

$$\|W\|_2 = \|C_{11}(\theta)^{-1}C_{12}(\theta)\|_2 = \|U_{11}(\theta)^{-1}U_{12}(\theta)\|_2 = \|R_{11}(\theta)^{-1}R_{12}(\theta)\|_2$$
$$\to \sqrt{\frac{1}{3}(n-r)(4^r-1)} \qquad \text{as } \theta \to 0,$$

from (2.7). Thus, from Lemma 2.2, for $E = \left(\begin{smallmatrix} \gamma I & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with $|\gamma|$ and θ sufficiently small,

$$\|S_r\big(\mathrm{cp}\big(C(\theta)+E\big)\big)\|_2 \approx \frac{1}{3}(n-r)(4^r-1)\|E\|_2.$$

3 Backward Error Analysis

In this section we present a backward error analysis for the Cholesky algorithm. Let A be a symmetric matrix of floating point numbers. Because of potential rounding errors in forming or storing A, it is unrealistic to assume that A is positive semi-definite and singular. Therefore we will write

$$A = \widetilde{A} + \Delta A,$$

where $\widetilde{A}$ is positive semi-definite of rank $r < n$, and ΔA is assumed "small". A natural choice for $\widetilde{A}$ is a nearest positive semi-definite matrix to A (Higham 1986).

The analysis makes no assumptions about the pivoting strategy, but to simplify the notation we will assume that any necessary interchanges are done at the start of the algorithm; thus $A := \Pi^T A \Pi$. For the analysis it is convenient to reorganise the equations (1.3) into the computationally equivalent form

$$r_{kk} = \left(a_{kk} - \sum_{i=1}^{k-1} r_{ik}^2 \right)^{\frac{1}{2}}, \tag{3.1}$$

$$r_{kj} = \left(a_{kj} - \sum_{i=1}^{k-1} r_{ik} r_{ij} \right) / r_{kk}, \qquad j = k+1, \ldots, n. \tag{3.2}$$

To analyse the evaluation of these expressions in floating point arithmetic we will use the following lemma. Here, and throughout, a hat is used to denote computed quantities.

Lemma 3.1. *Let*

$$s = \left(c - \sum_{i=1}^{k-1} a_i b_i \right) / d \tag{3.3}$$

be evaluated in floating point arithmetic. Assume

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \le u, \quad \operatorname{op} = +, -, *, /,$$

where u is the unit roundoff, and assume that $ku < 1$. Then

$$\widehat{s} d + \sum_{i=1}^{k-1} a_i b_i = c + e,$$

where

$$|e| \le \epsilon_k \left(\sum_{i=1}^{k-1} |a_i||b_i| + |\widehat{s}||d| \right), \tag{3.4}$$

and where $\epsilon_k = ku/(1-ku)$. If $d = 1$ then ϵ_k in (3.4) may be replaced by ϵ_{k-1}.

Assume also that

$$fl(\sqrt{x}) = \sqrt{x}(1+\delta), \qquad |\delta| \le u;$$

then, if the division in (3.3) is replaced by a square root,

$$\widehat{s}^2 + \sum_{i=1}^{k-1} a_i b_i = c + e,$$

where

$$|e| \le \epsilon_{k+1}\left(\sum_{i=1}^{k-1} |a_i||b_i| + \widehat{s}^2\right).$$

Proof. Straightforward. See, for example, Stoer and Bulirsch (1980, pp. 25–27). ∎

Applying Lemma 3.1 to (3.1) we obtain

$$\left.\begin{aligned} \sum_{i=1}^{k} \widehat{r}_{ik}^2 &= a_{kk} + e_{kk}, \\ |e_{kk}| &\le \epsilon_{k+1} \sum_{i=1}^{k} \widehat{r}_{ik}^2 \end{aligned}\right\} \quad k = 1, \ldots, r. \tag{3.5}$$

Similarly, for (3.2),

$$\left.\begin{aligned} \sum_{i=1}^{k} \widehat{r}_{ik}\widehat{r}_{ij} &= a_{kj} + e_{kj}, \\ |e_{kj}| &\le \epsilon_k \sum_{i=1}^{k} |\widehat{r}_{ik}||\widehat{r}_{ij}| \end{aligned}\right\} \quad j = k+1, \ldots, n, \quad k = 1, \ldots, r. \tag{3.6}$$

The elements in the Schur complement A_{r+1} (see (1.2)) are given by

$$a_{ij}^{(r+1)} = a_{ij} - \sum_{k=1}^{r} r_{ki} r_{kj}, \qquad i, j = r+1, \ldots, n.$$

Applying Lemma 3.1 to this expression we obtain

$$\left.\begin{aligned} \widehat{a}_{ij}^{(r+1)} + \sum_{k=1}^{r} \widehat{r}_{ki}\widehat{r}_{kj} &= a_{ij} + e_{ij}, \\ |e_{ij}| &\le \epsilon_r\left(\sum_{k=1}^{r} |\widehat{r}_{ki}||\widehat{r}_{kj}| + |\widehat{a}_{ij}^{(r+1)}|\right) \end{aligned}\right\} \quad i, j = r+1, \ldots, n. \tag{3.7}$$

Collecting (3.5–3.7) into one matrix equation we have

$$A - \widehat{R}_r^T \widehat{R}_r = E + \widehat{A}^{(r+1)} \tag{3.8}$$

where

$$\widehat{R}_r = \begin{array}{c} \\ r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \left(\begin{array}{cc} \widehat{R_{11}} & \widehat{R_{12}} \end{array}\right) \end{array},$$

$$\widehat{A}^{(r+1)} = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} 0 & 0 \\ 0 & \widehat{A}_{r+1} \end{pmatrix} \end{array},$$

and

$$|E| \le \epsilon_{r+1}\big(|\widehat{R}_r^T||\widehat{R}_r| + |\widehat{A}^{(r+1)}|\big). \tag{3.9}$$

Now we take norms in (3.9) and use the inequalities $\|B\|_2 \le \||B|\|_2 \le \sqrt{\mathrm{rank}(B)}\|B\|_2$. We obtain

$$\begin{aligned}
\|E\|_2 &\le \epsilon_{r+1}\big(r\|\widehat{R}_r^T\|_2\|\widehat{R}_r\|_2 + \sqrt{n-r}\|\widehat{A}^{(r+1)}\|_2\big) \\
&= \epsilon_{r+1}\big(r\|\widehat{R}_r^T\widehat{R}_r\|_2 + \sqrt{n-r}\|\widehat{A}^{(r+1)}\|_2\big) \\
&= \epsilon_{r+1}\big(r\|A - E - \widehat{A}^{(r+1)}\|_2 + \sqrt{n-r}\|\widehat{A}^{(r+1)}\|_2\big) \\
&\le \epsilon_{r+1}\big(r\|A\|_2 + r\|E\|_2 + n\|\widehat{A}^{(r+1)}\|_2\big),
\end{aligned}$$

which implies

$$\|E\|_2 \le \frac{\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\big(r\|A\|_2 + n\|\widehat{A}^{(r+1)}\|_2\big). \tag{3.10}$$

Our aim is to obtain an a priori bound for $\|A - \widehat{R}_r^T\widehat{R}_r\|_2$. It is clear from (3.8–3.10) that to do this we have only to bound $\|\widehat{A}^{(r+1)}\|_2$. To this end we interpret (3.8) and (3.9) in such a way that the perturbation theory of section 2 may be applied.

Equation (3.8) shows that $\widehat{A}^{(r+1)}$ is the true Schur complement for the matrix

$$A - E = \widetilde{A} + (\Delta A - E) =: \widetilde{A} + F. \tag{3.11}$$

Hence we can apply Lemma 2.1 to $\widetilde{A}$ to deduce that

$$\|\widehat{A}^{(r+1)}\|_2 = \|\widehat{A}_{r+1}\|_2 \le \|F_{22}\|_2 + 2\|F_{12}\|_2\|W\|_2 + \|W\|_2^2\|F_{11}\|_2 + O\big(\|F\|_2^2\big), \tag{3.12}$$

where $W = \widetilde{A}_{11}^{-1}\widetilde{A}_{12}$. We can weaken (3.12) to

$$\|\widehat{A}^{(r+1)}\|_2 \le \|F\|_2\big(\|W\|_2 + 1\big)^2 + O\big(\|F\|_2^2\big).$$

Using $\|F\|_2 \le \|\Delta A\|_2 + \|E\|_2$, substituting from (3.10), and rearranging, we find

$$\|\widehat{A}^{(r+1)}\|_2 \le \Omega\left(\frac{r\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\|A\|_2 + \|\Delta A\|_2\right)\big(\|W\|_2 + 1\big)^2 + O\big(\|F\|_2^2\big), \tag{3.13}$$

where

$$\Omega = \left(1 - \frac{n\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\big(\|W\|_2 + 1\big)^2\right)^{-1}. \tag{3.14}$$

Finally, using (3.8), (3.10) and (3.13), we have certainly

$$\|A - \widehat{R}_r^T\widehat{R}_r\|_2 \le \Omega\left(1 + \frac{n\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\right)\left(\frac{2r\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\|A\|_2 + \|\Delta A\|_2\right)\big(\|W\|_2 + 1\big)^2 + O\big(\|F\|_2^2\big). \tag{3.15}$$

On imposing conditions that ensure the above analysis is valid, we obtain the following backward error analysis result.

Theorem 3.1. *Let $A = \widetilde{A} + \Delta A$ be a symmetric $n \times n$ matrix of floating point numbers, where $\widetilde{A}$ is positive semi-definite of rank $r < n$, and partition $\widetilde{A}$ and ΔA conformally with (2.1) with $k = r$. Assume that*

$$\max\left\{ \frac{\|\Delta A_{11}\|_2}{\|A_{11}\|_2}, \frac{\|\Delta A\|_2}{\|A\|_2} \right\} = \theta u, \quad \text{where } \theta \text{ is a small constant;} \tag{3.16}$$

that A_{11} is positive definite with

$$\max\left\{ 20r^{3/2}u,\ 2\big(\theta u + \frac{r\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\big) \right\} \kappa_2(A_{11}) < 1; \tag{3.17}$$

and that

$$\frac{n\epsilon_{r+1}}{1 - r\epsilon_{r+1}} \big(\|W\|_2 + 1\big)^2 < \frac{1}{2}, \tag{3.18}$$

where $W = \widetilde{A}_{11}^{-1}\widetilde{A}_{12}$ and $\epsilon_{r+1} = (r+1)u/(1-(r+1)u)$. Then in floating point arithmetic with unit roundoff u the Cholesky algorithm applied to A successfully completes r stages, and the computed $r \times n$ Cholesky factor $\widehat{R}_r$ satisfies

$$\|A - \widehat{R}_r^T \widehat{R}_r\|_2 \le 2\big(2r(r+1) + \theta\big)\big(\|W\|_2 + 1\big)^2 u \|A\|_2 + O(u^2). \tag{3.19}$$

Proof. The assumptions are explained as follows. Condition (3.16) enables us to replace $O\big(\|F\|_2^2\big)$, and $\epsilon_{r+1}\|\Delta A\|_2$, by $O(u^2)$ in (3.15). The second condition serves two purposes. First, the definiteness of A_{11}, together with the "$20r^{3/2}u$" part of (3.17), ensures that Cholesky factorisation of A_{11} succeeds (Wilkinson 1968), that is, the Cholesky algorithm applied to A completes r stages without breakdown. As we show next, the second part of (3.17) ensures that Lemma 2.1 is applicable to (3.11), that is, that $\widetilde{A}_{11}$ is positive definite and $\|\widetilde{A}_{11}^{-1}F_{11}\|_2 < 1$. The definiteness of $\widetilde{A}_{11} = A_{11} - \Delta A_{11}$ is immediate since, certainly,

$$\kappa_2(A_{11})\frac{\|\Delta A_{11}\|_2}{\|A_{11}\|_2} \le \kappa_2(A_{11})\theta u < \frac{1}{2}.$$

To show that $\|\widetilde{A}_{11}^{-1}F_{11}\|_2 < 1$ we use the bounds

$$\|\widetilde{A}_{11}^{-1}\|_2 = \|(A_{11} - \Delta A_{11})^{-1}\|_2 \le \frac{\|A_{11}^{-1}\|_2}{1 - \|A_{11}^{-1}\Delta A_{11}\|_2} \le 2\|A_{11}^{-1}\|_2$$

and

$$\|E_{11}\|_2 \le \frac{r\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\|A_{11}\|_2$$

(which is proved in a similar way to (3.10)), obtaining, since $F = \Delta A - E$,

$$\|\widetilde{A}_{11}^{-1}F_{11}\|_2 \le 2\|A_{11}^{-1}\|_2\left(\theta u\|A_{11}\|_2 + \frac{r\epsilon_{r+1}}{1 - r\epsilon_{r+1}}\|A_{11}\|_2\right) < 1$$

by (3.17).

Finally, the bound (3.19) is obtained from (3.15) on using (3.16), (3.18) and $r\epsilon_{r+1}/(1 - r\epsilon_{r+1}) = r(r+1)u + O(u^2)$. ∎

4 Discussion

First, it is important to note that Theorem 3.1 is just about the best result that could have been expected. For the bound (3.19) is essentially the same as the bound obtained on taking norms in Lemma 2.1; in other words (3.19) simply reflects the inherent mathematical sensitivity of $A - R_r^T R_r$ to small perturbations in A.

We turn now to the issue of stability. Ideally, for A as defined in Theorem 3.1, the computed Cholesky factor $\widehat{R}_r$ produced after r stages of the algorithm would satisfy

$$\|A - \widehat{R}_r^T \widehat{R}_r\|_2 \leq \epsilon \|A\|_2,$$

where ϵ is a modest multiple of u. Theorem 3.1 shows that stability "depends" on the size of $\gamma = \|\widetilde{A}_{11}^{-1}\widetilde{A}_{12}\|_2$. Of course, because of the many inequalities used in its derivation we cannot say that the bound (3.19) will always be sharp when γ is large—but the analysis of section 2 shows that there certainly are perturbations E, which, if present in (3.8), would make (3.19) sharp.

If no form of pivoting is used then γ can be arbitrarily large for fixed n (see section 2) and the Cholesky algorithm must in this case be classed as unstable. But for complete pivoting we know from Lemma 2.3 that there holds the upper bound

$$\gamma \leq \sqrt{\frac{1}{3}(n-r)(4^r - 1)}.$$

Thus the Cholesky algorithm with complete pivoting is stable if r is small, but stability cannot be guaranteed, and seems unlikely in practice, if γ (and hence, necessarily, r and n) is large. We investigate the stability empirically in section 6.

Next we consider the implications of our analysis for LINPACK's SCHDC, assuming the use of the complete pivoting option. SCHDC follows the LINPACK philosophy of avoiding machine dependent constants and tests for "small" numbers, and leaving decisions about rank to the user. Consequently, SCHDC proceeds with the Cholesky algorithm until a nonpositive pivot is encountered, that is, up to and including stage $k-1$, where k is the smallest integer for which

$$\widehat{a}_{ii}^{(k)} \leq 0, \quad i = k, \ldots, n. \tag{4.1}$$

Usually, $k > r+1$, due to the effect of rounding errors. A potential danger is that continuing beyond the rth stage will lead to instability, induced by eliminating from indefinite submatrices consisting entirely of roundoff. To investigate this we consider the $(r+1)$st stage of the Cholesky algorithm and we write, using (1.3) and (3.3),

$$\widehat{a}_{ij}^{(r+2)} = \left(\widehat{a}_{ij}^{(r+1)} - \frac{\widehat{a}_{r+1,i}^{(r+1)}\widehat{a}_{r+1,j}^{(r+1)}}{\widehat{a}_{r+1,r+1}^{(r+1)}}(1+\delta_1)(1+\delta_2)\right)(1+\delta_3), \qquad |\delta_i| \leq u.$$

If $\max_{r+1\le i,j\le n} |\widehat{a}_{ij}^{(r+1)}| = c_{r+1}u$ then

$$\begin{aligned}|\widehat{a}_{ij}^{(r+2)}| &\le c_{r+1}u + O(u^2) + \frac{(c_{r+1}u)^2}{\widehat{a}_{r+1,r+1}^{(r+1)}}\big(1+O(u)\big)\\ &= \left(c_{r+1} + \frac{c_{r+1}^2}{d_r}\right)u + O(u^2),\end{aligned}$$

where $\widehat{a}_{r+1,r+1}^{(r+1)} = d_{r+1}u$. Thus

$$\|\widehat{A}_{r+2}\|_2 \le (n-r-1)\left(1+\frac{c_{r+1}}{d_{r+1}}\right)\|\widehat{A}_{r+1}\|_2 + O(u^2)$$

and so the factorisation remains stable provided that c_{r+1}/d_{r+1} is not too large. It does not seem possible to obtain an a priori bound for c_{r+1}/d_{r+1}. We note, however, that any instability that is encountered is confined to the submatrix of the residual consisting of the intersection of rows and columns $r+1,\ldots,n$.

A more sophisticated termination criterion is to stop as soon as

$$\|\widehat{A}_k\| \le \epsilon\|A\| \qquad \text{or} \qquad \widehat{a}_{ii}^{(k)} \le 0, \quad i = k,\ldots,n, \tag{4.2}$$

for some readily computed norm $\|\cdot\|$ and a suitable tolerance ϵ. This criterion terminates as soon as a stable factorisation is achieved, avoiding unnecessary work in eliminating negligible elements in the computed Schur complement $\widehat{A}_k$. Note that $\|\widehat{A}_k\|$ is indeed a reliable order-of-magnitude estimate of the true residual since by (3.8) and (3.10) $A - \widehat{R}_{k-1}^T\widehat{R}_{k-1} = E + \widehat{A}^{(k)}$ with $\|E\| = O(u)\big(\|A\| + \|\widehat{A}^{(k)}\|\big)$.

Another possible stopping criterion is

$$\max_{k\le i\le n} \widehat{a}_{ii}^{(k)} \le \epsilon\widehat{a}_{11}^{(1)}. \tag{4.3}$$

This is related to (4.2) in that if A and $\widehat{A}_k$ are positive semi-definite then $a_{11}^{(1)} = \max_{i,j}|a_{ij}| \approx \|A\|_2$, and similarly $\max_{k\le i\le n}\widehat{a}_{ii}^{(k)} \approx \|\widehat{A}_k\|_2$. Note that (4.3) bounds $\kappa_2(\widehat{R}_{k-1})$, since

$$\kappa_2(\widehat{R}_{k-1}) \le \left|\frac{\widehat{r}_{11}}{\widehat{r}_{k-1,k-1}}\right| = \left(\frac{\widehat{a}_{11}^{(1)}}{\widehat{a}_{k-1,k-1}^{(k-1)}}\right)^{1/2} \le \epsilon^{-1/2}.$$

The effectiveness of these three stopping criteria for obtaining a stable decomposition is investigated empirically in section 6.

5 Rank-Revealing Decompositions

5.1 The Cholesky Decomposition

As mentioned in the introduction, one use of the Cholesky algorithm with complete pivoting is for computing a rank-revealing Cholesky decomposition of a "nearly" positive

semi-definite matrix A. From the results of section 2, however, we know that the algorithm in general is unreliable, since the distance to a rank r matrix may be overestimated by as much as $\frac{1}{3}(n-r)(4^r-1)$ (see Example 2.1).

5.2 The QR Decomposition

Let $B \in \mathbb{R}^{m\times n}$ $(m \geq n)$ have the QR decomposition with column pivoting

$$B\Pi = \underbrace{Q}_{m\times n}\underbrace{R}_{n\times n}, \qquad Q^TQ = I, \quad r_{ii} \geq 0. \tag{5.2.1}$$

Then

$$\Pi^T B^T B \Pi = R^T R \tag{5.2.2}$$

is the Cholesky decomposition of B^TB with complete pivoting. In this section we apply the perturbation theory of section 2 to the Cholesky decomposition (5.2.2), in order to obtain a new perturbation result for the QR decomposition (5.2.1) in the case where B is rank-deficient.

Let $\text{rank}(B) = r < n$, so that

$$R = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \end{array}. \tag{5.2.3}$$

We wish to examine the effect of a perturbation F in B on the (2,2) block of R. Let $G = Q^TF\Pi$, so that

$$(B+F)\Pi = Q(R+G), \tag{5.2.4}$$

and let $B+F$ have the QR decomposition with column pivoting

$$(B+F)\overline{\Pi} = \overline{Q}\,\overline{R}.$$

Our aim is to bound $\|\overline{R}_{22}\|_2$. Define

$$A = \Pi^T B^T B \Pi = R^T R, \tag{5.2.5}$$

$$A + E = \overline{\Pi}^T (B+F)^T (B+F)\overline{\Pi} = \overline{R}^T\overline{R}. \tag{5.2.6}$$

It is easy to see that

$$\overline{R}_{22}^T\overline{R}_{22} = S_r(A+E), \tag{5.2.7}$$

and so our task is to bound $\|S_r(A+E)\|_2$. Assume that A satisfies conditions (2.4); then, for sufficiently small $\|E\|_2$, $\overline{\Pi} = \Pi$, and from (5.2.4–5.2.6),

$$E = R^TG + G^TR + G^TG. \tag{5.2.8}$$

We are now in a position to invoke the perturbation theory of section 2. On applying Lemma 2.1, we find that for the very special E in (5.2.8) the first order perturbation term vanishes, and the second order term is of a simple form.

Lemma 5.1. *Under the above assumptions, if* $\|A_{11}^{-1}E_{11}\|_2 < 1$ *then*

$$S_r(A+E) = -G_{12}^T G_{12} + G_{12}^T G_{11} W + W^T G_{11}^T G_{12} - W^T G_{11}^T G_{11} W + O\big(\|G\|_2^3\big),$$

where $W = A_{11}^{-1}A_{12} = R_{11}^{-1}R_{12}$.

Proof. The result is obtained from Lemma 2.1 on using $S_r(A) = 0$ and substituting for E from (5.2.8). We omit the tedious algebra. ∎

We obtain the following result.

Theorem 5.2. *Let* $B \in \mathbb{R}^{m\times n}$, *where* $m \geq n$ *and* $\operatorname{rank}(B) = r < n$. *Let* B *have the QR decomposition with column pivoting* $B\Pi = QR$, *where* R *is given by (5.2.3), and assume* $A = B^TB$ *satisfies the conditions (2.4). Then for sufficiently small* $\|F\|_2$, $B+F$ *has the QR decomposition with column pivoting* $(B+F)\Pi = \overline{Q}\,\overline{R}$, *and*

$$\frac{\|\overline{R}_{22}\|_2}{\|B\|_2} \leq \frac{\|F\|_2}{\|B\|_2}\big(1 + \|W\|_2\big) + O\left(\frac{\|F\|_2}{\|B\|_2}\right)^2. \tag{5.2.9}$$

Proof. Using Lemma 5.1 together with (5.2.7) and $\|G\|_2 = \|F\|_2$ one obtains

$$\|\overline{R}_{22}\|_2^2 \leq \|F\|_2^2\big(1 + \|W\|_2\big)^2 + O\big(\|F\|_2^3\big),$$

which, on dividing by $\|B\|_2^2$, is equivalent to (5.2.9). ∎

Theorem 5.2 sheds new light on the behaviour of the QR decomposition with column pivoting. For it shows that the quality of $\|\overline{R}_{22}\|_2$ as an estimate of $\|F\|_2$, which itself is an upper bound for the distance $\sigma_{r+1}(B+F)$ from $B+F$ to the rank r matrices, depends on the size of $W = W(B) = R_{11}^{-1}R_{12}$. (Here $\sigma_k(X)$ denotes the kth largest singular value of X.) If we regard $B+F$ as the given matrix, and we choose F so that $B := (B+F)-F$ has rank r with $\|F\|_2 = \sigma_{r+1}(B+F)$, then we obtain a bound similar to the following one, from Lawson and Hanson (1974, Theorem 6.31):

$$|\overline{r}_{r+1,r+1}| \leq \frac{1}{3}\sqrt{4^{r+1} + 6(r+1) - 1}\;\sigma_{r+1}(B+F) \qquad \big((B+F)\Pi = \overline{Q}\,\overline{R}\big). \tag{5.2.10}$$

Our result is stronger in the sense that for a particular matrix $\|W\|_2$ may be much smaller than its upper bound $\|W\|_2 \leq \sqrt{\frac{1}{3}(n-r)(4^r-1)}$ from Lemma 2.3 (see the test results of the next section); on the other hand (5.2.10) has the advantage of holding for all F.

While Theorem 5.2 is important theoretically, we do not feel that it leads to any new practical approaches to the use of the QR decomposition with column pivoting in rank estimation. The main reason for this is the difficulty of verifying that the perturbation F, which in practice must also include the backward error, is "sufficiently small" for one to be able to obtain a strict bound of the form

$$\frac{\|F\|_2}{\|B\|_2} \geq \left(\frac{\theta}{1+\|W\|_2}\right) \frac{\|\overline{R}_{22}\|_2}{\|B\|_2} \tag{5.2.11}$$

(say, for $\theta = 2$). For practical use, what is really required is a rigorous bound of the form (5.2.11), with θ a small constant, valid for all F. Such a bound would combine the best features of (5.2.9) and (5.2.10), but is, we suspect, impossible to achieve.

The results of our numerical testing in the next section do, however, enable us to draw some important conclusions from Theorem 5.2 about the practical effectiveness of the QR decomposition with column pivoting.

5.3 The LU Decomposition

Another interesting application of our perturbation theory is to the LU decomposition of a rank-r $m \times n$ matrix A ($m \geq n$), by Gaussian elimination with complete pivoting:

$$PAQ = \begin{array}{c} \\ r \\ m-r \end{array} \begin{array}{c} r \\ \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} \end{array} \quad \begin{array}{c} r \quad n-r \\ r \begin{pmatrix} U_{11} & U_{12} \end{pmatrix} \end{array}, \qquad l_{ii} \equiv 1.$$

Redefining $A := PAQ$, the relevant Schur complement is $S(A) = A_{22} - A_{21}A_{11}^{-1}A_{12}$, which is zero when A has rank r. A direct analogue of Lemma 2.1 shows that $S_r(A+E)$ contains a term $A_{21}A_{11}^{-1}E_{11}A_{11}^{-1}A_{12}$. The use of complete pivoting implies that

$$\begin{aligned} |u_{ii}| &\geq |u_{ij}|, \quad j \geq i, \\ |l_{ij}| &\leq 1, \quad i > j, \end{aligned}$$

using which it can be shown (cf. Lemma 2.3) that

$$\begin{aligned} \|A_{11}^{-1}A_{12}\|_{2,F} &\leq \sqrt{\frac{1}{3}(n-r)(4^r-1)}, \\ \|A_{21}A_{11}^{-1}\|_{2,F} &\leq \sqrt{\frac{1}{3}(m-r)(4^r-1)} \end{aligned}$$

(with equality for the matrix $A = LU$ where L and U have 1's on the diagonal and -1's everywhere below and above the diagonal respectively). From these results it follows that

although $A+E$ is within distance $\|E\|_2$ of the rank-r matrix A, the Schur complement after r stages of the elimination on $A+E$ is bounded approximately by

$$\|S_r(A+E)\|_2 \leq \frac{\|E\|_2}{3}(4^r-1)\sqrt{(n-r)(m-r)} + O(\|E\|_2^2).$$

(Note that, strictly, we should make an assumption similar to (2.4), and require $\|E\|_2$ to be sufficiently small, so as to ensure that the same pivot sequence is used for A as for $A+E$.) That equality is possible in this bound is shown by Example 2.1, with $A = C(\theta)$ (for which we have $P = Q = \Pi = I$, and $A = R^TR = R^TD^{-1} \cdot DR \equiv LU$ where $D = \mathrm{diag}(r_{ii})$).

It is interesting to note that by taking $r = n-1$ in the above example we obtain an $n \times n$ nonsingular matrix X ($\equiv C(\theta)+E$) for which in the LU factorisation by complete pivoting

$$\min_i |u_{ii}| \approx \left(\frac{4^{n-1}-1}{3}\right)\frac{1}{\|X^{-1}\|_2}$$

(note that $\min\{\|E\|_2 : X+E \text{ is singular}\} = \|X^{-1}\|_2^{-1}$). Thus the often quoted example

$$A = \begin{pmatrix} 1 & -1 & \dots & \dots & 1 \\ & 1 & -1 & \dots & 1 \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad \|A^{-1}\|_\infty = 2^{n-1},$$

which is left unchanged by Gaussian elimination with partial or complete pivoting, so that

$$\min_i |u_{ii}| = \frac{2^{n-1}}{\|A^{-1}\|_\infty},$$

is by no means a worst-case example of the failure of near rank-deficiency to be revealed by small diagonal elements of U! This second example is perhaps "psychologically" worse than the first, however, since the matrix X tends to be very ill-conditioned, so that, unlike for A, $\min_i |u_{ii}|$ for X always reveals some degree of ill-conditioning.

6 Numerical Experiments

We have carried out several numerical experiments in order to investigate the "typical" size in practice of $\|W\|_{\mathrm{F}}$ when pivoting is used in the Cholesky and QR decompositions, and to assess the effectiveness of the stopping criteria (4.1), (4.2) and (4.3) for the Cholesky decomposition.

Our first group of tests was implemented in Fortran 77 on a CDC Cyber machine, with $u = 2^{-48} \approx 3.55 \times 10^{-15}$. We used LINPACK's SCHDC with the complete pivoting

option to compute Cholesky decompositions of various random positive semi-definite matrices with pre-assigned spectra. Each matrix was constructed as $A = V\Lambda V^T$, where $\Lambda = \mathrm{diag}(\lambda_i)$ is rank-r and positive semi-definite, and where V is a random orthogonal matrix (different for each Λ) generated using the method of Stewart (1980). We used three distributions of the nonzero eigenvalues:

$$\begin{aligned} 1 = \lambda_1 = \lambda_2 = \cdots = \lambda_{r-1}, \qquad & \lambda_r = \alpha \le 1, \\ \lambda_1 = 1, \qquad \lambda_2 = \lambda_3 = \cdots = \lambda_r = \alpha \le 1, & \\ \lambda_i = \beta^{i-1}, \quad 1 \le i \le r, \qquad & \beta \le 1, \end{aligned}$$

α and β being used to vary $\kappa_2(A) = \lambda_1(A)/\lambda_r(A)$. For each distribution we generated 100 different matrices by taking all combinations of $n \in \{10, 15, 20, 25, 50\}$, $r \in \{2 + [i(n-3)/3] : i = 0, 1, 2, 3\}$, and $\kappa_2(A) \in \{1, 10^3, 10^6, 10^9, 10^{12}\}$.

Denote a computed Cholesky decomposition from SCHDC by $\Pi^T A \Pi \approx \widehat{R}_k^T \widehat{R}_k$, where $\widehat{R}_k$ is $k \times n$. Write, for $k \ge r$,

$$\widehat{R}_k = \begin{array}{c} \\ r \\ k-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} \widehat{R}_{11} & \widehat{R}_{12} \\ 0 & \widehat{R}_{22} \end{pmatrix} \end{array},$$

so that

$$\widehat{R}_r = (\, \widehat{R}_{11}, \quad \widehat{R}_{12} \,)$$

is the computed $r \times n$ Cholesky factor obtained after r stages of the decomposition. In all our tests $k \ge r$ was satisfied. For each decomposition we computed the following quantities:

$$\begin{aligned} \|W\|_{\mathrm{F}} &= \|\widehat{R}_{11}^{-1} \widehat{R}_{12}\|_{\mathrm{F}}, \\ \rho_k &= \|\Pi^T A \Pi - \widehat{R}_k^T \widehat{R}_k\|_{\mathrm{F}} / \big(u \|A\|_{\mathrm{F}}\big), \\ \rho_r &= \|\Pi^T A \Pi - \widehat{R}_r^T \widehat{R}_r\|_{\mathrm{F}} / \big(u \|A\|_{\mathrm{F}}\big), \\ \alpha_r &= \|\widehat{A}_{r+1}\|_{\mathrm{F}} / \big(u \|A\|_{\mathrm{F}}\big), \\ \beta_r &= \max_{r+1 \le i \le n} \widehat{a}_{ii}^{(r+1,r+1)} / (u \widehat{a}_{11}^{(1)}). \end{aligned}$$

The results were extremely consistent, showing no noticeable variations with n, r, $\kappa_2(A)$, or the eigenvalue distribution. They are summarised in Table 6.1.

The statistics for ρ_r show that throughout these tests the Cholesky algorithm had, after r stages, produced a remarkably stable factorisation. This stability is predicted by the bound (3.19), since $\|W\|_{\mathrm{F}} < 10$ throughout.

SCHDC continued for $k > r$ stages of the Cholesky algorithm in all but 18 of the 300 cases. For example, for the matrices with $n = 50$ and $r = 33$, k varied between 35

and 41. In a small number of cases these extra elimination stages led to instability, the largest relative residual being $\rho_k = 5814$.

The values of α_r, β_r and σ_r, together with values of α_{r-1} and β_{r-1} not reported here, show that for the termination criteria (4.2) (using the Frobenius norm with $\epsilon = 20u$) and (4.3) (with $\epsilon = 50u$), the Cholesky algorithm would in every case have been terminated after r stages, giving a stable decomposition.

A second group of tests was performed using various general, nonsingular matrices, including random matrices of several types, and Hilbert and Vandermonde matrices. For each matrix we computed the QR decomposition with column pivoting, and evaluated $\|W\|_{\mathrm{F}} = \|\widehat{R}_{11}^{-1}\widehat{R}_{12}\|_{\mathrm{F}}$ for each of the $n-2$ partitionings

$$\widehat{R} = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} \widehat{R}_{11} & \widehat{R}_{12} \\ 0 & \widehat{R}_{22} \end{pmatrix} \end{array}, \qquad 2 \le r \le n-1.$$

In this way we effectively generated, from a single QR decomposition, the values $\|W\|_{\mathrm{F}}$ corresponding to the Cholesky or QR decompositions (with pivoting) of a family of matrices, of ranks $2, \ldots, n-1$. The computations were done using MATLAB on a PC-AT compatible machine. The dimension n varied between 5 and 25. The largest value of $\|W\|_{\mathrm{F}}$ was 5.23.

Finally, we present examples of the failure of the Cholesky decomposition with complete pivoting (or equivalently, in these examples, Gaussian elimination with complete pivoting) to provide a rank-revealing decomposition. We used MATLAB to compute Cholesky decompositions of $C := C(\theta)/\|C(\theta)\|_2$, where $C(\theta) = \mathrm{cp}\big(C(\theta)\big)$ is defined in Example 2.1. For $n = 10, 20$ the "worst" results we obtained are as follows. Here $u \approx 2.22 \times 10^{-16}$.

$$\begin{aligned} n = r = 10, \quad \theta = 0.38, \quad |\widehat{r}_{nn}|^2 = 2.0 \times 10^{-11}, \quad \widehat{\sigma}_n(C) = 3.7 \times 10^{-16}, \\ n = r = 20, \quad \theta = 0.81, \quad |\widehat{r}_{nn}|^2 = 1.8 \times 10^{-9}, \quad \widehat{\sigma}_n(C) = 9.9 \times 10^{-18}. \end{aligned}$$

In both cases C is singular to working precision (since $\widehat{\sigma}_n(C) \approx u$) yet this is not revealed by the diagonal elements of $\widehat{R}$. In these examples $\kappa_2(C)u \ge 1$, so strictly the theory of section 2 is not applicable. Nevertheless, the ratio $|\widehat{r}_{nn}|^2/\widehat{\sigma}_n(C)$ is very close to the approximate theoretical maximum $(4^{n-1}-1)/3$ for $n = 10$, and within a factor ≈ 1000 of it for $n = 20$.

Several conclusions may be drawn from these numerical tests. First, it seems that $\|W\|_{\mathrm{F}}$ very rarely exceeds 10 in practice. Nevertheless, it is not difficult to generate matrices for which $\|W\|_{\mathrm{F}}$ is large. To see this, consider the effect of modifying the R

factors generated in the above tests by replacing r_{ij} by $-|r_{ij}|$ for $i \neq j$, and also replacing r_{ii} by $|r_{ii}|$ in the case of the QR decomposition. This modification has no effect on the inequalities (1.4) and so the new matrices R are still genuine triangular factors from the QR or Cholesky decompositions with pivoting. Moreover, it is easy to see that these sign changes do not decrease the value of $\|W\|_{\mathrm{F}}$. In the above tests $\|W\|_{\mathrm{F}}$ for the modified R was frequently bigger than 10, sometimes by several orders of magnitude. One implication of these results is that the "M-matrix sign pattern" occurs very rarely amongst the matrices R obtained in practice.

Another conclusion is that the termination criterion used in SCHDC does occasionally lead to a residual appreciably larger than would be obtained if the Cholesky algorithm were terminated at an earlier stage. Our test results indicate that either of the criteria (4.2) or (4.3) is preferable from the point of view of backward stability. An added benefit of terminating earlier is a reduction in the storage requirements for R.

Based on our numerical experience we suggest that $\epsilon = nu$ is a reasonable tolerance for use with (4.2) or (4.3). Because the term $\|\widehat{A}_k\|$ in (4.2) requires some non-trivial work for its evaluation we favour (4.3).

Certainly, the choice of stopping criterion in the Cholesky decomposition is a delicate matter, the "best" choice depending on many factors, such as the scaling of A, and possible *a priori* knowledge of the rank. We would suggest that in practice it is desirable to "prune" the $k \times n$ $\widehat{R}_k$ returned by SCHDC to a $p \times n$ $\widehat{R}_p$ with $p \leq k$; this could be done by removing each row 'k' of $\widehat{R}_k$ for which (4.3) is satisfied.

7 Concluding Remarks

Our error and perturbation analysis of the Cholesky decomposition of a semi-definite matrix has revealed the key role played by the matrix $W = R_{11}^{-1}R_{12}$, where $R = (R_{11}, R_{12})$ is the Cholesky factor of A. We have shown that in exact arithmetic the residual after an r-stage Cholesky decomposition of A can overestimate the distance of A from the rank-r semi-definite matrices by a factor $\approx \|W\|_2^2$. And we obtained a bound for the backward error in the computed Cholesky decomposition of a semi-definite A that is proportional to $\left(\|W\|_2 + 1\right)^2$. These results hold for any pivoting strategy and so a major objective of a pivoting strategy should be to keep $\|W\|_2$ small.

Our theoretical and numerical results indicate that complete pivoting—choosing as pivot the largest element from along the diagonal—is an excellent strategy. With complete pivoting $\|W\|_2 \leq \sqrt{\frac{1}{3}(n-r)(4^r-1)}$, and it appears that in practice $\|W\|_2$ rarely exceeds 10. Thus our overall conclusion is that for semi-definite matrices the Cholesky

algorithm with complete pivoting must be regarded as a stable algorithm.

A side product of a our analysis is further evidence for the reliability in practice of the QR decomposition with column pivoting as a means for computing a rank-revealing decomposition. Theorem 5.2, combined with the empirical observation that $\|W\|_2$ is usually small, leads to the conclusion that if $B\Pi = QR$ and B is close to a rank r matrix, then "nearly always" $R_{22} \in \mathbb{R}^{n-r \times n-r}$ will be appropriately small.

Acknowledgements

This work was begun during a visit to the Computer Science Department at Stanford University. I thank Gene Golub for financial support. It is a pleasure also to thank Ian Gladwell for stimulating discussions on this work, and Des Higham and Len Freeman for useful suggestions for improving the manuscript.

Table 6.1 Results for 300 test matrices.

Condition	Number of Cases
$1 \le \|W\|_{\mathrm{F}} < 9.7$	throughout
$1 \le \rho_k < 10$	192
$10 \le \rho_k < 100$	99
$100 \le \rho_k < 1000$	5
$1000 \le \rho_k < 10,000$	4
$1 \le \rho_r < 10$	275
$10 \le \rho_r < 20$	25
$\max \alpha_r = 16.7$	
$\beta_r < 10$	265
$10 \le \beta_r < 50$	35

REFERENCES

Å. Björck, Least Squares Methods, in *Handbook of Numerical Analysis, Volume 1: Solution of Equations in* $\mathbb{R}^n$, P.G. Ciarlet and J.L. Lions, eds., Elsevier/North Holland, 1987.

J.J. Dongarra, J.R. Bunch, C.B. Moler and G.W. Stewart, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

R. Fletcher, Expected conditioning, IMA Journal of Numerical Analysis, 5 (1985), pp. 247–273.

K. Fox and B.J. Krohn, Computation of cubic harmonics, J. Comput. Phys., 25 (1977), pp. 386–408.

G.H. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1983.

N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and Appl., 103:103–118, 1988.

A.S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

W. Kahan, Numerical linear algebra, Canadian Math. Bulletin, 9 (1966), pp. 757–801.

A. Kielbasinski, A note on rounding-error analysis of Cholesky factorization, Linear Algebra and Appl., 88/89 (1987), pp. 487–494.

C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

J. Meinguet, Refined error analyses of Cholesky factorization, SIAM J. Numer. Anal., 20 (1983), pp. 1243–1250.

C.B. Moler and G.W. Stewart, On the Householder-Fox algorithm for decomposing a projection, J. Comput. Phys., 28 (1978), pp. 82–91.

G.W. Stewart, The efficient generation of random orthogonal matrices with an application to condition estimators, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

G.W. Stewart, Rank degeneracy, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 403–413.

J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

J.H. Wilkinson, A priori error analysis of algebraic processes, in *Proc. International Congress of Mathematicians, Moscow 1966*, I.G. Petrovsky, ed., Mir Publishers, Moscow, 1968, pp. 629–640.