Stability of Householder QR Factorization for Weighted Least Squares Problems*

Anthony J. Cox† Nicholas J. Higham‡

September 5, 1997

Abstract

For least squares problems in which the rows of the coefficient matrix vary widely in norm, Householder QR factorization (without pivoting) has unsatisfactory backward stability properties. Powell and Reid showed in 1969 that the use of both row and column pivoting leads to a desirable row-wise backward error result. We give a reworked backward error analysis in modern notation and prove two new results. First, sorting the rows by decreasing ∞-norm at the start of the factorization obviates the need for row pivoting. Second, row-wise backward stability is obtained for only one of the two possible choices of sign in the Householder vector.

Key words. Weighted least squares problem, Householder matrix, QR factorization, row pivoting, column pivoting, backward error analysis

AMS subject classifications. 65F20, 65G05

1 Introduction

In many applications of the linear least squares (LS) problem $\min \|b - Ax\|_2$, where $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, the rows of A have widely varying norms (with corresponding variation in the size of the elements of b), typically because in the underlying model some

*In D. F. Griffiths, D. J. Higham, and G. A. Watson, editors, *Numerical Analysis 1997, Proceedings of the 17th Dundee Biennial Conference*, volume 380 of *Pitman Research Notes in Mathematics*, pages 57–73. Addison Wesley Longman, Harlow, Essex, UK, 1998.

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (coxtonyj@ma.man.ac.uk, http://www.ma.man.ac.uk/~coxtonyj/).

‡Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/).

observations have been given greater weight than others. Such weighted LS problems also arise when the solution to a linearly constrained LS problem

$$\min\{\, \|b - Ax\|_2 : Bx = d,\, B \in \mathbb{R}^{p\times n},\, \mathrm{rank}(B) = p \,\}$$

is approximated by the solution to the unconstrained problem

$$\min \left\| \begin{bmatrix} A \\ \mu B \end{bmatrix} x - \begin{bmatrix} b \\ \mu d \end{bmatrix} \right\|_2, \tag{1.1}$$

for a suitably large value of the parameter $\mu > 0$ [16].

For full rank A, QR factorization provides a standard way to solve the LS problem, and the QR factorization is perhaps most often computed using Householder transformations (as is done in LINPACK and LAPACK, for example). The backward stability of Householder QR factorization can be summarized by the following result [9, Thm. 18.4]. We make use of the standard model of floating point arithmetic:

$$fl(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1+\delta), \qquad |\delta| \le u, \quad \mathrm{op} = +, -, *, /, \tag{1.2}$$

where u is the unit roundoff. We introduce the constant

$$\gamma_k = \frac{ku}{1-ku}.$$

It also convenient to define the quantity

$$\tilde{\gamma}_k = \frac{cku}{1-cku},$$

in which c denotes a small integer constant whose exact value is unimportant. Thus we can write, for example, $3\tilde{\gamma}_k = \tilde{\gamma}_k$, and $m\tilde{\gamma}_n = n\tilde{\gamma}_m = \tilde{\gamma}_{mn}$. Absolute values and inequalities are interpreted componentwise.

Theorem 1.1 *Let $\widehat{R} \in \mathbb{R}^{m\times n}$ be the computed upper trapezoidal QR factor of $A \in \mathbb{R}^{m\times n}$ $(m \ge n)$ obtained via the Householder QR algorithm. Then there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

$$A + \Delta A = Q\widehat{R},$$

where $\|\Delta A\|_F \le n\tilde{\gamma}_m \|A\|_F$ and $|\Delta A| \le mn\tilde{\gamma}_m G|A|$, with $\|G\|_F = 1$. The matrix Q is given explicitly as $Q = P_1 P_2 \dots P_n$, where P_k is the Householder matrix that corresponds to the exact application of the kth step of the algorithm to the computed matrix produced after $k-1$ steps.

Theorem 1.1 shows that Householder QR factorization is normwise backward stable, and the componentwise bound shows that each column of the backward error matrix ΔA is nicely bounded relative to the corresponding column of A, that is, the computed factorization is columnwise backward stable. However, for weighted LS problems we would like the *row-wise backward error*

$$\max_i \frac{\|\Delta A(i,:)\|_2}{\|A(i,:)\|_2}$$

to be of order u. The following example of Powell and Reid [14] shows that this quantity can be large for Householder QR factorization. Consider the matrix

$$A = \begin{bmatrix} 0 & 2 & 1 \\ \lambda & \lambda & 0 \\ \lambda & 0 & \lambda \\ 0 & 1 & 1 \end{bmatrix}, \tag{1.3}$$

where $\lambda \gg 1$ is a parameter. The first step of the factorization produces the matrix

$$\begin{bmatrix} -\sqrt{2}\lambda & -\frac{\lambda}{\sqrt{2}} & \frac{-\lambda}{\sqrt{2}} \\ 0 & \frac{\lambda}{2} - \sqrt{2} & -\frac{\lambda}{2} - \frac{1}{\sqrt{2}} \\ 0 & -\frac{\lambda}{2} - \sqrt{2} & \frac{\lambda}{2} - \frac{1}{\sqrt{2}} \\ 0 & 1 & 1 \end{bmatrix}.$$

If $\lambda > 2\sqrt{2}u^{-1}$ then in the computed second and third rows the constants $\sqrt{2}$ and $1/\sqrt{2}$ will be lost. This loss can be shown to be equivalent to zeroing the first row of A and then carrying out exact computation, which corresponds to a row-wise backward error of order 1. The conclusion is that Householder QR factorization can be very unsatisfactory for weighted LS problems.

In this example there is a simple way to avoid the loss of row-wise stability, namely to bring an element of maximal absolute value into the pivot position by interchanging the first two rows of A:

$$\begin{bmatrix} \lambda & \lambda & 0 \\ 0 & 2 & 1 \\ \lambda & 0 & \lambda \\ 0 & 1 & 1 \end{bmatrix}.$$

Now, one step of the factorization yields the matrix

$$\begin{bmatrix} -\sqrt{2}\lambda & -\frac{\lambda}{\sqrt{2}} & \frac{-\lambda}{\sqrt{2}} \\ 0 & 2 & 1 \\ 0 & -\frac{\lambda}{\sqrt{2}} & \frac{\lambda}{\sqrt{2}} \\ 0 & 1 & 1 \end{bmatrix},$$

Pivoting:	None	Row	Col.	Row and col.
Normwise (η)	2.06e-16	3.47e-16	1.19e-16	3.41e-16
Row-wise (η_R)	1.27e-4	1.75e-8	1.27e-4	4.53e-16
$\rho_{m,n}$	1.41e+12	2.53e+7	1.41e+12	2.83e+0

Table 1.1: Backward errors for QR factorization with no pivoting, row pivoting and column pivoting on matrix (1.4).

and the previous loss of information has been avoided. Note that row interchanges before or during Householder QR factorization have no mathematical effect on the result, because they can be absorbed into the Q factor and the QR factorization is essentially unique. The effect of row interchanges is to change the intermediate numbers that arise during the factorization, and hence to alter the effects of rounding errors.

Row interchanges alone are not enough to make Householder QR factorization row-wise backward stable, as an example of Van Loan [16] shows. Let

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ \mu & \mu & \mu \\ \mu & \mu & -\mu \end{bmatrix}. \tag{1.4}$$

We applied Householder QR factorization in MATLAB with $\mu = 10^{12}$, using no pivoting, and combinations of row pivoting (defined later) and the standard column pivoting strategy introduced by Golub [7]. The unit roundoff $u \approx 1.1 \times 10^{-16}$. Table 1.1 reports the backward errors

$$\eta = \frac{\|A\Pi - \widehat{Q}\widehat{R}\|_2}{\|A\|_2}, \qquad \eta_R = \max_i \frac{\|(A\Pi - \widehat{Q}\widehat{R})(i,:)\|_2}{\|A(i,:)\|_2}.$$

In these expressions, $\widehat{Q}$ denotes the computed product of the Householder transformations and Π is the permutation matrix produced by column pivoting, or the identity if column pivoting was not used. The quantity $\rho_{m,n}$ in Table 1.1 is a row-wise growth factor that is explained at the end of Section 2. Normwise stability prevails in each case. But with no pivoting, or row or column pivoting alone, the computation is not row-wise stable. The combination of row and column pivoting, however, does yield row-wise stability.

The need for row and column pivoting in Householder QR factorization was established nearly two decades ago by Powell and Reid and soon became well known, being

reported in Lawson and Hanson's 1974 book for example [11, pp. 103–106, 149]. However, Powell and Reid's analysis is relatively inaccessible: the analysis is outlined in the conference proceedings paper [14] and the full details are given in the technical report [13]. Powell and Reid assumed that inner products are accumulated in extra precision and gave a first-order analysis. They also assumed, without comment, a particular choice of sign in the construction of the Householder matrix at each stage.

The contributions of this work are as follows.

- We give a reworked version of Powell and Reid's backward error analysis of QR factorization with column pivoting. Our analysis is shorter and easier to read because we use matrix and vector notation exclusively and are not concerned with obtaining explicit constants. Unlike in [13], [14], ours is not a first-order analysis and we do not assume that inner products are accumulated in extra precision.

- Björck [5, p. 169] conjectures that "there is no need to perform row pivoting in Householder QR, provided that the rows are sorted after decreasing row norm before the factorization". We prove that this conjecture is true, by showing that it leads to the same row-wise backward error bound as for row pivoting. A practical advantage of row sorting over row pivoting is that it enables us to use standard software for Householder QR factorization with column pivoting, such as that in LAPACK.

- We show the somewhat surprising result that with the alternative choice of sign in the vector defining the Householder matrix, row-wise backward stability is lost, and we explain the reasons for this behaviour.

Related investigations for QR factorizations computed using Givens transformations are given by Anda and Park [1] and Barlow [3]. Other work concerned with error analysis for the weighted least squares problem includes that of Barlow and Handy [4], Gulliksson [8], and Hough and Vavasis [10].

2 Error Analysis for the Factorization

In this section we derive a row-oriented backward error bound for Householder QR factorization with column pivoting.

First, we recall how a Householder matrix is constructed in Householder QR factorization. Let $A = A^{(1)} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ and let $a_j^{(k)}$ denote the jth column of $A^{(k)}$, the reduced matrix at the start of the kth stage of the reduction to trapezoidal form. We

form the Householder matrix

$$P_k = I - \beta_k v_k v_k^T \in \mathbb{R}^{m\times m}, \qquad \beta = \frac{2}{v_k^T v_k},$$

where $v_k(1\colon k-1) = 0$ and

$$v_k(k\colon m) = a_k^{(k)}(k\colon m) - \sigma_k e_1, \tag{2.1}$$

where $e_1 \in \mathbb{R}^{m-k+1}$ is the first unit vector and

$$\sigma_k = \pm\|a_k^{(k)}(k\colon m)\|_2.$$

This Householder matrix P_k has the property that $a_k^{(k+1)} = P_k a_k^{(k)}$ satisfies $a_k^{(k+1)}(k\colon m) = \sigma_k e_1$.

For the error analysis in this section and the next we assume that

$$\sigma_k = \ -\operatorname{sign}(a_{kk}^{(k)})\|a_k^{(k)}(k\colon m)\|_2, \tag{2.2}$$

which is the choice of sign recommended in most textbooks and the choice used by the QR factorization routines in LINPACK [6] and LAPACK [2]. In Section 5 we consider the other choice of sign.

In QR factorization with column pivoting, columns are exchanged at the start of the kth stage to ensure that

$$|\sigma_k| = \|a_k^{(k)}(k\colon m)\|_2 = \max_{j\ge k}\|a_j^{(k)}(k\colon m)\|_2. \tag{2.3}$$

It follows from (2.3) that

$$|\sigma_1| \ge |\sigma_2| \ge \cdots \ge |\sigma_n|. \tag{2.4}$$

To simplify the notation we assume, without loss of generality, that A is "pre-pivoted", that is, that no column interchanges are required in order to satisfy (2.3). The next result, from [14], is the key to obtaining a row-wise backward error bound.

Lemma 2.1 *Consider the application of the Householder matrix P_k to the vector $a_j^{(k)}$, where $j \ge k$:*

$$\begin{aligned} a_j^{(k+1)} &= P_k a_j^{(k)} = a_j^{(k)} - \beta_k v_k v_k^T a_j^{(k)} \\ &= a_j^{(k)} - \phi_j^{(k)} v_k. \end{aligned}$$

The scalar $\phi_j^{(k)} = \beta_k v_k^T a_j^{(k)} = \beta_k v_k(k\colon m)^T a_j^{(k)}(k\colon m)$ satisfies

$$|\phi_j^{(k)}| \le |\beta_k||v_k|^T|a_j^{(k)}| \le \sqrt{2}.$$

Proof. We have

$$\begin{aligned}|\phi_j^{(k)}| &\le |\beta_k||v_k|^T|a_j^{(k)}| \\ &\le |\beta_k|\|v_k\|_2\|a_j^{(k)}(k\colon m)\|_2 \\ &= \frac{2\|a_j^{(k)}(k\colon m)\|_2}{\|v_k\|_2}.\end{aligned}$$

For the choice of sign of σ_k in (2.2),

$$v_k^T v_k = |2\sigma_k(\sigma_k - a_{kk}^{(k)})| \ge 2\sigma_k^2. \tag{2.5}$$

Thus

$$|\phi_j^{(k)}| \le \frac{2\|a_j^{(k)}(k\colon m)\|_2}{\sqrt{2}|\sigma_k|} \le \sqrt{2},$$

using (2.3). □

Rounding errors in computing the quantities β and v that determine a Householder matrix are analyzed in [9, Lem. 18.1]. By absorbing the errors in β into the vector v we can assume that β is obtained exactly. Then the computed $\widehat{v}_k \in \mathbb{R}^m$ from the kth stage of the reduction satisfies

$$\widehat{v}_k = v_k + \Delta v_k, \qquad |\Delta v_k| \le \tilde{\gamma}_{m-k}|v_k|, \tag{2.6}$$

where

$$P_k = I - \beta_k v_k v_k^T$$

is *the Householder matrix corresponding to the exact application of the kth step of the algorithm to the computed matrix* $\widehat{A}^{(k)}$. We define

$$\alpha_i = \max_{j,k} |\widehat{a}_{ij}^{(k)}|, \qquad \Omega = \operatorname{diag}(\alpha_i). \tag{2.7}$$

Lemma 2.2 *Consider the computation of* $\widehat{a}_j^{(k+1)} = fl(\widehat{P}_k \widehat{a}_j^{(k)})$, *where* $\widehat{P}_k = I - \beta_k \widehat{v}_k \widehat{v}_k^T$ *and* $\widehat{v}_k$ *satisfies* (2.6). *We have*

$$\widehat{a}_j^{(k+1)} = P_k \widehat{a}_j^{(k)} + f_j^{(k)}, \tag{2.8}$$

where $f_j^{(k)}(1\colon k-1) = 0$ *and*

$$|f_j^{(k)}| \le u|\widehat{a}_j^{(k)}| + \tilde{\gamma}_{m-k}|v_k|. \tag{2.9}$$

Furthermore,

$$|f_j^{(k)}| \le \tilde{\gamma}_{m-k}\Omega e,$$

where $e = [1, 1, \ldots, 1]^T$.

Proof. It is straightforward to show using standard error analysis results (see the proof of Lemma 18.2 in [9]) that (2.8) holds with $f_j^{(k)}(1\colon k-1) = 0$ and

$$|f_j^{(k)}| \le u|\widehat{a}_j^{(k)}| + \tilde{\gamma}_{m-k}\big(|\beta_k||v_k|^T|\widehat{a}_j^{(k)}|\big)|v_k|.$$

But from Lemma 2.1 we have $|\beta_k||v_k|^T|\widehat{a}_j^{(k)}| \le \sqrt{2}$. For the last inequality, note that $|\widehat{a}_j^{(k)}| \le \Omega e$, trivially, and that, from (2.1),

$$|v_k|_i \le \begin{cases} \alpha_k + |\sigma_k| \le 2\alpha_k, & i = k, \\ \alpha_i, & i > k, \end{cases}$$

since $|\sigma_k| = |\widehat{a}_{kk}^{(k+1)}| \le \alpha_k$, so that

$$|v_k| \le 2\Omega e. \qquad \square \tag{2.10}$$

Now, using $P_k^2 = I$, we rewrite (2.8) as

$$\widehat{a}_j^{(k)} = P_k\widehat{a}_j^{(k+1)} - P_k f_j^{(k)}.$$

This gives

$$\begin{aligned} \widehat{a}_j^{(1)} &= P_1\widehat{a}_j^{(2)} - P_1 f_j^{(1)} \\ &= P_1(P_2\widehat{a}_j^{(3)} - P_2 f_j^{(2)}) - P_1 f_j^{(1)} \\ &\vdots \\ &= P_1P_2\dots P_j\widehat{a}_j^{(j+1)} - P_1P_2\dots P_j f_j^{(j)} - \dots - P_1 f_j^{(1)}. \end{aligned}$$

Since $a_j = \widehat{a}_j^{(1)}$ and $\widehat{a}_j^{(j+1)} = \widehat{a}_j^{(n+1)}$,

$$a_j = P_1P_2\dots P_j\widehat{a}_j^{(n+1)} - \sum_{i=1}^{j} P_1P_2\dots P_i f_j^{(i)}. \tag{2.11}$$

Consider a general term in the sum,

$$y_i = P_1P_2\dots P_i f_j^{(i)}, \qquad i \le j.$$

We have

$$\begin{aligned} y_i &= (I - \beta_1 v_1 v_1^T)P_2\dots P_i f_j^{(i)} = P_2\dots P_i f_j^{(i)} - \beta_1 v_1 v_1^T P_2\dots P_i f_j^{(i)} \\ &= (I - \beta_2 v_2 v_2^T)P_3\dots P_i f_j^{(i)} - \beta_1 v_1 v_1^T P_2\dots P_i f_j^{(i)} \\ &\vdots \\ &= f_j^{(i)} - \sum_{k=1}^{i} \beta_k v_k v_k^T P_{k+1}\dots P_i f_j^{(i)}. \end{aligned}$$

We use Lemma 2.2 to bound the first occurrence of $f_j^{(i)}$ in this equality by $\tilde{\gamma}_{m-i}\Omega e$ and the subsequent occurrences by $u|\widehat{a}_j^{(i)}| + \tilde{\gamma}_{m-i}|v_i|$. Writing

$$z_k = \beta_k v_k v_k^T P_{k+1} \dots P_i f_j^{(i)} = \frac{2v_k v_k^T}{v_k^T v_k} P_{k+1} \dots P_i f_j^{(i)}$$

and using (2.10), we have

$$|z_k| \le 4\Omega e \frac{\|f_j^{(i)}\|_2}{\|v_k\|_2}, \qquad k \le i.$$

Now, using $\|v_k\|_2 \ge \sqrt{2}|\sigma_k|$ from (2.5) and $\|v_i\|_2 \le 2|\sigma_i|$ from (2.1),

$$\begin{aligned}
\frac{\|f_j^{(i)}\|_2}{\|v_k\|_2} &\le u \frac{\|\widehat{a}_j^{(i)}(i\colon m)\|_2}{\|v_k\|_2} + \tilde{\gamma}_{m-i} \frac{\|v_i\|_2}{\|v_k\|_2} \\
&\le u \frac{|\sigma_i|}{\sqrt{2}|\sigma_k|} + \tilde{\gamma}_{m-i}\sqrt{2}\frac{|\sigma_i|}{|\sigma_k|} \\
&\le \frac{u}{\sqrt{2}} + \tilde{\gamma}_{m-i}\sqrt{2} = \tilde{\gamma}_{m-i}.
\end{aligned}$$

We conclude that

$$|y_i| \le \tilde{\gamma}_{m-i}\Omega e + 4i\Omega e\tilde{\gamma}_{m-i} = i\tilde{\gamma}_{m-i}\Omega e. \tag{2.12}$$

Thus

$$a_j = P_1 P_2 \dots P_j \widehat{a}_j^{(n+1)} + h_j, \tag{2.13}$$

where

$$|h_j| \le \sum_{i=1}^{j} i\tilde{\gamma}_{m-i}\Omega e = j^2\tilde{\gamma}_m \Omega e. \tag{2.14}$$

But

$$P_1 P_2 \dots P_j \widehat{a}_j^{(n+1)} = P_1 P_2 \dots P_n \widehat{a}_j^{(n+1)} =: Q\widehat{a}_j^{(n+1)} = Q\widehat{r}_j.$$

The conclusions of the analysis are summarized in the following theorem.

Theorem 2.3 *Let $\widehat{R} \in \mathbb{R}^{m\times n}$ be the computed upper trapezoidal QR factor of $A \in \mathbb{R}^{m\times n}$ $(m \ge n)$ obtained via the Householder QR algorithm with column pivoting. Then there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

$$(A + \Delta A)\Pi = Q\widehat{R},$$

where Π is a permutation matrix that describes the overall effect of the column interchanges and

$$|\Delta A| \le \tilde{\gamma}_m \Omega e e^T \operatorname{diag}(1, \dots, n)^2, \tag{2.15}$$

where Ω is defined in (2.7)*. The matrix Q is defined as in Theorem* 1.1*.*

Powell and Reid's bound is of the form $|\Delta A| \le p(n)u\Omega ee^T + O(u^2)$, where p is an explicitly given quadratic; the absence of a factor m is due to their assumption that inner products are accumulated in double-precision.

We can rewrite (2.15) in the slightly weakened form

$$\max_i \frac{\|\Delta A(i,:)\|_\infty}{\|A(i,:)\|_\infty} \le n^2 \tilde{\gamma}_m \rho_{m,n}, \tag{2.16}$$

where the row-wise growth factor

$$\rho_{m,n} = \max_i \frac{\alpha_i}{\|A(i,:)\|_\infty} = \max_i \left(\frac{\max_{j,k} |\widehat{a}_{ij}^{(k)}|}{\max_j |a_{ij}|} \right).$$

The $\rho_{m,n}$ values in Table 1.1 show that the bound (2.16) is reasonably sharp for the matrix (1.4).

3 Error Analysis for the Least Squares Solution

We now analyse the use of the QR factorization to solve the LS problem $\min \|b - Ax\|_2$. Given the QR factorization with column pivoting

$$A\Pi = QR, \qquad R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}, \quad R_1 \in \mathbb{R}^{n\times n},$$

the computation to be analysed is $c = Q^T b$ (application of the individual Householder transformations to the right-hand side), $R_1 y = c(1\colon n)$ (solution of a triangular system), and $x = \Pi y$. The standard analysis of this computation (see, e.g., [9, Thm. 19.3]) does not lead to row-wise backward error bounds, even though we have row-wise bounds for the factorization. Therefore we adopt a more specific approach, beginning by following the analysis of Powell and Reid.

First, consider the computation of c. For the purposes of the analysis it is convenient to regard b as an extra column of A. We cannot allow b to participate in the column interchanges, so we require that

$$\nu_k = \|b^{(k)}(k\colon m)\|_2 \le |\sigma_k|, \qquad k = 1\colon n. \tag{3.1}$$

In the special case of zero residual problems (where $b = Ax$), it is easy to show that, in fact, $\nu_k \le (n-k+1)|\sigma_k|$, $k = 1\colon n$. To ensure that b does not increase the α_i values we also require that

$$\mu_i = \max_k |b_i^{(k)}| \le \alpha_i, \qquad i = 1\colon m. \tag{3.2}$$

Since conditions (3.1) and (3.2) are not necessarily satisfied, we imagine multiplying b by ξ^{-1} before carrying out the analysis, where

$$\xi = \max\left\{ 1, \max_k \frac{\nu_k}{|\sigma_k|}, \max_i \frac{\mu_i}{\alpha_i} \right\} \geq 1. \tag{3.3}$$

The only consequence of this scaling is that we have to insert a factor ξ in the bound for Δb. Writing $b \equiv a_{n+1}$ and using (2.13) with $j = n+1$, we have

$$b = P_1 \dots P_n \widehat{c} - \Delta b, \qquad |\Delta b| \leq n^2 \tilde{\gamma}_m \xi \Omega e. \tag{3.4}$$

Next we turn to the triangular solve. Standard analysis yields [9, Thm. 8.5]

$$(\widehat{R}_1 + \Delta\widehat{R}_1)\widehat{y} = \widehat{c}(1\colon n), \qquad |\Delta\widehat{R}_1| \leq \gamma_n |\widehat{R}_1|. \tag{3.5}$$

Now $\widehat{R}_1(:, j) = \widehat{a}_j^{(j+1)}(1\colon n, j)$ and so

$$|\Delta\widehat{R}_1(:, j)| \leq \gamma_n |\widehat{a}_j^{(j+1)}(1\colon n)|. \tag{3.6}$$

We can try to incorporate the jth column of the backward error matrix $\Delta\widehat{R}_1$ into our analysis of the factorization by increasing the bound for $|f_j^{(j)}|$ in (2.9) by $\gamma_n|\widehat{a}_j^{(j+1)}|$. However, as Lemma 2.2 states, $f_j^{(j)}(1\colon j-1) = 0$, and this property is used in the proof of Theorem 2.3. Our proposed perturbation $\gamma_n|\widehat{a}_j^{(j+1)}|$ has nonzero leading entries in general. In the analysis of Powell and Reid, $\Delta\widehat{R}_1$ in (3.5) is diagonal (to first order), because inner products are assumed to be accumulated in double precision, and using this argument it is straightforward to obtain a backward error result analogous to that for the factorization. However, since we are making no assumption about the accumulation of inner products we must take a different approach.

It is not hard to see from the analysis above that the computed solution $\widehat{x}$ is the true solution to the LS problem with data $A + \widetilde{\Delta A}$ and $b + \Delta b$, where

$$\widetilde{\Delta A} = A + \Delta A + Q\Delta R, \qquad \Delta R = \begin{bmatrix} \Delta R_1 \\ 0 \end{bmatrix},$$

Δb is defined in (3.4), and ΔA and Q are as defined in Theorem 2.3. We therefore need a row-wise bound for the matrix $Q\Delta R$. Writing Δr_j for the jth column of ΔR we have

$$\begin{aligned}
Q\Delta r_j &= P_1 \dots P_n \Delta r_j \\
&= P_1 \dots P_j \Delta r_j \\
&= P_1 \dots P_{j-1}(I - \beta_j v_j v_j^T)\Delta r_j \\
&= P_1 \dots P_{j-1}\Delta r_j - P_1 \dots P_{j-1} v_j (\beta_j v_j^T \Delta r_j) \\
&\;\;\vdots \\
&= \Delta r_j - \sum_{i=1}^{j} P_1 \dots P_{i-1} v_i (\beta_i v_i^T \Delta r_j).
\end{aligned} \tag{3.7}$$

From (3.6),

$$|\Delta r_j| \le \gamma_n |\widehat{a}_j^{(j+1)}| \le \gamma_n \Omega e. \tag{3.8}$$

For $i \le j$, exploiting the key property that $v_i(1\colon i-1) = 0$, we have

$$\begin{aligned}
|v_i^T \Delta r_j| &\le |v_i^T||\Delta r_j(i\colon m)| \\
&\le \gamma_n |v_i^T||\widehat{a}_j^{(j+1)}(i\colon m)| \\
&\le \gamma_n \|v_i\|_2 \|\widehat{a}_j^{(j+1)}(i\colon m)\|_2 \\
&= \gamma_n \|v_i\|_2 \|\widehat{a}_j^{(i)}(i\colon m)\|_2 \quad \text{by orthogonality,} \\
&\le \gamma_n \|v_i\|_2 \|\widehat{a}_i^{(i)}(i\colon m)\|_2 \quad \text{by column pivoting,} \\
&\le \frac{\gamma_n}{\sqrt{2}} \|v_i\|_2^2, \quad \text{by (2.5).}
\end{aligned} \tag{3.9}$$

Now

$$\begin{aligned}
|P_1 \dots P_{i-1} v_i| &= |(I - \beta_1 v_1 v_1^T) P_2 \dots P_{i-1} v_i| \\
&\le |P_2 \dots P_{i-1} v_i| + |\beta_1 v_1 v_1^T P_2 \dots P_{i-1} v_i| \\
&\le |P_2 \dots P_{i-1} v_i| + 2|v_1| \frac{\|v_i\|_2}{\|v_1\|_2}.
\end{aligned}$$

Applying this inequality recursively, we obtain

$$|P_1 \dots P_{i-1} v_i| \le 2 \sum_{h=1}^{i-1} \frac{\|v_i\|_2}{\|v_h\|_2} |v_h|. \tag{3.10}$$

To bound the right-hand side of (3.10), we need a lemma.

Lemma 3.1 *The vectors v_k from Householder QR factorization with column pivoting satisfy*

$$\frac{\|v_i\|_2}{\|v_j\|_2} \le \sqrt{2}, \qquad i \ge j. \tag{3.11}$$

Proof. For $i \ge j$ we have, from (2.1),

$$\begin{aligned}
\|v_i\|_2 &\le 2\|a_i^{(i)}(i\colon m)\|_2 \\
&\le 2\|a_i^{(i)}(j\colon m)\|_2 \\
&= 2\|a_i^{(j)}(j\colon m)\|_2 \quad \text{by orthogonality,} \\
&\le 2\|a_j^{(j)}(j\colon m)\|_2 \quad \text{by column pivoting.}
\end{aligned}$$

Since (2.5) can be written as

$$\|v_j\|_2 \ge \sqrt{2}\|a_j^{(j)}(j\colon m)\|_2,$$

the result follows. □

Applying the lemma to (3.10) we obtain

$$|P_1 \dots P_{i-1} v_i| \le 2\sqrt{2} \sum_{h=1}^{i-1} |v_h|.$$

From (2.10), $|v_k| \le 2\Omega e$, so

$$\sum_{h=1}^{i-1} |v_h| \le 2(i-1)\Omega e.$$

Substituting into (3.7) and using (3.8) and (3.9) gives

$$\begin{aligned} |Q\Delta r_j| &\le \gamma_n \Omega e + \sum_{i=1}^{j} 2\sqrt{2}(2(i-1)\Omega e)\sqrt{2}\gamma_n \\ &= \tilde{\gamma}_n j^2 \Omega e. \end{aligned}$$

Hence

$$|Q\Delta R| \le \tilde{\gamma}_n \Omega e e^T \operatorname{diag}(1, \dots, n)^2.$$

We have proved the following result.

Theorem 3.2 *Let $A \in \mathbb{R}^{m\times n}$ $(m \ge n)$ have full rank and suppose the LS problem $\min \|b - Ax\|_2$ is solved using Householder QR factorization with column pivoting. Then the computed solution $\widehat{x}$ is the exact LS solution to*

$$\min_x \|(b + \Delta b) - (A + \Delta A)x\|_2,$$

where the perturbations satisfy

$$|\Delta A| \le \tilde{\gamma}_m \Omega e e^T \operatorname{diag}(1, \dots, n)^2, \qquad |\Delta b| \le n^2 \tilde{\gamma}_m \xi \Omega e, \tag{3.12}$$

where Ω is defined in (2.7) *and ξ in* (3.3).

Unfortunately, it is not possible to test the bounds in (3.12) empirically, because no computable expression is known for the row-wise backward error of an arbitrary approximate LS solution (formulae are known, however, for the normwise backward error [15], [17]).

4 Row Pivoting and Row Sorting

For Householder QR factorization with column pivoting, Theorem 2.3 bounds the rows of the backward error matrix ΔA in terms of the scalars α_i in (2.7). The matrix (1.3) shows that $\alpha_i / \max_j |a_{ij}|$ can be arbitrarily large for column pivoting, and the matrix

(1.4) shows that the row-wise backward error can also be large. As Powell and Reid discovered, the key to obtaining a small row-wise backward error is to incorporate row interchanges: at the start of the kth stage, after interchanging columns according to the column pivoting strategy, we interchange rows to ensure that

$$|a_{kk}^{(k)}| = \max_{i \geq k} |a_{ik}^{(k)}|. \tag{4.1}$$

(Note the importance of interchanging columns before rows.) This is what we did in Section 1 for the matrix (1.3) before the first stage of QR factorization. The next result slightly improves bounds of Powell and Reid [14].

Lemma 4.1 *With the row interchange strategy* (4.1) *and column pivoting we have*

$$\max_{j \geq k} |a_{ij}^{(k+1)}| \leq c_i^{(k)} \max_{j \geq k} |a_{ij}^{(k)}|, \tag{4.2}$$

where

$$c_i^{(k)} = \begin{cases} \sqrt{m-i+1}, & i = k, \\ 1+\sqrt{2}, & i > k. \end{cases}$$

Consequently,

$$\alpha_i \leq \begin{cases} \sqrt{m-i+1}(1+\sqrt{2})^{i-1} \max_j |a_{ij}|, & i \leq n, \\ (1+\sqrt{2})^{n-1} \max_j |a_{ij}|, & i > n. \end{cases}$$

Proof. Since $(v_k)_i = a_{ik}^{(k)}$ for $i > k$, from Lemma 2.1 we obtain

$$\max_{j \geq k} |a_{ij}^{(k+1)}| \leq (1+\sqrt{2}) \max_{j \geq k} |a_{ij}^{(k)}|, \qquad i > k. \tag{4.3}$$

Using the fact that premultiplication by P_k preserves 2-norms of columns and alters only elements with indices $k\colon m$ we have

$$\begin{aligned} |a_{kj}^{(k+1)}| &\leq \left(\sum_{i=k}^{m} |a_{ij}^{(k+1)}|^2 \right)^{1/2} = \left(\sum_{i=k}^{m} |a_{ij}^{(k)}|^2 \right)^{1/2} \\ &\leq |\sigma_k|. \end{aligned} \tag{4.4}$$

But $|\sigma_k| \leq \sqrt{m-k+1}\,|a_{kk}^{(k)}|$, using (4.1), which gives the formula for $c_k^{(k)}$. The bound for α_i follows from (4.2), since for $i \leq n$ row i is altered on the first $i-1$ stages and is a pivot row on the ith stage, while if $i > n$ then row i is altered on n stages (in the last of which it becomes the zero vector) and is never a pivot row. ∎

Lemma 4.1 shows that if both column pivoting and row pivoting are used then Householder QR factorization yields a bounded ratio $\alpha_i / \max_j |a_{ij}|$, although the bound can be of order $(1+\sqrt{2})^n$. Powell and Reid [14] give an example where the bound is nearly attained, but suggest that the ratio is usually of order 1.

As a result of Powell and Reid's analysis, it is standard practice to use row and column pivoting for badly row-scaled problems. However, as noted in Section 1, Björck [5, p. 169] suggests sorting the rows at the start of the factorization so that

$$\|A(1,:)\|_\infty \geq \|A(2,:)\|_\infty \geq \cdots \geq \|A(m,:)\|_\infty, \tag{4.5}$$

and then using column pivoting alone. This can be done in $O(m \log_2 m + mn)$ comparisons as opposed to the $O(m^2)$ required for row pivoting, and it may be preferable from the programming point of view to do all the row interchanges at once, rather than to spread them one per step through the factorization. Note that arranging for (4.5) to hold is trivial for a problem of the form (1.1) resulting from the method of weighting. The next lemma shows that row sorting yields the same bound on the α_i as row pivoting.

Lemma 4.2 *If the rows of A are ordered so that* (4.5) *holds then, with column pivoting,*

$$\alpha_i \leq \begin{cases} \sqrt{m-i+1}(1+\sqrt{2})^{i-1} \max_j |a_{ij}|, & i \leq n, \\ (1+\sqrt{2})^{n-1} \max_j |a_{ij}|, & i > n. \end{cases}$$

Proof. Note first that (4.3) is still valid, as it depends only on Lemma 2.1. For $i = k$ we have, using (4.4) and (4.3),

$$\begin{aligned} \max_{j\geq k} |a_{kj}^{(k+1)}| &\leq \max_{j\geq k} \left(\sum_{i=k}^{m} |a_{ij}^{(k)}|^2 \right)^{1/2} \\ &\leq \max_{j\geq k} \sqrt{m-k+1} \max_{i\geq k} |a_{ij}^{(k)}| \\ &= \sqrt{m-k+1} \max_{i\geq k} \max_{j\geq k} |a_{ij}^{(k)}| \\ &\leq \sqrt{m-k+1} \max_{i\geq k} (1+\sqrt{2})^{k-1} \max_j |a_{ij}| \\ &= \sqrt{m-k+1}(1+\sqrt{2})^{k-1} \max_j \max_{i\geq k} |a_{ij}| \\ &= \sqrt{m-k+1}(1+\sqrt{2})^{k-1} \max_j |a_{kj}|, \end{aligned}$$

by (4.5). □

In our numerical experiments we have found row sorting to give very similar row-wise backward errors to row pivoting.

5 Choice of Sign in Householder Matrix Construction

The error analysis in Section 2 rests on Lemma 2.1, which uses in its proof the assumption (2.2) on the choice of sign in the Householder matrices. To simplify the notation, consider

a Householder matrix $P = I - \beta vv^T$, where $\beta = 2/v^Tv$, and recall that if $v = x - \sigma e_1$ and $\sigma = \pm\|x\|_2$, then $Px = \sigma e_1$. Since $v_1 = x_1 - \sigma$, textbooks usually recommend the choice

$$\sigma = -\mathrm{sign}(x_1)\|x\|_2, \tag{5.1}$$

in order to avoid cancellation. The other choice of σ,

$$\sigma = \mathrm{sign}(x_1)\|x\|_2, \tag{5.2}$$

can be used in a numerically stable way provided that the computation of v_1 is rearranged as follows [12]:

$$v_1 = x_1 - \mathrm{sign}(x_1)\|x\|_2 = \frac{x_1^2 - \|x\|_2^2}{x_1 + \mathrm{sign}(x_1)\|x\|_2} = \frac{-(x_2^2 + \cdots + x_n^2)}{x_1 + \mathrm{sign}(x_1)\|x\|_2}.$$

In practice, we might want to take a positive σ at each step of the factorization in order to produce a matrix R normalized to have nonnegative diagonal elements, in which case we would need to switch between the choices (5.1) and (5.2).

Theorem 1.1 on the normwise and column-wise stability of Householder QR factorization holds no matter what choice of sign is made on each stage of the factorization. Lemma 2.1, however, is not valid for (5.2). This is illustrated by the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 0 & 1 \\ \epsilon & 1 \end{bmatrix}, \qquad \epsilon > 0.$$

Row pivoting, row sorting and column pivoting all leave this matrix unchanged, and for (5.2) we have

$$\phi_2^{(1)} = \beta_1 v_1^T a_2^{(1)} = O(1/\epsilon),$$

showing that $\phi_2^{(1)}$ is unbounded as $\epsilon \to 0$. This problem can occur whenever the leading column a_1 is such that $\|a_1\|_2 \approx |a_{11}|$, because then $v_1 = a_1 - \mathrm{sign}(a_{11})\|a_1\|_2 e_1$ is small and hence $\phi_j^{(1)} \le \|a_j^{(1)}\|/\|v_1\|_2$ can be large.

A numerical example confirms that the choice of sign (5.2) need not lead to row-wise stability. Let $A \in \mathbb{R}^{7\times 5}$ be the matrix with $a_{ij} = 1$ except that $a_{ii} = \lambda$ for $i = 1\colon 5$. With $\lambda = 10^8$ we obtain the backward errors and row-wise growth factors shown in Table 5.1.

This is the only situation we know in which the choice of sign in defining a Householder matrix affects the stability of an algorithm.

Acknowledgements

We thank Jesse Barlow for helpful discussions.

Pivoting:	None	Row	Col.	Row and col.
(5.2): Normwise (η)	2.90e-16	7.65e-16	1.32e-15	7.65e-16
Row-wise (η_R)	5.80e-9	1.08e-8	1.78e-8	1.08e-8
$\rho_{m,n}$	5.00e+7	5.00e+7	5.00e+7	5.00e+7
(5.1): Row-wise (η_R)	8.94e-16	8.94e-16	7.45e-16	8.94e-16
$\rho_{m,n}$	1.00	1.00	1.00	1.00

Table 5.1: Backward errors and row-wise growth factors for QR factorization with no pivoting, row pivoting and column pivoting, using (5.2) and (5.1).

References

[1] Andrew A. Anda and Haesun Park. Self-scaling fast rotations for stiff and equality-constrained linear least squares problems. *Linear Algebra and Appl.*, 234:137–161, 1996.

[2] E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen. *LAPACK Users' Guide, Release 2.0.* Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1995. xix+325 pp. ISBN 0-89871-345-5.

[3] Jesse L. Barlow. Stability analysis of the G-algorithm and a note on its application to sparse least squares problems. *BIT*, 25:507–520, 1985.

[4] Jesse L. Barlow and Susan L. Handy. The direct solution of weighted and equality constrained least-squares problems. *SIAM J. Sci. Stat. Comput.*, 9(4):704–716, 1988.

[5] Åke Björck. *Numerical Methods for Least Squares Problems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996. xvii+408 pp. ISBN 0-89871-360-9.

[6] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *LINPACK Users' Guide.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1979. ISBN 0-89871-172-X.

[7] G. H. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7:206–216, 1965.

[8] Mårten Gulliksson. Backward error analysis for the constrained and weighted linear least squares problem when using the weighted QR factorization. *SIAM J. Matrix Anal. Appl.*, 16(2):675–687, 1995.

[9] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996. xxviii+688 pp. ISBN 0-89871-355-2.

[10] Patricia D. Hough and Stephen A. Vavasis. Complete orthogonal decomposition for weighted least squares. *SIAM J. Matrix Anal. Appl.*, 18(2):369–392, 1997.

[11] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1995. xii+337

pp. Revised republication of work first published in 1974 by Prentice-Hall. ISBN 0-89871-356-0.

[12] Beresford N. Parlett. Analysis of algorithms for reflections in bisectors. *SIAM Review*, 13(2):197–208, 1971.

[13] M. J. D. Powell and J. K. Reid. On applying Householder transformations to linear least squares problems. Technical Report T.P. 322, Mathematics Branch, Theoretical Physics Division, Atomic Energy Research Establishment, Harwell, UK, February 1968. 20 pp.

[14] M. J. D. Powell and J. K. Reid. On applying Householder transformations to linear least squares problems. In *Proc. IFIP Congress 1968*, North-Holland, Amsterdam, The Netherlands, 1969, pages 122–126.

[15] Ji-guang Sun. Optimal backward perturbation bounds for the linear least-squares problem with multiple right-hand sides. *IMA J. Numer. Anal.*, 16(1):1–11, 1996.

[16] Charles F. Van Loan. On the method of weighting for equality-constrained least-squares problems. *SIAM J. Numer. Anal.*, 22(5):851–864, 1985.

[17] Bertil Waldén, Rune Karlson, and Ji-guang Sun. Optimal backward perturbation bounds for the linear least squares problem. *Numerical Linear Algebra with Applications*, 2(3):271–286, 1995.