BIT manuscript No.
(will be inserted by the editor)

Verified inclusions for a nearest matrix of specified rank deficiency via a generalization of Wedin's $\sin(\theta)$ theorem*

Marko Lange and Siegfried M. Rump

Received: date / Accepted: date

Abstract For an $m \times n$ matrix A, the mathematical property that the rank of A is equal to r for $0 < r < \min(m, n)$ is an ill-posed problem. In this note we show that, regardless of this circumstance, it is possible to solve the strongly related problem of computing a nearby matrix with at least rank deficiency k in a mathematically rigorous way and using only floating-point arithmetic.

Given an integer k and a real or complex matrix A, square or rectangular, we first present a verification algorithm to compute a narrow interval matrix $\varDelta$ with the property that there exists a matrix within $A - \varDelta$ with at least rank deficiency k.

Subsequently, we extend this algorithm for computing an inclusion for a specific perturbation E with that property but also a minimal distance with respect to any unitarily invariant norm. For this purpose, we generalize Wedin's $\sin(\theta)$ theorem by removing its orthogonality assumption. The corresponding result is the singular vector space counterpart to Davis and Kahan's generalized $\sin(\theta)$ theorem for eigenspaces.

The presented verification methods use only standard floating-point operations and are completely rigorous including all possible rounding errors and/or data dependencies.

* This research was partially supported by CREST, Japan Science and Technology Agency (JST), and by Université Paris Sorbonne.

M. Lange
Institute for Reliable Computing, Hamburg University of Technology, Am Schwarzenberg-Campus 3, Hamburg 21073, Germany E-mail: m.lange@tuhh.de

S. M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Am Schwarzenberg-Campus 3, Hamburg 21073, Germany, and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan E-mail: rump@tuhh.de

Keywords Verified error bounds, distance to rank deficiency, ill-posedness, $\sin(\Theta)$ theorem, unitarily invariant norms, separation of singular vector subspaces, INTLAB

Mathematics Subject Classification (2000) 65G20

1 Introduction and notation

Verification methods use standard floating-point arithmetic and estimate possible rounding errors rigorously so that computed results, i.e., error bounds, are true with mathematical rigor. As an advantage, verification methods are fast; as a disadvantage, their application is basically restricted to well-posed problems. For an overview of verification methods cf. [32] and [in Japanese] [28]. An easy-to-read introduction [in German] is [34].

The restriction to well-posed problems can be explained as follows. First, as a well-posed problem, consider the verification of regularity of a matrix. Let, for example, $A \in \mathbb{R}^{n \times n}$ be given and denote by I the identity matrix of suitable size. If $\|I - RA\|_\infty < 1$ for some $R \in \mathbb{R}^{n \times n}$, then no eigenvalue of A can be zero and A must be regular. An obvious choice for R is an approximate inverse of A.

It is important to observe that, mathematically, R may be an arbitrary matrix of appropriate size. Any requirement like "being sufficiently close to A^{-1}" would have to be verified to maintain mathematical rigor. Of course the verification is bound to fail for some randomly chosen R, but the assertion $\|I - RA\|_\infty < 1 \Rightarrow \det(A) \neq 0$ is correct for every R.

The verification of $\|I - RA\|_\infty < 1$ in floating-point poses the problem that in most practical situations rounding errors will inevitably occur when evaluating this expression. On the other hand, when using a floating-point arithmetic following the IEEE-754 floating-point standard [11,12], the maximum relative error of every floating-point operation is bounded by the relative rounding error unit [7]. Thus, all rounding errors can be rigorously estimated, and the property $\|I - RA\|_\infty < 1$ can be verified with mathematical rigor. Another possibility is the use of directed rounding for computing actual lower and upper bounds for the term on the left-hand side. The latter approach typically yields better results and it is implemented within INTLAB [31,32]. Usually, the entries of $I - RA$ are not representable as a floating-point matrix. And even if this would be the case, the same can be rarely said for all intermediate results. As a consequence, if successful, a verification method typically not only proves the regularity solely of the given matrix A but verifies regularity of all matrices in an ε-neighborhood of A.

Second, consider the problem to prove singularity of a matrix, which is mathematically the opposite of regularity. If a computation is not entirely exact, a certificate that A is singular cannot be mathematically sound because every ε-neighborhood of a singular matrix contains a regular one. Therefore the problem of deciding singularity of a matrix is ill-posed in the sense of

Hadamard [6] and outside the scope of verification methods. Similarly, the problem to verify that a given matrix has a certain rank deficiency is ill-posed.

The approach as described above may be applied to an interval matrix $\mathbf{A}$, in that case verifying that every real matrix A within $\mathbf{A}$ is regular. As has been shown by Poljak and Rohn [30] that problem is NP-hard; it is completely outside the scope of numerical methods.

Verification method produce mathematically rigorous results. That means that provided the computer arithmetic works to its specifications and the implementation is correct, the result of a verification method is true - like a mathematical theorem. The same is true, for example, for computer algebra methods. In verification methods, the benefit of being fast is traded against the limitation of scope to well-posed problems.

There are other methods to increase the reliability of numerical results. For example, in [40, 13], input data is stochastically perturbed to receive some information on the sensitivity of the problem. From that conclusions are drawn on the accuracy of the computed results. Also regarding rounding error analysis for standard numerical algorithms, it is well known that error bounds are worst case and hardly achieved in practice. A novel approach to a probabilistic rounding error analysis is given in [8]. All those methods are valuable, however, they do not provide mathematically rigorous error bounds.

In the following we shortly write "verified bounds" for mathematical rigorous bounds obtained by a verification method. Similarly, we use the term "inclusion for A" referring to an interval quantity that contains A. The lower and upper bounds of this interval are again obtained by a verification method.

It may come as a surprise that we approach the ill-posed problem of rank deficiency by verification methods based on standard floating-point arithmetic. However, we do not attempt to actually prove rank deficiency of a given matrix which, for the reasons discussed above, is outside of the scope of verification methods. Instead we discuss methods to compute tight inclusions for a nearby rank deficient matrix in the following sense: Given a matrix A and an integer k, we present verification methods to compute rigorous bounds for a perturbation of A such that the perturbed matrix has at least rank deficiency k.

By this different approach we circumvent the ill-posedness of the underlying problem. The concept of verifying that a nearby problem has a certain property is well known; for instance, in [42, 17, 35, 20] such techniques are described for systems of nonlinear equations and in [16] for saddle points.

Finally, we aim to compute an inclusion for a specific perturbation with minimal distance to A. To that end, we will see that a lower bound on a gap between singular values is necessary. Based on the first approach, we present a method for computing an inclusion for a perturbation E with minimal distance with respect to any unitarily invariant norm. The theoretical basis is a generalization of Wedin's $\sin(\theta)$ theorem [43] by removing its orthogonality assumption. Our result is the singular vector subspace counterpart to Davis and Kahan's generalized $\sin(\theta)$ theorem for eigenvector subspaces [2].

All of our results can be implemented solely using standard floating-point operations in rounding to nearest with suitable error estimates. Another convenient way is to use interval operations which will be used in the following.

Algorithms are explored by code written in INTLAB [31], the Matlab/Octave toolbox for Reliable Computing. The short code snippets given in the following section are basically self-explanatory. To understand them, not much knowledge of interval arithmetic is necessary. Basically, it suffices to know that for given A, B and an operation $\circ$, where at least one of A, B is an interval quantity, the result C of the induced interval operation $A \circ B$ satisfies $a \circ b \in C$ for all $a \in A$ and all $b \in B$. For more details, see [26,18,32,28].

Interval quantities are in bold-face, and an interval matrix $\mathbf{A}$ is said to have full rank if all $A \in \mathbf{A}$ have that property. For a matrix A, its Hermitian is denoted by A^*, and its Moore-Penrose pseudoinverse by A^+.

2 A nearby matrix of at least rank deficiency *k*

Let an $m \times n$ matrix A be given, and assume $m \geqslant n$. Furthermore, let k be an integer with $1 \leqslant k \leqslant n$. We aim to find a perturbation E of A such that $A - E$ has at least rank deficiency k.

The natural approach uses the singular value decomposition

$$A = U\Sigma V^* = \sum_{i=1}^{n} \sigma_i(A) \cdot (Ue_i)(Ve_i)^*$$

with $\sigma_1(A) \geqslant \cdots \geqslant \sigma_n(A)$ denoting the singular values of A in the usual order, and e_i denoting the ith column of the identity matrix of appropriate size. The following well-known result is referred to as Eckart-Young-Mirsky theorem [21]. According to [39] it was first published by E. Schmidt [37].

Theorem 2.1 *For some natural number $s \leqslant n$, consider the approximation problem*

$$\min_{B \in \mathbb{C}^{m\times n}} \quad \{\|A - B\| \colon \ \operatorname{rank}(B) \leqslant s\}. \tag{2.1}$$

Regardless of the choice of the unitarily invariant matrix norm $\|\cdot\|$,

$$\widehat{B} = \sum_{i=1}^{s} \sigma_i(A) \cdot (Ue_i)(Ve_i)^* \tag{2.2}$$

is a solution to problem (2.1).

Theorem 2.1 identifies

$$E = \sum_{i=n-k+1}^{n} \sigma_i(A) \cdot (Ue_i)(Ve_i)^* \tag{2.3}$$

as a possible choice of a minimum unitarily invariant norm perturbation of A giving at least rank deficiency k with $\|E\|_2 = \sigma_{n-k+1}$. Unitarily invariant norms of particular interest for us are the spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$. It is noteworthy that the choice given in (2.3) is the unique minimum regarding the Frobenius norm. On the other hand, if $\sigma_{n-k+1} > 0$ and $k < n$, it is easy to construct various perturbations of E with the same minimal spectral norm $\|E\|_2$ and $\operatorname{rank}(A - E) \leqslant n - k$.

2.1 Verification of the 2-norm distance to rank deficiency k

Standard perturbation results for singular values as in [29,5] can be used to obtain verified error bounds for the spectral norm of E in (2.3); refined bounds are given, for example, by [9, Theorem 4.5], [24, Theorem 3]. The following bound by Lange [19] is particularly suitable for our verification purpose for several reasons.

It is applicable for non-orthogonal approximations of the singular vectors and thereby avoids the computation of verified inclusions for actual orthogonal approximations. Moreover, the approach is comparably efficient since each component of the residuals R and S in the following theorem can be computed by a single dot product instead of triple products of matrices. If a higher accuracy is needed there are several methods to compute verified bounds for dot products accurately and efficiently, for instance, [27,4,14,3,23].

Theorem 2.2 *Let matrices $A \in \mathbb{C}^{m\times n}, H \in \mathbb{C}^{k\times k}$ be given with $m \geqslant n \geqslant k$. Denote the singular values of H by $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_k$. For some $Y \in \mathbb{C}^{m\times k}, X \in \mathbb{C}^{n\times k}$ define*

$$R := AX - YH \qquad \text{and} \qquad S := A^*Y - XH^*. \tag{2.4}$$

Then there is a subset of k singular values $\sigma_{i_1}, \ldots, \sigma_{i_k}$ of A such that

$$\max_{1\leqslant j\leqslant k} |\sigma_{i_j} - \theta_j| \leqslant \frac{\max\{\|R\|_2, \|S\|_2\}}{\min\{\sigma_{\min}(X), \sigma_{\min}(Y)\}}\ . \tag{2.5}$$

Natural choices for X, Y, H are suggested by an approximate singular value decomposition $A \approx \tilde{U}\tilde{\Sigma}\tilde{V}^*$: H is the diagonal matrix of the k smallest singular value approximations in the diagonal part of $\tilde{\Sigma}$, and X, Y are chosen accordingly.

The norms of the residuals R, S can be expected to be small, whereas X, Y are almost orthogonal with singular values close to 1. As a consequence, the upper bound in (2.5) can be expected to be small. Both the spectral norm from above and the smallest singular value from below can be bounded by standard verification methods as in [33]. For the former, $\|R\|_2 \leqslant \sqrt{\|R\|_\infty \|R\|_1}$ is faster to compute and may be sufficient. Moreover, the key quantities in (2.4) can be computed to high accuracy using compensated algorithms [15,25,38], double-double arithmetic [1] or other methods based on error-free transformations [36].

As a result we obtain narrow bounds for k distinct but not necessarily mutually different singular values of A close to the diagonal elements of B. These need not to be the set of k smallest singular values of A, but the largest of the singular values bounded via (2.5) is always an upper bound for $\sigma_{n-k+1} = \|E\|_2$. Moreover, the inclusions for k distinct singular values enable us to compute verified bounds for other unitarily invariant norms such as the Frobenius norm.

The 2-norm bound on E can be used as a radius around A, such that the corresponding interval matrix contains the solution $\hat{B}$ in (2.2). This approach can be sensible if A is nearby a matrix with rank deficiency k, i.e., $\|E\| = \|A-\hat{B}\|$ is very small. On the other hand, for larger distances to rank deficiency k, the inclusions based on norm bounds are too wide to be of practical use. In the following subsections we discuss suitable alternatives.

2.2 Verification of a perturbation for at least rank deficiency k

Rather than verifying only the distance to a nearest matrix of rank deficiency k in some measure, we are now interested in verified bounds for an actual perturbation producing a specified rank deficiency.

As before let a matrix $A \in \mathbb{C}^{m \times n}$ with $m \geqslant n$ and an integer k with $1 \leqslant k \leqslant n$ be given. One of the simplest though rarely sensible way to approximate A by a matrix with at least rank deficiency k is to choose k columns of A and set all corresponding elements to zero. Typically, one would choose the columns of A that have relatively small distance-to-zero measures. Better results may be obtained, for instance, by applying an approximate LU-decomposition with partial pivoting, resetting the k rows of U with smallest distance-to-zero measures, and computing the rank deficient approximation as the product of the approximate L and the reduced approximate U.

The two previous approaches can be realized efficiently even for very large and possibly sparse matrices. Nevertheless, in the context of a unitarily invariant norm measurement as in Theorem 2.1, the computed results will typically be far from optimal. For a better approximation we follow the natural approach and exploit the singular value decomposition of A.

Let $A \approx \tilde{U}\tilde{\Sigma}\tilde{V}^*$ be an economy-size approximate singular value decomposition, where $\tilde{U} \in \mathbb{C}^{m \times n}$, $\tilde{V} \in \mathbb{C}^{n \times n}$ with $\tilde{U}^*\tilde{U} \approx I \approx \tilde{V}^*\tilde{V}$ and $\tilde{\Sigma} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with non-negative entries. Denote by $\tilde{\Sigma}_k$ the matrix derived from $\tilde{\Sigma}$ by setting k smallest diagonal elements to zero, then $B = \tilde{U}\tilde{\Sigma}_k\tilde{V}^*$ has at least rank deficiency k.

How close the distance between A and our approximate B comes to the optimal distance $\|E\|$ depends on the quality of the approximate singular value decomposition. If a stable algorithm is used for the computation of $\tilde{U}, \tilde{\Sigma}, \tilde{V}$, then $\|A - B\| \gtrapprox \|A - \hat{B}\| = \|E\|$. An inclusion Δ_0 for an actual perturbation realizing rank deficiency k can be computed using the following INTLAB code.[1][2]

[1] The extra parameter zero in the call of `svd` indicates to compute the economy-size singular value decomposition.

[2] Note that the typecast of `S(I,I)` to interval forces all operations in the last line to be to be interval operations, so that `Delta0` is an inclusion for the desired quantity.

```
[U,S,V] = svd(A,0);
I = 1:n-k;
Delta0 = A - U(:,I)*intval(S(I,I))*V(:,I)';
```

A major issue of this approach is that B contains the significant parts of A by which the intermediate results in $\tilde{U}\tilde{\Sigma}_k\tilde{V}^*$ are relatively large. Hence by evaluating $\tilde{U}\tilde{\Sigma}_k\tilde{V}^*$ with all rounding errors taken into account, we introduce intervals with comparably large diameter. Moreover, the overall accuracy depends on the accuracy of all three approximates $\tilde{U}, \tilde{\Sigma}, \tilde{V}$. In the following we introduce a method circumventing these issues.

The concept originates from the following statement. If $(A - \Delta)X = 0$ for a matrix X with rank k and arbitrary $\Delta \in \mathbb{C}^{m\times n}$, then $A - \Delta$ has at least rank deficiency k.

In this regard, instead of looking for a solution to the approximation problem (2.1), we fix X and consider the alternative problem:

$$\min_{B\in\mathbb{C}^{m\times n}} \{\|A - B\| \colon BX = 0\}. \tag{2.6}$$

The following result is probably known; here we state it together with a short proof. Again $\|\cdot\|$ may be any unitarily invariant matrix norm.

Lemma 2.1 *Let $A \in \mathbb{C}^{m\times n}$ and $X \in \mathbb{C}^{n\times \ell}$ with $m \geqslant n \geqslant \ell$ be given. Then, regardless of the choice of the unitarily invariant matrix norm,*

$$\hat{B}_{\mathrm{x}} := A(I - XX^+) \tag{2.7}$$

is a solution to problem (2.6)*.*

Proof Let $Q = [Q_1\ Q_2]$ be a unitary matrix where the columns of Q_1 span the range of X, so that there exists a matrix Z satisfying $Q_1 = XZ$. Let B be a solution of (2.6). Then $BQ_1 = BXZ = 0$ and [21, Lemma 3] yield

$$\|A - B\| = \|(A - B)Q\| = \|[AQ_1\ \ (A - B)Q_2]\| \geqslant \|AQ_1\|.$$

On the other hand,

$$\|A - \hat{B}_{\mathrm{x}}\| = \|AXX^+\| = \|AQ_1Q_1^*\| = \|AQ_1\|,$$

so that $\hat{B}_{\mathrm{x}}$ satisfies the previous inequality with equality and is therefore optimal. □

By Lemma 2.1 a favorable choice for a perturbation of A is $\Delta := A - \hat{B}_{\mathrm{x}} = AXX^+$. The perturbation Δ can be computed as the minimum Frobenius norm solution to the underdetermined linear system $X^*\Delta^* = (AX)^*$, the solution of which is

$$\Delta = \left((X^*)^+(AX)^*\right)^* = \left((X^+)^*X^*A^*\right)^* = AXX^+$$

as desired. Given $A \in \mathbb{C}^{m\times n}$ and $X \in \mathbb{C}^{n\times k}$ with $m \geqslant n \geqslant k$, the INTLAB call

```
Delta = verifylss(X',(intval(A)*X)')'
```

computes a verified inclusion for $\varDelta$. If the INTLAB call `x = verifylss(A,b)` successfully returns an interval vector $\mathbf{x}$, this verifies a full rank of A and the returned interval vector contains the minimum 2-norm solution of $Ax = b$. For our code example this implies full rank of X, i.e., $\operatorname{rank}(X) = k$, and that the returned interval matrix contains the minimum Frobenius norm solution to the respective system.

Regarding our specific purpose, that can be simplified by computing error bounds for $\varDelta$ directly.

Lemma 2.2 *Let $A \in \mathbb{C}^{m\times n}$ and $X \in \mathbb{C}^{n\times k}$ with $m \geqslant n \geqslant k$ be given. Abbreviate $\varDelta := AXX^+$ as well as $G := I - X^*X$, and let $\|\cdot\|$ be a unitarily invariant norm. If $\|G\|_2 \leqslant \alpha < 1$, then*

$$\varDelta = AXX^* + F_1 = AX(I+G)X^* + F_2 \quad \text{with} \quad \|F_\nu\| \leqslant \frac{\alpha^\nu}{\sqrt{1-\alpha}}\,\|AX\| \tag{2.8}$$

for $\nu \in \{1, 2\}$. The matrix $A - \varDelta$ has at least rank deficiency k.

Proof The norm estimates are derived by exploiting the compatibility of unitarily invariant norms with the spectral norm, cf. [2]. An immediate consequence of this compatibility is the following inequality:

$$\forall C \in \mathbb{C}^{p\times q}, D \in \mathbb{C}^{q\times r}: \quad \|CD\| \leqslant \|C\| \cdot \|D\|_2. \tag{2.9}$$

The assumption $\|G\|_2 < 1$ implies that X has full rank such that

$$X^+ - X^* = (I - X^*X)(X^*X)^{-1}X^* = GX^+ = GX^* + G^2X^+$$

and therefore

$$\varDelta = AXX^* + F_1 = AX(I+G)X^* + F_2 \quad \text{with} \quad F_\nu := AXG^\nu X^+.$$

Then, using (2.9), we derive $\|F_\nu\| = \|AXG^\nu X^+\| \leqslant \|AX\| \cdot \|G^\nu\|_2 \cdot \|X^+\|_2$. Finally, [33, Lemma 2.2] gives $\|X^+\|_2 \leqslant \frac{1}{\sqrt{1-\alpha}}$ and finishes the proof. □

Note that Lemma 2.2 is true for any unitarily invariant norm; generic choices are the spectral or the Frobenius norm.

In the context of our initial problem (2.1), a natural choice for the columns of X are the right singular vectors of A corresponding to the k smallest singular values. In practice, we have only approximations of these singular vectors (i.e., the tailing k columns of $\tilde{V}$), but we can still expect that $X^*X \approx I$. Our choice of X is numerically orthogonal, so that α is of the order of the relative rounding error unit. It is computationally more efficient to use $\|I - X^*X\|_2 \leqslant \|I - X^*X\|_\infty =: \alpha$ without sacrificing much accuracy.

For element-wise bounds on the entries of $\varDelta$, we can exploit that $|(F_\nu)_{ij}| \leqslant \|F_\nu\|_2$. It is often sufficient to use $\|AX\|_2 \leqslant \sqrt{\|AX\|_1\|AX\|_\infty}$ rather than a tight estimate of the spectral norm based on singular value decomposition. On the other hand, for any matrix T of suitable dimension, we easily derive from the proof of Lemma 2.1 that

$$\|TF_\nu\| \leqslant \frac{\alpha^\nu}{\sqrt{1-\alpha}}\,\|TAX\|. \tag{2.10}$$

In particular

$$|(F_\nu)_{ij}| \leqslant \|e_i^T F_\nu\|_2 \leqslant \frac{\alpha^\nu}{\sqrt{1-\alpha}}\,\|e_i^T AX\|_2 \tag{2.11}$$

for all index pairs i,j. The final estimate is computationally cheap and always at least as good as the previous bound.

In any case, a computed inclusion for $\varDelta$ can be expected to be narrow. Given X, executable INTLAB code is as follows.

```
G = eye(k) - X'*intval(X);
alpha = norm(G,inf);
if alpha<1
    AX = A*intval(X);
    f1 = alpha/sqrt(1-alpha)*vecnorm(AX,2,2);
    Delta1 = AX*X' + midrad(0,mag(f1));
    Delta2 = AX*(X'+G*X') + midrad(0,mag(alpha*f1));
else
    Delta1 = intval(NaN(m,n));
    Delta2 = Delta1;
end
```

Here any expression involving interval quantities is computed with error bounds. Therefore `G` is an interval matrix, and `alpha` is an inclusion for the ∞-norm of all (real or complex) matrices within `G`, in particular of $I - X^*X$.

The ith entry of the vector `f1` is an inclusion for the right-most quantity in (2.11) for $\nu = 1$, and `mag(f1)` is a vector of upper bounds of the maximum absolute values of `f1`. Furthermore, `midrad(0,mag(f))` serves as a constructor for an interval vector that contains all real vectors v satisfying $-$`mag(f1)` $\leqslant v \leqslant$ `mag(f1)` with entry-wise comparison. That interval vector is added to the matrix `AX*X'` in the definition of `Delta1` by using MATLAB's implicit expansion of arrays into compatible sizes.[3] The computation of an inclusion for $\varDelta$ based on F_2 is accordingly.

Failure of the algorithm is only possible if `alpha<1` is not satisfied. In practice that means that the approximately computed matrix X of right singular vectors of A does not satisfy $\|I - X^*X\| < 1$, which seems extremely unlikely.

3 An inclusion for an optimal perturbation

The initial question of proving that a given matrix has at least rank deficiency k is ill-posed and therefore outside the scope of verification methods. By chang-

[3] In previous releases of Matlab `midrad(0,mag(f))` would be replaced by `repmat(midrad(0,mag(f)),1,size(A,2))`.

ing this question into asking for an inclusion of a nearby matrix with at least rank deficiency k, we constructed a well-posed problem that is strongly related to the initial problem. Methods for the computation of a tight inclusion for a matrix with the desired properties are given in the previous section.

However, we may ask not just for the inclusion for a nearby matrix but for a tight inclusion for an optimal solution to the minimization problem (2.1), or, more precisely, for an inclusion for the specific solution given in (2.2). Unlike the approaches in the previous section, such inclusions would always contain the matrix A if A has already rank deficiency k.

In order to compute a tight inclusion for this specific perturbation E defined in (2.3), it is necessary to somehow relate the approximate singular vectors with the actual singular vectors of A. However, the computation of verified bounds for the singular vectors of a matrix becomes ill-posed for multiple singular values.

As an example, consider the 4×4 Hadamard matrix[4] with a simultaneous ε-perturbation in the $(1,1)$- and $(4,4)$-entry, i.e.,

$$H_\varepsilon := \begin{pmatrix} 1+\varepsilon & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1+\varepsilon \end{pmatrix}.$$

For $\varepsilon \in \mathbb{R}$, the symbolic toolbox of Matlab computes the set of eigenvalues of Λ of H_ε to

$$\Lambda = \left\{-2,\ \frac{-\sqrt{16+\varepsilon^2}+\varepsilon}{2},\ \frac{\sqrt{16+\varepsilon^2}+\varepsilon}{2},\ 2+\varepsilon\right\}.$$

Since the absolute values of the eigenvalues are the singular values of H_ε, it follows

$$\sigma_1(H_{\varepsilon>0}) = 2+\varepsilon \quad \text{and} \quad \sigma_1(H_{\varepsilon<0}) = \frac{\sqrt{16+\varepsilon^2}+|\varepsilon|}{2}.$$

For small $\varepsilon \neq 0$ the singular values of H_ε are mutually different, hence the minimal norm perturbation E for $k = 3$ is uniquely determined by (2.3). Moreover, for most Ky-Fan-norms and the Frobenius norm, the solution to (2.1) is unique such that E is the only choice. Using symbolic linear algebra, the perturbations compute to

$$2E_{\varepsilon>0} = \begin{pmatrix} 0 & -2 & -2 & 0 \\ -2 & 2 & -2 & 2 \\ -2 & -2 & 2 & 2 \\ 0 & 2 & 2 & 0 \end{pmatrix} + \mathcal{O}(\varepsilon)$$

[4] The entries of a Hadamard matrix H of order n are in $\{-1,1\}$ such that $\frac{1}{n}H^T H$ is the identity matrix. It is known that the order n must be a multiple of 4 for $n \geqslant 4$, and a famous conjecture states that there exists a Hadamard matrix for all multiples of 4.

and

$$2E_{\varepsilon<0} = \begin{pmatrix} -3 & -1 & -1 & -1 \\ -1 & 1 & -3 & 1 \\ -1 & -3 & 1 & 1 \\ -1 & 1 & 1 & -3 \end{pmatrix} + \mathcal{O}(\varepsilon),$$

respectively. Hence an arbitrarily small simultaneous change of H_{11} and H_{44} causes a drastic change of the optimal perturbation E. Of course, the 2-norm of both perturbations differ by only $\mathcal{O}(\varepsilon)$. And the same statement is true for any unitarily invariant norm. Although rigorous bounds for $\|E\|$ are computed easily, obtaining tight inclusions for the actual perturbation E is much more challenging since the problem is ill-posed.

Clearly, the given example lies outside of the scope of verification methods relying on standard (approximate) floating-point operations. At a second glance, it becomes clear that the example above is very specific. The matrix H_ε for $\varepsilon = 0$ has not just some random clusters of multiple singular values, but the specific pair $\{\sigma_{n-k}(H_0), \sigma_{n-k+1}(H_0)\}$ is clustered. Because these two singular values are equal, the corresponding singular vectors are not unique. Hence there is an ambiguity in the singular vector space belonging to $\{\sigma_{n-k+1}(H_0), \ldots, \sigma_n(H_0)\}$, which leads to the discontinuous change of the optimal perturbation when moving from $H_{\varepsilon>0}$ to $H_{\varepsilon<0}$.

In the remainder of this section we will discuss what can be done in the absence of this ambiguity, i.e., if $\sigma_{n-k}(A)$ and $\sigma_{n-k+1}(A)$ are separated. More precisely, we saw in (2.6) and Lemma 2.1 that for a given X and a given unitarily invariant norm, $\Delta := AXX^+$ is a perturbation such that $B := A - \Delta$ is a minimal norm solution of the alternative problem $BX = 0$. Based on that we compute an inclusion for F_3 such that $E := \Delta + F_3$ is a minimum norm perturbation of A with at least rank deficiency k. We start with the underlying mathematical tool for connecting the given approximations with the actual singular vectors.

In order to do that we need a separation theorem between the singular vector subspaces spanned by the first $n - k$ and the remaining k singular vectors. The earliest and most well-known result for this purpose is Wedin's $\sin(\theta)$ theorem for singular value decomposition. It is inspired by and also implying Davis and Kahan's celebrated $\sin(\theta)$ theorem [2, Theorem 5.1]. In practice, however, we have to take care of the fact that "approximate" bases are not orthogonal. Wedin did not give a counterpart to Davis and Kahan's generalized version of their theorem [2, Theorem 6.1].

The following Lemma 3.1 fills this gap by extending the Davis/Kahan and Wedin $\sin(\theta)$ theorem to approximate left and right singular vectors.

Lemma 3.1 *Let $A \in \mathbb{C}^{m\times n}$, $H \in \mathbb{C}^{q\times q}$, $X \in \mathbb{C}^{n\times q}$, $Y \in \mathbb{C}^{m\times q}$ with $m \geqslant n$ be given. Define the residuals $R := AX - YH$, $S := A^*Y - XH^*$ and let $A = U\Sigma V^*$ be an economy-size singular value decomposition of A with $U \in \mathbb{C}^{m\times n}, \Sigma \in \mathbb{R}^{n\times n}, V \in \mathbb{C}^{n\times n}$ and non-increasing order of singular values (with possible ambiguities in the choice of singular vectors). Furthermore, for some*

$s \in \{1,\dots,n\}$, denote by U_{s} the matrix consisting of the first s columns of U and let V_{s} be accordingly. If there is a δ such that $\sigma_s(A) \geqslant \sigma_1(H)+\delta$, then

$$\delta \cdot \max\{\|V_{\mathrm{s}}^* X\|, \|U_{\mathrm{s}}^* Y\|\} \leqslant \max\{\|R\|, \|S\|\} \tag{3.1}$$

is satisfied for any unitarily invariant norm $\|\cdot\|$. In particular, if $\operatorname{rank}(X)=\ell$ and $P_{\mathrm{x}} := XX^+$ denotes the matrix for the orthogonal projection onto the column space of X, then

$$\delta \|P_{\mathrm{x}} V_{\mathrm{s}}\| \leqslant \frac{\max\{\|R\|, \|S\|\}}{\sigma_\ell(X)}. \tag{3.2}$$

Proof First we extend (2.9) by a lower bound for the unitarily invariant norm $\|\cdot\|$ of the product of two matrices $C \in \mathbb{C}^{p\times q}$ and $D \in \mathbb{C}^{q\times r}$. If $\sigma_q(D) \neq 0$ (and thereby $r \geqslant q$), then (2.9) yields

$$\|C\| = \|CDD^+\| \leqslant \|CD\| \cdot \|D^+\|_2 = \|CD\| \cdot \sigma_q^{-1}(D),$$

so that $\|C\| \cdot \sigma_q(D) \leqslant \|CD\|$. The latter inequality is also evident for $\sigma_q(D)=0$. By combining this inequality with (2.9), we derive

$$\forall C \in \mathbb{C}^{p\times q}, D \in \mathbb{C}^{q\times r}: \quad \|C\| \cdot \sigma_q(D) \leqslant \|CD\| \leqslant \|C\| \cdot \sigma_1(D). \tag{3.3}$$

Alternatively, this can be shown using the proof of [10, Theorem 3.3.16] for the respective singular value inequalities together with the fact that any unitarily invariant matrix norm is a symmetric gauge function of the singular values of the respective matrix [41].

For $\delta \leqslant 0$ the inequalities (3.1) and (3.2) are trivially true. Henceforth assume $\delta > 0$. Denote by $\Sigma_{\mathrm{s,s}} = U_{\mathrm{s}}^* A V_{\mathrm{s}} \in \mathbb{R}^{s\times s}$ the diagonal matrix consisting of the s largest singular values of A. Then $A = U\Sigma V^*$ implies $U_{\mathrm{s}}^* A = \Sigma_{\mathrm{s,s}} V_{\mathrm{s}}^*$ and $AV_{\mathrm{s}} = U_{\mathrm{s}}\Sigma_{\mathrm{s,s}}$. By (3.3) and the triangle inequality, we have

$$\|R\| = \sigma_1(U_{\mathrm{s}}^*) \cdot \|R\| \geqslant \|U_{\mathrm{s}}^* R\| = \|\Sigma_{\mathrm{s,s}} V_{\mathrm{s}}^* X - U_{\mathrm{s}}^* YH\| \geqslant \|\Sigma_{\mathrm{s,s}} V_{\mathrm{s}}^* X\| - \|U_{\mathrm{s}}^* YH\|.$$

If $\|V_{\mathrm{s}}^* X\| \geqslant \|U_{\mathrm{s}}^* Y\|$, then (3.3) yields

$$\|R\| \geqslant \sigma_s(A) \cdot \|V_{\mathrm{s}}^* X\| - \|U_{\mathrm{s}}^* Y\| \cdot \sigma_1(H) \geqslant \sigma_s(A) \cdot \|V_{\mathrm{s}}^* X\| - \|V_{\mathrm{s}}^* X\| \cdot \sigma_1(H),$$

such that, by the assumption $\sigma_s(A) \geqslant \sigma_1(H)+\delta$,

$$\|R\| \geqslant \delta \|V_{\mathrm{s}}^* X\| \geqslant \delta \|U_{\mathrm{s}}^* Y\|.$$

Similarly, for the case $\|V_{\mathrm{s}}^* X\| < \|U_{\mathrm{s}}^* Y\|$ one can show that

$$\|S\| \geqslant \|V_{\mathrm{s}}^* S\| \geqslant \|\Sigma_{\mathrm{s,s}} U_{\mathrm{s}}^* Y\| - \|V_{\mathrm{s}}^* X H^*\| \geqslant \delta \|U_{\mathrm{s}}^* Y\| \geqslant \delta \|V_{\mathrm{s}}^* X\|.$$

Combining theses inequalities proves (3.1). Finally, starting with (3.1),

$$\|P_{\mathrm{x}} V_{\mathrm{s}}\| = \|(P_{\mathrm{x}} V_{\mathrm{s}})^*\| = \|V_{\mathrm{s}}^* X X^+\| \leqslant \|V_{\mathrm{s}}^* X\| \cdot \sigma_1(X^+) = \|V_{\mathrm{s}}^* X\| \cdot \sigma_\ell^{-1}(X)$$

yields (3.2) and completes the proof. □

Following Theorem 2.1 we identified in (2.3) the matrix

$$E = \sum_{i=n-k+1}^{n} \sigma_i(A) \cdot (Ue_i)(Ve_i)^*$$

as a possible choice of a minimum norm perturbation of A giving at least rank deficiency k with $\|E\|_2 = \sigma_{n-k+1}$. The inequality in (3.2) is then the key to compute narrow error bounds for that matrix E. Note that theoretically the number q of columns of X may be larger than the desired rank deficiency k, but in practice always $q = \ell = k$.

Theorem 3.1 *With the notation for the matrices A, H, X, Y, R and S as in Lemma 3.1, abbreviate $\varDelta := AXX^+$. Assume X has rank k for some $k \in \{1, \dots, n-1\}$ and that there is a δ with $\sigma_{n-k}(A) \geqslant \sigma_1(H) + \delta$. Let further denote E a minimum norm perturbation of A according to* (2.3)*. Then*

$$E = \varDelta + F_3 \quad \textit{with} \quad \delta\|F_3\| \leqslant \frac{\max\{\|R\|_2, \|S\|_2\}}{\sigma_k(X)} \cdot \|A\| \tag{3.4}$$

is satisfied for any unitarily invariant norm $\|\cdot\|$.

Proof Define $s := n-k$ and denote by $A = U\varSigma V^*$ an economy-size singular value decomposition of A with non-increasing order of singular values. Let $V = [V_{\mathrm{s}}, V_{\bar{\mathrm{s}}}]$ denote the partitioning of V into the singular vectors corresponding to the s largest singular values and their orthogonal complement corresponding to the $n-s=k$ smallest singular values, respectively. Using $V_{\mathrm{s}}V_{\mathrm{s}}^* + V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^* = VV^* = I$, we derive

$$V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^* = P_{\mathrm{x}} - P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^* + (I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*$$

and, in particular,

$$E = U_{\bar{\mathrm{s}}}\varSigma_{\bar{\mathrm{s}},\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^* = AV_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^* = AP_{\mathrm{x}} \underbrace{- A(P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^* - (I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*)}_{=F_3}.$$

By orthogonality of the column and row spaces of $P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^*$ and $(I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*$, respectively, we have

$$\|P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^* - (I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*\|_2 = \max\{\|P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^*\|_2, \|(I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*\|_2\}.$$

Moreover, $\mathrm{rank}(X) = k = \mathrm{rank}(V_{\bar{\mathrm{s}}})$ and [43, Eq. (4.2)] yield

$$\|(I - P_{\mathrm{x}})V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*\| = \|P_{\mathrm{x}}(I - V_{\bar{\mathrm{s}}}V_{\bar{\mathrm{s}}}^*)\| = \|P_{\mathrm{x}}V_{\mathrm{s}}V_{\mathrm{s}}^*\| = \|P_{\mathrm{x}}V_{\mathrm{s}}\|.$$

Finally, using (2.9) and (3.2), the estimate in (3.4) follows. □

To compute verified bounds for the optimal perturbation E, we may now combine Lemma 2.2 to compute an inclusion for $\mathit{\Delta}$ with the bound on $F_3 = E - \mathit{\Delta}$ in Theorem 3.1. As in (2.8) the occurrence of $\sigma_k(X)$ in (3.4) can be replaced by $\sqrt{1-\alpha}$.

If we are just interested in element-wise bounds for the entries of E, we can modify the estimate (3.4) similar as in (2.11). To be precise,

$$\delta|(F_3)_{ij}| \leqslant \frac{\max\{\|R\|_2, \|S\|_2\}}{\sigma_k(X)} \cdot \|e_i^T A\|_2 \tag{3.5}$$

holds valid for all possible index pairs i, j. Rigorous inclusions for these bounds are typically easier and more efficient to compute. Additionally, they are at least as tight as the straightforward bounds using the spectral norm of A.

Another possible improvement arises from the specific form of H in the residuals R and S. In practice X, Y and H are extracted from an approximate singular value decomposition of A. Hence H is a diagonal matrix and Lemma 3.1 can be applied to each column of X and Y individually:

$$(\sigma_s(A) - e_i^* H e_i) \max\{\|V_\mathrm{s}^* X e_i\|_2, \|U_\mathrm{s}^* Y e_i\|_2\} \leqslant \max\{\|Re_i\|_2, \|Se_i\|_2\}$$

for all $i \in \{1, \ldots, k\}$. In particular, if $\sigma_s(A) > e_i^* H e_i$ for all indices i, then

$$\|V_\mathrm{s}^* X\|_2 \leqslant \|V_\mathrm{s}^* X\|_\mathrm{F} \leqslant \sqrt{\sum_{i=1}^{k} \frac{\max\{\|Re_i\|_2^2, \|Se_i\|_2^2\}}{(\sigma_s(A) - e_i^* H e_i)^2}}.$$

Depending on the distribution of the diagonal entries of H, the right-hand side may yield a better bound than $\max\{\|R\|_2, \|S\|_2\}/(\sigma_s(A) - \sigma_1(H))$. By combining this inequality with (3.5) using the argument for Theorem 3.1, we derive

$$|(F_3)_{ij}| \leqslant \sqrt{\sum_{q=1}^{k} \frac{\max\{\|Re_q\|_2^2, \|Se_q\|_2^2\}}{(\sigma_s(A) - e_q^* H e_q)^2}} \cdot \sigma_k^{-1}(X) \cdot \|e_i^T A\|_2, \tag{3.6}$$

which avoids the computation of tight spectral norm bounds and typically still leads to tighter inclusions.

The crucial part in the application of Theorem 3.1 as well as the modified estimate (3.6) is to compute a lower bound for $\sigma_s(A)$. The authors do not know of a method to realize this without computing lower bounds for all s largest singular values $\sigma_1(A), \ldots, \sigma_s(A)$. Once again, we can apply Theorem 2.2 for computing these bounds. Possible efficiency and accuracy improvements to this approach can be taken from [22]. The authors of [22] were concerned with verified bounds for eigenvalues of symmetric matrices, but the same ideas can be applied to our problem.

4 Numerical results

For fixed dimensions $m = 1000$ and $n = 300$ we construct floating-point matrices of numerical rank deficiency r and comput verified inclusions $\varDelta_\nu$ of a perturbation producing a matrix with at least rank deficiency k. We compare the straightforward approach based on an inclusion $\varDelta_0 \ni A - \tilde{U}\tilde{\Sigma}_k\tilde{V}^*$ with the two inclusions $\varDelta_1$ and $\varDelta_2$ determined by F_1 and F_2 in Lemma 2.2, respectively.

In Table 4.1 we consider the maximum imprecision of the computed inclusion $\varDelta_\nu$ relative to the magnitude of $\varDelta_\nu$ itself. For this purpose, we set the radii of the inclusions in relation to the maximum absolute value of the entries in the corresponding row and column of $\varDelta_\nu$. To be precise, for an interval inclusion $\varDelta$ we measure the inaccuracy via

$$\varrho(\varDelta) = \max_{ij} \left\{ \mathrm{rad}\varDelta_{ij} / \min\{\max_{\ell} |\varDelta_{i\ell}|, \max_{\ell} |\varDelta_{\ell j}|\} \right\}.$$

The presented results are the median of 100 samples. In column 6 and 7 of Table 4.1 we present the median of the ratios ϱ_0/ϱ_1 and ϱ_1/ϱ_2 (not the quotients of the respective medians). The bounds for F_ν are not computed by applying Lemma (2.2) directly, but instead we exploit the transformation (2.10) for the element-wise bounds given thereafter.

Table 4.1: Accuracy of the inclusions relative to $\varDelta$.

k	r	ϱ_0	ϱ_1	ϱ_2	ϱ_0/ϱ_1	ϱ_1/ϱ_2
1	0	2.25e-07	5.33e-09	5.33e-09	32.5	1.00
2	0	2.26e-09	1.13e-10	1.13e-10	18.9	1.00
2	1	6.52e-08	3.03e-09	3.03e-09	19.6	1.00
3	0	3.38e-10	2.17e-11	2.17e-11	14.6	1.00
3	1	1.17e-09	7.28e-11	7.28e-11	15.6	1.00
3	2	6.64e-08	3.71e-09	3.71e-09	15.0	1.00
4	0	1.12e-10	1.16e-11	1.15e-11	9.04	1.00
4	1	2.11e-10	2.54e-11	2.53e-11	8.89	1.00
4	2	6.95e-10	6.78e-11	6.77e-11	9.67	1.00
4	3	4.42e-08	4.32e-09	4.32e-09	9.96	1.00
5	0	5.31e-11	6.70e-12	6.65e-12	7.84	1.01
5	1	9.21e-11	1.22e-11	1.21e-11	7.83	1.00
5	2	1.51e-10	1.78e-11	1.77e-11	8.20	1.00
5	3	5.87e-10	8.04e-11	8.04e-11	7.93	1.00
5	4	3.48e-08	3.67e-09	3.67e-09	8.54	1.00

If the numerical rank deficiency r of A is greater than or equal to k, then $\varDelta_\nu$ will be very small and measuring the imprecision relative to $\varDelta_\nu$ makes little sense. In Table 4.2 we therefore measure the accuracy relative to the matrix A:

$$\mu(\varDelta, A) = \max_{ij} \left\{ \mathrm{rad}\varDelta_{ij} / \min\{\max_{\ell} |A_{i\ell}|, \max_{\ell} |A_{\ell j}|\} \right\}.$$

The inclusions based on Lemma 2.2 are roughly by a magnitude better than the bounds by the straightforward approach based on an inclusion

Table 4.2: Accuracy of the inclusions relative to A.

k	r	μ_0	μ_1	μ_2	μ_0/μ_1	μ_1/μ_2
1	1	1.42e-14	8.26e-16	8.26e-16	17.3	1.00
1	2	1.41e-14	8.39e-16	8.39e-16	16.9	1.00
1	3	1.41e-14	8.20e-16	8.20e-16	17.3	1.00
1	4	1.42e-14	8.50e-16	8.50e-16	16.6	1.00
2	2	1.41e-14	1.09e-15	1.09e-15	13.0	1.00
2	3	1.41e-14	1.09e-15	1.09e-15	13.0	1.00
2	4	1.42e-14	1.09e-15	1.09e-15	13.3	1.00
3	3	1.40e-14	1.29e-15	1.29e-15	10.9	1.00
3	4	1.42e-14	1.29e-15	1.29e-15	10.9	1.00
4	4	1.42e-14	2.35e-15	2.35e-15	6.06	1.00

$\varDelta_0 \ni A - \tilde{U}\tilde{\Sigma}_k\tilde{V}^*$. On the other hand, the two estimates in Lemma 2.2 lead to very similar bounds; only if k is strictly larger than the numerical rank deficiency of A computational tests suggest that the second bound is slightly superior. Although the additional computational effort for the estimate with F_2 is relatively small, the overall improvement is so insignificant that we suggest to use the simpler estimate obtained by the inclusion for F_1.

The estimate for the error term F_2 is approximately of the order α^2, which in turn is close to the relative rounding error unit squared. Thus the accuracy of the inclusion $\varDelta_2$ based on F_2 is practically equal to that of the approximate part $AX(I+G)X^*$. Our numerical results suggest that similarly the accuracy of the approximate part AXX^* dictates the overall accuracy ϱ_1 when using the estimate based on F_1.

In Table 4.3 we present numerical results for minimal distance perturbation inclusions based on Theorem 3.1 with the modified estimate (3.6). For the same dimensions as before, $m = 1000$ and $n = 300$, we constructed matrices with numerically fixed difference δ between the singular values σ_{n-4} and σ_{n-3}. The other singular values and the corresponding singular vectors are chosen randomly. The inclusion for E defined in (2.3) is computed for $k = 4$. As before the presented results are the median of 100 samples.

Table 4.3: Accuracy of the inclusions for optimal perturbation E from (2.3).

δ/σ_1	ϱ_3	δ/σ_1	ϱ_3	δ/σ_1	ϱ_3
1e-01	3.27e-10	1e-06	4.12e-08	1e-11	1.35e-05
1e-02	6.92e-10	1e-07	1.29e-07	1e-12	4.34e-05
1e-03	1.54e-09	1e-08	4.22e-07	1e-13	1.40e-04
1e-04	4.44e-09	1e-09	1.38e-06	1e-14	8.84e-04
1e-05	1.33e-08	1e-10	4.58e-06	1e-15	1.00

Our results demonstrate the correlation of the quality of the inclusion and the distance between the crucial singular values. If the relative distance $\frac{\sigma_{n-4}-\sigma_{n-3}}{\sigma_1}$ is about 10^{-15} or less, our verification method can no longer separate the respective singular values. Thus, Theorem 3.1 is not applicable and the inclusion falls back to a coarse interval inclusion based on the estimate $|E| \leqslant \sigma_{n-3}(A)$.

This happens also with the 4×4 Hadamard matrix example in the previous subsection. Consider a perturbation of H_{11} into $1-\varepsilon$ with $\varepsilon = 2^{-53}$. This is the smallest perturbation in double precision floating-point arithmetic.

The spectrum of the perturbed matrix computes symbolically, up to order $\mathcal{O}(\varepsilon^2)$, to $\Lambda = \{-2,\ -2-\frac{\varepsilon}{4},\ 2,\ 2-\frac{3\varepsilon}{4}\}$. Thus, for small enough ε, the smallest singular value is $2-\frac{3\varepsilon}{4}$, slightly less than 2. It has multiplicity one with (up to the sign) unique singular vector v, and according to (2.3) the minimum Frobenius norm perturbation for the nearest singular matrix is uniquely defined by $(2-\frac{3\varepsilon}{4})vv^T$. Symbolically it computes to (up to 4 figures):

$$E = \begin{pmatrix} 1.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.1667 & 0.1667 & 0.1667 \\ 0.5000 & 0.1667 & 0.1667 & 0.1667 \\ 0.5000 & 0.1667 & 0.1667 & 0.1667 \end{pmatrix}.$$

On the other hand, the verified inclusion Δ for $k = 1$ by Lemma 2.2 using MATLAB's approximate singular value decomposition and INTLAB is:

$$\Delta = \begin{pmatrix} 1-\varepsilon & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

In that specific case no rounding error occurs, so that lower and upper bound of the inclusion Δ coincide. The perturbation Δ produces obviously a rank deficient matrix, its norm is almost equal to $\|E\|_2$, but it is far away from the optimal perturbation E.

Our computed inclusion can be expected to compute narrow error bounds for two reasons. First, a computed approximation X is numerically orthogonal so that α is close to the relative rounding error unit. Even if that would not be the case, larger values of α up to about 10^{-8} do not influence the quality of the approximation too much.

Second, if the gap between σ_{n-k} and σ_{n-k+1} is small, the problem of finding a basis for the k-dimensional subspace to the smallest k singular values of A becomes ill-conditioned. Thus it becomes more and more difficult to compute tight inclusions for an optimal $\|\cdot\|$-norm perturbation. Nevertheless the error estimates (2.8) for the norm of F_ν remain small. That is because neither of the involved quantities is influenced by that fact.

5 Conclusion

Given a matrix and some integer k, the property that the rank deficiency is at least k is an ill-posed problem. Also without specifying the degree of rank deficiency the problem remains ill-posed. Beyond that, to compute a minimum norm perturbation realizing that rank deficiency can be ill-posed as well.

In this note we showed a simple method to compute narrow error bounds for a perturbation containing a matrix with at least rank deficiency k. It is shown that our method yields tighter inclusions than the straightforward approach and that usually narrow error bounds can be expected.

Moreover, we introduced a method to compute inclusions for an optimal perturbation with respect to some unitarily invariant norm. For this purpose we generalized Wedin's $\sin(\theta)$ theorem. The method fails if the respective singular vector spaces cannot be sufficiently separated from each other. Then the problem is too ill-conditioned to compute tight inclusions for an optimal perturbation via a verification method relying on standard floating-point arithmetic.

By computing bounds for a perturbation with the desired property, the principle problem of verification methods that ill-posed problems are outside their scope is circumvented. Similar techniques for systems of nonlinear equations are known, and we hope that other problems will follow.

Related to rank deficiency are other problems. For example: is there a method to compute verified error bounds for the k-th singular value of a matrix without calculating bounds for all singular values? The same question arises for eigenvalues of symmetric or Hermitian matrices.

Are there efficient methods for calculating bounds, in particular a lower bound for the smallest singular value of a matrix? For symmetric positive definite problems such methods are available [32], what about general symmetric matrices? That was posed as Challenge 10.15 in [32]. More precisely, the matrix should be large and sparse, and with condition number beyond 10^8 to avoid using A^TA as for binary64. If solved by an efficient algorithm, in particular for sparse matrices, that would be the key to the verified solution of sparse linear systems with not necessarily symmetric positive definite matrix.

Acknowledgements The authors wish to thank Michiel E. Hochstenbach for his extensive efforts and many fruitful and constructive comments.

References

1. David H. Bailey. A Fortran 90-based multiprecision system. *ACM Trans. Math. Softw.*, 21(4):379–387, 1995.
2. Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.
3. James Demmel, Peter Ahrens, and Hong Diep Nguyen. Efficient reproducible floating point summation and blas. Technical Report UCB/EECS-2016-121, EECS Department, University of California, Berkeley, 2016.
4. Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pelissier, and Paul Zimmermann. MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Trans. Math. Softw.*, 33(2), 2007.
5. Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, USA, 4th edition, 2013.
6. Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univ. Bull.*, 13:49–52, 1902.

7. Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*, volume 80 of *Other Titles in Applied Mathematics*. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, Pennsylvania, USA, 2nd edition, 2002.
8. Nicholas J. Higham and Theo Mary. A new approach to probabilistic rounding error analysis. *SIAM J. Sci. Comput.*, 41(5):A2815–A2835, 2019.
9. Michiel E. Hochstenbach. A Jacobi–Davidson type SVD method. *SIAM J. Sci. Comput.*, 23(2):606–628, 2001.
10. Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1st edition, 1991.
11. IEEE. IEEE standard for binary floating-point arithmetic. *ANSI/IEEE Std 754-1985*, pages 1–20, 1985.
12. IEEE. IEEE standard for floating-point arithmetic. *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pages 1–84, 2019.
13. Fabienne Jézéquel and Jean-Marie Chesneaux. CADNA: a library for estimating round-off error propagation. *Comput. Phys. Commun.*, 178(12):933–955, 2008.
14. Fredrik Johansson. Arb: A C library for ball arithmetic. *ACM Commun. Comput. Algebra*, 47(3/4):166–169, 2014.
15. William M. Kahan. A survey of error analysis. In Charles V. Freiman, John E. Griffith, and Jack L. Rosenfeld, editors, *7th IFIP Congress, Ljubljana*, Amsterdam, Netherlands, 1971. North-Holland Publishing Co.
16. Yuchi Kanzawa and Shin'ichi Oishi. Calculating bifurcation points with guaranteed accuracy. *IEICE Trans. Fundamentals*, E82-A(6):1055–1061, 1999.
17. Yuchi Kanzawa and Shin'ichi Oishi. Imperfect singular solutions of nonlinear equations and a numerical method of proving their existence. *IEICE Trans. Fundamentals*, E82-A(6):1062–1069, 1999.
18. Ralph Baker Kearfott, Mitsuhiro T. Nakao, Arnold Neumaier, Siegfried M. Rump, Sergey P. Shary, and Pascal Van Hentenryck. Standardized notation in interval analysis. *Comput. Technol.*, 15(1):7–13, 2010.
19. Marko Lange. Residual bounds for some or all singular values. *Linear Algebra Appl.*, 464:28–37, 2015. Special issue on eigenvalue problems.
20. Nan Li and Lihong Zhi. Verified error bounds for isolated singular solutions of polynomial systems. *SIAM J. Numer. Anal.*, 52(4):1623–1640, 2014.
21. Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *Q. J. Math.*, 11(1):50–59, 1960.
22. Shinya Miyajima, Takeshi Ogita, and Shin'ichi Oishi. Fast verification for respective eigenvalues of symmetric matrix. In Victor G. Ganzha, Ernst W. Mayr, and Evgenii V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing*, pages 306–317. Springer Berlin Heidelberg, 2005.
23. Daichi Mukunoki, Takeshi Ogita, and Katsuhisa Ozaki. Reproducible BLAS routines with tunable accuracy using Ozaki scheme for many-core architectures. In Roman Wyrzykowski, Ewa Deelman, Jack Dongarra, and Konrad Karczewski, editors, *Parallel Processing and Applied Mathematics*, pages 516–527, Cham, Switzerland, 2020. Springer Nature.
24. Yuji Nakatsukasa. Accuracy of singular vectors obtained by projection-based svd methods. *BIT Numer. Math.*, 57(4):1137–1152, 2017.
25. Arnold Neumaier. Rundungsfehleranalyse einiger Verfahren zur Summation endlicher Summen. *Z. Angew. Math. Mech.*, 54(1):39–51, 1974.
26. Arnold Neumaier. *Interval Methods for Systems of Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, England, 1991.
27. Takeshi Ogita, Siegfried M. Rump, and Shin'ichi Oishi. Accurate sum and dot product. *SIAM J. Sci. Comput.*, 26(6):1955–1988, 2005.
28. Shin'ichi Oishi, Kazuhiro Ichihara, Masahide Kashiwagi, Takuma Kimura, Xuefeng Liu, Hidetoshi Masai, Yusuke Morikura, Takeshi Ogita, Katsuhisa Ozaki, Siegfried M. Rump, Kouta Sekine, Akitoshi Takayasu, and Naoya Yamanaka. *Principle of Verified Numerical Computations*. Corona publisher, Tokyo, Japan, 2018. [in Japanese].
29. Beresford N. Parlett. *The Symmetric Eigenvalue Problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, USA, unabridged, corr. republication of the work 1st publ. by Prentice-Hall edition, 1998.

30. Svatopluk Poljak and Jiri Rohn. Checking robust nonsingularity is NP-Hard. *Math. of Control, Signals, and Systems 6*, pages 1–9, 1993.
31. Siegfried M. Rump. INTLAB – INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Springer Netherlands, Dordrecht, 1999.
32. Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numer.*, 19:287–449, 2010.
33. Siegfried M. Rump. Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse. *BIT Numer. Math.*, 51(2):367–384, 2011.
34. Siegfried M. Rump. Gleitkommaarithmetik auf dem Prüfstand. *Jahresber. Dtsch. Math. Ver.*, 118(3):179–226, 2016.
35. Siegfried M. Rump and Stef Graillat. Verified error bounds for multiple roots of systems of nonlinear equations. *Numer. Algorithms*, 54(3):359–377, 2010.
36. Siegfried M. Rump, Takeshi Ogita, and Shin'ichi Oishi. Accurate floating-point summation part I: Faithful rounding. *SIAM J. Sci. Comput.*, 31(1):189–224, 2008.
37. Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Math. Ann.*, 63(4):433–476, 1907.
38. Jonathan R. Shewchuk. Adaptive precision floating-point arithmetic and fast robust geometric predicates. *Discrete Comput. Geom.*, 18(3):305–363, 1997.
39. Gilbert W. Stewart. Perturbation theory for the singular value decomposition. Technical report, University of Maryland at College Park, USA, 1990.
40. Jean Vignes. A stochastic arithmetic for reliable scientific computation. *MATHCS*, 35(3):233–261, 1993.
41. John von Neumann. Some matrix-inequalities and metrization of matrix-space. *Tomsk Univ. Rev.*, 1:286–300, 1937.
42. Helmut Weber and Wilhelm Werner. On the accurate determination of nonisolated solutions of nonlinear equations. *Computing*, 26(4):315–326, 1981.
43. Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numer. Math.*, 12(1):99–111, 1972.