Row-wise backward stable elimination methods for the equality constrained least squares problem

Cox, Anthony J. and Higham, Nicholas J.

1999

MIMS EPrint: **2006.171**

ISSN 1749-9097

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 313–326

ROW-WISE BACKWARD STABLE ELIMINATION METHODS FOR THE EQUALITY CONSTRAINED LEAST SQUARES PROBLEM[*]

ANTHONY J. COX[†] AND NICHOLAS J. HIGHAM[‡]

Abstract. It is well known that the solution of the equality constrained least squares (LSE) problem $\min_{Bx=d} \|b - Ax\|_2$ is the limit of the solution of the unconstrained weighted least squares problem

$$\min_x \left\| \begin{bmatrix} \mu d \\ b \end{bmatrix} - \begin{bmatrix} \mu B \\ A \end{bmatrix} x \right\|_2$$

as the weight μ tends to infinity, assuming that $[\,B^T \quad A^T\,]^T$ has full rank. We derive a method for the LSE problem by applying Householder QR factorization with column pivoting to this weighted problem and taking the limit analytically, with an appropriate rescaling of rows. The method obtained is a type of direct elimination method. We adapt existing error analysis for the unconstrained problem to obtain a row-wise backward error bound for the method. The bound shows that, provided row pivoting or row sorting is used, the method is well-suited to problems in which the rows of A and B vary widely in norm. As a by-product of our analysis, we derive a row-wise backward error bound of precisely the same form for the standard elimination method for solving the LSE problem. We illustrate our results with numerical tests.

Key words. constrained least squares problem, weighted least squares problem, Householder QR factorization, Gaussian elimination, elimination method, rounding error analysis, backward stability, row pivoting, row sorting, column pivoting

AMS subject classifications. 65F20, 65G05

PII. S0895479898335957

1. Introduction.

Consider the equality constrained least squares (LSE) problem

$$\min_{Bx=d} \|b - Ax\|_2, \qquad A \in \mathbb{R}^{m\times n}, \quad B \in \mathbb{R}^{p\times n}, \quad m+p \geq n \geq p. \tag{1.1}$$

We assume throughout this work that $\operatorname{rank}(B) = p$, which ensures that the constraint system is consistent, and also that $\operatorname{rank}\begin{bmatrix} B \\ A \end{bmatrix} = n$, which ensures that the LSE problem has a unique solution. Two commonly used methods for solving the LSE problem are based on QR factorization. The null space method (of which there are several variations [6]) begins by factorizing $B^T = QR$ to obtain a basis for the null space of B. The elimination method, on the other hand, uses QR factorization with column pivoting to factorize

$$B\Pi = Q\,[\,R_1 \quad R_2\,], \qquad R_1 \in \mathbb{R}^{p\times p} \text{ upper triangular, nonsingular.} \tag{1.2}$$

Partitioning $\Pi^T x = [\widetilde{x}_1^T,\ \widetilde{x}_2^T]^T$, $\widetilde{x}_1 \in \mathbb{R}^p$ and substituting the factorization (1.2) into the constraints yields

$$R_1\widetilde{x}_1 = Q^T d - R_2\widetilde{x}_2.$$

[*] Received by the editors March 19, 1998; accepted for publication (in revised form) by P. Van Dooren August 31, 1998; published electronically September 30, 1999.
http://www.siam.org/journals/simax/21-1/33595.html

[†] Ground Processing Department, Mail Point 31, Matra Marconi Space UK, Anchorage Road, Portsmouth, PO3 5PU, UK (anthonyj.cox@mmsuk.co.uk). The work of this author was supported by an Engineering and Physical Sciences Research Council Research Studentship.

[‡] Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/). The work of this author was supported by Engineering and Physical Sciences Research Council grant GR/L76532.

By solving for $\widetilde{x}_1$ and partitioning $A\Pi = [\widetilde{A}_1,\ \widetilde{A}_2]$, $\widetilde{A}_1 \in \mathbb{R}^{m\times p}$ we reduce the LSE problem to the unconstrained problem

$$\min_{\widetilde{x}_2} \left\| (\widetilde{A}_2 - \widetilde{A}_1 R_1^{-1} R_2)\widetilde{x}_2 - (b - \widetilde{A}_1 R_1^{-1} Q^T d) \right\|_2 .$$

Solving this unconstrained problem by QR factorization completes the elimination method as originally presented by Björck and Golub [5] (see also [13, Chapter 21]). It is instructive to think of the method in terms of transformations on the matrix "B-over-A":

$$\begin{bmatrix} B \\ A \end{bmatrix} = \begin{array}{cc} & \begin{array}{cc} p & n-p \end{array} \\ \begin{array}{c} p \\ m \end{array} & \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} \end{array} \to \begin{bmatrix} R_1 & R_2 \\ \widetilde{A}_1 & \widetilde{A}_2 \end{bmatrix} \to \begin{bmatrix} R_1 & R_2 \\ 0 & \widetilde{A}_2 - \widetilde{A}_1 R_1^{-1} R_2 \end{bmatrix} \to \begin{bmatrix} R_1 & R_2 \\ 0 & R_3 \\ 0 & 0 \end{bmatrix},$$

where $R_3 \in \mathbb{R}^{(n-p)\times(n-p)}$ is upper triangular. Note that the penultimate transformation is simply the annihilation of $\widetilde{A}_1$ by Gaussian elimination. We will refer to this method as the *EG method*.

The B-over-A matrix also arises in the method of weighting for solving the LSE problem, which is based on the observation that the LSE solution is the limit of the solution of the unconstrained problem

$$\text{(1.3)} \qquad \min_x \left\| \begin{bmatrix} \mu d \\ b \end{bmatrix} - \begin{bmatrix} \mu B \\ A \end{bmatrix} x \right\|_2$$

as the weight μ tends to infinity. Van Loan [18] describes an algorithm that solves (1.3) for a single weight and uses a refinement procedure to approximate the required limit. The algorithm is analyzed further by Barlow and Vemulapati [1], [3].

In this work we derive a method for the LSE problem by taking the limit $\mu \to \infty$ analytically rather than numerically. To be precise, we apply Householder QR factorization with column pivoting to (1.3), rescale the first p rows, and then take the limit $\mu \to \infty$ to obtain a matrix that we view as an update of B-over-A. The new method is an elimination method very similar, but not identical, to the EG method. We show that it satisfies a pleasing row-wise backward error bound when row pivoting or row sorting is used—a result that makes the method attractive when the rows of A and B vary widely in norm. We also show that a row-wise backward error bound of exactly the same form holds for the EG method. We give some numerical experiments to confirm the stability properties of the methods.

After this paper was submitted for publication, we learned that Reid [15] has obtained the method derived here by a different argument involving infinite weights. He shows that the algorithm of Powell and Reid [14], including its error analysis, may be applied to the weighted problem with implicit scaling, that is, with the weights stored separately. By holding inverse weights, the constrained case can be accommodated. He obtains a backward error bound similar to our Theorem 4.3 and also shows that iterative refinement may be applied. Reid points out that the method is equivalent to a method of Gulliksson and Wedin [9], [10], which is expressed in the language of "M-invariant reflections."

2. A single stage of the method.

Define

$$\text{(2.1)} \qquad C = \begin{bmatrix} B \\ A \end{bmatrix}, \quad f = \begin{bmatrix} d \\ b \end{bmatrix}, \quad C_\mu = \begin{bmatrix} \mu B \\ A \end{bmatrix}, \quad f_\mu = \begin{bmatrix} \mu d \\ b \end{bmatrix},$$

where $\mu > 0$. We consider the first stage of Householder QR factorization applied to C_μ, examining what happens as $\mu \to \infty$. In outline, we write the transformed matrix as

$$C'_\mu = \begin{bmatrix} \mu B'_\mu \\ A'_\mu \end{bmatrix},$$

divide the first p rows of C'_μ by μ and take the limit as $\mu \to \infty$, obtaining a matrix

$$C' = \begin{bmatrix} B' \\ A' \end{bmatrix}$$

that is independent of μ. Then we view C' as the result of an update of C. We will show that all the elements of C' remain finite as $\mu \to \infty$. In this way, we set up a one-to-one correspondence between performing the first step of Householder QR factorization on C_μ,

$$\begin{bmatrix} \mu B \\ A \end{bmatrix} \to \begin{bmatrix} \mu B'_\mu \\ A'_\mu \end{bmatrix}, \tag{2.2}$$

and carrying out the update

$$\begin{bmatrix} B \\ A \end{bmatrix} \to \begin{bmatrix} B' \\ A' \end{bmatrix} \tag{2.3}$$

on the unscaled matrix C. Our aim is to determine the nature of this update.

We mention that Stewart [17] compares the equations for Householder QR factorization on C_μ with those for the EG method. While his analysis is related to ours, his aim (to understand the method of weighting) is different and he does not take limits.

We will denote by a_j, b_j, and c_j the jth columns of A, B, and C_μ, respectively (we will have no need to refer to the columns of C, so this slightly inconsistent notation should not cause any confusion); similarly, we write $C_\mu = (c_{ij})$. The first stage of Householder QR factorization applied to C_μ is

$$C'_\mu = C_\mu - 2v_\mu \frac{v_\mu^T C_\mu}{v_\mu^T v_\mu},$$

where

$$v_\mu = c_1 + \tau \|c_1\|_2 e_1, \qquad \tau = \operatorname{sign}(c_{11}), \tag{2.4}$$

which we can rewrite as

$$c'_{ij} = c_{ij} - 2v_\mu(i)\frac{v_\mu^T c_j}{v_\mu^T v_\mu}, \qquad i = 1{:}\,p+m, \quad j = 1{:}\,n. \tag{2.5}$$

We assume the use of column pivoting, so that columns are interchanged, if necessary, to ensure that

$$\|c_1\|_2 = \max_j \|c_j\|_2.$$

For notational simplicity, we assume that no interchanges are required by column pivoting. Our first concern is with the behavior of the multipliers

$$2\frac{v_\mu^T c_j}{v_\mu^T v_\mu}, \qquad j = 1\colon n,$$

which we will refer to as *Householder multipliers*, as $\mu \to \infty$. As a preliminary, we define ζ by

$$\|c_1\|_2 = \sqrt{\mu^2\|b_1\|_2^2 + \|a_1\|_2^2} = \mu\sqrt{\|b_1\|_2^2 + \frac{1}{\mu^2}\|a_1\|_2^2} = \mu\zeta,$$

and note that

$$\lim_{\mu\to\infty} \zeta = \|b_1\|_2.$$

We have

$$v_\mu^T c_j = c_1^T c_j + \tau\|c_1\|_2 c_{1j} = \mu^2 b_1^T b_j + a_1^T a_j + \mu^2\tau\zeta b_{1j}$$

and

$$v_\mu^T v_\mu = 2\|c_1\|_2(\|c_1\|_2 + |c_{11}|) = 2\mu^2\zeta(\zeta + |b_{11}|).$$

Hence

$$2\frac{v_\mu^T c_j}{v_\mu^T v_\mu} = \frac{b_1^T b_j + a_1^T a_j/\mu^2 + \tau\zeta b_{1j}}{\zeta(\zeta + |b_{11}|)},$$

which yields

$$\lim_{\mu\to\infty} 2\frac{v_\mu^T c_j}{v_\mu^T v_\mu} = \frac{b_1^T b_j + \tau\|b_1\|_2 b_{1j}}{\|b_1\|_2(\|b_1\|_2 + |b_{11}|)} = 2\frac{v_B^T b_j}{v_B^T v_B}, \tag{2.6}$$

where

$$v_B = b_1 + \tau\|b_1\|_2 e_1.$$

We see that, in the limit, A's contribution to the Householder multiplier $2v_\mu^T c_j / v_\mu^T v_\mu$ is lost.

Using (2.5) and (2.4) we have, for $i = 1\colon p$,

$$\begin{aligned} b'_{ij} := \lim_{\mu\to\infty} b_{ij}(\mu)' &= \lim_{\mu\to\infty} \frac{c'_{ij}}{\mu} \\ &= \lim_{\mu\to\infty}\left(b_{ij} - \frac{2v_\mu(i)}{\mu}\frac{v_\mu^T c_j}{v_\mu^T v_\mu}\right) \\ &= b_{ij} - 2v_B(i)\frac{v_B^T b_j}{v_B^T v_B}. \end{aligned}$$

This is just the first stage of Householder QR factorization applied to B, which corresponds to the application of the Householder update

$$C' = C - 2v_C\frac{v_C^T C}{v_C^T v_C}, \quad v_C = \begin{bmatrix} v_B \\ 0 \end{bmatrix} \tag{2.7}$$

to the unscaled matrix C. Moreover, $\|c_i\|_2/\|c_j\|_2 \to \|b_i\|_2/\|b_j\|_2$ so, in the limit, column pivoting based on the columns of C is equivalent to column pivoting based on the columns of B.

Now we turn to the rows of A. In the limit as $\mu \to \infty$, we have, from (2.4)–(2.6),

$$a'_{ij} = a_{ij} - 2a_{i1}\frac{v_B^T b_j}{v_B^T v_B},$$

which corresponds to the update

$$C' = C - 2v_A\frac{v_C^T C}{v_C^T v_C}, \quad v_A = \begin{bmatrix} 0 \\ a_1 \end{bmatrix}. \tag{2.8}$$

Combining (2.7) and (2.8) we see that, in the limit as $\mu \to \infty$, carrying out one step of Householder QR factorization on C_μ is equivalent to carrying out the outer product update

$$C' = PC = C - 2v\frac{v_C^T C}{v_C^T v_C}, \quad \text{where} \quad v = \begin{bmatrix} v_B \\ a_1 \end{bmatrix}, \tag{2.9}$$

on the unscaled matrix C. Note that the transformation matrix P is not a Householder matrix, because $v \neq v_C$.

It is natural to ask whether the equations we have derived are equivalent to those for the EG method described in section 1. Recall that the EG method QR factorizes B as in (1.2) and then eliminates the first p columns of $\widetilde{A} = A\Pi$ by Gaussian elimination. It is easily shown that the first row of $[\, R_1 \quad R_2 \,]$ is proportional to

$$[\, b_1^T b_1 \quad b_1^T b_2 \quad \dots \quad b_1^T b_n \,]. \tag{2.10}$$

The first column of $\widetilde{A}$ is therefore zeroed by performing one step of Gaussian elimination with this row as the pivot row. Using the fact that $v_B^T v_B = 2v_B^T b_1$, it is easy to show that for the new method we are effectively performing Gaussian elimination with the "virtual pivot row"

$$[\, v_B^T b_1 \quad v_B^T b_2 \quad \dots \quad v_B^T b_n \,], \tag{2.11}$$

which is not actually present in the matrix C. The vectors (2.10) and (2.11) are proportional only if B is upper trapezoidal, which shows that the new method and the EG method are mathematically different in general.

3. The complete method.

By carrying out p eliminations of the type described in the previous section, we reduce C to the form

$$\begin{bmatrix} R_1 & R_2 \\ 0 & A'_2 \end{bmatrix},$$

where $R_1 \in \mathbb{R}^{p\times p}$ is upper triangular and nonsingular. We mention that in place of column pivoting some other pivoting strategy that keeps $R_1^{-1}R_2$ small could be used instead; cf. [2].

The remaining $\min(n-p, m-1)$ steps of the method consist of applying standard Householder QR factorization with column pivoting to the matrix A'_2, after which the whole of C has been reduced to upper trapezoidal form. By applying the same sequence of updates to f and solving the resulting triangular system, we obtain the solution to the LSE problem (1.1). We summarize our method as follows.

ALGORITHM EH *This algorithm solves the LSE problem* (1.1).

```
1. Let
       $$C^{(1)} = \begin{bmatrix} B \\ A \end{bmatrix}, \qquad f^{(1)} = \begin{bmatrix} d \\ b \end{bmatrix}, \qquad q = p + m.$$
2. Stage I
   for $k = 1{:}\,p$
   Interchange the columns of $C^{(k)}$ so that $\|c_k^{(k)}(k{:}\,p)\|_2 = \max_{j \ge k} \|c_j^{(k)}(k{:}\,p)\|_2$.
   Update
       $$C^{(k+1)}(k{:}\,q, k{:}\,n) = C^{(k)}(k{:}\,q, k{:}\,n) - 2v_k(k{:}\,q)\frac{v_k(k{:}\,p)^T C^{(k)}(k{:}\,p, k{:}\,n)}{v_k(k{:}\,p)^T v_k(k{:}\,p)},$$
       $$f^{(k+1)}(k{:}\,q) = f^{(k)}(k{:}\,q) - 2v_k(k{:}\,q)\frac{v_k(k{:}\,p)^T f^{(k)}(k{:}\,p)}{v_k(k{:}\,p)^T v_k(k{:}\,p)},$$
   where
       $$v_k(i) = \begin{cases} 0, & i = 1{:}\,k-1, \\ c_{kk}^{(k)} + \operatorname{sign}\left(c_{kk}^{(k)}\right) \left\|c_k^{(k)}(k{:}\,p)\right\|_2, & i = k, \\ c_{ik}^{(k)}, & i = k+1{:}\,q. \end{cases}$$
   end
3. Stage II (standard Householder QR factorization with column pivoting)
   for $k = p+1{:}\,\min(n, m+p-1)$
   Interchange the columns of $C^{(k)}$ so that $\left\|c_k^{(k)}(k{:}\,q)\right\|_2 = \max_{j \ge k} \left\|c_j^{(k)}(k{:}\,q)\right\|_2$.
   Update
       $$C^{(k+1)}(k{:}\,q, k{:}\,n) = C^{(k)}(k{:}\,q, k{:}\,n) - 2v_k(k{:}\,q)\frac{v_k(k{:}\,q)^T C^{(k)}(k{:}\,q, k{:}\,n)}{v_k(k{:}\,q)^T v_k(k{:}\,q)},$$
       $$f^{(k+1)}(k{:}\,q) = f^{(k)}(k{:}\,q) - 2v_k(k{:}\,q)\frac{v_k(k{:}\,q)^T f^{(k)}(k{:}\,q)}{v_k(k{:}\,q)^T v_k(k{:}\,q)},$$
   where
       $$v_k(i) = \begin{cases} 0, & i = 1{:}\,k-1, \\ c_{kk}^{(k)} + \operatorname{sign}\left(c_{kk}^{(k)}\right) \left\|c_k^{(k)}(k{:}\,q)\right\|_2, & i = k, \\ c_{ik}^{(k)}, & i = k+1{:}\,q. \end{cases}$$
    end
4. Solve the triangular system $C^{(k+1)}(1{:}\,n, 1{:}\,n)y = f^{(k+1)}(1{:}\,n)$.
5. Obtain $x$ by permuting $y$ to take account of the column interchanges.
```

Algorithm EH implicitly computes a matrix factorization: ignoring column interchanges, we have

$$\underbrace{\begin{bmatrix} I & 0 \\ 0 & \widetilde{Q}_3 \end{bmatrix}}_{\text{Stage II}} \underbrace{\begin{bmatrix} \widetilde{Q}_1 & 0 \\ \widetilde{Q}_2 & I \end{bmatrix}}_{\text{Stage I}} \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} = \begin{bmatrix} R_1 & R_2 \\ 0 & R_3 \\ 0 & 0 \end{bmatrix},$$

where $\widetilde{Q}_1$ and $\widetilde{Q}_3$ are orthogonal and $R_1 \in \mathbb{R}^{p \times p}$ and $R_3 \in \mathbb{R}^{(n-p)\times(n-p)}$ are upper triangular. Rewriting, we have

$$\begin{bmatrix} \widetilde{Q}_1 & 0 \\ \widetilde{Q}_3\widetilde{Q}_2 & \widetilde{Q}_3 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix},$$

or, with $Q_1 = \widetilde{Q}_1^T$, $Q_2 = -\widetilde{Q}_2\widetilde{Q}_1^T$ and $Q_3 = \widetilde{Q}_3^T$,

$$\begin{array}{cc} & \begin{array}{cc} p & n-p \end{array} \\ \begin{array}{c} p \\ m \end{array} & \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} \end{array} = \begin{array}{cc} & \begin{array}{cc} p & m \end{array} \\ \begin{array}{c} p \\ m \end{array} & \begin{bmatrix} Q_1 & 0 \\ Q_2 & Q_3 \end{bmatrix} \end{array} \begin{array}{cc} n & \\ \begin{bmatrix} R \\ 0 \end{bmatrix} & \begin{array}{c} n \\ m-n+p \end{array} \end{array} . \tag{3.1}$$

The EG method produces exactly the same factorization; the difference in the methods lies in the intermediate quantities computed. Note that another way to derive both the EG method and Algorithm EH is to substitute the factorization into the associated augmented system; this is done in [4] and [5] for the EG method.

4. Numerical stability. Now we consider the stability of Algorithm EH. First, we recall what is known about the stability of Householder QR factorization for solving the LS problem $\min_x \|b - Ax\|_2$. The method is normwise backward stable. Moreover, when column pivoting is used the following row-wise backward error result holds, which is of interest in situations in which the rows of A vary widely in norm. We employ the standard model of floating point arithmetic with unit roundoff u [12, Section 2.2], and define the constant

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

where k is a positive integer and c is a small integer constant whose exact value is unimportant. We assume that the signs in the Householder vectors are chosen as in (2.4), which is the standard choice. For the other choice of sign, the following theorem is invalid (see [8] for more details).

THEOREM 4.1. *Let the LS problem* $\min_x \|b - Ax\|_2$*, where* $A \in \mathbb{R}^{m\times n}$ *is of full rank* n*, be solved using Householder QR factorization with column pivoting. Then the computed solution* $\widehat{x}$ *is the exact solution to*

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

where the perturbations satisfy

$$|\Delta a_{ij}| \le j^2 \widetilde{\gamma}_m \alpha_i \max_s |a_{is}|, \qquad |\Delta b_i| \le n^2 \widetilde{\gamma}_m \beta_i \max(\phi \max_j |a_{ij}|, |b_i|),$$

where

$$\alpha_i = \frac{\max_{j,k} |\widehat{a}_{ij}^{(k)}|}{\max_j |a_{ij}|}, \quad \beta_i = \frac{\max\bigl(\phi \max_{j,k} |\widehat{a}_{ij}^{(k)}|, |\widehat{b}_i^{(k)}|\bigr)}{\max(\phi \max_j |a_{ij}|, |b_i|)}, \quad \phi = \max_k \frac{\|\widehat{b}^{(k)}(k\colon m)\|_2}{\|\widehat{a}_k^{(k)}(k\colon m)\|_2}.$$

Proof. A similar result was originally proved under some additional assumptions by Powell and Reid [14]. The analysis was reworked by Cox and Higham [8]. The result as stated is a slightly improved version of the result in [8], with a minor error in the Δb bound corrected. □

Theorem 4.1 shows that, when column pivoting is used, the backward errors ΔA and Δb are bounded row-wise in terms of the element growth within each row, as measured by the α_i and β_i. Ideally, we would like the overall row-wise growth factor

$$\rho_{m,n} = \max_i \{\, \alpha_i, \beta_i \,\}$$

to be of order 1. In general, $\rho_{m,n}$ is unbounded. However, two techniques lead to a bounded $\rho_{m,n}$. Prior to carrying out the factorization, we can sort the rows so that

$$\|A(i,:)\|_\infty = \max_{j\ge i} \|A(j,:)\|_\infty, \qquad i = 1\colon m.$$

Alternatively, we can use row pivoting: at the kth stage of the factorization, after the column interchange has taken place, we interchange rows to ensure that

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

The following result was obtained by Powell and Reid [14] for row pivoting and by Cox and Higham [8] for row sorting. As usual for growth factor bounds, this one assumes exact arithmetic; however, the growth of exact and computed quantities differs by only $O(u)$.

THEOREM 4.2. *With row pivoting or row sorting in Householder QR factorization with column pivoting applied to* $A \in \mathbb{R}^{m\times n}$,

$$\rho_{m,n} \le \sqrt{m}(1+\sqrt{2})^{n-1}.$$

The bound of the theorem can be nearly attained, but $\rho_{m,n}$ is almost always small in practice.

The scalar ϕ in Theorem 4.1 is easily seen to be independent of the row ordering and so is beyond our control.

The limit technique used in the derivation of Algorithm EH can be used to obtain a row-wise backward error result. We apply Theorem 4.1 to $\min_x \|f_\mu - C_\mu x\|_2$ (see (2.1)) and then use a straightforward limit argument to deduce the following theorem. As a check on the plausibility of the theorem, we note that a key result used in the proof of Theorem 4.1 is that the Householder multipliers are all bounded by $\sqrt{2}$ (without column pivoting these multipliers are unbounded). For Stage I of Algorithm EH, the Householder multipliers

$$\eta_{k,j} = 2\frac{v_k(k\colon p)^T c_j^{(k)}(k\colon p)}{v_k^T(k\colon p)v_k(k\colon p)}$$

are precisely those determined by QR factorization with column pivoting applied to B, as we have already noted. Therefore $|\eta_{k,j}| \le \sqrt{2}$. Stage II of Algorithm EH is standard QR factorization with column pivoting, so again the multipliers are bounded by $\sqrt{2}$.

THEOREM 4.3. *Let the LSE problem* (1.1) *be solved by Algorithm EH. Then the computed solution* $\widehat{x}$ *is the exact solution to*

$$\min_{(B+\Delta B)x = d+\Delta d} \|(b+\Delta b) - (A+\Delta A)x\|_2,$$

where, defining

$$\Delta C = \begin{bmatrix} \Delta B \\ \Delta A \end{bmatrix}, \qquad \Delta f = \begin{bmatrix} \Delta d \\ \Delta b \end{bmatrix},$$

we have

$$|\Delta c_{ij}| \le j^2 \widetilde{\gamma}_{p+m} \alpha_i \max_s |c_{is}|, \qquad |\Delta f_i| \le n^2 \widetilde{\gamma}_{p+m} \beta_i \max(\phi \max_j |c_{ij}|, |f_i|),$$

with

$$\alpha_i = \frac{\max_{j,k} |\widehat{c}_{ij}^{(k)}|}{\max_j |c_{ij}|}, \quad \beta_i = \frac{\max\bigl(\phi \max_{j,k} |\widehat{c}_{ij}^{(k)}|, |\widehat{f}_i^{(k)}|\bigr)}{\max(\phi \max_j |c_{ij}|, |f_i|)},$$

$$\phi = \max \left\{ \max_{1 \le k \le p} \frac{\|\widehat{f}^{(k)}(k\colon p)\|_2}{\|\widehat{c}_k^{(k)}(k\colon p)\|_2}, \max_{p+1 \le k \le p+m} \frac{\|\widehat{f}^{(k)}(k\colon p+m)\|_2}{\|\widehat{c}_k^{(k)}(k\colon p+m)\|_2} \right\}.$$

Theorem 4.3 shows that Algorithm EH is row-wise backward stable provided that the row-wise growth factor $\rho_{m,n} = \max_i\{\alpha_i, \beta_i\}$ and the scalar ϕ are both of order 1. As for the LS problem, we can use row sorting or row pivoting in order to bound $\rho_{m,n}$, and we obtain a bound analogous to that in Theorem 4.2. For row sorting we sort the rows of A and B separately (that is, we do not sort the rows of C, which would change the LSE problem). For row pivoting, we interchange the rows of B in Stage I and the rows of A in Stage II.

The two stages of Algorithm EH can be implemented as a single loop with the help of a parameter that specifies the extent of the Householder vector, yielding the succinct pseudocode shown in Figure 4.1. The EG method and a variant of it based on modified Gram-Schmidt orthogonalization can be expressed in similarly concise fashions, as shown in [4], [5].

How does the stability of Algorithm EH compare with that of the EG method? The only published result for the EG method is one of Barlow and Handy [2], which bounds $\|C - \widehat{Q}\widehat{R}\|_2$ for the factorization (3.1) in terms of $\|C\|_2$; this result does not reveal the stability of the overall solution process or provide any information about row-wise stability.

From the remarks at the end of section 2 we know that if B is upper triangular, then Algorithm EH is identical to the EG method. We can therefore interpret the EG method as the two-stage process: (1) QR factorize B, using column pivoting, (2) apply Algorithm EH to the LSE problem with a triangular constraint matrix. Theorem 4.3 applies to stage (2), and an analogue of Theorem 4.1 for the QR factorization itself applies to Stage 1 [8] (essentially, we obtain $(B + \Delta B)\Pi = \widehat{Q}\widehat{R}$, with ΔB bounded in the same way as ΔA in Theorem 4.1). These two results can be combined, using the same techniques as in the proof of Theorem 4.1, to obtain a new result for the EG method.

Theorem 4.4. *Theorem* 4.3 *holds also for the EG method.*

5. Numerical experiments. Backward errors of an approximate solution y to the LSE problem (1.1) can be defined by minimizing suitable measures of the quadruple $(\Delta A, \Delta b, \Delta B, \Delta d)$ over all perturbations ΔA, Δb, ΔB, and Δd for which y solves $\min_{(B+\Delta B)y=d+\Delta d} \|b + \Delta b - (A + \Delta A)y\|_2$. Ideally, we would verify the row-wise backward stability results in Theorems 4.3 and 4.4 by comparing the actual row-wise backward errors with the bounds for a range of problems. Unfortunately, no computable expression is known for any backward error (even just the normwise backward error) of an arbitrary approximate LSE solution. We therefore compute upper bounds for the backward errors using a technique developed in [7].

Given an approximate LSE solution y, suppose we have determined perturbations ΔB and Δd such that $(B + \Delta B)y = d + \Delta d$. Let ΔA_* and Δb_* be solutions of the problem

$$(5.1)\quad \min\{\, \|\,[\,\Delta A \quad \theta\Delta b\,]\,\|_F : y \text{ solves } \min_{(B+\Delta B)x=d+\Delta d} \|b + \Delta b - (A + \Delta A)x\|_2 \,\},$$

This algorithm solves the LSE problem $\min_{Bx=d} \|b - Ax\|_2$*, where* $A \in \mathbb{R}^{m\times n}$*,* $B \in \mathbb{R}^{p\times n}$ *and* $m + p \geq n \geq p$*. A flag rp determines whether row pivoting is carried out.*

```
$C^{(1)} = \begin{bmatrix} B \\ A \end{bmatrix}$, $f^{(1)} = \begin{bmatrix} d \\ b \end{bmatrix}$, $top = p$, $q = p + m$.

for $k = 1\colon \min(n, m + p - 1)$

    Permute columns so that $\|C^{(k)}(k\colon top, k)\|_2 = \max_{j \geq k} \|C^{(k)}(k\colon top, j)\|_2$.

    If rp
       Permute rows so that $|c_{kk}^{(k)}| = \max_{i=k:top} |c_{ik}^{(k)}|$.
       Apply same row interchanges to $f^{(k)}$.
    end

    % Implicitly construct Householder-like matrix $I - \beta v v'^T$.
    $v = C^{(k)}(k\colon q, k)$
    $s = \mathrm{sign}(v_1)\|C^{(k)}(k\colon top, k)\|_2$
    $v_1 = v_1 + s$
    $\beta = 1/(s\, v_1)$

    $C^{(k+1)}(k\colon q, k\colon n) = C^{(k)}(k\colon q, k\colon n)$
                         $- \beta\, v(v(1\colon top - k + 1)^T C^{(k)}(k\colon top, k\colon n))$
    $f^{(k+1)}(k\colon q) = f^{(k)}(k\colon q) - \beta\, v(v(1\colon top - k + 1)^T f^{(k)}(k\colon top))$

    % Once $k > p$ we do standard Householder QR factorization.
    If $k = p$, $top = q$, end

end

Solve $C^{(k+1)}(1\colon n, 1\colon n)x = f^{(k+1)}(1\colon n)$ by substitution.
Permute $x$ to undo the effect of column pivoting.
```

FIG. 4.1. *Concise pseudocode for Algorithm EH.*

where θ is a parameter. How to compute ΔA_* and Δb_* is explained in [7]. Here we are minimizing over ΔA and Δb for fixed ΔB and Δd instead of minimizing over all ΔA, Δb, ΔB and Δd. Therefore by taking an appropriate measure of the quadruple $(\Delta A_*, \Delta b_*, \Delta B, \Delta d)$ we obtain an upper bound for the corresponding backward error. We choose the perturbations

$$\Delta B_* = \frac{\|B\|_2\|y\|_2}{(\|B\|_2\|y\|_2 + \|d\|_2)} \frac{ry^T}{y^T y}, \quad \Delta d_* = \frac{-\|d\|_2}{\|B\|_2\|y\|_2 + \|d\|_2} r,$$

where $r = d - By$. These are the perturbations corresponding to the relative 2-norm backward error for the constraint system [16]; in other words, they minimize $\max(\|\Delta B\|_2/\|B\|_2, \|\Delta d\|_2/\|d\|_2)$. We define two backward error bounds in terms of the perturbations ΔA_*, Δb_*, ΔB_* and Δd_*: the normwise backward error bound

$$\eta_N = \max\left\{ \frac{\|\Delta A_*\|_2}{\|A\|_2}, \frac{\|\Delta b_*\|_2}{\|b\|_2}, \frac{\|\Delta B_*\|_2}{\|B\|_2}, \frac{\|\Delta d_*\|_2}{\|d\|_2} \right\}$$

and the row-wise backward error bound

$$\eta_R = \max_i \frac{\|\Delta G(i,:)\|_2}{\|G(i,:)\|_2},$$

where

$$G = \begin{bmatrix} B & d \\ A & b \end{bmatrix}, \qquad \Delta G = \begin{bmatrix} \Delta B_* & \Delta d_* \\ \Delta A_* & \Delta b_* \end{bmatrix}.$$

In our experiments we took $\theta = 1$ in (5.1) since our chosen A, B, d, and f all have norms of order 1.

The experiments were performed in MATLAB, using simulated single precision arithmetic ($u = 2^{-24} \approx 5.96 \times 10^{-8}$). The backward error bounds were computed in double precision arithmetic.

We give results for four problems with $m = 16$, $n = 10$, $p = 6$. We denote by randn a matrix or vector from the normal(0,1) distribution and by randsvd(κ) a random matrix with 2-norm condition number κ and geometrically distributed singular values, generated by the routine `randsvd` in the Test Matrix Toolbox [11].

PROBLEM 1 $A = \text{randn}$, $b = \text{randn}$, $B = \text{randn}$, $d = \text{randn}$.

PROBLEM 2 $A = \text{randsvd}(10)$, $b = \text{randn}$, $B = \text{randsvd}(10^4)$, $d = \text{randn}$.

PROBLEM 3 $A = \text{randsvd}(10^6)$, $b = \text{randn}$, $B = \text{randsvd}(10)$, $d = \text{randn}$.

PROBLEM 4 $A = \text{randsvd}(10^4)$, $b = \text{randn}$, $B = \text{randsvd}(10^4)$, $d = \text{randn}$.

For each problem, a parameter tol $\in (0,1]$ determines a row scaling applied to the original data:

$$[\,A \quad b\,] \leftarrow D_{m,\text{tol}}\,[\,A \quad b\,], \quad [\,B \quad d\,] \leftarrow D_{p,\text{tol}}\,[\,B \quad d\,],$$

where

$$D_{k,\text{tol}} = \text{diag}(\theta^k, \theta^{k-1}, \dots, \theta, 1), \quad \theta^k = \text{tol}.$$

The backward error results are given in Tables 5.1–5.3. We give results for Algorithm EH with no row interchanges and with row sorting; the results for row pivoting were very similar and so are omitted. We do not report results for the EG method because they were very similar to those for Algorithm EH. For comparison, we also give results for the null space method (implemented using generalized QR factorization, as in LAPACK). Results for tol = 1 are given only for Problem 1; for Problems 2 and 3 with tol = 1 the backward error bounds were again all of order u.

The main observations from the results are as follows.

1. The importance of row pivoting or row sorting in Algorithm EH for badly row scaled problems is shown by Tables 5.1–5.3, in which when tol is small η_R is much smaller when row sorting is used than when it is not. Row sorting produces a growth factor $\rho_{m,n}$ of order 1 in each case.
2. Tables 5.2 and 5.3 show that a large value of ϕ does not necessarily lead to a large row-wise backward error. Although ϕ is theoretically independent of the row ordering, it varies between Algorithm EH with and without row sorting in some of the tests as a result of roundoff: in these cases ϕ is determined by ratios of quantities that are of order u and have large relative error.

Table 5.1
Results for Problem 1 *(random normal A and B).*

tol $= 1$

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	6.3e-8	6.4e-8	4.5e0	1.5e0
Alg. EH (row sorting)	4.8e-8	4.5e-8	2.6e0	1.5e0
Null space method	3.5e-8	3.9e-8		

tol $= 10^{-7}$

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	9.6e-4	2.1e-1	2.5e7	3.8e0
Alg. EH (row sorting)	3.3e-8	4.3e-7	3.0e0	1.7e1
Null space method	5.1e-8	2.7e-7		

Table 5.2
Results for Problem 2 *($\kappa_2(A) = 10$, $\kappa_2(B) = 10^4$),* tol $= 10^{-7}$.

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	4.3e-4	1.1e-2	6.6e6	4.6e3
Alg. EH (row sorting)	2.4e-8	1.6e-7	2.7e0	1.4e4
Null space method	7.2e-8	2.2e-7		

3. The normwise backward error bound η_N is small in every case except for Algorithm EH without row interchanges in two instances. However, using numerical optimization by direct search we found alternative choices of ΔB_* and Δd_* for which $\eta_N \approx u$ in these two examples, confirming normwise backward stability. In the cases where Algorithm EH without row interchanges produced a large value of η_R, we were unable to reduce this value significantly using direct search.
4. The null space method gives very similar values of η_N and η_R to Algorithm EH with row sorting. It is surprising that η_R is small for the null space method when tol $= 10^{-7}$ because our implementation does not incorporate any form of row interchanges. The reason for this excellent performance is not clear (it is not explained by the normwise backward error results in [6], for example).

Finally, although our concern in this work is with backward errors rather than forward errors, we carried out an experiment to show the effect of row interchanges on the forward error, $\|x - \widehat{x}\|_2 / \|x\|_2$. For Problems 1 and 4 we computed forward errors by taking as the exact solution the one computed in double precision by Algorithm EH with row sorting. Table 5.4 shows that row sorting can greatly decrease the forward error when A and B have badly scaled rows.

6. Concluding remarks. We have used a limit argument to derive a method, Algorithm EH, for solving the LSE problem and have derived a row-wise backward error bound for the method by exploiting existing error analysis. The row-wise backward error is guaranteed to be small when the growth factor $\rho_{m,n}$ is small (which is usually the case with row pivoting or sorting) and when the scalar ϕ is small (ϕ can be arbitrarily large, and is independent of the row interchanges).

The method is closely related to the EG method, which dates back to the 1960s, and this relation enabled us to obtain a row-wise backward error bound for that

TABLE 5.3
Results for Problem 3 *(*$\kappa_2(A) = 10^6$*,* $\kappa_2(B) = 10$*),* tol $= 10^{-7}$.

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	1.3e-8	2.8e-2	4.3e6	1.1e3
Alg. EH (row sorting)	2.8e-8	1.3e-7	2.2e0	1.2e3
Null space method	2.8e-8	7.3e-7		

TABLE 5.4
Forward errors $\|x - \widehat{x}\|_2 / \|x\|_2$.

	Problem 1	Problem 1	Problem 4	Problem 4
	tol = 1	tol $= 10^{-7}$	tol = 1	tol $= 10^{-7}$
Alg. EH (no row interchanges)	2.5e-7	6.6e-1	2.4e-5	1.3e0
Alg. EH (row sorting)	1.7e-7	1.2e-6	3.1e-6	2.1e-5
Null space method	1.2e-7	2.8e-6	4.9e-5	5.8e-6

method, too. Which method should be preferred? There is no difference between the two methods in terms of the backward error results or the numerical backward error bounds that we have evaluated. The methods also have similar operation counts. Algorithm EH can be coded slightly more concisely, since it carries out transformations on B-over-A as a whole while the EG method applies Householder transformations to B and separate elimination operations to A, and this difference should make Algorithm EH more efficient on high-performance computers.

REFERENCES

[1] J. L. BARLOW, *Error analysis and implementation aspects of deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 25 (1988), pp. 1340–1358.

[2] J. L. BARLOW AND S L. HANDY, *The direct solution of weighted and equality constrained least-squares problems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 704–716.

[3] J. L. BARLOW AND U. B. VEMULAPATI, *A note on deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 29 (1992), pp. 249–256.

[4] Å. BJÖRCK, *Iterative refinement of linear least squares solutions* II, BIT, 8 (1968), pp. 8–30.

[5] Å. BJÖRCK AND G. H. GOLUB, *Iterative refinement of linear least squares solutions by Householder transformation*, BIT, 7 (1967), pp. 322–337.

[6] A. J. COX AND N. J. HIGHAM, *Accuracy and Stability of the Null Space Method for Solving the Equality Constrained Least Squares Problem*, BIT, 39 (1999), pp. 34–50.

[7] A. J. COX AND N. J. HIGHAM, *Backward Error Bounds for Constrained Least Squares Problems*, BIT, 39 (1999), pp. 210–227.

[8] A. J. COX AND N. J. HIGHAM, *Stability of Householder QR factorization for weighted least squares problems*, Numerical Analysis 1997, Proceedings of the 17th Dundee Biennial Conference, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., Pitman Res. Notes Math. Ser., 380, Addison Wesley Longman, Harlow, UK, 1998, pp. 57–73.

[9] M. GULLIKSSON, *Backward error analysis for the constrained and weighted linear least squares problem when using the weighted QR factorization*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 675–687.

[10] M. GULLIKSSON AND P. WEDIN, *Modifying the QR-decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.

[11] N. J. HIGHAM, *The Test Matrix Toolbox for* MATLAB *(version* 3.0*)*, Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manchester, England, 1995.

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

[13] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, SIAM, Philadelphia, PA, 1995; revised republication of work first published by Prentice–Hall, 1974.

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

[15] J. K. Reid, *Implicit scaling of linear least squares problems*, Report RAL-TR-98-027, Atlas Centre, Rutherford Appleton Laboratory, Didcot, Oxon, UK, 1998. 12 pp.

[16] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14(3) (1967), pp. 543–548.

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

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