Reorthogonalization Methods Revisited

Csaba J. Hegedüs

Eötvös Loránd University, Dept. of Numerical Analysis
1117 Budapest, Pázmány P. sétány 1/C
hegedus@numanal.inf.elte.hu

Abstract: New theoretical background of Parlett-Kahan's "twice is enough" algorithm for computing accurate vectors in Gram-Schmidt orthogonalization is given. An unorthodox type of error analysis is applied by considering lost digits in cancellation. The resulting proof is simple and that makes it possible to calculate the number of accurate digits after all reorthogonalization steps. Self improving nature of projection matrices is found giving a possible explanation for the stability of some ABS methods. The numerical tests demonstrate the validity and applicability of the theoretical results for the CGS, MGS and rank revealing QR algorithms.

Keywords: twice is enough, Gram-Schmidt with reorthogonalization, self-improving nature of projections, ABS methods, rank-revealing QR.

1 Introduction

A new theoretical background of Parlett-Kahan's "twice is enough" algorithm for computing accurate vectors in Gram-Schmidt orthogonalization is given in this paper. To this aim Rutishauser's control parameter [20] – here called η – is used to decide if

i) some digits are lost, or

ii) the new vector to be processed is linearly dependent of the current base numerically, that is, up to machine precision.

Originally, the "twice is enough" algorithm was given for a one-vector projection, however, it works also for parallel multi-vector projections as in classical Gram-Schmidt (CGS). A useful by-product of our analysis is that an estimate for the number of accurate digits can be given in the course of the computation. That can be especially helpful when one has to decide linear dependence e.g. in pseudoinverse calculations.

When orthogonalizing numerically, one may have to face the problem, that the resulting vectors are not orthogonal to each other up to machine precision. The reason

can be attributed to rounding errors, however, cancellation errors are behind the phenomenon. In fact, the process of orthogonalizing two vectors is subject to cancellation errors if the vectors have nearly the same length and direction, in other words, their difference is small.

Wilkinson in his books [23], [22] already considered the problem of losing orthogonality and he identified the main cause as the presence of cancellation. In [22], pp. 382-387, he considered reorthogonalization in conjunction with the Arnoldi process. His numerical example showed that one reorthogonalization step was enough to get orthogonality up to machine precision.

Rice [19] and Hoffmann [17] did extensive numerical experimentations to find, how many reorthogonalization steps are needed. Hoffmann formulated the conjecture that one reorthogonalization step is enough for both – classical (CGS) and modified (MGS) Gram-Schmidt algorithms. On the other hand, Rice found that sometimes multiple reorthogonalizations were needed. For early theoretical investigations, see Daniel *at al.* [14] and Abdelmalek [2].

Parlett and Kahan [18] considered orthogonalization to one vector and gave their "twice is enough" algorithm. Having supposed that the starting vectors were accurate, they supplied an error analysis showing that two orthogonalization steps are practically enough to get a new accurate orthogonal vector.

The Parlett-Kahan (PK) algorithm is based on the following orthogonalization step. Let z be the vector to be orthogonalized to y. Then let

$$p = \left(I - \frac{yy^T}{\|y\|^2} \right) z = \mathrm{orth}(y,z) \tag{1}$$

denote the exact orthogonalization of z, where the 2-norm or Euclidean norm is used from now on. In reality, we have only a numerical approximation to p, say x'. Let the error $e' \equiv x' - p$ satisfy $\left\| e' \right\| = \varepsilon \|z\|$, where ε is a small number, practically close to the machine precision unit ε_M and let κ be any fixed value in the range $[1/(0.83-\varepsilon), 0.83/\varepsilon]$ then the "twice is enough" algorithm of Parlett and Kahan is given by

The PK algorithm

Calculate $x' = \mathrm{orth}(y,z)$, where orth is given in (1).

Case 1: If $\left\| x' \right\| \geq \|z\| / \kappa$ accept $x = x'$ and $e = e'$.otherwise compute

$x'' = \mathrm{orth}(y,x')$

with error

$$e'' \equiv x'' - \left(I - \frac{y * y^T}{\|y\|^2} \right) x'$$

satisfying $\left\| e'' \right\| = \varepsilon \left\| x' \right\|$ and go to Case 2

Case 2: If $\left\|x''\right\| \geq \left\|x'\right\| / \kappa$ accept $x = x''$ and $e = e" - p$.

Case 3: If $\left\|x''\right\| < \left\|x'\right\| / \kappa$ accept $x = 0$ and $e = -p$.

Theorem 1. *The vector x computed by the algorithm ensures that* $\|e\| \leq (1 + 1/\kappa)\varepsilon \|z\|$ *and* $\left|y^T x\right| \leq \kappa \varepsilon_M \|y\| \|x\|$.

For proof, see [18].

Remark 1. *Observe that if* x' *is machine zero then Case 2 will accept a zero vector. The equality sign should be moved to Case 3.*

One-vector projections are used in the MGS algorithm [4], [5], hence orthogonalizing twice solves the accuracy problem for MGS that is a sequential algorithm.

For the well parallelizing CGS the question if the "twice is enough" algorithm works well, was answered positively by Giraud et al [9], [10]. It is still worth mentioning that for computing the reduced norm of the orthogonalized vector, Smoktunowicz *et al* [21] suggest to compute $\sqrt{c^2 - a^2}$ by replacing the terms under the root sign with $(c-a)(c+a)$. They also supply an error analysis for justification. We shall compare this method with the standard computation and also, with another method by using trigonometric functions.

For a recent application of reorthogonalization in the Golub-Kahan-Lanczos bidiagonalization, see the paper by Barlow, [3].

The schedule of this paper is the following: We present our considerations in the next Section: conditions for reorthogonalization and a new short general proof.

The other sections are concerned with the comparison and testing of the new reorthogonalization algorithms.

It is also assumed that rounding errors and cancellation errors are such that there are some accurate digits in the computation.

2 Conditions for reorthogonalization

The "twice is enough" algorithm will be reformulated here from the point of view of cancellations. The theorem is stated for orthogonalizing with respect to a subspace in one step, such that the generalization given by [10] is also covered. The improvement of orthogonality is stated and we give a new short proof. Our analysis assumes that there are some accurate figures in the computation. The section is ended by accurate digits estimation and numerical experimentation.

2.1 The cancellation phenomenon

Cancellation happens if two numbers are nearly the same and they are subtracted from each other. For example, assume a 6-digit decimal arithmetic and compute: $126.426 - 126.411 = 0.015$. It is seen, the first four digits are lost, and the result,

if normalized, has the form: $0.150000 \cdot 10^{-1}$.Now the question is, how can we interpret the accuracy of the result. If there were 10 digits and the further 4 digits – which are not seen here – are the same, then the result is accurate to 6 decimals. If the missing four digits were not the same, then we have accuracy only for two figures. As seen, the number of accurate digits may range here from 2 to 6.

Wilkinson in his book [23] adopts the optimistic picture that accuracy is not lost, and that makes it possible to introduce the error postulate for floating point operations:

$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\varepsilon), \quad |\varepsilon| \leq \varepsilon_M,$$

where $\circ$ is any of the four arithmetic operations and ε_M is the machine precision unit. Higham [16] (Sec. 1.7) gives an example of computing $(1-\cos x)/x^2$ for $x = 1.2 \cdot 10^{-5}$ when there are 10 significant figures. The result is clearly in error and another formula is suggested to avoid subtraction. But such tricks are not always applicable.

Considering the relative precision, he states that "subtractive cancellation brings earlier errors into prominence". Without the postulate above, the error analysis of numerical algorithms can not be done or it can be overwhelmingly difficult. As a rule of thumb, the postulate is accepted and programmers are advised to avoid cancellation as much as possible.

In the following we shall consider cancellation as is and we shall be looking for the number of accurate figures.

Let the scalars α, β be nonzero and nearly the same. When subtracting, the cancellation can be characterized by the ratio

$$\eta = \frac{|\alpha - \beta|}{\max(|\alpha|, |\beta|)}. \tag{2}$$

If $\eta > 0.5$ we may say that there is no cancellation of binary digits, while in the case of $\eta < 10^{-\rho}$ – where ρ is the number of accurate digits – we say that the two numbers are the same to computational accuracy. Although 15 decimal digits are assumed in double precision computation, we should take into account that usually the last 2-3 digits are uncertain due to rounding errors. Therefore a practical choice for ρ is $\rho = 12$. We may loose digits by cancellation if the condition

$$10^{-\rho} \leq \eta < \eta_{\max} \tag{3}$$

holds, where $\eta_{\max} = 1/2$ may be chosen. The *worst case is assumed always*, therefore the number of lost decimals is estimated by $-\log_{10}\eta$. This value is 4.06... in the above example.

As a consequence, the number of accurate digits after subtraction is

$$\gamma = \rho + \log_{10}\eta \tag{4}$$

and the error of the difference $|\alpha - \beta|$ is $10^{-\gamma}|\alpha - \beta|$. Similarly, the error of η can be given by $10^{-\gamma}\eta$.

We shall see in the sequel that ρ – the number of accurate digits without cancellation – can be estimated after a reorthogonalization step.

2.2 Cancellation in Gram-Schmidt orthogonalization

Here we consider one step of Gram-Schmidt orthogonalization.

Introduce $Q = (q_1, q_2, \ldots, q_{k-1}) \in \Re^{n\times(k-1)}$ and $a \in \Re^n$ be known and accurate. Vector a is orthogonalized to the subspace spanned by the orthonormal columns of matrix Q in one Gram-Schmidt step

$$\theta_k q_k = (I - QQ^T)a, \tag{5}$$

where q_j-s are normalized that is, $\theta_k = \left\|(I - QQ^T)a\right\|_2$ holds and the subscript for the 2-norm will be omitted in the sequel.

Comparing the subtraction here with the case of cancellation from the previous subsection, θ_k of (5) refers to $|\alpha - \beta|$ and we identify $\max(|\alpha|, |\beta|)$ as the norm of $\|a\|$ because we may expect $\left\|Q^T a\right\|$ not larger than $\|a\|$. Hence we are led to the formula of

$$\eta = \frac{\theta_k}{\|a\|}, \tag{6}$$

a computable value for which (3) can be checked.

If $\eta \geq \eta_{\max}$ then q_k is accepted, else if $\eta < 10^{-\rho}$ the vectors $a, q_1, q_2, \ldots, q_{k-1}$ are considered linearly dependent – at least computationally – such that another vector a should be chosen.

Otherwise, if (3) is fulfilled then redo orthogonalization for q_k:

$$\widehat{\theta}_k \widehat{q}_k = (I - QQ^T)q_k. \tag{7}$$

The next theorem states that at most two orthogonalization steps are enough to get a new orthogonal vector to computational accuracy. The phenomenon was already observed by Wilkinson [22] and later formulated as a conjecture by Hoffmann [17]. Parlett in his book [18], with a reference to Kahan gave a proof for $k = 2$, (orthogonalization to one vector). Later Giraud *et al* [9], [10] gave proof for any k. We show here that the proof is much simpler using the above picture.

Theorem 2. *If there are accurate digits in the computation, then one may expect the fulfillment of condition $\eta_{\max} \leq \eta$ after the second orthogonalization step at most. The largest choice of such $\eta_{\max}$ is $1/\sqrt{2}$ to fulfill the condition. Hence the resulting vector $\widehat{q}_k$ can be considered orthogonal to $q_1, q_2, \ldots, q_{k-1}$ up to computational accuracy if $\eta_{\max}$ is not less than* 0.5.

Proof. Before giving the proof, recall that poor orthogonality after the first step is attributed to cancellation. The second orthogonalization step – if needed – may be interpreted as orthogonalizing the emerging error vector with respect to the columns of Q. Taking the square of the norm in (5), we get

$$\theta_k^2 = a^T(I - QQ^T)a = \|a\|^2(1 - \left\|Q^T\widetilde{a}\right\|^2), \tag{8}$$

where the normed vector $\widetilde{a} = a/\|a\|$ is used. Denote the angle between $\mathscr{R}(Q)$ (range of Q) and a by $\measuredangle(\mathscr{R}(Q), a)$, then we get the formula

$$\eta = \sin \measuredangle(\mathscr{R}(Q), a) \tag{9}$$

that can be obtained by considering the rectangular triangle with hypotenuse $\|\widetilde{a}\| = 1$, and legs $\left\|Q^T\widetilde{a}\right\|$ and $\eta = \sqrt{1 - \|Q^T\widetilde{a}\|^2}$, this latter is the distance of $\widetilde{a}$ from $\mathscr{R}(Q)$, see Figure 1. For more detailed informations on subspace angles, see [8]. A short proof for the smallest angle between a vector and a subspace can be found in [17].

Now assume $\eta \in [10^{-\rho}, \eta_{\max})$ holds such that reorthogonalization is needed. Then we have to show that after reorthogonalization $\eta_{\max} \leq \eta_r$ will succeed for the new η_r. Indeed, by replacing a with q_k in (5), we get for (9)

$$\eta_r = \sin \measuredangle(\mathscr{R}(Q), q_k). \tag{10}$$

This angle is $\pi/2$ accurately the sine of which is 1. Now it is simpler to estimate $\cos \measuredangle(\mathscr{R}(Q), q_k)$ instead of (10), where the computation is subject to errors. The cosine rule will be used for the almost rectangular triangle and the result is

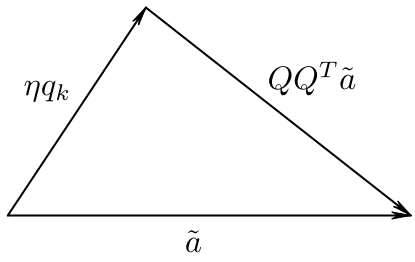

Figure 1
The projection triangle

$$\cos \measuredangle(\mathscr{R}(Q), q_k) = \frac{\eta^2 + \left\|Q^T\widetilde{a}\right\|^2 - 1}{2\eta \left\|Q^T\widetilde{a}\right\|}. \tag{11}$$

If there were no errors in the calculation, then the numerator would be zero as it can be checked from (8). Actually we have

$$\cos \measuredangle(\mathscr{R}(Q), q_k) = \frac{\eta^2 - \eta^2}{2\eta \left\|Q^T\widetilde{a}\right\|} = \frac{\eta - \eta}{2 \left\|Q^T\widetilde{a}\right\|}, \tag{12}$$

where numerator and denominator were divided by η. It was shown earlier that the error of η is $10^{-\gamma}\eta$ and we can assume that $\widetilde{a}$ is nearly the same as $QQ^T\widetilde{a}$ in case of cancellation. Therefore $\left\|Q^T\widetilde{a}\right\|$ is near to 1. We approximate the error of $\cos \measuredangle(\mathscr{R}(Q), q_k)$ by taking twice the error of η in the numerator and replace $\left\|Q^T\widetilde{a}\right\|$ by 1 in the denominator:

$$\cos \measuredangle(\mathscr{R}(Q), q_k) \approx \frac{2\eta 10^{-\gamma}}{2} \leq 10^{-\gamma}\eta_{\max} \leq \eta_{\max}, \tag{13}$$

as the largest possible value of η is $\eta_{\max}$ and the smallest value of γ is 0 (all accurate digits are lost). We are looking for an $\eta_{\max}$ for which the second inequality of

$$\sin \measuredangle(\mathscr{R}(Q), q_2) \gtrapprox \sqrt{1-\eta_{\max}^2} > \eta_{\max}$$

also holds. We have equality on the right if $\eta_{\max} = 1/\sqrt{2} \approx 0.707$. It is easily seen that if $10^{-\gamma}\eta$ with a positive γ is applied under the square root instead of $\eta_{\max}$ then the inequality on the right is fulfilled even better. That $\eta_{\max}$ should not be chosen below 0.5 was discussed in the first subsection here. □

Compare it with $1/\kappa$ that corresponds to our $\eta_{\max}$ in the PK algorithm. There the possible largest choice is $1/\kappa = 0.83 - \varepsilon$. That is near to the here found $\eta_{\max} = 0.707$. But $\kappa = 100$ is also suggested for less computational works. In that case one agrees to loose roughly two decimal digits of precision and computation to machine accuracy is abandoned. By choosing $10^{-k}/\sqrt{2} = \eta_{\max}$, one allows loosing k decimal digits.

On the other hand, the smallest possible choice in the PK algorithm for $1/\kappa$ is $\varepsilon_M/0.83$. It seems too small with respect to our criterion of acceptance.

2.3 Estimating the accuracy of computation

If we repeat orthogonalization then the new η_r can give a method to estimate ρ, the number of accurate digits.

For exact computation $\eta_r = 1$ should hold. We adopt the picture that when reorthogonalization is done, the error vector caused by cancellation is orthogonalized at the second step. The norm of the error vector of q_k can be estimated by $10^{-\gamma}\eta$ after the first step. We have in the second step:

$$\eta_r^2 = 1 - \left\|Q^T q_k\right\|^2 = 1 - \left(10^{-\gamma}\eta\right)^2$$

because the accurate part of q_k gives zero contribution. Consequently, see also (4)

$$\log_{10}\left\|Q^T q_k\right\| = -\gamma + \log_{10}\eta = -\rho - \log_{10}\eta + \log_{10}\eta$$

that is,

$$\rho = -\log_{10}\left\|Q^T q_k\right\|. \tag{14}$$

Observe that ρ depends on the step number k, therefore it should be calculated step by step such that

$$\rho_k = -\log_{10}\left\|Q^T q_k\right\|. \tag{15}$$

Comparing with the PK algorithm, there $\eta_r < \eta_{max} = 1/\kappa$ is used for stating zero for the projected vector. Now assume $\rho = 0.4$. The interpretation is that even half decimal digit accuracy can not be assumed after the first projection and that indicates

serious cancellation. Then $\eta_r \approx 0.92$ holds and using $\eta_{max} = 0.707$, the condition in Case 3 is not fulfilled. For this η_{max} the equivalent condition for Case 3 is:

$$\rho \leq 0.1505, \tag{16}$$

where the inclusion of equality was suggested in Remark 1. For smaller values of η_{max}, the upper bound here will be slightly diminishing, but it always remains positive. As seen, the PK algorithms allows the loss of almost all decimal digits for identifying a numerically zero vector.

If rounding errors are not negligible then observe that cancellation makes sense only if it is larger than rounding errors. An estimate for rounding error of a scalar product can be found in [12] :

$$\left|\delta(y^T x)\right| \leq 1.01 n \varepsilon_M \|y\| \|x\|, \tag{17}$$

where n is the length of vectors. For a rounding error analysis of the Gram-Schmidt process, see [2], [4], [5] and [11]. It is seen that for very large n the rounding errors may be so big that there are only few accurate digits, or in pathological cases, $\gamma < 0$ characterizes the situation.

Fig. 2 shows a picture to illustrate the behaviour of ρ. Vector a is orthogonalized to p with optional reorthogonalizing, where the distance of a and p is varying as 10^{-k}, such that k is between 1 and 14. In fact, $a - p$ was chosen to be perpendicular to p. It is seen that orthogonality holds for 16 figures in all cases and the number of accurate digits are diminishing as the two vectors are getting closer. Using a double precision arithmetic, normally one expects that the values in (15) are around 14-15. Smaller values may be considered as indicator for the events of serious cancellation.

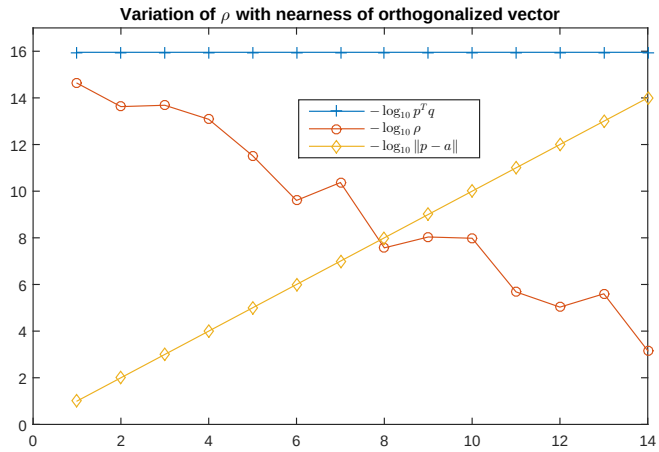

Figure 2
Orthogonalizing nearby vectors

2.4 Numerical experiments

Fig. 3 shows variation of precision in the function of $\eta_{\max}$. The matrix is an 80×80 random matrix having entries from $(-1,1)$ and the computation has been done under Matlab R2014b. The curve with + signs shows QR how well approximates matrix A. The values

$$-\log_{10} \frac{\max \left| (A - QR)_{ij} \right|}{\max \left| A_{ij} \right|} \tag{18}$$

are shown in the function of $\eta_{\max}$. Similarly, the values

$$-\log_{10} \max \left| \left(I - Q^T Q \right)_{ij} \right| \tag{19}$$

show the number of accurate digits for Q in the worst case. It is seen, we are below machine accuracy for $\eta_{\max} \leq 0.4$. But there is an improvement to machine accuracy when $\eta_{\max}$ reaches the value 0.6 which is in good agreement with the theory. That QR serves a good approximate to A even if the orthogonal system is less accurate was already stated in [5].

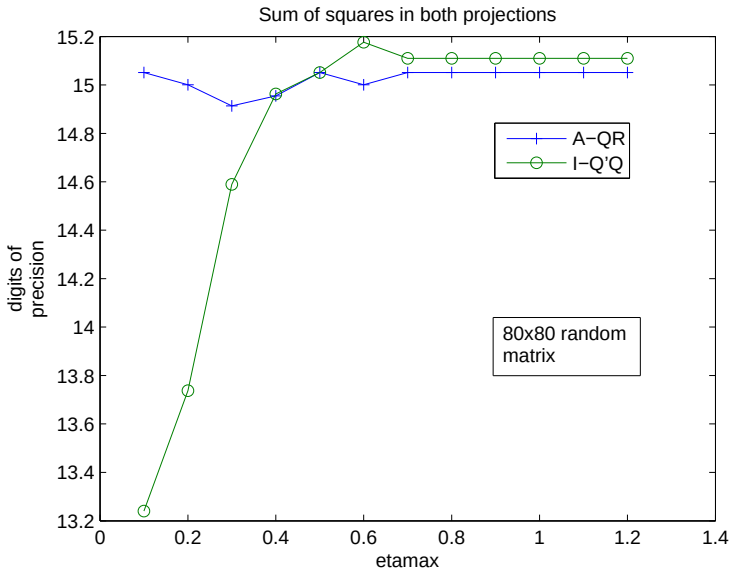

Figure 3
Sum of squares are computed in both cases

For computing θ_k of (5), we have some possibilities.

1) *Sum of squares.*

Here vector $\theta_k q_k$ is computed by (5) and the norm of the vector is taken by computing sum of squares as in 2-norm calculations. The first and second orthogonalization was cumputed by this approach in Fig. 3.

2) *Difference of squares.*

This way of computation uses

$$\theta_k = \sqrt{\|a\|^2 - \|Q^T a\|^2} = \|a\|^2 \sqrt{1 - \|Q^T \widetilde{a}\|^2}$$

analysed by Smoktunowicz *et al* [21], where the product form is taken for the difference of squares. The results of this second approach can be seen in Fig. 4 for the same matrix, where computation was done with difference of squares method in the first and second orthogonalization. Quite astonishingly the accuracy is poorer for small values of $\sigma = \left\|Q^T \widetilde{a}\right\|$, – that is the case for reorthogonalizations – and it occurs more frequently with increasing $\eta_{\max}$. The difference of squares method is not always better, in fact, more and more digits of σ are lost in the computation of $(1-\sigma)(1+\sigma)$ because of the relatively large value of 1. To check this statement,

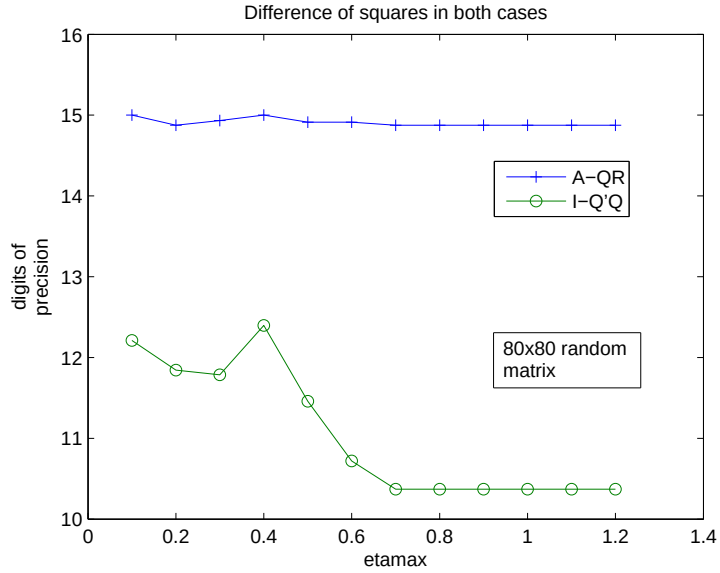

Figure 4
Difference of squares are computed in both cases

we show the results in Fig. 5, where the first approach is applied in the case of reorthogonalization.

3) *Trigonometric functions.*

A third approach is the use of trigonometric functions. We can use the formulas

$$\beta = arc\cos(\left\|Q^T z\right\|), \quad \eta = \sin\beta,$$

where $z = \widetilde{a}$ in the first step and $z = q_k$ in the second step. But then we get to a similar picture that could be seen in Fig. 4. And changing back to the first approach in the second step will result in the same situation that is shown in Fig. 5.

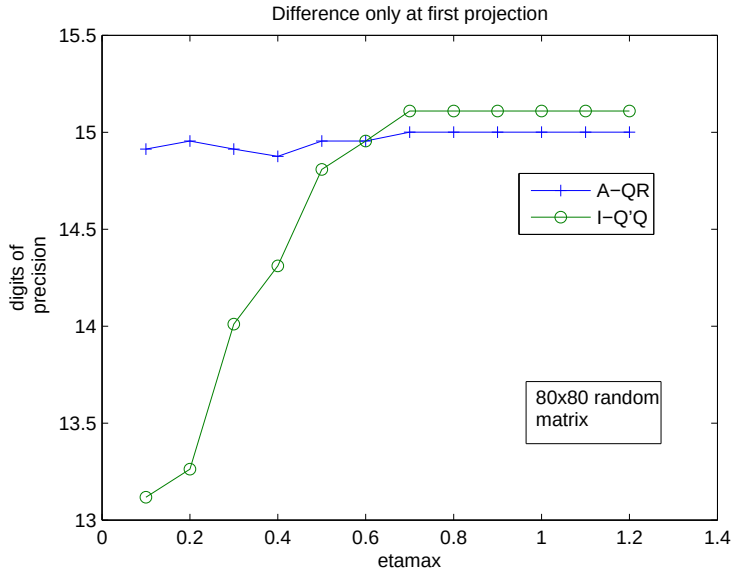

Figure 5
Difference at first, sum of squares in second case

The numerical experimentations have led us to the statement: numerical accuracy will be better using the first approach – compute projected vector and take norm – if $Q^T z$ is very small. This time, when looking into numbers, an additional surprise is that the numerical values of η are very close to 1 such that they may be even larger than 1 – a situation that contradicts to being a value of the sine function. It stresses the belief that rounding errors govern the situation here. Observe that the second and third approach force $\eta \leq 1$, therefore they can not handle the numerical case of $1 < \eta$ so well because division by η is needed for normalization.

2.5 Updating QR-decomposition in reorthogonalization

It still deserves some words how updating of matrix R can be done after the second orthogonalization step. At the firt step the kth column was given by

$$\left[\begin{array}{cc} a^T Q & \theta_k \end{array} \right]^T.$$

In the reorthogonalization step, one applies the projection to $\theta_k q_k$ giving

$$\theta_k(q_k - QQ^T q_k) = \theta_k \eta_r \widehat{q}_k,$$

where η_r is the norm of $(q_k - QQ^T q_k)$ such that the resulting vector $\widehat{q}_k$ is normed to 1. Now it is seen that the updated kth column of R is given by

$$\left[\begin{array}{cc} a^T Q + q_k^T Q & \theta_k \eta_r \end{array} \right]^T. \tag{20}$$

It was observed in numerical experiments that updating the nondiagonal column elements in matrix R ruins the quality of QR if there is a large loss of precision after the first orthogonalization step. Because of that updating was allowed only for large enough ρ of 15 in our rank finding program.

2.6 Orthogonal base algorithms

Minor modifications in the PK algorithm were suggested:

- Choose $\kappa = \sqrt{2} \cdot 10^i$ if the loss of i decimal digits are allowed.
- Move equality sign from Case 2 into Case 3.

We are in agreement with other authors that projection into an arbitrary subspace is also allowed.

Another variant of orthogonal base algorithm (OBA) can also be given that reflects the view of this paper. Now all three approaches may be applied for norm calculation in the first phase but for reorthogonalization only the first approach is suggested in accordance with the *Numerical experimentation* Subsection. Chose for ε a nearby value to ε_M and $\eta_{\max} = 0.707$. Assume that $k-1$ orthogonal vectors are ready, then the kth step can be given by

Algorithm 2. *One step of OBA*

Orthogonalize a_k to the first $k-1$ columns of Q by (5)
Compute θ_k by (5) and then η by (6)

If $\eta < \varepsilon$ **then** act for a linearly dependent vector
    else
        Compute the kth columns of Q and R
        if $\eta \leq \eta_{\max}$
            Perform reorthogonalization by (7)
            Update the kth columns of Q and R
        end_if
    end_else
end_if

One can also lower upper bound for loosing digits as in the PK algorithm. Projections can be done as in (5) and with explicitly computed matrices.

Another variant may be to apply reorthogonalizing always. There are signs that such an algorithm may show good performance, [7]. However, one should be cautious in that case, see the remark after (20).

3 Some further applications of OBA

First we remark that reorthogonalization may be applied to improve an orthogonal projection that is subject to numerical errors. If it is given in the form of QQ^T that can be considered a Choleski decomposition of a positive semidefinite matrix,

then the steps of OBA give a straightforward procedure to refine a vector q_i in the orthogonal system.

One can also give a quality improvement if the projection is given by a matrix P. Now say, column i should be corrected. Then form the projection

$$\widehat{P} = P - \frac{Pe_i e_i^T P}{e_i^T P e_i}. \tag{21}$$

It brings Pe_i into zero: $\widehat{P}Pe_i = 0$. Its direction may be corrected by

$$z = Pe_i - \widehat{P}Pe_i \tag{22}$$

and then the improved projection can be re-gained by

$$\widehat{P} + \frac{zz^T}{z^T z}. \tag{23}$$

Observe that all nonzero columns are eigenvectors of the projection matrix with eigenvalue 1. The eigenvectors with eigenvalue zero can be found in the zero space of the matrix. Taking the powers, the eigenvectors with zero eigenvalue will improve, while the eigenvectors with eigenvalue 1 may be slightly deteriorated, if an eigenvalue is not exactly 1. But we can change to $I - P$ such that the image space and zero space are interchanged. Then by taking the powers of $I - P$ improves the image space of P.

Also, observe that methods intensively using projections such as in ABS methods [1] will consecutively improve the quality of zero space, hence they have a self-improving nature. That explains, why some ABS methods can be unusually stable even in case of pathological matrices.

3.1 Rank revealing QR algorithm with reorthogonalization

Rank revealing by QR (RRQR) decompositions were introduced by Chan [6] and later investigated by many authors. Ch. 5 of [12] gives samples of such orthogonal algorithms. See also [13] for a good account of RRQR decompositions. We do not want to dwell much on such algorithms, our aim here is to show only some applications of repeated orthogonalization.

For rank revealing one permutes the columns of A so that the column having the maximal 2-norm comes first. An easy way of the algorithm is to reorder columns in decreasing order of length at the beginning and then apply QR factorization. A more demanding variant chooses the vector of maximal column norm in the ith step of the remaining projected vectors. Specifically, denote by Q_i the matrix of i orthonormal columns, the corresponding projection matrix by $P_i = I - Q_i Q_i^T$, then choose column $P_i A e_k$ for which $\|P_i A e_k\|$ is maximal among the so far not chosen vectors.

Program GSrank was written for rank revealing. The following choices were applied: $\eta_{\max} = 0.707$ and $\varepsilon = 4\varepsilon_M$. ρ_{out} was computed by (14) and reorthogonalization was done if $\eta < \eta_{\max}$ and $\rho_{out} < -\log_{10}(2\varepsilon_M)$ were satisfied and the current

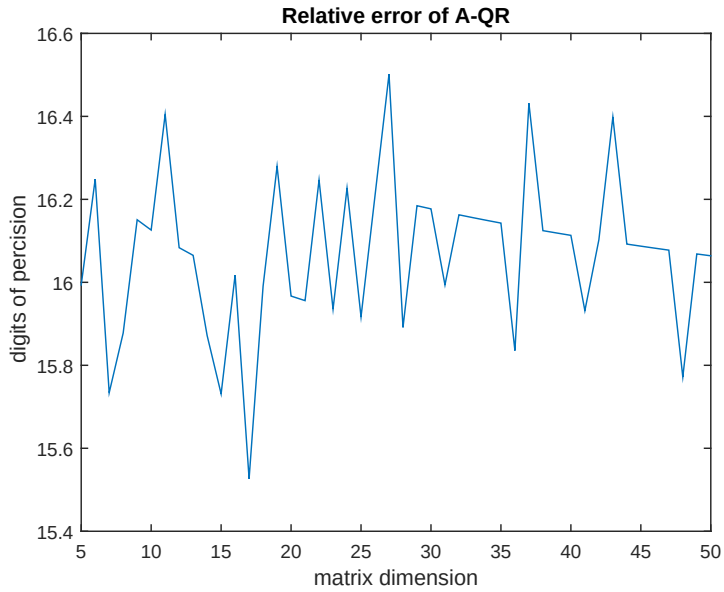

Figure 6
Errors of A - QR for Pascal matrices

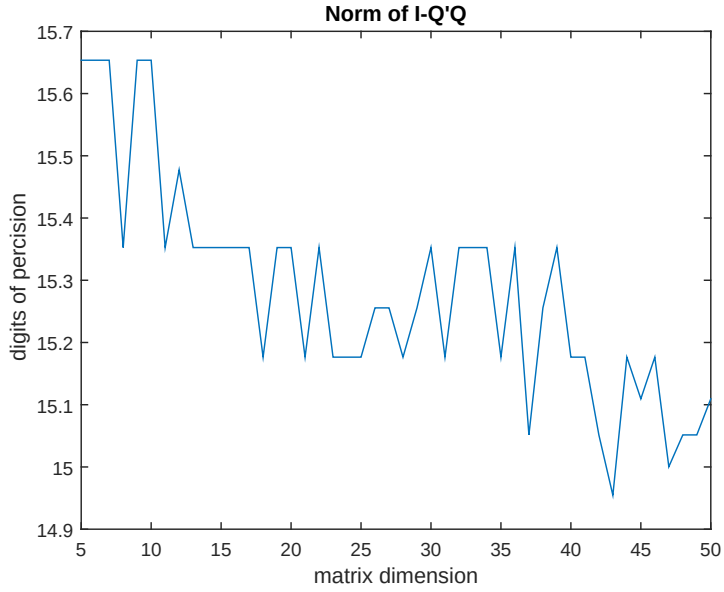

Figure 7
Goodness of orthogonal vectors for Pascal matrices

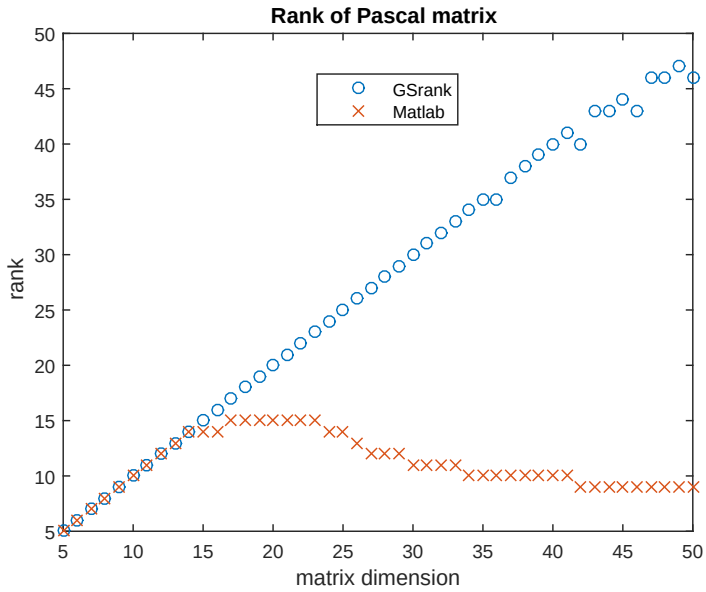

Figure 8
The found ranks of Pascal matrices

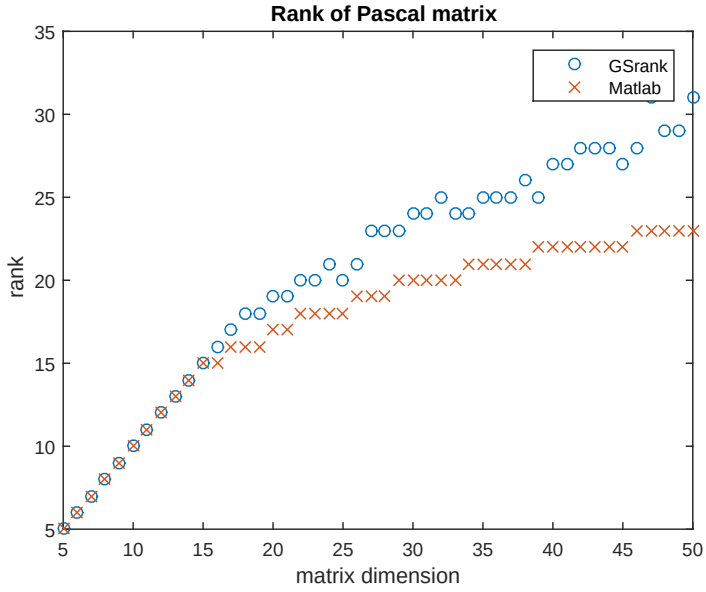

Figure 9
Rank results with normed rows

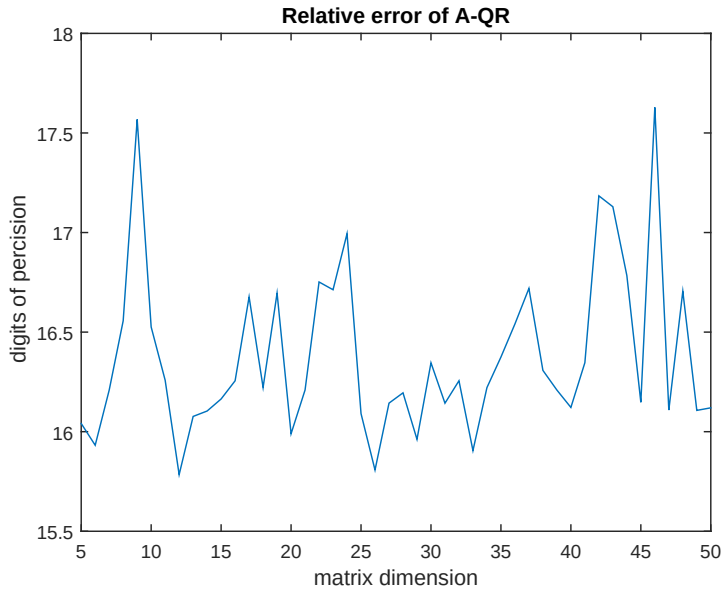

Figure 10
Vandermonde matrices

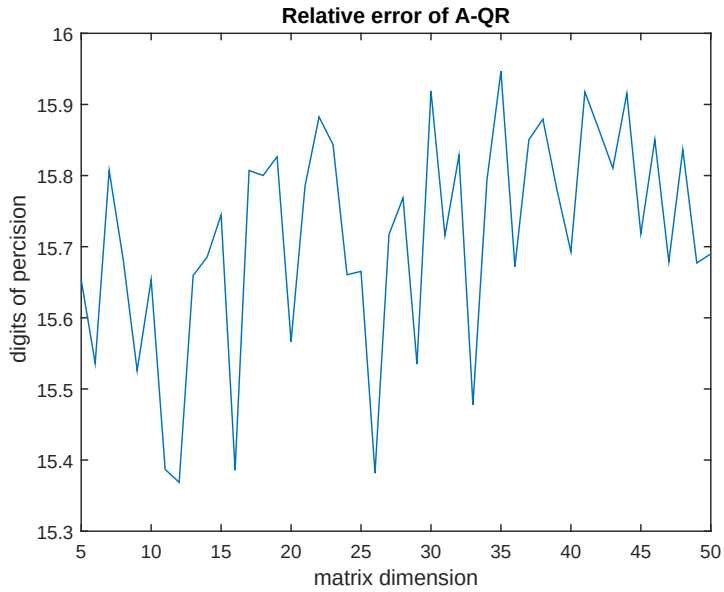

Figure 11
Corrected relative errors for Vandermonde matrices

vector was not considered linearly dependent. The update formula for the non-diagonal column elements of R was allowed only if $\rho_{out} \geq -\log_{10}(\varepsilon)$ had been satisfied. The projection P_i was calculated explicitly.

Such methods are working well for ordinary matrices. It is more interesting to show results for pathological cases. One example is the Pascal matrix that is found in Matlab's collection and can be called by the statement pascal(n). Figures 6 and 7 show the goodness of the factorization for Pascal matrices.

The attentive reader may observe in Fig. 6 that the relative precision can be as large as 16.4 decimal digits, though having a double precision arithmetic, that accuracy is impossible. Formula (18) was applied here. First one might think that it could be attributed to the chosen norm. However, a more probable explanation is the following: The absolute largest matrix element is so large that the next largest one is less by some orders of magnitude. Chances are good that the column having the largest element comes first, or it is among the firstly chosen vectors. As the explicit projection P_i is applied in all steps, then it follows that the direction of such vectors are projected out many times and finally it may happen that the error of some largest elements are machine zero. Then for a more reasonable relative error, only those largest elements should be taken, for which the error is not machine zero. Naturally, a smaller divisor applies in that case. An example for such kind of relative error computation will be shown for Vandermonde matrices.

For Fig. 7, formula (19) was applied. As seen, machine accuracy may be assumed for all Pascal matrices, the condition numbers of which are roughly proportional to 10^{n-1}. The entries in Pascal matrices are exactly representable by machine numbers up to the order of 23. It may be a question that the double precision form of higher order matrices still have rank equal to their size. Such matrices were converted and tested in quadruple precision arithhmetic. It was found that all of them have rank equal to their size [7].

The rank results can be seen in Fig. 8 as compared to those of Matlab.

In this example the 35th row was copied into the 45th row in order to test sensitivity. As seen, GSrank performs well, however, there are uncertainties in higher dimensions. Matlab's rank finder suffers if there are numbers very different in their order of magnitude. If all rows are normed to 1, then rounding errors are introduced into matrix data and that leads to another picture. Now Matlab performs better.

The other matrix tested is the Vandermonde matrix with base points $1, 2, ..., n$. The results are similar to those of the Pascal matrix.

Fig. 10 shows even "better" – but impossible – relative errors for the goodness of QR-decomposition. The remarks previously given to Fig. 6 apply here once again. According to that, a program was written for the relative error such that search for absolute maximal matrix element was done only for entries having a nonzero error. The corrected relative errors in Fig. 11 justify the supposed phenomenon. Figures 12 and 13 show the goodness of the orthogonal base and rank results for Vandermonde matrices.

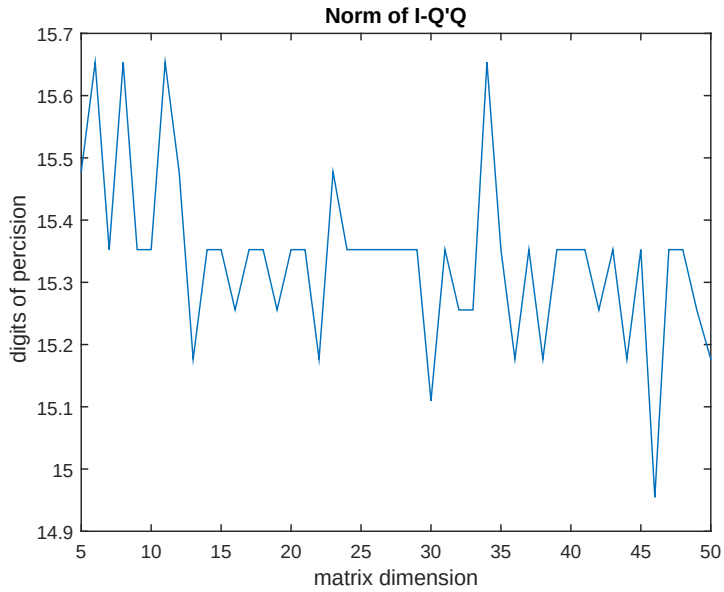

Figure 12
Quality of orthogonal systems for Vandermonde matrices

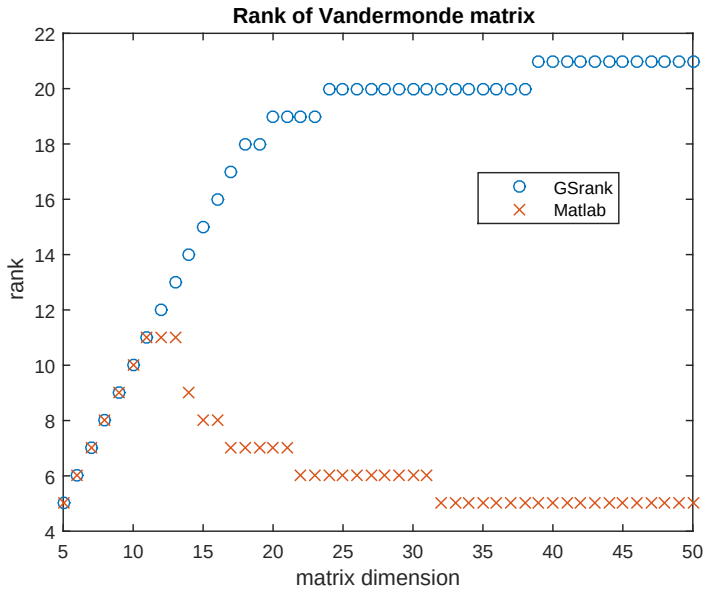

Figure 13
Found ranks of Vandermonde matrices

3.2 Programs to download

For checking and further tests, the following Matlab routines can be downloaded from: `http://numanal.inf.elte.hu/~hegedus/matlab.html`

GSrank: QR decomposition with pivoting and rank finding

Pproj: Performs one projection step, called by GSrank

relarel: Corrected residual error for matrices

lsqsol: Least squares solution for Ax=b, where A is decomposed by GSrank.

4 Conclusions

A new theoretical background and modified versions of the "twice is enough" algorithm are given. Quite surprisingly, cancellation error considerations lead to a simpler proof. The success may suggest a wider use of cancellation phenomena in error investigations. Another surprise is the possibility of estimating the number of accurate digits after the first projection with the help of second projection data (ρ_{out} from (14)). The analysis gives an explanation of the extraordinary stability of ABS methods in some cases. The test problems shown justify the given statements and also reveal some unexpected numerical phenomena. Further, it is demonstrated that orthogonalizing twice assures a good quality of rank revealing QR-decompositions.

Acknowledgement

Professor J. Abaffy is thanked for his constant encouragement and help when writing this paper. Also, Szabina Fodor is highly appreciated for helpful discussions.

References

[1] ABAFFY, J., AND SPEDICATO, E.: *ABS Projection Algorithms: Mathematical Techniques for Linear and Nonlinear Equations*, Ellis Horwood Limited, John Wiley and Sons, Chichester, England, (1989).

[2] ABDELMALEK, N. N.: *Round off error analysis for Gram-Schmidt method and solution of linear least squares problems,* BIT 11 (1971) pp. 345-368

[3] BARLOW, J. L.: *Reorthogonalization for the Golub-Kahan-Lanczos bidiagonal reduction,* Numerische Mathematik, 124 (2013) pp. 237-278

[4] BJÖRCK, A.: *Solving linear least squares problems by Gram-Schmidt orthogonalization,* BIT 7 (1967) pp. 1-21

[5] BJÖRCK, A. AND PAIGE, C.: *Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm,* SIAM J. Matrix Anal. Appl. 13(1) (1992) pp. 176-190

[6] CHAN, T. F.: *Rank revealing QR factorizations, Lin. Alg, and its Applic.* 88/89 (1987) pp. 67-82

[7] FODOR, SZ.: Private communication.

[8] Galántai, A. and Hegedüs, C. J.: *Jordan's principal angles in complex vector spaces*, Numerical Linear Algebra with Applications 13 (2006) pp. 589-598

[9] GIRAUD, L., LANGOU J. AND ROZLOZNIK, M.: *On the round-off error analysis of the Gram-Schmidt algorithm with reorthogonalization,* CERFACS Technical Report No. TR/PA/02/33 (2002) pp. 1-11

[10] GIRAUD, L., LANGOU J. AND ROZLOZNIK, M.: *The loss of orthogonality in the Gram-Schmidt orthogonalization process,* Computers and Mathematics with Applications, Vol. 51 (2005) pp. 1069-1075

[11] GIRAUD, L., LANGOU J. AND ROZLOZNIK, M. AND VAN DEN ESHOF, J.: *Rounding error analysis of the classical Gram-Schmidt orthogonalization process,* Numer. Math. 101 (2005) pp. 87-100

[12] GOLUB, G. AND VAN LOAN, C.: *Matrix Computations,* 3rd ed. John Hopkins Univ. Press, Baltimore, MD (1996)

[13] GU, MIND AND EISENSTAT, STANLEY C.: *Efficient algorithms for computing a strong rank revealing QR factorization,* SIAM J. Sci. Comput. **17**(4), (1996) pp. 848-869

[14] DANIEL, J.W, GRAGG, W.B., KAUFMAN L. AND STEWART G.W.: *Reorthogonalization and Stable Algorithms for Updating the Gram-Scmidt QR Factorization*, Mathematics of Computation 30(136) (1976) pp. 772-795

[15] HEGEDÜS, C. J.: *Short proofs for the pseudoinverse in linear algebra,* Annales Univ. Sci. Budapest, 44 (2001) pp. 115-121

[16] HIGHAM, N. J.: *Accuracy and Stabiility of Numerical Algorithms,* SIAM, Philadelphia, (1996)

[17] HOFFMANN, W.: *Iterative Algorithms for Gram_Schmidt orthogonalization*, Computing Vol 41 (1989) pp. 335-348

[18] PARLETT, B. N.: *The symmetric Eigenvalue Problem*, Englewood Cliffs, N. J. Prentice-Hall (1980)

[19] RICE, J. R.: *Experiments on Gram-Schmidt orthogonalization*, Math. Comp. 20 (1966) pp. 325-328

[20] RUTISHAUSER, H.: *Description of Algol 60,* Handbook for Automatic Computation, Vol 1a. Springer, Berlin, (1967)

[21] SMOKTUNOWICZ, A. BARLOW, J. L. AND LANGOU, J.: *A note on the error analysis of classical Gram-Schmidt,* Numer. Math. 105(2) (2006) pp. 299-313

[22] WILKINSON J. H.: *The Algebraic Eigenvalue Problem,* Oxford University Press (1965)

[23] WILKINSON J. H.: *Rounding Errors in Algebraic Processes*, Prentice-Hall (1963)