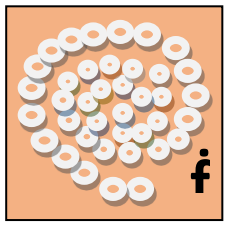

International Journal of Mathematical Analysis and Modelling

(Formerly Journal of the Nigerian Society for Mathematical Biology)

Volume 8 Issue 1, 2025

ISSN (Print): 2682 - 5694
ISSN (Online): 2682 – 5708
www.tnsmb.org

A stable algorithm for computing a nearby defective matrix

R.O. Akinola*

Abstract

Given that A is a matrix with n distinct eigenvalues, in this paper, we describe a stable numerical method for computing a nearby defective matrix to A and the distance between them. The method makes use of the Gauss-Newton method which converges quadratically and only works in the case where the nearby defective matrix is real. Numerical experiments are given which confirms the theory.

Keywords and phrases: defective matrix; Jordan block; sensitivity of eigenproblem; Newton's method.

1 Introduction

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix with n distinct eigenvalues. The aim of this paper, is to describe how to find a nearby defective matrix to $\mathbf{A}$. In more precise terms, we attempt to provide an answer to a question posed by Wilkinson [1] *i. e.*, given a simple matrix $\mathbf{A}$, find $\mathrm{d}(\mathbf{A})$ ([2, p. 274], [1, pp.90-93]);

$$\mathrm{d}(\mathbf{A}) = \inf\{\|\mathbf{A}-\mathbf{B}\| : \quad \mathbf{B} \quad \text{is a defective matrix}\},$$

such that $\mathbf{B}$ is defective. Hence, $\mathrm{d}(\mathbf{A})$ is the distance of $\mathbf{A}$ to the set of matrices which have a Jordan block of at least dimension 2. The main reason for computing a nearby defective matrix to $\mathbf{A}$ is because: if $\mathbf{A}$ has a nearby defective matrix, then it has ill-conditioned eigenvalues.

Alam and Bora [2, p. 284] proved that given $\mathbf{A} \in \mathbb{C}^{n\times n}$, with $\lambda \in \mathbb{C}\backslash\Lambda(\mathbf{A})$ which has to be found, $\mathbf{u}$ and $\mathbf{v}$ are a pair of normalized left and right singular vectors of $\mathbf{A}-\lambda I$ corresponding to the smallest singular value ε such that they are orthogonal, *e*.g., $\mathbf{u}^H\mathbf{v} = 0$, then the nearest defective matrix to $\mathbf{A}$ is given by the formula; $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^H$. It was shown that $\mathbf{u}$ and $\mathbf{v}$ are left and right eigenvectors of $\mathbf{B}$ corresponding to the eigenvalue λ *i*.e., $\mathbf{u}^H\mathbf{B} = \lambda\mathbf{u}^H$, $\mathbf{B}\mathbf{v} = \lambda\mathbf{v}$. Since $\mathbf{u}^H\mathbf{v} = 0$, this implies that λ is a multiple eigenvalue of $\mathbf{B}$, hence $\mathrm{d}(\mathbf{A}) = \varepsilon$. However, their analysis and computational procedure for finding the values of λ and ε and the nearest defective matrix to $\mathbf{A}$ relies on the ε-pseudospectrum of $\mathbf{A}$.

For any $\varepsilon > 0$, the ε-pseudospectrum of $\mathbf{A}$, $\Lambda_\varepsilon(\mathbf{A})$ (to be defined shortly) consists of nontrivial components and the interior of each of its component contains at least one eigenvalue of $\mathbf{A}$. As ε is increased, the components of $\Lambda_\varepsilon(\mathbf{A})$ coalesce and λ is found from the point of coalescence. Though the paper [2], provides the solution to Wilkinson's problem, the algorithm that they gave for the computation of $\mathrm{d}(\mathbf{A})$ and the nearest defective matrix is slow and requires human intervention. Before we continue, lets define what it mean for an eigenvalue of a matrix to be defective.

Definition 1.1. *An eigenvalue is said to be* ***defective*** *if its algebraic multiplicity is greater than its geometric multiplicity. A matrix is said to be defective if it has one or more defective eigenvalues.*

The distance of a simple matrix to a defective matrix is linked with the sensitivity analysis of eigenvalues. The condition number of a simple eigenvalue λ is given by $1/|\mathbf{y}^H\mathbf{x}|$, (see [1]) where $\mathbf{x}$ and $\mathbf{y}$ are normalised right and left eigenvectors respectively corresponding to λ. For a defective eigenvalue, we have $\mathbf{y}^H\mathbf{x} = 0$. Hence, the condition number of the eigenvalue λ, is infinite.

*Department of Mathematics, Faculty of Natural Sciences, University of Jos, Jos, Nigeria; E-mail: roakinola@gmail.com

However, it is well-known that even if the eigenvalues of a matrix are simple and well-separated from each other, they can be ill-conditioned [1]. Hence the measure of the distance $\mathrm{d}(\mathbf{A})$ of a matrix $\mathbf{A}$ to a defective matrix $\mathbf{B}$ is important for determining the sensitivity of an eigendecomposition. There is a very informative discussion and history of this problem in [3], where the contributions of Demmel [4, 5] and Wilkinson [6, 7] are discussed in detail. Another important paper is that by Lippert and Edelman [8], who use ideas from differential geometry and singularity theory to discuss the sensitivity of double eigenvalues. In particular, they present a condition that measures the ill-conditioning of a matrix with a 2-dimensional Jordan block.

Following Trefethen and Embree [9], the ε-pseudospectrum $\Lambda_\varepsilon(\mathbf{A})$ of a matrix $\mathbf{A}$ is given by

$$\Lambda_\varepsilon(\mathbf{A}) = \{\sigma_{\min}(\mathbf{A} - z\mathbf{I}) < \varepsilon\},$$

where $\varepsilon > 0$, $\sigma_{\min}$ denotes the smallest singular value and $z \in \mathbb{C}$. Equivalently,

$$\Lambda_\varepsilon(\mathbf{A}) = \{z \in \mathbb{C} \,|\, \det(\mathbf{A} + \mathbf{E} - z\mathbf{I}) = 0, \text{ for some } \mathbf{E} \in \mathbb{C}^{n\times n} \text{ with } \|\mathbf{E}\| < \varepsilon\}.$$

If $\Lambda_\varepsilon(\mathbf{A})$ has n components, then $\mathbf{A} + \mathbf{E}$ has n distinct eigenvalues for all perturbation matrices $\mathbf{E} \in \mathbb{C}^{n\times n}$ with $\|\mathbf{E}\| < \varepsilon$ and hence, $\mathbf{A} + \mathbf{E}$ is not defective. Alam and Bora [2] take these ideas and seek the smallest perturbation matrix $\mathbf{E}$ such that the pseudospectra of $\mathbf{A} + \mathbf{E}$ coalesce. They present the following theorem (see [2, Theorem 4.1] and [3, Lemma 1]) but the following proof is new, different, clearer and comprehensible.

Theorem 1.2. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $z \in \mathbb{C} \setminus \Lambda(\mathbf{A})$, so that $\mathbf{A} - z\mathbf{I}$ has a simple smallest singular value $\varepsilon > 0$ with corresponding left and right singular vectors $\mathbf{u}$ and $\mathbf{v}$ such that $(\mathbf{A} - z\mathbf{I})\mathbf{v} = \varepsilon\mathbf{u}$. Then z is an eigenvalue of $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^H$ with geometric multiplicity 1 and corresponding left and right eigenvectors $\mathbf{u}$ and $\mathbf{v}$ respectively. Furthermore, if $\mathbf{u}^H\mathbf{v} = 0$, then z has algebraic multiplicity greater than one, hence it is a defective eigenvalue of $\mathbf{B}$ and $\|\mathbf{A} - \mathbf{B}\| = \varepsilon$.*

Proof. First, we want to show that z is an eigenvalue of $\mathbf{B}$. To do this, we subtract $z\mathbf{I}$ from both sides of $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^H$ to obtain

$$\mathbf{B} - z\mathbf{I} = \mathbf{A} - z\mathbf{I} - \varepsilon\mathbf{u}\mathbf{v}^H,$$

and by post multiplying both sides by the right singular vector $\mathbf{v}$ we have,

$$(\mathbf{B} - z\mathbf{I})\mathbf{v} = (\mathbf{A} - z\mathbf{I} - \varepsilon\mathbf{u}\mathbf{v}^H)\mathbf{v},$$

After expanding the right hand side, we have $(\mathbf{B} - z\mathbf{I})\mathbf{v} = (\mathbf{A} - z\mathbf{I})\mathbf{v} - \varepsilon\mathbf{u} = \mathbf{0}$. We arrived at the last equality by using the fact that $(\mathbf{A} - z\mathbf{I})\mathbf{v} = \varepsilon\mathbf{u}$. Since $\mathbf{B} - z\mathbf{I} = \mathbf{0}$, hence, $\mathbf{B}\mathbf{v} = z\mathbf{v}$. In a similar fashion, it can be shown that $\mathbf{u}^H\mathbf{B} = z\mathbf{u}^H$. This shows that $\mathbf{u}$ and $\mathbf{v}$ are left and right eigenvectors corresponding to the eigenvalue z.

Next, we want to show that the geometric multiplicity of $\mathbf{B}$ is one. From

$$(\mathbf{B} - z\mathbf{I})\mathbf{v} = (\mathbf{A} - z\mathbf{I} - \varepsilon\mathbf{u}\mathbf{v}^H)\mathbf{v},$$

and using the singular value decomposition

$$\mathbf{A} - z\mathbf{I} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = \sum_{k=1}^{n} \sigma_k \mathbf{u}_k \mathbf{v}_k^H,$$

where $\mathbf{u} = \mathbf{u}_n$, $\sigma_n = \varepsilon$ and $\mathbf{v} = \mathbf{v}_n$. Thus,

$$\begin{aligned}
\mathbf{B} - z\mathbf{I} &= \sum_{k=1}^{n} \sigma_k \mathbf{u}_k \mathbf{v}_k^H - \varepsilon\mathbf{u}\mathbf{v}^H \\
&= \sum_{k=1}^{n-1} \sigma_k \mathbf{u}_k \mathbf{v}_k^H + \sigma_n \mathbf{u}_n \mathbf{v}_n^H - \varepsilon\mathbf{u}_n\mathbf{v}_n^H \\
&= \sum_{k=1}^{n-1} \sigma_k \mathbf{u}_k \mathbf{v}_k^H \\
&= \mathbf{U}_{n-1}\mathbf{\Sigma}_{n-1}\mathbf{V}_{n-1}^H.
\end{aligned}$$

This shows that the rank of $\mathbf{B} - z\mathbf{I}$ is $n-1$. Now, $\mathbf{B}$ has a geometric multiplicity of 1[1]. Furthermore, we want to show that $\mathbf{u}^H\mathbf{v} = 0$. Let $\hat{\mathbf{v}}$ be the right generalized eigenvector corresponding to the eigenvalue z such that $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$ with $\mathbf{v}^H\hat{\mathbf{v}} = 0$ and $\mathbf{u}^H\hat{\mathbf{v}} \neq 0$. The condition $\mathbf{u}^H\hat{\mathbf{v}} \neq 0$ ensures that $\mathbf{B}$ has a 2-dimensional Jordan block only. Premultiply both sides of $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$ by $\mathbf{u}^H$. This leads to $\mathbf{u}^H(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{u}^H\mathbf{v}$. But since $\mathbf{u}^H(\mathbf{B} - z\mathbf{I}) = \mathbf{0}^H$, this means that $\mathbf{u}^H\mathbf{v} = 0$.

The fact that z is an algebraically double eigenvalue of $\mathbf{B}$ can be seen by multiplying both sides of $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$ by $(\mathbf{B} - z\mathbf{I})$, which implies that

$$(\mathbf{B} - z\mathbf{I})^2\hat{\mathbf{v}} = (\mathbf{B} - z\mathbf{I})\mathbf{v} = \mathbf{0},$$

since $\mathbf{B}\mathbf{v} = z\mathbf{v}$. This shows that indeed the algebraic multiplicity of $\mathbf{B}$ is greater than its geometric multiplicity. Therefore, $\mathbf{B}$ is defective by definition because the algebraic multiplicity of z exceeds its geometric multiplicity.

Theorem 1.2 leads to the result $\mathbf{E} := -\varepsilon\mathbf{u}\mathbf{v}^H$ so that $\mathbf{B} = \mathbf{A} + \mathbf{E}$ is a defective matrix and $\mathrm{d}(\mathbf{A}) = \varepsilon$, since $\mathbf{v}^H\mathbf{v} = \mathbf{u}^H\mathbf{u} = 1$. One drawback of the algorithm in [2] is that it is rather expensive since it involves repeated calculation of pseudospectra. Also a decision on when two pseudospectral curves coalesce is required. In [3], a method based on calculating lowest generalised saddle points of singular values is described.

Our approach to find a nearby defective matrix to a real n by n matrix will be built around the same idea of Alam and Bora [2]. We consider $\mathbf{u}$ and $\mathbf{v}$ as left and right singular vectors of $\mathbf{A} - \lambda I$ corresponding to the smallest singular value ε. This implies that, for $\lambda \in \mathbb{R}$,
$(\mathbf{A} - \lambda I)\mathbf{v} = \varepsilon\mathbf{u}$, $(\mathbf{A} - \lambda I)^T\mathbf{u} = \varepsilon\mathbf{v}$. Since we want to find a nearby defective matrix to $\mathbf{A}$, we impose the further condition that the left and right singular vectors are orthogonal, *i*.e. $\mathbf{u}^T\mathbf{v} = 0$, and that $\mathbf{u}$ and $\mathbf{v}$ are each normalized to have norm one. This results in solving an over-determined nonlinear system of $2n+3$ real equations for the $2n+2$ real unknowns $\mathbf{z} = [\mathbf{v}, \varepsilon, \mathbf{u}, \lambda]^T$, which is a nonlinear least squares problem. The solution of which will be solved using a Gauss-Newton based method (see, for example, [10, ch. 10]).

A nearby[2] defective matrix will then be computed by the formula: $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^T$, where ε is the distance between $\mathbf{A}$ and $\mathbf{B}$. The proposed approach, is very fast, does not rely on the computation of the ε-pseudospectrum of $\mathbf{A}$ and converges quadratically. In addition, it gives rise to the first known algorithm for computing a nearby defective matrix to a simple matrix. However, it suffers the disadvantage that it relies on good initial guesses to ε and λ, and there is no proof that the computed defective matrix is the closest.

The plan of this paper is as follows: in Section 2.1, we present the nonlinear system of equations for finding $\mathrm{d}(\mathbf{A})$ and a nearby defective matrix. Main results are presented. In Section 3, results of numerical experiments are given which supports our theoretical discussion.

2 Methodology

2.1 Nonlinear system for finding $\mathrm{d}(\mathbf{A})$ and a nearby defective matrix

In this section, we present another approach for computing a nearby defective matrix $\mathbf{B}$ from $\mathbf{A}$ and the distance between them which is slightly different from the approach described in Akinola *et al.*, [11]. This approach involves solving an over-determined nonlinear system of $2n+3$ equations in $2n+2$ unknowns. A method based on Gauss-Newton theory for computing a nearby defective matrix to $\mathbf{A}$ and the distance between them, which is more efficient than the approach proposed by Alam of Bora [2], even though it does not guarantee that the computed defective matrix is the nearest. Be that as it may, the method only works when z is real.

[1] $\mathrm{rank}(\mathbf{B} - z\mathbf{I}) = n - \dim\mathcal{N}(\mathbf{B} - z\mathbf{I}) = n - 1$. This means that the dimension of the nullspace of $\mathbf{B} - z\mathbf{I}$ is 1 or $\mathbf{B}$ has a geometric multiplicity of 1.

[2] The term "nearby" is so used in this context rather than "nearest" because there is no proof in our analysis that the computed defective matrix is the nearest. Besides, Newton's method gives no such guarantee.

We begin by presenting the nonlinear system of equations (1), find analytic expressions for the Jacobian and present the key result; Theorem 2.1 which shows that the Jacobian is of full rank and so the method should have quadratic convergence with a close enough guess.

As mentioned earlier in the introduction, the new approach stems from the paper of Alam and Bora [2] in which they defined $\mathbf{u}$ and $\mathbf{v}$ as left and right singular vectors of $\mathbf{A} - z\mathbf{I}$ corresponding to the smallest singular value ε. We build on the same assumption by referring to both $\mathbf{u}$ and $\mathbf{v}$ as left and right singular vectors of $\mathbf{A} - z\mathbf{I}$ corresponding to the smallest singular value ε. This implies that $(\mathbf{A} - z\mathbf{I})\mathbf{v} = \varepsilon\mathbf{u}$ and $(\mathbf{A}^T - z\mathbf{I})\mathbf{u} = \varepsilon\mathbf{v}$. We also assume that the left and right singular vectors are orthogonal, *i.e.*, $\mathbf{u}^T\mathbf{v} = 0$, and that $\mathbf{u}$, $\mathbf{v}$ are each normalized to have norm one. These, leads to solving the following over-determined nonlinear system of $2n + 3$ real equations

$$\begin{aligned} (\mathbf{A} - z\mathbf{I})\mathbf{v} &= \varepsilon\mathbf{u} \\ \frac{1}{2}\mathbf{v}^T\mathbf{v} &= \frac{1}{2} \\ (\mathbf{A}^T - z\mathbf{I})\mathbf{u} &= \varepsilon\mathbf{v} \\ \frac{1}{2}\mathbf{u}^T\mathbf{u} &= \frac{1}{2} \\ \mathbf{u}^T\mathbf{v} &= 0, \end{aligned} \tag{1}$$

in $2n + 2$ real unknowns: $\mathbf{w} = [\mathbf{v}, \varepsilon, \mathbf{u}, z]^T$. After solving for $\mathbf{w}$ in (1), we will then compute a nearby defective matrix $\mathbf{B}$ by the formula $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^T$ in line with Alam and Bora [2]. The computed value of ε is then the distance between the simple matrix $\mathbf{A}$ and the defective $\mathbf{B}$.

In compact form $\mathbf{F}(\mathbf{w}) = \mathbf{0}$, (1) can be expressed as

$$\mathbf{F}(\mathbf{w}) = \begin{bmatrix} (\mathbf{A} - z\mathbf{I})\mathbf{v} - \varepsilon\mathbf{u} \\ -\frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2} \\ (\mathbf{A}^T - z\mathbf{I})\mathbf{u} - \varepsilon\mathbf{v} \\ -\frac{1}{2}\mathbf{u}^T\mathbf{u} + \frac{1}{2} \\ \mathbf{u}^T\mathbf{v} \end{bmatrix} = \mathbf{0}, \tag{2}$$

and the Jacobian of the system of equations (2) is given by

$$\mathbf{F}_{\mathbf{w}}(\mathbf{w}) = \begin{bmatrix} (\mathbf{A} - z\mathbf{I}) & -\mathbf{u} & -\varepsilon\mathbf{I} & -\mathbf{v} \\ -\mathbf{v}^T & 0 & \mathbf{0}^T & 0 \\ -\varepsilon\mathbf{I} & -\mathbf{v} & (\mathbf{A}^T - z\mathbf{I}) & -\mathbf{u} \\ \mathbf{0}^T & 0 & -\mathbf{u}^T & 0 \\ \mathbf{u}^T & 0 & \mathbf{v}^T & 0 \end{bmatrix}. \tag{3}$$

Let $\mathbf{A} - z\mathbf{I} = \mathbf{U}\Sigma\mathbf{V}^T$ be the singular value decomposition of $\mathbf{A} - z\mathbf{I}$ where

$$\Sigma = \begin{bmatrix} \Sigma_1 & \\ & \varepsilon \end{bmatrix}, \quad \text{with} \quad \Sigma_1 = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{n-1} \end{bmatrix}, \ \sigma_{n-1} > \varepsilon. \tag{4}$$

and ε is a simple smallest singular value of $\mathbf{A} - z\mathbf{I}$. Further, let the Jacobian $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$ in (3) be decomposed

as $\mathbf{F_w}(\mathbf{w}) = \mathbf{U_F}\mathbf{G_F}\mathbf{V_F}^T$,

$$\mathbf{F_w}(\mathbf{w}) = \begin{bmatrix} \mathbf{U}\Sigma\mathbf{V}^T & -\mathbf{U}\mathbf{e}_n & -\varepsilon\mathbf{I} & -\mathbf{U}\nu \\ -(\mathbf{V}\mathbf{e}_n)^T & 0 & \mathbf{0}^T & 0 \\ -\varepsilon\mathbf{I} & -\mathbf{V}\mathbf{e}_n & (\mathbf{U}\Sigma\mathbf{V}^T)^T & -\mathbf{V}\mu \\ \mathbf{0}^T & 0 & -(\mathbf{U}\mathbf{e}_n)^T & 0 \\ (\mathbf{V}\mu)^T & 0 & (\mathbf{U}\nu)^T & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{U} & & & & \\ & 1 & & & \\ & & \mathbf{V} & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} \Sigma & -\mathbf{e}_n & -\varepsilon\mathbf{I} & -\nu \\ -\mathbf{e}_n^T & 0 & \mathbf{0}^T & 0 \\ -\varepsilon\mathbf{I} & -\mathbf{e}_n & \Sigma & -\mu \\ \mathbf{0}^T & 0 & -\mathbf{e}_n^T & 0 \\ \mu^T & 0 & \nu^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}^T & & & & \\ & 1 & & & \\ & & \mathbf{U}^T & & \\ & & & 1 & \end{bmatrix}, \tag{5}$$

where $\mathbf{e}_n$ is the nth column of the n by n identity matrix, $\mathbf{U_F} \in \mathbb{R}^{(2n+3)\times(2n+3)}$, $\mathbf{G_F} \in \mathbb{R}^{(2n+3)\times(2n+2)}$, and $\mathbf{V_F} \in \mathbb{R}^{(2n+2)\times(2n+2)}$. Note that in the above factorization of the Jacobian (5), the vectors $\mathbf{u}$ and $\mathbf{v}$ of (3) each have two different expressions, i.e. $\mathbf{v} = \mathbf{V}\mathbf{e}_n, \mathbf{v} = \mathbf{U}\nu$ and $\mathbf{u} = \mathbf{U}\mathbf{e}_n, \mathbf{u} = \mathbf{V}\mu$. By using the fact that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal *i.e.*, $\mathbf{u}^T\mathbf{v} = 0$. This gives

$$\mathbf{u}^T\mathbf{v} = (\mathbf{U}\mathbf{e}_n)^T\mathbf{U}\nu = \mathbf{e}_n^T(\mathbf{U}^T\mathbf{U})\nu = \mathbf{e}_n^T\nu = \nu_n. \tag{6}$$

Which implies that $\nu_n = 0$. In the same vein, it can be shown that $\mu_n = 0$. So that ν and μ become respectively $[\nu_{n-1}, 0]^T$ and $[\mu_{n-1}, 0]^T$.

In expanded form, we rewrite the matrix $\mathbf{G_F}$ as

$$\mathbf{G_F} = \begin{bmatrix} \Sigma_1 & \mathbf{0}_{n-1} & \mathbf{0}_{n-1} & -\varepsilon\mathbf{I}_{n-1} & \mathbf{0}_{n-1} & -\nu_{n-1} \\ \mathbf{0}_{n-1}^T & \varepsilon & -1 & \mathbf{0}_{n-1}^T & -\varepsilon & 0 \\ \mathbf{0}_{n-1}^T & -1 & 0 & \mathbf{0}_{n-1}^T & 0 & 0 \\ -\varepsilon\mathbf{I}_{n-1} & \mathbf{0}_{n-1} & \mathbf{0}_{n-1} & \Sigma_1 & \mathbf{0}_{n-1} & -\mu_{n-1} \\ \mathbf{0}_{n-1}^T & -\varepsilon & -1 & \mathbf{0}_{n-1}^T & \varepsilon & 0 \\ \mathbf{0}_{n-1}^T & 0 & 0 & \mathbf{0}_{n-1}^T & -1 & 0 \\ \mu_{n-1}^T & 0 & 0 & \nu_{n-1}^T & 0 & 0 \end{bmatrix}. \tag{7}$$

The following preminary analysis contains some important relationships that will help in proving Theorem 2.1 and Lemma 2.2 shortly. We build on the assumption that z is a multiple eigenvalue of $\mathbf{B}$ by defining $\hat{\mathbf{v}} = \mathbf{V}\beta$ as the right generalized eigenvector corresponding to the eigenvalue z such that $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$ with $\mathbf{v}^T\hat{\mathbf{v}} = 0$ and $\mathbf{u}^T\hat{\mathbf{v}} \neq 0$. The condition $\mathbf{u}^T\hat{\mathbf{v}} \neq 0$ ensures that $\mathbf{B}$ has a 2-dimensional Jordan block only.

In the same vein, we define $\hat{\mathbf{u}} = \mathbf{U}\alpha$ as the left generalized eigenvector corresponding to z such that $(\mathbf{B}^T - z\mathbf{I})\hat{\mathbf{u}} = \mathbf{u}$ with $\mathbf{u}^T\hat{\mathbf{u}} = 0$ and $\mathbf{v}^T\hat{\mathbf{u}} \neq 0$. By taking tranpose of both sides of $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$, we have $\mathbf{v}^T = \hat{\mathbf{v}}^T(\mathbf{B}^T - z\mathbf{I})$, and postmultiplying by $\hat{\mathbf{u}}$ yields

$$\begin{aligned} \mathbf{v}^T\hat{\mathbf{u}} &= \hat{\mathbf{v}}^T(\mathbf{B}^T - z\mathbf{I})\hat{\mathbf{u}} \\ &= \hat{\mathbf{v}}^T\mathbf{u} \\ &= \mathbf{u}^T\hat{\mathbf{v}}. \end{aligned}$$

This shows that $\mathbf{v}^T\hat{\mathbf{u}} = \mathbf{u}^T\hat{\mathbf{v}}$. Since $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^T$, by post multiplying both sides of $\mathbf{B} - z\mathbf{I} = \mathbf{A} - z\mathbf{I} - \varepsilon\mathbf{u}\mathbf{v}^T$ by $\hat{\mathbf{v}}$, and after simplifying we get

$$(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = (\mathbf{A} - z\mathbf{I})\hat{\mathbf{v}}.$$

Similarly, it can be shown that $(\mathbf{B}^T - z\mathbf{I})\hat{\mathbf{u}} = (\mathbf{A}^T - z\mathbf{I})\hat{\mathbf{u}}$. Since $(\mathbf{B} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$ is the same as $(\mathbf{A} - z\mathbf{I})\hat{\mathbf{v}} = \mathbf{v}$, we have that $\mathbf{U}\Sigma(\mathbf{V}^T\mathbf{V})\beta = \mathbf{U}\nu$. Because $\mathbf{U}$, $\mathbf{V}$ are orthogonal, $\Sigma\beta = \nu$ and

$$\begin{bmatrix} \Sigma_1 & \\ & \varepsilon \end{bmatrix} \begin{bmatrix} \beta_{n-1} \\ \beta_n \end{bmatrix} = \begin{bmatrix} \nu_{n-1} \\ 0 \end{bmatrix},$$

hence, $\boldsymbol{\beta}_{n-1} = \Sigma_1^{-1}\boldsymbol{\nu}_{n-1}$. After premultiplying $\hat{\mathbf{v}} = \mathbf{V}\boldsymbol{\beta}$ by $\mathbf{V}^T$, we have $\boldsymbol{\beta} = \mathbf{V}^T\hat{\mathbf{v}}$. By following the steps that led to (6) using $\mathbf{v}^T\hat{\mathbf{v}} = 0$, with $\mathbf{v} = \mathbf{V}\mathbf{e}_n$ and $\hat{\mathbf{v}} = \mathbf{V}\boldsymbol{\beta}$, it can be shown that $\beta_n = 0$. This means that

$$\boldsymbol{\beta_{n-1}} = \mathbf{V}_{n-1}^T\hat{\mathbf{v}}_{n-1} = \Sigma_1^{-1}\boldsymbol{\nu}_{n-1}.$$

Similarly, taking $\hat{\mathbf{u}} = \mathbf{U}\boldsymbol{\alpha}$, and $(\mathbf{A}^T - z\mathbf{I})\hat{\mathbf{u}} = \mathbf{u}$, it can be shown as above that

$$\boldsymbol{\alpha_{n-1}} = \mathbf{U}_{n-1}^T\hat{\mathbf{u}}_{n-1} = \Sigma_1^{-1}\boldsymbol{\mu}_{n-1},$$

and by using the fact that $\mathbf{u}^T\hat{\mathbf{v}} \neq 0$,

$$\mathbf{u}^T\hat{\mathbf{v}} = \boldsymbol{\mu}^T\mathbf{V}^T\mathbf{V}\boldsymbol{\beta} = \boldsymbol{\mu}^T\boldsymbol{\beta} = \boldsymbol{\mu}_{n-1}^T\boldsymbol{\beta}_{n-1} = \boldsymbol{\mu}_{n-1}^T\Sigma_1^{-1}\boldsymbol{\nu}_{n-1} \neq 0. \tag{8}$$

In order to apply the Gauss-Newton theory to find the solution to the nonlinear least squares problem (2), we need to show that the Jacobian $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$ in (3) is of full rank. Before we prove that the Jacobian is of full rank, we define the matrix

$$\mathbf{M} = \begin{bmatrix} \Sigma_1 & -\varepsilon\mathbf{I} & -\boldsymbol{\nu}_{n-1} \\ -\varepsilon\mathbf{I} & \Sigma_1 & -\boldsymbol{\mu}_{n-1} \\ \boldsymbol{\mu}_{n-1}^T & \boldsymbol{\nu}_{n-1}^T & 0 \end{bmatrix}. \tag{9}$$

$\mathbf{M}$ is obtained from the expanded form of $\mathbf{G_F}$ in (7) by deleting appropriate rows and columns-reason of which will be made clear in the proof of Theorem 2.1.

Theorem 2.1. *The rank of the Jacobian* $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$ *in* (3) *is* $2n+2$ *or of full rank if* $\mathbf{M}$ *is nonsingular.*

Proof. Let $\mathbf{M}$ be nonsingular and $\mathbf{p} = [\mathbf{p}_{n-1}, p_n]^T$, $\mathbf{r} = [\mathbf{r}_{n-1}, r_n]^T$ be nonzero vectors, then we want to show that the rank of $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$ is $2n+2$. The rank of $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$ equals the rank of $\mathbf{G_F}$, because in the decomposition $\mathbf{F}_{\mathbf{w}}(\mathbf{w}) = \mathbf{U_F}\mathbf{G_F}\mathbf{V_F}^T$, the matrices $\mathbf{U_F} \in \mathbb{R}^{(2n+3)\times(2n+2)}$ and $\mathbf{V_F} \in \mathbb{R}^{(2n+2)\times(2n+2)}$ are orthogonal, while $\mathbf{G_F} \in \mathbb{R}^{(2n+3)\times(2n+2)}$ is of the same size as $\mathbf{F}_{\mathbf{w}}(\mathbf{w})$. So it is enough to show that the rank of $\mathbf{G_F}$ equals $2n+2$. This is the same as showing that the $2n+2$ vectors $[\mathbf{p}, q, \mathbf{r}, s]^T$ are zero in

$$\begin{bmatrix} \Sigma & -\mathbf{e}_n & -\varepsilon\mathbf{I} & -\boldsymbol{\nu} \\ -\mathbf{e}_n^T & 0 & \mathbf{0}^T & 0 \\ -\varepsilon\mathbf{I} & -\mathbf{e}_n & \Sigma & -\boldsymbol{\mu} \\ \mathbf{0}^T & 0 & -\mathbf{e}_n^T & 0 \\ \boldsymbol{\mu}^T & 0 & \boldsymbol{\nu}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ q \\ \mathbf{r} \\ s \end{bmatrix} = \mathbf{0}.$$

To make things easier, we will use the expanded form of $\mathbf{G_F}$ in (7) in our analysis. Multiply $\mathbf{G_F}$ from the right by the vector $[\mathbf{p}_{n-1}, p_n, q, \mathbf{r}_{n-1}, r_n, s]^T$, upon expanding the matrix vector multiplication and equating to the zero vector on the right hand side, we obtain $p_n = q = r_n = 0$. This is equivalent to deleting the nth, $(n+1)$th, $(2n+1)$th columns and the nth, $(n+1)$th, $(2n+1)$th, $(2n+2)$th rows of $\mathbf{G_F}$, the remaining nonzero entries constitute the matrix $\mathbf{M}$. Thus, we are left to show that $\mathbf{p}_{n-1} = \mathbf{r}_{n-1} = \mathbf{0}$ and $s = 0$ in

$$\begin{aligned} \Sigma_1\mathbf{p}_{n-1} - \varepsilon\mathbf{r}_{n-1} - \boldsymbol{\nu}_{n-1}s &= \mathbf{0} \\ -\varepsilon\mathbf{p}_{n-1} + \Sigma_1\mathbf{r}_{n-1} - \boldsymbol{\mu}_{n-1}s &= \mathbf{0} \\ \boldsymbol{\mu}_{n-1}^T\mathbf{p}_{n-1} + \boldsymbol{\nu}_{n-1}^T\mathbf{r}_{n-1} &= 0, \end{aligned} \tag{10}$$

which amounts to showing that $\mathbf{M}$ is nonsingular. But by assumption, $\mathbf{M}$ is nonsingular, hence, $\mathbf{p}_{n-1} = \mathbf{r}_{n-1} = \mathbf{0}$ and $s = 0$. Therefore, $[\mathbf{p}, q, \mathbf{r}, s] = [\mathbf{p}_{n-1}, p_n, q, \mathbf{r}_{n-1}, r_n, s] = \mathbf{0}$ and $\operatorname{rank}(\mathbf{G_F}) = 2n+2$. Since $\operatorname{rank}(\mathbf{F}_{\mathbf{w}}(\mathbf{w})) = \operatorname{rank}(\mathbf{G_F})$, thus $\operatorname{rank}(\mathbf{F}_{\mathbf{w}}(\mathbf{w})) = 2n+2$ or the Jacobian is of full rank. □

Because we will make use of Keller's [12] ABCD Lemma to prove that $\mathbf{M}$ nonsingular, we partition $\mathbf{M}$ as follows where in this case

$$\mathrm{A} = \begin{bmatrix} \Sigma_1 & -\varepsilon\mathbf{I} \\ -\varepsilon\mathbf{I} & \Sigma_1 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} -\boldsymbol{\nu}_{n-1} \\ -\boldsymbol{\mu}_{n-1} \end{bmatrix}, \quad \mathbf{c}^T = [\boldsymbol{\mu}_{n-1}^T, \boldsymbol{\nu}_{n-1}^T] \text{ and } d = 0. \tag{11}$$

One of the conditions of the ABCD Lemma [12] for showing that the partitioned matrix $\mathbf{M}$ is nonsingular, is that the matrix A must be nonsingular. So, we need to show that A is nonsingular. Which is equivalent to showing that $\mathbf{a}$ and $\mathbf{h}$ are both zero vectors in

$$\begin{bmatrix} \Sigma_1 & -\varepsilon\mathbf{I} \\ -\varepsilon\mathbf{I} & \Sigma_1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{h} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{12}$$

Multiply the first n rows by ε and the second by Σ_1 and after adding, we have

$$\begin{bmatrix} \Sigma_1 & -\varepsilon\mathbf{I} \\ 0\mathbf{I} & \Sigma_1^2 - \varepsilon^2\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{h} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

Using (13), the fact that $\Sigma_1^2 - \varepsilon^2\mathbf{I}$ is nonsingular if $\varepsilon < \sigma_{n-1}$, accordingly, $\mathbf{h} = \mathbf{0}$. In the same fashion, because Σ_1 is nonsingular we obtain $\mathbf{a} = \mathbf{0}$. Therefore, A is nonsingular.

The following lemma shows that under certain conditions on ε, the matrix $\mathbf{M}$ defined by (9) is nonsingular.

Lemma 2.2. *Assume that*

$$\varepsilon < \sigma_{n-1}. \tag{13}$$

Also, assume ε is so small that its second and higher powers can be neglected, and that

$$\varepsilon < \frac{2\mathbf{u}^T\hat{\mathbf{v}}}{\|\hat{\mathbf{u}}_{n-1}\|^2 + \|\hat{\mathbf{v}}_{n-1}\|^2}, \tag{14}$$

then the matrix $\mathbf{M}$ is nonsingular.

Proof. Since A has been shown to be nonsingular, in establishing the nonsingularity of $\mathbf{M}$ using the ABCD Lemma, all we need is to show that the expression $d - \mathbf{c}^T\mathsf{A}^{-1}\mathbf{b}$, is not equal to zero, where $\mathbf{b}$, $\mathbf{c}^T$ and d are as defined in (11). To begin, we note that $\mathsf{A}^{-1}\mathbf{b}$ is the same as solving for $[\mathbf{f}, \mathbf{g}]^T$ in

$$\begin{bmatrix} \Sigma_1 & -\varepsilon\mathbf{I} \\ -\varepsilon\mathbf{I} & \Sigma_1 \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{\nu}_{n-1} \\ -\boldsymbol{\mu}_{n-1} \end{bmatrix}. \tag{15}$$

Multiply the first n-equations by ε and the second one by Σ_1, add them together, solve for $\mathbf{g}$ to obtain

$$\mathbf{g} = -[\Sigma_1^2 - \varepsilon^2\mathbf{I}]^{-1}(\Sigma_1\boldsymbol{\mu}_{n-1} + \varepsilon\boldsymbol{\nu}_{n-1}). \tag{16}$$

Hence, by substituting $\mathbf{g}$ into $\boldsymbol{\Sigma}_1\mathbf{f} - \varepsilon\mathbf{g} = -\boldsymbol{\nu}_{n-1}$; we obtain

$$\mathbf{f} = \varepsilon\boldsymbol{\Sigma}_1^{-1}\mathbf{g} - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\nu}_{n-1}.$$

Which is equivalent to

$$\mathbf{f} = -\left\{\varepsilon\Sigma_1^{-1}[\Sigma_1^2 - \varepsilon^2\mathbf{I}]^{-1}(\Sigma_1\boldsymbol{\mu}_{n-1} + \varepsilon\boldsymbol{\nu}_{n-1}) + \Sigma_1^{-1}\boldsymbol{\nu}_{n-1}\right\}. \tag{17}$$

So that

$$\begin{aligned} d - \mathbf{c}^T\mathsf{A}^{-1}\mathbf{b} &= -[\boldsymbol{\mu}_{n-1}^T \quad \boldsymbol{\nu}_{n-1}^T] \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \\ &= \varepsilon\boldsymbol{\mu}_{n-1}^T\Sigma_1^{-1}[\Sigma_1^2 - \varepsilon^2\mathbf{I}]^{-1}(\Sigma_1\boldsymbol{\mu}_{n-1} + \varepsilon\boldsymbol{\nu}_{n-1}) + \boldsymbol{\mu}_{n-1}^T\Sigma_1^{-1}\boldsymbol{\nu}_{n-1} \\ &+ \boldsymbol{\nu}_{n-1}^T[\Sigma_1^2 - \varepsilon^2\mathbf{I}]^{-1}(\Sigma_1\boldsymbol{\mu}_{n-1} + \varepsilon\boldsymbol{\nu}_{n-1}). \end{aligned} \tag{18}$$

From the statement of the Lemma, if ε is so small that its second and higher powers can be neglected, with some simplifications, we have

$$\begin{aligned} d - \mathbf{c}^T \mathsf{A}^{-1}\mathbf{b} &= 2\boldsymbol{\mu}_{n-1}^T \Sigma_1^{-1} \boldsymbol{\nu}_{n-1} + \varepsilon(\boldsymbol{\mu}_{n-1}^T \Sigma_1^{-2} \boldsymbol{\mu}_{n-1} + \boldsymbol{\nu}_{n-1}^T \Sigma_1^{-2} \boldsymbol{\nu}_{n-1}) \\ &= 2\boldsymbol{\mu}_{n-1}^T \Sigma_1^{-1} \boldsymbol{\nu}_{n-1} + \varepsilon[(\Sigma_1^{-1}\boldsymbol{\mu}_{n-1})^T (\Sigma_1^{-1}\boldsymbol{\mu}_{n-1}) \\ &+ (\Sigma_1^{-1}\boldsymbol{\nu}_{n-1})^T (\Sigma_1^{-1}\boldsymbol{\nu}_{n-1})] \\ &= 2\boldsymbol{\mu}_{n-1}^T \Sigma_1^{-1} \boldsymbol{\nu}_{n-1} + \varepsilon(\|\Sigma_1^{-1}\boldsymbol{\mu}_{n-1}\|^2 + \|\Sigma_1^{-1}\boldsymbol{\nu}_{n-1}\|^2). \end{aligned} \tag{19}$$

It remains to be shown that the expression on the right side of (19) is nonzero provided (13) and (14) holds. With the simplified expression,

$$\mathbf{u}^T \hat{\mathbf{v}} = \boldsymbol{\mu}_{n-1}^T \Sigma_1^{-1} \boldsymbol{\nu}_{n-1} \neq 0,$$

in (8), (19) now becomes

$$\begin{aligned} d - \mathbf{c}^T \mathsf{A}^{-1}\mathbf{b} &= 2\mathbf{u}^T \hat{\mathbf{v}} + \varepsilon(\|\mathbf{U}_{n-1}^T \hat{\mathbf{u}}_{n-1}\|^2 + \|\mathbf{V}_{n-1}^T \hat{\mathbf{v}}_{n-1}\|^2) \\ &= 2\mathbf{u}^T \hat{\mathbf{v}} + \varepsilon(\|\hat{\mathbf{u}}_{n-1}\|^2 + \|\hat{\mathbf{v}}_{n-1}\|^2). \end{aligned} \tag{20}$$

In arriving at the expression on the right hand side of the above, we made use of the fact that the matrices $\mathbf{V}_{n-1}$ and $\mathbf{U}_{n-1}$ are orthogonal. Therefore, $d - \mathbf{c}^T \mathsf{A}^{-1}\mathbf{b}$ is not equal to zero if

$$\varepsilon < \frac{2\mathbf{u}^T \hat{\mathbf{v}}}{\|\hat{\mathbf{u}}_{n-1}\|^2 + \|\hat{\mathbf{v}}_{n-1}\|^2}. \tag{21}$$

This shows that the matrix $\mathbf{M}$ is nonsingular if (13) and (14) holds. □

Note that the Jacobian is of full rank under the conditions in which the matrix $\mathbf{M}$ is nonsingular. Since we have established that the Jacobian is of full rank, we can conveniently seek a solution for $\Delta\mathbf{w}^{(k)}$ using Gauss-Newton method. If we apply the standard Gauss-Newton method to the over-determined nonlinear system (2), then we obtain the following,

$$\mathbf{R}\Delta\mathbf{w}^{(k)} = -\mathbf{Q}^T \mathbf{F}(\mathbf{w}^{(k)}), \quad \text{and} \quad \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \Delta\mathbf{w}^{(k)},$$

where in this case, $\mathbf{Q} \in \mathbb{R}^{(2n+3)\times(2n+2)}$ and $\mathbf{R} \in \mathbb{R}^{(2n+2)\times(2n+2)}$. Now, we present a Gauss-Newton based algorithm for finding the parameters for computing a nearby defective matrix to a simple matrix. This is presented in Algorithm 2.3.

Algorithm 2.3. *Gauss-Newton algorithm for solving over-determined nonlinear systems*

1. *Input:* $\mathbf{w}^{(0)} = [\mathbf{v}^{(0)}, \varepsilon^{(0)}, \mathbf{u}^{(0)}, z^{(0)}]^T$, k_{max} *and tol.*
2. *For* $k = 0, 1, 2,$ *until convergence*
3. *Find the reduced QR factorization of* $\mathbf{F}_{\mathbf{w}}(\mathbf{w}^{(k)})$, $[Q, R] = qr(\mathbf{F}_{\mathbf{w}}(\mathbf{w}^{(k)}), 0)$.
4. *Compute the matrix-vector multiplication* $\mathbf{y}^{(k)} = -\mathbf{Q}^T \mathbf{F}(\mathbf{w}^{(k)})$.
5. *Solve the upper-triangular system* $\mathbf{R}\Delta\mathbf{w}^{(k)} = \mathbf{y}^{(k)}$ *for* $\Delta\mathbf{w}^{(k)}$.
6. *Update,* $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \Delta\mathbf{w}^{(k)}$.
7. *Output:* $\mathbf{w}^{(k_{max})}$.

Algorithm 2.3 should be stopped as soon as the norm of $\Delta\mathbf{w}^{(k)}$ is less than or equal to some user defined tolerance.

2.2 Optimal starting vectors

In this subsection, given a simple matrix $\mathbf{A}$, we discuss a schematic way to choose good starting guesses for computing $\mathrm{d}(\mathbf{A})$ and a nearby defective matrix to $\mathbf{A}$.

1. Take $z^{(0)}$ as the average of the two smallest diagonal elements, for example, for the Kahan matrix we could take $z^{(0)} = \frac{s^{n-1}+s^n}{2}$, or a value close to the average of two diagonal elements where the defective eigenvalues is suspected to be lurking (if they are known before hand), *e.g.*, Trefethen and Wilkinson matrix.
2. Find the `svd` of $(\mathbf{A} - z^{(0)}\mathbf{I})$.
3. Choose $\varepsilon^{(0)}$ as the minimum singular value of $(\mathbf{A} - z^{(0)}\mathbf{I})$.
4. Choose $\mathbf{u}^{(0)}$, $\mathbf{v}^{(0)}$, as the left and right singular vectors respectively corresponding to the smallest singular value $\varepsilon^{(0)}$ of $(\mathbf{A} - z^{(0)}\mathbf{I})$.

In the next section, we present the result of numerical experiments which confirms the theory.

3 Numerical simulation and discussion

We now illustrate the numerical performance of our method with several examples which are taken from [2]. As has been mentioned earlier, since our method is based on Newton's method it finds a nearby defective matrix. We cannot guarantee it finds the closest defective matrix. However, in all cases considered here our method found the nearest defective matrix according to Alam and Bora [2] (but at much less cost, of course). Throughout this section,
$\Delta\mathbf{y}^{(k)} = [\Delta\alpha^{(k)}, \Delta\beta^{(k)}, \Delta\varepsilon^{(k)}]^T$.

In this section, we present result of numerical experiments which confirms the theory discussed in the last section. We show that Algorithm 2.3 works for the Trefethen, Kahan, Frank and Wilkinson matrices.

Example 3.1. Consider the matrix $\mathbf{A} = \begin{bmatrix} -1 & 5 \\ 0 & -2 \end{bmatrix}$, (see [9]). As initial guesses we choose $\alpha^{(0)} = \beta^{(0)} = 0$, $\varepsilon^{(0)} = \sigma_{\min}$, $\mathbf{u}^{(0)} = \mathbf{u}_{\min}$ and $v^{(0)} = \mathbf{v}_{\min}$, where $\sigma_{\min}$ is the minimum singular value of $\mathbf{A}$ with corresponding left and right singular vectors $\mathbf{u}_{\min}$ and $\mathbf{v}_{\min}$. We stop the iteration once

$$\frac{\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|}{\|\mathbf{w}^{(k+1)}\|} < 6.0 \times 10^{-16}.$$

Table 1: Column seven shows quadratic convergence for Example 3.1.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000	3.6597e-01	1.0e+00	2.0e+00	5.1e-01	8.9e-01
1	-2.0400	-6.8361e-01	6.0e-01	5.0e-01	1.2e+00	3.7e-01
2	-1.5386	-8.5612e-02	3.6e-02	3.8e-02	7.0e-02	2.6e-02
3	-1.5001	-4.9585e-02	7.6e-05	8.1e-05	1.5e-04	5.4e-05
4	-1.5000	-4.9510e-02	1.6e-10	1.7e-10	3.3e-10	1.1e-10
5	-1.5000	-4.9510e-02	6.9e-18	0.0e+00	1.7e-17	5.0e-18

Table 1 shows the results for Example 3.1. Hence $z = -1.5$ is a degenerate common boundary point of the pseudospectrum, according to [2]. With $\varepsilon = -4.9510 \times 10^{-2}$,
$\mathbf{u} = [-9.8538 \times 10^{-2},\ -9.9513 \times 10^{-1}]^T$ and $\mathbf{v} = [9.9513 \times 10^{-1},\ -9.8538 \times 10^{-2}]^T$ we have that $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{u}\mathbf{v}^H$ is a defective matrix. The method converges quadratically in 5 iterations, as expected from Newton's method.

Example 3.2. [13] Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be the Kahan matrix [9], which is given by [13, p. 791], (see also, [2, p. 293]

$$\mathbf{A} = \begin{cases} 0, & \text{if } i > j \\ s^{i-1}, & \text{if } i = j \\ -cs^{i-1}, & \text{if } i < j, \end{cases}$$

$$= \begin{bmatrix} 1 & -c & -c & -c & \cdots & -c \\ & s & -sc & -sc & \cdots & -sc \\ & & s^2 & -s^2c & \cdots & -s^2c \\ & & & \ddots & \ddots & \vdots \\ & & & & \ddots & -s^{n-2}c \\ & & & & & s^{n-1} \end{bmatrix}, \tag{22}$$

for $i, j = 1, \ldots, n$, $s^2 + c^2 = 1$ and $a_{nn} = s^{n-1}$ is the smallest element on the diagonal. We consider this matrix for $n = 6, 15, 20$. The starting values and stopping condition are chosen as in Example 3.1.

Table 2: Results for Example 3.2, $n = 6$ using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000e+00	9.9694e-03	2.2e-02	1.4e-01	1.6e-01	2.5e-01
1	1.3643e-01	-1.2145e-02	1.2e-02	2.0e-02	7.1e-02	1.5e-01
2	1.1639e-01	-3.7277e-04	6.1e-05	9.5e-03	2.5e-02	3.9e-02
3	1.2590e-01	-4.3373e-04	3.7e-05	1.7e-03	1.5e-03	9.3e-03
4	1.2760e-01	-4.7082e-04	3.3e-07	3.2e-05	8.9e-05	2.5e-04
5	1.2763e-01	-4.7049e-04	2.9e-10	1.3e-08	6.3e-08	1.0e-07
6	1.2763e-01	-4.7049e-04	4.4e-17	2.2e-15	1.1e-14	1.6e-14
7	1.2763e-01	-4.7049e-04	5.4e-20	0.0e+00	9.9e-17	4.6e-17

Table 2 shows the results for $n = 6$. In this case the eigenvalues 1.5849×10^{-1} and 10^{-1} coalesce at 1.2763×10^{-1} for a value of $\varepsilon = -4.7049 \times 10^{-4}$. The quadratic convergence rate is observed in rows six and seven of column five.

Table 3: Results for Example 3.2, $n = 15$ using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000e+00	4.7454e-04	7.3e-04	5.9e-02	1.2e-02	1.4e-01
1	5.9261e-02	-2.5746e-04	2.6e-04	2.5e-02	2.1e-02	1.6e-01
2	8.4526e-02	4.7975e-07	8.7e-07	1.3e-02	2.5e-02	1.2e-01
3	9.7390e-02	1.3523e-06	2.0e-06	7.2e-03	1.5e-02	7.8e-02
4	1.0455e-01	-6.6553e-07	1.3e-07	2.4e-03	6.3e-03	3.1e-02
5	1.0699e-01	-5.3550e-07	1.8e-08	3.0e-04	9.7e-04	4.0e-03
6	1.0728e-01	-5.1778e-07	2.0e-10	4.3e-06	1.6e-05	5.8e-05
7	1.0729e-01	-5.1757e-07	4.1e-14	8.9e-10	3.4e-09	1.2e-08
8	1.0729e-01	-5.1757e-07	1.8e-21	2.8e-17	1.3e-16	4.9e-16

Table 3 shows the results for $n = 15$. In this case the eigenvalues 1.1788×10^{-1} and 10^{-1} coalesce at 1.0729×10^{-1} for a value of $\varepsilon = -5.1757 \times 10^{-7}$.

Table 4 shows the results for $n = 20$. In this case the eigenvalues 1.1288×10^{-1} and 10^{-1} coalesce at 1.0501×10^{-1} for a value of $\varepsilon = -2.8841 \times 10^{-8}$.

Example 3.3. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be the Frank matrix taken from the Matlab gallery $\mathbf{A} = gallery('frank', n)$, where $n = 6, 12$. As initial guesses we choose $\alpha^{(0)} = \beta^{(0)} = 0$, $\varepsilon^{(0)} = \sigma_{\min}$, $\mathbf{u}^{(0)} = \mathbf{u}_{\min}$ and $\mathbf{v}^{(0)} = \mathbf{v}_{\min}$, where $\sigma_{\min}$ is the minimum singular value of $\mathbf{A}$ with corresponding left and right singular vectors $\mathbf{u}_{\min}$ and $\mathbf{v}_{\min}$.

Table 4: Results for Example 3.2, $n = 20$ using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000e+00	1.3141e-04	1.9e-04	4.7e-02	4.0e-03	1.2e-01
1	4.7216e-02	-5.7554e-05	5.8e-05	2.7e-02	1.4e-02	1.4e-01
2	7.3921e-02	-3.1613e-08	4.3e-07	1.5e-02	1.9e-02	1.3e-01
3	8.9281e-02	4.0206e-07	4.5e-07	9.3e-03	1.6e-02	1.0e-01
4	9.8568e-02	-5.2530e-08	1.5e-08	4.7e-03	1.0e-02	6.3e-02
5	1.0329e-01	-3.7362e-08	7.9e-09	1.5e-03	4.0e-03	2.3e-02
6	1.0483e-01	-2.9487e-08	6.4e-10	1.8e-04	5.4e-04	2.7e-03
7	1.0501e-01	-2.8848e-08	7.3e-12	2.2e-06	7.5e-06	3.4e-05
8	1.0501e-01	-2.8841e-08	1.1e-15	3.3e-10	1.2e-09	5.1e-09
9	1.0501e-01	-2.8841e-08	3.0e-23	0.0e+00	1.6e-16	1.3e-16

Table 5: Results for Example 3.3, $n = 6$ using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000e+00	3.4855e-03	7.1e-03	1.0e-01	7.0e-02	1.1e-01
1	1.0137e-01	-3.5747e-03	3.0e-03	2.4e-02	1.4e-02	3.0e-02
2	1.2532e-01	-6.0255e-04	4.7e-05	2.6e-03	1.1e-03	2.9e-03
3	1.2788e-01	-5.5588e-04	3.9e-07	2.4e-05	1.0e-05	2.7e-05
4	1.2790e-01	-5.5549e-04	3.6e-11	2.1e-09	8.8e-10	2.4e-09
5	1.2790e-01	-5.5549e-04	1.4e-18	5.6e-16	1.0e-15	6.4e-16
6	1.2790e-01	-5.5549e-04	2.1e-17	3.9e-16	9.1e-16	5.0e-16

Table 5 shows the results for $n = 6$. In this case the eigenvalues 7.7080×10^{-2} and 1.8576×10^{-1} closest to zero coalesce at 1.2790×10^{-1} for a value of $\varepsilon = -5.5549 \times 10^{-4}$.

Table 6: Results for Example 3.3, $n = 12$ using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)} - \varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)} - \alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	0.0000e+00	1.1186e-08	1.6e-08	1.8e-02	8.6e-07	1.9e-02
1	1.8010e-02	-4.3454e-09	3.9e-09	1.3e-02	4.1e-04	1.4e-02
2	3.0849e-02	-4.1641e-10	2.0e-10	6.2e-03	2.1e-04	6.6e-03
3	3.7032e-02	-2.2016e-10	3.3e-11	1.5e-03	5.0e-05	1.6e-03
4	3.8559e-02	-1.8668e-10	1.7e-12	9.0e-05	3.1e-06	9.7e-05
5	3.8649e-02	-1.8499e-10	5.4e-15	3.0e-07	1.1e-08	3.3e-07
6	3.8649e-02	-1.8499e-10	3.4e-17	1.2e-11	1.2e-13	1.3e-11
7	3.8649e-02	-1.8499e-10	2.2e-17	1.2e-11	8.6e-16	1.2e-11
8	3.8649e-02	-1.8499e-10	1.5e-17	6.7e-12	1.6e-15	7.2e-12
9	3.8649e-02	-1.8499e-10	8.5e-18	3.7e-12	1.4e-15	4.0e-12
10	3.8649e-02	-1.8499e-10	1.0e-17	9.2e-13	7.4e-16	9.8e-13

Table 6 shows the results for $n = 12$. In this case the eigenvalues 3.1028×10^{-2} and 4.9509×10^{-2} closest to zero coalesce at 3.8649×10^{-2} for a value of $\varepsilon = -1.8499 \times 10^{-10}$.

Example 3.4. Consider the 20×20 bi-diagonal matrix whose diagonal entries are $20, 19, \ldots, 1$ and the super-diagonals are 20. This matrix was considered by Wilkinson [1] and has eigenvalues $1, 2, \ldots, 20$. Results are as tabulated in Table 7.

4 Conclusion

In an attempt to provide a partial answer to the question posed by Wilkinson, we have succeeded in presenting an algorithm based on Gauss-Newton method for computing a nearby defective matrix to

Table 7: Results for Example 3.4 using Algorithm 2.3.

k	$\alpha^{(k)}$	$\varepsilon^{(k)}$	$\lvert\varepsilon^{(k+1)}-\varepsilon^{(k)}\rvert$	$\lvert\alpha^{(k+1)}-\alpha^{(k)}\rvert$	$\lVert\mathbf{F}(\mathbf{w}^{(k)})\rVert$	$\frac{\lVert\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\rVert}{\lVert\mathbf{w}^{(k+1)}\rVert}$
0	10.200	3.6322e-14	6.5e-14	4.2e-01	1.6e-13	3.9e-02
1	10.619	1.0152e-13	3.9e-14	1.0e-01	1.5e-02	9.5e-03
2	10.519	6.2702e-14	1.4e-15	1.9e-02	8.5e-04	1.8e-03
3	10.500	6.1312e-14	4.7e-17	1.9e-04	3.0e-05	1.8e-05
4	10.500	6.1264e-14	4.4e-21	5.1e-08	2.9e-09	4.9e-09
5	10.500	6.1264e-14	1.8e-28	3.4e-13	2.3e-15	3.2e-14
6	10.500	6.1264e-14	1.3e-29	0.0e+00	1.5e-15	8.9e-18

A and the distance between them. Though the new approach relies on good initial guesses, however, we are guaranteed quadratic convergence and it is simple to apply and is significantly faster than the technique in [2]. Numerical examples show and confirm that this technique performs well.

Acknowledgements

The author acknowledges financial support from the University of Bath, United Kingdom, in the form of a University Research Studentship from April 2007 to June 2010. The author acknowledge financial support from the Tertiary Education Fund (TETfund) Nigeria and the Office of Research and Development (ORD) of the University of Jos for funds to publish this research.

References

[1] Wilkinson, J.H. (1965). The Algebraic Eigenvalue Problem, Oxford University Press, Ely House, London W.

[2] Alam, R. and Bora, S. (2005). On Sensitivity of Eigenvalues and Eigendecompositions of Matrices, Linear Algebra and its Applications 396, 273–301.

[3] Alam, R., Bora, S., Byers, R., and Overton, M.L. (2009). Characterisation and Construction of the Nearest Defective Matrix via Coalescence of Pseudospectral Components. Linear Algebra and its Applications 2011; 435(3):494–513.

[4] Demmel, J.W. (1983). A Numerical Analyst's Jordan Canonical Formxs, Ph.D. thesis, University of California, Berkeley.

[5] Demmel, J.W. (1986). Computing Stable Eigendecompositions of Matrices, Linear Algebra and its Applications 79, 163–193.

[6] Wilkinson, J.H. (1984). Sensitivity of Eigenvalues ii. Utilas Math. 25 (1984), 5–76.

[7] Wilkinson, J.H. (1986). Sensitivity of Eigenvalues ii, Utilas Math, no. 30, 243–286.

[8] Lippert, R.A. and Edelman, A. (1999). The Computation and Sensitivity of Double Eigenvalues, in Advances in Computational Mathematics (Guangzhou, 1997), 202 in Lecture Notes in Pure and Applied Mathematics, Dekker, New York. 353–393.

[9] Trefethen, L.N. and Embree, M. (2005). Spectra and Pseudospectra, Princeton University Press.

[10] Dennis, J.E. Jr. and Schnabel, R.B. (1996). Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Classics In Applied Mathematics, no. 16, SIAM Philadelphia.

[11] Akinola, R.O., Freitag, M.A. and Spence, A. (2013). The calculation of the distance to a nearby defective matrix, Numer. Linear Algebra Appl. **2**, no. 4, 367–371.

[12] Keller, H.B. (1977). Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems, Applications of Bifurcation Theory (in : P. Rabinowitz, ed.), Academic Press, New York, pp. 359–384.

[13] Kahan, W. (1966). Numerical Linear Algebra, Canadian Mathematical Bulletin 9, 757–801.