PERTURBATION ANALYSES FOR THE QR FACTORIZATION

XIAO-WEN CHANG*, CHRISTOPHER. C. PAIGE* AND G. W. STEWART†

Abstract. This paper gives perturbation analyses for Q_1 and R in the QR factorization $A = Q_1 R$, $Q_1^T Q_1 = I$, for a given real $m \times n$ matrix A of rank n. The analyses more accurately reflect the sensitivity of the problem than previous normwise results. The condition numbers here are altered by any column pivoting used in $AP = Q_1 R$, and the condition numbers for R are bounded for a fixed n when the standard column pivoting strategy is used. This strategy tends to improve the condition of Q_1, so the computed Q_1 and R will probably both have greatest accuracy when we use the standard column pivoting strategy.

First order normwise perturbation analyses are given for both Q_1 and R. It is seen that the analysis for R may be approached in two ways — a detailed "matrix–vector equation" analysis which provides tight bounds and resulting true condition numbers, which unfortunately are costly to compute and not very intuitive, and a perhaps simpler "matrix equation" analysis which provides results that are usually weaker but easier to interpret, and which allow efficient computation of a satisfactory estimate for the true condition number.

Key Words. QR factorization, perturbation analysis, condition estimation, matrix equations, pivoting

AMS Subject Classifications: 15A23, 65F35

1. Introduction.

The QR factorization is an important tool in matrix computations (see for example [4]): given an $m \times n$ real matrix A with full column rank, there exists a unique $m \times n$ real matrix Q_1 with orthonormal columns, and a unique nonsingular upper triangular $n \times n$ real matrix R with positive diagonal entries such that

$$A = Q_1 R.$$

The matrix Q_1 is referred to as the orthogonal factor, and R the triangular factor.

Suppose $A(t) \equiv A + tG$ has the unique QR factorization $A(t) = Q_1(t)R(t)$. If we differentiate $R(t)^T R(t) = A(t)^T A(t)$ with respect to t and set $t = 0$ we have

$$\begin{aligned} R^T \dot{R}(0) + \dot{R}^T(0) R &= A^T \dot{A}(0) + \dot{A}(0)^T A \\ &= R^T Q_1^T G + G^T Q_1 R, \end{aligned} \tag{1.1}$$

which with given A and G is a linear equation for the upper triangular matrix $\dot{R}(0)$. But $\dot{R}(0)$ determines the sensitivity of $R(t)$ at $t = 0$, and so the core of any perturbation analysis for the QR factorization effectively involves the use of (1.1) to determine bound $\dot{R}(0)$. There are essentially two main ways of approaching this problem. We now discuss these, together with some very recent history.

Xiao-Wen Chang [1, 2, 3] was the first to realize that most of the published results on the sensitivity of factorizations, such as LU, Cholesky, and QR, were extremely

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

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A., (stewart@cs.umd.edu). G. W. Stewart's research was supported in part by the US National Science Foundation under grant CCR 95503126.

weak for many matrices. He originated a general approach to obtaining provably tight results and sharp condition numbers for these problems. We will call this the "matrix-vector equation" approach. For the QR factorization this involves expressing (1.1) as a matrix-vector equation of the form

$$W_R \text{uvec}(\dot{R}(0)) = Z_R \text{vec}(Q_1^T G),$$

where W_R and Z_R are matrices involving the elements of R, and $\text{vec}(\cdot)$ transforms its argument into a vector of its elements. The notation $\text{uvec}(\cdot)$ denotes a variant of $\text{vec}(\cdot)$ defined in Section 5.2. Chang also realized that the condition of many such problems was significantly improved by pivoting, and provided the first firmly based theoretical explanations as to why this was so.

G. W. Stewart [11, 3] was stimulated by Chang's work 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) in their matrix-matrix form. Stewart [11] (also see [3]) used an elegant construct, partly illustrated by the "up" and "low" notation in Section 2, which makes the matrix equation approach a far more usable and intuitive tool. He combined this with deep insights on scaling to produce new matrix equation analyses which are appealingly clear, and provide excellent insight into the sensitivities of the problems. These new matrix equation analyses do 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.

It was the third author, Chris Paige, who insisted on writing this brief history, and delineating these exceptional contributions of Chang and Stewart. It requires a combination of the two analyses to provide a full understanding of the cases we have examined. The interplay of the two approaches was first revealed in [3], and we hope to convey it even more clearly here for the QR factorization.

The perturbation analysis for the QR factorization has been considered by several authors. The first norm-based result for R was presented by Stewart [9]. That was further modified and improved by Sun [13]. Using different approaches Sun [13] and Stewart [10] gave *first order* normwise perturbation analyses for R. A first order componentwise perturbation analysis for R has been given by Zha [17], and a *strict* componentwise analysis for R has been given by Sun [14]. These papers also gave analyses for Q_1. More recently Sun [15] gave strict perturbation bounds for Q_1 alone.

The purpose of this paper is to establish new first order perturbation bounds which are generally sharper than the equivalent results for the R factor in [10, 13], and more straightforward than the sharp result in [15] for the Q_1 factor.

In Section 2 we define some notation and give a result we will use throughout the paper. In Section 3 we will survey important key results on the sensitivity of R and Q_1. In Section 4 we give a refined perturbation analysis for Q_1, showing in a simple way why the standard column pivoting strategy for A can be beneficial for certain aspects of the sensitivity of Q_1. In Section 5 we analyze the perturbation in R, first by the straightforward matrix equation approach, then by the more detailed and tighter matrix–vector equation approach. We give numerical results and suggest practical condition estimators in Section 6, and summarize and comment on our findings in Section 7.

2. Notation and basics. To simplify the presentation, for any $n \times n$ matrix X, we define the upper and lower triangular matrices

$$\text{up}(X) \equiv \begin{pmatrix} \frac{1}{2}x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}x_{nn} \end{pmatrix}, \qquad \text{low}(X) \equiv \text{up}(X^T)^T, \tag{2.1}$$

so that $X = \text{low}(X) + \text{up}(X)$. For general X

$$\max\{\|\text{low}(X)\|_F, \|\text{up}(X)\|_F\} \leq \|X\|_F. \tag{2.2}$$

For symmetric X

$$2\|\text{low}(X)\|_F^2 = 2\|\text{up}(X)\|_F^2 = \|X\|_F^2 - \frac{1}{2}(x_{11}^2 + x_{22}^2 + \cdots + x_{nn}^2) \leq \|X\|_F^2. \tag{2.3}$$

To illustrate a basic use of "up", we show that for any given $n \times n$ nonsingular upper triangular R and any given $n \times n$ symmetric M, the equation of the form (cf. (1.1))

$$R^T U + U^T R = M \tag{2.4}$$

always has a unique upper triangular solution U. Since UR^{-1} is upper triangular in

$$UR^{-1} + (UR^{-1})^T = R^{-T} M R^{-1} \quad \text{symmetric},$$

we see immediately that

$$UR^{-1} = \text{up}(R^{-T} M R^{-1}),$$

so UR^{-1} and therefore U is uniquely defined. We will describe other uses later.

Our perturbation bounds will be tighter if we bound separately the perturbations along the columns 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 $\mathcal{R}(\mathcal{A})$, and P_2 be the orthogonal projector onto $\mathcal{R}(\mathcal{A})^\perp$. For real $m \times n$ ΔA define

$$\epsilon \equiv \|\Delta A\|_F / \|A\|_2, \quad \epsilon_1 \equiv \|P_1 \Delta A\|_F / \|A\|_2, \quad \epsilon_2 \equiv \|P_2 \Delta A\|_F / \|A\|_2, \tag{2.5}$$

so $\epsilon^2 = \epsilon_1^2 + \epsilon_2^2$. When $\epsilon > 0$ in (2.5) we also define

$$G \equiv \Delta A / \epsilon, \tag{2.6}$$

so for the QR factorization $A = Q_1 R$

$$\|G\|_F = \|A\|_2 = \|R\|_2. \tag{2.7}$$

We will use the following standard result.

LEMMA 2.1. *For real $m \times n$ A with rank n, real $A + \Delta A$ has rank n if*

$$\kappa_2(A) \frac{\|P_1 \Delta A\|_2}{\|A\|_2} < 1, \tag{2.8}$$

where $\kappa_2(A) \equiv \|A^\dagger\|_2 \|A\|_2$ and P_1 is the orthogonal projector onto $\mathcal{R}(\mathcal{A})$.

Proof. Let A have the QR factorization

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix} = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}; \quad \text{then} \quad Q^T(A+\Delta A) = \begin{pmatrix} R + Q_1^T \Delta A \\ Q_2^T \Delta A \end{pmatrix},$$

which necessarily has full column rank if $\|Q_1^T \Delta A\|_2 < \sigma_{\min}(A)$, the smallest singular value of A. But this inequality is just (2.8). □

COROLLARY 2.2. *If nonzero ΔA satisfies* (2.8)*, then for ϵ and G defined in* (2.5) *and* (2.6)*, $A + tG$ has full column rank and therefore a unique QR factorization for all $|t| \le \epsilon$.* □

3. Previous norm-based results. In this section we summarize the strongest norm-based results by previous authors. We first give a derivation of what is essentially Sun's [13] and Stewart's [10] first order normwise perturbation result for R, since their techniques and results will be useful later.

THEOREM 3.1. [13]. *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $A = Q_1 R$, and let ΔA be a real $m \times n$ matrix. Define $\epsilon \equiv \|\Delta A\|_F / \|A\|_2$ and $\epsilon_1 \equiv \|Q_1^T \Delta A\|_F/\|A\|_2$, see* (2.5)*. If* (2.8) *holds then $A + \Delta A$ has a unique QR factorization*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R), \tag{3.1}$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1 + O(\epsilon^2). \tag{3.2}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). From Corollary 2.2 $A + tG$ has the unique QR factorization

$$A(t) \equiv A + tG = Q_1(t)R(t), \tag{3.3}$$

for all $|t| \le \epsilon$, where

$$Q_1^T(t)Q_1(t) = I. \tag{3.4}$$

Notice that $R(0) = R$ and $R(\epsilon) = R + \Delta R$.

It is easy to verify that $Q_1(t)$ and $R(t)$ are twice continuously differentiable for $|t| \le \epsilon$ from the algorithm for the QR factorization. Thus as in (1.1) we have

$$R^T\dot{R}(0) + \dot{R}^T(0)R = R^T Q_1^T G + G^T Q_1 R, \tag{3.5}$$

which (see (2.4)) is a linear equation *uniquely* defining the elements of $\dot{R}(0)$ in terms of the elements of $Q_1^T G$. From upper triangular $\dot{R}(0)R^{-1}$ in

$$\dot{R}(0)R^{-1} + (\dot{R}(0)R^{-1})^T = Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T, \tag{3.6}$$

we see with (2.1) that

$$\dot{R}(0) = \mathrm{up}[Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T]R, \tag{3.7}$$

so with (2.3)

$$\begin{aligned}\|\dot{R}(0)\|_F &\le \frac{1}{\sqrt{2}}\|Q_1^T GR^{-1} + (Q_1^T GR^{-1})^T\|_F\|R\|_2\\ &\le \sqrt{2}\|Q_1^T GR^{-1}\|_F\|R\|_2\\ &\le \sqrt{2}\kappa_2(R)\|Q_1^T G\|_F,\end{aligned}$$

and since from (2.5) to (2.7) $\|Q_1^T G\|_F = \|A\|_2\epsilon_1/\epsilon$, and $\|R^{-1}\|_2 = \|A^\dagger\|_2$,

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1/\epsilon. \tag{3.8}$$

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

$$R + \Delta R = R(\epsilon) = R(0) + \epsilon\dot{R}(0) + O(\epsilon^2), \tag{3.9}$$

so that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\|\dot{R}(0)\|_F}{\|R\|_2}\epsilon_1 + O(\epsilon^2),$$

which, combined with (3.8), gives (3.2). □

This proof shows that the key point in deriving a first order perturbation bound for R is the use of (3.5) to give a good bound on the sensitivity $\|\dot{R}(0)\|_F/\|R\|_2$. Since we obtained the bounds directly from (3.7), this was a "matrix equation" analysis.

We now show how a recent perturbation result for Q_1 given by Sun [15] can be obtained in the present setting, since the analysis can easily be extended to obtain a more refined result in a simple way. Note the hypotheses of the following theorem are those of Theorem 3.1, so we can use results from the latter theorem.

THEOREM 3.2. [15]. *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $A = Q_1 R$, and let ΔA be a real $m \times n$ matrix. If $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ and* (2.8) *holds, then $A + \Delta A$ has a unique QR factorization*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R),$$

where

$$\|\Delta Q_1\|_F \le \sqrt{2}\kappa_2(A)\epsilon + O(\epsilon^2). \tag{3.10}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). From Corollary 2.2 $A + tG$ has the unique QR factorization $A(t) \equiv A + tG = Q_1(t)R(t)$ with $Q_1(t)^T Q_1(t) = I$, for all $|t| \le \epsilon$. Differentiating these at $t = 0$ gives

$$G = Q_1\dot{R}(0) + \dot{Q}_1(0)R, \qquad Q_1^T\dot{Q}_1(0) \quad \text{skew symmetric.}$$

It follows that

$$\dot{Q}_1(0) = GR^{-1} - Q_1\dot{R}(0)R^{-1},$$

so with any Q_2 such that $Q \equiv (Q_1, Q_2)$ is square and orthogonal,

$$Q_2^T\dot{Q}_1(0) = Q_2^T GR^{-1}, \tag{3.11}$$

$$Q_1^T\dot{Q}_1(0) = Q_1^T G R^{-1} - \dot{R}(0)R^{-1}.$$

Now using (2.1), we have with (3.7) in Theorem 3.1 that

$$\begin{aligned} Q_1^T\dot{Q}_1(0) &= \mathrm{low}(Q_1^T G R^{-1}) + \mathrm{up}(Q_1^T G R^{-1}) \\ &\quad -\mathrm{up}[Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T] \\ &= \mathrm{low}(Q_1^T G R^{-1}) - [\mathrm{low}(Q_1^T G R^{-1})]^T. \end{aligned} \tag{3.12}$$

We see from this, (2.2), (3.11), and $\|G\|_F = \|A\|_2$ from (2.7), that

$$\begin{aligned} \|\dot{Q}_1(0)\|_F^2 &= \|Q_1^T\dot{Q}_1(0)\|_F^2 + \|Q_2^T\dot{Q}_1(0)\|_F^2 \\ &\le 2\|Q_1^T G R^{-1}\|_F^2 + \|Q_2^T G R^{-1}\|_F^2 \le 2\|G R^{-1}\|_F^2, \end{aligned}$$

$$\|\dot{Q}_1(0)\|_F \le \sqrt{2}\|R^{-1}\|_2\|G\|_F = \sqrt{2}\kappa_2(A), \tag{3.13}$$

and from the Taylor expansion for $Q_1(t)$ about $t = 0$ at $t = \epsilon$,

$$Q_1 + \Delta Q_1 = Q_1(\epsilon) = Q_1(0) + \epsilon\dot{Q}_1(0) + O(\epsilon^2), \tag{3.14}$$

so that $\|\Delta Q_1\|_F \le \epsilon\|\dot{Q}_1(0)\|_F + O(\epsilon^2)$, which with (3.13) gives (3.10). $\square$

4. Refined analysis for Q_1. Note since both Q_1 and $Q_1+\Delta Q_1$ are orthonormal matrices in (3.1), and ΔQ_1 is small, $\Delta Q_1 = 0$ when $n = 1$, so the following analysis assumes $n > 1$. The results of Sun [15] give about as good as possible overall bounds on the change ΔQ_1 in Q_1. But by looking at how ΔQ_1 is distributed between $\mathcal{R}(Q_1)$ and its orthogonal complement, and following the ideas in Theorem 3.2, we are able to obtain a result which is tight but, unlike the related tight result in [15], easy to follow. It makes clear exactly where any ill-conditioning lies. From (3.14) with $Q = (Q_1, Q_2)$ square and orthogonal,

$$\Delta Q_1 = \epsilon Q_1 Q_1^T\dot{Q}_1(0) + \epsilon Q_2 Q_2^T\dot{Q}_1(0) + O(\epsilon^2),$$

and the key is to bound the first term on the right separately from the second. Note from (3.11) with (2.5) to (2.7) that

$$\|Q_2^T\dot{Q}_1(0)\|_F = \|Q_2^T G R^{-1}\|_F \le \|R^{-1}\|_2\|Q_2^T G\|_F = \kappa_2(A)\epsilon_2/\epsilon,$$

where G can be chosen to give equality here for any given A. Hence $\kappa_2(A)$ is the true condition number for that part of ΔQ_1 in $\mathcal{R}(Q_2)$:

$$\|Q_2^T\Delta Q_1\|_F \le \kappa_2(A)\epsilon_2 + O(\epsilon^2). \tag{4.1}$$

Now for the part of ΔQ_1 in $\mathcal{R}(Q_1)$. We see from (3.12) that

$$S \equiv Q_1^T\dot{Q}_1(0) = \mathrm{low}(Q_1^T G R^{-1}) - [\mathrm{low}(Q_1^T G R^{-1})]^T, \tag{4.2}$$

which is skew symmetric with clearly zero diagonal. Let R_j and S_j denote the leading $j \times j$ blocks of R and S respectively, G_j the matrix of the first j columns of G, and $Q_1 = [q_1, \ldots, q_n]$. If we write

$$S_1 = 0; \qquad S_j = \left(\begin{array}{c|c} S_{j-1} & -s_j \\ \hline s_j^T & 0 \end{array}\right), \quad j = 2, \ldots, n,$$

where s_j has j–1 elements, then from the upper triangular form of R in (4.2)

$$\begin{aligned} s_j^T &= q_j^T G_{j-1} R_{j-1}^{-1}, \\ \frac{1}{2}\|S\|_F^2 &= \|s_2\|_2^2 + \cdots + \|s_n\|_2^2 \\ &\le \|R_1^{-1}\|_2^2 \|G_1^T q_2\|_2^2 + \cdots + \|R_{n-1}^{-1}\|_2^2 \|G_{n-1}^T q_n\|_2^2 \\ &\le \|R_{n-1}^{-1}\|_2^2 \|G_{n-1}^T Q_1\|_F^2 \\ &\le \|R_{n-1}^{-1}\|_2^2 \|Q_1^T G\|_F^2. \end{aligned}$$

Clearly for any R_{n-1} equality is obtained by taking $G = (q_n y^T, 0)$, with y nonzero such that $\|R_{n-1}^{-T} y\|_2 = \|R_{n-1}^{-1}\|_2 \|y\|_2$. It follows that the bound is tight in

$$\|Q_1^T \dot{Q}_1(0)\|_F = \|S\|_F \le \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2 \epsilon_1/\epsilon,$$

$$\|Q_1^T \Delta Q_1\|_F \le \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2 \epsilon_1 + O(\epsilon^2), \tag{4.3}$$

so the true condition number for that part of ΔQ_1 in $\mathcal{R}(Q_1)$ is not $\sqrt{2}\kappa_2(A)$ but

$$\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2. \tag{4.4}$$

In some problems we are mainly interested in the change in Q_1 lying in $\mathcal{R}(Q_1)$, 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 be quite well-conditioned compared with R, and we will have $\|R_{n-1}^{-1}\|_2 \|A\|_2 << \kappa_2(A)$.

5. Perturbation analyses for R. In Section 3 we saw the key to deriving first order perturbation bounds for R in the QR factorization of full column rank A is the equation (3.5), which has the general form of finding (bounding) X in terms of given R and F in the matrix equation

$$R^T X + X^T R = R^T F + F^T R, \quad X \text{ and } R \text{ upper triangular, } R \text{ nonsingular.} \tag{5.1}$$

Sun [13] and Stewart [10] originally analyzed this using the matrix equation approach to give the result in Theorem 3.1. We will now analyze it in two new ways. The first, Stewart's [11, 3] refined matrix equation approach, gives a clear improvement on Theorem 3.1, while the second, Chang's [2, 3] matrix-vector equation approach, gives a further improvement still — provably tight bounds leading to the true condition number $\kappa_R(A)$ for R in the QR factorization of A. Both approaches provide efficient condition estimators (see [2] for the matrix–vector equation approach), and nice results for the special case of $AP = Q_1 R$, 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 [2], and only the results will be quoted here.

5.1. Refined matrix equation analysis for R. Our proof of Theorem 3.1 used (3.5) to produce the matrix equation (3.7), and derived the bounds directly from this. We now look at this approach more closely, but at first using the general form (5.1) to keep our thinking clear. From this we see

$$X = \mathrm{up}(FR^{-1} + R^{-T}F^T)R.$$

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

$$X = \mathrm{up}(F\bar{R}^{-1} + D^{-1}\bar{R}^{-T}F^TD)\bar{R} = [\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D]\bar{R}. \tag{5.2}$$

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

LEMMA 5.1. *For $n \times n$ B and $D \in \mathcal{D}_n$,*

$$\phi \equiv \|\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D\|_F \leq \sqrt{1 + \zeta_D^2}\|B\|_F, \tag{5.3}$$

where

$$\zeta_D \equiv \max_{1 \leq i < j \leq n} \{\delta_j/\delta_i\}. \tag{5.4}$$

Proof. Clearly

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

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

$$\begin{aligned}
\phi^2 &\leq \sum_{i=1}^{n} \beta_{ii}^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(\beta_{ij}^2 + \beta_{ji}^2)(1 + (\frac{\delta_j}{\delta_i})^2) \\
&= \|B\|_F^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(\beta_{ij}^2 + \beta_{ji}^2)(\frac{\delta_j}{\delta_i})^2 \\
&\leq \|B\|_F^2 + \zeta_D^2\|B\|_F^2. \qquad \square \qquad (5.5)
\end{aligned}$$

We can now bound the solution X of (5.1)

$$\begin{aligned}
\|X\|_F \leq \phi \cdot \|\bar{R}\|_2 &\leq \sqrt{1 + \zeta_D^2}\|B\|_F\|\bar{R}\|_2 \qquad (5.6)\\
&= \sqrt{1 + \zeta_D^2}\|F\bar{R}^{-1}\|_F\|\bar{R}\|_2 \\
&\leq \sqrt{1 + \zeta_D^2}\kappa_2(\bar{R})\|F\|_F. \qquad (5.7)
\end{aligned}$$

Since this is true for all $D \in \mathcal{D}_n$, we know that for the upper triangular solution X of (5.1)

$$\|X\|_F \leq \kappa_{ME}(A)\|F\|_F, \tag{5.8}$$

$$\kappa_{ME}(A) \equiv \inf_{D \in \mathcal{D}_n} \kappa(R, D), \tag{5.9}$$

$$\kappa(R, D) \equiv \sqrt{1 + \zeta_D^2}\kappa_2(D^{-1}R), \tag{5.10}$$

where ζ_D is defined in (5.4). This gives the encouraging result

$$\kappa_{ME}(A) \leq \kappa(R, I) = \sqrt{2}\kappa_2(R) = \sqrt{2}\kappa_2(A). \tag{5.11}$$

Comparing (3.5) with (5.1), we see for the QR factorization, with (2.5) to (2.7),

$$\|\dot{R}(0)\|_F \leq \kappa_{ME}(A)\|Q_1^T G\|_F = \kappa_{ME}(A)\|A\|_2\epsilon_1/\epsilon.$$

Hence

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \leq \kappa_{ME}(A)\epsilon_1/\epsilon, \tag{5.12}$$

and from (3.9) for a change $\Delta A = \epsilon G$ in A we have

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_{ME}(A)\epsilon_1 + O(\epsilon^2), \tag{5.13}$$

where from (5.11) these are never worse than the bounds in Theorem 3.1.

When we use the standard column pivoting strategy in $AP = Q_1 R$, P a permutation matrix, this analysis leads to a very nice result. Here the elements of R satisfy

$$r_{ii}^2 \geq \sum_{k=i}^{j} r_{kj}^2, \qquad i = 1, \dots, n, \quad j = i, \dots, n,$$

so $r_{11}^2 \geq r_{22}^2 \geq \dots \geq r_{nn}^2$. If D is the diagonal of R then $\zeta_D \leq 1$, and from (5.9) and (5.10)

$$\kappa_{ME}(AP) \leq \kappa(R, D) \leq \sqrt{2}\kappa_2(\bar{R}), \qquad \bar{R} = D^{-1}R.$$

But then $1 = |\bar{r}_{ii}| \geq |\bar{r}_{ij}|$ for all $j \geq i$, and it follows from [6, Theorem 8.13] that

$$1 \leq \|\bar{R}^{-1}\|_2 \leq 2^{n-1},$$

so since

$$\|\bar{R}\|_2^2 \leq \|\bar{R}\|_F^2 \leq n(n+1)/2,$$

$$\kappa_2(\bar{R}) \leq 2^{n-1} \cdot \sqrt{n(n+1)/2},$$

$$\kappa_{ME}(AP) \leq \sqrt{2}\kappa_2(\bar{R}) \leq 2^{n-1}\sqrt{n(n+1)}. \tag{5.14}$$

Thus when we use the standard pivoting strategy, we see the sensitivity of R is bounded for any n.

REMARK 5.1. *Clearly $\kappa_{ME}(A)$ is a potential candidate for the condition number of R in the QR factorization. From (5.9), $\kappa_{ME}(A)$ depends solely on R, but it will only be the true condition number if for* any *nonsingular upper triangular R we can find an F in (5.1) giving equality in (5.8). From (5.7) this can only be true if every column of F^T lies in the space of the right singular vectors corresponding to the maximum singular value of $\bar{R}^{-T}$. Such a restriction is in general too strong for (5.6) to be made an equality as well (see the lead up to (5.5)). However for a class of R this* is *possible. If R is diagonal, we can take $D = R$ giving $\bar{R} = I$, and the first restriction on F disappears. Let i and j be such that $\zeta_D = \delta_j/\delta_i$, $j > i$, and take $F = e_j e_i^T$, so from (5.2)*

$$X = \mathrm{up}(e_j e_i^T + \zeta_D e_i e_j^T),$$

$$\|X\|_F = \sqrt{1+\zeta_D^2} = \sqrt{1+\zeta_D^2}\kappa_2(\bar{R})\|F\|_F.$$

So we see that, at least for diagonal R, the bounds are tight, and in this restricted case $\kappa_{ME}(A)$ is *the true condition number.*

This refined matrix equation analysis shows to what extent the solution X of (5.1), and so the sensitivity of R in the QR factorization, is dependent on the row scaling in $R = D\bar{R}$. From the term $D^{-1}\mathrm{up}(\bar{R}^{-T}F^T)D$ in (5.2), we saw multipliers δ_j/δ_i occurred only with $j > i$. As a result we obtained ζ_D in our bounds rather than $\kappa_2(D)$, where

$$\zeta_D \leq \kappa_2(D),$$

with equality if and only if the minimum element comes before the maximum on the diagonal. Thus we obtained full cancellation of D^{-1} with D in the first term on the right hand side of (5.2), and partial cancellation in the second.

This 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 in $R = D\bar{R}$ can give $\kappa_2(\bar{R})$ very near one. If at the same time ζ_D is not large, then $\kappa(R, D)$ in (5.10) can be much smaller than $\sqrt{2}\kappa_2(R)$, see (5.11). Standard pivoting always ensures that such a D exists, and in fact gives (5.14). However if we do not use such pivoting, then Remark 5.1 suggests that any relatively small earlier elements on the diagonal of R could correspond to poor conditioning of the factorization.

We will return to $\kappa_{ME}(A)$ and $\kappa(R, D)$ when we seek practical estimates of the true condition number that we derive in the next section.

5.2. Matrix–vector equation analysis for R. We can now obtain provably sharp, but less intuitive results by viewing the matrix equation (5.1) as a large matrix–vector equation. For any matrix $C \equiv (c_{ij}) \equiv [c_1, \ldots, c_n] \in \mathcal{R}^{n\times n}$, denote by $c_j^{(i)}$ the vector of the first i elements of c_j. With this, we define ("u" denotes "upper")

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

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

To analyze (3.5) we again consider the general form (5.1), repeated here for clarity,

$$R^TX + X^TR = R^TF + F^TR, \quad X \text{ and } R \text{ upper triangular, } R \text{ nonsingular.} \tag{5.15}$$

which we saw via (2.4) has a unique upper triangular solution X. The upper and lower triangular parts of (5.1) contain identical information, and we now write the upper triangular part in matrix–vector, rather than matrix–matrix format. The first j elements of the jth column of (5.15) are given by

$$R_{jj}^Tx_j^{(j)} + X_{jj}^Tr_j^{(j)} = R_{jj}^Tf_j^{(j)} + \begin{bmatrix} f_1^{(j)T} \\ f_2^{(j)T} \\ \cdot \\ f_j^{(j)T} \end{bmatrix} r_j^{(j)},$$

and by rewriting this, we can see how to solve for $x_j^{(j)}$, $j = 1, \ldots, n$:

$$(R_{jj}^T + e_j r_j^{(j)T})x_j^{(j)} + \begin{bmatrix} r_j^{(1)T} x_1^{(1)} \\ r_j^{(2)T} x_2^{(2)} \\ \cdot \\ r_j^{(j-1)T} x_{j-1}^{(j-1)} \\ 0 \end{bmatrix} = (R_{jj}^T + e_j r_j^{(j)T})f_j^{(j)} + \begin{bmatrix} r_j^{(j)T} f_1^{(j)} \\ r_j^{(j)T} f_2^{(j)} \\ \cdot \\ r_j^{(j)T} f_{j-1}^{(j)} \\ 0 \end{bmatrix},$$

which, on dividing the last row of this by 2, gives

$$W_R \mathrm{uvec}(X) = Z_R \mathrm{vec}(F) \tag{5.16}$$

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

$$\left[\begin{array}{ccc|ccc|c|cccc}
r_{11} & & & & & & & & & & \\
r_{12} & r_{11} & & & & & & & & & \\
 & r_{12} & r_{22} & & & & & & & & \\
r_{13} & & & r_{11} & & & & & & & \\
 & r_{13} & r_{23} & r_{12} & r_{22} & & & & & & \\
 & & & r_{13} & r_{23} & r_{33} & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
r_{1n} & & & & & & & r_{11} & & & \\
 & r_{1n} & r_{2n} & & & & & r_{12} & r_{22} & & \\
 & & & r_{1n} & r_{2n} & r_{3n} & & r_{13} & r_{23} & r_{33} & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & r_{1n} & r_{2n} & r_{3n} \ \cdot & r_{nn}
\end{array}\right],$$

and $Z_R \in \mathcal{R}^{\frac{n(n+1)}{2} \times n^2}$ 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 (5.16)

$$\mathrm{uvec}(X) = W_R^{-1} Z_R \mathrm{vec}(F). \tag{5.17}$$

Remembering X is upper triangular, we see

$$\begin{aligned} \|X\|_F = \|\mathrm{uvec}(X)\|_2 &= \|W_R^{-1} Z_R \mathrm{vec}(F)\|_2 \\ &\le \|W_R^{-1} Z_R\|_2 \|\mathrm{vec}(F)\|_2 \\ &= \|W_R^{-1} Z_R\|_2 \|F\|_F, \end{aligned} \tag{5.18}$$

where for *any* nonsingular upper triangular R, equality can be obtained by choosing $\mathrm{vec}(F)$ to lie in the space spanned by the right singular vectors corresponding to the largest singular value of $W_R^{-1}Z_R$. It follows that (5.18) is tight, so from (5.8), derived from the matrix equation approach, and (5.11)

$$\|W_R^{-1}Z_R\|_2 \le \kappa_{ME}(A) \le \sqrt{2}\kappa_2(A). \tag{5.19}$$

REMARK 5.2. *Usually the first and second inequalities are strict. For example, let* $R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$*. Then we obtain* $\|W_R^{-1}Z_R\|_2 = 1.7321$*,* $\kappa_{ME}(A) = 2.8679$ *and* $\sqrt{2}\kappa_2(A) = 3.7025$ *by MATLAB (*$\kappa_{ME}(A) = 2.8679$ *was got through solving an optimization problem). But from Remark 5.1 the first inequality becomes an equality if* R *is diagonal, while the second also becomes an equality if* R *is an* $n \times n$ *identity matrix with* $n \ge 2$*, so the upper bound 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

$$\|W_R^{-1}Z_R\|_2 \ge 1. \tag{5.20}$$

REMARK 5.3. *This lower bound is approximately tight for any* n*, as can be seen by taking* $R = diag(1, \delta, \ldots, \delta^{n-1})$*,* $0 < \delta \le 1$*, for by taking* $D = R$ *in* (5.19)*,* (5.9) *and* (5.10)*, we see from Remark 5.1 that*

$$1 \le \|W_R^{-1}Z_R\|_2 = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1+\delta^2} \to 1 \quad as \quad \delta \to 0. \tag{5.21}$$

These results, and the analysis in Section 4 for Q_1, lead to our new first order normwise perturbation theorem.

THEOREM 5.2. *Let* $A = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}$ *be the QR factorization of* $A \in \mathcal{R}^{m\times n}$ *with full column rank, and let* ΔA *be a real* $m \times n$ *matrix. Let* $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$*,* $\epsilon_1 \equiv \|Q_1^T\Delta A\|_F/\|A\|_2$ *and* $\epsilon_2 \equiv \|Q_2^T\Delta A\|_F/\|A\|_2$*. If* (2.8) *holds, then there is a unique QR factorization*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R), \tag{5.22}$$

such that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_R(A)\epsilon_1 + O(\epsilon^2), \tag{5.23}$$

where with W_R *and* Z_R *as in* (5.16)*, and* $\kappa_{ME}(A)$ *as in* (5.9) *and* (5.10)*,*

$$1 \le \kappa_R(A) \equiv \|W_R^{-1}Z_R\|_2 \le \kappa_{ME}(A) \le \sqrt{2}\kappa_2(A), \tag{5.24}$$

and

$$\|Q_2^T\Delta Q_1\|_F \le \kappa_2(A)\epsilon_2 + O(\epsilon^2),$$
$$\|Q_1^T\Delta Q_1\|_F \le \kappa_{Q_1}(A)\epsilon_1 + O(\epsilon^2),$$

where

$$\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2\|A\|_2 \le \sqrt{2}\kappa_2(A).$$

Proof. From Corollary 2.2 $A+\Delta A$ has the unique QR factorization (5.22). From (3.5), (5.15), (5.17) with $G \equiv \Delta A/\epsilon$ and (2.5) we have

$$\mathrm{uvec}(\dot{R}(0)) = W_R^{-1} Z_R \mathrm{vec}(Q_1^T G),$$

so taking the 2-norm gives

$$\|\dot{R}(0)\|_F \leq \|W_R^{-1} Z_R\|_2 \|Q_1^T G\|_F = \|W_R^{-1} Z_R\|_2 \|A\|_2 \epsilon_1/\epsilon.$$

Combining this with $\|A\|_2 = \|R\|_2$, $\|A^\dagger\|_2 = \|R^{-1}\|_2$, (5.19) and (5.20) we get (5.24) and

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \leq \kappa_R(A)\epsilon_1/\epsilon. \tag{5.25}$$

Thus, from the Taylor series (3.9) of $R(t)$, (5.23) follows. The remaining results are restatements of (5.20), (5.19), (4.1) and (4.3). □

REMARK 5.4. *From* (5.24) *we know the first order perturbation bound* (5.23) *is at least as sharp as* (3.2)*, but it suggests it may be considerably sharper. In fact it can be sharper by an arbitrary factor. Consider the example in Remark 5.3, where taking $D = R$,*

$$\kappa_R(A) = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1+\delta^2}, \quad \kappa_2(A) = 1/\delta,$$

and

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

We see the first order perturbation bound (3.2) can severely overestimate the effect of a perturbation in A.

REMARK 5.5. *If we take $R = diag(\delta^{1-n}, \ldots, \delta, 1)$, $0 < \delta \leq 1$, we see $\kappa_2(R) = \kappa_2(A) = \delta^{1-n}$, while*

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

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.

When we use standard *pivoting*, we see from (5.24) and (5.14)

$$1 \leq \kappa_R(AP) \equiv \|W_R^{-1} Z_R\|_2 \leq \kappa_{ME}(AP) \leq 2^{n-1}\sqrt{n(n+1)},$$

but the following tighter result is shown in [2, Th. 2.2]

THEOREM 5.3. *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $AP = Q_1 R$ when the standard column pivoting strategy is used. Then*

$$1 \leq \kappa_R(AP) = \|W_R^{-1} Z_R\|_2 \leq \|W_R^{-1} Z_R\|_F \leq \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}. \tag{5.26}$$

There is a parametrized family of matrices $A(\theta)$, $\theta \in (0, \pi/2]$, for which

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

Theorem 5.3 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 we give an extreme example in Section 6 where it is not.

When we do not use pivoting, we have no such simple result for $\kappa_R(A)$, and it is, as far as we can see, unreasonably expensive to compute or approximate $\kappa_R(A)$ directly with the usual approach. Fortunately $\kappa_{ME}(A)$ is apparently an excellent approximation to $\kappa_R(A)$, and $\kappa_{ME}(A)$ *is* quite easy to estimate. All we need to do is choose a suitable D in $\kappa(R, D)$ in (5.10). We consider how to do this in the next section.

6. Numerical experiments and condition estimators. In Section 5 we presented new first order perturbation bounds for the R factor of the QR factorization using two different approaches, defined $\kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2$ as the true condition number for the R factor, and suggested $\kappa_R(A)$ could be estimated in practice by $\kappa(R, D)$. Our new first order results are sharper than previous results for R, and at least as sharp for Q_1, and we give some numerical tests to illustrate both this, and possible estimators for $\kappa_R(A)$.

We would like to choose D such that $\kappa(R, D)$ is a good approximation to the minimum $\kappa_{ME}(A)$ in (5.9), and show that this is a good estimate of the true 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 [5]) to estimate $\kappa_2(D^{-1}R)$, and take $\kappa(R, D)$ in (5.10) as the appropriate estimate.

By a well known result of van der Sluis [16], $\kappa_2(D^{-1}R)$ will be nearly minimal when the rows of $D^{-1}R$ are equilibrated. But this could lead to a large ζ_D in (5.10). 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. The first set of matrices are $n \times n$ Pascal matrices, (with elements $a_{1j} = a_{i1} = 1$, $a_{ij} = a_{i,j-1} + a_{i-1,j}$), $n = 1, 2, \ldots, 15$. The results are shown in Table 6.1 without pivoting, giving $A = Q_1 R$, and in Table 6.2 with pivoting, giving $AP = \tilde{Q}_1 \tilde{R}$. Note in Table 6.1 how $\sqrt{2}\kappa_2(A)$ can be far worse than the true condition number $\kappa_R(A)$, which itself can be much greater than its lower bound of 1. In Table 6.2 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 6.1 that both $\kappa(R, D_1)$ and $\kappa(R, D_2)$ are very good estimates for $\kappa_R(A)$. The latter is a little better than the former. In Table 6.2 $\kappa(\tilde{R}, D_1) = \kappa(\tilde{R}, D_2)$ (in fact $D_1 = D_2$), and they are also good estimates for $\kappa_R(AP)$.

Table 6.1
Results for Pascal matrices without pivoting, $A = Q_1 R$

n	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa(R, D2)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
1	1.0000e+00	1.4142e+00	1.4142e+00	—	1.4142e+00
2	1.8708e+00	3.4275e+00	1.8708e+00	2.6180e+00	9.6932e+00
3	4.6332e+00	1.4330e+01	1.4330e+01	1.8542e+01	8.7658e+01
4	1.4359e+01	6.1440e+01	6.1440e+01	1.5533e+02	9.7855e+02
5	4.9822e+01	2.5991e+02	2.5991e+02	1.5805e+03	1.2046e+04
6	1.8021e+02	1.0834e+03	1.0697e+03	1.7948e+04	1.5668e+05
7	6.6572e+02	4.4665e+03	4.2021e+03	2.1759e+05	2.1120e+06
8	2.4916e+03	1.8274e+04	1.6675e+04	2.7562e+06	2.9197e+07
9	9.4091e+03	7.4364e+04	6.6391e+04	3.6040e+07	4.1123e+08
10	3.5768e+04	3.0143e+05	2.6459e+05	4.8273e+08	5.8763e+09
11	1.3668e+05	1.2183e+06	1.0546e+06	6.5894e+09	8.4943e+10
12	5.2442e+05	4.9134e+06	4.1862e+06	9.1330e+10	1.2394e+12
13	2.0190e+06	1.9782e+07	1.6614e+07	1.2819e+12	1.8226e+13
14	7.7958e+06	7.9545e+07	6.6157e+07	1.8183e+13	2.6979e+14
15	3.0179e+07	3.1952e+08	2.6390e+08	2.6023e+14	4.0149e+15

Table 6.2
Results for Pascal matrices with pivoting, $AP = \tilde{Q}_1 \tilde{R}$

n	$\kappa_R(AP)$	$\kappa(\tilde{R}, D1)$	$\kappa(\tilde{R}, D2)$	$\kappa_{Q_1}(AP)$	$\sqrt{2}\kappa_2(A)$
1	1.0000e+00	1.4142e+00	1.4142e+00	—	1.4142e+00
2	1.1832e+00	1.7920e+00	1.7920e+00	1.6558e+00	9.6932e+00
3	1.2892e+00	2.1648e+00	2.1648e+00	1.2541e+01	8.7658e+01
4	1.6901e+00	3.3864e+00	3.3864e+00	1.0995e+02	9.7855e+02
5	1.8129e+00	4.1424e+00	4.1424e+00	1.0252e+03	1.2046e+04
6	2.2296e+00	4.7281e+00	4.7281e+00	7.5426e+03	1.5668e+05
7	2.0543e+00	5.1140e+00	5.1140e+00	8.4696e+04	2.1120e+06
8	2.6264e+00	6.5487e+00	6.5487e+00	1.1777e+06	2.9197e+07
9	3.4892e+00	8.8383e+00	8.8383e+00	1.4573e+07	4.1123e+08
10	3.3983e+00	9.4153e+00	9.4153e+00	2.3500e+08	5.8763e+09
11	3.3616e+00	9.2315e+00	9.2315e+00	2.3096e+09	8.4943e+10
12	3.3468e+00	9.7367e+00	9.7367e+00	2.9703e+10	1.2394e+12
13	3.3417e+00	1.1423e+01	1.1423e+01	3.5016e+11	1.8226e+13
14	3.6106e+00	1.2386e+01	1.2386e+01	5.3554e+12	2.6979e+14
15	3.3430e+00	1.1591e+01	1.1591e+01	8.5638e+13	4.0149e+15

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 6.3 without pivoting. All the results with pivoting are similar to that for $j = 8$ in Table 6.3, and so are not given here. For $j = 1, 2 \ldots, 7$, $\kappa_R(A)$ and $\kappa_{Q_1}(A)$ are both close to their upper bound $\sqrt{2}\kappa_2(A)$, but for $j = 8$, both $\kappa_R(A)$ and $\kappa_{Q_1}(A)$ 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 $\kappa_{Q_1}(A)$ 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_{88} 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 $\kappa(R, D1)$, $\kappa(R, D_2)$, 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 $\kappa_{Q_1}(A)$, the condition number for the change in Q_1 lying in the range of Q_1, 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_1 and R.

Table 6.3
Results for 10×8 matrix A_j, $j = 1, \ldots, 8$, without pivoting

j	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa(R, D2)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
1	1.9309e+08	4.0176e+08	3.0302e+08	3.0294e+08	4.7920e+08
2	1.2567e+08	2.8967e+08	2.7462e+08	2.6149e+08	3.8207e+08
3	1.9316e+08	4.4740e+08	3.9250e+08	4.6596e+08	5.4803e+08
4	1.3926e+08	3.0732e+08	2.6490e+08	2.8514e+08	4.4639e+08
5	1.2467e+08	3.1208e+08	2.3992e+08	3.9067e+08	4.2323e+08
6	8.8237e+07	2.2252e+08	1.6584e+08	3.4710e+08	3.9061e+08
7	9.2648e+07	2.1010e+08	1.7127e+08	4.4303e+08	5.4719e+08
8	2.2580e+00	5.4735e+00	4.9152e+00	6.6109e+00	6.2096e+08

The third set of matrices are $n \times n$ upper triangular

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

where $c = \cos(\theta)$, $s = \sin(\theta)$. These matrices were introduced by Kahan [7]. Of course $Q_1 = I$ 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 6.5. Again we found $D_1 = D_2$, and only list the results corresponding to D_1.

Table 6.5
Results for Kahan matrices, $\theta = \pi/8$

n	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
5	7.9795e+00	1.6993e+01	2.2367e+02	1.1075e+03
10	2.0908e+02	6.1182e+02	1.0139e+06	5.0970e+06
15	5.5106e+03	2.0703e+04	4.0256e+09	2.0238e+10
20	1.4524e+05	6.5195e+05	1.5006e+13	7.5439e+13
25	4.2936e+06	1.9672e+07	5.4053e+16	2.7174e+17

In all these examples we see $\kappa(R, D_1)$ and $\kappa(R, D_2)$ gave excellent estimates for $\kappa_R(A)$, with $\kappa(R, D_2)$ being marginally preferable. For the Kahan matrices, which correspond to correctly pivoted A, we see that in extreme cases, with large enough n, $\kappa_R(A)$ can be large even with pivoting. This is about as bad a result as we can get with pivoting (it gets a bit worse as $\theta \to 0$ in R), since the Kahan matrices are the parameterized family mentioned in Theorem 5.3. Nevertheless $\kappa(R, D_1)$ and $\kappa(R, D_2)$ still estimate $\kappa_R(A)$ excellently.

7. Summary and conclusion. The first order perturbation analyses presented here show just what the sensitivity (condition) of each of Q_1 and R is in the QR factorization of full column rank A, and in so doing provide their true 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 $A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R)$ are:

- $\kappa_2(A)$ for that part of ΔQ_1 in $\mathcal{R}(\mathcal{A})^\perp$, see (4.1),
- $\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2\|A\|_2$ for that part of ΔQ_1 in $\mathcal{R}(\mathcal{A})$, see (4.3),
- $\kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2$ for R, see Theorem 5.2,
- the estimate for $\kappa_R(A)$, that is $\kappa_{ME}(A) \equiv \inf_{D \in \mathcal{D}_n} \kappa(R, D)$,
  where $\kappa(R, D) \equiv \sqrt{1 + \zeta_D^2}\kappa_2(D^{-1}R)$, see (5.13).

The condition numbers obey

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

for Q_1, while for R

$$1 \leq \kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2 \leq \kappa_{ME}(A) \leq \sqrt{2}\kappa_2(A),$$

see (5.24). The numerical examples, and an analysis of the $n = 2$ case (not given here), suggest that $\kappa(R, D)$, with D chosen to equilibrate $\bar{R} \equiv D^{-1}R$ subject to $\zeta_D \leq 1$, 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 Remark 5.1 showed by taking $D = R$,

$$\kappa_R(A) = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1 + \zeta_D^2} \leq \sqrt{2}\kappa_2(D) = \sqrt{2}\kappa_2(A).$$

For general A when we use the standard column pivoting strategy in the QR factorization, $AP = Q_1 R$, we saw from (5.14) and [2] that

$$\kappa_{ME}(AP) \leq 2^{n-1}\sqrt{n(n+1)},$$
$$\kappa_R(AP) \leq \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}.$$

As a result of these analyses we see both R and in a certain sense Q_1 can be less sensitive than was thought from previous analyses. The true 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_1 in the range of Q_1, for a given size of perturbation in A.

The matrix equation analysis of Section 5.1 also provides a nice analysis of an interesting and possibly more general matrix equation (5.1).

By following the approach of Stewart [8, Th. 3.1], see also [12, 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 [3]. We could also provide new componentwise bounds, but we chose not to do either of these here, in order to keep the material and the basic ideas as brief and approachable as possible. Our condition numbers and resulting bounds are asymptotically sharp, so there is less need for strict bounds. A new componentwise bound for R is given in [2].

Acknowledgement. We would like to thank Ji-guang Sun for his suggestions and encouragement, and for providing us with draft versions of his work.

REFERENCES

[1] X. Chang, PhD Thesis, in preparation, McGill University, School of Computer Science, 1996.

[2] X.-W. Chang, C.C. Paige, A perturbation analysis for R in the QR factorization, Technical Report, **SOCS-95.7**, McGill University, School of Computer Science, 1995.

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

[4] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University press, 2nd ed., Baltimore, MD, 1989.

[5] N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596

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

[7] W. Kahan, *Numerical linear algebra*, Can. Math. Bull., 9 (1966), pp. 757–801.

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

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

[10] G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1141–1146.

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

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

[13] J.-G. Sun, *Perturbation bounds for the Cholesky and QR factorization*, BIT, 31 (1991), pp. 341–352.

[14] J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.

[15] J.-G. Sun, *On perturbation bounds for the QR factorization*, Linear Algebra and Appl., 215 (1995), pp. 95–112.

[16] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numerische Mathematik, 14 (1969), pp. 14–23.

[17] H. Zha, *A componentwise perturbation analysis of the QR decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1124–1131.