Proceedings of the 1. Conference on
Applied Mathematics and Computation
Dubrovnik, Croatia, September 13–18, 1999
pp. 105–112

Accuracy and Stability in Numerical Linear Algebra*

Zlatko Drmač†

Abstract. We give an overview of our recent work on accurate computation of generalized eigenvalue and singular value decompositions of matrix pairs and triplets. Our goal is to develop efficient and highly accurate algorithms and to produce high quality mathematical software. Using error analysis and perturbation theory, we develop templates for accurate floating point computation of the product and quotient induced singular value decompositions, canonical correlations and diagonalization of symmetric pencils $H - \lambda M$, $HM - \lambda I$, with positive definite H, M. The new algorithms are numerically robust. For instance, the eigenvalues of $H - \lambda M$ and $HM - \lambda I$ are computed with optimal relative error bound: each eigenvalue λ is computed with relative error $|\delta\lambda|/\lambda$ of order (up to a factor of the dimension) $\mathbf{u}\{\min_{\Delta\in\mathcal{D}} \kappa_2(\Delta H\Delta) + \min_{\Delta\in\mathcal{D}} \kappa_2(\Delta M\Delta)\}$, where $\mathbf{u}$ is the roundoff unit and $\mathcal{D}$ is the set of nonsingular diagonal matrices. Moreover, the backward error is elementwise small in the sense that finite precision computation corresponds to an exact computation with $H + \delta H$, $M + \delta M$, where for all i, j and some moderate functions f and g, $|\delta H_{ij}| \leq f(n)\mathbf{u}\sqrt{H_{ii}H_{jj}}$, $|\delta M_{ij}| \leq g(n)\mathbf{u}\sqrt{M_{ii}M_{jj}}$.

AMS subject classification: 65F15, 65F30, 65G50, 65Y20

Key words: accuracy, eigenvalues, singular values

1. Introduction

Various forms of decompositions of matrices, matrix pairs and triplets are powerful tools in theoretical and numerical treatment of problems in applied sciences (see, e.g., [10, 9, 8, 1]).

In this paper, we are particularly interested in decompositions related to generalized singular value and generalized symmetric eigenvalue problems. These include the ordinary singular value decomposition (SVD), the product induced SVD of matrix pairs (PSVD), the quotient SVD of matrix pairs (QSVD), the SVD with respect to pairs of elliptic norms ((H, K)–SVD), and spectral decomposition of symmetric pencils $H - \lambda M$, $HM - \lambda I$, with positive definite matrices H and M.

Our goal is to compute these decompositions with high accuracy whenever nu-

*This work was supported by the Croatian Ministry of Science and Technology grant 037012.

†Department of Mathematics, University of Zagreb, Bijenička cesta 30, 10000 Zagreb, Croatia, e–mail: drmac@math.hr

merically feasible. We consider high accuracy numerically feasible if small initial uncertainties in the data induce small relative uncertainties in the target values. In that case we say that those values are well determined by the data. So, for instance, the eigenvalues of $HMx = \lambda x$ are well determined if changing $n \times n$ matrices H and M to $H + \delta H$ and $M + \delta M$, where $|\delta H_{ij}| \leq \varepsilon |H_{ij}|$, $|\delta M_{ij}| \leq \varepsilon |M_{ij}|$, $1 \leq i, j \leq n$, changes any eigenvalue λ by $\delta\lambda$, $|\delta\lambda| \leq c(H, M) f(n) \varepsilon |\lambda|$, where $c(H, M)$ is a moderate condition number and $f(n)$ is a moderate polynomial. Analogously, with properly chosen metric and proper condition numbers, we can define well determined eigenvectors and eigenspaces. For the sake of simplicity, in this paper we consider only the singular value and eigenvalue computations.

A desirable property of an algorithm is that it approximates the well determined eigenvalues (and singular values) with high relative accuracy independent of their magnitudes. This is an important issue, because the smallest eigenvalues are in applications usually the most interesting ones and, unfortunately, the most sensitive ones in presence of numerical errors. To design such an algorithm, we need detailed knowledge of the structure of errors produced by finite precision implementation of the algorithm, as well as deep understanding of the sensitivity to perturbations of the original problem.

Consider, for simplicity, the ordinary symmetric eigenvalue problem $Hx = \lambda x$. The matrix H is diagonalized by an infinite number of orthogonal similarity transformations, $\cdots U_2^T (U_1^T H U_1) U_2 \cdots \longrightarrow \Lambda$, and in the limit $U^T H U = \Lambda$, where $U = U_1 U_2 \cdots$ and Λ is the diagonal matrix of H's eigenvalues. In finite precision computation, each transformation U_i is approximated by some $\widetilde{U}_i$, and applied with some error E_i. Moreover only a finite number of transformations is used:

$$\widetilde{H}_k = \widetilde{U}_k^T (\cdots (\widetilde{U}_2^T (\widetilde{U}_1^T H \widetilde{U}_1 + E_1) \widetilde{U}_2 + E_2) \cdots) \widetilde{U}_k + E_k.$$

The index k is chosen so that $\widetilde{H}_k$ is sufficiently close to diagonal matrix and its diagonals are taken as approximative eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ of H. Let $\widehat{U} = \widetilde{U}_1 \cdots \widetilde{U}_k$, and let $\widetilde{U}$ denote the computed matrix $\widehat{U}$. The columns of $\widetilde{U}$ are the computed approximations of the eigenvectors of H. To assess the error in the computed values, we prove the existence of symmetric perturbation δH such that $\widehat{U}^T (H + \delta H) \widehat{U} = \widetilde{H}_k$ exactly. The matrix δH is called backward error. Different algorithms produce backward errors of different structures and different sizes.

The problem of assessing the error is thus transformed into a perturbation problem: If H is changed to $H + \delta H$, estimate $|\delta\lambda|$. By the classical Weyl's theorem $|\delta\lambda| \leq \|\delta H\|_2$, where $\|H\|_2 = \max_{\|x\|_2 = 1} \|Hx\|_2$. The error in the output is not larger than the error in the input. Are we satisfied with this rather strong result?

Consider the following symmetric matrix and its eigenvalues computed by using MATLAB's function `eig`.

$$H = \begin{bmatrix} 10^{40} & 10^{29} & 10^{19} \\ 10^{29} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{bmatrix}, \quad \texttt{eig}(H) = \begin{bmatrix} 1.000000000000000 \cdot 10^{40} \\ 0 \\ -8.100009764062724 \cdot 10^{19} \end{bmatrix}. \tag{1}$$

Thus, the matrix H is indefinite and numerically singular.

How reliable is this conclusion? Let us for the sake of the experiment try to compute the Cholesky factorization $H = L^\tau L$, which is essentially unique for positive definite matrices. In that case, the eigenvalues of H are the squared singular values of L. So, let us try to compute L and its squared singular values. By using MATLAB's functions $L = \texttt{chol}(H)$ and $\texttt{svd}(L)$, we obtain[1]

$$L \approx \begin{bmatrix} 10^{20} & 10^{9} & 10^{-1} \\ 0 & 9.94 \cdot 10^{9} & 9.04 \cdot 10^{-2} \\ 0 & 0 & 9.90 \cdot 10^{-1} \end{bmatrix}, \quad \texttt{svd}(L).^2 = \begin{bmatrix} 1.000000000000000 \cdot 10^{40} \\ 9.900000000000002 \cdot 10^{19} \\ 9.818181818181819 \cdot 10^{-1} \end{bmatrix}. \quad (2)$$

The function `chol` declares the matrix H positive definite and different numbers appear as possible candidates for the two smallest eigenvalues of H. Does this mean that the two eigenvalues are not well determined by H? Are the two sets of values just computed merely *random*?

Let us experiment. Let us invert H numerically, compute the spectrum of the numerical inverse and take the reciprocals of the computed eigenvalues. Numerical inversion is done by using MATLAB's function `inv`. The computed eigenvalues are

$$\texttt{eig}(\texttt{inv}(H)).^{-1} = \begin{bmatrix} 1.000000000000000 \cdot 10^{40} \\ 9.900000000000002 \cdot 10^{19} \\ 9.818181818181817 \cdot 10^{-1} \end{bmatrix}. \quad (3)$$

Now consider this rather odd algorithm to compute the spectrum of H: Invert H numerically, then invert the computed inverse and then use the function `eig` to compute the eigenvalues of $\texttt{inv}(\texttt{inv}(H))$. We obtain

$$\texttt{eig}(\texttt{inv}(\texttt{inv}(H))) = \begin{bmatrix} 1.000000000000000 \cdot 10^{40} \\ 9.900000000000000 \cdot 10^{19} \\ 9.818181818181817 \cdot 10^{-1} \end{bmatrix}. \quad (4)$$

Our final experiment goes as follows. Write H as $H = DAD$ with

$$A = \begin{bmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 10^{20} & 0 & 0 \\ 0 & 10^{10} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The pencils $H - \lambda I$ and $A - \lambda D^{-2}$ are equivalent, and the spectrum of H can be computed as generalized eigenvalues of $A - \lambda D^{-2}$. The latter are computed in MATLAB as

$$\texttt{eig}(A, D^{-2}) = \begin{bmatrix} 1.000000000000000 \cdot 10^{40} \\ 9.900000000000002 \cdot 10^{19} \\ 9.818181818181818 \cdot 10^{-1} \end{bmatrix}. \quad (5)$$

[1]For simplicity, the matrix L is given only with few decimal places.

Comparing (1), (2), (3), (4) and (5), we see that the results of `eig`(H) are considerably different from others. In fact, one can show that only `eig`(H) failed to compute the two smallest eigenvalues to full machine precision!

An analysis of `eig` gives for the backward error $\|\delta H\|_2 \leq \alpha \cdot \texttt{eps} \cdot \|H\|_2$, where $\texttt{eps} \approx 2.22 \cdot 10^{-16}$, $\alpha \approx O(1)$. Thus, in the eigenvalues computed by `eig`(H) each eigenvalue has an uncertainty up to $\|\delta H\|_2 \leq \alpha \cdot 2.22 \cdot 10^{24}$. We have

$$\frac{|\delta\lambda|}{|\lambda|} \leq \frac{\|\delta H\|_2}{|\lambda|} \leq \frac{\|\delta H\|_2}{\|H\|_2} \|H\|_2 \|H^{-1}\|_2.$$

Here the quantity $\kappa_2(H) = \|H\|_2\|H^{-1}\|_2$ is the condition number. In our 3×3 example $\kappa_2(H) \approx 10^{40}$. This certainly means bad news for small eigenvalues. But the fact that several different algorithms in our example produce almost identical approximations of the eigenvalues indicates that the condition number $\kappa_2(H)$ is *artificial* and only a consequence of specific δH. For a detailed analysis see [3, 11, 2].

Similar difficulties occur in generalized eigenvalue and singular value problems where two or more matrices are involved. We use perturbation theory to identify classes of matrices with well determined decompositions. In such an analysis, we also determine the form of perturbation that yields smaller condition numbers. Finally, we try to design an algorithm capable of achieving optimal accuracy in previously defined classes. In the next section we briefly describe our approach, strategy, goals, and some recent results in algorithmic development. We avoid technical details and refer the reader to [4, 5, 6, 7, 2]. Also, for brevity, we only describe the PSVD and the $HMx = \lambda x$ problems.

2. Accurate computation of the PSVD

Our approach to numerical computation of the SVD of the product $B^\tau C$ is based on the following theorem.

Theorem 1. *Let $p \times m$ matrix B and $p \times n$ matrix C be of full row rank, and let $\widetilde{B} = B+\delta B$, $\widetilde{C} = C+\delta C$ 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 $\widetilde{B}^\tau \widetilde{C}$, respectively, then, for all i, either $\sigma_i = \tilde{\sigma}_i = 0$, or*

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

The crucial observation is the invariance of $B^\dagger \delta B$ and $C^\dagger \delta C$ under certain row scalings. That fact motivates the set of requirements listed below. With each requirement we briefly comment how our algorithm from [5] behaves with respect to that requirement.

(i) The backward error matrices δB, δC should be small not only in the sense[2] $\|\delta B\|/\|B\| \ll 1$, $\|\delta C\|/\|C\| \ll 1$, but also in the following stronger sense: for each row

[2]Here the symbol $\|\cdot\|$ denotes some matrix norm. On vectors, it denotes the Euclidean norm.

index i,

$$\|\delta B(i,:)\| \leq f_B(p,m)\mathbf{u}\|B(i,:)\|, \quad \|\delta C(i,:)\| \leq f_C(p,n)\mathbf{u}\|C(i,:)\|,$$

where f_B, f_C are modest functions of matrix dimensions. This means that the backward error in small rows is correspondingly small — small rows are preserved.

The proposed algorithm satisfies the row-wise backward error requirement, independent of any rank assumptions.

(*ii*) The computed results should be accurate in the class of *row-wise well scaled problems*. Let us explain: The current theory (cf., e.g., [2]) states that with the row-wise small backward errors δB, δC, and with full row rank B and C, the loss of accuracy is determined by the condition numbers $\|B_r^\dagger\|$, $\|C_r^\dagger\|$, where $B = D_B B_r$, $C = D_C C_r$, $D_B = \mathrm{diag}(\|B(i,:)\|)$, $D_C = \mathrm{diag}(\|C(i,:)\|)$. If $\|B_r^\dagger\|$ and $\|C_r^\dagger\|$ are moderate, we say that the PSVD of $B^\tau C$ is row-wise well scaled. Note that B_r and C_r have unit rows, and that the class of row-wise well scaled problems is closed under diagonal scalings — as long as the backward error is row-wise small, the accuracy of the algorithm is the same for all matrices $(D_1 B)^\tau (D_2 C)$, where D_1, D_2 are *arbitrary diagonal matrices*.

The proposed algorithm is accurate on row-wise well scaled problems. The singular values are computed with relative error bound

$$\max_i \frac{|\delta\sigma_i|}{\sigma_i} \leq f_{B,C}(m,n,p)\mathbf{u}(\|B_r^\dagger\| + \|C_r^\dagger\|).$$

Note that

$$\|B_r^\dagger\| \leq \sqrt{p} \min_{D=\mathrm{diag}} \kappa(DB),$$

that is, the algorithm computes as if B and C were optimally scaled! The code is equipped with a condition estimator and it optionally returns both the results and error estimates. Note: the algorithm can be modified to work well on wider class of input matrices. If, for example, the matrix C is structured as (diagonal)$\times$(well–conditioned)$\times$(diagonal), the modification requires only one sorting of the rows of C.

(*iii*) The algorithm should be simple and efficient, based on only a few building blocks which are common in singular value computation. These include, for example, the QR factorization (with column or complete pivoting), matrix multiplication, and an ordinary SVD procedure. Numerical analysis of the complete algorithm gives specific requirements on each building block. So, for instance, we require the classical matrix multiplication and avoid Strassen–like fast procedures. The Householder or Givens QR factorization with column pivoting (the best one is BLAS 3 based SGEQP3) satisfies all requirements. For numerically optimal results, the ordinary SVD procedure should have the accuracy of the Jacobi SVD algorithm. The modular structure of the algorithm makes it possible always to use the best currently available codes. This also opens possibilities for straightforward parallelizations in PBLAS and

ScaLAPACK styles. (We are currently working on a parallel implementation of the code.)

The proposed algorithm satisfies the simplicity and efficiency requirements. It uses two diagonal scalings, one QR factorization with column pivoting, one matrix multiplication [dense] × [triangular] (STRMM), and one SVD computation (by Jacobi SVD algorithm to enhance numerical precision).

3. Accurate eigenvalues of $HM - \lambda I$

The behaviour of the eigenvalues of $HM - \lambda I$ under symmetric elementwise perturbations $|\delta H| \leq \varepsilon |H|$, $|\delta M| \leq \varepsilon |M|$ is well understood. In the following two theorems the relevant condition numbers are related to the matrices H_s and M_s, where $H = D_H H_s D_H$, $M = D_M M_s D_M$, and $D_H = \mathrm{diag}(H_{ii})^{-1/2}$, $D_M = \mathrm{diag}(M_{ii})^{-1/2}$.

Theorem 2. *Let H and M be $n \times n$ real symmetric and positive definite matrices, and let δH, δM be symmetric perturbations such that $|\delta H| \leq \varepsilon |H|$, $|\delta M| \leq \varepsilon |M|$. Furthermore, let*

$$2n\varepsilon \max\{\|H_s^{-1}\|_2, \|M_s^{-1}\|_2\} < 1.$$

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

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

Theorem 3. *Let H and M be as in Theorem 2, and let $\kappa > 1$. If for all $\varepsilon < 1/\kappa$ and all symmetric perturbations as in Theorem 2, 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*

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

then

$$\max\{\|H_s^{-1}\|_2, \|M_s^{-1}\|_2\} \leq \frac{1+\kappa}{2}.$$

Since the elementwise rounding errors in H and M are generally unavoidable, we can hope to approximate the spectrum of $HM - \lambda I$ with high relative accuracy only if the matrices H_s and M_s are well conditioned. This and some other considerations yield the following road map for the algorithmic development (see [5]):

(i) The backward error should be elementwise small: the computed eigenvalues $\lambda_i + \delta\lambda_i$ correspond to the exact eigenvalues of $(H + \delta H)(M + \delta M)$, where $\max_{i,j} |\delta H_{ij}|/\sqrt{H_{ii}H_{jj}}$ and $\max_{i,j} |\delta M_{ij}|/\sqrt{M_{ii}M_{jj}}$ are, up to certain factors of the dimension, of the order of the roundoff $\mathbf{u}$. In other words, the backward perturbed

matrix can be represented as

$$H + \delta H = D_H(H_s + \delta H_s)D_H, \quad |(\delta H_s)_{ij}| \leq f_H(n)\mathbf{u},$$
$$M + \delta M = D_M(M_s + \delta M_s)D_M, \quad |(\delta M_s)_{ij}| \leq f_M(n)\mathbf{u}.$$

The proposed algorithm satisfies the elementwise backward error requirement. The only condition is that floating point Cholesky factorizations of H and M complete without breakdown. Note: if floating point Cholesky factorization fails, then the matrix is elementwise close to a symmetric nondefinite matrix and a different approach (natural factor formulation, implicit solution of Lyapunov equations) to the whole problem is strongly advised, because the accuracy is very likely lost at the very moment of storing H and M (and even exact computation wouldn't produce useful results).

(*ii*) The computed results should be as accurate as the data warrants. This is ensured by strong requirements on the backward error, and by the perturbation theory which precisely identifies matrices for which high accuracy is possible: symmetric positive definite matrix pencil $HM-\lambda I$ determines its eigenvalues well in the presence of rounding errors *if and only if* the condition numbers of the matrices H_s and M_s are moderate.

The proposed algorithm satisfies this high accuracy requirement. Once the matrices H and M are stored, the algorithm computes the eigenvalues with condition number of order of $\kappa(H_s) + \kappa(M_s)$, where $\kappa(X) = \|X\|\,\|X^{-1}\|$. That is,

$$\max_{1\leq i\leq n} \frac{|\delta\lambda_i|}{\lambda_i} \leq f_{H,M}(n)\mathbf{u}(\kappa(H_s) + \kappa(M_s)).$$

This means that this sharp error bound is the same for all pencils

$$(D_1HD_1)(D_2MD_2) - \lambda I,$$

where D_1, D_2 are arbitrary diagonal nonsingular matrices. Note that

$$\kappa(H_s) \leq n \min_{D=\mathrm{diag}} \kappa(DHD),$$

that is, the algorithm computes as if H and M were nearly optimally scaled! The code is equipped with a condition number estimator, and it optionally returns both the result and error estimates.

(*iii*) The algorithm should be simple and efficient, based on only a few building blocks which are common in symmetric matrix computation and for which reliable and optimized implementations already exist. These include, for instance, the Cholesky factorization (with pivoting), matrix multiplication with triangular matrices, ordinary eigenvalue (or singular value) solver.

The proposed algorithm satisfies the simplicity and efficiency requirements. It uses two symmetric diagonal scalings (multiplications with diagonal matrices), one

Cholesky factorization (pure), one Cholesky factorization with pivoting, one matrix multiplication [*square dense*] × [*square triangular*] *(STRMM), and one ordinary SVD (by Jacobi method to enhance numerical accuracy).*

Almost identical description applies to the QSVD and the $Hx = \lambda Mx$ problems, as well as to the (H, K)–SVD and SVD of matrix triplets (including canonical correlations problem). Our algorithms are implemented in an efficient, rigorously tested software.

Acknowledgement. The author thanks J. Barlow, J. Demmel, V. Hari, E. Jessup, I. Slapničar and K. Veselić for many interesting discussions and comments.

References

[1] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[2] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Technical report CS–97–348, Department of Computer Science, University of Tennessee, Knoxville (LAPACK Working Note 119), 1997, Linear Algebra Appl., to appear.

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

[4] Z. Drmač, *A tangent algorithm for computing the generalized singular value decomposition*, SIAM J. Numer. Anal., 35 (1998), pp. 1804–1832.

[5] Z. Drmač, *Accurate computation of the product induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.

[6] Z. Drmač, *A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm*, IMA J. Numer. Anal., 19 (1999), pp. 191–213.

[7] Z. Drmač, *New accurate algorithms for singular value decomposition of matrix triplets*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1026–1050.

[8] 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.

[9] 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*, IEEE Trans. Automat. Control, AC–32 (1987), pp. 115–122.

[10] C. C. Paige, *The general linear model and the generalized singular value decomposition*, Linear Algebra Appl., 70 (1985), pp. 269–284.

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