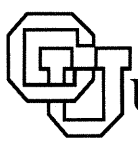
Accurate Computation of the Product Induced Singular Value Decomposition with Applications*

Zlatko Drmac

CU-CS-816-96



University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

* This research was supported by National Science Foundation grants ACS-9357812 and ASC-9625912, Department of Energy grant DE-FG03-94ER25215 and the Intel Corporation. Parts of this work were presented at the International Workshop on Accurate Eigensolving and Applications, July 11-17, 1996 Split, Croatia, and at the First Congress of the Croatian Mathematical Society, July 18-20, 1996 Zagreb, Croatia.

ANY OPINIONS, FINDINGS, AND CONCLUSIONS OR RECOMMENDATIONS EXPRESSED IN THIS PUBLICATION ARE THOSE OF THE AUTHOR(S) AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE AGENCIES NAMED IN THE ACKNOWLEDGMENTS SECTION.

ACCURATE COMPUTATION OF THE PRODUCT INDUCED SINGULAR VALUE DECOMPOSITION WITH APPLICATIONS*

ZLATKO DRMAČ†

November 19, 1996

Abstract. We present a new algorithm for floating–point computation of the singular value decomposition (SVD) of the product B^TC, where B and C are full row rank matrices. The algorithm replaces the pair (B,C) with an equivalent pair (B',C') and then it uses the Jacobi SVD algorithm to compute the SVD of the explicitly computed matrix B'^TC'. In this way, each nonzero singular value σ is approximated with some $\sigma+\delta\sigma$, where the relative error $|\delta\sigma|/\sigma$ is, up to a factor of the dimensions, of order $\varepsilon\{\min_{\Delta\in\mathcal{D}}\kappa_2(\Delta B)+\min_{\Delta\in\mathcal{D}}\kappa_2(\Delta C)\}$, where $\mathcal{D}$ denotes the set of diagonal nonsingular matrices, $\kappa_2(\cdot)$ denotes the spectral condition number and ε is the roundoff unit of floating–point arithmetic. The new algorithm is applied to the eigenvalue problem $HMx=\lambda x$ with symmetric positive definite H and M. It is shown that each eigenvalue λ is computed with high relative accuracy and that the relative error $|\delta\lambda|/\lambda$ of the computed approximation $\lambda+\delta\lambda$ is, up to factor of the dimension, of order $\varepsilon\{\min_{\Delta\in\mathcal{D}}\kappa_2(\Delta H\Delta)+\min_{\Delta\in\mathcal{D}}\kappa_2(\Delta M\Delta)\}$. The new algorithm can also be used for accurate SVD computation of a single matrix G that admits an accurate factorization $G=B^TC$.

Key words. contragredient transformation, eigenvalue problem, product induced singular value decomposition, relative accuracy, singular value decomposition, system balancing

AMS subject classifications. 65F15, 65F25, 65G05

1. Introduction. In this paper, we study floating–point computation of the singular value decomposition (SVD) of the product

$$A=B^TC,\quad B\in\mathbf{R}^{p\times m},\quad C\in\mathbf{R}^{p\times n},\quad \mathrm{rank}\,(B)=\mathrm{rank}\,(C)=p, \tag{1.1}$$

and floating–point solution of the eigenvalue problem

$$HMx=\lambda x,\quad H,M\in\mathbf{R}^{n\times n}\ \text{symmetric and positive definite.} \tag{1.2}$$

The singular value decomposition of the product of two matrices and the eigenvalue problem (1.2) arise in a variety of applications. For instance, consider the time invariant linear system

$$\dot{x}(t) = Ex(t)+Fu(t),\quad x(0)=x_0;\quad (x(t)\in\mathbf{R}^n,\ u(t)\in\mathbf{R}^m,\ E\ \text{stable}) \tag{1.3}$$

$$y(t) = Gx(t),\quad (y(t)\in\mathbf{R}^p). \tag{1.4}$$

By a change of state coordinates $x(t)=T\hat{x}(t)$, the reachability Gramian H and the observability Gramian M (at $t=\infty$) change to $\hat{H}=T^{-1}HT^{-\tau}$, $\hat{M}=T^\tau MT$, respectively. In designing reduced order model, it is of interest to find T that makes $\hat{H}$ and $\hat{M}$ diagonal, that is (cf. [27], [16])

$$T^{-1}HT^{-\tau}=T^\tau MT=\Sigma=\mathrm{diag}\,(\sigma_i). \tag{1.5}$$

The matrix T is called *contragredient* or *balancing* transformation, and it is the eigenvector matrix of HM, that is, $T^{-1}(HM)T=\Sigma^2$.

If $H=L_HL_H^\tau$, $M=L_ML_M^\tau$ are the Cholesky factorizations of H, M, respectively, then the eigenvalue problem (1.2) is equivalent to the singular value problem of A in (1.1) with $B=L_H$, $C=L_M$. This follows from the similarity of $HM=(L_HL_H^\tau)(L_ML_M^\tau)$ and $(L_M^\tau L_H)(L_H^\tau L_M)$. Hence, if $V\Sigma U^\tau=L_H^\tau L_M$ is the SVD of $L_H^\tau L_M$, then $T=L_HV\Sigma^{-1/2}$ satisfies (1.5).

---

* Technical report CU-CS-816-96, Department of Computer Science, University of Colorado at Boulder.

† Department of Computer Science, University of Colorado, Boulder CO 80309–0430. (zlatko@cs.colorado.edu)
This research was supported by National Science Foundation grants ACS–9357812 and ASC–9625912, Department of Energy grant DE–FG03–94ER25215, and the Intel Corporation. Parts of this work were presented at the International Workshop on Accurate Eigensolving and Applications, July 11–17, 1996 Split, Croatia, and at the First Congress of the Croatian Mathematical Society, July 18–20, 1996 Zagreb, Croatia.

Fernando and Hammarling [16] use the SVD of $A = B^T C$ to define the *product induced singular value decomposition* (HSVD) of the pair (B, C). They prove that there exist orthogonal matrices U, V and W, nonsingular triangular matrix R, and quasi–diagonal matrices Δ and Γ such that

$$B^T = U[\Gamma R, 0]W^T, \quad C^T = V[\Delta R^{-T}, 0]W^T. \tag{1.6}$$

Our goal is to compute the singular values of the product $B^T C$ in relation (1.1), and the eigenvalues of the positive definite pencil $HM - \lambda I$ with high relative accuracy whenever numerically feasible. We consider high relative accuracy numerically feasible if initial uncertainties in the matrices B and C, H and M induce small relative uncertainties in the singular values and the eigenvalues, respectively. In that case we also say that the singular values and the eigenvalues are well–determined by the data.

A desirable property of an algorithm is that it approximates the well–determined singular values and the eigenvalues with high relative accuracy independent of their magnitudes. This is important because in many applications it is of interest to compute even the smallest singular values of a matrix with certain relative accuracy. For example, Laub *et al* [27] emphasize the importance of computing small singular values accurately when a linear system is near–uncontrollable and near–unobservable.

The SVD computation of $A = B^T C$ in floating–point arithmetic is not straightforward because the explicit computation of the product $B^T C$ may cause large perturbations of the smallest singular values. The following example illustrates this. Let $\xi \neq 0$ and let

$$B^T = \begin{bmatrix} 1 & \xi \\ -1 & \xi \end{bmatrix}, \; C = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \; B^T C = \frac{1}{\sqrt{2}}\begin{bmatrix} 1-\xi & 1+\xi \\ -1-\xi & -1+\xi \end{bmatrix}. \tag{1.7}$$

C is orthogonal and the columns of B^T are mutually orthogonal. However, if $|\xi|$ is smaller than the roundoff unit ε, or if $|\xi| > 1/\varepsilon$, the floating–point product $B^T C$ is exactly singular matrix. Hence, the smaller of the two exact singular values, $\sigma_1 = \sqrt{2}$, $\sigma_2 = \sqrt{2}|\xi|$ is perturbed to zero.

To avoid the computation of the matrix $A = B^T C$, Heath *et al* [22] use a Kogbetliantz type algorithm and transform B and C separately. This algorithm is theoretically equivalent to the Kogbetliantz SVD algorithm applied to $A = B^T C$. An elegant implementation of the algorithm uses the QR factorization to transform B^T and C to upper triangular form, and it preserves the triangularity during the subsequent iterative phase, using suitable plane rotations. The convergence in the iterative phase may be slow, especially in the case of multiple or clustered singular values. The algorithm is backward stable and numerical experiments in [22] show that it usually has better accuracy properties than the SVD computation of the matrix $A = B^T C$.

In some cases, however, the algorithm from [22] produces similar errors as the computation of the product $B^T C$. To illustrate this, we use B and C from relation (1.7). In the phase of replacing (B, C) by a pair (B_1, C_1) of triangular matrices (cf. [22, § 5, *(i)*]), the matrix B_1 is computed by the QR factorization of the matrix $B^T Q$, where $C = QC_1$ is the QR factorization of C. Since in this example $C = Q$, the algorithm computes the QR factorization of the numericaly singular matrix $B^T C$.

In this paper, we propose a new approach. We replace the pair (B, C) with a new pair (B', C') such that $B^T C$ and $B'^T C'$ have the same singular values, and such that the SVD of the explicitly computed floating–point matrix $F = \boldsymbol{fl}(B'^T C')$ can be computed without introducing large perturbations of the singular values of $B^T C$. In the transition from (B, C) to F, we use row scalings, the QR factorization with pivoting, and the matrix product. The SVD of the matrix F is computed by the Jacobi SVD algorithm. We show that the computed singular values $\sigma_1 + \delta\sigma_1 \geq \cdots \geq \sigma_p + \delta\sigma_p$ approximate the exact singular values $\sigma_1 \geq \cdots \geq \sigma_p > 0$ with relative error bounded by

$$\max_{1 \leq i \leq p} \frac{|\delta\sigma_i|}{\sigma_i} \leq f(m, n, p)\varepsilon(K(C)\|B_r^\dagger\|_2 + \|C_r^\dagger\|_2), \tag{1.8}$$

where $B_r = \text{diag}\,(\|B^T e_i\|_2^{-1})B$, $C_r = \text{diag}\,(\|C^T e_i\|_2^{-1})C$, † denotes the generalized inverse, and $\|\cdot\|_2$ denotes the Euclidean vector norm and the corresponding induced operator norm. Furthermore, $K(C)$ is a constant bounded by $\|C_r^\dagger\|_2$ and $f(\cdot)$ is a modestly growing function of the matrix dimensions. An important feature that follows from relation (1.8) is that the new algorithm has the same accuracy properties for all pairs $\{(D_1 B, D_2 C),\ D_1, D_2$ diagonal nonsingular matrices$\}$.

We apply the new algorithm to the eigenvalue problem (1.2), using the Cholesky factors of H and M in places of B and C, respectively. The computed approximations $\lambda_1+\delta\lambda_1 \geq \cdots \geq \lambda_n+\delta\lambda_n$, of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ of HM satisfy

$$\max_{1\leq i\leq n} \frac{|\delta\lambda_i|}{\lambda_i} \leq g(n)\varepsilon(\|H_s^{-1}\|_2 + \|M_s^{-1}\|_2), \tag{1.9}$$

where $H_s = \text{diag}\,(H_{ii})^{-1/2} H \text{diag}\,(H_{ii})^{-1/2}$, $M_s = \text{diag}\,(M_{ii})^{-1/2} M \text{diag}\,(M_{ii})^{-1/2}$, and $g(n)$ is a modestly growing function of the matrix dimension. Hence, the new algorithm is accurate if the values of $\|H_s^{-1}\|_2$ and $\|M_s^{-1}\|_2$ are moderate. Using the perturbation estimates of Demmel and Veselić [8] and Veselić and Slapničar [36], we show that moderate $\|H_s^{-1}\|_2$ and $\|M_s^{-1}\|_2$ are also necessary for the computation with high relative accuracy. Furthermore, we show that the computed eigenvalues $\lambda_i + \delta\lambda_i$, $1 \leq i \leq n$, are the exact eigenvalues of the pencil $(H+\delta H)(M+\delta M) - \lambda I$, with symmetric perturbations δH, δM such that $|\delta H_{ij}|/\sqrt{H_{ii}H_{jj}}$, $|\delta M_{ij}|/\sqrt{M_{ii}M_{jj}}$ are small for all i, j. The only condition for the high accuracy is that the floating–point Cholesky factorizations of H and M are guaranteed to complete without breakdown.

The paper is organized as follows. In § 2, we give two illustrative low dimension examples. In § 3, we analyze accurate computation of the SVD of the product $B^T C$, and its application to the computation of the ordinary SVD. In § 4, we give detailed analysis of a new algorithm for solving the eigenvalue problem (1.2) with high relative accuracy. We derive relative error bounds for eigenvalues and individual entries of the eigenvector matrix. In § 5, we present results of extensive numerical testing of the new algorithm.

2. Two illustrative examples. Before we start with the presentation of the new algorithm, we analyze two simple but very instructive examples. In the first example, we illustrate the sensitivity of the singular values of the product of two matrices and we show that it is not always possible to compute the singular values with high relative accuracy. In the second example, we use two–by–two matrices to introduce a simple technique that is used in the new algorithm.

EXAMPLE 2.1. Let $B = [y] \in \mathbf{R}^{m\times 1}$, $C = [x] \in \mathbf{R}^{m\times 1}$. The only singular value of $B^T C$ is the inner product $(x, y) \equiv y^T x$. Thus, the perturbation analysis reduces to the analysis of the inner product in presence of perturbations. The sensitivity of the inner product to relative norm–wise perturbations is measured with the *relative asymptotic condition number*

$$\rho(x,y) \equiv \liminf_{\varepsilon\to 0}\{\xi \ :\ |(x+\delta x, y+\delta y) - (x,y)| \leq \xi\varepsilon|(x,y)|,\ \|\delta x\|_2 \leq \varepsilon\|x\|_2,\ \ \|\delta y\|_2 \leq \varepsilon\|y\|_2\},$$

introduced in [32]. Straightforward calculation yields (cf. [32])

$$\begin{aligned}
\rho(x,y) &= \limsup_{\varepsilon\to 0}\left\{\frac{|(x,\delta y)+(\delta x,y)+(\delta x,\delta y)|}{\varepsilon|(x,y)|},\ \ \|\delta x\|_2 \leq \varepsilon\|x\|_2,\ \ \|\delta y\|_2 \leq \varepsilon\|y\|_2\right\} \\
&= \limsup_{\varepsilon\to 0}\left\{\left|\frac{\|x\|_2\|\delta y\|_2\cos\angle(x,\delta y)}{\varepsilon|(x,y)|} + \frac{\|\delta x\|_2\|y\|_2\cos\angle(\delta x,y)}{\varepsilon|(x,y)|} + \right.\right. \\
&\quad \left.\left. + \frac{\|\delta x\|_2\|\delta y\|_2\cos\angle(\delta x,\delta y)}{\varepsilon|(x,y)|}\right|,\ \ \|\delta x\|_2 \leq \varepsilon\|x\|_2,\ \ \|\delta y\|_2 \leq \varepsilon\|y\|_2\right\} \\
&= \frac{2\|x\|_2\|y\|_2}{|(x,y)|} = \frac{2}{|\cos\angle(x,y)|}.
\end{aligned}$$

The relative asymptotic condition number with respect to element–wise perturbations $|\delta x_i| \leq \varepsilon|x_i|$, $|\delta y_i| \leq \varepsilon|y_i|$, $1 \leq i \leq m$ can be defined as

$$\rho_e(x,y) \equiv \liminf_{\varepsilon\to 0}\{\xi \ :\ |(x+\delta x, y+\delta y)-(x,y)| \leq \xi\varepsilon|(x,y)|,\ |\delta x_i| \leq \varepsilon|x_i|,\ \ |\delta y_i| \leq \varepsilon|y_i|,\ \ 1 \leq i \leq m\}.$$

An easy calculation yields

$$\rho_e(x,y) = 2\frac{(|x|,|y|)}{|(x,y)|} \leq \rho(x,y), \quad \text{where } |x|_i = |x_i|,\ |y|_i = |y_i|,\ 1 \leq i \leq m.$$

Hence, if the ranges of B and C are orthogonal up to working precision, we generally cannot expect to reveal any correct digit of the singular value of the product $B^T C$.

REMARK 2.2. Note that the singular values of BC^T in Example 2.1 ($m-1$ zeros and $\|x\|_2\|y\|_2$) are perfectly well determined by B and C. This is easily seen because no perturbation δx, δy can change the zero singular values and because $\rho(x,x) = \rho_e(x,x) = 2$ for all $x \in \mathbf{R}^m$.

EXAMPLE 2.3. This example is based on the well–known normal equation example in the least squares computation (cf. [3], [27]). Let

$$B = \begin{bmatrix} 0 & \xi \\ 1 & 1 \end{bmatrix}, \quad C = B, \quad A = B^T C = \begin{bmatrix} 0 & 1 \\ \xi & 1 \end{bmatrix}\begin{bmatrix} 0 & \xi \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1+\xi^2 \end{bmatrix}. \tag{2.10}$$

If $\xi^2 < \varepsilon$, the floating–point approximation $\tilde{A}$ of A is exactly singular. Hence, even an exact SVD computation of $\tilde{A}$ provides no useful information about the minimal singular value of A. This difficulty with the explicit computation of the matrix A can be avoided if we slightly change the approach to the problem. We use the fact that the singular values of the product $B^T C$ are invariants of the transformation

$$(B,C) \longmapsto (B',C') = (TBU, T^{-T}CV),\ U,\ V \text{ orthogonal},\ T \text{ nonsingular}. \tag{2.11}$$

Our goal is to find a pair (B',C') which is suitable for accurate singular value computation of the explicitly computed product $B'^T C'$. For example, define $D_B = \operatorname{diag}(1/\xi, 1)$ and compute the LQ factorization with row pivoting of $D_B^{-T} C$,

$$C' \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & \xi^2 \\ 1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ \frac{\xi^2}{\sqrt{2}} & \frac{\xi^2}{\sqrt{2}} \end{bmatrix}.$$

Then $B' \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_B B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and C' are obtained by a transformation as in (2.11), and

$$A' = B'^T C' = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{2} & 0 \\ \frac{\xi^2}{\sqrt{2}} & \frac{\xi^2}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ \sqrt{2}+\frac{\xi^2}{\sqrt{2}} & \frac{\xi^2}{\sqrt{2}} \end{bmatrix}. \tag{2.12}$$

From the perturbation theory of Demmel and Veselić [8], it follows that the singular values of the floating–point approximation $\tilde{A}'$ of A' approximate the singular values of A' to full machine precision.

3. The SVD of the product $B^T C$. In this section, we present the new algorithm for the computation of the SVD of the product $A = B^T C$, where $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{p\times n}$ are full row rank matrices. In this case, A has p nonzero singular values, and the remaining $\min\{m,n\} - p$ zero singular values can be deflated using the following procedure from [22]. Let

$$B^T = Q_B\begin{bmatrix} R_B \\ \mathbf{O} \end{bmatrix}, \quad C = [R_C, \mathbf{O}]Q_C$$

be the QR and the RQ factorization of B^T and C, respectively. Then

$$B^T C = Q_B\begin{bmatrix} R_B R_C & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} Q_C,$$

and the problem reduces to the computation of the SVD of the product $R_B R_C$. Heath *et al* [22] choose the RQ factorization of C to ensure that R_C is an upper triangular matrix. They compute

the SVD of the product $R_B R_C$ by the Kogbentliantz–type algorithm that iteratively transforms R_B and R_C.

Here we show that, in certain well–conditioned cases, we can safely replace the pair (B, C) by an explicitly computed single matrix. We start with § 3.1, where we give perturbation estimates for the singular values of the product $B^\tau C$ of full row rank matrices. In § 3.2, we describe the new algorithm. In § 3.3, we prove that the new algorithm computes the singular values with high relative accuracy. In 3.4, we derive sharp backward error estimate and use it to obtain an error bound for the computed singular values. In § 3.5, we analyze the errors in the singular vectors. In § 3.6, we show that in certain cases the new algorithm can be used for accurate computation of the SVD of a single matrix.

3.1. Sensitivity of the singular values. How the singular values of the product $B^\tau C$ change if B and C are subject to small perturbations δB, δC? If we assume that B and C are full row rank matrices and that δB and δC are sufficiently small, then $B + \delta B$ and $C + \delta C$ are also full row rank matrices and we can consider only the perturbations $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p$ of the p nonzero singular values $\sigma_1 \geq \cdots \geq \sigma_p$ of $B^\tau C$.

An application of the variational characterization to the singular values of the matrix

$$(B + \delta B)^\tau (C + \delta C) = B^\tau C + (\delta B)^\tau C + B^\tau \delta C + (\delta B)^\tau \delta C \tag{3.13}$$

does not always provide satisfactory error estimate because the bound

$$|\tilde{\sigma}_i - \sigma_i| \leq \|B\|_2 \|C\|_2 \left(\frac{\|\delta B\|_2}{\|B\|_2} + \frac{\|\delta C\|_2}{\|C\|_2} + \frac{\|\delta B\|_2}{\|B\|_2} \frac{\|\delta C\|_2}{\|C\|_2} \right), \quad 1 \leq i \leq p, \tag{3.14}$$

may be too pessimistic for $\tilde{\sigma}_i$ if $|\sigma_i| \ll \|B\|_2 \|C\|_2$. Instead of (3.13), we use another representation of the product $(B + \delta B)^\tau (C + \delta C)$, and then apply the variational characterization.

THEOREM 3.1. *Let $B \in \mathbf{R}^{p \times m}$, $C \in \mathbf{R}^{p \times n}$ have full row rank, and let $\tilde{B} = B + \delta B$, $\tilde{C} = C + \delta C$ be perturbed matrices such that $\|B^\dagger \delta B\|_2 < 1$, $\|C^\dagger \delta C\|_2 < 1$. If $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_{\min\{m,n\}}$ are the singular values of $B^\tau C$ and $\tilde{B}^\tau \tilde{C}$, respectively, then, for all i, either $\sigma_i = \tilde{\sigma}_i = 0$, or*

$$\frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq \|B^\dagger \delta B\|_2 + \|C^\dagger \delta C\|_2 + \|B^\dagger \delta B\|_2 \|C^\dagger \delta C\|_2. \tag{3.15}$$

Proof. Note that we can write $\tilde{B}^\tau \tilde{C} = (I + B^\dagger \delta B)^\tau B^\tau C (I + C^\dagger \delta C)$, and that for any nonzero vector x and $y = (I + C^\dagger \delta C)x$ it holds that $y \neq \mathbf{0}$ and

$$\begin{aligned} \frac{\|\tilde{B}^\tau \tilde{C} x\|_2}{\|x\|_2} &\leq (1 + \|B^\dagger \delta B\|_2)(1 + \|C^\dagger \delta C\|_2) \frac{\|B^\tau C y\|_2}{\|y\|_2}, \\ \frac{\|\tilde{B}^\tau \tilde{C} x\|_2}{\|x\|_2} &\geq (1 - \|B^\dagger \delta B\|_2)(1 - \|C^\dagger \delta C\|_2) \frac{\|B^\tau C y\|_2}{\|y\|_2}. \end{aligned}$$

Now an application of the variational characterization implies relation (3.15). □

REMARK 3.1. The proof of Theorem 3.1 is based on [18, Lemma 6.4 and Corollary 6.1] and on [25, Problem 12 in § 3.3]. Similar technique is used in [10] for perturbation estimates for the generalized singular values and in [28], [15] for development of an elegant theory of relative eigenvalue and singular value perturbations.

COROLLARY 3.2. *Let in Theorem 3.1, $\|(\delta B)^\tau e_i\|_2 \leq \epsilon_B \|B^\tau e_i\|_2$, $\|(\delta C)^\tau e_i\|_2 \leq \epsilon_C \|C^\tau e_i\|_2$, $1 \leq i \leq p$, and let*

$$B = \Delta_B B_r, \quad C = \Delta_C C_r, \quad \text{where } \Delta_B = \operatorname{diag}(\|B^\tau e_i\|_2), \quad \Delta_C = \operatorname{diag}(\|C^\tau e_i\|_2). \tag{3.16}$$

Then, for all i, either $\sigma_i = \tilde{\sigma}_i = 0$, or

$$\frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq \sqrt{p}(\epsilon_B \|B_r^\dagger\|_2 + \epsilon_C \|C_r^\dagger\|_2 + \sqrt{p}\epsilon_B\epsilon_C \|B_r^\dagger\|_2 \|C_r^\dagger\|_2). \tag{3.17}$$

Proof. Relation (3.17) follows from Theorem 3.1 since $B^\dagger \delta B = B^\tau (BB^\tau)^{-1}\delta B = B_r^\dagger \Delta_B^{-1}\delta B$. □

3.2. The algorithm. We now describe the new algorithm. Our goal is to achieve the relative accuracy from Corollary 3.2, that is to compute accurate approximations of the singular values of any product $B^\tau C$ in which $\|B_r^\dagger\|_2$ and $\|C_r^\dagger\|_2$ are moderate. The basic idea of the new algorithm is illustrated in Example 2.3.

ALGORITHM 3.3.

Input $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{p\times n}$, $\mathrm{rank}\,(B) = \mathrm{rank}\,(C) = p$.

Step 1 Compute $\Delta_B = \mathrm{diag}\,(\|B^\tau e_i\|_2)$ and $B_r = \Delta_B^{-1}B$, $C_1 = \Delta_B C$.

Step 2 Compute the QR factorization with column pivoting of C_1^τ,

$$C_1^\tau \Pi = Q \begin{bmatrix} R \\ \mathbf{O}_{n-p,p} \end{bmatrix}, \quad R \text{ upper triangular, } Q \text{ orthogonal.}$$

Step 3 Compute the matrix $F = B_r^\tau \Pi R^\tau$, using the standard matrix multiply algorithm.

Step 4 Use the Jacobi SVD algorithm, implemented as in [8], [9], to compute the SVD of F,

$$\begin{bmatrix} \Sigma \\ \mathbf{O}_{m-p,p} \end{bmatrix} = V^\tau F U.$$

Output The SVD of $B^\tau C$ is

$$\begin{bmatrix} \Sigma \oplus \mathbf{O} \\ \mathbf{O} \end{bmatrix} = V^\tau (B^\tau C)(Q(U \oplus I_{n-p})).$$

3.3. Error bound for singular values. Consider now the floating–point error analysis of Algorithm 3.3. We use the standard model of floating–point arithmetic,

$$\boldsymbol{fl}(a \odot b) = (a \odot b)(1+\xi), \quad \boldsymbol{fl}(\sqrt{c}) = \sqrt{c}(1+\zeta), \quad |\xi|, |\zeta| \leq \boldsymbol{\epsilon}, \tag{3.18}$$

where a, b and c are floating–point numbers, $\odot$ denotes any of the four elementary operations $+$, $-$, $\cdot$ and $\div$, and $\boldsymbol{\epsilon}$ is the round–off unit.
Let $\tilde{\Delta}_B$, $\tilde{B}_r$ and $\tilde{C}_1$ be the floating–point approximations of Δ_B, B_r and C_1, respectively. Then

$$\tilde{\Delta}_B = (I + \Psi)\Delta_B, \quad \Psi = \mathrm{diag}\,(\psi_i), \quad |\psi_i| \leq \boldsymbol{\epsilon}_{\ell_2}(m) = (1+\boldsymbol{\epsilon})^{(m+2)/2} - 1, \tag{3.19}$$

and

$$\tilde{B}_r = \Delta_B^{-1}(I+\Psi)^{-1}(B + \delta B_e), \quad |\delta B_e| \leq \boldsymbol{\epsilon}|B|, \tag{3.20}$$

$$\tilde{C}_1 = \Delta_B(I+\Psi)(C + \delta C_e), \quad |\delta C_e| \leq \boldsymbol{\epsilon}|C|. \tag{3.21}$$

Since $\tilde{B}_r^\tau \tilde{C}_1 = (B + \delta B_e)^\tau (C + \delta C_e)$, the only singular value perturbation in Step 1 is caused by small element–wise rounding errors.

In Step 2, we assume that the QR factorization is computed using Givens rotations. Using an idea of Gentleman [17], we can derive rather sharp backward error bound for each matrix column. More precisely, we have the following proposition. (Cf. [38], [17], [2], [10], [24].)

PROPOSITION 3.4. *Let $X \in \mathbf{R}^{m\times n}$, $m \geq n$, and let the QR factorization of X be computed by a sequence of Givens rotations in some prescribed order. Let all rotations be divided into $\wp$ sets,*

where each set contains rotations that can be applied simultaneously to different pairs of matrix rows. If the computation is performed in floating-point arithmetic, and if $\tilde{R}$ is the computed triangular factor, then there exist a backward error δX and an orthogonal matrix Q' such that

$$X+\delta X = Q'\begin{bmatrix}\tilde{R}\\ \mathbf{O}\end{bmatrix}, \text{ where } \|\delta X e_i\|_2 \le \varepsilon_{QR}(m,n)\|Xe_i\|_2,\ 1\le i\le n,\ \ \varepsilon_{QR}(m,n)\le((1+6\varepsilon)^{\wp}-1).$$

For the usual column-wise ordering of Givens rotations we have $\wp = mn - 3$. For some more sophisticated strategies as, for example, in [31], for large $m \gg n$ it holds that $\wp \approx \log_2 m + (n-1)\log_2\log_2 m$.

In Step 3, we use the standard matrix multiply algorithm. In that case, the floating point product Z of an $m\times n$ matrix X and an $n\times p$ matrix Y satisfies (cf. [19])

$$Z = XY + E,\quad |E| \le \varepsilon_{MM}(n)|X|\cdot|Y|,\quad 0\le\varepsilon_{MM}(n)\le(1+\varepsilon)^{n+1}-1, \tag{3.22}$$

where the absolute value and the inequality are taken element-wise. Using double precision accumulation, the bound for $\varepsilon_{MM}(n)$ can be reduced to $O(1)\varepsilon$ for all $n\le 1/\varepsilon$.
Using Proposition 3.4 and relation (3.22), we can analyze the product $\tilde{B}_r^T\Pi\tilde{R}^T$ in Step 3 of Algorithm 3.3. We first show that, under the assumption that C_r is well-conditioned, the computation of the matrix F in Step 3 is backward stable.

PROPOSITION 3.5. *Let $\tilde{R}$ be the computed triangular factor of the matrix $\tilde{C}_1$ from relation (3.21), and let $\eta_C = \varepsilon_{QR}(n,p)(1+\varepsilon)+\varepsilon$, where $\varepsilon_{QR}(n,p)$ is defined as in Proposition 3.4. If $\sqrt{p}\eta_C\|C_r^{\dagger}\|_2 < 1$, and if $\tilde{F}$ is the computed approximation of the product $\tilde{B}_r^T\Pi\tilde{R}^T$ in Step 3 of Algorithm 3.3, then there exist an orthogonal matrix Q_C and perturbations δB, δC such that*

$$[\tilde{F}, \mathbf{O}_{m,n-p}] = (B+\delta B)^T(C+\delta C)Q_C \tag{3.23}$$

and such that, for all i,

$$\begin{aligned}\|(\delta B)^T e_i\|_2 &\le \eta_B\|B^T e_i\|_2,\quad \eta_B = \varepsilon_{MM}(p)\frac{1+\varepsilon_{\ell_2(m)}}{1-\varepsilon_{\ell_2(m)}}(1+\varepsilon)\|\,|\tilde{R}^{-1}|\cdot|\tilde{R}|\,\|_\infty+\varepsilon,\\ \|(\delta C)^T e_i\|_2 &\le \eta_C\|C^T e_i\|_2,\end{aligned}$$

where $\|\cdot\|_\infty$ is the matrix norm induced by the ℓ_∞ vector norm.

Proof. From Proposition 3.4, it follows that for some orthogonal matrix Q_C and some perturbation $(\delta\tilde{C}_1)^T$ it holds that

$$(\tilde{C}_1^T + (\delta\tilde{C}_1)^T)\Pi = Q_C\begin{bmatrix}\tilde{R}\\ \mathbf{O}\end{bmatrix},\ \text{where } \|(\delta\tilde{C}_1)^T e_i\|_2 \le \varepsilon_{QR}(n,p)\|\tilde{C}_1^T e_i\|_2\ \ 1\le i\le p.$$

Note that $\tilde{R}$ is nonsingular. Using (3.22), we conclude that

$$\tilde{F} = \tilde{B}_r^T\Pi\tilde{R}^T + \mathcal{E},\quad |\mathcal{E}|\le\varepsilon_{MM}(p)|\tilde{B}_r^T|\cdot\Pi\cdot|\tilde{R}^T|. \tag{3.24}$$

Hence,

$$[\tilde{F}, \mathbf{O}_{m,n-p}] = (\tilde{B}_r^T + \mathcal{E}\tilde{R}^{-T}\Pi^T)(\tilde{C}_1+\delta\tilde{C}_1)Q_C. \tag{3.25}$$

From relation (3.24), we have, for all i,

$$\|\mathcal{E}\tilde{R}^{-T}\Pi^T e_i\|_2 \le \varepsilon_{MM}(p)\max_{1\le k\le p}\|\tilde{B}_r^T e_k\|_2\|\Pi\cdot|\tilde{R}^T|\cdot|\tilde{R}^{-T}|\cdot\Pi^T e_i\,\|_1,$$

where $\|\cdot\|_1$ is the ℓ_1 vector norm. Inserting $\tilde{\Delta}_B\tilde{\Delta}_B^{-1}$ in relation (3.25) and using (3.20), (3.21) we obtain

$$[\tilde{F}, \mathbf{O}_{m,n-p}] = (B^T + (\delta B_e)^T + \mathcal{E}\tilde{R}^{-T}\Pi^T\tilde{\Delta}_B)(C+\delta C_e+\tilde{\Delta}_B^{-1}\delta\tilde{C}_1)Q_C. \tag{3.26}$$

Finally, note that, for all i,

$$\|\tilde{B}_r^\tau e_i\|_2 \leq \frac{1+\varepsilon}{1-\varepsilon_{\ell_2}(m)}, \quad \|(\delta\tilde{C}_1)^\tau \tilde{\Delta}_B^{-1} e_i\|_2 \leq \varepsilon_{QR}(n,p)(1+\varepsilon)\|C^\tau e_i\|_2.$$

□

REMARK 3.6. Note that an estimate similar to Proposition 3.5 holds if the computed matrix $\tilde{R}$ is such that $\tilde{R}^\tau = [\tilde{L}, \mathbf{O}_{p-r}]$, where $\tilde{L}$ is lower trapezoidal $p \times r$ matrix with $\mathrm{rank}\,(L) = r < p$. In that case, we can replace $\tilde{R}^\tau$ with $\tilde{L}$, and the computed matrix in Step 3 of Algorithm 3.3 can be written as $[\tilde{F}', \mathbf{O}_{m,n-r}] = (B+\delta B)^\tau (C+\delta C) Q_C$. Furthermore, the factor $\| \, |\tilde{R}^{-1}| \cdot |\tilde{R}| \, \|_\infty$ in the definition of η_B is replaced with $\| \, |\tilde{L}| \cdot |\tilde{L}^\dagger| \, \|_1$. (Here $\|\cdot\|_1$ denotes the operator norm induced by the ℓ_1 vector norm.) Note, however, that in that case we generally have no assurance of the relative accuracy of the computed singular values.

COROLLARY 3.7. *Let the assumptions of Proposition 3.5 hold. Furthermore, let $\sigma_1 \geq \cdots \geq \sigma_p$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p$ be the singular values of $B^\tau C$ and $\tilde{F}$, respectively. If, in addition, $\sqrt{p}\eta_B \|B_r^\dagger\|_2 < 1$, then*

$$\max_{1\leq i\leq p} \frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq \sqrt{p}(\eta_B \|B_r^\dagger\|_2 + \eta_C \|C_r^\dagger\|_2 + \sqrt{p}\eta_B\eta_C \|B_r^\dagger\|_2 \|C_r^\dagger\|_2). \tag{3.27}$$

In the practice, the column pivoting in the QR factorization in Step 2 ensures that $\| \, |\tilde{R}^{-1}| \cdot |\tilde{R}| \, \|_\infty$ always remains bounded by a function of the dimension p. We pivot so that (cf. [4])

$$\tilde{R}_{ii}^2 \geq \sum_{k=i}^{j} \tilde{R}_{kj}^2, \quad 1 \leq i \leq j \leq p. \tag{3.28}$$

In that case, the following proposition (cf. [10], [11]) shows that $\| \, |\tilde{R}^{-1}| \cdot |\tilde{R}| \, \|_\infty$ is moderate if $\|C_r^\dagger\|_2$ is such.

PROPOSITION 3.8. *Let the pivoting in Step 2 of Algorithm 3.3 be such that relation (3.28) holds, and let $\tilde{R} = \mathrm{diag}\,(\|\tilde{R}^\tau e_i\|_2)\tilde{R}_r = \tilde{R}_c \mathrm{diag}\,(\|\tilde{R}e_i\|_2)$. Then $|\tilde{R}_r^{-1}| \leq \sqrt{n}|\tilde{R}_c^{-1}|$ and, thus,*

$$\| \, |\tilde{R}_r^{-1}| \, \| \leq \sqrt{n} \| \, |\tilde{R}_c^{-1}| \, \|, \quad \|\cdot\| \in \{\|\cdot\|_2, \|\cdot\|_F, \|\cdot\|_1, \|\cdot\|_\infty\}. \tag{3.29}$$

Furthermore,

$$\|\tilde{R}_c^{-1}\|_2 \leq \|C_r^\dagger\|_2 \frac{1+\sqrt{p}\eta_C}{1-\sqrt{p}\eta_C \|C_r^\dagger\|_2}. \tag{3.30}$$

Proof. Note that $|\tilde{R}_r^{-1}|_{ij} \leq \sqrt{n-j+1}(\tilde{R}_{jj}/\tilde{R}_{ii})|\tilde{R}_c^{-1}|_{ij}$. Relation (3.30) follows from Proposition 3.5 and Corollary 3.2. □

REMARK 3.9. Relation (3.28) ensures that $\| \, |\tilde{R}^{-1}| \cdot |\tilde{R}| \, \|_\infty$ is bounded by $O(2^n)$, independent of C. Using the column pivoting of Gu and Eisenstat [20], this bound reduces to the order of the Wilkinson's $O(n^{1+(1/4)\log n})$ bound for the pivot growth in the Gaussian elimination (cf. [37]). In the practice, $\| \, |\tilde{R}^{-1}| \cdot |\tilde{R}| \, \|_\infty$ is usually of the order of n. Similar bounds hold for the value of $\| \, |\tilde{L}| \cdot |\tilde{L}^\dagger| \, \|_1$ in Remark 3.6.

REMARK 3.10. If $m \gg p$, then we can improve the efficiency of Algorithm 3.3 by computing the QR factorization of the matrix $\tilde{B}_r^\tau$ (or $\tilde{B}_r^\tau \Pi$). If K is the computed triangular factor, then $F = K\Pi R^\tau$ (or $F = KR^\tau$). The relative perturbation of the singular values, caused by replacing B_r with K, can be bounded by $\sqrt{p}\varepsilon_{QR}(m,p)\|B_r^\dagger\|_2$, cf. Proposition 3.4 and Corollary 3.2.

In the last step of Algorithm 3.3, we compute the SVD of the matrix $\tilde{F}$. We use the Jacobi SVD algorithm, that is, the Hestenes [23] implicit variant of the Jacobi algorithm [26]. Floating–point implementation of the algorithm follows the lines of [5], [8], [9].

The Jacobi SVD algorithm generates in floating–point arithmetic a finite sequence

$$\tilde{F}^{(k+1)} = (\tilde{F}^{(k)} + \delta\tilde{F}^{(k)})U^{(k)}, \quad k = 0, 1, \ldots, \ell - 1 \quad (\tilde{F}^{(0)} = \tilde{F}), \tag{3.31}$$

where $U^{(k)}$, $0 \leq k \leq \ell-1$, are Jacobi plane rotations and $\delta\tilde{F}^{(k)}$, $0 \leq k \leq \ell-1$, are backward errors. The final matrix $\tilde{F}^{(\ell)}$ is chosen so that the computed approximation of $\max_{i,j} |\cos \angle(\tilde{F}^{(\ell)}e_i, \tilde{F}^{(\ell)}e_j)|$ is bounded by some given tolerance `tol`. Due to rounding errors, the actual bound is somewhat weaker, i.e.

$$\max_{i,j} |\cos \angle(\tilde{F}^{(\ell)}e_i, \tilde{F}^{(\ell)}e_j)| \leq \tau(m) = \mathtt{tol} + O(m\varepsilon). \tag{3.32}$$

Usually, $\mathtt{tol} \approx m\varepsilon$. The floating point values of the column norms of $\tilde{F}^{(\ell)}$ approximate its singular values with high relative accuracy. The relative error is bounded by

$$\varepsilon_{\tilde{F}^{(\ell)}}(m, p) = (1 + \frac{\sqrt{p(p-1)}}{\sqrt{1-(p-1)\tau(m)}}\tau(m))(1 + \varepsilon_{\ell_2}(m)) - 1. \tag{3.33}$$

The bound (3.33) is derived from the fact that in the QR factorization of the column scaled matrix $\tilde{F}_c^{(\ell)} = \tilde{F}^{(\ell)}(\Delta^{(\ell)})^{-1} = Q^{(\ell)}\begin{bmatrix} R_c^{(\ell)} \\ \mathbf{O} \end{bmatrix}$, $\Delta^{(\ell)} = \mathrm{diag}\,(\|\tilde{F}^{(\ell)}e_i\|_2)$, it holds that

$$(Q^{(\ell)})^\tau \tilde{F}^{(\ell)} = \begin{bmatrix} I + (R_c^{(\ell)} - I) & \mathbf{O} \\ \mathbf{O} & I \end{bmatrix}\begin{bmatrix} \Delta^{(\ell)} \\ \mathbf{O} \end{bmatrix}, \quad \|R_c^{(\ell)} - I\|_F \leq \frac{\sqrt{p(p-1)}}{\sqrt{1-(p-1)\tau(m)}}\tau(m),$$

and from the fact that each $\|\tilde{F}^{(\ell)}e_i\|_2$ is computed with relative error bounded by $\varepsilon_{\ell_2}(m)$.
If $\tilde{\sigma}'_1 \geq \cdots \geq \tilde{\sigma}'_p$ are the sorted floating–point values of the Euclidean column norms of $\tilde{F}^{(\ell)}$, then the eigenvalues $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p > 0$ of $\tilde{F}$ can be approximated by (cf. [8])

$$\max_{1\leq i\leq p} \frac{|\tilde{\sigma}'_i - \tilde{\sigma}_i|}{\tilde{\sigma}_i} \leq h(p)\varepsilon \max_{0\leq k\leq \ell} \kappa_2(\tilde{F}_c^{(k)}) + m\varepsilon + O(\varepsilon^2), \tag{3.34}$$

where $h(p)$ is modestly growing polynomial and $\tilde{F}_c^{(k)} = \tilde{F}^{(k)}\mathrm{diag}\,(\|\tilde{F}^{(k)}e_i\|_2)^{-1}$. Furthermore, Demmel and Veselić [8] observe that, in the practice, the condition number growth factor

$$\chi(\tilde{F}) = \max_{0\leq k\leq \ell} \frac{\kappa_2(\tilde{F}_c^{(k)})}{\kappa_2(\tilde{F}_c^{(0)})} \tag{3.35}$$

is never much larger than one. (See also [29], [12].) Hence, the accuracy of the Jacobi SVD algorithm is determined by the condition number of the column scaled matrix $\tilde{F}$,

$$\kappa_2(\tilde{F}_c) = \|\tilde{F}_c\|_2\|\tilde{F}_c^\dagger\|_2, \quad \text{where} \quad \tilde{F} = \tilde{F}_c D_F, \quad D_F = \mathrm{diag}\,(\|\tilde{F}e_i\|_2). \tag{3.36}$$

In the following proposition, we estimate $\kappa_2(\tilde{F}_c)$ and $\|\tilde{F}_c^\dagger\|_2$. For the sake of simplicity, we consider the exact matrix F.

PROPOSITION 3.11. *Let F, B_r and R be as in Algorithm 3.3, and let $F = F_c D_F$, $D_F = \mathrm{diag}\,(\|Fe_i\|_2)$ and $R = D_R R_{r,1}$, $D_R = \mathrm{diag}\,(\|R^\tau e_i\|_1)$. Then*

$$\|F_c^\dagger\|_2 \leq \|B_r^\dagger\|_2\|R_{r,1}^{-1}\|_2. \tag{3.37}$$

Furthermore,

$$\kappa_2(F_c) \leq \sqrt{p}\kappa_2(B_r) \min_{D=\mathrm{diag}} \kappa_2(DR). \tag{3.38}$$

Proof. Note that $\|Fe_i\|_2 \leq \|R^\tau e_i\|_1$, $1 \leq i \leq p$, and that $F_c^\dagger = D_F D_R^{-1} R_{r,1}^{-1} \Pi^\tau (B_r^\tau)^\dagger$. Hence, $\|F_c^\dagger\|_2 \leq \|D_F D_R^{-1}\|_2 \|R_{r,1}^{-1}\|_2 \|B_r^\dagger\|_2$. Let now $\kappa_2(\bar{D}R) = \min_{D=\mathrm{diag}} \kappa_2(DR)$. Then $\kappa_2(F\bar{D}) \leq \kappa_2(B_r)\kappa_2(\bar{D}R)$, and relation (3.38) follows from the fact that (cf. [34])

$$\kappa_2(F_c) \leq \sqrt{p} \min_{D=\mathrm{diag}} \kappa_2(FD). \tag{3.39}$$

□

Hence, modulo an assumption that $\chi(\tilde{F})$ in relation (3.35) is moderate, the relative errors in the singular values computed by the Jacobi SVD algorithm are of the same order as the uncertainty caused by $\tilde{F} - F$.

We can modify Step 4 of Algorithm 3.3 and remove the dependence on $\chi(\tilde{F})$. The modification is simple: *If $\tilde{F}$ is square, apply the Jacobi SVD algorithm to $G = \tilde{F}^\tau$, else compute the QR factorization $\tilde{F} = WK$ and then apply the Jacobi SVD algorithm to $G = K^\tau$.*
The error analysis of this modification is based on Proposition 3.4 and the following proposition (cf. [10], [11]).

PROPOSITION 3.12. *Let $\tilde{F} \in \mathbf{R}^{p\times p}$ and let the Jacobi SVD algorithm be applied on $G = \tilde{F}^\tau$. Let the stopping criterion (3.32) be satisfied by $\tilde{G}^{(\ell)}$ in the sth sweep. If one sweep can be implemented in $\wp$ parallel steps, as described in Proposition 3.4, then the matrix $\tilde{G}^{(\ell)}$ satisfies*

$$\tilde{G}^{(\ell)} = (G + \delta G)U, \quad \|(\delta G)^\tau e_i\|_2 \leq \varepsilon_J(p)\|G^\tau e_i\|_2, \quad 1 \leq i \leq p, \quad \varepsilon_J(p) \leq ((1+6\varepsilon)^{(\mathbf{s}-1)\wp} - 1),$$

where U is certain orthogonal matrix. Furthermore, let the matrix $\tilde{F}_c$ from relation (3.36) satisfy $\sqrt{p}\varepsilon_J(p)\|\tilde{F}_c^\dagger\|_2 < 1$. If $\tilde{\sigma}_1' \geq \cdots \geq \tilde{\sigma}_p'$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p > 0$ are the floating-point values of the Euclidean norms of the columns of $\tilde{G}^{(\ell)}$, and the singular values of $\tilde{F}$, respectively, then

$$\max_{1\leq i\leq p} \frac{|\tilde{\sigma}_i' - \tilde{\sigma}_i|}{\tilde{\sigma}_i} \leq (1 + \sqrt{p}\varepsilon_J(p)\|\tilde{F}_c^\dagger\|_2)(1 + \varepsilon_{\tilde{F}^{(\ell)}}(p,p)) - 1,$$

where $\varepsilon_{\tilde{F}^{(\ell)}}(p,p)$ is defined as in relation (3.33).

The importance of Proposition 3.12 is that it proves a very strong form of backward stability of the Jacobi SVD algorithm: The norm–wise relative backward error in each column of $\tilde{F}$ is small. Hence, if $\|\tilde{F}_c^\dagger\|_2$ is moderate, the singular values of $\tilde{F}$ are computed with high relative accuracy. Similarly, the QR factorization of $\tilde{F}$ introduces column–wise small backward error, see Proposition 3.4. If we combine Proposition 3.4 and Proposition 3.12, we conclude that the modified step is equivalent to exact computation with small relative norm–wise backward errors in each column of $\tilde{F}$.

REMARK 3.13. The efficiency and the accuracy of the modified Step 4 increase if the QR factorization of $\tilde{F}$ is computed with column pivoting, $\tilde{F}P = WK$. In that case, we have the accelerated Jacobi SVD algorithm of Veselić and Hari [35]. (See also [8].)

REMARK 3.14. The accuracy and the error estimates of Algorithm 3.3 can be further improved using Jacobi rotations after Step 1 as follows (cf. [11]). Apply only a few steps of the Jacobi SVD algorithm on C_1^τ and apply the same rotations to B_r^τ, to preserve the equivalence (cf. relation (2.11) in Example 2.3). The threshold for skipping the rotation is set high, larger that $1/\sqrt{p}$, say. After few rotations, scale the new matrices as in Step 1, and proceed with Steps 2–4. The goal is to reduce $\|C_r^\dagger\|_2$ and, hence, to reduce the condition number of the row scaled matrix $\tilde{R}$. At the same time, the spectral condition number of the row scaled *new* B_r can be increased at most $\sqrt{p}$ times. This preconditioning ensures more accurate QR factorization of C_1^τ as well as more accurate computation of $\tilde{F}$ and its singular values. After suitable scaling, a similar process can also be used to reduce $\|B_r^\dagger\|_2$.

3.4. Backward stability. Now we prove that Algorithm 3.3 with modified Step 4 is backward stable, i.e. the matrix $\tilde{F}^{(\ell)} = (\tilde{G}^{(\ell)})^\tau$ in Proposition 3.12 is the result of exact computation with some matrices $B+\delta B$ and $C+\delta C$ where for each i, $\|(\delta B)^\tau e_i\|_2/\|B^\tau e_i\|_2$ and $\|(\delta C)^\tau e_i\|_2/\|C^\tau e_i\|_2$ are bounded by ε times a modestly growing function of matrix dimensions.

PROPOSITION 3.15. *Let the assumptions of Proposition 3.5 hold, and let $\tilde{K}$ be the computed upper triangular factor in the floating-point QR factorization of $\tilde{F}$. Let the Jacobi SVD algorithm be applied on $G = \tilde{K}^\tau$, and let $\tilde{F}^{(\ell)} = (\tilde{G}^{(\ell)})^\tau$ be as in Proposition 3.12. Furthermore, let $\tilde{R}_{r,1} = \mathrm{diag}\,(\|\tilde{R}^\tau e_i\|_1)^{-1}\tilde{R}$ and $\eta(m,p) = \varepsilon_{QR}(m,p) + \varepsilon_J(p) + \varepsilon_{QR}(m,p)\varepsilon_J(p)$. There exist backward perturbations δB, δC such that the diagram in Figure 1 commutes. Furthermore, it holds, for all*

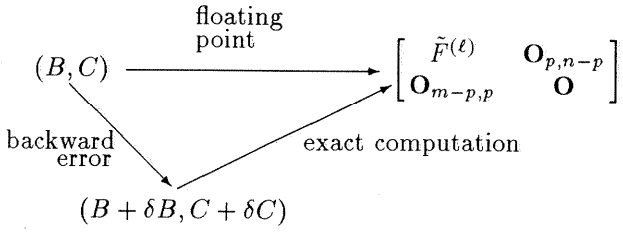


FIG. 1. *Commutative diagram for the modified Algorithm.*

i, that

$$\|(\delta B)^\tau e_i\|_2 \le \bar{\eta}_B \|B^\tau e_i\|_2, \quad \|(\delta C)^\tau e_i\|_2 \le \eta_C \|C^\tau e_i\|_2, \tag{3.40}$$

where η_C is as in Proposition 3.5 and

$$\bar{\eta}_B = \frac{1+\varepsilon_{\ell_2}(m)}{1-\varepsilon_{\ell_2}(m)}(1+\varepsilon)\left\{\varepsilon_{MM}(p)\| \,|\tilde{R}^{-1}|\cdot|\tilde{R}|\,\|_\infty + \eta(m,p)(1+\varepsilon_{MM}(p))\|\tilde{R}_{r,1}^{-1}\|_\infty\right\} + \varepsilon.$$

Hence, if $\sigma_1 \ge \cdots \ge \sigma_p$ are the singular values of $B^\tau C$ and if $\tilde{\sigma}'_1 \ge \cdots \ge \tilde{\sigma}'_p$ are the sorted floating-point approximations of the Euclidean norms of the rows of $\tilde{F}^{(\ell)}$, then

$$\max_{1\le i\le p} \frac{|\tilde{\sigma}'_i - \sigma_i|}{\sigma_i} \le (1+\sqrt{p}\bar{\eta}_B\|B_r^\dagger\|_2)(1+\sqrt{p}\eta_C\|C_r^\dagger\|_2)(1+\varepsilon_{\tilde{F}^{(\ell)}}(p,p)) - 1, \tag{3.41}$$

where $\varepsilon_{\tilde{F}^{(\ell)}}$ is defined in relation (3.33), and provided that $\sqrt{p}\bar{\eta}_B\|B_r^\dagger\|_2 < 1$.

Proof. From Proposition 3.4 and Proposition 3.12, it follows that

$$\begin{bmatrix} \tilde{K} \\ \mathbf{0} \end{bmatrix} = Q_F^\tau(\tilde{F}+\delta\tilde{F}), \quad \tilde{F}^{(\ell)} = U(\tilde{K}+\delta\tilde{K}),$$

where Q_F and U are certain orthogonal matrices and the backward errors $\delta\tilde{F}$ and $\delta\tilde{K}$ satisfy $\|\delta\tilde{F}e_i\|_2 \le \varepsilon_{QR}(m,p)\|\tilde{F}e_i\|_2$, $\|\delta\tilde{K}e_i\|_2 \le \varepsilon_J(p)\|\tilde{K}e_i\|_2$, $1 \le i \le p$. Hence,

$$\begin{bmatrix} \tilde{F}^{(\ell)} \\ \mathbf{O} \end{bmatrix} = \begin{bmatrix} U & \mathbf{O} \\ \mathbf{O} & I_{m-p} \end{bmatrix} Q_F^\tau(\tilde{F}+\delta\tilde{F}'), \quad \delta\tilde{F}' = \delta\tilde{F} + Q_F\begin{bmatrix} \delta\tilde{K} \\ \mathbf{0} \end{bmatrix},$$

where, for all i, $\|\delta\tilde{F}'e_i\|_2 \le (\varepsilon_{QR}(m,p)+\varepsilon_J(p)+\varepsilon_{QR}(m,p)\varepsilon_J(p))\|\tilde{F}e_i\|_2$. Furthermore, with the notation from the proof of Proposition 3.5, we have

$$\begin{bmatrix} \tilde{F}^{(\ell)} \\ \mathbf{O} \end{bmatrix} = \begin{bmatrix} U & \mathbf{O} \\ \mathbf{O} & I_{m-p} \end{bmatrix} Q_F^\tau(\tilde{B}_r^\tau + \mathcal{E}\tilde{R}^{-\tau}\Pi^\tau + \delta\tilde{F}'\tilde{R}^{-\tau}\Pi^\tau)\Pi\tilde{R}^\tau,$$

where, for all i and i' such that $\Pi^\tau e_i = e_{i'}$,

$$\|\delta\tilde{F}'\tilde{R}^{-\tau}\Pi^\tau e_i\|_2 \le (\varepsilon_{QR}(m,p)+\varepsilon_J(p)+\varepsilon_{QR}(m,p)\varepsilon_J(p))\sum_{k=i'}^{p}\|\tilde{F}e_k\|_2|(\tilde{R}^{-\tau})_{ki'}|.$$

Note that $\|\tilde{F}e_k\|_2 \leq (1+\varepsilon_{MM(p)})\max_{1\leq j\leq p}\|\tilde{B}_r^T e_j\|_2\|\tilde{R}^T e_k\|_1$, $1\leq k\leq p$. Hence, the backward error $\delta\tilde{B}_r$, defined by

$$(\delta\tilde{B}_r)^T = \mathcal{E}\tilde{R}^{-T}\Pi^T + \delta\tilde{F}'\tilde{R}^{-T}\Pi^T,$$

satisfies, for all i,

$$\begin{aligned}\|(\delta\tilde{B}_r)^T e_i\|_2 &\leq \varepsilon_{MM}(p)\frac{1+\varepsilon}{1-\varepsilon_{\ell_2(m)}}\| \,|\tilde{R}^{-1}|\cdot|\tilde{R}|\,\|_\infty \\ &+ \eta(m,p)(1+\varepsilon_{MM}(p))\frac{1+\varepsilon}{1-\varepsilon_{\ell_2(m)}}\sum_{k=i'}^{p}\|\tilde{R}^T e_k\|_1|(\tilde{R}^{-T})_{ki'}|.\end{aligned}$$

(Note that $\|\tilde{R}^T e_k\|_1|(\tilde{R}^{-T})_{ki'}| = |(\tilde{R}_{r,1}^{-T})_{ki'}|$.) On the other hand, as in Proposition 3.5, we can write $[\Pi\tilde{R}^T, \mathbf{O}] = [\tilde{C}_1+\delta\tilde{C}_1]Q_C$, where Q_C is an orthogonal matrix and $\|(\delta\tilde{C}_1)^T e_i\|_2 \leq \varepsilon_{QR}(n,p)\|\tilde{C}_1^T e_i\|_2$, $1\leq i\leq p$. Hence (cf. the proof of Proposition 3.5)

$$\begin{aligned}\begin{bmatrix}\tilde{F}^{(\ell)} & \mathbf{O}_{p,n-p}\\ \mathbf{O}_{m-p,p} & \mathbf{O}\end{bmatrix} &= \begin{bmatrix}U & \mathbf{O}\\ \mathbf{O} & I_{m-p}\end{bmatrix} Q_F^T(\tilde{B}_r^T\tilde{\Delta}_B + (\delta\tilde{B}_r)^T\tilde{\Delta}_B)(\tilde{\Delta}_B^{-1}\tilde{C}_1 + \tilde{\Delta}_B^{-1}\delta\tilde{C}_1)Q_C,\\ &= \begin{bmatrix}U & \mathbf{O}\\ \mathbf{O} & I_{m-p}\end{bmatrix} Q_F^T(B^T + (\delta B_e)^T + (\delta\tilde{B}_r)^T\tilde{\Delta}_B)(C+\delta C_e + \tilde{\Delta}_B^{-1}\delta\tilde{C}_1)Q_C\end{aligned}$$

where, for all i,

$$\|(\delta\tilde{B}_r)^T\tilde{\Delta}_B e_i\|_2 \leq \|(\delta\tilde{B}_r)^T e_i\|_2(1+\varepsilon_{\ell_2(m)})\|B^T e_i\|_2, \quad \|(\delta\tilde{C}_1)^T\tilde{\Delta}_B^{-1}e_i\|_2 \leq \varepsilon_{QR}(n,p)(1+\varepsilon)\|C^T e_i\|_2.$$

□

3.5. Error bound for singular vectors. In this section, we consider the errors in the singular vectors computed by the Jacobi SVD algorithm. Our analysis is based on the results from [8]. We include it here for the sake of completeness.

For the sake of simplicity we consider a simple singular value σ_j of F with corresponding left and right singular vectors v_j, u_j, respectively. Furthermore, we restrict the size of perturbation δF to ensure that the jth singular value $\sigma_j+\delta\sigma_j$ of the perturbed matrix $F+\delta F$ remains simple. In this way, the perturbed matrix has unique singular vectors $\tilde{v}_j$, $\tilde{u}_j$, that correspond to $\sigma_j+\delta\sigma_j$. Important condition number that determines the accuracy of a singular vector approximation is the *relative separation (gap)* in the set of singular values of F, introduced in in [8] as follows:

DEFINITION 3.16. *Let $\sigma_1,\ldots,\sigma_{\min\{m,n\}}$ be singular values of a nonzero $m\times n$ matrix. The relative gap of σ_i is defined by*

$$\gamma(\sigma_i) = \min_{\sigma_j\neq\sigma_i}\frac{|\sigma_i-\sigma_j|}{\sigma_i+\sigma_j}.$$

Our analysis is based on the following perturbation estimate from [8, Corollary 2.17].

PROPOSITION 3.17. *Let $F = F_c\Delta_F$ be an $m\times n$ full column rank matrix, $\Delta_F = \mathrm{diag}\,(\|Fe_i\|_2)$, and let $F+\delta F\equiv(F_c+\delta F')\Delta_F$. Let v_j and u_j be the left and right singular vectors of F, respectively, and let $\tilde{v}_j$, $\tilde{u}_j$ be the corresponding left and right singular vectors of $F+\delta F$. If*

$$\delta \equiv \|\delta F'\|_2\|F_c^\dagger\|_2 < \frac{\gamma(\sigma_j)}{1+\gamma(\sigma_j)}, \tag{3.42}$$

then

$$\max\{\|u_j-\tilde{u}_j\|_2,\ \|v_j-\tilde{v}_j\|_2\} \leq \frac{\sqrt{n-1/2}\,\delta}{(1-\delta)((1-\delta)\gamma(\sigma_j)-\delta)}. \tag{3.43}$$

In the Jacobi SVD algorithm, the limit matrix $F_\infty \equiv \lim_{k\to\infty} F^{(k)} = V\Sigma$ is the left singular vector matrix scaled by the diagonal matrix of the singular values. In the practice, the matrix F_∞ is replaced by a matrix with nearly orthogonal columns. In the next proposition, we estimate how the scaled columns of such a matrix approximate its left singular values.

PROPOSITION 3.18. *Let $F = F_c\Delta_F \in \mathbf{R}^{m\times n}$, $F_c^T F_c = I + E$, and let*

$$\delta \equiv \frac{\sqrt{2}\|E\|_F}{1+\sqrt{1-2\|E\|_F}} < \frac{1}{2}\frac{\gamma((\Delta_F)_{jj})}{1+\gamma((\Delta_F)_{jj})}, \tag{3.44}$$

where the relative gap $\gamma(\cdot)$ is computed in the set of the singular values of Δ_F. Then there is a left singular vector v_j of F such that

$$\|v_j - F_c e_j\|_2 \le \frac{\sqrt{n-1/2}\,\delta}{(1-\delta)((1-\delta)\gamma((\Delta_F)_{jj})-\delta)}. \tag{3.45}$$

Proof. For simplicity and without loss of generality we take $j = 1$. Let $F_c = Q(I+\Xi)$ be the QR factorization of F_c. Note that Ξ is upper triangular with $\Xi_{11} = 0$, hence $F_c e_1 = Qe_1$. From $E = \Xi + \Xi^T + \Xi^T\Xi$, using the continuity of the Cholesky factorization of $F_c^T F_c$ (cf. [13, Theorem 2.1]), we conclude

$$\|\Xi\|_F \le \frac{\sqrt{2}\|E\|_F}{1+\sqrt{1-2\|E\|_F}}.$$

Now write F as $F = (Q+\delta Q)\Delta_F$, $\delta Q = Q\Xi$, and apply Proposition 3.17. □

REMARK 3.19. In Proposition 3.18, we use $\|E\|_F$ instead of $\|E\|_2$ because, in the Jacobi SVD algorithm, we estimate $\|E\|_2$ by $\|E\|_2 \le \|E\|_F \le \sqrt{n}(\tau + O(m\varepsilon))$, where τ is the stopping criterion threshold. Using [13] we can estimate $\|\Xi\|_2$ also by

$$\|\Xi\|_2 \le \frac{2c_n\|E\|_2}{1+\sqrt{1-4c_n^2\|E\|_2}}, \quad c_n = 1/2 + [\log_2 n], \tag{3.46}$$

provided that $\|E\|_2 < 1/(4c_n^2)$.

The following proposition estimates norm–wise errors in computed left singular vector approximation $V + \delta V$.

PROPOSITION 3.20. *Let $V + \delta V$ be left singular vector approximation, computed in step 4) of Algorithm 3.3. Let the stopping criterion with threshold τ be satisfied after exactly* s *sweeps of some strategy. Furthermore, let one full sweep be divided into $\wp$ steps, where for $1 \le t \le \wp$ the t–th step consists of simultaneous application of $r(t)$ rotations with disjoint pivot indices. Let $\tilde{F}^{(s,t)}$ denote a floating–point matrix obtained after t such steps in the s–th sweep, and let $\sigma_1^{(s,t)} \ge \cdots \ge \sigma_n^{(s,t)}$ be the singular values of $\tilde{F}^{(s,t)}$. If for all (s,t)*

$$4\varepsilon\sqrt{2r(t)}\,\|(\tilde{F}^{(s,t)})_c^\dagger\|_2 < \frac{\gamma(\sigma_j^{(s,t)})}{1+\gamma(\sigma_j^{(s,t)})},$$

then

$$\|\delta V e_j\|_2 \le \sum_{s=1}^{\bar{s}}\sum_{t=1}^{\wp} \frac{\varepsilon\sqrt{16n-8}\,\eta^{(s,t)}}{(1-4\varepsilon\eta^{(s,t)})((1-4\varepsilon\eta^{(s,t)})\gamma(\sigma_j^{(t,s)})-4\varepsilon\eta^{(s,t)})} \tag{3.47}$$

$$+ \frac{\sqrt{2n^2-n}\,\tau'}{(1-\sqrt{2n}\,\tau')((1-\sqrt{2n}\,\tau')\gamma(\|\tilde{F}^{(\bar{s},t)}e_j\|_2)-\sqrt{2n}\tau')} \tag{3.48}$$

$$+ \varepsilon + (1+\varepsilon)(\sqrt{(1+\varepsilon)^{n+2}}-1), \tag{3.49}$$

where $\eta^{(s,t)} \leq \sqrt{2r(t)}\,\|(\tilde{F}^{(s,t)})_c^{\dagger}\|_2$, $\tau' \leq \tau + O(m\varepsilon)$, *and the relative gap* $\gamma(\|\tilde{F}^{(\bar{s},t)}e_j\|_2)$ *is computed with respect to the singular values of* $\mathrm{diag}\,(\|\tilde{F}^{(\bar{s},t)}e_i\|_2)$.

Proof. Let $U^{(s,t)}$ be product of computed Jacobi rotations in the tth step of the sth sweep. Since all rotation involved in $U^{(s,t)}$ have mutually different pivot indices, floating–point application of $U^{(s,t)}$ can be represented as

$$\tilde{F}^{(s,t+1)} = (\tilde{F}^{(s,t)} + \delta\tilde{F}^{(s,t)})U^{(s,t)},$$

where $\|\delta\tilde{F}^{(s,t)}\mathrm{diag}\,(\|\tilde{F}^{(s,t)}e_i\|_2^{-1})\|_2 \leq 4\varepsilon\sqrt{2r(t)}$. Under the assumptions of the proposition we can apply Proposition 3.17 for all (s,t). This gives the right hand side of the inequality in (3.47). The contribution (3.48) is due to application of Proposition 3.18 to the final matrix $\tilde{F}^{(\bar{s},p)}$. Note that the $O(m\varepsilon)$ correction of τ is due to possible underestimation of τ in floating–point arithmetic. Finally, the contribution (3.49) bounds errors caused by column scaling of $\tilde{F}^{(\bar{s},p)}$ in floating–point arithmetic. □

REMARK 3.21. For the row–cyclic strategy, $\wp = 2n-3$, which means that estimate in Proposition 3.20 contains an $O(n^2)$ factor, while the estimate in [8] contains an $O(n^{5/2})$ factor. Since in practice the number of sweeps is bounded by a constant, this is an $O(\sqrt{n})$ improvement over [8]. Another improvement by an $O(\sqrt{n})$ factor is possible using the results of Mathias [30].

3.6. Application to the ordinary SVD computation. The basic idea of Algorithm 3.3 is to reduce the computation of the SVD of the product of two matrices to the ordinary SVD of a single matrix. Now we show that in some cases we can use Algorithm 3.3 to compute an accurate SVD of a single matrix. The application of Algorithm 3.3 to SVD computation of a $G \in \mathbf{R}^{m\times n}$, $m \geq n$, is based on the following two important observations in [6]. First, floating–point LU factorization with complete pivoting

$$P_1GP_2 = L\Delta U,\quad L \in \mathbf{R}^{m\times p},\ U \in \mathbf{R}^{p\times n},\quad L_{ii} = U_{ii},\quad 1 \leq i \leq p = \mathrm{rank}\,(G), \tag{3.50}$$

usually computes very accurate factors $\tilde{L}$, $\tilde{\Delta}$ and $\tilde{U}$, and, if $\mathrm{rank}\,(G)$ can be determined exactly, the singular values of $\tilde{L}\tilde{\Delta}\tilde{U}$ approximate the singular values of G with high relative accuracy. Second, the values of $\|\tilde{L}^{\dagger}\|_2$ and $\|\tilde{U}^{\dagger}\|_2$ are moderate.
Hence, for an accurate SVD of G, it remains to compute the SVD of the product $\tilde{L}\tilde{\Delta}\tilde{U}$. This approach is especially attractive if the LU factorization of G can be computed with small element–wise relative errors as, e.g., in the case of network oscillator matrices where even the rank decision can be made correctly, see [14]. To compute the SVD of $\tilde{L}\tilde{\Delta}\tilde{U}$, we use Algorithm 3.3 with $B = \tilde{L}^{\tau}$, $C = \tilde{\Delta}\tilde{U}$.

ALGORITHM 3.22.
Input $G \in \mathbf{R}^{m\times n}$, $m \geq n$.
Step 1 Compute the LU factorization with complete pivoting, $P_1GP_2 = L(\Delta U) \equiv L\hat{U}$.
Step 2 Compute $\Delta_L = \mathrm{diag}\,(\|Le_i\|_2)$ and $L_c = L\Delta_L^{-1}$, $U_1 = \Delta_L\hat{U}$.
Step 3 Compute the LQ factorization with row pivoting, $\Pi^{\tau}U_1 = TQ$.
Step 4 (Optional) Compute the QR factorization $L_c\Pi = KR$.
Step 5 Compute $F = L_c\Pi T$ (optionally, $F = RT$) using the standard matrix multiply algorithm.
Step 6 Use the Jacobi SVD algorithm to compute the SVD of F, $\begin{bmatrix} \Sigma \\ \mathbf{O} \end{bmatrix} = V^{\tau}FW$.
Output The SVD of G reads

$$\begin{bmatrix} \Sigma \oplus \mathbf{O}_{n-p,n-p} \\ \mathbf{O} \end{bmatrix} = ((KV)^{\tau}P_1)G(P_2Q(W \oplus I_{n-p})).$$

REMARK 3.23. Note that $F = (K^{\tau}P_1)G(P_2Q^{\tau})$, which means that Algorithm 3.22 can be considered as an implicit way to precondition G by pre– and post–multiplication by orthogonal matrices.

The following error bound is a corollary of the analysis from § 3.3.

COROLLARY 3.24. *Let* $\mathrm{rank}\,(G) = p$ *and let* $\tilde{L}_c \in \mathbf{R}^{m\times p}$ *and* $\tilde{U}_1 \in \mathbf{R}^{p\times n}$ *be the computed (full rank) factors in Step 2 of Algorithm 3.22, and let* $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p$ *be the exact singular values of the product* $\tilde{L}_c\tilde{U}_1 = G + \delta G$, *with* δG *being the backward error in the first two steps of the algorithm. Furthermore, let* $\tilde{T}$, $\tilde{R}$ *and* $\tilde{F}$ *be the computed approximations of the matrices* T, R *and* F, *respectively, and let* $\tilde{U}_r = \mathrm{diag}\,(\|\tilde{U}_1^T e_i\|_2)^{-1}\tilde{U}_1$, $(\tilde{L}_c)_c = \tilde{L}_c\mathrm{diag}\,(\|\tilde{L}_c e_i\|_2)^{-1}$, $\tilde{T}_{c,1} = \tilde{T}\mathrm{diag}\,(\|\tilde{T}e_i\|_1)^{-1}$. *If* $\tilde{\sigma}'_1 \geq \cdots \geq \tilde{\sigma}'_p$ *are the singular values computed in Step 6 by applying the Jacobi SVD algorithm to* $\tilde{F}^T$, *then*

$$\max_{1\leq i\leq p} \frac{|\tilde{\sigma}'_i - \tilde{\sigma}_i|}{\tilde{\sigma}_i} \leq (1+\eta_1)(1+\eta_2)(1+\eta_3)(1+\eta_4) - 1, \tag{3.51}$$

where $0 \leq \eta_1 \leq \sqrt{p}\epsilon_{QR}(n,p)\|\tilde{U}_r^\dagger\|_2$ *bounds the relative perturbation from the LQ factorization in Step 3,* $0 \leq \eta_2 \leq \sqrt{p}\epsilon_{QR}(m,p)\|(\tilde{L}_c)_c^\dagger\|_2$ *bounds the relative perturbation from the QR factorization in Step 4,* $0 \leq \eta_3 \leq \epsilon_{MM}(p)\||\tilde{R}|\cdot|\tilde{T}|\cdot|\tilde{T}^{-1}|\cdot|\tilde{R}^{-1}|\,\|_2$ *bounds the relative perturbation from the floating-point matrix product in Step 5, and, finally,* $0 \leq \eta_4 \leq \sqrt{p}\epsilon_J(p)\|\tilde{L}_c^{-1}\|_2\|\tilde{T}_c^{-1}\|_2/(1-\eta_3)$ *bounds the relative error in the Jacobi SVD algorithm. We also assume that in relation (3.51)* $\max_{1\leq i\leq 4}\eta_i < 1$.

The result of Corollary 3.24 is attractive because it gives closed relative error bound that depends only on condition numbers of the matrices $\tilde{L}_c$, $\tilde{U}_r$ and $\tilde{T}_{c,1}$. All relevant condition numbers can be estimated using, e.g., LAPACK's `SPOCON()` procedure [1]. Note that the bound (3.51) does not depend on the condition growth factor in the Jacobi SVD algorithm.

4. The eigenvalue problem of the product HM. In this section, we apply the analysis from § 3 to the eigenvalue problem

$$HMx = \lambda x, \quad H, M \in \mathbf{R}^{n\times n} \text{ symmetric and positive definite.} \tag{4.52}$$

If $H = L_H L_H^T$, $M = L_M L_M^T$ are the Cholesky factorizations and if $L_H^T L_M = V\Sigma U^T$ is the SVD of $L_H^T L_M$, then the matrix $T = L_H V\Sigma^{-1/2}$ satisfies

$$T^{-1} = \Sigma^{-1/2}U^T L_M^T, \quad T^{-1}HT^{-T} = T^T MT = \Sigma, \text{ and } HMT = T\Sigma^2. \tag{4.53}$$

Hence, the Cholesky factorization of H and M, followed by the SVD of the product $L_H^T L_M$ gives the solution of the eigenvalue problem (4.52) and the solution of the balancing problem (1.5) for the linear system (1.3), (1.4). It is important to note that the balancing problem (1.5) can be solved using the SVD of $L_H^T L_M$ without computing the Gramians H and M. More precisely, we can use an algorithm of Hammarling [21] to compute the Cholesky factors L_H and L_M of H and M directly from the dual pair of the Lyapunov equations (cf. [27], [16])

$$EH + HE^T = -FF^T, \quad E^T M + ME = -G^T G. \tag{4.54}$$

In § 4.1, we analyze the sensitivity of the eigenvalues of HM to small element–wise relative perturbations δH, δM and we describe the set of positive definite pencils $HM - \lambda I$ for which the eigenvalues can be computed with high relative accuracy in floating–point arithmetic. In § 4.2, we use the natural connection between the eigenvalue problem (4.52) and the ΠSVD of (L_H, L_M) to define a new algorithm for eigenvalue computation and in § 4.3 we prove that the new algorithm is capable of achieving the high relative accuracy predicted in § 4.1. In § 4.4, we analyze errors in the computed eigenvectors.

4.1. Sensitivity of the eigenvalues. Let H and M from relation (4.52) be given with initial uncertainties δH, δM such that

$$|\delta H_{ij}| \leq \epsilon|H_{ij}|, \quad |\delta M_{ij}| \leq \epsilon|M_{ij}|; \quad 1 \leq i,j \leq n, \tag{4.55}$$

or, more generally, such that

$$|\delta H_{ij}| \leq \epsilon\sqrt{H_{ii}H_{jj}}, \quad |\delta M_{ij}| \leq \epsilon\sqrt{M_{ii}M_{jj}}, \quad 1 \leq i,j \leq n. \tag{4.56}$$

Our goal is to estimate the relative accuracy to which the eigenvalues of HM are determined by the data (H and M).

THEOREM 4.1. *Let H and M be $n \times n$ real symmetric and positive definite matrices, and let δH, δM be symmetric perturbations as in relation (4.56). Furthermore, let*

$$H_s = \mathrm{diag}\,(H_{ii})^{-1/2} H \mathrm{diag}\,(H_{ii})^{-1/2}, \quad M_s = \mathrm{diag}\,(M_{ii})^{-1/2} M \mathrm{diag}\,(M_{ii})^{-1/2}, \tag{4.57}$$

and let $2n\epsilon \max\{\|H_s^{-1}\|_2, \|M_s^{-1}\|_2\} < 1$. If $\lambda_1 \geq \cdots \geq \lambda_n$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ are the eigenvalues of HM and $(H+\delta H)(M+\delta M)$, respectively, then

$$\max_{1\leq i\leq n} \frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq 6\sqrt{2}n(\|H_s^{-1}\|_2 \max_{i,j} \frac{|\delta H_{ij}|}{\sqrt{H_{ii}H_{jj}}} + \|M_s^{-1}\|_2 \max_{i,j} \frac{|\delta M_{ij}|}{\sqrt{M_{ii}M_{jj}}}). \tag{4.58}$$

Proof. Let $H = (L_H + \delta L_H)(L_H + \delta L_H)^T$, $M = (L_M + \delta L_M)(L_M + \delta L_M)^T$ be the perturbed Cholesky factorizations. Using an estimate from [13], we can write

$$\begin{aligned} L_H + \delta L_H &= L_H(I + \Gamma_H), \quad \|\Gamma_H\|_2 \leq \sqrt{2}n\|H_s^{-1}\|_2 \max_{i,j} \frac{|\delta H_{ij}|}{\sqrt{H_{ii}H_{jj}}}, \\ L_M + \delta L_M &= L_M(I + \Gamma_M), \quad \|\Gamma_M\|_2 \leq \sqrt{2}n\|M_s^{-1}\|_2 \max_{i,j} \frac{|\delta M_{ij}|}{\sqrt{M_{ii}M_{jj}}}. \end{aligned}$$

Hence, the eigenvalues of $(H+\delta H)(M+\delta M)$ are the squares of the singular values of the product $(I+\Gamma_H)^T L_H^T L_M (I+\Gamma_M)$ and Theorem 3.1 implies that, for all i,

$$(1 - \|\Gamma_H\|_2)(1 - \|\Gamma_M\|_2) \leq \sqrt{\frac{\tilde{\lambda}_i}{\lambda_i}} \leq (1 + \|\Gamma_H\|_2)(1 + \|\Gamma_M\|_2). \tag{4.59}$$

□

We conclude that perturbations of the type (4.55), (4.56) cause small relative perturbations of the eigenvalues of $HM - \lambda I$, if $\|H_s^{-1}\|_2$ and $\|M_s^{-1}\|_2$ are moderate. Or next theorem, based on the results from [36], shows that moderate $\|H_s^{-1}\|_2$ and $\|M_s^{-1}\|_2$ are also necessary for accurate eigenvalue computation in the presence of the perturbations (4.55), (4.56).

THEOREM 4.2. *Let H and M be as in Theorem 4.1, and let $\kappa > 1$. If for all $\epsilon < 1/\kappa$ and all symmetric perturbations as in (4.55) the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ of HM and $(H+\delta H)(M+\delta M)$, respectively, satisfy*

$$\max_{1\leq i\leq n} \frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq \kappa\epsilon, \tag{4.60}$$

then $\max\{\|H_s^{-1}\|_2, \|M_s^{-1}\|_2\} \leq (1+\kappa)/2$.

Proof. Let $\delta H = \mathbf{O}$ and $|\delta M_{ij}| \leq \epsilon |M_{ij}|$, $1 \leq i,j \leq n$. Then $M + \delta M$ must remain positive definite and, for all $\epsilon < 1/\kappa$, $\|M_s^{-1}\|_2 \leq (1+\epsilon)/(2\epsilon)$ (cf. [36, Lemma 2.20]). This implies $\|M_s^{-1}\|_2 \leq (1+\kappa)/2$. Now choose $\delta M = \mathbf{O}$ and $|\delta H_{ij}| \leq \epsilon |H_{ij}|$, $1 \leq i,j \leq n$. □

4.2. The algorithm. The floating–point Cholesky factorizations of H and M are equivalent to the exact factorizations with symmetric backward perturbations δH, δM which satisfy (4.56) with $\epsilon \equiv \varepsilon_C(n) \leq (n+5)\varepsilon$, see [8]. Hence, if we use the computed Cholesky factors as input to Algorithm 3.3, we can compute the eigenvalues of HM with a relative error bound similar to the one from Theorem 4.1. It is more efficient, however, to perform the first two steps of Algorithm 3.3 implicitly, during the initial Cholesky factorizations. More precisely, we can use the following algorithm.

ALGORITHM 4.1.

Input $H, M \in \mathbf{R}^{n\times n}$ symmetric and positive definite.

Step 1 Compute $\Delta_H = \text{diag}\,(H_{ii})^{-1/2}$ and $H_s = \Delta_H H \Delta_H$, $M_1 = \Delta_H^{-1} M \Delta_H^{-1}$.

Step 2 Compute the Cholesky factorizations $BB^T = H_s$, $CC^T = \Pi^T M_1 \Pi$ (with complete pivoting).

Step 3 Compute the matrix $F = B^T \Pi C$ using the standard matrix multiply algorithm.

Step 4 Use the Jacobi SVD algorithm, as described in § 3, and compute the SVD of F, $\Sigma = V^T F U$.

Step 5 If needed, compute $T = \Delta_H^{-1} B V \Sigma^{-1/2}$, $T^{-1} = \Sigma^{-1/2} U^T C^T \Pi^T \Delta_H$.

Output The matrices T, T^{-1} and Σ satisfy

$$T^T M T = T^{-1} H T^{-T} = \Sigma, \quad HMT = T\Sigma^2, \quad MHT^{-T} = T^{-T}\Sigma^2.$$

For the computation of the eigenvector matrix T we need only the left singular vector matrix V. This means that in the Jacobi SVD algorithm in Step 4 we need not to accumulate the Jacobi rotations (the matrix U). If we use the modified Step 4 (cf. § 3) then we apply the Jacobi SVD algorithm on F^T and the matrix V is now the accumulated product of the Jacobi rotations. Hence, the price for better error estimate (cf. Proposition 3.12) in this case is extra work to accumulate the Jacobi rotations.

It is possible, however, to use the modified Step 4 and favorable error bounds from Proposition 3.12, and without the need to accumulate Jacobi rotations. First, note that in Algorithm 4.1 $Z \equiv T^{-T} = \Delta_H \Pi C U \Sigma^{-1/2}$ is the eigenvector matrix of MH. Next, note that U and V generally are not unique. Without loss of generality, we use U and V as generic symbols for the right and the left singular matrix of F, respectively. Hence, if we apply the Jacobi SVD algorithm to F^T, then V denotes the accumulated product of Jacobi rotations and $F^T V = U\Sigma$ is the limit matrix. This means that the eigenvectors of MH can be computed using the modified Step 4, and without accumulating of Jacobi rotations. So, for $T = Z^{-T}$ we simply apply the modified algorithm to (M, H) instead of (H, M).

4.3. Relative error bound for eigenvalues. Consider now floating–point errors in Algorithm 4.1. For the sake of simplicity, we first introduce some useful notation. For an arbitrary symmetric positive definite matrix X and an arbitrary matrix Y we define X_s, Y_r and Y_c as follows:

$$X_s = \text{diag}\,(X_{ii})^{-1/2} X \text{diag}\,(X_{ii})^{-1/2}, \quad Y = \text{diag}\,(\|Y^T e_i\|_2) Y_r = Y_c \text{diag}\,(\|Y e_i\|_2).$$

Let $\tilde{H}_s$ and $\tilde{M}_1$ be the computed approximations of H_s and M_1, respectively. Then

$$\tilde{H}_s = \Delta_H (H + \delta H_e) \Delta_H, \quad \tilde{M}_1 = \Delta_H^{-1} (M + \delta M_e) \Delta_H^{-1},$$

where, for all i, j,

$$|\delta H_e|_{ij} \le \varepsilon_1 |H_{ij}|, \quad |\delta M_e|_{ij} \le \varepsilon_1 |M_{ij}|, \quad \varepsilon_1 \le \frac{1+\varepsilon}{(1-\varepsilon)^{3/2}} - 1. \tag{4.61}$$

On the other hand, the computed Cholesky factors $\tilde{B}$ and $\tilde{C}$ satisfy

$$\begin{aligned} \tilde{B}\tilde{B}^T &= \tilde{H}_s + \delta\tilde{H}_s, \quad |\delta\tilde{H}_s|_{ij} \le \varepsilon_C(n), \\ \tilde{C}\tilde{C}^T &= \Pi^T(\tilde{M}_1 + \delta\tilde{M}_1)\Pi, \quad |\delta\tilde{M}_1|_{ij} \le \varepsilon_C(n)\sqrt{(\tilde{M}_1)_{ii}(\tilde{M}_1)_{jj}}, \end{aligned}$$

where $\varepsilon_C(n) \le (n+5)\varepsilon$ (cf. [8]). Hence,

$$\begin{aligned} \tilde{B}\tilde{B}^T &= \Delta_H(H + \delta H_e + \Delta_H^{-1}\delta\tilde{H}_s\Delta_H^{-1})\Delta_H, \quad |\Delta_H^{-1}\delta\tilde{H}_s\Delta_H^{-1}|_{ij} \le \varepsilon_C(n)\sqrt{H_{ii}H_{jj}}, \\ \Pi\tilde{C}\tilde{C}^T\Pi^T &= \Delta_H^{-1}(M + \delta M_e + \Delta_H\delta\tilde{M}_1\Delta_H)\Delta_H^{-1}, \quad |\Delta_H\delta\tilde{M}_1\Delta_H| \le \varepsilon_C(n)(1+\varepsilon_1)\sqrt{M_{ii}M_{jj}}. \end{aligned}$$

Now, using Theorem 4.1 we obtain the following error bound.

PROPOSITION 4.2. *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the true eigenvalues of HM and let $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ be the squared singular values of the exact product $\tilde{B}^T \Pi \tilde{C}$, where $\tilde{B}$ and $\tilde{C}$ are the computed Cholesky factors in Step 2 of Algorithm 4.1. Then*

$$\max_{1\leq i\leq n} \frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq 6\sqrt{2n}(\varepsilon_C(n) + \varepsilon_1 + \varepsilon_C(n)\varepsilon_1)(\|H_s^{-1}\|_2 + \|M_s^{-1}\|_2). \tag{4.62}$$

Using double precision accumulation in the Cholesky factorization, the bound for $\varepsilon_C(n)$ can be reduced to $O(1)\varepsilon$ for all $n \leq 1/\varepsilon$.

Consider now the errors in the eigenvalues obtained from the singular values of the computed product in Step 3.

PROPOSITION 4.3. *Let $\tilde{F}$ be the floating-point product $\tilde{B}^T \Pi \tilde{C}$ and let $\tilde{C}_{c,1} = \tilde{C}\mathrm{diag}\,(\|\tilde{C}e_i\|_1)^{-1}$. Furthermore, let the Jacobi SVD algorithm be applied to the matrix $\tilde{F}^T$ and let $\tilde{F}^{(\ell)}$ satisfies the stopping criterion (3.32). Let $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ be as in Proposition 4.2, and let $\tilde{\lambda}'_1 \geq \cdots \geq \tilde{\lambda}'_n$ be the squared singular values of the matrix $\tilde{F}^{(\ell)}$. Then, for all i,*

$$\frac{|\tilde{\lambda}'_i - \tilde{\lambda}_i|}{\tilde{\lambda}_i} \leq 2\eta + \eta^2, \tag{4.63}$$

where

$$\eta = \varepsilon_{MM}(n)\| \,|\tilde{B}|\, \|_2 \|\tilde{B}^{-1}\|_2 \| \,|\tilde{C}| \cdot |\tilde{C}^{-1}|\, \|_2 + \sqrt{n}\|\tilde{B}^{-1}\|_2 \|\tilde{C}_{c,1}^{-1}\|_1 \varepsilon_J(n)(1+\varepsilon_{MM}(n))\sqrt{1+\varepsilon_C(n)}.$$

Proof. From Proposition 3.12 and Proposition 3.5, it follows that $\tilde{F}^{(\ell)}$ is the result of exact application of the Jacobi SVD algorithm to the matrix $\tilde{F}' = \tilde{B}^T \Pi \tilde{C} + \mathcal{E} + \delta\tilde{F}$, where $\tilde{F} = \tilde{B}^T \Pi \tilde{C} + \mathcal{E}$ and

$$|\mathcal{E}| \leq \varepsilon_{MM}(n)|\tilde{B}^T|\Pi|\tilde{C}|, \quad \|\delta\tilde{F}e_i\|_2 \leq \varepsilon_J(n)\|(\tilde{B}^T\Pi\tilde{C} + \mathcal{E})e_i\|_2, \quad 1 \leq i \leq n.$$

The matrix $\tilde{F}'$ can be written as

$$\tilde{F}' = (I + \Omega)\tilde{B}^T \Pi \tilde{C}, \quad \Omega = \mathcal{E}\tilde{C}^{-1}\Pi^T\tilde{B}^{-T} + \delta\tilde{F}\tilde{C}^{-1}\Pi^T\tilde{B}^{-T},$$

where

$$\|\mathcal{E}\tilde{C}^{-1}\Pi^T\tilde{B}^{-T}\|_2 \leq \varepsilon_{MM}(n)\|\tilde{B}^{-1}\|_2 \| \,|\tilde{B}|\, \|_2 \| \,|\tilde{C}| \cdot |\tilde{C}^{-1}|\, \|_2 \tag{4.64}$$

and

$$\begin{aligned}
\|\delta\tilde{F}\tilde{C}^{-1}\Pi^T\tilde{B}^{-T}\|_F &\leq \|\tilde{B}^{-1}\|_2\sqrt{n}\max_{1\leq j\leq n}\|\delta\tilde{F}\tilde{C}^{-1}e_j\|_2 \\
&\leq \|\tilde{B}^{-1}\|_2\sqrt{n}\varepsilon_J(n)\max_{1\leq j\leq n}\sum_{k=j}^{n}\|\tilde{F}e_k\|_2|(\tilde{C}^{-1})_{kj}| \\
&\leq \|\tilde{B}^{-1}\|_2\sqrt{n}\varepsilon_J(n)(1+\varepsilon_{MM}(n))\max_{1\leq j\leq n}\|\tilde{B}^T e_j\|_2\sum_{k=j}^{n}\|\tilde{C}e_k\|_1|(\tilde{C}^{-1})_{kj}| \\
&\leq \|\tilde{B}^{-1}\|_2\sqrt{n}\varepsilon_J(n)(1+\varepsilon_{MM}(n))\max_{1\leq j\leq n}\|\tilde{B}^T e_j\|_2\|\tilde{C}_{c,1}^{-1}\|_1 \\
&\leq \sqrt{n}\|\tilde{B}^{-1}\|_2\|\tilde{C}_{c,1}^{-1}\|_1\varepsilon_J(n)(1+\varepsilon_{MM}(n))\sqrt{1+\varepsilon_C(n)}. \qquad (4.65)
\end{aligned}$$

On the other hand, from Theorem 3.1 it follows that

$$\max_{1\leq i\leq n}\frac{|\tilde{\lambda}'_i - \tilde{\lambda}_i|}{\tilde{\lambda}_i} \leq 2\|\Omega\|_2 + \|\Omega\|_2^2. \tag{4.66}$$

□

Now we show that the relative error bound in Proposition 4.3 is newer much larger and that it can be much smaller than the bound in Proposition 4.2. Define $\delta H = \delta H_e + \Delta_H^{-1}\delta\tilde{H}_s\Delta_H^{-1}$, $\delta M = \delta M_e + \Delta_H\delta\tilde{M}_1\Delta_H$ and note that

$$\tilde{C}\tilde{C}^\tau = (\Pi^\tau\Delta_H^{-1}\Pi)\Pi^\tau(M+\delta M)\Pi(\Pi^\tau\Delta_H^{-1}\Pi)$$

and $\|(\tilde{C}\tilde{C}^\tau)_s^{-1}\|_2 = \|\tilde{C}_r^{-1}\|_2^2 = \|(M+\delta M)_s^{-1}\|_2$. Furthermore, since $\tilde{C}$ is computed with complete pivoting, it holds (up to small relative error which we ignore) that

$$\tilde{C}_{ii}^2 \geq \sum_{k=i}^{j}\tilde{C}_{jk}^2, \quad 1 \leq i \leq j \leq n,$$

and from Proposition 3.8 it follows that $\| \, |\tilde{C}_c^{-1}| \, \|_2 \leq \sqrt{n}\| \, |\tilde{C}_r^{-1}| \, \|_2$. On the other hand,

$$\| \, |\tilde{C}| \cdot |\tilde{C}^{-1}| \, \|_2 = \| \, |\tilde{C}_c| \cdot |\tilde{C}_c^{-1}| \, \|_2 \leq n\| \, |\tilde{C}_r^{-1}| \, \|_2 \leq n^{3/2}\sqrt{\|(M+\delta M)_s^{-1}\|_2}. \tag{4.67}$$

Hence, we can bound $\|\mathcal{E}\tilde{C}^{-1}\Pi^\tau\tilde{B}^{-\tau}\|_2$ in relation (4.64) by

$$\|\mathcal{E}\tilde{C}^{-1}\Pi^\tau\tilde{B}^{-\tau}\|_2 \leq n^2\varepsilon_{MM}(n)\sqrt{\|(H+\delta H)_s^{-1}\|_2\|(M+\delta M)_s^{-1}\|_2}.$$

Similarly, since $\|\tilde{C}_{c,1}^{-1}\|_1 \leq n\|\tilde{C}_r^{-1}\|_1 \leq n^{3/2}\|\tilde{C}_r^{-1}\|_2$, the bound (4.65) for $\|\delta\tilde{F}\tilde{C}^{-1}\Pi^\tau\tilde{B}^{-\tau}\|_F$ can be replaced with

$$\|\delta\tilde{F}\tilde{C}^{-1}\Pi^\tau\tilde{B}^{-\tau}\|_F \leq n^2\varepsilon_J(n)(1+\varepsilon_{MM}(n))\sqrt{1+\varepsilon_C(n)}\sqrt{\|(H+\delta H)_s^{-1}\|_2\|(M+\delta M)_s^{-1}\|_2}.$$

REMARK 4.4. Note that $\tilde{C}^\tau\tilde{C} = \tilde{C}^{-1}(\tilde{C}\tilde{C}^\tau)\tilde{C}$ represents one step of Rutishauser's LR algorithm. It is well known that an LR step has nontrivial diagonalizing effect; see [33], [35], [8]. Hence, we may expect that $\|(\tilde{C}^\tau\tilde{C})_s^{-1}\|_2 \ll \|(M+\delta M)_s^{-1}\|_2$. Finally, since $\|\tilde{C}_c^{-1}\|_2 = \sqrt{\|(\tilde{C}^\tau\tilde{C})_s^{-1}\|_2}$, and since moderate $\|M_s^{-1}\|_2$ is necessary for accurate computation, we can take the factors $\| \, |\tilde{C}| \cdot |\tilde{C}^{-1}| \, \|_2$ and $\|\tilde{C}_{c,1}^{-1}\|_1$ as moderate functions of n.

We finish the eigenvalue analysis with an important observation. We show that the matrix $\tilde{F}^{(\ell)}$ in Proposition 4.3 is backward stable function of H and M. In other words, there are small backward perturbations δH_b, δM_b such that $\tilde{F}^{(\ell)}$ is the result of exact computation with the matrices $H+\delta H_b$, $M+\delta M_b$. To prove this, first note that the matrix $\tilde{F}'$ can be written as

$$\tilde{F}' = (\tilde{B}^\tau + (\delta\tilde{B})^\tau)\Pi\tilde{C}, \quad \|(\delta\tilde{B})^\tau e_i\|_2 \leq \zeta_B,$$

where $\zeta_B \leq \sqrt{1+\varepsilon_C(n)}(\varepsilon_{MM}(n)\| \, |\tilde{C}| \cdot |\tilde{C}^{-1}| \, \|_1 + \varepsilon_J(n)(1+\varepsilon_{MM}(n))\|\tilde{C}_{c,1}^{-1}\|_1)$. Hence,

$$(\tilde{B}+\delta\tilde{B})(\tilde{B}+\delta\tilde{B})^\tau = \tilde{H}_s + \delta\tilde{H}_s + \delta\tilde{H}_s', \quad |\delta\tilde{H}_s'|_{ij} \leq 2\sqrt{1+\varepsilon_C(n)}\zeta_B + \zeta_B^2.$$

If we define

$$\begin{aligned}\delta H_b &= \delta H_e + \Delta_H^{-1}\delta\tilde{H}_s\Delta_H^{-1} + \Delta_H^{-1}\delta\tilde{H}_s'\Delta_H^{-1},\\ \delta M_b &= \delta M_e + \Delta_H\delta\tilde{M}_1\Delta_H,\end{aligned}$$

then, for all i, j,

$$\begin{aligned}|\delta H_b|_{ij} &\leq (\varepsilon_1 + \varepsilon_C(n) + 2\sqrt{1+\varepsilon_C(n)}\zeta_B + \zeta_B^2)\sqrt{H_{ii}H_{jj}},\\ |\delta M_b|_{ij} &\leq (\varepsilon_1 + \varepsilon_C(n)(1+\varepsilon_1))\sqrt{M_{ii}M_{jj}},\end{aligned}$$

and $\tilde{F}^{(\ell)}$ is the result of an exact computation with the matrices $H+\delta H_b$ and $M+\delta M_b$. Note that the backward errors are given element–wise and that the error in M is $O(n\varepsilon)$, while the error in H contains an additional factor that depends on $\| \, |\tilde{C}| \cdot |\tilde{C}^{-1}| \, \|_1$ and on $\|\tilde{C}_{c,1}^{-1}\|_2$. However, Remark 4.4 indicates that the backward errors in H can be considered as $f(n)\varepsilon$, where $f(n)$ is moderate function of n. (Note that we can use the strong rank–revealing QR factorization [20] of $\tilde{C}$ to obtain $f(n)$ comparable with the Wilkinson's pivot growth factor in the Gaussian elimination with complete pivoting.) An important corollary of this backward stability analysis is the element–wise backward stability of the accelerated Jacobi algorithm for eigenvalue computation of symmetric positive definite matrix M. In that case, $H = I$, $\delta H_b = \mathbf{O}$, and the backward error is element–wise $O(n\varepsilon)$, see [11].

4.4. Error bound for eigenvectors. For an error estimate for the eigenvector matrix, we use the formula $T = \Delta_H^{-1} B V \Sigma^{-1/2}$ from Step 5 of Algorithm 4.1 and the results from § 3.5. As in § 3.5, we assume that the eigenvalues of HM are well separated and that they remain simple in the presence of floating–point errors. For simplicity, we let $\tilde{L}_H$ denote the computed product $\Delta_H^{-1}\tilde{B}$. In that case we can write $\tilde{L}_H = (I + \Gamma_H) L_H$, where L_H is the Cholesky factor of H, and Γ_H is bounded as in Theorem 4.1, with $\epsilon \approx O(n\varepsilon)$. (Note that in Algorithm 4.1 we can equivalently first compute the Cholesky factorization $H = L_H L_H^T$ and then define $B = \Delta_H L_H$.)

PROPOSITION 4.5. *Let $V + \delta V$ be as in Proposition 3.20, and let $\tilde{T}$ be the floating–point value of the product $\tilde{L}_H(V+\delta V)\tilde{\Sigma}^{-1/2}$, where $\tilde{\Sigma} = \operatorname{diag}(\sigma_j + \delta\sigma_j)$ is the matrix of the singular values computed in Step 4 of Algorithm 4.1. Then for $T_{ij} \neq 0$*

$$\tilde{T}_{ij} = T_{ij}\left(1 + \frac{\delta\sigma_j}{\sigma_j}\right)^{-1/2}(1+\epsilon_{ij})(1+\epsilon^{(j)})(1+\eta_{ij}), \tag{4.68}$$

where $|\epsilon_{ij}| \leq \varepsilon$, $|\epsilon^{(j)}| \leq (1+\varepsilon)^2 - 1$, and

$$|\eta_{ij}| \leq \frac{\|\Gamma_H\|_2 + (1+\|\Gamma_H\|_2)[\|\delta V e_j\|_2 + \varepsilon_{MM}(n)(1+\|\delta V e_j\|_2)]}{|\cos\angle(e_i^T L_H, V e_j)|}. \tag{4.69}$$

Proof. Note that

$$\tilde{T}_{ij} = (e_i^T L_H (I+\Gamma_H)(V+\delta V)e_j + E_{ij})(\sigma_j+\delta\sigma_j)^{-1/2}(1+\epsilon^{(j)})(1+\epsilon_{ij}),$$

where $(\sigma_j+\delta\sigma_j)^{-1/2}(1+\epsilon^{(j)})$, $|\epsilon^{(j)}| \leq (1+\varepsilon)^2 - 1$, is the computed value of $(\sigma_j+\delta\sigma_j)^{-1/2}$, $|\epsilon_{ij}| \leq \varepsilon$ are small rounding errors, and E is the error from matrix multiplication, $|E| \leq \varepsilon_{MM}(n)|L_H(I+\Gamma_H)| \cdot |(V+\delta V)|$. The rest of the proof is a tedious computation which we omit. □

5. Numerical tests. Numerical testing of Algorithm 3.3 is similar to testing of the generalized singular value computation algorithms in [11]. For the readers convenience, we give detailed description of the test.

5.1. Test matrix generation. We generate random matrices B_r and C_r with given $\kappa_2(B_r)$ and $\kappa_2(C_r)$, and apply scalings $B = \Delta_B B_r$, $C = \Delta_C C_r$, where Δ_B, Δ_C are random diagonal, nonsingular matrices with given spectral condition numbers. Recall that the subscript r means that the matrix has unit rows. The 4–tuple $(\kappa_2(B_r), \kappa_2(\Delta_B), \kappa_2(C_r), \kappa_2(\Delta_C))$ takes all values from the set

$$\mathcal{C} = \left\{ \kappa_{ijkl} = (10^i, 10^j, 10^k, 10^l) \ : \ (i,j,k,l) \in \mathcal{I} \times \mathcal{J} \times \mathcal{K} \times \mathcal{L} \subset \mathbf{N}^4 \right\},$$

where $\mathcal{I}, \mathcal{J}, \mathcal{K}, \mathcal{L}$ are determined at the very beginning of the test and kept fixed. For each fixed κ_{ijkl}, we generate a set of test pairs using the LAPACK's `DLATM1` procedure [7] as follows. We let the 4–tuple $(\mu_{i'}, \mu_{j'}, \mu_{k'}, \mu_{l'})$ of distributions of the singular values of $(B_r, \Delta_B, C_r, \Delta_C)$ take all values from the set

$$\mathcal{M} = \{\mu_{i'j'k'l'} = (\mu_{i'}, \mu_{j'}, \mu_{k'}, \mu_{l'})\} \subseteq \mathcal{P}_1 \times \mathcal{P}_2 \times \mathcal{P}_3 \times \mathcal{P}_4 \subseteq \{\pm 1, \ldots, \pm 6\}^4,$$

where the sets of indices $\mathcal{P}_1, \ldots, \mathcal{P}_4$ contain admissible values of parameter **MODE** in the procedure **DLATM1**. For each fixed $(\kappa_{ijkl}, \mu_{i'j'k'l'})$ we generate random pairs using random number generators with distributions chosen from the set $\mathcal{R} \subseteq \{\mathcal{U}(-1,1), \mathcal{U}(0,1), \mathcal{N}(0,1)\}$. For each fixed distribution $\chi \in \mathcal{R}$ we generate a set $\mathcal{E}^{\chi}_{\kappa_{ijkl},\mu_{i'j'k'l'}}$ of different pairs, with the cardinality of $\mathcal{E}^{\chi}_{\kappa_{ijkl},\mu_{i'j'k'l'}}$ being fixed at the beginning of the test. Each test pair is generated in double precision and its generalized singular values are computed using a double precision procedure. The generalized singular values computed by double precision procedure are then taken as reference for single precision procedure that runs on original pair rounded to single precision.

A random matrix $B = \Delta_B B_r$ with given $\kappa_2(B_r)$ and $\kappa_2(\Delta_B)$ is generated using the following algorithm. (Cf. [19, P.8.5.3 and P.8.5.4], [8], [11].)

ALGORITHM 5.1.

1. $B := \mathrm{diag}\,(b_{ii})$, where $b_{11}, \ldots, b_{nn}$ are generated using **DLATM1()** from [7] with parameters chosen accordingly to the current node in $\mathcal{C} \times \mathcal{M} \times \mathcal{R}$.
2. $B := \cdots (U_i(\cdots(U_1 B V_1)\cdots)V_j)\cdots$, where U_i, V_j are random plane rotations.
3. $B := \cdots (W_k(\cdots(W_1 B)\cdots))\cdots$, where W_k, $k = 1, \ldots$ are plane rotations designed to equilibrate the rows of B. On output, all rows of B have about the same Euclidean length.
4. The diagonal matrix Δ_B is generated by **DLATM1()** with parameters chosen accordingly.
5. $B := \Delta_B B$.

Before PSVD computation, both $\kappa_2(B_r)$ and $\kappa_2(C_r)$ are computed using singular values computed by LAPACK's **SGESVD()** procedure applied to B_r, C_r, respectively. Computed condition numbers are compared with desired values in κ_{ijkl}. The two sets of values usually differ by a small factor (cf. [11]).

5.2. Test results. All tests were done on an Intel 486DX processor. We used Microsoft Fortran Power–Station with *improve floating–point consistency* compiler option.

EXAMPLE 5.2. We use our double precision procedure as reference for testing our single precision procedure **SGPSVD()**. We do not use preconditioning explained in Remark 3.14. For each test pair (B, C), we compute

$$\theta(B,C) = \frac{\max\limits_{\sigma \in \sigma(B,C)} \dfrac{|\delta\sigma|}{\sigma}}{\max\{\kappa_2(B_r), \kappa_2(C_r)\}},$$

where σ and $\sigma + \delta\sigma$ are the double and the corresponding single precision approximations of a singular value of $B^T C$. Our analysis predicts values $\theta(\cdot)$ of order of single precision roundoff unit. The input parameters for the test are

$$\begin{aligned}
\mathcal{I} &= \{1, \ldots, 7\}, \ \mathcal{K} = \mathcal{I}, \\
\mathcal{J} &= \{4, 8, 12, 8, 10\}, \ \mathcal{L} = \{3, 5, 7, 9, 11\}, \\
\mathcal{M} &= \{(5, 4, -5, 3), (3, -4, 5, -3), (4, 5, 3, -4)\}, \ \mathcal{R} = \{\mathcal{U}(-1,1), \mathcal{U}(0,1), \mathcal{N}(0,1)\}.
\end{aligned}$$

For each node of $\mathcal{C} \times \mathcal{M} \times \mathcal{R}$ we perform three tests on randomly generated pairs. This makes the total of 33075 test pairs.

In Figure 2, we display the values of $\theta(\cdot,\cdot)$ for all test pairs. In Figure 3, we display in $\log_{10}$ scale the values of

$$\varepsilon(i,k) = \max_{\mathcal{J},\mathcal{L}} \max_{\mathcal{M}} \max_{(B,C) \in \bigcup_{\chi \in \mathcal{R}} \mathcal{E}^{\chi}_{\kappa_{ijkl},\mu_{i'j'k'l'}}} \max_{\sigma \in \sigma(B,C)} \frac{|\delta\sigma|}{\sigma}, \quad (i,k) \in \mathcal{I} \times \mathcal{K}.$$

Note that relative accuracy depends on $\max\{\kappa_2(B_r), \kappa_2(C_r)\}$, and not on $\kappa_2(\Delta_B)$, $\kappa_2(\Delta_C)$.

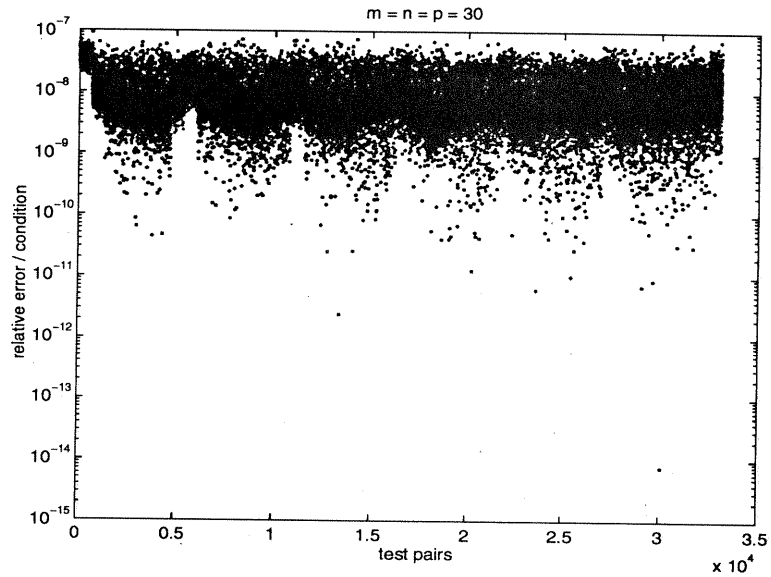


FIG. 2. *The values of $\theta(B,C)$ for all test pairs (B,C). Note that the values of $\theta(\cdot,\cdot)$ below 10^{-8} indicate that $\max\{\kappa_2(B_r),\kappa_2(C_r)\}$ overestimates the real condition number.*

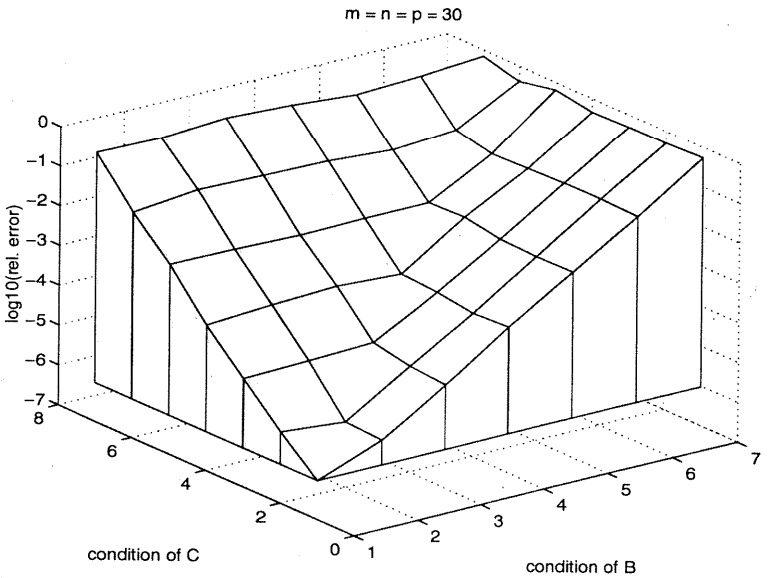


FIG. 3. *The logarithms of maximal relative errors $\varepsilon(i,k)$ ($-\log_{10}\varepsilon(i,k) \approx$ the minimal number of correct digits for all test pairs (B,C) with $\kappa_2(B_r) = 10^i$, $\kappa_2(C_r) = 10^k$) for $(i,k) \in \mathcal{I} \times \mathcal{K}$. Note that $\varepsilon(i,k) \approx 7 - \max\{i,k\}$, as predicted by the theory.*

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenny, S. Ostrouchov, and D. Sorensen, *LAPACK users' guide, second edition*, SIAM, 1992.

[2] J. Barlow, *Stability analysis of the G-algorithm and a note on its application to sparse least squares problems*, BIT, 25 (1985), pp. 507–520.

[3] Å. Björk, *Numerical Methods for Least Squares Problems*, SIAM, 1996.

[4] P. A. Businger and G. H. Golub, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[5] P. P. M. de Rijk, *A one-sided Jacobi algorithm for computing the singular value decomposition on a vector computer*, SIAM J. Sci. Stat. Comp., 10 (1989), pp. 359–371.

[6] J. Demmel, S. Eisenstat, M. Gu, I. Slapničar, and K. Veselić, *Notes on computing the SVD with high relative accuracy.* Preprint, 1995.

[7] J. Demmel and A. McKenney, *A test matrix generation suite*, LAPACK Working Note 9, Courant Institute, New York, March 1989.

[8] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[9] Z. Drmač, *Implementation of Jacobi rotations for accurate singular value computation in floating point arithmetic.* SIAM J. Sci. Comp., to appear.

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

[11] ———, *A tangent algorithm for computing the generalized singular value decomposition.* Department of Computer Science, University of Colorado at Boulder, Technical report CU-CS-815-96, submitted to SIAM J. Numer. Anal., July 1995.

[12] ———, *On the condition behaviour in the Jacobi method*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 509–514.

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

[14] Z. Drmač and K. Veselić, *Note on the accuracy of the eigenvalues or singular values of matrices generated by finite elements.* Preprint, April 1995.

[15] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Num. Anal., 32 (1995), pp. 1–2.

[16] K. V. Fernando and S. Hammarling, *A product induced singular value decomposition (ΠSVD) for two matrices and balanced realization*, in Linear Algebra in Signals, Systems, and Control, SIAM, Philadelphia, 1988, pp. 128–140.

[17] W. M. Gentleman, *Error analysis of QR decompositions by Givens transformations*, Linear Algebra Appl., 10 (1975), pp. 189–197.

[18] S. K. Godunov, A. G. Antonov, O. P. Kirilyuk, and V. I. Kostin, *Garantirovannaya tochnost resheniya sistem lineĭnykh uravneniĭ v evklidovykh prostranstvakh*, Novosibirsk Nauka, Sibirskoe Otdelenie, 1988.

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations, second edition*, The Johns Hopkins University Press, 1989.

[20] M. Gu and S. Eisenstat, *An efficient algorithm for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848 – 869.

[21] S. Hammarling, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.

[22] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM J. Sci. Stat. Comp., 7 (1986), pp. 1147–1159.

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

[24] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, 1996.

[25] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[26] 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.

[27] A. J. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, *Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms*, IEE Trans. Automat. Contr, AC–32 (1987), pp. 115–122.

[28] R.-C. Li, *Relative perturbation theory: (I) Eigenvalue and singular value variations*, technical report, Mathematical Science Section, Oak Ridge National Laboratory, Oak Ridge, TN 37831–6367, January 1996.

[29] W. F. Mascarenhas, *A note on Jacobi being more accurate than QR*, SIAM J. Matrix Anal. Appl., 15 (1993), pp. 215–218.

[30] R. Mathias, *Spectral perturbation bounds for positive definite matrices*, tech. report, Department of Mathematics, College of William and Mary, Williamsburg, VA 23187, September 1995.

[31] J. J. Modi and M. R. B. Clarke, *An alternative Givens ordering*, Numer. Math., 43 (1984), pp. 83–90.

[32] J. H. Rice, *A theory of condition*, SIAM J. Num. Anal., 3 (1966), pp. 287–310.

[33] H. Rutishauser, *Vorlesungen über numerische Mathematik, Band 2., Differentialgleichungen und Eigen-*

wertprobleme, Birkhäuser Verlag, Basel und Stuttgart, 1976. Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften, Math. Reihe, Band 57.

[34] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

[35] K. VESELIĆ AND V. HARI, *A note on a one-sided Jacobi algorithm*, Numer. Math., 56 (1989), pp. 627–633.

[36] K. VESELIĆ AND I. SLAPNIČAR, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

[37] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1962), pp. 281–330.

[38] ———, *The Algebraic Eigenvalue Problem*, Springer, Berlin Heidelberg New York, 1965.