IMA Journal of Numerical Analysis (2014) Page 1 of 26
doi:10.1093/imanum/dru004

A structurally backward stable algorithm for solving the indefinite least squares problem with equality constraints

NICOLA MASTRONARDI*
Istituto per le Applicazioni del Calcolo 'M. Picone', Consiglio Nazionale delle Ricerche, sede di Bari, via Amendola 112D, 70126 Bari, Italy
*Corresponding author: n.mastronardi@ba.iac.cnr.it

AND

PAUL VAN DOOREN
Department of Mathematical Engineering, Catholic University of Louvain, Louvain-la-Neuve, Belgium
paul.vandooren@uclouvain.be

[Received on 8 May 2013; revised on 22 January 2014]

The equality constrained indefinite least squares problem involves the minimization of an indefinite quadratic form subject to a linear equality constraint. In this paper, we study this problem and present a numerical method that is proved to be backward stable in a strict sense, i.e., that the computed solution satisfies a slightly perturbed equality constrained indefinite least squares problem. We also perform a sensitivity analysis of this problem and derive bounds for the accuracy of the computed solution. We give several numerical experiments to illustrate these results.

Keywords: indefinite matrices; matrix decomposition; stability; sensitivity; least squares.

1. Introduction

It is well known (see, e.g., Bojanczyk *et al.*, 2003a) that the indefinite least squares problem with linear equality constraints (abbreviated as ILSE)

$$\min_{\mathbf{x}} \|A\mathbf{x}-\mathbf{b}\|^2_{\Sigma_{pq}}, \quad \text{subject to } B\mathbf{x}=\mathbf{d}, \tag{1.1}$$

where $A \in \mathbb{R}^{(p+q)\times n}$, $B \in \mathbb{R}^{s\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^{(p+q)}$, $\mathbf{d} \in \mathbb{R}^s$ and

$$\|A\mathbf{x}-\mathbf{b}\|^2_{\Sigma_{pq}} := (A\mathbf{x}-\mathbf{b})^{\mathrm{T}}\Sigma_{pq}(A\mathbf{x}-\mathbf{b}), \quad \Sigma_{pq} := \mathrm{diag}\{I_p, -I_q\}$$

has a unique solution if B has full row rank s and if the matrix $A^{\mathrm{T}}\Sigma_{pq}A$ is positive definite on $\mathrm{Ker}B$. Moreover, this unique solution is also obtained from the unique solution of the augmented linear system

$$\begin{bmatrix} 0 & 0 & B \\ 0 & \Sigma_{pq} & A \\ B^{\mathrm{T}} & A^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{s} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \\ 0 \end{bmatrix}, \tag{1.2}$$

which we will denote for short by

$$M\mathbf{y} = \mathbf{f}. \tag{1.3}$$

This can be seen as follows. Let $\mathbf{r} := \mathbf{b} - A\mathbf{x}$ be the residual of the linear system being minimized and let λ be the Lagrange multiplier in the unconstrained problem

$$\min_{\mathbf{x}} \frac{1}{2}(A\mathbf{x} - \mathbf{b})^{\mathrm{T}} \Sigma_{pq}(A\mathbf{x} - \mathbf{b}) - \lambda^{\mathrm{T}}(B\mathbf{x} - \mathbf{d});$$

then defining $\mathbf{s} := \Sigma_{pq}\mathbf{r}$ yields the equations

$$B\mathbf{x} = \mathbf{d}, \quad \Sigma_{pq}\mathbf{s} + A\mathbf{x} = \mathbf{b}, \quad B^{\mathrm{T}}\lambda = -A^{\mathrm{T}}\mathbf{s} = A^{\mathrm{T}}\Sigma_{pq}(A\mathbf{x} - \mathbf{b}).$$

This problem has been analysed by several authors, including Chandrasekaran *et al.* (1998) and Bojanczyk *et al.* (2003b) (without constraints), Liu & Wang (2010), Liu *et al.* (2010) and Mastronardi & Van Dooren (2013a). Two different algorithms were proposed in Bojanczyk *et al.* (2003a). The first one, called GQR–Cholesky, was shown to be backward stable. The second one, called GHQR, was shown to be forward stable, that is, the forward error is bounded in the same way as if the algorithm were backward stable. The algorithm proposed in Mastronardi & Van Dooren (2013a) transforms the block matrix in (1.2) to lower antitriangular form, mixing up the blocks. The algorithm proposed here transforms the block matrix in (1.2) to lower antitriangular form preserving the original block structure. Although independently derived, the algorithm is closely related to GQR–Cholesky, but slightly more efficient. Moreover, we give a new proof of backward numerical stability in a strict sense. We also give a new sensitivity analysis, which we believe brings new insights with the problem. Finally, we show that a proper scaling of the problem may yield significant improvements in estimating the sensitivity of the problem.

The paper is organized as follows. The definitions and the notation used in the paper, followed by the description on how to scale the augmented system, are described in Section 2. The proposed algorithm is described in Section 3. The sensitivity of the solution computed by the proposed method is studied in Section 4, followed by the section on the numerical experiments and by the conclusions.

2. Background

2.1 *Definitions and notation*

The method we will describe here is heavily based on a so-called antitriangular decomposition presented in Mastronardi & Van Dooren (2012, 2013b), and hence it is useful to recall some basic definitions here. A matrix $A \in \mathbb{R}^{n\times n}$ is called upper (respectively, lower) antitriangular if all the entries below (respectively, above the main antidiagonal are zero. An antitriangular matrix is obviously nonsingular if all the entries in the main antidiagonal are different from zero. Moreover, the solution of a square linear system, with coefficient matrix a nonsingular upper (respectively, lower) antitriangular matrix of order n, can be computed by forward (respectively, backward) substitution in n^2 floating point operations (see, e.g., Mastronardi & Van Dooren, 2012, 2013b).

The right lower anti-QR factorization of a matrix $A \in \mathbb{R}^{n\times m}$, $m \geqslant n$, $\operatorname{rank}(A) = n$, is as

$$A = [\mathbf{0} \quad Y]V^{\mathrm{T}}, \quad [\mathbf{0} \quad Y] = \begin{bmatrix} & \\ & \diagup \end{bmatrix},$$

with $Y \in \mathbb{R}^{n\times n}$ lower antitriangular and $V \in \mathbb{R}^{m\times m}$ orthogonal. The left upper anti-QR factorizations of a matrix $A \in \mathbb{R}^{m\times n}$, $m \geqslant n$, $\operatorname{rank}(A) = n$, is defined as Q5

$$A = U \begin{bmatrix} Y \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} Y \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \diagup \end{bmatrix},$$

with $Y \in \mathbb{R}^{n\times n}$ upper antitriangular and $U \in \mathbb{R}^{m\times m}$ orthogonal.

We will say that a symmetric indefinite matrix $A \in \mathbb{R}^{n\times n}$ is symmetric block upper antitriangular if

$$A = \begin{bmatrix} X & Z & Y \\ Z^{\mathrm{T}} & W & \\ Y^{\mathrm{T}} & & \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_1 \end{matrix}, \tag{2.1}$$

with $Y \in \mathbb{R}^{n_1\times n_1}$ nonsingular upper antitriangular, $X \in \mathbb{R}^{n_1\times n_1}$ symmetric, $W \in \mathbb{R}^{n_2\times n_2}$ symmetric, $Z \in \mathbb{R}^{n_1\times n_2}$ and $2n_1 + n_2 = n$. We say that it is symmetric block lower antitriangular if

$$A = \begin{bmatrix} & & Y \\ & W & Z \\ Y^{\mathrm{T}} & Z^{\mathrm{T}} & X \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_1 \end{matrix}, \tag{2.2}$$

with $Y \in \mathbb{R}^{n_1\times n_1}$ nonsingular lower antitriangular, $X \in \mathbb{R}^{n_1\times n_1}$ symmetric, $W \in \mathbb{R}^{n_2\times n_2}$ symmetric, $Z \in \mathbb{R}^{n_2\times n_1}$ and $2n_1 + n_2 = n$. These forms are convenient to determine the inertia of the matrix A, which is defined as the triple $\mathbf{In}(A) = (n_-, n_0, n_+)$, where n_-, n_0 and n_+ are, respectively, the number of negative, zero and positive eigenvalues of the symmetric matrix A. Indeed, if A is in one of the two symmetric block antitriangular forms, then the following theorem holds (see, e.g., Gould, 1985; Mastronardi & Van Dooren, 2013b).

THEOREM 2.1 Let A be a symmetric matrix partitioned either as in (2.1) or in (2.2). Then,

$$\mathbf{In}(A) = (n_1, 0, n_1) + \mathbf{In}(W).$$

Finally, let $\mathbf{v} \in \mathbb{R}^n$ be partitioned in subvectors as follows:

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_j \end{bmatrix} \begin{matrix} \}\iota_1 \\ \}\iota_2 \\ \vdots \\ \}\iota_j \end{matrix}, \quad j \leqslant n, \quad \sum_{l=1}^{j} \iota_l = n.$$

Then, the vector containing the following subvectors:

$$\begin{bmatrix} \mathbf{v}_{\iota_1} \\ \mathbf{v}_{\iota_1+1} \\ \vdots \\ \mathbf{v}_{\iota_2} \end{bmatrix}, \quad 1 \leqslant \iota_1 < \iota_2 \leqslant j$$

is denoted by $\mathbf{v}_{\iota_1:\iota_2}$.

2.2 *Scaling the problem*

In order to study the sensitivity of this problem, we first show that we can construct a simple scaling such that all nonzero blocks and the vector $\mathbf{f}$ in (1.3) have spectral norm equal to 1. If we multiply the system $A\mathbf{x} = \mathbf{b}$ by an appropriate scalar α and the system $B\mathbf{x} = \mathbf{d}$ by an appropriate scalar β, we can impose $\|\alpha A\|_2 = 1$ and $\|\beta B\|_2 = 1$. Finally, we can multiply the right-hand side and the solution with a scalar γ so that the new right-hand side has norm 1. If we apply these scalings to the equation $M\mathbf{y} = \mathbf{f}$ as $(DMD)(D^{-1}\mathbf{y}\gamma) = (D\mathbf{f}\gamma)$ with $D = \mathrm{diag}\{(\beta/\alpha)I_s, I_{p+q}, \alpha I_n\}$ we obtain the equivalent scaled system

$$\begin{bmatrix} 0 & 0 & \beta B \\ 0 & \Sigma_{pq} & \alpha A \\ \beta B^{\mathrm{T}} & \alpha A^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \frac{\alpha\gamma}{\beta}\lambda \\ \gamma\mathbf{s} \\ \frac{\gamma}{\alpha}\mathbf{x} \end{bmatrix} = \begin{bmatrix} \frac{\beta\gamma}{\alpha}\mathbf{d} \\ \gamma\mathbf{b} \\ 0 \end{bmatrix}, \tag{2.3}$$

which shows that this yields the same solution $\mathbf{x}$ up to a scaling as well. We will therefore assume in the future that this scaling has been applied on the above problem and associated system $M\mathbf{y} = \mathbf{f}$, and hence that $\|A\|_2 = \|B\|_2 = \|\Sigma_{pq}\|_2 = \|\mathbf{f}\|_2 = 1$, which implies that $1 \leqslant \|M\|_2 \leqslant 2$, since

$$M = \begin{bmatrix} 0 & 0 & B \\ 0 & \Sigma_{pq} & 0 \\ B^{\mathrm{T}} & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & A \\ 0 & A^{\mathrm{T}} & 0 \end{bmatrix}.$$

In the algorithm that we develop in this paper, that scaling does not appear to affect the accuracy of the computed solution, but it does improve significantly the error bounds for the computed solution.

3. Solution of the augmented linear system

Since the solution of the ILSE problem can be described in terms of the augmented system

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} & B \\ \mathbf{0} & \Sigma_{pq} & A \\ B^{\mathrm{T}} & A^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{s} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \\ \mathbf{0} \end{bmatrix} \quad \Leftrightarrow \quad M\mathbf{y} = \mathbf{f}, \tag{3.1}$$

it is convenient to use it to describe our algorithm. Here λ is the vector of the Lagrange multipliers and $N := p + q + n + s$ is the order of the matrix M. It can then be shown (see, e.g., Mastronardi & Van Dooren, 2013a) that if B has full row rank and $A^{\mathrm{T}}\Sigma_{pq}A$ is positive definite on $\mathrm{Ker}B$, then its inertia is given by

$$\mathbf{In}(M) = (q + n, 0, p + s). \tag{3.2}$$

Let $M^{(0)} = M$, $\mathbf{y}^{(0)} = \mathbf{y}$, $\mathbf{f}^{(0)} = \mathbf{f}$, where $\mathbf{y}^{(0)}$ and $\mathbf{f}^{(0)}$ are partitioned as follows:

$$\mathbf{y}^{(0)} = \begin{bmatrix} \lambda \\ \mathbf{s} \\ \mathbf{x} \end{bmatrix} \begin{matrix} \}s \\ \}p+q, \\ \}n \end{matrix} \quad \mathbf{f}^{(0)} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}s \\ \}p+q. \\ \}n \end{matrix}$$

We then perform orthogonal congruence transformations to reduce $M^{(0)}$ to a more convenient form.

Step 1. Reduction of the constraint $B\mathbf{x} = \mathbf{d}$

Here we eliminate that part of the vector $\mathbf{x}$ that can be assigned via the constraint $B\mathbf{x} = \mathbf{d}$. For this, we first compute the left lower antitriangular factorization of B, i.e.,

$$B = [0,\ Y_1]Q_B^{\mathrm{T}}, \tag{3.3}$$

with $Y_1 \in \mathbb{R}^{s\times s}$ nonsingular lower antitriangular and $Q_B \in \mathbb{R}^{n\times n}$ orthogonal, partitioned as follows:

$$Q_B^{\mathrm{T}} = \begin{bmatrix} Q_{B1}^{\mathrm{T}} \\ Q_{B2}^{\mathrm{T}} \end{bmatrix} \begin{matrix} \}n-s \\ \}s \end{matrix}, \tag{3.4}$$

with the columns of $Q_{B1} \in \mathbb{R}^{n\times(n-s)}$ spanning the nullspace of B. Let

$$Q^{(1)} = \begin{bmatrix} I_{p+q+s} & \\ & Q_B \end{bmatrix};$$

then the linear system (3.1) is transformed into the equivalent system

$$\tilde{M}^{(0)}\tilde{\mathbf{y}}^{(0)} = \tilde{\mathbf{f}}^{(0)}, \tag{3.5}$$

where

$$\tilde{M}^{(0)} = {Q^{(1)}}^{\mathrm{T}} M^{(0)} Q^{(1)} = \left[\begin{array}{c|c|c|c} & & & Y_1 \\ \hline & \Sigma_{pq} & A_1 & A_2 \\ \hline & {A_1}^{\mathrm{T}} & \mathbf{0} & \mathbf{0} \\ \hline {Y_1}^{\mathrm{T}} & {A_2}^{\mathrm{T}} & \mathbf{0} & \mathbf{0} \end{array}\right],$$

with

$$A_1 = AQ_{B1} \in \mathbb{R}^{(p+q)\times(n-s)}, \quad A_2 = AQ_{B2} \in \mathbb{R}^{(p+q)\times s} \tag{3.6}$$

and

$$\tilde{\mathbf{y}}^{(0)} = {Q^{(1)}}^{\mathrm{T}}\mathbf{y}^{(0)} = \begin{bmatrix} \lambda \\ \mathbf{s} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \begin{matrix} \}s \\ \}p+q \\ \}n-s \\ \}s \end{matrix}, \quad \tilde{\mathbf{f}}^{(0)} = {Q^{(1)}}^{\mathrm{T}}\mathbf{f}^{(0)} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}s \\ \}p+q \\ \}n-s \\ \}s \end{matrix}.$$

This thus allows us to solve (by backward substitution) the nonsingular lower antitriangular linear system

$$Y_1\mathbf{x}_2 = d, \tag{3.7}$$

and eliminate $\mathbf{x}_2$. Therefore, we can also update the right-hand side with $\hat{\mathbf{b}} = \mathbf{b} - A_2\mathbf{x}_2$ as follows:

$$\hat{\mathbf{f}}^{(0)} = \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{b}} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and extract from this a new (smaller) system:

$$M^{(1)}\mathbf{y}^{(1)} = \mathbf{f}^{(1)}, \tag{3.8}$$

where

$$M^{(1)} = \left[\begin{array}{c|c} \Sigma_{pq} & A_1 \\ \hline A_1{}^{\mathrm{T}} & \end{array}\right], \quad \mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{s} \\ \mathbf{x}_1 \end{bmatrix}, \quad \mathbf{f}^{(1)} = \begin{bmatrix} \hat{\mathbf{b}} \\ \mathbf{0} \end{bmatrix}. \tag{3.9}$$

Note that in this smaller system we have not only eliminated $\mathbf{x}_2$, but also the associated Lagrange multiplier λ. Moreover, since $\Sigma_{pq}^{-1} = \Sigma_{pq}$,

$$M^{(1)} = \left[\begin{array}{c|c} I & \\ \hline A_1{}^{\mathrm{T}}\Sigma_{pq} & I \end{array}\right] \left[\begin{array}{c|c} \Sigma_{pq} & \\ \hline & -A_1{}^{\mathrm{T}}\Sigma_{pq}A_1 \end{array}\right] \left[\begin{array}{c|c} I & \Sigma_{pq}A_1 \\ \hline & I \end{array}\right].$$

Therefore, by (3.2), it follows that its inertia is given by

$$\mathbf{In}(M^{(1)}) = (q + n - s, 0, p). \tag{3.10}$$

Step 2. Breaking up the variable $\mathbf{s}$
We first point out that $p \geqslant n - s$ since $A_1^{\mathrm{T}}\Sigma_{pq}A_1$ is supposed to be positive definite (Bojanczyk *et al.*, 2003a). Let us then partition A_1 as follows:

$$A_1 = \begin{bmatrix} A_{11} \\ A_{12} \end{bmatrix} \begin{matrix} \}p \\ \}q \end{matrix},$$

and $\mathbf{y}^{(1)}$ and $\mathbf{f}^{(1)}$ as follows:

$$\mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{x}_1 \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \\ \}q \\ \}n-s \end{matrix}, \quad \mathbf{f}^{(1)} = \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \\ \hat{\mathbf{b}}_3 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \\ \}q \\ \}n-s \end{matrix}.$$

Compute the left upper antitriangular factorization of A_{11},

$$A_{11} = Q_p \begin{bmatrix} \tilde{Y}_2 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \end{matrix}, \tag{3.11}$$

with $\tilde{Y}_2$ nonsingular upper antitriangular and let

$$Q^{(2)} = \begin{bmatrix} Q_p & \\ & I_{n+q-s} \end{bmatrix}.$$

Then, the linear system (3.8) is transformed into the following equivalent system:

$$\tilde{M}^{(1)}\tilde{\mathbf{y}}^{(1)} = \tilde{\mathbf{f}}^{(1)}, \tag{3.12}$$

where

$$\tilde{M}^{(1)} = Q^{(2)\mathrm{T}} M^{(1)} Q^{(2)} = \left[\begin{array}{c|c|c|c} I_{n-s} & & & \tilde{Y}_2 \\ \hline & I_{p-n+s} & & \mathbf{0} \\ \hline & & -I_q & A_{12} \\ \hline \tilde{Y}_2^{\mathrm{T}} & \mathbf{0} & A_{12}^{\mathrm{T}} & \end{array}\right],$$

$$\tilde{\mathbf{y}}^{(1)} = Q^{(2)\mathrm{T}}\mathbf{y}^{(1)} = \begin{bmatrix} Q_p{}^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} \\ \begin{bmatrix} \mathbf{s}_3 \\ \mathbf{x}_1 \end{bmatrix} \end{bmatrix}, \quad \tilde{\mathbf{f}}^{(1)} = Q^{(1)\mathrm{T}}\mathbf{f}^{(1)} = \begin{bmatrix} Q_p{}^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} \\ \begin{bmatrix} \hat{\mathbf{b}}_3 \\ \mathbf{0} \end{bmatrix} \end{bmatrix} \tag{3.13}$$

120 are partitioned as $\mathbf{y}^{(1)}$ and $\mathbf{f}^{(1)}$, respectively.

Then compute the left upper antitriangular factorization of

$$\begin{bmatrix} \tilde{Y}_2 \\ A_{12} \end{bmatrix} = U \begin{bmatrix} Y_2 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}q \end{matrix}, \tag{3.14}$$

with $U \in \mathbb{R}^{(q+n-s)\times(q+n-s)}$ orthogonal and Y_2 nonsingular upper antitriangular. We partition U as follows:

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

with $U_{11} \in \mathbb{R}^{(n-s)\times(n-s)}$, $U_{12} \in \mathbb{R}^{(n-s)\times q}$, $U_{21} \in \mathbb{R}^{q\times(n-s)}$ and $U_{22} \in \mathbb{R}^{q\times q}$, and let

$$Q_\Sigma = \begin{bmatrix} U_{11} & & U_{12} \\ & I_{p-n+s} & \\ U_{21} & & U_{22} \end{bmatrix}, \quad Q^{(3)} = \begin{bmatrix} Q_\Sigma & \\ & I_{n-s} \end{bmatrix}.$$

125 Observe that, due to the particular structure of the left-hand side of (3.14), the matrix U_{11} is upper triangular. Then, the linear system (3.12) is transformed into the following equivalent one:

$$M^{(2)}\mathbf{y}^{(2)} = \mathbf{f}^{(2)}, \tag{3.15}$$

where

$$M^{(2)} = Q^{(3)\mathrm{T}}\tilde{M}^{(1)}Q^{(3)} = \begin{bmatrix} X_{11} & & X_{12} & Y_2 \\ & I_{p-n+s} & & \\ X_{12}{}^{\mathrm{T}} & & X_{22} & \\ Y_2^{\mathrm{T}} & & & \end{bmatrix},$$

with

$$\begin{aligned} X_{11} &= {U_{11}}^{\mathrm{T}} U_{11} - {U_{21}}^{\mathrm{T}} U_{21}, \\ X_{12} &= {U_{11}}^{\mathrm{T}} U_{12} - {U_{21}}^{\mathrm{T}} U_{22}, \\ X_{22} &= {U_{12}}^{\mathrm{T}} U_{12} - {U_{22}}^{\mathrm{T}} U_{22}, \end{aligned} \tag{3.16}$$

$$\mathbf{y}^{(2)} = \begin{bmatrix} \tilde{\mathbf{s}}_1 \\ \tilde{\mathbf{s}}_2 \\ \tilde{\mathbf{s}}_3 \\ \mathbf{x}_1 \end{bmatrix} = {Q^{(3)}}^{\mathrm{T}} \tilde{\mathbf{y}}^{(1)} = \begin{bmatrix} {U_{11}}^{\mathrm{T}} \tilde{\mathbf{s}}_1 - {U_{21}}^{\mathrm{T}} \mathbf{s}_3 \\ \bar{\mathbf{s}}_2 \\ {U_{12}}^{\mathrm{T}} \tilde{\mathbf{s}}_1 - {U_{22}}^{\mathrm{T}} \mathbf{s}_3 \\ \mathbf{x}_1 \end{bmatrix}, \tag{3.17}$$

$$\mathbf{f}^{(2)} = \begin{bmatrix} \tilde{\mathbf{b}}_1 \\ \tilde{\mathbf{b}}_2 \\ \tilde{\mathbf{b}}_3 \\ \mathbf{0} \end{bmatrix} = {Q^{(3)}}^{\mathrm{T}} \tilde{\mathbf{f}}^{(1)} = \begin{bmatrix} {U_{11}}^{\mathrm{T}} \bar{\mathbf{b}}_1 - {U_{21}}^{\mathrm{T}} \hat{\mathbf{b}}_3 \\ \bar{\mathbf{b}}_2 \\ {U_{12}}^{\mathrm{T}} \bar{\mathbf{b}}_1 - {U_{22}}^{\mathrm{T}} \hat{\mathbf{b}}_3 \\ \mathbf{0} \end{bmatrix} \tag{3.18}$$

and

$$\begin{bmatrix} \bar{\mathbf{s}}_1 \\ \bar{\mathbf{s}}_2 \end{bmatrix} = Q_p^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{b}}_1 \\ \bar{\mathbf{b}}_2 \end{bmatrix} = Q_p^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix}.$$

Observe that, by (3.10), $\mathbf{In}(M^{(2)}) = \mathbf{In}(M^{(1)}) = (q+n-s, 0, p)$. Therefore, by Theorem 2.1,

$$\mathbf{In}\left(\begin{bmatrix} I_{p-n+s} & \\ & X_{22} \end{bmatrix} \right) = (q, 0, p-n+s).$$

Hence, X_{22} is symmetric negative definite and has a Cholesky factorization

$$X_{22} = -L_{22} L_{22}^{\mathrm{T}}. \tag{3.19}$$

Step 3. Computing the solution
We can now solve for $\mathbf{y}^{(2)}$ from the linear system (3.15) as follows:

- $\tilde{\mathbf{s}}_2 = \tilde{\mathbf{b}}_2$;
- solve the lower antitriangular linear system $Y_2^{\mathrm{T}} \tilde{\mathbf{s}}_1 = \mathbf{0}$ which implies $\tilde{\mathbf{s}}_1 = \mathbf{0}$;
- solve the negative definite linear system $X_{22} \tilde{\mathbf{s}}_3 = -L_{22} L_{22}^{\mathrm{T}} \tilde{\mathbf{s}}_3 = \tilde{\mathbf{b}}_3$;
- solve the upper antitriangular linear system $Y_2 \mathbf{x}_1 = (\tilde{\mathbf{b}}_1 - X_{12} \tilde{\mathbf{s}}_3)$ by forward substitution.

Once $\mathbf{y}^{(2)}$ has been obtained, $\tilde{\mathbf{y}}^{(1)}$ can be computed as well using

$$\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = U \begin{bmatrix} \mathbf{y}_1^{(2)} \\ \mathbf{y}_3^{(2)} \end{bmatrix},$$

TABLE 1 *Computation complexity of the algorithm described in Section* 3

	Flops
(3.3)	$2s^2(n-s/3)$
(3.11)	$2(n-s)^2(p-(n-s)/3)$
(3.14)	$4q(n-s)(q+n-s)$
(3.16)	$(n-s)^3/3+q(n-s)^2+q^3$

and, because of (3.17),

$$\tilde{\mathbf{y}}^{(1)} = Q^{(3)}\mathbf{y}^{(2)} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{y}_2^{(1)} \\ \mathbf{v}_2 \\ \mathbf{y}_4^{(1)} \end{bmatrix}.$$

Furthermore, by (3.9) and (3.13),

$$\tilde{\mathbf{y}}_{2:3}^{(0)} = \mathbf{y}^{(1)} = Q^{(2)}\tilde{\mathbf{y}}^{(1)} = \begin{bmatrix} Q_p\tilde{\mathbf{y}}_{1:2}^{(1)} \\ \tilde{\mathbf{y}}_{3:4}^{(1)} \end{bmatrix}. \tag{3.20}$$

The solution of the linear system (3.5) can then completely be computed. Since, by (3.7) and (3.20), $\tilde{\mathbf{y}}_{2:4}^{(0)}$ has been already computed, we only need to compute $\tilde{\mathbf{y}}_1^{(0)}$. We first update the right-hand side:

$$\hat{\mathbf{f}}^{(0)} = \begin{bmatrix} \tilde{\mathbf{f}}_{1:3}^{(0)} \\ \tilde{\mathbf{f}}_4^{(0)} - A^{\mathrm{T}}\tilde{\mathbf{y}}_2^{(0)} \end{bmatrix}, \tag{3.21}$$

then solve the lower antitriangular linear system

$$Y_1^{\mathrm{T}}\tilde{\mathbf{y}}_1^{(0)} = \hat{\mathbf{f}}_4^{(0)}, \tag{3.22}$$

and finally construct

$$\mathbf{y}^{(0)} = Q^{(1)}\tilde{\mathbf{y}}^{(0)} = \begin{bmatrix} \tilde{\mathbf{y}}_{1:2}^{(0)} \\ Q_B\tilde{\mathbf{y}}_{3:4}^{(0)} \end{bmatrix}.$$

We observe that if one is interested only in the solution $\mathbf{x}$ of the initial problem, it is not needed to compute (3.21) and (3.22). These computations are needed only if one wants to know the Lagrange multipliers of (3.1).

In Table 1, we give the number of floating point operations needed in this algorithm, thereby neglecting the lower-order terms.

155 REMARK 3.1 The algorithm just described can be used for any values of q. When $q > n - s$, steps 2 and 3 can be modified to yield a more efficient algorithm. Such a modification is described in the appendix.

Let us now make the link between the above algorithm and GQR–Cholesky described in Bojanczyk *et al.* (2003a). The orthogonal transformation matrix constructed by this algorithm is in fact given by

$$Q = \begin{bmatrix} I_s & & \\ & Q_A & \\ & & Q_B \end{bmatrix} \quad \text{where } Q_A = \begin{bmatrix} Q_p & \\ & I_q \end{bmatrix} Q_\Sigma.$$

It is important to note here that this block transformation does not mix variables λ, $\mathbf{s}$ and $\mathbf{x}$ or $\mathbf{b}$ and $\mathbf{d}$,
160 or matrices A, B and Σ_{pq}. This is why the block scaling performed earlier will not affect the round-off errors of the algorithm, but we will see in the next section that it does affect the condition estimates.

The orthogonal transformation matrix implicitly constructed by GQR–Cholesky is

$$U = \begin{bmatrix} I_s & & \\ & U_A & \\ & & Q_B \end{bmatrix},$$

where U_A is the orthogonal factor of the QL factorization of A_1. The computational complexity of GQR–Cholesky, neglecting the terms of lower order, is $(n-s)^2(7(p+q) - 2(n-s))$.

165 Let us now show how the algorithm described here can be considered as an algorithm to compute the lower block antitriangular factorization of the coefficient matrix in (1.2). The algorithm just described transforms problem (1.2) into the following equivalent one:

$$M_q \mathbf{y}_q = \mathbf{f}_q,$$

with

$$M_q := Q^{\mathrm{T}} M Q = \begin{bmatrix} & & & & & Y_1 \\ & X_{11} & & X_{12} & Y_2 & A_1^{(2)} \\ & & I_{p-n+s} & & & A_2^{(2)} \\ & X_{12}^{\mathrm{T}} & & X_{22} & & A_3^{(2)} \\ & Y_2^{\mathrm{T}} & & & & \\ Y_1^{\mathrm{T}} & A_1^{(2)\mathrm{T}} & A_2^{(2)\mathrm{T}} & A_3^{(2)\mathrm{T}} & & \end{bmatrix}, \tag{3.23}$$

170
$$\begin{bmatrix} \mathbf{y}_q & \mathbf{f}_q \end{bmatrix} := Q^{\mathrm{T}} \begin{bmatrix} \mathbf{y} & \mathbf{f} \end{bmatrix} = \begin{bmatrix} \lambda & \mathbf{d} \\ \mathbf{s}_1 & \mathbf{b}_1 \\ \mathbf{s}_2 & \mathbf{b}_2 \\ \mathbf{s}_3 & \mathbf{b}_3 \\ \mathbf{x}_1 & 0 \\ \mathbf{x}_2 & 0 \end{bmatrix}, \tag{3.24}$$

where

$$\begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} := Q_A^{\mathrm{T}} A_2 = Q_\Sigma^{\mathrm{T}} \begin{bmatrix} Q_p^{\mathrm{T}} & \\ & I_q \end{bmatrix}^{\mathrm{T}} A_2.$$

Observe that the latter product is not explicitly computed. Moreover, after a suitable symmetric permutation, (3.23) becomes a block antitriangular system:

$$\begin{bmatrix} & & & & Y_1 \\ & & & Y_2^{\mathrm{T}} & \\ & I_{p-n+s} & & & A_2^{(2)} \\ & & X_{22} & X_{12}^{\mathrm{T}} & A_3^{(2)} \\ Y_2 & & X_{12} & X_{11} & A_1^{(2)} \\ Y_1^{\mathrm{T}} & A_2^{(2)\mathrm{T}} & A_3^{(2)\mathrm{T}} & A_1^{(2)\mathrm{T}} & \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{x}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{s}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_1 \\ \mathbf{0} \end{bmatrix}.$$

4. Stability

We consider the solution of the (normalized) symmetric system of Equation (1.2) using the last block antitriangular form, obtained from an orthogonal transformation which we will call Q_M:

$$Q_M \begin{bmatrix} 0 & 0 & B \\ 0 & \Sigma_{pq} & A \\ B^{\mathrm{T}} & A^{\mathrm{T}} & 0 \end{bmatrix} Q_M^{\mathrm{T}} = \begin{bmatrix} & & & & Y_1 \\ & & & Y_2^{\mathrm{T}} & \\ & I_{p-n+s} & & & A_2^{(2)} \\ & & X_{22} & X_{12}^{\mathrm{T}} & A_3^{(2)} \\ Y_2 & & X_{12} & X_{11} & A_1^{(2)} \\ Y_1^{\mathrm{T}} & A_2^{(2)\mathrm{T}} & A_3^{(2)\mathrm{T}} & A_1^{(2)\mathrm{T}} & \end{bmatrix}. \tag{4.1}$$

Since the algorithm produces an orthogonal congruence transformation, the backward error will be symmetric, and since the transformation has a block form compatible with the block structure of M, the backward error will also have this structure. Hence, if we denote the *computed* quantities with a hat, we have that

$$\hat{Q}_M (M + \Delta M) \hat{Q}_M^{\mathrm{T}} = \begin{bmatrix} & & & & \hat{Y}_1 \\ & & & \hat{Y}_2^{\mathrm{T}} & \\ & I_{p-n+s} & & & \hat{A}_2^{(2)} \\ & & \hat{X}_{22} & \hat{X}_{12}^{\mathrm{T}} & \hat{A}_3^{(2)} \\ \hat{Y}_2 & & \hat{X}_{12} & \hat{X}_{11} & \hat{A}_1^{(2)} \\ \hat{Y}_1^{\mathrm{T}} & \hat{A}_2^{(2)\mathrm{T}} & \hat{A}_3^{(2)\mathrm{T}} & \hat{A}_1^{(2)\mathrm{T}} & \end{bmatrix}, \tag{4.2}$$

with

$$\Delta M := \begin{bmatrix} 0 & 0 & \Delta B \\ 0 & \Delta\Sigma_{pq} & \Delta A \\ \Delta B^{\mathrm{T}} & \Delta A^{\mathrm{T}} & 0 \end{bmatrix}, \quad \Delta\Sigma_{pq} = \Delta\Sigma_{pq}^{\mathrm{T}}, \tag{4.3}$$

and also $\|\Delta M\|_2, \|\hat{Q}_M^{\mathrm{T}}\hat{Q}_M - I\|_2 \approx \epsilon$, with ϵ the machine precision, because this algorithm is known to be backward stable. Let $\|\Delta\Sigma_{pq}\|_2 := c\epsilon$. There exists, according to Lemma A.1 in the appendix, an updating transformation Q_{up} such that

$$Q_{up}(\Sigma_{pq} + \Delta\Sigma_{pq})Q_{up}^{\mathrm{T}} = \Sigma_{pq}, \quad \|Q_{up} - I\|_2 \leqslant \frac{c}{2}\epsilon.$$

Incorporating this transformation into $\hat{Q}_M$ then yields again the result (4.2) and (4.3), but now with $\Delta\Sigma_{pq} = 0$. We now turn to the perturbation of the solution of the linear system

$$\begin{bmatrix} & & & & \hat{Y}_1 \\ & & & \hat{Y}_2^{\mathrm{T}} & \\ & I_{p-n+s} & & & \hat{A}_2^{(2)} \\ & & \hat{X}_{22} & \hat{X}_{12}^{\mathrm{T}} & \hat{A}_3^{(2)} \\ & \hat{Y}_2 & \hat{X}_{12} & \hat{X}_{11} & \hat{A}_1^{(2)} \\ \hat{Y}_1^{\mathrm{T}} & \hat{A}_2^{(2)\mathrm{T}} & \hat{A}_3^{(2)\mathrm{T}} & \hat{A}_1^{(2)\mathrm{T}} & \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{x}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{s}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_1 \\ \mathbf{0} \end{bmatrix}.$$

It follows from this system that $\mathbf{s}_1 = 0$, and hence the computed solution $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \hat{\mathbf{s}}_3$ satisfies $\hat{\mathbf{s}}_1 = 0$ and the perturbed system of equations

$$\begin{bmatrix} & & \tilde{Y}_1 \\ & I_{p-n+s} & \tilde{A}_2^{(2)} \\ & \tilde{X}_{22} & \tilde{A}_3^{(2)} \\ \tilde{Y}_2 & \tilde{X}_{12} & \tilde{A}_1^{(2)} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{s}}_2 \\ \hat{\mathbf{s}}_3 \\ \hat{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_1 \end{bmatrix}, \tag{4.4}$$

where standard error analysis of triangular linear system solvers (Higham, 2002, Chapter 8) says that $\|\tilde{Y}_1 - \hat{Y}_1\|_2$, $\|\tilde{Y}_2 - \hat{Y}_2\|_2$, $\|\tilde{A}_1^{(2)} - \hat{A}_1^{(2)}\|_2$, $\|\tilde{A}_2^{(2)} - \hat{A}_2^{(2)}\|_2$, $\|\tilde{A}_3^{(2)} - \hat{A}_3^{(2)}\|_2$, $\|\tilde{X}_{22} - \hat{X}_{22}\|_2$ and $\|\tilde{X}_{12} - \hat{X}_{12}\|_2$ are all of the order of ϵ. Note that we did not include in this system the equation used to solve for $\hat{\lambda}$, since it does not affect the solution for $\mathbf{x}$ and $\mathbf{s}$. This is important, since the structure of (4.4) has no repeated blocks and only the constraint that $\tilde{X}_{22}$ ought to be symmetric; but it is shown in (Mackey *et al.*, 2008) that this is easy to satisfy with a similar backward error.

It thus follows from this discussion that the computed solution $(\hat{\mathbf{x}}, \hat{\mathbf{s}})$ corresponds exactly to the solution of a nearby problem with data $(\tilde{A}, \tilde{B}, \Sigma_{pq}, \tilde{\mathbf{b}}, \mathbf{d})$. Note that we cannot say this for the computed solution $\hat{\lambda}$. It is also important to note that Σ_{pq} and $\mathbf{b}$ do not get perturbed.

5. Sensitivity

In this section, we analyse the condition number of this problem. Since there is an orthogonal transfor-
200 mation between the system $M\mathbf{y} = \mathbf{f}$ and the permuted system $M_p\mathbf{y}_p = \mathbf{f}_p$:

$$\left[\begin{array}{c|ccc|c} \mathbf{0} & & \mathbf{0} & \mathbf{0} & \hat{Y}_2^{\mathrm{T}} \\ \hline & \mathbf{0} & & \hat{Y}_1 & \\ & \mathbf{0} & I_{p-n+s} & \hat{A}_2^{(2)} & \\ & \mathbf{0} & & \hat{X}_{22} \quad \hat{A}_3^{(2)} & \hat{X}_{12}^{\mathrm{T}} \\ & \hat{Y}_2 & & \hat{X}_{12} \quad \hat{A}_1^{(2)} & \hat{X}_{11} \\ \hline \hat{Y}_1^{\mathrm{T}} & & \hat{A}_2^{(2)^{\mathrm{T}}} & \hat{A}_3^{(2)^{\mathrm{T}}} & \hat{A}_1^{(2)^{\mathrm{T}}} \end{array}\right] \left[\begin{array}{c} \lambda \\ \hline \mathbf{x}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{x}_2 \\ \hline \mathbf{s}_1 \end{array}\right] = \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{d} \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_1 \\ \hline \mathbf{0} \end{array}\right], \tag{5.1}$$

we can as well look at that one. Note that we have displayed all sub-blocks that can be perturbed. The missing blocks are thus hard zeros that are not perturbed by our algorithm.

We first look at the effect of perturbation in $\mathbf{b}$ and $\mathbf{d}$ only. In this case, the zero blocks displayed in (5.1) remain zero as well, and we have thus $M_p\delta\mathbf{y}_p = \delta\mathbf{f}_p$. From these two equations we observe that
205 $\mathbf{s}_1 = 0$ and $\delta\mathbf{s}_1 = 0$. We then finally obtain the equations:

$$\hat{M}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{s}_2\\ \mathbf{s}_3\\ \mathbf{x}_2\end{bmatrix} = \begin{bmatrix}\mathbf{d}\\ \mathbf{b}_2\\ \mathbf{b}_3\\ \mathbf{b}_1\end{bmatrix}, \quad \hat{M}\begin{bmatrix}\delta\mathbf{x}_1\\ \delta\mathbf{s}_2\\ \delta\mathbf{s}_3\\ \delta\mathbf{x}_2\end{bmatrix} = \begin{bmatrix}\delta\mathbf{d}\\ \delta\mathbf{b}_2\\ \delta\mathbf{b}_3\\ \delta\mathbf{b}_1\end{bmatrix}, \tag{5.2}$$

for both the original data and the perturbation, where $\hat{M}$ is the submatrix

$$\hat{M} := \begin{bmatrix} & & & \hat{Y}_1 \\ & I_{p-n+s} & & \hat{A}_2^{(2)} \\ & & \hat{X}_{22} & \hat{A}_3^{(2)} \\ \hat{Y}_2 & & \hat{X}_{12} & \hat{A}_1^{(2)} \end{bmatrix}.$$

It immediately follows from this that

$$\left\|\begin{bmatrix}\delta\mathbf{x}_1\\ \delta\mathbf{s}_2\\ \delta\mathbf{s}_3\\ \delta\mathbf{x}_2\end{bmatrix}\right\|_2 \Big/ \left\|\begin{bmatrix}\mathbf{x}_1\\ \mathbf{s}_2\\ \mathbf{s}_3\\ \mathbf{x}_2\end{bmatrix}\right\|_2 \leqslant \|\hat{M}\|_2\|\hat{M}^{-1}\|_2 \left\|\begin{bmatrix}\delta\mathbf{d}\\ \delta\mathbf{b}_2\\ \delta\mathbf{b}_3\\ \delta\mathbf{b}_1\end{bmatrix}\right\|_2 \Big/ \left\|\begin{bmatrix}\mathbf{d}\\ \mathbf{b}_2\\ \mathbf{b}_3\\ \mathbf{b}_1\end{bmatrix}\right\|_2,$$

and since the vector $\mathbf{s}_1$ is always zero, we also obtain

$$\left\|\begin{bmatrix}\delta\mathbf{x}\\ \delta\mathbf{s}\end{bmatrix}\right\|_2 \Big/ \left\|\begin{bmatrix}\mathbf{x}\\ \mathbf{s}\end{bmatrix}\right\|_2 \leqslant \|\hat{M}\|_2\|\hat{M}^{-1}\|_2 \left\|\begin{bmatrix}\delta\mathbf{d}\\ \delta\mathbf{b}\end{bmatrix}\right\|_2 \Big/ \left\|\begin{bmatrix}\mathbf{d}\\ \mathbf{b}\end{bmatrix}\right\|_2.$$

We point out that since $\hat{M}$ and $\hat{M}^{-1}$ are submatrices of, respectively, M and M^{-1}, the condition number
210 of $\hat{M}$ can be no larger than M.

We now turn to the analysis of perturbations in M only, but this is a more complicated issue. From the equations $(M_p + \Delta M_p)(\mathbf{y}_p + \delta\mathbf{y}_p) = \mathbf{f}_p$ and $M_p\mathbf{y}_p = \mathbf{f}_p$, we obtain (up to second-order terms) the identity

$$M_p\delta\mathbf{y}_p + \Delta M_p\mathbf{y}_p = 0. \tag{5.3}$$

Selecting the middle block of equations from (5.3), we then obtain from (5.1) the identity (up to second-order terms)

$$\hat{M}\begin{bmatrix}\delta\mathbf{x}_1\\ \delta\mathbf{s}_2\\ \delta\mathbf{s}_3\\ \delta\mathbf{x}_2\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \hat{X}_{12}^{\mathrm{T}}\\ \hat{X}_{11}\end{bmatrix}\delta\mathbf{s}_1 + \Delta\hat{M}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{s}_2\\ \mathbf{s}_3\\ \mathbf{x}_2\end{bmatrix} = 0.$$

The vector $\delta\mathbf{s}_1$ can also be computed from (5.3) and yields

$$\delta\mathbf{s}_1 = -Y_2^{-\mathrm{T}}[\Delta_{11}, \Delta_{13}, \Delta_{14}]\begin{bmatrix}\lambda\\ \mathbf{s}_2\\ \mathbf{s}_3\end{bmatrix},$$

where Δ_{ij} are the nonzero blocks of ΔM_p, partitioned as in (5.1). Since

$$\hat{M}^{-1}\begin{bmatrix}0\\ 0\\ \hat{X}_{12}^{\mathrm{T}}\\ \hat{X}_{11}\end{bmatrix}Y_2^{-\mathrm{T}} = -\begin{bmatrix}I & 0 & 0 & 0\\ 0 & I & 0 & 0\\ 0 & 0 & I & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\tilde{M}^{-1}\begin{bmatrix}I\\ 0\\ 0\\ 0\end{bmatrix}, \quad \tilde{M} := \begin{bmatrix} & I_{p-n+s} & & Y_2^{\mathrm{T}}\\ & & \hat{X}_{22} & \hat{X}_{12}^{\mathrm{T}}\\ Y_2 & & \hat{X}_{12} & \hat{X}_{11}\end{bmatrix},$$

this then yields the bound

$$\left\|\begin{bmatrix}\delta\mathbf{x}\\ \delta\mathbf{s}\end{bmatrix}\right\|_2 \leqslant \|\hat{M}^{-1}\|_2\left\|\Delta M\begin{bmatrix}0\\ \mathbf{s}\\ \mathbf{x}\end{bmatrix}\right\|_2 + \|\tilde{M}^{-1}\|_2\left\|\Delta M\begin{bmatrix}\lambda\\ \mathbf{s}\\ 0\end{bmatrix}\right\|_2. \tag{5.4}$$

It is obvious to see that, in general, $\|\hat{M}^{-1}\|_2, \|\tilde{M}^{-1}\|_2 \leqslant \|M^{-1}\|_2$, which indicates that the ILSE problem is less sensitive than the problem of solving the linear system $M\mathbf{y} = \mathbf{f}$. In order to better understand the difference between both terms in the above expression, let us look at the case when $\Sigma_{pq} = I$. Then, it easily follows that $\hat{X}_{22}$ and $\hat{X}_{12}$ are void and $\hat{X}_{11} = I_p$, and hence

$$\|\hat{M}^{-1}\|_2 = \left\|\begin{bmatrix}0 & 0 & \hat{Y}_1\\ 0 & I & \hat{A}_2^{(2)}\\ \hat{Y}_2 & 0 & \hat{A}_1^{(2)}\end{bmatrix}^{-1}\right\|_2 \approx \left\|\begin{bmatrix}0 & \hat{Y}_1\\ \hat{Y}_2 & \hat{A}_1^{(2)}\end{bmatrix}^{-1}\right\|_2 \approx \|\hat{Y}_2^{-1}\hat{A}_1^{(2)}\hat{Y}_1^{-1}\|_2,$$

$$\|\tilde{M}^{-1}\|_2 = \left\|\begin{bmatrix}0 & 0 & \hat{Y}_2^{\mathrm{T}}\\ 0 & I & 0\\ \hat{Y}_2 & 0 & I\end{bmatrix}^{-1}\right\|_2 \approx \left\|\begin{bmatrix}0 & \hat{Y}_2^{\mathrm{T}}\\ \hat{Y}_2 & I\end{bmatrix}^{-1}\right\|_2 \approx \|\hat{Y}_2^{-1}\hat{Y}_2^{-\mathrm{T}}\|_2.$$

If, moreover, there are no constraints, then $\hat{Y}_1$, $\hat{A}_1^{(2)}$ and $\hat{A}_2^{(2)}$ are also void and these respective matrix norms become

$$\|\hat{M}^{-1}\|_2 \approx \|Y_2^{-1}\|_2, \quad \|\tilde{M}^{-1}\|_2 \approx \|Y_2^{-1}Y_2^{-1}\|_2.$$

We can recognize here the two components of the sensitivity analysis of the classical linear least squares problem as analysed, e.g., in Golub & Van Loan (1996) and in Elden (1980). Note also that when the residual s is nearly zero, then also λ is zero, and the second term of the sensitivity bound disappears. This was also observed by Grcar (2011) who showed that the conditioning problem of our problem is much better when the residual vector is nearly zero (i.e., when the systems are nearly consistent).

6. Numerical experiments

In this section, we report the results of some numerical experiments performed in `Matlab©R2010b` with machine precision $\varepsilon \approx 2.2 \times 10^{-16}$. For each example, the matrix $B \in \mathbb{R}^{s\times n}$, given its condition number[1] κ_B, is constructed by using the matlab command $B = \kappa_B \times$ `gallery('randsvd',`$[s,n],\kappa_B$`)`, so that $\|B\|_2 = \kappa_B$ and its singular values are geometrically distributed between κ_B and 1. Moreover, Q6
given κ_A, $A = Q_{\Sigma_{pq}} D U Q_B^{\mathrm{T}}$, with $Q_{\Sigma_{pq}} \in \mathbb{R}^{(p+q)\times(p+q)}$ Σ_{pq}-orthogonal, i.e., such that $Q_{\Sigma_{pq}}^{\mathrm{T}} \Sigma_{pq} Q_{\Sigma_{pq}} = \Sigma_{pq}$, generated by the method described in Higham (2003), $D \in \mathbb{R}^{(p+q)\times n}$ a diagonal matrix with decreasing diagonal values geometrically distributed between κ_A and 1, $U \in \mathbb{R}^{n\times n}$ random orthogonal generated by the function `gallery('qmult',`n`)`, and with $Q_B \in \mathbb{R}^{n\times n}$ the orthogonal factor of the right lower antitriangular factorization (3.3) of B, $B = [\mathbf{0},\ Y_1]Q_B^{\mathrm{T}}$, so that $\kappa(A) \approx \|A\| \approx \kappa_A$. This construction guarantees that $A^{\mathrm{T}}\Sigma_{pq}A > 0$. The solution $\mathbf{x}$ of problem (1.1), depending on a parameter $c_1 \in \mathbb{R}$, is chosen as $\mathbf{x} = Q_B\mathbf{v}_1$, with $\mathbf{v}_1 = c_1 \times$ `randn`$(n,1)$. Furthermore, partitioning Q_B and Q7
A as in (3.4) and (3.6), respectively, $Q_B = [Q_{B1},\ Q_{B2}]$, with $Q_{B1} \in \mathbb{R}^{n\times(n-s)}$ spanning the nullspace of B, $A_1 = AQ_{B1} \in \mathbb{R}^{(p+q)\times(n-s)}$, $A_2 = AQ_{B2} \in \mathbb{R}^{(p+q)\times s}$, the vector $\mathbf{s}$, depending on a parameter $c_2 \in \mathbb{R}$, is chosen as and $\mathbf{s} = c_2 V_2 \times$ `randn`$(s,1)$, where $V_2 \in \mathbb{R}^{(p+q)\times s}$ has orthogonal columns spanning the nullspace of A_1. Hence, $\mathbf{d} = Y_1\mathbf{v}_1(n-s+1:n)$, $\mathbf{b} = \Sigma_{pq}\mathbf{s} + AQ_B^{\mathrm{T}}\mathbf{v}_1$ and $\lambda = -Y_1^{\mathrm{T}}\backslash(A_2^{\mathrm{T}}\mathbf{s})$. Then, the solution of the augmented system (3.1) is given by $\mathbf{y} = [\lambda^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{x}^{\mathrm{T}}]^{\mathrm{T}}$. For all the experiments, we choose $n = 50$, $s = 20$, $p = 60$, $q = 40$.

Each set of experiments consists of 16 runs; in each of these κ_A and κ_B are taken from the set $\{1e+1, 1e+2, 1e+4, 1e+8\}$. The solution computed by the proposed method, denoted by $\mathbf{x}_{MV}$, is compared to the one computed by `matlab` using the command '\' applied to (1.2), denoted by $\mathbf{x}_B$, to the solutions yielded by the methods `GQR-Cholesky` and `GHQR` described in Bojanczyk *et al.* (2003a), denoted by $\mathbf{x}_{BHP1}$ and $\mathbf{x}_{BHP2}$, respectively, and to the solution computed by the method proposed in Mastronardi & Van Dooren (2013a) and denoted by x_{MVB}. Moreover, we add an 'S' to the name of Q8
the considered methods when they are used to solve the corresponding scaled augmented system (2.3). Moreover, the corresponding scaled matrix is denoted by M_S.

[1] The condition number of a rectangular matrix $A \in \mathbb{R}^{m\times n}$, $m \geqslant n$, $\mathrm{rank}(A) = n$, is defined as $\sigma_{\max}(A)/\sigma_{\min}(A)$ (Golub & Van Loan, 1996).

We denote by

$$r(\mathbf{w}) = \frac{\left\| \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \end{bmatrix} - \begin{bmatrix} 0 & B \\ \Sigma_{pq} & A \end{bmatrix} \mathbf{w} \right\|_2}{\|w\|_2},$$

the 'restricted' relative residual, where $\mathbf{w} = [\mathbf{s}^{\mathrm{T}}\ \mathbf{x}^{\mathrm{T}}]^{\mathrm{T}}$.

Since the `GQR-Cholesky` and `GHQR` methods compute only the solution $\mathbf{x}$ of (1.1), then the missing $\mathbf{s}$ is computed as $\mathbf{s} = \Sigma_{pq}(\mathbf{b} - A\mathbf{x})$ for these two methods. Q9

6.1 *First set*

For this set of experiments, we choose $c_1 = c_2 = 1$.

In columns 2, 3, 4 and 5 of Table 2, the condition numbers of A, B, M and M_S are displayed, respectively. The condition numbers of the matrices $\hat{M}, \hat{M}_S, \tilde{M}$ and $\tilde{M}_S$ are displayed in columns 6, 7, 8 and 9, respectively. We can observe that the condition number of the scaled matrix M_S is often smaller than the condition number of matrix M. This also holds for the condition numbers of the matrices $\hat{M}_S$ and $\hat{M}$.

In Table 3, the restricted relative residuals of the vector $\mathbf{w}$ computed by the considered methods are reported. It can be observed that solving the scaled problem using the `matlab` '\' gives more accurate results than '\' without scaling. Moreover, the considered methods have a similar behaviour and this is independent of the use of scaling since these methods work on the individual blocks of M.

In Table 4, the relative residuals of the $\mathbf{w}$ and $\mathbf{y}$ computed by the proposed method are reported in columns 2 and 3, respectively. Since c_1 and c_2 are of the same size, the vectors $\mathbf{s}$ and $\mathbf{x}$ have approximately the same norm. Therefore, in this case we have that the error of the computed solution depends on both $\|\hat{M}_S\|$ and $\|\tilde{M}_S\|$, as shown by (5.4). In Table 5, columns 2, 3, 4, 5 and 6, the relative error of the solutions computed by the considered methods are displayed. These experiments show that the proposed method has often a smaller relative error.

6.2 *Second set*

For this second set of experiments, we choose $c_1 = 1e10$, and $c_2 = 1$. We omit a table similar to Table 1 because the construction of M does not depend on the coefficients c_1 and c_2.

We can see in Table 6 that the backslash of `matlab` now behaves even worse when using the unscaled matrix (see also the discussion at the end of Section 4) (Table 7). Q10

Since $c_1 \gg c_2$, it follows that $\|\mathbf{x}\| \gg \|\mathbf{s}\|$. Therefore, taking (5.4) into account, the error in the computed solution depends more on $\|\hat{M}_S\|$. This can be clearly seen in comparing columns 2, 3, 4 and 5 of Table 8.

6.3 *Third set*

For this third set of experiments, we choose $c_1 = 1$, and $c_2 = 1e10$. Also in this case, we omit a table similar to Table 2 because the construction of M does not depend on the coefficients c_1 and c_2 (Table 9). Q10

Since $c_2 \gg c_1$, it follows that $\|\mathbf{s}\| \gg \|\mathbf{x}\|$. Therefore, taking (5.4) into account, the error in the computed solution depends more on $\|\tilde{M}_S\|$. This can be clearly seen in comparing columns 2, 3, 4 and 5 of Tables 10 and 11. Q10

TABLE 2 *Condition numbers of the matrix of the augmented system, of the scaled one and of their related submatrices*

#	$\kappa(A)$	$\kappa(B)$	$\kappa(M)$	$\kappa(M_S)$	$\kappa(\hat{M})$	$\kappa(\hat{M}_S)$	$\kappa(\tilde{M})$	$\kappa(\tilde{M}_S)$
1	4.8e+01	1.0e+01	3.0e+03	1.8e+03	3.0e+03	2.0e+02	2.0e+03	1.4e+03
2	2.3e+02	1.0e+01	4.6e+04	1.5e+04	1.9e+04	6.2e+02	1.5e+04	1.3e+04
3	2.3e+04	1.0e+01	3.9e+08	8.6e+06	5.4e+06	6.1e+03	3.6e+05	7.4e+06
4	3.5e+08	1.0e+01	5.2e+15	9.2e+11	4.7e+12	4.2e+06	2.0e+11	8.1e+11
5	2.5e+01	1.0e+02	1.6e+03	5.3e+02	1.6e+03	2.2e+02	4.0e+02	4.9e+02
6	1.6e+02	1.0e+02	1.8e+04	9.4e+03	3.3e+03	3.9e+02	1.9e+03	8.5e+03
7	1.5e+04	1.0e+02	6.8e+07	5.2e+06	2.1e+06	4.4e+03	3.6e+05	4.6e+06
8	2.1e+08	1.0e+02	6.3e+15	4.8e+11	3.8e+12	1.8e+06	8.6e+09	4.4e+11
9	2.2e+01	1.0e+04	1.4e+05	1.7e+06	1.3e+05	2.2e+04	3.0e+02	5.0e+02
10	7.1e+02	1.0e+04	1.8e+06	1.6e+05	1.8e+06	3.2e+04	1.2e+05	1.4e+05
11	1.4e+04	1.0e+04	3.2e+07	6.1e+06	1.2e+06	4.0e+04	2.3e+05	5.6e+06
12	1.6e+08	1.0e+04	4.2e+14	1.5e+11	4.9e+11	6.0e+05	1.5e+09	1.3e+11
13	2.5e+01	1.0e+08	2.0e+09	1.1e+14	2.0e+09	1.8e+08	5.0e+02	5.3e+02
14	4.4e+02	1.0e+08	9.3e+09	4.4e+12	9.2e+09	4.2e+08	3.1e+04	3.9e+04
15	2.5e+04	1.0e+08	4.3e+11	6.1e+10	1.0e+10	6.2e+08	1.0e+06	6.4e+06
16	1.8e+08	1.0e+08	9.2e+14	6.6e+11	9.2e+11	5.9e+08	8.0e+09	5.9e+11

TABLE 3 *Relative residuals* ($c_1 = c_2$)

#	$r(\mathbf{x}_B)$	$r(\mathbf{x}_{BS})$	$r(\mathbf{x}_{MVS})$	$r(\mathbf{x}_{BHP1S})$	$r(\mathbf{x}_{BHP2S})$	$r(\mathbf{x}_{MVBS})$
1	5.741e−15	1.79e−16	5.43e−16	2.43e−16	2.25e−16	4.49e−16
2	3.878e−14	2.36e−16	5.44e−16	2.05e−16	2.38e−16	4.45e−16
3	2.912e−14	1.95e−16	4.46e−16	2.07e−16	2.28e−16	3.45e−16
4	1.113e−12	2.61e−16	4.71e−16	2.88e−16	2.39e−16	4.19e−16
5	3.821e−15	2.92e−16	5.67e−16	2.17e−16	2.77e−16	5.48e−16
6	1.629e−14	3.17e−16	5.69e−16	1.85e−16	2.36e−16	5.42e−16
7	1.136e−13	2.48e−16	5.36e−16	2.48e−16	2.51e−16	5.57e−16
8	5.817e−12	2.18e−16	4.12e−16	1.94e−16	1.97e−16	4.07e−16
9	5.727e−14	1.14e−14	4.64e−16	2.11e−16	2.58e−16	4.41e−16
10	8.307e−15	1.87e−16	4.40e−16	2.02e−16	2.04e−16	4.85e−16
11	5.011e−14	2.49e−16	5.07e−16	2.34e−16	2.18e−16	4.32e−16
12	4.201e−13	2.05e−16	5.19e−16	2.07e−16	2.35e−16	6.17e−16
13	2.534e−10	2.61e−11	4.70e−16	2.61e−16	2.18e−16	4.39e−16
14	6.324e−12	1.12e−12	4.81e−16	2.02e−16	2.18e−16	4.41e−16
15	1.434e−13	5.15e−14	7.61e−16	2.09e−16	2.55e−16	5.34e−16
16	2.926e−12	7.33e−16	3.90e−16	1.54e−16	2.06e−16	4.08e−16

TABLE 4 *Relative residuals and 'modified' condition numbers of* $\kappa(\hat{M}_S)$ *and* $\kappa(\tilde{M}_S)$

#	$\frac{\|\mathbf{w}-\mathbf{w}_{MVS}\|}{\|\mathbf{w}\|}$	$\frac{\|\mathbf{y}-\mathbf{y}_{MVS}\|}{\|\mathbf{y}\|}$	$\varepsilon \times \kappa(\hat{M}_S)$	$\varepsilon \times \kappa(\tilde{M}_S)$
1	1.48e−14	1.47e−14	4.45e−14	3.26e−13
2	1.27e−13	1.27e−13	1.40e−13	3.07e−12
3	1.08e−11	1.08e−11	1.35e−12	1.66e−09
4	1.03e−06	1.03e−06	9.34e−10	1.81e−04
5	5.68e−15	3.42e−15	4.99e−14	1.10e−13
6	3.18e−14	2.42e−14	8.69e−14	1.90e−12
7	9.79e−12	9.79e−12	9.92e−13	1.03e−09
8	9.02e−07	9.02e−07	4.19e−10	9.80e−05
9	2.11e−13	1.31e−13	5.04e−12	1.13e−13
10	3.26e−13	3.63e−14	7.28e−12	3.20e−11
11	2.05e−11	7.40e−12	9.06e−12	1.25e−09
12	1.06e−07	1.06e−07	1.35e−10	2.92e−05
13	9.73e−10	5.60e−10	4.02e−08	1.19e−13
14	4.41e−10	2.84e−11	9.44e−08	8.71e−12
15	7.10e−10	3.70e−12	1.40e−07	1.42e−09
16	5.25e−07	3.30e−09	1.33e−07	1.31e−04

TABLE 5 *Relative errors of the solutions*

#	$\frac{\|\mathbf{x}_{BS}-\mathbf{x}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVS}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVBS}\|}{\|\mathbf{x}\|}$
1	2.84e−14	1.64e−14	2.16e−14	1.73e−14	2.58e−14
2	3.48e−13	1.45e−13	1.30e−13	1.05e−13	3.74e−13
3	6.44e−11	1.20e−11	1.73e−11	7.96e−12	9.22e−11
4	2.03e−06	1.16e−06	2.08e−06	2.68e−06	5.22e−06
5	2.88e−14	6.99e−15	2.39e−14	2.33e−14	1.10e−14
6	1.63e−13	3.80e−14	8.68e−14	9.38e−14	4.04e−13
7	7.35e−11	1.26e−11	5.06e−12	1.13e−11	5.98e−11
8	2.30e−06	1.04e−06	1.83e−06	2.90e−06	5.63e−06
9	2.69e−12	2.38e−13	1.64e−12	1.64e−12	2.39e−13
10	9.57e−12	3.88e−13	1.19e−11	1.20e−11	1.66e−12
11	2.39e−11	2.34e−11	4.49e−11	5.09e−11	8.57e−11
12	7.73e−07	1.33e−07	5.82e−07	4.67e−07	4.03e−07
13	1.22e−05	1.15e−09	9.61e−09	9.61e−09	1.15e−09
14	4.33e−05	5.24e−10	3.08e−07	3.08e−07	5.24e−10
15	2.57e−06	8.76e−10	5.45e−07	5.45e−07	8.56e−10
16	8.79e−04	5.92e−07	1.15e−03	1.15e−03	7.47e−06

TABLE 6 *Relative residuals* ($c_1 \gg c_2$)

#	$r(\mathbf{x}_B)$	$r(\mathbf{x}_{BS})$	$r(\mathbf{x}_{MVS})$	$r(\mathbf{x}_{BHP1S})$	$r(\mathbf{x}_{BHP2S})$	$r(\mathbf{x}_{MVBS})$
1	4.96e−15	1.27e−16	1.68e−16	1.82e−16	1.68e−16	2.36e−16
2	3.20e−14	1.96e−16	2.39e−16	2.24e−16	1.93e−16	4.04e−16
3	1.17e−12	1.43e−16	2.02e−16	1.47e−16	1.44e−16	2.31e−16
4	2.23e−08	1.70e−16	2.32e−16	2.24e−16	1.71e−16	2.83e−16
5	1.35e−14	1.83e−16	2.16e−16	1.22e−16	1.02e−16	2.99e−16
6	3.15e−14	2.40e−16	2.30e−16	1.15e−16	1.36e−16	2.58e−16
7	1.66e−12	1.73e−16	1.95e−16	1.11e−16	9.85e−17	2.91e−16
8	1.25e−08	1.89e−16	1.90e−16	1.19e−16	1.62e−16	2.17e−16
9	4.38e−13	1.62e−16	2.57e−16	6.28e−17	7.50e−17	2.98e−16
10	6.11e−13	1.44e−16	2.51e−16	1.11e−16	8.41e−17	2.26e−16
11	2.45e−12	1.28e−16	1.88e−16	7.12e−17	6.60e−17	1.87e−16
12	1.61e−08	1.02e−16	1.37e−16	7.13e−17	6.16e−17	1.83e−16
13	5.60e−09	2.08e−16	3.21e−16	9.39e−17	1.49e−16	2.95e−16
14	2.46e−09	1.75e−16	2.60e−16	6.48e−17	7.33e−17	2.64e−16
15	5.18e−09	1.34e−16	1.22e−16	7.01e−17	5.59e−17	2.12e−16
16	1.55e−08	4.48e−17	9.02e−17	8.08e−17	4.45e−17	1.92e−16

TABLE 7 *Relative residuals and 'modified' condition numbers of* $\kappa(\hat{M}_S)$ *and* $\kappa(\tilde{M}_S)$

#	$\frac{\|\mathbf{w}-\mathbf{w}_{MVS}\|}{\|\mathbf{w}\|}$	$\frac{\|\mathbf{y}-\mathbf{y}_{MVS}\|}{\|\mathbf{y}\|}$	$\varepsilon \times \kappa(\hat{M}_S)$	$\varepsilon \times \kappa(\tilde{M}_S)$
1	1.34e−15	1.34e−15	3.60e−14	1.97e−13
2	8.99e−15	8.99e−15	1.11e−13	5.96e−12
3	5.10e−14	5.10e−14	7.27e−13	6.00e−10
4	4.58e−12	4.58e−12	1.06e−10	1.87e−05
5	2.32e−15	2.59e−15	4.42e−14	1.25e−13
6	1.16e−14	1.18e−14	9.06e−14	1.15e−12
7	2.31e−14	2.31e−14	4.14e−13	2.40e−10
8	1.25e−11	1.25e−11	2.21e−10	2.09e−05
9	2.16e−14	1.36e−12	4.77e−12	4.93e−14
10	1.54e−13	4.13e−13	7.78e−12	4.19e−12
11	4.56e−13	4.56e−13	1.38e−11	8.44e−10
12	1.04e−11	1.04e−11	4.15e−10	9.40e−05
13	2.28e−10	3.25e−04	4.62e−08	7.11e−14
14	7.08e−11	6.80e−06	5.33e−08	1.20e−12
15	1.03e−09	2.33e−07	1.02e−07	1.67e−09
16	8.62e−10	8.62e−10	1.06e−07	3.04e−04

TABLE 8 *Relative errors of the solutions*

#	$\frac{\|\mathbf{x}_{BS}-\mathbf{x}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVS}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVBS}\|}{\|\mathbf{x}\|}$
1	1.52e−14	1.34e−15	1.64e−15	1.73e−15	5.99e−15
2	1.40e−13	8.83e−15	1.12e−14	4.48e−15	7.36e−13
3	4.60e−11	5.10e−14	5.07e−14	3.60e−14	3.40e−11
4	1.77e−06	4.58e−12	1.31e−11	6.37e−12	5.01e−07
5	6.24e−15	2.30e−15	4.77e−15	4.39e−15	6.97e−15
6	1.13e−13	1.16e−14	1.41e−14	1.42e−14	3.68e−14
7	1.99e−12	2.31e−14	3.58e−14	3.71e−14	7.97e−12
8	6.95e−07	1.25e−11	1.60e−11	2.01e−11	3.71e−07
9	1.18e−13	2.13e−14	2.05e−13	2.05e−13	2.13e−14
10	3.06e−13	1.54e−13	3.72e−13	3.69e−13	1.99e−13
11	1.53e−11	4.56e−13	7.25e−13	7.67e−13	1.39e−11
12	2.10e−06	1.04e−11	2.10e−11	1.08e−11	2.01e−06
13	3.59e−09	2.22e−10	3.91e−09	3.91e−09	2.22e−10
14	1.37e−09	7.08e−11	5.66e−10	5.66e−10	7.08e−11
15	2.45e−09	1.03e−09	2.20e−09	2.20e−09	1.04e−09
16	7.32e−07	8.62e−10	1.23e−09	1.25e−09	6.49e−06

TABLE 9 *Relative residuals* ($c_1 \ll c_2$)

#	$r(\mathbf{x}_B)$	$r(\mathbf{x}_{BS})$	$r(\mathbf{x}_{MVS})$	$r(\mathbf{x}_{BHP1S})$	$r(\mathbf{x}_{BHP2S})$	$r(\mathbf{x}_{MVBS})$
1	3.664e−015	4.50e−016	8.11e−016	3.91e−016	3.06e−016	7.19e−016
2	1.575e−014	4.24e−016	9.48e−016	2.80e−016	3.10e−016	8.32e−016
3	5.872e−014	3.31e−016	7.54e−016	2.60e−016	3.27e−016	6.40e−016
4	1.730e−012	4.59e−016	7.90e−016	3.73e−016	2.57e−016	6.54e−016
5	9.230e−015	3.84e−016	8.21e−016	3.22e−016	3.62e−016	7.77e−016
6	2.798e−014	3.90e−016	1.01e−015	3.30e−016	3.35e−016	7.97e−016
7	3.893e−014	5.87e−016	9.34e−016	4.18e−016	3.07e−016	8.58e−016
8	9.193e−013	3.13e−016	8.64e−016	3.07e−016	3.59e−016	8.09e−016
9	3.179e−015	6.63e−015	8.30e−016	3.76e−016	4.76e−016	8.58e−016
10	7.444e−015	6.43e−015	8.95e−016	2.78e−016	2.91e−016	6.82e−016
11	1.406e−014	3.24e−016	7.58e−016	3.22e−016	3.55e−016	7.60e−016
12	3.054e−012	3.42e−016	1.02e−015	3.55e−016	2.94e−016	7.68e−016
13	3.665e−015	6.54e−011	8.39e−016	2.89e−016	2.99e−016	7.46e−016
14	1.084e−014	4.21e−011	6.91e−016	2.32e−016	4.44e−016	7.68e−016
15	9.280e−014	3.18e−012	8.82e−016	3.73e−016	2.56e−016	6.82e−016
16	7.577e−013	3.45e−016	8.27e−016	2.76e−016	3.65e−016	8.68e−016

TABLE 10 *Relative residuals and 'modified' condition numbers of $\kappa(\hat{M}_S)$ and $\kappa(\tilde{M}_S)$*

#	$\frac{\|\mathbf{w}-\mathbf{w}_{MVS}\|}{\|\mathbf{w}\|}$	$\frac{\|\mathbf{y}-\mathbf{y}_{MVS}\|}{\|\mathbf{y}\|}$	$\varepsilon \times \kappa(\hat{M}_S)$	$\varepsilon \times \kappa(\tilde{M}_S)$
1	4.50e−015	3.84e−015	6.83e−015	4.40e−014
2	9.21e−014	9.02e−014	3.51e−014	9.75e−013
3	4.02e−011	4.02e−011	9.55e−013	7.06e−010
4	2.18e−006	2.18e−006	2.38e−010	8.31e−005
5	2.02e−014	6.96e−015	4.40e−014	2.57e−013
6	3.41e−013	1.70e−013	1.38e−013	4.30e−012
7	1.59e−011	1.58e−011	4.94e−013	3.60e−010
8	5.16e−006	5.16e−006	5.94e−010	2.49e−004
9	1.73e−014	9.10e−016	4.82e−012	2.29e−013
10	1.48e−013	4.33e−015	5.65e−012	1.75e−012
11	2.10e−011	1.09e−011	8.35e−012	1.50e−009
12	3.66e−006	3.65e−006	2.68e−010	9.10e−005
13	1.52e−014	1.05e−015	4.74e−008	7.45e−014
14	6.22e−014	1.71e−015	6.30e−008	2.10e−012
15	5.20e−011	4.30e−014	6.04e−008	1.10e−009
16	5.18e−007	7.65e−009	2.17e−007	1.43e−004

TABLE 11 *Relative errors of the solutions*

#	$\frac{\|\mathbf{x}_{BS}-\mathbf{x}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVS}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{BHP1S}\|}{\|\mathbf{x}\|}$	$\frac{\|\mathbf{x}-\mathbf{x}_{MVBS}\|}{\|\mathbf{x}\|}$
1	3.31e−005	2.27e−005	6.18e−005	5.31e−005	4.14e−005
2	9.26e−004	7.72e−004	6.21e−004	8.78e−004	8.50e−004
3	1.34e−001	2.27e−001	1.99e−001	7.75e−002	3.29e−001
4	4.59e+004	1.49e+004	6.37e+003	1.33e+004	9.99e+003
5	3.97e−004	1.49e−004	6.15e−004	6.72e−004	1.06e−004
6	7.13e−003	2.86e−003	6.59e−003	7.58e−003	2.67e−003
7	4.28e−002	7.51e−002	2.70e−002	9.20e−003	5.07e−002
8	2.81e+003	2.83e+004	2.10e+004	9.00e+002	5.33e+004
9	1.29e−001	1.48e−004	1.93e−002	1.93e−002	2.49e−004
10	3.30e−002	5.69e−004	7.96e−003	8.03e−003	1.14e−003
11	4.08e−001	1.44e−001	1.43e+000	1.43e+000	2.28e+000
12	7.29e+003	2.31e+004	1.05e+003	3.58e+003	1.30e+004
13	1.71e+006	9.56e−005	4.42e+001	4.42e+001	1.13e−004
14	1.20e+007	3.39e−004	8.98e+002	8.98e+002	3.58e−004
15	1.45e+005	3.09e−001	2.22e+004	2.22e+004	2.77e−001
16	1.56e+006	3.10e+003	1.69e+006	1.68e+006	9.02e+004

7. Conclusions

An algorithm for computing the solution of indefinite least squares problems with equality constrained is described in the paper. The proposed algorithm is related to the so-called `GQR-Cholesky` method proposed in Bojanczyk *et al.* (2003a), even though it turns out to be slightly more accurate and efficient. Moreover, the proof of the backward numerical stability is given in a strict sense. A sensitivity analysis, bringing a new insight in the problem, is also given.

Funding

N.M. was partly supported by the GNCS INdAM project 'Strategie risolutive per sistemi lineari di tipo KKT con uso di informazioni strutturali'. P.V.D. was partly supported by the Belgian Network DYSCO (Dynamical Systems, Control and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with the authors.

References

Bojanczyk, A., Higham, N. J. & Patel, H. (2003a) The equality constrained indefinite least squares problem: theory and algorithms. *BIT*, **43**, 505–517.

Bojanczyk, A., Higham, N. J. & Patel, H. (2003b) Solving the indefinite least squares problem by hyperbolic QR factorization. *SIAM J. Matrix Anal. Appl.*, **24**, 914–931.

Chandrasekaran, S., Gu, M. & Sayed, A. H. (1998) A stable and efficient algorithm for the indefinite linear least squares problem. *SIAM J. Matrix Anal. Appl.*, **20**, 354–362.

Elden, L. (1980) Perturbation theory for the least squares problem with linearity constraints. *SIAM J. Numer. Anal.*, **17**, 338–350.

Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computations*, 3rd edn. Baltimore: Johns Hopkins University Press.

Gould, N. I. M. (1985) On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem. *Math. Program.*, **32**, 90–99.

Grcar, J. F. (2011) Unattainability of a perturbation bound for indefinite least squares problems. Preprint, arXiv: 1004.4921v5.

Higham, N. J. (2002) *Accuracy and Stability of Numerical Algorithms*, 2nd edn. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Higham, N. J. (2003) J-orthogonal matrices: properties and generation. *SIAM Rev.*, **45**, 504–519.

Liu, Q., Pan, B. & Wang, Q. (2010) The hyperbolic elimination method for solving the equality constrained indefinite least squares problem. *Int. J. Comput. Math.*, **87**, 2953–2966.

Liu, Q. & Wang, M. (2010) Algebraic properties and perturbation results for the indefinite least squares problem with equality constraints. *Int. J. Comput. Math.*, **87**, 425–434.

Mackey, S., Mackey, N. & Tisseur, F. (2008) Structured mapping problems for matrices associated with scalar products part I: Lie and Jordan algebras. *SIAM J. Matrix Anal. Appl.*, **29**, 1389–1410.

Mastronardi, N. & Van Dooren, P. (2012) Recursive approximation of the dominant eigenspace of an indefinite matrix. *J. Comput. Appl. Math.*, **236**, 4090–4104.

Mastronardi, N. & Van Dooren, P. (2013a) An algorithm for solving the indefinite least squares problem with equality constraints. *BIT*. doi:10.1007/s10543-013-0452-2 (in press).

Mastronardi, N. & Van Dooren, P. (2013b) The antitriangular factorization of symmetric matrices. *SIAM J. Matrix Anal. Appl.*, **34**, 173–196.

Appendix

A.1 *Congruence transformation lemma*

LEMMA A.1 Let $M := \Sigma + \Delta$ be an ϵ-perturbation of $\Sigma := \mathrm{diag}\{I_p, -I_q\}$; then there exists a congruence transformation restoring the original matrix: $\Sigma = (I + C)(\Sigma + \Delta)(I + C)^{\mathrm{T}}$ where $(I + C)$ is an ϵ-perturbation of the identity.

Proof. Let the matrices M and C be denoted by

$$M := \begin{bmatrix} I_p + \Delta_{11} & \Delta_{12} \\ \Delta_{12}^{\mathrm{T}} & -I_q + \Delta_{22} \end{bmatrix}, \quad C := \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

If the symmetric perturbation matrix Δ has 2-norm $c\epsilon \ll 1$, then an approximate solution is given by $C_0 = -\Delta\Sigma/2$ since

$$(I + C_0)(\Sigma + \Delta)(I + C_0)^{\mathrm{T}} = \Sigma - \tfrac{3}{4}\Delta\Sigma\Delta + \tfrac{1}{4}\Delta\Sigma\Delta\Sigma\Delta,$$

which is $\mathcal{O}(\epsilon^2)$-close to the solution. Since the scalar function $(1 + \delta)^{-1/2}$ has a converging Taylor series for $\delta < 1$, one then shows that C has a converging Taylor series in powers of $\Delta\Sigma$ of which C_0 is the first term. It then follows that $\|C\|_2 = (c/2)\epsilon + \mathcal{O}(\epsilon^2)$, which completes the proof. □

A.2 *Case $q > n - s$*

Here we describe an algorithm to handle the case $q > n - s$. Partition A_1 as follows:

$$A_1 = \begin{bmatrix} A_{11} \\ A_{12} \end{bmatrix} \begin{matrix} \}p \\ \}q \end{matrix},$$

and $\mathbf{y}^{(1)}$ and $\mathbf{f}^{(1)}$ as

$$\mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{s}_4 \\ \mathbf{x}_1 \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \\ \}n-s \\ \}q-n+s \\ \}n-s \end{matrix}, \quad \mathbf{f}^{(1)} = \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \\ \hat{\mathbf{b}}_3 \\ \hat{\mathbf{b}}_4 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \\ \}n-s \\ \}q-n+s \\ \}n-s \end{matrix}.$$

Compute the left upper antitriangular factorization of A_{11} and A_{12}:

$$A_{11} = Q_p \begin{bmatrix} \tilde{Y}_2 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}p-n+s \end{matrix}, \quad A_{12} = Q_q \begin{bmatrix} \tilde{Y}_3 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}q-n+s \end{matrix}, \tag{A.1}$$

with $\tilde{Y}_2 \in \mathbb{R}^{(n-s)\times(n-s)}$ nonsingular upper antitriangular, $\tilde{Y}_3 \in \mathbb{R}^{(n-s)\times(n-s)}$ upper antitriangular and $Q_p \in \mathbb{R}^{p\times p}$ $Q_q \in \mathbb{R}^{q\times q}$ orthogonal, and let

$$Q^{(2)} = \begin{bmatrix} Q_p & & \\ & Q_q & \\ & & I_{n-s} \end{bmatrix}.$$

Then, the linear system (3.8) is transformed into the following equivalent one:

$$\tilde{M}^{(1)}\tilde{\mathbf{y}}^{(1)} = \tilde{\mathbf{f}}^{(1)}, \tag{A.2}$$

where

$$\tilde{M}^{(1)} = {Q^{(2)}}^{\mathrm{T}} M^{(1)} Q^{(2)} = \left[\begin{array}{c|c|c|c|c} I_{n-s} & & & & \tilde{Y}_2 \\ \hline & I_{p-n+s} & & & \mathbf{0} \\ \hline & & -I_{n-s} & & \tilde{Y}_3 \\ \hline & & & -I_{q-n+s} & \mathbf{0} \\ \hline \tilde{Y}_2^{\mathrm{T}} & \mathbf{0} & \tilde{Y}_3^{\mathrm{T}} & \mathbf{0} & \end{array}\right],$$

$$\tilde{\mathbf{y}}^{(1)} = {Q^{(2)}}^{\mathrm{T}} \mathbf{y}^{(1)} = \begin{bmatrix} {Q_p}^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} \\ {Q_q}^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_3 \\ \mathbf{s}_4 \end{bmatrix} \\ \mathbf{x}_1 \end{bmatrix}, \quad \tilde{\mathbf{f}}^{(1)} = {Q^{(1)}}^{\mathrm{T}} \mathbf{f}^{(1)} = \begin{bmatrix} {Q_p}^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} \\ {Q_q}^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_3 \\ \hat{\mathbf{b}}_4 \end{bmatrix} \\ \mathbf{0} \end{bmatrix} \tag{A.3}$$

partitioned as $\mathbf{y}^{(1)}$ and $\mathbf{f}^{(1)}$, respectively. Q11

Then, compute the left upper antitriangular factorization of

$$\begin{bmatrix} \tilde{Y}_2 \\ \tilde{Y}_3 \end{bmatrix} = U \begin{bmatrix} Y_2 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}n-s \\ \}n-s' \end{matrix} \tag{A.4}$$

with $U \in \mathbb{R}^{(q+n-s)\times(q+n-s)}$ orthogonal and $R_{11} \in \mathbb{R}^{(n-s)\times(n-s)}$ nonsingular upper antitriangular. Let us partition U as follows:

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

with $U_{11} \in \mathbb{R}^{(n-s)\times(n-s)}$, $U_{12} \in \mathbb{R}^{(n-s)\times q}$, $U_{21} \in \mathbb{R}^{q\times(n-s)}$ and $U_{22} \in \mathbb{R}^{q\times q}$, and let

$$Q^{(3)} = \begin{bmatrix} U_{11} & & U_{12} & \\ & I_{p-n+s} & & \\ U_{21} & & U_{22} & \\ & & & I_q \end{bmatrix}.$$

Observe that, due to the particular structure of (A.4), U_{11} and U_{21} are upper triangular. Then, the linear system (A.2) is transformed into the following equivalent one:

$$M^{(2)}\mathbf{y}^{(2)} = \mathbf{f}^{(2)}, \tag{A.5}$$

where

$$M^{(2)} = Q^{(3)^{\mathrm{T}}}\tilde{M}^{(1)}Q^{(3)} = \begin{bmatrix} X_{11} & & X_{12} & & R_{11} \\ & I_{p-n+s} & & & \\ X_{12}{}^{\mathrm{T}} & & X_{22} & & \\ & & & -I_{q-n+s} & \\ R_{11}^{\mathrm{T}} & & & & \end{bmatrix},$$

with

$$\begin{aligned} X_{11} &= U_{11}{}^{\mathrm{T}}U_{11} - U_{21}{}^{\mathrm{T}}U_{21}, \\ X_{12} &= U_{11}{}^{\mathrm{T}}U_{12} - U_{21}{}^{\mathrm{T}}U_{22}, \\ X_{22} &= U_{12}{}^{\mathrm{T}}U_{12} - U_{22}{}^{\mathrm{T}}U_{22}, \end{aligned} \tag{A.6}$$

$$\mathbf{y}^{(2)} = \begin{bmatrix} \tilde{\mathbf{s}}_1 \\ \tilde{\mathbf{s}}_2 \\ \tilde{\mathbf{s}}_3 \\ \tilde{\mathbf{s}}_4 \\ \mathbf{x}_1 \end{bmatrix} = Q^{(3)^{\mathrm{T}}}\tilde{\mathbf{y}}^{(1)} = \begin{bmatrix} U_{11}{}^{\mathrm{T}}\bar{\mathbf{s}}_1 - U_{21}{}^{\mathrm{T}}\bar{\mathbf{s}}_3 \\ \bar{\mathbf{s}}_2 \\ U_{12}{}^{\mathrm{T}}\bar{\mathbf{s}}_1 - U_{22}{}^{\mathrm{T}}\bar{\mathbf{s}}_3 \\ \bar{\mathbf{s}}_4 \\ \mathbf{x}_1 \end{bmatrix}, \quad \mathbf{f}^{(2)} = Q^{(3)^{\mathrm{T}}}\tilde{\mathbf{f}}^{(1)} = \begin{bmatrix} U_{11}{}^{\mathrm{T}}\bar{\mathbf{b}}_1 - U_{21}{}^{\mathrm{T}}\bar{\mathbf{b}}_3 \\ \bar{\mathbf{b}}_2 \\ U_{12}{}^{\mathrm{T}}\bar{\mathbf{b}}_1 - U_{22}{}^{\mathrm{T}}\bar{\mathbf{b}}_3 \\ \bar{\mathbf{b}}_4 \\ \mathbf{0} \end{bmatrix}$$

and

$$\begin{bmatrix} \bar{\mathbf{s}}_1 \\ \bar{\mathbf{s}}_2 \end{bmatrix} = Q_p^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{s}}_3 \\ \bar{\mathbf{s}}_4 \end{bmatrix} = Q_q^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_3 \\ \mathbf{s}_4 \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{b}}_1 \\ \bar{\mathbf{b}}_2 \end{bmatrix} = Q_p^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{b}}_3 \\ \bar{\mathbf{b}}_4 \end{bmatrix} = Q_q^{\mathrm{T}} \begin{bmatrix} \hat{\mathbf{b}}_3 \\ \hat{\mathbf{b}}_4 \end{bmatrix}.$$

$\mathbf{In}(M^{(2)}) = \mathbf{In}(M^{(1)}) = (q+p-s, 0, p)$. Therefore, by Gould (1985, Theorem 2.2),

$$\mathbf{In}\left(\begin{bmatrix} I_{p-n+s} & & \\ & X_{22} & \\ & & -I_{q-n+s} \end{bmatrix}\right) = (q, 0, p-n+s).$$

Hence, $X_{22} \in \mathbb{R}^{(n-s)\times(n-s)}$ is symmetric negative definite with Cholesky factorization

$$X_{22} = -L_{22}L_{22}^{\mathrm{T}}. \tag{A.7}$$

The linear system (A.5) can now be solved in the same way the system (3.15) in Step 3 of Section 3 is solved. For the sake of brevity, we omit the details.

Q13 TABLE A1 *Computation complexity of the algorithm described in Appendix* A.3

	Flops
(3.3)	$2s^2(n-s/3)$
(A.1)	$2(n-s)^2(p+q-2(n-s)/3)$
(A.4)	$2(n-s)^3$
(A.6)	$11/6(n-s)^3$

375 In Table A1, we report the number of floating point operation due to compute the most expensive parts of this variant of the algorithm, neglecting the terms of lower order. Q12