BIT
1999, Vol. 39, No. 1, pp. 34–50
0006-3835/98/3901-0034 $15.00


# ACCURACY AND STABILITY OF THE NULL SPACE METHOD FOR SOLVING THE EQUALITY CONSTRAINED LEAST SQUARES PROBLEM *

ANTHONY J. COX[1] and NICHOLAS J. HIGHAM[2]

[1] *Department of Mathematics, University of Manchester, Manchester, M13 9PL, England. email: higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham*

[1] *Ground Processing Department, Mail Point 31, Matra Marconi Space UK, Anchorage Road Portsmouth, PO3 5PU, UK. email: anthonyj.cox@mmsuk.co.uk*

**Abstract.**

The null space method is a standard method for solving the linear least squares problem subject to equality constraints (the LSE problem). We show that three variants of the method, including one used in LAPACK that is based on the generalized QR factorization, are numerically stable. We derive two perturbation bounds for the LSE problem: one of standard form that is not attainable, and a bound that yields the condition number of the LSE problem to within a small constant factor. By combining the backward error analysis and perturbation bounds we derive an approximate forward error bound suitable for practical computation. Numerical experiments are given to illustrate the sharpness of this bound.

*Key words:* Constrained least squares problem, null space method, rounding error analysis, condition number, generalized QR factorization, LAPACK

*AMS subject classification:* 65F20, 65G05

## 1 The LSE problem.

We consider the least squares problem with equality constraints

$$\text{LSE}: \qquad \min_{Bx=d} \|b - Ax\|_2, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$, with $m + p \geq n \geq p$. We will assume that

$$\text{rank}(B) = p, \qquad \text{null}(A) \cap \text{null}(B) = \{0\}. \tag{1.2}$$

The assumption that $B$ is of full rank ensures that the system $Bx = d$ is consistent and hence that the LSE problem has a solution. The second condition in (1.2), which is equivalent to the condition that the matrix $[A^T,\ B^T]^T$ has full rank $n$, then guarantees that there is a unique solution [7, Section 5.1]. Note

*Received August 1997. Revised August 1998. Communicated by Axel Ruhe.

that the condition $m \geq n - p$ ensures that the LSE problem is overdetermined and is more general than the common assumption $m \geq n$.

The LSE problem arises in various applications, including the analysis of large-scale structures [4] and the solution of the inequality constrained least square problem [16, Chap. 23].

A standard method for solving the LSE problem is the null space method, of which there are at least three versions in the literature. Existing error analysis applies to just one of the methods and is unnecessarily pessimistic. We prove that all three versions are numerically stable. We also give normwise perturbation theory for the LSE problem. Combining the error analysis and the perturbation theory we obtain an error bound that can be estimated efficiently in practice. The main motivation for our work comes from LAPACK, which has a driver routine `xgglse.f` for solving the LSE problem. There is no existing error analysis for `xgglse.f` and this driver does not currently compute error bounds.

In Section 2 we introduce the generalized QR factorization and describe the three null space methods. The stability of the methods is proved in Section 3 via detailed rounding error analysis. In Section 4 we present perturbation theory for the LSE problem. We derive a bound essentially the same as one of Eldén [8]. We explain why the bound is not sharp and, by modifying the analysis, derive an almost sharp bound that yields a quantity that is within a small constant factor of the condition number for the LSE problem. Finally, in Section 5 we describe a practical error bound and present some numerical experiments to show the sharpness of the bound.

## 2 The null space method.

We describe three versions of the null space method for solving the LSE problem, so-called because it employs an orthogonal basis for the null space of the constraint matrix. We begin with a version based on the generalized QR factorization. The generalized QR factorization was introduced by Hammarling [12] and Paige [17] and further analyzed by Anderson, Bai and Dongarra [2].

THEOREM 2.1. (GENERALIZED QR FACTORIZATION) *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$ with $m + p \geq n \geq p$. There are orthogonal matrices $Q \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ such that*

$$U^T A Q = \begin{array}{c} \\ m-n+p \\ n-p \end{array} \begin{array}{c} \begin{array}{cc} p & n-p \end{array} \\ \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \end{array}, \qquad BQ = \begin{array}{c} \\ p \end{array} \begin{array}{c} \begin{array}{cc} p & n-p \end{array} \\ \begin{pmatrix} S & 0 \end{pmatrix} \end{array}, \tag{2.1}$$

*where $L_{22}$ and $S$ are lower triangular. More precisely, we have*

$$U^T A Q = \begin{cases} \begin{array}{c} \\ m-n \\ n \end{array} \begin{array}{c} n \\ \begin{pmatrix} 0 \\ L \end{pmatrix} \end{array} & \text{if } m \geq n, \\ \begin{array}{c} \\ m \end{array} \begin{array}{c} \begin{array}{cc} n-m & m \end{array} \\ \begin{pmatrix} X & L \end{pmatrix} \end{array} & \text{if } m < n, \end{cases} \tag{2.2}$$

*where $L$ is lower triangular. The assumptions* (1.2) *are equivalent to $S$ and $L_{22}$ being nonsingular.*

PROOF. Let

$$Q^T B^T = \begin{bmatrix} S^T \\ 0 \end{bmatrix}$$

be a QR factorization of $B^T$. We can determine an orthogonal $U$ so that $U^T(AQ)$ has the form (2.2), where $L$ is lower triangular (for example, we can construct $U$ as a product of suitably chosen Householder transformations). Clearly, $B$ has full rank if and only if $S$ is nonsingular. Partition $Q = [Q_1 \;\; Q_2]$ conformably with $[S \;\; 0]$ and assume $S$ is nonsingular. Then, clearly, $\mathrm{null}(B) = \mathrm{range}(Q_2)$. We can write

$$A\,[\,Q_1 \quad Q_2\,] = [\,U_1 \quad U_2\,] \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix},$$

so that $AQ_2 = U_2 L_{22}$. It follows that $\mathrm{null}(A) \cap \mathrm{null}(B) = \{0\}$ is equivalent to $L_{22}$ being nonsingular. □

While (2.2) is needed to define the generalized QR factorization precisely, the partitioning of $U^T AQ$ in (2.1) enables us to explain the application to the LSE problem without treating the cases $m \geq n$ and $m < n$ separately.

Using (2.1) the constraint $Bx = d$ may be written

$$Sy_1 = [\,S \quad 0\,] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = d, \qquad y = Q^T x.$$

Hence the constraint determines $y_1 \in \mathbb{R}^p$ as the solution of the triangular system $Sy_1 = d$ and leaves $y_2 \in \mathbb{R}^{n-p}$ arbitrary. Since

$$\|b - Ax\|_2 = \|c - U^T AQy\|_2, \qquad c = U^T b,$$

we see that we have to find

$$\min_{y_2} \left\| \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} - \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\|_2 = \min_{y_2} \left\| \begin{bmatrix} c_1 - L_{11}y_1 \\ (c_2 - L_{21}y_1) - L_{22}y_2 \end{bmatrix} \right\|_2.$$

Therefore $y_2$ is the solution to the triangular system $L_{22}y_2 = (c_2 - L_{21}y_1)$. The solution $x$ is recovered from $x = Qy$. We refer to this particular solution process as the GQR method.

Note that the GQR method carries out slightly more computation than necessary, since if the first block column of $U^T AQ$ in (2.1) is full a solution can be computed in a similar way. We can therefore redefine $U$ to reduce just the last $n-p$ columns of $AQ$ to lower triangular form (so that $U$ is a product of just $n-p$ Householder transformations) and not explicitly apply $U$ to the first $p$ columns of $AQ$:

$$[\,W_1 \quad W_2\,] = AQ = A \overset{\begin{matrix} p & n-p \end{matrix}}{\begin{pmatrix} Q_1 & Q_2 \end{pmatrix}}, \qquad U^T W_2 = \begin{bmatrix} 0 \\ L_{22} \end{bmatrix}.$$

Then

$$\begin{aligned}\|b - Ax\|_2 &= \left\| b - [\, W_1 \quad W_2 \,] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\|_2 = \|(b - W_1 y_1) - W_2 y_2\|_2 \\ &= \left\| U^T (b - W_1 y_1) - \begin{bmatrix} 0 \\ L_{22} \end{bmatrix} y_2 \right\|_2 \\ &= \left\| \begin{bmatrix} g_1 \\ g_2 - L_{22} y_2 \end{bmatrix} \right\|_2,\end{aligned}$$

where

$$y = Q^T x, \qquad \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = U^T (b - W_1 y_1),$$

and we solve $L_{22} y_2 = g_2$. This is the form of the null space method described by Lawson and Hanson [16, Chap. 20] and Golub and Van Loan [11, Sec. 12.1.4]; we will call it Method 2.

The computational cost can be reduced further by observing that we do not need to form $W_1$ explicitly, since $W_1 y_1 = A(Q_1 y_1)$; it suffices to compute matrix-vector products with $A$ and $Q_1$. Exploiting this observation we obtain the form of the null space method described by Björck [7, Section 5.1.3] and Van Loan [18], which we will call Method 3.

The distinction between Methods 2 and 3 has not, to our knowledge, previously been pointed out in the literature. Approximate operation counts for all three methods are given in Table 2 assuming the use of Householder transformations and where a flop denotes a floating point operation.

Table 2.1: Approximate flop counts for the three null space methods.

| | |
|---|---|
| GQR Method | $2mn^2 + 4mnp + 2np^2 - 2mp^2 - 2n^3/3 - 2p^3/3$ |
| Method 2 | $2m(n-p)^2 + 4mnp + 2np^2 - 2mp^2 - 2p^3/3 - 2(n-p)^3/3$ |
| Method 3 | $4n^2p + 4p^3/3 + 2mn(n-p) + 2m(n-p)^2 - 2p^3/3$ $- 2np^2 - 2(n-p)^3/3$ |

## 3 Error analysis.

We now investigate the stability of the three variations of the null space method. We use the standard model of floating point arithmetic [15, Section 2.2], namely

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

where $u$ is the unit roundoff, together with the (equally valid) variation

$$fl(x \operatorname{op} y) = \frac{x \operatorname{op} y}{1+\delta}, \qquad |\delta| \le u, \quad \operatorname{op} = +, -, *, /. \tag{3.1}$$

We define the constants

$$\gamma_n = \frac{nu}{1-nu}, \qquad \widetilde{\gamma}_n = \frac{cnu}{1-cnu},$$

where $c$ denotes a small integer constant whose precise value is unimportant. First, we examine the GQR method.

THEOREM 3.1. *Suppose the LSE problem* (1.1) *is solved using the GQR method, where the generalized QR factorization is computed using Householder transformations and where the assumptions* (1.2) *are satisfied. The computed solution* $\widehat{x} = \overline{x} + \Delta\overline{x}$, *where* $\overline{x}$ *solves* $\min\{\|b+\Delta b-(A+\Delta A)x\|_2 : (B+\Delta B)x = d\}$, *and where*

$$\begin{aligned} \|\Delta\overline{x}\|_2 &\le \widetilde{\gamma}_{np}\|\overline{x}\|_2, & \|\Delta b\|_2 &\le \widetilde{\gamma}_{mn}\|b\|_2, \\ \|\Delta A\|_F &\le \widetilde{\gamma}_{mn}\|A\|_F, & \|\Delta B\|_F &\le \widetilde{\gamma}_{np}\|B\|_F. \end{aligned}$$

PROOF. The proof involves careful combination of standard results on matrix–vector multiplication, solution of triangular systems, and application of Householder transformations.

Let $\widehat{L}$ and $\widehat{S}$ be the computed triangular matrices in the generalized QR factorization (2.1). We deduce from [15, Lemma 18.3, Thm. 18.4] that there exist orthogonal $\widetilde{Q} \in \mathbb{R}^{n\times n}$ and $\widetilde{U} \in \mathbb{R}^{m\times m}$ such that

$$(B + \Delta B_1)\widetilde{Q} = \begin{bmatrix} \widehat{S} & 0 \end{bmatrix}, \qquad \|\Delta B_1\|_F \le \widetilde{\gamma}_{np}\|B\|_F, \tag{3.2}$$

$$\widetilde{U}^T(A + \Delta A_1)\widetilde{Q} = \begin{bmatrix} \widehat{L}_{11} & 0 \\ \widehat{L}_{21} & \widehat{L}_{22} \end{bmatrix}, \qquad \|\Delta A_1\|_F \le \widetilde{\gamma}_{mn}\|A\|_F, \tag{3.3}$$

where $\Delta A_1$ includes the effect of the errors in forming $fl(AQ)$ and the errors in the reduction to lower triangular form. The results that we state for multiplication by $Q$ and $U$ hold whether the multiplications are done using the representation of each matrix as a product of Householder transformations or by computing $Q$ and $U$ explicitly and carrying out a full matrix–matrix or matrix–vector multiplication.

The computed $\widehat{y}_1$ satisfies [15, Thm. 8.5]

$$(\widehat{S} + \Delta S)\widehat{y}_1 = d, \qquad \|\Delta S\|_F \le \gamma_p\|\widehat{S}\|_F, \tag{3.4}$$

while we have [15, Lemma 18.3]

$$\widehat{c} = \widetilde{U}^T(b + \Delta b), \qquad \|\Delta b\|_2 \le \widetilde{\gamma}_{mn}\|b\|_2. \tag{3.5}$$

Now we consider the computation of $\widehat{y}_2$; care is required here to avoid introducing unnecessary condition number terms. First, we form the vector $g = c_2 - L_{21}y_1$. The computed vector can be written, using (3.1) and [15, Section 3.5]

$$\begin{aligned} \widehat{g} &= (I + D)^{-1}\big(\widehat{c}_2 - (\widehat{L}_{21} + \Delta L_{21})\widehat{y}_1\big), \\ D &= \operatorname{diag}(\delta_i), \quad |\delta_i| \le u, \qquad \|\Delta L_{21}\|_F \le \gamma_p\|\widehat{L}_{21}\|_F. \end{aligned}$$

Then we have

$$(\widehat{L}_{22} + \widetilde{\Delta L_{22}})\widehat{y}_2 = \widehat{g}, \qquad \|\widetilde{\Delta L_{22}}\|_F \le \gamma_{n-p}\|\widehat{L}_{22}\|_F.$$

Hence

$$\widehat{c}_2 - (\widehat{L}_{21} + \Delta L_{21})\widehat{y}_1 = (I + D)(\widehat{L}_{22} + \widetilde{\Delta L}_{22})\widehat{y}_2 = (\widehat{L}_{22} + \Delta L_{22})\widehat{y}_2, \tag{3.6}$$

where

$$\|\Delta L_{22}\|_F \le (1+u)\gamma_{n-p}\|\widehat{L}_{22}\|_F + u\|\widehat{L}_{22}\|_F = \widetilde{\gamma}_{n-p}\|\widehat{L}_{22}\|_F.$$

From (3.2), (3.3), (3.5) and (3.6) it follows that $\overline{x} = \widetilde{Q}\widehat{y}$ is the exact LSE solution corresponding to the perturbed data $A + \Delta A$, $B + \Delta B$ and $b + \Delta b$, where $\Delta b$ is defined in (3.5) and

$$\begin{aligned} \Delta B &= \Delta B_1 + [\,\Delta S \quad 0\,]\,\widetilde{Q}^T, \\ \Delta A &= \Delta A_1 + \widetilde{U}\begin{bmatrix} 0 & 0 \\ \Delta L_{21} & \Delta L_{22} \end{bmatrix}\widetilde{Q}^T. \end{aligned}$$

The bounds on the norms of $\Delta b$, $\Delta A$ and $\Delta B$ are immediate. Finally, we have

$$\widehat{x} = fl(Q\widehat{y}) = \widetilde{Q}\widehat{y} + \Delta\overline{x}, \qquad \|\Delta\overline{x}\|_2 \le \widetilde{\gamma}_{np}\|\widehat{y}\|_2 = \widetilde{\gamma}_{np}\|\overline{x}\|_2, \tag{3.7}$$

using [15, Lemma 18.3] again. □

Theorem 3.1 shows that the computed LSE solution is close to the solution of a slightly perturbed problem, using normwise measures; in other words, the GQR method is numerically stable in a mixed forward/backward sense. We can obtain a genuine backward error result at the cost of weakening the bound on $\Delta b$ and perturbing $d$.

THEOREM 3.2. *Under the same assumptions as in Theorem* 3.1*, the computed solution* $\widehat{x}$ *from the GQR method solves* $\min\{\|b + \Delta b - (A + \Delta A)x\|_2 : (B + \Delta B)x = d + \Delta d\}$*, where*

$$\begin{aligned} &\|\Delta b\|_2 \le \widetilde{\gamma}_{mn}\|b\|_2 + \widetilde{\gamma}_{np}\|A\|_F\|\widehat{x}\|_2, \quad &&\|\Delta A\|_F \le \widetilde{\gamma}_{mn}\|A\|_F, \\ &\|\Delta B\|_F \le \widetilde{\gamma}_{np}\|B\|_F, &&\|\Delta d\|_2 \le \widetilde{\gamma}_{np}\|B\|_F\|\widehat{x}\|_2. \end{aligned}$$

PROOF. We modify the proof of Theorem 3.1 by rewriting $\widehat{x}$ as

$$\widehat{x} = \widetilde{Q}\overline{y}, \qquad \overline{y} = \widehat{y} + \Delta\widehat{y}, \qquad \|\Delta\widehat{y}\|_2 \le \widetilde{\gamma}_{np}\|\widehat{y}\|_2.$$

The error $\Delta\widehat{y}$ now has to be accounted for by perturbations to $d$ and $b$. Equation (3.4) may be rewritten

$$\begin{aligned} (\widehat{S} + \Delta S)\overline{y}_1 = d + \Delta d, \qquad \Delta d = (\widehat{S} + \Delta S)\Delta\widehat{y}_1, \\ \|\Delta d\|_2 \le \widetilde{\gamma}_{np}\|B\|_F\|\widehat{y}\|_2 = \widetilde{\gamma}_{np}\|B\|_F\|\widehat{x}\|_2. \end{aligned}$$

Equation (3.6) may be rewritten

$$(\widehat{L}_{22} + \Delta L_{22})\overline{y}_2 = \left[\widehat{c}_2 + (\widehat{L}_{22} + \Delta L_{22})\Delta\widehat{y}_2\right] - (\widehat{L}_{21} + \Delta L_{21})\widehat{y}_1$$

and we can redefine $\widehat{c}_2$ to be the term in square brackets if we add

$$\widetilde{U}\begin{bmatrix} 0 \\ (\widehat{L}_{22}+\Delta L_{22})\Delta\widehat{y}_2 \end{bmatrix}$$

to $\Delta b$ in (3.5). We then have

$$\begin{aligned}\|\Delta b\|_2 &\le \widetilde{\gamma}_{mn}\|b\|_2 + \|A\|_F\widetilde{\gamma}_{np}\|\widehat{y}\|_2\\ &= \widetilde{\gamma}_{mn}\|b\|_2 + \widetilde{\gamma}_{np}\|A\|_F\|\widehat{x}\|_2.\end{aligned}$$

The result follows. □

Note that Theorem 3.2 does not prove normwise backward stability since the perturbation in $b$ can be much larger than $u\|b\|_2$, and likewise for $d$.

Whether the conclusions that we have drawn for the GQR method hold also for Method 2 is not immediately obvious. Indeed, Galligani and Laratta [10] give an error analysis of Method 2 in which they obtain a result of the form in Theorem 3.1 but with an extra factor $\kappa_2(B)\|B^+d\|_2$ in the bound for $\|\Delta b\|_2$, where $\kappa_2(B) = \|B\|_2\|B^+\|_2$. As we now show, this factor is unnecessary. The following lemma will be needed.

LEMMA 3.3. *Let $s = b - Ax + e$, for any conformable matrix $A$ and vectors $s, b, x, e$, where $\|e\|_2 \le \epsilon\|b - Ax\|_2$. Then $s = b + \Delta b - (A + \Delta A)x$, where $\|\Delta b\|_2 \le \epsilon\|b\|_2$ and $\|\Delta A\|_2 \le \epsilon\|A\|_2$.*

PROOF. Let $E = er^T/(r^Tr)$, where $r = b - Ax$. Then $\|E\|_2 = \|e\|_2/\|r\|_2 \le \epsilon$ and $e = Er$. Hence $s = b + Eb - (A + EA)x$, so the result holds with $\Delta b = Eb$ and $\Delta A = EA$. □

THEOREM 3.4. *Suppose the LSE problem* (1.1) *is solved by Method* 2*, using Householder transformations, and let the assumptions* (1.2) *be satisfied. The computed solution $\widehat{x} = \overline{x} + \Delta\overline{x}$, where $\overline{x}$ solves* $\min\{\|b + \Delta b - (A + \Delta A)x\|_2 : (B + \Delta B)x = d\}$*, and where*

$$\begin{aligned}\|\Delta\overline{x}\|_2 &\le \widetilde{\gamma}_{np}\|\overline{x}\|_2, & \|\Delta b\|_2 &\le \widetilde{\gamma}_{1+m(n-p)}\|b\|_2,\\ \|\Delta A\|_F &\le \widetilde{\gamma}_{np+m(n-p)}\|A\|_F, & \|\Delta B\|_F &\le \widetilde{\gamma}_{np}\|B\|_F.\end{aligned}$$

PROOF. As in the proof of Theorem 3.1, there exist orthogonal $\widetilde{Q} = \begin{bmatrix}\widetilde{Q}_1 & \widetilde{Q}_2\end{bmatrix} \in \mathbb{R}^{n\times n}$ and $\widetilde{U} \in \mathbb{R}^{m\times m}$ such that

$$\begin{aligned}(B + \Delta B_1)\widetilde{Q} &= \begin{bmatrix}\widehat{S} & 0\end{bmatrix}, & \|\Delta B_1\|_F &\le \widetilde{\gamma}_{np}\|B\|_F,\\ \begin{bmatrix}\widehat{W}_1 & \widehat{W}_2\end{bmatrix} &= (A + \Delta A_1)\widetilde{Q}, & \|\Delta A_1\|_F &\le \widetilde{\gamma}_{np}\|A\|_F,\\ \widetilde{U}^T(\widehat{W}_2 + \widetilde{\Delta W}_2) &= \begin{bmatrix}0\\ \widehat{L}_{22}\end{bmatrix}, & \|\widetilde{\Delta W}_2\|_F &\le \widetilde{\gamma}_{m(n-p)}\|\widehat{W}_2\|_F \le \widetilde{\gamma}_{m(n-p)}\|A\|_F.\end{aligned}$$

Again, the computed $\widehat{y}_1$ satisfies

$$(\widehat{S} + \Delta S)\widehat{y}_1 = d, \qquad \|\Delta S\|_F \le \gamma_p\|\widehat{S}\|_F.$$

For the formation of $h = b - W_1 y_1$, we obtain, using (3.1) and [15, Section 3.5],

$$\widehat{h} = (I + D)(b - (\widehat{W}_1 + E_1)\widehat{y}_1), \quad D = \mathrm{diag}(\delta_i), \quad |\delta_i| \le u, \quad \|E_1\|_F \le \gamma_p \|\widehat{W}_1\|_F$$

and

$$\widehat{h} = b + \Delta b_1 - (\widehat{W}_1 + E_2)\widehat{y}_1, \qquad \|\Delta b_1\|_2 \le u\|b\|_2, \quad \|E_2\|_F \le \widetilde{\gamma}_p \|\widehat{W}_1\|_F.$$

Then we form $g = U^T h$, obtaining [15, Lemma 18.3]

$$\widehat{g} = \widetilde{U}^T(\widehat{h} + \Delta h), \qquad \|\Delta h\|_2 \le \widetilde{\gamma}_{m(n-p)} \|\widehat{h}\|_2.$$

Invoking Lemma 3.3 we obtain

$$\widehat{g} = \widetilde{U}^T(b + \Delta b - (\widehat{W}_1 + \Delta W_1)\widehat{y}_1), \qquad \|\Delta b\|_2 \le \widetilde{\gamma}_{1+m(n-p)} \|b\|_2, \tag{3.8}$$

$$\|\Delta W_1\|_F \le \widetilde{\gamma}_{p+m(n-p)} \|\widehat{W}_1\|_F \le \widetilde{\gamma}_{p+m(n-p)} \|A\|_F.$$

The computed solution $\widehat{y}_2$ satisfies

$$(\widehat{L}_{22} + \Delta L_{22})\widehat{y}_2 = \widehat{g}_2, \qquad \|\Delta L_{22}\|_F \le \gamma_{n-p} \|\widehat{L}_{22}\|_F.$$

Hence $\widehat{y}_2$ is the solution to

$$\begin{aligned}
&\min_z \left\| \begin{bmatrix} \widehat{g}_1 \\ \widehat{g}_2 \end{bmatrix} - \begin{bmatrix} 0 \\ \widehat{L}_{22} + \Delta L_{22} \end{bmatrix} z \right\|_2 \\
&\quad = \min_z \left\| b + \Delta b - (\widehat{W}_1 + \Delta W_1)\widehat{y}_1 - \widetilde{U} \begin{bmatrix} 0 \\ \widehat{L}_{22} + \Delta L_{22} \end{bmatrix} z \right\|_2 \\
&\quad = \min_z \left\| b + \Delta b - (\widehat{W}_1 + \Delta W_1)\widehat{y}_1 - \left( \widehat{W}_2 + \widetilde{\Delta W}_2 + \widetilde{U} \begin{bmatrix} 0 \\ \Delta L_{22} \end{bmatrix} \right) z \right\|_2 \\
&\quad = \min_z \left\| b + \Delta b - (\widehat{W}_1 + \Delta W_1)\widehat{y}_1 - (\widehat{W}_2 + \Delta W_2) z \right\|_2,
\end{aligned}$$

where

$$\|\Delta W_2\|_F \le \widetilde{\gamma}_{m(n-p)} \|A\|_F.$$

Putting these results together, we find that $\overline{x} = \widetilde{Q}^T \widehat{y}$ is the exact LSE solution corresponding to the perturbed data $A + \Delta A$, $B + \Delta B$ and $b + \Delta b$, where $\Delta b$ is defined in (3.8) and

$$\begin{aligned}
\Delta B &= \Delta B_1 + [\, \Delta S \quad 0 \,] \widetilde{Q}^T, \\
\Delta A &= \Delta A_1 + [\, \Delta W_1 \quad \Delta W_2 \,] \widetilde{Q}^T.
\end{aligned}$$

The normwise bounds for $\Delta b$, $\Delta A$ and $\Delta B$ follow. The last part of the proof is the same as for Theorem 3.1. □

Theorem 3.4 shows that Method 2 enjoys the same stability result as the GQR method, with slightly improved dependence of the constants on the dimensions. It is straightforward to show that the result of Theorem 3.4 applies also to Method 3. Analogues of Theorem 3.2 hold for Methods 2 and 3.

Having obtained backward error bounds for the three versions of the null space method, we now wish to determine forward error bounds, that is, bounds for $\|x - \widehat{x}\|_2 / \|x\|_2$. The necessary perturbation theory is the subject of the next section.

## 4 Perturbation theory.

We wish to determine the sensitivity of the LSE solution to perturbations $\Delta A$, $\Delta b$, $\Delta B$ and $\Delta d$ in $A$, $b$, $B$ and $d$, respectively. We assume that the conditions (1.2) hold. We also assume that (1.2) holds for the perturbed data $A + \Delta A$ and $B + \Delta B$, which will certainly be true if $\Delta A$ and $\Delta B$ are sufficiently small. Perturbation theory under weaker rank assumptions is presented by Wei [20].

Perturbation theory for the LSE problem is given in several earlier references [2, 5, 8]; we give a full analysis in order to investigate the sharpness of the bounds.

We will measure the perturbations normwise by the smallest $\epsilon$ for which

$$\|\Delta A\|_F \leq \epsilon \|\mathbf{A}\|_F, \qquad \|\Delta b\|_2 \leq \epsilon \|\mathbf{b}\|_2, \tag{4.1a}$$

$$\|\Delta B\|_F \leq \epsilon \|\mathbf{B}\|_F, \qquad \|\Delta d\|_2 \leq \epsilon \|\mathbf{d}\|_2, \tag{4.1b}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{b}$ and $\mathbf{d}$ are matrices and vectors of tolerances. We use the Frobenius norm for the matrices since the Frobenius norm is used in the error analysis of Section 3 and also because it leads to more cheaply computable bounds than the 2-norm. A tilde will be used to denote vectors associated with the perturbed problem.

Before beginning the analysis we note that some key matrices that will appear later can be expressed in terms of the generalized QR factorization as follows:

$$B^+ = Q \begin{bmatrix} S^{-1} \\ 0 \end{bmatrix}, \tag{4.2a}$$

$$P = I_p - B^+ B = Q \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix} Q^T, \tag{4.2b}$$

$$(AP)^+ = Q \begin{bmatrix} 0 & 0 \\ 0 & L_{22}^{-1} \end{bmatrix} U^T, \tag{4.2c}$$

$$B_A^+ = (I_n - (AP)^+ A)B^+ = Q \begin{bmatrix} I_p \\ -L_{22}^{-1} L_{21} \end{bmatrix} S^{-1}. \tag{4.2d}$$

Let

$$\|b - Ax\|_2 = \min\{\, \|b - Az\|_2 : Bz = d \,\}, \tag{4.3a}$$

$$\|(b + \Delta b) - (A + \Delta A)\widetilde{x}\|_2 = \min\{\, \|(b + \Delta b) - (A + \Delta A)z\|_2 : (B + \Delta B)z = d + \Delta d \,\} \tag{4.3b}$$

and define the residuals

$$r = b - Ax, \qquad \widetilde{r} = b + \Delta b - (A + \Delta A)\widetilde{x}.$$

We note for later use that a condition number for the LSE problem can be defined by

$$\operatorname{cond}(A, B, b, d) = \lim_{\epsilon \to 0} \sup \left\{ \frac{\|x - \widetilde{x}\|_2}{\|x\|_2} : (4.3) \text{ and } (4.1) \text{ hold} \right\}.$$

One way to obtain perturbation bounds is to invoke perturbation theory for the generalized QR factorization [6]. However, the bounds in [6] are rather complicated and it is more satisfactory to work directly from the easily derived *augmented system*

$$\begin{bmatrix} 0 & 0 & B \\ 0 & I & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ r \\ x \end{bmatrix} = \begin{bmatrix} d \\ b \\ 0 \end{bmatrix}, \tag{4.4}$$

where $\lambda \in \mathbb{R}^p$ is a vector of Lagrange multipliers. For the perturbed problem, the augmented system can be rewritten as

$$\begin{bmatrix} 0 & 0 & B \\ 0 & I & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \widetilde{\lambda} \\ \widetilde{r} \\ \widetilde{x} \end{bmatrix} = \begin{bmatrix} d + \Delta d - \Delta B\widetilde{x} \\ b + \Delta b - \Delta A\widetilde{x} \\ -\Delta B^T\widetilde{\lambda} - \Delta A^T\widetilde{r} \end{bmatrix}. \tag{4.5}$$

Defining $\Delta\lambda$, $\Delta r$ and $\Delta x$ by

$$\widetilde{\lambda} = \lambda + \Delta\lambda, \quad \widetilde{r} = r + \Delta r, \quad \widetilde{x} = x + \Delta x,$$

we subtract (4.4) from (4.5) to obtain

$$\begin{bmatrix} 0 & 0 & B \\ 0 & I & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta r \\ \Delta x \end{bmatrix} = \begin{bmatrix} \Delta d - \Delta B\widetilde{x} \\ \Delta b - \Delta A\widetilde{x} \\ -\Delta B^T\widetilde{\lambda} - \Delta A^T\widetilde{r} \end{bmatrix}. \tag{4.6}$$

We can solve for the desired perturbations with the aid of the following lemma.

LEMMA 4.1. *Under the assumption* (1.2)*, the inverse of the matrix on the left-hand side of* (4.6) *is*

$$\begin{bmatrix} (AB_A^+)^T AB_A^+ & -(AB_A^+)^T & B_A^{+T} \\ -AB_A^+ & I - AP(AP)^+ & (AP)^{+T} \\ B_A^+ & (AP)^+ & -((AP)^T AP)^+ \end{bmatrix},$$

*where*

$$P = I - B^+B, \qquad B_A^+ = (I - (AP)^+A)B^+.$$

PROOF. Eldén [8] shows that the inverse in question is

$$\begin{bmatrix} (AB_A^+)^T AB_A^+ & -(AB_A^+)^T & B_A^{+T} \\ -AB_A^+ & I - AP(AP)^+ & (P(AP)^+)^T \\ B_A^+ & P(AP)^+ & -P((AP)^T AP)^+P^T \end{bmatrix}.$$

Using the generalized QR factorization (2.1) it is straightforward to show that

$$\begin{aligned} P(AP)^+ &= (AP)^+, \\ P((AP)^T AP)^+P^T &= ((AP)^T AP)^+, \end{aligned}$$

which yields the claimed expression for the inverse. □

Using Lemma 4.1 we obtain from (4.6) the following expression for $\Delta x$:

$$\begin{aligned}\Delta x = B_A^+(\Delta d - \Delta B\widetilde{x}) + (AP)^+(\Delta b - \Delta A\widetilde{x})\\ + ((AP)^T AP)^+(\Delta B^T\widetilde{\lambda} + \Delta A^T\widetilde{r}).\end{aligned} \tag{4.7}$$

Since $\Delta\lambda$, $\Delta r$ and $\Delta x$ are all of order $\epsilon$, we can replace $\widetilde{\lambda}$, $\widetilde{r}$ and $\widetilde{x}$ by their unperturbed counterparts to obtain first order expressions. We then need to remove the Lagrange multiplier $\lambda$ from the expressions, since we require expressions involving only $x$ and $r$. From (4.4) we have $B^T\lambda + A^T r = 0$. Since $B$ has full rank, $BB^+ = I$ and so $\lambda = -(AB^+)^T r$. The following lemma provides an alternative expression for $\lambda$.

LEMMA 4.2. *Under the assumptions* (1.2),

$$(AB^+)^T r = (AB_A^+)^T r.$$

PROOF. Using the generalized QR factorization (2.1) and (4.2) we find that

$$AB^+ = U\begin{bmatrix} L_{11} \\ L_{21}\end{bmatrix} S^{-1}, \qquad AB_A^+ = U\begin{bmatrix} L_{11} \\ 0\end{bmatrix} S^{-1}. \tag{4.8}$$

The derivation of the GQR method shows that the LSE residual

$$r = U\begin{bmatrix} c_1 - L_{11}y_1 \\ 0\end{bmatrix}.$$

The result follows. □

We will use the expression $\lambda = -(AB_A^+)^T r$, since

$$\|AB_A^+\|_2 \le \|AB^+\|_2 \tag{4.9}$$

from (4.8), and the norm of the matrix multiplying $r$ is a factor in our final bound. (Inequality (4.9) also follows from the characterization of $B_A^+$ as a weighted pseudo-inverse [8, 9].) With these simplifications we obtain

$$\begin{aligned}\Delta x = B_A^+(\Delta d - \Delta B x) + (AP)^+(\Delta b - \Delta A x)\\ + ((AP)^T AP)^+(-\Delta B^T(AB_A^+)^T + \Delta A^T)r + O(\epsilon^2).\end{aligned} \tag{4.10}$$

Taking 2-norms and using (4.1) we obtain

$$\begin{aligned}\|\Delta x\|_2 \le \epsilon\Big[&\|B_A^+\|_2(\|\mathbf{d}\|_2 + \|\mathbf{B}\|_F\|x\|_2) + \|(AP)^+\|_2(\|\mathbf{b}\|_2 + \|\mathbf{A}\|_F\|x\|_2)\\ &+ \|((AP)^T AP)^+\|_2(\|\mathbf{B}\|_F\|AB_A^+\|_2 + \|\mathbf{A}\|_F)\|r\|_2\Big] + O(\epsilon^2).\end{aligned}$$

Using the equality, for arbitrary $X$, $\|(X^T X)^+\|_2 = \|X^+\|_2^2$, we have

$$\begin{aligned}\frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon\Bigg[&\|B_A^+\|_2\left(\frac{\|\mathbf{d}\|_2}{\|x\|_2} + \|\mathbf{B}\|_F\right) + \|(AP)^+\|_2\left(\frac{\|\mathbf{b}\|_2}{\|x\|_2} + \|\mathbf{A}\|_F\right)\\ &+ \|(AP)^+\|_2^2(\|\mathbf{B}\|_F\|AB_A^+\|_2 + \|\mathbf{A}\|_F)\|r\|_2/\|x\|_2\Bigg] + O(\epsilon^2).\end{aligned}$$

Defining

$$\kappa_B(A) = \|A\|_F \|(AP)^+\|_2, \qquad \kappa_A(B) = \|B\|_F \|B_A^+\|_2,$$

the bound can be rewritten as

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon \Bigg[ \kappa_A(B) \left( \frac{\|\mathbf{d}\|_2}{\|B\|_F \|x\|_2} + \frac{\|\mathbf{B}\|_F}{\|B\|_F} \right) + \kappa_B(A) \left( \frac{\|\mathbf{b}\|_2}{\|A\|_F \|x\|_2} + \frac{\|\mathbf{A}\|_F}{\|A\|_F} \right)$$

$$+ \kappa_B(A)^2 \left( \frac{\|\mathbf{B}\|_F}{\|B\|_F} \frac{\|B\|_F}{\|A\|_F} \|AB_A^+\|_2 + \frac{\|\mathbf{A}\|_F}{\|A\|_F} \right) \frac{\|r\|_2}{\|A\|_F \|x\|_2} \Bigg] + O(\epsilon^2). \tag{4.11}$$

This is essentially the same bound as that obtained by Eldén [8], the differences stemming from different assumptions on the perturbations in (4.1). Unlike Anderson, Bai and Dongarra [2], we will not simplify the bound using the inequality

$$\frac{\|B\|_F}{\|A\|_F} \|AB_A^+\|_2 \le \kappa_A(B),$$

because this is potentially a very weak inequality, as can be seen by noting that (see (4.2) and (4.8))

$$\|AB_A^+\|_2 = \|L_{11} S^{-1}\|_2, \qquad \|B_A^+\|_2 = \|L_{22}^{-1} L_{21} S^{-1}\|_2.$$

As a check, we can recover a perturbation bound for the standard LS problem by setting $B = 0$, $\mathbf{B} = 0$ and $\mathbf{d} = 0$. We obtain

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon \left[ \widetilde{\kappa}(A) \left( \frac{\|\mathbf{b}\|_2}{\|A\|_F \|x\|_2} + \frac{\|\mathbf{A}\|_F}{\|A\|_F} \right) + \widetilde{\kappa}(A)^2 \frac{\|\mathbf{A}\|_F}{\|A\|_F} \frac{\|r\|_2}{\|A\|_F \|x\|_2} \right] + O(\epsilon^2),$$

where $\widetilde{\kappa}(A) = \|A\|_F \|A^+\|_2$. If we take $\mathbf{A} = A$ and $\mathbf{b} = b$ and use $\|b\|_2 \le \|r\|_2 + \|A\|_F \|x\|_2$ the bound can be rewritten as

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon \widetilde{\kappa}(A) \left( 2 + (\widetilde{\kappa}(A) + 1) \frac{\|r\|_2}{\|A\|_F \|x\|_2} \right) + O(\epsilon^2),$$

which is essentially the standard bound of Wedin [19, Thm. 5.1], [15, Thm. 19.1].

The bound (4.11) shows that if the residual $r$ is small or zero, the sensitivity is governed by $\kappa_A(B)$ and $\kappa_B(A)$, otherwise by $\kappa_A(B)$ and $\kappa_B(A)^2 \|B\|_F \|AB_A^+\|_2 / \|A\|_F$. A *sufficient* condition for the LSE problem to be well conditioned is that $B$ and $AP$ are both well conditioned, as can be seen by using the relations (4.2).

The bound (4.11) does not yield a condition number for the LSE problem, since it is not attainable. To obtain a sharp bound we must combine the two $\Delta A$ terms before taking norms, and likewise for $\Delta B$. This can be achieved with the aid of the vec operator, which stacks the columns of a matrix into one long vector, together with the Kronecker product $A \otimes B = (a_{ij} B)$ [13]. Applying the vec operator to (4.10), and using the property that $\operatorname{vec}(AXB) = (B^T \otimes A) \operatorname{vec}(X)$, we obtain

$$\begin{aligned}\Delta x = {} & B_A^+\Delta d - (x^T \otimes B_A^+)\,\mathrm{vec}(\Delta B) + (AP)^+\Delta b - (x^T \otimes (AP)^+)\,\mathrm{vec}(\Delta A) \\ & - \left(r^T A B_A^+ \otimes ((AP)^T AP)^+\right)\mathrm{vec}(\Delta B^T) \\ & + \left(r^T \otimes ((AP)^T AP)^+\right)\mathrm{vec}(\Delta A^T) + O(\epsilon^2).\end{aligned}$$

Employing the relation $\mathrm{vec}(\Delta A^T) = \Pi\,\mathrm{vec}(\Delta A)$ where $\Pi$ is the vec-permutation matrix [13], we obtain

$$\begin{aligned}\Delta x = {} & B_A^+\Delta d + (AP)^+\Delta b - \left(x^T \otimes B_A^+ + [r^T A B_A^+ \otimes ((AP)^T AP)^+]\Pi\right)\mathrm{vec}(\Delta B) \\ & + \left(-x^T \otimes (AP)^+ + [r^T \otimes ((AP)^T AP)^+]\Pi\right)\mathrm{vec}(\Delta A) + O(\epsilon^2).\end{aligned} \tag{4.12}$$

By taking 2-norms and using $\|\,\mathrm{vec}(\Delta A)\|_2 = \|\Delta A\|_F$ and (4.1) we deduce that

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \Psi\epsilon + O(\epsilon^2), \tag{4.13}$$

where

$$\begin{aligned}\Psi = {} & \big(\|B_A^+\|_2\|\mathbf{d}\|_2 + \|(AP)^+\|_2\|\mathbf{b}\|_2 \\ & + \|x^T \otimes B_A^+ + [r^T A B_A^+ \otimes ((AP)^T AP)^+]\Pi\|_2\|\mathbf{B}\|_F \\ & + \|-x^T \otimes (AP)^+ + [r^T \otimes ((AP)^T AP)^+]\Pi\|_2\|\mathbf{A}\|_F\big)/\|x\|_2\end{aligned}$$

and

$$\mathrm{cond}(A, B, b, d) \le \Psi \le 4\,\mathrm{cond}(A, B, b, d).$$

The bound (4.13) is much more difficult to interpret than (4.11) because of the complicated expression for $\Psi$. In the next section we compare the bounds (4.13) and (4.11) on some numerical examples.

A bound can be derived for $\Delta r$ in just the same way as for $\Delta x$:

$$\frac{\|\Delta r\|_2}{\|r\|_2} \le \epsilon\Bigg[\frac{\|B\|_F}{\|A\|_F}\|AB_A^+\|_2\left(\frac{\|\mathbf{d}\|_2}{\|r\|_2}\frac{\|A\|_F}{\|B\|_F} + \frac{\|\mathbf{B}\|_F}{\|B\|_F}\frac{\|A\|_F\|x\|_2}{\|r\|_2}\right) \tag{4.14}$$

$$+ \min\{1, m+1-n\}\left(\frac{\|\mathbf{b}\|_2}{\|r\|_2} + \frac{\|\mathbf{A}\|_F}{\|A\|_F}\frac{\|A\|_F\|x\|_2}{\|r\|_2}\right) \tag{4.15}$$

$$+ \kappa_B(A)\left(\frac{\|\mathbf{B}\|_F}{\|B\|_F}\frac{\|B\|_F}{\|A\|_F}\|AB_A^+\|_2 + \frac{\|\mathbf{A}\|_F}{\|A\|_F}\right)\Bigg] + O(\epsilon^2). \tag{4.16}$$

In contrast to the bound for $\Delta x$, that for $\Delta r$ shows no direct dependence on $\kappa_A(B)$, and $\kappa_B(A)$ appears only to the first power.

Finally, we note that componentwise perturbation bounds for $\Delta x$ (and, similarly, for $\Delta r$) can be obtained by replacing the norms in (4.1) by absolute values and taking absolute values in (4.10) and (4.12). These bounds are of limited practical use since we do not have componentwise backward error bounds for any of the existing numerical methods for solving the LSE problem.

## 5 Practical error bound and numerical experiments.

Some of the LAPACK expert driver routines return error bounds, while for certain other routines the LAPACK Users' Guide [1] explains how the user can compute approximate error bounds. No error bounds are implemented or described for the driver routine `xgglse.f`, which solves the LSE problem by the GQR method. We can derive a bound by combining Theorem 3.1 and the perturbation bound (4.11). We find that the computed solution $\widehat{x}$ satisfies

$$\frac{\|x-\widehat{x}\|_2}{\|x\|_2} \le \widetilde{\gamma}_{np}\kappa_A(B) + \widetilde{\gamma}_{mn}\kappa_B(A)\left(\frac{\|b\|_2}{\|A\|_F\|x\|_2}+1\right) + \kappa_B(A)^2\left(\widetilde{\gamma}_{np}\frac{\|B\|_F}{\|A\|_F}\|AB_A^+\|_2 + \widetilde{\gamma}_{mn}\right)\frac{\|r\|_2}{\|A\|_F\|x\|_2} + O(u^2).$$

This bound is unsatisfactory for two reasons. First, we do not know the constants implicit in the $\widetilde{\gamma}$ terms. Following the LAPACK Users' Guide we adopt the radical step of replacing each $\widetilde{\gamma}$ term by $u$, which is partly justified by the fact that the dimension-dependent terms are obtained by many applications of the triangular and submultiplicative inequalities and so are pessimistic (see [1, pp. 70–73] for a more detailed discussion). The second problem is that the bound requires computation of the quantities $\|B_A^+\|_2$, $\|(AP)^+\|_2$ and $\|AB_A^+\|_2$. Using the generalized QR factorization (2.1) these quantities may be expressed as (see (4.2) and (4.8))

$$\|B_A^+\|_2 = \left\| \begin{bmatrix} I_p \\ -L_{22}^{-1}L_{21} \end{bmatrix} S^{-1} \right\|_2,$$
$$\|(AP)^+\|_2 = \|L_{22}^{-1}\|_2, \qquad \|AB_A^+\|_2 = \|L_{11}S^{-1}\|_2.$$

To avoid the possibly expensive formation of the matrices involving $L_{22}^{-1}$ and $S^{-1}$ the norms of these matrices are *estimated* using the LAPACK norm estimator [1], [14], which estimates $\|B\|_1$ given only the ability to form matrix-vector products $Bx$ and $B^Ty$. We can compute the required products by solving triangular systems and we accept the 1-norm estimate as an approximation to the 2-norm. A further computational saving is to compute $\|A\|_F$ and $\|B\|_F$ as

$$\|A\|_F = \left(\|L_{11}\|_F^2 + \|L_{21}\|_F^2 + \|L_{22}\|_F^2\right)^{1/2}, \qquad \|B\|_F = \|S\|_F.$$

To summarize, the approximate bound that we estimate is

$$\frac{\|x-\widehat{x}\|_2}{\|x\|_2} \lesssim u\left[\kappa_A(B) + \kappa_B(A)\left(\frac{\|b\|_2}{\|A\|_F\|x\|_2}+1\right) + \kappa_B(A)^2\left(\frac{\|B\|_F}{\|A\|_F}\|AB_A^+\|_2+1\right)\frac{\|r\|_2}{\|A\|_F\|x\|_2}\right]$$
$$= \text{lse\_err}. \tag{5.1}$$

We have carried out numerical experiments to test the sharpness and reliability of the bound (5.1) and to compare it with the smaller bound $\Psi u$ (with

Table 5.1: Results for the GQR Method.
$\kappa_A(B) = 2.16\text{e}1$, $\kappa_B(A) = 1.99\text{e}1$, $\|AB_A^+\|_2 = 6.17$.

| err | lse_err | $\psi u$ | res | $\|x\|_2$ |
|---|---|---|---|---|
| 5.85e-7 | 2.75e-6 | 2.75e-6 | 5.05e-8 | 4.45e0 |
| 8.14e-7 | 1.37e-5 | 7.45e-6 | 2.26e-1 | 4.45e0 |

Table 5.2: Results for the GQR Method.
$\kappa_A(B) = 9.33\text{e}4$, $\kappa_B(A) = 4.16\text{e}1$, $\|AB_A^+\|_2 = 3.83$.

| err | lse_err | $\psi u$ | res | $\|x\|_2$ |
|---|---|---|---|---|
| 5.13e-4 | 5.56e-3 | 5.56e-3 | 2.12e-8 | 7.99e4 |
| 3.49e-4 | 5.60e-3 | 5.59e-3 | 2.70e-1 | 4.45e0 |

$\mathbf{A} = A$, $\mathbf{b} = b$, $\mathbf{B} = B$, $\mathbf{d} = 0$) obtained in an analogous way from (4.13). Our experiments were performed in MATLAB. We simulated single precision arithmetic by rounding the result of every arithmetic operation to 24 bits; thus the unit roundoff $u = 2^{-24} \approx 6 \times 10^{-8}$. To compute the forward error $\|x - \widehat{x}\|_2/\|x\|_2$ we used for $x$ the solution computed in double precision arithmetic. The GQR Method and Methods 2 and 3 were found to give very similar errors (usually agreeing to at least one significant figure), so we report only the errors for the GQR Method.

We generated random LSE problems with various prescribed conditioning properties by choosing the generalized QR factors and working backwards.

Results are reported in Tables 5.1–5.4, with the notation

$$\text{err} = \frac{\|x - \widehat{x}\|_2}{\|x\|_2}, \qquad \text{res} = \frac{\|r\|_2}{\|A\|_F \|x\|_2}$$

for the forward error and relative residual, respectively. In these experiments we computed the quantities $\kappa_A(B)$, $\kappa_B(A)$ and $\|AB_A^+\|_2$, required to evaluate lse_err, exactly. We took $m = 25$, $n = 15$ and $p = 5$.

Each table reports two examples with the same conditioning parameters, one with a small relative residual and the other with a large relative residual. In Tables 5.1 and 5.2 the size of the residual has little effect on the error. In Tables 5.3 and 5.4 the error is much larger when the residual is large than when it is small, reflecting the effect of a large term $\kappa_B(A)^2\|AB_A^+\|_F$ premultiplying the relative residual in the bound (5.1).

We see from the tables that the bound lse_err exceeds the forward error in every case, by a factor varying from approximately 10 to 500. The quantity $\Psi u$ from the sharp bound (4.13) is no larger than lse_err, but is smaller by at most a factor 10. We conclude that the overestimation of the error produced by lse_err (which can, of course, be much more severe than in these experiments, since the actual error can be made zero or very small by suitable choice of the problem) is largely inherent in the use of a worst-case bound, and cannot be improved

Table 5.3: Results for the GQR Method.
$\kappa_A(B) = 1.01\text{e}5$, $\kappa_B(A) = 9.72\text{e}3$, $\|AB_A^+\|_2 = 1.88\text{e}2$.

| err | lse_err | $\psi u$ | res | $\|x\|_2$ |
|---|---|---|---|---|
| 1.28e-4 | 6.73e-3 | 6.72e-3 | 6.20e-8 | 4.45e0 |
| 5.17e-2 | 2.58e1 | 3.71e0 | 2.14e-1 | 4.76e0 |

Table 5.4: Results for the GQR Method.
$\kappa_A(B) = 9.43\text{e}0$, $\kappa_B(A) = 5.76\text{e}3$, $\|AB_A^+\|_2 = 1.24\text{e}1$.

| err | lse_err | $\psi u$ | res | $\|x\|_2$ |
|---|---|---|---|---|
| 7.39e-6 | 3.45e-4 | 3.44e-4 | 1.62e-8 | 1.29e3 |
| 2.51e-2 | 3.02e0 | 1.42e0 | 3.96e-1 | 4.42e0 |

significantly by changes to the derivation of the bound. Because of the extra expense of computing or estimating $\Psi$, there is no reason to prefer $\Psi u$ to lse_err as a practical error bound.

The quantity lse_err can be recommended as an approximate forward error bound for the null space method for solving the LSE problem and, in particular, for use with the driver routine `xgglse.f` in LAPACK. This bound has been implemented for LAPACK by Bai and Fahey [3], who also develop and implement a practical error bound for LAPACK's generalized QR factorization-based method for solving the generalized linear model problem $\min\{y^T y : d = Ax + By\}$.

**Acknowledgements.**

We thank Zhaojun Bai for helpful comments on this work. The work of Cox was supported by an Engineering and Physical Sciences Research Council Research Studentship.

## REFERENCES

1. E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen, *LAPACK Users' Guide, Release 2.0*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.
2. E. Anderson, Z. Bai, and J. Dongarra, *Generalized QR factorization and its applications*, Linear Algebra Appl., pp. 162–164 (1992), pp. 243–271.
3. Z. Bai and M. Fahey, *Computation of error bounds in linear least squares problems with equality constraints and generalized linear model problems*, Manuscript, June 1997.
4. J. L. Barlow, N. K. Nichols, and R. J. Plemmons, *Iterative methods for equality-constrained least squares problems*, SIAM J. Sci. Stat. Comput., 9:5 (1988), pp. 892–906.

5. J. L. Barlow. *Error analysis and implementation aspects of deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 25:6 (1988), pp. 1340–1358.
6. A. Barrlund, *Perturbation bounds for the generalized QR factorization*, Linear Algebra Appl., 207 (1994), pp. 251–271.
7. Å. Björck, *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.
8. L. Eldén, *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17:3 (1980), pp. 338–350.
9. L. Eldén, *A weighted pseudoinverse, generalized singular values, and constrained least squares problems*, BIT, 22 (1982), pp. 487–502.
10. E. Galligani and A. Laratta, *Error analysis of null space algorithm for linear equality constrained least squares problems*, Computing, 52 (1994), pp. 161–176.
11. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, USA, 1996.
12. S. J. Hammarling, *The numerical solution of the general Gauss–Markov linear model*, in Mathematics in Signal Processing, T. S. Durrani, J. B. Abbiss, and J. E. Hudson, eds., Oxford University Press, 1987, pp. 451–456.
13. H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear and Multilinear Algebra, 9 (1981), pp. 271–288.
14. N. J. Higham, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm 674)*, ACM Trans. Math. Software, 14:4 (1988), pp. 381–396.
15. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.
16. C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995. Revised republication of work first published in 1974 by Prentice-Hall.
17. C. C. Paige, *Some aspects of generalized QR factorizations*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, 1990, pp. 73–91.
18. C. F. Van Loan, *On the method of weighting for equality-constrained least-squares problems*, SIAM J. Numer. Anal., 22:5 (1985), pp. 851–864.
19. P.-Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT, 13 (1973), pp. 217–232.
20. M. Wei, *Perturbation theory for the rank-deficient equality constrained least squares problems*, SIAM J. Numer. Anal., 29:5 (1992), pp. 1462–1481.