Componentwise Analysis of Direct Factorization of Real Symmetric and Hermitian Matrices

Ivan Slapničar*

University of Split
Faculty of Electrical Engineering, Mechanical Engineering
and Naval Architecture
R. Boškovića b.b
21000 Split, Croatia
e-mail: ivan.slapnicar@fesb.hr

Submitted by Ludwig Elsner

ABSTRACT

We derive componentwise backward error bound for the factorization $H = GJG^T$, where H is a real symmetric matrix, G has full column rank, and J is diagonal with ±1's on the diagonal. We also derive componentwise forward error bound, that is we bound the difference between the exact and the computed factor G, in the cases where such bound is possible. We extend these results to the Hermitian case, and to the well-known Bunch–Parlett factorization. Finally, we prove bounds for the scaled condition of the matrix G, and show that the factorization can have rank revealing property.

1. INTRODUCTION

The $n \times n$ real symmetric matrix H can be decomposed as

$$H = GJG^T, \tag{1.1}$$

where G has full column rank, and $J = \mathrm{diag}\,(\pm 1)$. Further, there is a permutation matrix P such that the matrix PG is lower block triangular

*Part of this work is contained in the author's Ph.D. thesis which was done at the Fernuniversität Hagen. This work was also supported by the Grant No. 1-01-252 from the Croatian Ministry of Science and Technology.

655 Avenue of the Americas, New York, NY 10010 0024- 3795/93/$6.00

matrix with 1 × 1 and 2 × 2 diagonal blocks. The factorization (1.1) is a natural extension of the Cholesky factorization of a positive definite matrix,

$$H = LL^T \equiv LIL^T, \tag{1.2}$$

where L is lower triangular matrix, and I is the identity matrix. The indefinite factorization differs from the Cholesky factorization in three aspects: J instead of I, 2×2 diagonal blocks, and the permutation matrix P. The number of positive (negative) diagonal elements of J is equal to the number of positive (negative) eigenvalues of H. Existence of 2×2 diagonal blocks in the matrix PG is necessary since, in general, an indefinite matrix does not allow the factorization (1.2), even with J instead of I. As an example consider the matrix $H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The permutation matrix P ensures stability of the factorization, as we shall see later.

The factorization (1.1) is a modification of the well-known method by Bunch and Parlett [8]. The relationship between these two factorizations is as follows [27, 16]: the Bunch–Parlett method decomposes H as

$$PHP^T = LTL^T, \tag{1.3}$$

where P is permutation matrix, L is unit lower triangular matrix with full column rank, T is block-diagonal matrix with 1×1 and 2×2 blocks, and the diagonal blocks of L which correspond to 2×2 diagonal blocks of T are 2×2 identity matrices. This factorization is an extension of the LDL^T factorization of a positive definite matrix [19, 30]. Let $U^T T U = \Delta \equiv |\Delta|^{1/2} J |\Delta|^{1/2}$ be the eigenvalue factorization of T. Then

$$G = P^T LU|\Delta|^{1/2}. \tag{1.4}$$

The Bunch–Parlett method is well suited for solving symmetric systems of linear equations. In particular, the version of the Bunch–Parlett method with partial pivoting known as the Bunch–Kaufman method [6] is implemented in LAPACK [1]. The factorization (1.1) has recently attracted attention in two ways: first, eigenvalues of the pair $(G^T G, J)$ are the non-zero eigenvalues of H, and the factorization (1.1) followed by one-sided Jacobi type method on the pair (G, J) makes highly accurate eigenreduction algorithm [13, 31, 27]. Second, a version of the matrix GG^T is used as a good preconditioner for some indefinite systems of linear equations [16]. Factoring real symmetric and Hermitian matrices also has other important applications in eigenvalue problems, optimization and control. The inverse iteration method [26] which solves a sequence of linear systems by factoring $H - \lambda I$ is used to determine few eigenvectors of Hermitian matrix H. In optimization the so called augmented systems (or the Karush–Kuhn–Tucker systems) of the form $\begin{bmatrix} A & B^* \\ B & 0 \end{bmatrix}$ are used in several cases: in unconstrained least squares problems [2] where the augmented system approach has better numerical properties than the normal equation approach, in constrained least squares problems [14], and in general quadratic programming [17, 33]. The last application naturally extends to the minimization of general function with linear constraints, since the Newton step is computed from the local quadratic problem (see e.g. the review paper [33]). In control theory the above factorizations are used in solving algebraic Riccati equations, where the matrix sign function of the corresponding Hamiltonian matrix is computed by symmetric iterations [9, 24].

In this paper we give componentwise error bounds for the factorization (1.1). Our main result is the componentwise backward error bound: the computed G and J are the exact factors of the perturbed matrix $H + E$,

$$GJG^T = H + E \ , \qquad |E| \leq 91n(|H| + |G||G|^T)\varepsilon, \tag{1.5}$$

Here ε is the machine precision, and $|\cdot|$ stands for the elementwise absolute value. This bound compares well to the existing bound for the Bunch-Parlett method by Bunch [5]. Maximal predicted errors are in both cases similar and close to actual errors. Our bound reveals better the error structure, and has simpler form which is more suitable for further applications. For example, as a part of the error analysis of the above mentioned accurate eigenreduction algorithm, we can apply the relative perturbation theory for indefinite eigenvalue problems by Veselić and Slapničar [32, 28] to the bound (1.5), thus obtaining error bounds for eigenvalues and eigenvectors of H after the factorization [27]. Demmel and Veselić [13] used the same approach for positive definite matrices.

The bound (1.5) holds for complete pivoting. We also prove similar bound for the partial pivoting strategy which is used in the Bunch–Kaufman method [6, 7] and in the LAPACK routine *dsyjf2.f*. Further, we discuss normwise stability of the method.

If the matrix PG has only 1×1 diagonal blocks, then the bound (1.5) reduces to $|E| \leq 3n(|H| + |G||G|^T)\varepsilon$. This is, for example, always the case for positive definite and scaled diagonally dominant matrices [4]. Moreover, in both these cases the factorization (1.1) can be performed without pivoting, that is, with $P = I$. If H is positive definite and $P = I$, then (1.1) reduces to the Cholesky factorization, and the above bound is similar to the bounds by Demmel [11] and Sun [30].

Our second result is the componentwise forward error bound. First we need the forward perturbation result. Let $H = \widetilde{G}\widetilde{J}\widetilde{G}^T$ and $H + E = GJG^T$ be the factorizations of the unperturbed and perturbed matrix H. Since the factorization (1.1) involves pivoting, it is generally not possible to give reasonable bounds for the forward perturbation matrix $\delta G = G - \widetilde{G}$.

However, if we make additional assumptions that in both factorizations the same pivoting sequence and only 1×1 pivots have been used, and that $\widetilde{J} = J$, then we can bound the elements of δG in terms of G and E. This result generalizes the result by Sun [30, Theorem 2.2.1] for the Cholesky factorization. The forward error bound follows by inserting the backward error bound into the forward perturbation bound.

If H is Hermitian, then the algorithm and the error bounds for the factorization are similar to the ones for the real symmetric case. As a special case, we obtain componentwise backward and forward error bound for the Cholesky factorization of a Hermitian matrix.

Further, we derive similar results for the Bunch–Parlett factorization (1.3). In particular, these results hold for the LAPACK implementations the Bunch–Kaufman method [6], *dsytf2.f* and *chetf2.f* [1].

All above results can be viewed as generalizations of the results for LL^T and LDL^T factorizations of positive definite matrices by Sun [30] to indefinite real symmetric and Hermitian matrices.

Finally, we prove bounds for the *scaled condition* of the matrix G. The scaled matrix of G is defined by $\mathrm{scal}\,(G) = GD^{-1}$, where D is diagonal such that the columns of $\mathrm{scal}\,(G)$ have unit 2-norms. We prove a remarkable fact that $\kappa(\mathrm{scal}\,(G)) \leq O(n3.781^n)$ *irrespective* of the condition or even singularity of H. Here κ is the spectral condition number. If H is positive definite, then the bound is of order $O(n2^n)$, which can be almost attained. Both bounds hold for the Hermitian case, as well. As an application, we show that the factorization usually has non-trivial diagonalization effect and, consequently, rank reveling property.

The rest of the paper is organized as follows: in Section 2 we describe the algorithm of the factorization (1.1) in detail. In Section 3 we give the backward error analysis and discuss some special cases and normwise stability. In Section 4 we prove the forward error bound. In Section 5 we give the algorithm and the error bounds for the Hermitian case. In Section 6 we derive bounds for the scaled condition. In Section 7 we derive similar results for the Bunch–Parlett factorization. In Section 8 we summarize our results, and compare our backward error bounds with the existing analysis of the Bunch–Parlett method by Bunch [5]. We also describe results of numerical experiments, and illustrate our results by numerical example.

2. ALGORITHM

We shall now derive the algorithm of the factorization (1.1). We begin by describing the first step of the algorithm. Let H be a non-zero real symmetric matrix of order n. Let $\hat{P}$ be the permutation matrix such that

$$\hat{P}H\hat{P}^T = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}, \tag{2.1}$$

where X is nonsingular $k \times k$ matrix, $k \in \{1, 2\}$, C is a $(n-k) \times k$ matrix, and Y is a $(n-k) \times (n-k)$ matrix. Such $\hat{P}$ always exists because H is non-zero. Let $Q^T XQ = D$ be the eigenvalue factorization of X. If $k = 1$ then $Q = I_1$, and if $k = 2$ then

$$Q = \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix}, \qquad cs^2 + sn^2 = 1. \tag{2.2}$$

Thus, $X = Q|D|^{1/2}J_k|D|^{1/2}Q^T$, where $J_k = \mathrm{diag}\,(\pm 1)$, and we have

$$\hat{P}H\hat{P}^T = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix} = \begin{bmatrix} B & 0 \\ Z & I_{n-k} \end{bmatrix} \begin{bmatrix} J_k & 0 \\ 0 & \hat{H} \end{bmatrix} \begin{bmatrix} B & 0 \\ Z & I_{n-k} \end{bmatrix}^T, \tag{2.3}$$

where

$$B = Q|D|^{1/2}, \qquad Z = CQ|D|^{-1/2}J_k, \qquad \hat{H} = Y - ZJ_kZ^T.$$

The pivoting strategy is as follows: according to [8] we choose 1×1 pivot if and only if $\nu_1 \geq \alpha\nu_0$, where

$$\alpha = \frac{1+\sqrt{17}}{8}, \qquad \nu_0 = \max_{i \neq j} |H_{ij}|, \qquad \nu_1 = \max_i |H_{ii}|. \tag{2.4}$$

If we are performing 1×1 pivot, then we choose $\hat{P}$ in (2.1) to interchange row and column 1 with s, where s is the least integer such that $\nu_1 = |H_{ss}|$. Therefore, $|X| = \nu_1$. If we are performing 2×2 pivot, we can choose $\hat{P}$ in (2.1) according to several complete and partial pivoting strategies which are described in [8, 6, 7]. We shall use the unequilibrated diagonal pivoting from [8], that is, we choose $\hat{P}$ to interchange rows and columns 1 with q and 2 with p, where q is the least column integer and p is the least row integer in the qth column such that $\nu_0 = |H_{pq}|$. Note that $p > q$. This pivoting strategy implies that in the 1×1 case $J_1 = \mathrm{sign}\,(X)$, and in the 2×2 case X has one positive and one negative eigenvalue, that is, either $J_2 = \mathrm{diag}\,(1, -1)$ or $J_2 = \mathrm{diag}\,(-1, 1)$.

If H is non-singular, then by recursive application of (2.3) in the obvious manner we obtain the factorization

$$PHP^T = (PG)J(PG)^T, \tag{2.5}$$

where PG is a lower block triangular matrix, $J = \mathrm{diag}\,(\pm 1)$, and P is a permutation matrix. This, in turn, implies the factorization (1.1).

Pivoting strategy can be defined with some other $\alpha \in (0,1)$, as well. The case $\alpha \to 0$ ($\alpha \to 1$) corresponds to the use of 1×1 (2×2) pivot at each step [8], and both of these cases are clearly unstable. As shown in [8], the choice of α from (2.4) minimizes the element growth which can take place in the transition from H to $\hat{H}$ in (2.3), and the elements of the strict lower triangle of the matrix L from (1.3) are bounded as follows:

$$|L_{ij}| \leq \begin{cases} 1/\alpha & \text{for a } 1 \times 1 \text{ pivot,} \\ 1/(1-\alpha) & \text{for a } 2 \times 2 \text{ pivot.} \end{cases} \tag{2.6}$$

We now present algorithm in the Matlab notation:

Algorithm 2.1 (Symmetric indefinite factorization) *On entry, array H contains an $n \times n$ real symmetric matrix. On exit the first $r =$* rank(H) *columns of the array H are overwritten by the factor matrix G. Vector J contains the diagonal of the matrix J. Vector P describes the pivoting.*

```
/* Initialize starting values. */
$alpha = (1 + \mathbf{sqrt}(17))/8$
$r = n$
$P = 1 : n$
/* Main loop. */
$i = 1$
while $i <= n$
/* Find the current $\nu_0$ and $\nu_1$, and the indices $p$, $q$ and $s$. */
    $[temp, p] = \mathbf{max}(\mathbf{abs}(H(i:n, i:n) - \mathbf{diag}(\mathbf{diag}(H(i:n, i:n)))))$
    $[nu0, q] = \mathbf{max}(temp)$
    $p = p(q)$
    $[nu1, s] = \mathbf{max}(\mathbf{abs}(\mathbf{diag}(H(i:n, i:n))))$
    if $nu1 \geq alpha * nu0$
/* $1 \times 1$ pivot. If the current block is singular, then finish. */
        if $nu1 == 0$
            $r = i - 1$
            $i = n + 1$
        else
/* Permute $H$ such that $H(s,s)$ comes to the position $(i,i)$, and notify
    this in $P$. */
            $s = s + i - 1$
            $H([i\ s], :) = H([s\ i], :)$
            $H(i:n, [i\ s]) = H(i:n, [s\ i])$
            $P([i\ s]) = P([s\ i])$
/* Update $H$. */
            $J(i,i) = \mathbf{sign}(H(i,i))$
            $H(i,i) = \textbf{sqrt}(\textbf{abs}(H(i,i)))$
            **if** $i < n$
                $H(i+1:n,i) = H(i+1:n,i) * (J(i,i)/H(i,i))$
                $H(i,i+1:n) = \textbf{zeros}(1,n-i)$
                $H(i+1:n,i+1:n) = H(i+1:n,i+1:n) - \ldots$
                    $J(i,i) * H(i+1:n,i) * H(i+1:n,i)'$
            **end**
            $i = i+1$
        **end** /* $1 \times 1$ *pivot.* */
    **else**
/* $2 \times 2$ *pivot. Permute* $H$ *such that* $H(p,p)$ *and* $H(q,q)$ *come to the*
    *position* $(i+1,i+1)$ *and* $(i,i)$*, respectively. Notify this in* $P$*.* */
        $p = p+i-1$
        $q = q+i-1$
        $H([i\ q],:) = H([q\ i],:)$
        $H([i+1\ p],:) = H([p\ i+1],:)$
        $H(i:n,[i\ q]) = H(i:n,[q\ i])$
        $H(i:n,[i+1\ p]) = H(i:n,[p\ i+1])$
        $P([i\ q]) = P([q\ i])$
        $P([i+1\ p]) = P([p\ i+1])$
/* *Compute the* $2 \times 2$ *orthogonal matrix* $Q$*.* */
        $zeta = (H(i+1,i+1) - H(i,i))/(2 * H(i+1,i))$
        **if** $z == 0$
            $t = 1$
        **else**
            $t = \textbf{sign}(zeta)/(\textbf{abs}(zeta) + \textbf{sqrt}(zeta^2+1))$
        **end**
        $h = \textbf{sqrt}(1+t^2)$
        $cs = 1/h$
        $sn = t/h$
        $Q = [cs\ sn; -sn\ cs]$
/* *Update* $H$*.* */
        $a = H(i,i) - H(i+1,i) * t$
        $b = H(i+1,i+1) + H(i+1,i) * t$
        $J(i:i+1,i:i+1) = \textbf{diag}(\textbf{sign}([a\ b]))$
        $D = \textbf{sqrt}(\textbf{diag}(\textbf{abs}([a\ b])))$
        $H(i:i+1,i:i+1) = Q * D$
        **if** $i < n-1$
            $H(i+2:n,i:i+1) = H(i+2:n,i:i+1) * Q * \ldots$
                $J(i:i+1,i:i+1) * \textbf{inv}(D)$
            $H(i:i+1,i+2:n) = \textbf{zeros}(2,n-i-1)$
            $H(i+2:n,i+2:n) = H(i+2:n,i+2:n) - \ldots$
                $H(i+2:n,i:i+1) * J(i:i+1,i:i+1) * \ldots$
            $H(i+2:n,i:i+1)'$
        **end**
        $i = i + 2$
    **end** /* $2 \times 2$ pivot. */
**end** /* Main loop. */
/* Permute rows of $H$ to obtain the final factor. */
$H(P,:) = H$
```

The fact that the symmetry of the submatrices is lost in the above algorithm, does not influence the subsequent error analysis. The algorithm can easily be redefined to preserve symmetry, and to use only lower or upper part of the matrix H, which saves storage and reduces the operation count. We omit these enhancements for the sake of simplicity.

In some applications [31, 27] it is convenient to have the diagonal of J sorted, that is, first $+1$'s, then -1's, or vice versa. This is easily achieved by appropriately permuting the columns of G in (1.1). This permutation does not influence the error analysis.

If H is singular, then at some stage of the algorithm we shall have $\hat{H} = 0$. By taking only those columns of G and the elements of J which have so far been computed, we obtain the desired factorization (1.1).

If H is positive definite, then Algorithm 2.1 reduces to the Cholesky factorization with diagonal pivoting (see e.g. [13]).

3. BACKWARD ERROR ANALYSIS

In this section we give the backward error analysis of the symmetric factorization defined by Algorithm 2.1. In Section 3.1 we prove the error bound for partial pivoting. In Section 3.2 we give some comments about different implementations of the algorithm. In Section 3.3 we discuss normwise stability of the algorithm, and in Section 3.4 we specialize or main result for the case when only 1×1 pivots are used.

We first present our model of the finite precision floating-point arithmetic: floating-point result $\mathrm{fl}\,(\cdot)$ of the operation $(\cdot)$ is given by [13, 18, 19]

$$\begin{aligned} \mathrm{fl}\,(a \pm b) &= a(1+\varepsilon_1) \pm b(1+\varepsilon_2) \\ \mathrm{fl}\,(a \times b) &= (a \times b)(1+\varepsilon_3) \\ \mathrm{fl}\,(a/b) &= (a/b)(1+\varepsilon_4) \\ \mathrm{fl}\,(\sqrt{a}) &= \sqrt{a}(1+\varepsilon_5) \end{aligned} \tag{3.1}$$

where $|\varepsilon_i| \leq \varepsilon$, and $\varepsilon \ll 1$ is the machine precision. This is somewhat more general than the usual model which uses $\mathrm{fl}\,(a \pm b) = (a \pm b)(1+\varepsilon_1)$ and includes machines like the Cray which do not have a guard digit. If a and b have the same sign, then in our model we also have $\mathrm{fl}\,(a+b) = (a+b)(1+\varepsilon_1)$.

To make the analysis simpler we shall ignore the terms of order $O(\varepsilon^2)$, that is, we shall make the usual assumptions

$$(1+\varepsilon_1)(1+\varepsilon_2) = 1+\varepsilon_1+\varepsilon_2 = 1+\varepsilon', \quad \frac{1+\varepsilon_1}{1+\varepsilon_2} = 1+\varepsilon_1-\varepsilon_2 = 1+\varepsilon'',$$

where $|\varepsilon'|, |\varepsilon''| \leq 2\varepsilon$. Under additional realistic assumption on ε, say $\varepsilon \leq 0.0001$, we can bound the second order terms in terms of $O(\varepsilon)$, and the bound of the following theorem holds exactly but with slightly larger constant.

We shall also assume that no underflow or overflow occurs.

THEOREM 3.1. *Let G and J be the factors of a real symmetric matrix H computed by Algorithm 2.1 in floating-point arithmetic with precision ε. Then, with the relative error of order $O(\varepsilon)$,*

$$GJG^T = H + E\ , \qquad |E| \leq 91n(|H| + |G||G|^T)\varepsilon\ .$$

Proof. The proof is by induction on n. We use the approach from [19, Theorem 3.3.1]. It is easy to see that the theorem holds for all matrices of order 1. To start the induction, we must also analyze the case of the 2×2 pivot for $n=2$. Let $\widetilde{\zeta}$, $\widetilde{t}$, $\widetilde{cs}$, $\widetilde{sn}$, $\widetilde{a}$, $\widetilde{b}$, and $\widetilde{G}_{ij}$ denote the quantities computed by Algorithm 2.1 in exact arithmetic. We shall show that in the floating-point arithmetic these quantities are computed with small relative errors. From now on we assume that $|\varepsilon_i| \leq \varepsilon$ for all i.

We have

$$\zeta = \mathrm{fl}\left(\frac{H_{22}-H_{11}}{2H_{21}}\right) = \frac{H_{22}(1+\varepsilon_1) - H_{11}(1+\varepsilon_2)}{2H_{21}(1+\varepsilon_3)}(1+\varepsilon_4) = \widetilde{\zeta} + \varepsilon_\zeta,$$

where $|\varepsilon_\zeta| \leq 3\alpha\varepsilon$. The bound on ε_ζ follows from the fact that our pivoting strategy implies

$$|H_{21}| = \nu_0, \qquad \max\{|H_{11}|, |H_{22}|\} \leq \nu_1, \tag{3.2}$$

which, in turn, implies

$$|\widetilde{\zeta}| \leq \begin{cases} \alpha & \text{if } \mathrm{sign}\,(H_{11}) = -\mathrm{sign}\,(H_{22})\ , \\ \alpha/2 & \text{otherwise.} \end{cases} \tag{3.3}$$

Therefore, we have

$$\mathrm{fl}\,(1+\zeta^2) = (1+\varepsilon_5)(1+(\widetilde{\zeta}+\varepsilon_\zeta)^2(1+\varepsilon_6)) = (1+\widetilde{\zeta}^2)(1+\varepsilon'),$$

where, by solving the above equation for ε', and then bounding $|\varepsilon'|$ from above,

$$|\varepsilon'| \le 2|\varepsilon_\zeta\widetilde{\zeta}| + (|\varepsilon_5| + |\varepsilon_6|)\widetilde{\zeta}^2 + |\varepsilon_5| \le (8\alpha^2 + 1)\varepsilon \le 4.3\varepsilon.$$

Further, the equality

$$\begin{aligned} \mathrm{fl}\,(|\zeta| + \sqrt{1+\zeta^2}) &= (1+\varepsilon_7)(|\widetilde{\zeta} + \varepsilon_\zeta| + (1+\varepsilon_8)(1+\varepsilon'/2)\sqrt{1+\widetilde{\zeta}^2}) \\ &= (1+\varepsilon'')(|\widetilde{\zeta}| + \sqrt{1+\widetilde{\zeta}^2}) \end{aligned}$$

holds for some $|\varepsilon''| \le 7\varepsilon$, so that finally

$$\begin{aligned} t &= \mathrm{fl}\left(\frac{\mathrm{sign}\,(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}\right) = \widetilde{t}(1+\varepsilon_t), && |\varepsilon_t| \le 8\varepsilon, \\ cs &= \mathrm{fl}\,(1/\sqrt{1+t^2}) = \widetilde{cs}(1+\varepsilon_{cs}), && |\varepsilon_{cs}| \le 11\varepsilon, \\ sn &= \mathrm{fl}\,(t/\sqrt{1+t^2}) = \widetilde{sn}(1+\varepsilon_{sn}), && |\varepsilon_{sn}| \le 11\varepsilon. \end{aligned} \tag{3.4}$$

Let

$$a = \mathrm{fl}\,(H_{11} - H_{21}t) = \widetilde{a}(1+\varepsilon_a), \qquad b = \mathrm{fl}\,(H_{22} + H_{21}t) = \widetilde{b}(1+\varepsilon_b). \tag{3.5}$$

If $H_{11} = 0$ and/or $H_{22} = 0$ or $\mathrm{sign}\,(H_{11}) \ne \mathrm{sign}\,(H_{22})$, then both a and b are computed by adding numbers of the same sign, thus

$$|\varepsilon_a|, |\varepsilon_b| \le |\varepsilon_t| + 2\varepsilon \le 10\varepsilon. \tag{3.6}$$

If $H_{11} \ge H_{22} > 0$ or $0 > H_{22} \ge H_{11}$, then a is again computed by adding numbers of the same sign, so $|\varepsilon_a| \le 10\varepsilon$. By using (3.2 – 3.4), since $|H_{22}| < |H_{21}\widetilde{t}|$, we have

$$\begin{aligned} b &= H_{22}(1+\varepsilon_9) + (1+\varepsilon_{10})(1+\varepsilon_{11})(1+\varepsilon_t)H_{21}\widetilde{t} \equiv \widetilde{b}(1+\varepsilon_b), \\ |\varepsilon_b| &\le \frac{|H_{22}| + 10|H_{21}\widetilde{t}|}{|H_{21}\widetilde{t}| - |H_{22}|}\varepsilon \le \frac{11|H_{21}\widetilde{t}|}{|H_{21}\widetilde{t}| - |H_{22}|}\varepsilon \equiv \gamma_b\varepsilon \le 90\varepsilon. \end{aligned} \tag{3.7}$$

Here

$$\gamma_b \le \frac{11}{1 - \frac{\alpha}{|\widetilde{t}|}} \le \frac{11}{1 - \alpha(\alpha/2 + \sqrt{1+\alpha^2/4})}.$$

Similarly, if $H_{22} \ge H_{11} > 0$ or $0 > H_{11} \ge H_{22}$, then $|\varepsilon_b| \le 10\varepsilon$, and

$$|\varepsilon_a| \le \frac{|H_{11}| + 10|H_{21}\widetilde{t}|}{|H_{21}\widetilde{t}| - |H_{11}|} \equiv \gamma_a\varepsilon \le 90\varepsilon. \tag{3.8}$$

Thus, we conclude that in any case

$$|\varepsilon_a|, |\varepsilon_b| \leq \max\{10, \gamma_a, \gamma_b\}\varepsilon \leq 90\varepsilon. \tag{3.9}$$

This, for example, implies

$$\begin{aligned} G_{21} &= \mathrm{fl}\,(-sn\sqrt{|a|}) = \widetilde{G}_{21}(1+\varepsilon_G), \\ |\varepsilon_G| &\leq |\varepsilon_{sn}| + |\varepsilon_a|/2 + 2\varepsilon \leq 58\varepsilon, \end{aligned} \tag{3.10}$$

so we have

$$G = \widetilde{G} + \delta G, \qquad |\delta G| \leq 58|\widetilde{G}|\varepsilon. \tag{3.11}$$

Therefore,

$$\begin{aligned} GJG^T &= (\widetilde{G}+\delta G)J(\widetilde{G}+\delta G)^T = H + E, \\ |E| &\leq 2\cdot 58|\widetilde{G}||\widetilde{G}|^T\varepsilon + O(\varepsilon^2) = 116|G||G|^T\varepsilon + O(\varepsilon^2), \end{aligned} \tag{3.12}$$

and the theorem holds.

The induction step must also be analyzed separately for 1 × 1 and 2 × 2 pivot. We assume without loss of generality that the permutation matrix $\hat{P}$ from (2.1) and (2.3) is the identity. Let us first consider a 1 × 1 pivot[†], that is $k = 1$. Then (2.3) holds with

$$\begin{aligned} & B = \mathrm{fl}\,(|H_{11}|^{1/2}) = |H_{11}|^{1/2} + \delta B, \\ & |\delta B| \leq |H_{11}|^{1/2}\varepsilon, \\ & Z = \mathrm{fl}\,(CJ_1/B) = CJ_1|H_{11}|^{-1/2} + \delta Z, \\ & |\delta Z| \leq 2\varepsilon|C||H_{11}|^{-1/2}, \\ & \hat{H} = \mathrm{fl}\,(Y - ZJ_1Z^T) = Y - ZJ_1Z^T + \hat{F}, \\ & |\hat{F}| \leq 2\varepsilon(|Y| + |Z||Z|^T). \end{aligned} \tag{3.13}$$

By assumption the computed factors $\hat{G}$ and $\hat{J}$ of $\hat{H}$ satisfy

$$\hat{G}\hat{J}\hat{G}^T = \hat{H} + \hat{E}, \qquad |\hat{E}| \leq 91(n-k)\varepsilon(|\hat{H}| + |\hat{G}||\hat{G}|^T). \tag{3.14}$$

By setting $G = \begin{bmatrix} B & 0 \\ Z & \hat{G} \end{bmatrix}$, we have

$$G\begin{bmatrix} J_1 & \\ & \hat{J} \end{bmatrix}G^T = \begin{bmatrix} BJ_kB^T & BJ_kZ^T \\ ZJ_kB^T & ZJ_kZ^T + \hat{G}\hat{J}\hat{G}^T \end{bmatrix}. \tag{3.15}$$

[†] The analysis of the 1 × 1 case is similar to the one of [30, Theorem 2.1.1], although here the matrix H need not be positive definite.

By setting $J = J_k \oplus \hat{J}$ and using (3.13), we obtain

$$GJG^T = H + E, \qquad |E| \leq \left[\begin{array}{cc} 2|H_{11}| & 3|C|^T \\ 3|C| & |\hat{E} + \hat{F}| \end{array} \right] \varepsilon. \tag{3.16}$$

From (3.13) it also follows that

$$|\hat{H}| \leq (1 + 2\varepsilon)(|Y| + |Z||Z|^T).$$

By inserting this into (3.14), and adding the bound for $|\hat{F}|$ from (3.13), we have

$$|\hat{E} + \hat{F}| \leq (91(n-1) + 2)(|Y| + |Z||Z|^T + |\hat{G}||\hat{G}|^T)\varepsilon. \tag{3.17}$$

By inserting the above inequality into (3.16) we finally obtain

$$|E| \leq (91(n-1) + 3)(|H| + |G||G|^T)\varepsilon,$$

and the theorem holds.

Let us now consider a 2×2 pivot, that is, $k = 2$. Let H be partitioned as in (2.3). Let $\widetilde{Q}^T X \widetilde{Q} = \widetilde{D}$ be the exact spectral factorization of X, and let Q and D be the computed matrices $\widetilde{Q}$ and $\widetilde{D}$, respectively. The analysis for $n = 2$ also applies to Q and D, that is, (3.4) and (3.9) imply that

$$\begin{array}{rcll} Q & = & \widetilde{Q} + \delta Q, & |\delta Q| \leq 11|\widetilde{Q}|\varepsilon, \\ D & = & \widetilde{D} + \delta D, & |\delta D| \leq 90|\widetilde{D}|\varepsilon. \end{array} \tag{3.18}$$

Similarly to the 1×1 case, from (3.11) and (3.18) we conclude that (2.3) holds with

$$\begin{aligned} & B = \mathrm{fl}\,(Q|D|^{1/2}) = \widetilde{Q}|\widetilde{D}|^{1/2} + \delta B, \\ & |\delta B| \leq 58|\widetilde{Q}||\widetilde{D}|^{1/2}\varepsilon, \\ & Z = \mathrm{fl}\,(CQ|D|^{-1/2}J_2) = C\widetilde{Q}|\widetilde{D}|^{-1/2}J_2 + \delta Z, \\ & |\delta Z| \leq 60|C||\widetilde{Q}||\widetilde{D}|^{-1/2}\varepsilon, \\ & \hat{H} = \mathrm{fl}\,(Y - ZJ_2Z^T) = Y - ZJ_2Z^T + \hat{F}, \\ & |\hat{F}| \leq 3\varepsilon(|Y| + |Z||Z|^T). \end{aligned} \tag{3.19}$$

The induction assumption (3.14), (3.15), and (3.19) imply that

$$GJG^T = \left[\begin{array}{cc} X & C^T \\ C & Y \end{array} \right] + \left[\begin{array}{cc} \delta X & \delta C^T \\ \delta C & \hat{E} + \hat{F} \end{array} \right] \equiv H + E, \tag{3.20}$$

where

$$\delta C = \delta Z J_2 |\widetilde{D}|^{1/2} \widetilde{Q}^T + C\widetilde{Q}|\widetilde{D}|^{-1/2} J_2 \delta B^T. \tag{3.21}$$

From (3.12) it follows directly that

$$|\delta X| \leq 116|B||B|^T\varepsilon. \tag{3.22}$$

Further, as in the proof of (3.17), we have

$$|\hat{E}+\hat{F}| \leq (110(n-2)+3)(|Y|+|Z||Z|^T+|\hat{G}||\hat{G}|^T)\varepsilon, \tag{3.23}$$

so it remains to bound $|\delta C|$ in terms of $|Z||B|^T$ and $|C|$. From (3.21) and (3.19) we have

$$|\delta C| \leq 118|C||\widetilde{Q}||\widetilde{Q}|^T\varepsilon \leq 118(|C|+2\widetilde{cs}|\widetilde{sn}|\left[\begin{array}{cc} |C_{:2}| & |C_{:1}| \end{array}\right])\varepsilon, \tag{3.24}$$

where $C_{:j}$ denotes the $j-$th column of C, and

$$|Z||B|^T \geq |ZB^T| \geq |C\widetilde{Q}J_2\widetilde{Q}^T| - 118|C||\widetilde{Q}||\widetilde{Q}|^T\varepsilon + O(\varepsilon^2). \tag{3.25}$$

In both cases, $J_2 = \mathrm{diag}\,(1,-1)$ or $J_2 = \mathrm{diag}\,(-1,1)$, we have

$$\begin{aligned} |C\widetilde{Q}J_2\widetilde{Q}^T|_{:1} &= |(\widetilde{cs}^2-\widetilde{sn}^2)C_{:1} - 2\widetilde{cs}\widetilde{sn}C_{:2}| \\ |C\widetilde{Q}J_2\widetilde{Q}^T|_{:2} &= |-2\widetilde{cs}\widetilde{sn}C_{:1} - (\widetilde{cs}^2-\widetilde{sn}^2)C_{:2}|. \end{aligned}$$

Therefore,

$$|C\widetilde{Q}J_2\widetilde{Q}^T| \geq 2\widetilde{cs}|\widetilde{sn}|\left[\begin{array}{cc} |C_{:2}| & |C_{:1}| \end{array}\right] - (\widetilde{cs}^2-\widetilde{sn}^2)|C|. \tag{3.26}$$

Now (3.3) implies $|\tilde{t}| \geq 1/(\alpha+\sqrt{1+\alpha^2})$ which, in turn, implies

$$0 \leq \widetilde{cs}^2 - \widetilde{sn}^2 \leq \frac{\alpha}{\sqrt{1+\alpha^2}} \leq 0.539. \tag{3.27}$$

By inserting this, (3.26), and (3.25) into (3.24), and ignoring terms of order $O(\varepsilon^2)$, we obtain

$$\begin{aligned} |\delta C| &\leq 118\left(|C|+|Z||B|^T+0.539|C|\right)\varepsilon \\ &\leq 182(|C|+|Z||B|^T)\varepsilon. \end{aligned} \tag{3.28}$$

The theorem now follows by inserting this, (3.22), and (3.23) into (3.20). ■

Note that the theorem also holds if H is singular, that is, if Algorithm 2.1 encounters a zero submatrix at some step. In that case the error matrix $\hat{E}$ from the induction step of the proof equals zero at some stage of the factorization.

We can further reduce the bound of Theorem 3.1 as follows: for each 2×2 step, instead of using the worst-case bounds, we can compute the actual values of γ from (3.9), and $\widetilde{cs}^2 - \widetilde{sn}^2$ from (3.27). The fact that these quantities are computed by using t instead of $\tilde{t}$ is not important since this only contributes an error of $O(\varepsilon^2)$. By inserting these quantities into the rest of the proof we have

$$|\delta C| \leq \varepsilon_C \equiv \left(\left(13 + \frac{\gamma}{2}\right) \cdot 2 + 2\right) (1 + \widetilde{cs}^2 - \widetilde{sn}^2)\, \varepsilon,$$

and the theorem holds with 91 replaced by $\max\{\varepsilon_C/2\}$, where maximum is taken over all 2×2 steps. In numerical experiments this procedure usually reduces the constant 91 by four times.

3.1. Other Pivoting Strategies

We can easily obtain bounds for some α other than the one defined in (2.4). For example, if we set $\alpha = 1/2$, then Theorem 3.1 holds with 91 replaced by 41.

Theorem 3.1 holds for any pivoting strategy for which (3.3) holds when we apply a 2×2 step. Moreover, Theorem 3.1 holds for any pivoting strategy for which the tangents in the 2×2 steps can be accurately computed, although with different constants. In particular, we shall show that the theorem holds for the partial pivoting strategy used in the Bunch–Kaufman method [6, 7] which is implemented in the LAPACK routine *dsytf2.f* [1]. This strategy is of interest since it requires only $O(n^2)$ search, contrary to the unequilibrated diagonal pivoting that we use which requires $O(n^3)$ search. We have chosen the unequilibrated diagonal pivoting since (as already mentioned) it has better bounds for the element growth, and (2.6) makes it possible to bound the scaled condition in Section 6.

Let us now prove the backward error bound for the Bunch–Kaufman partial pivoting strategy. We first describe the pivoting strategy. Let λ be the absolute value of the absolutely largest off-diagonal element in the first column,

$$\lambda = \max_{i \geq 2} |H_{i1}|,$$

and let s be the least integer such that $\lambda = |H_{si}|$. Further, let σ be the absolute value of the absolutely largest off-diagonal element in the s-th column,

$$\sigma = \max_{i \neq s} |H_{is}|.$$

We have the following algorithm:

Algorithm 3.1 (Partial pivoting) *We describe only the first step of the pivoting strategy. The complete algorithm is obtained by combining this algorithm with Algorithm 2.1.*

```
determine $\lambda$
if $\lambda = 0$
    go to the next step
else
    if $|H_{11}| \geq \alpha\lambda$
        perform $1 \times 1$ pivot
    else
        determine $s$ and $\sigma$
        if $|H_{11}|\sigma \geq \alpha\lambda^2$
            perform $1 \times 1$ pivot
        else if $|H_{ss}| \geq \alpha\sigma$
            interchange rows and columns 1 with $s$
            perform $1 \times 1$ pivot
        else
            interchange rows and columns 2 with $s$
            perform $2 \times 2$ pivot
        end
    end
    go to the next step
end
```

THEOREM 3.2. *Let G and J be the factors of a real symmetric matrix H computed by symmetric indefinite decomposition with partial pivoting of Algorithm 3.1 in floating-point arithmetic with precision ε. Then Theorem 3.1 holds.*

Proof. The induction for 1×1 pivots is proved as in Theorem 3.1. Let us assume that we are performing 2×2 step. The conditions from Algorithm 3.1 imply

$$\begin{aligned} |H_{21}| &= \lambda, \\ |H_{11}| &< \alpha\frac{\lambda^2}{\sigma} \leq \alpha\lambda, \\ |H_{22}| &< \alpha\sigma, \\ |H_{11}||H_{22}| &< \alpha^2\lambda^2. \end{aligned} \tag{3.29}$$

If $n = 2$, then

$$\lambda = \sigma = |H_{21}|, \quad |H_{11}|, |H_{22}| < \alpha\lambda,$$

and the start of the induction is proved as in Theorem 3.1. Let us now assume that $n \geq 3$. If, in addition to (3.29), $|H_{22}| < \alpha\lambda$, then the induction step is proved as in Theorem 3.1. Let

$$|H_{11}| < \alpha\frac{\lambda^2}{\sigma} < \alpha\lambda \leq |H_{22}| < \alpha\sigma.$$

We consider two cases.

Case 1. If $\operatorname{sign}(H_{11}) \neq \operatorname{sign}(H_{22})$, then $\zeta = \widetilde{\zeta}(1+\varepsilon'_\zeta)$, where $|\varepsilon_\zeta| \leq 3\varepsilon$, thus the relations (3.4), (3.5), and (3.6) hold. We proceed as in the proof of Theorem 3.1, with the exception that (3.27) is replaced by the trivial bound $\widetilde{cs}^2 - \widetilde{sn}^2 \leq 1$. We finally obtain that (3.28) holds with 76 instead of 182, which completes the proof of this case.

Case 2. Let $\operatorname{sign}(H_{11}) = \operatorname{sign}(H_{22})$. Then

$$|\widetilde{\zeta}| \leq \frac{|H_{22}|}{2\lambda} \leq \frac{\alpha\sigma}{2\lambda}.$$

We consider two sub-cases.

Case 2A. If $\sigma/\lambda \leq 2$, then $|\widetilde{\zeta}| \leq \alpha$ and (3.4) holds. In (3.5), b is computed by adding numbers of the same sign, thus $|\varepsilon_b| \leq 10\varepsilon$. Further, ε_a is bounded by (3.8), where

$$\begin{aligned} \gamma_a &\leq \frac{11}{1-\frac{|H_{11}|}{\lambda|\widetilde{t}|}} \leq \frac{11}{1-\frac{\alpha\lambda}{\sigma}\left(\frac{\alpha\sigma}{2\lambda}+\sqrt{1+\left(\frac{\alpha\sigma}{2\lambda}\right)^2}\right)} \\ &= \frac{11}{1-\alpha\left(\frac{\alpha}{2}+\sqrt{\frac{\lambda^2}{\sigma^2}+\frac{\alpha^2}{4}}\right)} \leq 90. \end{aligned} \tag{3.30}$$

The rest of the proof is the same as the proof of Theorem 3.1.

Case 2B. If $\sigma/\lambda > 2$, then $|H_{11}|/|H_{22}| < 1/2$, and $\zeta = \widetilde{\zeta} + \varepsilon_\zeta$, where

$$|\varepsilon_\zeta| \leq \left(2 + \frac{|H_{22}|+|H_{11}|}{2\lambda}\cdot\frac{1}{|\widetilde{\zeta}|}\right)|\widetilde{\zeta}|\varepsilon \leq 5|\widetilde{\zeta}|\varepsilon.$$

Therefore, $\zeta = \widetilde{\zeta}(1+\varepsilon'_\zeta)$, where $|\varepsilon_\zeta| \leq 5\varepsilon$, and the relations (3.4) hold again. As above, $|\varepsilon_b| \leq 10\varepsilon$, and by inserting $\lambda/\sigma < 1/2$ into (3.30), we obtain $|\varepsilon_a| \leq \gamma_a\varepsilon \leq 27\varepsilon$. The rest of the proof is as in the proof of Theorem 3.1, with the exception that again (3.27) is replaced by $\widetilde{cs}^2 - \widetilde{sn}^2 \leq 1$. We obtain that (3.28) holds with 110 instead of 182, which completes the proof of the theorem. ■

Several pivoting strategies which ensure the normwise stability of the method (see Section 3.3) have recently been derived by Ashcraft, Grimes and Lewis in [3]. These strategies perform the number of searches for the pivot element that lies between complete pivoting of Algorithm 2.1 and partial pivoting of Algorithm 3.1. By combining Theorems 3.1 and 3.2 we see that our bound holds for these strategies, as well.

3.2. Different Implementations

In the case of 2 × 2 pivots, cs, sn, a and b can be computed by different formulas than those used in Algorithm 2.1. One can, for example, use the formulas which are used in the LAPACK auxiliary routine *dlaev2.f* which solves the 2×2 symmetric eigenvalue problem. Also, in the case of 2×2 pivots, the Schur complement $\hat{H}$ from (2.3) can be computed by using one symmetric rank two update

$$\hat{H} = CX^{-1}C^T,$$

instead of using two rank one updates $\hat{H} = Y - ZJ_kZ^T$, as in Algorithm 2.1. These modifications only slightly change the error analysis, so Theorems 3.1 and 3.2 still hold, but with slightly different constants. However, as noticed in [3], the use of two rank one updates can in real computations lead to unnecessary errors in some cases (see the illustrative example in [3]). Similarly, sometimes it is better to compute X^{-1} by using the direct inversion formula, $X^{-1} = \frac{1}{\det(X)} \begin{bmatrix} X_{22} & -X_{12} \\ -X_{12} & X_{11} \end{bmatrix}$, which is componentwise more accurate than the approach via the eigenvalue decomposition. We have chosen to use the rank one updates, which are also used by the current LAPACK implementation of the Bunch–Kaufman method *dsytf2.f*. This is due to the fact that BLAS [1] does not implement a symmetric rank two update yet. This will be cured in the next version of BLAS [12].

3.3. Normwise Stability

The standard definition of normwise stability is the following: factorization is considered normwise stable if the computed factorization is equal to the exact factorization of some matrix $H + E$ and $\|E\|/\|H\|$ is small in some norm. Such bounds have been proved for the Bunch–Parlett factorization with complete and partial pivoting. Bunch [5] proved that for the Bunch–Parlett method (1.3) with complete pivoting

$$\|E\|_1 \leq O(115n^3\varepsilon)\rho_n\|H\|_M, \tag{3.31}$$

where the M-norm is defined by $\|H\|_M = \max_{i,j}|H_{ij}|$, and ρ_n is the growth factor. The growth factor is defined by $\rho_n = \max_k \|H^{(k)}\|_M/\|H\|_M$, where $H^{(k)}$ is the Schur complement arising in the k-th stage of the factorization. The a priori upper bound for the element growth in this case is [5]

$$\rho_n \leq 3.07\sqrt{n}(n-1)^{0.446}f(n), \quad f(n) = \sqrt{\prod_{k=2}^{n} k^{\frac{1}{k-1}}} \leq 1.8n^{\frac{\log n}{4}}. \tag{3.32}$$

The analogous bound for the Bunch–Kaufman factorization (1.3) ([6], see also Section 7) with partial pivoting of Algorithm 3.1 has recently been proved by Higham [21]. He proved that

$$\||L||T||L|^T\|_M \leq 36n\rho_n\|H\|_M,$$

which, in turn, implies the normwise stability. For the partial pivoting the element growth is a priori bounded by $\rho_n \leq 2.57^{n-1}$ [6]. However, such large element growth is very rare in practice, and in [6] a simple and inexpensive algorithm for monitoring element growth is given.

Another possibility to ensure the normwise stability and have less search for pivot elements than complete pivoting, is to use some pivoting strategy which ensures that the elements of the matrix L from (1.3) are bounded. Namely, for complete pivoting (2.6) holds, but it is possible to have bounds similar to (2.6) without complete pivoting, as well. Such approach is used in [3] and in [14] for sparse matrices.

We shall now analyze normwise stability of the symmetric indefinite decompositions of Algorithms 2.1 and 3.1. We shall use the technique from [21]. Let us first analyze Algorithm 2.1. For simplicity, we assume that G and J are the exact factors of H, since by Theorem 3.1 this contributes only $O(\varepsilon^2)$ term in the final bound for $\|E\|$. Let

$$\begin{aligned} |G||G|^T &\equiv \begin{bmatrix} |B| & 0 \\ |Z| & |\hat{G}| \end{bmatrix} \begin{bmatrix} |B| & 0 \\ |Z| & |\hat{G}| \end{bmatrix}^T \\ &= \begin{bmatrix} |B||B|^T & |B||Z|^T \\ |Z||B|^T & |Z||Z|^T + |\hat{G}||\hat{G}|^T \end{bmatrix}, \end{aligned}$$

where B and Z are defined by (2.2) and (2.3), and $\hat{H} = \hat{G}\hat{J}\hat{G}$ is the factorization of the Schur complement $\hat{H}$. If the first pivot is 1×1, then $\nu_0/\nu_1 \leq 1/\alpha$ and

$$\begin{aligned} \||B||B|^T\|_M &= \|X\|_M \leq \|H\|_M, \\ \||B||Z|^T\|_M &= \|C\|_M \leq \|H\|_M, \\ \||Z||Z|^T\|_M &\leq \||C||D|^{-1}|C|^T\|_M \\ &\leq \frac{1}{\alpha}\|C\|_M \leq 2\|H\|_M. \end{aligned} \tag{3.33}$$

In the last inequality we have used the fact that $|D| = \nu_1$. If the first pivot is 2×2, then $|H_{21}| = \nu_0$ and

$$\begin{aligned} \||B||B|^T\|_M &\leq \||Q||D||Q|^T\|_M \leq \max\{|D_{11}|, |D_{22}|\} \\ &\leq \|X\|_1 \leq 2\|H\|_M, \\ \||B||Z|^T\|_M &\leq \||Q||Q|^T|C|\|_M \leq 2\|H\|_M, \\ \||Z||Z|^T\|_M &\leq \||C||Q||D|^{-1}|Q|^T|C|^T\|_M. \end{aligned} \tag{3.34}$$

It is easy to see that

$$[|Z||Z|^T]_{ij} \leq \frac{1}{|D_{11}||D_{22}|} \begin{bmatrix} \nu_0 & \nu_0 \end{bmatrix} \cdot \tag{3.35}$$

$$
\begin{aligned}
&\cdot \begin{bmatrix} cs^2|D_{22}| + sn^2|D_{11}| & \nu_0 \\ \nu_0 & cs^2|D_{11}| + sn^2|D_{22}| \end{bmatrix} \begin{bmatrix} \nu_0 \\ \nu_0 \end{bmatrix} \\
\leq\ & \frac{\nu_0^2}{|D_{11}||D_{22}|}(2\nu_0 + |D_{11}| + |D_{22}|) \leq 11\|H\|_M. \qquad (3.36)
\end{aligned}
$$

The last inequality follows from

$$D_{11}D_{22} = H_{11}H_{22} - \nu_0^2, \qquad |H_{11}||H_{22}| \leq \alpha^2\nu_0^2,$$

which combined gives $\nu_0^2/(|D_{11}||D_{22}|) \leq 1/(1-\alpha^2)$.

By applying the above bounds recursively to $\hat{G}$ and $\hat{H}$ and by noting that every Schur complement $\hat{H}$ satisfies $\|\hat{H}\|_M \leq \rho_n\|H\|_M$, where ρ_n is the growth factor, we conclude that $\||G||G|^T\|_M \leq 11n\rho_n\|H\|_M$. By using Theorem 3.1 and $|H| \leq |G||G|^T$, we finally have

$$\|E\|_M \leq 182 \cdot 11n^2\varepsilon\rho_n\|H\|_M,$$

where ρ_n is bounded by (3.32). Note that this bound is of the same order as (3.31).

Let us now analyze Algorithm 3.1. If the first pivot is 1 × 1, then (3.33) holds, with the exception that now

$$\||Z||Z|^T\|_M \leq \frac{\lambda^2}{|H_{11}|} \leq \max\left\{\frac{\lambda}{\alpha}, \frac{\sigma}{\alpha}\right\} \leq 2\|H\|_M,$$

if $|H_{11}| \geq \alpha\lambda$ or $|H_{11}|\sigma \geq \alpha\lambda^2$, and

$$\||Z||Z|^T\|_M \leq \frac{\sigma^2}{|H_{ss}|} \leq \frac{\sigma}{\alpha} \leq 2\|H\|_M,$$

if $|H_{ss}| \geq \alpha\sigma$. Therefore, we conclude that the method of Algorithm 3.1 is normwise stable if only 1 × 1 pivots are used. This includes some important classes of matrices which are described in Section 3.4.

However, if 2 × 2 pivots are used, the method of Algorithm 3.1 is not as stable as Algorithm 2.1 or the Bunch–Kaufman method. Namely, (3.34) holds, but with the exception that now

$$
\begin{aligned}
[|Z||Z|^T]_{ij} \ \leq\ & \frac{1}{|D_{11}||D_{22}|}\begin{bmatrix} \lambda & \sigma \end{bmatrix} \cdot \\
&\cdot \begin{bmatrix} cs^2|D_{22}| + sn^2|D_{11}| & \lambda \\ \lambda & cs^2|D_{11}| + sn^2|D_{22}| \end{bmatrix} \begin{bmatrix} \lambda \\ \sigma \end{bmatrix} \\
\leq\ & \frac{1}{|D_{11}||D_{22}|}((\lambda^2 + \sigma^2)\max\{|D_{11}|, |D_{22}|\} + 2\sigma\lambda^2),
\end{aligned}
$$

and even though $\lambda^2/(|D_{11}||D_{22}|) \leq 1/(1-\alpha^2)$, the bound such as (3.35) does not exist. The examples when such worst case is attained can be easily constructed by taking σ large enough.

3.4. Lower Triangular Factor

In the proof of Theorem 3.1 we see that 2×2 steps contribute much more to the error bound than 1×1 steps. If only 1×1 steps are performed, that is, if the factor PG from (2.5) is lower triangular, then the bound of Theorem 3.1 reduces to

$$|E| \leq 3n(|H| + |G||G|^T)\varepsilon. \tag{3.37}$$

Indeed, if only 1×1 steps are performed, then the constant 91 from the induction assumption (3.14) can be changed to 3, which combined with (3.17) gives (3.37). Also note that (3.37) holds always when only 1×1 pivots are used, *even without pivoting*, so long as the algorithm does not break down. Such factorization may, however, lead to large element growth, which will then be included in the $|G||G|^T$ term.

Two important classes of matrices which can be decomposed by performing only 1×1 steps without pivoting are positive definite matrices and scaled diagonally dominant matrices [4].

If H is positive definite, then Algorithm 2.1 reduces to the Cholesky factorization with diagonal pivoting, and only 1×1 steps are performed, so (3.37) holds. From the proof we see that (3.37) holds even if we do not use pivoting, in which case it closely resembles the results by Sun [30, Section 2]. Even more, by analyzing the proofs, we see that these results by Sun, which are slightly stronger than (3.37), hold for all indefinite matrices which can be decomposed by using only 1×1 pivots. Further, since $|H| + |G||G|^T \leq 2\sqrt{H_{ii}H_{jj}}$, which holds with the relative error of $O(\varepsilon)$, we have

$$|E_{ij}| \leq 6n(H_{ii}H_{jj})^{1/2}\varepsilon.$$

This is similar to the result by Demmel [11]. There the constant $6n$ is replaced by $(n+1)/(1-(n+1)\varepsilon)$, which is slightly better. Note, however, that the above bound holds for the outer product version of the Cholesky factorization [19, Algorithm 4.2.2] with or without diagonal pivoting, while Demmel analyzed the Gaxpy version [19, Algorithm 4.2.1].

Scaled diagonally dominant matrix is defined as $H = D(J+N)D$, where D is diagonal positive definite, $J = \mathrm{diag}\,(\pm 1)$, and N has zero diagonal with $\|N\|_2 < 1$ [4]. We shall now show that such matrix can be decomposed by performing only 1×1 steps both with or without pivoting, so that (3.37) holds. Indeed, the form of H implies that $H_{11} \neq 0$, thus we can start by performing a 1×1 step. Therefore, (2.3) implies

$$H = D(J+N)D = \begin{bmatrix} B & 0 \\ Z & I_{n-1} \end{bmatrix} \begin{bmatrix} J_1 & 0 \\ 0 & \hat{H} \end{bmatrix} \begin{bmatrix} B & 0 \\ Z & I_{n-1} \end{bmatrix}^T \equiv \tilde{L}\tilde{H}\tilde{L}^T.$$

Also, $\mathrm{sign}\,(H_{ii}) = \mathrm{sign}\,(\tilde{H}_{ii})$: for $i = 1$ this is obvious; for $i = 2, \cdots, n$ we have

$$\widetilde{H}_{ii} = D_{ii}^2 J_{ii} - N_{i1}^2 D_{ii}^2 J_{11} = D_{ii}^2 (J_{ii} - N_{i1}^2 J_{11}), \qquad (3.38)$$

and the statement follows from the fact that $|N_{i1}| < 1$. Thus, we can write $\widetilde{H} = \widetilde{D}(J + \widetilde{N})\widetilde{D}$, where $\widetilde{D}$ is diagonal positive definite, and $\widetilde{N}$ has zero diagonal. Since $J + N$ and $J + \widetilde{N}$ are congruent, they have the same inertia, and we conclude that $\|\widetilde{N}\|_2 < 1$. Therefore, $\widetilde{H}$ and, consequently, $\hat{H}$ are scaled diagonally dominant matrices, and by induction we conclude that the factorization can be continued by performing only 1×1 steps without pivoting.

Similar problem of performing LU factorization without pivoting was analyzed by Funderlic, Neumann and Plemmons in [15]. They showed that the Gaussian elimination can be performed without pivoting for generalized diagonally dominant matrices, that is, the matrices which can be row scaled to be diagonally dominant. By definition, H is such matrix if there exists a vector y such that

$$y > 0, \quad y^T (|\mathrm{diag}\,(H)| - |H - \mathrm{diag}\,(H)|) \geq 0, \qquad (3.39)$$

where the last two inequalities are interpreted componentwise. We shall restrict ourselves to strictly generalized diagonally dominant matrices, that is to the case where "$\geq$" in (3.39) is replaced by "$>$". Let now $H = D(J + N)D$, be a scaled diagonally dominant matrix. Then (3.39) (with "$>$" instead of "$\geq$") is equivalent to

$$y > 0, \quad y^T (I - |N|) > 0.$$

Notice that such y exists if and only if $I - |N|$ is an M-matrix [23, p. 114]. This implies that if $\bar{H}$ is strictly generalized diagonally dominant, then $\||N|\|_2 < 1$, which, in turn, implies that $\bar{H}$ is scaled diagonally dominant. The converse is not true, that is, there exist scaled diagonally dominant matrices which are not (strictly) generalized diagonally dominant. Indeed, let

$$\bar{H} = D(I + N)D, \quad N = \begin{bmatrix} 0 & 0.4 & -0.4 & -0.4 \\ 0.4 & 0 & 0.4 & -0.4 \\ -0.4 & 0.4 & 0 & 0 \\ -0.4 & -0.4 & 0 & 0 \end{bmatrix}.$$

Then $\|N\|_2 = 0.8$ and $\||N|\|_2 > 1$. We have therefore enlarged the class of symmetric matrices from [15] for which the Gaussian elimination can be performed without pivoting.

4. FORWARD ERROR BOUND

Forward error is defined as the matrix $\delta G = G - \widetilde{G}$, where $\widetilde{G}$ and G are the exact and the computed factors of a given matrix H, respectively. In this section we shall derive the componentwise forward perturbation bound, and then combine it with Theorem 3.1 to obtain the componentwise forward error bound. An example is given in Section 8.

Since the decomposition (1.1) and Algorithm 2.1 require pivoting, small relative componentwise perturbations of H can cause different permutations and different choices of 1×1 and 2×2 pivots. This implies that it is not, in general, possible to obtain useful bounds for δG. We illustrate this by a simple example: let $\widetilde{H} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and $H = \begin{bmatrix} 1+\varepsilon & 2 \\ 2 & 1 \end{bmatrix}$. Then $\widetilde{G} \approx \begin{bmatrix} 0.70711 & 1.22474 \\ -0.70711 & 1.22474 \end{bmatrix}$ and $G \approx \begin{bmatrix} 1.22474 & -0.70711 \\ 1.22474 & 0.70711 \end{bmatrix}$, and we see that even J need not remain the same. It is also easy to construct examples, where 1×1 pivots are used, but the permutation sequence changes, or where the inertia changes.

Our results are generalizations of the results by Sun [30, Sections 2.2 and 2.3] to indefinite matrices. In order to prove the following theorems we need some additional assumptions. These assumptions are the weakest possible, and our results apply to large class of matrices which includes e.g. scaled diagonally dominant matrices. We shall first prove the perturbation theorem:

THEOREM 4.1. *Let H and $H+E$ be non-singular matrices with the same inertia, and let $H = \widetilde{G} J \widetilde{G}^T$ and $H + E = GJG^T$. Set $\delta G = G - \widetilde{G}$. If the permutation sequences are in both decompositions the same, and if both matrices $P\widetilde{G}$ and PG from (2.5) are lower triangular, then*

$$|\delta G| \leq |G| \text{tril}\left(|G^{-1}||E||\widetilde{G}^{-1}|^T \right). \tag{4.1}$$

Here tril(A) *denotes the lower triangle of the matrix A. Moreover, if*

$$\epsilon \equiv \min_{D \in \mathcal{D}} \|(D\widetilde{G})^{-1}\|_2 \|(DG)^{-1}\|_2 \|DED\|_F < 1, \tag{4.2}$$

where $\mathcal{D}$ is the set of all $n \times n$ diagonal positive definite matrices, then

$$\begin{aligned} |\delta G| \;\leq\; & |G| \text{tril}\left(|G^{-1}||E||G^{-1}|^T \right) \\ & + |G| \text{tril}\left(|G^{-1}||E||G^{-1}|^T \Phi |G^{-1}||E||G^{-1}|^T \right), \end{aligned} \tag{4.3}$$

where $\Phi_{ij} = 1/(1-\epsilon^2)$.

Proof. Our proof is very similar to the proof by Sun [30, Theorem 2.2.1]. Matrices PG and $\widetilde{P}G$ are by assumption lower triangular, which implies that only 1×1 pivots have been used in both decompositions.

From Algorithm 2.1 we see that both matrices have all positive diagonal elements. From

$$PEP^T = (P\delta G)J(P\widetilde{G})^T + (PG)J(P\delta G)^T$$

we have

$$(PG)^{-1}PEP^T(P\widetilde{G})^{-T} = (PG)^{-1}(P\delta G)J + J(P\delta G)^T(P\widetilde{G})^{-T}.$$

Further, $(PG)^{-1}(P\delta G)J$ is lower triangular, $J(P\delta G)^T(P\widetilde{G})^{-T}$ is upper triangular, and

$$\text{sign}\,([(PG)^{-1}(P\delta G)J]_{ii}) = \text{sign}\,([J(P\delta G)^T(P\widetilde{G})^{-T}]_{ii}).$$

This implies that

$$|(PG)^{-1}(P\delta G)| \leq \text{tril}\,\left(|(PG)^{-1}PEP^T(P\widetilde{G})^{-T}|\right). \tag{4.4}$$

Finally, (4.1) follows by inserting this into $|\delta G| \leq |G||G^{-1}\delta G|$.

The inequality (4.4) also implies that for any $D \in \mathcal{D}$

$$\begin{aligned} \|G^{-1}\delta G\|_F &\leq \|G^{-1}E\widetilde{G}^{-T}\|_F \leq \|(DG)^{-1}DED(D\widetilde{G})^{-T}\|_F \\ &\leq \|(DG)^{-1}\|_2\|(D\widetilde{G})^{-T}\|_2\|DED\|_F \end{aligned}$$

The assumption (4.2) gives $\|G^{-1}\delta G\|_F \leq \epsilon < 1$, which implies that the matrix $Z = I - |G^{-1}\delta G|$ is invertible. The rest of the proof is as in [30, Theorem 2.2.1]. ■

As in [30, Section 2.3], the componentwise forward error bound is now obtained by inserting (3.37) into Theorem 4.1:

THEOREM 4.2. *Let $H = \widetilde{G}J\widetilde{G}^T$ be the decomposition of H computed by Algorithm 2.1 in exact arithmetic. Let G and J be the factors of H computed by Algorithm 2.1 in floating-point arithmetic with precision ε. Set $\delta G = G - \widetilde{G}$. If $\widetilde{G}$ and G satisfy the assumptions of Theorem 4.1, and if*

$$3n\varepsilon \min_{D\in\mathcal{D}} \|(D\widetilde{G})^{-1}\|_2\|(DG)^{-1}\|_2\|D(|H| + |G||G^T|)D\|_F < 1,$$

then

$$|\delta G| \leq 3n|G|\text{tril}\,\left(|G^{-1}|(|H| + |G||G^T|)|G^{-1}|^T\right)\varepsilon + O(\varepsilon^2).$$

5. HERMITIAN CASE

In this section we consider the decomposition of a Hermitian matrix H,

$$H = GJG^*, \tag{5.1}$$

where G has full column rank and $J = \mathrm{diag}\,(\pm 1)$. We derive the algorithm, and show that all results from previous sections hold here, as well.

The description of the algorithm is as in Section 2, except that the transposed matrices C^T, Q^T and Z^T are replaced by conjugate transposed matrices C^*, Q^*, and Z^*, respectively. Also, the matrix Q from (2.2) is now

$$Q = \begin{bmatrix} cs & \overline{sn} \\ -sn & cs \end{bmatrix}, \qquad cs^2 + |sn|^2 = 1 \ . \tag{5.2}$$

The elements of Q and D are computed similarly as in the auxiliary routine *claev2.f*, which is used in the Hermitian implementation of the Bunch-Parlett method (1.3) in the Lapack routine *chetf2.f* [1]. Since in Matlab $H' = H^T$ if H is real, and $H' = H^*$ if H is complex, the only parts in Algorithm 2.1 which change are the computation of Q, a, and b.

ALGORITHM 5.1 (HERMITIAN INDEFINITE DECOMPOSITION) *On entry array H contains an $n \times n$ complex Hermitian matrix. On exit the first $r = \mathrm{rank}\,(H)$ columns of the array H are overwritten by the factor matrix G. Vector J contains the diagonal of the matrix J. Vector P describes the pivoting. Only part which differs from Algorithm 2.1 is displayed.*

```
/* Compute the $2 \times 2$ orthogonal matrix $Q$. */
    $phi = H(i+1,i)/\mathbf{abs}(H(i+1,i))$
    $zeta = (H(i+1,i+1) - H(i,i))/(2 * \mathbf{abs}(H(i+1,i)))$
    if $z == 0$
        $t = 1$
    else
        $t = \mathbf{sign}(zeta)/(\mathbf{abs}(zeta) + \mathbf{sqrt}(zeta^2 + 1))$
    end
    $h = \mathbf{sqrt}(1 + t^2)$
    $cs = 1/h$
    $sn = t * phi/h$
    $Q = [cs \ \mathbf{conj}(sn); -sn \ cs]$
/* Update $H$. */
    $a = H(i,i) - \mathbf{abs}(H(i+1,i)) * t$
    $b = H(i+1,i+1) + \mathbf{abs}(H(i+1,i)) * t$
```

All comments about Algorithm 2.1 hold here, as well.

In order to prove error bounds, we first need to describe complex finite precision floating-point arithmetic. All subscribed and superscribed ε's denote complex numbers, ε denotes the machine precision, and we assume that $|\varepsilon_i| \leq \varepsilon$ for all i. It is easy to see that for $\gamma \in \mathbb{R}$ and $a, b \in \mathbb{C}$ the real model (3.1) implies

$$\mathrm{fl}\,(a \times \gamma) \quad = \quad (a \times \gamma)(1 + \varepsilon_1)$$

$$\begin{aligned} \mathrm{fl}\,(a/\gamma) &= (a/\gamma)(1+\varepsilon_2) \\ \mathrm{fl}\,(a \pm b) &= a(1+\varepsilon_3) \pm b(1+\varepsilon_4) \\ \mathrm{fl}\,(a \times b) &= (a \times b)(1+2\varepsilon_5) \end{aligned}$$

The backward error bound is given by the following theorem.

THEOREM 5.1. *Let G and J be the factors of a Hermitian matrix H computed by Algorithm 5.1 in floating-point arithmetic with precision ε. Then, with the relative error of order $O(\varepsilon)$,*

$$GJG^* = H + E\ , \qquad |E| \le 126n(|H| + |G||G|^T)\varepsilon\ .$$

Proof. The proof is very similar to the proof of Theorem 3.1, so we state only parts where the proofs differ. First, all transposed matrices should be replaced by conjugate transposed matrices, where applicable.

Let

$$H_{21} = |H_{21}|\phi = \sqrt{[\mathrm{Re}(H_{21})]^2 + [\mathrm{Im}(H_{21})]^2}.$$

Then

$$\begin{aligned} \mathrm{fl}\,(|H_{21}|) &= |H_{21}|(1+\varepsilon_{|H_{21}|}), \qquad |\varepsilon_{|H_{21}|}| \le 2\varepsilon, \\ \mathrm{fl}\,(\phi) &= \mathrm{fl}\left(\frac{H_{21}}{|H_{21}|}\right) = \widetilde{\phi}(1+\varepsilon_\phi), \qquad |\varepsilon_\phi| \le 3\varepsilon. \end{aligned}$$

This, (3.2), and (3.3) imply that

$$\zeta = \mathrm{fl}\left(\frac{H_{22} - H_{11}}{2|H_{21}|}\right) = \widetilde{\zeta} + \varepsilon_\zeta, \qquad |\varepsilon_\zeta| \le 5\alpha\varepsilon,$$

which further implies $|\varepsilon'| \le (12\alpha^2 + 1)\varepsilon$ and $|\varepsilon''| \le 9\varepsilon$. Thus, we conclude that (3.4) holds with

$$|\varepsilon_t| \le 10\varepsilon, \qquad |\varepsilon_{cs}| \le 13\varepsilon, \qquad |\varepsilon_{sn}| \le 13\varepsilon + |\varepsilon_\phi| + \varepsilon \le 17\varepsilon.$$

If $H_{11} = 0$ and/or $H_{22} = 0$ or $\mathrm{sign}\,(H_{11}) \ne \mathrm{sign}\,(H_{22})$, then

$$a = \mathrm{fl}\,(H_{11} - |H_{21}|t) = \widetilde{a}(1+\varepsilon_a), \qquad b = \mathrm{fl}\,(H_{22} + |H_{21}|t) = \widetilde{b}(1+\varepsilon_b), \quad (5.3)$$

holds with $|\varepsilon_a|, |\varepsilon_b| \le |\varepsilon_t| + |\varepsilon_{|H_{21}|}| + 2\varepsilon \le 14\varepsilon$. If $H_{11} \ge H_{22} > 0$ or $0 > H_{22} \ge H_{11}$, then (5.3) holds with

$$|\varepsilon_a| \le 14\varepsilon, \qquad |\varepsilon_b| \le \frac{|H_{22}| + 14|H_{21}||\widetilde{t}|}{|H_{21}||\widetilde{t}| - |H_{22}|}\varepsilon \le 122\varepsilon,$$

and if $H_{22} \ge H_{11} > 0$ or $0 > H_{11} \ge H_{22}$, then (5.3) holds with

$$|\varepsilon_b| \le 14\varepsilon, \qquad |\varepsilon_a| \le \frac{|H_{11}| + 14|H_{21}||\widetilde{t}|}{|H_{21}||\widetilde{t}| - |H_{11}|} \le 122\varepsilon.$$

Therefore, (3.11) holds with

$$|\delta G| \leq (|\varepsilon_{sn}| + \max\{|\varepsilon_a|, |\varepsilon_b|\}/2 + 2\varepsilon)|\widetilde{G}| \leq 80|\widetilde{G}|\varepsilon,$$

and (3.12) holds with $|E| \leq 160|G||G|^T\varepsilon + O(\varepsilon^2)$, which completes the start of the induction.

The proof of the induction step for a 1 × 1 pivot is as in Theorem 3.1, except that now (3.13) holds with $|\hat{F}| \leq 3\varepsilon(|Y| + |Z||Z|^T)$.

The proof of the induction step for a 2 × 2 pivot is as in Theorem 3.1 with the following changes: (3.19) holds with

$$|\delta B| \leq 80|\widetilde{Q}||\widetilde{D}|^{1/2}\varepsilon, \quad |\delta Z| \leq 83|C||\widetilde{Q}||\widetilde{D}|^{-1/2}\varepsilon, \quad |\hat{F}| \leq 4\varepsilon(|Y|+|Z||Z|^T),$$

and (3.26) and (3.27) hold with $\widetilde{sn}^2$ replaced by $|\widetilde{sn}|^2$, so that finally

$$|\delta C| \leq 163(1 + 0.539)(|C| + |Z||B|^T)\varepsilon \leq 251(|C| + |Z||B|^T)\varepsilon. \quad \blacksquare$$

The computational effort in searching for $\nu_0 = \max_{i\neq j}|H_{ij}|$ can be reduced by using 1-norm instead of 2-norm. That is, we can set $\nu_0 = \max_{i\neq j}(|\mathrm{Re}(H_{ij})| + |\mathrm{Im}(H_{ij})|)$. Such approach is used in the Lapack routine *chetf2.f*. Since the two norms differ by at most factor $\sqrt{2}$, Theorem 5.1 also holds for the above choice of ν_0, but with slightly larger constant.

All comments from Sections 3.1, 3.2, 3.3 and 3.4 apply to Theorem 5.1, as well. From the proof we see that if the matrix is decomposed by using only 1 × 1 pivots, then the elements of E are again bounded by (3.37). In particular (3.37) bounds the componentwise backward error for the Cholesky decomposition of a Hermitian semi-definite matrix. Also, scaled diagonally dominant Hermitian matrix can be decomposed by using only 1 × 1 steps with or without pivoting. The proof of this fact is as in Section 3.4, with the exception that in (3.38) the term N_{i1}^2 should be substituted by $|N_{i1}|^2$.

Finally, let us consider forward error bounds. From the proof of Theorem 4.1 we see that the forward componentwise perturbation bounds (4.1) and (4.3) hold for the Hermitian decomposition (5.1). By combining (4.3) with (3.37), we see that the Theorem 4.2 also holds for the Hermitian decomposition. For example, these results hold for the Hermitian Cholesky decomposition and scaled diagonally dominant Hermitian matrices.

6. BOUNDS FOR THE SCALED CONDITION

Let $H = GJG^T$ be the factorization of a real symmetric $n \times n$ matrix H, where G has full column rank and $J = \mathrm{diag}(\pm 1)$, and let us define the matrix $\mathrm{scal}(G)$ by

$$G = \mathrm{scal}(G)D, \qquad D_{ii} = \|G_{:i}\|_2, \qquad D_{ij} = 0, \text{ for } i \neq j. \tag{6.1}$$

The matrix scal (G) is the scaled matrix of the matrix G, and its condition $\kappa(\mathrm{scal}\,(G)) \equiv ||\mathrm{scal}\,(G)||_2||(\mathrm{scal}\,(G))^{-1}||_2$ is the scaled condition of the matrix G. Note that the columns of scal (G) have unit norms. According to the result by van der Sluis [29, Theorem 4.1], such scaling is almost the best possible over all diagonal scalings, that is,

$$\kappa(\mathrm{scal}\,(G)) \leq \sqrt{n} \min_{\Delta=\mathrm{diag}} \kappa(G\Delta). \tag{6.2}$$

Demmel and Veselić [13] proved a remarkable fact that if H is positive semi-definite and the factorization $H = GG^T$ is obtained by the Cholesky decomposition with complete pivoting, $\kappa(\mathrm{scal}\,(G))$ is bounded by a function of n irrespective of the condition or even singularity of H. However, their bound is, as they stated, a large overestimate. Here we show that a much better bound from [25, (6.13)], which is essentially almost attainable, readily applies here, and extend the result to indefinite, possibly singular, matrices. By combining these results, numerical evidence, and the perturbation results of [13] and [32], we show that the indefinite decomposition usually has diagonalization effect and rank revealing property. All results also hold for Hermitian matrices.

In the positive semi-definite case we have the bound

$$\kappa(\mathrm{scal}\,(G)) \leq \frac{\sqrt{n}}{3}(4^2 + 6n - 1)^{-1/2}. \tag{6.3}$$

which follows from [25, (6.13)]. Indeed, if H in positive definite, then the fact that we are performing Cholesky decomposition with complete pivoting implies that the matrix PG is such that

$$[PG]_{ii}^2 \geq \sum_{k=i}^{j} [PG]_{jk}^2, \quad i = 1, \ldots, n-1, \quad , j > i. \tag{6.4}$$

Therefore, scal (PG) is equal to both matrices A and R from [25, (6.13)], so

$$||(\mathrm{scal}\,(PG))^{-1}||_2 \leq \frac{1}{3}(4^2 + 6n - 1)^{-1/2}.$$

Here we have also used the fact that $[\mathrm{scal}\,(PG)]_{nn} = 1$. Combining the above inequality with $||\mathrm{scal}\,(PG)||_2 \leq \sqrt{n}$ gives (6.3).

By inspecting the proof of [25, (6.13)] it can be seen that the proof also applies to singular H, and the bounds are even better since some summations have fewer terms. The full proof of this result is in [27]. Similar proof was also used by Higham [20] in a different context.

The bound (6.3) is almost attained, as we see in the following example due to Kahan [25]: let $H = LL^T$, where

$$L = \begin{bmatrix} 1 & & & & \\ -c & 1 & & & \\ -c & -c & 1 & & \\ \vdots & & & \ddots & \\ -c & -c & \cdots & -c & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ & s & & & \\ & & s^2 & & \\ & & & \ddots & \\ & & & & s^{(n-1)} \end{bmatrix}, \quad c^2 + s^2 = 1.$$

Then L is itself the Cholesky factor with complete pivoting of $\widetilde{H}$, and when $c \to 1$, then $\kappa(H) \to \infty$, while $(\text{scal}\,(L))^{-1}$ tends to the matrix $\widetilde{D}\widetilde{L}$ from the proof of the theorem. The examples like this are, however, very rare, and $\kappa(\text{scal}\,(G))$ is usually much smaller, typically $O(n)$. Even the above example can be improved: if we apply Algorithm 2.1 to the permuted matrix PHP^T, where P swaps the first and the last row and column, we get $\kappa(\text{scal}\,(G)) \leq n$ – see also the related result of Hong and Pan [22].

Demmel and Veselić [13, Proposition 2.10] proved the following result for a positive definite matrix H: if $H = DAD$, where D is diagonal such that $A_{ii} = 1$, then

$$\lambda_{min}(A) \leq \frac{\lambda_i}{h_i} \leq \lambda_{max}(A),$$

where λ_i are the eigenvalues of H and h_i are the diagonal entries of H, both sorted in the ascending order.

If H is positive semi-definite, then matrices $H = GG^T$ and G^TG have the same nonzero eigenvalues. By applying the above inequality to the matrix G^TG, we obtain

$$\sigma^2_{min}(\text{scal}\,(G)) \leq \frac{\lambda_i}{h_i} \leq \sigma^2_{max}(\text{scal}\,(G)), \tag{6.5}$$

where σ_{min} and σ_{max} are the minimal and the maximal elements from the spectrum of $\text{scal}\,(G)$, λ_i are the nonzero eigenvalues of H, and h_i are the diagonal entries of G^TG (squares of the norms of the columns of G), both sorted in ascending order. The above relation holds, of course, for any factor G. If G is obtained by the Cholesky decomposition with complete pivoting, then, by combining the above relation with (6.3) and the fact mentioned above that $\kappa(\text{scal}\,(G))$ is usually very small (even if H is singular), we conclude that the decomposition with complete pivoting usually has strong diagonalization effect. Also, by looking only at the small eigenvalues, we conclude that such decomposition usually has rank reveling property. This property is similar to the one of QR factorization with complete column pivoting as described by Chan [10] and Hong and Pan [22]. From the previous example we also conclude that in some cases the complete diagonal pivoting does not produce good results. This, too, corresponds to the results from [10, 22], where rank revealing QR factorization requires some additional information about singular vectors of small singular values in order to find satisfactory pivoting sequence.

Let us now turn to the indefinite case.

THEOREM 6.1. *Let $H = GJG^T$ be the decomposition of a symmetric matrix H obtained by Algorithm 2.1 in exact arithmetic. Then*

$$\kappa(\mathrm{scal}\,(G)) \le \sqrt{n + 15n^2}\, 3.781^n. \tag{6.6}$$

Proof. Assume that H is non-singular. From (1.4) we see that the matrix $PG = LU|\Delta|^{1/2}$ is lower block triangular with 1×1 and 2×2 diagonal blocks. Here U is orthogonal and block diagonal, and L is unit lower triangular. According to (2.6), the under diagonal elements of L are bounded by $|L_{ij}| \le 2.781 \equiv \mu$. By using the monotonicity property of the 2-norm,

$$|A_{ij}| \le B_{ij} \Longrightarrow \|A\|_2 \le \|B\|_2,$$

we have

$$\|(GD^{-1})^{-1}\|_2 = \|(PGD^{-1})^{-1}\|_2 \le \|\widetilde{D}|U|^T\bar{L}\|_2 \le \|\widetilde{D}\widetilde{L}\|_2,$$

where

$$\begin{aligned} \bar{L}_{ij} &= \begin{cases} 1, & i = j, \\ \mu(1+\mu)^{i-1-j}, & i > j, \\ 0, & i < j, \end{cases} \\ \widetilde{L}_{ij} &= \begin{cases} 1+\mu, & i = j, \\ \mu(2+\mu)(1+\mu)^{i-1-j}, & i > j, \\ 1, & i = j-1, \\ 0, & i < j-1, \end{cases} \end{aligned}$$

and $\widetilde{D}$ is diagonal with $\widetilde{D}_{ii} = \sqrt{1 + 2(n-i)\mu^2}$. Therefore,

$$\begin{aligned} &\|(GD^{-1})^{-1}\|_2^2 \le \mathrm{trace}(\widetilde{D}\widetilde{L}\widetilde{L}^T\widetilde{D}) \\ &= \sum_{i=1}^{n} \left[1 + (1+\mu)^2 + \sum_{j=1}^{i-1} \Big(\mu(1+\mu)^{(i-1-j)}(2+\mu)\Big)^2 \right] (1 + 2(n-i)\mu^2) \\ &= \sum_{i=1}^{n} \Big[1 + \mu(2+\mu)((1+\mu)^{2(i-1)} - 1) \Big] (1 + 2(n-i)\mu^2) \\ &\le (1 + 2n\mu^2)(1+\mu)^{2n}, \end{aligned}$$

and the theorem follows by using this and $\|GD^{-1}\|_2 \le \sqrt{n}$. It is easy to see that the theorem holds for singular H, as well. ■

Note that the optimal value of α in (2.6) is $1/2$, in which case the theorem holds with 3.781 replaced by 3. As in the positive definite case, numerical experiments show that $\kappa(\mathrm{scal}\,(G)$ is usually very small, typically $O(n)$.

We shall generalize (6.5) to the indefinite case.

THEOREM 6.2. *Let $H = GJG^T$ be the decomposition of a symmetric matrix H obtained by Algorithm 2.1 in exact arithmetic. Then (6.5) holds, where now λ_i are the nonzero eigenvalues of H, and h_i are the diagonal elements of G^TGJ, both sorted in ascending order.*

Proof. Let $r = \mathrm{rank}(H)$, and let

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k < 0 < \lambda_{k+1} \leq \cdots \lambda_r$$

be the nonzero eigenvalues of H. Let $\lambda_i < 0$, and let us without loss of generality assume that the columns of G are permuted such that $J = -I_k \oplus I_{r-k}$, and

$$-\frac{1}{[G^TG]_{11}} \leq -\frac{1}{[G^TG]_{22}} \leq \cdots \leq -\frac{1}{[G^TG]_{kk}}.$$

Let $B \equiv \mathrm{scal}\,(G)$. Since the nonzero eigenvalues of H are the inverses of the eigenvalues of the pair (J, G^TG), by applying the Courant–Fischer Minimax Theorem we have

$$\frac{1}{\lambda_i} = \min_{dim(S)=k-i+1} \max_{0\neq x\in S} \frac{x^TJx}{x^TG^TGx} \leq \max_{0\neq x\in S_0} \frac{x^TJx}{x^TDB^TBDx}.$$

Here S_0 is spanned by the first $k-i+1$ standard basis vectors, and D and B are defined by (6.1). By setting $y = Dx$ we have

$$\frac{1}{\lambda_i} \leq \max_{0\neq y\in S_0} \frac{y^TD^{-1}JD^{-1}y}{y^TB^TBy} \leq \frac{-\frac{1}{[G^TG]_{k-i+1,k-i+1}}}{\max_{\|z\|_2=1} z^TB^TBz} = \frac{\frac{1}{h_i}}{\lambda_{max}(B^TB)},$$

which proves the right hand side of (6.5). Further,

$$\frac{1}{\lambda_i} = \max_{dim(S)=r-k+i} \min_{0\neq x\in S} \frac{x^TJx}{x^TG^TGx} \geq \min_{0\neq y\in S_0} \frac{y^TD^{-1}JD^{-1}y}{y^TB^TBy} \geq \frac{\frac{1}{h_i}}{\lambda_{min}(B^TB)},$$

where S_0 is spanned by the last $r-k+i$ standard basis vectors. Thus, the theorem is proved for $\lambda_i < 0$. For $\lambda_i > 0$ consider the matrix $-H$. ∎

If G is obtained by Algorithm 2.1, then by combining the above relation with Theorem 6.1 and the fact that $\kappa(\mathrm{scal}\,(G))$ is usually very small, we conclude that such decomposition usually has strong diagonalization effect and rank revealing property.

Finally, note that, since both key properties (6.4) and (2.6) hold for Hermitian matrices (for the latter see [8]), the results of this section also hold for the Hermitian decomposition from Section 5.

7. ANALYSIS OF THE BUNCH–PARLETT METHOD

In this section we prove results similar to the results of Sections 3, 4, and 5 for the Bunch–Parlett method. By combining (2.1-2.3) with (1.4) we see that one step of the decomposition (1.3) is given by

$$\hat{P}H\hat{P}^T = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ Z & I_{n-k} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & \hat{H} \end{bmatrix} \begin{bmatrix} I_k & Z^T \\ 0 & I_{n-k} \end{bmatrix}, \quad (7.1)$$

where

$$W = CQ, \quad \hat{H} = Y - WD^{-1}W^T, \quad Z = WD^{-1}Q^T,$$

and $X = QDQ^T$ is the eigenvalue decomposition of X. These formulas are used in the Lapack implementation of the method, *dsytf2.f*, and are formally slightly different from the original formulas from [8] or [19, Section 4.4.4]. The remarks from Section 3.2 hold here, as well.

Let us begin by the backward error analysis.

THEOREM 7.1. *Let L and T be the factors of a real symmetric matrix H computed by the Bunch–Parlett method with unequilibrated diagonal pivoting from Section 2 in floating-point arithmetic with precision ε. Then, with the relative error of order $O(\varepsilon)$,*

$$LTL^T = PHP^T + E, \qquad |E| \leq 5148n(P|H|P^T + |L||U||\Delta||U|^T|L|^T)\varepsilon,$$

where $U\Delta U^T$ is the computed eigenvalue decomposition of T.

Proof. We are using the notation from the proof of Theorem 3.1. The theorem holds for $n = 1$, and for $n = 2$ for a 2×2 pivot since in both cases $E = 0$.

We must analyze the induction step separately for 1×1 and 2×2 pivot. We assume without loss of generality that the permutation matrices $\hat{P}$ from (7.1) and P from (1.3) are the identity matrices.

Let us first consider a 1×1 pivot, that is, $k = 1$, $W = C$, $Q = 1$, and $X = H_{11}$. The analysis is similar to the one of [30, Theorem 3.1.1], although here the matrix H need not be positive definite. We have

$$\begin{aligned} \hat{H} &= \mathrm{fl}\,(Y - CX^{-1}C^T) = Y - CX^{-1}C^T + \hat{F}, \\ Z &= \mathrm{fl}\,(WX^{-1}) = CX^{-1} + \delta Z, \end{aligned} \quad (7.2)$$

where

$$|\hat{F}| \leq 3\varepsilon(|Y| + |C||X^{-1}||C|^T), \qquad |\delta Z| \leq \varepsilon|C||X^{-1}|.$$

The induction assumption is

$$\hat{L}\hat{T}\hat{L}^T = \hat{H} + \hat{E}, \qquad |\hat{E}| \leq 5148(n-k)\varepsilon(|\hat{H}| + |\hat{L}||\hat{U}||\hat{\Delta}||\hat{U}|^T|\hat{L}|^T), \quad (7.3)$$

where $\hat{U}\hat{\Delta}\hat{U}^T$ is the computed eigenvalue decomposition of $\hat{T}$. By setting

$$L = \begin{bmatrix} I_k & 0 \\ Z & \hat{L} \end{bmatrix}, \qquad T = \begin{bmatrix} X & \\ & \hat{T} \end{bmatrix},$$

we have

$$LTL^T = \begin{bmatrix} X & XZ^T \\ ZX & ZXZ^T + \hat{L}\hat{T}\hat{L}^T \end{bmatrix} \equiv H + E, \qquad E = \begin{bmatrix} 0 & (\delta C)^T \\ \delta C & \delta Y \end{bmatrix}. \tag{7.4}$$

From (7.2) we have $C = (Z - \delta Z)X$ and

$$|\delta C| = |\delta Z X| \leq \varepsilon |C||X^{-1}||X| = \varepsilon |C|. \tag{7.5}$$

From (7.2) and (7.3), by ignoring the terms of $O(\varepsilon^2)$, we have

$$\begin{aligned} |\delta Y| &= |\hat{E} + \hat{F} + ZXZ^T - CX^{-1}C^T| \\ &\leq 5148(n-1)\varepsilon(|Y| + |C||X^{-1}||C|^T + |\hat{L}||\hat{U}||\hat{\Delta}||\hat{U}|^T|\hat{L}|^T) \\ &\quad +3\varepsilon(|Y| + |C||X^{-1}||C|^T) + 2\varepsilon|C||X^{-1}||C|^T \\ &\leq (5148(n-1)+5)(|Y| + |Z||X||Z|^T + |\hat{L}||\hat{U}||\hat{\Delta}||\hat{U}|^T|\hat{L}|^T)\varepsilon \end{aligned} \tag{7.6}$$

The theorem now follows by inserting this and (7.5) into (7.4) and setting

$$U = \begin{bmatrix} 1 & \\ & \hat{U} \end{bmatrix}, \qquad \Delta = \begin{bmatrix} X & \\ & \hat{\Delta} \end{bmatrix}.$$

Let us now consider a 2×2 pivot, that is, $k = 2$. Let $X = \widetilde{Q}\widetilde{D}\widetilde{Q}^T$ and QDQ^T be the exact and the computed eigenvalue decompositions of X, respectively. The relationship between these two decompositions is given by (3.18). Now (7.1) holds with

$$\begin{aligned} W &= \mathrm{fl}(CQ) = C\widetilde{Q} + \delta W, \\ \hat{H} &= \mathrm{fl}(Y - WD^{-1}W^T) = Y - C\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}^T C^T + \hat{F}, \\ Z &= \mathrm{fl}(WD^{-1}Q^T) = C\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}^T + \delta Z, \end{aligned} \tag{7.7}$$

where

$$\begin{aligned} |\delta W| &\leq 13|C||\widetilde{Q}|\varepsilon, \\ |\hat{F}| &\leq 120(|Y| + |C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|C|^T)\varepsilon, \\ |\delta Z| &\leq 117|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T\varepsilon. \end{aligned}$$

Therefore, (7.4) holds with

$$|\delta C| = |\delta Z \widetilde{Q}\widetilde{D}\widetilde{Q}^T| \leq 117|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T\varepsilon, \tag{7.8}$$

and we have to bound the right hand side in terms of $|C|$ and $|Z||Q||D||Q|^T$. After a tedious computation we obtain

$$|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T = \left(1 + 2\widetilde{cs}^2\widetilde{sn}^2(\frac{|\widetilde{a}|}{|\widetilde{b}|} + \frac{|\widetilde{b}|}{|\widetilde{a}|})\right)|C| +$$

$$+2\widetilde{cs}|\widetilde{sn}|\left[\ \left(1 + \widetilde{cs}^2\tfrac{|\widetilde{a}|}{|\widetilde{b}|} + \widetilde{sn}^2\tfrac{|\widetilde{b}|}{|\widetilde{a}|}\right)|C_{:2}| \quad \left(1 + \widetilde{cs}^2\tfrac{|\widetilde{b}|}{|\widetilde{a}|} + \widetilde{sn}^2\tfrac{|\widetilde{a}|}{|\widetilde{b}|}\right)|C_{:1}|\ \right],$$

and

$$|C\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T \geq -\left(1 + 2\widetilde{cs}^2\widetilde{sn}^2(\frac{|\widetilde{a}|}{|\widetilde{b}|} + \frac{|\widetilde{b}|}{|\widetilde{a}|} + 1)\right)|C| +$$

$$+\widetilde{cs}|\widetilde{sn}|\left[\ \left(1 + \widetilde{cs}^2\tfrac{|\widetilde{a}|}{|\widetilde{b}|} + \widetilde{sn}^2\tfrac{|\widetilde{b}|}{|\widetilde{a}|}\right)|C_{:2}| \quad \left(1 + \widetilde{cs}^2\tfrac{|\widetilde{b}|}{|\widetilde{a}|} + \widetilde{sn}^2\tfrac{|\widetilde{a}|}{|\widetilde{b}|}\right)|C_{:1}|\ \right].$$

By combining these two relations and using $\widetilde{cs}^2\widetilde{sn}^2 \leq 1/4$, we get

$$\begin{aligned} |C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T \quad &\leq \quad 2|C\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T \\ &+ \left(4 + \frac{3}{2}(\frac{|\widetilde{a}|}{|\widetilde{b}|} + \frac{|\widetilde{b}|}{|\widetilde{a}|})\right)|C|. \end{aligned} \tag{7.9}$$

Further, (3.2-3.5) imply that

$$\max\left\{\frac{|\widetilde{a}|}{|\widetilde{b}|}, \frac{|\widetilde{b}|}{|\widetilde{a}|}\right\} \leq \frac{|\widetilde{t}|\nu_0 + \nu_1}{|\widetilde{t}|\nu_0} \leq \alpha(\alpha + \sqrt{1+\alpha^2}) + 1 \leq 2.171,$$

for sign $(H_{11}) = -$sign (H_{22}), and

$$\max\left\{\frac{|\widetilde{a}|}{|\widetilde{b}|}, \frac{|\widetilde{b}|}{|\widetilde{a}|}\right\} \leq \frac{|\widetilde{t}|\nu_0 + \nu_1}{|\widetilde{t}|\nu_0 - \nu_1} \leq 1 + \frac{2\alpha}{\frac{1}{\alpha/2+\sqrt{1+\alpha^2/4}} - \alpha} \leq 15.322$$

otherwise, so that

$$\frac{|\widetilde{a}|}{|\widetilde{b}|} + \frac{|\widetilde{b}|}{|\widetilde{a}|} \leq 15.387.$$

By inserting this and (7.7) into (7.9), we have

$$|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T \leq 2|Z||Q||D||Q|^T + 27.081|C| + O(\varepsilon). \tag{7.10}$$

By inserting this into (7.8) and ignoring the $O(\varepsilon^2)$ term, we finally have

$$|\delta C| \leq 3168.5(|Z||Q||D||Q|^T + |C|)\varepsilon. \tag{7.11}$$

To complete the proof it remains to bound $|\delta Y|$ from (7.4) in terms of $|Y|$ and $|Z||Q||D||Q|^T|Z|^T$. Indeed, (7.3) and (7.7) imply that

$$\begin{aligned} |\delta Y| &= |\hat{E}+\hat{F}+ZXZ^T-CX^{-1}C^T| \\ &\leq |\hat{E}|+|\hat{F}|+|\delta Z||C|^T+|C||\delta Z|^T+O(\varepsilon) \\ &\leq 5148(n-2)\varepsilon(|\hat{H}|+|\hat{L}||\hat{U}||\hat{\Delta}||\hat{U}|^T|\hat{L}|^T) \\ &\quad +120\varepsilon|Y|+(120+2\cdot 117)\varepsilon|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|C|^T. \quad (7.12) \end{aligned}$$

Since

$$\hat{H}=Y-ZQDQ^TZ^T+\hat{F}+O(\varepsilon),$$

we have

$$|\hat{H}|\leq|Y|+|Z||Q||D||Q|^T|Z|^T+O(\varepsilon). \quad (7.13)$$

Further, (7.10) and (7.7) imply that

$$\begin{aligned} &|C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|C|^T \leq |C||\widetilde{Q}||\widetilde{D}^{-1}||\widetilde{Q}|^T|\widetilde{Q}||\widetilde{D}||\widetilde{Q}|^T|\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}^TC^T| \\ &\leq (2|Z||Q||D||Q|^T+27.081|(Z-\delta Z)\widetilde{Q}\widetilde{D}\widetilde{Q}^T|)|\widetilde{Q}\widetilde{D}^{-1}\widetilde{Q}^TC^T| \\ &\leq 29.081|Z||Q||D||Q|^T|Z|^T+O(\varepsilon). \end{aligned}$$

By inserting this and (7.13) into (7.12) and ignoring the $O(\varepsilon^2)$ term we have

$$|\delta Y|\leq(5148(n-2)+10296)(|Y|+|Z||Q||D||Q|^T|Z|^T+|\hat{L}||\hat{U}||\hat{\Delta}||\hat{U}|^T|\hat{L}|^T)\varepsilon.$$

The theorem now follows by inserting this and (7.11) into (7.4) and setting

$$U=\begin{bmatrix} Q & \\ & \hat{U}\end{bmatrix},\qquad \Delta=\begin{bmatrix} D & \\ & \hat{\Delta}\end{bmatrix}. \qquad \blacksquare$$

Even though the constant of the theorem is larger than the constant from Theorem 3.1, numerical experiments show that the entire factor $O(n)$ is usually an overestimate. All remarks from Sections 3.1, 3.2 and 3.4 hold here, as well. In particular, if the matrix T from the theorem is diagonal, then (7.6) implies that the error is bounded by

$$|E|\leq 5(P|H|P^T+|L||T||L|^T)\varepsilon. \quad (7.14)$$

This bound holds e.g. for positive definite and scaled diagonally dominant matrices. For positive definite matrices this bound is slightly worse than the bound of [30, Theorem 3.1.1]. Further, by inspecting its proof, we see that this theorem gives the backward error for the Bunch–Parlett decomposition, too, if T is diagonal. The only exception is that the matrix $\widetilde{D}$ from [30, (3.1.4)] should be replaced by $|\widetilde{D}|$.

Normwise stability has been proved by Bunch [5] for the Bunch–Parlett method with complete pivoting, and recently by Higham [21] for the Bunch–Kaufman method with partial pivoting (see also Section 3.3).

Let us now consider forward error. Let us make assumptions similar to the ones in Section 4: non-singular unperturbed and perturbed problems are decomposed by using the same permutation sequence, resulting in matrices T from (1.3) being diagonal and having the corresponding diagonal elements of the same sign. Then we see that the componentwise forward perturbation and error bounds are given by [30, Theorem 3.2.1] and [30, Theorem 3.3.1], respectively. Here, too, in the statements of the theorems the matrices D and $\widetilde{D}$ should be replaced by $|D|$ and $|\widetilde{D}|$, respectively.

Similarly as in Section 5 we conclude that all above results hold for Hermitian matrices, with the exception that the constant in Theorem 7.1 is slightly larger due to the complex arithmetic.

Finally, in view of Section 3.1, all results of this section hold for the Lapack implementations of the real symmetric and Hermitian versions of the Bunch–Kaufmann–Parlett method [6, 7], *dsytf2.f* and *chetf2.f*.

8. CONCLUDING REMARKS

In this section we summarize our contributions, describe results of numerical experiments, and compare our results with the existing analysis by Bunch [5]. We also illustrate our results by a small example.

We have proved componentwise backward error bounds for two versions of the real symmetric and Hermitian decomposition, the $H = GJG^T$ decomposition and the Bunch–Parlett decomposition $PHP^T = LTL^T$. The bounds hold for the outer product version of the algorithms. The bounds are easy to compute, and simple to use in further applications. Numerical experiments show that the bounds reveal well the structure of actual errors, and that the factors of order $O(n)$ are usually an overestimate. More precisely, the bounds of Theorems 3.1 and 5.1 can usually be replaced by the simpler bound $|E| \leq |G||G|^T\varepsilon$, and the bound of Theorem 7.1 can be replaced by $|E| \leq |L||U||\Delta||U|^T|L|^T\varepsilon$.

For non-singular real or Hermitian matrices which have lower triangular factor G or diagonal factor T, we proved componentwise forward error bound, that is, we are able to estimate the precision of the computed factors.

Our results extend the results by Sun [30] by enlarging the class of matrices to indefinite matrices and by including the Hermitian case.

We proved attainable bounds for the scaled condition of the matrix G, and showed that the decomposition $H = GJG^T$ usually has non-trivial diagonalization effect and rank revealing property.

It is interesting to compare our result with the analysis of the Bunch–Parlett decomposition (1.3). Bunch [5, (2.3.4)] showed that the factors L and T computed with the unequilibrated diagonal pivoting in floating-point arithmetic with precision ε satisfy $LTL^T = PHP^T + E$, where elements of the backward error matrix E are bounded in terms of absolutely maximal elements of the reduced matrices:

$$|E_{jk}| = |E_{kj}| \leq C_{jk}\varepsilon, \qquad \text{for } \ j \geq k, \tag{8.1}$$

$$C_{jk} = 5.71 \sum_{\substack{i=1 \\ p_i=1}}^{k-1} \nu_0^{(i)} + 31.65 \sum_{\substack{i=1 \\ p_i=2}}^{k} \nu_0^{(i)} + \begin{cases} \nu_0^{(k)} & \text{if } p_k = 1 \\ 13.7\nu_0^{(k)} & \text{if } p_k = 2 \\ 13.7\nu_0^{(k-1)} & \text{if } p_k = 0 \end{cases}$$

Here $\nu_0^{(i)}$ denotes the value of ν_0 in the ith step of Algorithm 2.1. If in the step $i-1$ a 2×2 pivot was chosen, then $\nu_0^{(i)} = 0$. The quantities p_i have the following meaning: $p_i = 1$ if in the ith step a 1×1 pivot was chosen; $p_i = 2$ if in the ith step a 2×2 pivot was chosen; and $p_i = 0$ if in the step $i-1$ a 2×2 pivot was chosen (in this case $\nu_0^{(i)}$ does not exist). Although Bunch [5] gives no explicit error bound for the Hermitian version of the Bunch–Parlett decomposition, the nature of his proof is such that it holds for the Hermitian version, as well.

Our bounds and (8.2) are all a posteriori bounds since they are computed after the decomposition is completed. The bounds for maximal elements, $\nu_0^{(i)}$, are implicitly included in the $|G||G|^T$ or $|L||U||\Delta||U|^T|L|^T$ terms of our bounds. Note that our bounds are more convenient for further applications such as in [27]. The comparison of the bound (8.2), our bounds, and actual errors is as follows: the maximal elementwise bounds are almost the same; our bounds often reveal better the actual error structure (note that in (8.2) all elements E_{jk}, $j \geq k$, have the same bound, and the bound grows with k); bounds for particular elements of E can vary by even several orders of magnitude, although our bounds are on average better; for smaller dimensions all bounds approximate actual errors well, for larger dimensions all bounds overestimate actual errors by a factor of order $O(n)$.

We conclude the paper by illustrating our results with the following example: let

$$H = \begin{bmatrix} 3207938000 & 300000 & -423212 & 19800 \\ 300000 & 1600 & -300 & 14 \\ -423212 & -300 & 43.5 & -4.75 \\ 19800 & 14 & -4.75 & 0.1875 \end{bmatrix}.$$

Note that H is stored exactly on machines with base 2 [18]. The decomposition (1.1) computed by Algorithm 2.1 in single precision, $\varepsilon \approx 10^{-8}$, is

$$G = \begin{bmatrix} 56638.662 & 0 & 0 & 0 \\ 5.2967353 & 39.6477567 & 0 & 0 \\ -7.4721398 & -6.568393 & 7.4482656 & 0 \\ 0.34958453 & 0.30640682 & 0.01681668 & 0.16826074 \end{bmatrix},$$

with $J = \operatorname{diag}(-1, 1, 1, -1)$ and $P = I$. The backward error matrix $E = GJG^T - H$ is

$$E = \begin{bmatrix} 3.32e+01 & 3.60e-04 & -5.49e-04 & 3.51e-05 \\ 3.60e-04 & 4.00e-05 & 3.81e-06 & -1.37e-07 \\ -5.49e-04 & 3.81e-06 & -6.55e-07 & 9.13e-09 \\ 3.51e-05 & -1.37e-07 & 9.13e-09 & 5.61e-09 \end{bmatrix},$$

and its elements are bounded by (3.37) as follows:

$$|E| \leq \begin{bmatrix} 7.70e+02 & 7.20e-02 & 1.02e-01 & 4.75e-03 \\ 7.20e-02 & 3.84e-04 & 7.20e-05 & 3.36e-06 \\ 1.02e-01 & 7.20e-05 & 2.38e-05 & 1.14e-06 \\ 4.75e-03 & 3.36e-06 & 1.14e-06 & 5.19e-08 \end{bmatrix}.$$

Further, we assume that the factor $\widetilde{G}$ computed by Algorithm 2.1 in double precision, $\varepsilon \approx 10^{-16}$, is exact. The forward error matrix $\delta G = G - \widetilde{G}$ is

$$\delta G = \begin{bmatrix} 2.93e-04 & 0 & 0 & 0 \\ -2.10e-08 & 5.07e-07 & 0 & 0 \\ 2.89e-08 & 1.72e-07 & -1.37e-07 & 0 \\ -1.19e-09 & -7.03e-09 & 1.49e-08 & -3.34e-08 \end{bmatrix},$$

and its elements are bounded by Theorem 4.2 as follows:

$$|\delta G| \leq \begin{bmatrix} 1.36e-02 & 0 & 0 & 0 \\ 3.81e-06 & 1.02e-05 & 0 & 0 \\ 5.80e-06 & 5.80e-06 & 1.28e-05 & 0 \\ 2.88e-07 & 2.89e-07 & 6.58e-07 & 1.36e-06 \end{bmatrix}.$$

This also illustrates the perturbation bound of Theorem 4.2. Finally, note that $\sigma^2_{min}(\operatorname{scal}(G)) = 0.83628307$, $\sigma^2_{max}(\operatorname{scal}(G)) = 1.1635069$, while the eigenvalues of H, the diagonal elements of the matrix $G^T GJ$, and their respective quotients from (6.5) are

$$\begin{aligned} \lambda_i &= -54.043364,\ -0.028309685,\ 1613.7487,\ 3207938084 \\ h_i &= -55.476945,\ -0.028311688,\ 1615.1823,\ 3207938040 \\ \lambda_i/h_i &= 0.97415898,\ 0.99992925,\ 0.99911242,\ 1 \end{aligned}$$

This illustrates Theorem 6.2, and diagonalization effect and rank revealing property of the symmetric indefinite decomposition.

The Bunch–Parlett decomposition defined by (1.3) and (7.1) computed in single precision is

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 9.3518017e-05 & 1 & 0 & 0 \\ -1.3192649e-04 & -0.16566872 & 1 & 0 \\ 6.1721891e-06 & 7.7282259e-03 & 2.257798e-03 & 1 \end{bmatrix},$$

$$T = \operatorname{diag}(3207938000,\ 1571.9446,\ -55.476663,\ -0.02831168),$$

and $P = I$. The backward error matrix $E = LTL^T - H$ is

$$E = \begin{bmatrix} 0.00e+00 & 4.19e-04 & -4.78e-04 & -4.29e-05 \\ 4.19e-04 & 5.14e-06 & 3.15e-06 & -2.96e-07 \\ -4.78e-04 & 3.15e-06 & -1.60e-06 & -2.52e-08 \\ -4.29e-05 & -2.96e-07 & -2.52e-08 & 1.99e-09 \end{bmatrix},$$

and its elements are bounded by (7.14) as follows

$$|E| \leq \begin{bmatrix} 1.28e+03 & 1.20e-01 & 1.69e-01 & 7.92e-03 \\ 1.20e-01 & 6.40e-04 & 1.20e-04 & 5.60e-06 \\ 1.69e-01 & 1.20e-04 & 3.96e-05 & 1.90e-06 \\ 7.92e-03 & 5.60e-06 & 1.90e-06 & 8.64e-08 \end{bmatrix}.$$

On the other hand, the bound by Bunch (8.2) is

$$|E| \leq \begin{bmatrix} 4.23e-03 & 4.23e-03 & 4.23e-03 & 4.23e-03 \\ 4.23e-03 & 2.42e-02 & 2.42e-02 & 2.42e-02 \\ 4.23e-03 & 2.42e-02 & 4.42e-02 & 2.42e-02 \\ 4.23e-03 & 2.42e-02 & 2.42e-02 & 4.42e-02 \end{bmatrix},$$

and we see that in this example both our bounds reveal the error structure much better. Further, we assume that the factors $\widetilde{L}$ and $\widetilde{T}$ computed in double precision are exact. The forward error matrices $\delta L = L - \widetilde{L}$ and $\delta T = T - \widetilde{T}$ are

$$\delta L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1.31e-13 & 0 & 0 & 0 \\ -1.49e-13 & 2.60e-09 & 0 & 0 \\ -1.34e-14 & -2.12e-10 & 1.93e-09 & 0 \end{bmatrix},$$

$$\delta T = \operatorname{diag}(0,\ 5.06 \cdot 10^{-6},\ -5.12e \cdot 10^{-7},\ 7.87 \cdot 10^{-9}),$$

and their elements are bounded by [30, Theorem 3.3.1] (modified as described in Section 7) as follows:

$$|\delta L| \leq \begin{bmatrix} 0 & 0 & 0 & 0 \\ 5.05e-11 & 0 & 0 & 0 \\ 7.68e-11 & 1.10e-07 & 0 & 0 \\ 3.81e-12 & 5.47e-09 & 6.62e-08 & 0 \end{bmatrix},$$

$$|\delta T| \leq \operatorname{diag}(577,\ 3.03 \cdot 10^{-4},\ 7.13 \cdot 10^{-5},\ 1.71 \cdot 10^{-7}).$$

This also illustrates the perturbation bound of [30, Theorem 3.2.1].

I am grateful to my Ph. D. thesis supervisor Professor Krešimir Veselić for his great help and advice. I would also like to thank my thesis referee Professor James Demmel for his valuable and detailed remarks, and Professor Jesse Barlow, Eberhard Pietzsch, Zlatko Drmač, and Xiaofeng Wang for their comments. Finally, I would like to thank Professor Nick Higham and the referee for careful reading of the manuscript and their remarks which helped me correct some errors and led to some improvements in the final version of the paper.

REFERENCES

1 E. Anderson et al., *LAPACK Users' Guide, Second Edition*, SIAM, Philadelphia, 1995.

2 M. Arioli, I. S. Duff, and P. P. M. de Rijk, On the augmented system approach to sparse least–squares problems, *Numer. Math.*, 55:667–684, (1989).

3 C. Ashcraft, R. G. Grimes, and J. G. Lewis, Accurate symmetric indefinite linear equation solvers, Manuscript, May 1995.

4 J. Barlow and J. Demmel, Computing accurate eigensystems of scaled diagonally dominant matrices, *SIAM J. Numer. Anal.*, 27:762–791, (1990).

5 J. R. Bunch, Analysis of the diagonal pivoting method, *SIAM J. Numer. Anal.*, 8:656–680, (1971).

6 J. R. Bunch and L. Kaufman, Some stable methods for calculating inertia and solving symmetric linear systems, *Math. Comp.*, 31:163–179, (1977).

7 J. R. Bunch, L. Kaufman, and B. N. Parlett, Decomposition of a symmetric matrix, *Numer. Math.*, 27:95–109, (1976).

8 J. R. Bunch and B. N. Parlett, Direct methods for solving symmetric indefinite systems of linear equations, *SIAM J. Numer. Anal.*, 8:639–655, (1971).

9 R. Byers, Solving the algebraic Riccati equation with the matrix sign function, *Linear Algebra Appl.*, 85:267–279, (1987).

10 T. F. Chan, Rank revealing QR factorizations, *Linear Algebra Appl.*, 88/89:67–82, (1987).

11 J. Demmel, On floating point errors in Cholesky, LAPACK Working Note 14, Computer Science Dept. Report, University of Tennessee, Knoxville, October 1989.

12 J. Demmel, personal communication, 1996.

13 J. Demmel and K. Veselić, Jacobi's method is more accurate than QR, *SIAM J. Matrix Anal. Appl.*, 13:1204–1243, (1992).

14 I. S. Duff et al., The factorization of sparse symmetric indefinite matrices, *IMA J. Numerical Analysis*, 11:181–204, (1991).

15 R. E. Funderlic, M. Neumann, and R. J. Plemmons, *LU* decompositions of generalized diagonally dominant matrices, *Numer. Math.*, 40:57–69, (1982).

16 P. E. Gill, W. Murray, D. B. Poncaleón, and M. A. Saunders, Preconditioners for indefinite systems arising in optimization, *SIAM J. Matrix Anal. Appl.*, 13:292–311, (1992).

17 P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, Inertia–controlling methods for general quadratic programming, *SIAM Review*, 33:1–36, (1991).

18 D. Goldberg, What every computer scientist should know about floating-point arithmetic, *ACM Computing Surveys*, 23:5–48, (1991).

19 G. H. Golub and C. F. Van Loan, *Matrix Computations, Second ed.*, The John Hopkins University Press, Baltimore, MD, 1989.

20 N. Higham, Analysis of the Cholesky decomposition of a semi-definite matrix, in *Reliable Numerical Computation*, M. G. Cox and S. Hammarling, eds., Clarendon Press, 1990.

21 N. Higham, Stability of the diagonal pivoting method with partial pivoting, Numerical Analysis Report No. 265, University of Manchester, July 1995.

22 Y. P. Hong and C.-T. Pan, Rank-revealing QR factorization and SVD, Technical report, Department of Mathematical Sciences, Northern Illinois University, October 1990.

23 R. A. Horn and C. J. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

24 C. S. Kenney and A. J. Laub, The matrix sign function, *IEEE Trans. Automatic Control*, 40:1330–1348, (1995).

25 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM, Philadelphia, 1995.

26 B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, N. J., 1980.

27 I. Slapničar, Accurate Symmetric Eigenreduction by a Jacobi Method, Ph.d. thesis, Fernuniversität, Hagen 1992.

28 I. Slapničar and K. Veselić, Perturbations of the eigenprojections of a factorized Hermitian matrix, *Linear Algebra Appl.*, 218:273–280, (1995).

29 A. van der Sluis, Condition numbers and equilibration of matrices, *Numer. Math.*, 14:14–23, (1969).

30 J. Sun, Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations, *Linear Algebra Appl.*, 173:77–97, (1992).

31 K. Veselić, A Jacobi eigenreduction algorithm for definite matrix pairs, *Numer. Math.*, 64:241–269, (1993).

32 K. Veselić and I. Slapničar, Floating–point perturbations of Hermitian matrices, *Linear Algebra Appl.*, 195:81-116, (1993).

33 M. H. Wright, Interior methods for constrained optimization, in *Acta Numerica*, A. Iserles, ed., Cambridge University Press, New York, 1992.