INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Perturbation Analysis of Some Matrix Factorizations

by

Xiao-Wen Chang

School of Computer Science

McGill University

Montréal, Québec

Canada

February 1997

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

OF MCGILL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

Abstract

Matrix factorizations are among the most important and basic tools in numerical linear algebra. Perturbation analyses of matrix factorizations are not only important in their own right, but also useful in many applications, e.g. in estimation, control and statistics. The aim of such analyses is to show what effects changes in the data will have on the factors. This thesis is concerned with developing new general purpose perturbation analyses, and applying them to the Cholesky, QR and LU factorizations, and the Cholesky downdating problem.

We develop a new approach, the so called 'matrix-vector equation' approach, to obtain sharp results and true condition numbers for the above problems. Our perturbation bounds give significant improvements on previous results, and could not be sharper. Also we use the so called 'matrix equation' approach originated by G. W. Stewart to derive perturbation bounds that are usually weaker but easier to interpret. This approach allows efficient computation of satisfactory estimates for the true condition numbers derived by our approach. The combination of these two approaches gives a powerful understanding of these problems. Although first-order perturbation bounds are satisfactory for all but the most delicate work, we also give some rigorous perturbation bounds for some factorizations.

We show that the condition of many such factorizations is significantly improved by the standard pivoting strategies (except the L factor in the LU factorization), and provide firmly based theoretical explanations as to why this is so. This extremely important information is very useful for designing more reliable matrix algorithms.

Our approach is a powerful general tool, and appears to be applicable to the perturbation analysis of any matrix factorization.

Résumé

Les factorisations de matrices sont parmi les outils les plus importants et les plus fondamentaux de l'algèbre linéaire numérique. Les analyses de perturbation des factorisations de matrices sont non seulement importantes en elles-mêmes, mais ils sont aussi utiles dans maintes applications, par exemple dans les domaines de l'estimation, du contrôle et des statistiques. Ces analyses ont pour but de démontrer quels effets les changements dans les données produiront sur les facteurs. Cette thèse s'intéresse au développement de nouvelles analyses genénérales des perturbations, et à leur application aux factorisations de Cholesky, QR et LU, et au problème de modification de factorisation de Cholesky.

Nous développons une nouvelle approche que nous nommons l'approche équation matrice-vecteur, afin d'obtenir des résultats précis et des vrais nombres de conditionnement pour les problèmes mentionnés ci-dessus. Nos bornes sur les perturbations apportent des améliorations significatives aux résultats antérieurs et ne pourraient être plus précises. De plus, nous utilisons l'approche équation matrice développée par G. W. Stewart pour dériver des bornes sur les perturbations qui sont généralement plus faibles mais plus faciles à interpréter. Cette approche permet des calculs efficaces d'estimés satisfaisants pour les vrais nombres de conditionnement qui dérivent de notre approche. La combinaison de ces deux approches donne une compréhension profonde de ces problèmes. Bien que des bornes sur les perturbations de premier ordre soient satisfaisantes pour tout travail, sauf pour le plus délicat, nous donnons également des bornes rigoureuses sur les perturbations pour certa ines factorisations.

Nous démontrons que la condition de plusieurs de ces factorisations est améliorée de façon significative par les stratégies usuelles de pivotage (sauf en ce qui concerne le facteur L dans la factorisation LU), et nous fournissons des explications théoriques solidement fondées pour démontrer pourquoi il en est ainsi. Cette information extrêmement importante est des plus utiles pour construire des algorithmes

plus sûrs pour les matrices.

Notre approche est un outil général puissant, et semble pouvoir s'appliquer aux analyses de perturbation de n'importe quelle factorisation de matrice.

Acknowledgements

I would like to extend my sincere thanks and deep gratitude to my supervisor, Professor Chris Paige, for his invaluable guidance, support, encouragement and care through the years of Ph.D studies. He displayed extreme generosity with his ideas and time, and generated a continuous enthusiasm for this work. He made all his research facilities available to me. My association with him is, for me, a rare privilege. I hope he will find this thesis to be a small reward for the many days of inspired collaboration as well as for his efforts to ensure that I received the best possible training. I also wish to thank Dr. Françoise Paige for translating the abstract of this thesis into French.

I thank Professor G. W. Stewart for his insightful ideas, which provided one of the approaches used in this thesis. My thanks to Professor Ji-guang Sun for his suggestions and encouragement, and for providing me draft versions of his work. Their pioneering work was a major inspiration for this work.

I am thankful to Professor Jim Varah, external reviewer of this thesis, for his careful reading. This thesis has been improved by his helpful comments and suggestions.

I am indebted to Professors Gene Golub, Nick Higham, and George Styan for their interest and encouragement in this work.

I am grateful to Professor Jia-Song Wang, my Master's Thesis supervisor, for introducing me to numerical analysis and continuous encouragement.

The School of Computer Science at McGill University provided me with a comfortable environment for my thesis work. Many people including my fellow graduate students gave me various help. My heartfelt thanks go to Penny Anderson, Franca Cianci, Naixun Pei, Josée Turgeon, Josie Vallelonga, Xiaoyan Zhao and Binghai Zhu.

My deepest gratitude goes to my parents, brothers and sister. Their love and care have always been a constant source of my enthusiasm to complete this work. I deeply regret that my father, who died during my Ph.D. studies, cannot see the completion of this thesis. Also my thanks to my uncle and aunt for their strong support.

Finally I would like to give special thanks to my wife, Su, whose love, patience and understanding made it possible for me to finally finish my work on this thesis.

To my wife and my parents

Contents

List of Tables

Chapter 1

Introduction and Preliminaries

1.1 Introduction

This thesis is concerned with the perturbation analysis for the Cholesky, QR, and LU factorizations, and for the Cholesky downdating problem. These matrix factorizations are among the most fundamental and important tools in numerical linear algebra (see for example Golub and Van Loan [26, 1996]). The goal of such an analysis is to determine bounds for the changes in the factors of a matrix when the matrix is perturbed.

We first give some motivation for our concerns. Suppose A is a given matrix, and has a factorization

$$A = BC, \tag{1.1.1}$$

where B and C are the factors of A. As in any topic of matrix perturbation theory, there are three main considerations in perturbation theory for matrix factorizations. First, the elements of A may be determined from physical measurement, and therefore be subject to errors of observation. The true matrix is $A + \Delta A$, where ΔA is the observation error. Suppose the same factorization for $A + \Delta A$ is

$$A + \Delta A = (B + \Delta B)(C + \Delta C). \tag{1.1.2}$$

We are thus led immediately to the consideration of the perturbations ΔB and ΔC. Second, even if the elements of A can be defined exactly by mathematical formulae, usually these can not efficiently be represented exactly by a digital computer due to its finite precision. The matrix stored in a computer is $A + \Delta A$, with $|\Delta A| \leq u|A|$, where u is the unit roundoff. So we are faced with much the same problem as before. Finally, *backward rounding error analysis* throws back errors made in executing an algorithm on the original data, see Wilkinson [52, 1963]. Suppose for a stored matrix A, a backward rounding error analysis shows the computed factors $\tilde{B}$ and $\tilde{C}$ of A are the exact factors of $A + \Delta A$, i.e.,

$$A + \Delta A = \tilde{B}\tilde{C},$$

where a bound on $\|\Delta A\|$ (or $|\Delta A|$) is known, then perturbation theory is used to assess the effects of these backward errors on the accuracy of the computed factors, i.e., give bounds on $\|\tilde{B} - B\|$ and $\|\tilde{C} - C\|$ (or $|\tilde{B} - B|$ and $|\tilde{C} - C|$).

Although in solving linear equations the sensitivity of factors may not be of central interest, it is important when the factors have significance. For example in the estimation problem with $m \times n$ A of full column rank and m dimensional y given (where $\mathcal{E}(\cdot)$ indicates the expected value),

$$y = Ax + v, \qquad \mathcal{E}(v) = 0, \qquad \mathcal{E}(vv^T) = \sigma^2 I,$$

if we obtain the QR factorization $A = QR$ then solving $R\hat{x} = Q^T y$ gives the best linear unbiased estimate (BLUE) $\hat{x}$ of x, and

$$R\mathcal{E}\{(\hat{x} - x)(\hat{x} - x)^T\}R^T = \sigma^2 I,$$

so R is the factor of what has been called in the engineering literature the 'information matrix'. This is important in its own right (see for example Paige [34, 1985]), and we are interested in how changes in A affect R.

In general we regularly use the fact that the columns of Q in the QR factorization of A form an orthonormal basis for $\mathcal{R}(A)$, and we are concerned with how changes in A affect Q.

In some statistical applications, if certain matrices A and B have QR factorizations

$$A = Q_A R_A, \qquad B = Q_B R_B,$$

then the singular values of $Q_A^T Q_B$ give what are called the 'canonical correlations' (more generally these give the angles between the subspaces $\mathcal{R}(A)$ and $\mathcal{R}(B)$), see for example Björck and Golub [4, 1973]. Thus the sensitivity of Q in the QR factorization can be used directly to answer the following important problems: "How do changes in A and B affect $\mathcal{R}(A)$ and $\mathcal{R}(B)$ and the angles between these (or the canonical correlations)".

We thus see the area is an interesting and useful one to study in general. This area has been an active area of research in recent years. Most of the existing results have been incorporated in Higham [30, 1996].

Realizing most of the published results on the sensitivity of factorizations, such as LU, Cholesky, and QR, were extremely weak for certain classes of matrices, Chang, under the supervision of Chris Paige, see the commentary in Chang, Paige and Stewart [14, 1996], originated an approach to obtaining provably sharp results and corresponding condition numbers for the Cholesky factorization. He also realized that the condition of the problem was significantly improved by pivoting, and provided the first firmly based theoretical explanations as to why this was so. Even though the original work was only about the Cholesky factorization, the approach is a general approach, and thus can be applied to almost all well-known matrix factorizations. From (1.1.1) and (1.1.2) we have by dropping the second-order term that

$$\Delta A \approx B\Delta C + \Delta B C, \tag{1.1.3}$$

The basic idea of this approach is to write the approximate matrix equation (1.1.3) as a matrix-vector equation by using the special structure and properties of B and

C, then get the vector-type expressions for ΔB and ΔC. So we will call this the 'matrix-vector equation' approach.

Stewart [44, 1995] was stimulated by Chang's work on the Cholesky factorization to understand this more deeply, and present simple explanations for what was going on. Before Chang's work, the most used approach to perturbation analyses of factorizations was what we will call the 'matrix equation' approach, which keeps equations like (1.1.3) in their matrix-matrix form. Stewart [44] (also see Chang, Paige and Stewart [13, 1996]) used an elegant construct, partly illustrated by the 'up' and 'low' notation in Section 1.2, which makes the matrix equation approach a far more usable and intuitive tool. He combined this with deep insights on scaling to produce the new *matrix equation analysis* which is appealingly clear, and provides excellent insights into the sensitivities of the LU and Cholesky factorizations. This new matrix equation analysis does not in general provide tight results like the matrix-vector equation analyses do, but they are usually more simple, and provide practical estimates for the true condition numbers obtained from the latter. This approach is also fairly general, but for each factorization a particular treatment is needed. This is different from the matrix-vector equation approach, which can be applied to any factorization directly without any difficulty.

We combined these two approaches to give a deep understanding of the sensitivity of the Cholesky factorization, see Chang, Paige and Stewart [13, 1996]. We also applied the two approaches to the QR factorization and the Cholesky downdating problem, see Chang, Paige and Stewart [14, 1996], and, Chang and Paige [10, 1996]. The interplay of the two approaches goes through the whole thesis.

The main purpose of this thesis is to establish *first-order* perturbation bounds that are as tight as possible for the factorizations mentioned above, present the corresponding condition numbers, give some condition estimators, and shed light on the effect of the standard pivoting on the conditioning. Although first-order perturbation bounds are satisfactory for all but the most delicate work, we also give some rigorous

perturbation bounds for some factorizations. Some results in this thesis have been presented in the papers mentioned above. Some other new results here have not yet been published.

1.2 Notation and basics

First we describe some mostly standard notation, and define some elementary concepts used throughout this thesis.

- $\mathbf{R}^{m\times n}$ denotes the vector space of all $m \times n$ real matrices, and $\mathbf{R}^n = \mathbf{R}^{n\times 1}$.
- A matrix is always denoted by a capital letter, e.g. A. The corresponding lowercase letter with the subscript j and ij refers to the the jth column and (i,j)th entry respectively, e.g. a_j, a_{ij}. Also the notation $(A)_{ij}$ designates the (i,j)th entry. $A(i,:)$ denotes the ith row of A and $A(:,j)$ the jth column.
- A vector is represented by a lowercase letter, e.g. b. The individual components are denoted with single subscripts, e.g. b_i.
- $\mathcal{R}(A)$ denotes the space spanned by the columns of A.
- $\lambda(A)$ denotes an eigenvalue of a matrix A; $\rho(A)$ denotes the spectral radius of A, i.e. $\rho(A) = \max|\lambda(A)|$.
- $\sigma(A)$ denotes a nonzero singular value of a matrix A; $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the largest and smallest nonzero singular values of A, respectively.
- Let $A = (a_{ij})$ be an $m \times n$ matrix, then $|A|$ is defined by $|A| = (|a_{ij}|)$.
- Let t be a scalar and let $A(t) = (a_{ij}(t))$ be an $m \times n$ matrix. If $a_{ij}(t)$ is a differentiable function of t for all i and j, then we say $A(t)$ is differentiable with

respect to t and define

$$\dot{A}(t) = \frac{d}{dt}A(t) = \left(\frac{d}{dt}a_{ij}(t)\right) = (\dot{a}_{ij}(t)).$$

- $\|\cdot\|$ denotes a vector norm or matrix norm.
- $\|\cdot\|$ is a monotone and consistent matrix norm if $|A| \leq |B|$ implies $\|A\| \leq \|B\|$, and $\|AB\| \leq \|A\|\|B\|$.
- The 1-norm, 2-norm (or Euclidean norm), and ∞-norm of an n-dimension vector x are defined respectively by

$$\|x\|_1 = \sum_{i=1}^{n}|x_i|, \qquad \|x\|_2 = (\sum_{i=1}^{n}|x_i|^2)^{1/2}, \qquad \|x\|_\infty = \max_{1\leq i\leq n}|x_i|.$$

- The S-norm (S for summation), F-norm (or Frobenius norm), and M-norm (M for maximum) of an $m \times n$ matrix A are defined respectively by

$$\|A\|_S = \sum_{i,j}|a_{ij}|, \qquad \|A\|_F = (\sum_{i,j}|a_{ij}|^2)^{1/2}, \qquad \|A\|_M = \max_{i,j}|a_{ij}|.$$

- The 1-norm, 2-norm (or spectral norm), and ∞-norm of an $m \times n$ matrix A defined by

$$\|A\|_p = \sup_{x\neq 0}(\|Ax\|_p/\|x\|_p), \quad p = 1, 2, \infty$$

are given respectively by

$$\|A\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|, \quad \|A\|_2 = \sigma_{\max}(A), \quad \|A\|_\infty = \max_{1\leq i\leq m}\sum_{j=1}^{n}|a_{ij}|.$$

- $\kappa_\nu(A) = \|A^\dagger\|_\nu \|A\|_\nu$ denotes the standard condition number of matrix A and $\mathrm{cond}_\nu(A) = \| |A^\dagger||A| \|_\nu$ the Bauer-Skeel condition number of matrix A when norm $\|\cdot\|_\nu$ is used, where $A^\dagger$ is the Moore-Penrose generalized inverse of A. If A is nonsingular, then $\kappa_\nu(A) = \|A^{-1}\|_\nu\|A\|_\nu$ and $\mathrm{cond}_\nu(A) = \| |A^{-1}||A| \|_\nu$.

- Let $C = [c_1, c_2, \cdots, c_n]$ be an $m \times n$ matrix, then $\mathrm{vec}(C)$ is defined by

$$\mathrm{vec}(C) = \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ c_n \end{bmatrix}.$$

Now we describe some special notation used throughout this thesis.

- $\mathbf{D}_n$ always denotes the set of all $n \times n$ real positive definite diagonal matrices.

- Let $X = (x_{ij})$ be an $n \times n$ matrix. The upper triangular part, strictly lower triangular part and strictly upper triangular part of X are denoted respectively by

$$\mathrm{ut}(X) = \begin{bmatrix} x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & x_{nn} \end{bmatrix}, \quad \mathrm{slt}(X) = X - \mathrm{ut}(X), \quad \mathrm{sut}(X) = \mathrm{slt}(X^T)^T, \tag{1.2.1}$$

and the diagonal of X is denoted by

$$\mathrm{diag}(X) = \mathrm{diag}(x_{11}, x_{22}, \ldots, x_{nn}). \tag{1.2.2}$$

- For any $n \times n$ matrix $X = (x_{ij})$, we define the upper and lower triangular matrices

$$\mathrm{up}(X) = \begin{bmatrix} \frac{1}{2}x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}x_{nn} \end{bmatrix}, \quad \mathrm{low}(X) = \mathrm{up}(X^T)^T = X - \mathrm{up}(X), \tag{1.2.3}$$

- For any $n \times n$ matrix $C = [c_1, \ldots, c_n]$, denote by $c_j^{(i)}$ the vector of the first i elements of c_j, and by $c_j^{(\underline{i})}$ the vector of the last i elements of c_j. With these, we define ('u' denotes 'upper', 'sl' denotes 'strictly lower')

$$\mathrm{uvec}(C) = \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ \cdot \\ c_n^{(n)} \end{bmatrix}, \qquad \mathrm{slvec}(C) = \begin{bmatrix} c_1^{(\underline{n-1})} \\ c_2^{(\underline{n-2})} \\ \cdot \\ c_{n-1}^{(\underline{1})} \end{bmatrix}. \tag{1.2.4}$$

  They are the vectors formed by stacking the columns of the upper triangular part of C into one long vector and by stacking the columns of the strictly lower triangular part of C into one long vector, respectively.

The 'low' and 'up' have the following basic properties. For general $X \in \mathbf{R}^{n\times n}$

$$\max\{\|\,\mathrm{low}(X)\|_F, \|\,\mathrm{up}(X)\|_F\} \le \|X\|_F, \tag{1.2.5}$$

$$\|X - \mathrm{up}(X + X^T)\|_F = \|\,\mathrm{low}(X) - [\mathrm{low}(X)]^T\|_F \le \sqrt{2}\,\|X\|_F. \tag{1.2.6}$$

For symmetric $X \in \mathbf{R}^{n\times n}$

$$2\,\|\,\mathrm{up}(X)\|_F^2 = 2\,\|\,\mathrm{low}(X)\|_F^2 = \|X\|_F^2 - \frac{1}{2}(x_{11}^2 + x_{22}^2 + \cdots + x_{nn}^2) \le \|X\|_F^2. \tag{1.2.7}$$

The following well-known theorem obtained by van der Sluis [51, 1969] will often be referred to when we discuss the effect of scaling on the condition estimators in this thesis.

Theorem 1.2.1 *Let $S, T \in \mathbf{R}^{n\times n}$ and let S be nonsingular, and define*

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

Then

$$\|\,|T||S|\,\|_\infty = \|TD_{r1}\|_\infty\,\|D_{r1}^{-1}S\|_\infty = \min_{D\in\mathbf{D}_n}\|TD\|_\infty\,\|D^{-1}S\|_\infty, \tag{1.2.8}$$

$$\|\,|S||T|\,\|_1 = \|SD_{c1}^{-1}\|_1\,\|D_{c1}T\|_1 = \min_{D\in\mathbf{D}_n}\|SD^{-1}\|_1\,\|DT\|_1, \tag{1.2.9}$$

$$\|TD_{r2}\|_2\,\|D_{r2}^{-1}S\|_2 \le \sqrt{n}\,\inf_{D\in\mathbf{D}_n}\|TD\|_2\,\|D^{-1}S\|_2, \tag{1.2.10}$$

$$\|SD_{c2}^{-1}\|_2\,\|D_{c2}T\|_2 \le \sqrt{n}\,\inf_{D\in\mathbf{D}_n}\|SD^{-1}\|_2\,\|DT\|_2. \tag{1.2.11}$$

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

$$\mathrm{cond}_\infty(S) = \kappa_\infty(D_{r1}^{-1}S) = \min_{D\in\mathbf{D}_n} \kappa_\infty(D^{-1}S), \tag{1.2.12}$$

$$\mathrm{cond}_1(S^{-1}) = \kappa_1(SD_{c1}^{-1}) = \min_{D\in\mathbf{D}_n} \kappa_1(SD^{-1}), \tag{1.2.13}$$

$$\kappa_2(D_{r2}^{-1}S) \le \sqrt{n}\, \inf_{D\in\mathbf{D}_n} \kappa_2(D^{-1}S), \tag{1.2.14}$$

$$\kappa_2(SD_{c2}^{-1}) \le \sqrt{n}\, \inf_{D\in\mathbf{D}_n} \kappa_2(SD^{-1}). \tag{1.2.15}$$

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

$$\kappa_2(D_*^{-1}SD_*^{-1}) \le n \inf_{D\in\mathbf{D}_n} \kappa_2(D^{-1}SD^{-1}). \qquad \square \tag{1.2.16}$$

We will often use the following results when discussing the effect of standard pivoting on the condition numbers.

Theorem 1.2.2 *Let* $T \in \mathbf{R}^{n\times n}$ *be a nonsingular upper triangular matrix satisfying*

$$|t_{ii}| \ge |t_{ij}| \quad \textit{for all } j > i. \tag{1.2.17}$$

Then with $D \equiv \mathrm{diag}(T)$,

$$\kappa_F(D^{-1}T) \equiv \|(D^{-1}T)^{-1}\|_F \, \|D^{-1}T\|_F \le \sqrt{2n(n+1)(4^n+6n-1)}/6, \tag{1.2.18}$$

$$\kappa_{1,\infty}(D^{-1}T) \equiv \|(D^{-1}T)^{-1}\|_{1,\infty} \, \|D^{-1}T\|_{1,\infty} \le n2^{n-1}, \tag{1.2.19}$$

$$\mathrm{cond}_F(T) \equiv \| \, |T^{-1}||T| \, \|_F \le \sqrt{4^{n+1}-3n-4}/3, \tag{1.2.20}$$

$$\mathrm{cond}_{1,\infty}(T) \equiv \| \, |T^{-1}||T| \, \|_{1,\infty} \le 2^n - 1. \tag{1.2.21}$$

All of the upper bounds above can be reached for the $n \times n$ *matrix*

$$T = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ & 1 & -1 & \cdots & -1 \\ & & \ddots & \vdots & \vdots \\ & & & 1 & -1 \\ & & & & 1 \end{bmatrix}. \tag{1.2.22}$$

Proof. Let $\bar{T} \equiv D^{-1}T$. Then we have $1 = |\bar{t}_{ii}| \geq |\bar{t}_{ij}|$. It is easy to show that

$$|(\bar{T}^{-1})_{ij}| \leq 2^{j-i-1} \quad \text{for } j > i. \tag{1.2.23}$$

Thus

$$\|\bar{T}\|_F^2 \leq n(n+1)/2, \qquad \|\bar{T}\|_{1,\infty} \leq n,$$

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

$$\|\bar{T}^{-1}\|_1 \leq \sum_{i=1}^{n-1} 2^{n-i-1} + 1 = 2^{n-1}, \quad \|\bar{T}^{-1}\|_\infty \leq 1 + \sum_{j=2}^{n} 2^{j-2} = 2^{n-1}.$$

From these (1.2.18) and (1.2.19) follow.

By (1.2.23) we have for $j > i$,

$$(|T^{-1}||T|)_{ij} = (|\bar{T}^{-1}||\bar{T}|)_{ij} = \sum_{k=i}^{j} |(\bar{T}^{-1})_{ik}||\bar{t}_{kj}| \leq 1 + \sum_{k=i+1}^{j} 2^{k-i-1} \cdot 1 = 2^{j-i}.$$

Thus

$$\| \, |T^{-1}||T| \, \|_F^2 \leq \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2^{j-i})^2 + \sum_{i=1}^{n} 1 = (4^{n+1} - 3n - 4)/9,$$

$$\| \, |T^{-1}||T| \, \|_1 \leq \sum_{i=1}^{n} 2^{n-i} = 2^n - 1, \quad \| \, |T^{-1}||T| \, \|_\infty \leq \sum_{j=1}^{n} 2^{n-j} = 2^n - 1,$$

which give (1.2.20) and (1.2.21).

If T has the form of (1.2.22), then we easily verify $T^{-1} \doteq \begin{bmatrix} 1 & 1 & 2 & \cdots & 2^{n-2} \\ & 1 & 1 & \cdots & 2^{n-3} \\ & & \ddots & \vdots & \vdots \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix}$,

i.e., $|(T^{-1})_{ij}| = 2^{j-i-1}$ for $j > i$. It is easy to see from the foregoing proof that all of the inequalities in (1.2.18), (1.2.19), (1.2.20) and (1.2.21) become equalities with $D = I$. □

Chapter 2

The Cholesky Factorization

2.1 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

$$A = R^T R.$$

R is called the Cholesky factor.

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

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

for $j = 1 : n$
  for $i = 1 : j - 1$
    $r_{ij} = (a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}) / r_{ii}$
  end
  $r_{jj} = (a_{jj} - \sum_{k=1}^{j-1} r_{kj}^2)^{1/2}$
end

Let $\Delta A \in \mathbf{R}^{n\times n}$ be a symmetric matrix such that $A + \Delta A$ is still symmetric positive definite, then $A + \Delta A$ has the unique Cholesky factorization

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

The goal of the perturbation analysis for the Cholesky factorization is to determine a bound on $\|\Delta R\|$ (or $|\Delta R|$) in terms of (a bound on) $\|\Delta A\|$ (or $|\Delta A|$).

The rest of this chapter is organized as follows. In Section 2.2 we consider the case where only a bound on $\|\Delta A\|$ is known. We refer to these as 'norm-bounded changes in A'. First-order and rigorous tight bounds are presented by the so called matrix-vector equation approach, and somewhat weaker but more insightful and computationally applicable bounds are also given by the so called matrix equation approach. In Section 2.3 we make a similar analysis to that in Section 2.2 for the case where a bound on $|\Delta A|$ is known. We refer to these as 'component-bounded changes in A'. In both of the sections, we derive useful upper bounds on the condition of the problem when we use pivoting, and give numerical results to confirm our theoretical analyses. Finally we summarize our findings and point out future work in Section 2.4.

This Cholesky analyses (and particularly Section 2.2.1) may also be taken as an introduction to the general approach to perturbation analysis of factorizations proposed by this thesis.

2.2 Perturbation analysis with norm-bounded changes in A

There have been several papers dealing with the perturbation analysis for the Cholesky factorization with norm-bounded changes in A. The first result was that of Stewart [39, 1977]. It was further modified and improved by Sun [46, 1991], who included a *first-order* perturbation result. Using a different approach, Stewart [41, 1993] obtained the same first-order perturbation result. Recently Drmač, Omladič

and Veselić [20, 1994] presented perturbation results of a different flavor. They made a perturbation analysis for the Cholesky factorization of $H = D_c^{-1}AD_c^{-1}$ with $D_c = \mathrm{diag}(a_{ii}^{1/2})$, instead of A. The advantage of their approach is that a good bound can be obtained when ΔA corresponds to backward rounding errors. So their result will be referred to in the next section.

The main goal of this section is to establish new first-order bounds on the norm of the perturbation in the Cholesky factor, smaller than those of Sun [46, 1991] and Stewart [41, 1993], and present a condition number which more closely reflects the true sensitivity of the problem. Also, we give *rigorous* perturbation bounds. Many of the results have been presented in Chang, Paige and Stewart [13, 1996].

2.2.1 First-order perturbation bounds

We first obtain an equation and an expression for $\dot{R}(0)$ in the Cholesky factorization $A + tG = R^T(t)R(t)$, then we use these to obtain our new first-order perturbation bounds by the matrix-vector equation approach and the matrix equation approach. The first approach will provide a sharp bound, resulting in the condition number for the Cholesky factorization with norm-bounded changes in A, while the second approach provides results that are usually weaker but easier to interpret, and allows efficient computation of satisfactory estimates for the actual condition number.

Rate of change of R

Here we derive the basic results on how R changes as A changes. We then derived Sun's [46, 1991] results. The following theorem summarizes the results we use later.

Theorem 2.2.1 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^TR$, let $G \in \mathbf{R}^{n\times n}$ be symmetric, and let $\Delta A = \epsilon G$, for some $\epsilon \geq 0$. If*

$$\rho(A^{-1}\Delta A) < 1, \tag{2.2.1}$$

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

$$A + \Delta A = (R + \Delta R)^T (R + \Delta R), \tag{2.2.2}$$

with ΔR satisfying

$$\Delta R = \epsilon \dot{R}(0) + O(\epsilon^2), \tag{2.2.3}$$

where $\dot{R}(0)$ is defined by the unique Cholesky factorization

$$A + tG = R^T(t)R(t), \qquad |t| \leq \epsilon, \tag{2.2.4}$$

and so satisfies the equations

$$R^T \dot{R}(0) + \dot{R}^T(0)R = G, \tag{2.2.5}$$

$$\dot{R}(0) = \mathrm{up}(R^{-T}GR^{-1})R, \tag{2.2.6}$$

where the 'up' notation is defined by (1.2.3).

Proof. If (2.2.1) holds, then for all $|t| \leq \epsilon$ the spectral radius of $tR^{-T}GR^{-1}$ satisfies

$$\rho(tR^{-T}GR^{-1}) = \rho(tR^{-1}R^{-T}G) = \rho(tA^{-1}G) < 1.$$

Therefore for all $|t| \leq \epsilon$, $A + tG = R^T(I + tR^{-T}GR^{-1})R$ is symmetric positive definite and so has the unique Cholesky factorization (2.2.4). Notice that $R(0) = R$ and $R(\epsilon) = R + \Delta R$, so (2.2.2) holds.

It is easy to verify that $R(t)$ is twice continuously differentiable for $|t| \leq \epsilon$ from the algorithm for the Cholesky factorization. If we differentiate (2.2.4) and set $t = 0$ in the result, we obtain (2.2.5) which we will see is a linear equation *uniquely* defining the elements of upper triangular $\dot{R}(0)$ in terms of the elements of G. From upper triangular $\dot{R}(0)R^{-1}$ in

$$(\dot{R}(0)R^{-1})^T + \dot{R}(0)R^{-1} = R^{-T}GR^{-1},$$

we see with the 'up' notation in (1.2.3) that (2.2.6) holds. Finally the Taylor expansion for $R(t)$ about $t = 0$ gives (2.2.3) at $t = \epsilon$. □

Using Theorem 2.2.1 we can now easily obtain the first-order perturbation bound due to Sun [46, 1991], and also proved by Stewart [41, 1993] by a different approach.

Theorem 2.2.2 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^TR$, and let ΔA be a real symmetric $n \times n$ matrix satisfying $\|\Delta A\|_F \le \epsilon \|A\|_2$. If*

$$\kappa_2(A)\epsilon < 1, \tag{2.2.7}$$

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

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

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{1}{\sqrt{2}}\kappa_2(A)\epsilon + O(\epsilon^2). \tag{2.2.8}$$

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

$$\|G\|_F \le \|A\|_2. \tag{2.2.9}$$

Since

$$\rho(A^{-1}\Delta A) \le \|A^{-1}\Delta A\|_2 \le \kappa_2(A)\epsilon,$$

the assumption (2.2.7) implies that (2.2.1) holds. So the conclusion of Theorem 2.2.1 holds here. By using the fact that $\|\operatorname{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F$ for any symmetric X (see (1.2.7)), we have from (2.2.6) that

$$\|\dot{R}(0)\|_F \le \frac{1}{\sqrt{2}}\|R^{-T}GR^{-1}\|_F\|R\|_2 \le \frac{1}{\sqrt{2}}\|R^{-1}\|_2^2\|R\|_2\|G\|_F, \tag{2.2.10}$$

which, with (2.2.9) and $\|R^{-1}\|_2^2 = \|A^{-1}\|_2$, gives

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \frac{1}{\sqrt{2}}\kappa_2(A). \tag{2.2.11}$$

Then (2.2.8) follows immediately from the Taylor expansion (2.2.3). □

Clearly from (2.2.8) we see $\frac{1}{\sqrt{2}}\kappa_2(A)$ can be regarded as a measure of the sensitivity of the Cholesky factorization. Since a condition number as a function of a matrix of a certain class has to be from a bound which is attainable to first-order for any matrix in the given class (see (2.2.19) for a more formal definition of the condition number of the Cholesky factorization with norm-bounded changes in A) we will use this rigorous terminology, and use a qualified term *condition estimator* when this criterion is not met. For general A the first-order bound in (2.2.8) is not attainable, in other words, we are not always able to choose a symmetric ΔA satisfying $\|\Delta A\|_F \leq \epsilon \|A\|_2$ to make (2.2.11) an equality. We could use a simple example to illustrate this, but we choose not to do so here, as it will be obvious after we obtain the actual condition number. Therefore we say $\frac{1}{\sqrt{2}}\kappa_2(A)$ is a *condition estimator* for the Cholesky factorization.

We have seen the basis for deriving first-order perturbation bounds for R is the equation (2.2.5) (or the expression (2.2.6) of its solution), which will be used later. Our following analyses will be based on the same assumptions as in Theorem 2.2.2.

Matrix-vector equation analysis

Now we would like to derive an attainable first-order perturbation bound.

The upper and lower triangular parts of the matrix equation (2.2.5) contain identical information. The upper triangular part can be rewritten in the following form by using the 'uvec' notation in (1.2.4) (for the derivation, see the Appendix of Chang, Paige and Stewart [13, 1996]):

$$W_R \,\mathrm{uvec}(\dot{R}(0)) = \mathrm{uvec}(G), \tag{2.2.12}$$

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

$$\begin{bmatrix}
2r_{11} & & & & & & & & & & \\
r_{12} & r_{11} & & & & & & & & & \\
 & 2r_{12} & 2r_{22} & & & & & & & & \\
r_{13} & & & r_{11} & & & & & & & \\
 & r_{13} & r_{23} & r_{12} & r_{22} & & & & & & \\
 & & & 2r_{13} & 2r_{23} & 2r_{33} & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
r_{1n} & & & & & & \cdot & r_{11} & & & \\
 & r_{1n} & r_{2n} & & & & & r_{12} & r_{22} & & \\
 & & & r_{1n} & r_{2n} & r_{3n} & & r_{13} & r_{23} & r_{33} & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \\
 & & & & & & & 2r_{1n} & 2r_{2n} & 2r_{3n} \cdot & 2r_{nn}
\end{bmatrix}. \tag{2.2.13}$$

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

$$\mathrm{duvec}(C) \equiv \mathcal{D}_F \, \mathrm{uvec}(C), \tag{2.2.14}$$

where

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

Thus for any *symmetric* matrix G we have $\|\mathrm{duvec}(G)\|_2 = \|G\|_F$. For our norm-based analysis, we rewrite (2.2.12) as

$$\widehat{W}_R \, \mathrm{uvec}(\dot{R}(0)) = \mathrm{duvec}(G), \qquad \widehat{W}_R \equiv \mathcal{D}_F W_R. \tag{2.2.15}$$

Since R is nonsingular, $\widehat{W}_R$ is also, and from (2.2.15)

$$\mathrm{uvec}(\dot{R}(0)) = \widehat{W}_R^{-1} \, \mathrm{duvec}(G). \tag{2.2.16}$$

so taking the 2-norm and using $\|G\|_F \leq \|A\|_2$ from (2.2.9), we obtain

$$\begin{aligned} \|\dot{R}(0)\|_F &= \|\widehat{W}_R^{-1}\,\mathrm{duvec}(G)\|_2 \\ &\leq \|\widehat{W}_R^{-1}\|_2\,\|\mathrm{duvec}(G)\|_F \\ &= \|\widehat{W}_R^{-1}\|_2\,\|G\|_F \\ &\leq \|\widehat{W}_R^{-1}\|_2\,\|A\|_2, \end{aligned} \tag{2.2.17}$$

where for *any* nonsingular upper triangular R, equalities can be obtained by choosing G such that duvec(G) lies in the space spanned by the right singular vectors corresponding to the largest singular value of $\widehat{W}_R^{-1}$ and $\|G\|_F = \|A\|_2$. Using the Taylor expansion (2.2.3) and $\|A\|_2 = \|R\|_2^2$, we see

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \|\widehat{W}_R^{-1}\|_2\,\|A\|_2^{1/2}\epsilon + O(\epsilon^2), \tag{2.2.18}$$

and this bound is attainable to first-order in ϵ. Thus for the Cholesky factorization with norm-bounded changes in A the condition number (with respect to the combination of the F- and 2-norms)

$$\kappa_C(A) \equiv \lim_{\epsilon\to 0}\sup\left\{\frac{\|\Delta R\|_F}{\epsilon\,\|R\|_2} : A+\Delta A = (R+\Delta R)^T(R+\Delta R),\ \|\Delta A\|_F \leq \epsilon\,\|A\|_2\right\} \tag{2.2.19}$$

is given by

$$\kappa_C(A) = \|\widehat{W}_R^{-1}\|_2\,\|A\|_2^{1/2}. \tag{2.2.20}$$

Obviously with the definition of $\kappa_C(A)$ we have from (2.2.8) that

$$\kappa_C(A) \leq \frac{1}{\sqrt{2}}\,\kappa_2(A). \tag{2.2.21}$$

This upper bound on $\kappa_C(A)$ is achieved if R is an $n \times n$ identity matrix with $n \geq 2$, and so is tight.

We now derive a lower bound on $\kappa_C(A)$. Observe that the $n \times n$ bottom right hand corner of W_R is just $\mathrm{diag}(1,1,\ldots,1,2)R^T$, so that $\widehat{W}_R$ has the form

$$\widehat{W}_R = \begin{bmatrix} \times & 0 \\ \times & \hat{D}R^T \end{bmatrix}, \tag{2.2.22}$$

where

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

Therefore we have

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

It follows that

$$\|\widehat{W}_R^{-1}\|_2 \geq \|R^{-T}\hat{D}^{-1}\|_2 \geq \frac{1}{2}\|R^{-1}\|_2, \tag{2.2.23}$$

thus

$$\kappa_C(A) = \|\widehat{W}_R^{-1}\|_2 \|A\|_2^{1/2} \geq \frac{1}{2}\|R^{-1}\|_2 \|R\|_2 = \frac{1}{2}\kappa_2^{1/2}(A). \tag{2.2.24}$$

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

We summarize these results as the following theorem.

Theorem 2.2.3 *With the same assumptions as in Theorem* 2.2.2, $A + \Delta A$ *has the unique Cholesky factorization*

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

such that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_C(A)\epsilon + O(\epsilon^2), \tag{2.2.25}$$

$$\frac{1}{2}\kappa_2^{1/2}(A) \leq \kappa_C(A) \leq \frac{1}{\sqrt{2}}\kappa_2(A), \tag{2.2.26}$$

where $\kappa_C(A) = \|\widehat{W}_R^{-1}\|_2 \|A\|_2^{1/2}$, *and the first-order bound in* (2.2.25) *is attainable.* □

From (2.2.26) we know the new first-order bound in (2.2.25) is at least as good as that in (2.2.8), but it suggests the former may be considerably smaller than the latter. Consider the following example.

Example 1: Let $A = \mathrm{diag}(1, \delta^2)$. Then $R = \mathrm{diag}(1, \delta)$, $\widehat{W}_R = \mathrm{diag}(2, \sqrt{2}, 2\delta)$. When $0 < \delta \leq 1/\sqrt{2}$, we obtain

$$\kappa_C(A) = \frac{1}{2\delta}, \qquad \kappa_2(A) = \frac{1}{\delta^2}. \qquad \square$$

We see that the first-order perturbation bound (2.2.8) can severely overestimate the effect of a perturbation in A.

But it is possible that $\kappa_C(A)$ has the same order as $\kappa_2(A)$, as we now show.

Example 2: If $A = \begin{bmatrix} \delta^2 & \delta \\ \delta & 2 \end{bmatrix}$ with small $\delta > 0$, then $R = \begin{bmatrix} \delta & 1 \\ 0 & 1 \end{bmatrix}$. Some simple computations give

$$\kappa_C(A) = O(\frac{1}{\delta^2}), \qquad \kappa_2(A) = O(\frac{1}{\delta^2}). \qquad \square$$

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 2.2.3 we have

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

and

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

Note that the first-order bound in (2.2.8) does not change when the Cholesky factorization of A is approached by using any pivoting strategy. Clearly the perturbation bound (2.2.25) 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

$\kappa_2^{1/2}(A)\sqrt{n(n+1)(4^n+6n-1)}/6$, see (2.2.39). Using the standard symmetric pivoting strategy in Example 2 gives

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}}\delta \\ 0 & \frac{1}{\sqrt{2}}\delta \end{bmatrix},$$

and $\kappa_C(PAP^T) = O(1/\delta)$, showing how pivoting can improve the condition of the problem (as measured by our condition number) by an arbitrary amount.

There have been several techniques for estimating the 2-norm of the inverse of a triangular matrix, e.g., Cline et al [16, 1979], Cline et al [15, 1982] and Dixon [18, 1983]. A comprehensive, comparative survey on these techniques has been given by Higham [27, 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 $O(n^3)$. 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 the following matrix equation analysis.

Matrix equation analysis

As we saw, $\kappa_C(A)$ is unreasonably expensive to compute or estimate directly with the usual approach, except when we use pivoting, in which case $\kappa_C(PAP^T)$ usually approaches its lower bound $\kappa_2^{1/2}(A)/2$, see (2.2.39). Fortunately, the approach of Stewart [44, 1995] can be extended to obtain an excellent upper bound on $\kappa_C(A)$, 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.

In Theorem 2.2.2 we used the expression of $\dot{R}(0)$ in (2.2.6) to derive Sun's first-order perturbation bound. Now we again look at (2.2.6), repeated here for clarity.

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

A useful observation is that for any matrix $B \in \mathbf{R}^{n\times n}$ and diagonal matrix $D \in \mathbf{R}^{n\times n}$,

$$\mathrm{up}(B)D = \mathrm{up}(BD).$$

Let $\mathbf{D}_n$ be the set of all $n \times n$ real positive definite diagonal matrices. For any $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathbf{D}_n$ we can take $R = D\bar{R}$, giving

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

which leads to cancellation of the D^{-1} with D:

$$\dot{R}(0) = \mathrm{up}(R^{-T}G\bar{R}^{-1})\bar{R}, \tag{2.2.27}$$

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

$$\|\dot{R}(0)\|_F \le \|\mathrm{up}(R^{-T}G\bar{R}^{-1})\|_F \, \|\bar{R}\|_2 \le \|R^{-1}\|_2 \, \|G\|_F \, \kappa_2(\bar{R}). \tag{2.2.28}$$

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

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \kappa_2(R)\kappa_2(\bar{R}) = \kappa_2(R)\kappa_2(D^{-1}R) \equiv \kappa_C'(A, D), \qquad \text{say.} \tag{2.2.29}$$

Since this is true for all $D \in \mathbf{D}_n$, we may choose D to minimize $\kappa_C'(A, D)$,

$$\kappa_C'(A) \equiv \inf_{D\in\mathcal{D}_n} \kappa_C'(A, D), \tag{2.2.30}$$

which gives the encouraging result

$$\kappa_C'(A) \le \kappa_C'(A, I) = \kappa_2^2(R) = \kappa_2(A). \tag{2.2.31}$$

Then from the Taylor expansion (2.2.3) we have

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_C'(A)\epsilon + O(\epsilon^2). \tag{2.2.32}$$

(2.2.31) shows $1/\sqrt{2}$ times the first-order bound in (2.2.32) is at least as good as that in (2.2.8), the first-order perturbation bound in Sun [46, 1991] and Stewart [41, 1993].

With (2.2.19), the definition of $\kappa_C(A)$, we see from (2.2.32) that $\kappa_C(A) \leq \kappa'_C(A)$. But it is useful to prove this more delicately, and so obtain some indication of how weak $\kappa'_C(A)$ is as an approximation to $\kappa_C(A)$. We know G can be chosen to make (2.2.17) an equality, so that for such a G we have

$$\|\dot{R}(0)\|_F = \|\widehat{W}_R^{-1}\|_2 \|G\|_F = \|\operatorname{up}(R^{-T}GR^{-1})R\|_F,$$

Thus for any $D \in \mathbf{D}_n$ in $R = D\bar{R}$

$$\begin{aligned} \|\widehat{W}_R^{-1}\|_2 \|G\|_F &\leq \|\operatorname{up}(R^{-T}G\bar{R}^{-1})\|_F \|\bar{R}\|_2 && (2.2.33) \\ &\leq \kappa_2(\bar{R})\|R^{-1}\|_2 \|G\|_F, && (2.2.34) \end{aligned}$$

or

$$\|\widehat{W}_R^{-1}\|_2 \leq \kappa_2(\bar{R})\|R^{-1}\|_2, \tag{2.2.35}$$

which implies

$$\kappa_C(A) = \|\widehat{W}_R^{-1}\|_2 \|R\|_2 \leq \kappa_2(\bar{R})\kappa_2(R) \equiv \kappa'_C(A, D), \tag{2.2.36}$$

for any $D \in \mathbf{D}_n$. Note the two inequalities (2.2.33) and (2.2.34) in going from $\kappa_C(A)$ to $\kappa'_C(A, D)$.

Now we summarize the above results as the following theorem.

Theorem 2.2.4 *With the same the assumptions as in Theorem* 2.2.2, $A + \Delta A$ *has the Cholesky factorization*

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

such that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa'_C(A)\epsilon + O(\epsilon^2). \tag{2.2.37}$$

$$\kappa_C(A) \leq \kappa'_C(A) \leq \kappa_2(A) \tag{2.2.38}$$

where $\kappa'_C(A)$ *is as in* (2.2.29) *and* (2.2.30). □

This matrix equation 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 D in $R = D\bar{R}$ can give $\kappa_2(\bar{R})$ very near one, so $\kappa_C'(A, D)$ will approach twice the lower bound $\kappa_2^{1/2}(A)/2$ on $\kappa_C(A)$. As an illustration suppose $R = \begin{bmatrix} 1 & \delta \\ 0 & \delta \end{bmatrix}$, then for small $\delta > 0$, $\kappa_2(R) = O(1/\delta)$. But if we set $D = \mathrm{diag}(1, \delta)$, then $\kappa_2(D) = 1/\delta$ and $\kappa_2(\bar{R}) = O(1)$, so that $\kappa_C'(A, D)$ 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 symmetric 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 standard symmetric pivoting, the ill-conditioning of A will usually reveal itself in the diagonal elements of R. Stewart [43, 1995] 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 D. This implies that $\kappa_C'(PAP^T, D)$, and therefore (as we shall show) $\kappa_C(PAP^T)$, will approach its lower bound. We support this mathematically in the following.

Theorem 2.2.5 *Let $A \in \mathrm{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $PAP^T = R^TR$ when the standard symmetric pivoting strategy is used. Then*

$$\frac{1}{2}\kappa_2^{1/2}(A) \le \kappa_C(PAP^T) \le \kappa_C'(PAP^T) \le \kappa_2^{1/2}(A)\sqrt{2n(n+1)(4^n+6n-1)}/6, \tag{2.2.39}$$

and from this

$$\frac{1}{2}\|A^{-1}\|_2^{1/2} \le \|\widehat{W}_R^{-1}\|_2 \le \|A^{-1}\|_2^{1/2}\sqrt{2n(n+1)(4^n+6n-1)}/6. \tag{2.2.40}$$

Proof. We need only prove the last inequality of (2.2.39). In fact standard pivoting ensures $r_{ii}^2 \geq \sum_{k=i}^{n} r_{ij}^2$ for all $j \geq i$, so $|r_{ii}| \geq |r_{ij}|$. Since for any $D \in \mathbf{D}_n$,

$$\kappa_C'(PAP^T) \leq \kappa_C'(PAP^T, D) \equiv \kappa_2(R)\kappa_2(D^{-1}R) \leq \kappa_2(R)\kappa_F(D^{-1}R),$$

the inequality follows immediately from $\kappa_2(R) = \kappa_2^{1/2}(A)$ and (1.2.18) in Theorem 1.2.2 with $D = \mathrm{diag}(R)$. □

One may not be impressed by the 4^n factor in the upper bound in (2.2.39), and may wonder if the upper bound can significantly be improved. In fact the upper bound can nearly be approximated by a parametrized family of matrices $A = K_n^T(\theta)K_n(\theta)$, where

$$K_n(\theta) = \mathrm{diag}(1, s, \cdots, s^{n-1}) \begin{bmatrix} 1 & -c & -c & \cdot & -c \\ & 1 & -c & \cdot & -c \\ & & 1 & \cdot & -c \\ & & & \cdot & \cdot \\ & & & & 1 \end{bmatrix} \tag{2.2.41}$$

with $c = \cos(\theta)$ and $s = \sin(\theta)$, were introduced by Kahan [32, 1966]. Notice here the permutation P corresponding to the standard symmetric pivoting strategy is the identity. Taking $D = \mathrm{diag}(K_n(\theta)) = \mathrm{diag}(1, s, \cdots, s^{n-1})$, then by the last part of Theorem 1.2.2 we have

$$\kappa_F(D^{-1}K_n(\theta)) \to \sqrt{2n(n+1)(4^n + 6n - 1)}/6 \qquad \text{as } \theta \to 0. \tag{2.2.42}$$

Let $D_r = \mathrm{diag}(\|K_n(\theta)(i,:)\|_2)$, then

$$D_r = D\,\mathrm{diag}(\sqrt{1 + (n-1)c^2}, \cdots, \sqrt{1 + c^2}, 1).$$

Hence

$$\begin{aligned} \kappa_F(D^{-1}K_n(\theta)) &= \kappa_F(D^{-1}D_rD_r^{-1}K_n(\theta)) \leq \kappa_F(D^{-1}D_r)\,\kappa_2(D_r^{-1}K_n(\theta)) \\ &\leq \sqrt{n} \cdot \sqrt{n(n+1)/2} \cdot \kappa_2(D_r^{-1}K_n(\theta)), \end{aligned}$$

or

$$\kappa_2(D_r^{-1}K_n(\theta)) \geq \frac{1}{\sqrt{n}\cdot\sqrt{n(n+1)/2}}\kappa_F(D^{-1}K_n(\theta)).$$

But by (1.2.14) in van der Sluis's Theorem 1.2.1 we have

$$\kappa_c'(PAP^T) \geq \frac{1}{\sqrt{n}}\kappa_2^{1/2}(A)\,\kappa_2(D_r^{-1}K_n(\theta)).$$

Thus

$$\kappa_c'(PAP^T) \geq \frac{1}{n\sqrt{n(n+1)/2}}\kappa_2^{1/2}(A)\,\kappa_F(D^{-1}K_n(\theta)).$$

Then it follows from (2.2.42) that as $\theta \to 0$,

$$\kappa_c'(PAP^T) \gtrsim \kappa_2^{1/2}(A)\frac{\sqrt{4^n+6n-1}}{6n}.$$

This indicates the upper bound in (2.2.39) is nearly tight.

Many computational experiments show with standard symmetric pivoting that $\kappa_c(PAP^T)$ is usually quite close to the lower bound of $\kappa_2^{1/2}(A)$, see Section 2.2.3 Tables 2.2.1 and 2.2.2, but can significantly larger for the matrices whose Cholesky factors are Kahan matrices, see Section 2.2.3 Table 2.2.3 as well as the comments. In the latter case, something like a rank-revealing pivoting strategy such as that in Hong and Pan [31, 1992]) will most likely be required to make the condition number close to its lower bound.

The practical outcome of this simple analysis is that we now have an $O(n^2)$ condition estimator for the Cholesky factor. By (1.2.14) in van der Sluis's Theorem 1.2.1, $\kappa_2(\bar{R})$ will be nearly optimal when the rows of $\bar{R}$ are equilibrated in the 2-norm. Thus the estimation procedure for the condition of the Cholesky factorization is to choose $D = D_r = \mathrm{diag}(\|R(i,:)\|_2)$ in $R = D\bar{R}$, and use a standard condition estimator (for matrix inversion) to estimate $\kappa_2(\bar{R})$ and $\kappa_2(R)$ in (2.2.29).

Finally we give a new perturbation bound which does not involve any scaling matrix D, by using a monotone and consistent matrix norm $\|\cdot\|$ (see Section 1.2).

Theorem 2.2.6 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^T R$, and let $\Delta A \in \mathbf{R}^{n\times n}$ be a symmetric matrix satisfying $\|\Delta A\| \leq \epsilon \|A\|$ for some monotone and consistent matrix norm. If $\kappa(A)\epsilon < 1$, then*

$$\frac{\|\Delta R\|}{\|R\|} \leq \kappa(R^T)\mathrm{cond}(R)\epsilon + O(\epsilon^2). \tag{2.2.43}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). Since

$$\rho(A^{-1}\Delta A) \leq \|A^{-1}\Delta A\| \leq \kappa(A)\epsilon < 1,$$

Theorem 2.2.1 is applicable here. If we take $\|\cdot\|$ on both sides of (2.2.6), we have

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

Combining this with $\|G\| \leq \|A\| \leq \|R^T\|\,\|R\|$, we obtain

$$\frac{\|\dot{R}(0)\|}{\|R\|} \leq \kappa(R^T)\mathrm{cond}(R),$$

which with the Taylor expansion (2.2.3) gives (2.2.43). □

Note $\mathrm{cond}(R)$ is invariant under the row scaling of R, in other words, the perturbation bound (2.2.43) provides the scaling automatically. This makes (2.2.43) look simpler than (2.2.37). Also from (1.2.12) in van der Sluis's Theorem 1.2.1, we know for the ∞-norm,

$$\mathrm{cond}_\infty(R) = \min_{D\in\mathbf{D}_n} \kappa_\infty(D^{-1}R) = \kappa_\infty(D_{r1}^{-1}R),$$

where $D_{r1}^{-1}R$ has rows of unit 1-norm ($D_{r1} = \mathrm{diag}(\|R(i,:)\|_1)$. This gives the condition estimator $\kappa_1(R)\kappa_\infty(D_{r1}^{-1}R)$ with respect to the ∞-norm.

2.2.2 Rigorous perturbation bounds

Usually a first-order bound is satisfactory, but sometimes more careful work is needed. In this section, we will present rigorous perturbation bounds (with no higher order

terms) for the Cholesky factor by the matrix-vector equation approach and the matrix equation approach.

Let $A = R^T R$. If $A + \Delta A = (R + \Delta R)^T (R + \Delta R)$, then we have

$$R^T \Delta R + \Delta R^T R = \Delta A - \Delta R^T \Delta R, \tag{2.2.44}$$

or

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

which gives

$$\Delta R R^{-1} = \mathrm{up}[R^{-T}(\Delta A - \Delta R^T \Delta R) R^{-1}]. \tag{2.2.45}$$

We will use (2.2.44) and (2.2.45) in deriving rigorous perturbation bounds as well as the following lemma.

Lemma 2.2.1 (*A trivial variant of Theorem* 3.1 *in Stewart* [38, 1973]. *and Theorem* 2.11 *in Stewart and Sun* [45, 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* $\rho = \tau\eta/\delta < 1/4$, *then the equation*

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

has a unique solution x *that satisfies*

$$\|x\| \leq \frac{2\tau}{1 + \sqrt{1 - 4\rho}} \leq 2\tau. \qquad \square$$

First we give a rigorous perturbation bound by the matrix-vector equation approach.

Let $X = \Delta R$, and define $\mathbf{T}_R X = R^T X + X^T R$ and $\phi(X) = X^T X$. Then (2.2.44) can be rewritten as

$$\mathbf{T}_R X = \Delta A - \phi(X). \tag{2.2.46}$$

By using Lemma 2.2.1 we can prove the following theorem.

Theorem 2.2.7 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^T R$. Let $\Delta A \in \mathbf{R}^{n\times n}$ be symmetric. If*

$$\|\widehat{W}_R^{-1}\|_2 \, \|\widehat{W}_R^{-1} \operatorname{duvec}(\Delta A)\|_2 < 1/4, \tag{2.2.47}$$

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

$$A + \Delta A = (R + \Delta R)^T (R + \Delta R), \tag{2.2.48}$$

where

$$\begin{aligned} \|\Delta R\|_F &\leq \frac{2\,\|\widehat{W}_R^{-1} \operatorname{duvec}(\Delta A)\|_2}{1 + \sqrt{1 - 4\,\|\widehat{W}_R^{-1}\|_2 \, \|\widehat{W}_R^{-1} \operatorname{duvec}(\Delta A)\|_2}} \\ &\leq 2\,\|\widehat{W}_R^{-1} \operatorname{duvec}(\Delta A)\|_2 \leq 2\,\|\widehat{W}_R^{-1}\|_2 \, \|\Delta A\|_F. \end{aligned} \tag{2.2.49}$$

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

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq 2\kappa_C(A) \frac{\|\Delta A\|_F}{\|A\|_2}. \tag{2.2.50}$$

Proof. See Chang, Paige and Stewart [13, 1996]. □

From (2.2.23) and (2.2.35), it follows that for any $D \in \mathbf{D}_n$

$$\frac{1}{4} \|A^{-1}\|_2 \, \|\Delta A\|_F \leq \|\widehat{W}_R^{-1}\|_2^2 \, \|\Delta A\|_F \leq \kappa_2^2(D^{-1}R) \|A^{-1}\|_2 \, \|\Delta A\|_F.$$

In randomly perturbed problems we expect

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

Thus the assumption (2.2.47) is generally stronger than (2.2.7), and may be greatly so. Following the same argument as Stewart [41, 1993], however, (2.2.47) is needed

to guarantee that the bound on $\|\Delta R\|_F$ will not explode. Furthermore, if the ill-conditioning of R is mostly due to bad scaling of the rows, then correct choice of D can give $\kappa_2(D^{-1}R)$ very near one. In particular, if the standard symmetric pivoting is used in computing the Cholesky factorization, then $\|\widehat{W}_R^{-1}\|_2$ and $\frac{1}{2}\|A^{-1}\|_2^{1/2}$ will usually be of similar magnitude, see (2.2.40); that is, $\|\widehat{W}_R^{-1}\|_2 \|\widehat{W}_R^{-1}\,\mathrm{duvec}(\Delta A)\|_2$ and $\frac{1}{4}\|A^{-1}\|_2\|\Delta A\|_F$ will usually have similar magnitude. So the condition (2.2.47) is not too constraining. Numerical experiments suggest that (2.2.49) is better than the equivalent result in Sun [46, 1991], see Chang, Paige and Stewart [13, Section 3.2].

Now we use the matrix equation approach to derive a weaker but practical rigorous perturbation bound. Let $R \equiv D\bar{R}$. then from (2.2.45) we have

$$\Delta R\bar{R}^{-1} = \mathrm{up}(R^{-T}\Delta A\bar{R}^{-1}) - \mathrm{up}(R^{-T}\Delta R^T\Delta R\bar{R}^{-1}). \tag{2.2.51}$$

Let $X = \Delta R\bar{R}^{-1}$, and define $\phi(X) = \mathrm{up}(D^{-1}X^TX)$. Then (2.2.51) can be rewritten as

$$X = \mathrm{up}(R^{-T}\Delta A\bar{R}^{-1}) - \phi(X). \tag{2.2.52}$$

Applying Lemma 2.2.1 to (2.2.52) we obtain the following result.

Theorem 2.2.8 *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, and assume $D \in \mathbf{D}_n$. If*

$$\|R^{-T}\Delta AR^{-1}D\|_F\|D^{-1}\|_2 < 1/4, \tag{2.2.53}$$

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

$$A + \Delta A = (R + \Delta R)^T(R + \Delta R), \tag{2.2.54}$$

where

$$\|\Delta R\|_F \le \frac{2\|R^{-T}\Delta AR^{-1}D\|_F\|D^{-1}R\|_2}{1 + \sqrt{1 - 4\|R^{-T}\Delta AR^{-1}D\|_F\|D^{-1}\|_2}} \tag{2.2.55}$$

If the assumption (2.2.53) *is strengthened to*

$$\kappa_2(R)\|R\|_2\,\|R^{-1}D\|_2\,\|D^{-1}\|_2\frac{\|\Delta A\|_F}{\|A\|_2} < 1/4, \tag{2.2.56}$$

then

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{2\kappa_2(R)\kappa_2(D^{-1}R)\frac{\|\Delta A\|_F}{\|A\|_2}}{1+\sqrt{1-4\kappa_2(R)\|R\|_2\|R^{-1}D\|_2\|D^{-1}\|_2\frac{\|\Delta A\|_F}{\|A\|_2}}} \tag{2.2.57}$$

$$\leq 2\kappa_c'(A,D)\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{2.2.58}$$

Proof. It is easy to see that the function $\phi(X) = \text{up}(D^{-1}X^TX)$ satisfies

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

and for any upper triangular matrices X and Y,

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

so we take $\eta = \|D^{-1}\|_2$. By the assumption (2.2.53), we have $\rho = \tau\eta/\delta < 1/4$, with $\tau = \|\text{up}(R^{-T}\Delta AR^{-1}D)\|_F \leq \|R^{-T}\Delta AR^{-1}D\|_F$, $\eta = \|D^{-1}\|_2$, and $\delta = 1$ since here $\mathbf{T}$ is an identity operator. Thus, by Lemma 2.2.1, (2.2.52) has a unique upper triangular solution, say $\Delta R\bar{R}^{-1}$, where $\bar{R} \equiv RD^{-1}$, that satisfies

$$\|\Delta R\bar{R}^{-1}\|_F \leq \frac{2\|\text{up}(R^{-T}\Delta A\bar{R}^{-1})\|_2}{1+\sqrt{1-4\|\text{up}(R^{-T}\Delta A\bar{R}^{-1})\|_F\|D^{-1}\|_2}}$$

$$\leq \frac{2\|R^{-T}\Delta A\bar{R}^{-1}\|_2}{1+\sqrt{1-4\|R^{-T}\Delta A\bar{R}^{-1}\|_F\|D^{-1}\|_2}} \tag{2.2.59}$$

thus (2.2.54) and (2.2.55) follow, the latter using $\|\Delta R\|_F \leq \|\Delta R\bar{R}^{-1}\|_F\|\bar{R}\|_2$.

But $R+\Delta R$ in (2.2.54) must have positive diagonal to satisfy our definition of the Cholesky factorization, where R was given and $\Delta R\bar{R}^{-1}$ solves (2.2.52). We now prove the positivity. From (2.2.59) and (2.2.53) it follows that

$$\|\Delta RR^{-1}\|_F \leq \|\Delta R\bar{R}^{-1}\|_F\|D^{-1}\|_2 \leq 2\|R^{-T}\Delta A\bar{R}^{-1}\|_F\|D^{-1}\|_2 \leq 1/2.$$

Thus $R+\Delta R$ is nonsingular for any $t \in [0,1]$, and by continuity of elements, $R+\Delta R$ has positive diagonal.

If (2.2.56) holds, then (2.2.57) can easily be obtained from (2.2.59) and $\|\Delta R\|_F \leq \|\Delta R\bar{R}^{-1}\|_F\|\bar{R}\|_2$ by using $\|A\|_2 = \|R\|_2^2$. □

If we take $D = I$ in Theorem 2.2.8, the assumption (2.2.56) can be weakened to

$$\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2} < 1/2, \tag{2.2.60}$$

and we have the following perturbation bound, which is due to Sun [46, 1991],

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}{1+\sqrt{1-2\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}}} \leq \sqrt{2}\kappa_2(A)\frac{\|\Delta A\|_F}{\|A\|_2}. \tag{2.2.61}$$

This is a slightly stronger than (2.2.57) where D is replaced by I. The proof is similar to that of Theorem 2.2.8. The only difference is using the fact that $\|\operatorname{up}(X)\|_F \leq \frac{1}{\sqrt{2}}\|X\|_F$ for any symmetric $X \in \mathbf{R}^{n\times n}$.

As we know from (1.2.14) in van der Sluis's Theorem 1.2.1 that if we take $D = D_r = \operatorname{diag}(\|R(i,:)\|_2)$ in $\kappa_C'(A,D)$, $\kappa_C'(A,D_r)$ will be nearly minimal. Thus possibly the bound (2.2.57) is much smaller than (2.2.61). But the assumption (2.2.56) with $D = D_r$ is possibly much more constraining than (2.2.60).

2.2.3 Numerical experiments

In Section 2.2.1, we made first-order perturbation analyses for the Cholesky factorization with norm-bounded changes in A using two different approaches, presented $\kappa_C(A) = \|\widehat{W}_R^{-1}\|_2\|A\|_2^{1/2}$ as the corresponding condition number, and suggested $\kappa_C(A)$ could be approximated in practice by $\kappa_C'(A,D_r) = \kappa_2(R)\kappa_2(D_r^{-1}R)$ with $D_r = \operatorname{diag}(\|R(i,:)\|_2)$, which could be estimated by a standard condition estimator (see for example Higham [30, 1996, Ch. 14]) in $O(n^2)$. Also we showed the condition of the problem can usually be (significantly) improved by standard symmetric pivoting. In Section 2.2.2 rigorous perturbation bounds were obtained. In order to confirm our theoretical analyses, we have carried out several numerical experiments to compute the following measures of the sensitivity of the Cholesky factorization, which satisfy (see (1.2.14), (2.2.26), (2.2.30) and (2.2.31))

$$\tfrac{1}{2}\kappa_2^{1/2}(A) \leq \kappa_C(A) \leq \kappa_C'(A) \leq \kappa_C'(A,D_r),$$

$$\kappa_C(A) \le \tfrac{1}{\sqrt{2}}\kappa_2(A), \qquad \tfrac{1}{\sqrt{n}}\kappa'_C(A, D_r) \le \kappa'_C(A) \le \kappa_2(A).$$

The computations were performed in Matlab. Here we give three sets of numerical examples.

(1) The matrices in the first set have the form

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

where $Q \in \mathbf{R}^{n\times n}$ is an orthogonal matrix obtained from the QR factorization of a random $n \times n$ matrix, Λ=diag$(r_1, \ldots, r_{n-2}, \delta, \delta)$ with r_1, ..., r_{n-2} random positive numbers and $0 < \delta \le 1$. We generated different matrices by taking all combinations of $n \in \{5, 10, 15, 20, 25\}$ and $\delta \in \{1, 10^{-1}, \ldots, 10^{-10}\}$. The results for $n = 25$, $\delta = 10^{-i}$, i=0, 1, ..., 10 are shown in Table 2.2.1, where P is a permutation corresponding to the standard symmetric pivoting. Results obtained by putting the two δ's at the top of Λ were mainly similar.

(2) The second set of matrices are $n \times n$ Pascal matrices (with elements $a_{1j} = a_{i1} = 1$, $a_{ij} = a_{i,j-1} + a_{i-1,j}$), $n = 1, \ldots, 15$. The results are shown in Table 2.2.2, where P is a permutation corresponding to the standard symmetric pivoting.

(3) The third set of matrices are $n \times n$ $A = K_n^T(\theta)K_n(\theta)$, where $K_n(\theta)$ are Kahan matrices, see (2.2.41). The results for $n = 5, 10, 15, 20, 25$ with $\theta = \pi/4$ are shown in Table 2.2.3, where Π is a permutation such that the first column and row are moved to the last column and row positions, and the remaining columns and rows are moved left and up one position — this permutation Π corresponds to the rank-revealing pivoting strategy, for details, see Hong and Pan [31, 1992]. The permutation P corresponding to the standard symmetric pivoting is the identity, so the standard symmetric pivoting does not bring any changes to our condition number and condition estimators.

We give some comments on the results.

- The experiments confirm that $\kappa_2(A)/\sqrt{2}$ can be much larger than $\kappa_C(A)$ for ill-conditioned problems, so the first-order bound in (2.2.25) may be much smaller

Table 2.2.1: Results for matrix $A = Q\Lambda Q^T$ of order 25, $\tilde{A} = PAP^T$, $D = D_r$

δ	$\frac{\kappa_2^{1/2}(A)}{2}$	$\kappa_C(\tilde{A})$	$\kappa'_C(\tilde{A}, D)$	$\kappa_C(A)$	$\kappa'_C(A, D)$	$\frac{\kappa_2(A)}{\sqrt{2}}$	$\frac{2\kappa_C(\tilde{A})}{\kappa_2^{1/2}(A)}$
1.0	2.0e+00	4.3e+00	1.3e+01	4.8e+00	1.5e+01	1.2e+01	2.1
1.0e-01	4.0e+00	1.1e+01	4.3e+01	1.2e+01	4.8e+01	4.5e+01	2.8
1.0e-02	4.9e+00	1.5e+01	6.8e+01	2.2e+01	8.0e+01	6.8e+01	3.1
1.0e-03	1.6e+01	4.2e+01	1.8e+02	1.3e+02	4.7e+02	7.0e+02	2.7
1.0e-04	4.9e+01	1.5e+02	6.8e+02	1.1e+03	5.2e+03	6.8e+03	3.0
1.0e-05	1.5e+02	4.5e+02	1.7e+03	5.6e+02	2.2e+03	6.7e+04	2.9
1.0e-06	5.0e+02	1.4e+03	5.0e+03	1.4e+04	4.9e+04	7.0e+05	2.9
1.0e-07	1.6e+03	4.4e+03	2.0e+04	1.7e+04	6.3e+04	7.0e+06	2.8
1.0e-08	5.0e+03	1.3e+04	9.4e+04	1.2e+05	6.5e+05	7.1e+07	2.6
1.0e-09	1.5e+04	5.3e+04	2.4e+05	1.9e+05	8.5e+05	6.7e+08	3.4
1.0e-10	5.0e+04	1.4e+05	5.4e+05	2.0e+06	1.0e+07	7.0e+09	2.8

Table 2.2.2: Results for Pascal matrices, $\tilde{A} = PAP^T$, $D = D_r$

n	$\frac{\kappa_2^{1/2}(A)}{2}$	$\kappa_C(\tilde{A})$	$\kappa'_C(\tilde{A}, D)$	$\kappa_C(A)$	$\kappa'_C(A, D)$	$\frac{\kappa_2(A)}{\sqrt{2}}$	$\frac{2\kappa_C(\tilde{A})}{\kappa_2^{1/2}(A)}$
1	5.0e−01	5.0e−01	1.0e+00	5.0e−01	1.0e+00	7.1e−01	1.0
2	1.3e+00	1.5e+00	4.2e+00	2.1e+00	6.3e+00	4.8e+00	1.2
3	3.9e+00	5.1e+00	1.6e+01	9.7e+00	5.0e+01	4.4e+01	1.3
4	1.3e+01	2.2e+01	8.0e+01	5.5e+01	4.8e+02	4.9e+02	1.7
5	4.6e+01	8.3e+01	3.3e+02	3.5e+02	6.0e+03	6.0e+03	1.8
6	1.7e+02	2.5e+02	1.3e+03	2.5e+03	5.2e+04	7.8e+04	1.5
7	6.1e+02	9.4e+02	5.1e+03	1.9e+04	5.7e+05	1.1e+06	1.5
8	2.3e+03	4.0e+03	2.4e+04	1.5e+05	6.3e+06	1.5e+07	1.8
9	8.5e+03	1.6e+04	1.0e+05	1.3e+06	7.0e+07	2.1e+08	1.9
10	3.2e+04	7.6e+04	4.7e+05	1.1e+07	7.9e+08	2.9e+09	2.4
11	1.2e+05	2.4e+05	1.8e+06	9.8e+07	9.0e+09	4.2e+10	1.9
12	4.7e+05	8.3e+05	8.2e+06	8.7e+08	1.0e+11	6.2e+11	1.8
13	1.8e+06	3.2e+06	3.1e+07	7.8e+09	1.2e+12	9.1e+12	1.8
14	6.9e+06	1.3e+07	1.2e+08	7.1e+10	1.4e+13	1.3e+14	1.9
15	2.7e+07	5.4e+07	4.9e+08	6.5e+11	1.6e+14	2.0e+15	2.0

Table 2.2.3: Results for $A = K_n^T(\theta)K_n(\theta)$, $\theta = \pi/4$, $\tilde{A} = \Pi A \Pi^T$, $D = D_r$

n	$\frac{\kappa_2^{1/2}(A)}{2}$	$\kappa_C(\tilde{A})$	$\kappa'_C(\tilde{A}, D)$	$\kappa_C(A)$	$\kappa'_C(A, D)$	$\frac{\kappa_2(A)}{\sqrt{2}}$	$\frac{2\kappa_C(\tilde{A})}{\kappa_2^{1/2}(A)}$
5	1.7e+01	2.2e+01	1.0e+02	8.7e+01	3.1e+02	8.5e+02	1.3
10	2.1e+03	2.6e+03	2.8e+04	1.5e+05	7.3e+05	1.3e+07	1.2
15	2.2e+05	2.8e+05	4.8e+06	2.3e+08	1.4e+09	1.4e+11	1.2
20	2.2e+07	2.7e+07	6.6e+08	3.2e+11	2.5e+12	1.3e+15	1.2
25	2.0e+09	2.5e+09	8.0e+10	4.4e+14	3.8e+15	1.2e+19	1.2

than that in (2.2.8).

- The standard symmetric pivoting almost always gives an improvement on $\kappa_C(A)$ and $\kappa'_C(A, D_r)$. Table 2.2.2 indicates the improvement can be significant. Our experiments 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 usually have the same order as its lower limit $\kappa_2^{1/2}(A)/2$ (the ratio in the last columns of Table 2.2.1 and Table 2.2.2 was never larger than 4). But Table 2.2.3 shows the ratio can be large. However such examples are rare in practice, and furthermore if we adopt the rank-revealing pivoting strategy, we see from Table 2.2.3 the ratio $2\kappa_C(\Pi A \Pi^T)/\kappa_2^{1/2}(A)$ is again small.

- Note in Table 2.2.1 and Table 2.2.3 $\kappa'_C(A, D_r)$ ($\kappa'_C(PAP^T, D_r)$, $\kappa'_C(\Pi A \Pi^T, D_r)$) is a very good approximation of $\kappa_C(A)$ ($\kappa_C(PAP^T)$, $\kappa'_C(\Pi A \Pi^T)$). In Table 2.2.2, the results with pivoting also show this. For $n = 15$ without pivoting in Table 2.2.2 $\kappa'_C(A, D_r)$ overestimates $\kappa_C(A)$ by a factor of about 250, but is much better for low n. A study of the $n = 2$ case shows $\kappa'_C(A)$ can never be much larger than $\kappa_C(A)$, and we have not found an example which shows $\kappa'_C(A, D_r)$ can be much larger than $\kappa_C(A)$ for $n > 2$, so we suspect $\kappa'_C(A)$ and $\kappa'_C(A, D_r)$ are at worst $\kappa_C(A)$ times some function of n alone (probably involving some-

thing like 2^n) in general. From Theorem 2.2.5 we know this is true when the standard symmetric pivoting strategy is used.

2.3 Perturbation analysis with component-bounded changes in A

In this section we consider the case where a bound on $|\Delta A|$ is given. There have been a few papers dealing with such problems. Sun [47, 1992] first presented a rigorous perturbation bound on $|\Delta R|$ in terms of $|\Delta A|$, which was improved by Sun [48, 1992]. For ΔA corresponding to the backward rounding error in A resulting from numerical stable computations, e.g. Algorithm CHOL, on finite precision floating point computers, Drmač, Omladič and Veselić [20, 1994] presented a nice norm-based perturbation result using their $H = D_c^{-1} A D_c^{-1}$ approach. Sun in [47] and [48] also included component perturbation bounds for a different and a somewhat complicated form of the backward rounding error in A.

The main purpose of this section is to establish new perturbation bounds when ΔA corresponds to the backward rounding error, which are sharper than the equivalent results in [20]. Most of the results have been presented in Chang [8, 1996]. Also we present first-order perturbation bounds for some other kinds of bounds on $|\Delta A|$.

In Section 2.3.1 we establish first-order perturbation bounds with a general bound $|\Delta A| \leq \epsilon E$, and apply such results to a special case: $|\Delta A| \leq \epsilon |A|$. The motivation for considering this special form is that possibly the relative error in each element of A has the same reasonable known bound, for example, often the elements of A can not be stored exactly on a digital computer, and the matrix actually stored is $A + \Delta A$ with $|\Delta A| \leq u|A|$, where u is the unit roundoff. In Section 2.3.2 and Section 2.3.3 we respectively derive first-order and rigorous perturbation bounds with $|\Delta A| \leq \epsilon\, dd^T$, where $d_i = a_{ii}^{1/2}$, which comes from the backward rounding error

analysis, see Demmel [17, 1989]. In our analyses we use both the matrix-vector equation approach and the matrix equation approach.

2.3.1 First-order perturbation bound with $|\Delta A| \leq \epsilon |A|$

First we give the following result by using the matrix-vector equation approach.

Theorem 2.3.1 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite, with the Cholesky factorization $A = R^T R$. Let $\Delta A \in \mathbf{R}^{n\times n}$ be a symmetric matrix satisfying $|\Delta A| \leq \epsilon E$ for some nonnegative matrix E with nonnegative ϵ. If*

$$\epsilon \| \, |A^{-1}|E| \, \| \leq 1, \tag{2.3.1}$$

where $\|\cdot\|$ denotes a monotone and consistent matrix norm, then $A + \Delta A$ has the Cholesky factorization

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

such that

$$\operatorname{uvec}(|\Delta R|)| \leq \epsilon \, |W_R^{-1}| \operatorname{uvec}(E) + O(\epsilon^2), \tag{2.3.2}$$

$$\frac{\|\Delta R\|_\nu}{\|R\|_\nu} \leq \frac{\| \, |W_R^{-1}| \operatorname{uvec}(|E|)\|_\nu}{\|R\|_\nu} \epsilon + O(\epsilon^2), \quad \nu = F, \, M, \, S, \tag{2.3.3}$$

where $\|X\|_M = \max_{i,j} |x_{ij}|$ and $\|X\|_S = \sum_{i,j} |x_{ij}|$, and for the M-norm the first-order bound in (2.3.3) *is attainable. Thus, in particular, if $E = |A|$, then*

$$\operatorname{uvec}(|\Delta R|)| \leq \epsilon \, |W_R^{-1}| \operatorname{uvec}(|A|) + O(\epsilon^2), \tag{2.3.4}$$

$$\frac{\|\Delta R\|_\nu}{\|R\|_\nu} \leq \frac{\| \, |W_R^{-1}| \operatorname{uvec}(|A|)\|_\nu}{\|R\|_\nu} \epsilon + O(\epsilon^2), \quad \nu = F, \, M, \, S, \tag{2.3.5}$$

and for the M-norm the first-order bound in (2.3.5) *is attainable.*

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). Since

$$\rho(A^{-1}\Delta A) \leq \|A^{-1}\Delta A\| \leq \| \, |A^{-1}|E| \, \| \, \epsilon < 1,$$

the conclusion of Theorem 2.2.1 holds. Then from (2.2.12), the matrix-vector equation form of (2.2.5), we have

$$|\mathrm{uvec}(\dot{R}(0))| = |W_R^{-1}\mathrm{uvec}(G)| \le |W_R^{-1}|\,\mathrm{uvec}(|G|) \le |W_R^{-1}|\,\mathrm{uvec}(E).$$

This with the Taylor expansion (2.2.3) gives (2.3.2), from which (2.3.3) follows. For the M-norm the first-order bound is attained for ΔA satisfying

$$\mathrm{uvec}(\Delta A) = \epsilon\,\mathcal{D}\,\mathrm{uvec}(E), \qquad \mathcal{D} = \mathrm{diag}(\zeta_j),$$

where $\zeta_j = \mathrm{sign}(W_R^{-1})_{kj}$ and $\| |W_R^{-1}|\,\mathrm{uvec}(E)\|_M = (|W_R^{-1}|\,\mathrm{uvec}(E))_k$. □

Theorem 2.3.1 implies that for the Cholesky factorization under relative changes in the elements of A the condition number (with respect to the M-norm)

$$\mu_C(A) \equiv \limsup_{\epsilon\to 0}\left\{\frac{\|\Delta R\|_M}{\epsilon\|R\|_M} :\ (A+\Delta A) = (R+\Delta R)^T(R+\Delta R),\ |\Delta A| \le \epsilon|A|\right\}$$

is given by

$$\mu_C(A) \equiv \frac{\| |W_R^{-1}|\,\mathrm{uvec}(|A|)\|_M}{\|R\|_M}. \tag{2.3.6}$$

We see $\mu_C(A)$ is not very intuitive, and is expensive to estimate directly by any presently known approach. Fortunately by the matrix-equation approach we have the following practical results.

Theorem 2.3.2 *With the same assumptions as in Theorem* 2.3.1, $A + \Delta A$ *has the Cholesky factorization*

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

where

$$|\Delta R| \le \epsilon\ \mathrm{up}(|R^{-T}|\,E\,|R^{-1}|)|R| + O(\epsilon^2). \tag{2.3.7}$$

In particular, if $E = |A|$, *then*

$$|\Delta R| \le \epsilon\ \mathrm{up}(|R^{-T}||R^T||R||R^{-1}|)|R| + O(\epsilon^2), \tag{2.3.8}$$

and for a monotone and consistent matrix norm $\|\cdot\|$,

$$\frac{\|\Delta R\|}{\|R\|} \leq \min\{\mathrm{cond}(R^T)\,\mathrm{cond}(R),\ \mathrm{cond}(R^T)\,\mathrm{cond}(R^{-1})\}\epsilon + O(\epsilon^2), \tag{2.3.9}$$

and for the M-norm,

$$\frac{\|\Delta R\|_M}{\|R\|_M} \leq \mu_C'(A)\epsilon + O(\epsilon^2), \tag{2.3.10}$$

where

$$\mu_C'(A) \equiv \min\{\mathrm{cond}_\infty(R^T)\,\mathrm{cond}_1(R),\ \mathrm{cond}_\infty(R^T)\,\mathrm{cond}_1(R^T)\}, \tag{2.3.11}$$

$$\mu_C(A) \leq \mu_C'(A). \tag{2.3.12}$$

Proof. (2.3.7) can easily be proved by using (2.2.6) and (2.2.3). Notice $|A| \leq |R^T||R|$, then (2.3.8) follows immediately from (2.3.7), and (2.3.9) is obtained by taking the norm $\|\cdot\|$ and using

$$\begin{aligned} \|\,\mathrm{up}(|R^{-T}||R^T||R||R^{-1}|)|R|\,\| &\leq \|\,|R^{-T}||R^T||R||R^{-1}||R|\,\| \\ &\leq \begin{cases} \|\,|R^{-T}||R^T|\,\|\cdot\|R\|\cdot\|\,|R^{-1}||R|\,\|, \\ \|\,|R^{-T}||R^T|\,\|\cdot\|\,|R||R^{-1}|\,\|\cdot\|R\|. \end{cases} \end{aligned}$$

Similarly (2.3.10) can be obtained by taking the M-norm and using

$$\|\,\mathrm{up}(|R^{-T}||R^T||R||R^{-1}|)|R|\,\|_M \leq \|\,|R^{-T}||R^T||R||R^{-1}||R|\,\|_M \tag{2.3.13}$$

$$\leq \begin{cases} \|\,|R^{-T}||R^T|\,\|_\infty\|R\|_M\|\,|R^{-1}||R|\,\|_1, \\ \|\,|R^{-T}||R^T|\,\|_\infty\|\,|R||R^{-1}|\,\|_\infty\|R\|_M, \end{cases} \tag{2.3.14}$$

where we used the fact that $\|AB\|_M \leq \|A\|_M\|B\|_1$, $\|A\|_\infty\|B\|_M$ for any $A \in \mathbf{R}^{m\times p}$ and $B \in \mathbf{R}^{p\times n}$, which can easily be verified. From the definition of $\mu_C(A)$ and (2.3.10), we see the inequality (2.3.12) holds. □

The quintuple matrix product in (2.3.8) is nice, as it counteracts the effect of poor column scaling in R (the first product and the third product), and poor row scaling (the fourth product). This is reflected in both of the first-order bounds in

(2.3.9) and (2.3.10). The significance of this theorem is now we can estimate a bound on the relative change in R in $O(n^2)$ by using the standard condition estimators. Numerical experiments suggest usually $\mu'_c(A)$ is a reasonable upper bound on the condition number $\mu_c(A)$. But the former can be arbitrarily larger than the latter. For example, if $A = \begin{bmatrix} 1 & \delta \\ \delta & \delta^2 + \delta^4 \end{bmatrix}$ with small $\delta > 0$, then $R = \begin{bmatrix} 1 & \delta \\ 0 & \delta^2 \end{bmatrix}$. Simple computations give

$$\mu_c(A) = O(1), \qquad \mu'_c(A) = O(\frac{1}{\delta}).$$

It is easy to check the overestimation was caused by the inequality $|\operatorname{up}(B)| \leq |B|$ used in deriving (2.3.10). Clearly the strictly lower triangular of $|B|$ can be arbitrarily larger than that of $|\operatorname{up}(B)|$. Hence (2.3.10) can sometimes overestimate the true sensitivity of the problem, so can (2.3.9) for the same reason. We also can give an example to show $\mu'_c(A)$ can overestimate $\mu_c(A)$ due to the inequality (2.3.14).

A careful reader may have noticed the following fact: in the proof of Theorem 2.2.37 we also used the inequality $|\operatorname{up}(B)| \leq |B|$ (see (2.2.28)) but we mentioned in Section 2.2.3 that we have not found an example to show $\kappa'_c(A)$ can be arbitrarily larger than $\kappa_c(A)$, and furthermore initial investigations suggest probably $\kappa'_c(A)/\kappa_c(A)$ can be bounded above by a function of n. Why does the inequality appear to have different effects? The reason is that here B is the function of only R and so has a special structure, whereas in (2.2.28) B has a parameter matrix G, and for any R possibly G can be chosen such that $\|B\|$ is close to $\|\operatorname{up}(B)\|$.

Comparing the first-order bound in (2.3.9) with that in (2.2.43), we see the former is at least as small as the latter since $\operatorname{cond}(R^T) \leq \kappa(R^T)$. If the ill-conditioning of R is mostly due to bad scaling of the columns, then $\operatorname{cond}(R^T) \ll \kappa(R^T)$, that is to say the former can be much smaller than the latter. That is not surprising since the assumption $|\Delta A| \leq \epsilon |A|$ in Theorem 2.3.2 provides more information about the perturbations in data than the assumption $\|\Delta A\| \leq \epsilon \|A\|$ in Theorem 2.2.6.

The standard pivoting strategy can usually improve $\mu_c(A)$, just as it can usually improve $\kappa_c(A)$. In fact we have the following theorem.

Theorem 2.3.3 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $PAP^T = R^T R$ when the standard pivoting strategy is used. Then*

$$\mu_c(PAP^T) \le \mu'_c(PAP^T) \le (2^n - 1)\,\mathrm{cond}_\infty(R^T). \tag{2.3.15}$$

Proof. Standard pivoting ensures $|r_{ii}| \ge |r_{ij}|$ for all $j \ge i$. Then from the definition of $\mu'_c(A)$ in (2.3.11), (2.3.15) is immediately obtained by using (1.2.21) in Theorem 1.2.2. □

From (2.3.15) it is natural to raise the following question: is it possible that the standard pivoting makes $\mathrm{cond}_\infty(R^T)$ much worse than that without pivoting, so that $\mu'_c(PAP^T)$ is actually much worse than $\mu'_c(A)$? The answer is no. In fact we can show any pivoting can not bring an essential change to $\mathrm{cond}_\infty(R^T)$. Let $PAP^T = R^T R$, where P is any permutation matrix. Let $D_1 = \mathrm{diag}(\|R^T(i,:)\|_1)$ and let $D_2 = \mathrm{diag}(\|R^T(i,:)\|_2)$. Then $\|D_2^{-1}D_1\|_\infty \le \sqrt{n}$ and $\|D_1^{-1}D_2\|_\infty \le 1$. By using van der Sluis's Theorem 1.2.1, we have

$$\begin{aligned}\mathrm{cond}_\infty(R^T) &= \|R^{-T}D_1\|_\infty \|D_1^{-1}R^T\|_\infty = \|R^{-T}D_2D_2^{-1}D_1\|_\infty \\ &\le n\,\|R^{-T}D_2\|_2 = n\,\|D_2PA^{-1}P^TD_2\|_2^{1/2}.\end{aligned}$$

Notice that $D_2 = \mathrm{diag}(\|R^T(i,:)\|_2) = \mathrm{diag}(PAP^T)^{1/2} = P\,\mathrm{diag}(A)^{1/2}\,P^T$, then

$$\mathrm{cond}_\infty(R^T) \le n\,\|P\,\mathrm{diag}(A)^{1/2}\,A^{-1}\,\mathrm{diag}(A)^{1/2}\,P^T\|_2^{1/2} = n\,\|H^{-1}\|_2^{1/2},$$

where $H \equiv \mathrm{diag}(A)^{-1/2}\,A\,\mathrm{diag}(A)^{-1/2}$. On the other hand, we have

$$\mathrm{cond}_\infty(R^T) = \|R^{-T}D_2D_2^{-1}D_1\|_\infty \ge \frac{1}{\sqrt{n}}\,\|R^{-T}D_2\|_2/\|D_1^{-1}D_2\|_\infty \ge \frac{1}{\sqrt{n}}\,\|H^{-1}\|_2^{1/2}.$$

Notice $\|H^{-1}\|_2$ is independent of P, thus permutation has no significant effect on $\mathrm{cond}_\infty(R^T)$.

Using the same approaches as in Section 2.2.2 we could easily provide rigorous perturbation bounds with $|\Delta A| \leq \epsilon |A|$. But we choose not to do so here in order to keep the material and the basic ideas as brief as possible.

2.3.2 First-order perturbation bounds with backward rounding errors

In this section we first use the matrix-vector equation approach to derive tight perturbation bounds, leading to the condition number $\chi_C(A)$ for perturbations having bounds of the form of the equivalent backward rounding error for the Cholesky factorization, then use the matrix equation approach to derive a practical perturbation bound, leading to a practical estimator $\chi'_C(A)$ of $\chi_C(A)$. We also compare $\chi_C(A)$ with $\kappa_C(A)$, and $\chi'_C(A)$ with $\kappa'_C(A)$. Finally we show how standard pivoting improves the condition number $\chi_C(A)$.

Before proceeding we introduce the following result due to Demmel [17, 1989], also see Higham [30, 1996, Theorems 10.5 and 10.7].

Lemma 2.3.1 *Let $A \equiv D_c H D_c \in \mathbf{R}^{n\times n}$ be a symmetric positive definite floating point matrix, where $D_c = \mathrm{diag}(A)^{1/2}$. If*

$$n\epsilon \|H^{-1}\|_2 < 1, \tag{2.3.16}$$

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

$$A + \Delta A = \tilde{R}^T \tilde{R}, \qquad |\Delta A| \leq \epsilon\, dd^T, \tag{2.3.17}$$

where $d_i = a_{ii}^{1/2}$. □

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

Based on Lemma 2.3.1, we can establish the following bound for the computed Cholesky factor $\tilde{R}$.

Theorem 2.3.4 *With $A = R^T R$ and the same assumptions as in Lemma 2.3.1, for the perturbation ΔA and result $\tilde{R}$ in (2.3.17) we have*

$$\operatorname{uvec}(|\Delta R|) \leq \epsilon\, |W_R^{-1}|\, \operatorname{uvec}(dd^T) + O(\epsilon^2), \tag{2.3.18}$$

$$\frac{\|\Delta R\|_\nu}{\|R\|_\nu} \leq \frac{\| \, |W_R^{-1}|\, \operatorname{uvec}(dd^T)\|_\nu}{\|R\|_\nu}\, \epsilon + O(\epsilon^2), \quad \nu = M,\, S, \tag{2.3.19}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{n\|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2}{\|A\|_2^{1/2}}\, \epsilon + O(\epsilon^2), \tag{2.3.20}$$

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

$$\mathcal{D}_c \equiv \operatorname{diag}(a_{11}^{1/2}, \underbrace{a_{22}^{1/2}, a_{22}^{1/2}}_{2}, \ldots, \underbrace{a_{nn}^{1/2}, a_{nn}^{1/2}, \ldots, a_{nn}^{1/2}}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}, \tag{2.3.21}$$

$\hat{R} \equiv R D_c^{-1}$, and $\widehat{W}_{\hat{R}}$ is just $\widehat{W}_R$ in (2.2.15) with each entry r_{ij} replaced by $\hat{r}_{ij}$. The first-order bound in (2.3.19) is attainable for the M-norm, and the first-order bound in (2.3.20) is approximately attainable.

Proof. Let $G \equiv \Delta A/\epsilon$. By (2.3.17) and (2.3.16), we have

$$\begin{aligned} \rho(A^{-1}\Delta A) &\leq \rho(D_c^{-1}H^{-1}D_c^{-1}\Delta A) = \rho(H^{-1}D_c^{-1}\Delta A D_c^{-1}) \\ &\leq \|H^{-1}\|_2\, \|D_c^{-1}\Delta A D_c^{-1}\|_2 \leq \epsilon\, \|H^{-1}\|_2\, \|D_c^{-1} dd^T D_c^{-1}\|_F \\ &\leq n\epsilon\, \|H^{-1}\|_2 < 1. \end{aligned}$$

Then as we did in the proof of Theorem 2.3.1 we can apply Theorem 2.2.1 to show (2.3.18) and (2.3.19) hold and the first-order bound in (2.3.19) is attainable for the M-norm.

It remains to show (2.3.20) and its attainability. From (2.2.5) in Theorem 2.2.1 and $R = \hat{R} D_c$, we have

$$\hat{R}^T \dot{R}(0) D_c^{-1} + D_c^{-1} \dot{R}^T(0) \hat{R} = D_c^{-1} G D_c^{-1}. \tag{2.3.22}$$

As we know (2.2.5) can be rewritten as the matrix-vector equation form (2.2.15), so similarly (2.3.22) can be rewritten as the following matrix-vector equation form

$$\widehat{W}_{\hat{R}}\,\mathrm{uvec}(\dot{R}(0)D_c^{-1}) = \mathrm{duvec}(D_c^{-1}GD_c^{-1}). \tag{2.3.23}$$

It is easy to verify that $\mathrm{uvec}(\dot{R}(0)D_c^{-1}) = \mathcal{D}_c^{-1}\,\mathrm{uvec}(\dot{R}(0))$ with $\mathcal{D}_c$ as in (2.3.21), then from (2.3.23) we have

$$\mathrm{uvec}(\dot{R}(0)) = \mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\,\mathrm{duvec}(D_c^{-1}GD_c^{-1}), \tag{2.3.24}$$

which with $|G| = |\Delta A/\epsilon| \le dd^T$ gives

$$\|\dot{R}(0)\|_F \le \|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2\,\|D_c^{-1}dd^TD_c^{-1}\|_F = n\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2. \tag{2.3.25}$$

Thus (2.3.20) is obtained immediately from the Taylor expansion (2.2.3).

Obviously there exists a symmetric matrix $F \in \mathbf{R}^{n\times n}$ such that

$$\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\mathrm{duvec}(D_c^{-1}FD_c^{-1})\|_2 = \|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2\,\|D_c^{-1}FD_c^{-1}\|_F.$$

Then by taking $G = (\min_{f_{ij}\neq 0} d_id_j/|f_{ij}|)F$, we have $|\Delta A| \le \epsilon\, dd^T$ and from (2.3.24) that

$$\|\dot{R}(0)\|_F = (\min_{f_{ij}\neq 0} d_id_j/|f_{ij}|)\,\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2\,\|D_c^{-1}FD_c^{-1}\|_F \ge \|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2,$$

which shows the first-order bound in (2.3.20) is approximately attained for such G. □

It is easy to verify that $D_c \equiv \mathrm{diag}(a_{ii}^{1/2}) = \mathrm{diag}(\|R(:,j)\|_2)$. Thus $\hat{R}$, the Cholesky factor of H ($H \equiv D_c^{-1}AD_c^{-1} = D_c^{-1}R^TRD_c^{-1} \equiv \hat{R}^T\hat{R}$), has columns of unit Euclidean length. That is the reason that we use the notation D_c, where 'c' denote 'column'.

Since the first-order bound in (2.3.20) is *approximately* attainable, the quantity

$$\chi_C(A) \equiv \frac{n\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2}{\|A\|_2^{1/2}} \tag{2.3.26}$$

can be thought of as the the condition number for the Cholesky factorization with the form of backward rounding error satisfying (2.3.17) when the combination of F-

and 2-norms is used. Notice here we have relaxed a little the requirement a *condition number* should satisfy. Strictly the condition number should be defined by

$$\tilde{\chi}_C(A) \equiv \limsup_{\epsilon \to 0} \left\{ \frac{\|\Delta R\|_F}{\epsilon \|R\|_2} : A+\Delta A = (R+\Delta R)^T (R+\Delta R),\ |\Delta A| \le \epsilon d d^T,\ d_i = a_{ii}^{1/2} \right\}.$$

But from the proof of Theorem 2.3.4 we see

$$\frac{\|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2}{\|A\|_2^{1/2}} \le \tilde{\chi}_C(A) \le \frac{n \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2}{\|A\|_2^{1/2}},$$

so such relaxing is harmless. The main reason for introducing the condition number with respect to the combination of the F- and 2-norms rather than the M-norm is that we are interested in the comparison of the results given in this section with those given in Section 2.2.

We now give a lower bound on $\chi_C(A)$. Since $\widehat{W}_{\hat{R}}$ has the form (cf. (2.2.22))

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

where

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

we have

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

$$\|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2 \ge a_{nn}^{1/2} \|\hat{R}^{-T} \hat{D}^{-1}\|_2. \tag{2.3.27}$$

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

$$\chi_C(A) \ge \frac{n a_{nn}^{1/2} \|\hat{R}^{-1} \hat{D}^{-1}\|_2}{\|A\|_2^{1/2}}. \tag{2.3.28}$$

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

$$\|\hat{R}^{-1} \hat{D}^{-1}\|_2 \ge \frac{1}{2} \|\hat{R}^{-1}\|_2 = \frac{1}{2} \|H^{-1}\|^{1/2}, \tag{2.3.29}$$

we have from (2.3.28) that

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

Numerical experiments suggest that usually $\chi_C(A)$ is smaller or much smaller than $\kappa_C(A)$, the condition number for a general perturbation ΔA, defined by (2.2.20). We now relate these two condition numbers mathematically for the general case. For the case where standard pivoting is used in computing the Cholesky factorization, see the comment following Theorem 2.3.9.

Theorem 2.3.5

$$\frac{1}{n} \chi_C(A) \leq \kappa_C(A) \leq \frac{\max_i a_{ii}}{\min_i a_{ii}} \chi_C(A). \tag{2.3.30}$$

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

$$\widehat{W}_R = \hat{\mathcal{D}}_c \widehat{W}_{\hat{R}} \mathcal{D}_c^{-1}, \tag{2.3.31}$$

where $\mathcal{D}_c$ is defined by (2.3.21) and

$$\hat{\mathcal{D}}_c \equiv \operatorname{diag}(\sqrt{a_{11}a_{11}}, \underbrace{\sqrt{a_{11}a_{22}}, \sqrt{a_{22}a_{22}}}_{2}, \ldots, \underbrace{\sqrt{a_{11}a_{nn}}, \sqrt{a_{22}a_{nn}}, \ldots, \sqrt{a_{nn}a_{nn}}}_{n}). \tag{2.3.32}$$

Thus using (2.3.31) and $\max_i a_{ii} \leq \|A\|_2$, we have

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

We now prove the second inequality. Using (2.3.31) and $\|A\|_2 \leq n \max_i a_{ii}$, we obtain

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

The proof is complete. □

The first inequality in (2.3.30) is attainable, since equality will hold by taking $A = cI$ with $c > 0$. The second inequality is at least nearly attainable. In fact by taking $A = R^T R$, with $R = \mathrm{diag}(\delta^{n-1}, \delta^{n-2}, \ldots, \delta, 1) + e_1 e_n^T$ with small $\delta > 0$, we easily obtain

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

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

A reader might want to know why the first inequality in (2.3.30) is not $\chi_C(A) \leq \kappa_C(A)$. This can easily be explained. For a general ΔA, we have by Theorem 2.2.3 that $\|\Delta R\|_F / \|R\|_2 \lesssim \kappa_C(A)\epsilon$, where ϵ satisfies $\|\Delta A\|_F \leq \epsilon \|A\|_2$. For the backward rounding error ΔA, we have by Theorem 2.3.4 that $\|\Delta R\|_F / \|R\|_2 \lesssim \chi_C(A)\epsilon$, where ϵ satisfies $|\Delta A| \leq \epsilon\, dd^T$ (see (2.3.17)), from which it follows that $\|\Delta A\|_F \leq \epsilon \|dd^T\|_F = \epsilon \|R\|_F^2$, where $\|R\|_F$ satisfies attainable inequalities $\|A\|_2 \leq \|R\|_F^2 \leq n\|A\|_2$.

Like $\kappa_C(A)$, $\chi_C(A)$ is difficult to estimate directly. Now we derive practical perturbation bounds by using the matrix equation approach.

Theorem 2.3.6 *With $A = R^T R$ and the same assumptions as in Lemma* 2.3.1, *for the perturbation ΔA and result $\tilde{R}$ in* (2.3.17) *we have*

$$|\Delta R| \leq \epsilon \ \mathrm{up}(|\hat{R}^{-T}| e e^T |\hat{R}^{-1}|)|R| + O(\epsilon^2), \tag{2.3.33}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \chi'_C(A)\epsilon + O(\epsilon^2), \tag{2.3.34}$$

$$\chi_C(A) \leq \chi'_C(A), \tag{2.3.35}$$

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

$$\chi'_C(A) \equiv \inf_{D \in \mathbf{D}_n} \chi'_C(A, D), \tag{2.3.36}$$

$$\chi'_C(A, D) \equiv n \|\hat{R}^{-1}\|_2 \, \|\hat{R}^{-1} D\|_2 \, \|D^{-1} R\|_2 / \|R\|_2. \tag{2.3.37}$$

Proof. Let $G \equiv \Delta A/\epsilon$. Since $\rho(A^{-1}\Delta A) < 1$ (see the proof of Theorem 2.3.4), we can apply Theorem 2.2.1 here. From (2.2.6) with $R = \hat{R}D_c$, it follows that

$$\dot{R}(0) = \mathrm{up}(\hat{R}^{-T}D_c^{-1}GD_c^{-1}\hat{R}^{-1})R, \tag{2.3.38}$$

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

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

Then by the Taylor expansion (2.2.3), (2.3.33) follows immediately.

Also from (2.3.38) we have for any $D \in \mathbf{D}_n$ that

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

Thus taking the Frobenius norm we have

$$\|\dot{R}(0)\|_F \leq \|\hat{R}^{-1}\|_2 \|D_c^{-1}GD_c^{-1}\|_F \|\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2, \tag{2.3.39}$$

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

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

Since this holds for any $D \in \mathbf{D}_n$, (2.3.34) follows by using Taylor expansion (2.2.3).

It remains to prove (2.3.35). From (2.3.24) and (2.3.39) we have

$$\|\mathcal{D}_c\widehat{W}_R^{-1}\mathrm{duvec}(D_c^{-1}GD_c^{-1})\|_2 \leq \|\hat{R}^{-1}\|_2 \|D_c^{-1}GD_c^{-1}\|_F \|\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2. \tag{2.3.40}$$

Actually this holds for any symmetric $G \in \mathbf{R}^{n\times n}$ since it was essentially obtained from the matrix equation $R^TX + X^TR = G$ with X triangular by the two different approaches. Notice $\|\mathrm{duvec}(D_c^{-1}GD_c^{-1})\|_2 = \|D_c^{-1}GD_c^{-1}\|_F$, thus from (2.3.40) we must have

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

which implies (2.3.35). □

Notice that $\|\operatorname{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F$ for any *symmetric* X (see (1.2.7)), and $\|\hat{R}\|_2^2 = \|H\|_2$, then from (2.3.38) it is easy to obtain

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{1}{\sqrt{2}} n\|H^{-1}\|_2\epsilon + O(\epsilon^2), \tag{2.3.41}$$

which is essentially the first-order bound of Drmač, Omladič and Veselić [20, Theorem 3.1], except that their bound is rigorous and only the 2-norm is used. But this bound can severely overestimate the true relative errors of the computed Cholesky factor. In fact

$$\chi'_C(A) \le \chi'_C(A, I) = n\|H^{-1}\|_2, \tag{2.3.42}$$

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

$$\chi'_C(A) \le \chi'_C(A, D) = O(1/\delta), \qquad \|H^{-1}\|_2 = O(1/\delta^2).$$

Thus the new approximation $\chi'_C(A)$ to the condition number $\chi_C(A)$ is a significant improvement on that of Drmač et al. Furthermore, it is easy to see $\|H^{-1}\|_2$ is invariant if pivoting is used in computing the Cholesky factorization of A, whereas $\chi_C(A)$ and $\chi'_C(A)$ depend on any pivoting. Thus the new bounds (2.3.20) and (2.3.34) more closely reflect the true sensitivity of the Cholesky factorization than (2.3.41). However, if the ill-conditioning of R is mostly due to bad scaling of its columns, then $n\|H^{-1}\|_2$ is small, and is as good as $\chi'_C(A)$.

We see (2.3.42) implies $n\|H^{-1}\|_2$ is also an upper bound on $\chi_C(A)$. In fact we have the following stronger result

$$\chi_C(A) \le \frac{1}{\sqrt{2}} n\|H^{-1}\|_2, \tag{2.3.43}$$

which can easily be proved by using (2.3.24), (2.3.38) and the fact that for any *symmetric* X, $\|\operatorname{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F$. That is to say the first-order bound in (2.3.20)

is at least as small as that in (2.3.41). The example above suggests the former can be much smaller than the latter.

The practical outcome of Theorem 2.3.6 is that $\chi'_C(A)$ is quite easy to estimate. According to (1.2.10) in van der Sluis's Theorem 1.2.1, all we need to do is to choose $D = D_r = \text{diag}(\|R(i,:)\|_2)$ in $\chi'_C(A, D)$ in (2.3.37), then use norm estimators (see for example Higham [30, 1996, Ch. 14]) to estimate $\chi'_C(A, D_r)$ in $O(n^2)$.

Numerical experiments suggest usually $\chi'_C(A)$ is a reasonable approximation to $\chi_C(A)$. But the following example shows $\chi'_C(A)$ can still be much larger than $\chi_C(A)$, even though it can be much smaller than $n\|H^{-1}\|_2$. Let $A = \begin{bmatrix} 1 & \delta \\ \delta & \delta^2 + \delta^4 \end{bmatrix}$, then $R = \begin{bmatrix} 1 & \delta \\ 0 & \delta^2 \end{bmatrix}$. It is easy to show

$$\chi_C(A) = O(1), \qquad \chi'_C(A) = O(1/\delta), \qquad \|H^{-1}\|_2 = O(1/\delta^2).$$

In Theorem 2.3.6 the F- and 2-norms are used. If we use a monotone and consistent matrix norm $\|\cdot\|$, then from (2.3.38) it is straightforward to show we have the following perturbation bound instead of (2.3.34),

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

The advantage here is that the bound does not involve the scaling matrix D.

In Theorem 2.3.5 we established a relationship (2.3.30) between $\chi_C(A)$ and $\kappa_C(A)$ defined in Theorem 2.2.3. Is there a similar relationship between $\chi'_C(A, D)$ (or $\chi'_C(A)$) and $\kappa'_C(A, D)$ (or $\kappa'_C(A)$) defined in Theorem 2.2.4? The answer is yes.

Theorem 2.3.7

$$\frac{1}{n}\chi'_C(A, D) \leq \kappa'_C(A, D) \leq \frac{\max_i a_{ii}}{\min_i a_{ii}}\chi'_C(A, D), \tag{2.3.44}$$

and from this,

$$\frac{1}{n}\chi'_C(A) \leq \kappa'_C(A) \leq \frac{\max_i a_{ii}}{\min_i a_{ii}}\chi'_C(A). \tag{2.3.45}$$

Proof.

$$\begin{aligned}
\frac{1}{n}\chi'_C(A,D) &\equiv \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2/\|R\|_2 \\
&= \|D_c R^{-1}\|_2 \|D_c R^{-1} D\|_2 \|D^{-1}R\|_2/\|R\|_2 \\
&\leq \|D_c\|_2^2 \|R^{-1}\|_2 \|R^{-1}D\|_2 \|D^{-1}R\|_2/\|R\|_2 \\
&\leq \|R\|_2 \|R^{-1}\|_2 \|R^{-1}D\|_2 \|D^{-1}R\|_2 \quad (\text{using } \|D_c\|_2 \leq \|R\|_2) \\
&\equiv \kappa'_C(A,D).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\kappa'_C(A,D) &\equiv \|R\|_2 \|R^{-1}\|_2 \|R^{-1}D\|_2 \|D^{-1}R\|_2 \\
&= \|R\|_2 \|D_c^{-1}\hat{R}^{-1}\|_2 \|D_c^{-1}\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2 \\
&\leq \|R\|_2^2 \|D_c^{-1}\|_2^2 \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2/\|R\|_2 \\
&= \frac{1}{n}\|A\|_2 \|D_c^{-1}\|_2^2 \chi'_C(A,D) \\
&\leq \frac{\max_i a_{ii}}{\min_i a_{ii}} \chi'_C(A,D), \quad (\text{using } \|A\|_2 \leq n \max_i a_{ii}).
\end{aligned}$$

The proof is complete. □

The standard pivoting strategy can usually improve $\chi_C(A)$ too. In fact we have the following theorem.

Theorem 2.3.8 *Let $A \in \mathbf{R}^{n\times n}$ be symmetric positive definite with the Cholesky factorization $PAP^T = R^T R$ when the standard pivoting strategy is used. Then*

$$\chi_C(PAP^T) \leq \chi'_C(PAP^T) \leq \|H^{-1}\|_2^{1/2} n\sqrt{2n(n+1)(4^n+6n-1)/6}, \qquad (2.3.46)$$

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

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

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

and so from (2.3.37) we obtain

$$\chi'_C(PAP^T) \le n \inf_{D\in\mathbf{D}_n} \|\hat{R}^{-1}\|_2 \kappa_2(D^{-1}R). \tag{2.3.47}$$

Standard pivoting ensures $|r_{ii}| \ge |r_{ij}|$ for all $j \ge i$. Then (2.3.46) follows immediately by $\|\hat{R}^{-1}\|_2 = \|H^{-1}\|_2^{1/2}$ and (1.2.18) in Theorem 1.2.2. □

In Theorem 2.2.5 we have

$$\kappa_C(PAP^T) \le \kappa'_C(PAP^T) \le \kappa_2^{1/2}(A)\sqrt{2n(n+1)(4^n+6n-1)}/6.$$

By van der Sluis's result (1.2.16),

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

where it is possible that

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

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

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

and it is possible that

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

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

Suppose the Cholesky factorization of A be approached by using the standard symmetric pivoting strategy: $PAP^T = R^TR$. If the permutation matrix P is known beforehand and the Cholesky factorization is applied to PAP^T directly, it is easy to observe that Lemma 2.3.1 still holds except that now D_c and H should be redefined as

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

and $A + \Delta A = \tilde{R}^T \tilde{R}$ in (2.3.17) should be replaced by

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

Then by Theorem 2.3.4 we have

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

This with (2.3.46) suggests that the computed Cholesky factor $\tilde{R}$ has high accuracy. However, usually the permutation matrix P for A is not known beforehand. The permutation matrix for $A + \Delta A$, which is produced in the computing process, is possibly different from that for A. Fortunately, Higham [30, 1996, Lemma 10.11] showed that if there are no ties in the pivoting strategy for $PAP^T = R^T R$, then for sufficiently small ΔA, the two permutation matrices are the same. Thus the Cholesky factorization with standard symmetric pivoting will most likely gives an $\tilde{R}$ which is about as accurate as possible.

2.3.3 Rigorous perturbation bound with backward rounding errors

Drmač, Omladič and Veselić [20, 1994] obtained rigorous perturbation bounds. For comparison here we also present our rigorous perturbation bounds, which can be obtained by applying the results in Section 2.2.2.

Theorem 2.3.9 *Let $A \in \mathbf{R}^{n\times n}$ be a symmetric positive definite floating point matrix, with the exact Cholesky factorization $A = R^T R$. Define $D_c \equiv \mathrm{diag}(A)^{1/2}$ and $\hat{R} \equiv RD_c^{-1}$. If*

$$n\epsilon \, \|\mathcal{D}_c \widehat{W}_{\hat{R}}^{-1}\|_2^2 / \min_i a_{ii} < 1/4, \tag{2.3.48}$$

where $\epsilon \equiv (n+1)u/(1-2(n+1)u)$ with u being the unit round-off, and $\mathcal{D}_c$ is defined by (2.3.21), *then the Cholesky factorization applied to A succeeds (barring underflow*

and overflow) and produces a nonsingular $\tilde{R}$, which satisfies

$$\begin{aligned} \|\Delta R\|_2 &\leq \frac{2n\epsilon \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2}{1 + \sqrt{1 - 4n\epsilon \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2^2 / \min_i a_{ii}}} \\ &\leq 2n\epsilon \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2, \end{aligned} \tag{2.3.49}$$

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

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq 2\chi_C(A)\epsilon. \tag{2.3.50}$$

Proof. Let $H \equiv D_c^{-1} A D_c^{-1}$ as before. Then $H = \hat{R}^T \hat{R}$. From (2.3.29) and (2.3.27), we get

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

which with the assumption (2.3.48) implies that (2.3.16) holds. Thus by Lemma 2.3.1 the Cholesky factorization applied to A succeeds and the computed Cholesky factor $\tilde{R}$ satisfies

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

where $d_i = a_{ii}^{1/2}$. Then using $\widehat{W}_R = \widehat{\mathcal{D}}_c \widehat{W}_{\hat{R}} \mathcal{D}_c^{-1}$ (see (2.3.31)) and $\mathrm{uvec}(D_c^{-1} \Delta A D_c^{-1}) = \widehat{\mathcal{D}}_c^{-1} \mathrm{uvec}(\Delta A)$, which is easily verified, we have

$$\begin{aligned} \|\widehat{W}_R^{-1}\|_2 \|\widehat{W}_R^{-1} \mathrm{duvec}(\Delta A)\|_2 &\leq \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1} \widehat{\mathcal{D}}_c^{-1}\|_2 \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2 \|\widehat{\mathcal{D}}_c^{-1} \mathrm{duvec}(\Delta A)\|_2 \\ &\leq \epsilon \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2^2 \|\widehat{\mathcal{D}}_c^{-1}\|_2 \|D_c^{-1} dd^T D_c^{-1}\|_F \\ &\leq n\epsilon \|\mathcal{D}_c \widehat{W}_{\tilde{R}}^{-1}\|_2^2 / \min_i a_{ii}, \end{aligned}$$

which with the assumption (2.3.48) implies that the condition (2.2.47) of Theorem 2.2.7 is satisfied. Then (2.3.49) is immediately obtained by using Theorem 2.2.7. □

Theorem 2.3.10 *Let $A \in \mathbf{R}^{n \times n}$ be a symmetric positive definite floating point matrix with the exact Cholesky factorization $A = R^T R$. Define $D_c \equiv \mathrm{diag}(A)^{1/2}$ and $\hat{R} \equiv$*

RD_c^{-1}. *Assume* $D \in \mathbf{D}_n$. *If*

$$n\epsilon \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}\|_2 < 1/4, \tag{2.3.51}$$

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

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{2n\epsilon \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}R\|_2/\|R\|_2}{1+\sqrt{1-4n\epsilon \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}\|_2}} \tag{2.3.52}$$

$$\leq 2\chi_c'(A,D)\epsilon, \tag{2.3.53}$$

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

Proof. Let $H \equiv D_c^{-1}AD_c^{-1}$. Since

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

which with (2.3.51) implies that (2.3.16) holds. Thus by Lemma 2.3.1 Cholesky factorization applied to A succeeds and the computed Cholesky factor $\tilde{R}$ satisfies

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

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

$$\begin{aligned} \|R^{-T}\Delta A R^{-1}D\|_F \|D^{-1}\|_2 &\leq \|\hat{R}^{-1}\|_2 \|D_c^{-1}\Delta A D_c^{-1}\|_F \|\hat{R}^{-1}D\|_F \|D^{-1}\|_2 \\ &\leq n\epsilon \|\hat{R}^{-1}\|_2 \|\hat{R}^{-1}D\|_2 \|D^{-1}\|_2 < 1/4. \end{aligned}$$

Hence the bound (2.3.52) can be easily obtained from (2.2.55) in Theorem 2.2.8. □

If we take $D = I$, it is easy to show by using the fact that $\|\operatorname{up}(X)\|_F \leq \frac{1}{\sqrt{2}}\|X\|_F$ for any symmetric $X \in \mathbf{R}^{n\times n}$ that the assumption (2.3.51) can be weaken to

$$n\|H^{-1}\|_2 \epsilon \leq 1/2, \tag{2.3.54}$$

Table 2.3.1: Results for Pascal matrices without pivoting

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

and we have the following bound:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \frac{\sqrt{2}n\|H^{-1}\|_2\epsilon}{1+\sqrt{1-2n\|H^{-1}\|_2\epsilon}} \leq \sqrt{2}n\|H^{-1}\|_2\epsilon, \tag{2.3.55}$$

which is a slightly stronger than (2.3.54) where $D = I$. An equivalent rigorous bound where only the 2-norm is used was obtained by Drmač, Omladič and Veselić [20, 1994].

As we pointed out in the comment following Theorem 2.3.6, with a correct choice of D, e.g., $D = D_r \equiv \mathrm{diag}(\|(R(i,:)\|_2)$, possibly $\chi'_C(A,D)$ can be much smaller than $\|H^{-1}\|_2$. So the bound (2.3.55) is potentially weak, although the condition (2.3.54) is not as constraining as (2.3.51).

2.3.4 Numerical experiments

In Sections 2.3.2 and 2.3.3 we presented new first-order and rigorous perturbation bounds for the changes caused by backward rounding errors for the Cholesky fac-

Table 2.3.2: Results for Pascal matrices with pivoting, $\tilde{A} \equiv PAP^T$

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

torization using two different approaches, defined $\chi_C(A) \equiv n\|\mathcal{D}_c\widehat{W}_{\hat{R}}^{-1}\|_2/\|A\|_2^{1/2}$ as the condition number of the problem, and suggested that usually $\chi_C(A)$ could be estimated in practice by $\chi'_C(A, D_r) \equiv n\|\hat{R}^{-1}\|_2\,\|\hat{R}^{-1}D_r\|_2\,\|D_r^{-1}R\|_2/\|R\|_2$, with $D_r = \mathrm{diag}(\|R(i,:)\|_2)$, which can be estimated by standard norm estimators in $O(n^2)$. Our bounds are potentially much smaller than the equivalent bound in Drmač, Omladič and Veselić [20, 1994]. Also we compare $\chi_C(A)$ with $\kappa_C(A)$, and compare corresponding estimators $\chi'_C(A, D)$ with $\kappa'_C(A, D)$ as well. These condition numbers and condition estimators satisfy the following inequalities (see (1.2.10), (2.3.30), (2.3.35), (2.3.36), (2.3.42), (2.3.43), and (2.3.44)):

$$\chi_C(A) \le \chi'_C(A) \le \chi'_C(A, D_r),$$

$$\chi_C(A) \le \tfrac{1}{\sqrt{2}}n\|H^{-1}\|_2, \quad \tfrac{1}{\sqrt{n}}\chi'_C(A, D_r) \le \chi'_C(A) \le n\|H^{-1}\|_2,$$

$$\tfrac{1}{n}\chi_C(A) \le \kappa_C(A) \le \tfrac{\max a_{ii}}{\min a_{ii}}\chi_C(A), \quad \tfrac{1}{n}\chi'_C(A, D_r) \le \kappa'_C(A, D_r) \le \tfrac{\max a_{ii}}{\min a_{ii}}\chi'_C(A, D_r).$$

Now we give a set of examples to show our findings. The matrices are $n \times n$ Pascal matrices, $n = 1, 2, \ldots, n$. The results are shown in Table 2.3.1 without pivoting and in Table 2.3.2 with pivoting.

Note in Tables 2.3.1 and 2.3.2 how $\frac{1}{\sqrt{2}} n \|H^{-1}\|_2$ can be worse than $\chi_C(A)$. In Table 2.3.2 pivoting is seen to give a significant improvement on $\chi_C(A)$. Also we observe from both Tables 2.3.1 and 2.3.2 that $\chi'_C(A)$ is a reasonable approximation of $\chi_C(A)$. We see $\chi_C(A)$ is smaller than $\kappa_C(A)$ for $n > 2$.

2.4 Summary and future work

Although with norm-bounded changes in A the Sun [46, 1991] and Stewart [41, 1993] first-order perturbation bound (2.2.8) 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 (2.2.25) obtained by the matrix-vector equation approach 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 usually leads to a near optimally conditioned factorization for a given A in $PAP^T = R^T R$. Because of the difficulty in understanding and computing $\kappa_C(A)$, there was need for, and fortunately we have been able to give by the matrix-equation approach, a simpler bound. Although the new bound (2.2.29) is somewhat weaker, it provides a computationally practical and useful estimate $\kappa'_C(A, D_r)$ 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 its lower bound $\frac{1}{2}\kappa_2^{1/2}(A)$.

For the perturbation ΔA which comes from the backward rounding error analysis, we first presented first-order (nearly) attainable bounds (see Theorems 2.3.4) by the

matrix-vector equation approach, then gave computationally practical bounds (see Theorem 2.3.6) by the matrix equation approach. Even though the latter are weaker than the former, both of them are (potentially) stronger than the corresponding equivalent bound (2.3.41) of Drmač et al. Our condition number $\chi_C(A)$ more closely reflects the true sensitivity of the problem. Also numerical experiments and analysis show that usually the standard symmetric pivoting strategy can significantly improve the condition number $\chi_C(A)$. So the computed Cholesky factor most likely has high accuracy when the standard symmetric pivoting is used.

With the relative changes in the elements of A (i.e. $|\Delta A| \leq \epsilon|A|$) we presented first-order perturbation analyses, which resulted in the condition number $\mu_C(A)$ and a practical and useful upper bound $\mu'_C(A)$ on $\mu_C(A)$.

In the future we would like to

- Investigate the ratio $\kappa_C(A)/\kappa'_C(A)$, which we suspect is bounded by the function of n alone, probably involving something like 2^n.
- Explore the effect of rank-revealing pivoting on κ_C in both theory and computations, and study the optimization problem $\min_P \kappa_C(PAP^T)$.
- Give a better approximation to $\chi_C(A)$ than our current $\chi'_C(A)$, which can sometimes overestimate $\chi_C(A)$, or alternatively look for other methods to estimate $\chi_C(A)$ efficiently.
- Give a better approximation to $\mu_C(A)$ than our current $\mu'_C(A)$, which can sometimes overestimate $\mu_C(A)$, or alternatively look for other methods to estimate $\mu_C(A)$ efficiently.

Chapter 3

The QR factorization

3.1 Introduction

The QR factorization is an important tool in matrix computations: given an $m \times n$ real matrix A with full column rank, there exists a unique $m \times n$ real matrix Q with orthonormal columns, and a unique nonsingular upper triangular $n \times n$ real matrix R with positive diagonal entries such that

$$A = QR.$$

The matrix Q is referred to as the orthogonal factor, and R the triangular factor.

Let ΔA be a real $m \times n$ matrix such that $A + \Delta A$ is still of full column rank, then $A + \Delta A$ has a unique QR factorization

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R).$$

The goal of the perturbation analysis for the QR factorization is to determine a bound on $\|\Delta Q\|$ (or $|\Delta Q|$) and $\|\Delta R\|$ (or $|\Delta R|$) in terms of (a bound on) $\|\Delta A\|$ (or $|\Delta A|$).

The perturbation analysis for the QR factorization has been considered by several authors. Given $\|\Delta A\|$, the first result was presented by Stewart [39, 1977]. That was further modified and improved by Sun [46, 1991]. Using different approaches Sun [46,

1991] and Stewart [41, 1993] gave *first-order* perturbation analyses. Recently a new rigorous perturbation bound for Q alone was given by Sun [49, 1995]. Given $|\Delta A|$, Sun [48, 1992] presented a *rigorous* analysis for the components of Q and R. For ΔA which has the form of the equivalent backward rounding error (componentwise form) from a numerically stable computation of the QR factorization, Zha [55, 1993] gave *a first-order* analysis.

The main goal of this chapter is to establish new first-order perturbation bounds given a bound on $\|\Delta A\|$, which are sharper than the equivalent results for the R factor in Sun [46, 1991] and Stewart [41, 1993], and more straightforward than the sharp result in Sun [49, 1995] for the Q factor, and present the corresponding condition numbers which more closely reflect the true sensitivity of the problem.

The rest of this chapter is organized as follows. In Section 3.2 we obtain expressions for $\dot{Q}(0)$ and $\dot{R}(0)$ in the QR factorization $A + tG = Q(t)R(t)$. These basic sensitivity expressions will be used to obtain our new perturbation bounds in Sections 3.3 and 3.4, but in Section 3.2 they are also used to derive Sun's results on the sensitivity of R and Q. In Section 3.3 we give a refined perturbation analysis for Q, showing in a simple way why the standard column pivoting strategy for A can be beneficial for certain aspects of the sensitivity of Q. In Section 3.4 we analyze the perturbation in R, first by the detailed and tight matrix-vector equation approach, then by the straightforward matrix equation approach. We give numerical results and suggest practical condition estimators in Section 3.5. Finally we summarize our findings and point out future work in Section 3.6.

3.2 Rate of change of Q and R, and previous results

Our perturbation bounds for Q will be tighter if we bound separately the perturbations along the column space of A and along its orthogonal complement. Thus we introduce the following notation. For real $m \times n$ A, let P_1 be the orthogonal projector onto $\mathcal{R}(A)$, and P_2 be the orthogonal projector onto $\mathcal{R}(A)^\perp$. For real $m \times n$ ΔA define

$$\epsilon \equiv \|\Delta A\|_F/\|A\|_2, \quad \epsilon_1 \equiv \|P_1\Delta A\|_F/\|A\|_2, \quad \epsilon_2 \equiv \|P_2\Delta A\|_F/\|A\|_2, \tag{3.2.1}$$

so $\epsilon^2 = \epsilon_1^2 + \epsilon_2^2$.

Here we derive the basic results on how Q and R change as A changes. We then derive the first-order results obtained by Sun [46, 1991][49, 1995]. The following theorem summarizes the results we use later.

Theorem 3.2.1 *Let $A \in \mathbf{R}^{m\times n}$ be of full column rank n with the QR factorization $A = QR$, let G be a real $m \times n$ matrix, and let $\Delta A = \epsilon G$, for some $\epsilon \geq 0$. If*

$$\kappa_2(A)\frac{\|P_1\Delta A\|_2}{\|A\|_2} < 1, \tag{3.2.2}$$

where $\kappa_2(A) \equiv \|A^\dagger\|_2\|A\|_2$ and P_1 is the orthogonal projector onto $\mathcal{R}(A)$, then $A+\Delta A$ has the unique QR factorization

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R), \tag{3.2.3}$$

with ΔR and ΔQ satisfying

$$\Delta R = \epsilon\dot{R}(0) + O(\epsilon^2), \tag{3.2.4}$$

$$\Delta Q = \epsilon\dot{Q}(0) + O(\epsilon^2), \tag{3.2.5}$$

where $\dot{R}(0)$ and $\dot{Q}(0)$ are defined by the unique QR factorization

$$A(t) \equiv A + tG = Q(t)R(t), \quad Q^T(t)Q(t) = I, \quad |t| \leq \epsilon, \tag{3.2.6}$$

and so satisfy the equations

$$R^T\dot{R}(0) + \dot{R}^T(0)R = R^TQ^TG + G^TQR, \tag{3.2.7}$$

$$\dot{R}(0) = \text{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]R, \tag{3.2.8}$$

$$\dot{Q}(0) = GR^{-1} - Q\,\text{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T], \tag{3.2.9}$$

where the 'up' notation is defined by (1.2.3).

Proof. Take any $\bar{Q}$ such that $[Q, \bar{Q}]$ is square and orthogonal, then for all $|t| \leq \epsilon$

$$A + tG = [Q, \bar{Q}]\begin{bmatrix} R \\ 0 \end{bmatrix} + tG = [Q, \bar{Q}]\begin{bmatrix} R + tQ^TG \\ t\bar{Q}^TG \end{bmatrix}.$$

From the inequality (3.2.2) we see $\|tQ^TG\|_2 \leq \sigma_{\min}(A) = \sigma_{\min}(R)$ for all $|t| \leq \epsilon$. Thus $A + tG$ has full column rank and the unique QR factorization (3.2.6). Notice that $R(0) = R$, $R(\epsilon) = R + \Delta R$, $Q(0) = Q$ and $Q(\epsilon) = Q + \Delta Q$, so (3.2.3) holds.

It is easy to verify that $Q(t)$ and $R(t)$ are twice continuously differentiable for $|t| \leq \epsilon$ from the algorithm for the QR factorization. If we differentiate $R(t)^TR(t) = A(t)^TA(t)$ with respect to t and set $t = 0$, and use $A = QR$, we obtain (3.2.7) which we will see is a linear equation ***uniquely*** defining the elements of upper triangular $\dot{R}(0)$ in terms of the elements of Q^TG. From upper triangular $\dot{R}(0)R^{-1}$ in

$$\dot{R}(0)R^{-1} + (\dot{R}(0)R^{-1})^T = Q^TGR^{-1} + (Q^TGR^{-1})^T,$$

we see with the 'up' notation (see (1.2.3)) that (3.2.8) holds. Next differentiating (3.2.6) at $t = 0$ gives

$$G = Q\dot{R}(0) + \dot{Q}(0)R,$$

and combining this with (3.2.8) gives (3.2.9). Finally the Taylor expansions for $R(t)$ and $Q(t)$ about $t = 0$ give (3.2.4) and (3.2.5) at $t = \epsilon$. □

By Theorem 3.2.1 we can easily obtain the first-order perturbation bound for R given by Sun [46, 1991] and also by Stewart [41, 1993], and the first-order bound for Q given by Sun [49, 1995].

Theorem 3.2.2 *Let $A \in \mathbf{R}^{m\times n}$ be of full column rank n with the QR factorization $A = QR$, and let ΔA be a real $m \times n$ matrix. Define $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ and $\epsilon_1 \equiv \|QQ^T\Delta A\|_F/\|A\|_2$, see (3.2.1). If (3.2.2) holds, then $A + \Delta A$ has the unique QR factorization*

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R),$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1 + O(\epsilon^2). \tag{3.2.10}$$

$$\|\Delta Q\|_F \le \sqrt{2}\kappa_2(A)\epsilon + O(\epsilon^2). \tag{3.2.11}$$

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

$$\|G\|_F = \|A\|_2 = \|R\|_2. \tag{3.2.12}$$

Clearly all the conclusions of Theorem 3.2.1 hold here. From (3.2.8) and the fact that for symmetric X, $\|\operatorname{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F$ (see (1.2.7)) we have

$$\|\dot{R}(0)\|_F \le \frac{1}{\sqrt{2}}\|Q^TGR^{-1} + (Q^TGR^{-1})^T\|_F\,\|R\|_2 \tag{3.2.13}$$

$$\le \sqrt{2}\,\|Q^TGR^{-1}\|_F\,\|R\|_2 \le \sqrt{2}\,\kappa_2(R)\|Q^TG\|_F, \tag{3.2.14}$$

and since

$$\|Q^TG\|_F = \|Q^T\Delta A\|_F/\epsilon = \|A\|_2\,\epsilon_1/\epsilon = \|R\|_2\,\epsilon_1/\epsilon \tag{3.2.15}$$

and $\kappa_2(R) = \kappa_2(A)$,

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1/\epsilon.$$

Thus (3.2.10) follows from the Taylor expression (3.2.4).

If $[Q, \bar{Q}]$ is square and orthogonal, then from (3.2.9) we have

$$\begin{aligned}
\|\dot{Q}(0)\|_F^2 &= \|Q^T\dot{Q}(0)\|_F^2 + \|\bar{Q}^T\dot{Q}(0)\|_F^2 \\
&= \|Q^TGR^{-1} - \operatorname{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]\|_F^2 + \|\bar{Q}^TGR^{-1}\|_F^2 \\
&\le 2\,\|Q^TGR^{-1}\|_F^2 + \|\bar{Q}^TGR^{-1}\|_F^2 \qquad \text{(using (1.2.6))} \\
&\le 2\,\|GR^{-1}\|_F^2,
\end{aligned}$$

thus with (3.2.12),

$$\|\dot{Q}(0)\|_F \leq \sqrt{2}\kappa_2(A),$$

which, with the Taylor expression (3.2.5), gives the bound (3.2.11) for the Q factor. □

We see $\sqrt{2}\kappa_2(A)$ is a measure for the sensitivity of both R and Q, but it is not the actual condition number since for general A the first-order bounds in (3.2.10) and (3.2.11) are not attainable. Thus $\sqrt{2}\kappa_2(A)$ is a condition estimator for both R and Q in the QR factorization.

3.3 Refined analysis for Q

The results of Sun [49, 1995] give about as good as possible overall bounds on the change ΔQ in Q. But by looking at how ΔQ is distributed between $\mathcal{R}(Q)$ and its orthogonal complement, and following the ideas in the proof of Theorem 3.2.2, we are able to obtain a result which is tight but, unlike the related tight result in [49], easy to follow. It makes clear exactly where any ill-conditioning lies. From (3.2.5) with $Q = [Q, \bar{Q}]$ square and orthogonal,

$$\Delta Q = \epsilon Q Q^T \dot{Q}(0) + \epsilon \bar{Q}\bar{Q}^T \dot{Q}(0) + O(\epsilon^2),$$

and the key is to bound the first term on the right separately from the second. Note from (3.2.9) with $G = \Delta A/\epsilon$ and (3.2.1) that

$$\|\bar{Q}\bar{Q}^T \dot{Q}_1(0)\|_F = \|\bar{Q}\bar{Q}^T G R^{-1}\|_F \leq \|R^{-1}\|_2 \, \|\bar{Q}\bar{Q}^T G\|_F = \kappa_2(A)\epsilon_2/\epsilon,$$

where G can be chosen to give equality here for any given A. Hence

$$\|\bar{Q}\bar{Q}^T \Delta Q\|_F \leq \kappa_2(A)\epsilon_2 + O(\epsilon^2), \tag{3.3.1}$$

and for that part of ΔQ in $\mathcal{R}(\bar{Q})$ the condition number (with respect to the combination of the F- and 2-norms)

$$\kappa_{Q^\perp}(A) \equiv \lim_{\epsilon\to 0}\sup\Big\{\frac{\|P_2\Delta Q\|_F}{\epsilon_2\|Q\|_2}: \quad A+\Delta A=(Q+\Delta Q)(R+\Delta R), \\ \epsilon=\|\Delta A\|_F/\|A\|_2,\ \epsilon_2=\|P_2\Delta A\|_F/\|A\|_2\Big\}$$

is given by

$$\kappa_{Q^\perp}(A)=\kappa_2(A).$$

Now for the part of ΔQ in $\mathcal{R}(Q)$. We see from (3.2.9) that

$$\begin{aligned} Q^T\dot{Q}_1(0) &= Q^TGR^{-1}-\mathrm{up}[Q^TGR^{-1}+(Q^TGR^{-1})^T] \\ &= \mathrm{low}(Q^TGR^{-1})-[\mathrm{low}(Q^TGR^{-1})]^T, \end{aligned} \tag{3.3.2}$$

which is skew symmetric with clearly zero diagonal. If we partition Q, G and R as follows

$$Q=\overset{n-1}{[}\overset{1}{Q_{n-1},q}],\quad G=\overset{n-1}{[}\overset{1}{G_{n-1},g}],\quad R=\begin{bmatrix} \overset{n-1}{R_{n-1}} & \overset{1}{r} \\ & r_{nn}\end{bmatrix},$$

then from (3.3.2) we have

$$\begin{aligned} Q^T\dot{Q}(0) &= \mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1},Q^T(-G_{n-1}R_{n-1}^{-1}r+g)/r_{nn}]) \\ &\quad -\{\mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1},Q^T(-G_{n-1}R_{n-1}^{-1}r+g)/r_{nn}])\}^T \\ &= \mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1},0])-\{\mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1},0])\}^T. \end{aligned} \tag{3.3.3}$$

Thus taking the Frobenius norm and using (1.2.6) and (3.2.15) gives

$$\begin{aligned} \|QQ^T\dot{Q}(0)\|_F &= \|Q^T\dot{Q}(0)\|_F\le\sqrt{2}\,\|Q^TG_{n-1}R_{n-1}^{-1}\|_F \\ &\le \sqrt{2}\,\|Q^TG\|_F\,\|R_{n-1}^{-1}\|_F=\sqrt{2}\,\|R_{n-1}^{-1}\|_2\,\|A\|_2\,\epsilon_1/\epsilon. \end{aligned}$$

It is easy to verify for any R_{n-1} that equalities are obtained by taking $G=(q_ny^T,0)$, with y nonzero such that $\|R_{n-1}^{-T}y\|_2=\|R_{n-1}^{-1}\|_2\,\|y\|_2$. It follows that the first-order bound is attainable in

$$\|QQ^T\Delta Q\|_F\le\sqrt{2}\,\|R_{n-1}^{-1}\|_2\,\|A\|_2\,\epsilon_1+O(\epsilon^2), \tag{3.3.4}$$

so for that part of ΔQ in $\mathcal{R}(Q)$ the condition number

$$\kappa_Q(A) \equiv \lim_{\epsilon \to 0} \sup \Big\{ \frac{\|P_1 \Delta Q\|_F}{\epsilon_1 \|Q\|_2} : \quad A + \Delta A = (Q + \Delta Q)(R + \Delta R), \\ \epsilon = \|\Delta A\|_F / \|A\|_2,\ \epsilon_1 = \|P_1 \Delta A\|_F / \|A\|_2 \Big\}$$

is given by

$$\kappa_Q(A) = \sqrt{2}\, \|R_{n-1}^{-1}\|_2\, \|A\|_2. \tag{3.3.5}$$

In some problems we are mainly (in fact *only*, if A is square and nonsingular) interested in the change in Q lying in $\mathcal{R}(Q)$, and this result shows its bound can be smaller than we previously thought. In particular if A has only one small singular value, and we use the standard column pivoting strategy in computing the QR factorization, then R_{n-1} will usually be quite well-conditioned compared with R, and we will have $\|R_{n-1}^{-1}\|_2 \|A\|_2 \ll \kappa_2(A)$. However for some special cases this may not be true, for example the Kahan matrix in Section 3.5, and then a rank revealing pivoting strategy such as in Hong and Pan [31, 1992] would be required to obtain such an improvement.

We now summarize the above results as the following theorem.

Theorem 3.3.1 *Let* $A = [Q, \bar{Q}] \begin{bmatrix} R \\ 0 \end{bmatrix}$ *be the QR factorization of* $A \in \mathbf{R}^{m \times n}$ *with full column rank, and let* ΔA *be a real* $m \times n$ *matrix. Let* $\epsilon \equiv \|\Delta A\|_F / \|A\|_2$, $\epsilon_1 \equiv \|QQ^T \Delta A\|_F / \|A\|_2$ *and* $\epsilon_2 \equiv \|\bar{Q}\bar{Q}^T \Delta A\|_F / \|A\|_2$. *If* (3.2.2) *holds, then there is a unique QR factorization satisfying*

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R),$$

where

$$\|QQ^T \Delta Q\|_F \le \kappa_Q(A)\epsilon_1 + O(\epsilon^2),$$
$$\|\bar{Q}\bar{Q}^T \Delta Q\|_F \le \kappa_2(A)\epsilon_2 + O(\epsilon^2),$$

with

$$\kappa_Q(A) = \sqrt{2}\, \|R_{n-1}^{-1}\|_2\, \|A\|_2 \le \sqrt{2}\, \kappa_2(A). \qquad \square$$

3.4 Perturbation analyses for R

In Section 3.2 we saw the key to deriving first-order perturbation bounds for R in the QR factorization of full column rank A is the equation (3.2.7). Like in Chapter 2 for the Cholesky factor, we will now analyze the equation in two ways. The first, the matrix-vector equation approach, gives a sharp bound leading to the condition number $\kappa_R(A)$ for R in the QR factorization of A; while the second, the matrix equation approach, gives a clear improvement on (3.2.10), and provides an upper bound on $\kappa_R(A)$. Both approaches provide efficient condition estimators (see Chang and Paige [9, 1995] for the matrix–vector equation approach), and nice results for the special case of $AP = QR$, where P is a permutation matrix giving the standard column pivoting, but we will only derive the matrix equation versions of these. The tighter but more complicated matrix–vector equation analysis for the case of pivoting is given in [9], and only the results will be quoted here. All our analyses in this section are based on the same assumptions as in Theorem 3.2.2. Most of the results to be given here have been presented in Chang, Paige and Stewart [14, 1996].

3.4.1 Matrix-vector equation analysis for R

The matrix-vector equation approach views the matrix equation (3.2.7) as a large matrix-vector equation. The upper and lower triangular parts of (3.2.7) contain identical information. By using the 'uvec' notation defined by (1.2.4) and 'vec' notation, we can easily show the upper triangular part can be rewritten in the following form (for the derivation, see [14]):

$$W_R \,\mathrm{uvec}(\dot{R}(0)) = Z_R \,\mathrm{vec}(Q^T G), \tag{3.4.1}$$

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

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

and $Z_R \in \mathcal{R}^{\frac{n(n+1)}{2} \times n^2}$ is

$$\left[\begin{array}{cccc|cccc|c|cccc}
r_{11} & & & & & & & & & & & \\
\hline
r_{12} & r_{22} & & & r_{11} & & & & & & & \\
 & & & & r_{12} & r_{22} & & & & & & \\
\hline
\cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot & & & \\
\hline
r_{1n} & r_{2n} & \cdot & r_{nn} & & & & & & r_{11} & & \\
 & & & & r_{1n} & r_{2n} & \cdot & r_{nn} & & r_{12} & r_{22} & \\
 & & & & & & & & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & r_{1n} & r_{2n} & \cdot r_{nn}
\end{array}\right].$$

Since R is nonsingular, W_R is also, and from (3.4.1)

$$\text{uvec}(\dot{R}(0)) = W_R^{-1} Z_R \, \text{vec}(Q^T G). \tag{3.4.2}$$

Remembering $\dot{R}(0)$ is upper triangular, we see

$$\|\dot{R}(0)\|_F = \|\text{uvec}(\dot{R}(0))\|_2 = \|W_R^{-1} Z_R \, \text{vec}(Q^T G)\|_2$$

$$\leq \|W_R^{-1}Z_R\|_2 \|\mathrm{vec}(Q^TG)\|_2 = \|W_R^{-1}Z_R\|_2 \|Q^TG\|_F$$
$$= \|W_R^{-1}Z_R\|_2 \|R\|_2 \epsilon_1/\epsilon, \qquad \text{(using (3.2.15))}$$

where for *any* nonsingular upper triangular R, equality can be obtained by choosing G such that $\mathrm{vec}(Q^TG)$ lies in the space spanned by the right singular vectors corresponding to the largest singular value of $W_R^{-1}Z_R$. Therefore we see from the Taylor expansion (3.2.4),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \|W_R^{-1}Z_R\|_2 \epsilon_1 + O(\epsilon^2),$$

and this bound is attainable to first order in ϵ. This implies for R in the QR factorization of A the condition number (with respect to the combination of the F- and 2-norms) defined by

$$\kappa_R(A) \equiv \lim_{\epsilon \to 0} \sup \Big\{ \frac{\|\Delta R\|_F}{\epsilon_1 \|R\|_2} : \quad (A + \Delta A) = (Q + \Delta Q)(R + \Delta R)$$
$$\epsilon = \|\Delta A\|_F / \|A\|_2,\ \epsilon_1 = \|P_1 \Delta A\|_F / \|A\|_2 \Big\} \quad (3.4.3)$$

is given by

$$\kappa_R(A) = \|W_R^{-1}Z_R\|_2.$$

From the definition of $\kappa_R(A)$ and the Sun's first-order perturbation bound (3.2.10) we easily observe

$$\kappa_R(A) \leq \sqrt{2}\kappa_2(A).$$

This upper bound is achieved if R is an identity matrix, and so is tight.

The structure of W_R and Z_R reveals that each column of W_R is one of the columns of Z_R, and so $W_R^{-1}Z_R$ has an $n(n+1)/2$ square identity submatrix, giving

$$\|W_R^{-1}Z_R\|_2 \geq 1. \qquad (3.4.4)$$

We now summarize these results as the following theorem.

Theorem 3.4.1 *With the same assumptions as in Theorem* 3.2.2, $A + \Delta A$ *has the unique QR factorization*

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R),$$

where with W_R and Z_R as in (3.4.1),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_R(A)\epsilon_1 + O(\epsilon^2), \tag{3.4.5}$$

$$1 \leq \kappa_R(A) = \|W_R^{-1} Z_R\|_2 \leq \sqrt{2}\kappa_2(A), \tag{3.4.6}$$

and the first-order bound in (3.4.5) *is attainable.* □

Unity is the best constant lower bound on $\kappa_R(A)$ we can obtain, as can be seen from the following example.

Example 1. Let $R = \text{diag}(1, \delta, \ldots, \delta^{n-1})$, $0 < \delta \leq 1$. We can easily show

$$1 \leq \|W_R^{-1} Z_R\|_2 = \sqrt{1+\delta^2} \to 1 \quad \text{as} \quad \delta \to 0. \qquad \square \tag{3.4.7}$$

From (3.4.6) we know the first-order perturbation bound in (3.4.5) is at least as good as as that in (3.2.10). In fact it can be better by an arbitrary factor. Consider Example 1,

$$\kappa_R(A) = \sqrt{1+\delta^2}, \qquad \kappa_2(A) = 1/\delta,$$

and

$$\frac{\sqrt{2}\kappa_2(A)}{\kappa_R(A)} \sim \frac{\sqrt{2}}{\delta} \quad \text{as} \quad \delta \to 0.$$

We see the first-order perturbation bound (3.2.10) can severely overestimate the effect of a perturbation in A.

Suppose we use the standard column pivoting strategy in $AP = QR$, where P is a permutation matrix designed so that columns of A are interchanged, during the computation of the reduction, to make the leading diagonal elements of R as large as possible, see Golub and Van Loan [26, 1996, §5.4] for details. We see $\kappa_2(A)$ in (3.2.10) does not change, but $\kappa_R(A)$ does change. The following result was shown by Chang and Paige [9, 1995].

Theorem 3.4.2 *Let $A \in \mathbf{R}^{m\times n}$ be of full column rank, with the QR factorization $AP = QR$ when the standard column pivoting strategy is used. Then*

$$1 \le \kappa_R(AP) = \|W_R^{-1}Z_R\|_2 \le \|W_R^{-1}Z_R\|_F \le \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}. \quad (3.4.8)$$

If $R = K_n(\theta)$, where $K_n(\theta)$ are the Kahan matrices (see (2.2.41)), then

$$\|W_R^{-1}Z_R\|_F \to \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}} \quad as \quad \theta \to 0. \qquad \square$$

Theorem 3.4.2 shows that when the standard column pivoting strategy is used, $\kappa_R(AP)$ is bounded for fixed n no matter how large $\kappa_2(A)$ is. Many numerical experiments with this strategy suggest that $\kappa_R(AP)$ is usually close to its lower bound of one. But it is not for the Kahan matrices. Fortunately such examples are rare in practice, and furthermore if we adopt the rank-revealing pivoting strategy, the condition number will most likely be close to its lower bound, see Section 3.5.

3.4.2 Matrix equation analysis for R

As far as we can see, $\kappa_R(A)$ is unreasonablely expensive to compute or estimate directly with the usual approach, except when we use pivoting, in which case $\kappa_R(AP)$ usually approaches its lower bound of 1. Fortunately, by the matrix equation approach we can obtain an excellent upper bound on $\kappa_R(A)$.

In the proof of Theorem 3.2.2 we used the expression of $\dot{R}(0)$ in (3.2.8) to derive Sun's first-order perturbation bound. Now we again look at (3.2.8), repeated here for clarity:

$$\dot{R}(0) = \mathrm{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]R.$$

Let $\mathbf{D}_n$ be the set of all $n \times n$ real positive definite diagonal matrices. For any $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathbf{D}_n$, let $R = D\bar{R}$. Note that for any matrix B we have $\mathrm{up}(B)D = \mathrm{up}(BD)$. Hence if we define $B \equiv Q^TG\bar{R}^{-1}$, then

$$\dot{R}(0) = \mathrm{up}[Q^TG\bar{R}^{-1} + D^{-1}(Q^TG\bar{R}^{-1})^TD]\bar{R} = [\mathrm{up}(B) + D^{-1}\,\mathrm{up}(B^T)D]\bar{R}. \quad (3.4.9)$$

With obvious notation, the upper triangular matrix $\mathrm{up}(B) + D^{-1}\,\mathrm{up}(B^T)D$ has (i,j) element $b_{ij} + b_{ji}\delta_j/\delta_i$ for $i < j$, and (i,i) element b_{ii}. To bound this, we use:

Lemma 3.4.1 *For $B \in \mathbf{R}^{n\times n}$ and $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathbf{D}_n$,*

$$\phi \equiv \|\,\mathrm{up}(B) + D^{-1}\,\mathrm{up}(B^T)D\|_F \le \sqrt{1+\zeta_D^2}\,\|B\|_F, \tag{3.4.10}$$

where

$$\zeta_D \equiv \max_{1\le i<j\le n}\{\delta_j/\delta_i\}. \tag{3.4.11}$$

Proof. Clearly

$$\phi^2 = \sum_{i=1}^{n} b_{ii}^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(b_{ij} + \frac{\delta_j}{\delta_i}b_{ji})^2.$$

But by the Cauchy-Schwarz theorem, $(b_{ij} + \frac{\delta_j}{\delta_i}b_{ji})^2 \le (b_{ij}^2 + b_{ji}^2)(1 + (\frac{\delta_j}{\delta_i})^2)$, so

$$\begin{aligned}
\phi^2 &\le \sum_{i=1}^{n} b_{ii}^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(b_{ij}^2 + b_{ji}^2)(1 + (\frac{\delta_j}{\delta_i})^2) \\
&= \|B\|_F^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(b_{ij}^2 + b_{ji}^2)(\frac{\delta_j}{\delta_i})^2 \\
&\le \|B\|_F^2 + \zeta_D^2\|B\|_F^2. \qquad (3.4.12)
\end{aligned}$$

From this (3.4.10) follows. □

We can now bound $\dot{R}(0)$ of (3.4.9)

$$\begin{aligned}
\|\dot{R}(0)\|_F &\le \phi\cdot\|\bar{R}\|_2 \le \sqrt{1+\zeta_D^2}\,\|B\|_F\|\bar{R}\|_2 \\
&= \sqrt{1+\zeta_D^2}\,\|Q^TG\bar{R}^{-1}\|_F\|\bar{R}\|_2 \le \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{R})\|Q^TG\|_F \qquad (3.4.13) \\
&= \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R)\|R\|_2\,\epsilon_1/\epsilon, \qquad \text{(using (3.2.15))}
\end{aligned}$$

or

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R)\epsilon_1/\epsilon.$$

But this is true for all $D \in \mathbf{D}_n$, so that

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \leq \kappa'_R(A)\epsilon_1/\epsilon, \tag{3.4.14}$$

$$\kappa'_R(A) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_R(A,D), \tag{3.4.15}$$

$$\kappa'_R(A,D) \equiv \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R), \tag{3.4.16}$$

where ζ_D is defined in (3.4.11). This gives the encouraging result

$$\kappa'_R(A) \leq \kappa'_R(A,I) = \sqrt{2}\kappa_2(R) = \sqrt{2}\kappa_2(A). \tag{3.4.17}$$

Hence from the Taylor expansion (3.2.4) we have

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa'_R(A)\epsilon_1 + O(\epsilon^2), \tag{3.4.18}$$

where from (3.4.17) this is never worse than the bound (3.2.10).

Clearly $\kappa'_R(A)$ is a measure of the sensitivity of R in the QR factorization. Since $\kappa_R(A)$ is the condition number for R (see the definition (3.4.3)), certainly we have

$$\kappa_R(A) \leq \kappa'_R(A), \tag{3.4.19}$$

Usually (3.4.19) is a strict inequality, but if R is diagonal equality will hold. In fact, take $D = R$, and let $\zeta_D = r_{jj}/r_{ii}$, $j > i$, then $\kappa'_R(A,D) = \sqrt{1+\zeta_D^2}$. On the other hand, it is straightforward to show $\kappa_R(A) = \sqrt{1+\zeta_D^2}$. So $\kappa_R(A) = \kappa'_R(A,D)$, which implies $\kappa_R(A) = \kappa'_R(A)$. For another proof for this, see Chang, Paige and Stewart [14, 1996, Remark 5.1].

Now we summarize the above results as the following theorem.

Theorem 3.4.3 *With the same assumptions as in Theorem* 3.2.2, $A + \Delta A$ *has the unique QR factorization satisfying*

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R),$$

where with $\kappa'_R(A)$ *as in* (3.4.15) *and* (3.4.16),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa'_R(A)\epsilon_1 + O(\epsilon^2), \tag{3.4.20}$$

$$\kappa_R(A) \le \kappa'_R(A) \le \sqrt{2}\kappa_2(A), \tag{3.4.21}$$

and if R is diagonal the first inequality in (3.4.21) *will become an equality.* □

From (3.4.21) we know the first-order perturbation bound (3.4.20) is at least as good as (3.2.10). In fact it can be better by an arbitrary factor, as can also be seen from Example 1. Taking $D = R$, we have

$$\kappa_R(A) = \kappa'_R(A) = \kappa'_R(A, D) = \sqrt{1+\delta^2}, \qquad \kappa_2(A) = 1/\delta.$$

If we take $R = \mathrm{diag}(\delta^{1-n}, \ldots, \delta, 1)$, $0 < \delta \le 1$, we see $\kappa_2(R) = \kappa_2(A) = \delta^{1-n}$, while

$$\kappa_R(A) = \kappa'_R(A) = \kappa'_R(A, D) = \sqrt{1+\delta^{2-2n}},$$

which is close to the upper bound $\sqrt{2}\kappa_2(A)$ for small δ. This shows that relatively small early diagonal elements of R cause poor condition, and suggests if we do not use pivoting, then there is a significant chance that the condition of the problem will approach its upper bound, at least for randomly chosen matrices.

With the standard column pivoting strategy in $AP = QR$, P a permutation matrix, this analysis also leads simply to a very nice result, even though it is a bit weaker than the tight result (3.4.8).

Theorem 3.4.4 *Let $A \in \mathbf{R}^{m\times n}$ be of full column rank, with the QR factorization $AP = QR$ when the standard column pivoting strategy is used. Then*

$$\kappa_R(AP) \le \kappa'_R(AP) \le \sqrt{n(n+1)(4^n + 6n - 1)}/3. \tag{3.4.22}$$

Proof. Standard column pivoting ensures $|r_{ii}| \ge |r_{ij}|$ and $|r_{ii}| \ge |r_{jj}|$ for all $j \ge i$. Since for any $D \in \mathbf{D}_n$,

$$\kappa'_R(AP) \le \kappa'_R(A, D) = \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R),$$

(3.4.22) is immediately obtained from (1.2.18) in Theorem 1.2.2 by taking $D = \mathrm{diag}(R)$. □

This analysis gives some insight as to why R in the QR factorization is less sensitive than the earlier condition estimator $\sqrt{2}\kappa_2(A)$ indicated. If the ill-conditioning of R is mostly due to bad scaling of its rows, then correct choice of D can give $\kappa_2(D^{-1}R)$ very near one. If at the same time ζ_D is not large, then $\kappa'_R(A,D)$ in (3.4.16) can be much smaller than $\sqrt{2}\kappa_2(R)$, see (3.4.17). Standard pivoting always ensures that such a D exists, and in fact gives (3.4.22).

The significance of this analysis is that it provides an excellent upper bound on $\kappa_R(A)$. $\kappa'_R(A)$ is quite easy to estimate. All we need to do is choose a suitable D in $\kappa'_R(A,D)$ in (3.4.16). We consider how to do this in the next section.

3.5 Numerical experiments

In Section 3.4 we presented new first-order perturbation bounds for the R factor of the QR factorization using two different approaches, obtained the condition number $\kappa_R(A) = \|W_R^{-1}Z_R\|_2$ for the R factor, and suggested $\kappa_R(A)$ could be estimated in practice by $\kappa'_R(A,D)$. Our new first-order results are sharper than previous results for R, and at least as sharp for Q, and we give some numerical experiments to illustrate both this, and the possible estimators for $\kappa_R(A)$.

We would like to choose D such that $\kappa'_R(A,D)$ is a good approximation to the minimum $\kappa'_R(A)$ in (3.4.15), and show that this is a good estimate of the condition number $\kappa_R(A)$. Then a procedure for obtaining an $O(n^2)$ condition estimator for R in the QR factorization (i.e. an estimator for $\kappa_R(A)$), is to choose such a D, use a standard condition estimator (see for example Higham [27, 1987]) to estimate $\kappa_2(D^{-1}R)$, and take $\kappa'_R(A,D)$ in (3.4.16) as the appropriate estimate.

By van der Sluis's Theorem 1.2.1, $\kappa_2(D^{-1}R)$ will be nearly minimal when the rows of $D^{-1}R$ are equilibrated. But this could lead to a large ζ_D in (3.4.16). There are three obvious possibilities for D. The first one is choosing D to equilibrate R precisely. Specifically, take $\delta_i = \sqrt{\sum_{j=i}^{n} r_{ij}^2}$ for $i = 1,\ldots,n$. The second one is

choosing D to equilibrate R as far as possible while keeping $\zeta_D \leq 1$. Specifically, take $\delta_1 = \sqrt{\sum_{j=1}^{n} r_{1j}^2}$, $\delta_i = \sqrt{\sum_{j=i}^{n} r_{ij}^2}$ if $\sqrt{\sum_{j=i}^{n} r_{ij}^2} \leq \delta_{i-1}$ otherwise $\delta_i = \delta_{i-1}$, for $i = 2, \ldots, n$. The third one is choosing $\delta_i = r_{ii}$. Computations show that the third choice can sometimes cause unnecessarily large estimates, so we will not give any results for that choice. We specify the diagonal matrix D obtained by the first method and the second method by D_1 and D_2 respectively in the following.

We give three sets of examples.

(1) The first set of matrices are $n \times n$ Pascal matrices, $n = 1, 2, \ldots, 15$. The results are shown in Table 3.5.1 without pivoting, and in Table 3.5.2 with standard column pivoting. Table 3.5.1 illustrate how the upper bound $\sqrt{2}\kappa_2(A)$ can be far worse than the condition number $\kappa_R(A)$, which itself can be much greater than its lower bound of 1. In Table 3.5.2 standard column pivoting is seen to give a significant improvement on $\kappa_R(A)$, bringing $\kappa_R(AP)$ very close to its lower bound, but of course $\sqrt{2}\kappa_2(AP) = \sqrt{2}\kappa_2(A)$ still. Also we observe from Table 3.5.1 that both $\kappa'_R(A, D_1)$ and $\kappa'_R(A, D_2)$ are very good estimates for $\kappa_R(A)$. The latter is a little better than the former. In Table 3.5.2 $\kappa_R(AP, D_1) = \kappa_R(AP, D_2)$ (in fact $D_1 = D_2$), and they are also good estimates for $\kappa_R(AP)$.

(2) The second set of matrices are 10×8 A_j, $j = 1, 2, \ldots, 8$, which are all obtained from the same random 10×8 matrix (produced by the MATLAB function randn), but with its jth column multiplied by 10^{-8} to give A_j. The results are shown in Table 3.5.3 without pivoting. All the results with pivoting are similar to that for $j = 8$ in Table 3.5.3, and so are not given here. For $j = 1, 2 \ldots, 7$, $\kappa_R(A)$ and $\kappa_Q(A)$ are both close to their upper bound $\sqrt{2}\kappa_2(A)$, but for $j = 8$, both $\kappa_R(A)$ and $\kappa_Q(A)$ are significantly smaller than $\sqrt{2}\kappa_2(A)$. All these results are what we expected, since the matrix R is ill-conditioned due to the fact that r_{jj} is very small, but for $j = 1, 2, \ldots, 7$ the rows of R are already essentially equilibrated, and we do not expect $\kappa_R(A)$ to be much better than $\sqrt{2}\kappa_2(A)$. Also for the first seven cases the smallest singular value of the leading part R_{n-1} is close to that of R, so that

Table 3.5.1: Results for Pascal matrices without pivoting, $A = QR$

n	$\kappa_R(A)$	$\kappa'_R(A, D_1)$	$\kappa'_R(A, D_2)$	$\kappa_Q(A)$	$\sqrt{2}\kappa_2(A)$
1	1.0e+00	1.4e+00	1.4e+00	—	1.4e+00
2	1.9e+00	3.4e+00	1.9e+00	2.6e+00	9.7e+00
3	4.6e+00	1.4e+01	1.4e+01	1.9e+01	8.8e+01
4	1.4e+01	6.1e+01	6.1e+01	1.6e+02	9.8e+02
5	5.0e+01	2.6e+02	2.6e+02	1.6e+03	1.2e+04
6	1.8e+02	1.1e+03	1.1e+03	1.8e+04	1.6e+05
7	6.7e+02	4.5e+03	4.2e+03	2.2e+05	2.1e+06
8	2.5e+03	1.8e+04	1.7e+04	2.8e+06	2.9e+07
9	9.4e+03	7.4e+04	6.6e+04	3.6e+07	4.1e+08
10	3.6e+04	3.0e+05	2.6e+05	4.8e+08	5.9e+09
11	1.4e+05	1.2e+06	1.1e+06	6.6e+09	8.5e+10
12	5.2e+05	4.9e+06	4.2e+06	9.1e+10	1.2e+12
13	2.0e+06	2.0e+07	1.7e+07	1.3e+12	1.8e+13
14	7.8e+06	8.0e+07	6.6e+07	1.8e+13	2.7e+14
15	3.0e+07	3.2e+08	2.6e+08	2.6e+14	4.0e+15

$\kappa_Q(A)$ could not be much better than $\sqrt{2}\kappa_2(A)$. For $j = 8$, even though R is still ill-conditioned due to the fact that $r_{8,8}$ is very small, it is not at all equilibrated, and becomes well-conditioned by row scaling. Notice at the same time ζ_D is close to 1, so $\kappa'_R(A, D_1)$, $\kappa'_R(A, D_2)$, and therefore $\kappa_R(A)$ are much better than $\sqrt{2}\kappa_2(A)$. In this case, the smallest singular value of R is significantly smaller than that of R_{n-1}. Thus $\kappa_Q(A)$, the condition number for the change in Q lying in the range of Q, is spectacularly better than $\sqrt{2}\kappa_2(A)$. This is a contrived example, but serves to emphasize the benefits of pivoting for the condition of both Q and R.

(3) The third set of matrices are $n \times n$ Kahan matrices $A = K_n(\theta)$; see (2.2.41). Of course without pivoting $Q = I$ here, but the condition numbers depend on R only, and these are all we are interested in. The results for $n = 5, 10, 15, 20, 25$ with $\theta = \pi/8$ are shown in Table 3.5.4, where Π is a permutation such that the first column is moved to the last column position, and the remaining columns are moved to left

Table 3.5.2: Results for Pascal matrices with pivoting, $AP = \tilde{Q}\tilde{R}$

n	$\kappa_R(AP)$	$\kappa'_R(AP, D_1)$	$\kappa'_R(AP, D_2)$	$\kappa_Q(AP)$	$\sqrt{2}\kappa_2(A)$
1	1.0e+00	1.4e+00	1.4e+00	—	1.4e+00
2	1.2e+00	1.8e+00	1.8e+00	1.7e+00	9.7e+00
3	1.3e+00	2.2e+00	2.2e+00	1.3e+01	8.8e+01
4	1.7e+00	3.4e+00	3.4e+00	1.1e+02	9.8e+02
5	1.8e+00	4.1e+00	4.1e+00	1.0e+03	1.2e+04
6	2.2e+00	4.7e+00	4.7e+00	7.5e+03	1.6e+05
7	2.1e+00	5.1e+00	5.1e+00	8.5e+04	2.1e+06
8	2.6e+00	6.5e+00	6.5e+00	1.2e+06	2.9e+07
9	3.5e+00	8.8e+00	8.8e+00	1.5e+07	4.1e+08
10	3.4e+00	9.4e+00	9.4e+00	2.4e+08	5.9e+09
11	3.4e+00	9.2e+00	9.2e+00	2.3e+09	8.5e+10
12	3.3e+00	9.7e+00	9.7e+00	3.0e+10	1.2e+12
13	3.3e+00	1.1e+01	1.1e+01	3.5e+11	1.8e+13
14	3.6e+00	1.2e+01	1.2e+01	5.4e+12	2.7e+14
15	3.3e+00	1.2e+01	1.2e+01	8.6e+13	4.0e+15

Table 3.5.3: Results for 10×8 matrix A_j, $j = 1, \ldots, 8$, without pivoting

j	$\kappa_R(A)$	$\kappa'_R(A, D_1)$	$\kappa'_R(A, D_2)$	$\kappa_Q(A)$	$\sqrt{2}\kappa_2(A)$
1	1.9e+08	4.0e+08	3.0e+08	3.0e+08	4.8e+08
2	1.3e+08	2.9e+08	2.7e+08	2.6e+08	3.8e+08
3	1.9e+08	4.5e+08	3.9e+08	4.7e+08	5.5e+08
4	1.4e+08	3.1e+08	2.6e+08	2.9e+08	4.5e+08
5	1.2e+08	3.1e+08	2.4e+08	3.9e+08	4.2e+08
6	8.8e+07	2.2e+08	1.7e+08	3.5e+08	3.9e+08
7	9.3e+07	2.1e+08	1.7e+08	4.4e+08	5.5e+08
8	2.3e+00	5.5e+00	4.9e+00	6.6e+00	6.2e+08

Table 3.5.4: Results for Kahan matrices, $\theta = \pi/8$, $A\Pi = \hat{Q}\hat{R}$

n	$\kappa_R(A\Pi)$	$\kappa'_R(A\Pi, D_1)$	$\kappa_Q(A\Pi)$	$\kappa_R(A)$	$\kappa'_R(A, D_1)$	$\kappa_Q(A)$	$\sqrt{2}\kappa_2(A)$
5	2.3e+00	5.6e+00	3.9e+00	8.0e+00	1.7e+01	2.2e+02	1.1e+03
10	3.5e+00	1.4e+01	6.8e+03	2.1e+02	6.1e+02	1.0e+06	5.1e+06
15	4.9e+00	2.3e+01	1.0e+06	5.5e+03	2.1e+04	4.0e+09	2.0e+10
20	5.3e+00	3.2e+01	1.4e+08	1.5e+05	6.5e+05	1.5e+13	7.5e+13
25	6.1e+00	4.1e+01	2.0e+10	4.3e+06	2.0e+07	5.4e+16	2.7e+17

one position — this permutation Π is adopted in the rank-revealing QR factorization for Kahan matrices, see Hong and Pan [31, 1992]. Again we found $D_1 = D_2$, and only list the results corresponding to D_1. As we know the Kahan matrices correspond to correctly pivoted A by standard column pivoting. From Table 3.5.4 we see that in these extreme cases, with large enough n, $\kappa_R(A)$ can be large even with standard pivoting. This is about as bad a result as we can get with standard column pivoting (it gets a bit worse as $\theta \to 0$ in R), since the Kahan matrices make the upper bound on $\|W_R^{-1} Z_R\|_F$ approximately reachable, see Theorem 3.4.2. However if we use the rank-revealing pivoting strategy, we see from Table 3.5.4 $\kappa_R(A\Pi)$ is again close to its lower bound of 1. Also we see $\kappa_Q(A\Pi)$ is significantly smaller than $\kappa_Q(A)$. This is due to the fact that the smallest singular value of R_{n-1} is much small than that of $\hat{R}_{n-1}$, the leading $n-1$ square part of $\hat{R}$ in $A\Pi = \hat{Q}\hat{R}$. For both of the cases with and without rank-revealing pivoting, $\kappa'_R(A, D_1)$ still estimates $\kappa_R(A)$ excellently.

In all these examples we see $\kappa'_R(A, D_1)$ and $\kappa'_R(A, D_2)$ gave excellent estimates for $\kappa_R(A)$, with $\kappa'_R(A, D_2)$ being marginally preferable.

3.6 Summary and future work

The first-order perturbation analyses presented here show just what the sensitivity (condition) of each of Q and R is in the QR factorization of full column rank A, and

in so doing provide their condition numbers (with respect to the measures used, and for sufficiently small ΔA), as well as efficient ways of approximating these. The key norm-based condition numbers we derived for $A + \Delta A = (Q + \Delta Q)(R + \Delta R)$ are:

- $\kappa_{Q^\perp} = \kappa_2(A)$ for that part of ΔQ in $\mathcal{R}(A)^\perp$, see (3.3.1),
- $\kappa_Q(A) = \sqrt{2}\,\|R_{n-1}^{-1}\|_2\|A\|_2$ for that part of ΔQ in $\mathcal{R}(A)$, see (3.3.4),
- $\kappa_R(A) = \|W_R^{-1}Z_R\|_2$ for R, see Theorem 3.4.1,
- the estimate for $\kappa_R(A)$, that is $\kappa'_R(A) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_R(A, D)$, where $\kappa'_R(A, D) \equiv \sqrt{1 + \zeta_D^2}\,\kappa_2(D^{-1}R)$, see (3.4.3).

The condition numbers obey

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

for Q, while for R

$$1 \le \kappa_R(A) = \|W_R^{-1}Z_R\|_2 \le \kappa'_R(A) \le \sqrt{2}\kappa_2(A),$$

see (3.4.6) and (3.4.21). The numerical examples, and an analysis of the $n = 2$ case (not given here), suggest that $\kappa'_R(A, D)$, with D chosen to equilibrate $\bar{R} \equiv D^{-1}R$ subject to $\zeta_D \le 1$, gives an excellent approximation to $\kappa_R(A)$ in the general case. In the special case of A with orthogonal columns, so R is diagonal, then by taking $D = R$,

$$\kappa_R(A) = \kappa'_R(A) = \kappa'_R(A, D) = \sqrt{1 + \zeta_D^2} \le \sqrt{2}\kappa_2(D) = \sqrt{2}\kappa_2(A),$$

see Theorem 3.4.3. For general A when we use the standard column pivoting strategy in the QR factorization, $AP = QR$, we saw from (3.4.8) and (3.4.22) that

$$\begin{aligned} \kappa_R(AP) &\le \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}, \\ \kappa'_R(AP) &\le \sqrt{n(n+1)(4^n + 6n - 1)/3}. \end{aligned}$$

As a result of these analyses we see both R and in a certain sense Q can be less sensitive than was thought from previous analyses. The condition numbers depend on any column pivoting of A, and show that the standard pivoting strategy often results in much less sensitive R, and sometimes leads to a much smaller possible change of Q in the range of Q, for a given size of perturbation in A.

By following the approach of Stewart [38, 1973,Th. 3.1], see also [45, 1990,Th. 2.11], it would be straightforward, but detailed and lengthy, to extend our first-order results to provide strict perturbation bounds, as was done in Chapter 2. Our condition numbers and resulting bounds are asymptotically sharp, so there is less need for strict bounds.

In the future we would like to

- Investigate the ratio $\kappa_R(A)/\kappa'_R(A)$.
- Explore the effect of rank-revealing pivoting on κ_R in both theory and computations, and study the optimization problem $\min_P \kappa_R(PAP^T)$.
- Extend our analysis to the case where ΔA has the equivalent componentwise form of backward rounding errors. In fact a new perturbation bound has been given by Chang and Paige [9, 1995]. Also some other results have been obtained by Chang and Paige [11].

Chapter 4

The LU factorization

4.1 Introduction

The LU factorization is a basic and effective tool in numerical linear algebra: given a real $n \times n$ matrix A whose leading principal submatrices are all nonsingular, there exist a unique unit lower triangular matrix L and an upper triangular matrix U such that

$$A = LU.$$

Notice here we require the diagonal elements of L to be 1. L and U are referred to as the LU factors. The LU factorization is a "high-level" algebraic description of the Gaussian elimination. Simple examples shows the standard algorithms for the LU factorization are not numerically stable. In order to repair this shortcoming of the algorithms, partial pivoting or complete pivoting is introduced in the computation. For all of these details, see for example Wilkinson [53, 1965,Chap.4], Higham [30, 1996,Chap.9] and Golub and Van Loan [26, 1996,Chap.3]).

Let ΔA be a sufficiently small $n \times n$ matrix such that the leading principal submatrices of $A+\Delta A$ are still all nonsingular, then $A+\Delta A$ has the unique LU factorization

$$A + \Delta A = (L + \Delta L)(U + \Delta U).$$

The goal of the sensitivity analysis for the LU factorization is to determine a bound on $\|\Delta L\|$ (or $|\Delta L|$) and a bound on $\|\Delta U\|$ (or $|\Delta U|$) in terms of (a bound on) $\|\Delta A\|$ (or $|\Delta A|$).

The perturbation analysis of the LU factorization has been considered by a few authors. Given $\|\Delta A\|$, the first rigorous normwise perturbation bound was presented by Barrland [2, 1991]. Using a different approach, Stewart [41, 1993] gave first-order perturbation bounds, which recently were improved by Stewart [42, 1995]. In [42], L was not assumed to be unit lower triangular, and a parameter p was used to control how much of the perturbation is attached to the diagonals of L and U. Given $|\Delta A|$, the first rigorous componentwise perturbation bounds were given by Sun [48, 1992].

The main purpose of this chapter is to establish new first-order perturbation bounds given a bound on $\|\Delta A\|$, present the condition numbers, give the condition estimators, and shed light on the effect of the partial pivoting and complete pivoting on the sensitivity of the problem.

The rest of this chapter is organized as follows. In Section 4.2 we obtain expressions for $\dot{L}(0)$ and $\dot{U}(0)$ in the LU factorization $A + tG = L(t)U(t)$. These basic sensitivity expressions will be used to obtain our new perturbation bounds in Section 4.3. In Section 4.3 we present perturbation results, first by the so called matrix-vector equation approach, which leads to sharp bounds, then by the so called matrix equation approach, which leads to weaker but practical bounds. We give numerical examples in Section 4.4. Finally we briefly summarize our findings and point out future work in Section 4.5.

4.2 Rate of change of L and U

Here we derive the basic results on how L and U change as A changes, which will be used later.

Theorem 4.2.1 *Let $A \in \mathbf{R}^{n\times n}$ have nonsingular leading $k \times k$ principal submatrices for $k = 1, \ldots, n$ with the LU factorization $A = LU$, let G is a real $n \times n$ matrix, and let $\Delta A = \epsilon G$, for some $\epsilon \geq 0$. If ϵ is sufficiently small such that all leading principal submatrices of $A + tG$ are nonsingular for all $|t| \leq \epsilon$, then $A + \Delta A$ has the LU factorization*

$$A + \Delta A = (L + \Delta L)(U + \Delta U), \tag{4.2.1}$$

with ΔL and ΔU satisfying

$$\Delta L = \epsilon \dot{L}(0) + O(\epsilon^2), \tag{4.2.2}$$

$$\Delta U = \epsilon \dot{U}(0) + O(\epsilon^2), \tag{4.2.3}$$

where $\dot{L}(0)$ and $\dot{U}(0)$ are defined by the unique LU factorization

$$A + tG = L(t)U(t), \qquad |t| \leq \epsilon, \tag{4.2.4}$$

and so satisfy the equations

$$L\dot{U}(0) + \dot{L}(0)U = G, \tag{4.2.5}$$

$$\dot{L}(0) = L\,\mathrm{slt}(L^{-1}GU^{-1}), \tag{4.2.6}$$

$$\dot{U}(0) = \mathrm{ut}(L^{-1}GU^{-1})U. \tag{4.2.7}$$

Proof. Since all leading principal submatrices of $A+tG$ are nonsingular for all $|t| \leq \epsilon$, $A+tG$ has the unique LU factorization (4.2.4). Note that $L(0) = L$, $L(\epsilon) = L + \Delta L$, $U(0) = U$ and $U(\epsilon) = U + \Delta U$. When $t = \epsilon$, (4.2.4) becomes (4.2.1). It is easy to observe that $L(t)$ and $U(t)$ are twice continuously differentiable for $|t| \leq \epsilon$ from a standard algorithm for the LU factorization. If we differentiate (4.2.4) and set $t = 0$ in the result, we obtain (4.2.5), which we will see is a linear equation *uniquely* defining the elements of strictly lower triangular $\dot{L}(0)$ and and upper triangular $\dot{U}(0)$ in terms of the elements of G. From (4.2.5) we have

$$L^{-1}\dot{L}(0) + \dot{U}(0)U^{-1} = L^{-1}GU^{-1}.$$

Note that $L^{-1}\dot{L}(0)$ is strictly lower triangular and $\dot{U}(0)U^{-1}$ is upper triangular, thus we have

$$L^{-1}\dot{L}(0) = \mathrm{slt}(L^{-1}GU^{-1}), \quad \dot{U}(0)U^{-1} = \mathrm{ut}(L^{-1}GU^{-1}),$$

which give (4.2.6) and (4.2.7). Finally the Taylor expansions for $L(t)$ and $U(t)$ about $t = 0$ give (4.2.2) and (4.2.3). □

4.3 New perturbation results

The basis for deriving first-order perturbation bounds is the equation (4.2.5) (or the expressions (4.2.6) and (4.2.7) of its solutions). As in the proceeding chapters, we will now analyze the equation in two ways. The first, the matrix-vector equation approach, provides sharp bounds, resulting in the condition numbers of the problem; while the second, the matrix equation approach, gives practical bounds, resulting in condition estimators. Throughout this section we suppose all assumptions in Theorem 4.2.1 hold, so we can use its conclusions. Also we assume $\|\Delta A\|_F \leq \epsilon \|A\|_F$, hence $\|G\|_F \equiv \|\Delta A\|_F/\epsilon \leq \|A\|_F$ (if $\epsilon = 0$, all results we will present are obviously true). One exception is in Theorem 4.3.3 we assume $\|\Delta A\|_{1,\infty} \leq \epsilon \|A\|_{1,\infty}$.

4.3.1 Matrix-vector equation analysis

It is not difficult to show that with the 'uvec' notation and 'slvec' notation in (1.2.4) the matrix equation (4.2.5) can be rewritten in the following form:

$$W \begin{bmatrix} \mathrm{uvec}(\dot{U}(0)) \\ \mathrm{slvec}(\dot{L}(0)) \end{bmatrix} = \mathrm{vec}(G), \tag{4.3.1}$$

where $W \equiv [W_L, W_U]$ with $W_L \in \mathcal{R}^{n^2 \times \frac{n(n+1)}{2}}$ being

$$\begin{bmatrix}
1 & & & & \\
l_{21} & & & & \\
\vdots & & & & \\
l_{n1} & & & & \\
 & 1 & & & \\
 & l_{21} & 1 & & \\
 & \vdots & \vdots & & \\
 & l_{n1} & l_{n2} & & \\
 & & & \ddots & \\
 & & & 1 & \\
 & & & l_{n-1,1} & l_{n-1,2} \quad \cdots \quad 1 \\
 & & & l_{n1} & l_{n2} \quad \cdots \quad l_{n,n-1} \quad 1
\end{bmatrix}$$

and $W_U \in \mathcal{R}^{n^2 \times \frac{n(n-1)}{2}}$ being

$$\left[\begin{array}{cccc|cccc|c|c}
u_{11} & & & & & & & & & \\
 & u_{11} & & & & & & & & \\
 & & \ddots & & & & & & & \\
 & & & u_{11} & & & & & & \\
\hline
u_{12} & & & & & & & & & \\
 & u_{12} & & & u_{22} & & & & & \\
 & & \ddots & & & \ddots & & & & \\
 & & & u_{12} & & & u_{22} & & & \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\
\hline
u_{1n} & & & & & & & & & \\
 & u_{1n} & & & u_{2n} & & & & & \\
 & & \ddots & & & \ddots & & & \cdots & \\
 & & & u_{1n} & & & u_{2n} & & & u_{n-1,n}
\end{array}\right]$$

It is easy to observe that after appropriate column permutations, $[W_L, W_U]$ will become lower triangular with diagonal elements

$$\underbrace{1, u_{11}, u_{11}, \cdots, u_{11}}_{n}, \underbrace{1, 1, u_{22}, \cdots, u_{22}}_{n}, \cdots\cdots, \underbrace{1, 1, \cdots, 1, 1}_{n}.$$

Since U is nonsingular, W is also, and from (4.3.1)

$$\begin{bmatrix} \text{uvec}(\dot{U}(0)) \\ \text{slvec}(\dot{L}(0)) \end{bmatrix} = W^{-1} \text{vec}(G). \tag{4.3.2}$$

Partitioning W^{-1} into two blocks, $W^{-1} \equiv \begin{bmatrix} Y_U \\ Y_L \end{bmatrix}$, we have

$$\text{slvec}(\dot{L}(0)) = Y_L \text{vec}(G), \qquad \text{uvec}(\dot{U}(0)) = Y_U \text{vec}(G). \tag{4.3.3}$$

so taking the 2-norm and using $\|G\|_F \le \|A\|_F$ gives

$$\|\dot{L}(0)\|_F \le \|Y_L\|_2 \|G\|_F \le \|Y_L\|_2 \|A\|_F, \tag{4.3.4}$$

$$\|\dot{U}(0)\|_F \le \|Y_U\|_2 \|G\|_F \le \|Y_U\|_2 \|A\|_F, \tag{4.3.5}$$

where equalities can be obtained by choosing G such that $\mathrm{vec}(G)$ lies in the space spanned by the right singular vectors corresponding to the largest singular value of Y_L and Y_U, respectively, and $\|G\|_F = \|A\|_F$. Therefore we see from the Taylor expansions (4.2.2) and (4.2.3) that

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \frac{\|Y_L\|_2 \|A\|_F}{\|L\|_F}\epsilon + O(\epsilon^2), \tag{4.3.6}$$

$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \frac{\|Y_U\|_2 \|A\|_F}{\|U\|_F}\epsilon + O(\epsilon^2), \tag{4.3.7}$$

and for the L factor and the U factor the condition numbers (with respect to the F-norm)

$$\kappa_L(A) \equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta L\|_F}{\epsilon\|L\|_F} : (A+\Delta A) = (L+\Delta L)(U+\Delta U),\ \|\Delta A\|_F \le \epsilon\|A\|_F\Big\},$$

$$\kappa_U(A) \equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta U\|_F}{\epsilon\|U\|_F} : (A+\Delta A) = (L+\Delta L)(U+\Delta U),\ \|\Delta A\|_F \le \epsilon\|A\|_F\Big\}$$

are respectively given by

$$\kappa_L(A) = \frac{\|Y_L\|_2 \|A\|_F}{\|L\|_F}, \qquad \kappa_U(A) = \frac{\|Y_U\|_2 \|A\|_F}{\|U\|_F}.$$

We summarize these results as the following theorem.

Theorem 4.3.1 *Suppose all the assumptions of Theorem* 4.2.1 *hold, and let* $\|\Delta A\|_F \le \epsilon\|A\|_F$, *then* $A+\Delta A$ *has the unique LU factorization*

$$A + \Delta A = (L+\Delta L)(U+\Delta U),$$

where with $\kappa_L(A) = \frac{\|Y_L\|_2\|A\|_F}{\|L\|_F}$ *and* $\kappa_U(A) = \frac{\|Y_U\|_2\|A\|_F}{\|U\|_F}$,

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \kappa_L(A)\epsilon + O(\epsilon^2), \tag{4.3.8}$$

$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \kappa_U(A)\epsilon + O(\epsilon^2), \tag{4.3.9}$$

and these bounds are attainable to first-order in ϵ. □

4.3.2 Matrix equation analysis

In Section 4.3.1 we derived sharp perturbation bounds for the L factor and U factor, and presented the corresponding condition numbers. But it is difficult to estimate the condition numbers by using the usual approach. Now we use the matrix equation approach to derive practical perturbation bounds, leading to the condition estimators.

First we derive a perturbation bound for the L factor. Let U_{n-1} denote the leading $(n-1)\times(n-1)$ block of U. If we write $U = \begin{bmatrix} U_{n-1} & u \\ 0 & u_{nn} \end{bmatrix}$, then from (4.2.6)

$$\dot{L}(0) = L\,\mathrm{slt}(L^{-1}G\begin{bmatrix} U_{n-1}^{-1} & -U_{n-1}^{-1}u/u_{nn} \\ 0 & 1/u_{nn} \end{bmatrix}) = L\,\mathrm{slt}(L^{-1}G\begin{bmatrix} U_{n-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}). \quad (4.3.10)$$

Denote by $\mathbf{D}_n$ the set of all $n\times n$ real positive definite diagonal matrices. Let $D = \mathrm{diag}(\delta_1,\ldots,\delta_n) \in \mathbf{D}_n$. Note that for any $n\times n$ matrix B we have $D\,\mathrm{slt}(B) = \mathrm{slt}(DB)$, then from (4.3.10) we obtain

$$\dot{L}(0) = LD^{-1}\mathrm{slt}(DL^{-1}G\begin{bmatrix} U_{n-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}).$$

Note $\|\mathrm{slt}(B)\|_F \le \|B\|_F$ for any $B \in \mathbf{R}^{n\times n}$, so we have

$$\|\dot{L}(0)\|_F \le \|LD^{-1}\|_2\,\|DL^{-1}\|_2\,\|U_{n-1}^{-1}\|_2\,\|G\|_F, \quad (4.3.11)$$

which with $\|G\|_F \le \|A\|_F$ gives

$$\frac{\|\dot{L}(0)\|_F}{\|L\|_F} \le \kappa_2(LD^{-1})\frac{\|U_{n-1}^{-1}\|_2\,\|A\|_F}{\|L\|_F}.$$

Since this is true for all $D \in \mathbf{D}_n$, by the Taylor expansion (4.2.2) we have

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \kappa_L'(A)\epsilon + O(\epsilon^2), \quad (4.3.12)$$

where

$$\kappa_L'(A) = \inf_{D\in\mathbf{D}_n} \kappa_L'(A,D), \quad (4.3.13)$$

$$\kappa_L'(A, D) \equiv \kappa_2(LD^{-1}) \frac{\|U_{n-1}^{-1}\|_2 \|A\|_F}{\|L\|_F}. \tag{4.3.14}$$

From the definition of $\kappa_L(A)$ and the perturbation bound (4.3.12) we easily see

$$\kappa_L(A) \leq \kappa_L'(A). \tag{4.3.15}$$

Also we can obtain a lower bound on $\kappa_L(A)$. Let $v \in \mathbf{R}^{n-1}$ be such that $\|U_{n-1}^{-T} v\|_2 = \|U_{n-1}^{-1}\|_2 \|v\|_2$. In (4.3.10), take $G = [e_n v^T, 0]$, where $e_n = (0, \ldots, 0, 1)^T \in \mathbf{R}^n$. Then it is easy to verify that

$$\dot{L}(0) = e_n [v^T U_{n-1}^{-1}, 0],$$

so

$$\|\dot{L}(0)\|_F = \|v^T U_{n-1}^{-1}\|_2 = \|U_{n-1}^{-1}\|_2 \|v\|_2 = \|U_{n-1}^{-1}\|_2 \|G\|_F.$$

Combining this with the first equality of (4.3.3), we have for this special G that

$$\|Y_L \mathrm{vec}(G)\|_2 = \|U_{n-1}^{-1}\|_2 \|G\|_F,$$

which gives

$$\|Y_L\|_2 \geq \|U_{n-1}^{-1}\|_2, \tag{4.3.16}$$

or

$$\kappa_L(A) \equiv \frac{\|Y_L\|_2 \|A\|_F}{\|L\|_F} \geq \frac{\|U_{n-1}^{-1}\|_2 \|A\|_F}{\|L\|_F}. \tag{4.3.17}$$

We would like to point out that (4.3.16) can also be derived directly from the structure of W in (4.3.1).

Now we derive a practical perturbation bound for the U factor. Notice for any $B \in \mathbf{R}^{n \times n}$ and $D \in \mathbf{D}_n$ we have $\mathrm{ut}(BD) = \mathrm{ut}(B)D$, then from (4.2.7) we obtain

$$\dot{U}(0) = \mathrm{ut}(L^{-1} G U^{-1} D) D^{-1} U.$$

Thus

$$\|\dot{U}(0)\|_F \leq \|L^{-1}\|_2 \|U^{-1} D\|_2 \|D^{-1} U\|_2 \|G\|_F, \tag{4.3.18}$$

which with $\|G\|_F \leq \|A\|_F$ gives

$$\frac{\|\dot{U}(0)\|_F}{\|U\|_F} \leq \kappa_2(D^{-1}U)\frac{\|L^{-1}\|_2\,\|A\|_F}{\|U\|_F}.$$

Since this is true for all $D \in \mathbf{D}_n$, by the Taylor expansion (4.2.2) we have

$$\frac{\|\Delta U\|_F}{\|U\|_F} \leq \kappa'_U(A)\epsilon + O(\epsilon^2), \tag{4.3.19}$$

where

$$\kappa'_U(A) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_U(A,D), \tag{4.3.20}$$

$$\kappa'_U(A,D) \equiv \kappa_2(D^{-1}U)\frac{\|L^{-1}\|_2\,\|A\|_F}{\|U\|_F}. \tag{4.3.21}$$

From the definition of $\kappa_U(A)$ and the perturbation bound (4.3.19) we see

$$\kappa_U(A) \leq \kappa'_U(A). \tag{4.3.22}$$

Also we can get a lower bound on $\kappa_U(A)$. Let $v \in \mathcal{R}^n$ be such that $\|L^{-1}v\|_2 = \|L^{-1}\|_2\,\|v\|_2$, and take $G = ve_n^T$ in (4.2.7), then combining (4.2.7) and the second equality of (4.3.3) we can easily show

$$\|Y_U\|_2 \geq \|L^{-1}\|_2, \tag{4.3.23}$$

or

$$\kappa_U(A) \geq \frac{\|L^{-1}\|_2\,\|A\|_F}{\|U\|_F}. \tag{4.3.24}$$

Like (4.3.16), (4.3.23) can also be showed directly from the structure of W in (4.3.1).

These results lead to the following theorem.

Theorem 4.3.2 *With the same assumptions as in Theorem* 4.3.1, $A + \Delta A$ *has the unique LU factorization*

$$A + \Delta A = (L + \Delta L)(U + \Delta U),$$

where for the L factor,

$$\frac{\|\Delta L\|_F}{\|U\|_F} \leq \kappa'_L(A)\epsilon + O(\epsilon^2), \tag{4.3.25}$$

$$\frac{\|U_{n-1}^{-1}\|_2\|A\|_F}{\|L\|_F} \le \kappa_L(A) \le \kappa_L'(A) \equiv \inf_{D\in\mathbf{D}_n} \kappa_L'(A,D), \tag{4.3.26}$$

with $\kappa_L'(A,D) \equiv \kappa_2(LD^{-1})\|U_{n-1}^{-1}\|_2\,\|A\|_F/\|L\|_F$, *and for the* U *factor,*

$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \kappa_U'(A)\epsilon + O(\epsilon^2), \tag{4.3.27}$$

$$\frac{\|L^{-1}\|_2\|A\|_F}{\|U\|_F} \le \kappa_U(A) \le \kappa_U'(A) \equiv \inf_{D\in\mathbf{D}_n} \kappa_U'(A,D), \tag{4.3.28}$$

with $\kappa_U'(A,D) \equiv \kappa_2(D^{-1}U)\|L^{-1}\|_2\,\|A\|_F/\|U\|_F$. □

We might want to simplify $\kappa_L'(A,D)$ and $\kappa_U'(A,D)$. If we use

$$\|A\|_F \le \|L\|_F\,\|U\|_2,\ \|L\|_2\,\|U\|_F, \tag{4.3.29}$$

then we have

$$\kappa_L'(A,D) \le \kappa_2(LD^{-1})\|U\|_2\,\|U_{n-1}^{-1}\|_2, \tag{4.3.30}$$

$$\kappa_U'(A,D) \le \kappa_2(L)\kappa_2(D^{-1}U). \tag{4.3.31}$$

As we know in practice $\kappa_2(L)$ is usually smaller or much smaller than $\kappa_2(U)$. So these bounds suggest the L factor may be more sensitive than the U factor in practice. However both of the right hand sides of (4.3.30) and (4.3.31) can be arbitrarily larger than corresponding left hand sides due to the inequality (4.3.29).

If we take $D = I$ in both $\kappa_L'(A,D)$ and $\kappa_U'(A,D)$ and use $\|L\|_2 \le \|L\|_F$, $\|U\|_2 \le \|U\|_F$ and $\|U_{n-1}^{-1}\|_2 \le \|U^{-1}\|_2$, we have

$$\kappa_L'(A) \le \kappa_L'(A,I) \le \kappa_2(L)\|U_{n-1}^{-1}\|_2\,\|A\|_F/\|L\|_F \le \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F, \tag{4.3.32}$$

$$\kappa_U'(A) \le \kappa_U'(A,I) \le \kappa_2(U)\|L^{-1}\|_2\,\|A\|_F/\|U\|_F \le \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F. \tag{4.3.33}$$

Thus from (4.3.25) and (4.3.27) we have

$$\frac{\|\Delta L\|_F}{\|L\|_F} \lesssim \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F\,\epsilon, \tag{4.3.34}$$

$$\frac{\|\Delta U\|_F}{\|U\|_F} \lesssim \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F\,\epsilon, \tag{4.3.35}$$

which are due to Stewart [41, 1993]. These perturbation bounds are simple, but can overestimate the true sensitivity of the problem. By using the scaling technology, Stewart [42, 1995] obtained significant improvements on the above results. In [42], the diagonal elements of L were not assumed to be 1's, and the diagonal elements of ΔL may not be 0's, and a parameter p was used to control how much of the perturbation is attached to the diagonals of L and U. The perturbation bounds given in [42] are equivalent to

$$\frac{\|\Delta L\|_F}{\|L\|_F} \lesssim \kappa_2(LD^{-1})\kappa_2(U), \tag{4.3.36}$$

$$\frac{\|\Delta U\|_F}{\|U\|_F} \lesssim \kappa_2(L)\kappa_2(D^{-1}U). \tag{4.3.37}$$

These bounds were derived by using the inequality (4.3.29), so they are unnecessarily weak as we pointed out in the preceding comment. (4.3.30) suggests that under the usual assumption that the diagonal elements of L are always 1's, a better bound than (4.3.36) could be obtained.

As we know it is expensive to estimate $\kappa_L(A)$ and $\kappa_U(A)$ directly by the usual approach. Fortunately we now have other methods to do this. By van der Sluis's Theorem 1.2.1, $\kappa_2(LD^{-1})$ will be nearly minimum when each column of LD^{-1} has unit 2-norm, so in practice we choose $D = D_L = \text{diag}(\|L(:,j)\|_2)$, then use a standard condition estimator and a norm estimator to estimate $\kappa'_L(A, D_L)$, which costs $O(n^2)$. Similarly, $\kappa_2(D^{-1}U)$ will be nearly minimum when each row of $D^{-1}U$ has unit 2-norm, then we choose $D = D_U = \text{diag}(\|U(i,:)\|_2)$, and use a standard condition estimator and a norm estimator to estimate $\kappa'_U(A, D_U)$, which costs $O(n^2)$. Numerical experiments showed $\kappa'_L(A, D_L)$ and $\kappa'_U(A, D_U)$ are good approximations of $\kappa_L(A)$ and $\kappa_U(A)$.

When we use the 1- and ∞-norms, we can get perturbation bounds without involving the scaling matrix D.

Theorem 4.3.3 *Suppose all the assumptions of Theorem* 4.2.1 *hold and let* $\|\Delta A\|_p \leq$

$\epsilon\|A\|_p$, $p = 1, \infty$, *then* $A + \Delta A$ *has the unique LU factorization*

$$A + \Delta A = (L + \Delta L)(U + \Delta U),$$

where

$$\frac{\|\Delta L\|_p}{\|L\|_p} \leq \text{cond}_p(L^{-1}) \frac{\|U_{n-1}^{-1}\|_p \|A\|_p}{\|L\|_p} \epsilon + O(\epsilon^2) \tag{4.3.38}$$

$$\leq \text{cond}_p(L^{-1}) \|U\|_p \|U_{n-1}^{-1}\|_p \epsilon + O(\epsilon^2), \tag{4.3.39}$$

$$\frac{\|\Delta U\|_p}{\|U\|_p} \leq \text{cond}_p(U) \frac{\|L^{-1}\|_p \|A\|_p}{\|U\|_p} \epsilon + O(\epsilon^2), \tag{4.3.40}$$

$$\leq \kappa_p(L) \text{cond}_p(U) \epsilon + O(\epsilon^2), \tag{4.3.41}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$, the theorem is trivial), then $\|G\|_p \leq \|A\|_p$, $p = 1, \infty$. From (4.3.10) we have

$$|\dot{L}(0)| \leq |L||L^{-1}||G| \begin{bmatrix} |U_{n-1}^{-1}| & 0 \\ 0 & 0 \end{bmatrix},$$

so taking the p-norm ($p = 1, \infty$) gives

$$\|\dot{L}(0)\|_p \leq \text{cond}_p(L^{-1}) \|U_{n-1}\|_p \|G\|_p \leq \text{cond}_p(L^{-1}) \|U_{n-1}\|_p \|A\|_p.$$

Then (4.3.38) follows immediately from the Taylor expansion (4.2.2). By using $\|A\|_p \leq \|L\|_p \|U\|_p$, (4.3.39) follows.

The results (4.3.40) and (4.3.41) can similarly be proved. □

Note $\text{cond}_p(L^{-1})$ is invariant under the column scaling of L and $\text{cond}_p(U)$ is invariant under the row scaling of U. These make (4.3.38) and (4.3.40) look simpler than (4.3.8) and $(4.3.9)_t$ respectively, where the Frobenius norm is used.

As we know the standard algorithms for LU factorization with no pivoting are not numerically stable. In order to repair this shortcoming of the algorithms, partial or complete pivoting should be incorporated in the computation. Do these two pivoting strategies have effects on the sensitivity of the factorization? Let us see the following theorem.

Theorem 4.3.4 *If partial pivoting is used in the LU factorization: $PA = LU$, where P is a permutation matrix. Then*

$$\sqrt{\frac{2}{n(n+1)}}\,\|U_{n-1}^{-1}\|_2\,\|A\|_F \le \kappa_L(PA) \le \kappa_L'(PA) \le \frac{1}{3}\sqrt{4^n+6n-1}\,\|U_{n-1}^{-1}\|_2\,\|A\|_F, \tag{4.3.42}$$

$$1 \le \kappa_U(PA) \le \kappa_U'(PA) \le \frac{1}{6}\sqrt{2n(n+1)(4^n+6n-1)}\,\inf_{D\in\mathbf{D}_n}\kappa_2(D^{-1}U). \tag{4.3.43}$$

If complete pivoting is used in the LU factorization: $PAQ = LU$, where P and Q are permutation matrices. Then

$$\sqrt{\frac{2}{n(n+1)}}\,\|U_{n-1}^{-1}\|_2\,\|A\|_F \le \kappa_L(PAQ) \le \kappa_L'(PAQ) \le \frac{1}{3}\sqrt{4^n+6n-1}\,\|U_{n-1}^{-1}\|_2\,\|A\|_F, \tag{4.3.44}$$

$$1 \le \kappa_U(PAQ) \le \kappa_U'(PAQ) \le n(n+1)(4^n+6n-1)/18. \tag{4.3.45}$$

Proof. If partial pivoting or complete pivoting is used in computing the LU factorization, then $l_{ij} \le 1$ for $i > j$. Since $l_{ii} = 1$, $|(L^T)_{ii}| \ge |(L^T)_{ij}|$ for all $j > i$. By the proof of Theorem 1.2.2 we see $\|L^{-1}\|_2 \le \sqrt{4^n+6n-1}/3$. Also note $\|L\|_F \le \sqrt{n(n+1)/2}$. Then (4.3.42) and (4.3.44) follow from (4.3.26) by taking $D = I$.

Note $\kappa_U'(A,D) \le \kappa_2(L)\kappa_2(D^{-1}U)$ (see (4.3.31)) and $\|U\|_F = \|L^{-1}A\|_F \le \|L^{-1}\|_2\,\|A\|_F$, then (4.3.43) follows from (4.3.28). If complete pivoting is used, then $|u_{ii}| \ge |u_{ij}|$ for all $j > i$. Thus by Theorem 1.2.2 we have $\kappa_2(D^{-1}U) \le \sqrt{2n(n+1)(4^n+6n-1)}/6$ with $D = \mathrm{diag}(U)$. Then from (4.3.43) we obtain the much better result (4.3.45). □

When partial pivoting or complete pivoting is used, $\kappa_2(L)$ is bounded by a function of n, but possibly $\|U_{n-1}^{-1}\|$ may become larger, thus from (4.3.42) and (4.3.44) we cannot see that $\kappa_L(PA)$ and $\kappa_L(PAQ)$ are larger or smaller than $\kappa_L(A)$. (4.3.42) and (4.3.44) also suggest there is no big difference between the effects of partial pivoting and complete pivoting on the sensitivity of the L factor. Similarly from (4.3.43) we are not quite sure if $\kappa_U(A)$ will become larger or smaller when partial pivoting is used. But note there is an essential difference between the upper bound on $\kappa_U(PA)$ and

that on $\kappa_L(PA)$ — the former has a choice D, which may make $\inf_{D\in\mathbf{D}_n}\kappa_2(D^{-1}U)$ not increase much, so the possibility that $\kappa_U(A)$ will become smaller seems high. From (4.3.45) we see complete pivoting can give a significant improvement on $\kappa_U(A)$.

4.4 Numerical experiments

In Section 4.3 we presented first-order sharp perturbation bounds for the LU factors, obtained the corresponding condition numbers $\kappa_L(A)$ and $\kappa_U(A)$, and suggested $\kappa_L(A)$ and $\kappa_U(A)$ could be respectively estimated in practice by $\kappa'_L(A, D_L)$ and $\kappa'_U(A, D_U)$ with $D_L = \mathrm{diag}(\|L(:,j)\|_2)$ and $D_U = \mathrm{diag}(\|U(i,:)\|_2)$. The condition numbers and condition estimators satisfy the following inequalities (see (1.2.14), (1.2.15), (4.3.26), (4.3.28), (4.3.32), and (4.3.33)):

$$\|U_{n-1}^{-1}\|_2\,\|A\|_F/\|L\|_F \le \kappa_L(A) \le \kappa'_L(A) \le \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F,$$
$$\kappa'_L(A) \le \kappa'_L(A, D_L) \le \sqrt{n}\,\kappa'_L(A),$$
$$\|L^{-1}\|_2\,\|A\|_F/\|U\|_F \le \kappa_U(A) \le \kappa'_U(A) \le \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F,$$
$$\kappa'_U(A) \le \kappa'_U(A, D_U) \le \sqrt{n}\,\kappa'_U(A).$$

Also we discussed the effects of partial pivoting and complete pivoting on those condition numbers.

Now we give some numerical experiments to illustrate our theoretical analyses.

The matrices have the form $A = D_1 B D_2$, where $D_1 = \mathrm{diag}(1, d_1, \ldots, d_1^{n-1})$, $D_2 = \mathrm{diag}(1, d_2, \ldots, d_2^{n-1})$ and B is an $n \times n$ random matrix (produced by the MATLAB function `randn`). The results for $n = 10$, $d_1, d_2 \in \{0.2, 1, 2\}$ and the same matrix B are shown in Table 4.4.1 without pivoting, in Table 4.4.2 with partial pivoting, and in Table 4.4.3 with complete pivoting, where $\beta_L \equiv \|U_{n-1}^{-1}\|_2\,\|A\|_F/\|L\|_F$, $\beta_U \equiv \|L^{-1}\|_2\,\|A\|_F/\|U\|_F$ and $\beta \equiv \|L^{-1}\|_2\,\|U^{-1}\|_2\,\|A\|_F$.

We give some comments on the results.

Table 4.4.1: Results without pivoting, $A = LU$

d_1	d_2	β_L	$\kappa_L(A)$	$\kappa'_L(A, D_L)$	β_U	$\kappa_U(A)$	$\kappa'_U(A, D_U)$	β
0.2	0.2	2.4e+09	2.4e+09	1.7e+10	3.8e+00	8.7e+00	1.8e+01	1.5e+12
0.2	1	1.4e+05	1.7e+05	9.7e+05	2.8e+00	2.0e+02	7.3e+02	1.7e+07
0.2	2	5.3e+06	7.2e+06	3.8e+07	1.2e+00	2.7e+04	1.2e+05	7.9e+08
1	0.2	2.8e+03	3.4e+04	4.2e+05	4.9e+01	1.1e+02	2.3e+02	1.1e+07
1	1	4.5e+00	2.2e+02	6.7e+02	2.8e+00	1.9e+02	7.3e+02	4.1e+03
1	2	7.9e+02	3.7e+04	1.2e+05	1.5e+00	3.2e+04	1.4e+05	7.2e+05
2	0.2	2.6e+01	3.4e+04	1.0e+06	3.1e+05	3.2e+05	1.4e+06	2.6e+08
2	1	3.2e+00	3.2e+04	1.3e+05	3.2e+02	3.6e+03	8.2e+04	3.2e+07
2	2	2.7e+02	2.7e+06	1.0e+07	6.7e+01	1.6e+05	6.4e+06	2.7e+09

Table 4.4.2: Results with partial pivoting, $\tilde{A} \equiv PA = \tilde{L}\tilde{U}$

d_1	d_2	β_L	$\kappa_L(\tilde{A})$	$\kappa'_L(\tilde{A}, D_L)$	β_U	$\kappa_U(\tilde{A})$	$\kappa'_U(\tilde{A}, D_U)$	β
0.2	0.2	3.5e+09	3.5e+09	8.5e+09	1.6e+00	1.7e+00	3.1e+00	6.6e+11
0.2	1	2.0e+05	2.4e+05	4.8e+05	1.6e+00	2.2e+01	8.8e+01	7.5e+06
0.2	2	7.7e+06	1.0e+07	1.9e+07	1.6e+00	3.5e+03	2.1e+04	3.4e+08
1	0.2	1.1e+05	2.1e+05	6.2e+05	4.7e+00	4.7e+00	6.7e+00	3.9e+06
1	1	7.1e+00	1.3e+01	4.0e+01	2.5e+00	1.2e+01	4.3e+01	8.5e+01
1	2	8.0e+02	1.4e+03	4.5e+03	1.6e+00	1.5e+03	6.0e+03	9.8e+03
2	0.2	8.2e+06	1.2e+07	2.7e+07	2.1e+00	2.1e+00	3.2e+00	2.6e+08
2	1	4.9e+02	6.3e+02	1.6e+03	1.7e+00	1.8e+01	7.8e+01	5.0e+03
2	2	2.2e+04	2.7e+04	7.4e+04	1.7e+00	2.7e+03	1.4e+04	1.7e+05

Table 4.4.3: Results with complete pivoting, $\hat{A} \equiv PAQ = \hat{L}\hat{U}$

d_1	d_2	β_L	$\kappa_L(\hat{A})$	$\kappa'_L(\hat{A}, D_L)$	β_U	$\kappa_U(\hat{A})$	$\kappa'_U(\hat{A}, D_U)$	β
0.2	0.2	3.5e+09	3.5e+09	8.5e+09	1.6e+00	1.7e+00	3.1e+00	6.6e+11
0.2	1	1.4e+05	1.4e+05	2.0e+05	1.2e+00	2.5e+00	6.6e+00	5.6e+06
0.2	2	2.6e+06	2.6e+06	5.8e+06	1.4e+00	1.5e+00	4.4e+00	2.9e+08
1	0.2	1.1e+05	2.1e+05	6.2e+05	4.7e+00	4.7e+00	6.7e+00	3.9e+06
1	1	2.8e+00	5.0e+00	1.9e+01	3.4e+00	4.9e+00	1.4e+01	6.7e+01
1	2	1.4e+02	3.3e+02	1.2e+03	5.0e+00	7.0e+00	1.6e+01	5.8e+03
2	0.2	8.2e+06	1.2e+07	2.7e+07	2.1e+00	2.1e+00	3.2e+00	2.6e+08
2	1	3.1e+02	3.4e+02	9.6e+02	1.8e+00	3.0e+00	1.3e+01	3.9e+03
2	2	1.1e+04	1.2e+04	4.4e+04	2.2e+00	3.0e+00	7.5e+00	1.3e+05

- The results confirm that $\beta = \|L^{-1}\|_2 \|U^{-1}\|_2 \|A\|_F$ can be much larger than $\kappa_L(A)$ and $\kappa_U(A)$, especially for the latter, so the first-order bounds (4.3.34) and (4.3.35) can significantly overestimate the true sensitivity of the LU factorization.

- $\kappa'_L(A, D_L)$ and $\kappa'_U(A, D_U)$ are good approximations of $\kappa_L(A)$ and $\kappa_U(A)$, respectively, no matter whether pivoting is used or not. This is also confirmed by our other numerical experiments.

- Both $\kappa_L(PA)$ and $\kappa_L(PAQ)$ can be much larger or smaller than $\kappa_L(A)$. So partial pivoting and complete pivoting can make the L factor more sensitive or less sensitive. But from Tables 4.4.1–4.4.2 we see partial pivoting can give a significant improvement on the condition of the U factor. In fact here $\kappa_U(PA) \leq \kappa_U(A)$ for all cases. From Table 4.4.3 we see that complete pivoting can give a more significant improvement.

- It can be seen for most cases the L factor is more sensitive than the U factor no matter whether pivoting is used or not.

- When partial pivoting or complete pivoting is used, we see both κ_L and κ_U are

close to their lower bounds β_L and β_U, respectively.

4.5 Summary and future work

The first-order perturbation analyses presented here show what the sensitivity of each of L and U is in the LU factorization of A, and in so doing provide their condition numbers $\kappa_L(A)$ and $\kappa_U(A)$ (with respect to the measures used, and for sufficiently small ΔA), as well as efficient ways of approximating these.

As we know $\kappa_2(L)$ is usually (much) smaller than $\kappa_2(U)$, especially in practice when we use partial pivoting in computing the LU factorization. So we can expect that the computed solution of the linear system $Lx = b$ will usually be more accurate than that of the linear system $Uy = b$. However our analysis and numerical experiments suggest that usually the L factor is more sensitive than the U factor in the LU factorization, so we expect U is more accurate than L. This is an interesting phenomenon. Also we see the effect of partial pivoting and complete pivoting on the sensitivity of L is uncertain — both $\kappa_L(PA)$ and $\kappa_U(PAQ)$ can be much larger or smaller than $\kappa_L(A)$. But partial pivoting can usually improve the condition of U, and complete pivoting can give significant improvement.

In the future we would like to

- Investigate the ratios $\kappa_L(A)/\kappa'_L(A)$ and $\kappa_L(A)/\kappa'_L(A)$.
- Extend our analysis to the case where $|\Delta A| \leq \epsilon|A|$. In fact some results have been obtained by Chang and Paige [12].

Chapter 5

The Cholesky downdating problem

5.1 Introduction

In this chapter we give perturbation analyses of the following problem: given an upper triangular matrix $R \in \mathbf{R}^{n\times n}$ and a matrix $X \in \mathbf{R}^{k\times n}$ such that $R^TR - X^TX$ is positive definite, find an upper triangular matrix $U \in \mathbf{R}^{n\times n}$ with positive diagonal elements such that

$$U^TU = R^TR - X^TX. \tag{5.1.1}$$

This problem is called the *block Cholesky downdating problem*, and the matrix U is referred to as the downdated Cholesky factor. The block Cholesky downdating problem has many important applications, and the case for k=1 has been extensively studied in the literature (see [1, 5, 6, 19, 25, 24, 36, 37, 40]).

Let ΔR and ΔX be real $n \times n$ and $k \times n$ matrices, respectively, such that $(R + \Delta R)^T(R + \Delta R) - (X + \Delta X)^T(X + \Delta X)$ is still positive definite, then this has the unique Cholesky factorization

$$(U + \Delta U)^T(U + \Delta U) = (R + \Delta R)^T(R + \Delta R) - (X + \Delta X)^T(X + \Delta X).$$

The goal of the perturbation analysis for the Cholesky downdating problem is to determine a bound on $\|\Delta U\|$ (or $|\Delta U|$) in terms of $\|\Delta R\|$ (or $|\Delta R|$) and $\|\Delta X\|$ (or $|\Delta X|$).

The perturbation analysis for the Cholesky downdating problem with norm-bounded changes in R and X has been considered by several authors. Stewart [40, 1979] presented perturbation results for single downdating (i.e. $k = 1$). Eldén and Park [22, 1994] made an analysis for block downdating. But these two papers just considered the case that only R or X is perturbed. More complete analyses, with both R and X being perturbed, were given by Pan [35, 1993] and Sun [50, 1995]. Pan [35, 1993] gave first-order perturbation bounds for single downdating. Sun [50, 1995] gave rigorous, also first-order perturbation bounds for single downdating and first-order perturbation bounds for block downdating. Recently Eldén and Park [23, 1996] gave new first-order perturbation bounds for block downdating. Unfortunately there was an error in their paper when the result of Sun [46, 1991] was applied in deriving the perturbation bound. Because of this, the results presented in [23] will not be cited in this chapter.

The main purpose of this chapter is to establish new first-order perturbation results and present new condition numbers which more closely reflect the true sensitivity of the problem. In Section 5.2 we will give the key result of Sun [50, 1995], and a new result using the approach of these earlier papers. In Section 5.3 we present new perturbation results, first by the matrix-vector equation approach, then by the matrix equation approach. We give numerical results and suggest practical condition estimators in Section 5.4. Finally we briefly summarize our findings and point out future work in Section 5.5. Most of the results have been presented in Chang and Paige [10, 1996].

Previous work by others implied the change ΔR in R was upper triangular, and Sun [50, 1995] said this, but neither he nor the others made use of this fact. In fact a backward stable algorithm for computing U given R and X would produce the exact

result $U_c = U + \Delta U$ for nearby data $R + \Delta R$ and $X + \Delta X$, where it is not clear that ΔR would be upper triangular — the form of the equivalent backward rounding error ΔR would depend on the algorithm, and if it were upper triangular, it would require a rounding error analysis to show this. Thus for completeness it seems necessary to consider two separate cases — general ΔR and upper triangular ΔR. We do this throughout Sections 5.3–5.4, and get stronger results for upper triangular ΔR than in the general case.

In any perturbation analysis it is important to examine how good the results are. In Section 5.3.1 we produce provably tight bounds, leading to the true condition numbers (for the norms chosen). The numerical example in Section 5.4 indicates how much better the results of this new analysis can be compared with some earlier ones, but a theoretical understanding is also desirable. By considering the asymptotic case as $X \to 0$, the results simplify, and are easily understandable. We show the new results have the correct properties as $X \to 0$, in contrast to earlier results.

5.2 Basics, previous results, and an improvement

Let Γ satisfy $\Gamma^T\Gamma = I_n - R^{-T}X^TXR^{-1}$ (so Γ would be the Cholesky factor of $I_n - R^{-T}X^TXR^{-1}$), and let $\sigma_n(\Gamma)$ be the smallest singular value of Γ. Notice that for fixed R, $\Gamma^T\Gamma \to I_n$ as $X \to 0$, so $\sigma_n(\Gamma) \to 1$. First we derive some relationships among U, R, X and Γ.

1) From (5.1.1) obviously we have

$$\|U\|_2 \le \|R\|_2, \qquad \|X\|_2 \le \|R\|_2. \tag{5.2.1}$$

2) From (5.1.1) it follows that

$$RU^{-1}U^{-T}R^T = (I_n - R^{-T}X^TXR^{-1})^{-1},$$

so that taking the 2-norm gives

$$\|RU^{-1}\|_2 = \frac{1}{\sigma_n(\Gamma)}. \tag{5.2.2}$$

3) From (5.1.1) we have

$$U^{-T}X^TXU^{-1} = U^{-T}R^TRU^{-1} - I_n,$$

which, combined with (5.2.2), gives

$$\|XU^{-1}\|_2 = \sqrt{\|RU^{-1}\|_2^2 - 1} = \frac{\sqrt{1-\sigma_n^2(\Gamma)}}{\sigma_n(\Gamma)} = \sqrt{1-\sigma_n^2(\Gamma)}\|RU^{-1}\|_2. \tag{5.2.3}$$

4) From (5.1.1) we have

$$R^{-T}X^TXR^{-1} = I_n - R^{-T}U^TUR^{-1},$$

which, combined with (5.2.2), gives

$$\|XR^{-1}\|_2 = \sqrt{1-\sigma_{\min}^2(UR^{-1})} = \sqrt{1-\frac{1}{\|RU^{-1}\|_2^2}} = \sqrt{1-\sigma_n^2(\Gamma)}. \tag{5.2.4}$$

5) By (5.2.2) we have

$$\|R\|_2 = \|RU^{-1}U\|_2 \le \|RU^{-1}\|_2\|U\|_2 = \frac{\|U\|_2}{\sigma_n(\Gamma)}. \tag{5.2.5}$$

6) Finally from (5.2.4) we see

$$\frac{\|X\|_2}{\|R\|_2} \le \|XR^{-1}\|_2 = \sqrt{1-\sigma_n^2(\Gamma)}. \tag{5.2.6}$$

Now we derive the basic result on how U changes as R and X change.

Theorem 5.2.1 *Suppose we have an upper triangular matrix $R \in \mathbf{R}^{n\times n}$ and a matrix $X \in \mathbf{R}^{k\times n}$ with the Cholesky factorization $U^TU = R^TR - X^TX$, where $U \in \mathbf{R}^{n\times n}$ is upper triangular with positive diagonal elements. Let G be a real $n \times n$ matrix, and let F be a real $k \times n$ matrix. Assume $\Delta R = \epsilon G$ and $\Delta X = \epsilon F$, for some $\epsilon \ge 0$. If*

$$\|\Delta RR^{-1}\|_2 < 1, \qquad \frac{\|XR^{-1}\|_2 + \|\Delta XR^{-1}\|_2}{1 - \|\Delta RR^{-1}\|_2} < 1, \tag{5.2.7}$$

then there is a unique Cholesky factorization

$$(U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R) - (X+\Delta X)^T(X+\Delta X), \quad (5.2.8)$$

with ΔU *satisfying*

$$\Delta U = \epsilon \dot{U}(0) + O(\epsilon^2), \quad (5.2.9)$$

where $\dot{U}(0)$ *is defined by the unique Cholesky factorization*

$$U^T(t)U(t) = (R+tG)^T(R+tG) - (X+tF)^T(X+tF), \qquad |t| \le \epsilon, \quad (5.2.10)$$

and so satisfies the equations

$$U^T\dot{U}(0) + \dot{U}^T(0)U = R^TG + G^TR - X^TF - F^TX, \quad (5.2.11)$$

$$\dot{U}(0) = \text{up}[U^{-T}(R^TG + G^TR - X^TF - F^TX)U^{-1}]U, \quad (5.2.12)$$

where the 'up' notation is defined by (1.2.3).

Proof. If $\|\Delta R R^{-1}\|_2 \le 1$, then it is easy to show $R+tG$ is nonsingular for all $|t| \le \epsilon$. Notice for all $|t| \le \epsilon$,

$$\begin{aligned}&(R+tG)^T(R+tG) - (X+tF)^T(X+tF)\\ &= (R+tG)^T[I_n - (R+tG)^{-T}(X+tF)^T(X+tF)(R+tG)^{-1}](R+tG),\end{aligned}$$

and

$$\begin{aligned}\|(X+tF)(R+tG)^{-1}\|_2 &= \|(XR^{-1}+tFR^{-1})(I+tGR^{-1})^{-1}\|_2\\ &\le \frac{\|XR^{-1}\|_2 + \|\Delta X R^{-1}\|_2}{1-\|\Delta R R^{-1}\|_2},\end{aligned}$$

then if (5.2.7) holds, $(R+tG)^T(R+tG)-(X+tF)^T(X+tF)$ is positive definite and has the unique Cholesky factorization (5.2.10). Notice that $U(0) = U$ and $U(\epsilon) = U+\Delta U$, so (5.2.8) holds.

It is easy to verify that $U(t)$ is twice continuously differentiable for $|t| \le \epsilon$ from the algorithm for the Cholesky factorization. If we differentiate (5.2.10) and set $t = 0$

in the result, we obtain (5.2.11), which, like (2.2.5), is a linear equation uniquely defining the elements of upper triangular $\dot{U}(0)$ in terms of the elements of G and F. With the 'up' notation in (1.2.3) we see (5.2.12) holds. Finally the Taylor expansion for $U(t)$ about $t = 0$ gives (5.2.9) at $t = \epsilon$. □

By Theorem 5.2.1 we derive a new first-order perturbation bounds for the block Cholesky downdating problem, from which the first-order perturbation bound given by Sun [50, 1995] follows.

Theorem 5.2.2 *Suppose we have an upper triangular matrix $R \in \mathbf{R}^{n\times n}$ and a matrix $X \in \mathbf{R}^{k\times n}$ with the Cholesky factorization $U^TU = R^TR - X^TX$, where $U \in \mathbf{R}^{n\times n}$ is upper triangular with positive diagonal elements. Let ΔR be a real $n \times n$ matrix, and let ΔX be a real $k \times n$ matrix. Define $\epsilon_R \equiv \|\Delta R\|_F/\|R\|_2$ and $\epsilon_X \equiv \|\Delta X\|_F/\|X\|_2$. Set $\epsilon \equiv \max\{\epsilon_R, \epsilon_X\}$. If*

$$\kappa_2(R)\epsilon_R < 1, \qquad \frac{(1+\epsilon_X)\|X\|_2\,\|R^{-1}\|_2}{1-\kappa_2(R)\epsilon_R} < 1, \tag{5.2.13}$$

then there is a unique Cholesky factorization

$$(U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R) - (X+\Delta X)^T(X+\Delta X),$$

where

$$\frac{\|\Delta U\|_F}{\|U\|_2} \le \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sigma_n(\Gamma)}\,\epsilon_R + \sqrt{2}\,\frac{\sqrt{1-\sigma_n^2(\Gamma)}\|U^{-1}\|_2\|X\|_2}{\sigma_n(\Gamma)}\,\epsilon_X + O(\epsilon^2). \tag{5.2.14}$$

Proof. Let $G \equiv \Delta R/\epsilon$ and $F \equiv \Delta X/\epsilon$ (if $\epsilon = 0$, the theorem is trivial), then

$$\|G\|_F = \|R\|_2\,\epsilon_R/\epsilon, \qquad \|F\|_F = \|X\|_2\,\epsilon_X/\epsilon. \tag{5.2.15}$$

It is easy to verify that (5.2.13) implies that (5.2.7) holds, so Theorem 5.2.1 is applicable here. From (5.2.12) and the fact that for any symmetric B, $\|\operatorname{up}(B)\|_F \le \frac{1}{\sqrt{2}}\|B\|_F$ (see (1.2.7)) we have with (5.2.15) that

$$\begin{aligned}
\|\dot{U}(0)\|_F &\le \frac{1}{\sqrt{2}}\|U^{-T}(R^TG + G^TR - X^TF - F^TX)U^{-1}\|_F\,\|U\|_2 \\
&\le \sqrt{2}\,\|U\|_2\,\|U^{-1}\|_2\,(\|RU^{-1}\|_2\,\|G\|_F + \|XU^{-1}\|_2\,\|F\|_F), \\
&= \sqrt{2}\,\|U\|_2\,\|U^{-1}\|_2\,(\|RU^{-1}\|_2\,\|R\|_2\,\epsilon_R/\epsilon + \|XU^{-1}\|_2\,\|X\|_2\,\epsilon_X/\epsilon)
\end{aligned}$$

which, combined with (5.2.2) and (5.2.3), gives

$$\|\dot{U}(0)\|_F \leq \sqrt{2}\,\frac{\|U\|_2\,\|U^{-1}\|_2}{\sigma_n(\Gamma)}(\|R\|_2\,\epsilon_R/\epsilon + \sqrt{1-\sigma_n^2(\Gamma)}\|X\|_2\,\epsilon_X/\epsilon)$$

Then (5.2.14) follows from the Taylor expansion (5.2.9). □

From (5.2.14) we see

$$\phi_R \equiv \sqrt{2}\,\frac{\|U^{-1}\|_2\,\|R\|_2}{\sigma_n(\Gamma)}, \qquad \phi_X \equiv \sqrt{2}\,\frac{\sqrt{1-\sigma_n^2(\Gamma)}\,\|U^{-1}\|_2\,\|X\|_2}{\sigma_n(\Gamma)} \tag{5.2.16}$$

can be regarded as the condition estimators for U with respect to relative changes in R and X, respectively. Notice from (5.2.1) we see $\phi_R > \phi_X$, so we can define a new overall condition estimator

$$\phi \equiv \phi_R = \sqrt{2}\,\frac{\|U^{-1}\|_2\,\|R\|_2}{\sigma_n(\Gamma)}. \tag{5.2.17}$$

If we rewrite (5.2.14) as

$$\frac{\|\Delta U\|_F}{\|U\|_2} \leq \sqrt{2}\,\frac{\|U^{-1}\|_2\,\|R\|_2}{\sigma_n(\Gamma)}(\epsilon_R + \sqrt{1-\sigma_n^2(\Gamma)}\,\frac{\|X\|_2}{\|R\|_2}\,\epsilon_X) + O(\epsilon^2),$$

and combine it with (5.2.5) and (5.2.6), then we obtain Sun's bound

$$\frac{\|\Delta U\|_F}{\|U\|_2} \leq \sqrt{2}\,\frac{\kappa_2(U)}{\sigma_n^2(\Gamma)}(\epsilon_R + (1-\sigma_n^2(\Gamma))\epsilon_X) + O(\epsilon^2), \tag{5.2.18}$$

which leads to the overall condition estimator proposed by Sun:

$$\beta = \sqrt{2}\,\frac{\kappa_2(U)}{\sigma_n^2(\Gamma)}, \tag{5.2.19}$$

We have seen the right hand side of (5.2.14) is never worse than that of (5.2.18), and also

$$\phi \leq \beta. \tag{5.2.20}$$

Although ϕ is a minor improvement on β, it is still not what we want. We can see this from the asymptotic behavior of these condition estimators. The Cholesky factorization is unique, so as $X \to 0$, $U \to R$, and $X^T\Delta X \to 0$ in (5.2.8). Now for any upper triangular perturbation ΔR in R, $\Delta U \to \Delta R$, so the true condition number should approach unity. Here $\beta, \phi \to \sqrt{2}\kappa_2(R)$. The next section shows how we can overcome this inadequacy.

5.3 New perturbation results

In Section 5.2 we saw the key to deriving first-order perturbation bounds for U in the block Cholesky downdating problem is the equation (5.2.11). We will now analyze it by two approaches. The first approach, the matrix-vector equation approach, gives sharp perturbation bounds, which lead to the condition numbers for the block Cholesky downdating problem, while the second, the the matrix equation approach, gives a clear improvement on other earlier results, and provides practical condition estimators for the true condition numbers. All our discussion is based on the same assumptions as in Theorem 5.2.2.

5.3.1 Matrix-vector equation analysis

The matrix-vector equation approach views the matrix equation (5.2.11) as a large matrix-vector equation.

First assume ΔR is a *general* real $n \times n$ matrix. It is easy to show (5.2.11) can be rewritten in the following matrix-vector form (cf. Chapter3):

$$W_U \operatorname{uvec}(\dot{U}(0)) = Z_R \operatorname{vec}(G) - Y_X \operatorname{vec}(F), \tag{5.3.1}$$

where $W_U \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ is

$$\left[\begin{array}{ccccccc|ccccc}
u_{11} & & & & & & & & & & \\
u_{12} & u_{11} & & & & & & & & & \\
 & u_{12} & u_{22} & & & & & & & & \\
u_{13} & & & u_{11} & & & & & & & \\
 & u_{13} & u_{23} & u_{12} & u_{22} & & & & & & \\
 & & & u_{13} & u_{23} & u_{33} & & & & & \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
\hline
u_{1n} & & & & & & & u_{11} & & & \\
 & u_{1n} & u_{2n} & & & & & u_{12} & u_{22} & & \\
 & & & u_{1n} & u_{2n} & u_{3n} & & u_{13} & u_{23} & u_{33} & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & u_{1n} & u_{2n} & u_{3n} & \cdot \; u_{nn}
\end{array}\right],$$

$Z_R \in \mathbf{R}^{\frac{n(n+1)}{2} \times n^2}$ is

$$\left[\begin{array}{ccc|ccc|c|ccc}
r_{11} & & & & & & & & & \\
\hline
r_{12} & r_{22} & & r_{11} & & & & & & \\
 & & & r_{12} & r_{22} & & & & & \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & \\
\hline
r_{1n} & r_{2n} & \cdot \; r_{nn} & & & & & r_{11} & & \\
 & & & r_{1n} & r_{2n} & \cdot \; r_{nn} & & r_{12} & r_{22} & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & r_{1n} & r_{2n} & \cdot \; r_{nn}
\end{array}\right],$$

and $Y_X \in \mathbf{R}^{\frac{n(n+1)}{2} \times kn}$ is

$$\left[\begin{array}{cccc|cccc|c|cccc}
x_{11} & x_{21} & \cdot & x_{k1} & & & & & & & & & \\
\hline
x_{12} & x_{22} & \cdot & x_{k2} & x_{11} & x_{21} & \cdot & x_{k1} & & & & & \\
 & & & & x_{12} & x_{22} & \cdot & x_{k2} & & & & & \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
\hline
x_{1n} & x_{2n} & \cdot & x_{kn} & & & & & & x_{11} & x_{21} & \cdot & x_{k1} \\
 & & & & x_{1n} & x_{2n} & \cdot & x_{kn} & & x_{12} & x_{22} & \cdot & x_{kn} \\
 & & & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & x_{1n} & x_{2n} & \cdot & x_{nn}
\end{array}\right]$$

Since U is nonsingular, W_U is also, and from (5.3.1)

$$\text{uvec}(\dot{U}(0)) = W_U^{-1} Z_R \,\text{vec}(G) - W_U^{-1} Y_X \,\text{vec}(F). \tag{5.3.2}$$

Remembering $\dot{U}(0)$ is upper triangular, we see

$$\begin{aligned}
\|\dot{U}(0)\|_F &\leq \|W_U^{-1} Z_R\|_2 \,\|G\|_F + \|W_U^{-1} Y_X\|_2 \,\|F\|_F, \qquad (5.3.3)\\
&= \|W_U^{-1} Z_R\|_2 \,\|R\|_2 \,\epsilon_R/\epsilon + \|W_U^{-1} Y_X\|_2 \,\|X\|_2 \,\epsilon_X/\epsilon, \quad \text{(using (5.2.15))}
\end{aligned}$$

where for any R and X equality can be made by choosing G and F such that

$$\|W_U^{-1} Z_R \,\text{vec}(G)\|_2 = \|W_U^{-1} Z_R\|_2 \,\|G\|_F, \qquad F = 0, \tag{5.3.4}$$

$$\text{or} \qquad G = 0, \qquad \|W_U^{-1} Y_X \,\text{vec}(F)\|_2 = \|W_U^{-1} Y_X\|_2 \,\|F\|_F. \tag{5.3.5}$$

Then from the Taylor expansion (5.2.9), we see

$$\frac{\|\Delta U\|_F}{\|U\|_2} \leq \frac{\|W_U^{-1} Z_R\|_2 \|R\|_2}{\|U\|_2} \epsilon_R + \frac{\|W_U^{-1} Y_X\|_2 \|X\|_2}{\|U\|_2} \epsilon_X + O(\epsilon^2), \tag{5.3.6}$$

and the condition numbers for U with respect to relative changes in R and X are (here subscript $_G$ refers to *general* ΔR, and later the subscript $_T$ will refer to *upper triangular* ΔR)

$$\kappa_{RG}(R,X) \equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta U\|_F}{\epsilon\|U\|_2} : (U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R) - X^TX,\ \epsilon = \|\Delta R\|_F/\|R\|_2\Big\} \tag{5.3.7}$$

$$= \frac{\|W_U^{-1}Z_R\|_2\|R\|_2}{\|U\|_2},$$

and

$$\kappa_X(R,X) \equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta U\|_F}{\epsilon\|U\|_2} : (U+\Delta U)^T(U+\Delta U) = R^TR - (X+\Delta X)^T(X+\Delta X),\ \epsilon = \|\Delta X\|_F/\|X\|_2\Big\} \tag{5.3.8}$$

$$= \frac{\|W_U^{-1}Y_X\|_2\|X\|_2}{\|U\|_2},$$

respectively. Then a whole condition number for the *Cholesky downdating* problem with *general* ΔR can be defined as

$$\kappa_{CDG}(R,X) \equiv \max\{\kappa_{RG}(R,X),\ \kappa_X(R,X)\}. \tag{5.3.9}$$

By the definitions of $\kappa_{RG}(R,X)$ and $\kappa_X(X,R)$, it is easy to verify from (5.2.14) and (5.2.16) that

$$\kappa_{RG}(R,X) \le \phi_R, \quad \kappa_X(R,X) \le \phi_X, \tag{5.3.10}$$

therefore

$$\kappa_{CDG}(R,X) \le \phi. \tag{5.3.11}$$

It is easy to observe that if $X \to 0$, $\kappa_{CDG}(R,X) \to \|W_R^{-1}Z_R\|_2$, where W_R is just W_U with each entry u_{ij} replaced by r_{ij}. If R was found using the standard pivoting strategy in the Cholesky factorization, then $\|W_R^{-1}Z_R\|_2$ has a bound which is a function of n alone (see Theorem 3.4.2). So in this case our condition number $\kappa_{CDG}(R,X)$ also has a bound which is a function of n alone as $X \to 0$.

Now we consider the case where ΔR is ***upper triangular***. (5.2.11) can now be rewritten in the following matrix-vector form:

$$W_U\,\mathrm{uvec}(\dot{U}(0)) = W_R\,\mathrm{uvec}(G) - Y_X\,\mathrm{vec}(F), \tag{5.3.12}$$

where $W_U \in \mathbf{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ and $Y_X \in \mathbf{R}^{\frac{n(n+1)}{2}\times kn}$ are defined as before, and $W_R \in \mathbf{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is just W_U with each entry u_{ij} replaced by r_{ij}. Since U is nonsingular, W_U is also, and from (5.3.12)

$$\operatorname{uvec}(\dot{U}(0)) = W_U^{-1} W_R \operatorname{uvec}(G) - W_U^{-1} Y_X \operatorname{vec}(F), \tag{5.3.13}$$

so taking the 2-norm gives

$$\begin{aligned} \|\dot{U}(0)\|_F &\leq \|W_U^{-1}W_R\|_2 \|G\|_F + \|W_U^{-1}Y_X\|_2 \|F\|_F, \qquad (5.3.14)\\ &= \|W_U^{-1}W_R\|_2 \|R\|_2 \epsilon_X/\epsilon + \|W_U^{-1}Y_X\|_2 \epsilon_R/\epsilon, \quad \text{(using (5.2.15))} \end{aligned}$$

where, like (5.3.3), (5.3.14) will become an equality if we choose G and F as in (5.3.4) and (5.3.5) with Z_R there replaced by W_R. Then from the Taylor expansion (5.2.9), we see

$$\frac{\|\Delta U\|_F}{\|U\|_2} \leq \frac{\|W_U^{-1}W_R\|_2 \|R\|_2}{\|U\|_2}\epsilon_R + \frac{\|W_U^{-1}Y_X\|_2 \|X\|_2}{\|U\|_2}\epsilon_X + O(\epsilon^2). \tag{5.3.15}$$

and the condition numbers for U with respect to relative changes in R and X are (subscript τ indicates upper triangular ΔR)

$$\begin{aligned} \kappa_{R\tau}(R,X) &\equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta U\|_F}{\epsilon\|U\|_2} : (U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R)\\ &\qquad - X^TX,\ \epsilon = \|\Delta R\|_F/\|R\|_2\Big\} \qquad (5.3.16)\\ &= \frac{\|W_U^{-1}W_R\|_2\|R\|_2}{\|U\|_2}, \end{aligned}$$

and

$$\begin{aligned} \kappa_X(R,X) &\equiv \limsup_{\epsilon\to 0}\Big\{\frac{\|\Delta U\|_F}{\epsilon\|U\|_2} : (U+\Delta U)^T(U+\Delta U) = R^TR\\ &\qquad - (X+\Delta X)^T(X+\Delta X),\ \epsilon = \|\Delta X\|_F/\|X\|_2\Big\} \qquad (5.3.17)\\ &= \frac{\|W_U^{-1}Y_X\|_2\|X\|_2}{\|U\|_2}, \end{aligned}$$

respectively. Note $\kappa_X(R,X)$ is the same as that defined in (5.3.8). Then a whole condition number for the Cholesky downdating problem with *upper triangular* ΔR

can be defined as

$$\kappa_{CDT}(R,X) \equiv \max\{\kappa_{RT}(R,X), \kappa_X(R,X)\}. \tag{5.3.18}$$

Since $\kappa_{RG}(R,X)$ is for a general ΔR, certainly we have

$$\kappa_{RT}(R,X) \leq \kappa_{RG}(R,X), \tag{5.3.19}$$

which can also be proved directly by the fact that the columns of W_R form a proper subset of the columns of Z_R. Thus from (5.3.9) and (5.3.18) we see

$$\kappa_{CDT}(R,X) \leq \kappa_{CDG}(R,X). \tag{5.3.20}$$

If as well $X \to 0$, then since $U \to R$, $W_U^{-1}W_R \to I_{\frac{n(n+1)}{2}}$, and $\kappa_{CDT}(R,X) \to 1$. So in this case the Cholesky downdating problem becomes very well conditioned no matter how ill-conditioned R or U is.

Finally we summarize the results above as the following theorem.

Theorem 5.3.1 *With the same assumptions as in Theorem* 5.2.2, *there is a unique Cholesky factorization*

$$(U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R) - (X+\Delta X)^T(X+\Delta X),$$

where for general ΔR,

$$\begin{aligned} \frac{\|\Delta U\|_F}{\|U\|_2} &\leq \kappa_{RG}(R,X)\epsilon_R + \kappa_X(R,X)\epsilon_X + O(\epsilon^2) \\ &\leq \kappa_{CDG}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2), \end{aligned}$$

and for upper triangular ΔR,

$$\begin{aligned} \frac{\|\Delta U\|_F}{\|U\|_2} &\leq \kappa_{RT}(R,X)\epsilon_R + \kappa_X(R,X)\epsilon_X + O(\epsilon^2) \\ &\leq \kappa_{CDT}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2). \end{aligned}$$

There are the following relationships among the various measures of sensitivity of the problem (see (5.3.10), (5.3.11), (5.3.19) and (5.3.20)) :

$$\kappa_{RT}(R,X) \leq \kappa_{RG}(R,X) \leq \phi_R, \qquad \kappa_X(R,X) \leq \phi_X,$$
$$\kappa_{CDT}(R,X) \leq \kappa_{CDG}(R,X) \leq \phi. \quad \square$$

5.3.2 Matrix equation analysis

As far as we see, the condition numbers obtained in the last section are expensive to compute or estimate directly with the usual approach. We now use the matrix equation approach to obtain practical condition estimators.

In Theorem 5.2.2 we used the expression of $\dot{U}(0)$ in (5.2.12) to derive a new first-order perturbation bound (5.2.14), from which Sun's bound was derived. Now we again look at (5.2.12), repeated here for clarity:

$$\dot{U}(0) = \mathrm{up}[U^{-T}(R^TG + G^TR - X^TF - F^TX)U^{-1}]U. \tag{5.3.21}$$

Let $\mathbf{D}_n$ be the set of all $n \times n$ real positive definite diagonal matrices. For any $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathbf{D}_n$, let $U = D\bar{U}$. Note that for any matrix B we have $\mathrm{up}(BD^{-1}) = \mathrm{up}(B)D^{-1}$ and $\mathrm{up}(D^{-1}B) = D^{-1}\,\mathrm{up}(B)$.

First with *general* ΔR we have from (5.3.21) that

$$\begin{aligned}\dot{U}(0) &= \{\mathrm{up}(U^{-T}R^TG\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}G^TRU^{-1})D\}\bar{U} \\ &\quad - \{\mathrm{up}(U^{-T}X^TF\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}F^TXU^{-1})D\}\bar{U},\end{aligned}$$

so taking the F-norm gives

$$\begin{aligned}\|\dot{U}(0)\|_F &\le \|\mathrm{up}(U^{-T}R^TG\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}G^TRU^{-1})D\|_F\,\|\bar{U}\|_2 \\ &\quad + \|\mathrm{up}(U^{-T}X^TF\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}F^TXU^{-1})D\|_F\,\|\bar{U}\|_2. \end{aligned}\tag{5.3.22}$$

Lemma 3.4.1 shows for any $B \in \mathbf{R}^{n\times n}$

$$\|\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D\|_F \le \sqrt{1+\zeta_D^2}\,\|B\|_F,$$

where $\zeta_D = \max_{1\le i<j\le n}\{\delta_j/\delta_i\}$. Thus from (5.3.22) we have

$$\begin{aligned}\|\dot{U}(0)\|_F &\le \sqrt{1+\zeta_D^2}\,(\|U^{-T}R^TG\bar{U}^{-1}\|_F + \|\dot{U}^{-T}X^TF\bar{U}^{-1}\|_2)\|\bar{U}\|_2 \\ &\le \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{U})(\|RU^{-1}\|_2\,\|G\|_F + \|XU^{-1}\|_2\,\|F\|_F). \\ &= \sqrt{1+\zeta_D^2}\,\frac{\kappa_2(\bar{U})}{\sigma_n(\Gamma)}(\|G\|_F + \sqrt{1-\sigma_n^2(\Gamma)}\,\|F\|_F), \quad \text{(using (5.2.2), (5.2.3))} \\ &= \sqrt{1+\zeta_D^2}\,\frac{\kappa_2(\bar{U})}{\sigma_n(\Gamma)}(\|R\|_2\,\epsilon_R/\epsilon + \sqrt{1-\sigma_n^2(\Gamma)}\,\|X\|_2\,\epsilon_X/\epsilon) \quad \text{(using (5.2.15))}\end{aligned}$$

which leads to the following perturbation bound in terms of relative changes

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} &\leq \sqrt{1+\zeta_D^2}\,\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2}\,\epsilon_R \\ &\quad + \sqrt{1+\zeta_D^2}\,\sqrt{1-\sigma_n^2(\Gamma)}\,\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|X\|_2}{\|U\|_2}\,\epsilon_X + O(\epsilon^2). \quad (5.3.23)\end{aligned}$$

Naturally we define the following two quantities as condition estimators for U with respect to relative changes in R and X, respectively:

$$\kappa'_{RG}(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_{RG}(R,X,D), \qquad \kappa'_X(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_X(R,X,D), \quad (5.3.24)$$

where

$$\kappa'_{RG}(R,X,D) \equiv \sqrt{1+\zeta_D^2}\,\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2}, \quad (5.3.25)$$

$$\kappa'_X(R,X,D) \equiv \sqrt{1+\zeta_D^2}\,\sqrt{1-\sigma_n^2(\Gamma)}\,\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|X\|_2}{\|U\|_2}. \quad (5.3.26)$$

Then an overall condition estimator can be defined as

$$\kappa'_{CDG}(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_{CDG}(R,X,D), \quad (5.3.27)$$

where

$$\kappa'_{CDG}(R,X,D) \equiv \max\{\kappa'_{RG}(R,X,D),\, \kappa'_X(R,X,D)\}.$$

Since $\|X\|_2 \leq \|R\|_2$, we see

$$\kappa'_{CDG}(R,X,D) = \kappa'_{RG}(R,X,D) \geq \kappa'_X(R,X,D),$$

which gives

$$\kappa'_{CDG}(R,X) = \kappa'_{RG}(R,X) \geq \kappa'_X(R,X). \quad (5.3.28)$$

Therefore with these, we have from (5.3.23) that

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} &\leq \kappa'_{RG}(R,X)\,\epsilon_R + \kappa'_X(R,X)\,\epsilon_X + O(\epsilon^2) \quad (5.3.29)\\ &\leq \kappa'_{CDG}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2).\end{aligned}$$

Clearly if we take $D = I_n$, (5.3.23) will become (5.2.14), and

$$\kappa'_{RG}(R,X) \le \kappa'_{RG}(R,X,I_n) = \phi_R, \quad \kappa'_X(R,X) \le \kappa'_X(R,X,I_n) = \phi_X, \quad (5.3.30)$$

$$\kappa'_{CDG}(R,X) \le \kappa'_{CDG}(R,X,I_n) = \phi. \quad (5.3.31)$$

It is not difficult to give an example to show ϕ can be arbitrarily larger than $\kappa'_{CDG}(R,X)$, as can be seen from the following asymptotic behaviour.

If $X \to 0$ we saw $U \to R$ and $\sigma_n(\Gamma) \to 1$, so

$$\kappa'_{CDG}(R,X,D) \to \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R).$$

It is shown in Theorem 3.4.4 that with an appropriate choice of D, $\sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}R)$ has a bound which is a function of n only, if R was found using the standard pivoting strategy in the Cholesky factorization, and in this case, we see $\kappa'_{CDG}(R,X)$ is bounded independently of $\kappa_2(R)$ as $X \to 0$, for *general* ΔR. At the end of this section we give an even stronger result when $X \to 0$ for the case of upper triangular ΔR. Note in the case here that ϕ in (5.2.17) can be made as large as we like, and thus arbitrarily larger than $\kappa'_{CDG}(R,X)$.

By the definitions of $\kappa_{RG}(R,X)$ and $\kappa_X(R,X)$ respectively in (5.3.7) and (5.3.8), we can easily verify from (5.3.29) that

$$\kappa_{RG}(R,X) \le \kappa'_{RG}(R,X), \qquad \kappa_X(R,X) \le \kappa'_X(R,X), \quad (5.3.32)$$

therefore

$$\kappa_{CDG}(R,X) \le \kappa'_{CDG}(R,X). \quad (5.3.33)$$

In the case where ΔR is *upper triangular* (so G is upper triangular), we can refine the analysis further. From (5.3.21) we have

$$\begin{aligned} \dot{U}(0) \;=\; & [\,\mathrm{up}(U^{-T}R^TGU^{-1} + U^{-T}G^TRU^{-1}) \\ & - \mathrm{up}(U^{-T}X^TFU^{-1} + U^{-T}F^TXU^{-1})]U. \end{aligned} \quad (5.3.34)$$

Notice with the 'slt', 'sut' and 'diag' notation defined in (1.2.1) and (1.2.2),

$$\begin{aligned} & U^{-T}R^TGU^{-1} + U^{-T}G^TRU^{-1} \\ = \; & [\mathrm{slt}(U^{-T}R^T) + \mathrm{diag}(U^{-T}R^T)]GU^{-1} + U^{-T}G^T[\mathrm{sut}(RU^{-1}) + \mathrm{diag}(RU^{-1})] \\ = \; & \mathrm{diag}(U^{-T}R^T)\cdot GU^{-1} + U^{-T}G^T\cdot\mathrm{diag}(RU^{-1}) \\ & + \mathrm{slt}(U^{-T}R^T)\cdot GU^{-1} + U^{-T}G^T\cdot\mathrm{sut}(RU^{-1}). \end{aligned} \tag{5.3.35}$$

But for any *upper* triangular matrix T we have

$$\mathrm{up}(T) + \mathrm{up}(T^T) = T,$$

so that if we define $T \equiv \mathrm{diag}(U^{-T}R^T)\cdot GU^{-1}$, then

$$\mathrm{up}[\mathrm{diag}(U^{-T}R^T)\cdot GU^{-1} + U^{-T}G^T\cdot\mathrm{diag}(RU^{-1})] = \mathrm{diag}(U^{-T}R^T)\cdot GU^{-1}. \tag{5.3.36}$$

Thus from (5.3.34), (5.3.35) and (5.3.36) we obtain

$$\begin{aligned} \dot{U}(0) \; = \; & \mathrm{diag}(U^{-T}R^T)\cdot G + \{\,\mathrm{up}[\mathrm{slt}(U^{-T}R^T)\cdot GU^{-1} + U^{-T}G^T\cdot\mathrm{sut}(RU^{-1})] \\ & - \mathrm{up}(U^{-T}X^TFU^{-1} + U^{-T}F^TXU^{-1})\}U. \end{aligned} \tag{5.3.37}$$

As before, let $U = D\bar{U}$, where $D = \mathrm{diag}(\delta_1,\ldots,\delta_n) \in \mathbf{D}_n$. From (5.3.37) it follows that

$$\begin{aligned} & \|\dot{U}(0)\|_F \le \|\mathrm{diag}(U^{-T}R^T)\|_2\,\|G\|_F \\ & + \|\mathrm{up}[\mathrm{slt}(U^{-T}R^T)\cdot G\bar{U}^{-1}] + D^{-1}\mathrm{up}[\bar{U}^{-T}G^T\cdot\mathrm{sut}(RU^{-1})]D\|_F\,\|\bar{U}\|_2 \\ & + \|\mathrm{up}(U^{-T}X^TF\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}F^TXU^{-1})D\|_F\,\|\bar{U}\|_2. \end{aligned}$$

Then, applying (5.3.2) to this, we have

$$\begin{aligned} \|\dot{U}(0)\|_F \; \le \; & \|\mathrm{diag}(U^{-T}R^T)\|_2\,\|G\|_F + \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{U})\|\mathrm{sut}(RU^{-1})\|_2\,\|G\|_F \\ & + \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{U})\|XU^{-1}\|_2\,\|F\|_F \\ \le \; & (\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{U})\|\mathrm{sut}(RU^{-1})\|_2)\|R\|_2\,\epsilon_R/\epsilon \\ & + \sqrt{1+\zeta_D^2}\,\kappa_2(\bar{U})\|XU^{-1}\|_2\,\|X\|_2\,\epsilon_X/\epsilon, \quad \text{(using (5.2.15))} \end{aligned}$$

which leads to the following perturbation bound

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} \leq\ & (\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2)\frac{\|R\|_2}{\|U\|_2}\epsilon_R \\ & + \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)\|XU^{-1}\|_2\,\frac{\|X\|_2}{\|U\|_2}\epsilon_X + O(\epsilon^2) \qquad (5.3.38)\end{aligned}$$

Comparing this with (5.3.23) and noticing (5.2.3), we see the coefficient multiplying ϵ_X does not change, so $\kappa'_X(R,X)$ defined in (5.3.24) can still be regarded as a condition estimator for U with respect to changes in X. But we now need to define a new condition estimator for U with respect to *upper triangular* changes in R, that is

$$\kappa'_{RT}(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_{RT}(R,X,D),$$

where

$$\kappa'_{RT}(R,X,D) \equiv (\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2)\frac{\|R\|_2}{\|U\|_2}. \quad (5.3.39)$$

Thus an overall condition estimator can be defined as

$$\kappa'_{CDT}(R,X) = \inf_{D\in\mathbf{D}_n} \kappa'_{CDT}(R,X,D), \qquad (5.3.40)$$

where

$$\kappa'_{CDT}(R,X,D) = \max\{\kappa'_{RT}(R,X,D),\ \kappa'_X(R,X,D)\}.$$

Obviously we have

$$\kappa'_{CDT}(R,X) = \max\{\kappa'_{RT}(R,X),\ \kappa'_X(R,X)\}. \qquad (5.3.41)$$

With these, we have from (5.3.38) that

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} &\leq \kappa'_{RT}(R,X)\epsilon_R + \kappa'_X(R,X)\epsilon_X + O(\epsilon^2) \qquad (5.3.42)\\ &\leq \kappa'_{CDT}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2).\end{aligned}$$

What is the relationship between $\kappa'_{CDT}(R,X)$ and $\kappa'_{CDG}(R,X) = \kappa'_{RG}(R,X)$? For any $n\times n$ upper triangular matrix $T=(t_{ij})$, observe the following two facts:

1) t_{ii}, $i = 1, 2, \ldots, n$ are the eigenvalues of T, so that $|t_{ii}| \leq \|T\|_2$. From this it follows that

$$\|\mathrm{diag}(T)\|_2 \leq \|T\|_2.$$

2)

$$\|\mathrm{sut}(T)\|_2 \leq \|\mathrm{sut}(T)\|_F \leq \|T\|_F \leq \sqrt{n}\,\|T\|_2.$$

(Note: In fact we can prove a slightly sharper inequality $\|\mathrm{sut}(T)\|_2 \leq \sqrt{n-1}\,\|T\|_2$).

Therefore

$$\begin{aligned}
\kappa'_{RT}(R,X,D) &= (\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2)\frac{\|R\|_2}{\|U\|_2} \\
&\leq (\|RU^{-1}\|_2 + \sqrt{n}\sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)\|RU^{-1}\|_2)\frac{\|R\|_2}{\|U\|_2} \\
&< (1+\sqrt{n})\sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)\|RU^{-1}\|_2\,\frac{\|R\|_2}{\|U\|_2} \\
&= (1+\sqrt{n})\sqrt{1+\zeta_D^2}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2} \quad (\text{ using (5.2.2)}) \\
&= (1+\sqrt{n})\,\kappa'_{RG}(R,X,D),
\end{aligned}$$

so that

$$\kappa'_{RT}(R,X) \leq (1+\sqrt{n})\kappa'_{RG}(R,X). \tag{5.3.43}$$

Thus we have from (5.3.28) and (5.3.41) that

$$\kappa'_{CDT}(R,X) \leq (1+\sqrt{n})\kappa'_{CDG}(R,X). \tag{5.3.44}$$

On the other hand, $\kappa'_{CDT}(R,X)$ can be arbitrarily smaller than $\kappa'_{CDG}(R,X)$. This can be seen from the asymptotic behaviour, which is important in its own right. As $X \to 0$, since $U \to R$, $\sigma_n(\Gamma) \to 1$ and $RU^{-1} \to I_n$, we have

$$\kappa'_{CDT}(R,X,I_n) \to 1,$$

so for upper triangular changes in R, whether pivoting was used in finding R or not,

$$\kappa'_{CDT}(R,X) \to 1.$$

Thus when $X \to 0$, the bound in (5.3.42) reflects the true sensitivity of the problem. For the case of general ΔR, if we do not use pivoting it is straightforward to make $\kappa'_{CDG}(R,X)$ in (5.3.27) arbitrarily large even with $X = 0$, see (5.3.25).

By the definition of $\kappa_{RT}(R,X)$ in (5.3.16), we can easily verify from (5.3.42) that

$$\kappa_{RT}(R,X) \le \kappa'_{RT}(R,X). \tag{5.3.45}$$

Thus from this and the second inequality in (5.3.32), it follows with (5.3.18) and (5.3.41) that that

$$\kappa_{CDT}(R,X) \le \kappa'_{CDT}(R,X). \tag{5.3.46}$$

Now we summarize these results as the following theorem.

Theorem 5.3.2 *With the same assumptions as in Theorem* 5.2.2, *there is a unique Cholesky factorization*

$$(U+\Delta U)^T(U+\Delta U) = (R+\Delta R)^T(R+\Delta R) - (X+\Delta X)^T(X+\Delta X),$$

where for general ΔR,

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} &\le \kappa'_{RG}(R,X)\epsilon_R + \kappa'_X(R,X)\epsilon_X + O(\epsilon^2)\\ &\le \kappa'_{CDG}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2),\end{aligned}$$

and for upper triangular ΔR,

$$\begin{aligned}\frac{\|\Delta U\|_F}{\|U\|_2} &\le \kappa'_{RT}(R,X)\epsilon_R + \kappa'_X(R,X)\epsilon_X + O(\epsilon^2)\\ &\le \kappa'_{CDT}(R,X)(\epsilon_R+\epsilon_X) + O(\epsilon^2).\end{aligned}$$

There are the following relationships among the various measures of sensitivity of the problem (see (5.3.28), (5.3.30), (5.3.32), (5.3.33), (5.3.43), (5.3.44), (5.3.45) and (5.3.46)) :

$$\kappa_{RG}(R,X) \le \kappa'_{RG}(R,X) \le \kappa'_{RG}(R,X,I_n) = \phi_R, \quad \kappa_{RT}(R,X) \le \kappa'_{RT}(R,X),$$

$$\kappa_X(R,X) \le \kappa'_X(R,X) \le \kappa'_X(R,X,I_n) = \phi_X, \quad \kappa'_X(R,X) \le \kappa'_{RG}(R,X),$$

$$\kappa'_{RT}(R,X) \le (1+\sqrt{n})\,\kappa'_{RG}(R,X), \quad \kappa'_{CDT}(R,X) \le (1+\sqrt{n})\,\kappa'_{CDG}(R,X),$$

$$\kappa_{CDG}(R,X) \le \kappa'_{CDG}(R,X), \quad \kappa_{CDT}(R,X) \le \kappa'_{CDT}(R,X). \qquad \square$$

Our numerical experiments suggest $\kappa'_{CDG}(R,X)$ is usually a good approximation to $\kappa_{CDG}(R,X)$. But the following example shows $\kappa'_{CDG}(R,X)$ can sometimes be arbitrarily larger than $\kappa_{CDG}(R,X)$.

$$R = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \delta^3 & 0 \\ 0 & 0 & 0 & \delta^2 \end{bmatrix}, \quad X = \begin{bmatrix} \sqrt{3} & 2/\sqrt{3} & 0 & 0 \\ 0 & \sqrt{2/3-\delta^2} & 0 & 0 \end{bmatrix}, \quad U = \mathrm{diag}(1,\delta,\delta^3,\delta^2),$$

where δ is a small positive number. It is not difficult to show

$$\kappa'_{CDG}(R,X) = O(\frac{1}{\delta^2}), \qquad \kappa_{CDG}(R,X) = O(\frac{1}{\delta}).$$

But $\kappa'_{CDG}(R,X)$ has an advantage over $\kappa_{CDG}(R,X)$ — it can be quite easy to estimate — all we need to do is to choose a suitable D in $\kappa'_{CDG}(R,X,D)$. We consider how to do this in the next section. In contrast $\kappa_{CDG}(R,X)$ is, as far as we can see, unreasonably expensive to compute or estimate.

Numerical experiments also suggest $\kappa'_{CDT}(R,X)$ is usually a good approximation to $\kappa_{CDT}(R,X)$. But sometimes $\kappa'_{CDT}(R,X)$ can be arbitrarily larger than $\kappa_{CDT}(R,X)$. This can also be seen from the example above. In fact, it is not difficult to obtain

$$\kappa'_{CDT}(R,X) = O(\frac{1}{\delta^2}), \qquad \kappa_{CDT}(R,X) = O(\frac{1}{\delta}).$$

Like $\kappa_{CDG}(R,X)$, $\kappa_{CDT}(R,X)$ is difficult to compute or estimate. But $\kappa'_{CDT}(R,X)$ is easy to estimate, which is discussed in the next section.

5.4 Numerical experiments

In Section 5.3 we presented new first-order perturbation bounds for the the downdated Cholesky factor U using first the matrix-vector equation approach, and then the matrix equation approach. We defined $\kappa_{CDG}(R,X)$ for general ΔR, and $\kappa_{CDT}(R,X)$ for upper triangular ΔR, as the overall condition numbers of the problem. Also we

gave two corresponding practical but weaker condition estimators $\kappa'_{CDG}(R,X)$ and $\kappa'_{CDT}(R,X)$ for the two ΔR cases.

We would like to choose D such that $\kappa'_{CDG}(R,X,D)$ and $\kappa'_{CDT}(R,X,D)$ are good approximations to $\kappa'_{CDG}(R,X)$ and $\kappa'_{CDT}(R,X)$, respectively. We see from (5.3.25), (5.3.26) and (5.3.39) that we want to find D such that $\sqrt{1+\zeta_D^2}\,\kappa_2(D^{-1}U)$ approximates its infimum. That is the same problem we faced in Section 3.5. We adopt the best method of choosing D proposed there. Specifically take $\zeta_1 = \sqrt{\sum_{j=1}^n u_{1j}^2}$, $\zeta_i = \sqrt{\sum_{j=i}^n u_{ij}^2}$ if $\sqrt{\sum_{j=i}^n u_{ij}^2} \le \zeta_{i-1}$ otherwise $\zeta_i = \zeta_{i-1}$, for $i = 2,\ldots,n$. Then we use a standard condition estimator to estimate $\kappa_2(D^{-1}U)$ in $O(n^2)$ operations.

Notice from (5.2.4) we have $\sigma_n(\Gamma) = \sqrt{1-\|XR^{-1}\|_2^2}$. Usually k, the number of rows of X, is much smaller than n, so $\sigma_n(\Gamma)$ can be computed in $O(n^2)$. If k is not much smaller than n, then we use a standard norm estimator to estimate $\|XR^{-1}\|_2$ in $O(n^2)$. Similarly $\|U\|_2$ and $\|R\|_2$ can be estimated in $O(n^2)$. So finally $\kappa'_{CDG}(R,X,D)$ can be estimated in $O(n^2)$. Estimating $\kappa'_{CDT}(R,X,D)$ is not as easy as estimating $\kappa'_{CDG}(R,X,D)$. The part $\|\mathrm{diag}(RU^{-1})\|_2$ in $\kappa'_{RT}(R,X,D)$ can easily be computed in $O(n)$, since $\mathrm{diag}(RU^{-1}) = \mathrm{diag}(r_{11}/u_{11},\ldots,r_{nn}/u_{nn})$. The part $\|\mathrm{sut}(RU^{-1})\|_2$ in $\kappa'_{RT}(R,X,D)$ can roughly be estimated in $O(n^2)$, based on

$$\frac{1}{\sqrt{n-1}}\|\mathrm{sut}(RU^{-1})\|_F \le \|\mathrm{sut}(RU^{-1})\|_2 \le \|\mathrm{sut}(RU^{-1})\|_F,$$
$$\|\mathrm{sut}(RU^{-1})\|_F = \sqrt{\|RU^{-1}\|_F^2 - \|\mathrm{diag}(RU^{-1})\|_F^2},$$

and the fact that $\|RU^{-1}\|_F$ can be estimated by a standard norm estimator in $O(n^2)$. The value of $\|XU^{-1}\|_2$ in $\kappa'_X(R,X,D)$ can be calculated (if $k \ll n$) or estimated by a standard estimator in $O(n^2)$. All the remaining values $\|R\|_2$, $\|X\|_2$ and $\|U\|_2$ can also be estimated by a standard norm estimator in $O(n^2)$. Hence $\kappa'_{RT}(R,X,D)$, $\kappa'_X(R,X,D)$, and thus $\kappa'_{CDT}(R,X,D)$ can be estimated in $O(n^2)$. For standard condition estimators and norm estimators, see Chapter 14 of Higham [30, 1996].

The relationships among the various overall measures of sensitivity of the Cholesky

downdating problem presented in Section 5.2 and Section 5.3 are as follows.

$$\beta \geq \phi \geq \kappa'_{CDG}(R,X) \geq \kappa_{CDG}(R,X) \geq \kappa_{CDT}(R,X),$$
$$(1+\sqrt{n})\kappa'_{CDG}(R,X) \geq \kappa'_{CDT}(R,X) \geq \kappa_{CDT}(R,X).$$

Now we give one numerical example to illustrate these. The example, quoted from Sun [50, 1995], is as follows.

$$R = \mathrm{diag}(1,s,s^2,s^3,s^4)\begin{bmatrix} 1 & -c & -c & -c & -c \\ 0 & 1 & -c & -c & -c \\ 0 & 0 & 1 & -c & -c \\ 0 & 0 & 0 & 1 & -c \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad X^T = \tau \begin{bmatrix} 0.240 \\ -0.899 \\ 0.899 \\ 1.560 \\ 2.390 \end{bmatrix},$$

where $c = 0.95$, $s = \sqrt{1-c^2}$. The results obtained using MATLAB are shown in Table 5.4.1 for various values of τ:

$$\tau_1 = 1.004015006005433e-2, \quad \tau_2 = 1.003021021209640e-2,$$
$$\tau_3 = 9.036225416303058e-3,$$

and $\tau_4 = \tau_3 \cdot e-01$, $\tau_5 = \tau_3 \cdot e-03$, $\tau_6 = \tau_3 \cdot e-5$.

Table 5.4.1: Results for the example in Sun's paper

τ	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
$\|XR^{-1}\|_2$	0.99999	0.999	0.9	0.09	0.0009	0.000009
β	2.25e+10	2.25e+07	2.60e+04	2.72e+03	2.69e+03	2.69e+03
ϕ	1.01e+08	1.01e+06	1.14e+04	2.71e+03	2.69e+03	2.69e+03
$\kappa'_{CDG}(R,X,D)$	3.60e+03	3.61e+02	3.79e+01	1.79e+01	1.78e+01	1.78e+01
$\kappa_{CDG}(R,X)$	1.66e+03	1.66e+02	1.71e+01	8.42e+00	8.41e+00	8.41e+00
$\kappa'_{CDT}(R,X,D)$	2.12e+03	2.12e+02	1.79e+01	1.07e+00	1.00e+00	1.00e+00
$\kappa_{CDT}(R,X)$	2.43e+02	2.43e+01	2.44e+00	1.01e+00	1.00e+00	1.00e+00

Note in Table 5.4.1 how β and ϕ can be far worse than the condition numbers $\kappa_{CDG}(R,X)$ and $\kappa_{CDT}(R,X)$, although ϕ is not as bad as β. Also we observe

that $\kappa'_{CDG}(R,X,D)$ and $\kappa'_{CDT}(R,X,D)$ are very good approximations to $\kappa_{CDG}(R,X)$ and $\kappa_{CDT}(R,X)$, respectively. When X become small, all of the condition numbers and condition estimators decrease. The asymptotic behavior of $\kappa'_{CDG}(R,X,D)$, $\kappa'_{CDT}(R,X,D)$, $\kappa_{CDG}(R,X)$ and $\kappa_{CDT}(R,X)$ coincides with our theoretical results: when $X \to 0$, $\kappa'_{CDG}(R,X)$ and $\kappa_{CDG}(R,X)$ will be bounded in terms of n since here R is actually $K_5(\arccos(0.95))$, a Kahan matrix, which corresponds to the Cholesky factor of a correctly pivoted A, and $\kappa'_{CDT}(R,X), \kappa_{CDT}(R,X) \to 1$.

5.5 Summary and future work

The first-order perturbation analyses presented here show just what the sensitivity of the Cholesky downdating problem is, and in so doing provide the condition numbers, as well as efficient ways of approximating these. The key measures of the sensitivity of the problem we derived are:

- For general ΔR:
  - overall condition number: $\kappa_{CDG}(R,X)$, see (5.3.9),
  - overall condition estimator: $\kappa'_{CDG}(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_{CDG}(R,X,D)$, see (5.3.27),
- For triangular ΔR:
  - overall condition number: $\kappa_{CDT}(R,X)$, see (5.3.18),
  - overall condition estimator: $\kappa'_{CDT}(R,X) \equiv \inf_{D\in\mathbf{D}_n} \kappa'_{CDT}(R,X,D)$, see (5.3.40).

These quantities and the condition estimators ϕ (see (5.2.17)) and β (see (5.2.19)) obey

$$\kappa_{CDT}(R,X) \le \kappa_{CDG}(R,X) \le \kappa'_{CDG}(R,X) \le \phi \le \beta,$$
$$\kappa_{CDT}(R,X) \le \kappa'_{CDT}(R,X) \le (1+\sqrt{n})\kappa'_{CDG}(R,X).$$

For the asymptotic case as $X \to 0$, $\kappa_{CDG}(R, X)$ and $\kappa'_{CDG}(R, X)$ will be bounded in terms of n, and $\kappa_{CDT}(R, X)$, $\kappa'_{CDT}(R, D) \to 1$, while β and ϕ have no such properties.

Recently Stewart [42, 1995] presented a backward rounding error analysis for the block downdating algorithm presented by Eldén and Park. It would be straightforward to combine our results here with Stewart's result to give a forward error estimate for the computed U. But we choose not to do this here in order to keep the material as simple as possible.

In the future we would like to

- Give better approximations to $\kappa_{CDG}(R, X)$ and $\kappa_{CDT}(R, X)$ than $\kappa'_{CDG}(R, X)$ and $\kappa'_{CDT}(R, X)$.
- Extend our analyses here to other cases, such as that when ΔR and ΔX come from a componentwise backward rounding error analysis.

Chapter 6

Conclusions and future research

A new approach, the so called 'matrix-vector equation approach' has been developed here for the perturbation analysis of matrix factorizations. The basic idea of this approach is to write the perturbation matrix equation as a matrix-vector equation by using the special structure and properties of the factors. Using this approach we obtained tight first-order perturbation results and condition numbers for the Cholesky, QR and LU factorizations, and for the Cholesky downdating problem. Our perturbation bounds give significant improvements on the previous results, and could not be sharper.

Also we used the so called 'matrix equation approach' originated by G. W. Stewart to obtain perturbation bounds that are usually weaker but easier to interpret, leading to condition estimators which are easily estimated by the standard condition estimators (for matrix inversion) or norm estimators. Our experiments suggested that for the Cholesky, QR and LU factorizations with norm-bounded changes in the original matrices the condition estimators are very good approximations of the corresponding condition numbers. Also our numerical experiments suggested for the Cholesky factorization with component-bounded changes in the original matrix and the Cholesky downdating problem with norm-bounded changes in the original matrices, the condition estimators are usually good approximations of the corresponding

condition numbers, even though some counter-examples were found.

The matrix-vector equation approach is a powerful general tool, and appears to be applicable to the perturbation analysis of any matrix factorization. The matrix equation approach is also fairly general, but for each factorization a particular treatment is needed. The combination of these two approaches gives a deep understanding of these problems. Although first-order perturbation bounds are satisfactory for all but the most delicate work, we also gave some rigorous perturbation bounds for the Cholesky factorization.

In computing these factorizations, standard pivoting is often used to improve the stability of the algorithms. We showed that the condition of these factorizations is significantly improved by the standard pivoting strategies (except the L factor in the LU factorization), and provided firmly based theoretical explanations as to why this is so. This extremely important information is very useful for designing more reliable matrix algorithms.

In the future we hope to continue this research in several directions:

- To analyze the Cholesky, LU and QR factorizations, and Cholesky downdating of general matrices, where perturbations have special structure, for example, by assuming the perturbation has the form of the equivalent backward rounding error from a numerically stable computation of the factorization (some results for the Cholesky factorization have been given in this thesis). Such structure leads to improved sensitivity results.

- To analyze other factorizations of general matrices for both general perturbations and structured perturbations.

- To extend our approach to the factorizations of special matrices. In many applications matrices and the resulting factorizations used to solve the problems in a numerically stable way have some special structure. Applying the existing

general perturbation results to these special problems will result in overestimation of the true sensitivity. Our new approach to such perturbation analyses can make full use of the structure, so should lead to results which closely reflect the true sensitivity of the problems.

References

[1] S. T. Alexander, C.-T. Pan and R. J. Plemmons, *Analysis of a recursive least squares hyperbolic rotation algorithm for signal processing*. Linear Algebra Appl., 98 (1988), pp. 3–40.

[2] A. Barrland. *Perturbation bounds for the LDL^H and the LU factorizations*. BIT, 31 (1991), pp. 358–363.

[3] R. Bhatia. *Matrix factorizations and their perturbations*. Linear Algebra and Appl., 197,198 (1994), pp.245–276.

[4] Å. Björck and G. H. Golub. *Numerical methods for computing angles between linear subspaces*. Math. Comp., 27 (1973), pp. 579–594.

[5] A. Björck, H. Park and L. Eldén. *Accurate downdating of least squares solutions*. SIAM J. Matrix Anal. Appl., 15 (1994), pp. 549–568.

[6] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog. *A note on downdating the Cholesky factorization*. SIAM J. Sci. Stat. Comput., 8 (1987), pp. 210–221.

[7] P. A. Businger and G. H. Golub, *Linear least squares solutions by Householder Transformations*. Numer. Math., 7 (1965), pp. 269–276.

[8] X.-W. Chang. *Perturbation analyses for the Cholesky factorization with backward rounding errors.* Technical Report SOCS-96.3, School of Computer Science, McGill University, 1996.

[9] X.-W. Chang and C. C. Paige. *A perturbation analysis for R in the QR factorization.* Technical Report SOCS-95.7, School of Computer Science, McGill University, 1995.

[10] X.-W. Chang and C. C. Paige. *Perturbation analyses for the Cholesky downdating problem.* Submitted to SIAM J. Matrix Anal. Appl., April 1996. 15 pp.

[11] X.-W. Chang and C. C. Paige. *Perturbation analyses for the QR factorization with bounds on changes in the elements of A.* In preparation.

[12] X.-W. Chang and C. C. Paige. *Perturbation analyses for the LU factorization.* In preparation.

[13] X.-W. Chang, C. C. Paige and G. W. Stewart. *New perturbation analyses for the Cholesky factorization.* IMA J. Numer. Anal., 16 (1996), pp. 457–484.

[14] X.-W. Chang, C. C. Paige and G. W. Stewart. Perturbation analyses for the QR factorization. SIAM J. Matrix Anal. Appl., to appear. 18 pp.

[15] A. K. Cline, A. R. Conn and C. F. Van Loan. *Generalizing the LINPACK condition estimator.* Numerical Analysis, ed. J. P. Hennart. Lecture Notes in Mathematics, no. 909, Springer-Verlag, NY. 1982.

[16] A. K. Cline, C. B. Moler. G. W. Stewart and J. H. Wilkinson. *An estimate for the condition number of a matrix.* SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

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

[18] J. D. Dixon. Estimating extremal eigenvalues and condition number of matrices. *SIAM J. Numer. Anal.*, 20(4):812–814, 1983.

[19] J. J. Dongarra, J. R. Bunch, C. B. Moler and G. W. Stewart. *LINPACK User's Guide*, SIAM, Philadelphia, 1979.

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

[21] L. Eldén and H. Park. *Block downdating of least squares solutions.* SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1018–1034.

[22] L. Eldén and H. Park. *Perturbation analysis for block downdating of a Cholesky decomposition.* Numerische Mathematik, 68 (1994), pp. 457–467.

[23] L. Eldén and H. Park. *Perturbation and error analyses for block downdating of a Cholesky decomposition.* BIT, 36 (1996), pp. 247–263.

[24] P. E. Gill, G. H. Golub, W. Murray and M. A. Saunders. *Methods for modifying matrix factorizations.* Math. Comp., 28 (1974), pp. 505–535.

[25] G. H. Golub and G. P. Styan, *Numerical computations for univariate linear models.* J. Stat. Comp. Simul., 2 (1973), pp. 253–272.

[26] G. H. Golub and C. F. Van Loan. *Matrix computations.* Third Edition. The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[27] N. J. Higham. *A survey of condition number estimation for triangular matrices.* SIAM Rev., 29 (1987), pp. 575–596.

[28] N. J. Higham. *Analysis of the Cholesky decomposition of a semi-definite matrix.* Reliable Numerical Computation, ed. M. G. Cox & S. J. Hammarling. Oxford University Press, 1990, pp. 161–186.

[29] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra.* In Mathematica of Computation 1943-1993: A Half Century of Computational Mathematics, Walter Gautschi, editor, volume 48 of Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 1994, pp. 49–77.

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

[31] Y. P. Hong and C.-T. Pan. *Rank-revealing QR factorizations and the singular value decomposition.* Mathematics of Computation, 58 (1992), pp. 213–232.

[32] W. Kahan. *Numerical linear algebra.* Can. Math. Bull., 9 (1966), pp. 757–801.

[33] J. Meinguet. *Refined Error Analyses of Cholesky Factorization.* SIAM J. Numer. Anal., 20 (1983), pp. 1243–1250.

[34] C. C. Paige. *Covariance matrix representation in linear filtering.* In Linear Algebra and Its Role in Systems Theory, B.N. Datta (ed.), AMS, Providence, RI, 1985, pp. 309–321.

[35] C.-T. Pan. *A perturbation analysis of the problem of downdating a Cholesky factorization.* Linear Algebra Appl., 183 (1993), pp. 103–116.

[36] C:-T. Pan and R. Plemmons. *Least squares modification with inverse factorizations: parallel implications.* J. Comp. Appl. Math., 27 (1989), pp. 109–127.

[37] M. A. Saunders. *Large-scale linear programming using the Cholesky factorization.* Technical Report CS252, Computer Science Department, Stanford University, 1972.

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

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

[40] G. W. Stewart. *The effects of rounding error on an algorithm for downdating a Cholesky factorization.* J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[41] G. W. Stewart. On the perturbation of LU, Cholesky, and QR factorizations. *SIAM J. Matrix Anal. Appl.*, 14 (1993), pp. 1141–1145.

[42] G. W. Stewart. *On sequential updates and downdates.* IEEE Transactions on Signal Processing, 43 (1995), pp. 2642–2648.

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

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

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

[46] J. G. Sun. *Perturbation bounds for the Cholesky and QR factorization.* BIT, 31 (1991), 341–352.

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

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

[49] J. G. Sun. *On perturbation bounds for the QR factorization.* Linear Algebra and Appl., 215 (1995), pp. 95–112.

[50] J.-G. Sun. *Perturbation analysis of the Cholesky downdating and QR updating problems*. SIAM J. Matrix Anal. Appl., 16 (1995), pp. 760–775.

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

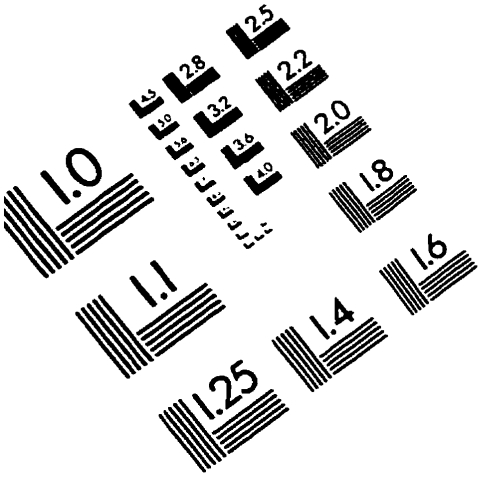
[52] J. H. Wilkinson. *Rounding errors in algebraic processes*. Prentice-Hall, Englewood Cliffs, NJ, 1963.

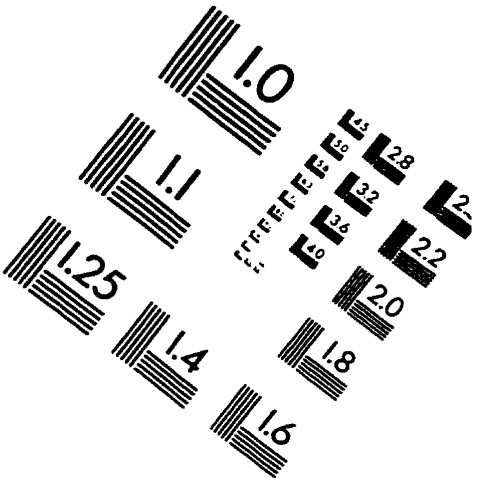
[53] J. H. Wilkinson. *The algebraic eigenvalue problem*. Oxford University Press, 1965.

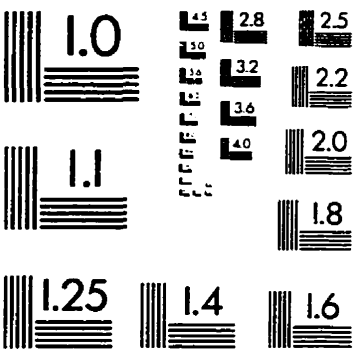
[54] J. H. Wilkinson. *A prior error analysis of algebraic processes*. In Proc. International Congress of Mathematicians, Moscow 1966, I. G. Petrovsky, ed., Mir Publishers, Moscow, 1968, pp. 629–640.

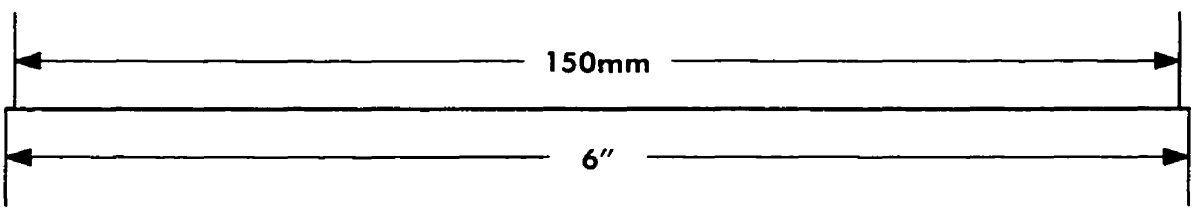
[55] H. Zha. *A componentwise perturbation analysis of the QR decomposition*. SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1124–1131.

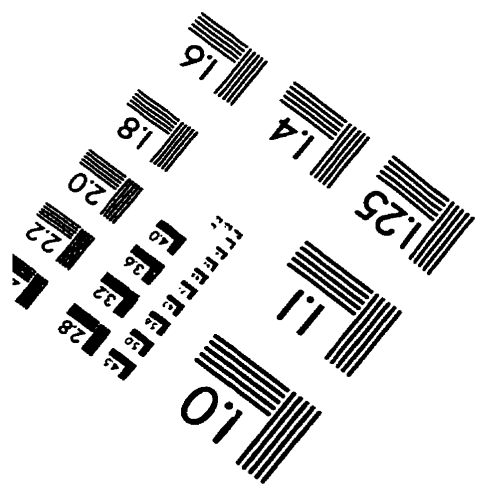
IMAGE EVALUATION
TEST TARGET (QA-3)

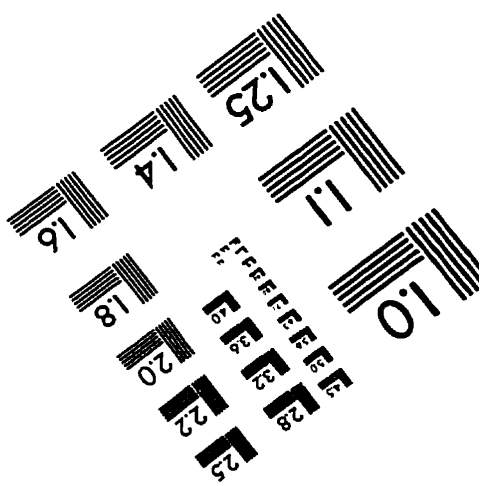












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989