MATHEMATICAL ENGINEERING TECHNICAL REPORTS

Roundoff Error Analysis of the CholeskyQR2 Algorithm

Yusaku YAMAMOTO, Yuji NAKATSUKASA, Yuka YANAGISAWA and Takeshi FUKAYA

METR 2014–36 December 2014

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

ROUNDOFF ERROR ANALYSIS OF THE CHOLESKYQR2 ALGORITHM

YUSAKU YAMAMOTO*, YUJI NAKATSUKASA†, YUKA YANAGISAWA‡, AND TAKESHI FUKAYA§

Abstract. We consider the QR decomposition of an $m \times n$ matrix X with full column rank, where $m \geq n$. Among the many algorithms available, the Cholesky QR algorithm is ideal from the viewpoint of high performance computing, since it consists entirely of standard level 3 BLAS operations with large matrix sizes and requires only one allreduce and broadcast in parallel environments. Unfortunately, it is well known that the algorithm is not numerically stable and the deviation from orthogonality of the computed Q factor is of $O((\kappa_2(X))^2\mathbf{u})$, where $\kappa_2(X)$ is the 2-norm condition number of X and $\mathbf{u}$ is the unit roundoff. In this paper, we show that if the condition number of X is not too large, we can improve the stability greatly by iterating the Choelsky QR algorithm twice. More specifically, if $\kappa_2(X)$ is at most $O(\mathbf{u}^{-\frac{1}{2}})$, both the residual and deviation from orthogonality are shown to be of $O(\mathbf{u})$. Numerical results support our theoretical analysis.

Key words. QR decomposition, Cholesky QR, commnication-avoiding algorithms, roundoff error analysis

AMS subject classifications. 15A23, 65F25, 65G50

1. Introduction.

Let $X \in \mathbb{R}^{m\times n}$ be an m by n matrix with $m \geq n$ of full column rank. We consider computing its QR decomposition, $X = QR$, where $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns and $R \in \mathbb{R}^{n\times n}$ is upper triangular. This is one of the most fundamental matrix decompositions and is used in various scientific computations. Examples include linear least squares, preprocessing for the singular value decomposition of a rectangular matrix [10], and orthogonalization of vectors arising in block Krylov methods [2, 17] or electronic structure calculations [3, 22]. Often, the matrix size is very large, so an algorithm suited for modern high performance computers is desired.

One important feature of modern high performance architectures is that communication is much slower than arithmetic. Here, communication refers to both data transfer between processors or nodes and data movement between memory hierarchies. Thus it is essential for higher performance to minimize the frequency and amount of these communications [1]. To minimize interprocessor communications, the algorithm has to have a large grain parallelism. To minimize data movement between memory hierarchies, it is effective to reorganize the algorithm to use level 3 BLAS operations as much as possible [10]. Of course, the benefit of using level 3 BLAS operations increases as the size of matrices used there becomes larger.

Conventionally, three major algorithms have been used to compute the QR decomposition: the Householder QR algorithm, the classical Gram-Schmidt (CGS) algorithm and the modified Gram-Schmidt (MGS) algorithm. The Householder QR algorithm is widely used due to its excellent numerical stability [11]. MGS, which is less stable, is often preferred when the Q factor is needed explicitly, because it requires

*The University of Electro-Communications, Tokyo, Japan / JST CREST, Tokyo, Japan (yusaku.yamamoto@uec.ac.jp).

† Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (nakatsukasa@mist.i.u-tokyo.ac.jp).

‡Waseda University, Tokyo, Japan.

§RIKEN Advanced Institute for Computational Science, Kobe, Japan / JST CREST, Tokyo, Japan.

only half as much work as the Householder QR in that case. When the matrix A is well conditioned, CGS is also sometimes used since it provides more parallelism. Note that for matrices with the 2-norm condition number $\kappa_2(X)$ at most $O(\mathbf{u}^{-1})$, where $\mathbf{u}$ is the unit roundoff, repeating CGS or MGS twice leads to algorithms that are as stable as Householder QR [9]. They are known as CGS2 and MGS2, respectively.

For each of these algorithms, variants that can better exploit modern high performance architectures have been developed. There are block versions and recursive versions of Householder QR [6, 18], MGS [13] and CGS [12] that can perform most of the computations in the form of level 3 BLAS. There is also a variant of Householder QR called the tall-and-skinny QR (TSQR) [5], which has large grain parallelism and requires only one allreduce and broadcast in a distributed environment.

While these variants have been quite successful, they are not completely satisfactory from the viewpoint of high performance computing. In the block and recursive versions mentioned above, the sizes of matrices appearing in the level 3 BLAS are generally smaller than that of X, and become even smaller as the level goes down in the case of recursive algorithms. For the TSQR algorithm, though only one allreduce is required throughout the algorithm, the reduction operation is a non-standard one, which corresponds to computing the QR decomposition of a $2n \times n$ matrix formed by concatenating two upper triangular matrices [5]. Thus each reduction step requires $O(n^3)$ work and this tends to become a bottleneck in parallel environments [7]. In addition, the TSQR algorithm requires non-standard level 3 BLAS operations such as multiplication of two triangular matrices [5], for which no optimized routines are available on most machines.

There is another algorithm for the QR decomposition, namely, the Cholesky QR algorithm. In this algorithm, one first forms the Gram matrix $A = X^\top X$, computes its Cholesky factororization $A = R^\top R$, and then finds the Q factor by $Q = XR^{-1}$. This algorithm is ideal from the viewpoint of high performance computing because (1) its computational cost is $2mn^2$ (in the case where $m \gg n$), which is equivalent to the cost of CGS and MGS and half that of Householder QR, (2) it consists entirely of standard level 3 BLAS operations, (3) the first and third steps are highly parallel large size level 3 BLAS operations in which two of the three matrices are of size $m \times n$, (4) the second step, which is the only sequential part, requires only $O(n^3)$ work, as opposed to the $O(mn^2)$ work in the first and the third steps, and (5) it requires only one allreduce and one broadcast if X is partitioned horizontally. Unfortunately, it is well known that Cholesky QR is not stable. In fact, deviation from orthogonality of the Q factor computed by Cholesky QR is proportional to $(\kappa_2(X))^2$ [19]. Accordingly, standard textbooks like [21] describe the method as "quite unstable and is to be avoided unless we know a priori that R is well conditioned".

In this paper, we show that the Cholesky QR algorithm can be applied to matrices with a larger condition number to give a stable QR factorization if it is repeated *twice*. More specifically, we show that if $\kappa_2(X)$ is at most $O(\mathbf{u}^{-\frac{1}{2}})$, the Q and R factors obtained by applying Cholesky QR twice satisfy $\|Q^\top Q - I\|_F = O(\mathbf{u})$ and $\|X - QR\|_F = O(\mathbf{u})$. Furthermore, we give the coefficients of $\mathbf{u}$ in these bounds explicitly as simple low degree polynomials in m and n. In the following, we call this method *CholeskyQR2*. Of course, the arithmetic cost of CholeskyQR2 is twice that of Cholesky QR, CGS and MGS, but it is equivalent to the cost of Householder QR, CGS2 and MGS2. Given the advantages stated above, the increase in the computational work might be more than compensated in some cases. Hence, for matrices with $\kappa_2(X) \sim O(\mathbf{u}^{-\frac{1}{2}})$, CholeskyQR2 can be the method of choice in terms of both numerical stability and efficiency on high performance architectures.

The idea of performing the QR decomposition twice to get better stability is not new. In his textbook [15], Parlett analyses Gram-Schmidt orthogonalization of two vectors and introduces the principle of "twice is enough", which he attributes to Kahan. There is also a classical paper by Daniel, Gragg, Kaufman and Stewart [4], which deals with the effect of reorthogonalization on the update of the Gram-Schmidt QR decomposition. More recently, Giraud et al. performed a detailed error analysis of CGS2 and MGS2 and showed that they give numerically orthogonal Q factor and small residual for matrices with $\kappa_2(X) \sim O(\mathbf{u}^{-1})$ [9]. Stathopoulas et al. experimentally show that the Cholesky QR algorithm can be applied to matrices with a large condition number, if it is applied twice (or more) [19]. Rozložník et al. analyze the CholeskyQR2 algorithm in a more general setting of orthogonalization under indefinite inner product and derive bounds on the residual and deviation from orthogonality [16]. However, their bounds are expressed in terms of the computed Q and R factors, along with the matrix B that defines the inner product, and do not constitute a priori error bounds, in contrast to the bounds derived in this paper. Also, the coefficients of $\mathbf{u}$ are not given explicitly.

Even though the underlying idea of repeating an unstable algorithm twice to improve stability is the same, it is worth noting the inherent disadvantage of CholeskyQR2 when compared with CGS2 and MGS2: numerical breakdown. Specifically, if $\kappa_2(X) \gg \mathbf{u}^{-\frac{1}{2}}$ then the Cholesky factorization of X^TX can break down, and so does CholeskyQR2. By contrast, Gram-Schmidt type algorithms are free from such breakdowns (except for very obvious breakdowns due to division by zeros in the normalization), and as shown in [9], gives stable QR factorizations for a much wider class of matrices $\kappa_2(X) \sim O(\mathbf{u}^{-1})$ when repeated twice.

The rest of this paper is organized as follows. In section 2, after giving some definitions and assumptions, we introduce the CholeskyQR2 algorithm. Detailed error analysis of CholeskyQR2 is presented in section 3. Numerical results that support our analysis is provided in Section 4. Section 5 gives some discussion on our results. Finally, some concluding remarks are given in section 6.

2. The ChoelskyQR2 algorithm.

2.1. Notation and assumptions. In the following, we consider computing the QR decomposition of an m by n real matrix X, where $m \geq n$. Throughout this paper, we assume that computations are performed using IEEE 754 floating point standard and denote the unit roundoff by $\mathbf{u}$. Let $\sigma_i(X)$ be the ith largest singular value of X and $\kappa_2(X) = \sigma_1(X)/\sigma_n(X)$ be its condition number. We further assume that

$$\delta \equiv 8\kappa_2(X)\sqrt{mn\mathbf{u} + n(n+1)\mathbf{u}} \leq 1. \tag{2.1}$$

This means that the condition number of X is at most $O(\mathbf{u}^{-\frac{1}{2}})$. From this assumption and $\kappa_2(X) \geq 1$, we also have

$$mn\mathbf{u} \leq \frac{1}{64}, \quad n(n+1)\mathbf{u} \leq \frac{1}{64}. \tag{2.2}$$

Following [11], let us define a quantity γ_k for a positive integer k by

$$\gamma_k = \frac{k\mathbf{u}}{1-k\mathbf{u}}.$$

Then it is easy to show that under the assumption (2.1)

$$\gamma_m = \frac{m\mathbf{u}}{1-m\mathbf{u}} \le 1.1m\mathbf{u}, \quad \gamma_{n+1} = \frac{(n+1)\mathbf{u}}{1-(n+1)\mathbf{u}} \le 1.1(n+1)\mathbf{u}. \tag{2.3}$$

2.2. The algorithm. In the Cholesky QR algorithm, we compute the QR decomposition of X by the following procedure.

$$\begin{aligned} A &= X^\top X, \\ R &= \mathrm{chol}(A), \\ Y &= XR^{-1}, \end{aligned}$$

where $\mathrm{chol}(A)$ is a function that computes the (upper triangular) Cholesky factor of A. Then, $X = YR$ can be regarded as the QR decomposition of X.

In the CholeskyQR2 algorithm, after obtaining Y and R by the above procedure, we further compute the following.

$$\begin{aligned} B &= Y^\top Y, \\ S &= \mathrm{chol}(B), \\ Z &= YS^{-1} \ \ (= X(SR)^{-1}), \\ U &= SR \end{aligned}$$

If the columns of Y are exactly orthonormal, B becomes the identity and $Z = Y$. However, in finite precision arithmetic, this does not hold in general and $Z \neq Y$. In the CholeskyQR2 algorithm, the QR decomposition of X is given by $X = ZU$.

3. Error analysis of the CholeskyQR2 algorithm. Our objective is to show that under assumption (2.1), the CholeskyQR2 algorithm delivers an orthogonal factor Z and an upper triangular factor U for which both the orthogonality $\|Z^\top Z - I\|_F$ and residual $\|X - ZU\|_F/\|X\|_2$ are of $O(\mathbf{u})$. Here, the constants in $O(\mathbf{u})$ contain lower order terms in m and n, but not in $\kappa_2(X)$.

This section is structured as follows. In subsection 3.1, we formulate the CholeskyQR2 algorithm in floating point arithmetic and prepare several bounds that are necessary to evaluate the orthogonality of the computed orthogonal factor. Using these bounds, the bound on the orthogonality is derived in subsection 3.2. In subsection 3.3, several bounds that are needed to evaluate the residual are provided, and they are used in subsection 3.4 to give a bound on the residual.

3.1. Preparation for evaluating the orthogonality. Let us denote the matrices A, R and Y computed using floating point arithmetic by $\hat{A} = fl(X^\top X)$, $\hat{R} = fl(\mathrm{chol}(\hat{A}))$ and $\hat{Y} = fl(X\hat{R}^{-1})$, respectively. Taking rounding errors into account, the computed quantities satisfy

$$\hat{A} = X^\top X + E_1, \tag{3.1}$$

$$\hat{R}^\top \hat{R} = \hat{A} + E_2 \ = \ X^\top X + E_1 + E_2, \tag{3.2}$$

$$\hat{\mathbf{y}}_i^\top = \mathbf{x}_i^\top (\hat{R} + \Delta\hat{R}_i)^{-1} \quad (i = 1, 2, \ldots, m). \tag{3.3}$$

Here, $\mathbf{x}_i^\top$ and $\hat{\mathbf{y}}_i^\top$ are the ith row vectors of X and $\hat{Y}$, respectively. E_1 is the forward error of the matrix-matrix multiplication $X^\top X$, while E_2 is the backward error of the Cholesky decomposition of $\hat{A}$. $\Delta\hat{R}_i$ denotes the backward error arising from solving the linear simultaneous equation $\mathbf{y}_i^\top \hat{R} = \mathbf{x}_i^\top$ by forward substitution. It would be easier if we could express the backward error of the forward substitution as

$$\hat{Y} = X(\hat{R} + \Delta\hat{R})^{-1},$$

but we have to use the row-wise expression (3.3) instead because the backward error $\Delta\hat{R}$ depends on the right-hand side vector $\mathbf{x}_i^\top$.

In the following, we evaluate each of E_1, E_2 and $\Delta\hat{R}_i$. We also give bounds on the 2-norms of $\hat{R}^{-1}$ and $X\hat{R}^{-1}$ for later use. Furthermore, we derive an alternative form of Eq. (3.3):

$$\hat{\mathbf{y}}_i^\top = (\mathbf{x}_i^\top + \Delta\mathbf{x}_i^\top)\hat{R}^{-1}, \tag{3.4}$$

in which the backward error enters in the right-hand side vector instead of the coefficient matrix. Equivalently, $\Delta\mathbf{x}_i^\top$ is the residual of the linear system $\mathbf{y}_i^\top R = \mathbf{x}_i^\top$. Then, by letting $\Delta X = (\Delta\mathbf{x}_1, \Delta\mathbf{x}_2, \ldots, \Delta\mathbf{x}_m)^\top$, we can rewrite (3.3) as

$$\hat{Y} = (X + \Delta X)\hat{R}^{-1}, \tag{3.5}$$

which is more convenient to use. We also evaluate the norm of ΔX.

Forward error in the matrix-matrix multiplication $X^\top X$. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Then the componentwise forward error of the matrix-matrix multiplication $C = AB$ can be evaluated as

$$|C - \hat{C}| \le \gamma_n |A||B|, \tag{3.6}$$

where $\hat{C} = fl(AB)$, $|A|$ denotes the matrix whose (i,j)th element is $|a_{ij}|$, and the inequality means componentwise inequality [11]. The 2-norm of the ith column of X, which we denote by $\tilde{\mathbf{x}}_i$, is clearly less than or equal to $\|X\|_2$. Hence,

$$|E_1|_{ij} = |A - \hat{A}|_{ij} \le \gamma_m(|X|^\top|X|)_{ij} = \gamma_m|\tilde{\mathbf{x}}_i|^\top|\tilde{\mathbf{x}}_j| \le \gamma_m\|\tilde{\mathbf{x}}_i\|\|\tilde{\mathbf{x}}_j\| \le \gamma_m\|X\|_2^2. \tag{3.7}$$

(Throughout, a vector norm is always the 2-norm.) Thus we have

$$\|E_1\|_2 \le \|E_1\|_F \le \gamma_m n\|X\|_2^2. \tag{3.8}$$

Simplifying this result using (2.3) leads to

$$\|E_1\|_2 \le 1.1mn\mathbf{u}\|X\|_2^2. \tag{3.9}$$

Backward error of the Cholesky decomposition of $\hat{A}$. Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite and assume that the Cholesky decomposition of A in floating point arithmetic runs to completion and the upper triangular Cholesky factor $\hat{R}$ is obtained. Then, there exists $\Delta A \in \mathbb{R}^{n\times n}$ satisfying

$$\hat{R}^\top\hat{R} = A + \Delta A, \quad |\Delta A| \le \gamma_{n+1}|\hat{R}|^\top|\hat{R}|.$$

See Theorem 10.3 of [11] for details. In our case, we take $A \leftarrow \hat{A}$ in (3.2) to obtain

$$|E_2| \le \gamma_{n+1}|\hat{R}|^\top|\hat{R}|. \tag{3.10}$$

Hence,

$$\|E_2\|_2 \le \|E_2\|_F = \|\,|E_2|\,\|_F \le \gamma_{n+1}\|\,|\hat{R}|^\top|\hat{R}|\,\|_F \le \gamma_{n+1}\|\,|\hat{R}|\,\|_F^2 = \gamma_{n+1}\|\hat{R}\|_F^2 \le \gamma_{n+1}n\|\hat{R}\|_2^2. \tag{3.11}$$

On the other hand, we have from Eq. (3.2),

$$\|\hat{R}\|_2^2 = \|\hat{R}^\top \hat{R}\|_2 \le \|\hat{A}\|_2 + \|E_2\|_2. \tag{3.12}$$

Substituting Eq. (3.12) into the rightmost hand side of Eq. (3.11) leads to

$$\|E_2\|_2 \le \gamma_{n+1} n (\|\hat{A}\|_2 + \|E_2\|_2),$$

or,

$$\|E_2\|_2 \le \frac{\gamma_{n+1} n}{1 - \gamma_{n+1} n} \|\hat{A}\|_2. \tag{3.13}$$

Noting that

$$\|\hat{A}\|_2 \le \|X^\top X\|_2 + \|E_1\|_2 \le \|X\|_2^2 + \gamma_m n \|X\|_2^2 = (1 + \gamma_m n)\|X\|_2^2, \tag{3.14}$$

from Eqs. (3.1) and (3.8), we have

$$\|E_2\|_2 \le \frac{\gamma_{n+1} n (1 + \gamma_m n)}{1 - \gamma_{n+1} n} \|X\|_2^2. \tag{3.15}$$

This result can be simplified using (2.2) and (2.3) as

$$\begin{aligned} \|E_2\|_2 &\le \frac{1.1(n+1)\mathbf{u} \cdot n \cdot (1 + 1.1mn\mathbf{u})}{1 - 1.1n(n+1)\mathbf{u}} \|X\|_2^2 \\ &\le \frac{1.1(n+1)\mathbf{u} \cdot n \cdot (1 + \frac{1.1}{64})}{1 - \frac{1.1}{64}} \|X\|_2^2 = \frac{7161}{6290} n(n+1)\mathbf{u}\|X\|_2^2 \le 1.2n(n+1)\mathbf{u}\|X\|_2^2. \end{aligned} \tag{3.16}$$

Backward error of the forward substitution. Let $U \in \mathbb{R}^{n\times n}$ be a nonsingular triangular matrix. Then, the solution $\hat{\mathbf{x}}$ obtained by solving the linear simultaneous equation $U\mathbf{x} = \mathbf{b}$ by substitution in floating point arithmetic satisfies

$$(U + \Delta U)\hat{\mathbf{x}} = \mathbf{b}, \quad |\Delta U| \le \gamma_n |U|. \tag{3.17}$$

See Theorem 8.5 of [11]. Note that ΔU depends both on U and $\mathbf{b}$, although the bound in (3.17) does not. In our case, $U = \hat{R}$, so we have for $1 \le i \le m$,

$$\|\Delta \hat{R}_i\|_2 \le \|\Delta \hat{R}_i\|_F = \|\,|\Delta \hat{R}_i|\,\|_F \le \gamma_n \|\,|\hat{R}|\,\|_F \le \gamma_n \sqrt{n} \|\hat{R}\|_2. \tag{3.18}$$

By inserting Eq. (3.13) into Eq. (3.12) and using (3.14), we have

$$\|\hat{R}\|_2^2 \le \frac{1}{1 - \gamma_{n+1} n} \|\hat{A}\|_2 \le \frac{1 + \gamma_m n}{1 - \gamma_{n+1} n} \|X\|_2^2. \tag{3.19}$$

Inserting this into Eq. (3.18) leads to

$$\|\Delta \hat{R}_i\|_2 \le \gamma_n \sqrt{\frac{n(1 + \gamma_m n)}{1 - \gamma_{n+1} n}} \|X\|_2.$$

Simplifying the right-hand side in the same way as in Eq. (3.16), we obtain

$$\begin{aligned} \|\Delta \hat{R}_i\|_2 &\le 1.1n\mathbf{u} \sqrt{\frac{n(1 + 1.1mn\mathbf{u})}{1 - 1.1n(n+1)\mathbf{u}}} \|X\|_2 \\ &\le 1.1n\mathbf{u} \sqrt{\frac{n \cdot (1 + \frac{1.1}{64})}{1 - \frac{1.1}{64}}} \|X\|_2 \le 1.2n\sqrt{n}\mathbf{u}\|X\|_2. \end{aligned} \tag{3.20}$$

Bounding the 2-norm of $\hat{R}^{-1}$. Next we evaluate the 2-norm of $\hat{R}^{-1}$. Noting that all the matrices appearing in Eq. (3.2) are symmetric, we can apply the Bauer-Fike theorem (or Weyl's theorem) to obtain

$$(\sigma_n(X))^2 - (\|E_1\|_2 + \|E_2\|_2) \leq (\sigma_n(\hat{R}))^2.$$

Using assumption (2.1), Eqs. (3.9) and (3.16), we have $\|E_1\|_2 + \|E_2\|_2 \leq \frac{1.2}{64}(\sigma_n(X))^2 \leq (1 - \frac{1}{1.1^2})(\sigma_n(X))^2$. Hence,

$$\frac{1}{1.1^2}(\sigma_n(X))^2 \leq (\sigma_n(\hat{R}))^2,$$

leading to the bound on $\hat{R}^{-1}$ as

$$\|\hat{R}^{-1}\|_2 = (\sigma_n(\hat{R}))^{-1} \leq 1.1(\sigma_n(X))^{-1}. \tag{3.21}$$

Bounding the 2-norm of $X\hat{R}^{-1}$. From Eq. (3.2), we have

$$\hat{R}^{-\top} X^\top X \hat{R}^{-1} = I - \hat{R}^{-\top}(E_1 + E_2)\hat{R}^{-1}. \tag{3.22}$$

Thus,

$$\|X\hat{R}^{-1}\|_2^2 \leq 1 + \|\hat{R}^{-1}\|_2^2 (\|E_1\|_2 + \|E_2\|_2).$$

By using $\|E_1\|_2 + \|E_2\|_2 \leq \frac{1.2}{64}(\sigma_n(X))^2$ again and inserting Eq. (3.21), we obtain

$$\|X\hat{R}^{-1}\|_2 \leq 1.1. \tag{3.23}$$

Evaluation of the backward error ΔX. From Eq. (3.3), we have

$$\begin{aligned}\hat{\mathbf{y}}_i^\top &= \mathbf{x}_i^\top(\hat{R} + \Delta\hat{R}_i)^{-1} \\ &= \mathbf{x}_i^\top(I + \hat{R}^{-1}\Delta\hat{R}_i)^{-1}\hat{R}^{-1}.\end{aligned}$$

Now, let

$$(I + \hat{R}^{-1}\Delta\hat{R}_i)^{-1} = I + \check{R}_i.$$

Then, since $\check{R}_i = \sum_{k=1}^{\infty}(-\hat{R}^{-1}\Delta\hat{R}_i)^k$, we obtain the bound on $\|\check{R}_i\|_2$ as

$$\begin{aligned}\|\check{R}_i\|_2 &\leq \sum_{k=1}^{\infty}(\|\hat{R}^{-1}\|_2\|\Delta\hat{R}_i\|_2)^k \\ &= \frac{\|\hat{R}^{-1}\|_2\|\Delta\hat{R}_i\|_2}{1 - \|\hat{R}^{-1}\|_2\|\Delta\hat{R}_i\|_2} \\ &\leq \frac{1.1(\sigma_n(X)^{-1}) \cdot 1.2n\sqrt{n}\mathbf{u}\|X\|_2}{1 - 1.1(\sigma_n(X)^{-1}) \cdot 1.2n\sqrt{n}\mathbf{u}\|X\|_2},\end{aligned} \tag{3.24}$$

where we used Eq. (3.20) and (3.21) in the last inequality. The denominator of Eq. (3.24) can be evaluated as

$$\begin{aligned}1 - 1.1(\sigma_n(X))^{-1} \cdot 1.2n\sqrt{n}\mathbf{u}\|X\|_2 &\geq 1 - \frac{1.1 \cdot 1.2n\sqrt{n}\mathbf{u}}{8\sqrt{mn\mathbf{u} + n(n+1)\mathbf{u}}} \\ &\geq 1 - \frac{1.32}{8}\sqrt{n\mathbf{u}} \\ &\geq 1 - \frac{1.32}{8}\sqrt{\frac{1}{11}} \geq 0.95.\end{aligned}$$

Inserting this into Eq. (3.24) and evaluating the numerator using Eq. (3.20) again, we have

$$\|\breve{R}_i\|_2 \;\leq\; \frac{1}{0.95} \cdot 1.1\kappa_2(X) \cdot 1.2n\sqrt{n}\mathbf{u} \;\leq\; 1.4\kappa_2(X)n\sqrt{n}\mathbf{u}.$$

Now, let

$$\Delta\mathbf{x}_i^\top = \mathbf{x}_i^\top \breve{R}_i.$$

Then,

$$\hat{\mathbf{y}}_i^\top = (\mathbf{x}_i^\top + \Delta\mathbf{x}_i^\top)\hat{R}^{-1}. \tag{3.25}$$

By defining the matrix $\Delta X \in \mathbb{R}^{m\times n}$ as $\Delta X = (\Delta\mathbf{x}_1, \Delta\mathbf{x}_2, \ldots, \Delta\mathbf{x}_m)^\top$, we can rewrite Eq. (3.25) as

$$\hat{Y} = (X + \Delta X)\hat{R}^{-1}. \tag{3.26}$$

The bound on $\|\Delta X\|_F$ can be given as

$$\begin{aligned}\|\Delta X\|_F = \sqrt{\sum_{i=1}^m \|\Delta\mathbf{x}_i^\top\|^2} \;&\leq\; \sqrt{\sum_{i=1}^m \|\mathbf{x}_i^\top\|^2\|\breve{R}_i\|_2^2}\\ &\leq 1.4\kappa_2(X)n\sqrt{n}\mathbf{u}\sqrt{\sum_{i=1}^m \|\mathbf{x}_i^\top\|^2} \;\leq\; 1.4\kappa_2(X)\|X\|_2 n^2\mathbf{u},\end{aligned} \tag{3.27}$$

where the relationship $\sqrt{\sum_{i=1}^m \|\mathbf{x}_i^\top\|^2} = \|X\|_F \leq \sqrt{n}\,\|X\|_2$ is used to derive the last inequality.

3.2. Orthogonality of $\hat{Y}$ and $\hat{Z}$. Based on the preparations given in the previous subsection, we evaluate the orthogonality of $\hat{Y}$ and $\hat{Z}$ computed by the Cholesky QR and CholeskyQR2 algorithms. The following lemma holds.

Lemma 3.1. *Suppose that $X \in \mathbb{R}^{m\times n}$ with $m \geq n$ satisfies Eq. (2.1) . Then, the matrix $\hat{Y}$ obtained by applying the Cholesky QR algorithm in floating point arithmetic to X satisfies the following inequality. With δ as defined in (2.1),*

$$\|\hat{Y}^\top\hat{Y} - I\|_2 \leq \frac{5}{64}\delta^2.$$

Proof. By expanding $\hat{Y}^\top\hat{Y}$ using Eq. (3.26), we have

$$\begin{aligned}\hat{Y}^\top\hat{Y} &= \hat{R}^{-\top}(X+\Delta X)^\top(X+\Delta X)\hat{R}^{-1}\\ &= \hat{R}^{-\top}X^\top X\hat{R}^{-1} + \hat{R}^{-\top}X^\top\Delta X\hat{R}^{-1} + \hat{R}^{-\top}\Delta X^\top X\hat{R}^{-1} + \hat{R}^{-\top}\Delta X^\top\Delta X\hat{R}^{-1}\\ &= I - \hat{R}^{-\top}(E_1+E_2)\hat{R}^{-1} + (X\hat{R}^{-1})^\top\Delta X\hat{R}^{-1} + \hat{R}^{-\top}\Delta X^\top(X\hat{R}^{-1}) + \hat{R}^{-\top}\Delta X^\top\Delta X\hat{R}^{-1}.\end{aligned}$$

Here, we used Eq. (3.22) to derive the last equality. Thus,

$$
\begin{aligned}
\|\hat{Y}^\top \hat{Y} - I\|_2 &\le \|\hat{R}^{-\top}(E_1+E_2)\hat{R}^{-1}\|_2 + 2\|\hat{R}^{-\top}\Delta X^\top (X\hat{R}^{-1})\|_2 + \|\hat{R}^{-\top}\Delta X^\top \Delta X \hat{R}^{-1}\|_2 \\
&\le \|\hat{R}^{-1}\|_2^2(\|E_1\|_2 + \|E_2\|_2) + 2\|\hat{R}^{-1}\|_2 \|X\hat{R}^{-1}\|_2 \|\Delta X\|_2 + \|\hat{R}^{-1}\|_2^2 \|\Delta X\|_2^2 \\
&\le \|\hat{R}^{-1}\|_2^2(\|E_1\|_2 + \|E_2\|_2) + 2\|\hat{R}^{-1}\|_2 \|X\hat{R}^{-1}\|_2 \|\Delta X\|_F + \|\hat{R}^{-1}\|_2^2 \|\Delta X\|_F^2 \\
&\le (1.1(\sigma_n(X))^{-1})^2 (1.1mn\mathbf{u} + 1.2n(n+1)\mathbf{u})\|X\|_2^2 \\
&\qquad + 2\cdot 1.1(\sigma_n(X))^{-1}\cdot 1.1\cdot 1.4\kappa_2(X)\|X\|_2 n^2\mathbf{u} \\
&\qquad + (1.1(\sigma_n(X))^{-1}\cdot 1.4\kappa_2(X)\|X\|_2 n^2\mathbf{u})^2 \\
&\le \frac{1.1^2\cdot 1.2}{64}\delta^2 + \frac{2\cdot 1.1^2\cdot 1.4}{64}\delta^2 + \left(\frac{1.1\cdot 1.4}{64}\delta^2\right)^2 \\
&\le \frac{5}{64}\delta^2. \qquad (3.28)
\end{aligned}
$$

In the fourth inequality, we used Eqs. (3.9), (3.16), (3.21), (3.23) and (3.27). In the last inequality, we simplified the expression using the assumption $\delta \le 1$. □

The next corollary follows immediately from Lemma 3.1.

COROLLARY 3.2. *The condition number of $\hat{Y}$ satisfies $\kappa_2(\hat{Y}) \le 1.1$.*

Proof. By Lemma 3.1, every eigenvalue λ_i of $\hat{Y}^\top \hat{Y}$ satisfies

$$1 - \frac{5}{64} \le \lambda_i \le 1 + \frac{5}{64}.$$

Hence, every singular value $\sigma_i(\hat{Y})$ of $\hat{Y}$ satisfies

$$\frac{\sqrt{59}}{8} \le \sigma_i(\hat{Y}) \le \frac{\sqrt{69}}{8}. \qquad (3.29)$$

Thus it follows that

$$\kappa_2(\hat{Y}) = \frac{\sigma_1(\hat{Y})}{\sigma_n(\hat{Y})} \le \sqrt{\frac{69}{59}} \le 1.1. \quad □$$

In other words, the matrix $\hat{Y}$ obtained by applying the Cholesky QR algorithm once is extremely well-conditioned, though its deviation from orthogonality, $\|\hat{Y}^\top \hat{Y} - I\|_2$, is still of order 0.1.

Combining Lemma 3.1 and Corollary 3.2, we obtain one of the main results of this paper.

THEOREM 3.3. *The matrix $\hat{Z}$ obtained by applying CholeskyQR2 in floating point arithmetic to X satisfies the following inequality.*

$$\|\hat{Z}^\top \hat{Z} - I\|_2 \le 6(mn\mathbf{u} + n(n+1)\mathbf{u}). \qquad (3.30)$$

Proof. Noting that $\kappa_2(\hat{Y}) \le \sqrt{\frac{69}{59}}$ from Corollary 3.2 and applying Lemma 3.1 again to $\hat{Y}$, we have

$$
\begin{aligned}
\|\hat{Z}^\top \hat{Z} - I\|_2 &\le \frac{5}{64}\delta^2 \le \frac{5}{64}\cdot\frac{69}{59}\cdot 64(mn\mathbf{u} + n(n+1)\mathbf{u}) \\
&\le 6(mn\mathbf{u} + n(n+1)\mathbf{u}). \quad □ \qquad (3.31)
\end{aligned}
$$

Orthogonality error in the Frobenius norm. Above, we derived the bound on the orthogonality error in terms of the 2-norm, because we wanted to give a bound on the 2-norm based condition number of $\hat{Y}$. However, by tracing the derivation of Eq. (3.28), we can also derive the following bound in the Frobenius norm,

$$\|\hat{Y}^\top\hat{Y}-I\|_F \le \|\hat{R}^{-1}\|_2^2(\|E_1\|_F+\|E_2\|_F)+2\|\hat{R}^{-1}\|_2\|X\hat{R}^{-1}\|_2\|\Delta X\|_F+\|\hat{R}^{-1}\|_2^2\|\Delta X\|_F^2.$$

As is clear from Eqs. (3.8) and (3.11), the upper bounds on $\|E_1\|_2$ and $\|E_2\|_2$ that were used in Eq. (3.28) are also bounds on $\|E_1\|_F$ and $\|E_2\|_F$. Thus, the same bound given in Eq. (3.31) holds for the Frobenius norm as well. We summarize this observation as a corollary as follows.

Corollary 3.4. *The matrix $\hat{Z}$ obtained by applying CholeskyQR2 in floating point arithmetic to X satisfies the following inequality.*

$$\|\hat{Z}^\top\hat{Z}-I\|_F \le 6(mn\mathbf{u}+n(n+1)\mathbf{u}). \tag{3.32}$$

3.3. Preparation for evaluating the residual. Let the matrices B, S, Z and U computed by floating point arithmetic be denoted by $\hat{B}=fl(\hat{Y}^\top\hat{Y})$, $\hat{S}=fl(\text{chol}(\hat{B}))$, $\hat{Z}=fl(\hat{Y}\hat{S}^{-1})$ and $\hat{U}=fl(\hat{S}\hat{R})$, resepctively. Then we have

$$\hat{B}=\hat{Y}^\top\hat{Y}+E_3, \tag{3.33}$$
$$\hat{S}^\top\hat{S}=\hat{B}+E_4 \;=\; \hat{Y}^\top\hat{Y}+E_3+E_4, \tag{3.34}$$
$$\hat{\mathbf{z}}_i^\top=\hat{\mathbf{y}}_i^\top(\hat{S}+\Delta\hat{S}_i)^{-1} \quad (i=1,2,\ldots,m), \tag{3.35}$$
$$\hat{U}=\hat{S}\hat{R}+E_5. \tag{3.36}$$

Here, $\hat{\mathbf{z}}_i^\top$ is the ith row vector of $\hat{Z}$. E_3 and E_5 are the forward errors of the matrix multiplications $\hat{Y}^\top\hat{Y}$ and $\hat{S}\hat{R}$, respectively, while E_4 is the backward error of the Cholesky decomposition of $\hat{B}$. $\Delta\hat{S}_i$ is the backward error introduced in solving the linear simultaneous equation $\mathbf{z}_i^\top\hat{S}=\hat{\mathbf{y}}_i^\top$ by forward substitution.

As a preparation of evaluating the residual, we first evaluate the norms of $\hat{R}$, $\hat{S}$, $\Delta\hat{S}_i$, E_5 and $\hat{Z}$.

Evaluation of $\hat{R}$. From Eq. (3.19), we have

$$\frac{\|\hat{R}\|_2}{\|X\|_2} \le \sqrt{\frac{1+\gamma_m n}{1-\gamma_{n+1}n}} \le \sqrt{\frac{1+\frac{mn\mathbf{u}}{1-m\mathbf{u}}}{1-\frac{n(n+1)\mathbf{u}}{1-(n+1)\mathbf{u}}}} \le \sqrt{\frac{1+\frac{\frac{1}{64}}{1-\frac{1}{11}}}{1-\frac{\frac{1}{64}}{1-\frac{1}{11}}}} = \sqrt{\frac{651}{629}} \le 1.1. \tag{3.37}$$

Evaluation of $\hat{S}$. Noticing that $\|\hat{Y}\|_2 \le \frac{\sqrt{69}}{8}$ from Eq. (3.29), we can obtain an upper bound on the norm of $\hat{S}$ by multiplying the bound of Eq. (3.37) by $\frac{\sqrt{69}}{8}$. Thus,

$$\|\hat{S}\|_2 \le \sqrt{\frac{651}{629}}\cdot\frac{\sqrt{69}}{8} \le 1.1.$$

Evaluation of $\Delta\hat{S}_i$. Similarly, multiplying the bound of Eq. (3.20) by $\frac{\sqrt{69}}{8}$ leads to the following bound on $\Delta\hat{S}_i$.

$$\|\Delta\hat{S}_i\|_2 \le \frac{\sqrt{69}}{8}\cdot 1.2n\sqrt{n}\mathbf{u} \le 1.3n\sqrt{n}\mathbf{u}.$$

Evaluation of E_5. By using the error bound on matrix multiplication given in Eq. (3.6), we have

$$|E_5| \leq \gamma_n |\hat{S}||\hat{R}|.$$

Hence,

$$\|E_5\|_2 \leq \| \, |E_5| \, \|_F \leq \gamma_n \| \, |\hat{S}||\hat{R}| \, \|_F \leq \gamma_n \| \, |\hat{S}| \, \|_F \| \, |\hat{R}| \, \|_F = \gamma_n \|\hat{S}\|_F \|\hat{R}\|_F \leq n\gamma_n \|\hat{S}\|_2 \|\hat{R}\|_2$$
$$\leq n \cdot 1.1 n\mathbf{u} \cdot \sqrt{\frac{651}{629}} \cdot \frac{\sqrt{69}}{8} \cdot \sqrt{\frac{651}{629}} \|X\|_2 \leq 1.2 n^2 \mathbf{u} \|X\|_2.$$

Evaluation of $\hat{Z}$. From Eq. (3.29), we have $\|\hat{Y}\|_F \leq \frac{\sqrt{69}}{8}\sqrt{n}$. Multiplying this by $\frac{\sqrt{69}}{8}$ yields the following upper bound on $\hat{Z}$.

$$\|\hat{Z}\|_F \leq \frac{69}{64}\sqrt{n} \leq 1.1\sqrt{n}.$$

3.4. Bounding the residual. Based on the above results, we evaluate the residual of the pair $(\hat{Z}, \hat{U})$. The following theorem holds, which is also one of our main results.

THEOREM 3.5. *Assume that an* $m \times n$ *real matrix* X $(m \geq n)$ *satisfies Eq. (2.1). Then the matrices* $\hat{Z}$ *and* $\hat{U}$ *obtained by applying the CholeskyQR2 algorithm in floating point arithmetic to* X *satisfy the following inequality.*

$$\frac{\|\hat{Z}\hat{U} - X\|_F}{\|X\|_2} \leq 5n^2\sqrt{n}\mathbf{u}. \tag{3.38}$$

Proof. Expanding $\hat{\mathbf{z}}_i^\top \hat{U} - \mathbf{x}_i^\top$ using Eqs. (3.36), (3.35) and (3.3) leads to

$$\begin{aligned}
\|\hat{\mathbf{z}}_i^\top \hat{U} - \mathbf{x}_i^\top\| &= \|\hat{\mathbf{z}}_i^\top(\hat{S}\hat{R} + E_5) - \hat{\mathbf{z}}_i^\top(\hat{S} + \Delta\hat{S}_i)(\hat{R} + \Delta\hat{R}_i)\| \\
&= \|\hat{\mathbf{z}}_i^\top \hat{S}\hat{R} + \hat{\mathbf{z}}_i^\top E_5 - \hat{\mathbf{z}}_i^\top \hat{S}\hat{R} - \hat{\mathbf{z}}_i^\top \hat{S}\Delta\hat{R}_i - \hat{\mathbf{z}}_i^\top \Delta\hat{S}_i\hat{R} - \hat{\mathbf{z}}_i^\top \Delta\hat{S}_i\Delta\hat{R}_i\| \\
&\leq \|\hat{\mathbf{z}}_i^\top\|(\|E_5\|_2 + \|\hat{S}\|_2\|\Delta\hat{R}_i\|_2 + \|\Delta\hat{S}_i\|_2\|\hat{R}\|_2 + \|\Delta\hat{S}_i\|_2\|\Delta\hat{R}_i\|_2) \\
&\leq \|\hat{\mathbf{z}}_i^\top\|(1.2n^2\mathbf{u} + 1.1 \cdot 1.2n\sqrt{n}\mathbf{u} + 1.3n\sqrt{n}\mathbf{u} \cdot 1.1 + 1.3n\sqrt{n}\mathbf{u} \cdot 1.2n\sqrt{n}\mathbf{u})\|X\|_2 \\
&\leq \|\hat{\mathbf{z}}_i^\top\| \, \|X\|_2 \cdot 4n^2\mathbf{u}. \qquad (3.39)
\end{aligned}$$

Hence,

$$\frac{\|\hat{Z}\hat{U} - X\|_F}{\|X\|_2} = \frac{\sqrt{\sum_{i=1}^n \|\hat{\mathbf{z}}_i^\top \hat{U} - \hat{\mathbf{x}}_i^\top\|^2}}{\|X\|_2} \leq 4n^2\mathbf{u}\sqrt{\sum_{i=1}^n \|\hat{\mathbf{z}}_i^\top\|^2} = 4n^2\mathbf{u}\|\hat{Z}\|_F \leq 5n^2\sqrt{n}\mathbf{u}. \quad \square$$

4. Numerical results. Here we evaluate the numerical stability of CholeskyQR2 and compare it with the stability of other popular QR decomposition algorithms, namely, Householder QR, classical and modified Gram-Schmidt (CGS and MGS; we also run them twice, shown as CGS2 and MGS2) and Cholesky QR. To this end, we generated test matrices with a specified condition number by $X := U\Sigma V \in \mathbb{R}^{m\times n}$, where U is an $m \times n$ random orthogonal matrix, V is an $n \times n$ random orthogonal matrix and

$$\Sigma = \mathrm{diag}(1, \sigma^{\frac{1}{n-1}}, \cdots, \sigma^{\frac{n-2}{n-1}}, \sigma).$$

Here, $0 < \sigma < 1$ is some constant. Thus $\|X\|_2 = 1$ and the 2-norm condition number of X is $\kappa_2(X) = 1/\sigma$. We varied $\kappa_2(X)$, m and n and investigated the dependence of the orthogonality and residual on them. All computations were done on Matlab 2012b using IEEE standard 754 binary64 (double precision) on Mac OS X version 10.8 with 2 GHz Intel Core i7 Duo processor, so that $\mathbf{u} = 2^{-53} \approx 1.11 \times 10^{-16}$.

We show the orthogonality and residual measured by the Frobenius norm under various conditions in Figures 4.1 through 4.6. Figures 4.1 and 4.2 show the orthogonality $\|\hat{Z}^T\hat{Z} - I\|_F$ and residual $\|\hat{Z}\hat{U} - X\|_F$, respectively, for the case $m = 10,000$, $n = 100$ and varying $\kappa_2(X)$. In Figures 4.3 and 4.4, $\kappa_2(X) = 10^5$, $n = 100$ and m was varied from 1,000 to 10,000. In Figures 4.5 and 4.6, $\kappa_2(X) = 10^5$, $m = 1,000$ and n was varied from 100 to 1,000.

It is clear from Figures 4.1 and 4.2 that both the orthogonality and residual are independent of $\kappa_2(X)$ and are of $O(\mathbf{u})$, as long as $\kappa_2(X)$ is at most $O(\mathbf{u}^{-\frac{1}{2}})$. This is in good agreement with the theoretical prediction and is in marked contrast to the results of CGS, MGS and Cholesky QR, for which the deviation from orthogonality increases in proportional to $\kappa_2(X)$ and $(\kappa_2(X))^2$, respectively. As can be seen from Figures 4.3 through 4.6, the orthogonality and residual increase only mildly with m and n, which is also in agreement with the theoretical results, although they are inevitably overestimates. Compared with Householder QR, it was observed that CholeskyQR2 generally produces smaller orthogonality and residual. From these results, we can conclude that CholeskyQR2 is stable for matrices with condition number at most $O(\mathbf{u}^{\frac{1}{2}})$. As is well known, Gram-Schmidt type algorithms perform well when repeated twice.

5. Discussion. In this section, we discuss four topics related to the stability of CholeskyQR2. First, we compare the orthogonality and residual bounds of CholeskyQR2 given in Theorems 3.4 and 3.5, respectively, with known bounds for Householder QR [11] and CGS2 [9]. Second, we consider how to examine the applicability of CholeskyQR2 for a given matrix. Third, we show that CholeskyQR2 is not only norm-wise stable, but also column-wise stable. Finally, we discuss row-wise stability of CholeskyQR2, which cannot be proved but is nearly always observed in practice.

5.1. Comparison with the error bounds of Householder QR and CGS2.

Orthogonality. For Householder QR, the Q factor is computed by applying n Householder transformations to $I_{1:m,1:n}$, an $m \times n$ matrix consisting of the first n columns of the identity matrix of order m. Hence, from Lemma 19.3 of [11], the computed Q factor satisfies

$$\hat{Q} = P^\top(I_{1:m,1:n} + \Delta I),$$

where P is some $m \times m$ exactly orthogonal matrix and ΔI is an $m \times n$ matrix whose each column vector has a norm bounded by $n\gamma_{cm}$, where c is a small positive constant. From this, it is easy to derive the bound

$$\|\hat{Q}^\top\hat{Q} - I\|_F \le n\sqrt{n}\gamma_{c'm} \simeq c'mn\sqrt{n}\mathbf{u}.$$

For CGS2, Giraud et al. show the following bound for deviation from orthogonality under the assumption that $\kappa_2(X)m^2n^3\mathbf{u} = O(1)$ [9].

$$\|\hat{Q}^\top\hat{Q} - I\|_2 \le c''mn\sqrt{n}\mathbf{u}.$$

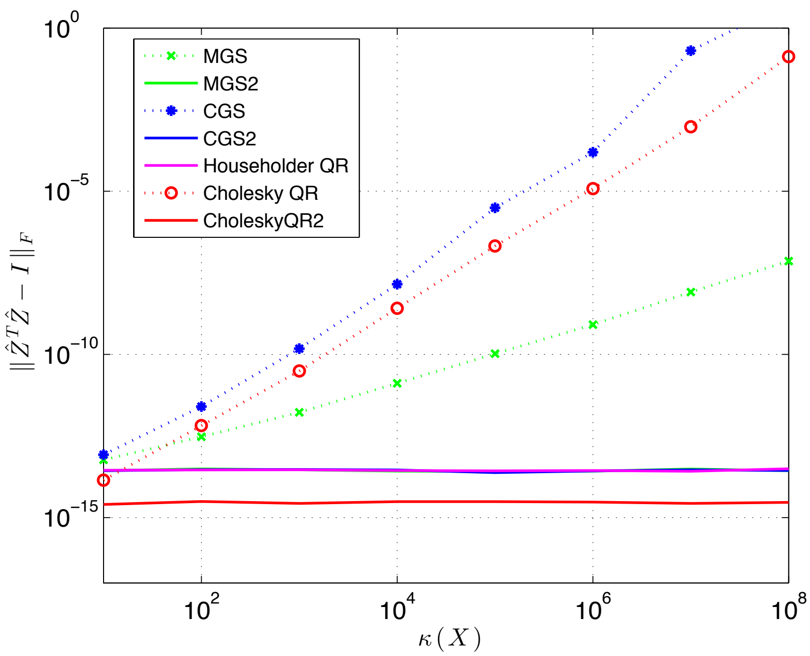

FIG. 4.1. *Orthogonality* $\|\hat{Z}^\top \hat{Z} - I\|_F$ *for test matrices with* $m = 10,000$*,* $n = 100$*, varying* $\kappa_2(X)$*.*

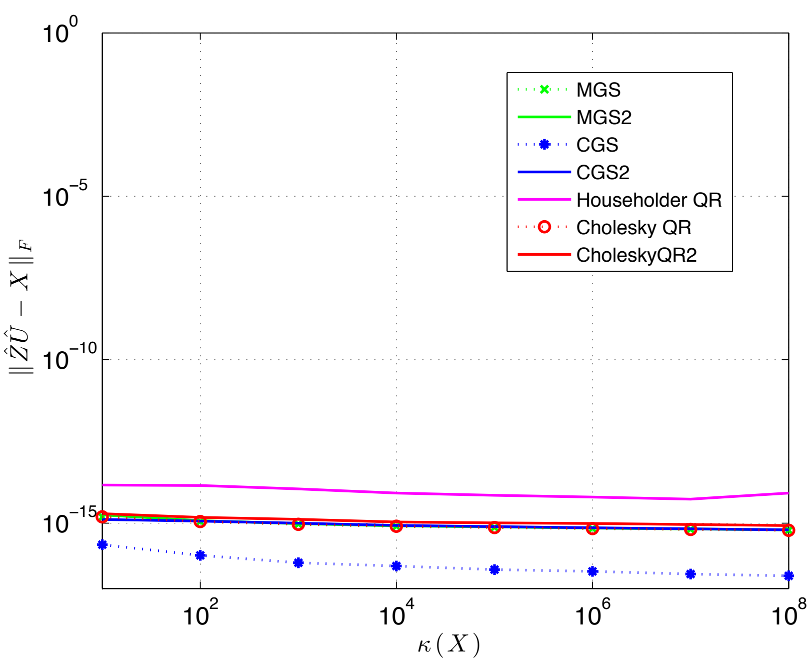

FIG. 4.2. *Residual* $\|\hat{Z}\hat{U} - X\|_F$ *for test matrices with* $m = 10,000$*,* $n = 100$*, varying* $\kappa_2(X)$*.*

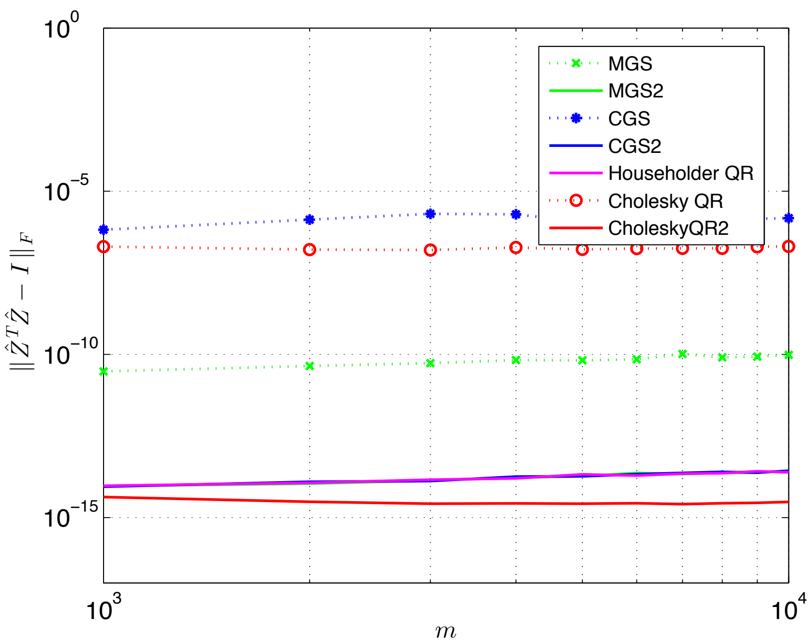

FIG. 4.3. *Orthogonality* $\|\hat{Z}^\top \hat{Z} - I\|_F$ *for test matrices with* $\kappa_2(X) = 10^5$, $n = 100$, *varying* m.

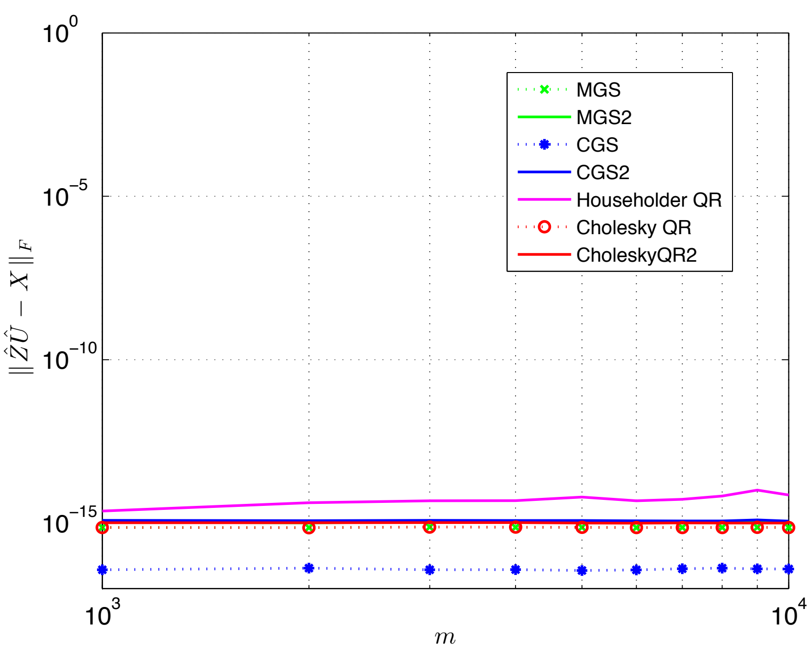

FIG. 4.4. *Residual* $\|\hat{Z}\hat{U} - X\|_F$ *for test matrices with* $\kappa_2(X) = 10^5$, $n = 100$, *varying* m.

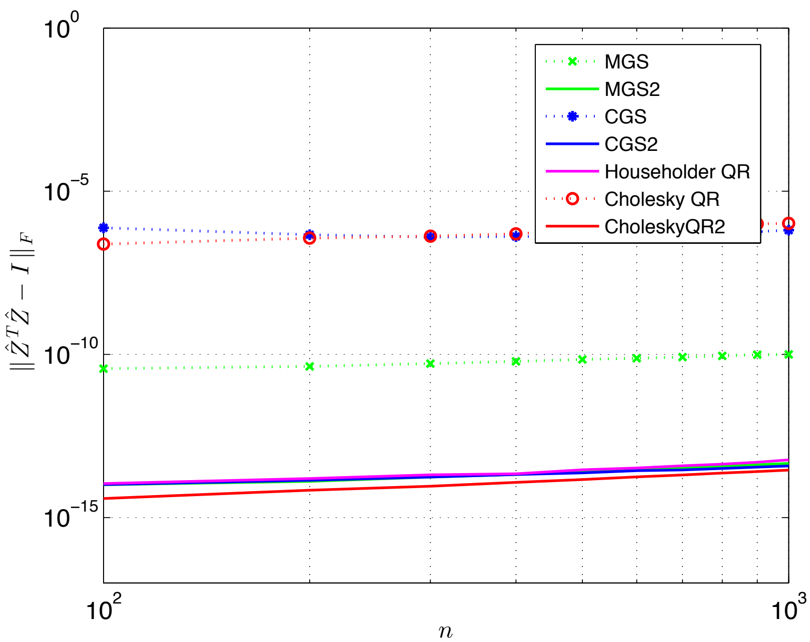

FIG. 4.5. *Orthogonality* $\|\hat{Z}^\top \hat{Z} - I\|_F$ *for test matrices with* $\kappa_2(A) = 10^5$*,* $m = 1,000$*, varying* n*.*

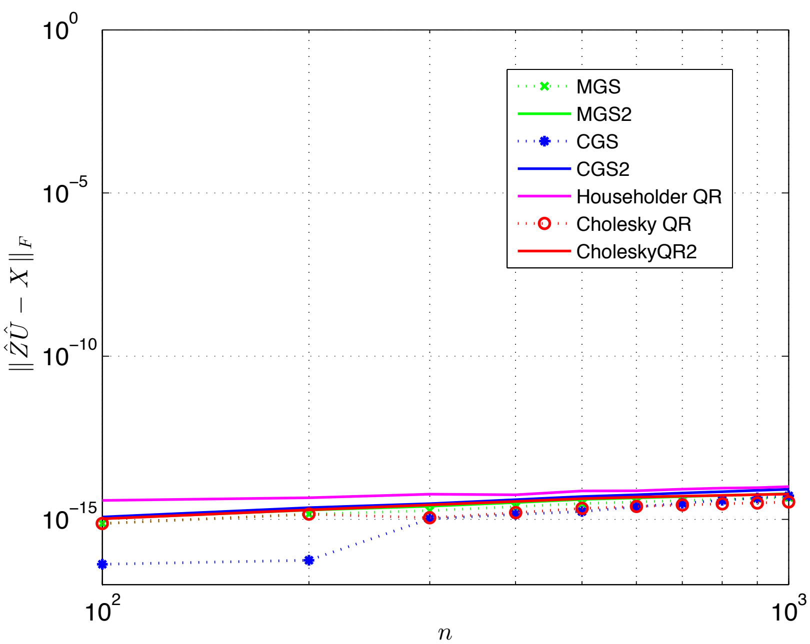

FIG. 4.6. *Residual* $\|\hat{Z}\hat{U} - X\|_F$ *for test matrices with* $\kappa_2(A) = 10^5$*,* $m = 1,000$*, varying* n*.*

Although this assumption is hard to satisfy for large matrices (notice that $\kappa_2(X)m^2n^3$ is 10^{19} for the largest matrix appearing in Fig. 4.3), it has been observed that CGS2 produces near-orthogonal matrices in many cases where this condition is violated [14].

Comparing these bounds with Eq. (3.32), we observe that the error bound of CholeskyQR2 is smaller by a factor of $\sqrt{n}$. This is in qualitative agreement with the results of numerical experiments given in the previous section.

Note however that this difference should not be overemphasized, because these are merely upper bounds. In fact, the Givens QR algorithm admits error bound that is smaller than that of Householder QR by a factor of n, but it is observed that no difference in accuracy is seen in practice [11, p.368].

Residual. According to [11, Sec. 19.3], the upper bound on the residual of the Householder QR algorithm can be evaluated as $O(mn\sqrt{n}\mathbf{u})$. As for CGS2, it is not difficult to derive a bound of the same order using the results given in [9]. Thus, we can say that the CholsekyQR2 algorithm has a smaller bound also in terms of the residual. This is related to the fact that in the CholeskyQR2 algorithm, the computation of Y from X and Z from Y is done by row-wise forward substitution. Thus the backward errors introduced there or their sum of squares, which is one of the main sources of the residual, do not depend on m when $\|X\|_2$ is fixed. In addition, the forward error in the computation of $\hat{S}\hat{R}$, which is another source of residual, also involves only n. Thus the residual depends only on n, which is in marked contrast to Householder QR.

A few more comments are in order regarding the bound (3.38). A close examination of Eq. (3.39) shows that the highest order term in the residual comes from the forward error of the matrix multiplication $\hat{S}\hat{R}$, which we denoted by E_5. This implies that if we compute this matrix multiplication using extended precision arithmetic, we can reduce the upper bound on the residual to $O(n^2\mathbf{u})$ with virtually no increase in the computational cost (when $m \gg n$). Moreover, in a situation where only the orthogonal factor $\hat{Z}$ is needed, as in orthogonalization of vectors, we can leave the product $\hat{S}\hat{R}$ uncomputed and say that the triplet $(\hat{Z}, \hat{S}, \hat{R})$ has residual of $O(n^2\mathbf{u})$.

5.2. Applicability of CholeskyQR2 for a given matrix. There are some cases in which the condition number of X is known in advance to be moderate. An example is orthogonalization of vectors in first-principles molecular dynamics [3]. In this application, we are interested in the time evolution of an orthogonal matrix $X(t) \in \mathbb{R}^{m\times n}$, whose column vectors are orthogonal basis of the space of occupied-state wave functions. To obtain $X(t+\Delta t)$, we first compute $\tilde{X} = X(t) - F(X)\Delta t$, where $F(X) \in \mathbb{R}^{m\times n}$ is some nonlinear matrix function of X, and then compute $X(t+\Delta t)$ by orthogonalizing the columns of $\tilde{X}$. Since $X(t)$ is orthogonal, we can easily evaluate the deviation from orthogonality of $\tilde{X}$ by computing the norm of $F(X)\Delta t$. Usually, the time step Δt is small enough to ensure that $\kappa_2(\tilde{X}) \ll \mathbf{u}^{-\frac{1}{2}}$.

In some cases, however, the condition number of X cannot be estimated in advance and one may want to examine the applicability of CholeskyQR2 from intermediate quantities that are computed in the algorithm. This is possible if $\hat{R}$ has been computed without breakdown in the Cholesky decomposition. Given $\hat{R}$, one can estimate its largest and smallest singular values using the power method and inverse power method on $R^\top R$, respectively. Indeed the MATLAB condition number estimator `condest` first computes the LU factorization of the input matrix, then applies a few iterations of power method to obtain a reliable estimate of the 1-norm condition number. This should not cost too much because $\hat{R}$ is triangular and each step of both methods requires only $O(n^2)$ work. After that, one can evaluate the condition number of X by using the relations (3.1) and (3.2), the bounds (3.9) and (3.16) on $\|E_1\|_2$ and $\|E_2\|_2$, respectively, and the Bauer-Fike theorem.

5.3. Column-wise stability of CholeskyQR2. Thus far we have investigated the normwise residual of CholeskyQR2. Sometimes the columns of X have widely varying norms, and one may wish to obtain the more stringent column-wise backward stability, which requires

$$\|\tilde{\mathbf{x}}_j - \hat{Q}\hat{\mathbf{r}}_j\| / \|\tilde{\mathbf{x}}_j\| = O(\mathbf{u}), \quad j = 1, \ldots, n.$$

Here, $\tilde{\mathbf{x}}_j$ and $\hat{\mathbf{r}}_j$ denote the jth columns of X and $\hat{R}$, respectively. In this subsection, we prove that CholeskyQR2 is indeed column-wise backward stable.

To see this, we first consider a single Cholesky QR and show that the computed $\|\hat{\mathbf{r}}_j\|$ is of the same order as $\|\tilde{\mathbf{x}}_j\|$. Let us recall Eqs. (3.1) through (3.3). From Eq. (3.7), we have

$$|E_1|_{jj} \leq \gamma_m |\tilde{\mathbf{x}}_j|^\top |\tilde{\mathbf{x}}_j| = \gamma_m \|\tilde{\mathbf{x}}_j\|^2. \tag{5.1}$$

By considering the (j,j)th element of Eq. (3.2) and substituting Eqs. (5.1) and (3.10), we obtain

$$\begin{aligned} \|\hat{\mathbf{r}}_j\|^2 &\leq |\hat{A}_{jj}| + \gamma_{n+1} |\hat{\mathbf{r}}_j|^\top |\hat{\mathbf{r}}_j| \\ &\leq \|\tilde{\mathbf{x}}_j\|^2 + \gamma_m \|\tilde{\mathbf{x}}_j\|^2 + \gamma_{n+1} \|\hat{\mathbf{r}}_j\|^2. \end{aligned}$$

Hence,

$$\|\hat{\mathbf{r}}_j\| \leq \sqrt{\frac{1+\gamma_m}{1-\gamma_{n+1}}}\, \|\tilde{\mathbf{x}}_j\| = \|\tilde{\mathbf{x}}_j\| \cdot O(1). \tag{5.2}$$

Now we demonstrate the column-wise backward stability of a single Cholesky QR. Let the jth column of $\hat{Y}$ be denoted by $\tilde{\mathbf{y}}_j$. From Eq. (3.3), we have

$$X_{ij} = \hat{\mathbf{y}}_i^\top (\hat{\mathbf{r}}_j + \Delta \hat{\mathbf{r}}_j^{(i)}). \tag{5.3}$$

Here, $\Delta \hat{\mathbf{r}}_j^{(i)}$ is the jth column of $\Delta \hat{R}_i$. Thus,

$$\begin{aligned} |X_{ij} - \hat{\mathbf{y}}_i^\top \hat{\mathbf{r}}_j| &\leq |\hat{\mathbf{y}}_i^\top \Delta \hat{\mathbf{r}}_j^{(i)}| \\ &\leq \|\hat{\mathbf{y}}_i\| \, \|\Delta \hat{\mathbf{r}}_j^{(i)}\| \\ &\leq \gamma_n \|\hat{\mathbf{y}}_i\| \, \|\hat{\mathbf{r}}_j\|. \end{aligned}$$

Squaring both sides and summing over i leads to

$$\|\tilde{\mathbf{x}}_j - \hat{Y}\hat{\mathbf{r}}_j\|^2 \leq \gamma_n^2 \|\hat{Y}\|_F^2 \|\hat{\mathbf{r}}_j\|^2.$$

By using $\|\hat{Y}\|_F = O(1)$ (see Lemma 3.1) and Eq. (5.2), we can establish the column-wise backward stability of Cholesky QR as follows.

$$\frac{\|\tilde{\mathbf{x}}_j - \hat{Y}\hat{\mathbf{r}}_j\|}{\|\tilde{\mathbf{x}}_j\|} \leq \gamma_n \|\hat{Y}\|_F \cdot \frac{\|\hat{\mathbf{r}}_j\|}{\|\tilde{\mathbf{x}}_j\|} = \gamma_n \cdot O(1) \cdot \sqrt{\frac{1+\gamma_m}{1-\gamma_{n+1}}}. \tag{5.4}$$

To apply the above result to CholeskyQR2, we consider the backward errors in the second QR decomposition $Y = ZS$ and the product of the two upper triangular factors $U = SR$. These backward errors, which we denote by $\Delta\hat{Y}$ and $\Delta\hat{S}$, respectively, satisfy

$$\hat{Y} + \Delta\hat{Y} = \hat{Z}\hat{S},$$
$$\hat{\mathbf{u}}_j = (\hat{S} + \Delta\hat{S})\hat{\mathbf{r}}_j.$$

Here, $\hat{\mathbf{u}}_j$ is the j th column of $\hat{U}$. To evaluate $\Delta\hat{Y}$, we note the following inequality, which can be obtained in the same way as Eq. (5.4).

$$\|\tilde{\mathbf{y}}_j - \hat{Z}\hat{\mathbf{s}}_j\|^2 \leq \gamma_n^2 \|\hat{Z}\|_F^2 \|\hat{\mathbf{s}}_j\|^2.$$

Summing both sides over i and taking the square root gives

$$\|\Delta\hat{Y}\|_F = \|\hat{Y} - \hat{Z}\hat{S}\|_F \leq \gamma_n \|\hat{Z}\|_F \|\hat{S}\|_F = \gamma_n \cdot O(1). \tag{5.5}$$

As for $\Delta\hat{S}$, the standard result on the error analysis of matrix-vector product, combined with $\|\hat{S}\|_F \simeq \|\hat{Y}\|_F = O(1)$, leads to

$$\|\Delta\hat{S}\|_F \leq \gamma_n \|\hat{S}\|_F = \gamma_n \cdot O(1). \tag{5.6}$$

On the other hand,

$$\begin{aligned} \tilde{\mathbf{x}}_j - \hat{Z}\hat{\mathbf{u}}_j &= \tilde{\mathbf{x}}_j - \hat{Z}(\hat{S} + \Delta\hat{S})\hat{\mathbf{r}}_j \\ &= \tilde{\mathbf{x}}_j - (\hat{Y} + \Delta\hat{Y} + \hat{Z}\Delta\hat{S})\hat{\mathbf{r}}_j \\ &= (\tilde{\mathbf{x}}_j - \hat{Y}\hat{\mathbf{r}}_j) - (\Delta\hat{Y} + \hat{Z}\Delta\hat{S})\hat{\mathbf{r}}_j. \end{aligned} \tag{5.7}$$

By substituting Eqs. (5.4), (5.5) and (5.6) into Eq. (5.7), we finally obtain the column-wise backward stability of CholeskyQR2 as follows.

$$\frac{\|\tilde{\mathbf{x}}_j - \hat{Z}\hat{\mathbf{u}}_j\|}{\|\tilde{\mathbf{x}}_j\|} \leq \frac{\|\tilde{\mathbf{x}}_j - \hat{Y}\hat{\mathbf{r}}_j\|}{\|\tilde{\mathbf{x}}_j\|} + (\|\Delta\hat{Y}\|_F + \|\hat{Z}\|_F \|\Delta\hat{S}\|_F) \cdot \frac{\|\hat{\mathbf{r}}_j\|}{\|\tilde{\mathbf{x}}_j\|} = \gamma_n \cdot O(1).$$

5.4. Row-wise stability of CholeskyQR2. In this subsection, we investigate the row-wise stability of CholeskyQR2, which is defined as

$$\|\mathbf{x}_i^\top - \hat{\mathbf{q}}_i^\top \hat{R}\| / \|\mathbf{x}_i^\top\| = O(\mathbf{u}), \quad i = 1, \ldots, m. \tag{5.8}$$

Here $\mathbf{x}_i^\top$ and $\hat{\mathbf{q}}_i^\top$ denote the ith rows of the matrices.

The requirement (5.8) is strictly more stringent than the normwise stability, and indeed the standard Householder QR factorization does not always achieve (5.8). It is known [11, Ch.19] that when row sorting and column pivoting are used, Householder QR factorization gives row-wise stability. However, pivoting involves an increased communication cost and is best avoided in high-performance computing.

Having established the normwise and column-wise stability of CholeskyQR2, we now examine its row-wise stability. To gain some insight we first run experiments with a semi-randomly generated matrix X, whose row norms vary widely. Specifically, we generate a random $m \times n$ matrix via the MATLAB command `X = randn(m,n)`, then left-multiply a diagonal matrix $X := DX$ with $D_{jj} = 2^{\frac{j}{2}}$ for $j = 1, \ldots, m$. Here we took $m = 100$ and $n = 50$; the matrix thus has rows of exponentially growing norms and $\kappa_2(X) \approx \mathbf{u}^{-\frac{1}{2}}$. Figure 5.1 shows the row-wise residuals of three algorithms: standard Householder QR, Householder QR employing row sorting and column pivoting, and CholeskyQR2.

We make several observations from Figure 5.1. First, we confirm the known fact that the standard Householder QR factorization is not row-wise backward stable, but this can be cured by employing row sorting and column pivoting. Second, CholeskyQR2 gave row-wise stability comparable to Householder QR with row sorting and column piroving; this is perhaps surprising considering the fact that CholeskyQR2 employs no pivoting or sorting.

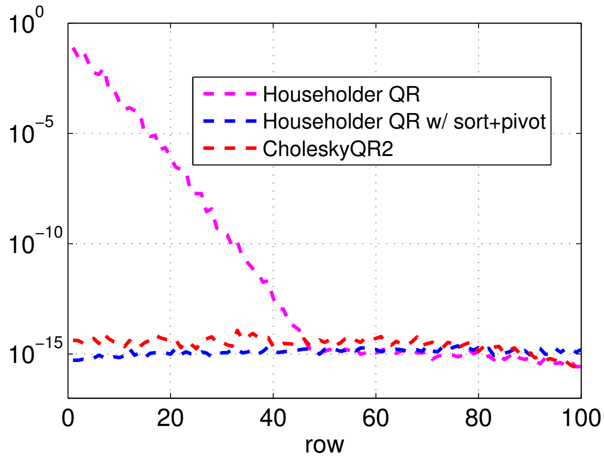

FIG. 5.1. *Row-wise residual* $\|\mathbf{x}_i^\top - \hat{\mathbf{q}}_i^\top \hat{R}\|_2/\|\mathbf{x}_i^\top\|_2$.

To illustrate the situation, we examine the first step of CholeskyQR2. Recall that $\hat{\mathbf{q}}_i^\top = fl(\mathbf{x}_i^\top \hat{R}^{-1})$. Hence $\|\mathbf{x}_i^\top - \hat{\mathbf{q}}_i^\top \hat{R}\|_2 = \|\mathbf{x}_i^\top - fl(\mathbf{x}_i^\top \hat{R}^{-1})\hat{R}\|_2$, and by standard triangular solve there exist ΔR_i for $i = 1, \dots, m$ such that

$$fl(\mathbf{x}_i^\top \hat{R}^{-1})(\hat{R} + \Delta R_i) = \mathbf{x}_i^\top, \quad \|\Delta R_i\| = O(\mathbf{u})\|\hat{R}\|.$$

Hence for row-wise stability we need $\|fl(\mathbf{x}_i^\top \hat{R}^{-1})\Delta R_i\| = O(\mathbf{u})\|\mathbf{x}_i^\top\|$. Since $\|fl(\mathbf{x}_i^\top \hat{R}^{-1})\Delta R_i\| \leq O(\mathbf{u})\|\hat{R}\|\|fl(\mathbf{x}_i^\top \hat{R}^{-1})\|$, a sufficient condition is

$$\|fl(\mathbf{x}_i^\top \hat{R}^{-1})\| = O(\|\mathbf{x}_i^\top\|/\|\hat{R}\|). \tag{5.9}$$

Since the general normwise bound for $\|\mathbf{y}^\top R^{-1}\|$ is $\|\mathbf{y}^\top R^{-1}\| \leq \|\mathbf{y}\|/\|R\|\kappa_2(R)$, the condition (5.9) is significantly more stringent when R is ill-conditioned.

Even so, as illustrated in the example above, in all our experiments with random matrices the condition (5.9) was satisfied with $\|\mathbf{x}_i^\top - \hat{\mathbf{q}}_i^\top \hat{R}\|/\|\mathbf{x}_i^\top\| < n\mathbf{u}$ for all i. We suspect that this is due to the observation known to experts that triangular linear systems are usually solved to much higher accuracy than the theoretical bound suggests [11][20]. However, as with this classical observation, counterexamples do exist in our case: For example, taking R to be the Kahan matrix [11], which are ill-conditioned triangular matrices known to have special properties, the bound $\|fl(\mathbf{y}^\top R^{-1})\| = O(\|\mathbf{y}^\top\|/\|R\|)$ is typically tight for a randomly generated $\mathbf{y}^\top$, which means (5.9) is significantly violated. In view of this we form X so that the Cholesky factor R of $X^\top X$ is the Kahan matrix. This can be done by taking $X = QR$ for an $m \times n$ orthogonal matrix Q. To introduce large variation in the row norms of X we construct Q as the orthogonal factor of a matrix as generated in the example above.

For every such X with varying size n, (5.9) was still satisfied. Finally, we then appended a row at the bottom of X of random elements with much smaller norm than the rest, and repeated the experiment. Now the row-wise residual for the last row was significantly larger than $O(\mathbf{u}\|\mathbf{x}_i^\top\|)$, indicating row-wise stability does not always hold. Employing pivoting in the Cholesky factorization did not improve the residual.

A referee has suggested more examples for which CholeskyQR2 fails to have row-wise backward stability. One example is as follows: take X to be the off-diagonal parts of the 6×6 Hilbert matrix and setting the $(3, 3)$ element to $5e6$.

Experiments suggest nonetheless that cases in which CholeskyQR is not row-wise stable is extremely rare.

6. Conclusion. In this paper, we performed roundoff error analysis of the CholeskyQR2 algorithm for computing the QR decomposition of an $m \times n$ real matrix X, where $m \geq n$. We showed that if X satisfies Eq. (2.1), the computed Q and R factors, which we denote by $\hat{Z}$ and $\hat{U}$, respectively, satisfy the following error bounds.

$$\begin{aligned} \|\hat{Z}^\top \hat{Z} - I\|_F &\leq 6(mn\mathbf{u} + n(n+1)\mathbf{u}), \\ \|\hat{Z}\hat{U} - X\|_F / \|X\|_2 &\leq 5n^2\sqrt{n}\mathbf{u}. \end{aligned} \tag{6.1}$$

The bounds shown here is of a smaller order than the corresponding bounds for the Householder QR algorithm. Furthermore, it was shown that when only the Q factor is required, the right hand side of Eq. (6.1) can be reduced to $O(n^2\mathbf{u})$. Numerical experiments support our theoretical analysis. CholeskyQR2 is also column-wise backward stable, as Householder QR. We also observed that the row-wise stability, which is a more stringent condition than the norm-wise stability shown by Eq. (6.1), nearly always holds in practice, though it cannot be proved theoretically.

In this paper, we focused on the stability of CholeskyQR2. Performance results of CholeskyQR2 on large scale parallel machines, along with comparison with other QR decomposition algorithms and detailed performance analysis, is given in our recent paper [8].

When the matrix is near square, it might be more efficient to partition the matrix into panels and apply the CholeskyQR2 algorithm to each panel successively. Development of such an algorithm remains as future work.

Acknowledgments. We thank Professor Nicholas J. Higham for suggesting the Kahan matrix for a possible counterexample of the row-wise backward stability of CholeskyQR2. We also thank Professor Masaaki Sugihara for valuable comments on the first version of our manuscript. We are grateful to the referees for their suggestions, especially for another counterexample for the row-wise backward stability and a comment on columnwise stability.

REFERENCES

[1] G. Ballard, J. Demmel, O. Holtz and O. Schwartz, *Minimizing communication in numerical linear algebra*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 866–901.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, 2000.

[3] R. Car and M. Parrinello, *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[4] J. W. Daniel, W. B. Gragg, L. Kaufman and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.

[5] J. Demmel, L. Grigori, M. Hoemmen and J. Langou, *Communication-optimal parallel and sequential QR and LU factorizations*, SIAM J. Sci. Comput, 34 (2012), pp. 206–239.
[6] E. Elmroth and F. G. Gustavson, *Applying recursion to serial and parallel QR factorization leads to better performance*, IBM J. Res. Dev., 44 (2000), pp. 605–624.
[7] T. Fukaya, T. Imamura and Y. Yamamoto, *Performance analysis of the Householder-type parallel tall-skinny QR factorizations toward automatic algorithm selection*, The Ninth International Workshop on Automatic Performance Tuning (iWAPT2014), accepted.
[8] T. Fukaya, Y. Nakatsukasa, Y. Yanagisawa and Y. Yamamoto, *CholeskyQR2: A practical and communication-avoiding algorithm for tall-skinny QR factorizations*, to appear in Proceedings of ScalA2014.
[9] L. Giraud, J. Langou, M. Rozložník and J. van den Eshof, *Rounding error analysis of the classical Gram-Schmidt orthogonalization process*, Numer. Math., 101 (2005), pp. 87–100.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th Ed., Johns Hopkins University Press, 2012.
[11] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd Edition, SIAM, 2002.
[12] J. Iwata, D. Takahashi, A. Oshiyama, T. Boku, K. Shiraishi, S. Okada and K. Yabana, *A massively-parallel electronic-structure calculations based on real-space density functional theory*, J. Comput. Phys., 229 (2010), pp. 2339–2363.
[13] W. Jalby and B. Philippe, *Stability analysis and improvement of the block Gram-Schmidt algorithm*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 1058–1073.
[14] S. J. Leon, A. Björck and W. Gander *Gram-Schmidt orthogonalization: 100 years and more*, Numer. Lin. Alg. Appl., 20 (2013), pp. 492–532.
[15] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, 1998.
[16] M. Rozložník, F. Okulicka-Dłużewska and A. Smoktunowicz, *Indefinite orthogonalization with rounding errors*, submitted.
[17] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, 1992.
[18] R. Schreiber and C. F. van Loan, *A storage-efficient WY representation for products of Householder transformations*, SIAM J. Sci. Stat. Comp., 10 (1989), pp. 53–57.
[19] A. Stathopoulos and K. Wu, *A block orthogonalization procedure with constant synchronization requirements*, SIAM J. Sci. Comput., 23 (2002), pp. 2165–2182.
[20] G. W. Stewart: *Introduction to Matrix Computations*, Academic Press, 1973.
[21] G. W. Stewart, *Matrix Algorithms, Vol. 1: Basic Decompositions*, SIAM, 1998.
[22] S. Toledo and E. Rabani, *Very large electronic structure calculations using an out-of-core filter-diagonalization method*, J. Comput. Phys., 180 (2002), pp. 256–269.