NEW FAST AND ACCURATE JACOBI SVD ALGORITHM: II.

ZLATKO DRMAČ* AND KREŠIMIR VESELIĆ†

Abstract. This paper presents new implementation of one–sided Jacobi SVD for triangular matrices and its use as the core routine in a new preconditioned Jacobi SVD algorithm, recently proposed by the authors. New pivot strategy exploits the triangular form and uses the fact that the input triangular matrix is the result of rank revealing QR factorization. If used in the preconditioned Jacobi SVD algorithm, it delivers superior performance leading to the currently fastest method for computing SVD decomposition with high relative accuracy. Furthermore, the efficiency of the new algorithm is comparable to the less accurate bidiagonalization based methods. The paper also discusses underflow issues in floating point implementation, and shows how to use perturbation theory to fix the imperfectness of machine arithmetic on some systems.

Key words. Jacobi method, singular value decomposition, eigenvalues

AMS subject classifications. 15A09, 15A12, 15A18, 15A23, 65F15, 65F22, 65F35

1. Introduction.

Jacobi iteration is one of the time-honored methods to compute the spectral decomposition $H = V\Lambda V^T$ of a real symmetric matrix H. The early discovery in 1846. is certainly due to the simplicity and the elegance of the method as well as to the geniality of the 19'th century computing matador C. G. J. Jacobi who called it 'Ein leichtes Verfahren' and applied it to compute the secular perturbations of the planets. Jacobi's original article [18] is a very masterpiece of applied mathematics and may even today be read with profit by both students and scientists. The simplicity of the Jacobi method is not only theoretical (it may even be used to prove the existence of the spectral decomposition) but also computational, in this aspect it may well be compared with the Gaussian elimination. Thus, upon coming of automatic computation Jacobi method was soon rediscovered by Goldstine, Murray and von Neumann [14] who provided first detailed implementation and error analysis.

In our recent work [13] we reviewed Hestenes variant [17] of the Jacobi method for SVD computation of general matrix A by implicit diagonalization of $H = A^T A$. Briefly, if $V^T H V = \Lambda$, then $AV = U\Sigma$, where $\Sigma = \sqrt{\Lambda}$ is diagonal and U is orthogonal. Orthogonal matrix V is the limit product of Jacobi rotations. In [13] we have presented a novel approach to the Jacobi SVD computation with the goal of making the method more efficient, while retaining superior numerical properties. We have shown that rank revealing QR factorization can serve as versatile preconditioner which enables efficient execution of the iterative part (Jacobi rotations) on a structured triangular matrix.

The idea of QR iterations as preconditioner for SVD computation is well known, but it is not fully exploited in the context of Jacobi method. It is both simple and powerful: If $AP = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$, and $R^T P_1 = Q_1 R_1$ are rank revealing QR factorizations of A and R^T respectively, then the Hestenes one–sided Jacobi algorithm applied to $X = R^T$ or $X = R_1^T$ converges much faster than applied to A. This iterative part on triangular matrix is in [13] used as a black–box procedure: starting with $X^{(0)} =$

[*] Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia. The work of the author is supported by the Croatian Ministry of Science and Technology under grant 0037120 (*Numerical Analysis and Matrix Theory*), and by the Volkswagen–Stiftung grant *Designing Highly Accurate Algorithms for Eigenvalue and Singular Value Decompositions.*

[†] Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Postfach 940, D–58084 Hagen, Germany. The work of the author is supported by the Volkswagen–Stiftung grant *Designing Highly Accurate Algorithms for Eigenvalue and Singular Value Decompositions.*

X, the sequence $X^{(k+1)} = X^{(k)}V^{(k)}$ converges to $X_\infty = U\Sigma$ and the product of Jacobi rotations $V^{(0)}V^{(1)}\cdots$ converges to V. The SVD of X is $X = U\Sigma V^T$, where the matrix V is obtained not from the accumulated product of Jacobi rotations but rather a posteriori, based on the relation $V = X^{-1}X_\infty$. Assembling the SVD of A is straightforward.

In this report we unwrap the black–box and show how it performs in the framework of the new preconditioned Jacobi SVD algorithm [13]. In that sense, this report is a continuation of [13], and it is organized as follows.

In §2 we describe a new pivot strategy for the Hestenes one–sided Jacobi SVD on triangular matrices. It is based on the state of the art knowledge of the asymptotic convergence of Jacobi iterations. We use triangular and scaled diagonal dominance structures to reduce flop count and memory traffic, as well as to achieve faster convergence. It should be stress here that the global structure of the algorithm in [13] is such that we are completely free to choose pivot strategy (for both serial and parallel computing, open for blocking to enhance efficient use of cache memory) and have all numerical properties described in [13]. In §3 we explore certain functions of column and row norms of A, which can be used to guess the off-diagonal structure of A^TA and AA^T. Such functions can be useful for advanced control of the flow of the algorithm. The results of detailed numerical testing of mathematical software implementing the new algorithm, as well as modifications prompted by the tests are presented in §4. The material of §4.2.1, 4.2.2 is an example of that tricky part of scientific computing where we have to deal with the peculiarities of the floating point arithmetic. One has to use many little tricks and facts that perhaps are not worth to be wrapped up in forms of propositions, not to mention theorems, but all of them together make a big difference. Concluding remarks and comments on the future development of SVD computations are given in §5.

The authors acknowledge generous support by the Volkswagen Science Foundation and the Croatian Ministry of Science and Technology. We are also indebted to P. Arbenz (Zürich), J. Barlow (State College), J. Demmel (Berkeley), F. Dopico (Madrid), V. Hari (Zagreb), W. Kahan (Berkeley), J. Moro (Madrid), B. Parlett (Berkeley), I. Slapničar (Split) for their comments, criticisms and many fruitful discussions.

2. One–sided Jacobi on preconditioned triangular matrices. Since in our algorithm Jacobi iterations start with preconditioned matrix with special structure (triangular with certain diagonal dominance properties), pivot strategy should not blindly control the order of rotations following hardwired rule. Instead, pivoting should be highly adaptive, with optimal use of current structure of the matrix.

We have several structural adaptations in mind. First, we will use triangular form and avoid trivial computations with zero blocks as long as possible. This will save unnecessary computation before eventual fill in. Secondly, proper pivoting could lead to faster convergence, given the fact that the matrix has been preconditioned. This means that the matrix nearly satisfies certain asymptotic assumptions for higher order convergence (quadratic or even cubic) of some pivot strategies, and that the positions of the largest off–diagonal entries are predictable, see [13]. The question is how to choose pivot strategy to exploit this structure. Next, since one dot product is necessary to determine if pivot columns should be rotated, dot products not followed by rotation are the overwhelming cost of the later stages as the algorithm approaches numerical convergence. Proper pivot strategy should monitor and use the structure of the matrix to predict (and thus avoid) many unnecessary dot products.

Complete description and full explanation of our new pivot strategy is rather technical, and we will not present it here with all details. Instead, we will just sketch the main ideas and discuss relevant issues.

2.1. Review of asymptotic convergence. For the reader's convenience, we first offer a brief review of the main facts about the asymptotic convergence of the Jacobi method. Understanding asymptotic behavior of the Jacobi process helps in devising efficient stoping criteria and it also indicates which mechanisms create higher order convergence and thus helps in constructing better pivot strategies.

For simplicity, we state all results in equivalent form of diagonalization of symmetric matrices. Our symmetric matrix H will always be given implicitly as $H = X^T X$, and the Jacobi rotations are applied to X from the right–hand side in usual way. As basic strategy we take the row–cyclic one, which is periodic and in one full sweep of $n(n-1)/2$ rotations it rotates row–by–row at the pivot positions $(1,2), (1,3), \ldots (1,n)$; $(2,3), \ldots, (2,n)$; $(3,4), \ldots, (3,n); \ldots, (n-2,n)$; $(n-1,n)$. There is a particular reason why we place the row–cyclic pivoting as the basic strategy for future development – the transformation pattern mimics the one of the modified Gram–Schmidt (MGS) orthogonalization, which is finite (one sweep) algorithm. In fact, in the case of extremely graded matrices, the one–sided Jacobi SVD behaves almost as the MGS.[1]

2.1.1. Quadratic convergence. The symmetric Jacobi algorithm with row–cyclic strategy is quadratically convergent. This is a well known fact, and the proof of the general case of multiple eigenvalues is given by Hari [15]. Using the off-norm $\Omega(\cdot) = \| \cdot - \mathrm{diag}(\cdot)\|_F$, the quadratic convergence is stated as follows:

Theorem 2.1. *Let $\{H^{(k)} \,:\, k = 0, 1, 2, \ldots\}$ be the sequence of matrices generated by the symmetric Jacobi algorithm with row–cyclic strategy and with the initial matrix $H = H^{(0)}$ with diagonal entries ordered from large to small. Let the eigenvalues of H be*[2] *$\lambda_1 \geq \cdots \geq \lambda_n$, and let the absolute gap in the spectrum be $\mathbf{d} = \min\limits_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j|$. Then $\Omega(H^{(0)}) \leq \dfrac{\mathbf{d}}{3} \Longrightarrow \Omega(H^{(n(n-1)/2)}) \leq \dfrac{9}{5} \dfrac{\Omega(H^{(0)})^2}{\mathbf{d}}$.*

For the high relative accuracy, we need small off–diagonal part of scaled matrices $(H_s^{(k)})_{ij} = (H^{(k)})_{ij} / \sqrt{(H^{(k)})_{ii} (H^{(k)})_{jj}}$. The following theorem from [21] describes the asymptotic behavior of scaled matrices in the Jacobi algorithm.

Theorem 2.2. *Let $\{H^{(k)} \,:\, k = 0, 1, 2, \ldots\}$ be the sequence of matrices generated by the symmetric Jacobi algorithm with row–cyclic strategy and with the initial positive definite matrix $H = H^{(0)}$. Let the corresponding scaled matrices be $H_s^{(k)}$, $k \geq 0$. If H has only simple eigenvalues $\lambda_1 > \cdots > \lambda_n > 0$ and if $h_{11} \geq \cdots \geq h_{nn}$, then*

$$\|H_s^{(0)} - I\|_F \leq \frac{1}{4} \min \left\{ \frac{1}{n}, \mathbf{g} \right\} \Longrightarrow \|H_s^{(n(n-1)/2)} - I\|_F \leq 0.715 \frac{\|H_s^{(0)} - I\|_F^2}{\mathbf{g}},$$

where $\mathbf{g} = \min\limits_{i \neq j} \dfrac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j}$ is the minimal relative gap in the spectrum.

One important ingredient of the quadratic convergence is proper affiliation of the diagonal entries of H with the corresponding eigenvalues. The assumption on the off–diagonal part of H, combined with the Weyl's theorem, implies that the diagonals of H are partitioned by belonging to disjoint intervals around the eigenvalues, where each interval contains the number of affiliated diagonals equal the multiplicity of that eigenvalue. Diagonal entries affiliated with the same eigenvalue should occupy

[1] That is, symmetric Jacobi becomes similar to Gaussian eliminations.

[2] We always assume that H has at least two different eigenvalues.

successive positions along the diagonal. *We shall always try to enforce this condition, independent of other modifications we may undertake.*

Remark 2.1. It is important to keep in mind that the reality of floating point implementation of the implicit Jacobi SVD looks quite differently from the above description of the convergence. The problem is that the zeros are introduced implicitly, thus the higher order corrections are not added to zeroes but to small quantities at the round–off level. (Floating point rotation in general makes the pivot columns at most numerically orthogonal, and not exactly orthogonal.)

2.1.2. Cubic convergence. Mascarenhas [20] used the fact that off–diagonal entries converge to zero at different rates and showed that using special quasi-cyclic strategies Jacobi method can perform cubic asymptotic convergence per quasi-cycle. Here the term quasi–cycle refers to a modified row–cyclic strategy where in a cycle some (slowly convergent) positions are visited more than once. The complexity of the quasi-cycle is about 1.25 times the complexity of classical cycle. Rhee and Hari [23] proved, under certain asymptotic conditions, the global and the cubic (per quasi-cycle) convergence of such modified method.

The cubic convergence is asymptotic, which means that the matrix $X^T X$ needs to be sufficiently close to diagonal form. Since our initial X has been preconditioned twice, we can expect that $X^T X$ is such that the effects of the higher order convergence will be non–negligible. Let the initial matrix be $H = X^T X$ and let us introduce two levels of partitions of H:

$$H = \left(\begin{array}{c||c} H^{[11]} & H^{[12]} \\ \hline\hline H^{[21]} & H^{[22]} \end{array}\right) = \left(\begin{array}{c|c||c|c} H_{[11]} & H_{[12]} & H_{[13]} & H_{[14]} \\ \hline H_{[21]} & H_{[22]} & H_{[23]} & H_{[24]} \\ \hline\hline H_{[31]} & H_{[32]} & H_{[33]} & H_{[34]} \\ \hline H_{[41]} & H_{[42]} & H_{[43]} & H_{[44]} \end{array}\right). \tag{2.1}$$

The quasi–cyclic strategy means, choosing pivot positions in row–cyclic fashion inside the submatrices, respectively, $H_{[33]}$, $H_{[44]}$, $H_{[34]}$, $H_{[33]}$, $H_{[44]}$, $H_{[11]}$, $H_{[22]}$, $H_{[12]}$, $H_{[11]}$, $H_{[22]}$, $H^{[12]}$. Let H' be the matrix computed after such quasi–cycle.

Theorem 2.3. *Let the diagonal entries of H in (2.1) be ordered form large to small, and let no two diagonal entries from different blocks $H_{[ii]}$, $H_{[jj]}$ be affiliated with the same eigenvalue. Let*

$$\Gamma_1 = \frac{\sqrt{\Omega(H^{[11]})^2 + \Omega(H^{[22]})^2}}{\mathbf{d}/3}, \quad \Gamma_2 = \left(\frac{\|H^{[12]}\|_F}{\mathbf{d}/3}\right)^{2/3}, \quad \Gamma \equiv \Gamma(H) = \max\{\Gamma_1, \Gamma_2\}.$$

If $\Gamma(H) < 1/4$, then $\Gamma(H') < (49/25)\Gamma(H)^3$. Further, if $\Gamma(H) = \Gamma_1 < 1/4$, then

$$\frac{\Omega(H')}{\mathbf{d}} \leq 18 \left(\frac{\Omega(H)}{\mathbf{d}}\right)^3.$$

Theorem 2.3 states that the convergence is cubic in the Γ measure, and in the Ω measure things look slightly differently. In fact, Rhee and Hari pointed out that the convergence of $\Omega(H)$ is only quadratic if Γ_2 dominates Γ_1. *One of the key points in our development is that preconditioning makes the dominance of Γ_2 over Γ_1 very unlikely, see [13, §3.]. Intuitively, it is then reasonable strategy to take care of Γ_1 first, thus reducing Γ down to the level which initiates cubic convergence. Briefly, we dare to hope for effects of cubic convergence very early in the process.*

Here we also note that the presented results on cubic convergence hold for the unscaled off–norm. That the principle can be extended to the scaled off–norm $\Omega(H_s)$ is only a technical matter.

2.2. How to exploit triangular form. Strategy of choosing Jacobi rotations is usually not designed to preserve any zero pattern of the input matrix. In fact, the incapability to preserve created zeros is the main reason for the poor performance, as compared with tridiagonalization and bidiagonalization based methods. Strong criticism of this is given in [22, §9.6]. In our new algorithm, we exploit a possibility to partially use and preserve the zero structure during one–sided transformations, if the initial matrix is triangular. We illustrate this in the case of lower triangular matrix.

In the scheme (2.2), X is the array in the memory occupied by the iterates in the Jacobi SVD algorithm. (That is, the input matrix X and all matrices computed by application of right–handed Jacobi rotations to X are denoted by X.) Let the columns of X be partitioned in four blocks X_1, X_2, X_3, X_4, of dimensions $n \times n_i$, respectively, where each n_i is approximately $n/4$.

$$X = \left(\begin{array}{cc|cc||cc|cc}
\blacksquare & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\blacksquare & \blacksquare & 0 & 0 & 0 & 0 & 0 & 0 \\ \cline{3-8}
\blacksquare & \blacksquare & \blacksquare & 0 & 0 & 0 & 0 & 0 \\
\blacksquare & \blacksquare & \blacksquare & \blacksquare & 0 & 0 & 0 & 0 \\ \hline\hline
\blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & 0 & 0 & 0 \\
\blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & 0 & 0 \\ \hline
\blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & 0 \\
\blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare
\end{array}\right) \equiv (X_1, X_2, X_3, X_4)\,. \tag{2.2}$$

We will say that rotations are applied in the column space of X_i if we implicitly transform $X_i^T X_i$, following some pivot strategy. Further, for two blocks of columns Y_i and Y_j of X, rotating in $Y_i \leftrightarrow Y_j$ means implicit (one sided) transformations inside the off–diagonal blocks in the matrix $\begin{pmatrix} Y_i^T Y_i & \boxed{Y_i^T Y_j} \\ \boxed{Y_j^T Y_i} & Y_j^T Y_j \end{pmatrix}$, following some pivot strategy. In other words, each pivot pair has one column from Y_i and one from Y_j. Here Y_k can be some X_i or e.g. (X_1, X_2).

Consider the most natural greedy approach. Rotating in the column space of X_4 is efficient because all columns are and remain (during rotations) in a canonical subspace of dimension $n_4 \approx n/4$. Thus, only the submatrix $X_4(n_1 + n_2 + n_3 + 1 : n, 1 : n_4)$ is transformed, and for simplicity pivot pairs can be chosen in the row–cyclic manner. In the same way, transformation of the columns of X_3 by any strategy is computation in a subspace of dimension of approximately $n/2$, as well as the transformation of the columns of the sub–array (X_3, X_4). It is trivial to see that the savings in floating point operations and the effects of better use of memory are nontrivial. Repeated transformations of the columns of X_3 (and independently X_4) are still in the lower dimensional subspace, thus very efficient. Savings in the transformations of the columns of X_2 are modest, but not worthless.

This greedy approach leads to another nontrivial improvement of the convergence. Namely, rotating, respectively, in spaces of X_3, X_4, $X_3 \leftrightarrow X_4$, X_3, X_4, X_1, X_2, $X_1 \leftrightarrow X_2$, X_1, X_2, and finally in $(X_1, X_2) \leftrightarrow (X_3, X_4)$ leads, as discussed in §2.1.2, to *cubically* convergent process.

Remark 2.2. Our goal is to initiate synergy of the positive effects of preconditioning, improved memory traffic, lower flop count (due to triangular form) and higher order convergence. Because of the preconditioning, we expect that with some partition (2.1) $H = X^T X$ is a good candidate for cubic reduction of the off–diagonal entries. However, the block partition (2.2) is motivated by flop count savings and regardless of the almost diagonal structure of $X^T X$ and the distribution of its eigenvalues, which is important for the cubic convergence. A more careful partitioning of X would try to estimate the distribution of its singular values, match the column norms and the corresponding singular values (by column permutations if needed), and then partition so that multiple or clustered singular values are approximated from single column block. This can be done, but we choose not to change the initial partition too much. We combine different devices and therefore reasonable trade–offs are unavoidable.

Obviously, X becomes dense after rotating a full sweep, and this block–quasi–cyclic strategy cannot use any zero structure of X in the next sweep. However, there is another, more subtle, structure which can be exploited and which becomes clearly visible if we study the Gram matrix $X^T X$.

2.3. How to adapt to the nearly band structure?. Now, X is dense, but $X_c = X\mathrm{diag}(1/\|X(:,i)\|)_{i=1}^n$ is close to orthogonal and the largest off-diagonal entries of $X_c^T X_c$ are located close to the diagonal. That is, we expect that $X^T X$ is γ–s.d.d. matrix in the sense of [2], and the off–diagonal mass is distributed close to the diagonal. In fact, $X^T X$ is also expected to be shifted quasi–definite matrix, see [13]. This is typical non–pathological situation. Pathological case occurs if the singular values are poorly separated, that is, the spectrum is composed of many clusters of singular values. We first discuss the case of well separated singular values.

The scheme (2.3) gives the main idea how pivot strategy can dynamically adapt to the structure of $X^T X$. The basic strategy is row–cyclic, but if in a row i certain number of consecutive rotations is skipped (because the stopping criterion is satisfied), the control of the row–cycling moves to the next row. This means that the remaining pivot positions of the i–th row are not tested against the tolerance. This strategy is motivated by the following reasons. First, it is very likely that the $\otimes$–positions in the scheme (2.3) will pass the tolerance check, so we save unnecessary dot products. (See [13].) Secondly, even if the $\otimes$–positions would not satisfy the tolerance criterion, they are expected to be much smaller than the pivot positions closer to the diagonal and it is more useful for the overall convergence to reduce those positions close to the diagonal. Certainly, this modification of the row–cyclic strategy may bring no substantial savings if the initial matrix has close singular values. But, it does no harm – in that case it simply reduces to the classical strategy. Also, this modification is in general not convergent, so an additional switch returns the control to a convergent strategy after at most 3 or 4 modified sweeps. In a non–pathological case the expected total number of rotations in 3 or 4 sweeps of this predict–and–skip strategy is $O(n)$.

$$X^T X = \left(\begin{array}{cccc||cccc}
\blacksquare & \boxplus & \boxminus & \boxminus & \otimes & \otimes & \otimes & \otimes \\
 & \blacksquare & \boxminus & \boxplus & \boxminus & \boxminus & \otimes & \otimes \\
 & & \blacksquare & \boxminus & \boxminus & \otimes & \otimes & \otimes \\
 & & & \blacksquare & \boxplus & \boxminus & \boxminus & \otimes \\
\hline\hline
 & & & & \blacksquare & \boxplus & \boxminus & \boxplus \\
 & & & & & \blacksquare & \boxplus & \boxminus \\
 & & & & & & \blacksquare & \boxplus \\
 & & & & & & & \blacksquare
\end{array}\right), \tag{2.3}$$

$\boxplus$	rotated
$\boxminus$	rotation skipped after test
$\otimes$	dot product test skipped

FIG. 2.1. Example of modified row–cyclic strategy: if two consequtive rotations in a row are skipped, then the remaining pivot positions in that row are not even tested against the threshold.

There are more good reasons to avoid even to probe the pivots at the $\otimes$ positions. It is known that the convergence of Jacobi iterations is improved if the threshold for the rotation is set higher at the beginning of the process and then gradually reduced to the final level. Such strategy is not suitable for the implicit Jacobi SVD algorithm because checking the pivot against the threshold value is extremely expensive – some $O(n)$ flops per pivot. Our predict–and–skip strategy produces the positive effect of variable threshold, but avoids dot products.

Our next rationale is related to the convergence of the scaled matrices $H_s^{(k)}$. Note that in general the sequence $\|H_s^{(\ell)}-I\|_F$, $\ell \geq 0$, does not converge monotonically. The transition from $H_s^{(\ell)}$ to $H_s^{(\ell+1)}$ introduces two zeros at the pivot positions, but the transformation is not orthogonal. Let us analyze one step. To simplify notation, consider the transformation $H' = V^T H V$, where V is the Jacobi rotation that annihilates the position (p,q). Let the corresponding scalings be $H = DH_sD$, $H' = D'H_s'D'$. Then $H_s' = (D')^{-1}V^T DH_sDV(D')^{-1}$. The matrix $S = (D')^{-1}V^T D$ differs from the identity only at the pivot positions where it takes the form

$$\hat{S} = \begin{pmatrix} \sqrt{\frac{h_{pp}}{h'_{pp}}}\cos\phi & -\sqrt{\frac{h_{qq}}{h'_{pp}}}\sin\phi \\ \sqrt{\frac{h_{pp}}{h'_{qq}}}\sin\phi & \sqrt{\frac{h_{qq}}{h'_{qq}}}\cos\phi \end{pmatrix} = \begin{pmatrix} s_{pp} & s_{pq} \\ s_{qp} & s_{qq} \end{pmatrix}. \tag{2.4}$$

Let $\Delta = \|H_s' - I\|_F^2 - \|H_s - I\|_F^2$, and let y_p and y_q be the p–th and the q–th row of the $(n-2)\times(n-2)$ matrix obtained from H_s by deleting its pivot rows and columns. Direct calculation, similar as in [12], gives

$$\frac{\Delta}{2} = \|y_p\|^2(s_{pp}^2 + s_{qp}^2 - 1) + \|y_q\|^2(s_{pq}^2 + s_{qq}^2 - 1) + 2y_p^T y_q(s_{pp}s_{pq} + s_{qp}s_{qq}) - \frac{h_{pq}^2}{h_{pp}h_{qq}}.$$

Using the formulas for the Jacobi rotation and the definition of S, we obtain

$$s_{pp}^2 + s_{qp}^2 - 1 = \frac{\frac{h_{pq}^2}{h_{pp}h_{qq}}}{1 - \frac{h_{pq}^2}{h_{pp}h_{qq}}} = s_{pq}^2 + s_{qq}^2 - 1,$$

$$s_{pp}s_{pq} + s_{qp}s_{qq} = \frac{1}{2}\sqrt{h_{pp}h_{qq}}\left(\frac{1}{h'_{qq}} - \frac{1}{h'_{pp}}\right)\sin 2\phi = -\frac{\frac{h_{pq}}{\sqrt{h_{pp}h_{qq}}}}{1 - \frac{h_{pq}^2}{h_{pp}h_{qq}}}.$$

Proposition 2.4. *Let $H^{(\ell+1)}$ be computed from $H^{(\ell)}$ by rotating at pivot position (p,q). Then the difference $\|H_s^{(\ell+1)} - I\|_F^2 - \|H_s^{(\ell)} - I\|_F^2$ equals*
$2\left(\frac{(H_s^{(\ell)})_{pq}^2}{1-(H_s^{(\ell)})_{pq}^2}\sum'((H_s^{(\ell)})_{pk}^2 + (H_s^{(\ell)})_{qk}^2) - \frac{2(H_s^{(\ell)})_{pq}}{1-(H_s^{(\ell)})_{pq}^2}\sum'(H_s^{(\ell)})_{pk}(H_s^{(\ell)})_{qk} - (H_s^{(\ell)})_{pq}^2\right)$,
where the sums $\sum'$ are $\sum_{\substack{k=1\\k\neq p,q}}^n$.

The next corollary gives a necessary condition for the growth of $\Omega(H_s^{(k)})$.

Corollary 2.5. *Let in one step of the Jacobi algorithm $\Omega(H_s^{(\ell+1)}) > \Omega(H_s^{(\ell)})$, and let $\omega = \sum_{\substack{k=1\\k\neq p,q}}^{n} ((H_s^{(\ell)})_{pk}^2 + (H_s^{(\ell)})_{qk}^2) + (H_s^{(\ell)})_{pq}^2 < 1$, where (p,q) is the pivot position. Then $\left|(H_s^{(\ell)})_{pq}\right| \leq \frac{2}{1-\omega}\sum_{\substack{k=1\\k\neq p,q}}^{n}(H_s^{(\ell)})_{pk}(H_s^{(\ell)})_{qk}$, which means that the pivot was quadratically small. The increase of the scaled off norm is also at most quadratic.*

Fairly simple proof is omitted. The corollary suggests that only too small pivot (relative to the remaining entries in pivot rows) can lead to an increase of the scaled off–norm. But even if the rotation does not increase the scaled off–norm, too small pivot cannot make the reduction substantial. That is exactly what we can expect in the $\otimes$–positions in (2.3).

The predict–and–skip idea can be transferred to checking the numerical convergence which occurs after $n(n-1)/2$ consecutive rotations have been skipped. Note that this means computation of $n(n-1)/2$ dot products not followed by rotations, an $O(n^3)$ cost that is not negligible. Of course, all dot products must be computed in order to declare numerical convergence, there is no possibility in saving flops by skipping some of them. However, if the dot products are computed as composite BLAS 3 [9] operation xSYRK, information needed to check the convergence is gathered much faster. For that, we need to predict that such an empty sweep is ahead and then we can interrupt the pivot strategy by the call to xSYRK. If the predictor is right, this means drastic reduction of the cost of the last sweep. In the case of false alarm, more rotations are needed, but we can use the information on the structure of X^TX. If extra rotations are non–overlapping, they will lead to numerical convergence. If not, the control is returned to the main pivot strategy until the predictor issues next interrupt. This schema can be constructed in many different ways (with two or more collaborating predictors) and we omit the details which will be given elsewhere, together with the details of the predictor based on convergence theory of Jacobi processes.

We finally note that the tolerance in the pivot strategy can be modified to give better performance in the case that only the singular values are needed, and with the classical absolute error bounds, see [13].

2.4. Cache–aware pivot strategy. From the early days of scientific computing it was clear that memory references play significant role in the run time. Having to read from (or even first manually replace) the magnetic tape to access matrix elements needed for operations but at the moment not in computer memory certainly kills the performance of any matrix algorithm. Nowadays, we do not use magnetic tapes in matrix computations, but even for matrices that fit into the main memory the penalty of having to perform operations on data not in fast cache memory is still severe.

Since the CPU speed is not matched by the memory speed, the patterns of memory traffic during the run time are of the utmost importance. Well designed algorithm will use the cache memory wisely and avoid various sorts of cache misses. The basic principle to improve temporal and spatial data locality has led to development of a bag of tricks and tips for cache efficient matrix computations. Of course, optimal performance on a particular machine is obtained by an additional profiling with given tools and technical data of the cache system.

Row cyclic pivoting in one–sided Jacobi SVD algorithm is an embarrassing example of violation of main principles of good cache–aware matrix computation. Fairly simple reasoning reveals that the processor is very often stalled due to cache misses. Fortunately, a simple idea of tiling taken from the bag of tricks changes that.

Introduce a parameter b (block size expressed in number of columns) and partition the columns of X. This introduces $\lceil n/b \rceil \times \lceil n/b \rceil$ block partition in $H = X^TX$ and the new strategy is to visit all blocks in the usual row–cyclic fashion, and inside each block all positions are visited row by row. This strategy is equivalent (in the sense of equivalence of pivot strategies in convergence theory of Jacobi processes) to the row–cyclic strategy: both strategies compute the same matrix after $\binom{n}{2}$ rotations.

This block–cyclic strategy can be modified following the lines of §2.2, 2.3 and we omit the technical details which can be found in the software source code. Let us just mention one out of many additional options. At the beginning of the r–th block row, after rotating in the diagonal block (r, r) we allow the possibility to transform next k diagonal blocks before entering the block $(r, r + 1)$. The parameter k is small integer (typically $0, 1, 2$) depending on X, n, b and cache parameters. It influences convergence rate (this is easily seen, take e.g. $k = 1$) and memory access patterns. The proof of global convergence of this strategy is probably only a technical matter.

Algorithm 1 Example of tiling for a modified row–cyclic strategy with tile size b.

```
{Simplified description of one full sweep}
$N_{bl} = \lceil n/b \rceil$
for $r = 1$ to $N_{bl}$ do
  $i = (r-1) \cdot b + 1$
  for $d = 0$ to $k$ do {Do the blocks $(r, r), \ldots, (r+k, r+k)$}
    $i = i + d \cdot b$
    for $p = i$ to $\min\{i + b - 1, n\}$ do
      for $q = p + 1$ to $\min\{i + b - 1, n\}$ do
        rotate pivot pair $(p, q)$
      end for
    end for
  end for
  $i = (r-1) \cdot b + 1$
  for $c = r + 1$ to $N_{bl}$ do
    $j = (c-1) \cdot b + 1$
    for $p = i$ to $\min\{i + b - 1, n\}$ do
      for $q = j$ to $\min\{j + b - 1, n\}$ do
        rotate pivot pair $(p, q)$
      end for
    end for
  end for
end for
```

3. Guessing the off–diagonal structure.

In [13], we used the diagonal entries of $A^T A$ and AA^T (that is, the column norms $\gamma_1, \ldots, \gamma_n$ and the row norms $\rho_1, \ldots, \rho_n$ of A) to decide which of the two matrices, A or A^T, would be better input to the new Jacobi SVD algorithm. This was useful in case of non–normal square matrix A. The general idea is to use certain functions of A which hare computed in at most quadratic time, and which give useful information about the off–diagonal structures of $A^T A$ and AA^T. As shown in [13], the Shannon entropy computed at the γ_i's and ρ_i's (normalized by the trace) can provide useful information. Here we continue exploration of that idea, try new interesting functions and in particular consider application to triangular matrices.

3.1. Hadamard measure.

The Hadamard measure of a positive definite matrix $H = (h_{ij})$, defined as $\chi_+(H) = \dfrac{\prod_{i=1}^n h_{ii}}{\det(H)}$ has the property of being one for a diagonal

H, otherwise it is greater than one. We then define

$$\chi(A) = \sqrt{\chi_+(A^T A)} = \frac{\prod_{i=1}^n \gamma_i}{|\det(A)|}. \tag{3.1}$$

If we compute the QR factorization of A, $A = QR$, then obviously $\chi(A) = \chi(R) = \prod_{i=1}^n (\gamma_i/|r_{ii}|)$. Now we immediately conclude that $\chi(A) \geq 1$ and $\chi(A) = 1$ **iff** the columns of A are mutually orthogonal. Further, $\chi(A)$ does not change if we scale the columns of A, permute them, or premultiply A by an orthogonal matrix.

PROPOSITION 3.1. *Let D be diagonal $n \times n$ matrix, $\det(D) \neq 0$, and let $A' = AD$. Then $\sqrt[n-1]{\chi(A)} \leq \frac{1}{2}(\kappa_2(A') + \frac{1}{\kappa_2(A')})$. Further, if $A_c = A\mathrm{diag}(1/\gamma_i)$, then $\chi(A) \geq \|A_c^{-1}\|/\sqrt{\exp(1)}$.*

Proof. The first claim uses the column scaling invariance of $\chi(A)$ and the fact that $q \equiv \dfrac{2\kappa_2(A')}{\kappa_2^2(A') + 1}$ satisfies the inequality $\sqrt{\det((A')^T A')} \geq q^{n-1} \prod_{i=1}^n \|A'(:,i)\|$. □

Moreover, a geometric point of view reveals that

$$\frac{1}{\chi(A)} = \prod_{i=2}^n \|(I - \mathbf{P}_{\mathrm{Span}\{a_1,..,a_{i-1}\}})\frac{a_i}{\|a_i\|}\| = \prod_{i=2}^n \sin\angle(a_i, \mathrm{Span}\{a_1,..,a_{i-1}\}), \tag{3.2}$$

where $a_i = A(:,i)$ and $\mathbf{P}$ is the orthogonal projection onto the indicated space. Relation (3.2) shows that linear dependencies (small angles) between the columns of A will be reflected through large value of $\chi(A)$.

Note that the values γ_i/r_{ii} are easily computed after we have the QR factorization of A. They can then be used to estimate the angles between the columns of R (that is, of A), which is relevant for the decision whether or not the second preconditioning step is necessary (cf. Algorithm 4. in [13]).

PROPOSITION 3.2. *Let $\tilde{A} = (\,A(:,1:k-1), \tilde{A}(:,k:l), A(:,l+1:n)\,)$, where $A(:,k:l) = \tilde{A}(:,k:l)\hat{T}$ is the QR factorization of $A(:,k:l)$. Then*

$$\chi(\tilde{A}) = \chi(A) \prod_{j=2}^{l-k+1} \|(I - \mathbf{P}_{\mathrm{Span}\{a_k,..,a_{k+j-2}\}})\frac{a_{k+j-1}}{\|a_{k+j-1}\|}\|. \tag{3.3}$$

Let $A = [a_1, \ldots, a_k, a_{k+1}, \ldots, a_n]$, $\hat{A} = [a_1, \ldots, a_k]$ and let $\tilde{A} = A(S \bigoplus I_{n-k})$. If the columns of $\hat{A}S$ are mutually orthogonal, then $\chi(\tilde{A}) = \chi(A)/\chi(\hat{A})$.

Proof. Obviously, $|\hat{T}_{11}| = \|a_k\|$, and $|\hat{T}_{jj}| = \|(I - \mathbf{P}_{\mathrm{Span}\{a_k,..,a_{k+j-2}\}})a_{k+j-1}\|$ for $2 \leq j \leq l-k+1$. Let $T = I_{k-1} \oplus \hat{T} \oplus I_{n-l}$, $A = \tilde{A}T$, and let $A = QR$, $\tilde{A} = \tilde{Q}\tilde{R}$ be the QR factorizations of A, $\tilde{A}$. From $\tilde{Q}^T QR = \tilde{R}T$ follows $\tilde{Q}^T Q = \mathrm{diag}(\pm 1)$, since $\tilde{Q}^T Q$ is orthogonal and the QR factorization is essentially unique. Now $|R_{jj}| = |\tilde{R}_{jj}||T_{jj}|$, $1 \leq j \leq n$, implies (3.3). □

COROLLARY 3.1. *Let $\tilde{A} = AU$, where U is the Jacobi plane transformation which makes the columns a_k and a_l orthogonal. Then $\chi(\tilde{A}) = \chi(A)\sqrt{1 - \cos^2\angle(a_k, a_l)}$.*

These properties of the function $\chi(A)$ indicate that it is a reasonable measure for the diagonality of $A^T A$. Note that for our purposes we do need to compute $\det(A)$ because we need to compare $\chi(A)$ and $\chi(A^T)$, eg. it is enough to have $\chi(A)/\chi(A^T) = \prod_i (\gamma_i/\rho_i)$. (Note that this expression is independent of the column and row orderings and that in floating point it is safely computed in the log scale.)

3.1.1. Estimating the scaled off–norm. There is more. Once the first QR-decomposition of A is made, we have the Hadamard measure of both A^TA and AA^T — cheaply! Equally cheaply we may compute the Hadamard measure before each Jacobi cycle (or even update it after each rotation!) and decide how diagonal the, otherwise invisible, matrix A^TA is. Note that after Jacobi has started, the matrix A is not triangular anymore. We will show that the Hadamard measure bounds the scaled diagonal dominance of the QR triangular factor of A. More precisely, we have

PROPOSITION 3.3. *Let $A^TA = DBD$ such that B has unit diagonals. Let $B = R^TR$, $R = (r_{ij})$, be the Cholesky decomposition of B. Then*

$$\|\Omega(R)\|_E^2 \le (n-1)\left(1 - \left(\frac{1}{(\chi_+(A^TA))}\right)^{1/(n-1)}\right) \le \ln\left(\chi_+(A^TA)\right), \tag{3.4}$$

where $\Omega(R)$ is the off-diagonal part of R. This estimate is asymptotically sharp for $\chi_+(A^TA)$ close to one.

Proof. By $r_{1j}^2 + \cdots + r_{jj}^2 = 1$, $j = 1, \ldots, n$, we have

$$\chi_+(A^TA) = \chi_+(B) = \chi_+(R^TR) = \frac{1}{(1-\alpha_2)\cdots(1-\alpha_n)} \tag{3.5}$$

with $\alpha_j = r_{1j}^2 + \cdots + r_{j-1j}^2$, $j = 2, \ldots, n$. Obviously, $0 \le \alpha_i < 1$. Now,

$$\|\Omega(R)\|_E^2 = \alpha_2 + \cdots + \alpha_n = (n-1)\left(\frac{n-1-((1-\alpha_2)+\cdots+(1-\alpha_n))}{n-1}\right)$$
$$\le (n-1)\left(1 - \sqrt[n-1]{\prod_{j=2}^{n}(1-\alpha_j)}\right) \le (n-1)\left(1 - \sqrt[n-1]{\frac{1}{\chi_+(A^TA)}}\right).$$

Here we have used (3.5) as well as the fact that the arithmetic means dominates the geometric one. Thus, the first inequality from (3.4) follows. To prove the second one we consider the function $0 < x \mapsto g(x) = x(1 - b^{\frac{1}{x}})$, where $b = 1/\chi_+(A^TA) < 1$. By use of the L'Hospital rule we see that $\lim_{x\to\infty} g(x) = -\ln b$. Also, this function is increasing with x. This is seen from its derivative $g'(x) = 1 - e^{\frac{\ln b}{x}}\left(1 - \frac{\ln b}{x}\right)$ which is positive by the known inequality $1 + y < e^y$ with $y = -(\ln b)/x > 0$. Thus, the second inequality from (3.4) follows. Finally, observe that $\chi_+(A^TA)$ is close to one, if and only if all α_i are small, in this case (3.5) reads asymptotically $\chi_+(A^TA) \approx 1 + \alpha_2 + \cdots + \alpha_n = 1 + \|\Omega(R)\|_E^2$. Hence $\|\Omega(R)\|_E^2 \approx \ln(\chi_+(A^TA))$, and the last assertion follows as well. □

There are now various possibilities: we may easily check/update $\chi_+(A^TA)$ at every desired stage of the process (it may be necessary to work with logarithms) and stop when it is close enough to one. Or, if we fear rounding errors in its computed value (they will be fairly small on Jacobi after a pivoted QR decomposition, as our preliminary experiments have indicated) we may make a final check by computing A^TA explicitly. Another difficulty is that the Hadamard measure must be ε^2-close to one in order to insure the ε-diagonal dominance.

3.2. Diagonal means measure. The normalization by the invariant $\det(\cdot)$ may be replaced by another one, which is easier to compute and this is, for instance, a function of the trace. With $H = A^TA$ we set

$$\mu_+(H) = \frac{\sqrt[n]{\prod_{i=1}^n h_{ii}}}{\mathrm{Trace}(H)/n}, \quad \mu(A) = \sqrt{\mu_+(H)}.$$

Obviously, as long we run over a set of mutually similar matrices the two measures $\chi(\cdot)$ and $\mu(\cdot)$ are just proportional – A^TA, AA^T are orthogonally similar! $\mu(A)$ is equal to one if and only if the diagonal of H is constant. Otherwise $\mu(A)$ is less than one and takes its minimum, if and only if H is diagonal.

The difference between $\mu(\cdot)$ and the Hadamard measure lies in the fact that $\mu(\cdot)$ cannot be used absolutely i.e. on a given single matrix. It has the value one on any diagonally constant matrix while its minimum depends on the (generally unknown) spectrum. On the contrary, the Hadamard measure equals to one, iff the matrix is diagonal. Our, 'purely diagonal', measure $\mu(\cdot)$ just helps to decide which of the mutually similar matrices is better diagonal. There are many such diagonal measures.

Theorem 3.4. *Let f be a twice continuously differentiable function on $[0,\infty)$ with $f, -f'' > 0$. Let $\mathcal{S}$ be set of mutually orthogonally similar positive definite real symmetric matrices of order n. With $A = (a_{ij}) \in \mathcal{S}$ set $h(A) = g\left(\sum_{i=1}^n f(a_{ii})\right)$, where g is any strictly increasing continuous positive valued function. Then $h(A)$ takes its minimum, if and only if A is diagonal and its maximum, if and only if A has constant diagonal i.e. $a_{ii} = \mathrm{Trace}(A)/n$, $i = 1,\ldots,n$.*

Proof. Without loss of generality we may assume that $g(x) \equiv x$. We consider first the case $n = 2$. Again, without loss of generality we may assume that the eigenvalues of A are $\mu_1 \neq \mu_2$. We are looking for the extrema of the function $h(A) = f(a_{11}) + f(a_{22})$ under the conditions

$$a_{11} + a_{22} = const._1 > 0 \qquad a_{11}a_{22} - a_{12}^2 = const._2 > 0. \tag{3.6}$$

By the compactness argument there exist a minimum and a maximum. We apply the Lagrange formalism. Let $L = f(a_{11}) + f(a_{22}) + \lambda_1 (a_{11} + a_{22}) + \lambda_2 \left(a_{11}a_{22} - a_{12}^2\right)$. The differential of the conditions (3.6) is $\begin{pmatrix} 1 & a_{22} \\ 1 & a_{11} \\ 0 & -2a_{12} \end{pmatrix}$. This has rank 2 (except in the case $a_{12} = 0$, $a_{11} = a_{22}$ which is excluded). Setting grad $L = 0$ gives $f'(a_{11}) + \lambda_1 + \lambda_2 a_{22} = 0$, $f'(a_{22}) + \lambda_1 + \lambda_2 a_{11} = 0$, and $-2\lambda_2 a_{12} = 0$. Thus, we have the alternative:

1. $\lambda_2 = 0$: this implies $a_{11} = a_{22} = \dfrac{\mu_1 + \mu_2}{2}$, and $f(a_{11}) + f(a_{22}) = 2f(\dfrac{\mu_1 + \mu_2}{2})$.
2. $\lambda_2 \neq 0$: this implies $a_{12} = 0$ and $f(a_{11}) + f(a_{22}) = f(\mu_1) + f(\mu_2)$.

By $\mu_1 \neq \mu_2$ and $f'' < 0$ we conclude $\dfrac{f(\mu_1) + f(\mu_2)}{2} < f\left(\dfrac{\mu_1 + \mu_2}{2}\right)$. Thus, in the first case we have a maximum and in the second a minimum.

Consider now the case of a general n. On minimum the matrix must be diagonal. Indeed, if e.g. $a_{12} \neq 0$ then we can annihilate a_{12} by a plane rotation $A' = U^TAU$. Then $h(A)$ goes into $h(A') = f(a'_{11}) + f(a'_{22}) + f(a_{33}) + \cdots + f(a_{nn})$ and $f_0(a'_{11}) + f_0(a'_{22}) < f_0(a_{11}) + f_0(a_{22})$ as we have proved above. Similarly, on maximum the diagonal is seen to be constant. The proof is completed by recalling that the function h is constant on all diagonal matrices from $\mathcal{S}$ and also on all matrices from $\mathcal{S}$ which have constant diagonal. □

4. Numerical testing. In this section we present mathematical software that implements the new algorithm described in [13, Algorithm 4] and in this report. We give the results of preliminary testing of the algorithm with respect to numerical accuracy and efficiency (run time compared with existing algorithms). Our goal is numerically reliable software implemented to reach reasonable fraction of the efficiency of the less accurate bidiagonalization based methods. In other words, to make the high accuracy of the Jacobi SVD algorithm so affordable that the new algorithm becomes attractive as one of the methods of choice for dense full SVD computation.

High performance matrix computation is usually achieved by using machine optimized BLAS library. It is our impression that the developers of these libraries are mainly interested in gaining high flop count per second at any price, and that some numerical properties can be silently lost in the release after release after release fixes and improvements. It seems that the potentials of the processors with respect to numerical accuracy are sometimes sacrificed in favor of potential for optimization for speed. The criticism of such tendencies in hardware and software developments is best expressed by Kahan [19]. Memory leaks are sometimes considered acceptable risk for gaining higher execution speed. Bugs in such libraries are considered as unlucky accidents to be resolved in the next releases.

Here we assume that we can trust that the compiler, the implemented machine arithmetic and the BLAS library are bug free. One should be aware that this is not always warranted, and that implementing and testing a theoretically completed algorithm into mathematical software is far from being a routine software engineering task, even if we use building blocks from the BLAS and the LAPACK libraries. Sometimes, it requires tedious and frustrating work even to get commercial compiler suite to link properly. The criticism expressed in the 1973. review paper of F. Bauer [3] applies today as well.

4.1. Why and how to test an SVD algorithm. Why do we need to test software implementation of a numerical algorithm if we have numerical analysis proving its accuracy? Once we are sure that there are no pure programming errors in the software and if we assume that the arithmetic is implemented according to the model (standard) used in the analysis, everything should run as predicted by the theory.

The problem is that the assumptions of the developed theory are not always met in the practice. Typical example is expectation that implemented arithmetic behaves following simplified model which ignores (perhaps of really low probability but not impossible) occurrences of underflow (flush to zero or denormalized) in the error analysis. Good starting point in the study of these problems is [6]. Further, there is a dangerous possibility of a gap between numerical properties of standard implementation o some elementary matrix factorization (used in the theoretical analysis) and its highly optimized version used to get optimal performance.

Carefully designed testing of the software is also test of the theory. It shows how sharp are the theoretical bounds and it also gives new insights in the cases of input matrices which are on the boundaries of the theoretical assumptions. Good test cases (accidentally found or cleverly designed) give insights into the behavior of the algorithm and may induce modifications which improve the efficiency of the algorithm. The feedback loop created in this way is part of the research process. In fact, before giving any test result, in §4.2 we describe modification of the algorithm, prompted by numerical tests of an early version of the code.

Finally, testing different algorithms can help in deciding which one is the method of choice in particular situations, with respect to numerical properties (accuracy, stability) and efficiency (speed, memory requirements).

We test single precision (32 bit representation, $\varepsilon \approx 5.3 \cdot 10^{-8}$) implementation because lower precision computation increases the probability of encountering unforeseen numerical difficulties related to finite precision arithmetic. Underflow threshold is denoted by $\boldsymbol{\nu}$ and overflow by $\boldsymbol{\omega}$. It is always assumed that the nonzero entries of input matrix are normalized floating point numbers.

4.1.1. Measuring error – distance to what. Once we compute $A \approx \tilde{U}\tilde{\Sigma}\tilde{V}^T$, $\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n)$, we can immediately estimate the quality of the computed decomposition using the following computed quantities

$$\begin{aligned}
\mathbf{r} &= \mathit{computed}(\frac{\|A - \tilde{U}\tilde{\Sigma}\tilde{V}^T\|_F}{\|A\|_F}) \text{ (should be at most } f(m,n)\varepsilon,\ f \text{ moderate);} \\
\mathbf{o}_U &= \mathit{computed}(\max_{i,j} |(\tilde{U}^T\tilde{U})_{ij} - \delta_{ij}|) \text{ (should be at most } O(m\varepsilon)); \\
\mathbf{o}_V &= \mathit{computed}(\max_{i,j} |(\tilde{V}^T\tilde{V})_{ij} - \delta_{ij}|) \text{ (should be at most } O(n\varepsilon)).
\end{aligned}$$

These measures are useful to test the correctness of the code (bug free in the usual sense) and backward stability in the matrix norm sense. It is easy to show that $\mathbf{r}$, $\mathbf{o}_U$ and $\mathbf{o}_V$ can be computed sufficiently accurately (best using higher precision) to be used as relevant measures of the quality of the computed decomposition. Thus, the standard error bound can be a posteriori numerically checked. On the other hand, the column–wise backward error (that is, residual bounds for each column of A) cannot be easily checked. Namely, it holds that

$$\mathit{computed}(\tilde{U}\tilde{\Sigma}\tilde{V}^T) = \tilde{U}\tilde{\Sigma}\tilde{V}^T + E, \quad |E| \leq O(n\varepsilon)|\tilde{U}| \cdot \tilde{\Sigma} \cdot |\tilde{V}|^T, \tag{4.1}$$

which means that small columns of $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ are not necessarily computed to sufficient accuracy, so even if the backward error in some column is small, we are not able to detect that fact by using computed residuals. Note that even double precision computation of the residual may not be enough if the condition number of A is large.

One difficulty in testing a new SVD software on a large set of pseudo–random matrices is how to provide reference (exact) values of Σ, U and V which are used to estimate the accuracy of the computed approximations $\tilde{\Sigma}$, $\tilde{U}$, $\tilde{V}$. One could start by generating pseudo–random numerically orthogonal $\overline{U}$, diagonal $\overline{\Sigma}$ and numerically orthogonal $\overline{V}$ and then define $A = \mathit{computed}(\overline{U}\ \overline{\Sigma}\ \overline{V}^T)$. But as the relation (4.1) shows, the numerical SVD of A may be very much different from $\overline{U}\ \overline{\Sigma}\ \overline{V}^T$. Using the same algorithm in higher precision is useful but not always – depending on the matrix, it is possible that both procedures compute with large errors. We should also keep in mind that working in higher precision generates different numerical process in finite precision environment and that double precision procedure also needs to be tested. The alternative is to use existing, tested (!) and trusted double precision software to compute the SVD of a given test matrix A. In our case, this means DGESVD and/or DGESDD[3] from LAPACK, but this will be useful only as long these procedures guarantee at least eight digits of accuracy, that is for (roughly) $\kappa(A) < 1/\varepsilon \approx 10^8$. Our choice of the reference procedure is classical one–sided Jacobi SVD with de Rijk's pivoting, implemented in double precision.

If $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_n$ and $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_n$ are the computed and the reference singular values computed in higher precision, then the forward errors of interest are[4]

$$\mathbf{e}_i = \frac{|\tilde{\sigma}_i - \hat{\sigma}_i|}{\hat{\sigma}_i}, \quad i = 1, \ldots, n, \quad \mathbf{e} = \max_{i=1:n} \mathbf{e}_i. \tag{4.2}$$

[3] During the testing we have accidentally found an example of serious failure of DGESDD procedure from the SUN performance library – a ghost singular value of the size of the largest one appeared in the dominant part of the spectrum.

[4] Here is by definition $0/0 = 0$.

We should not forget that the double precision and the single precision procedure do not have the same matrix on input, as the single precision version A_{single} of a double precision matrix A_{double} is generally from an $\boldsymbol{\varepsilon}$–neighborhood of A_{double}.

4.1.2. Test matrices. Our primary target are matrices of the form $A = BD$, where D is diagonal and B is well conditioned with equilibrated (unit in Euclidean norm) columns. In that case the relative error in the output is governed by the condition number $\kappa(B)$ independent of D. To illustrate this property we need to generate test matrices $A = BD$ where B has given $\kappa(B)$ and unit columns. Moreover, the matrices should be generated so systematically that the maximal measured forward errors attain the predicted theoretical bounds, and that experimental data show that no accuracy can be guaranteed if the assumptions of the theory are not satisfied. In that case we will have experimental evidence that both the theory and the numerical testing are done properly.

We use the algorithm of Stewart [24] to generate random orthogonal matrices distributed uniformly according to the Haar measure over the orthogonal group $\mathcal{O}(n)$. If W_1 and W_2 are two such matrices, and if S is diagonal with given condition number $\kappa(S)$, we compute $C = W_1SW_2$. Then we use the fact that for the matrix C^TC there always exists an orthogonal W_3 such that the diagonal entries of $W_3^T(C^TC)W_3$ are all equal to $\mathrm{Trace}(C^TC)/n$. Then the matrix $B = CW_3$ has equilibrated columns and condition number κ. If we generate diagonal D, then $A = BD$. There are several ways to generate the matrix W_3, see e.g. [4], [5]. The distributions of the diagonal entries of S and D can be chosen in different ways. We use the modes provided in the LAPACK test matrix generators (parameter MODE in DLATM1) and the chosen modes are denoted by $\mu(S)$ and $\mu(D)$. Thus, each generated matrix A has four parameters, $p(A) = (\kappa(S), \mu(S), \kappa(D), \mu(D))$. For each fixed $p(A)$, we use three different random number generators provided in LAPACK testing library (LAPACK/TESTING/MATGEN/), and with each of them we generate certain number of samples (test matrices). In this way, we have an automated generator of pseudo–random matrices with certain relevant parameters varying systematically in given range, for instance $\kappa(S)(= \kappa(B))$ is set to take the values 10, 10^2, $10^3, \ldots, 10^8$.

4.2. Underflow and overflow – problems and solutions by perturbation theory. In error analysis of matrix algorithms we usually assume that the data are so scaled that no overflow and underflow exceptions occur during the computation. So, for instance, LAPACK's driver routine SGESVD computes $\alpha = \max_{i,j} |A_{ij}|$ and scales A with $(1/\alpha)\sqrt{\boldsymbol{\nu}}/\boldsymbol{\varepsilon}$ (if $\alpha < \sqrt{\boldsymbol{\nu}}/\boldsymbol{\varepsilon}$) or $(1/\alpha)\boldsymbol{\varepsilon}\sqrt{\boldsymbol{\omega}}$ (if $\alpha > \boldsymbol{\varepsilon}\sqrt{\boldsymbol{\omega}}$).

Since our implementation of Jacobi rotations uses the column norms and the cosines of the angles between the columns, we can compute the singular values almost in the entire range $(\boldsymbol{\nu}, \boldsymbol{\omega})$ of floating point numbers. This means that we need to insure that A is scaled so that its maximal column is not larger than $\boldsymbol{\omega}/\sqrt{n}$. (Note that largest value of any column norm of an $m \times n$ A is $\sqrt{m}\boldsymbol{\omega}$.) However, since we use computational routines from other libraries (such as BLAS and LAPACK) which may implicitly use elements of A^TA, we choose to be on the safe side and scale A to make sure that its maximal column is not larger than $\sqrt{\boldsymbol{\omega}}/\sqrt{n}$ in Euclidean norm. If the spectrum of singular values spreads over the full range of normalized numbers (which can be detected for instance by discovering columns with very small (near $\boldsymbol{\nu}$) as well as very large (near $\boldsymbol{\omega}$) Euclidean length) and if all of them are wanted to high relative accuracy then enforcing $\sqrt{\boldsymbol{\omega}}/\sqrt{n}$ as the maximum column norm wipes out smallest singular values. In that case $\boldsymbol{\omega}/\sqrt{n}$ is better choice, provided that no computational routine in linked libraries squares the column norms. In any case, the scaling factor c

can also be modified to represent the closest exact power of the base of the arithmetic.

An extension of this is to use diagonal scalings – we can think of $A = A_0 D_0$ with diagonal D_0 and such that A cannot be stored because of underflow/overflow, but both A_0 and D_0 can. Fast scaled rotations work on the pair A_0, D_0 and deliver the result in factored form.

Underflow problems can be dangerous, tricky and not removable by scaling. In the case of Jacobi algorithm, underflow can cause non–convergence and loss of accuracy. A catastrophic instance of this is flushing Jacobi rotation to identity, even if its action is needed and substantial. Solution of this problem, independent of the mode the denormals are treated (gentle or flush to zero underflow) is given in [11], and we present it here for the readers convenience.

4.2.1. What if rotation underflows?. Recall that in each step the Jacobi rotation implicitly diagonalizes the Gram matrix

$$G = \begin{pmatrix} a & c \\ c & b \end{pmatrix} = \begin{pmatrix} x^T \\ y^T \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix}, \quad x, y \in \mathbf{R}^m \setminus \{0\}. \tag{4.3}$$

The rotation is defined by

$$J = \frac{1}{\sqrt{1+t^2}} \begin{pmatrix} 1 & t \\ -t & 1 \end{pmatrix}, \;\; \text{where} \;\; t = \frac{\mathrm{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}, \;\; \text{and} \;\; \zeta = \frac{b-a}{2c}. \tag{4.4}$$

In practice, one first computes the parameters a, b, c, and then carefully implements the formulas (4.4). We use slightly modified strategy for the following reasons:

• The range of the computational routine should be the largest possible. This means that singular values should be computed with theoretical accuracy if they are between the underflow and the overflow thresholds. Since the computation of the Gram matrix (4.3) squares the column norms, using the formulae (4.3), (4.4) in general reduces the computation range to the interval $(\sqrt{underflow}, \sqrt{overflow})$. We find this an unnecessary restriction that should be removed. One obvious way to do that is to use higher precision (single to double, double to extended) to compute G. We prefer more intrinsic modification of the formula for ζ, namely

$$\zeta = \frac{\sqrt{\dfrac{b}{a}} - \sqrt{\dfrac{a}{b}}}{2\dfrac{c}{\sqrt{ab}}} = \frac{\dfrac{\|y\|_2}{\|x\|_2} - \dfrac{\|x\|_2}{\|y\|_2}}{2\cos\angle(x,y)}. \tag{4.5}$$

Since $\sigma_{\min} \leq \|x\|, \|y\| \leq \sigma_{\max}$, the column norms will not overflow (underflow) unless $\sigma_{\max}$ ($\sigma_{\min}$) overflows (underflows).

• One sided algorithm introduces the zeros implicitly by orthogonalizing the pivot columns. This means that inaccuracies in the rotation parameters (due to roundoff) do not affect the accuracy of a single floating point rotation – the constructed rotation is numerically orthogonal and its application is stable. But, to reach the eventual convergence, each rotation must increase toward $\pi/2$ the acute angle between pivot columns. The formula (4.5) clearly reveals potential source of problem. If e.g. $\|x\| \gg \|y\|$ so that $\|x\|/\|y\|$ overflows, then $\tilde{\zeta} = computed(\zeta)$ overflows and the computed t is flushed to zero (thus, the computed rotation is identity) or denormalized (thus, with loss of relative accuracy). Note that denormalized nonzero $\tilde{t} = computed(t)$ is possible if it is first computed in higher precision arithmetic and then rounded back to working precision. From the formulas

$$\zeta = \frac{\sqrt{\frac{b}{a}} - \sqrt{\frac{a}{b}}}{2\frac{c}{\sqrt{ab}}} = \frac{\frac{\|y\|_2}{\|x\|_2} - \frac{\|x\|_2}{\|y\|_2}}{2\cos\angle(x,y)}, \quad t = \frac{\operatorname{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}$$

we concluded that, for $\|x\| \gg \|y\|$, $\tilde{\zeta}$ or $\tilde{\zeta}^2$ may overflow, thus causing $\tilde{t}$ to underflow, independent of the value of $\cos\angle(x,y)$. Here we do not mean only the trivial mistake to use the straightforward formula for t as function of ζ in cases of big ζ (e.g. if $\tilde{\zeta}^2$ overflows, but $\tilde{\zeta}$ does not) – the correct formula is then $\tilde{t} = \mathit{computed}(0.5/\tilde{\zeta})$.

If $\tilde{\zeta}$ overflows, the problem is subtle. For, even if we go over to the higher precision to compute the rotation, its satisfactory working precision representation may not exist. If we keep it in higher precision, then its application to the working precision data would be inefficient because of the type conversion.

Important thing to realize here is that floating point matrix representation of the rotation is identity, but the rotation itself performs nontrivial action. To solve this problem, we first analyze the geometry behind the rotation as $\|x\|/\|y\|$ grows toward infinity. In fact, useful insight into the geometric structure of the transformation in such situation is already possible for $\|x\| > \|y\|/\sqrt{\boldsymbol{\varepsilon}}$. In that case

$$\cot 2\phi = \frac{\frac{\|y\|}{\|x\|} - \frac{\|x\|}{\|y\|}}{2\frac{y^T x}{\|x\|\|y\|}} \approx \frac{-\frac{\|x\|}{\|y\|}}{2\frac{y^T x}{\|x\|\|y\|}}, \quad |\cot 2\phi| \geq \frac{1}{2\sqrt{\boldsymbol{\varepsilon}}}$$

and, with relative error of order $\boldsymbol{\varepsilon}$, $\tan\phi \approx \frac{1}{2}\frac{1}{\cot 2\phi} \approx -\frac{y^T x}{\|x\|^2} = -\frac{y^T x}{\|x\|\|y\|}\frac{\|y\|}{\|x\|}$. Thus, $|\tan\phi| \leq \sqrt{\boldsymbol{\varepsilon}}$ and $\cos\phi$ will be computed as one. Note that in fact the transformation $y' = y + x\tan\phi \approx y - x\frac{y^T x}{\|x\|^2}$ performs the Gram–Schmidt orthogonalization of y against x and that the transformation $x' = x - y\tan\phi$ changes x only by

$$|\tan\phi|\|y\| \approx \frac{|y^T x|}{\|x\|\|y\|}\frac{\|y\|^2}{\|x\|} \leq \frac{\|y\|^2}{\|x\|^2}\|x\| \leq \boldsymbol{\varepsilon}\|x\|.$$

This modification can be rigorously analyzed and it works very well in practice. See [11] for more details on this and other issues related to floating point Jacobi rotation.

Now we discuss a different sort of problems – denormalized values are just annoying small numbers which have no influence on the result (they are added to larger values, or can be treated as tiny perturbation) but do have influence on the execution time if not properly handled by the computer.

4.2.2. Just annoying underflows. Denormalized numbers are a clever way to fill the interval around zero free of normalized numbers. From the mathematical point of view this is much more than a neat idea, see [6]. Unfortunately, the implementation of this extension (and some other exceptions) on modern processors and operating systems is not very efficient and it is possible that many denormalized numbers, which in the algorithm may be as good as zeros, drastically slowdown the computation. Depending on the computational task, the *set to zero* option may solve the problem, provided it exists and works properly. It also helps to scale the matrix with maximal factor which will not cause overflow. (For a discussion on how to deal with floating point exceptions in numerical algebra see [6], [7].) However, if the mechanism that creates denormalized numbers is invariant under scaling, the problem persists.

Let us first describe the genesis of the unwanted denormalized numbers in our algorithm. Their first occurrence is possible in the QR factorization of A, even if all non–zero entries of A are deep in the range of the normalized numbers. The easiest illustrative example is sparse A where many Householder reflections hit a very sparse vector. (Think of a 1000×1000 lower triangular matrix (plus some random very sparse upper triangular) and consider application of a series of Householder reflections to its last column. In the same way, the reconstruction of the Q factor can have the same problem.) Once we have computed $\tilde{R}$ with some denormalized entries we transpose it and compute the QR factorization again. The same mechanism generates new denormals and there it is! The second QR factorization can be extremely slow if the arithmetic with denormalized numbers is not well implemented. (The first QR factorization too, but this kind of problem with general dense A occurs with lower probability.) From the point of view of numerical accuracy, these denormalized numbers are as good as zeros – backward stability and the forward errors are given with respect to column norms and these are well preserved if the entries of the initial A are normalized numbers.

To make things worse, there is another mechanism that produces denormals – the QR iteration which we used as preconditioner. If the matrix A is such that at some index k it holds $\sigma_k \gg \sigma_{k+1}$ and if the RRQR factorization $AP = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$ is good, then we expect separation to be seen in R. Thus, the second QR factorization on R^T may produce even more denormalized numbers.

An finally, the Jacobi rotations of the columns of lower triangular factor are the last victims of the small villains. The dot product needed to compute the rotation angle can be extremely slow (many of the summands can be denormalized and are as good as zeros in the final result), and rotations with small angles may generate additional denormalized matrix entries, especially if sparse vectors are involved.

The problems described above are very likely to appear if the columns of A are heavily graded or if A is large and sparse with most nonzero entries in the lower triangle. In that case the convergence of the Jacobi iterations is swift, the total number of rotation is very small and the run time is unacceptably long. We have found this situation rather frustrating – the worst performance in cases of swift convergence!

How to deal with this problem? The first natural idea would be to set small off–diagonal matrix entries to zero, but it is immediately clear that this is not the solution of the problem. First, inspecting individual entries after each transformation in QR factorization in order to set denormalized numbers to zero kills the performance of highly optimized blocked code. Secondly, the denormals would keep reappearing because they actually grow in places of zero entries. We choose quite the opposite approach – using artificial perturbation we destroy all zeros and increase small entries.

Let X be $n \times n$ lower triangular and nonsingular transposed upper triangular factor from the QR factorization. Our goal is to replace X with $X + \delta X$ where the perturbation δX is: *(i)* small enough so that it does not introduce errors larger than the initial uncertainty of the SVD caused by computing X; *(ii)* big enough to prevent underflows in the next QR factorization or Jacobi iterations; *(iii)* small enough so that it does not interfere with the preconditioner and that it does not prevent the use of the lower triangular structure; *(iv)* small enough so that it does not preclude stable a posteriori computation of the right singular vectors. Clearly, the set of active constrains depends on the moment in the algorithm and the task to be completed with given X. We list the relevant cases and discuss the structure of perturbation that fits our needs. We use ζ to denote appropriate threshold value used in construction of $\delta X \equiv \delta X_\zeta$, for instance $\zeta = \sqrt{\boldsymbol{\nu}/\boldsymbol{\varepsilon}}$, or $\zeta = \sqrt{\boldsymbol{\nu}}$, or $\zeta = \boldsymbol{\varepsilon}/n$.

Case 1. *In the next step we need QR factorization of X (with or without column pivoting) or we apply the right–handed Jacobi SVD algorithm but without computation of the right singular vectors.* Here X is the result of the first or the second QR factorization with pivoting, thus with known nontrivial structure. (Some of the ideas apply to general lower triangular matrix with no other known structure.) We first describe the lower triangle of δX_ζ. Each X_{ij} with the property $|X_{ij}| < \zeta |X_{jj}|$ is replaced with $\mathrm{sign}(X_{ij})\zeta|X_{jj}|$, thus $(\delta X_\zeta)_{ij} = -X_{ij} + \mathrm{sign}(X_{ij})\zeta|X_{jj}|$ for all $i > j$. Simultaneously, the position X_{ji} in the upper triangle is set to $(\delta X_\zeta)_{ji} = -\mathrm{sign}(X_{ij})\zeta|X_{ii}|$. (Here ζ can be replaced with a random number of the order of magnitude of ζ.) Note that $\|\delta X_\zeta(:,j)\| \leq \sqrt{n}\zeta|X_{jj}|$ which means that computations with X and $X + \delta X_\zeta$ are almost indistinguishable from the backward and forward error points of view.

Case 2. *In the next step we compute the SVD of X using right–handed Jacobi rotations and the right singular vectors are computed a posteriori from matrix equation.* In this case we also have to preserve the stability of the matrix equation solved for the right singular vectors. This means that the perturbation δX_ζ has to be also row–wise small. This is achieved by taking $(\delta X_\zeta)_{ij} = -X_{ij} + \mathrm{sign}(X_{ij})\zeta\min\{|X_{ii}|, |X_{jj}|\}$, $(\delta X_\zeta)_{ji} = -\mathrm{sign}(X_{ij})\zeta\min\{|X_{ii}|, |X_{jj}|\}$ for all $i > j$. It is interesting that after introducing δX_ζ we simply ignore it and treat $\tilde{X} = X + \delta X_\zeta$ as lower triangular. This is because the pivot strategy tailored for lower triangular matrix is important for the overall performance and losing its effects is not an option. To justify this manipulation with δX_ζ, we first note that the upper triangle of δX_ζ contains tiny row–wise relative error in X. Indeed, for $i < j$ we have

$$\frac{|(\delta X_\zeta)_{ij}|}{\|X(i,:)\|} \leq \zeta \frac{|X_{jj}|}{\|X(i,:)\|} \leq \zeta \;\; (\text{since } |X_{jj}| \leq |X_{ii}|)$$

where in case of graded matrix $|X_{jj}|/\|X(i,:)\| \ll 1$. Consider now any $(k, n-k)$ block partition of $\tilde{X}$,

$$\tilde{X} = \begin{pmatrix} \tilde{X}_{11} & \tilde{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} \end{pmatrix} = \begin{pmatrix} \tilde{X}_{11} & 0 \\ \tilde{X}_{21} & \tilde{X}_{22} \end{pmatrix} + \begin{pmatrix} 0 & \tilde{X}_{12} \\ 0 & 0 \end{pmatrix} \equiv \tilde{X}_{\text{L}} + \delta\tilde{X}_{\text{L}}.$$

Obviously, $\|\tilde{X}_{12}(i,:)\| \leq \sqrt{n-k}\zeta\|X(i,:)\|$ for all $i = 1, \ldots, k$. Suppose we are to rotate last $n-k$ columns of $\tilde{X}$, following some pivot strategy, but decide to think of $\tilde{X}$ as if it was $\tilde{X}_{\text{L}}$ (that is, we do not transform $\tilde{X}_{12}$). The computed matrix is represented by the backward perturbation analysis as

$$\begin{aligned} \tilde{X}' &= \begin{pmatrix} \tilde{X}_{11} & 0 \\ \tilde{X}_{21} & \tilde{X}_{22} + \delta\tilde{X}_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \hat{W} \end{pmatrix} + \begin{pmatrix} 0 & \tilde{X}_{12} \\ 0 & 0 \end{pmatrix} = \\ &= \begin{pmatrix} \tilde{X}_{11} & \tilde{X}_{12}\hat{W}^T \\ \tilde{X}_{21} & \tilde{X}_{22} + \delta\tilde{X}_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \hat{W} \end{pmatrix}, \;\; \hat{W}^T\hat{W} = I, \\ \text{where} \quad & \|\delta\tilde{X}_{22}(i,:)\| \leq \varepsilon_J \|\tilde{X}_{22}(i,:)\|, \;\; i = 1, \ldots, n-k, \end{aligned}$$

which means that ignoring $\tilde{X}_{12}$ is equivalent to replacing it with $\tilde{X}_{12}\hat{W}^T$. Since right–handed orthogonal transformation does not change the Euclidean lengths in the row space of the involved matrices, the row–wise backward stability is preserved.

Remark 4.1. Adding artificial perturbation to avoid underflows can be applied for instance in computing the orthogonal Q from stored reflections by multiplying perturbation of the identity by the reflections constructed during the QR factorization or bidiagonalization. Note that, also in this case, the perturbation is constructed, applied and ignored in exploiting the special structure of the unperturbed matrix.

4.3. Test results. Our new algorithm is implemented in a LAPACK–style routine SGEPVD, which is in an early stage of the development. We have done no serious profiling in order to optimize it for a particular architecture. The most consuming part of the code (rotations) is still on BLAS 1 level and we have plans to change this in near future. Nevertheless, the obtained results are surprisingly good and encouraging.

4.3.1. Computing the full SVD. The test matrices of the form $A = BD$ are generated as follows. We take $A = ((W_1 S W_2) W_3) D$ as described in §4.1.2, and $\kappa(S) = 10^i$, $i = 1, \ldots, 8$ and $\kappa(D) = 10^{2j}$, $j = 0, \ldots, 7$. For each fixed pair (i, j) we generate diagonal S and D each with four different distributions of the diagonal entries (as specified by the parameter MODE). This gives for each fixed $(\kappa(S) = 10^i, \kappa(D) = 10^{2j})$ 16 different types of matrices, giving the total of $64 \cdot 16 = 1024$ classes. The matrices are generated in four nested loops, the outer loop controls $\kappa(B) = \kappa(S)$. Hence, the matrices are divided in eight groups with fixed $\kappa(B)$.[5] Finally, we choose the row and the column dimensions, m and n, and the test procedure is ready.

Before we go over to the comparison of the new procedure SGEPVD[6] with SGESVD and SGESDD from LAPACK, we should point out that SGEPVD computes the SVD to higher accuracy and it also provides an estimate of the maximal relative error by computing an approximation of $\|B^\dagger\|$. Also, after computing the singular vectors, SGEPVD computes their norms using doubly accumulated dot products and explicitly normalizes those with Euclidean length deviating from unity by more than 3ε. Note that SGEPVD returns the singular vectors numerically orthogonal up to $m\varepsilon$ which, as in other algorithms dealing with numerical orthogonality, for large m and n may not look satisfactory in single precision ($\varepsilon \approx 10^{-7}$ and e.g. $m = n = 4000$).

We show only two out of many tests performed during code development. Or first test ran with $m = 1500$, $n = 1300$, and with one pseudo–random matrix in each class. Compare the maximal relative errors in computed singular values for all 1024 test cases, shown on Figure 4.1. It is clearly seen that the accuracy of SGEPVD depends on $\kappa(B)$, while the other two methods depend on $\kappa(A)$. Any SVD algorithm that starts with bidiagonalization is at risk to have error behaving like the SGESVD and SGESDD errors in this Figure. The best caption for this figure is the title of [8]. Note that SGESVD returns much better results than SGESDD. To our best knowledge, this fact is not mentioned elsewhere in the literature. Also, note the considerable upward bias in the relative errors of the bidiagonalization based procedures (cf. [25]).

The timings for this example are shown on Figure 4.2. We immediately note that the new Jacobi SVD algorithm is not at all much slower than the bidiagonalization based fast methods. In fact, it outperforms the QR algorithm and it is on average less than twice slower than the divide and conquer algorithm. The worst case performance for SGEPVD is on matrices with weak column scaling and singular spectrum composed of many tight clusters (examples above the 1.5 mark on Figure 4.2). In all

[5] This helps in the interpretation of the results of the experiments and explains the shapes of the graphs on the figures given here.

[6] PVD is the acronym for Principal Value Decomposition, an old name of the SVD.

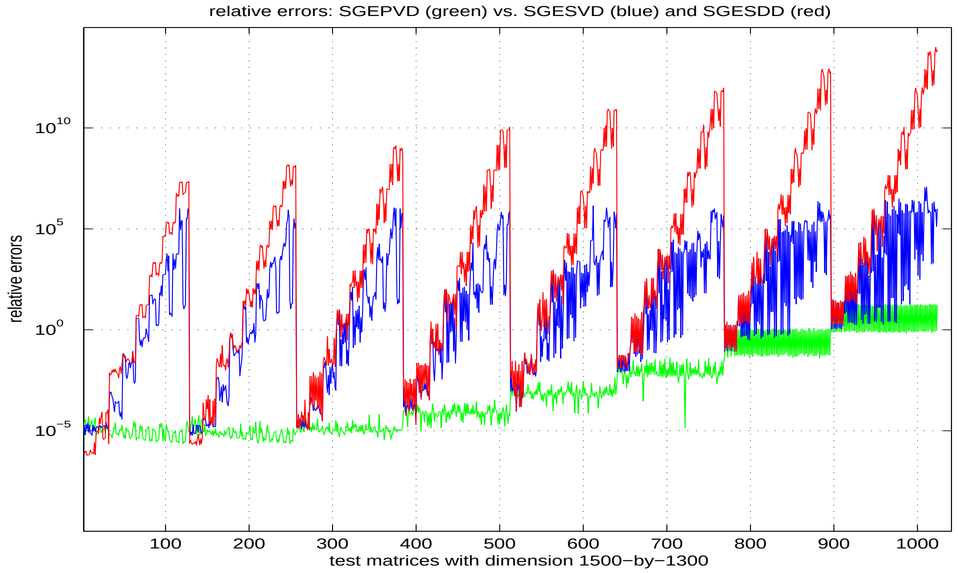

Fig. 4.1. Maximal relative errors for 1500×1300 matrices. The top curve (worst case) describes the accuracy of SGESDD. The middle curve represents the errors of SGESVD, and the lowest curve (smallest relative errors) belongs to SGEPVD.

other cases the time of SGEPVD is on average 1.5 times the time of SGESDD. Here, again, we stress the fact that the results obtained by SGEPVD enjoy much better numerical properties and that the time of SGEPVD includes computed error bound – better result and additional information are computed in reasonable time. Thus, for a fair comparison one should consider both Figure 4.1 and Figure 4.2.

In the second test we have 500×350 matrices, with two examples in each of 1024 classes. The results are on Figure 4.3 and Figure 4.4.

Remark 4.2. Note few outliers above mark 2 on Figures 4.2, 4.4. They correspond to matrices on which SGEPVD actually performed very well, with low number of rotations and swift convergence. (We have checked that by inspecting the details of those particular runs. In fact, on these matrices even the classical one–sided Jacobi, usually much slower, comes close to our new method.) However, since our threshold for perturbation used to trap denormals was set to $\zeta = \sqrt{\boldsymbol{\nu}} \approx 10^{-19}$, some of them were not captured and imperfect denormalized arithmetic caused considerable slowdown. We used this value of ζ to illustrate the problem. In practice the threshold can be set higher, e.g. if in those cases $\zeta = \boldsymbol{\varepsilon}/n$ the run time reflects the actual flop count. If *set to zero* underflow is in function, this kind of problems disappears.

Remark 4.3. We have noted that using double accumulated dot product in preparation of Jacobi rotation reduces the total number of rotations, especially in case of multiple or tightly clustered singular values. Unfortunately, in optimized BLAS libraries DSDOT performs poorly in comparison to SDOT and the saving in number of rotations does not reduce the total run time.

Remark 4.4. The results may vary on the same machine with different BLAS libraries. We note that with the GOTO BLAS all routines run faster (as compared to MKL library), but the relative speedup is smallest in our algorithm. This is because GOTO BLAS 3 outperforms MKL 6.1 BLAS 3, but GOTO BLAS 1 is no match

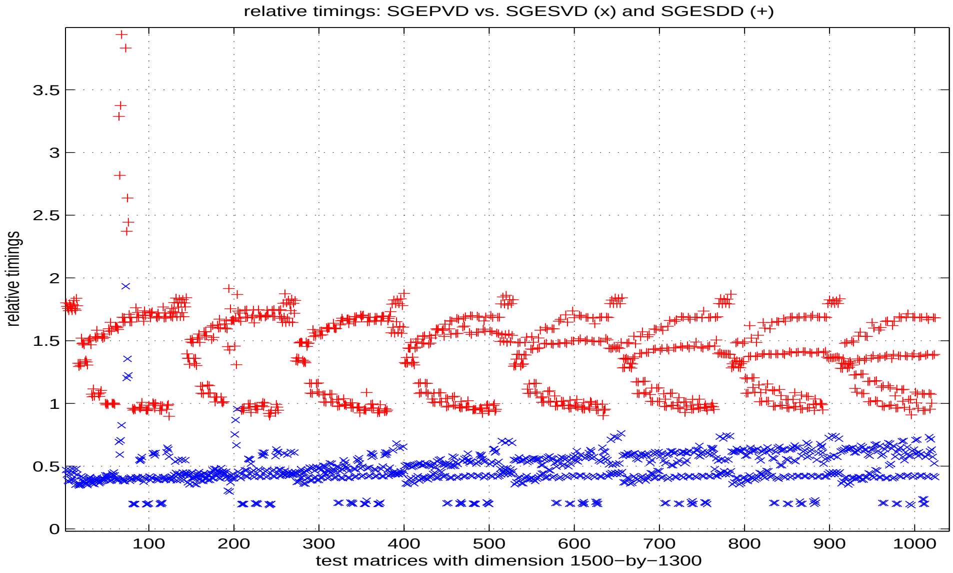

FIG. 4.2. Computing full SVD: Relative timings for 1500 × 1300 matrices on a Pentium 4 machine with Intel MKL 6.1 library. The crosses denote $time_{SGEPVD}/time_{SGESVD}$ and the pluses are $time_{SGEPVD}/time_{SGESDD}$.

for really efficient MKL 6.1 BLAS 1. Since our code still depends on BLAS 1 (dot products and plane rotations), switching to GOTO BLAS has mixed consequences. A hybrid of the two libraries would be much better for our algorithm.

4.3.2. Computing only Σ. We have established that the singular values are computed by the new algorithm as predicted by the theory – our variant of the Jacobi SVD complies with [8], [10]. In the previous section we showed that it computes the full SVD with efficiency comparable with fast bidiagonalization–based approaches.

If only the singular values are needed, reaching similar level of relative efficiency seems to be mission impossible. For, we need Jacobi–based algorithm that computes Σ in time comparable to the time needed to bidiagonalize the matrix!

We immediately note that in the case $m \gg n$ both methods start with the QR factorization which is the most expensive part of the computation. So, this is one example where Jacobi–based approach can be competitive. Further, our approach should have some advantage in case of matrices of low numerical rank, because we start with a rank revealing factorization. (Bidiagonalization is not rank revealing unless enhanced with pivoting which would make it more expensive.) Also, if only the standard absolute error bound is required, then Jacobi iterations can be controlled by a loosened stopping criterion, thus allowing satisfactory approximation with less computational effort.

The test matrices of the form $A = BD$ are generated as in §4.3.1, with 2048 examples in each test run. We show the results of two tests, in the first the matrices were 500×350 and in the second $m = 1000$, $n = 700$. The maximal measured relative errors $\mathbf{e}$ are not shown because they behave as shown in §4.3.1. We display the relative timings and compare SGEPVD with SGESDD (or SGESVD) and with the classical one–sided Jacobi SVD with de Rijk's pivoting (SGESVJ). See Figures 4.5, 4.6. (We note that SGEPVD computes Σ and also an estimate of the scaled condition number.)

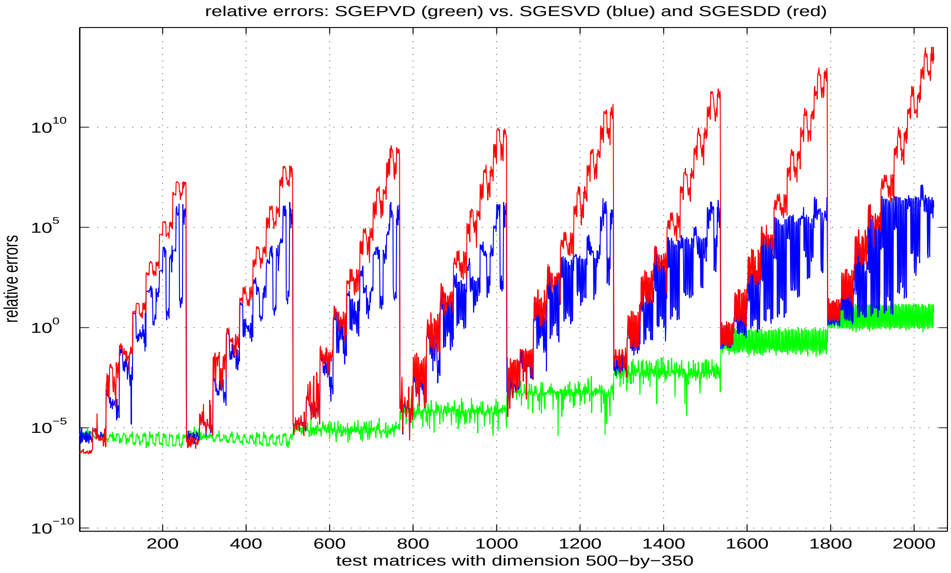

FIG. 4.3. Maximal relative errors for 500 × 350 matrices. The top curve (worst case) describes the accuracy of SGESDD. The middle curve represents the errors of SGESVD, and the lowest curve (smallest relative errors) belongs to SGEPVD.

5. Conclusion and future work. The most important message of the results in this report is that the question of the ultimate dense non–structured full SVD method is still open. Wanted is the most efficient algorithm capable of computing the SVD to optimal (numerically feasible) accuracy warranted by the data. In the first stage of our program we have shown that a new preconditioned Jacobi–type SVD algorithm can be competitive with fast bidiagonalization based methods without trading accuracy for speed. Moreover, if only the absolute accuracy suffices, our algorithm can take advantage of the rank revealing property of the preconditioner and thus run very efficiently. Because of that, our quest for ultimate dense non–structured SVD algorithm will follow the Jacobi idea.

Another approach could be to try to improve the accuracy of the bidiagonalization. An improvement suggested by Barlow preprocesses the bidiagonalization using QR factorization with complete pivoting, the bidiagonalization itself must use pivoting and in some cases Givens rotations instead of reflections. This in fact works very well in practice, and it is only partially understood, see Barlow [1]. Hence, it improves the accuracy of the bidiagonalization and in many cases it enables bidiagonalization based SVD methods to deliver the decomposition as accurately as the Jacobi SVD method. However, these modifications increase the total time of any bidiagonalization based algorithm by (i) the time of the QR factorization with complete pivoting; (ii) the time of multiplication by the orthogonal factor of this initial QR factorization, to get the left singular vectors; (iii) the additional time of pivoting and replacing reflections by rotations in the bidiagonalization. The question is how these improvements of accuracy impact the efficiency.

On the other hand, the most expensive part of our method are BLAS 1 Jacobi rotations which means that there is more potential for optimization. This and other relevant issues will be explored in the second stage of our program. One simple

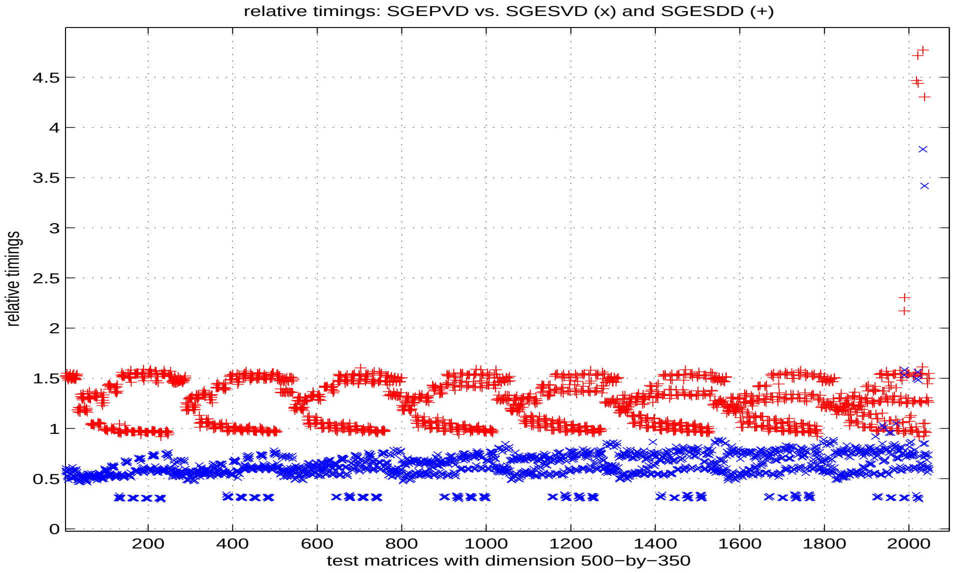

FIG. 4.4. Computing full SVD: Relative timings for 500 × 350 matrices on a Pentium 4 machine with Intel MKL 6.1 library. The crosses denote $time_{SGEPVD}/time_{SGESVD}$ and the pluses are $time_{SGEPVD}/time_{SGESDD}$.

improvement will be possible once optimized combinations of AXPY and DOT, and combination of two linked AXPY's operations are available in single calls. Nontrivial improvement will be obtained by using block rotations. We expect that the fast scaled block rotations designed by Hari [16] and currently under further development and implementation at the Department of Mathematics in Zagreb will fully exploit the potential of our approach. Other issues include new rank–revealing QR factorization and using shifts in the second QR (and third in some cases) factorization if only classical absolute error bounds are required.

The second stage will complete with the release of a high performance LAPACK–style software. The question is whether or not we will be able to reach the efficiency of SGESDD. Time will tell.

REFERENCES

[1] J. Barlow, *More accurate bidiagonal reduction for computing the singular value decomposition*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 761–798.

[2] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Num. Anal., 27 (1990), pp. 762–791.

[3] F. Bauer, *Software and software engineering*, SIAM Review, 15 (1973), pp. 469–480.

[4] N. N. Chan and Kim-Hung Li, *Diagonal elements and eigenvalues of a real symmetric matrix*, SIAM J. Matrix Anal. Appl., 91 (1983), pp. 562–566.

[5] P. I. Davies and N. J. Higham, *Numerically stable generation of correlation matrices and their factors*, BIT, 40 (2000), pp. 640–651.

[6] J. Demmel, *Underflow and the reliability of numerical software*, SIAM J. Sci. Stat. Comp., (1984), pp. 887–919.

[7] J. Demmel and X. Li, *Faster numerical algorithms via exception handling.* Technical report UT–CS–93–192 Department of Computer Science, University of Tennessee, Knoxville (LAPACK Working Note 59), 1993.

[8] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal.

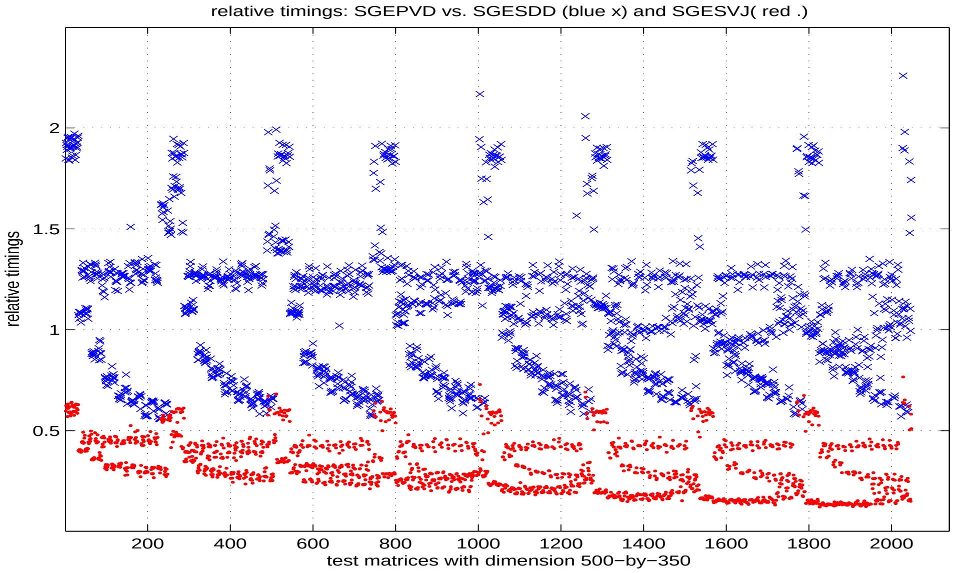

FIG. 4.5. Computing only Σ: Relative timings for 500×350 matrices on a Pentium 4 machine with Intel MKL 6.1 library. The blue crosses denote $time_{SGEPVD}/time_{SGESVD}$ and the red dots are $time_{SGEPVD}/time_{SGESVJ}$.

Appl., 13 (1992), pp. 1204–1245.

[9] J. J. Dongarra, J. J. Du Croz, I. Duff, and S. Hammarling, *A set of Level 3 Basic Linear Algebra Subprograms*, ACM Trans. Math. Soft., (1990), pp. 1–17.

[10] Z. Drmač, *Computing the Singular and the Generalized Singular Values*, PhD thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, 1994.

[11] ———, *Implementation of Jacobi rotations for accurate singular value computation in floating point arithmetic*, SIAM J. Sci. Comp., 18 (1997), pp. 1200–1222.

[12] Z. Drmač and V. Hari, *On the quadratic convergence of the J–symmetric Jacobi method*, Numer. Math., 64 (1993), pp. 147–180.

[13] Z. Drmač and K. Veselić, *New fast and accurate Jacobi SVD algorithm: I.*, tech. report, Department of Mathematics, University of Zagreb, Croatia, June 2005.

[14] H. H. Goldstine, H. H. Murray, and J. von Neumann, *The Jacobi method for real symmetric matrices*, J. Assoc. Comp. Mach., 6 (1959), pp. 59–96. (Also in J. von Neumann,*Collected Works*, vol. V, pages 573-610, Pergamon Press, New York,1973).

[15] V. Hari, *On sharp quadratic convergence bounds for the serial Jacobi methods*, Numer. Math., 60 (1991), pp. 375–406.

[16] ———, *Fast scaled block Jacobi rotations*, Computing, to appear, (2005).

[17] M. R. Hestenes, *Inversion of matrices by biorthogonalization and related results*, J. SIAM, 6 (1958), pp. 51–90.

[18] C. G. J. Jacobi, *Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen*, Crelle's Journal für reine und angew. Math., 30 (1846), pp. 51–95.

[19] W. Kahan, *The baleful effect of computer benchmarks upon applied mathematics, physics and chemistry*, tech. report, 1995.

[20] W. F. Mascarenhas, *On the convergence of the Jacobi method for arbitrary orderings*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1197–1209.

[21] J. Matejaš, *Quadratic convergence of scaled matrices in Jacobi method*, Numer. Math., 87 (2000), pp. 171–199.

[22] B. N. Parlett, *The Symmetric Eigenvalue Problem, Classics In Applied Mathematics 20*, SIAM, Philadelphia, PA, 1998.

[23] N. H. Rhee and V. Hari, *On the global and cubic convergence of a quasi–cyclic Jacobi method*, Numer. Math., 66 (1993), pp. 97–122.

[24] G. W. Stewart, *The efficient generation of random orthogonal matrices with an application*

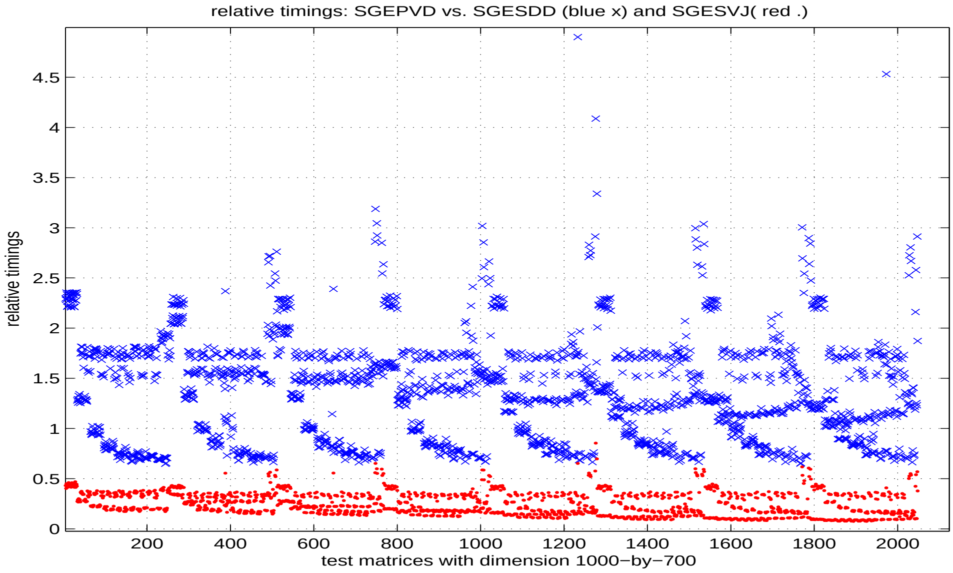

FIG. 4.6. Computing only Σ: Relative timings for 1000×700 matrices on a Pentium 4 machine with Intel MKL 6.1 library. The blue crosses denote $time_{SGEPVD}/time_{SGESVD}$ and the red dots are $time_{SGEPVD}/time_{SGESVJ}$.

to condition estimators, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

[25] ———, *Perturbation theory for the singular value decomposition*, Technical Report UMIACS–TR–90–124, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, 1990.