Numerische Mathematik manuscript No.
(will be inserted by the editor)

Componentwise Perturbation Analyses for the QR Factorization *

Xiao-Wen Chang, Chris Paige

School of Computer Science, McGill University
Montreal, Quebec, Canada, H3A 2A7
e-mail: chang@cs.mcgill.ca, paige@cs.mcgill.ca

Received: / Revised version:

Summary This paper gives componentwise perturbation analyses for Q and R in the QR factorization $A = QR$, $Q^TQ = I$, R upper triangular, for a given real $m \times n$ matrix A of rank n. Such specific analyses are important for example when the columns of A are badly scaled. First order perturbation bounds are given for both Q and R. The analyses more accurately reflect the sensitivity of the problem than previous such results. The condition number for R is bounded for a fixed n when the standard column pivoting strategy is used. This strategy also tends to improve the condition of Q, so usually the computed Q and R will both have higher accuracy when we use the standard column pivoting strategy. Practical condition estimators are derived. The assumptions on the form of the perturbation ΔA are explained and extended. Weaker rigorous bounds are also given.

Key words QR factorization – sensitivity analysis – perturbation bounds – condition numbers – pivoting strategies

Mathematics Subject Classification (1991): 15A23, 65F35

1 Introduction

The QR factorization is an important tool in matrix computations (see for example [6, Chap. 5]): given an $m \times n$ real matrix A with

* This research was supported by NSERC of Canada Grants RGPIN217191-99 for Xiao-Wen Chang, and OGP0009236 for Chris Paige.
Correspondence to: Xiao-Wen Chang

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 an $m \times n$ real matrix such that $A + \Delta A$ still has full column rank, then $A + \Delta A$ has the unique QR factorization

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

where $(Q+\Delta Q)^T(Q+\Delta Q) = I$ and $R+\Delta R$ is upper triangular with positive diagonal elements. The goal of the perturbation analysis for the QR factorization is to determine bounds on $\|\Delta Q\|$ (or $|\Delta Q|$) and $\|\Delta R\|$ (or $|\Delta R|$) in terms of (a bound on) $\|\Delta A\|$ (or $|\Delta A|$), where for a matrix $C = (c_{ij})$, $|C|$ is defined by $(|c_{ij}|)$.

The perturbation analysis for the QR factorization has been considered by several authors. Given (a bound on) $\|\Delta A\|$, the first results were presented by Stewart [12]. Analyses based on bounds on $\|\Delta A\|$ are sometimes called normwise or norm-based perturbation analyses. Stewart's results were modified and improved by Sun [14]. Similar results to those of Sun [14] were obtained by Stewart [13] by a different approach. Later Sun [16] gave new strict perturbation bounds for Q alone. More recently Chang *et al.* [4] gave new first-order perturbation analyses using the so called refined matrix equation and matrix-vector equation approaches. Analyses based on bounds on $|\Delta A|$ have been called componentwise analyses. Given a bound on $|\Delta A|$, Sun [15] presented *strict* but somewhat complicated bounds on $|\Delta Q|$ and $|\Delta R|$. In [18], Zha considered the following class of perturbations:

$$|\Delta A| \le \epsilon\, C|A|; \qquad C \in \mathcal{R}^{m\times m}, \quad 0 \le c_{ij} \le 1, \tag{1.1}$$

and presented first-order perturbation bounds on $\|\Delta Q\|$ and $\|\Delta R\|$. An important motivation for considering such a class of perturbations is that the equivalent backward rounding error from a rounding error analysis of the standard QR factorization fits in this class, see Higham [7, Chap. 18] and the last paragraph of Section 2 here.

The main purpose of this paper is to establish new first-order perturbation analyses under the condition (1.1). The perturbation bounds that are derived here are significantly sharper than the equivalent results in Zha [18, Theorem 2.1]. Simple *rigorous* perturbation bounds are also presented. Thus the present paper will, among other

things, increase our understanding of the errors we can expect in computing Q and R in $A = QR$.

In Section 2 we discuss the generality of the class of perturbations (1.1), how this class may be extended, and how the equivalent backward rounding error for the Householder QR factorization belongs to this class. In Section 3 we define our notation. In Section 4 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 7 and 8, but in Section 5 they are used to derive simple 2- and F-norm versions of Zha's results [18, Theorem 2.1] on the sensitivity of R and Q. Section 6 derives basic rigorous bounds that will help us understand some of the more refined first-order bounds. In Section 7 we give a refined perturbation analysis for Q, showing why the standard column pivoting strategy for A can be beneficial for certain aspects of the sensitivity of Q. In Section 8 we analyze the perturbation in R by the matrix–vector equation approach, then we combine this with the matrix equation approach to get useful estimates. The ideas behind these two approaches were discussed in the norm-based perturbation analysis for the QR factorization [4]. Here these approaches show that the sensitivity of R can be significantly improved by pivoting. We give numerical results and suggest practical condition estimators in Section 9. We summarize our findings and point out possible future work in Section 10.

2 The class of perturbations, and rounding effects

We now discuss the generality of the assumption (1.1). Taking $C = I$ in (1.1) gives bounds on each element $|\Delta a_{ij}|$ of the form

$$|\Delta A| \le \epsilon\,|A|,$$

which covers the case of small relative errors in the elements. Now suppose that we only have the column information

$$\|\Delta a_j\|_1 \le \epsilon\,\|a_j\|_1, \quad j = 1, \ldots, n.$$

This implies $|\Delta a_{ij}| \le \|\Delta a_j\|_1 \le \epsilon\, e^T |a_j|$ with $e = [1, 1, \ldots, 1]^T$, which implies (1.1) with $C = ee^T$. Similarly (1.1) with $C = ee^T$ implies $\|\Delta a_j\|_1 \le \epsilon m \|a_j\|_1$. Since for any $v \in \mathcal{R}^m$,

$$\|v\|_\infty \le \|v\|_1 \le m^{1/2}\|v\|_2 \le m^{1/2}\|v\|_1 \le m^{3/2}\|v\|_\infty,$$

etc., we see that (1.1) essentially handles any information of the form

$$\|\Delta a_j\|_p \le \epsilon\,\|a_j\|_p, \quad j = 1, \ldots, n, \quad p = 1, 2, \infty. \tag{2.1}$$

Thus (1.1) is an elegant way of handling most bounds on the elements or the columns of A. However to cover cases where some columns of ΔA have different relative bounds than others, as might happen when the columns of A are obtained by experimental observation at different times or with different instruments, we can extend (1.1) to

$$|\Delta A| \le \epsilon\, C|A|D;\ C \in \mathcal{R}^{m\times m},\ \ 0 \le c_{ij} \le 1;\ \ D = \mathrm{diag}(\delta_1, \ldots, \delta_n) > 0. \tag{2.2}$$

This then includes the extension of (2.1)

$$\|\Delta a_j\|_p \le \epsilon\, \delta_j \|a_j\|_p, \quad j = 1, \ldots, n. \tag{2.3}$$

For the QR factorization $A = QR$, (2.2) leads to

$$|\Delta A| \le \epsilon\, C|Q|\cdot|R|D \quad \text{(for clarity "$\cdot$" indicates multiplication)},$$

and where Zha [18] considered $\| \, |R|\cdot|R^{-1}| \, \|_p$, which is independent of column scaling, we can define the (extended) condition number

$$\mathrm{cond}_p(R, D) \equiv \| \, |R|\cdot D\cdot|R^{-1}| \, \|_p, \quad p = 1, 2, \infty. \tag{2.4}$$

This condition number is also independent of column scaling, and can be arbitrarily smaller than $\|RD\|_p\|R^{-1}\|_p$. For brevity we give the analysis without D and assume (1.1) only, but all the results can trivially be extended to changes in A of the form (2.2).

The equivalent backward error for a numerically stable QR factorization is important for this exposition. For an $m \times n$ matrix A of rank n let $A = QR$ be the exact, and $A \approx Q_c R_c$ the computed QR factorization of A obtained via Householder transformations. Higham [7, Theorem 18.4]) showed

$$A + \Delta A = \widetilde{Q} R_c, \quad |\Delta A| \le \epsilon\, C|A|, \quad \epsilon = \gamma_{m,n} u, \tag{2.5}$$

where $\widetilde{Q}^T\widetilde{Q} = I$, $\gamma_{m,n}$ is a moderate constant depending on m and n, u is the unit roundoff, $C \ge 0$ and $\|C\|_F = 1$. The bound on ΔA has the form (1.1), so the perturbation analyses here will allow us to obtain good bounds on the errors $\widetilde{Q} - Q$ and $R_c - R$. Also the computed Q_c satisfies $Q_c = \widetilde{Q} + \Delta$, where $|\Delta| \le \gamma_{m,n} u C_2 |\widetilde{Q}|$ with $C_2 \ge 0$, $\|C_2\|_F = 1$. Since $Q_c - Q = \Delta + (\widetilde{Q} - Q)$ we have

$$\|Q_c - Q\|_F \le n^{1/2}\gamma_{m,n} u + \|\widetilde{Q} - Q\|_F. \tag{2.6}$$

For the whole of this paper we assume perturbations satisfying (1.1).

3 Notation

In this paper, for any matrix $X \in \mathcal{R}^{m\times n}$, we denote by $(X)_{i,:}$ the ith row of X, and by $(X)_{:,j}$ the jth column of X.

For any nonsingular matrix X we define

$$\kappa_2(X) \equiv \|X\|_2\|X^{-1}\|_2, \quad \mathrm{cond}_2(X) \equiv \| \, |X|\cdot|X^{-1}| \, \|_2. \tag{3.1}$$

Notice $\| \, |X^{-1}|\cdot|X| \, \|$ is the standard Bauer–Skeel condition number, but the present variant seems more intuitive for column scaling.

For an $m \times n$ matrix Q such that $Q^TQ = I$ we can find $\bar{Q}$ such that $[Q, \bar{Q}]$ is square and orthogonal, then define

$$\eta_1 \equiv \| \, |Q^T|\cdot C\cdot|Q| \, \|_F, \quad \eta_2 \equiv \| \, |\bar{Q}^T|\cdot C\cdot|Q| \, \|_F, \quad \eta_3 \equiv \||C|Q| \, \|_F. \tag{3.2}$$

Since $\| \, |Q^T| \, \|_F = \| \, |Q| \, \|_F = n^{1/2}$, $\| \, |\bar{Q}^T| \, \|_F = (m-n)^{1/2}$, and in (1.1) $\|C\|_F \le m$, if we use the fact that $\|AB\|_F \le \|A\|_F\|B\|_F$ we obtain

$$\eta_1 \le mn, \quad \eta_2 \le ((m-n)n)^{1/2}m, \quad \eta_3 \le mn^{1/2}.$$

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

$$\mathrm{up}(X) \equiv \begin{bmatrix} \frac{1}{2}x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}x_{nn} \end{bmatrix}, \qquad \mathrm{low}(X) \equiv \mathrm{up}(X^T)^T, \tag{3.3}$$

so that $X = \mathrm{low}(X) + \mathrm{up}(X)$. For any $n \times n$ ($n > 1$) matrix X and positive definite $D = \mathrm{diag}(\delta_1, \ldots, \delta_n)$, we can show (for a proof, see Lemma 5.1 in [4]; it is straightforward by considering elements)

$$\|\mathrm{up}(X)+D^{-1}\mathrm{up}(X^T)D\|_F \le \rho_D\|X\|_F, \quad \rho_D \equiv \Big[1+\max_{1\le i<j\le n}(\delta_j/\delta_i)^2\Big]^{1/2}. \tag{3.4}$$

In particular with $D = I$,

$$\|\mathrm{up}(X + X^T)\|_F \le \sqrt{2}\,\|X\|_F; \ \|\mathrm{up}(X)\|_F \le \frac{1}{\sqrt{2}}\|X\|_F \text{ if } X = X^T. \tag{3.5}$$

It is also easy to see that for any $n \times n$ ($n > 1$) matrix X

$$\|X - \mathrm{up}(X + X^T)\|_F = \|\mathrm{low}(X) - [\mathrm{low}(X)]^T\|_F \le \sqrt{2}\,\|X\|_F. \tag{3.6}$$

When $n = 1$, we have the following equalities:

$$\|\mathrm{up}(X)+D^{-1}\mathrm{up}(X^T)D\|_F = \|\mathrm{up}(X+X^T)\|_F = 2\,\|\mathrm{up}(X)\|_F = \|X\|_F,$$
$$\|X - \mathrm{up}(X + X^T)\|_F = 0.$$

In this paper we assume that the matrix A has more than one column, i.e., $n > 1$. The case $n = 1$ is trivial and straightforward bounds can be derived by using these last equalities.

4 Rate of change of Q and R

Here we derive the basic results on how Q and R change as A changes. The following theorem summarizes several results that we use later.

Theorem 4.1 *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank n with the QR factorization $A = QR$, and let ΔA be a real $m \times n$ matrix satisfying*

$$\Delta A = \epsilon\, G;\ \epsilon \geq 0,\ |G| \leq C|A|,\ C \in \mathcal{R}^{m\times m},\ 0 \leq c_{ij} \leq 1. \tag{4.1}$$

Let $A^\dagger$ denote the Moore-Penrose inverse of A. If

$$\epsilon \,\| \,|A^\dagger|\cdot C\cdot|A|\,\|_2 < 1, \tag{4.2}$$

then $A + tG$ has the unique QR factorization

$$A(t) \equiv A + tG = Q(t)R(t), \quad Q^T(t)Q(t) = I, \quad |t| \leq \epsilon, \tag{4.3}$$

where

$$R^T\dot{R}(0) + \dot{R}^T(0)R = R^TQ^TG + G^TQR, \tag{4.4}$$

$$\dot{R}(0) = \mathrm{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]R, \tag{4.5}$$

$$\dot{Q}(0) = GR^{-1} - Q\,\mathrm{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]. \tag{4.6}$$

In particular, $A + \Delta A$ has the unique QR factorization

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R), \tag{4.7}$$

where ΔR and ΔQ satisfy

$$\Delta R = \epsilon\dot{R}(0) + O(\epsilon^2), \tag{4.8}$$

$$\Delta Q = \epsilon\dot{Q}(0) + O(\epsilon^2). \tag{4.9}$$

Proof Since $\|X\|_2 \leq \||X|\|_2$, if (4.2) holds, then for all $|t| \leq \epsilon$,

$$\|tA^\dagger G\|_2 \leq \epsilon\| \,|A^\dagger|\cdot C\cdot|A|\,\|_2 < 1.$$

Also from

$$Q^T(A + tG) = R + tQ^TG = R(I + tR^{-1}Q^TG) = R(I + tA^\dagger G)$$

we see that $Q^T(A + tG)$ is nonsingular. Hence for all $|t| \leq \epsilon$, $A + tG$ has full column rank and the unique QR factorization (4.3). Taking $t = \epsilon$ shows that (4.7) is unique, and then $R(0) = R$, $R(\epsilon) = R + \Delta R$, $Q(0) = Q$ and $Q(\epsilon) = Q + \Delta Q$.

It is easy to verify that $Q(t)$ and $R(t)$ are twice continuously differentiable for $|t| \leq \epsilon$ from a standard 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 (4.4). This we will see is a linear equation *uniquely* defining the elements of upper triangular $\dot{R}(0)$ in terms of the elements of $Q^T G$. From upper triangular $\dot{R}(0)R^{-1}$ in

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

we see with (3.3) that (4.5) holds. Differentiating (4.3) at $t = 0$ gives

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

which with (4.5) gives (4.6). Finally the Taylor expansions for $R(t)$ and $Q(t)$ about $t = 0$ give (4.8) and (4.9) at $t = \epsilon$. $\square$

The perturbation ΔA in (4.1) satisfies (1.1), and we will always assume (4.2) holds, so the results of this theorem will apply for the rest of this paper.

5 Zha's first-order bounds

We can use the notation of (3.1) and (3.2) to derive the combined 2-norm and F-norm results which are analogous to Zha's [18] first-order consistent monotone norm results, but are a little simpler in form and derivation. We then give examples to show how these can be too pessimistic. From (4.5), we have by using (3.5) and (4.1) that

$$\begin{aligned}
\|\dot{R}(0)\|_F &\le \|\mathrm{up}[Q^T G R^{-1} + (Q^T G R^{-1})^T]\|_F \|R\|_2 \\
&\le \sqrt{2}\,\|Q^T G R^{-1}\|_F \|R\|_2 \le \sqrt{2}\,\| \,|Q^T|\cdot C \cdot |A| \cdot |R^{-1}| \,\|_F \|R\|_2 \\
&\le \sqrt{2}\,\| \,|Q^T|\cdot C \cdot |Q| \cdot |R| \cdot |R^{-1}| \,\|_F \|R\|_2 \le \sqrt{2}\,\eta_1 \mathrm{cond}_2(R)\|R\|_2.
\end{aligned}$$

Similarly, from (4.6), (3.6), (4.1), if $[Q, \bar{Q}]$ is square and orthogonal,

$$\begin{aligned}
\|\dot{Q}(0)\|_F^2 &= \|Q^T \dot{Q}(0)\|_F^2 + \|\bar{Q}^T \dot{Q}(0)\|_F^2 \\
&= \|Q^T G R^{-1} - \mathrm{up}[Q^T G R^{-1} + (Q^T G R^{-1})^T]\|_F^2 + \|\bar{Q}^T G R^{-1}\|_F^2 \\
&\le 2\,\|Q^T G R^{-1}\|_F^2 + \|\bar{Q}^T G R^{-1}\|_F^2 \le 2\,\|G R^{-1}\|_F^2, \\
\|\dot{Q}(0)\|_F &\le \sqrt{2}\,\| \,C|Q|\cdot|R|\cdot|R^{-1}| \,\|_F \le \sqrt{2}\,\eta_3 \mathrm{cond}_2(R).
\end{aligned}$$

Thus with (4.8) and (4.9) we get the following bounds:

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \eta_1 \varphi(A)\epsilon + O(\epsilon^2), \tag{5.1}$$

$$\|\Delta Q\|_F \le \eta_3 \varphi(A)\epsilon + O(\epsilon^2), \tag{5.2}$$

$$\varphi(A) \equiv \sqrt{2}\,\mathrm{cond}_2(R). \tag{5.3}$$

Apart from the multipliers η_1 and η_3 (see also (6.6), (6.5)), $\varphi(A)$ can be thought of as (an upper bound on) the condition number for

both Q and R (for small enough ΔA) when we use the combination of 2 and F norms. The constant $\sqrt{2}$ involved in the definition of $\varphi(A)$ may be removed, but we keep it here since it is useful for comparison with the modified results to be given in Section 8. Notice that $\varphi(A)$ is invariant under any column scaling of A. This is a significant improvement on the normwise perturbation results published before [4] when the perturbation ΔA satisfies (1.1), but sometimes these perturbation bounds do not reflect the true sensitivity of the QR factorization very well, as we see from the following example.

Example 5.1 Consider the following two matrices (the first one is quoted from [18], the new one gives even worse results).

$$A_1 = \begin{bmatrix} 1 & 1 \\ 0 & 10^{-10} \\ 1 & 1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 1 & 1-10^{-10} \\ 1 & 1+10^{-10} \end{bmatrix}. \tag{5.4}$$

Computing the QR factorization of A_1 and A_2 in MATLAB with unit roundoff $u \approx 1.11 \times 10^{-16}$ (all our computations were performed in MATLAB 5.2 on a Pentium-II running LINUX), we obtained the following computed factors, shown here to 5 figures (to make the diagonal elements of R_{1c} and R_{2c} positive, some signs have been altered).

$$Q_{1c} = \begin{bmatrix} \texttt{7.0711e-01} & \texttt{-1.2539e-06} \\ \texttt{0} & \texttt{1.0000e+00} \\ \texttt{7.0711e-01} & \texttt{1.2539e-06} \end{bmatrix}, \; R_{1c} = \begin{bmatrix} \texttt{1.4142} & \texttt{1.4142e+00} \\ \texttt{0} & \texttt{1.0000e-10} \end{bmatrix},$$

$$Q_{2c} = \begin{bmatrix} \texttt{7.0711e-01} & \texttt{-7.0711e-01} \\ \texttt{7.0711e-01} & \texttt{7.0711e-01} \end{bmatrix}, \; R_{2c} = \begin{bmatrix} \texttt{1.4142} & \texttt{1.4142e+00} \\ \texttt{0} & \texttt{1.4142e-10} \end{bmatrix}.$$

These have errors

$$\begin{aligned} e_{Q_1} &= \|Q_{1c} - Q_1\|_F \approx 1.7 \times 10^{-6}, \quad e_{R_1} = \tfrac{\|R_{1c} - R_1\|_F}{\|R_1\|_2} \approx 2.2 \times 10^{-16}, \\ e_{Q_2} &= \|Q_{2c} - Q_2\|_F \approx 1.9 \times 10^{-16}, \quad e_{R_2} = \tfrac{\|R_{2c} - R_2\|_F}{\|R_2\|_2} \approx 3.3 \times 10^{-17}, \end{aligned} \tag{5.5}$$

where $A_1 = Q_1 R_1$ and $A_2 = Q_2 R_2$ are the exact QR factorizations. The condition numbers (5.3) are

$$\varphi(A_1) \approx 4.0 \times 10^{10}, \qquad \varphi(A_2) \approx 2.8 \times 10^{10}. \tag{5.6}$$

MATLAB computes the QR factorization using Householder transformations. Comparing (2.5) with (4.7) we see $\Delta Q = \widetilde{Q} - Q$ and $\Delta R = R_c - R$, so (2.6) and (5.2) with $\epsilon = \gamma_{m,n} u$ in (2.5) show that

$$\|Q_c - Q\|_F \le \gamma'_{m,n} \varphi(A) u + O(u^2), \tag{5.7}$$

where $\gamma'_{m,n}$ is a moderate constant depending on m and n. Finally (5.1) with $\epsilon = \gamma_{m,n}u$ in (2.5) shows that

$$\frac{\|R_c - R\|_F}{\|R\|_2} \leq \gamma''_{m,n}\varphi(A)u + O(u^2), \tag{5.8}$$

where again $\gamma''_{m,n}$ is a moderate constant depending on m and n. For more details of such arguments, see [7, pp.367–368, 382] and [18]. From the computed results, we see for Q_1 the bound (5.7) matches the actual error e_{Q_1} in (5.5) very well, but for Q_2, R_1 and R_2 the bounds (5.7) and (5.8) are bad overestimates of the corresponding errors. □

Although the matrices (5.4) are contrived, they do represent a fairly common case when A has a very large condition number: each matrix has only one very small singular value. By choosing such examples with small dimensions we are able to illustrate the drawbacks of the bounds in [18] simply and directly, showing that it is necessary to obtain stronger perturbation bounds.

6 Rigorous bounds

Later we will derive tighter first-order bounds, but in order to explain some subtleties of these we first obtain some simple but weak rigorous bounds. From the QR factorization (4.7), with $A = QR$,

$$R^T\Delta R + \Delta R^T R + \Delta R^T\Delta R = R^TQ^T\Delta A + \Delta A^TQR + \Delta A^T\Delta A.$$

Multiplying on the left by R^{-T} and the right by R^{-1} we see that

$$\begin{aligned}\Delta RR^{-1} + R^{-T}\Delta R^T = {} & Q^T\Delta AR^{-1} + R^{-T}\Delta A^TQ \\ & + R^{-T}(\Delta A^T\Delta A - \Delta R^T\Delta R)R^{-1}.\end{aligned}$$

Since ΔRR^{-1} is upper triangular, this gives with (3.3)

$$\Delta RR^{-1} = \mathrm{up}[Q^T\Delta AR^{-1} + R^{-T}\Delta A^TQ + R^{-T}(\Delta A^T\Delta A - \Delta R^T\Delta R)R^{-1}].$$

Using (3.5) we obtain

$$\begin{aligned}\|\Delta RR^{-1}\|_F &\leq \sqrt{2}\,\|Q^T\Delta AR^{-1}\|_F + (\|\Delta AR^{-1}\|_F^2 + \|\Delta RR^{-1}\|_F^2)/\sqrt{2}, \\ \sqrt{2}\,\|\Delta RR^{-1}\|_F &\leq \|\Delta AR^{-1}\|_F(2 + \|\Delta AR^{-1}\|_F) + \|\Delta RR^{-1}\|_F^2. \end{aligned} \tag{6.1}$$

Also from $(Q + \Delta Q)(R + \Delta R) = QR + \Delta A$ we see that

$$\begin{aligned}&\Delta QR + (Q + \Delta Q)\Delta R = \Delta A, \\ &\Delta Q = \Delta AR^{-1} - (Q + \Delta Q)\Delta RR^{-1}, \\ &\|\Delta Q\|_F \leq \|\Delta AR^{-1}\|_F + \|\Delta RR^{-1}\|_F. \end{aligned} \tag{6.2}$$

If we replace ΔA here by tG, $|t| \le \epsilon$ satisfying (4.2) as in Theorem 4.1, then (6.1) and (6.2) still hold with ΔQ and ΔR continuous functions of t. Let $\rho(t) \equiv \|\Delta R R^{-1}\|_F$, $\delta(t) \equiv \|\Delta A R^{-1}\|_F$, $\beta(t) \equiv \delta(t)(2+\delta(t))$, then $\rho(0) = \delta(0) = \beta(0) = 0$, and from (6.1)

$$\rho(t)(\sqrt{2} - \rho(t)) \le \beta(t).$$

Here the left hand side has its maximum of $1/2$ with $\rho(t) = 1/\sqrt{2}$. If $\beta(t) < 1/2$ for $|t| \le \epsilon$ then the left hand side cannot attain its maximum, and so for $|t| \le \epsilon$, $\rho(t) < 1/\sqrt{2}$. This means that $\sqrt{2} - \rho(t) > 1/\sqrt{2}$, and

$$\rho(t) \le \beta(t)/(\sqrt{2} - \rho(t)) \le \sqrt{2}\beta(t) = \sqrt{2}\delta(t)(2+\delta(t)), \qquad |t| \le \epsilon. \tag{6.3}$$

But with $\delta(t) \ge 0$, $\beta(t) \equiv \delta(t)(2+\delta(t)) = (\delta(t)+1)^2 - 1 < 1/2$ if and only if $\delta(t) < \sqrt{3}/\sqrt{2} - 1$, and the following rigorous bounds hold.

Theorem 6.1 *Assume that the conditions and assumptions of Theorem 4.1 hold together with*

$$\epsilon\|GR^{-1}\|_F \equiv \|\Delta A R^{-1}\|_F < \sqrt{3}/\sqrt{2} - 1 \approx .3178. \tag{6.4}$$

Then $A + \Delta A = A + \epsilon G$ has the unique QR factorization

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

where with the notation of (3.1) and (3.2),

$$\|\Delta Q\|_F \le (1 + \sqrt{2} + \sqrt{3})\eta_3 \mathrm{cond}_2(R)\epsilon, \tag{6.5}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le (\sqrt{2} + \sqrt{3})\eta_3 \mathrm{cond}_2(R)\epsilon. \tag{6.6}$$

Proof Since $|G| \le C \cdot |Q| \cdot |R|$ from (4.1), (6.3) and (6.4) show

$$\begin{aligned}\|\Delta R R^{-1}\|_F &\le \sqrt{2}\,\|\Delta A R^{-1}\|_F(2+\sqrt{3/2}-1) = (\sqrt{2}+\sqrt{3})\|\Delta A R^{-1}\|_F \\ &\le (\sqrt{2}+\sqrt{3})\epsilon\,\||C|Q|\cdot|R|\cdot|R^{-1}|\|_F \le (\sqrt{2}+\sqrt{3})\eta_3 \mathrm{cond}_2(R)\epsilon,\end{aligned}$$

This result with (6.2) gives (6.5). Finally (6.6) follows since $\|\Delta R\|_F \le \|\Delta R R^{-1}\|_F \|R\|_2$. □

Remark 6.1 The bounds (6.5) and (6.6) are clearly the rigorous versions of the first order bounds (5.2) and (5.1), which were analogous to Zha's [18, Theorem 2.1] results. Thus (6.5) and (6.6) are just as weak as (5.2) and (5.1) were shown to be by Example 5.1. □

7 Refined analysis for Q

The expression (5.2) gives an important *overall* bound on the change ΔQ in Q. But by looking at how ΔQ is distributed between $\mathcal{R}(Q)$, the range of Q, and its orthogonal complement $\mathcal{R}^{\perp}(Q)$, we will obtain better results. These show more clearly where any ill-conditioning lies.

Take a matrix $\bar{Q} \in \mathcal{R}^{m\times(m-n)}$ such that $[Q, \bar{Q}]$ is square and orthogonal. Then from (4.9) we see that

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

and the key is to bound the first term on the right separately from the second. Since Q is an orthonormal matrix, $Q^T\dot{Q}(0) = 0$ when $n = 1$, and results involving $Q^T\dot{Q}(0)$ will only be nontrivial when $n > 1$.

For the part of $\dot{Q}(0)$ orthogonal to $\mathcal{R}(Q)$, we see from (4.6) that

$$\bar{Q}^T\dot{Q}(0) = \bar{Q}^T GR^{-1}, \tag{7.1}$$

and combining this with (4.1) gives

$$\|\bar{Q}\bar{Q}^T\dot{Q}(0)\|_F = \|\bar{Q}\bar{Q}^T GR^{-1}\|_F \le \| \,|\bar{Q}^T|\cdot C\cdot|Q|\,\|_F \,\| \,|R|\cdot|R^{-1}|\,\|_2.$$

Thus with (3.2) and (3.1) we have

$$\begin{aligned} \bar{Q}\bar{Q}^T\Delta Q &= \epsilon\bar{Q}\bar{Q}^T\dot{Q}(0) + O(\epsilon^2), \\ \|\bar{Q}\bar{Q}^T\Delta Q\|_F &\le \eta_2 \mathrm{cond}_2(R)\epsilon + O(\epsilon^2), \end{aligned} \tag{7.2}$$

and $\mathrm{cond}_2(R)$ can be regarded as the condition number for the part of ΔQ in $\mathcal{R}(\bar{Q})$. Note the similarity with (5.2).

Now we deal with that part of ΔQ lying in $\mathcal{R}(Q)$, first we show there is an important *lower* bound on $\|QQ^T\Delta Q\|_2$. Since $Q+\Delta Q$ has orthonormal columns,

$$\begin{aligned} &(Q+\Delta Q)^T(Q+\Delta Q) = I + Q^T\Delta Q + \Delta Q^T Q + \Delta Q^T\Delta Q = I, \\ &\|\Delta Q^T\Delta Q\|_2 = \|\Delta Q\|_2^2 = \|Q^T\Delta Q + \Delta Q^T Q\|_2 \le 2\,\|Q^T\Delta Q\|_2, \end{aligned} \tag{7.3}$$

and we have the useful lower bound

$$\|QQ^T\Delta Q\|_2 = \|Q^T\Delta Q\|_2 \ge \frac{1}{2}\|\Delta Q\|_2^2.$$

To obtain a good upper bound, we will manipulate the equations to avoid using the triangle equality ($\|X+Y\| \le \|X\| + \|Y\|$) *etc.* where possible. We see from (4.6) and (3.3) that

$$\begin{aligned} Q^T\dot{Q}(0) &= Q^T GR^{-1} - \mathrm{up}[Q^T GR^{-1} + (Q^T GR^{-1})^T] \\ &= \mathrm{low}(Q^T GR^{-1}) - [\mathrm{low}(Q^T GR^{-1})]^T, \end{aligned} \tag{7.4}$$

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

$$\begin{array}{c} \\ Q = \end{array}\begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ [Q_{n-1}, q], \end{array} \qquad \begin{array}{c} \\ G = \end{array}\begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ [G_{n-1}, g], \end{array} \qquad \begin{array}{c} \\ R = \end{array}\begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ \begin{bmatrix} R_{n-1} & r \\ 0 & r_{nn} \end{bmatrix}. \end{array}$$

This allows us to rewrite (7.4) as

$$\begin{aligned} Q^T\dot{Q}(0) &= \mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1}, Q^T(-G_{n-1}R_{n-1}^{-1}r+g)/r_{nn}]) \\ &\quad - \{\mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1}, Q^T(-G_{n-1}R_{n-1}^{-1}r+g)/r_{nn}])\}^T \\ &= \mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1}, 0]) - \{\mathrm{low}([Q^TG_{n-1}R_{n-1}^{-1}, 0])\}^T. \quad (7.5) \end{aligned}$$

From (4.1) it follows that

$$|G_{n-1}| \le C|Q_{n-1}|\cdot|R_{n-1}|, \tag{7.6}$$

and using (3.6), we have from (7.5) and (7.6) that

$$\begin{aligned} \|QQ^T\dot{Q}(0)\|_F &\le \sqrt{2}\,\|Q^TG_{n-1}R_{n-1}^{-1}\|_F \\ &\le \sqrt{2}\,\|\,|Q^T|\cdot C\cdot|Q_{n-1}|\,\|_F\cdot\|\,|R_{n-1}|\cdot|R_{n-1}^{-1}|\,\|_2 \\ &\le \sqrt{2}\,\|\,|Q^T|\cdot C\cdot|Q|\,\|_F\|\,|R_{n-1}|\cdot|R_{n-1}^{-1}|\,\|_2. \end{aligned}$$

This with (4.9), (3.2) and (3.1) gives our bound

$$\|QQ^T\Delta Q\|_F \le \sqrt{2}\,\eta_1\mathrm{cond}_2(R_{n-1})\epsilon + O(\epsilon^2). \tag{7.7}$$

If we define

$$\kappa_Q(A) \equiv \sqrt{2}\,\mathrm{cond}_2(R_{n-1}), \tag{7.8}$$

then we can regard this as the the condition number (for small enough ΔA) for that part of ΔQ in $\mathcal{R}(Q)$, and summarize the results for Q.

Theorem 7.1 *Suppose all the assumptions of Theorem 4.1 hold, and $\bar{Q} \in \mathcal{R}^{m\times(m-n)}$ is a matrix such that $[Q, \bar{Q}]$ is orthogonal. Then $A + \Delta A = A + \epsilon G$ has the unique QR factorization*

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

such that

$$\|\bar{Q}\bar{Q}^T\Delta Q\|_F \le \eta_2\,\mathrm{cond}_2(R)\epsilon + O(\epsilon^2), \tag{7.9}$$

$$\frac{1}{2}\|\Delta Q\|_2^2 \le \|QQ^T\Delta Q\|_F \le \eta_1\kappa_Q(A)\epsilon + O(\epsilon^2). \tag{7.10}$$

If $\epsilon\|GR^{-1}\|_F \equiv \|\Delta AR^{-1}\|_F < \sqrt{3}/\sqrt{2} - 1$ holds as well, then

$$\|\Delta Q\|_F = [\|QQ^T\Delta Q\|_F^2 + \|\bar{Q}\bar{Q}^T\Delta Q\|_F^2]^{1/2} \le (1+\sqrt{2}+\sqrt{3})\eta_3\mathrm{cond}_2(R)\epsilon. \tag{7.11}$$

Here η_1, η_2 and η_3 are defined in (3.2), *$\mathrm{cond}_2(\cdot)$ is defined in* (3.1), *and $\kappa_Q(A)$ is defined by* (7.8). □

Proof The unique QR factorization follows from Theorem 4.1, (7.9) is just (7.2), (7.10) follows from (7.3) and (7.7)–(7.8), while (7.11) is just (6.5) in Theorem 6.1, since (6.4) holds. □

In some problems we are interested in the change in Q lying in $\mathcal{R}(Q)$, that is $QQ^T\Delta Q$. For example when A is square $\bar{Q}$ is nonexistent, and the change in Q necessarily lies in $\mathcal{R}(Q)$. Theorem 7.1 shows the upper bound on $\|QQ^T\Delta Q\|_F$ can be smaller than was previously thought in for example [18], see (5.2). In particular if A has only one small singular value and its columns are not badly scaled (both matrices in Example 5.1 are of this form), and we use the standard column pivoting strategy in computing the QR factorization (see, for example, [6, p248]), then usually we will have $\mathrm{cond}_2(R_{n-1}) \ll \mathrm{cond}_2(R)$. For the two matrices in Example 5.1, the values of $\mathrm{cond}_2(R_{n-1})$ are 1 and 1, while the values of $\mathrm{cond}_2(R)$ are about 3×10^{10} and 2×10^{10}, with or without column pivoting.

In some special cases standard column pivoting may not give $\mathrm{cond}_2(R_{n-1}) \ll \mathrm{cond}_2(R)$, for example the Kahan matrices (see [10]). For these a rank revealing pivoting strategy such as in [9] is required to obtain a significant improvement of $\mathrm{cond}_2(R_{n-1})$ over $\mathrm{cond}_2(R)$, see the $\kappa_Q(A)$ or $\kappa_Q(AP)$ ($\sqrt{2}\,\mathrm{cond}_2(R_{n-1})$) and $\varphi(A)$ or $\varphi(AP)$ ($\sqrt{2}\,\mathrm{cond}_2(R)$) columns in Tables 9.9 and 9.10 of Section 9.

Now we return to the error e_{Q_2} in (5.5) for the example with A_2 in (5.4). When $m = n$, $\bar{Q}$ does not exist, so (7.10) gives

$$\|\Delta Q\|_F \le \eta_1\kappa_Q(A)\epsilon + O(\epsilon^2),$$

and by a similar argument to that for (5.7), we have for the MATLAB QR factorization

$$\|Q_c - Q\|_F \le \gamma'_{n,n}\kappa_Q(A)u + O(u^2). \tag{7.12}$$

For A_2, $m = n = 2$, so $\kappa_Q(A_2) = \sqrt{2}$ in (7.8). We see for Q_2 the bound of $O(10^{-16})$ using (7.12) matches the observed e_{Q_2} of 1.9×10^{-16} in (5.5) well, whereas the bound of $O(10^{-6})$ using (5.7) was very weak.

Remark 7.1 When $m > n$ it is possible to have $\|\Delta Q\|_F \gg \|QQ^T\Delta Q\|_F$, and (7.10) must be used carefully. Of course (7.10) is *asymptotically* correct, but when ϵ is large enough, the $O(\epsilon^2)$ term can dominate in the upper bound in (7.10) when the overall $\|\Delta Q\|_F$ is large. That is, the *multiplier* in the $O(\epsilon^2)$ term can be very large. This is illustrated nicely in the computational example with A_1 in (5.4), for there $m = 3$, but $n = 2$ so $\kappa_Q(A_1) = \sqrt{2}$, also from (2.5)–(2.6)

$$A_1 + \Delta A_1 = \widetilde{Q}_1 R_{1c}, \quad \widetilde{Q}_1^T\widetilde{Q}_1 = I,$$
$$\|\widetilde{Q}_1 - Q_{1c}\|_F = O(10^{-16}), \quad \|\Delta A_1\|_F = O(10^{-16}).$$

We see from the argument preceding (2.6), and e_{Q_1} in (5.5), that

$$\Delta Q_1 \equiv \widetilde{Q}_1 - Q_1 \approx Q_{1c} - Q_1, \quad e_{Q_1} \equiv \|Q_{1c} - Q_1\|_F \approx 1.7 \times 10^{-6},$$
$$\varphi(A_1) \equiv \sqrt{2}\,\mathrm{cond}_2(R_1) \approx 4.0 \times 10^{10},$$

see (5.6). However we also found using MATLAB that

$$\|Q_1 Q_1^T (Q_{1c} - Q_1)\|_F \approx 2.5 \times 10^{-12},$$

so that necessarily

$$\|Q_1 Q_1^T \Delta Q_1\|_F = \|Q_1 Q_1^T (\widetilde{Q}_1 - Q_1)\|_F \approx 2.5 \times 10^{-12},$$

which is much larger than the first-order term in the upper bound in (7.10), whose value is $O(10^{-16})$. But from our computations the lowest bound in (7.10) is $\frac{1}{2}\|\Delta Q_1\|_2^2 \approx 2.4 \times 10^{-12}$, which is also much larger than the first-order term, so the $O(\epsilon^2)$ term dominates the ϵ term in (7.10) even though $\epsilon \approx 10^{-16}$, explaining this result. □

Theorem 7.1 can be used effectively as follows. Estimate $\mathrm{cond}_2(R)\epsilon$ and $\kappa_Q(A)\epsilon$. Since (7.11) is rigorous, the $O(\epsilon^2)$ term in (7.9) can never obscure the $\eta_2 \mathrm{cond}_2(R)\epsilon$ term, so use this latter as the (approximate) bound. $(\eta_3 \mathrm{cond}_2(R)\epsilon)^2$ gives an indication of how large the lower bound in (7.10) could be. The $O(\epsilon)$ term in the upper bound of (7.10) is an excellent asymptotic bound, but if $(\eta_3 \mathrm{cond}_2(R)\epsilon)^2 \gg \eta_1\, \kappa_Q(A)\epsilon$, then the $O(\epsilon^2)$ term may dominate in (7.10), and we are forced to use $\eta_3 \mathrm{cond}_2(R)\epsilon$ as the (approximate) upper bound for $\|QQ^T \Delta Q\|_F$ as well.

8 Perturbation analyses for R

In Section 4 we saw (4.4) gives a basis for deriving first-order perturbation bounds for R in the QR factorization of a full column rank matrix A. In [4] it was shown there were two effective approaches to carrying out such derivations. These were named the matrix–vector equation approach, and the (refined) matrix equation approach. The matrix–vector equation approach will be used to provide a good measure of the conditioning of R, and a tight lower bound on the resulting condition number. We will then use ideas from the refined matrix equation approach to obtain an upper bound on this condition number, and a useful approach to estimating it in practice.

8.1 Matrix–vector equation analysis for R

The first approach views the matrix equation (4.4) 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 column 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.

The matrix equation (4.4) can be rewritten as the following form (for the derivation, see [2,4], or just write down the "uvec" of (4.4) for the $n = 2$ case).

$$W_R \mathrm{uvec}(\dot{R}(0)) = Z_R \mathrm{vec}(Q^T G), \tag{8.1}$$

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

$$W_R \equiv \begin{bmatrix} r_{11} & & & & & & \\ r_{12} & r_{11} & & & & & \\ & r_{12} & r_{22} & & & & \\ \cdot & \cdot & \cdot & \cdot & & & \\ r_{1n} & & & r_{11} & & & \\ & r_{1n} & r_{2n} & r_{12} & r_{22} & & \\ & & & \cdot & \cdot & \cdot & \\ & & & r_{1n} & r_{2n} & \cdot & r_{nn} \end{bmatrix},$$

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

Since R is nonsingular, W_R is also, and from (8.1)

$$\mathrm{uvec}(\dot{R}(0)) = W_R^{-1} Z_R \mathrm{vec}(Q^T G). \tag{8.2}$$

In [4] we only assumed a bound on $\|G\|_F$, and the tight condition number for R was immediately seen to be $\|W_R^{-1}Z_R\|_2$. Here the analysis has to be more subtle to take account of the important additional information in (4.1).

From (4.1)

$$|Q^TG| \le |Q^T|\cdot|G| \le |Q^T|\cdot C\cdot|A| \le |Q^T|\cdot C\cdot|Q|\cdot|R|,$$

and with this (8.2) gives

$$|\mathrm{uvec}(\dot{R}(0))| \le |W_R^{-1}Z_R|\mathrm{vec}(|Q^TG|) \le |W_R^{-1}Z_R|\mathrm{vec}(|Q^T|\cdot C\cdot|Q|\cdot|R|).$$

The second inequality here appears to lead to upper bounds which are not in general tight, but this seems unavoidable. Note for any matrix $Y \in \mathcal{R}^{p\times m}$ and $N \in \mathcal{R}^{m\times n}$,

$$\mathrm{vec}(YN) = (N^T \otimes I_p)\mathrm{vec}(Y), \tag{8.3}$$

where $\otimes$ denotes the Kronecker product (see for example [11, p. 410]). It follows that

$$|\mathrm{uvec}(\dot{R}(0))| \le |W_R^{-1}Z_R|\cdot|R^T \otimes I_n|\mathrm{vec}(|Q^T|\cdot C\cdot|Q|). \tag{8.4}$$

Taking the 2-norm, we obtain

$$\|\dot{R}(0)\|_F \le \| \, |W_R^{-1}Z_R|\cdot|R^T \otimes I_n| \, \|_2 \, \| \, |Q^T|\cdot C\cdot|Q| \, \|_F.$$

Finally using the Taylor expansion (4.8) and the notation in (3.2),

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \eta_1 \frac{\| \, |W_R^{-1}Z_R|\cdot|R^T \otimes I_n| \, \|_2}{\|R\|_2}\epsilon + O(\epsilon^2). \tag{8.5}$$

Thus for perturbations of the form (4.1) we can regard the quantity

$$\kappa_R(A) \equiv \frac{\| \, |W_R^{-1}Z_R|\cdot|R^T \otimes I_n| \, \|_2}{\|R\|_2} \tag{8.6}$$

as the condition number for R in the QR factorization of A.

Now we obtain a lower bound for $\kappa_R(A)$. It is not difficulty to verify (see the Appendix of [2]) from the structures of W_R and Z_R that the first row of $W_R^{-1}Z_R$ is

$$(\underbrace{1,0,\ldots,0}_{n},\underbrace{0,\ldots,0}_{(n-1)n}),$$

and the $i(i-1)/2+1$-th row of $W_R^{-1}Z_R$ is, for $i = 2,\ldots,n$,

$$(\underbrace{0,r_{2i}/r_{11},\ldots,r_{ii}/r_{11},0,\ldots,0}_{n},\underbrace{0,\ldots,0}_{(i-2)n},\underbrace{1,0,\ldots,0}_{n},\underbrace{0,\ldots,0}_{(n-i)n}).$$

Thus by simple multiplications, we see that the first n elements of the $i(i-1)/2+1$-th row of $|W_R^{-1}Z_R|\cdot|R^T\otimes I_n|$ are

$$(|r_{1i}|,|r_{2i}|,\ldots,|r_{ii}|,\underbrace{0,\ldots,0}_{n-i}),\quad i=1,\ldots,n.$$

That is to say there exists a permutation matrix P such that

$$P|W_R^{-1}Z_R|\cdot|R^T\otimes I_n| = \begin{pmatrix} |R^T| & \times \\ \times & \times \end{pmatrix}.$$

Hence we have

$$\kappa_R(A) = \frac{\| \, |W_R^{-1}Z_R|\cdot|R^T\otimes I_n| \, \|_2}{\|R\|_2} \geq 1, \tag{8.7}$$

where it can be seen from the matrices W_R and Z_R following (8.1) that this lower bound is attained with $R=I$.

The difficulty with the condition number $\kappa_R(A)$ is that it is expensive to compute or even estimate directly. In Section 8.2 we will obtain bounds suggesting a practical condition estimator.

8.2 Obtaining upper bounds using matrix equation ideas

The matrix equation description (4.5) showing $\dot{R}(0) = \mathrm{up}[Q^TGR^{-1} + (Q^TGR^{-1})^T]R$, is just another way of saying the same thing as $\mathrm{uvec}(\dot{R}(0)) = W_R^{-1}Z_R\mathrm{vec}(Q^TG)$ in (8.2). So for any $X\in\mathcal{R}^{n\times n}$,

$$W_R^{-1}Z_R\mathrm{vec}(X) = \mathrm{uvec}\{\mathrm{up}[XR^{-1}+(XR^{-1})^T]R\}. \tag{8.8}$$

It is clear from the right hand side of this that each element of $W_R^{-1}Z_R$ is a sum of terms, where each term is a product of an element of R^{-1} with an element of R. It follows that for any $X\geq 0\in\mathcal{R}^{n\times n}$,

$$|W_R^{-1}Z_R|\mathrm{vec}(X) \leq \mathrm{uvec}\{\mathrm{up}[X|R^{-1}|+(X|R^{-1}|)^T]|R|\}. \tag{8.9}$$

This can also be proven by comparing the ith elements of both sides of (8.9) ($i=1,2,\ldots,n(n+1)/2$). Let $D_i = \mathrm{diag}(\mathrm{sign}((W_R^{-1}Z_R)_{i,:}))$ and define $X_i\in\mathcal{R}^{n\times n}$ by $\mathrm{vec}(X_i) = D_i\cdot\mathrm{vec}(X)$. Then for $X\geq 0$

$$\begin{aligned}
(|W_R^{-1}Z_R|\mathrm{vec}(X))_i &= (W_R^{-1}Z_R)_{i,:}\cdot D_i\cdot\mathrm{vec}(X)\\
&= (W_R^{-1}Z_R)_{i,:}\cdot\mathrm{vec}(X_i)\\
&= (\mathrm{uvec}\{\mathrm{up}[X_iR^{-1}+(X_iR^{-1})^T]R\})_i,\quad \text{see (8.8)},\\
&\leq (\mathrm{uvec}\{\mathrm{up}[|X_i|\cdot|R^{-1}|+(|X_i|\cdot|R^{-1}|)^T]|R|\})_i.
\end{aligned}$$

Notice that $X \geq 0$ and $|X_i| = X$, so (8.9) indeed holds. Now we define M to be our matrix of interest in (8.6), that is

$$M \equiv |W_R^{-1} Z_R| \cdot |R^T \otimes I_n|, \qquad \kappa_R(A) \equiv \|M\|_2 / \|R\|_2. \tag{8.10}$$

For later use, notice from (8.3) that for any $Y \in \mathcal{R}^{n \times n}$

$$|W_R^{-1} Z_R| \mathrm{vec}(Y|R|) = M \mathrm{vec}(Y).$$

We want to find practical bounds for $\|M\|_2$. Write $\mathcal{D}_n$ for the set of all $n \times n$ real positive definite diagonal matrices. For any $D \in \mathcal{D}_n$, let $R = D\bar{R}$. Notice that for any matrix B we have $\mathrm{up}(B)D = \mathrm{up}(BD)$. Now from (8.9) with $X \equiv Y|R|$ and $Y \geq 0$, it follows that

$$\begin{aligned}
\frac{\|M\mathrm{vec}(Y)\|_2}{\|\mathrm{vec}(Y)\|_2} &= \frac{\||W_R^{-1} Z_R| \mathrm{vec}(Y|R|)\|_2}{\|\mathrm{vec}(Y)\|_2} \\
&\leq \frac{\|\mathrm{uvec}\{\mathrm{up}[Y|R| \cdot |R^{-1}| + (Y|R| \cdot |R^{-1}|)^T]|R|\}\|_2}{\|\mathrm{vec}(Y)\|_2} \\
&= \|\mathrm{up}[Y|R| \cdot |R^{-1}| + (Y|R| \cdot |R^{-1}|)^T]|R|\,\|_F / \|Y\|_F \\
&= \|\mathrm{up}[Y|R| \cdot |\bar{R}^{-1}| + D^{-1}(Y|R| \cdot |\bar{R}^{-1}|)^T D]|\bar{R}|\,\|_F / \|Y\|_F \\
&\leq \rho_D \|Y|R| \cdot |\bar{R}^{-1}|\,\|_F \|\bar{R}\|_2 / \|Y\|_F, \quad \text{see (3.4)} \\
&\leq \rho_D \| \, |R| \cdot |R^{-1}| D \|_2 \|D^{-1} R\|_2.
\end{aligned}$$

But since $M \geq 0$, $\|M\|_2 = \max_{Y \geq 0, Y \neq 0} \|M\mathrm{vec}(Y)\|_2 / \|\mathrm{vec}(Y)\|_2$, so

$$\|M\|_2 \leq \rho_D \| \, |R| \cdot |R^{-1}| D \|_2 \|D^{-1} R\|_2, \quad \forall D \in \mathcal{D}_n. \tag{8.11}$$

When $D = I$, $\rho_D = \sqrt{2}$, and this last with (8.5) and (8.10) gives

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \sqrt{2} \eta_1 \mathrm{cond}_2(R) \epsilon + O(\epsilon^2),$$

which is exactly our 2– and F–norm analogy (5.1) of Zha's [18, Theorem 2.1] bound. Clearly by choosing D carefully we will usually get a better result than this. Now we define (with ρ_D as in (3.4))

$$\kappa'_R(A) \equiv \inf_{D \in \mathcal{D}_n} \kappa(R, D), \tag{8.12}$$

$$\kappa(R, D) \equiv \rho_D \| \, |R| \cdot |R^{-1}| D \|_2 \|D^{-1} R\|_2 / \|R\|_2. \tag{8.13}$$

This gives bounds on the true condition number $\kappa_R(A)$ in (8.6) and (8.7) (with $\varphi(A)$ in (5.3)) of

$$1 \leq \kappa_R(A) \leq \kappa'_R(A) \leq \kappa(R, I) = \sqrt{2}\,\mathrm{cond}_2(R) = \varphi(A). \tag{8.14}$$

The above analysis, with (8.5), leads to the following theorem.

Theorem 8.1 *Assume that the conditions and assumptions of Theorem 4.1 hold, then $A+\Delta A = A+\epsilon G$ has the unique QR factorization*

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

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \eta_1 \kappa_R(A)\epsilon + O(\epsilon^2), \tag{8.15}$$

$$1 \leq \kappa_R(A) \leq \kappa'_R(A) \leq \varphi(A), \tag{8.16}$$

with η_1, $\kappa_R(A)$, $\kappa'_R(A)$ and $\varphi(A)$ defined by (3.2), (8.6), (8.12) *with* (8.13), *and* (5.3). □

When we use the standard column pivoting strategy in $AP = QR$, where P is the permutation matrix, we get a very nice result. Here the elements of the resulting R satisfy, for $i = 1, \ldots, n$,

$$r_{ii}^2 \geq \sum_{k=i}^{j} r_{kj}^2, \qquad \text{for each} \quad j = i, \ldots, n.$$

It follows that $r_{11} \geq r_{22} \geq \ldots \geq r_{nn}$, and $r_{ii} \geq |r_{ij}|$. Take $D = \mathrm{diag}(r_{ii})$, then $\rho_D \leq \sqrt{2}$, and $\bar{R} \equiv D^{-1}R$ is unit upper triangular with $1 = \bar{r}_{ii} \geq |\bar{r}_{ij}|$ for all $j \geq i$, and it is easy to show that for $j > i$, $|(\bar{R}^{-1})_{ij}| \leq 2^{j-i-1}$ (see, for example, [7, Lemma 8.6]). Thus from (8.13) we have

$$\kappa'_R(AP) \leq \kappa(R, D) \leq \sqrt{2n}\, \|\bar{R}^{-1}\|_F \|\bar{R}\|_F \leq \sqrt{n^2(1+n)(4^n+6n-1)}/3.$$

We see that when the standard pivoting strategy is used, the sensitivity of R is bounded for any n. We summarize this as a theorem.

Theorem 8.2 *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $AP = QR$ when the standard column pivoting strategy is used. Then in (8.16)*

$$1 \leq \kappa_R(AP) \leq \kappa'_R(AP) \leq \sqrt{n^2(1+n)(4^n+6n-1)}/3. \quad \square \tag{8.17}$$

In contrast, $\varphi(A)$ in (5.3) can be arbitrarily large for fixed n, even when standard column pivoting is used. For example, $A = R = \begin{bmatrix} 1 & 1/2 \\ 0 & \epsilon \end{bmatrix}$, with very small $\epsilon > 0$. It is easy to see $\varphi(A) = O(\frac{1}{\epsilon})$. So the bounds (5.1) can severely overestimate the sensitivity of R.

Clearly $\kappa'_R(A)$ in (8.12) is a candidate for estimating the condition number $\kappa_R(A)$ of R in the QR factorization. We now give some insight as to why R in the QR factorization is often less sensitive than the

earlier condition estimates $\varphi(A) = \sqrt{2}\,\mathrm{cond}_2(R)$ suggested. We know that $\mathrm{cond}_2(R) = \| \, |R| \cdot |R^{-1}| \, \|_2$ only takes out the effect of bad column scaling in R, whereas according to [17, Thm. 3.3], we have

$$\| \, |R| \cdot |R^{-1}| D_r \|_2 \| D_r^{-1} R \|_2 \le \sqrt{n} \inf_{D \in \mathcal{D}_n} \| \, |R| \cdot |R^{-1}| D \|_2 \| D^{-1} R \|_2,$$

where $D_r = \mathrm{diag}(\|(R)_{i,:}\|_2)$. Thus $\| \, |R| \cdot |R^{-1}| D_r \|_2 \| D_r^{-1} R \|_2$ takes out the effect of bad column *and* row scaling in R. Thus for an R with poor row scaling, if ρ_D is not large, then $\kappa'_R(A)$ will be much smaller than $\varphi(A)$.

Now we return to the example in Section 5. By a similar argument to that for (5.8), we have for the MATLAB QR factorization

$$\frac{\|R_c - R\|_F}{\|R\|_2} \le \gamma''_{m,n} \kappa_R(A) u + O(u^2). \tag{8.18}$$

For the matrices A_1 and A_2, we take row scaling $D_1 = \mathrm{diag}(2, 10^{-10})$ and $D_2 = \mathrm{diag}(2, \sqrt{2} \times 10^{-10})$, respectively, then with (8.12)–(8.14)

$$\kappa_R(A_1) \le \kappa(R_1, D_1) \approx 2.3, \quad \kappa_R(A_2) \le \kappa(R_2, D_2) \approx 2.3.$$

The analysis by Zha [18] leads to condition numbers of about $\varphi(A_1) \approx 4.0 \times 10^{10}$, $\varphi(A_2) \approx 2.8 \times 10^{10}$, see (5.1), (5.2). Obviously the new error bound (8.18) gives good estimates for both e_{R_1} and e_{R_2} in (5.5), in contrast to [18].

9 Condition estimation and numerical experiments

In Section 7 we gave first-order perturbation bounds for the change in Q lying in $\mathcal{R}(Q)$, and the change in Q lying in the orthogonal complement of $\mathcal{R}(Q)$, and defined $\kappa_Q(A) = \sqrt{2} \, \| \, |R_{n-1}| \cdot |R_{n-1}^{-1}| \, \|_2$ as the (asymptotic) condition number for the former.

In Section 8.2 we presented a perturbation bound for the R factor, and defined $\kappa_R(A)$ as a corresponding condition number. In practice we would like to choose D such that

$$\kappa(R, D) \equiv \rho_D \| \, |R| \cdot |R^{-1}| D \|_2 \| D^{-1} R \|_2 / \|R\|_2$$

is a good approximation to the infimum $\kappa'_R(A)$ of $\kappa(R, D)$, where we know $\kappa_R(A) \le \kappa'_R(A)$.

It seems there is no obvious way to cheaply find the best D. Our numerical experiments indicate if we choose D to equilibrate the rows of $D^{-1}R$, i.e., $D = D_r \equiv \mathrm{diag}(\|(R)_{i,:}\|_2)$, we usually obtain a good estimate of $\kappa_R(A)$. But sometimes it can lead to large ρ_D and result

in an overestimate. However in such situations we found the previous estimate $\varphi(A) = \kappa(R, I)$ gave a good approximation. So we may use $\min\{\kappa(R, D_r), \kappa(R, I)\}$ as an estimate of the condition number $\kappa_R(A)$. The remaining problem is in practice how to cheaply estimate $\| |R| \cdot |R^{-1}|D\|_2$ with $D = D_r$ or $D = I$. Since $\| |R| \cdot |R^{-1}|D\|_2$ and $\| |R| \cdot |R^{-1}|D\|_1$ differ by at most a factor of $\sqrt{n}$, we can estimate the latter instead of the former. Following van der Sluis [17, Thm. 2.6],

$$\| |R| \cdot |R^{-1}|D\|_1 = \|RD_c^{-1}\|_1 \|D_c R^{-1} D\|_1, \quad D_c \equiv \mathrm{diag}(\|(R)_{:,j}\|_1). \tag{9.1}$$

Notice $\|D_c R^{-1} D\|_1$ can be estimated in $O(n^2)$ flops (see for example Higham [7, Chap. 14]). Thus both $\kappa(R, D_r)$ and $\kappa(R, I)$ can be estimated in a total of $O(n^2)$ flops.

Our numerical experiments also suggest that another option for D may give a good approximation. With D_c as given in (9.1), choose $D = \mathrm{diag}(\delta_j)$ to approximately equilibrate the columns of $D_c R^{-1}$ in (9.1) while keeping $\rho_D \le \sqrt{2}$. To do this, take $\delta_1 = 1/\|(D_c R^{-1})_{:,1}\|_2$, then for $j = 2, \ldots, n$ take $\delta_j = 1/\|(D_c R^{-1})_{:,j}\|_2$ if $\|(D_c R^{-1})_{:,j}\|_2 \ge \|(D_c R^{-1})_{:,j-1}\|_2$; otherwise $\delta_j = \delta_{j-1}$. For this D we write

$$D_e \equiv D \tag{9.2}$$

to indicate this choice in the tables ("e" denotes "equilibration"). The problem is that to our knowledge there is no known way to estimate the 2-norm of each column of the inverse of an upper triangular matrix in $O(n^2)$ flops yet. This is an interesting problem in itself.

We showed with (5.4) that it is easy to construct artificial examples where the previous bounds (5.1)–(5.3) are exceptionally poor, while the new ones here are very good. The examples we now give are intended to develop a feeling for more usual behaviour.

We give three sets of examples, each with and without pivoting, to show how good the new condition numbers are compared with the previous ones, how well the new error bounds match the actual errors in the computed factors, and how well the condition estimates approximate these new condition numbers. In all these experiments we computed the condition numbers and condition estimates. For the first and second sets of examples, see Tables 9.1–9.8, we also computed (c.f. (7.12) and (8.18), with $P = I$ for no pivoting):

$$\begin{aligned} e_Q &= \|Q_c - Q\|_F, \quad & b_Q &= \kappa_Q(AP)u, \\ e_R &= \frac{\|R_c - R\|_F}{\|R\|_2}, \quad & b_R &= \kappa_R(AP)u. \end{aligned} \tag{9.3}$$

Here Q_c and R_c are the computed QR factors of A in single precision by means of a MATLAB program `chop.m` provided by Higham [8],

Q and R are the computed QR factors of A by MATLAB (in double precision), and $u \approx 5.96 \times 10^{-8}$ is the single precision unit roundoff. So e_Q and e_R are very good approximations to the norms of the actual errors in the computed Q_c and R_c, and b_Q and b_R are approximations to the error bounds (note the matrices in our examples are square and here we ignore the constant $\gamma'_{n,n}$ in (7.12) and $\gamma''_{m,n}$ in (8.18)).

Each matrix in the first set has 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`, so these A are graded. The results for $n = 20$, $d_1, d_2 \in \{0.8, 1, 2\}$, and the same matrix B, are shown in Tables 9.1–9.4. The results are given for Q, then for R, first without, then with standard pivoting.

Each matrix in the second set has the form $A = Q(D_1 U D_2)$, where U is the upper triangular part of a random matrix produced also by `randn`, and D_1 and D_2 are the same as in the first set of examples. Q is a random orthogonal matrix produced by `qmult.m` in [8]. This gives the less likely case of graded R when no pivoting is used. The results for $n = 20$, $d_1, d_2 \in \{0.8, 1, 2\}$, the same matrix U, but different Q for each case, are shown in Tables 9.5–9.8.

The third set involves $n \times n$ Kahan matrices (see [10]):

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

where $c = \cos(\theta)$, $s = \sin(\theta)$. The results for $n = 5, 10, 15, 20, 25$ with $\theta = \pi/8$ are shown in Table 9.9 without pivoting, and in Table 9.10 for AP where P is a permutation such that the first column moves to the last column position and the remaining columns are moved left one position — this permutation is adopted in the rank-revealing QR factorization for Kahan matrices, see for example Hong and Pan [9].

We now comment on the tables. Ideally we have $AP = QR$, with $P = I$ for no pivoting. Remember that $\varphi(AP) = \kappa(R, I)$, see (8.14).

1. We used square matrices in the examples, so $QQ^T \Delta Q = \Delta Q$. The new measure of sensitivity $\kappa_Q(AP)$ of Q, see (7.8), (7.10), is never larger than the old $\varphi(AP) = \kappa(R, I)$, see (5.2), (5.3). In Tables 9.7, 9.10 the new measure gives a significant improvement over the old, and even in the other tables the difference makes it worthwhile using the new measure.
2. For the sensitivity of R, the old measure $\kappa(R, I)$ compares poorly with the new measure $\kappa_R(A)$ and estimates $\kappa(R, D_r)$ and $\kappa(R, D_e)$ in all except Table 9.6, where it is clearly superior to $\kappa(R, D_r)$ in

the last three cases (it is also marginally better in the fifth and sixth cases). In all but these three anomalous cases $\kappa(R, D_r)$ gave an excellent estimate of $\kappa_R(AP)$, while in these anomalous cases $\kappa(R, I)$ gave a good estimate, supporting the suggestion of taking $\min\{\kappa(R, D_r), \kappa(R, I)\}$. In all the cases where $\kappa(R, I)$ is very large, the new measure $\kappa_R(AP)$ is significantly smaller. This suggests R in the QR factorization is not nearly as sensitive to perturbations of the form (1.1) as was previously thought, see [18].

3. When standard column pivoting is used $\kappa_R(AP)$ can be improved significantly, and in fact in Tables 9.4 and 9.8, $\kappa_R(AP)$ is $O(1)$. $\kappa_Q(AP)$ can also be improved significantly, compare Table 9.5 with 9.7. But the old measure for both, $\kappa(R, I)$, does not change much.
4. The error bounds b_Q and b_R match the corresponding actual errors e_Q and e_R very well, see (9.3). The numerical results show that standard pivoting can significantly improve the accuracy of both the QR factors, compare Table 9.5 with 9.7, and 9.6 with 9.8.
5. The Kahan matrix is theoretically already in standard column pivoting form, and $\kappa_R(A)$ grows significantly as n increases, though not as fast as the bound in (8.17). But rank-revealing pivoting brought $\kappa_R(AP)$ back down to $O(1)$.

Table 9.1. $A = D_1 B D_2$ without pivoting, sensitivity of Q

d_1	d_2	$\kappa_Q(A)$	$\varphi(A) = \kappa(R, I)$	e_Q	b_Q
0.8	0.8	3.5e+02	3.9e+03	1.3e-06	2.1e-05
0.8	1	3.5e+02	3.9e+03	1.5e-06	2.1e-05
0.8	2	3.5e+02	3.9e+03	1.5e-06	2.1e-05
1	0.8	3.5e+01	3.5e+02	1.1e-06	2.1e-06
1	1	3.5e+01	3.5e+02	1.1e-06	2.1e-06
1	2	3.5e+01	3.5e+02	1.1e-06	2.1e-06
2	0.8	1.1e+06	7.4e+06	1.7e-02	6.6e-02
2	1	1.1e+06	7.4e+06	2.8e-02	6.6e-02
2	2	1.1e+06	7.4e+06	2.8e-02	6.6e-02

10 Summary and future work

Componentwise perturbation analyses have been given for the QR factorization of a matrix A with the form of perturbations we could expect from the equivalent backward error in A resulting from numerically stable computations. For the Q factor we derived the condition numbers for that part of ΔQ in $\mathcal{R}(A)$, and that part in $\mathcal{R}^{\perp}(A)$. Both

Table 9.2. $A = D_1BD_2$ without pivoting, sensitivity of R

d_1	d_2	$\kappa_R(A)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\kappa(R, I)$	e_R	b_R
0.8	0.8	1.9e+00	5.8e+00	7.7e+00	3.9e+03	2.0e-07	1.1e-07
0.8	1	1.2e+01	5.0e+01	4.2e+01	3.9e+03	2.6e-07	7.1e-07
0.8	2	1.3e+01	1.9e+02	6.2e+01	3.9e+03	2.0e-07	7.7e-07
1	0.8	1.1e+00	2.8e+00	3.8e+00	3.5e+02	1.3e-07	6.6e-08
1	1	9.2e+00	2.8e+01	3.1e+01	3.5e+02	4.1e-07	5.5e-07
1	2	1.1e+01	2.2e+02	4.8e+01	3.5e+02	3.2e-07	6.3e-07
2	0.8	2.0e+00	1.2e+01	1.6e+01	7.4e+06	6.2e-08	1.2e-07
2	1	1.7e+01	2.1e+02	9.0e+01	7.4e+06	1.2e-07	1.0e-06
2	2	5.2e+01	1.1e+03	4.1e+02	7.4e+06	4.2e-07	3.1e-06

Table 9.3. $A = D_1BD_2$ with standard pivoting, sensitivity of Q

d_1	d_2	$\kappa_Q(AP)$	$\varphi(AP) = \kappa(R, I)$	e_Q	b_Q
0.8	0.8	3.3e+02	3.7e+03	1.0e-06	2.0e-05
0.8	1	3.5e+02	3.8e+03	1.2e-06	2.1e-05
0.8	2	5.0e+02	3.7e+03	1.9e-06	3.0e-05
1	0.8	3.4e+01	3.5e+02	9.4e-07	2.0e-06
1	1	4.2e+01	3.5e+02	1.0e-06	2.5e-06
1	2	5.5e+01	3.4e+02	1.4e-06	3.3e-06
2	0.8	1.1e+06	7.4e+06	1.6e-02	6.6e-02
2	1	1.4e+06	7.5e+06	2.2e-02	8.2e-02
2	2	1.5e+06	7.4e+06	2.0e-02	9.1e-02

Table 9.4. $A = D_1BD_2$ with standard pivoting, sensitivity of R

d_1	d_2	$\kappa_R(AP)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\kappa(R, I)$	e_R	b_R
0.8	0.8	1.6e+00	4.2e+00	6.5e+00	3.7e+03	2.1e-07	9.6e-08
0.8	1	5.6e+00	1.5e+01	1.7e+01	3.8e+03	2.4e-07	3.4e-07
0.8	2	1.0e+00	1.9e+00	5.8e+01	3.7e+03	1.1e-07	6.0e-08
1	0.8	1.1e+00	2.4e+00	3.7e+00	3.5e+02	9.2e-08	6.5e-08
1	1	6.1e+00	1.5e+01	1.7e+01	3.5e+02	3.7e-07	3.6e-07
1	2	1.0e+00	1.7e+00	4.3e+00	3.4e+02	7.2e-08	6.0e-08
2	0.8	1.3e+00	5.2e+00	1.0e+01	7.4e+06	6.1e-08	7.5e-08
2	1	3.8e+00	2.1e+01	2.4e+01	7.5e+06	7.0e-08	2.3e-07
2	2	1.0e+00	2.6e+00	7.6e+02	7.4e+06	8.5e-08	6.2e-08

can be estimated in $O(n^2)$ flops. For the R factor, we first derived the new condition number, and then suggested practical condition estimators. These provide estimates in $O(n^2)$ flops. The analyses more accurately reflect the sensitivity of the problem than previous such results. Both the analysis and numerical results show that standard column pivoting can significantly decrease the sensitivity of R, and Q as well, and so give more accurate QR factors. We also found a

Table 9.5. $A = Q(D_1 U D_2)$ without pivoting, sensitivity of Q

d_1	d_2	$\kappa_Q(A)$	$\varphi(A) = \kappa(R, I)$	e_Q	b_Q
0.8	0.8	2.3e+07	2.3e+07	4.6e-02	1.4e+00
0.8	1	2.3e+07	2.3e+07	1.4e-01	1.4e+00
0.8	2	2.5e+07	2.5e+07	2.8e+00	1.5e+00
1	0.8	4.8e+05	4.8e+05	8.2e-03	2.9e-02
1	1	4.8e+05	4.8e+05	5.0e-04	2.9e-02
1	2	4.8e+05	4.8e+05	4.7e-03	2.9e-02
2	0.8	7.3e+01	7.3e+01	1.9e-06	4.3e-06
2	1	7.3e+01	7.3e+01	4.1e-06	4.3e-06
2	2	7.3e+01	7.3e+01	3.2e-06	4.3e-06

Table 9.6. $A = Q(D_1 U D_2)$ without pivoting, sensitivity of R

d_1	d_2	$\kappa_R(A)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\kappa(R, I)$	e_R	b_R
0.8	0.8	3.2e+03	1.5e+04	1.6e+04	2.3e+07	1.2e-05	1.9e-04
0.8	1	8.0e+04	5.3e+05	4.0e+05	2.3e+07	4.7e-04	4.8e-03
0.8	2	2.1e+05	1.4e+06	1.0e+06	2.5e+07	2.3e-02	1.2e-02
1	0.8	4.2e+03	2.6e+04	2.0e+04	4.8e+05	7.1e-05	2.5e-04
1	1	9.9e+04	6.5e+05	4.7e+05	4.8e+05	8.9e-05	5.9e-03
1	2	1.6e+05	1.8e+06	7.3e+05	4.8e+05	1.5e-03	9.2e-03
2	0.8	5.2e+01	1.9e+04	1.0e+02	7.3e+01	5.6e-07	3.1e-06
2	1	5.2e+01	4.6e+05	1.0e+02	7.3e+01	3.4e-07	3.1e-06
2	2	5.2e+01	2.8e+06	1.0e+02	7.3e+01	9.4e-07	3.1e-06

Table 9.7. $A = Q(D_1 U D_2)$ with standard pivoting, sensitivity of Q

d_1	d_2	$\kappa_Q(AP)$	$\varphi(AP) = \kappa(R, I)$	e_Q	b_Q
0.8	0.8	9.2e+04	2.5e+07	2.1e-04	5.5e-03
0.8	1	9.5e+04	2.5e+07	3.0e-04	5.6e-03
0.8	2	9.5e+04	2.6e+07	2.8e-04	5.7e-03
1	0.8	2.9e+03	5.0e+05	2.0e-05	1.7e-04
1	1	2.9e+03	5.1e+05	1.8e-05	1.7e-04
1	2	3.0e+03	5.2e+05	3.8e-05	1.8e-04
2	0.8	3.2e+01	7.5e+01	2.3e-06	1.9e-06
2	1	3.2e+01	7.6e+01	1.6e-06	1.9e-06
2	2	3.1e+01	8.0e+01	1.4e-06	1.9e-06

rank-revealing pivoting strategy could improve the condition of R and Q.

The Q and R factors have practical meanings in several applications, see for example [3]. Certainly in these cases it can be important to know how accurate, or sensitive, these factors are. One contribution of this paper has been to provide the theory whereby the sensitivity of *computed* QR factors may be obtained. We propose that standard

Table 9.8. $A = Q(D_1 U D_2)$ with standard pivoting, sensitivity of R

d_1	d_2	$\kappa_R(AP)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\kappa(R, I)$	e_R	b_R
0.8	0.8	2.3e+00	9.2e+00	1.3e+01	2.5e+07	1.0e-07	1.4e-07
0.8	1	5.4e+00	1.9e+01	1.9e+01	2.5e+07	1.1e-07	3.2e-07
0.8	2	1.0e+00	2.0e+00	6.1e+00	2.6e+07	3.2e-08	6.0e-08
1	0.8	3.5e+00	1.4e+01	1.6e+01	5.0e+05	1.2e-07	2.1e-07
1	1	4.6e+00	1.3e+01	1.4e+01	5.1e+05	2.0e-07	2.8e-07
1	2	1.0e+00	1.8e+00	3.8e+00	5.2e+05	3.8e-08	6.0e-08
2	0.8	1.0e+00	2.0e+00	3.1e+00	7.5e+01	9.2e-08	6.0e-08
2	1	1.0e+00	2.0e+00	3.1e+00	7.6e+01	3.1e-08	6.0e-08
2	2	1.0e+00	1.7e+00	3.6e+00	8.0e+01	1.6e-07	6.0e-08

Table 9.9. $n \times n$ Kahan matrices without pivoting

n	$\kappa_Q(A)$	$\kappa_R(A)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\varphi(A) = \kappa(R, I)$
5	1.8e+02	6.5e+00	1.4e+01	1.5e+01	9.0e+02
10	5.8e+05	1.2e+02	3.5e+02	4.0e+02	2.9e+06
15	1.9e+09	2.5e+03	9.5e+03	1.1e+04	9.3e+09
20	6.0e+12	5.8e+04	2.6e+05	2.9e+05	3.0e+13
25	1.9e+16	1.4e+06	7.0e+06	7.6e+06	9.6e+16

Table 9.10. $n \times n$ Kahan matrices with rank-revealing pivoting

n	$\kappa_Q(AP)$	$\kappa_R(AP)$	$\kappa(R, D_r)$	$\kappa(R, D_e)$	$\varphi(AP) = \kappa(R, I)$
5	2.8e+01	1.8e+00	4.9e+00	5.0e+00	8.9e+02
10	3.5e+03	2.3e+00	1.1e+01	1.1e+01	2.8e+06
15	4.2e+05	2.5e+00	1.8e+01	1.8e+01	9.1e+09
20	5.1e+07	2.6e+00	2.6e+01	2.4e+01	2.9e+13
25	6.3e+09	2.7e+00	3.3e+01	3.1e+01	9.4e+16

numerical linear algebra software packages include the option of estimating the condition estimates given in this paper for the QR factors. This can be done by using standard norm estimators that are already available in the packages. These estimates can then be used to supply measures of the accuracy of the computed factors. As pointed out in section 9, this estimation can be done in $O(n^2)$ flops. Compared with the $O(mn^2)$ flops required for the computation of the QR factorization itself by the Householder or givens transformations, this extra computation does not cost much. Such information will be very helpful to users interested in either the accuracy of the computed QR factors, or their sensitivity to other perturbations of the form (1.1).

In the future we would like to investigate how well the suggested practical condition estimates approximate the corresponding condition numbers. Also we would like to apply the approaches here to the

Householder QR factorization with complete pivoting (see Cox and Higham [5]).

Acknowledgements The referees' suggestions improved the presentation greatly.

References

1. X.-W. Chang, Ph.D. Thesis, School of Computer Science, McGill University, Montreal, Quebec, Canada, February 1997.
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, *Sensitivity Analyses for Factorizations of Sparse or Structured Matrices*, Linear Algebra and Appl., **284** (1998), pp. 53-71.
4. X.-W. Chang, C. C. Paige and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., **18** (1997), pp. 775–791.
5. A. J. Cox and N. J. Higham, *Stability of Householder QR factorization for weighted least squares problem*, In D. F. Griffiths, D. J. Higham, and G. A. Watson, editors, Numerical Analysis 1997, Proceeding of the 17th Dundee Biennial Conference, Volume 380 of Pitman Research Notes in Mathematics, pp. 57–73. Addison Wesley Longman, Harlow, Essex, UK, 1998.
6. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (The Johns Hopkins University press, Baltimore, MD, 1996).
7. N. J. Higham, *Accuracy and Stability of Numerical Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996).
8. N. J. Higham, *The Test Matrix Toolbox for Matlab, version 3.0*, Numerical Analysis Report No. 265, University of Manchester, Manchester, England, 1995.
9. Y. P. Hong and C.-T. Pan, *Rank-revealing QR factorizations and the singular value decomposition*, Math. Comp., **58** (1992), pp. 213-232.
10. W. Kahan, *Numerical linear algebra*, Can. Math. Bull., **9** (1966), pp. 757–801.
11. P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed. (Academic Press, New York, 1984).
12. G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., **14** (1977), pp. 509–518.
13. G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., **14** (1993), pp. 1141–1146.
14. J.-G. Sun, *Perturbation bounds for the Cholesky and QR factorization*, BIT, **31** (1991), pp. 341–352.
15. J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, **32** (1992), pp. 702–714.
16. J.-G. Sun, *On perturbation bounds for the QR factorization*, Linear Algebra and Appl., **215** (1995), pp. 95–112.
17. A. van der Sluis, *Condition numbers and equilibration of matrices*, Numerische Mathematik, **14** (1969), pp. 14–23.
18. H. Zha, *A componentwise perturbation analysis of the QR decomposition*, SIAM J. Matrix Anal. Appl., **14** (1993), pp. 1124–1131.