Library
Bedford

UNCLASSIFIED

(Approved for Publication)

T.P. 322

Copy

"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)

A.E.R.E.
LIBRARY

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

February, 1968.
HL68/1628

16087
2025

UNCLASSIFIED
(Approved for Publication)

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 Establishment,
Harwell,
Berkshire,
England.

February, 1968.
HL68/1628

1. Introduction

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

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

($m > n$) 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.

$$QA \underset{\sim}{x} = Q\underset{\sim}{b}, \qquad \ldots\ldots\ldots\ldots(2)$$

where Q is any mxm 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:

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

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 equal to zero. Each matrix $P^{(k)}$ is of the form

$$P^{(k)} = I - \beta_k \underset{\sim}{u}^{(k)} \underset{\sim}{u}^{(k)T}, \qquad \ldots\ldots\ldots\ldots(4)$$

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

$$\beta_k \left\| \underset{\sim}{u}^{(k)} \right\|_2^2 = 2. \qquad \ldots\ldots\ldots\ldots(5)$$

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 is 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 on 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

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

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

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

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

$$\sigma_k = \{ \sum_{i=k}^{m} (a_{ik}^{(k)})^2 \}^{\frac{1}{2}} \qquad \ldots\ldots\ldots(8)$$

and

$$\beta_k = (\sigma_k^2 + \sigma_k |a_{kk}^{(k)}|)^{-1} \qquad \text{.....(9)}$$

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

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

We obtain the numbers

$$y_j^{(k)} = \beta_k \, \underset{\sim}{u}^{(k)T} A_j^{(k)} \quad , \quad j > k, \qquad \text{.....(11)}$$

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

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

For the right-hand sides of the equations we let

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

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

$$\gamma_k = \beta_k \, \underset{\sim}{u}^{(k)T} \underset{\sim}{b}^{(k)}, \qquad \text{.....(14)}$$

and then we use the equations

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

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

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

having an upper triangular matrix. The equations indicated by the values $i = n + 1, n + 2,\ldots, m$ are ignored, and the required vector $\underset{\sim}{x}$ 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

$$\left| y_j^{(k)} \right| \leq \sqrt{2}. \qquad \text{.....(16)}$$

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

$$\begin{aligned} \left| y_j^{(k)} \right| &\leq \beta_k \left\| \underset{\sim}{u}^{(k)} \right\|_2 \left\{ \sum_{i=k}^{m} (a_{ij}^{(k)})^2 \right\}^{\frac{1}{2}} \\ &\leq \sigma_k \beta_k \left\| \underset{\sim}{u}^{(k)} \right\|_2 \\ &= 2\sigma_k \Big/ \left\| \underset{\sim}{u}^{(k)} \right\|_2. \end{aligned} \qquad \text{.....(17)}$$

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

$$\left\| \underset{\sim}{u}^{(k)} \right\|_2^2 = 2\sigma_k \, (\sigma_k + \left| a_{kk}^{(k)} \right|), \qquad \text{.....(18)}$$

which implies the inequality

$$\| \underset{\sim}{u}^{(k)} \|_2 \geq \sqrt{2}\ \sigma_k. \qquad \ldots\ldots\ldots(19)$$

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

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

$$A_j^{(k+1)} = A_j^{(k)} - y_j^{(k)}\ \underset{\sim}{u}^{(k)}, \qquad \ldots\ldots\ldots(20)$$

so the theorem bounds the multiple of $\underset{\sim}{u}^{(k)}$ that is added to the $j^{\underline{th}}$ 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 $\underset{\sim}{\bar{u}}^{(k)}$ 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

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

We find that the calculated vector $\underset{\sim}{x}$ 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

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

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

$$| \bar{\sigma}_k - \sigma_k | \leq \frac{3}{2} \varepsilon \sigma_k . \qquad \ldots\ldots\ldots(23)$$

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

$$| \bar{\beta}_k^{-1} - \beta_k^{-1} | \leq \varepsilon \beta_k^{-1} + \varepsilon \sigma_k^2 + \varepsilon \sigma_k | a_{kk}^{(k)} | + \frac{3}{2} \varepsilon \sigma_k | a_{kk}^{(k)} |$$

$$= \varepsilon \sigma_k (2\sigma_k + \frac{7}{2} | a_{kk}^{(k)} |) , \qquad \ldots\ldots\ldots(24)$$

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

$$| \bar{u}_k^{(k)} - u_k^{(k)} | \leq \varepsilon | u_k^{(k)} | + \frac{3}{2} \varepsilon \sigma_k$$

$$= \varepsilon (\frac{5}{2} \sigma_k + | a_{kk}^{(k)} |). \qquad \ldots\ldots\ldots(25)$$

However in bounding the error in $y_j^{(k)}$ we depart from the Algol listing of Businger and Goluh (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

$$| \bar{y}_j^{(k)} - y_j^{(k)} | \leq \varepsilon | y_j^{(k)} | \{2 + \beta_k \sigma_k (2\sigma_k + \frac{7}{2} | a_{kk}^{(k)} |)\}$$

$$+ \varepsilon \beta_k (\frac{5}{2} \sigma_k + | a_{kk}^{(k)} |) | a_{kj}^{(k)} | , \qquad \ldots\ldots\ldots(26)$$

the second term being a consequence of the error in $\underset{\sim}{u}^{(k)}$. Therefore, from

Theorem 1, the inequality (7) and the definition (8), we obtain the result

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

which, because of the inequality

$$\left| a_{kk}^{(k)} \right| \leq \sigma_k \quad \ldots\ldots(28)$$

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

$$\left| \bar{y}_j^{(k)} - y_j^{(k)} \right| \leq \varepsilon(\frac{19}{4}\sqrt{2} + \frac{7}{4}) . \quad \ldots\ldots(29)$$

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:

$$\Delta_{ij}^{(k)} = 0 \qquad \begin{cases} j < k \\ i < k,\ j \geq k \\ i > k,\ j = k. \end{cases} \quad \ldots\ldots(30)$$

Also equation (23) gives the result

$$\left| \Delta_{kk}^{(k)} \right| \leq \frac{3}{2}\varepsilon\sigma_k, \quad \ldots\ldots(31)$$

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

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

$$= \varepsilon\left\{ \left| a_{ij}^{(k+1)} \right| + (\frac{23}{4}\sqrt{2} + \frac{7}{4})\left| u_i^{(k)} \right| \right\}. \quad \ldots\ldots(32)$$

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

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

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

$$\sigma_k \leq \alpha_k , \quad \ldots\ldots(34)$$

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

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

(derived from equation (10)), and

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

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

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

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

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

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

$$\|\Delta_j^{(k)}\|_2 \leq \varepsilon\Big[\Big\{\sum_{i=k}^{m}(a_{ij}^{(k+1)})^2\Big\}^{\frac{1}{2}} + (\tfrac{23}{4}\sqrt{2} + \tfrac{7}{4})\,\|\underset{\sim}{u}^{(k)}\|_2\Big] + \tfrac{7}{2}\sqrt{2}\,\varepsilon\,\sigma_k, \qquad \ldots\ldots(39)$$

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

$$\begin{aligned} \|\Delta_j^{(k)}\|_2 &\leq \varepsilon\,\{(\tfrac{23}{4}\sqrt{2} + \tfrac{7}{4})\,\|\underset{\sim}{u}^{(k)}\|_2 + (\tfrac{7}{2}\sqrt{2} + 1)\sigma_k\} \\ &\leq \varepsilon\ (\tfrac{25}{4}\sqrt{2} + \tfrac{21}{4})\,\|\underset{\sim}{u}^{(k)}\|_2\,, \quad j > k. \end{aligned} \qquad \ldots\ldots\ldots\ldots(40)$$

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 $P^{(k)}$ $(k=1,2,\ldots,n)$ 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

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

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

$$(A + \Delta)\ \underset{\sim}{x} = \underset{\sim}{b} + \underset{\sim}{\delta}. \qquad \ldots\ldots\ldots\ldots (42)$$

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

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

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

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

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 $\Delta^{(k)}$ $(k=1,2,\ldots,n)$, which gives the identity

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

from which we deduce the equation

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

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 $k^{\underline{th}}$ term of the sum (46), we substitute expression (4) in place of some of the orthogonal matrices, obtaining the result

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

$\ldots\ldots\ldots\ldots(47)$

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

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

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

$$\sigma_q \geq \sigma_k \ , \ q < k, \qquad \ldots\ldots\ldots\ldots(49)$$

so from statement (19) we deduce the inequality

$$\left\|\underset{\sim}{u}^{(q)}\right\|_2 \geq \sqrt{2}\ \sigma_k \geq \left\|\underset{\sim}{u}^{(k)}\right\|_2/\sqrt{2} \ , \ q < k, \qquad \ldots\ldots\ldots\ldots(50)$$

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

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

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

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

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

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

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

$$\left| \Delta_{ij} \right| \leq \{U_{ij} + V_{ij}\}\varepsilon\alpha_i, \qquad \ldots\ldots\ldots\ldots(54)$$

where

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

and

$$V_{ij} = \tfrac{3}{4}\sqrt{2}\ \lambda_1^{(j)} + (\frac{25}{4}\sqrt{2} + \frac{21}{4}) \sum_{k=1}^{j-1} \lambda_i^{(k)}. \qquad \ldots\ldots\ldots\ldots(56)$$

The theorem of this section states that $\left| \Delta_{ij} \right|$ 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 $i \leq n-1$, and it is not difficult to show that if expression (53) is an equality then V_{in} is an increasing function of i for $i \leq n-2$. Therefore we consider the details of four separate cases and derive the results

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

where

$$\phi(n) = n^2(\frac{25}{2} + \frac{21}{4}\sqrt{2}) + n(-\frac{85}{4} + \frac{5}{2}\sqrt{2}). \quad \ldots\ldots\ldots\ldots(58)$$

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

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

are bounded by the inequality

$$|\Delta_{ij}| \leq \{n^2(\frac{25}{2} + \frac{21}{4}\sqrt{2}) - n(\frac{85}{4} - \frac{5}{2}\sqrt{2}) + (24 + 14\sqrt{2})\}\varepsilon\, a_i. \quad \ldots\ldots(60)$$

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 $\bar{\underset{\sim}{b}}^{(1)}, \bar{\underset{\sim}{b}}^{(2)}, \ldots, \bar{\underset{\sim}{b}}^{(n+1)}$ (see equation (15)), and we must treat the back-substitution stage of the algorithm, in which $\underset{\sim}{x}$ is determined from the equation

$$\bar{A}^{(n+1)}\,\underset{\sim}{x} = \bar{\underset{\sim}{b}}^{(n+1)}. \quad \ldots\ldots\ldots\ldots(61)$$

If it happens that $\|\underset{\sim}{b}\|_2$ is so small that both the inequalities

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

and

$$\nu_k = \left\{ \sum_{i=k}^{m} (b_i^{(k)})^2 \right\}^{\frac{1}{2}} \leq \sigma_k , \quad k = 1,2,\ldots,n \qquad \ldots\ldots\ldots(63)$$

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

$$\rho = \max \left[\max_i(\mu_i/\alpha_i) , \max_k(\nu_k/\sigma_k) \right] \qquad \ldots\ldots\ldots(64)$$

exceeds one, then to make the inequalities (62) and (63) hold we could scale the original right-hand side vector $\underset{\sim}{b}$ 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 $\underset{\sim}{b}^{(k)}$ 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

$$\underset{\sim}{\delta}^{(k)} = P^{(k)} \underset{\sim}{b}^{(k)} - \underset{\sim}{b}^{(k+1)}, \qquad \ldots\ldots\ldots(65)$$

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

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

and from expression (40) we find the inequality

$$\left\| \underset{\sim}{\delta}^{(k)} \right\|_2 \leq \rho (\frac{25}{4}\sqrt{2} + \frac{21}{4}) \left\| \underset{\sim}{u}^{(k)} \right\|_2 . \qquad \ldots\ldots\ldots(67)$$

To calculate bounds on the components of $\underset{\sim}{\delta}$, defined in equation (42), we find by the argument that led to equation (46) the result

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

and instead of equation (51) we obtain the bound

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

Therefore the inequality corresponding to statement (54) is

$$|\delta_i| \leq \{ \beta_i + T_i \} \varepsilon \rho \alpha_i, \qquad \ldots\ldots\ldots(70)$$

where

$$\beta_i = \begin{cases} (\frac{23}{4}\sqrt{2} + \frac{11}{4})(i-1) + (15\sqrt{2} + \frac{9}{2}), & i \leq n \\ (\frac{23}{4}\sqrt{2} + \frac{11}{4})\, n, & i > n \end{cases} \qquad \ldots\ldots\ldots(71)$$

and

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

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

$$|\delta_i| \leq \{ n^2(\frac{25}{2} + \frac{21\sqrt{2}}{4}) + n(\frac{3}{4} + 13\sqrt{2}) + (24 + 14\sqrt{2}) \} \varepsilon \rho \alpha_i. \qquad \ldots\ldots\ldots(73)$$

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

$$(\bar{\underset{\sim}{A}}^{(n+1)} + E)\, \underset{\sim}{x} = \bar{\underset{\sim}{b}}^{(n+1)}, \qquad \ldots\ldots\ldots(74)$$

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

$$|E_{ij}| \leq \varepsilon \; |\bar{a}_{ii}^{(n+1)}| \; \delta_{ij}, \qquad \ldots\ldots\ldots(75)$$

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 equalities (37) and (38) the case $j = k$ becomes

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

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

$$\left.\begin{aligned} U_{ii} &= (\tfrac{23}{4}\sqrt{2} + \tfrac{11}{4})(i-1) + \tfrac{5}{2} \\ V_{ij} &= \tfrac{5}{4}\sqrt{2}\,\lambda_1^{(j)} + (\tfrac{25}{4}\sqrt{2} + \tfrac{21}{4}) \sum_{k=1}^{j-1} \lambda_1^{(k)} \end{aligned}\right\}. \qquad \ldots\ldots\ldots(77)$$

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

$$(A + \Delta)\,\underset{\sim}{x} = \underset{\sim}{b} + \underset{\sim}{\delta}, \qquad \ldots\ldots\ldots(78)$$

where the elements of $\underset{\sim}{\delta}$ are bounded by the inequality (73), and where the elements of Δ are bounded by the inequality

$$|\Delta_{ij}| \leq \{ n^2(\tfrac{25}{2} + \tfrac{21}{4}\sqrt{2}) - n(\tfrac{77}{4} - \tfrac{5}{2}\sqrt{2}) + (24 + 14\sqrt{2}) \}\, \varepsilon\alpha_i, \qquad \ldots\ldots\ldots(79)$$

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

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 10^6 & 10^6 & 0 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{pmatrix}. \qquad \ldots\ldots\ldots(80)$$

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

$$A^{(2)} = \begin{pmatrix} -10^6\sqrt{2} & -10^6/\sqrt{2} & -10^6/\sqrt{2} \\ 0 & \tfrac{1}{2}10^6-\sqrt{2} & -\tfrac{1}{2}10^6-1/\sqrt{2} \\ 0 & -\tfrac{1}{2}10^6-\sqrt{2} & \tfrac{1}{2}10^6-1/\sqrt{2} \\ 0 & 1 & 1 \end{pmatrix} \quad \ldots(81)$$

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 $\underset{\sim}{x}$, 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

$$\alpha_1 = 10^6\sqrt{2} \quad , \qquad \ldots\ldots\ldots\ldots(82)$$

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

$$\max_j \left| a_{ij}^{(k+1)} \right| = \max_j \left| a_{ij}^{(k)} \right|, \quad i < k \ , \qquad \ldots\ldots\ldots(83)$$

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

$$\max_j \left| a_{ij}^{(k+1)} \right| \leq (\sqrt{2}+1) \max_j \left| a_{ij}^{(k)} \right| \quad , \quad i > k. \qquad \ldots\ldots(84)$$

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

Theorem 4

If the inequality

$$\left| a_{kk}^{(k)} \right| \geq \left| a_{ik}^{(k)} \right|, \quad i > k, \qquad \text{.....(85)}$$

holds, then we have the bound

$$\max_j \left| a_{kj}^{(k+1)} \right| \leq \sqrt{m} \max_j \left| a_{kj}^{(k)} \right|. \qquad \text{.....(86)}$$

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

$$\left| a_{kj}^{(k+1)} \right| \leq \left[\sum_{i=k}^{m} \left| a_{ij}^{(k+1)} \right|^2 \right]^{\frac{1}{2}}$$

$$= \left[\sum_{i=k}^{m} \left| a_{ij}^{(k)} \right|^2 \right]^{\frac{1}{2}}. \qquad \text{.....(87)}$$

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

$$\left| a_{kj}^{(k+1)} \right| \leq \sigma_k. \qquad \text{.....(88)}$$

From statements (8) and (85) the inequality

$$\sigma_k \leq \sqrt{m} \left| a_{kk}^{(k)} \right| \qquad \text{.....(89)}$$

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 $\{|a_{kk}^{(k)}|, |a_{k+1\,k}^{(k)}|, \ldots, |a_{mk}^{(k)}|\}$, say it is $|a_{qk}^{(k)}|$, and we interchange the $k^{\underline{th}}$ and $q^{\underline{th}}$ rows of $A^{(k)}$ if $q \neq k$.

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

$$A^{(1)} = \begin{pmatrix} 10^6 & 10^6 & 0 \\ 0 & 2 & 1 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{pmatrix} \qquad \text{.....(90)}$$

and

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

so the previous loss of accuracy is avoided.

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

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

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

$$\beta = \max_i [\alpha_i / \max_j |a_{ij}^{(1)}|]. \qquad \ldots\ldots\ldots(93)$$

We used one hundred 20x10 matrices whose elements were

$$a_{ij} = 10^{10p_i} q_{ij} , \qquad \ldots\ldots\ldots(94)$$

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

$$\begin{pmatrix} 1 & -0.99 & -0.99 & -0.99 & \ldots & -0.99 & -0.99 \\ 0 & 0.1 & -0.099 & -0.099 & \ldots & -0.099 & -0.099 \\ 0 & 0 & 0.01 & -0.0099 & \ldots & -0.0099 & -0.0099 \\ 0 & 0 & 0 & 0.001 & \ldots & -0.00099 & -0.00099 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 10^{-n+2} & -0.99\times10^{-n+2} \\ 10^{-n-10} & 10^{-n-10} & 10^{-n-10} & 10^{-n-10} & \ldots & 10^{-n-10} & 10^{-n-10} \end{pmatrix}$$

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

$$\Big[\sum_{i=k}^{m} \{a_{ik}^{(k)}\}^2\Big]^{\frac{1}{2}} + \max_{k\leq i\leq m} |a_{ik}^{(k)}|$$

$$\geq \Big[\sum_{i=k}^{m} \{a_{ij}^{(k)}\}^2\Big]^{\frac{1}{2}} + \max_{k\leq i\leq m} |a_{ij}^{(k)}|, \quad j > k, \qquad \ldots.(95)$$

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

$$|y_j^{(k)}| \leq 1 , \qquad \ldots\ldots\ldots(96)$$

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 $i^{\underline{th}}$ row of the matrices $\bar{A}^{(1)}, \bar{A}^{(2)}, \ldots, \bar{A}^{(n+1)}$. Therefore if x_j is scaled so that for $i = 1, 2, \ldots, m$ the element a_{ij} is much smaller than the other elements of the $i^{\underline{th}}$ 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

$$\max_i \left[|a_{ij}| / \max_k |a_{ik}|\right] , \quad j = 1,2,\ldots,n, \qquad \ldots\ldots\ldots(97)$$

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. and GOLUB, G.H. (1965). "Linear least squares solutions by Householder transformations", Numer. Math., Vol. 7, pp. 269-276.

GOLUB, G.H. (1965). "Numerical methods for solving linear least squares problems", Numer. Math., Vol. 7, pp. 206-216.

WILKINSON, J.H. (1965). "The algebraic eigenvalue problem", Oxford University Press.