RAL-TR-98-027

Implicit scaling of linear least squares problems[1]

by

J. K. Reid[2]

Abstract

We consider the solution of weighted linear least squares problems by Householder transformations with implicit scaling, that is, with the weights stored separately. By holding inverse weights, the constrained case can be accommodated. The error analysis of the weighted and unconstrained case is readily extended and we show that iterative refinement may be applied.

Categories and subject descriptors: G.1.3 [**Numerical Linear Algebra**]: Linear systems (direct methods), Sparse and very large systems.

General Terms: Algorithms, Error analysis

Additional Key Words and Phrases: weighted linear least squares, QR factorization, backward errors, iterative refinement.

Department for Computation and Information,
Rutherford Appleton Laboratory,
Oxon OX11 0QX.
UK

March 1998.

---

[1] Available by anonymous ftp from `matisa.cc.rl.ac.uk` in directory `pub/reports` in the file `rRAL98027.ps.gz`

[2] Email address: `jkr@rl.ac.uk`

CONTENTS

1 Introduction

Householder (1958) proposed the use of orthogonal transformations for the solution of the overdetermined system of linear equations

$$A\,x = b, \tag{1.1}$$

where A is an $m \times n$ matrix with $m > n$. The method was discussed further by Golub (1965), who proposed the use of column interchanges. As refined by Golub, the algorithm involves the use of Householder transformations to form a QR factorization of a column permutation of A.

The method was implemented in ALGOL 60 by Businger and Golub (1965). The factorization is computed in n major steps, each of which may involve a column interchange. Powell and Reid (1968, 1969) proposed row interchanges, too. Apart from organizational aspects and the application of the inverse column permutation to x, the computation is identical if the interchanges are applied first to A and b, and then the algorithm is performed without interchanges. We will make this assumption, which allows us to simplify our notation greatly. The factorization obtained is

$$A = Q\,R, \tag{1.2}$$

where Q is orthogonal and R is upper triangular.

The weighted problem

$$\min\,(b - Ax)^T W^2 (b - Ax), \tag{1.3}$$

where $W = \mathrm{diag}(w_k)$ is a diagonal matrix, can be expressed as the unweighted least squares solution of the system

$$W A\,x = W\,b. \tag{1.4}$$

Powell and Reid (1969) formed this set of equations explicitly and performed a backward error analysis that showed the solution obtained to be exact for a perturbed system where the perturbations in each row were small compared with the largest element in the row.

We will show that the algorithm can be extended to the constrained case by use of implicit scaling, that is, without ever forming $W A$ or $W\,b$ explicitly. Corresponding to the factorization (1.2), we obtain the factorization

$$W A = Q\,W R. \tag{1.5}$$

By storing the inverse weights W^{-1}, we can include the case of infinite weights without the need to represent ∞. We set out the resulting algorithm in Section 2. The backward error analysis of Powell and Reid is applicable, as we explain in Section 3. The detailed proof of their bounds was given only in a report (Powell and Reid 1969), which is now out of print. Therefore, it is reproduced in the Appendix.

Gulliksson and Wedin (1992) obtain the same algorithm by introducing the concept of an '*M*–invariant reflection' and developing associated mathematical theory, which obscures the fact that it is really a very straightforward extension and that the error analysis is applicable. *M* is the diagonal matrix

$$M = W^{-2} \tag{1.6}$$

and a matrix Q_M is said to be *M*–invariant if the equation

$$Q_M M Q_M^T = M \tag{1.7}$$

holds. It is readily verified that if *Q* is orthogonal, the matrix

$$Q_M = W^{-1} Q W \tag{1.8}$$

is *M*–invariant, and the factorization (1.5) can be written

$$A = Q_M R. \tag{1.9}$$

Cox and Higham (1997) provide an alternative error analysis that is shorter and does not rely on accurate accumulation of inner products. We summarize their results in Section 4. They also show that their bounds are available if row pivoting is replaced by initial sorting of the rows so that they have descending infinity norm. The same is true for the Powell and Reid result, as shown at the end of the section.

Iterative refinement in the weighted case is considered in Section 5 and the extension to infinite weights is considered in Section 6. The resulting iteration is the same as that of Gulliksson (1994), but is not dependent on the theory of *M*–invariant reflections. An error analysis has been provided by Gulliksson (1995).

2 Householder reduction for the weighted problem

In this section, we set out the details of the Householder reduction with implicit scaling. We start with the system (1.4) and end with the factorization (1.5). Thus, *A*, *R*, and *b* are implicitly scaled by *W*. Other quantities are implicitly scaled as shown in Table 1. Where both sides of an equation are weighted in the same way, we omit the weights. All we are actually doing is taking the algorithm and the variant of Powell and Reid and writing down what happens when the quantities are held with implicit scaling. For each of the equations in this and the next section, the equivalent original equation may be recovered easily by making the substitution $W = I$.

Factor	Applied to
W	A, $A^{(k)}$, R, b, $b^{(k)}$, α, Δ, δ
w_k	σ_k
w_k^{-2}	β_k
1	H_k, $y^{(k)}$, $\gamma^{(k)}$, ρ

Table 1. Implicit scaling factors.

The reduction involves n steps starting from $A^{(1)} = A$ and ending with $A^{(n+1)} = R$. The k-th step involves the transformation

$$WA^{(k+1)} = H^{(k)}\, WA^{(k)}, \quad k = 1, 2, ..., n, \tag{2.1}$$

where $H^{(k)}$ is the matrix

$$H^{(k)} = I - w_k^{-2} \beta_k\, W u^{(k)} u^{(k)\,T} W, \tag{2.2}$$

the vector $u^{(k)}$ has components

$$\left.\begin{aligned} u_i^{(k)} &= 0, \quad i < k \\ u_k^{(k)} &= \left(\sigma_k + |a_{kk}^{(k)}|\right) \mathrm{sign}\left(a_{kk}^{(k)}\right) \\ u_i^{(k)} &= a_{ik}^{(k)}, \quad i > k, \end{aligned}\right\} \tag{2.3}$$

and σ_k and β_k are given by the equations

$$w_k \sigma_k = \sqrt{\sum_{i=k}^{m} w_i^2\, a_{ik}^{(k)\,2}} \tag{2.4}$$

and

$$\beta_k = \left(\sigma_k^2 + \sigma_k |a_{kk}^{(k)}|\right)^{-1}. \tag{2.5}$$

This choice of β_k ensures that $H^{(k)}$ is orthogonal and that the k-th column of $WA^{(k+1)} = H^{(k)}\, WA^{(k)}$ is zero below the diagonal. Later multiplication by $H^{(j)}$, $j > k$, preserves this property since $u_i^{(j)}$, $i \le k$, is chosen to be zero. Therefore, finally, $A^{(n+1)}$ is the upper triangular matrix R of the factorization (1.5).

For the right-hand side, we start with $b^{(1)} = b$ and the k-th step involves the formula

$$Wb^{(k+1)} = H^{(k)}\, Wb^{(k)}, \quad k = 1, 2, ..., n, \tag{2.6}$$

which requires the computation of

$$\gamma_k = w_k^{-2} \beta_k\, u^{(k)\,T} W^2 b^{(k)} \tag{2.7}$$

and

$$b^{(k+1)} = b^{(k)} - \gamma_k u^{(k)}. \tag{2.8}$$

Finally, the upper-triangular system

$$\sum_{j=i}^{n} a_{ij}^{(n+1)} x_j = b_i^{(n+1)}, \quad i = 1, 2, ..., n \tag{2.9}$$

is solved by back-substitution.

The column interchanges are chosen so that the inequalities

$$w_k \sigma_k = \sqrt{\sum_{i=k}^{m} w_i^2 a_{ik}^{(k)\,2}} \geq \sqrt{\sum_{i=k}^{m} w_i^2 a_{ij}^{(k)\,2}}, \quad j > k \tag{2.10}$$

hold for each k.

Computing $A^{(k+1)}$, see equation (2.1), requires the computation of

$$y^{(k)\,T} = w_k^{-2} \beta_k \, u^{(k)\,T} W^2 A^{(k)}, \tag{2.11}$$

whose nonzero components are

$$y_j^{(k)} = \beta_k \sum_{i=k}^{m} u_i \left(\frac{w_i}{w_k} \right)^2 a_{ij}^{(k)}, \quad j = k, \ldots, n \tag{2.12}$$

and

$$A^{(k+1)} = A^{(k)} - u^{(k)} y^{(k)\,T}. \tag{2.13}$$

A consequence of the column interchanges is the bound

$$|y_j^{(k)}| \leq \sqrt{2} \tag{2.14}$$

on the components of $y^{(k)}$, which is important for the stability of the algorithm, see equation (3.6). This is proved by applying Schwarz's inequality to the j-th component in equation (2.11). For reasons explained in the next section, Powell and Reid introduced row interchanges to ensure that the inequality

$$|w_k a_{kk}^{(k)}| \geq |w_i a_{ik}^{(k)}|, \quad i > k \tag{2.15}$$

is always true.

3 The error analysis of Powell and Reid

Powell and Reid (1968, 1969) performed a backward error analysis for the explicitly weighted case, assuming that all floating-point operations (including square roots and inner products) are performed with relative accuracy bounded by ε and terms of order ε^2 are ignored. The results hold equally well when every number is scaled by a power of a weight, provided the scaling can be performed without error, which will be the case when the weights are powers of the base. We summarize their results

here for this case. The effect of other weights is to increase the numerical values of the factors in the bounds, but it does not affect the qualititive result.

Using bars for computed quantities, Powell and Reid introduced the row factors

$$\alpha_i = \max_{j,k} |\bar{a}_{ij}^{(k)}|, \quad i = 1, 2, ..., m, \tag{3.1}$$

and showed that if all terms of order ε^2 are ignored, the computed solution $\bar{x}$ is the exact weighted least squares solution of the system

$$(A + \Delta)\, x = b + \delta \tag{3.2}$$

where the elements of Δ are bounded by the inequality

$$|\Delta_{ij}| \le \left[n^2 \left(\tfrac{25}{2} + \tfrac{21\sqrt{2}}{4} \right) - n \left(\tfrac{85}{4} - \tfrac{5\sqrt{2}}{2} \right) + (24 + 14\sqrt{2}) \right] \varepsilon\, \alpha_i \tag{3.3}$$

and the elements of δ are bounded by the inequality

$$|\delta_i| \le \left[n^2 \left(\tfrac{25}{2} + \tfrac{21\sqrt{2}}{4} \right) + n \left(\tfrac{3}{4} + 13\sqrt{2} \right) + (24 + 14\sqrt{2}) \right] \varepsilon\, \rho\, \alpha_i, \tag{3.4}$$

where

$$\rho = \max \left[\max_{i,k} \frac{|b_i^{(k)}|}{\alpha_i}, \max_k \frac{\sqrt{\sum_{i=k}^{m} w_i^2 b_i^{(k)\,2}}}{w_k \sigma_k} \right]. \tag{3.5}$$

These results show that Golub's algorithm is backwards stable in the presence of rows with widely varying norms provided there is no significant growth of these norms during the reduction. The column interchanges ensure that, outside the pivot row, the growth is limited by the factor $1 + \sqrt{2}$ in a single step:

$$\max_j |a_{ij}^{(k+1)}| \le (1 + \sqrt{2}) \max_j |a_{ij}^{(k)}|, \quad i > k. \tag{3.6}$$

This is an immediate consequence of the bound (2.14) and the equations (2.3) and (2.13). However, there may be growth in the pivot row unless row interchanges are included, too. Powell and Reid used the simple example

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 10^6 & 10^6 & 0 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{pmatrix} \tag{3.7}$$

for illustration. Without row interchanges, there is disastrous growth in the first row, with $\alpha_1 = 10^6\sqrt{2}$. With row interchanges to ensure that inequality (2.15) holds they were able to prove the inequality

$$\max_j |a_{kj}^{(k+1)}| \le \sqrt{m}\,|a_{kk}^{(k)}|. \tag{3.8}$$

To prove this, we note that since orthogonal transformations preserve the Euclidean norms of columns, the inequality

$$w_k |a_{kj}^{(k+1)}| \le \sqrt{\sum_{i=k}^{m} w_i^2 a_{ij}^{(k+1)\,2}} = \sqrt{\sum_{i=k}^{m} w_i^2 a_{ij}^{(k)\,2}}. \tag{3.9}$$

holds. The result now follows from the inequality

$$\sqrt{\sum_{i=k}^{m} w_i^2 a_{ij}^{(k)\,2}} \le w_k \sigma_k \tag{3.10}$$

which is a consequence of the column interchanges, and the inequality

$$\sigma_k \le \sqrt{m}\,|a_{kk}^{(k)}|, \tag{3.11}$$

which is a consequence of of the row interchanges.

The results (3.6) and (3.8) allowed Powell and Reid to prove that the inequality

$$\alpha_i \le (1+\sqrt{2})^{n-1}\sqrt{m}\max_j |a_{ij}| \tag{3.12}$$

holds.

4 The work of Cox and Higham

Cox and Higham (1997) provide an alternative error analysis that is shorter and easier to read because they use vector and matrix notation exclusively and are not concerned with obtaining explicit constants. It is not a first-order analysis and does not assume that inner products are accumulated in extra precision. They use the notation $\bar{\gamma}_k$ for the quantity

$$\bar{\gamma}_k = \frac{ck\varepsilon}{1-ck\varepsilon}, \tag{4.1}$$

where c is a small integer constant whose exact value is unimportant. We will again express their results in terms of the weighted problem. Corresponding to the bound (3.3), they find the bound

$$|\Delta_{ij}| \le \bar{\gamma}_m j^2 \varepsilon \alpha_i. \tag{4.2}$$

The presence of the factor m is attributable to ordinary accumulation of inner products.

Corresponding to the bound (3.4), they find the bound

$$|\delta_i| \le \bar{\gamma}_m n^2 \varepsilon \max(1,\rho)\,\alpha_i. \tag{4.3}$$

The row interchanges in general ensure that rows with large norms are pivoted first, but Cox and Higham show that their bounds hold if the weighted rows are ordered initially to have decreasing

infinity norm, that is, so that the relations

$$\max_j |w_i a_{ij}| \le \max_j |w_k a_{kj}|, \quad i > k \tag{4.4}$$

hold. This rule is also applicable to the analysis of Powell and Reid. The bounds (3.8) and (3.11) will no longer always be true, but relationship (3.12) does still hold. To establish this, we note that relationship (3.9) gives the inequality

$$\max_{j\ge k} |w_k a_{kj}^{(k+1)}| \le \sqrt{m} \max_{j\ge k} \max_{i\ge k} |w_i a_{ij}^{(k)}| = \sqrt{m} \max_{i\ge k} \max_{j\ge k} |w_i a_{ij}^{(k)}|. \tag{4.5}$$

Now using relationship (3.6) repeatedly gives the inequality

$$\max_{j\ge k} |w_k a_{kj}^{(k+1)}| \le \sqrt{m} \max_{i\ge k} (1+\sqrt{2})^{k-1} \max_j |w_i a_{ij}| \tag{4.6}$$

from which relationship (3.12) is a consequence of the row ordering, see equation (4.4).

5 Iterative refinement

Björck (1967, 1968), see also Björck (1996), showed that the key for successful iterative refinement of the least squares solution of equation (1.1) is to work with the augmented system

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{5.1}$$

For our weighted case, this becomes

$$\begin{pmatrix} W^{-2} & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{5.2}$$

The residual for this system is

$$\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} - \begin{pmatrix} W^{-2} & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} \tag{5.3}$$

and iterative refinement involves solving the equation

$$\begin{pmatrix} W^{-2} & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \delta r \\ \delta x \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{5.4}$$

Using equation (1.5) to substitute for A gives the equation

$$\begin{pmatrix} W^{-2} & W^{-1} Q W R \\ R^T W Q^T W^{-1} & 0 \end{pmatrix} \begin{pmatrix} \delta r \\ \delta x \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{5.5}$$

which can be rewritten as the equation

$$\begin{pmatrix} W^{-2} & R \\ R^T & 0 \end{pmatrix} \begin{pmatrix} WQ^T W^{-1} \delta r \\ \delta x \end{pmatrix} = \begin{pmatrix} W^{-1} Q^T W f \\ g \end{pmatrix}. \tag{5.6}$$

Since R is upper triangular, this can be rewritten in the form

$$\begin{pmatrix} W_1^{-2} & 0 & U \\ 0 & W_2^{-2} & 0 \\ U^T & 0 & 0 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \delta x \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ g \end{pmatrix}, \tag{5.7}$$

where U is square and upper triangular. The matrix of equation (5.7) is a permutation of a triangular matrix and can therefore be solved by first solving the equation

$$U^T h_1 = g \tag{5.8}$$

by forward substitution, then solving the diagonal system

$$W_2^{-2} h_2 = d_2, \tag{5.9}$$

and then solving the equation

$$U \delta x = d_1 - W_1^{-2} h_1 \tag{5.10}$$

by back-substitution. The vector $d = W^{-1} Q^T W f$ may be calculated as $f^{(n+1)}$ from $f = f^{(1)}$ by the sequence of steps

$$f^{(k+1)} = W^{-1} H^{(k)} W f^{(k)}, \quad k = 1, 2, ..., n, \tag{5.11}$$

which are exactly like those applied to b in equations (2.6) to (2.8). Similarly, $\delta r = W Q W^{-1} h$ may be calculated as $h^{(0)}$ from $h = h^{(n)}$ in the backward sweep:

$$h^{(k-1)} = W H^{(k)} W^{-1} h^{(k)} = \left(I - \beta_k w_k^{-2} W^2 u^{(k)} u^{(k)\,T} \right) h^{(k)}, \quad k = n, n-1, ..., 1, \tag{5.12}$$

which requires the computation of

$$\eta_k = \beta_k u^{(k)\,T} h^{(k)} \tag{5.13}$$

and

$$h^{(k-1)} = h^{(k)} - w_k^{-2} \eta_k W^2 u^{(k)}. \tag{5.14}$$

It is usual to start the iteration with $r = 0$ and $x = 0$, so that in the first iteration equation (5.3) gives $f = b$ and $g = 0$. It is interesting that the computations applied to f to form d are exactly the same as those applied to b in the method without refinement. Further, since g is zero, so is h_1 (see equation (5.8)) and equation (5.10) simplifies to

$$U \delta x = d_1, \tag{5.15}$$

which has the same matrix and right-hand side as equation (2.9). Therefore the first iteration will yield

exactly the same approximation for x as the non-refined method.

6 The constrained case

We now consider the constrained weighted least squares problem. We suppose that there are p equality constraints and that they are linearly independent. Infinite weights are needed to implement the Powell-Reid method, so we store W^{-1} rather than W. The weights always appear relative to each other, which means that if a weighted sum involves a nonzero with an infinite weight, we can ignore all nonzeros without infinite weights. We take the infinite weights to be equal to each other, that is, the ratio of any two of them to be unity.

The column and row interchanges, see inequalities (2.10) and (2.15), are sure to bring the constraint rows to the front if they are not already there. The steps of the algorithm become

(i) Choose next column and calculate σ_k. If $k \le p$, inequality (2.10) reduces to

$$\sigma_k = \sqrt{\sum_{i=k}^{p} a_{ik}^{(k)\,2}} \ge \sqrt{\sum_{i=k}^{p} a_{ij}^{(k)\,2}}, \quad j > k, \tag{6.1}$$

For $k > p$, inequality (2.10) is used as before.

(ii) Choose next row. If $k \le p$, use row interchanges to ensure that the inequality

$$|a_{kk}^{(k)}| \ge |a_{ik}^{(k)}|, \quad k < i \le p \tag{6.2}$$

is always true. For $k > p$, we return to inequality (2.15).

(iii) Calculate u and β_k. Equations (2.3) and (2.5) are used unchanged since they do not involve the weights.

(iv) Calculate $y^{(k)\,T}$. If $k \le p$, equation (2.12) reduces to

$$y_j^{(k)} = \beta_k \sum_{i=k}^{p} u_i a_{ij}^{(k)}, \quad j = k, \ldots, n. \tag{6.3}$$

For $k > p$, we return to equation (2.12).

(v) Calculate $A^{(k+1)}$. Equation (2.13) is used unchanged since it does not involve the weights.

The resulting algorithm is equivalent to that of Gulliksson and Wedin (1992). The error analysis of Powell and Reid is still applicable. We will have performed a weighted and constrained solution of the perturbed system (3.2) with bounds given by inequalities (3.3) and (3.4), where ρ has the value

$$\rho = \max\left[\max_{i,k} \frac{|b_i^{(k)}|}{\alpha_i}, \max_{k \le p} \frac{\sqrt{\sum_{i=k}^{m} b_i^{(k)\,2}}}{\sigma_k}, \max_{k>p} \frac{\sqrt{\sum_{i=k}^{m} (w_i b_i^{(k)})^2}}{w_k \sigma_k} \right]. \tag{6.4}$$

We illustrate this behaviour with the example considered by Gulliksson and Wedin (1992). Here, A is the matrix

$$A = \begin{pmatrix} 1 & 1 & 5 & 4 \\ 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 1 \\ 1 & 0 & 6 & 1 \\ 1 & 6 & 10 & 2 \end{pmatrix}$$

and the right-hand side is defined by the equation $b = W^{-2}\lambda + Ax$ for $\lambda = [3, -9, 5, 1, 0]^T$ and $x = [-12, 1, 3, 3]^T$. Since $A^T\lambda = 0$, this yields a problem whose solution is x. Using a SUN Ultra 1 workstation, we held the given problem and the vectors x and r in double precision and the other quantities in single precision. Taking $W^{-1} = [\mu, \mu, \mu, 1, 1]^T$ for $\mu = 1, 10^{-3}, 10^{-6}, 0$, we found residuals with norms as shown in Table 2. Convergence was good in all these cases and only 3 iterations were needed to obtain as much accuracy as can be expected.

	μ	1	10^{-3}	10^{-6}	0
Iteration 1	$\|f\|_\infty$	5.0e-06	2.3e-06	2.3e-06	2.3e-06
	$\|g\|_\infty$	4.4e-07	1.7e-06	1.2e-06	1.2e-06
Iteration 2	$\|f\|_\infty$	8.8e-13	4.5e-13	3.3e-13	4.3e-14
	$\|g\|_\infty$	3.3e-13	3.5e-12	1.3e-12	2.2e-12
Iteration 3	$\|f\|_\infty$	0	3.6e-15	7.1e-15	0
	$\|g\|_\infty$	0	1.8e-15	1.3e-19	1.7e-19

Table 2. Residual norms after 1, 2, and 3 iterations.

For the constrained but unweighted case, the usual approach (Björck and Golub 1967) is to perform a QR decomposition of the p leading (constraint) rows, with column interchanges, to yield the system

$$\begin{pmatrix} Q^T A_{11} & Q^T A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} Q^T b_1 \\ b_2 \end{pmatrix}, \tag{6.5}$$

that is,

$$\begin{pmatrix} R_{11} & A'_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ b_2 \end{pmatrix}, \tag{6.6}$$

where R_{11} is upper triangular. Now Gaussian elimination is applied to yield the equivalent system

$$\begin{pmatrix} R_{11} & A'_{12} \\ 0 & A'_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}. \tag{6.7}$$

QR decomposition is used for the least squares solution of

$$A'_{22} x_2 = c_2 \tag{6.8}$$

and x_1 is found by back-substitution

$$R_{11} x_1 = c_1 - A'_{12} x_2. \tag{6.9}$$

Using infinite weights in the Powell-Reid method yields an algorithm that is very closely related. The infinite weights mean that the computation in the leading rows ignores all data in the remaining rows, so that R_{11}, A'_{12}, and c_1 are computed as in (6.6). In fact, when replacing $A^{(k)}$ by $A^{(k+1)}$, every row of the matrix is altered by a multiple of $y^{(k)\,T}$, see equation (2.11), which is now a linear combination of rows k to p since $w_i/w_k = 0$ for $i > p$. Thus, a form of Gaussian elimination is applied. The number of arithmetic operations is the same and the end result is the same,

$$A'_{22} = A_{22} - A_{21} R_{11}^{-1} A'_{12} = A_{22} - A_{21} A_{11}^{-1} Q\, Q^T A_{12} = A_{22} - A_{21} A_{11}^{-1} A_{12}, \tag{6.10}$$

apart from roundoff effects.

Stewart (1997) considers the constrained case and compares elimination methods with explicitly weighted methods. He suggests weighting so that the ratio of the norm of the non-constraint part of the matrix to the norm of the constraint part is less than the relative precision. However, his analysis is based on the use of norms and does not provide a row-wise backward error result comparable with inequalities (3.3) and (3.4).

7 Conclusion

The Powell and Reid algorithm, with minor changes, is available for constrained weighted linear least squares problems and their backward error bounds apply. Those error bounds apply also if row interchanges are applied a priori to order the rows by decreasing weighted infinity norms. The algorithm can be applied with iterative refinement.

8 Acknowledgements

I would like to thank Iain Duff, Nick Gould, and Michael Powell for reading drafts of this paper and making suggestions that have substantially improved the presentation.

9 References

Björck, Å. (1967). Iterative refinement of linear least squares solutions I. *BIT* **7**, 257-278.

Björck, Å. (1996). Numerical methods for least squares problems. SIAM, Philadelphia.

Björck, Å. (1968). Iterative refinement of linear least squares solutions II. *BIT* **8**, 8-30.

Björck, Å. and Golub, G. H. (1967). Iterative refinement of linear least squares solutions by Householder transformations. *BIT* **7**, 322-337.

Businger, P. A. and Golub, G. H. (1965). Linear least squares solutions by Householder transformations. *Numerische Math.* **7**, 269-276.

Cox, A. J. and Higham, N. J. (1997). Stability of Householder QR factorizations for weighted least squares problems. Report 301, Department of Mathematics, University of Manchester.

Golub, G. H. (1965). Numerical methods for solving linear least squares problems. *Numerische Math.* **7**, 206-216.

Gulliksson, M. (1994). Iterative refinement for constrained and weighted linear least squares. *BIT* **34**, 239-253.

Gulliksson, M. (1995). Backward error analysis for the constrained and weighted linear least squares when using the weighted QR decomposition. *SIAM J. Matrix Anal. Appl.* **16**, 675-687.

Gulliksson, M. and Wedin, P-A (1992). Modifying the QR-decomposition to constrained and weighted linear least squares. *SIAM J. Matrix Anal. Appl.* **13**, 1298-1313.

Householder, A. S. (1958). Unitary triangularization of a nonsymmetric matrix. *J. ACM* **5**, 339-342.

Powell, M. J. D. and Reid, J. K. (1968). On applying Householder transformations to linear least squares problems. Report TP 322, Theoretical Physics Division, Harwell Laboratory. Reproduced as the appendix to this work.

Powell, M. J. D. and Reid, J. K. (1969). On applying Householder transformations to linear least squares problems. In Information Processing 68, Ed. Morrell, A.J.H., North-Holland, Amsterdam, New York, and London, 122-126.

Stewart, G. W. (1997). On the weighting method for least squares problems with linear equality constraints. *BIT* **37**, 961-967.

Appendix: Harwell report of Powell and Reid

In this appendix, we provide a typeset version of the report of Powell and Reid (1968), which is out of print.

T.P. 322

ON APPLYING HOUSEHOLDER TRANSFORMATIONS TO LINEAR LEAST SQUARES PROBLEMS

by

M. J. D. Powell (A. E. R. E. Harwell)

and

J. K. Reid (Mathematics Division, University of Sussex)

ABSTRACT

We derive some new error bounds for Golub's (1965) algorithm for calculating the least squares solution of an overdetermined system of linear equations, which are useful when the equations have widely differing weights. We show that improved accuracy can sometimes be obtained if Golub's algorithm is extended to include row interchanges.

Mathematics Branch,
Theoretical Physics Division,
Atomic Energy Research Establishent,
Harwell,
Berkshire,
England.

February, 1968.

1. Introduction

An excellent algorithm for calculating the least squares solution of the overdetermined system of linear equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, ..., m, \tag{1}$$

($m > n$) is described by Golub (1965), and an ALGOL listing is given by Businger and Golub (1965). It exploits the fact that the required solution is also the least squares solution of the system

$$QAx = Qb, \tag{2}$$

where Q is any $m \times m$ orthogonal matrix, by finding an orthogonal transformation that causes QA to be an upper triangular matrix. This upper triangular matrix is obtained by a sequence of n elementary transformations, which we write as

$$\left.\begin{aligned} A^{(1)} &= A \\ A^{(k+1)} &= P^{(k)} A^{(k)} \quad (k = 1, 2, ..., n) \\ QA &= A^{(n+1)} \end{aligned}\right\} \tag{3}$$

and the matrix $P^{(k)}$ is calculated so that all the elements of the first k columns of $A^{(k+1)}$ that are below the diagonal are zero. Each matrix $P^{(k)}$ is of the form

$$P^{(k)} = I - \beta_k u^{(k)} u^{(k)\,T}, \tag{4}$$

the orthogonality of $P^{(k)}$ being obtained by the condition

$$\beta_k \| u^{(k)} \|_2^2 = 2. \tag{5}$$

Wilkinson (1965) gives an error analysis of this type of calculation, and shows in his equation (45.3) on page 160 that the calculated components of QA differ from their true values by small multiples (depending on the precision of the computer) of $\|A\|_E$. One purpose of this paper is to extend Wilkinson's results, because they are not suitable for a situation that occurs frequently in data fitting problems. We are referring to the case when some of the data to be fitted are much more accurate than the remaining data, so, to take account of the difference in precision, some of the rows of A are scaled so that their elements are much larger than those of the remaining rows. In this case the value of the number $\|A\|_E$ is dominated by the large rows, but, if the number of very accurate observations is less than n, the required solution has an important dependence on the less precise data. Therefore we would prefer any error bounds or estimates to reflect the scaling of the rows of A; such bounds are derived in Sections 3, 4 and 5 of this paper.

In obtaining these bounds we find that the ordering of the columns of A is important; our results depend of the strategy that Golub recommends. A discussion of the ordering of both rows and

columns is given in Section 6, and it indicates that Golub's algorithm should be extended to include some row interchanges. Although this result is presented as a conclusion of the theoretical analysis, really the theoretical analysis is a consequence of the need for row interchanges, for the work in this paper was begun when Golub's algorithm failed on a real problem.

2. Golub's algorithm

We quote the details of Golub's algorithm that are needed for our error analysis.

The strategy for ordering the columns of the matrix A is applied before each elementary transformation $P^{(k)}$ is calculated. It depends on the numbers

$$\tau_j^{(k)} = \sum_{i=k}^{m} a_{ij}^{(k)\,2}, \quad j = k, k+1, ..., n, \tag{6}$$

and we let the largest be $\tau_q^{(k)}$. If $q = k$ then no interchanges take place, but otherwise the unknowns x_j are reordered so that the k-th and q-th columns of $A^{(k)}$ are interchanged. This process does not introduce any errors so, in order to simplify our notation, we suppose that the matrix A is such that no column interchanges are necessary. Therefore we have the inequalities

$$\sum_{i=k}^{m} a_{ik}^{(k)\,2} \geq \sum_{i=k}^{m} a_{ij}^{(k)\,2}, \quad j > k. \tag{7}$$

Next the transformation $P^{(k)}$ is applied both to $A^{(k)}$ and to the current right-hand side vector of the equations. The numbers

$$\sigma_k = \sqrt{\sum_{i=k}^{m} a_{ik}^{(k)\,2}} \tag{8}$$

and

$$\beta_k = \left(\sigma_k^2 + \sigma_k |a_{kk}^{(k)}|\right)^{-1} \tag{9}$$

are evaluated, and the components of $u^{(k)}$ (see equation (4)) are set to

$$\left.\begin{aligned} u_i^{(k)} &= 0, \quad i < k \\ u_k^{(k)} &= \left(\sigma_k + |a_{kk}^{(k)}|\right) \operatorname{sign}\left(a_{kk}^{(k)}\right) \\ u_i^{(k)} &= a_{ik}^{(k)}, \quad i > k. \end{aligned}\right\} \tag{10}$$

We obtain the numbers

$$y_j^{(k)} = \beta_k u^{(k)\,T} A_j^{(k)}, \quad j > k, \tag{11}$$

where the notation $A_j^{(k)}$ represents the j-th column of $A^{(k)}$, and we calculate the elements of $A^{(k+1)}$ from

the equations

$$\left.\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)},\ j<k \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)},\ i<k,\ j\geq k \\ a_{kk}^{(k+1)} &= -\sigma_k \operatorname{sign}\left(a_{kk}^{(k)}\right) \\ a_{ik}^{(k+1)} &= 0,\ i>k \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - u_i^{(k)} y_j^{(k)},\ i\geq k,\ j>k. \end{aligned}\right\} \tag{12}$$

For the right-hand sides of the equations we let

$$\left.\begin{aligned} b^{(1)} &= b \\ b^{(k+1)} &= P^{(k)}\, b^{(k)}. \end{aligned}\right\} \tag{13}$$

To calculate $b^{(k+1)}$ we obtain the number

$$\gamma_k = \beta_k u^{(k)\,T} b^{(k)}, \tag{14}$$

and then use the equations

$$\left.\begin{aligned} b_i^{(k+1)} &= b_i^{(k)},\ i<k \\ b_i^{(k+1)} &= b_i^{(k)} - \gamma_k u_i^{(k)},\ i\geq k. \end{aligned}\right\} \tag{15}$$

Thus, by applying the sequence of transformations $P^{(1)}$, $P^{(2)}$, ..., $P^{(n)}$, we obtain a system of linear equations

$$\sum_{j=1}^{n} a_{ij}^{(n+1)} x_j = b_i^{(n+1)},\ i=1,2,...,m,$$

having an upper triangular matrix. The equations indicated by the values $i = n+1, n+2, ..., m$ are ignored, and the required vector x is obtained by back substitution.

To conclude this section we derive a result that indicates why Golub's strategy for interchanging columns is important. It is a bound on the numbers $y_j^{(k)}$, defined by equation (11), and we will use it many times in the error analysis.

Theorem 1

$$|y_j^{(k)}| \leq \sqrt{2}. \tag{16}$$

Proof By applying Schwartz's inequality to the definition of $y_j^{(k)}$, and by using the statements (7), (8) and (5) we obtain the inequality

$$|y_j^{(k)}| \le \beta_k \|u^{(k)}\|_2 \sqrt{\sum_{i=k}^{m} a_{ij}^{(k)\,2}}$$

$$\le \sigma_k \beta_k \|u^{(k)}\|_2$$

$$= \frac{2\sigma_k}{\|u^{(k)}\|_2}. \tag{17}$$

Also from the definitions (10) and (8) we have the identity

$$\|u^{(k)}\|_2^2 = 2\sigma_k \left(\sigma_k + |a_{kk}^{(k)}|\right), \tag{18}$$

which implies the inequality

$$\|u^{(k)}\|_2 \ge \sqrt{2}\sigma_k. \tag{19}$$

Therefore the theorem is an immediate consequence of the inequality (17). □

Note that for $j > k$ Golub's algorithm includes the equation

$$A_j^{(k+1)} = A_j^{(k)} - y_j^{(k)} u^{(k)}, \tag{20}$$

so the theorem bounds the multiple of $u^{(k)}$ that is added to the j-th column of $A^{(k)}$. An operation like equation (20) can cause any errors to grow persistently as k ranges from 1 to n, so the column interchanges are justified by the fact that they limit the multipliers $y_j^{(k)}$.

3. The errors of the transformation $P^{(k)}$

In the error analysis we assume that we are using a floating-point computer on which all the operations of addition, subtraction, multiplication, division, extraction of square roots and rounding to single precision are performed with relative errors no greater than ε, and we omit terms of order ε^2. We also assume that a subroutine is available for the double-length accumulation of inner products, followed by roundoff to single precision. To distinguish computed numbers from those that would result from exact arithmetic, we attach "bars" to numbers that are calculated; for example $\bar{\sigma}_k$ and $\bar{u}^{(k)}$ are computed quantities.

The intention of Sections 3, 4, and 5 is to bound the total error of the calculation in a way that reflects the scaling of the rows of A. We state our results in terms of the greatest numbers that occur in each row of A during the operation of the algorithm, namely

$$\alpha_i = \max_{j,k} |\bar{a}_{ij}^{(k)}|, \quad i = 1, 2, ..., m. \tag{21}$$

We find that the calculated vector x is the exact least squares solution of a system of linear equations

that is little different from the system (1), and we bound the differences between corresponding matrix elements a_{ij} by multiples of α_i. In fact these multipliers are of order n^2, which is pleasing because equations (16) and (20) suggest that we might have had terms of order $(1+\sqrt{2})^n$.

In this section we bound the errors in $A^{(k+1)}$ that are caused by the calculation when the matrix $A^{(k)}$ is exact, and we use the notation

$$\Delta^{(k)} = \bar{A}^{(k+1)} - A^{(k+1)}. \tag{22}$$

The effect of errors in $A^{(k)}$ is treated in Section 4.

By using the double-length scalar product routine, we calculate σ_k^2 (see equation (8)) to a relative accuracy of ε, so, because a square root halves a relative error, we obtain the result

$$|\bar{\sigma}_k - \sigma_k| \le \tfrac{3}{2}\varepsilon\sigma_k. \tag{23}$$

The error in β_k^{-1} (see equation (9)) is bounded by the inequality

$$|\bar{\beta}_k^{-1} - \beta_k^{-1}| \le \varepsilon\beta_k^{-1} + \varepsilon\sigma_k^2 + \varepsilon\sigma_k|a_{kk}^{(k)}| + \tfrac{3}{2}\varepsilon\sigma_k|a_{kk}^{(k)}| = \varepsilon\sigma_k\left(2\sigma_k + \tfrac{7}{2}|a_{kk}^{(k)}|\right), \tag{24}$$

and for the error in $u_k^{(k)}$ (see equation (10)) we have the bound

$$|\bar{u}_k^{(k)} - u_k^{(k)}| \le \varepsilon|u_k^{(k)}| + \tfrac{3}{2}\varepsilon\sigma_k = \varepsilon\left(\tfrac{5}{2}\sigma_k + |a_{kk}^{(k)}|\right). \tag{25}$$

However in bounding the error in $y_j^{(k)}$ we depart from the ALGOL listing of Businger and Golub (1965), because we can gain some accuracy by dividing by β_k^{-1} in expression (11), instead of calculating β_k and multiplying. Thus, using the inner product routine, the error in $y_j^{(k)}$ is at most

$$|\bar{y}_j^{(k)} - y_j^{(k)}| \le \varepsilon|y_j^{(k)}|\left\{2 + \beta_k\sigma_k\left(2\sigma_k + \tfrac{7}{2}|a_{kk}^{(k)}|\right)\right\} + \varepsilon\beta_k\left(\tfrac{5}{2}\sigma_k + |a_{kk}^{(k)}|\right)|a_{kj}^{(k)}|, \tag{26}$$

the second term being a consequence of the error in $u^{(k)}$. Therefore, from Theorem 1, the inequality (7) and the definition (8), we obtain the result

$$|\bar{y}_j^{(k)} - y_j^{(k)}| \le 2\sqrt{2}\varepsilon + \varepsilon\beta_k\sigma_k\left\{\left(2\sqrt{2} + \tfrac{5}{2}\right)\sigma_k + \left(\tfrac{7}{2}\sqrt{2} + 1\right)|a_{kk}^{(k)}|)\right\}, \tag{27}$$

which, because of the inequality

$$|a_{kk}^{(k)}| \le \sigma_k \tag{28}$$

and the definition (9), gives the very simple bound

$$|\bar{y}_j^{(k)} - y_j^{(k)}| \le \varepsilon\left(\tfrac{19}{4}\sqrt{2} + \tfrac{7}{4}\right). \tag{29}$$

We now obtain bounds for the elements of the error matrix $\Delta^{(k)}$ (see equation (22)), and find immediately from equation (12) that several elements are zero:

$$\Delta_{ij}^{(k)} = 0 \begin{cases} j < k \\ i < k, \ j \ge k \\ i > k, \ j = k. \end{cases} \tag{30}$$

Also equation (23) gives the result

$$|\Delta_{kk}^{(k)}| \leq \tfrac{3}{2}\varepsilon\sigma_k, \tag{31}$$

and from equation (29) and Theorem 1 we obtain (for $i>k$, $j>k$) the bound

$$|\Delta_{ij}^{(k)}| \leq \varepsilon|a_{ij}^{(k+1)}|+\sqrt{2}\varepsilon|u_i^{(k)}|+\varepsilon\left(\tfrac{19}{4}\sqrt{2}+\tfrac{7}{4}\right)|u_i^{(k)}| = \varepsilon\left\{|a_{ij}^{(k+1)}|+\left(\tfrac{23}{4}\sqrt{2}+\tfrac{7}{4}\right)|u_i^{(k)}|\right\} \tag{32}$$

For the case $i=k$, $j>k$ there is also an error from $u_i^{(k)}$, so using equation (25) as well we find the inequality

$$|\Delta_{kj}^{(k)}| \leq \varepsilon\left\{|a_{kj}^{(k+1)}|+\left(\tfrac{23}{4}\sqrt{2}+\tfrac{7}{4}\right)|u_k^{(k)}|\right\}+\varepsilon\sqrt{2}\left(\tfrac{5}{2}\sigma_k+|a_{kk}^{(k)}|\right). \tag{33}$$

We express the results (31) – (33) in terms of the numbers α_i, defined by equation (21), using the inequalities

$$\sigma_k \leq \alpha_k\ , \tag{34}$$

(derived from the expression for $a_{kk}^{(k+1)}$ in equation (12)),

$$|u_i^{(k)}| \leq \alpha_i\ ,\ i>k \tag{35}$$

(derived from equation (10)), and

$$|u_k^{(k)}| \leq 2\alpha_k, \tag{36}$$

which follows from the statements (10) and (34). Thus we obtain the bounds

$$\left.\begin{aligned} |\Delta_{kk}^{(k)}| &\leq \tfrac{3}{2}\varepsilon\alpha_k \\ |\Delta_{kj}^{(k)}| &\leq \varepsilon\left(15\sqrt{2}+\tfrac{9}{2}\right)\alpha_k\ ,\ j>k \\ |\Delta_{ij}^{(k)}| &\leq \varepsilon\left(\tfrac{23}{4}\sqrt{2}+\tfrac{11}{4}\right)\alpha_i\ ,\ i>k,\ j>k. \end{aligned}\right\} \tag{37}$$

In the next section we also require bounds on the numbers $\|\Delta_j^{(k)}\|_2$, where again the single subscript indicates a column of a matrix. From equations (30), (31) and (19) we obtain the results

$$\left.\begin{aligned} \|\Delta_j^{(k)}\|_2 &= 0\ ,\ j<k \\ \|\Delta_k^{(k)}\|_2 &\leq \tfrac{3}{4}\sqrt{2}\varepsilon\|u^{(k)}\|_2\ , \end{aligned}\right\} \tag{38}$$

but it is more difficult to derive our bound for $j>k$. We use equations (30), (32), (33) and (28) to calculate the inequality

$$\|\Delta_j^{(k)}\|_2 \leq \varepsilon\left[\sqrt{\sum_{i=k}^{m} a_{ij}^{(k+1)\,2}}+\left(\tfrac{23}{4}\sqrt{2}+\tfrac{7}{4}\right)\|u^{(k)}\|_2\right]+\tfrac{7}{2}\sqrt{2}\varepsilon\sigma_k, \tag{39}$$

and from the fact that Euclidean norms are invariant under orthogonal transformations, and from equations (7), (8) and (19), we find the bound

$$\|\Delta_j^{(k)}\|_2 \le \varepsilon\left\{\left(\tfrac{23}{4}\sqrt{2}+\tfrac{7}{4}\right)\|u^{(k)}\|_2+\left(\tfrac{7}{2}\sqrt{2}+1\right)\sigma_k\right\}$$
$$\le \varepsilon\left(\tfrac{25}{4}\sqrt{2}+\tfrac{21}{4}\right)\|u^{(k)}\|_2\,,\ j>k. \qquad (40)$$

This completes the analysis of a single transformation $P^{(k)}$, and we use the results (30), (37), (38) and (40) to obtain the bounds for the whole calculation.

4. The error of the sequence of transformations

In our notation for the analysis of the errors of the sequence of transformations $P^{(n)}\ P^{(n-1)}\ \ldots\ P^{(1)}$, we make a slight change from the nomenclature of the last section. Now we let $P^{(k)}$ (k =1, 2, ..., n) be the orthogonal transformation that would be obtained from the computed matrix $\bar{A}^{(k)}$ if exact arithmetic were used, and in place of equation (22) we write

$$\Delta^{(k)} = \bar{A}^{(k+1)} - P^{(k)}\bar{A}^{(k)}. \qquad (41)$$

The purpose of this section is to bound the elements of a matrix Δ having the property that the final computed matrix $\bar{A}^{(n+1)}$ would be obtained by an exact application of the algorithm to the overdetermined system of linear equations

$$(A+\Delta)x = b+\delta. \qquad (42)$$

Therefore Δ is related to A and $\bar{A}^{(n+1)}$ by the equation

$$A+\Delta = \Omega\bar{A}^{(n+1)}, \qquad (43)$$

where Ω is an exactly orthogonal matrix. Different choices of Ω provide different error matrices Δ, and, in order that Δ is zero if the calculation of $\bar{A}^{(n+1)}$ is exact, we define Δ by the equation

$$A+\Delta=\left\{P^{(n)}\ P^{(n-1)}\ \ldots P^{(1)}\right\}^{-1}\bar{A}^{(n+1)}$$
$$=P^{(1)}\ P^{(2)}\ \ldots P^{(n)}\bar{A}^{(n+1)}, \qquad (44)$$

the last line being a consequence of the symmetry of $P^{(k)}$. It is possible that the error bounds of this section can be improved by a different choice of Ω.

To bound the elements of Δ we use equation (41) to express the right-hand side of equation (44) in terms of A and $\Delta^{(k)}$ (k = 1, 2, ..., n), which gives the identity

$$A+\Delta=P^{(1)}\ P^{(2)}\ \ldots P^{(n)}\Delta^{(n)}+P^{(1)}\ P^{(2)}\ \ldots P^{(n-1)}\bar{A}^{(n)}$$
$$=\ldots$$
$$=\sum_{k=1}^{n}P^{(1)}\ P^{(2)}\ \ldots P^{(k)}\Delta^{(k)}+\bar{A}^{(1)}, \qquad (45)$$

from which we deduce the equation

$$\Delta = \sum_{k=1}^{n} P^{(1)}\, P^{(2)} \ldots P^{(k)} \Delta^{(k)}. \tag{46}$$

Our results for the total error from the sequence of transformations are obtained from equation (46) and the inequalities of the last section.

In the *k*-th term of the sum (46), we substitute expression (4) in place of some of the orthogonal matrices, obtaining the result

$$\begin{aligned} P^{(1)}\, P^{(2)} \ldots P^{(k)}\, \Delta^{(k)} &= \left\{ I - \beta_1 u^{(1)} u^{(1)\,T} \right\} \left\{ P^{(2)}\ P^{(3)} \ldots P^{(k)} \Delta^{(k)} \right\} \\ &= P^{(2)}\, P^{(3)} \ldots P^{(k)} \Delta^{(k)} - \beta_1 u^{(1)} u^{(1)\,T} P^{(2)}\, P^{(3)} \ldots P^{(k)} \Delta^{(k)} \\ &= \ldots \\ &= \Delta^{(k)} - \sum_{q=1}^{k} \beta_q u^{(q)} u^{(q)\,T} P^{(q+1)}\, P^{(q+2)} \ldots P^{(k)} \Delta^{(k)}. \end{aligned} \tag{47}$$

Thus, using equation (5) and the orthogonality of the transformations $P^{(k)}$, we deduce the inequality

$$\begin{aligned} |P^{(1)}\, P^{(2)} \ldots P^{(k)}\, \Delta^{(k)}|_{ij} &\le |\Delta^{(k)}|_{ij} + \sum_{q=1}^{k} \beta_q |u_i^{(q)}| \left| u^{(q)\,T} P^{(q+1)}\, P^{(q+2)} \ldots P^{(k)} \Delta_j^{(k)} \right| \\ &\le |\Delta^{(k)}|_{ij} + \sum_{q=1}^{k} \beta_q |u_i^{(q)}|\, \|u^{(q)}\|_2 \left\| P^{(q+1)}\, P^{(q+2)} \ldots P^{(k)} \Delta_j^{(k)} \right\|_2 \\ &= |\Delta^{(k)}|_{ij} + 2\|\Delta_j^{(k)}\|_2 \sum_{q=1}^{k} |u_i^{(q)}| / \|u^{(q)}\|_2. \end{aligned} \tag{48}$$

We simplify the inequality (48) by removing the term $\|u^{(q)}\|_2$ from the summation. We do this by noting that inequality (7) gives the result

$$\sigma_q \ge \sigma_k\,,\ q < k, \tag{49}$$

so from statement (19) we deduce the inequality

$$\|u^{(q)}\|_2 \ge \sqrt{2}\sigma_k \ge \|u^{(k)}\|_2 / \sqrt{2}\,,\ q < k, \tag{50}$$

the last line being a consequence of equations (8) and (18). Thus from expression (48) we find the bound

$$\left| P^{(1)}\, P^{(2)} \ldots P^{(k)} \Delta^{(k)} \right|_{ij} \le |\Delta_{ij}^{(k)}| + \lambda_i^{(k)} \alpha_i \|\Delta_j^{(k)}\|_2 / \|u^{(k)}\|_2, \tag{51}$$

where $\lambda_i^{(k)}$ is defined by the equation

$$\alpha_i \lambda_i^{(k)} = 2\sqrt{2} \sum_{q=1}^{k-1} |u_i^{(q)}| + 2|u_i^{(k)}|; \tag{52}$$

equations (10), (21) and (36) give the result

$$\lambda_i^{(k)} \le \begin{cases} 2\sqrt{2}(i+1), & i<k \\ 2\sqrt{2}(k-1)+4, & i=k \\ 2\sqrt{2}(k-1)+2, & i>k. \end{cases} \tag{53}$$

To summarize the inequalities that we have obtained so far, we combine expressions (46), (51), (30), (37), (38) and (40) and write

$$|\Delta_{ij}| \le (U_{ij}+V_{ij})\varepsilon\alpha_i, \tag{54}$$

where

$$U_{ij} = \begin{cases} \left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)(i-1)+\left(15\sqrt{2}+\frac{9}{2}\right), & i<j \\ \left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)(i-1)+\frac{3}{2}, & i=j \\ \left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)j, & i>j, \end{cases} \tag{55}$$

and

$$V_{ij} = \tfrac{3}{4}\sqrt{2}\,\lambda_i^{(j)} + \left(\tfrac{25}{4}\sqrt{2}+\tfrac{21}{4}\right)\sum_{k=1}^{j-1}\lambda_i^{(k)}. \tag{56}$$

The theorem of this section states that $|\Delta_{ij}|$ is not greater than a certain multiple of $\varepsilon\alpha_i$, and for simplicity the multiplier is independent of i and j. Therefore we now seek the best value of this multiplier that can be obtained from expressions (53), (54), (55) and (56).

Clearly both U_{ij} and V_{ij} are greatest when $j = n$, but the value of i that yields the required multiplier is not obvious. However it is apparent that U_{in} is an increasing function of i for $i \le n-1$, and it is not difficult to show that if expression (53) is an equality then V_{in} is an increasing function of i for $i \le n-2$. Therefore we consider the details of four separate cases and derive the results

$$(U_{in}+V_{in}) \le \begin{cases} \phi(n)+\left(\frac{131}{4}+\frac{25}{4}\sqrt{2}\right), & i=n-2 \\ \phi(n)+\left(24+14\sqrt{2}\right), & i=n-1 \\ \phi(n)+\left(\frac{41}{4}-\frac{19}{4}\sqrt{2}\right), & i=n \\ \phi(n)+\left(\frac{23}{2}-\frac{1}{2}\sqrt{2}\right), & i>n \end{cases} \tag{57}$$

where

$$\phi(n) = n^2\left(\tfrac{25}{2}+\tfrac{21}{4}\sqrt{2}\right)+n\left(-\tfrac{85}{4}+\tfrac{5}{2}\sqrt{2}\right). \tag{58}$$

Thus we obtain the theorem

Theorem 2

To first order in ε the accumulation of errors in calculating $\bar{A}^{(n+1)}$ by Golub's algorithm is so small that the elements of the matrix Δ, defined by the equation

$$\bar{A}^{(n+1)} = \{P^{(n)} P^{(n-1)} \ldots P^{(1)}\}(A+\Delta), \tag{59}$$

are bounded by the inequality

$$|\Delta_{ij}| \leq \left\{n^2\left(\tfrac{25}{2}+\tfrac{21}{4}\sqrt{2}\right) - n\left(\tfrac{85}{4}-\tfrac{5}{2}\sqrt{2}\right) + (24+14\sqrt{2})\right\}\varepsilon\alpha_i. \tag{60}$$

5. The error in the solution of the equations

To complete the error analysis we must consider the sequence of calculated right-hand side vectors $\bar{b}^{(1)}, \bar{b}^{(2)}, \ldots, \bar{b}^{(n+1)}$ (see equation (15)), and we must treat the back-substitution stage of the algorithm, in which x is determined from the equation

$$\bar{A}^{(n+1)} x = \bar{b}^{(n+1)}. \tag{61}$$

If it happens that $\|b\|_2$ is so small that both the inequalities

$$\mu_i = \max_k |b_i^{(k)}| \leq \alpha_i\,, \quad i = 1, 2, \ldots, m \tag{62}$$

and

$$\nu_k = \sqrt{\sum_{i=k}^{m} b_i^{(k)\,2}} \leq \sigma_k\,, \quad k = 1, 2, \ldots, n \tag{63}$$

hold, then we have already carried out much of the analysis of the errors of the vectors $\bar{b}^{(k)}$, because we can regard $b^{(k)}$ as an additional column of $A^{(k)}$. However if the number

$$\rho = \max\left[\max_i \frac{\mu_i}{\alpha_i}, \max_k \frac{\nu_k}{\sigma_k}\right] \tag{64}$$

exceeds one, then to make the inequalities (62) and (63) hold we could scale the original right-hand side vector b by the factor ρ^{-1}. As a result the numbers γ_k (see equation (14)) and $b_i^{(k+1)}$ would be scaled by ρ^{-1}, and the size of any errors in the vectors $b^{(k)}$ would also be scaled by the same amount. Therefore, instead of carrying out this scaling, we may anticipate its effect by including the factor ρ in our error bounds. For example, using the definition

$$\delta^{(k)} = P^{(k)}\bar{b}^{(k)} - \bar{b}^{(k+1)}, \tag{65}$$

we obtain from equations (30) and (37) the bounds

$$|\delta_i^{(k)}| \le \begin{cases} 0, & i<k \\ \varepsilon\rho\left(15\sqrt{2}+\frac{9}{2}\right)\alpha_k, & i=k \\ \varepsilon\rho\left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)\alpha_i, & i>k, \end{cases} \tag{66}$$

and from expression (40) we find the inequality

$$\|\delta^{(k)}\|_2 \le \rho\left(\frac{25}{4}\sqrt{2}+\frac{21}{4}\right)\|u^{(k)}\|_2. \tag{67}$$

To calculate bounds on the components of δ, defined in equation (42), we find by the argument that led to equation (46) the result

$$\delta = \sum_{k=1}^{n} P^{(1)}P^{(2)}\ldots P^{(k)}\delta^{(k)}, \tag{68}$$

and instead of equation (51) we obtain the bound

$$\left|P^{(1)}P^{(2)}\ldots P^{(k)}\delta^{(k)}\right|_i \le |\delta_i^{(k)}| + \lambda_i^{(k)}\alpha_i\|\delta^{(k)}\|_2/\|u^{(k)}\|_2. \tag{69}$$

Therefore the inequality corresponding to statement (54) is

$$|\delta_i| \le (S_i + T_i)\varepsilon\rho\alpha_i, \tag{70}$$

where

$$S_i = \begin{cases} \left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)(i-1) + \left(15\sqrt{2}+\frac{9}{2}\right), & i\le n \\ \left(\frac{23}{4}\sqrt{2}+\frac{11}{4}\right)n, & i>n \end{cases} \tag{71}$$

and

$$T_i = \left(\frac{25}{4}\sqrt{2}+\frac{21}{4}\right)\sum_{k=1}^{n}\lambda_i^{(k)}. \tag{72}$$

Again it happens that our final bound is derived from the case $i = n-1$, and we calculate that the elements of the vector δ are bounded by the inequality

$$|\delta_i| \le \left\{n^2\left(\frac{25}{2}+\frac{21\sqrt{2}}{4}\right) + n\left(\frac{3}{4}+13\sqrt{2}\right) + (24+14\sqrt{2})\right\}\varepsilon\rho\alpha_i. \tag{73}$$

Wilkinson (1965) gives the error analysis of a back-substitution process on pages 247 and 248 of his book. From his work we conclude that the computed solution of the equations (61) is the exact least squares solution of a system

$$(\bar{A}^{(n+1)} + E)\,x = \bar{b}^{(n+1)}, \tag{74}$$

where, to first order in ε, the elements of E are bounded by the inequality

$$|E_{ij}| \le \varepsilon|\bar{a}_{ii}^{(n+1)}|\,\delta_{ij}, \tag{75}$$

δ_{ij} being the Kronecker delta.

We absorb these errors into our analysis by supposing that each orthogonal transformation $P^{(k)}$

causes an extra error of $\varepsilon\sigma_k$ in the matrix element $\bar{a}_{kk}^{(k+1)}$. Therefore in the inequalities (37) and (38) the case $j = k$ becomes

$$\left.\begin{aligned} |\Delta_{kk}^{(k)}| &\le \tfrac{5}{2}\varepsilon\alpha_k \\ \|\Delta_k^{(k)}\|_2 &\le \tfrac{5}{4}\sqrt{2}\varepsilon\|u^{(k)}\|_2, \end{aligned}\right\} \tag{76}$$

so, instead of the middle line of expression (55) and expression (56), we now have the equations

$$\left.\begin{aligned} U_{ii} &= \left(\tfrac{23}{4}\sqrt{2}+\tfrac{11}{4}\right)(i-1)+\tfrac{5}{2}, \\ V_{ij} &= \tfrac{5}{4}\sqrt{2}\lambda_i^{(j)} + \left(\tfrac{25}{4}\sqrt{2}+\tfrac{21}{4}\right)\sum_{k=1}^{j-1}\lambda_i^{(k)}. \end{aligned}\right\} \tag{77}$$

More calculation shows that in the new bounds that replace the inequality (57), the case $i = n-1$ remains dominant. Thus we obtain the main theorem of the error analysis.

Theorem 3

The calculated vector $\bar{x}$ obtained by Golub's algorithm is the exact least squares solution of a system of equations

$$(A+\Delta)\,x = b+\delta, \tag{78}$$

where the elements of δ are bounded by the inequality (73), and where the elements of Δ are bounded by the inequality

$$|\Delta_{ij}| \le \left\{ n^2\left(\tfrac{25}{2}+\tfrac{21\sqrt{2}}{4}\right) - n\left(\tfrac{77}{4}-\tfrac{5\sqrt{2}}{2}\right) + (24+14\sqrt{2}) \right\}\varepsilon\alpha_i\,, \tag{79}$$

α_i being defined by equation (21).

6. The need for row interchanges

As we said in the introduction, the theorems we have given were derived because Golub's algorithm failed on a real problem, so the main purpose of this section is to recommend a modification to the algorithm. This modification is a strategy for interchanging rows of the matrix $A^{(k)}$, and we note that the theorems proved so far do not depend on any particular ordering of the rows.

The fact that Golub's algorithm will sometimes give poor accuracy is illustrated by the matrix

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 10^6 & 10^6 & 0 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{pmatrix}. \tag{80}$$

Using exact arithmetic we calculate that $A^{(2)}$ is the matrix

$$A^{(2)} = \begin{pmatrix} -10^6\sqrt{2} & -10^6/\sqrt{2} & -10^6/\sqrt{2} \\ 0 & \frac{1}{2}10^6 - \sqrt{2} & -\frac{1}{2}10^6 - 1/\sqrt{2} \\ 0 & -\frac{1}{2}10^6 - \sqrt{2} & \frac{1}{2}10^6 - 1/\sqrt{2} \\ 0 & 1 & 1 \end{pmatrix}. \tag{81}$$

However, if five-decimal floating-point computation is used, the terms $-\sqrt{2}$ and $-1/\sqrt{2}$ in the second and third rows are lost, which is equivalent to the loss of all the information present in the first row of *A*. This loss of information is disastrous because the number of rows of *A* containing large elements is less than the number of components of *x*, so there is a substantial dependence of the required vector on the first and fourth rows of *A*.

Theorem 3 shows that Golub's algorithm would have worked well if the numbers α_1, α_2, α_3 and α_4, were of an acceptable size, but in the case of the example

$$\alpha_1 = 10^6\sqrt{2}, \tag{82}$$

which is much larger than the elements in the first row of *A*. Therefore the theorem suggests, correctly, that there may be loss of accuracy. It also shows that the difficulty would not occur if we can prevent the elements of every row of $A^{(k+1)}$ from being much larger than those of the corresponding rows of $A^{(k)}$; fortunately we can achieve this aim by making some row interchanges.

Already from equations (10) and (20) we have the result

$$\max_j |a_{ij}^{(k+1)}| = \max_j |a_{ij}^{(k)}|, \ i<k, \tag{83}$$

and from Theorem 1 and equations (10) and (20) we deduce the inequality

$$\max_j |a_{ij}^{(k+1)}| \le (\sqrt{2}+1) \max_j |a_{ij}^{(k)}|, \ i>k. \tag{84}$$

Therefore just the *k*-th row of $A^{(k+1)}$ is critical. We ensure that it is not much larger than the *k*-th row of $A^{(k)}$ by exploiting the following theorem:

Theorem 4

If the inequality

$$|a_{kk}^{(k)}| \ge |a_{ik}^{(k)}|, \ i>k, \tag{85}$$

holds, then we have the bound

$$\max_j |a_{kj}^{(k+1)}| \le \sqrt{m} \max_j |a_{kj}^{(k)}|. \tag{86}$$

Proof. Since $P^{(k)}$ is an orthogonal transformation that leaves the first (k–1) components of a vector unchanged, we find the inequality

$$|a_{kj}^{(k+1)}| \le \sqrt{\sum_{i=k}^{m} a_{ij}^{(k+1)\,2}} = \sqrt{\sum_{i=k}^{m} a_{ij}^{(k)\,2}} \,. \tag{87}$$

All the components of the sum are zero if $j<k$, so from equations (7) and (8) we obtain the result

$$|a_{kj}^{(k+1)}| \le \sigma_k \,. \tag{88}$$

From statements (8) and (85) the inequality

$$\sigma_k \le \sqrt{m}\,|a_{kk}^{(k)}| \tag{89}$$

holds, so the theorem is a consequence of statement (88). □

Therefore the modification that we recommend just provides the inequality (85). After the columns of $A^{(k)}$ have been ordered for the calculation of $P^{(k)}$, we obtain the largest number in the sequence $\{|a_{kk}^{(k)}|, |a_{k+1\,k}^{(k)}|, ..., |a_{mk}^{(k)}|\}$, say it is $|a_{qk}^{(k)}|$, and we interchange the k-th and q-th rows of $A^{(k)}$ if $q \ne k$.

Thus in place of the matrices (80) and (81) we have

$$A = \begin{pmatrix} 10^6 & 10^6 & 0 \\ 0 & 2 & 1 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{pmatrix} \tag{90}$$

and

$$A^{(2)} = \begin{pmatrix} -10^6\sqrt{2} & -10^6/\sqrt{2} & -10^6/\sqrt{2} \\ 0 & 2 & 1 \\ 0 & -10^6/\sqrt{2} & 10^6/\sqrt{2} \\ 0 & 1 & 1 \end{pmatrix}, \tag{91}$$

so the previous loss of accuracy is avoided.

In the modified algorithm the inequalities (83), (84) and (86) provide the bound

$$\alpha_i \le (1+\sqrt{2})^{n-1}\sqrt{m}\,\max_j |a_{ij}^{(1)}|\,, \tag{92}$$

but if this bound is attained and n is large, Theorem 3 is not very useful. Therefore we carried out some numerical experiments to estimate typical values of the ratio

$$\beta = \max_i \left[\alpha_i \,/ \max_j |a_{ij}^{(1)}| \right]. \tag{93}$$

We used one hundred 20×10 matrices whose elements were

$$a_{ij} = 10^{10p_i} q_{ij}\,, \tag{94}$$

where p_i and q_{ij} are pseudo-random numbers from the distribution that is uniform over $[-1, 1]$. We found that in all cases the value of the ratio (93) was less than five, so it seems that the error bounds are sufficiently small to be useful in many real calculations. However the last row of the pathological matrix

$$\begin{bmatrix}
1 & -0.99 & -0.99 & -0.99 & . & . & . & -0.99 & -0.99 \\
0 & 0.1 & -0.099 & -0.099 & . & . & . & -0.099 & -0.099 \\
0 & 0 & 0.01 & -0.0099 & . & . & . & -0.0099 & -0.0099 \\
0 & 0 & 0 & 0.001 & . & . & . & -0.00099 & -0.00099 \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
0 & 0 & 0 & 0 & . & . & . & 10^{-n+2} & -0.99\times 10^{-n+2} \\
10^{-n-10} & 10^{-n-10} & 10^{-n-10} & 10^{-n-10} & . & . & . & 10^{-n-10} & 10^{-n-10}
\end{bmatrix}$$

shows that the ratio can approach the value 2^n.

The same matrices (94) were also used to try a different strategy for column interchanges, namely to arrange the columns so that in place of statement (7) we have the inequality

$$\sqrt{\sum_{i=k}^{m} a_{ik}^{(k)\,2}} + \max_{k\le i\le m} |a_{ik}^{(k)}| \ge \sqrt{\sum_{i=k}^{m} a_{ij}^{(k)\,2}} + \max_{k\le i\le m} |a_{ij}^{(k)}|\,,\ j>k\,, \tag{95}$$

but the results did not justify the extra work required to follow this alternative. The reason we tried it is that if inequality (95) holds, and if the recommended row interchanges are made, then in place of Theorem 1 we can derive the result

$$|y_j^{(k)}| \le 1\,, \tag{96}$$

so the theoretical results corresponding to inequalities (60), (73), (79) and (92) would contain smaller numbers.

To complete this paper we must remark on the importance of the scaling of the columns of the matrix A. The point to notice is that the error bounds of Theorems 2 and 3 are moderate multiples of the numbers $\varepsilon\alpha_i$, and α_i is governed by the largest elements of the i-th row of the matrices $\bar{A}^{(1)}$, $\bar{A}^{(2)}$, ..., $\bar{A}^{(n+1)}$. Therefore if x_j is scaled so that for $i = 1, 2, ..., m$ the element a_{ij} is much smaller than the other elements of the i-th row of A, then the bounds on Δ_{ij} will be rather unsatisfactory. Careful scaling of columns can avoid this happening, and before applying Golub's algorithm the variables x_j should be chosen so that the n numbers

$$\max_i \frac{|a_{ij}|}{\max_k |a_{ik}|}, \quad j = 1, 2, ..., n, \tag{97}$$

are all close to one.

Remember that in a least squares problem there is no freedom to scale the separate rows of *A*, which is the motivation for the character of our error bounds.

References

Businger, P. A. and Golub, G. H. (1965). Linear least squares solutions by Householder transformations. *Numerische Math.* **7**, 269-276.

Golub, G. H. (1965). Numerical methods for solving linear least squares problems. *Numerische Math.* **7**, 206-216.

Wilkinson, J. H. (1965). *The algebraic eigenvalue problem.* Oxford University Press, London.