ETH Library

On positive semidefinite matrices with known null space

Report

Author(s):
Arbenz, Peter; Drmač, Zlatko

Publication date:
2000

Permanent link:
https://doi.org/10.3929/ethz-a-006654183

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
[Technical report / Computer Science Department, ETH Zürich] 352

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

ON POSITIVE SEMIDEFINITE MATRICES WITH KNOWN NULL SPACE

PETER ARBENZ* AND ZLATKO DRMAČ†

Abstract. We show how the zero structure of a basis of the null space of a positive semidefinite matrix can be exploited to very accurately compute its Cholesky factorization. We discuss consequences of this result for the solution of (constrained) linear systems and eigenvalue problems. The results are of particular interest if A and the null space basis are sparse.

Key words. Positive semidefinite matrices, Cholesky factorization, null space basis.

AMS subject classifications. 65F05, 65F50

1. Introduction.

The Cholesky factorization $A = R^TR$, R upper-triangular, exists for any symmetric positive semidefinite matrix A. In fact, R is the upper triangular factor of the QR factorization of $A^{1/2}$ [11, §10.3]. R can be computed with the well-known algorithm for positive definite matrices. However, zero pivots may appear. As zero pivots come with a zero row/column in the reduced A, a zero pivot implies a zero row in R. To actually compute a numerically stable Cholesky factorization of a positive semidefinite matrix one is advised to apply diagonal pivoting [11].

A semidefinite matrix A may be given implicitly, in factored form $A = F^TF$, where $F \in \mathbb{R}^{p\times n}$ is of full row rank $r = \text{rank}(A)$. F, that does not need to be a Cholesky factor, exposes the singularity of A explicitly as $\mathcal{N}(A) = \mathcal{N}(F)$. In this case both the linear system and the eigenvalue problem can be solved efficiently and elegantly by working directly on the matrix F, never forming the matrix A explicitly. In fact, in some applications, not assembling the matrix A but its factor F is the most important step in the overall process of the numerical computation. One obvious reason is that the (spectral) condition number of F is the square root of the condition number of A. In finite element computation, F is the so called natural factor of the stiffness matrix A [2]. In the framework of linear algebra, every symmetric positive semidefinite matrix is the Gram matrix of some set of vectors, the columns of F.

Another possibility to have the singularity of A explicit is to have available a basis of its null space $\mathcal{N}(A)$. This is the situation that we want to investigate in this note. We will see that knowing a basis of $\mathcal{N}(A)$ allows to determine *a priori* when the zero pivots will occur in the Cholesky factorization. It also permits to give a positive definite submatrix of A right away. These results are of particular interest if A and the null space basis are sparse. This is the case in the application from electromagnetics that prompted this study [1]. There, a vector that is orthogonal to the null space corresponds to a discrete electric field that is divergence-free.

Our findings permit to work with the positive definite part of A and to compute a rank revealing Cholesky factorization $A = R^TR$ where the upper trapezoidal R has full row rank. What is straightforward in exact arithmetic amounts to simply *replacing by zero* potentially inaccurate small numbers. We analyze the error that is introduced by this procedure.

[*] Swiss Federal Institute of Technology (ETH), Institute of Scientific Computing, CH-8092 Zurich, Switzerland (arbenz@inf.ethz.ch)

[†] University of Zagreb, Departement of Mathematics, Bijenička 30, HR-10000 Zagreb, Croatia (drmac@math.hr). The work of this author was supported by the Croatian Ministry of Science and Technology grant 037012.

ETH Zürich, Computer Science Department, Technical Report #352, November 17, 2000

We complement this note with some implications of the above for solving eigenvalue problems and constrained systems of equations.

2. Cholesky factorization of a positive semidefinite matrix with known null space.

In this section we consider joint structures of a semidefinite matrix A and its null space.

THEOREM 2.1. *Let $A = R^T R$ be the Cholesky factorization of the positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$. Let $Y \in \mathbb{R}^{n\times m}$ with $\mathcal{R}(Y) = \mathcal{N}(A)$ and, for $i = 1,\ldots,m$, set $n_i := \max\{k \mid y_{ki} \neq 0\}$. If $n_1 < n_2 < \cdots < n_m$ then $r_{n_i n_i} = 0$, $i = 1,\ldots,m$. These are the only zero entries on the diagonal of R.*

Proof. Notice that the assumptions imply that $Y := [\mathbf{y}_1,\ldots,\mathbf{y}_m]$ has full rank. By Sylvester's law of inertia R has precisely m zeros on its diagonal. Further,

$$(R\mathbf{y}_i)_{n_i} = r_{n_i n_i} y_{n_i i} = 0,$$

whence $r_{n_i n_i} = 0$ as $y_{n_i i} \neq 0$. □

If only $n_1 \leq n_2 \leq \cdots \leq n_m$, Y, flipped upside-down, can be transformed into column-echelon form in order to obtain strong inequalities.

The Cholesky factor R appearing in Theorem 2.1 is an $n \times n$ upper triangular matrix with m zero rows. These rows do not affect the product $R^T R$. Therefore, they can be removed from R to yield an $(n-m) \times n$ matrix $\widehat{R}$ with $\widehat{R}^T \widehat{R} = A$.

If the numbers n_i are known, it is convenient to permute the rows of Y and accordingly the rows and columns n_i of A to the end. Then Theorem 2.1 can be applied with $n_i = n-m+i$. The last m rows of R in Theorem 2.1 vanish. So, $\widehat{R}$ is upper trapezoidal.

After the just mentioned permutation the lowest $m\times m$ block of Y is non-singular, in fact, upper triangular. This consideration leads to an alternative formulation of Theorem 2.1.

THEOREM 2.2. *Let $A = R^T R$ be the Cholesky factorization of the positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$. Let $Y \in \mathbb{R}^{n\times m}$ with $\mathcal{R}(Y) = \mathcal{N}(A)$. If the last m rows of Y are linearly independent, then the leading principal $(n-m) \times (n-m)$ submatrix of A is positive definite and R can be taken $(n-m) \times n$ upper triangular.*

Proof. Let

$$W := \begin{bmatrix} I_{n-m} & Y_1 \\ O & Y_2 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \qquad Y_2 \in \mathbb{R}^{m\times m}. \tag{2.1}$$

Y_2 consists of the last m rows of Y. W is therefore invertible. Applying a congruence transformation with W on A gives

$$W^T A W = \begin{bmatrix} I_{n-m} & O \\ Y_1^T & Y_2^T \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I_{n-m} & Y_1 \\ O & Y_2 \end{bmatrix} = \begin{bmatrix} A_{11} & O \\ O & O \end{bmatrix}. \tag{2.2}$$

By Sylvester's law of inertia A_{11} must be positive definite.

Let $A_{11} = R_{11}^T R_{11}$ be the Cholesky factorization of A_{11}. Then, the Cholesky factor of the matrix in (2.2) is $[R_{11}, O] \in \mathbb{R}^{(n-m)\times n}$. Therefore, the Cholesky factor of A is $[R_{11}, O]W^{-1} = [R_{11}, -R_{11}Y_1Y_2^{-1}]$. □

Theorem 2.2 is applicable as long as the last m rows of Y form an invertible matrix. If rows $i_1,\ldots,i_m$ of Y are linearly independent, we can permute Y such that these rows become the last ones. In particular, if we want A_{11} to be as sparse as possible, we may choose $i_1, i_2, \ldots$ to be the m most densely populated rows/columns of A with the following greedy algorithm: If we have determined $i_1,\ldots,i_k$ we choose

i_{k+1} to be the index of the densest column of A such that rows $i_1, \ldots, i_{k+1}$ of Y are linearly independent. In this way we can hope for an A_{11} with sparse Cholesky factors.

Remark 2.1. The equation

$$-\Delta u(\mathbf{x}) = 0 \text{ in } \Omega \subset \mathbb{R}^n, \qquad \partial_n u(\mathbf{x}) = 0 \text{ on } \partial\Omega, \tag{2.3}$$

in a simply connected domain Ω is satisfied by all constant functions u. The discretization of (2.3) with finite elements of Lagrange type [4] leads to a positive semidefinite matrix A with a one dimensional null space spanned by the vector $\mathbf{e}$ with all entries equal to 1. Theorem 2.1 now implies that, no matter how we permute A, in the Cholesky factorization the single zero on the diagonal of R will not appear before the very last elimination step. $\square$

Example 2.1. Let A and Y be given by

$$A := \begin{bmatrix} 1 & 0 & 1 & 1 & 3 \\ 0 & 9 & 3 & 9 & 9 \\ 1 & 3 & 3 & 6 & 8 \\ 1 & 9 & 6 & 14 & 16 \\ 3 & 9 & 8 & 16 & 22 \end{bmatrix}, \qquad Y := \begin{bmatrix} 2 & 3 \\ 0 & 1 \\ 0 & 6 \\ 1 & 0 \\ -1 & -3 \end{bmatrix}.$$

Then $AY = O$. As the last two rows of Y are linearly independent, Theorem 2.2 states that the principal 3×3 submatrix of A is positive definite and that its Cholesky factor is 3×5 upper triangular. In fact,

$$R = \begin{bmatrix} 1 & 0 & 1 & 1 & 3 \\ 0 & 3 & 1 & 3 & 3 \\ 0 & 0 & 1 & 2 & 2 \end{bmatrix}.$$

Let P be the permutation matrix, that exchanges 2nd with 4th and 3rd with 5th entry of a 5-vector. Then,

$$A_1 := PAP^T = \begin{bmatrix} 1 & 1 & 3 & 0 & 1 \\ 1 & 14 & 16 & 9 & 6 \\ 3 & 16 & 22 & 9 & 8 \\ 0 & 9 & 9 & 9 & 3 \\ 1 & 6 & 8 & 3 & 3 \end{bmatrix}, \qquad Y_1 := PY = \begin{bmatrix} 2 & 3 \\ 1 & 0 \\ -1 & -3 \\ 0 & 1 \\ 0 & 6 \end{bmatrix}.$$

Now we have $n_1 = 3 < n_2 = 5$ and according to Theorem 2.1 the Cholesky factor R_1 of A_1 has zero diagonal elements at positions 3 and 5. Indeed,

$$R_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} \sqrt{13} & \sqrt{13} & 3\sqrt{13} & 0 & \sqrt{13} \\ 0 & 13 & 13 & 9 & 5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3. Consistent semidefinite systems. In this section we discuss how to solve

$$A\mathbf{x} = R^T R\mathbf{x} = \mathbf{b} \in \mathcal{R}(A), \tag{3.1}$$

where A, R, and Y are as in Theorem 2.1. Without loss of generality, we can assume that $n_i = r + i$, $r := n - m$. We split matrices and vectors in (3.1),

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{bmatrix} R_{11}^T \\ R_{12}^T \end{bmatrix} [R_{11},\ R_{12}] \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \tag{3.2}$$

with $\mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^r$ and $\mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^m$. So, A_{11} is obtained from A by deleting rows and columns n_i, $i = 1, \ldots, m$. The factorization (3.2) yields

$$A_{11} = R_{11}^T R_{11}, \qquad A_{12} = R_{11}^T R_{12}. \tag{3.3}$$

Although A_{11} is invertible, its condition number can be arbitrarily high. To reduce fill-in during factorization [8] any symmetric permutations can be applied to A_{11} without affecting the sequel. As R^T has full rank, $AY = O$ implies $RY = O$ or

$$R_{11} Y_1 + R_{12} Y_2 = O. \tag{3.4}$$

Since $n_i = r + i$, the $m \times m$ matrix Y_2 is upper triangular with non-zero diagonal elements. Because $\mathcal{R}(A) = \mathcal{N}(Y)^\perp$ the right side $\mathbf{b}$ of (3.1) has to satisfy

$$Y_1^T \mathbf{b}_1 + Y_2^T \mathbf{b}_2 = \mathbf{0}. \tag{3.5}$$

It is now easy to show that a particular solution of (3.1) is given by $\mathbf{x}$ with components

$$\mathbf{x}_1 = A_{11}^{-1} \mathbf{b}_1 = R_{11}^{-1} R_{11}^{-T} \mathbf{b}_1, \qquad \mathbf{x}_2 = \mathbf{0}.$$

In fact, employing (3.3)–(3.5) the second block row in (3.2) is

$$A_{12}^T \mathbf{x}_1 - \mathbf{b}_2 = R_{12}^T R_{11} \mathbf{x}_1 + Y_2^{-T} Y_1^T \mathbf{b}_1 = R_{12}^T R_{11}^{-T} (A_{11} \mathbf{x}_1 - \mathbf{b}_1) = \mathbf{0}.$$

The manifold $\mathcal{S}$ of the solutions of (3.1)–(3.2) is

$$\mathcal{S} = \left\{ \mathbf{x} = \begin{pmatrix} A_{11}^{-1} \mathbf{b}_1 \\ \mathbf{0} \end{pmatrix} + Y\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^m \right\}.$$

The vector $\mathbf{a}$ can be determined such that the solution $\mathbf{x}$ satisfies some constraints $C^T \mathbf{x} = \mathbf{0}$ with $C \in \mathbb{R}^{n \times m}$ provided that $C^T Y$ is invertible. In particular, if $C = Y$ then $\mathbf{x}$ is perpendicular to the null space of A.

Let now A be given implicitly as a Gram matrix $A = F^T F$ with $F \in \mathbb{R}^{p \times n}$, $p \geq n$, and $Y \in \mathbb{R}^{n \times m}$ be as above. (This may require renumbering the columns of F.) As

$$FY = F_1 Y_1 + F_2 Y_2 = O,$$

and as Y_2 is nonsingular, the block F_2 depends linearly on F_1. Therefore, the QR factorization of F has the form

$$F = [F_1,\ F_2] = [Q_1,\ Q_2] \begin{bmatrix} R_{11} & R_{12} \\ O & O \end{bmatrix} = Q_1 [R_{11},\ R_{12}].$$

Since $A = F^T F = RR^T$, the factor $R = [R_{11},\ R_{12}]$ equals the upper trapezoidal Cholesky factor in (3.2).

4. Error Analysis. In this section we give backward error analyses for the semidefinite Cholesky factorization and for the null space basis.

4.1. Semidefinite Cholesky factorization. The floating-point computation of the Cholesky factorization of a semidefinite matrix is classified as unstable by Higham [11, §10.3.2]. The principal problem is the determination of the rank of the matrix.

If we assume, as we do in this note, that a basis of the null space of the matrix under consideration is known *a priori* then, of course, its rank is known. Let A be partitioned as in (3.2). We assume that $A_{11} \in \mathbb{R}^{r\times r}$ is positive definite numerically, i.e. that the Cholesky factorization does not break down in floating point arithmetic with round-off unit $\mathbf{u}$. Due to a result by Demmel [5] (see also [11, Thm.10.14]) this is the case if,

$$\lambda_{\min}((A_{11})_s) \equiv \|(A_{11})_s^{-1}\|^{-1} > 2rf(r)\mathbf{u}, \qquad f(r) = \frac{r+1}{1-2(r+1)\mathbf{u}}, \tag{4.1}$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue, $\|\cdot\|$ is the spectral norm, and

$$(A_{11})_s = \mathrm{diag}(A_{11})^{-1/2} A_{11} \mathrm{diag}(A_{11})^{-1/2}.$$

If (4.1) does not hold, A_{11} is not numerically definite. Note that $(A_{11})_s$ is symmetric positive definite with unit diagonal. The assumption on $\lambda_{\min}((A_{11})_s)$ can be relaxed if, for instance, we use double precision accumulation during the factorization. Then $f(r)$ can be replaced by a small integer for all r not larger than $1/\mathbf{u}$. We assume, however, that $2rf(r)\mathbf{u} < 1$.

The Cholesky decomposition of A is computed as indicated in (3.3). The Cholesky factor of A_{11} is computed first. Then the matrix R_{12} is obtained as the solution of the matrix equation $R_{11}^T X = A_{12}$.

Let $\widetilde{R}_{11}$ denote the computed floating-point Cholesky factor of A_{11}. Then the following two important facts are well-known.

(1) There exists a symmetric δA_{11} such that $A_{11} + \delta A_{11} = \widetilde{R}_{11}^T \widetilde{R}_{11}$ and

$$\max_{1\le i,j\le r} \frac{|(\delta A_{11})_{ij}|}{\sqrt{(A_{11})_{ii}(A_{11})_{jj}}} \le f(r)\mathbf{u}. \tag{4.2}$$

This is the backward error bound by Demmel [5], [11, Theorem 10.5].

(2) Let $(\delta A_{11})_s := \mathrm{diag}(A_{11})^{-1/2}\delta A_{11}\mathrm{diag}(A_{11})^{-1/2}$. (4.1) and (4.2) imply that the Frobenius norm of $(\delta A_{11})_s$ satisfies $\|(\delta A_{11})_s\|_F \le rf(r)\mathbf{u} < \lambda_{\min}((A_{11})_s)$. Since assumption (4.1) implies $2\|(\delta A_{11})_s\|_F < \lambda_{\min}((A_{11})_s)$, one can show [7] that there exists an upper triangular matrix Γ such that

$$\widetilde{R}_{11} = (I+\Gamma)R_{11}, \qquad \|\Gamma\|_F \le \frac{\sqrt{2}\|(A_{11})_s^{-1}\|\|(\delta A_{11})_s\|_F}{1+\sqrt{1-2\|(A_{11})_s^{-1}\|\|(\delta A_{11})_s\|_F}} < \frac{1}{\sqrt{2}}.$$

Let $\widetilde{R}_{12}$ be the floating-point solution of the matrix equation $\widetilde{R}_{11}^T X = A_{12}$. Then $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ is the computed approximation of the exact Cholesky factor $R = [R_{11}, R_{12}]$. Let $\widetilde{A} = A + \delta A = \widetilde{R}^T\widetilde{R}$ be partitioned conforming with (3.2). Since $A+\delta A$ is positive semidefinite and of rank r by construction, the equation $\widetilde{A}_{22} = \widetilde{A}_{12}^T \widetilde{A}_{11}^{-1}\widetilde{A}_{12}$ holds.

If we compute $\widetilde{R}_{12}$ column by column, then, using Wilkinson's analysis of triangular linear systems [11, Theorem 8.5],

$$|\widetilde{R}_{11}^T\widetilde{R}_{12} - A_{12}| \le t(r)\mathbf{u}|\widetilde{R}_{11}|^T|\widetilde{R}_{12}|, \qquad t(r) = \frac{r}{1-r\mathbf{u}},$$

where the matrix absolute values and the inequality are to be understood entry-wise. Thus, we can write $\widetilde{R}_{12}$ as

$$\widetilde{R}_{12} = \widetilde{R}_{11}^{-T}(A_{12} + \delta A_{12}), \qquad |\delta A_{12}| \leq t(r)\mathbf{u}|\widetilde{R}_{11}|^T|\widetilde{R}_{12}|. \tag{4.3}$$

Also, if we define $\Psi = (I+\Gamma)^{-T} - I$, $\Omega = t(r)\mathbf{u}|\widetilde{R}_{11}^{-T}||\widetilde{R}_{11}^T|$, we have

$$\widetilde{R}_{12} = (I+\Psi)R_{12} + \widetilde{R}_{11}^{-T}\delta A_{12}, \qquad |\widetilde{R}_{11}^{-T}\delta A_{12}| \leq \Omega|\widetilde{R}_{12}|. \tag{4.4}$$

Further, from the inequality $|\widetilde{R}_{12}| \leq (I + |\Psi|)|R_{12}| + \Omega|\widetilde{R}_{12}|$ and using the M-matrix property of $I - \Omega$ we obtain

$$|\widetilde{R}_{12}| \leq (I-\Omega)^{-1}(I+|\Psi|)|R_{12}|. \tag{4.5}$$

Hence, relations (4.2), (4.3), (4.5) imply that the backward error for all (i,j) in the $(1,2)$ block in (3.2) is bounded by

$$\begin{aligned} |(\delta A_{12})_{ij}| &\leq t(r)\mathbf{u}\|\widetilde{R}_{11}\mathbf{e}_i\|\|\widetilde{R}_{12}\mathbf{e}_{j'}\| \leq t(r)\mathbf{u}\sqrt{(\widetilde{A}_{11})_{ii}(\widetilde{A}_{22})_{j'j'}}, \qquad j' = j - r, \\ &\leq t(r)\mathbf{u}(1+f(r)\mathbf{u})\frac{1 + \|\,|\Psi|\,\|}{1-\|\Omega\|}\sqrt{(A_{11})_{ii}(A_{22})_{j'j'}}. \end{aligned}$$

We first observe that $\|\,|\Psi|\,\| \leq \sqrt{r}\|\Gamma\|/(1-\|\Gamma\|)$ and that $\|\Omega\| \leq rt(r)\mathbf{u}\sqrt{\|(\widetilde{A}_{11})_s^{-1}\|}$. Note that our assumptions imply that

$$\|\Omega\| \leq \frac{1}{\sqrt{2}}\frac{\sqrt{r}}{r+1} < 1/2, \qquad \frac{\|\Gamma\|}{1-\|\Gamma\|} < 1 + \sqrt{2}.$$

It remains to estimate the backward error in the $(2,2)$ block of the partition (3.2). Using relation (4.4), we compute $\delta A_{22} = \widetilde{R}_{12}^T\widetilde{R}_{12} - R_{12}^T R_{12}$ as follows:

$$\begin{aligned} \delta A_{22} = R_{12}^T(\Psi^T + \Psi + \Psi^T\Psi)R_{12} &+ R_{12}^T(I+\Psi^T)\widetilde{R}_{11}^{-T}\delta A_{12} \\ &+ \delta A_{12}^T\widetilde{R}_{11}^{-1}(I+\Psi)R_{12} + \delta A_{12}^T\widetilde{R}_{11}^{-1}\widetilde{R}_{11}^{-T}\delta A_{12}. \end{aligned}$$

Using the inequalities from relations (4.4), (4.5) we obtain, for all (i,j),

$$|(\delta A_{22})_{ij}| \leq \sqrt{(A_{22})_{ii}(A_{22})_{jj}}\left(2\psi + 2\omega\frac{1+\psi'}{1-\omega} + \psi^2 + \omega^2\frac{(1+\psi)^2}{(1-\omega)^2}\right),$$

where $\omega = \|\Omega\|$, $\psi = \|\Psi\|$, $1 + \psi' = (1+\psi)(1 + \|\,|\Psi|\,\|)$.

We summarize the above analysis in the following

THEOREM 4.1. *Let A be a $n \times n$ positive semidefinite matrix of rank r with block partition* (3.2)*, where the $r \times r$ matrix A_{11} is positive definite with the property* (4.1)*. Then the floating-point Cholesky factorization with roundoff* $\mathbf{u}$ *will compute an upper trapezoidal matrix $\widetilde{R}$ of rank r such that $\widetilde{R}^T\widetilde{R} = A + \delta A$ where δA is a symmetric backward perturbation with the following bounds:*

$$\begin{aligned} &|\delta A_{ij}| \leq f(r)\mathbf{u}\sqrt{A_{ii}A_{jj}}, \quad 1 \leq i,j \leq r, \\ &|\delta A_{ij}| \leq \left\{2t(r)(1 + (1+\sqrt{2})\sqrt{r})(1+f(r)\mathbf{u})\right\}\mathbf{u}\sqrt{A_{ii}A_{jj}}, \quad 1 \leq i \leq r < j \leq n, \\ &|\delta A_{ij}| \leq \left\{2rt(r)\sqrt{\tilde{\kappa}} + \sqrt{8}rf(r)\kappa + O(\mathbf{u})\right\}\mathbf{u}\sqrt{A_{ii}A_{jj}}, \quad r < i,j \leq n. \end{aligned}$$

In the last estimate, $\kappa = \|(A_{11})_s^{-1}\|$, $\tilde{\kappa} = \|(\widetilde{A}_{11})_s^{-1}\|$. *Further, if* $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ *and if* $R = [R_{11}, R_{12}]$ *is the exact Cholesky factor of* A*, then*

$$\begin{aligned} \widetilde{R}_{11} - R_{11} &= \Gamma R_{11}, & \|\Gamma\| &\le \sqrt{2} r f(r) \kappa \mathbf{u}, \\ |\widetilde{R}_{12} - R_{12}| &\le \Xi |R_{12}|, & \|\Xi\| &\le r t(r) \sqrt{\tilde{\kappa}} \mathbf{u} + \sqrt{2} r f(r) \kappa \mathbf{u} + O(\mathbf{u}^2). \end{aligned}$$

Here, the matrix Γ *is upper triangular and* Ξ *is to the first order* $|\Psi| + \Omega$.

Further, let the Cholesky factorization of A_{11} *be computed with pivoting so that* $(R_{11})_{ii} \ge \sum_{k=i}^{j} (R_{11})_{kj}^2$, $1 \le i \le j \le r$. *Then, the error* $\delta R_{11} = \widetilde{R}_{11} - R_{11}$ *is also row-wise small, that is*

$$\|\mathbf{e}_i^T \delta R_{11}\| \le \|\mathbf{e}_i^T \Gamma\| \sqrt{r - i + 1} (R_{11})_{ii}, \quad i = 1, \ldots, r. \tag{4.6}$$

Remark 4.1. Note that Theorem 4.1 also states that in the positive definite case the Cholesky factorization with pivoting computes the triangular factor with small column- and row-wise relative errors. This affects the accuracy of the linear equation solver (forward and backward substitutions following the Cholesky factorization) not only by ensuring favorable condition numbers but also by ensuring that the errors in the coefficients of the triangular systems are small. □

4.2. Null space error. We now derive a backward error for the null space Y of A. We seek an $n \times (n - r)$ full rank matrix $\widetilde{Y} = Y + \delta Y$ such that δY is small and $\widetilde{A}\widetilde{Y} = 0$. As the null space and the range of A change simultaneously (being orthogonal complements of each other), the size of δY necessarily depends on a certain condition number of A; and the relevant condition number will depend on the form of the perturbation δA.

The equation that we investigate is $\widetilde{R}(Y + \delta Y) = 0$ or, equivalently, $\widetilde{R}\delta Y = -\delta R Y$. If $\widetilde{R}$ is sufficiently close to R (to guarantee invertibility of $\widetilde{R}R^+$) we can write

$$\delta Y = R^+ (\widetilde{R} R^+)^{-1} \delta R Y = R^T (\widetilde{R} R^T)^{-1} \delta R Y. \tag{4.7}$$

Though simple this equation is instructive. First of all, only the components of the columns of δR that lie in the null space $\mathcal{N}(A)$ affect the value of δY. Also, $Y + \delta Y$ keeps the full column rank of Y. Finally, $Y^T \delta Y = O$. Therefore, $\tan\angle(\mathcal{R}(Y), \mathcal{R}(\widetilde{Y})) = \|\delta Y\| / \sigma_{\min}(Y)$. It is easy to modify Y such that $\sigma_{\min}(Y) \ge 1$, e.g., if $Y_2 = I_m$. Thus, $\|\delta Y\|$ measures the angle between the true null space and the null space of the perturbed matrix $\widetilde{A}$. In the sequel we try to bound $\|\delta Y\|$.

If we rewrite (4.7) as

$$\delta Y = R^+ (I + \delta R R^+)^{-1} \delta R Y = (R')^+ \left(I + \delta R' (R')^+ \right)^{-1} \delta R' Y,$$

we get, after some manipulations,

Proposition 4.2. *Let* D *be nonsingular matrix and let* $R = DR'$, $\delta R = D \delta R'$. *If* $\|\delta R' (R')^+\| < 1$ *then, for* $i = 1, \ldots, n - r$,

$$\|\delta \mathbf{y}_i\| \le \frac{\|(R')^+\|}{1 - \|\delta R' (R')^+\|} \|\delta R' \mathbf{y}_i\| \le \frac{\|\delta R' P_{\mathcal{N}(A)}\| \|(R')^+\|}{1 - \|\delta R' (R')^+\|} \|\mathbf{y}_i\|. \tag{4.8}$$

Here, $\mathbf{y}_i = Y \mathbf{e}_i$, $\delta \mathbf{y}_i = Y \delta \mathbf{e}_i$, *and* $P_{\mathcal{N}(A)}$ *denotes the orthogonal projection onto the null space of* A. □

We will discuss choices for D later. The Proposition indicates that the crucial quantity for bounding $\|\delta Y\|$ is $\|\delta R'Y\|$. The following two examples detail this fact.

Example 4.1. Let β be big, of the order of $1/\mathbf{u}$, and let

$$A = R^T R = \begin{bmatrix} \sqrt{3} & 0 \\ 1 & 1 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 1 & \beta \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & \sqrt{3} & \beta\sqrt{3} \\ \sqrt{3} & 2 & \beta+1 \\ \beta\sqrt{3} & \beta+1 & \beta^2+1 \end{bmatrix}.$$

The null space of A is spanned by $Y = [(1-\beta)/\sqrt{3}, -1, 1]^T$, which means that deleting any row and column of A leaves a nonsingular 2×2 matrix. Let's choose it be the last one, and let us follow the algorithm. For the sake of simplicity, let the only error be committed in the computation of the $(1,1)$ entry of $\widetilde{R}_{11}$ which is $\sqrt{3}(1+\varepsilon_1)$, $|\varepsilon_1| \leq \mathbf{u}$, instead of $\sqrt{3}$. Then we solve the lower triangular system for $\widetilde{R}_{12}$ and obtain

$$\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}] = \begin{bmatrix} \sqrt{3}\,(1+\varepsilon_1) & 1 & \beta(1+\varepsilon_2) \\ 0 & 1 & 1-\beta\varepsilon_2 \end{bmatrix}, \qquad |\varepsilon_2| \leq \mathbf{u} + O(\mathbf{u}^2).$$

Thus,

$$\delta R = \begin{bmatrix} \sqrt{3}\,\varepsilon_1 & 0 & \beta\varepsilon_2 \\ 0 & 0 & -\beta\varepsilon_2 \end{bmatrix}, \qquad \delta R Y = \begin{pmatrix} (1-\beta)\varepsilon_1 + \beta\varepsilon_2 \\ -\beta\varepsilon_2 \end{pmatrix}.$$

If we take $\beta = 10^{15}$ and perform the computation in MATLAB where $\mathbf{u} \approx 2.22 \cdot 10^{-16}$ then $\beta\varepsilon_2 = 0.25$. Thus, $\|\delta R Y\| = \mathcal{O}(1)$. However, $\sigma_{\min}(Y) = \|Y\| = \mathcal{O}(\beta)$ such that the angle between Y and δY is small. □

Example 4.2. We alter the $(1,1)$ entry $\sqrt{3}$ of R of the previous example to get β,

$$A = R^T R = \begin{bmatrix} \beta & 0 \\ 1 & 1 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \beta & 1 & \beta \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \beta^2 & \beta & \beta^2 \\ \beta & 2 & \beta+1 \\ \beta^2 & \beta+1 & \beta^2+1 \end{bmatrix}.$$

Now, $Y = [(1-\beta)/\beta, -1, 1]^T$. Again, we delete the last row and column of A and proceed as in Example 4.1. Let us again assume that the only error occurs in the $(1,1)$ entry of R_{11} which becomes $\beta/(1+\varepsilon_1)$. Then,

$$\widetilde{R} = \begin{bmatrix} \beta/(1+\varepsilon_1) & 1 & \beta(1+\varepsilon_1) \\ 0 & 1 & 1-\beta\varepsilon_1 \end{bmatrix},$$

and

$$\delta R = \begin{bmatrix} -\beta\varepsilon_1/(1+\varepsilon_1) & 0 & \beta\varepsilon_1 \\ 0 & 0 & -\beta\varepsilon_1 \end{bmatrix}, \qquad \delta R Y = \begin{pmatrix} \varepsilon_1(1+\beta\varepsilon_1)/(1+\varepsilon_1) \\ -\beta\varepsilon_1 \end{pmatrix}.$$

Again, $\|\delta R Y\| = \mathcal{O}(1)$. But now also $\|Y\| = \mathcal{O}(1)$. In fact, in computations with MATLAB, we observe an angle as large as $\mathcal{O}(10^{-2})$ between Y and δY. □

Remark 4.2. Interestingly, if we set $\beta = 10^5$ in Example 4.1, the MATLAB function `chol()` computes the Cholesky factor

$$\widetilde{R} = \begin{bmatrix} \mathtt{1.7321e+000} & \mathtt{1.0000e+000} & \mathtt{1.0000e+005} \\ 0 & \mathtt{1.0000e+000} & \mathtt{1.0000e+000} \\ 0 & 0 & \mathtt{1.9531e-003} \end{bmatrix}.$$

It is clear that the computed and stored A is a perturbation of the true A. Therefore, numerically, it can be positive definite. It is therefore quite possible to know the rank

$r < n$ of A exactly, to have a basis of the null space of A and a numerically stored positive definite floating-point A. Strictly speaking, this is a contradiction. Certainly, from an application or numerical point of view, it is advisable to be very careful when dealing with semidefiniteness. □

In Examples 4.1 and 4.2 we excluded the largest diagonal entry of A. In fact, we can give an estimate that relates the error in R_{12} to the size of the deleted entries. Suppose we managed the deleted diagonal entries of A to be the $m = n - r$ smallest ones. Can we then guarantee that the relevant error in R will be small, and can we check the stability by a simple, inexpensive test?

According to Theorem 4.1, the matrix R_{11} is computed with row-wise small relative error, provided that the Cholesky factorization of A_{11} is computed with pivoting. If that is the case, then it remains to estimate the row-wise perturbations of R_{12}. If Ξ is as in Theorem 4.1, then the inequality

$$\|\mathbf{e}_i^T \delta R_{12}\| \leq \|\mathbf{e}_i^T \Xi\| \sqrt{\mathrm{trace}(A_{22})} \leq \|\mathbf{e}_i^T \Xi\| \left(\frac{\mathrm{trace}(A_{22})}{(A_{11})_{ii}} \right)^{1/2} \frac{\|R_{11}\mathbf{e}_i\|}{\|\mathbf{e}_i^T R\|} \|\mathbf{e}_i^T R\|, \tag{4.9}$$

holds for all $i = 1, \ldots, r$ and

$$\frac{\|R_{11}\mathbf{e}_i\|}{\|\mathbf{e}_i^T R\|} = \frac{\|R_{11}\mathbf{e}_i\|}{|(R_{11})_{ii}|} \frac{|(R_{11})_{ii}|}{\|\mathbf{e}_i^T R\|} \leq \frac{\|R_{11}\mathbf{e}_i\|}{|(R_{11})_{ii}|} = \frac{1}{\sin \phi_i} \leq \sqrt{\|(A_{11})_s^{-1}\|} \tag{4.10}$$

with some $\phi_i \in (0, \pi/2]$. The angle ϕ_i has a nice interpretation. Let $A = F^T F$ be any factorization of A, with $F = [F_1, F_2]$ where F_1 has full column rank and $F_1^T F_1 = A_{11}$. Then ϕ_i is the angle between $F_1 \mathbf{e}_i$ and the span of $\{F_1\mathbf{e}_1, \ldots, F_1\mathbf{e}_{i-1}\}$. (This is easily seen from the QR factorization of F_1.)

The following Proposition states that well-conditioned $(A_{11})_s$ and a certain dominance of A_{11} over A_{22} ensure accurate rows of the computed matrix $\widetilde{R}$.

Proposition 4.3. *With the notation of Theorem 4.1, let A (and accordingly Y) be arranged such that*

$$\max_i (A_{22})_{ii} \leq \min_i (A_{11})_{ii}. \tag{4.11}$$

If the Cholesky factorization of A_{11} is computed with (standard) pivoting, then

$$\|\mathbf{e}_i^T \delta R\| \leq \max\{\|\mathbf{e}_i^T \Gamma\|, \|\mathbf{e}_i^T \Xi\|\} \frac{\sqrt{n-i+1}}{\sin \phi_i} \|\mathbf{e}_i^T R_{11}\|, \quad i = 1, \ldots, r, \tag{4.12}$$

where $\sin\phi_i$ *is defined in* (4.10).

Proof. This follows from relations (4.6), (4.9), (4.10) and the assumption (4.11). We only note that in (4.9) and (4.10) we can replace $\|\mathbf{e}_i^T R\|$ by $\|\mathbf{e}_i^T R_{11}\|$. □

Remark 4.3. If $A = SA_sS$ with $S^2 = \mathrm{diag}(A_{ii})$, then SY spans $\mathcal{N}(A_s)$, and any partition of A_s satisfies condition (4.11). If we apply the preceeding analysis to A_s and SY, we get an estimate for δY in the elliptic norm generated by S. □

Note that Proposition 4.2 is true for any diagonal D as long as $\|(R')^+\|$ is moderately big and $\|\delta R'\|$ is small. We have just seen that $\delta R'$ is nicely bounded if we choose $D = \mathrm{diag}(\|\mathbf{e}_i^T R_{11}\|)$. Moreover, $R' = D^{-1}R$ has an inverse nicely bounded independent of A_{11} because [11, §10]

$$\|(R')^+\| \leq \|(D^{-1}R_{11})^{-1}\| \leq h(r).$$

Here the function $h(r)$ is in the worst case dominated by 2^r and in practice one usually observes an $O(r)$ behaviour. In any case, $\|(D^{-1}R_{11})^{-1}\|$ is at most r times larger than $\|(A_{11})_s^{-1}\|^{1/2}$. More sophisticated pivoting can make sure that the behaviour of $h(r)$ is not worse than Wilkinson's pivot growth factor. We skip the details for the sake of brevity.

To conclude, if the Cholesky factorization of A_{11} is computed with pivoting and relation (4.11) holds, then the backward error in Y can be estimated using (4.8) and (4.12), where $D = \mathrm{diag}(\|\mathbf{e}_i^T R_{11}\|)$.

4.3. Computation with implicit A**.** We consider now the backward stability of the computation with A given implicitly as $A = F^T F$, where $F \in \mathbb{R}^{p\times n}$ has rank r. Thus, the Cholesky factorization of A is accomplished by computing the QR factorization of F.

In the numerical analysis of the QR factorization we use the standard, well-known backward error analysis which can be found e.g. in [11, §18]. The simplest form of this analysis states that the backward error in the QR factorization is column-wise small. For instance, if we compute the Householder (or Givens) QR factorization of F in floating point arithmetic with roundoff $\mathbf{u}$, then the backward error δF satisfies

$$\|\delta F\mathbf{e}_i\| \le \varepsilon_1 \|F\mathbf{e}_i\|, \quad \varepsilon_1 \le f_1(p,n)\mathbf{u}, \qquad 1 \le i \le n,$$

where $f_1(p,n)$ is a polynomial of moderated degree in the matrix dimensions.

Our algorithm follows the same ideas as in the direct computation of R from A. The knowledge of a null space basis admits that we can assume that F is in the form $F = [F_1, F_2]$ the $p \times r$ matrix F_1 is of rank r, see section 3. We then apply r Householder reflections to F which yields, in exact arithmetic, the matrix

$$Q^T F = R = \begin{pmatrix} R_{11} & R_{12} \\ O & R_{22} \end{pmatrix}, \qquad R_{22} = O,$$

where $R_{11} \in \mathbb{R}^{r\times r}$ is upper triangular and nonsingular. If $Q = [Q_1, Q_2]$ is partitioned conforming with F, then $F_1 = Q_1 R_{11}$ is the QR factorization of F_1.

In floating point computation, R_{22} is unlikely to be zero. Our algorithm simply sets to zero whatever is computed as approximation of R_{22}. As we shall see, the backward error (in F) of this procedure depends on a certain condition number of the matrix F_1.

THEOREM 4.4. *Let $F \in \mathbb{R}^{p\times n}$ have rank r and be partitioned in the form $F = [F_1, F_2]$, where $F_1 \in \mathbb{R}^{p\times r}$ has the numerically well determined full rank r. More specifically, if $(F_1)_c$ is obtained from F_1 by scaling columns to have unit Euclidean norm, then we assume that $\sqrt{r}\varepsilon_1\|(F_1)_c^{+}\| < 1/5$.*

Let the QR factorization of F be computed as described above, and let $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ be the computed upper trapezoidal factor.

Then there exist a backward perturbation ΔF and an orthogonal matrix $\widehat{Q}$ such that $F + \Delta F = \widehat{Q}\widetilde{R}$ is the QR factorization of $F + \Delta F$. The matrix $F + \Delta F$ has rank r. If $\Delta F = [\Delta F_1, \Delta F_2]$ and $\widehat{Q} = [\widehat{Q}_1, \widehat{Q}_2]$ are partitioned as F, and $\delta Q_1 := \widehat{Q}_1 - Q_1$, then

$$\begin{aligned}
&\|\Delta F\mathbf{e}_i\| \le \varepsilon_1\|F\mathbf{e}_i\|, && 1 \le i \le r,\\
&\|\delta Q_1\|_F \le 11\eta + O(\eta^2), && \eta = \|\Delta F_1 R_{11}^{-1}\|_F \le \sqrt{r}\varepsilon_1\|(F_1)_c^{+}\|,\\
&\|\Delta F\mathbf{e}_i\| \le (\varepsilon_1 + \|\delta Q_1\|)\|F\mathbf{e}_i\|, && r+1 \le i \le n,\\
&\widetilde{R}_{11} - R_{11} = GR_{11}, && \|G\|_F \le \|\delta Q_1\|_F + \eta,
\end{aligned}$$

where $\varepsilon_1 \leq f_1(p,r)\mathbf{u}$ bounds the roundoff.

Proof. Let $\widetilde{F}^{(r)}$ be the matrix obtained after r steps of the Householder QR factorization. Then there exist an orthogonal matrix $\widehat{Q}$ and a backward perturbation δF such that

$$\begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ O & \widetilde{R}_{22} \end{bmatrix} \equiv \widetilde{F}^{(r)} = \widehat{Q}^T(F+\delta F), \qquad \|\delta F\mathbf{e}_i\| \leq \varepsilon_1 \|F\mathbf{e}_i\|, \qquad 1 \leq i \leq n.$$

Our assumption on the numerical rank of F_1 implies that $F_1 + \delta F_1 = \widehat{Q}_1\widetilde{R}_{11}$ is the QR factorization with nonsingular $\widetilde{R}_{11}$. Now, setting $\widetilde{R}_{22}$ to zero is, in the backward error sense, equivalent to the QR factorization of a rank r matrix,

$$\widehat{Q}\begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ O & O \end{bmatrix} = F + \Delta F, \qquad \Delta F = \delta F - \widehat{Q}\begin{bmatrix} O & O \\ O & \widetilde{R}_{22} \end{bmatrix}.$$

It remains to estimate $\widehat{Q}_2\widetilde{R}_{22} = \widehat{Q}_2\widehat{Q}_2^T(F_2+\delta F_2)$. First note that $F_2 = Q_1R_{12}$, where the i-th column of R_{12} has the same norm as the corresponding column of F_2. Then,

$$\widehat{Q}_2\widehat{Q}_2^T F_2 = \widehat{Q}_2\widehat{Q}_2^T Q_1R_{12} = \widehat{Q}_2\widehat{Q}_2^T(\widehat{Q}_1 - \delta Q_1)R_{12} = -\widehat{Q}_2\widehat{Q}_2^T\delta Q_1R_{12}$$

and we can write

$$\|\widehat{Q}_2\widehat{Q}_2^T F_2\mathbf{e}_i\| \leq \|\delta Q_1\|\,\|F_2\mathbf{e}_i\|, \qquad 1 \leq i \leq n-r.$$

To estimate δQ_1, we first note that $F_1 = Q_1R_{11}$ and $F_1 + \Delta F_1 = \widehat{Q}_1\widetilde{R}_{11}$ imply that

$$\widehat{Q}_1 = (I + \Delta F_1F_1^+)Q_1(R_{11}\widetilde{R}_{11}^{-1}),$$

and that

$$R_{11}^{-T}\widetilde{R}_{11}^T\widetilde{R}_{11}R_{11}^{-1} = I + Q_1^T\Delta F_1R_{11}^{-1} + R_{11}^{-T}\Delta F_1^TQ_1 + R_{11}^{-T}\Delta F_1^T\Delta F_1R_{11}^{-1}.$$

Thus, $\widetilde{R}_{11}R_{11}^{-1}$ is the Cholesky factor of $I+E$, where

$$\|E\|_F \leq 2\|\Delta F_1R_{11}^{-1}\|_F + \|\Delta F_1R_{11}^{-1}\|_F^2.$$

Now, by [7], $\|E\|_F < 1/2$ implies that $\widetilde{R}_{11}R_{11}^{-1} = I + \Gamma$, where Γ is upper triangular and

$$\|\Gamma\|_F \leq \frac{\sqrt{2}\|E\|_F}{1+\sqrt{1-2\|E\|_F}} < \frac{1}{\sqrt{2}}.$$

Hence, $R_{11}\widetilde{R}_{11}^{-1} = I + \widehat{\Gamma}$, where $\|\widehat{\Gamma}\|_F \leq \|\Gamma\|_F/(1-\|\Gamma\|_F) < (2+\sqrt{2})\|\Gamma\|_F$. Since $\widehat{Q}_1 = Q_1 + Q_1\widehat{\Gamma} + \Delta F_1R_{11}^{-1} + \Delta F_1R_{11}^{-1}\widehat{\Gamma}$ we obtain

$$\|\delta Q_1\|_F \leq \|\widehat{\Gamma}\|_F + \|\Delta F_1R_{11}^{-1}\|_F + \|\widehat{\Gamma}\|\,\|\Delta F_1R_{11}^{-1}\|_F.$$

Finally note that $\widetilde{R}_{11} - R_{11} = (\widehat{Q}_1^T\delta F_1R_{11}^{-1} - \delta Q_1^TQ_1)R_{11}$. □

We remark that

$$\widetilde{R}_{12} = R_{12} + \delta Q_1^TQ_1R_{12} + \widehat{Q}_1^T\Delta F_2,$$

which means that we can nicely bound $\delta R_{12} = \widetilde{R}_{12} - R_{12}$. We have, for instance,

$$\|\delta R_{12}\mathbf{e}_i\| \leq (2\|\delta Q_1\| + \varepsilon_1)\|R_{12}\mathbf{e}_i\|, \qquad 1 \leq i \leq n-r.$$

If we use entry-wise backward analysis of the QR factorization ($|\delta F| \leq \varepsilon_2 \mathbf{e}\mathbf{e}^T|F_2|$, $\mathbf{e} = (1,\dots,1)^T$), then we can also write

$$|\delta R_{12}| \leq (|\delta Q_1^T Q_1| + \varepsilon_2 |\widehat{Q}_1|^T \mathbf{e}\mathbf{e}^T |Q_1|)|R_{12}|,$$

where the matrix absoulute values and inequalities are understood entry-wise, and ε_2 is defined similarly as ε_1.

From the above analysis we see that the error in the computed matrix $\widetilde{R}$ is bounded in the same way as in Theorem 4.1. Also, the QR factorization can be computed with the standard column pivoting and R_{11} can have additional structure just as in the Cholesky factorization of A_{11}. Therefore, the analysis of the backward null space perturbation based on $\widetilde{R}^T$ holds in this case as well. However, the bounds of Theorem 4.4 are sharper than those of Theorem 4.1.

5. Constrained systems of equations.

Let again be $\mathcal{N}(A) = \mathcal{R}(Y)$ with $Y \in \mathbb{R}^{n\times m}$ having full rank. Let $C \in \mathbb{R}^{n\times m}$ be a matrix with full rank. Systems of equations of the form

$$\begin{bmatrix} A & C \\ C^T & O \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix}, \tag{5.1}$$

appear at many occasions, e.g. in mixed finite element methods [3], or constrained optimization [12]. They have a solution for every right side if $\mathbb{R}^n = \mathcal{R}(A) \oplus \mathcal{R}(C)$ which is the case if $H := Y^T C$ is nonsingular. In computations of Stokes [3] or Maxwell equations [1] the second equation in (5.1) with $\mathbf{c} = \mathbf{0}$ imposes a divergence-free condition on the flow or electric field, respectively.

To obtain a solution of (5.1) we first construct a particular solution of the first block row. Pre-multiplying it by Y^T yields $\mathbf{y} = H^{-1}Y^T\mathbf{b}$. As $\mathbf{b} - C\mathbf{y} \in \mathcal{R}(A)$ we can proceed as in section 3 to obtain a vector $\tilde{\mathbf{x}}$ with $A\tilde{\mathbf{x}} = \mathbf{b} - C\mathbf{y}$. The solution $\mathbf{x}$ of (5.1) is obtained by setting $\mathbf{x} = \tilde{\mathbf{x}} + Y\mathbf{a}$ and determining $\mathbf{a}$ such that $C^T\mathbf{x} = \mathbf{c}$. Thus, $\mathbf{a} = H^{-T}(\mathbf{c} - C^T\tilde{\mathbf{x}})$.

This procedure can be described in an elegant way if a congruence transformation as in (6.2) is applied. Multiplying (5.1) by $W^T \oplus I_m$, cf. (2.2), yields

$$\begin{bmatrix} A_{11} & O & C_1 \\ O & O & H \\ C_1^T & H^T & O \end{bmatrix} \begin{pmatrix} \tilde{\mathbf{x}}_1 \\ \mathbf{a} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ Y^T\mathbf{b} \\ \mathbf{c} \end{pmatrix}, \qquad \begin{aligned} \tilde{\mathbf{x}}_1 &= \mathbf{x}_1 - Y_1 Y_2^{-1}\mathbf{x}_2, \\ \mathbf{a} &= Y_2^{-1}\mathbf{x}_2, \\ \mathbf{b}_1 &= I_{n,r}^T\mathbf{b}. \end{aligned} \tag{5.2}$$

Notice that $\tilde{\mathbf{x}}_1 \in \mathbb{R}^r$. From (5.2) we read that

$$\begin{array}{llll} \text{(i)} & \mathbf{y} = H^{-1}Y^T\mathbf{b}, & & \\ \text{(ii)} & \tilde{\mathbf{x}}_1 = A_{11}^{-1}(\mathbf{b}_1 - C_1\mathbf{y}), & \text{(iv)} & \mathbf{x}_1 = \tilde{\mathbf{x}}_1 + Y_1\mathbf{a}, \\ \text{(iii)} & \mathbf{a} = H^{-T}(\mathbf{c} - C_1^T\tilde{\mathbf{x}}_1), & \text{(v)} & \mathbf{x}_2 = Y_2\mathbf{a}. \end{array} \tag{5.3}$$

This geometric approach differs from the algebraic one based on the factorization

$$\begin{bmatrix} A_{11} & A_{12} & C_1 \\ A_{12}^T & A_{22} & C_2 \\ C_1^T & C_2^T & O \end{bmatrix} = \begin{bmatrix} R_{11}^T & O & O \\ R_{12}^T & I_m & O \\ C_1^T R_{11}^{-1} & O & I_m \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{11}^{-T}C_1 \\ O & O & C_2 - R_{12}^T R_{11}^{-T} C_1 \\ O & C_2^T - C_1^T R_{11}^{-1} R_{12} & -C_1^T R_{11}^{-1} R_{11}^{-T} C_1 \end{bmatrix}$$

where the LU factorization of $C_2 - R_{12}^T R_{11}^{-T} C_1$ is employed to solve (5.1). In the geometric approach the LU factorization of H is used instead. Of course, there is a close connection between the two approaches: Using (3.4) we get $C_2^T - C_1^T R_{11}^{-1} R_{12} = H^T Y_2^{-1}$. Notice that the columns of C or Y can be scaled such that the condition numbers of H or $C_2 - R_{12}^T R_{11}^{-T} C_1$ are not too big. Notice also that Y can be chosen such that $Y_2 = I_m$ in which case $C_2^T - C_1^T R_{11}^{-1} R_{12} = H^T$. A thorough perturbation analysis of (5.1)–(5.3) remains to be done in our future work.

Golub and Greif [9] use the algebraic approach to solve systems of the form (5.1) if the positive semidefinite A has a low-dimensional null space. As they do not have available a basis for the null space they apply a trial-and-error strategy for finding a permutation of A such that the leading $r \times r$ principal submatrix becomes nonsingular. They report that usually the first trial is successful. This is intelligible because $n_i = r + i = n - m + i$ if the basis of the null space is dense which is often the case.

If the null space of A is high-dimensional then Golub and Greif use an augmented Lagrangian approach. They modify (5.1) such that the $(1,1)$ block becomes positive definite,

$$\begin{bmatrix} A + C\Delta C^T & C \\ C^T & O \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b} + C\Delta\mathbf{c} \\ \mathbf{c} \end{pmatrix}.$$

Here, Δ is some symmetric positive definite matrix, e.g. a multiple of the identity. $A + C\Delta C^T$ is positive definite if $Y^T C$ is nonsingular. The determiniation of a good Δ is difficult. Golub and Greiff thoroughly discuss how to choose Δ and how the 'penalty term' $C\Delta C^T$ affects the condition of the problem. In contrast to this approach where a term is added to A that is positive definite on the null space of A, $\mathcal{N}(A)$ can be avoided right away if a basis of it is known.

6. Eigenvalue problems.

Let us consider the eigenvalue problem

$$A\mathbf{x} = \lambda M \mathbf{x}, \tag{6.1}$$

where A is symmetric positive semidefinite with $\mathcal{N}(A) = \mathcal{R}(Y)$ and M is symmetric positive definite. We assume that the last m rows of Y are linearly independent such that W in (2.1) is nonsingular. Then,

$$W^T A W = \begin{bmatrix} A_{11} & O \\ O & O \end{bmatrix}, \qquad W^T M W = \begin{bmatrix} M_{11} & C_1 \\ C_1^T & H \end{bmatrix} \tag{6.2}$$

where

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \qquad H = Y^T M Y = Y^T C.$$

Using the decomposition

$$W^T M W = \begin{bmatrix} M_{11} & C_1 \\ C_1^T & H \end{bmatrix} = P^T \begin{bmatrix} S & O \\ O & H \end{bmatrix} P, \qquad P = \begin{bmatrix} I & O \\ H^{-1} C_1^T & I \end{bmatrix}, \tag{6.3}$$

with the Schur complement $S := M_{11} - C_1 H^{-1} C_1^T$, and noting that $P^T W^T A W P = W^T A W$, it is easy to see that the positive eigenvalues of (6.1) are the eigenvalues of

$$A_{11}\mathbf{y} = \lambda (M_{11} - C_1 H^{-1} C_1^T)\mathbf{y} = \lambda S \mathbf{y}. \tag{6.4}$$

Notice that S is dense, in general, whence, in sparse matrix computations, it should not be formed explicitly.

If $\mathbf{y}$ is an eigenvector of (6.4) then

$$\mathbf{x} = P^{-1}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{y}\\ -H^{-1}C_1^T\mathbf{y}\end{pmatrix} \tag{6.5}$$

is an eigenvector of (6.1). By construction, $C^T\mathbf{x} = Y^TM\mathbf{x} = \mathbf{0}$, i.e., $\mathbf{x}$ is M-orthogonal to the null space of A.

We now consider the situation when A and M are given in factored form, $A = F^TF$ and $M = B^TB$, with $F = [F_1, F_2]$ and $B = [B_1, B_2]$ such that the rank of F_1 equals the rank of A. Let us find an implicit formulation of the reduced problem (6.4). With W from (2.1) we have $[F_1, F_2]W = [F_1, O]$. As before, $A_{11} = R_{11}^TR_{11}$, where R_{11} is computed by the QR factorization of F_1. It remains to compute a Cholesky factor of the Schur complement S, but directly from the matrix B. To that end we employ the QL factorization ('backward' QR factorization) of BW,

$$BW = QL = [Q_1, Q_2]\begin{bmatrix}L_{11} & O\\ L_{21} & L_{22}\end{bmatrix}, \qquad Q^TQ = I_n, \tag{6.6}$$

whence, with (6.3),

$$W^TMW = W^TB^TBW = \begin{bmatrix}L_{11}^TL_{11} + L_{21}^TL_{21} & L_{21}^TL_{22}\\ L_{22}^TL_{21} & L_{22}^TL_{22}\end{bmatrix} = \begin{bmatrix}M_{11} & C_1\\ C_1^T & H\end{bmatrix}. \tag{6.7}$$

Straightforward calculation now reveals that

$$S = M_{11} - C_1H^{-1}C_1^T = L_{11}^TL_{11}.$$

Thus, the eigenvalues of the matrix pencil (A_{11}, S) are the squares of the generalized singular values [10] of the matrix pair (R_{11}, L_{11}) or, equivalently, the squares of the singular values of $R_{11}L_{11}^{-1}$. An eigenvector $\mathbf{y}$ corresponds to a right singular vector $L_{11}\mathbf{y}$. The blocks L_{21} and L_{22} come into play when the eigenvectors of (6.1) are to be computed: using (6.7) equation (6.5) becomes

$$\mathbf{x} = \begin{pmatrix}\mathbf{y}\\ -L_{22}^{-1}L_{21}\mathbf{y}\end{pmatrix}.$$

It is known that the GSVD of (R_{11}, L_{11}) can be computed with high relative accuracy if the matrices $(R_{11})_c$ and $(L_{11})_c$ are well conditioned [6]. Here, $(R_{11})_c$ and $(L_{11})_c$ are obtained by R_{11} and L_{11}, respectively, by scaling their columns to make them of unit length. Obviously, $\kappa_2((R_{11})_c) = \kappa_2((F_1)_c)$, where $\kappa_2(\cdot)$ is the spectral condition number. It remains to determine $\kappa_2((L_{11})_c)$. From (6.6) we get

$$Q_1^TBW = Q_1^T[B_1, BY] = [L_{11}, O_{r,m}],$$

whence $Q_1^TB_1 = L_{11}$. Let the diagonal matrix D_1 be such that $(B_1)_c := B_1D_1^{-1}$ has columns of unit length. Further, let $(B_1)_c = U_1G_1$ be the QR factorization of $(B_1)_c$ and let $(L_{11})_s = L_{11}D_1^{-1} = Q_1^TU_1G_1$. As Q_1 is orthogonal we have $\|(L_{11})_s\| \leq \|(B_1)_c\| = \sigma_{\max}((B_1)_c)$. Further,

$$\|(L_{11})_s^{-1}\| \leq \|G_1^{-1}\|\|(Q_1^TU_1)^{-1}\| = \frac{1}{\sigma_{\min}((B_1)_c)\cos\Phi},$$

where Φ is the largest principal angle [10] between $\mathcal{R}(B_1)$ and $\mathcal{R}(B_2)^\perp \bigcap \mathcal{R}(B)$. Therefore,

$$\kappa_2((L_{11})_s) \leq \frac{\sigma_{\max}((B_1)_c)}{\sigma_{\min}((B_1)_c)\cos\Phi} = \frac{\kappa_2((B_1)_c)}{\cos\Phi}.$$

Since $\kappa_2((L_{11})_c) \leq \sqrt{r}\min_{D=\text{diagonal}}\kappa_2(L_{11}D)$ [13] [11, Thm.7.5], we have

$$\kappa_2((L_{11})_c) \leq \sqrt{r}\,\kappa_2((L_{11})_s) \leq \sqrt{r}\,\kappa_2((B_1)_c)/\cos\Phi. \tag{6.8}$$

So, we have identified condition numbers that do not depend on column scalings and that have a nice geometric interpretation. If the perturbations are column-wise small, then these condition number are the relevant ones.

7. Concluding remarks.

In this paper we have investigated ways to exploit the knowledge of an explicit basis of the null space of a symmetric positive semidefinite matrix.

We have considered consistent systems of equations, constrained systems of equations and generalized eigenvalue problems. First of all, the knowledge of a basis of the null space of a matrix A permits to extract *a priori* a maximal positive semidefinite submatrix. The rest of the matrix is redundant information and is needed neither for the solution of systems of equations nor for the eigenvalue computation. The order of the problem is reduced by the dimension of the null space. In iterative solvers it is not necessary to complement preconditioners with projections onto the complement of the null space.

Our error analysis shows that a backward stable positive semidefinite Cholesky factorization exists if the principal $r \times r$ submatrix, $r = \text{rank}(A)$, is well conditioned. This does however not mean that the computed Cholesky factor $\tilde{R}$ has a null space that is close to the known null space of R, $A = R^TR$. We observed that the backward error in the null space is small if the error in the Cholesky factor is (almost) orthogonal to the null space of A. We show that this is the case if the positive definite principal $r \times r$ submatrix after scaling is well conditioned and if its diagonal elements dominate those of the remaining diagonal block.

For systems of equations and eigenvalue problems, we considered the case when $A = F^TF$, where F is rectangular. This leads to interesting variants of the original algorithms and most of all leads to more accurate results.

What remains to be investigated is the relation between extraction of a positive definite matrix and fill-in during the Cholesky factorization. In future work we will use the new techniques in applications and, if possible, extend the theory to matrix classes more general than positive semidefinite ones.

REFERENCES

[1] P. Arbenz and R. Geus, *A comparison of solvers for large eigenvalue problems originating from Maxwell's equations*, Numer. Lin. Alg. Appl., 6 (1999), pp. 3–16.

[2] J. H. Argyris and O. E. Brønlund, *The natural factor formulation of the stiffness for the matrix displacement method*, Comput. Methods Appl. Mech. Engrg., 5 (1975), pp. 97–119.

[3] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991. (Springer Series in Computational Mathematics, 15).

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978. (Studies in Mathematics and its Applications, 4).

[5] J. Demmel, *On floating point errors in Cholesky*, Techn. Report CS-89-87, Computer Science Department, University of Tennessee, Knoxville, TN, October 1989. LAPACK Working Note 14. (Available at URL `http://www.netlib.org/lapack/lawns/`).

[6] Z. Drmač, *A tangent algorithm for computing the generalized singular value decomposition*, SIAM J. Numer. Anal., 35 (1998), pp. 1804–1832.

[7] Z. Drmač, M. Omladič, and K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

[8] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[9] G. H. Golub and C. Greif, *Techniques for solving general KKT systems*, Techn. Report SCCM-00-05, Stanford University, Scientific Computing/Computational Mathematics Program, July 2000. (Available at URL `http://www-sccm.stanford.edu/pub/sccm/`).

[10] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 3rd ed., 1996.

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

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999. (Springer Series in Operations Research).

[13] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.