BIT Numerical Mathematics **43:** 505–517, 2003.

THE EQUALITY CONSTRAINED INDEFINITE LEAST SQUARES PROBLEM: THEORY AND ALGORITHMS*†

ADAM BOJANCZYK[1,‡], NICHOLAS J. HIGHAM[1,$] and HARIKRISHNA PATEL[1,¶]

[1] *School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853-3801, USA. email: adamb@ee.cornell.edu, http://www.ee.cornell.edu/~adamb.*

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

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

Abstract.

We present theory and algorithms for the equality constrained indefinite least squares problem, which requires minimization of an indefinite quadratic form subject to a linear equality constraint. A generalized hyperbolic QR factorization is introduced and used in the derivation of perturbation bounds and to construct a numerical method. An alternative method is obtained by employing a generalized QR factorization in combination with a Cholesky factorization. Rounding error analysis is given to show that both methods have satisfactory numerical stability properties and numerical experiments are given for illustration. This work builds on recent work on the unconstrained indefinite least squares problem by Chandrasekaran, Gu, and Sayed and by the present authors.

AMS subject classification (2000): 65F20, 65G05.

Key words: equality constrained indefinite least squares problem, J-orthogonal matrix, hyperbolic rotation, hyperbolic QR factorization, generalized hyperbolic QR factorization, rounding error analysis, forward stability, perturbation theory, Cholesky factorization.

* Received October 2002. Revised January 2003. Communicated by Åke Björck.

† Numerical Analysis Report 413, Manchester Centre for Computational Mathematics, Oct. 2002; revised Jan. 2003. This work was begun with the support of Engineering and Physical Sciences Research Council Visiting Fellowship GR/R22414. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defence Advanced Research Project Agency (DARPA), Rome Laboratory or the U.S. Government.

‡ Effort sponsored by the Defence Advanced Research Project Agency (DARPA) and Rome Laboratory, Air Force Material Command, USAF, under agreement number F30602-97-1-0292. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon

$ This work was supported by Engineering and Physical Sciences Research Council grant GR/R22612.

¶ This work was supported by an Engineering and Physical Sciences Research Council Ph.D. Studentship.

1 Introduction.

The indefinite least squares problem has the form

$$\text{ILS}: \qquad \min_x (b-Ax)^T J(b-Ax), \tag{1.1}$$

where $A \in \mathbb{R}^{(p+q)\times n}$, $b \in \mathbb{R}^{p+q}$ and the signature matrix

$$J = \begin{bmatrix} -I_q & 0 \\ 0 & I_p \end{bmatrix}. \tag{1.2}$$

This problem was introduced by Chandrasekaran, Gu, and Sayed [2] and further studied by Bojanczyk, Higham, and Patel [1]. We are concerned here with the extension of the ILS problem to include equality constraints:

$$\text{ILSE}: \qquad \min_x (b-Ax)^T J(b-Ax) \quad \text{subject to } Bx = d, \tag{1.3}$$

where $B \in \mathbb{R}^{s\times n}$ and $d \in \mathbb{R}^s$. A major difference between indefinite and standard least squares problems is that indefinite problems do not always have a solution.

We will assume that

$$\operatorname{rank}(B) = s, \qquad x^T(A^TJA)x > 0 \text{ for all nonzero } x \in \operatorname{null}(B). \tag{1.4}$$

The first condition ensures that there is a solution to the constraint equations. The second condition, which says that A^TJA is positive definite on the null space of B, then ensures that there is a unique solution to the ILSE problem. The uniqueness can be shown by manipulating the normal equations

$$A^TJ(b-Ax) = B^T\mu, \qquad Bx = d, \tag{1.5}$$

where μ is a vector of Lagrange multipliers, and it is also established by our analysis in terms of the generalized hyperbolic QR factorization (see Section 4).

We note for later use that A^TJA has rank at most p, and so since $\operatorname{null}(B)$ has dimension $n-s$, the second condition in (1.4) implies that

$$p \geq n - s. \tag{1.6}$$

An important role in the theory and numerical solution of indefinite least squares problems is played by J-orthogonal transformations. The matrix $H \in \mathbb{R}^{(p+q)\times(p+q)}$ is J-orthogonal if

$$H^TJH = J. \tag{1.7}$$

Central to this work is a new factorization that we call the generalized hyperbolic QR (GHQR) factorization. The following result defines the factorization and establishes its existence.

THEOREM 1.1 (GENERALIZED HYPERBOLIC QR FACTORIZATION). *Let $A \in \mathbb{R}^{(p+q)\times n}$ and $B \in \mathbb{R}^{s\times n}$, let J be given by* (1.2)*, and assume that* (1.4) *holds.*

Then there exist an orthogonal $Q \in \mathbb{R}^{n\times n}$ and a J-orthogonal $H \in \mathbb{R}^{(p+q)\times(p+q)}$ such that

$$HAQ = \begin{array}{c} \\ p+q-(n-s) \\ n-s \end{array} \begin{array}{c} \begin{array}{cc} s & n-s \end{array} \\ \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \end{array}, \qquad BQ = \begin{array}{c} \\ s \end{array} \begin{array}{c} \begin{array}{cc} s & n-s \end{array} \\ \begin{pmatrix} K & 0 \end{pmatrix} \end{array}, \tag{1.8}$$

where L_{22} and K are lower triangular and nonsingular.

PROOF. Let E_m denote the $m \times m$ exchange matrix, that is, the identity matrix I_m with its columns in reverse order: $E_m = I_m(:, m\colon -1 : 1)$. Let

$$Q_B^T B^T = \begin{bmatrix} K^T \\ 0 \end{bmatrix}$$

be a QR factorization of B^T. The lower triangular matrix K is nonsingular since B has full rank. Write

$$E_{p+q} A Q_B = \begin{array}{c} \begin{array}{cc} s & n-s \end{array} \\ \begin{pmatrix} \widetilde{A}_1 & \widetilde{A}_2 \end{pmatrix} \end{array}$$

and set

$$J_E = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}.$$

Note that $E_{p+q} = E_{p+q}^T$ and $E_{p+q} J_E E_{p+q} = J$. Now, from the second assumption in (1.4) it follows that $\widetilde{A}_2^T J_E \widetilde{A}_2$ is positive definite. Hence, there exists a hyperbolic QR factorization of $\widetilde{A}_2$ [1]

$$W \widetilde{A}_2 = \begin{bmatrix} R \\ 0 \end{bmatrix},$$

where R is a nonsingular upper triangular matrix and W is J_E-orthogonal. Moreover we have

$$E_{p+q} W \widetilde{A}_2 E_{n-s} = \begin{bmatrix} 0 \\ L \end{bmatrix},$$

where $L = E_{n-s} R E_{n-s}$ is lower triangular. Define

$$Q = Q_B \begin{bmatrix} I_s & 0 \\ 0 & E_{n-s} \end{bmatrix}, \quad H = E_{p+q} W E_{p+q}, \quad L_{22} = L.$$

Then (1.8) holds. □

As we will show below, the GHQR factorization plays a similar role for the ILSE problem as the generalized QR factorization plays for the standard least squares problem with equality constraints.

The outline of this paper is as follows. In Section 2 we give perturbation theory for the ILSE problem, obtaining a bound that is cheaply estimable but possibly non-sharp, and another bound based on Kronecker products that is

sharp but more expensive to compute or estimate. In Section 3 we show how the ILSE problem can be solved using a generalized QR factorization and Cholesky factorization, in what is an extension of the method of Chandrasekaran, Gu, and Sayed [2] for the ILS problem. We prove that this method is mixed forward–backward stable. A method based on the GHQR factorization is presented in Section 4 and the method is proved to be forward stable under a reasonable assumption. Numerical experiments are described in Section 5 that corroborate the error analysis.

2 Perturbation theory.

We begin by examining the sensitivity of the ILSE problem to perturbations in the data. We will consider perturbations ΔA, Δb, ΔB and Δd to the data A, b, B and d measured normwise by the smallest ϵ for which

$$\|\Delta A\|_F \le \epsilon\|A\|_F, \quad \|\Delta b\|_2 \le \epsilon\|b\|_2, \tag{2.1a}$$

$$\|\Delta B\|_F \le \epsilon\|B\|_F, \quad \|\Delta d\|_2 \le \epsilon\|d\|_2. \tag{2.1b}$$

We assume that the conditions (1.4) continue to hold for the perturbed problem.

The solution to the ILSE problem (1.3) satisfies the augmented system

$$\begin{bmatrix} 0 & 0 & B \\ 0 & J & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ s \\ x \end{bmatrix} = \begin{bmatrix} d \\ b \\ 0 \end{bmatrix}, \tag{2.2}$$

where $s = J(b - Ax) = Jr$, which is a rewritten version of the normal equations (1.5) with $\lambda = -\mu$. The augmented system for the perturbed problem is

$$\begin{bmatrix} 0 & 0 & B + \Delta B \\ 0 & J & A + \Delta A \\ B^T + \Delta B^T & A^T + \Delta A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda + \Delta\lambda \\ s + \Delta s \\ x + \Delta x \end{bmatrix} = \begin{bmatrix} d + \Delta d \\ b + \Delta b \\ 0 \end{bmatrix},$$

which leads to the relationship for the perturbations

$$\begin{bmatrix} 0 & 0 & B \\ 0 & J & A \\ B^T & A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta s \\ \Delta x \end{bmatrix} = \begin{bmatrix} \Delta d - \Delta B x \\ \Delta b - \Delta A x \\ -\Delta B^T\lambda - \Delta A^T s \end{bmatrix} + \mathrm{O}(\epsilon^2). \tag{2.3}$$

It is possible to solve this system for Δx and to continue working with expressions involving the original data. However, these expressions become rather complicated and are not very amenable to interpretation or computation. We therefore use the GHQR factorization to simplify the analysis.

Let A and B have the GHQR factorization

$$HAQ = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}, \qquad BQ = \begin{bmatrix} K & 0 \end{bmatrix},$$

where H is J-orthogonal and Q is orthogonal.

Premultiplying (2.3) by $\operatorname{diag}(I, H, Q^T)$, we obtain

$$\begin{bmatrix} 0 & 0 & BQ \\ 0 & HJH^T & HAQ \\ Q^TB^T & Q^TA^TH^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ JHJ\Delta s \\ Q^T\Delta x \end{bmatrix} = \begin{bmatrix} \Delta d - \Delta Bx \\ H(\Delta b - \Delta Ax) \\ -Q^T(\Delta B^T\lambda + \Delta A^Ts) \end{bmatrix} + \mathrm{O}(\epsilon^2). \tag{2.4}$$

Using the GHQR factorization the matrix on the left can be rewritten as

$$Z = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & K & 0 \\ \hline 0 & \widetilde{J} & 0 & L_{11} & 0 \\ 0 & 0 & I_{n-s} & L_{21} & L_{22} \\ \hline K^T & L_{11}^T & L_{21}^T & 0 & 0 \\ 0 & 0 & L_{22}^T & 0 & 0 \end{array}\right],$$

where $\widetilde{J} = \operatorname{diag}(-I_q, I_{p-(n-s)})$. The inverse of Z is

$$Z^{-1} = \left[\begin{array}{c|cc|cc} K^{-T}L_{11}^T\widetilde{J}L_{11}K^{-1} & -K^{-T}L_{11}^T\widetilde{J} & 0 & K^{-T} & -K^{-T}L_{21}^TL_{22}^{-T} \\ \hline -\widetilde{J}L_{11}K^{-1} & \widetilde{J} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & L_{22}^{-T} \\ \hline K^{-1} & 0 & 0 & 0 & 0 \\ -L_{22}^{-1}L_{21}K^{-1} & 0 & L_{22}^{-1} & 0 & -L_{22}^{-1}L_{22}^{-T} \end{array}\right]. \tag{2.5}$$

From (2.4) we therefore obtain

$$\begin{aligned} Q^T\Delta x \;=\; & \begin{bmatrix} I \\ -L_{22}^{-1}L_{21} \end{bmatrix} K^{-1}(\Delta d - \Delta Bx) + \begin{bmatrix} 0 & 0 \\ 0 & L_{22}^{-1} \end{bmatrix} H(\Delta b - \Delta Ax) \\ & + \begin{bmatrix} 0 & 0 \\ 0 & L_{22}^{-1}L_{22}^{-T} \end{bmatrix} Q^T(\Delta B^T\lambda + \Delta A^Ts) + \mathrm{O}(\epsilon^2). \end{aligned} \tag{2.6}$$

The third equation in (2.2) is $B^T\lambda = -A^Ts$. Using the GHQR factorization this equation can be written

$$\begin{bmatrix} K^T \\ 0 \end{bmatrix} \lambda = -\begin{bmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix} JHJs,$$

and by examining both block components it can be seen that

$$\lambda = -K^{-T}\begin{bmatrix} L_{11}^T & 0 \end{bmatrix} JHJs.$$

Hence the penultimate term in (2.6) is

$$\begin{bmatrix} 0 & 0 \\ 0 & L_{22}^{-1}L_{22}^{-T} \end{bmatrix} Q^T\big(-\Delta B^TK^{-T}\begin{bmatrix} L_{11}^T & 0 \end{bmatrix} JHJ + \Delta A^T\big)s. \tag{2.7}$$

Taking norms in (2.6) we therefore obtain the perturbation bound

$$\begin{aligned}\frac{\|\Delta x\|_2}{\|x\|_2} \;\le\; \epsilon\Bigg[&\|B\|_F \left\| \begin{bmatrix} I \\ -L_{22}^{-1}L_{21} \end{bmatrix} K^{-1} \right\|_2 \left(\frac{\|d\|_2}{\|B\|_F\|x\|_2} + 1 \right) \\ &+ \|L_{22}^{-1}H(h+1\colon p+q,:)\|_2 \|A\|_F \left(\frac{\|b\|_2}{\|A\|_F\|x\|_2} + 1 \right) \\ &+ \|L_{22}^{-1}\|_2^2 \|A\|_F \big(\|B\|_F \|K^{-T} \begin{bmatrix} L_{11}^T & 0 \end{bmatrix} JH\|_2 + \|A\|_F \big) \frac{\|r\|_2}{\|A\|_F\|x\|_2} \Bigg] \\ &+ \mathrm{O}(\epsilon^2),\end{aligned} \tag{2.8}$$

where $h = p + q - (n - s)$. An interesting feature of this bound is that the first $p + q - (n - s)$ rows of H contribute to the third term of the bound and the last $n - s$ rows contribute to the second term. This bound includes as special cases bounds for standard LS problem, the equality constrained LS problem [3], [4], and the unconstrained ILS problem [1].

An advantage of expressing (2.8) in terms of the generalized hyperbolic QR factorization is that only triangular matrices are inverted in the formula and hence the bound can be cheaply estimated using standard condition estimation techniques [5, Chap. 15].

As for the results in [1] and [3], it is unclear how close (2.8) is to being attainable for all sets of A, b, B and d. We obtain a sharp perturbation bound by making use of the vec operator and the Kronecker product [7, Chap. 4], as in the latter references. First, we rewrite (2.6), using (2.7), as

$$Q^T\Delta x = G_1\Delta d - G_1\Delta B x + G_2\Delta b - G_2\Delta A x + (G_3\Delta B^T G_4 - G_3\Delta A^T)s + \mathrm{O}(\epsilon^2),$$

where the G_i notation is used to simplify the ensuing expressions. Applying the vec operator, using the relation $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$, and recalling that $s = Jr$, we obtain

$$\begin{aligned}Q^T\Delta x \;=\; & G_1\Delta d - (x^T \otimes G_1)\mathrm{vec}(\Delta B) + G_2\Delta b - (x^T \otimes G_2)\mathrm{vec}(\Delta A) \\ & + (r^T J G_4^T \otimes G_3)\mathrm{vec}(\Delta B^T) - (r^T J \otimes G_3)\mathrm{vec}(\Delta A^T) + \mathrm{O}(\epsilon^2).\end{aligned}$$

Using the relations $\mathrm{vec}(\Delta A^T) = \Pi_1\mathrm{vec}(\Delta A)$ and $\mathrm{vec}(\Delta B^T) = \Pi_2\mathrm{vec}(\Delta B)$, where Π_1 and Π_2 are vec-permutation matrices [7, Chap. 4], gives

$$\begin{aligned}Q^T\Delta x \;=\; & G_1\Delta d + G_2\Delta b - [(x^T \otimes G_1) - (r^T J G_4^T \otimes G_3)\Pi_2]\mathrm{vec}(\Delta B) \\ & - [(x^T \otimes G_2) + (r^T J \otimes G_3)\Pi_1]\mathrm{vec}(\Delta A) + \mathrm{O}(\epsilon^2),\end{aligned}$$

which is expressed in terms of the four independent perturbations Δd, Δb, ΔA and ΔB. Now we take 2-norms. Using (2.1) and the fact that $\|\mathrm{vec}(\Delta A)\|_2 = \|\Delta A\|_F$, we deduce that

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \psi\epsilon + \mathrm{O}(\epsilon^2), \tag{2.9}$$

where

$$\psi = \|x\|_2^{-1}\Big(\|G_1\|_2\|d\|_2 + \|G_2\|_2\|b\|_2 + \|(x^T \otimes G_1) - (r^T J G_4^T \otimes G_3)\Pi_2\|_2\|B\|_F + \|(x^T \otimes G_2) + (r^T J \otimes G_3)\Pi_1\|_2\|A\|_F\Big),$$

and this bound is attainable to within a factor 4. This bound will be used in the experiments of Section 5 to test the sharpness of (2.8).

3 The GQR-Cholesky method.

Our first method for solving the ILSE problem does not involve any hyperbolic transformations. It is an extension of the method of Chandrasekaran, Gu, and Sayed [2] to handle equality constraints.

Consider a generalized QR factorization of A and B (see, e.g., [5, Thm. 20.9])

$$U^T A Q = \begin{matrix} \\ p+q-(n-s) \\ n-s \end{matrix} \begin{matrix} \begin{matrix} s & n-s \end{matrix} \\ \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \end{matrix}, \qquad BQ = \begin{matrix} \\ s \end{matrix} \begin{matrix} \begin{matrix} s & n-s \end{matrix} \\ \begin{pmatrix} K & 0 \end{pmatrix} \end{matrix}, \tag{3.1}$$

where U and Q are orthogonal and L_{22} and K are lower triangular and, in view of (1.4), nonsingular. (Note that this factorization differs from (1.8) only in the U factor, which is replaced by a J-orthogonal matrix in (1.8).) Let

$$\begin{matrix} s \\ n-s \end{matrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y := Q^T x, \qquad C = \begin{matrix} \begin{matrix} s & n-s \end{matrix} \\ \begin{pmatrix} C_1 & C_2 \end{pmatrix} \end{matrix} := AQ. \tag{3.2}$$

Then the constraint $Bx = d$ is equivalent to $Ky_1 = d$, which determines y_1. Thus

$$\begin{aligned} b - Ax &= b - AQy \\ &= b - [\,C_1 \quad C_2\,]\,y \\ &= (b - C_1 y_1) - C_2 y_2 \\ &=: f - C_2 y_2, \end{aligned}$$

and so minimizing $(b - Ax)^T J (b - Ax)$ subject to $Bx = d$ is equivalent to the unconstrained ILS problem

$$\min_{y_2} (f - C_2 y_2)^T J (f - C_2 y_2). \tag{3.3}$$

It is easy to see that (1.4) implies that $C_2^T J C_2$ is positive definite. Noting from (3.2) and (3.1) that

$$C_2 = U \begin{bmatrix} 0 \\ L_{22} \end{bmatrix} =: \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} 0 \\ L_{22} \end{bmatrix} = \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix} L_{22},$$

the normal equations $C_2^T J C_2 y_2 = C_2^T J f$ for (3.3) can be rewritten as

$$(U_{22}^T U_{22} - U_{12}^T U_{12}) L_{22} y_2 = \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix}^T J f.$$

The matrix in parentheses is positive definite and so this system can be solved via a Cholesky factorization. Effectively, we are solving (3.3) by the method of [2]. The entire procedure is summarized as follows.

ALGORITHM 3.1 (GQR-CHOLESKY METHOD). *This algorithm solves the ILSE problem* (1.3) *with the aid of a generalized QR factorization of A and B.*

1. Compute the generalized QR factorization (3.1).
2. Solve the lower triangular system $Ky_1 = d$.
3. $f = b - AQ(:, 1\colon s)y_1$.
4. Form the symmetric matrix $W = U_{22}^T U_{22} - U_{12}^T U_{12}$.
5. Compute the Cholesky factorization $W = R^T R$.
6. Solve $R^T R L_{22} y_2 = [\, U_{12}^T \quad U_{22}^T \,]\, Jf$ by one back and two forward substitutions.
7. $x = Qy$.

Note that the GQR-Cholesky method does not require Q to be formed explicitly, and only the last $n - s$ columns of U need to be formed explicitly. The operation count of the method is a complicated function of p, q, n and s, and when $p + q \gg n \gg s$ it is approximately $7(p + q)n^2$ flops, where a flop denotes one of the four elementary operations.

3.1 Error analysis.

The GQR-Cholesky method has very satisfactory numerical stability properties. It is mixed forward–backward stable, in the sense that the computed solution is very close to the solution of a slightly perturbed problem. The next result makes this statement precise.

In our error analysis we will use the standard model of floating point arithmetic [5, Sec. 2.2]. We denote by u the unit roundoff and by $c_i(u)$ a term of the form $g(p, q, n, s)u + \mathrm{O}(u^2)$, where g is a polynomial; thus we are not concerned with the precise values of constants.

THEOREM 3.1. *Suppose the GQR-Cholesky method is implemented with the generalized QR factorization computed using Householder transformations. The computed solution $\widehat{x} = \overline{x} + \Delta\overline{x}$, where $\overline{x}$ solves a perturbed ILSE problem with data $A + \Delta A$, $B + \Delta B$, $b + \Delta b$ and d, where*

$$\begin{aligned} \|\Delta A\|_F &\le c_1(u)\|A\|_F, \quad \|\Delta B\|_F \le c_2(u)\|B\|_F, \\ \|\Delta b\|_2 &\le c_3(u)\|b\|_2, \qquad \|\Delta\overline{x}\|_2 \le c_4(u)\|\overline{x}\|_2. \end{aligned}$$

PROOF. Standard error analysis of Householder transformations [5, Lem. 19.3, Thm. 19.4] shows that the computed $\widehat{K}$ and $\widehat{C}$ satisfy

$$\begin{aligned} (B + \Delta B_1)\widetilde{Q} &= \begin{bmatrix} \widehat{K} & 0 \end{bmatrix}, & \|\Delta B_1\|_F &\le c_1(u)\|B\|_F, \\ \widehat{C} &= (A + \Delta A_1)\widetilde{Q}, & \|\Delta A_1\|_F &\le c_2(u)\|A\|_F, \end{aligned}$$

where $\widetilde{Q}$ is orthogonal. The computed solution of the triangular system $\widehat{K}y_1 = d$ satisfies [5, Thm. 8.5]

$$(\widehat{K} + \Delta K)\widehat{y}_1 = d, \qquad \|\Delta K\|_F \le c_3(u)\|\widehat{K}\|_F.$$

The computed coefficient vector $\widehat{f} = fl(b - \widehat{C}_1\widehat{y}_1)$ of the ILS problem (3.3) satisfies

$$\widehat{f} = b + \Delta b_0 - (\widehat{C}_1 + \Delta\widehat{C}_0)\widehat{y}_1,$$
$$\|\Delta b_0\|_2 \le u\|b\|_2, \qquad \|\Delta\widehat{C}_0\|_F \le c_4(u)\|\widehat{C}_1\|_F.$$

(For simplicity, we are assuming that C_1 is explicitly formed and applied to y_1; the analysis remains valid if $C_1 y_1 = AQ(:, 1\colon s)y_1$ is computed without forming C_1 or $Q(:, 1\colon s)$.) We can apply to (3.3) the error analysis of [2] to deduce that the computed $\widehat{y}_2$ solves

$$\min_{y_2}(\widehat{f} + \Delta f - (\widehat{C}_2 + \Delta\widehat{C}_2)y_2)^T J(\widehat{f} + \Delta f - (\widehat{C}_2 + \Delta\widehat{C}_2)y_2),$$

where

$$\|\Delta\widehat{C}_2\|_F \le c_5(u)\|\widehat{C}_2\|_F, \qquad \|\Delta f\|_2 \le c_6(u)\|\widehat{f}\|_2.$$

In view of the latter inequality we can write, using [3, Lem. 3.3],

$$\widehat{f} + \Delta f = b + \Delta b - (\widehat{C}_1 + \Delta\widehat{C}_1)\widehat{y}_1,$$
$$\|\Delta b\|_2 \le c_7(u)\|b\|_2, \quad \|\Delta C_1\|_2 \le c_7(u)\|\widehat{C}_1\|_F.$$

Putting all these results together, we conclude that $\overline{x} = \widetilde{Q}\widehat{y}$ is the true solution to the perturbed problem with data $A + \Delta A$, $B + \Delta B$ and $b + \Delta b$, where

$$\Delta A = \Delta A_1 + \begin{bmatrix} \Delta\widehat{C}_1 & \Delta\widehat{C}_2 \end{bmatrix}\widetilde{Q}^T, \quad \Delta B = \Delta B_1 + \begin{bmatrix} \Delta K & 0 \end{bmatrix}\widetilde{Q}^T.$$

The bounds on $\|\Delta A\|_F$ and $\|\Delta B\|_F$ follow readily. Finally,

$$\widehat{x} = fl(Q\widehat{y}) = \widetilde{Q}\widehat{y} + \Delta\overline{x}, \qquad \|\Delta\overline{x}\|_2 \le c_8(u)\|\widehat{y}\|_2 = c_8(u)\|\overline{x}\|_2,$$

using [5, Lem. 19.3] again. □

4 The GHQR method.

Another method for solving the ILSE problem can be derived using the generalized hyperbolic QR factorization introduced in Theorem 1.1. Using (1.8) the constraint $Bx = d$ can be written

$$Ky_1 = \begin{bmatrix} K & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = d, \qquad y = Q^T x.$$

This triangular system uniquely determines y_1. Then, with $c = Hb$,

$$\begin{aligned}(b-Ax)^T J(b-Ax) &= (c-HAQy)^T J(c-HAQy)\\ &= \begin{bmatrix} c_1 - L_{11}y_1 \\ c_2 - L_{21}y_1 - L_{22}y_2 \end{bmatrix}^T J \begin{bmatrix} c_1 - L_{11}y_1 \\ c_2 - L_{21}y_1 - L_{22}y_2 \end{bmatrix} \\ &= \|c_2 - L_{21}y_1 - L_{22}y_2\|_2^2 + (c_1 - L_{11}y_1)^T \widetilde{J}(c_1 - L_{11}y_1),\end{aligned} \tag{4.1}$$

where $\widetilde{J} = J(1\colon p+q-(n-s), 1\colon p+q-(n-s)) = \operatorname{diag}(-I_q, I_{p-(n-s)})$. Therefore y_2 is uniquely determined as the solution of the triangular system $L_{22}y_2 = c_2 - L_{21}y_1$. This use of the GHQR factorization forms the basis for our next method.

ALGORITHM 4.1 (GHQR METHOD). *This algorithm solves the ILSE problem* (1.3) *using a GHQR factorization of A and B.*

1. Compute the Householder QR factorization of B^T: $BQ = B\,[\,Q_1 \quad Q_2\,] = [\,K \quad 0\,]$.
2. Solve the lower triangular system $Ky_1 = d$.
3. $C_2 = AQ_2$.
4. Compute the hyperbolic QL factorization $HC_2 = [\,0 \quad L_{22}^T\,]^T$, where L_{22} is lower triangular and H is J-orthogonal, using the natural modification of the method of [1].
5. $f = [\,f_1^T \quad f_2^T\,]^T = H(b - AQ_1y_1)$ (compute by applying H in factored form).
6. Solve the lower triangular system $L_{22}y_2 = f_2$.
7. $x = Qy$.

Again, the cost of the method is a complicated function of p, q, n and s. For $p+q \gg n \gg s$ it is approximately $4(p+q)n^2$ flops, compared with $7(p+q)n^2$ flops for the GQR-Cholesky method. It is clear that the GHQR method always requires fewer flops than the GQR-Cholesky method, since the latter method requires significant additional computation beyond a generalized (hyperbolic) QR factorization while the former does not.

4.1 Error analysis.

We re-interpret the GHQR method in order to be able to make use of existing error analysis in [1]. The method essentially recasts the ILSE problem as the unconstrained ILS problem

$$\min_{y_2}(g - C_2y_2)^T J(g - C_2y_2), \qquad C_2 = AQ_2, \quad g = b - AQ_1y_1, \tag{4.2}$$

where y_1 is determined by the constraints, and then solves this problem by the hyperbolic QR factorization method of [1].

The analysis for steps 1–3 of Algorithm 4.1 is given in the proof of Theorem 3.1: we recall that

$$\begin{aligned}(B+\Delta B_1)\widetilde{Q} &= \begin{bmatrix}\widehat{K} & 0\end{bmatrix}, & \|\Delta B_1\|_F &\le c_1(u)\|B\|_F,\\ \widehat{C}_2 &= (A+\Delta A_2)\widetilde{Q}_2, & \|\Delta A_2\|_F &\le c_2(u)\|A\|_F,\\ (\widehat{K}+\Delta K)\widehat{y}_1 &= d, & \|\Delta K\|_F &\le c_3(u)\|\widehat{K}\|_F,\end{aligned}$$

where $\widetilde{Q} = \begin{bmatrix}\widetilde{Q}_1 & \widetilde{Q}_2\end{bmatrix}$ is orthogonal. In addition,

$$\widehat{g} = b+\Delta b-(A+\Delta A_1)\widetilde{Q}_1\widehat{y}_1, \qquad \|\Delta A_1\|_F \le c_4(u)\|A\|_F, \quad \|\Delta b\|_2 \le u\|b\|_2.$$

The computed data $\widehat{g}$ and $\widehat{C}_2$ is therefore exact for the perturbed ILSE problem with data $A+\Delta A$, $B+\Delta B$ and $b+\Delta b$ with

$$\Delta A = \begin{bmatrix}\Delta A_1\widetilde{Q}_1 & \Delta A_2\widetilde{Q}_2\end{bmatrix}\widetilde{Q}^T, \qquad \Delta B = \Delta B_1 + [\,\Delta K \quad 0\,]\,\widetilde{Q}^T.$$

Clearly, these are small normwise relative perturbations.

We now need to invoke the results of [1] for the unconstrained ILS problem, which we summarize in the next theorem.

THEOREM 4.1. *Consider the ILS problem*

$$\min_x (b-Ax)^T J(b-Ax), \tag{4.3}$$

were $A \in \mathbb{R}^{(p+q)\times n}$, $b \in \mathbb{R}^{p+q}$, and the perturbed ILS problem

$$\min_x \big(b+\Delta b-(A+\Delta A)x\big)^T J\big(b+\Delta b-(A+\Delta A)x\big),$$

where

$$\|\Delta A\|_F \le \epsilon\|A\|_F, \quad \|\Delta b\|_2 \le \epsilon\|b\|_2.$$

We have

$$\begin{aligned}\frac{\|\Delta x\|_2}{\|x\|_2} \;\le\; \epsilon\Bigg[&\|M^{-1}A^T\|_2\|A\|_F\left(\frac{\|b\|_2}{\|A\|_F\|x\|_2}+1\right)\\ &+\|M^{-1}\|_2\|A\|_F^2\frac{\|r\|_2}{\|A\|_F\|x\|_2}\Bigg]+\mathrm{O}(\epsilon^2),\end{aligned}$$

where $M = A^TJA$.

THEOREM 4.2. *Let the ILS problem* (4.3) *be solved using a hyperbolic QR factorization of A by the method in* [1] *based on Householder transformations and hyperbolic rotations applied in mixed* (*or factored*) *form. Under the assumption that the perturbation bound of Theorem* 4.1 *is approximately attainable, this method is forward stable* (*that is, the forward error is bounded in the same way as for a backward stable method*).

The assumption in Theorem 4.2 is reasonable in that in extensive numerical experiments reported in [1], for every A and b the perturbation bound was approximately attained for some ΔA and Δb.

Using these results from [1] we conclude that under the assumption in Theorem 4.2, (4.2) is solved in a forward stable manner. It is not too hard to see that a normwise relative perturbation of g and C_2 in (4.2) corresponds to a normwise relative perturbation of the same size of A, B and b in the original ILSE problem. Therefore forward errors introduced by a forward stable method applied to (4.2) are no larger than those that can be produced by small normwise relative perturbations in the ILSE problem.

Combining these two parts of analysis leads to the next result.

THEOREM 4.3. *Under the assumption that the perturbation bound of Theorem* 4.1 *is approximately attainable, the computed solution $\widehat{x}$ from the GHQR method is forward stable; therefore the forward error $\Delta x = \widehat{x} - x$ satisfies* (2.8) *with $\epsilon = c(u)$.*

5 Numerical experiments.

The aim of this section is to compare the predictions of the error analysis with the errors observed in practice. We do not have any way to test the mixed forward–backward stability of the GQR-Cholesky method, so we concentrate on testing forward stability.

As well as the GQR-Cholesky method (Algorithm 3.1) and the GHQR method (Algorithm 4.1), we also test the "augmented system method" that forms the augmented system (2.2) and solves it by LU factorization with partial pivoting. This method requires $2(s+p+q+n)^3/3$ flops, which is much more than the other two methods, but its ease of coding makes it of possible interest for small dimensions.

We generate test problems for which different terms of the perturbation bound (2.8) dominate. This is achieved by choosing the matrices in the GHQR factorization (1.8) and thereby defining A and B. Crucial here are the conditioning of the triangular matrices L_{22} and K and the norm of H. The triangular matrices are generated via the QR factorizations of random matrices with pre-assigned singular value distributions (generated via MATLAB's `gallery('randsvd',...)`). Random J-orthogonal matrices H of specified norm are generated using the method of Higham [6]. The orthogonal factor Q in (1.8) is also generated randomly (via MATLAB's `gallery('qmult',...)`). Given that we know the GHQR factorization we are able to control the size of the residual $r = b - Ax$ by choosing c_1 (and hence b) in (4.1) appropriately.

Our experiments are carried out in MATLAB, with unit roundoff $u \approx 1.1 \times 10^{-16}$. We show results for five different problems with $p = 8$, $q = 6$, $n = 6$ and $s = 4$ in Table 5.1. The first three columns show the relative errors $\|x - \widehat{x}\|_2 / \|x\|_2$ for the three methods of interest; for the true solution x we take a solution computed at high precision using MATLAB's Symbolic Math Toolbox. The fourth column tabulates the first order part of the perturbation bound (2.8),

Table 5.1: Errors for the three methods, and perturbation bounds, for five problems.

GQR-Cholesky	GHQR	Aug. system	Bound (2.8)	terms	Kronecker bound (2.9)
2.2e-8	1.9e-8	1.1e-8	8.0e-6	3.2e3, 9.6e6, 7.2e10	7.1e-7
9.5e-14	9.7e-14	3.1e-13	1.0e-12	1.9e3, 7.4e3, 1.4e-9	1.0e-12
2.3e-8	2.3e-8	2.5e-9	2.3e-7	2.1e9, 4.6e3, 3.2e2	2.3e-7
1.2e-3	1.4e-3	2.7e-4	2.2e-2	6.8e2, 1.9e14, 7.8e11	1.4e-2
2.3e-7	2.2e-7	3.6e-7	8.1e-6	2.4e5, 7.3e10, 2.9e5	5.3e-6

with $\epsilon = u$, and the next three columns show the sizes of the three first order terms in that bound (without the ϵ factor). The final column is the first order part of the Kronecker-based perturbation bound (2.9), with $\epsilon = u$, which we know is sharp.

Two main features are notable in the results. First, all three methods behave in a forward stable manner, since the errors are no larger than the bound (2.9). Our results are therefore consistent with our error analysis of the GQR-Cholesky and GHQR methods. Second, the bound (2.8) is of similar size to (2.9), being a most a factor 10 bigger, showing that it is reasonably sharp in these tests.

Finally, we emphasize that computations with J-orthogonal matrices are delicate, and must be carefully arranged in order to avoid instability. To illustrate, we repeated the tests with step 5 of Algorithm 4.1 carried out by explicitly forming H and then multiplying the result into $b - A_1 y_1$ (recall that the algorithm requires H to be applied in factored form). The results for the GHQR method were quantitatively unchanged *except* for the fourth test problem: here $\|H\|_2 \approx 10^6$ and the error was 2.4, so the algorithm performed unstably with this modified implementation.

REFERENCES

1. A. Bojanczyk, N. J. Higham, and H. Patel, *Solving the indefinite least squares problem by hyperbolic QR factorization*, Numerical Analysis Report No. 397, Manchester Centre for Computational Mathematics, Manchester, England, 2002. Revised April 2002. To appear in SIAM J. Matrix Anal. Appl.
2. S. Chandrasekaran, M. Gu, and A. H. Sayed, *A stable and efficient algorithm for the indefinite linear least-squares problem*, SIAM J. Matrix Anal. Appl., 20:2 (1998), pp. 354–362.
3. A. J. Cox and N. J. Higham, *Accuracy and stability of the null space method for solving the equality constrained least squares problem*, BIT, 39:1 (1999), pp. 34–50.
4. L. Eldén *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17:3 (1980), pp. 338–350.
5. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd ed., 2002.
6. N. J. Higham, *J-orthogonal matrices: Properties and generation*, Numerical Analysis Report No. 408, Manchester Centre for Computational Mathematics, Manchester, England, 2002.
7. R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, UK, 1991.