Sensitivity and Computation of a Defective Eigenvalue

Zhonggang Zeng*

February 8, 2016

Abstract

A defective eigenvalues is well documented to be hypersensitive to data perturbations and round-off errors, making it a formidable challenge in numerical computation particularly when the matrix is known through approximate data. This paper establishes a finitely bounded sensitivity of a defective eigenvalue with respect to perturbations that preserve the geometric multiplicity and the smallest Jordan block size. Based on this perturbation theory, numerical computation of a defective eigenvalue is regularized as a well-posed least squares problem so that it can be accurately carried out using floating point arithmetic even if the matrix is perturbed.

1 Introduction

Computing matrix eigenvalues is one of the fundamental problems in theoretical and numerical linear algebra. Remarkable advancement has been achieved since the advent of the Francis QR algorithm in 1960s. However, it is well documented that multiple and defective eigenvalues are hypersensitive to both data perturbations and the inevitable round-off. For an eigenvalue of a matrix A associated with the largest Jordan block size $l \times l$ while A is perturbed by ΔA, the error bound [2, p. 58][3, 13] on the eigenvalue deviation is proportional to $\|\Delta A\|_2^{1/l}$, implying that the accuracy of the computed eigenvalue in number of digits is a fraction $\frac{1}{l}$ of the accuracy of the matrix data. As a result, numerical computation of defective eigenvalues remains a formidable challenge.

On the other hand, it has been known that a defective eigenvalue disperses into a cluster when the matrix is under arbitray perturbations but the mean of the cluster is not hypersensitive [11, 17]. In his seminal technical report[10], Kahan proved that the sensitivity of an m-fold eigenvalue is actually bounded by $\frac{1}{m}\|P\|_2$ where P is the spectral projector associated with the eigenvalue as long as the perturbation is constrained to preserve the algebraic multiplicity. The same proof and the same sensitivity also apply to the mean of the eigenvalue cluster

*Department of Mathematics, Northeastern Illinois University, Chicago, Il 60625. *Email:* zzeng@neiu.edu

emanating from the m-fold eigenvalue with respect to perturbations. Indeed, using cluster means as approximations to defective eigenvalues has been extensively applied to numerical computation of Jordan Canonical Forms and staircase forms, provided that the clusters can be sorted out from the spectrum. This approach includes works of Ruhe [16], Sdridhar and Jordan [20], and culminated in Golub and Wilkinson's review [7] as well as Kågström and Ruhe's JNF [8, 9]. Theoretical issues have been analyzed in, e.g. works of Demmel [4, 5] and Wilkinson [22, 23]. Perturbations on eigenvalue clusters are also studied as pseudospectra of matrices in works of Trefethon and Embree [21] as well as Rump [18, 19].

In this paper we elaborate a different measurement of the sensitivity of a defective eigenvalue with respect to perturbations constrained to preserve the geometric multiplicity and the smallest Jordan block size. We prove that such sensitivity is also finitely bounded even if the multiplicity is not preserved, and it is large only if either the geometric multiplicity or the smallest Jordan block size can be increased by a small perturbation on the matrix. This sensitivity can be small even if the spectral projector norm is large, or vice versa.

In computation, perturbations are expected to be arbitrary without preserving either the multiplicity or what we refer to as the multiplicity support. We prove that a certain type of pseudo-eigenvalue uniquely exists, is Lipschitz continuous, is backward accurate and approximates the defective eigenvalue with a forward accuracy in the same order of the data accuracy, making it a well-posed problem for computing a defective eigenvalue via solving a least squares problem. Based on this analysis, we develop an iterative algorithm PseudoEig[1] that is capable of accurate computation of defective eigenvalues using floating point arithmetic from empirical matrix data even if the spectral projector norm is large and thus the cluster mean is inaccurate.

2 Notation

The space of dimension n vectors is $\mathbb{C}^n$ and the space of $m \times n$ matrices is $\mathbb{C}^{m\times n}$. Matrices are denoted by upper case letters A, X, and G, etc, with O representing a zero matrix whose dimensions can be derived from the context. Boldface lower case letters such as $\mathbf{x}$ and $\mathbf{y}$ represent vectors. Particularly, the zero vector in $\mathbb{C}^n$ is denoted by $\mathbf{0}_n$ or simply $\mathbf{0}$ if the dimension is clear. The conjugate transpose of a matrix or vector $(\cdot)$ is denoted by $(\cdot)^{\mathsf{H}}$, and the Moore-Penrose inverse of a matrix $(\cdot)$ is $(\cdot)^{\dagger}$. The submatrix formed by entries in rows $i_1, \ldots, i_2$ and columns $j_1, \ldots, j_2$ of a matrix A is denoted by $A_{i_1:i_2, j_1:j_2}$. The kernel and range of a matrix $(\cdot)$ are denoted by $\mathscr{Kernel}(\cdot)$ and $\mathscr{Range}(\cdot)$ respectively. The notation $\mathscr{eig}(\cdot)$ represents the spectrum of a matrix $(\cdot)$.

We also consider vectors in product spaces such as $\mathbb{C} \times \mathbb{C}^{m\times k}$. In such cases, the vector 2-norm is the square root of the sum of squares of all components. For instance, a vector $(\lambda, X) \in \mathbb{C} \times \mathbb{C}^{n\times k}$ can be arranged as a column vector $\mathbf{u}$ in $\mathbb{C}^{nk+1}$ and $\|(\lambda, X)\|_2 = \|\mathbf{u}\|_2$ regardless of the ordering. A zero vector in such a vector space is also denoted by $\mathbf{0}$.

Let λ_* be an eigenvalue of a matrix $A \in \mathbb{C}^{n\times n}$. Its algebraic multiplicity can be partitioned

[1] A permanent website `homepages.neiu.edu/∼zzeng/pseudoeig.html` is set up to provide Matlab source codes and other resources for Algorithm PseudoEig.

into a non-increasing sequence $\{l_1, l_2, \ldots\}$ of integers called the *Segre characteristic* [6] that are the sizes of elementary Jordan blocks, and there is a matrix $X_* \in \mathbb{C}^{n\times m}$ such that

$$A\,X_* \;=\; X_* \begin{bmatrix} J_{l_1}(\lambda_*) & & \\ & J_{l_2}(\lambda_*) & \\ & & \ddots \end{bmatrix} \quad \text{where} \quad J_k(\lambda_*) \;=\; \begin{bmatrix} \lambda_* & 1 & & \\ & \lambda_* & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_* \end{bmatrix}_{k\times k}.$$

For convenience, a Segre characteristic is infinite in formality and the number of nonzero entries is the geometric multiplicity. The last nonzero component of a Segre characteristic, namely the size of the smallest Jordan block associated with λ_*, is of particular importance in our analysis and we shall call it the *Segre characteristic anchor* or simply *Segre anchor*.

For instance, if λ_* is an eigenvalue of A associated with elementary Jordan blocks $J_5(\lambda_*)$, $J_5(\lambda_*)$, $J_4(\lambda_*)$, $J_4(\lambda_*)$ and $J_3(\lambda_*)$, its Segre characteristic is $\{5, 5, 4, 4, 3, 0, \ldots\}$ with a Segre anchor 3. The geometric multiplicity is 5. A Segre characteristic along with its conjugate that is called the Weyr characteristic can be illustrated by a Ferrer's diagram [6] in Fig. 1, where the geometric multiplicity and the Segre anchor represent the dimensions of the base rectangle occupied by the equal leading entries of the Weyr characteristic.

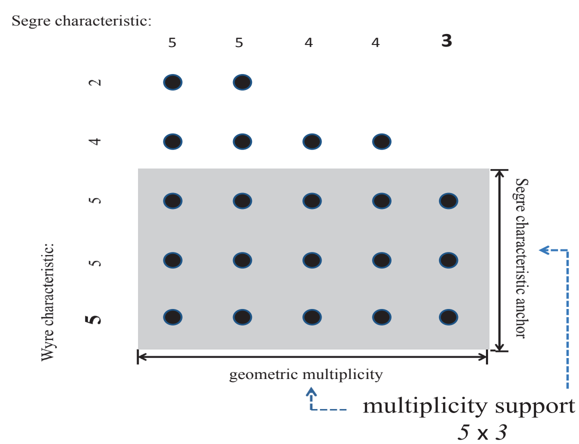

Figure 1: Illustration of the multiplicity support for a defective eigenvalue

For a matrix A, we shall say the *multiplicity support* of its eigenvalue λ_* is $m \times k$ if the geometric multiplicity of λ_* is m and the Segre anchor is k. In this case, there is a unique $X_* \in \mathbb{C}^{n\times k}$ satisfying the equations

$$\begin{aligned} (A - \lambda_* I)\, X_* &= X_*\, J_k(0) \\ C^{\mathsf{H}}\, X_* &= T \end{aligned}$$

with proper choices of $C \in \mathbb{C}^{n\times m}$ and

$$T \;=\; \begin{bmatrix} 1 & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{m-1} & O_{(m-1)\times(k-1)} \end{bmatrix} \;\in\; \mathbb{C}^{m\times k}. \tag{1}$$

as we shall prove in Lemma 3.2. Here $J_k(0)$ is a nilpotent upper-triangular matrix of rank $k-1$ and can be replaced with any matrix of such kind. For integers $m, k \leq n$, we define a holomorphic mapping

$$\begin{array}{rcccc} \mathbf{g} & : & \mathbb{C}^{n\times n} \times \mathbb{C} \times \mathbb{C}^{n\times k} & \longrightarrow & \mathbb{C}^{n\times k} \times \mathbb{C}^{m\times k} \\ & & (G, \lambda, X) & \longmapsto & \begin{pmatrix} (G - \lambda I)\, X - X\, S \\ C^{\mathsf{H}}\, X - T \end{pmatrix} \end{array} \tag{2}$$

that depends on parameters $C \in \mathbb{C}^{n\times m}$ and an upper-triangular nilpotent matrix

$$S \;=\; \begin{bmatrix} 0 & s_{12} & \cdots & s_{1k} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & s_{k-1,k} \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \quad \text{with} \quad s_{12}s_{23}\cdots s_{k-1,k} \neq 0 \tag{3}$$

of rank $k-1$. We shall denote the Jacobian and partial Jacobian

$$\begin{aligned} \mathbf{g}_{G\lambda X}(G_0,\lambda_0,X_0) &= \left.\frac{\partial \mathbf{g}(G,\lambda,X)}{\partial(G,\lambda,X)}\right|_{(G,\lambda,X)=(G_0,\lambda_0,X_0)} \\ \mathbf{g}_{\lambda X}(G_0,\lambda_0,X_0) &= \left.\frac{\partial \mathbf{g}(G_0,\lambda,X)}{\partial(\lambda,X)}\right|_{(\lambda,X)=(\lambda_0,X_0)} \end{aligned}$$

at particular G_0, λ_0 and X_0 that can be considered linear transformations

$$\begin{aligned} \mathbf{g}_{G\lambda X}(G_0,\lambda_0,X_0) \;:\; \mathbb{C}^{n\times n}\times\mathbb{C}\times\mathbb{C}^{n\times k} &\longrightarrow \mathbb{C}^{n\times k}\times\mathbb{C}^{m\times k} \\ (G,\lambda,X) &\longmapsto \begin{pmatrix} (G-\lambda I)\,X_0 + (G_0-\lambda_0 I)\,X - X\,S \\ C^{\mathsf{H}}\,X \end{pmatrix} \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathbf{g}_{\lambda X}(G_0,\lambda_0,X_0) \;:\; \mathbb{C}\times\mathbb{C}^{n\times k} &\longrightarrow \mathbb{C}^{n\times k}\times\mathbb{C}^{m\times k} \\ (\lambda,X) &\longmapsto \begin{pmatrix} -\lambda\,X_0 + (G_0-\lambda_0 I)\,X - X\,S \\ C^{\mathsf{H}}\,X \end{pmatrix} \end{aligned} \tag{5}$$

respectively. The actual matrices representing the Jacobians depend on the ordering of the bases for the domains and codomains of those linear transformations. The Moore-Penrose inverse of a linear transformation such as $\mathbf{g}_{\lambda X}(G_0,\lambda_0,X_0)^{\dagger}$ is the linear transformation whose matrix representation is the Moore-Penrose inverse matrix of the matrix representation for $\mathbf{g}_{\lambda X}(G_0,\lambda_0,X_0)$ corresponding to the same bases.

3 Properties of the multiplicity support

The following lemma asserts a basic property of the multiplicity support.

Lemma 3.1 *Let $A \in \mathbb{C}^{n\times n}$ with an eigenvalue λ_* of multiplicity support $m\times k$. Then*

$$\mathit{Kernel}\big((A-\lambda_* I)^j\big) \;\subset\; \mathit{Range}(A-\lambda_* I) \quad \text{for} \quad j = 1,2,\ldots,k-1. \tag{6}$$

Furthermore, there is an open and dense subset $\mathcal{C}$ of $\mathbb{C}^{n\times m}$ such that, for every $C \in \mathcal{C}$, the solution $\mathbf{x}_*$ of the equation

$$C^{\mathsf{H}}\,\mathbf{x} \;=\; \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \quad \text{for} \quad \mathbf{x} \in \mathit{Kernel}(A-\lambda_* I) \tag{7}$$

uniquely exists and satisfies $\mathbf{x}_* \in \Big(\bigcap_{j=1}^{k-1}\mathit{Range}\big((A-\lambda_* I)^j\big)\Big) \;\setminus\; \mathit{Range}\big((A-\lambda_* I)^k\big)$.

Proof. From the multiplicity support of λ_*, there are m elementary Jordan blocks of sizes $\ell_1 \geq \cdots \geq \ell_m$ respectively with $\ell_m = k$ along with m sequences of generalized eigenvectors $\big\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{\ell_i}^{(i)}\big\}_{i=1}^{m}$ such that $(A - \lambda_* I)\mathbf{x}_{j+1}^{(i)} = \mathbf{x}_j^{(i)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, \ell_i - 1$. Moreover, $\mathit{Kernel}\big((A-\lambda_* I)^j\big) = \mathit{span}\big\{\mathbf{x}_l^{(i)} \;\big|\; 1 \leq l \leq j,\ 1 \leq i \leq m\big\}$ and thus (6) holds. Furthermore $(A - \lambda_* I)^j\, \mathbf{x}_{j+1}^{(i)} = \mathbf{x}_1^{(i)}$ for $j = 1, \ldots, \ell_i - 1$ and $i = 1, \ldots, m$. Namely every $\mathbf{z} \in \mathit{Kernel}(A - \lambda_* I)$ is in $\bigcap_{j=1}^{k-1} \mathit{Range}\big((A-\lambda_* I)^j\big)$ since $\ell_i \geq k$ for all i. However, $\mathbf{x}_1^{(m)} \notin \mathit{Range}\big((A - \lambda_* I)^k\big)$ since

$$(A - \lambda_* I)^k \left[\mathbf{x}_1^{(m)}, \cdots, \mathbf{x}_{\ell_m}^{(m)}\right] \;=\; \left[\mathbf{x}_1^{(m)}, \cdots, \mathbf{x}_{\ell_m}^{(m)}\right] J_k(0)^k \;=\; O$$

and $\mathbb{C}^n$ is the direct sum of those invariant subspaces, implying at least one vector in the basis of $\mathit{Kernel}(A - \lambda_* I)$ is not in $\mathit{Range}\big((A - \lambda_* I)^k\big)$ so the dimension of the subspace $\mathcal{K} = \mathit{Kernel}(A - \lambda_* I) \cap \mathit{Range}\big((A - \lambda_* I)^k\big)$ is less than m.

Let columns of $N \in \mathbb{C}^{n\times m}$ form an orthonormal basis for $\mathit{Kernel}(A - \lambda_* I)$ and denote $\mathcal{C}_0 = \big\{C \in \mathbb{C}^{n\times m} \;\big|\; (C^{\mathsf{H}} N)^{-1} \text{ exists}\big\}$, which is open since $(C + \Delta C)^{\mathsf{H}} N$ is invertible if $C^{\mathsf{H}} N$ is invertible and $\|\Delta C\|_2$ is sufficiently small. For any $C \notin \mathcal{C}_0$ so that $C^{\mathsf{H}} N$ is rank-deficient, we have $(C - \varepsilon\, N)^{\mathsf{H}} N = C^{\mathsf{H}} N - \varepsilon\, I$ is invertible for all $\varepsilon \notin \mathit{eig}(C^{\mathsf{H}} N)$ so $C - \varepsilon\, N \in \mathcal{C}_0$ and $\mathcal{C}_0$ is thus dense. The equation (7) then has a unique solution

$$\mathbf{x}_* \;=\; N\,(C^{\mathsf{H}}\, N)^{-1} \left[\begin{smallmatrix} 1 \\ \mathbf{0} \end{smallmatrix}\right]$$

for every $C \in \mathcal{C}_0$. Let $\mathcal{C} \subset \mathcal{C}_0$ such that the $\mathbf{x}_* \notin \mathcal{K}$ for every $C \in \mathcal{C}$. Clearly $\mathcal{C}$ is open since, for every $C \in \mathcal{C}$, we have $\hat{\mathbf{x}} = N\left((C + \Delta C)^{\mathsf{H}}\, N\right)^{-1} \left[\begin{smallmatrix} 1 \\ \mathbf{0} \end{smallmatrix}\right] \notin \mathcal{K}$ for sufficiently small $\|\Delta C\|_2$ and thus $C + \Delta C \in \mathcal{C}$. To show $\mathcal{C}$ is dense in $\mathcal{C}_0$, let $C \in \mathcal{C}_0$ with the corresponding $\mathbf{x}_* \in \mathcal{K}$. Since $\dim(\mathcal{K}) < m$, there is a unit vector $\hat{\mathbf{x}} \in \mathit{Kernel}(A - \lambda_* I) \setminus \mathcal{K}$. For any $\varepsilon \geq 0$, let $D^{(\varepsilon)} = -\frac{\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}}{\|\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}\|_2^2}\,\hat{\mathbf{x}}^{\mathsf{H}}\, C$. There is a $\mu > 0$ such that $\big\|D^{(\varepsilon)}\big\|_2 \leq \mu$ for all $\varepsilon \in [0, 1]$ since $\min_{\varepsilon\in[0,1]} \|\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}\|_2 > 0$. Then

$$\big(C + \varepsilon\, D^{(\varepsilon)}\big)^{\mathsf{H}}(\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}) \;=\; C^{\mathsf{H}}\,\mathbf{x}_* + \varepsilon\, C^{\mathsf{H}}\hat{\mathbf{x}} - \varepsilon\, C^{\mathsf{H}}\,\hat{\mathbf{x}}\, \frac{(\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}})^{\mathsf{H}}}{\|\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}\|_2^2}(\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}}) \;=\; \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$$

with $\mathbf{x}_* + \varepsilon\,\hat{\mathbf{x}} \in \mathit{Kernel}(A + \lambda_* I) \setminus \mathcal{K}$ for all $\varepsilon \in (0, 1)$ and $\big\|\varepsilon\, D^{(\varepsilon)}\big\|_2 < \varepsilon\,\mu$. Namely, the matrix $C + \varepsilon\, D^{(\varepsilon)} \in \mathcal{C}$ for sufficiently small ε and approaches to C when $\varepsilon \to 0$, implying $\mathcal{C}$ is dense in $\mathcal{C}_0$ that is dense in $\mathbb{C}^{n\times m}$ so the lemma is proved. □

The following lemma sets the foundation for our sensitivity analysis and algorithm development on a defective eigenvalue by laying out critical properties of the mapping (2).

Lemma 3.2 *Let $A \in \mathbb{C}^{n\times n}$ with $\lambda_* \in \mathit{eig}(A)$ of multiplicity support $m_* \times k_*$ and $\mathbf{g}$ be as in* (2) *with S and T as in* (3) *and* (1) *respectively. The following assertions hold.*

(i) *For almost all $C \in \mathbb{C}^{n\times m}$, there is an $X_* \in \mathbb{C}^{n\times k}$ such that $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ if and only if $m \leq m_*$ and $k \leq k_*$. Such an X_* is unique if and only if $m = m_*$.*

(ii) *Let $m \leq m_*$ and $k \leq k_*$. For almost all $C \in \mathbb{C}^{n\times m}$ in $\mathbf{g}$ with $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$, the linear transformation $\mathbf{g}_{G\lambda X}(A, \lambda_*, X_*)$ is surjective, and $\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)$ is injective if and only if $m = m_*$ and $k = k_*$.*

(iii) *Let* $m = m_*$, $k = k_*$ *and* $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$. *Then* C *and* S *can be modified so that the columns of* X_* *are orthonormal.*

Proof. Let $N \in \mathbb{C}^{n\times m_*}$ be a matrix whose columns span the kernel $\mathscr{Kernel}(A - \lambda_* I)$. We shall prove the assertion (i) by an induction. For almost all $C \in \mathbb{C}^{n\times m}$, the matrix $C^{\mathsf{H}} N$ is of full row rank if $m \leq m_*$ so that there is a $\mathbf{u} \in \mathbb{C}^m$ such that $(C^{\mathsf{H}} N)\,\mathbf{u} = T_{1:m,1}$ while $\mathbf{u}$ is unique if and only if $m = m_*$. For $m \leq m_*$, let $\mathbf{x}_1 = N\,\mathbf{u}$ and assume vectors $\mathbf{x}_1, \ldots, \mathbf{x}_j \in \mathbb{C}^n$ are obtained such that $1 \leq j < k$ and

$$\begin{aligned} (A - \lambda_* I)\big[\mathbf{x}_1, \cdots, \mathbf{x}_j\big] &= \big[\mathbf{x}_1, \cdots, \mathbf{x}_j\big]\, S_{1:j,1:j} \\ C^{\mathsf{H}}\big[\mathbf{x}_1, \cdots, \mathbf{x}_j\big] &= T_{1:m,1:j} \end{aligned}$$

Then $\mathbf{x}_1, \ldots, \mathbf{x}_j \in \mathscr{Kernel}\big((A - \lambda_* I)^j\big)$ from $(S_{1:j,1:j})^j = O$, and (6) implies the equation

$$(A - \lambda_* I)\,\mathbf{x} \;=\; s_{1,j+1}\mathbf{x}_1 + \cdots + s_{j,j+1}\mathbf{x}_j \;\equiv\; [\mathbf{x}_1, \ldots, \mathbf{x}_j]\, S_{1:j,j+1}$$

has a particular solution $\mathbf{u} \in \mathbb{C}^n$ and a unique solution $\mathbf{x}_{j+1} = \mathbf{u} - N\,(C^{\mathsf{H}} N)^{-1} C^{\mathsf{H}}\mathbf{u}$ such that $C^{\mathsf{H}}\,\mathbf{x}_{j+1} = \mathbf{0}$ when $m = m_*$. By induction, there is an $X_* = \big[\mathbf{x}_1, \cdots, \mathbf{x}_k\big] \in \mathbb{C}^{n\times k}$ such that (λ_*, X_*) is a solution to the system $\mathbf{g}(A, \lambda, X) = \mathbf{0}$ and X_* is unique if and only if $m = m_*$. The assertion (i) is proved.

Assume $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ and write $X_* = \big[\,\mathbf{x}_1, \cdots, \mathbf{x}_k\,\big]$. Then $\mathbf{x}_1 \neq \mathbf{0}$ and, by $S_{1:j,1:j}$ being upper triangular nilpotent, $\mathbf{x}_j \in \mathscr{Kernel}\big((A - \lambda_* I)^j\big) \setminus \mathscr{Kernel}\big((A - \lambda_* I)^{j-1}\big)$ for $j = 1, \ldots, k$ so X_* is of full column rank and $X_*^{\dagger} X_* = I$. Furthermore, the Jacobian $\mathbf{g}_{G\lambda X}(A, \lambda_*, X_*)$ is surjective since, for any $U \in \mathbb{C}^{n\times k}$ and $V \in \mathbb{C}^{m\times k}$, a straightforward calculation using (4) yields

$$\mathbf{g}_{G\lambda X}(A, \lambda_*, X_*)\big((U - (A - \lambda_* I)\, C^{\mathsf{H}\dagger} V + C^{\mathsf{H}\dagger} V S)\, X_*^{\dagger},\, 0,\, C^{\mathsf{H}\dagger}\, V\big) \;=\; \begin{pmatrix} U \\ V \end{pmatrix}$$

using $C^{\mathsf{H}}\, C^{\mathsf{H}\dagger} = I$ when C is of full column rank. Let $(A, \lambda_*, \hat{X})$ be a zero of $\mathbf{g}$ and assume $m = m_*$ and $k = k_*$. Then, for almost all $C \in \mathbb{C}^{n\times m}$, the solution $\mathbf{u} = \hat{\mathbf{u}}$ of the equation $(C^{\mathsf{H}}\, N)\,\mathbf{u} = T_{1:m,1}$ is unique and the first column of $\hat{X}$, from Lemma 3.1, is

$$\hat{\mathbf{x}}_1 \;=\; N\,\hat{\mathbf{u}} \;\in\; \Big(\textstyle\bigcap_{j=1}^{k-1} \mathscr{Range}\big((A - \lambda_* I)^j\big)\Big) \;\setminus\; \mathscr{Range}\big((A - \lambda_* I)^k\big). \tag{8}$$

Assume, for certain $(\sigma, Y) \in \mathbb{C} \times \mathbb{C}^{n\times k}$, its image $\mathbf{g}_{\lambda X}(A, \lambda_*, \hat{X})\big(\sigma, Y\big) \;=\; \mathbf{0}$. By (5),

$$-\,\sigma\,\hat{X} + (A - \lambda_* I)\, Y - Y\, S \;=\; O \tag{9}$$

$$C^{\mathsf{H}}\, Y \;=\; O. \tag{10}$$

Right-multiplying both sides of the equation (9) by S yields

$$\begin{aligned} Y\, S^2 + \sigma\,\hat{X}\, S &= (A - \lambda_* I)\, Y\, S \\ &= (A - \lambda_* I)^2\, Y - \sigma\,(A - \lambda_* I)\,\hat{X} && \text{(by (9))} \\ &= (A - \lambda_* I)^2\, Y - \sigma\,\hat{X}\, S, && \text{(by } (A - \lambda_* I)\hat{X} = \hat{X} S\text{)} \end{aligned}$$

namely

$$(A - \lambda_* I)^2\, Y \;=\; Y\, S^2 + 2\,\sigma\, \hat{X}\, S.$$

Continuing the process of recursive right-multiplying the above equation by S leads to

$$(A - \lambda_* I)^k\, Y \quad = \quad Y\, S^k + k\,\sigma\,\hat{X}\,S^{k-1} \quad = \quad k\,\sigma\, s_{12}\, s_{23} \,\cdots\, s_{k-1,k}\big[\, O_{n\times(k-1)},\; \hat{\mathbf{x}}_1 \,\big]$$

with $s_{12}\, s_{23} \cdots s_{k-1,k} \neq 0$. Hence $\sigma = 0$ due to (8). Denote columns of Y as $\mathbf{y}_1, \ldots, \mathbf{y}_k \in \mathbb{C}^n$. Then the first columns of the equations (9) and (10) are $(A - \lambda_* I)\,\mathbf{y}_1 = \mathbf{0}$ and $C^{\mathsf{H}}\,\mathbf{y}_1 = 0$ that imply $\mathbf{y}_1 = \mathbf{0}$. For $1 \leq j < k$, using $\sigma = 0$ and $\mathbf{y}_1 = \cdots = \mathbf{y}_j = \mathbf{0}$ on the $(j+1)$-th columns of the equations (9) and (10) we have $\mathbf{y}_{j+1} = \mathbf{0}$. Thus $Y = O$. As a result, $(A, \lambda_*, \hat{X})$ is a zero of $\mathbf{g}$ with injective partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_*, \hat{X})$.

If $m < m_*$, the solution (λ_*, X_*) of $\mathbf{g}(A, \lambda, X) = \mathbf{0}$ is on an algebraic variety of a positive dimension and thus $\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)$ is not injective. Let $m = m_*$, we now prove the partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_*, \hat{X})$ is injective *only if* $k = k_*$. Assume $k < k_*$ and write $\hat{X} = \big[\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_k\big]$. Since $\mathbf{g}(A, \lambda_*, \hat{X}) = \mathbf{0}$ and S is upper-triangular nilpotent, hence $\hat{\mathbf{x}}_j \in \mathcal{Kernel}\big((A - \lambda_* I)^j\big)$ for $j = 1, \ldots, k$. Then $k < k_*$ implies $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_k \in \mathcal{Range}(A - \lambda_* I)$. For almost all $C \in \mathbb{C}^{n\times m}$, the matrix $\left[\begin{smallmatrix} A-\lambda_* I \\ C^{\mathsf{H}} \end{smallmatrix}\right]$ is of full column rank and the vector $\mathbf{y}_1 = \frac{1}{s_{12}}\,\hat{\mathbf{x}}_2$ is the unique solution to the linear system $\left[\begin{smallmatrix} A-\lambda_* I \\ C^{\mathsf{H}} \end{smallmatrix}\right]\mathbf{z} = \left[\begin{smallmatrix} \hat{\mathbf{x}}_1 \\ \mathbf{0} \end{smallmatrix}\right]$. Using an induction, assume $\mathbf{y}_1, \ldots, \mathbf{y}_j \in \mathcal{span}\{\hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_{j+1}\}$ for any $j < k$ such that

$$\begin{aligned} -\big[\,\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_j\,\big] + (A - \lambda_* I)\big[\,\mathbf{y}_1, \cdots, \mathbf{y}_j\,\big] - \big[\,\mathbf{y}_1, \cdots, \mathbf{y}_j\,\big]\, S_{1:j,1:j} \quad &= \quad O \\ C^{\mathsf{H}}\big[\,\mathbf{y}_1, \cdots, \mathbf{y}_j\,\big] \quad &= \quad O. \end{aligned}$$

Then $\mathbf{y}_1, \ldots, \mathbf{y}_j \in \mathcal{Kernel}\big((A - \lambda_* I)^{j+1}\big)$ and (6) imply that there is a unique vector $\mathbf{z} = \mathbf{y}_{j+1} \in \mathcal{span}\{\mathbf{x}_1, \ldots, \hat{\mathbf{x}}_{j+1}\}$ satisfying

$$(A - \lambda_* I)\mathbf{z} \quad = \quad \hat{\mathbf{x}}_{j+1} + s_{1,j+1}\mathbf{y}_1 + \cdots + s_{j,j+1}\mathbf{y}_j \qquad \text{and} \qquad C^{\mathsf{H}}\,\mathbf{z} \quad = \quad \mathbf{0}.$$

Write $Y = \big[\,\mathbf{y}_1, \cdots, \mathbf{y}_k\,\big]$. We have $\mathbf{g}_{\lambda X}(A, \lambda_*, \hat{X})\big(1, Y\big) = \mathbf{0}$ and thus the partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_*, \hat{X})$ is not injective. As a result, the assertion (ii) is proved.

We now prove (iii). Let $\mathbf{g}(A, \lambda_*, \hat{X}) = \mathbf{0}$ for certain parameters C and S. We can assume C and S are properly scaled so that $\|\hat{X}_{1:n,1}\|_2 = 1$. Reset $C_{1:n,1}$ as $\hat{X}_{1:n,1}$, $\hat{X}_{1:n,2:k}$ as $\hat{X}_{1:n,2:k} - \hat{X}_{1:n,1}\,(\hat{X}_{1:n,1})^{\mathsf{H}}\hat{X}_{1:n,2:k}$ and $S_{1,1:k}$ as $S_{1,1:k} + (\hat{X}_{1:n,1})^{\mathsf{H}}\hat{X}_{1:n,2:k}\,S_{2:k,1:k}$ so that $\mathbf{g}(A, \lambda_*, \hat{X}) = \mathbf{0}$ still holds and $(\hat{X}_{1:n,2:k})^{\mathsf{H}}\hat{X}_{1:n,1} = \mathbf{0}$. As a result, there is a thin QR decomposition $\hat{X} = Q\,R$ with $R_{1,1:k} = [1, 0, \cdots, 0]$. Reset $X_* = Q$ and S as $R\,S\,R^{-1}$. It is thus a straightforward verification that $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ with $(X_*)^{\mathsf{H}}X_* = I$. $\square$

4 Sensitivity of a defective eigenvalue

Based on Lemma 3.2 and the Implicit Function Theorem, the following lemma establishes the defective eigenvalue as a holomorphic function of certain entries of the matrix.

Lemma 4.1 *Assume* $A \in \mathbb{C}^{n\times n}$ *and* $\lambda_* \in \mathit{eig}(A)$ *of multiplicity support* $m\times k$. *Let* $\mathbf{g}$ *be defined in* (2) *using proper parameters* $C \in \mathbb{C}^{n\times m}$ *and* $S \in \mathbb{C}^{n\times k}$ *so that* $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ *with a surjective* $\mathbf{g}_{G\lambda X}(A, \lambda_*, X_*)$ *and an injective* $\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)$. *There is a neighborhood* Ω *of certain* $\mathbf{z}_*$ *in* $\mathbb{C}^{n^2-m\,k+1}$ *and a neighborhood* Σ *of* (A, λ_*, X_*) *in* $\mathbb{C}^{n\times n}\times\mathbb{C}\times\mathbb{C}^{n\times k}$ *along with holomorphic mappings* $G : \Omega \longrightarrow \mathbb{C}^{n\times n}$, $\lambda : \Omega \longrightarrow \mathbb{C}$ *and* $X : \Omega \longrightarrow \mathbb{C}^{n\times k}$ *with* $\big(G(\mathbf{z}_*), \lambda(\mathbf{z}_*), X(\mathbf{z}_*)\big) = (A, \lambda_*, X_*)$ *such that* $\mathbf{g}(G_0, \lambda_0, X_0) = \mathbf{0}$ *for* $(G_0, \lambda_0, X_0) \in \Sigma$ *if and only if* $(G_0, \lambda_0, X_0) = (G(\mathbf{z}_0), \lambda(\mathbf{z}_0), X(\mathbf{z}_0))$ *for certain* $\mathbf{z}_0 \in \Omega$.

Proof. Since the mapping $(G, \lambda, X) \mapsto \mathbf{g}(G, \lambda, X)$ has a surjective $\mathbf{g}_{G\lambda X}(A, \lambda_*, X_*)$ to $\mathbb{C}^{n\times k}\times\mathbb{C}^{m\times k}$ and an injective $\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)$ from $\mathbb{C}\times\mathbb{C}^{n\times k}$, there are $m\,k-1$ entries of the variable $G \in \mathbb{C}^{n\times n}$ forming a variable $\mathbf{y}$ such that the partial Jacobian $\mathbf{g}_{\mathbf{y}\lambda X}(A, \lambda_*, X_*)$ is invertible. By the Implicit Function Theorem, the remaining entries of G excluding $\mathbf{y}$ form a variable vector $\mathbf{z} \in \mathbb{C}^{n^2-m\,k+1}$ so that the assertion holds. □

From the proof of Lemma 4.1, the components of the variable $\mathbf{z}$ are identical to $n^2-m\,k+1$ entries of the matrix $G(\mathbf{z})$. We can now establish one of the main theorems of this paper.

Theorem 4.2 (Eigenvalue Sensitivity Theorem) *The sensitivity of an eigenvalue is finitely bounded if its multiplicity support is preserved. More precisely, let* $A \in \mathbb{C}^{n\times n}$ *and* $\lambda_* \in \mathit{eig}(A)$ *with a multiplicity support* $m\times k$. *There is a neighborhood* Φ *of* (A, λ_*) *in* $\mathbb{C}^{n\times n}\times\mathbb{C}$ *and a neighborhood* Ω *of certain* $\mathbf{z}_*$ *in* $\mathbb{C}^{n^2-m\,k+1}$ *along with holomorphic mappings* $G : \Omega \to \mathbb{C}^{n\times n}$ *and* $\lambda : \Omega \to \mathbb{C}$ *with* $(A, \lambda_*) = \big(G(\mathbf{z}_*), \lambda(\mathbf{z}_*)\big)$ *such that every* $(\tilde{A}, \tilde{\lambda}) \in \Phi$ *with* $\tilde{\lambda} \in \mathit{eig}\big(\tilde{A}\big)$ *of multiplicity support* $m\times k$ *is equal to* $\big(G(\tilde{\mathbf{z}}), \lambda(\tilde{\mathbf{z}})\big)$ *for certain* $\tilde{\mathbf{z}} \in \Omega$. *Furthermore,*

$$\limsup_{\mathbf{z}\to\mathbf{z}_*} \frac{\big|\lambda(\mathbf{z}) - \lambda_*\big|}{\|G(\mathbf{z}) - A\|_F} \quad\le\quad \big\|\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)^\dagger\big\|_2 \quad<\quad \infty \tag{11}$$

where $X_* \in \mathbb{C}^{n\times k}$ *satisfies* $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ *for the mapping* $\mathbf{g}$ *defined in* (2) *that renders columns of* X_* *orthonormal.*

Proof. Let Σ and Ω be the neighborhoods specified in Lemma 4.1 along with the holomorphic mappings G and λ. For any $(\tilde{A}, \tilde{\lambda})$ sufficiently close to (A, λ_*) with $\tilde{\lambda} \in \mathit{eig}(A)$ of multiplicity support $m\times k$, the matrix $\left[\begin{smallmatrix}\tilde{A}-\tilde{\lambda}I\\ C^{\mathsf{H}}\end{smallmatrix}\right]$ is of full rank so there is a unique $\tilde{X}$ such that $\mathbf{g}(\tilde{A}, \tilde{\lambda}, \tilde{X}) = \mathbf{0}$. Furthermore, the linear transformation $X \mapsto \big((\tilde{A}-\tilde{\lambda}I)\,X - X\,S,\ C^{\mathsf{H}}X\big)$ is injective from $\mathbb{C}^{n\times k}$ to $\mathbb{C}^{n\times k}\times\mathbb{C}^{m\times k}$, implying $\|\tilde{X}-X_*\|_F$ can be as small as needed so that $(\tilde{A}, \tilde{\lambda}, \tilde{X}) \in \Sigma$ and thus $(\tilde{A}, \tilde{\lambda}) = (G(\mathbf{z}), \lambda(\mathbf{z}))$ for certain $\mathbf{z} \in \Omega$. Consequently, the neighborhood Φ of (A, λ_*) exists.

From Lemma 4.1, we have $\mathbf{g}\big(G(\mathbf{z}), \lambda(\mathbf{z}), X(\mathbf{z})\big) \equiv \mathbf{0}$ for all $\mathbf{z} \in \Omega$. As a result,

$$\begin{aligned}\mathbf{0} \quad&=\quad \left(\frac{\partial \mathbf{g}(G(\mathbf{z}),\lambda(\mathbf{z}),X(\mathbf{z}))}{\partial \mathbf{z}}\bigg|_{\mathbf{z}=\mathbf{z}_*}\right)(\mathbf{z}-\mathbf{z}_*)\\ &=\quad \mathbf{g}_G(A, \lambda_*, X_*)\,G_{\mathbf{z}}(\mathbf{z}_*)\,(\mathbf{z}-\mathbf{z}_*) + \mathbf{g}_{\lambda X}(A, \lambda_*, X_*)\left(\frac{\partial(\lambda(\mathbf{z}),X(\mathbf{z}))}{\partial \mathbf{z}}\bigg|_{\mathbf{z}=\mathbf{z}_*}\right)(\mathbf{z}-\mathbf{z}_*)\end{aligned}$$

implying

$$\begin{aligned} \big|\lambda(\mathbf{z}) - \lambda_*\big| &\le \big\|\big(\lambda(\mathbf{z}), X(\mathbf{z})\big) - \big(\lambda_*, X_*\big)\big\|_2 \\ &= \Big\|\tfrac{\partial(\lambda(\mathbf{z}),X(\mathbf{z}))}{\partial \mathbf{z}}\Big|_{\mathbf{z}=\mathbf{z}_*}(\mathbf{z}-\mathbf{z}_*)\Big\|_2 + O\big(\|\mathbf{z}-\mathbf{z}_*\|_2^2\big) \\ &= \big\|\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)^\dagger\,\mathbf{g}_G(A,\lambda_*,X_*)\,G_{\mathbf{z}}(\mathbf{z}_*)\,(\mathbf{z}-\mathbf{z}_*)\big\|_2 + O\big(\|\mathbf{z}-\mathbf{z}_*\|_2^2\big) \\ &\le \big\|\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)^\dagger\big\|_2\,\big\|G(\mathbf{z})-A\big\|_F + O\big(\|\mathbf{z}-\mathbf{z}_*\|_2^2\big) \end{aligned}$$

since the partial Jacobian $\mathbf{g}_G(A,\lambda_*,X_*)$ is the linear transformation $G \mapsto G\,X_*$ with a unit operator norm due to orthonormal columns of X_*, leading to (11). The norm $\big\|\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)^\dagger\big\|_2$ is finite because $\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)$ is injective by Lemma 3.2. □

In light of Theorem 4.2, we introduce the $m\times k$ *condition number*

$$\tau_{A,m\times k}(\lambda_*) \quad := \quad \inf_{C,S} \big\|\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)^\dagger\big\|_2 \tag{12}$$

of an eigenvalue $\lambda_* \in \mathit{eig}(A)$ where $\mathbf{g}$ is as in (2) and the infimum is taken over all the proper choices of matrix parameters C and S that render the columns of the unique X_* orthonormal so that $\mathbf{g}(A,\lambda_*,X_*) = \mathbf{0}$. We shall refer to $\tau_{A,m\times k}(\lambda_*)$ as the *multiplicity support condition number* if the specific m and k are irrelevant in the discussion. From Lemma 3.2, the $m\times k$ condition number is infinity *only if* either m is less than the actual geometric multiplicity or k is less than the Segre anchor. Consequently, the condition number $\tau_{A,m\times k}(\lambda_*)$ is large only if A is close to a matrix $\tilde{A}$ that possesses an eigenvalue $\tilde{\lambda} \approx \lambda_*$ whose multiplicity support is $\tilde{m}\times\tilde{k}$ with either $\tilde{m} > m$ or $\tilde{k} > k$. As a special case, the condition number $\tau_{A,1\times 1}(\lambda_*)$ measures the sensitivity of a simple eigenvalue λ_*.

We can now revisit the old question:

Is a defective eigenvalue hypersensitive to perturbations?

The answer is not as simple as the question may seem to be. It is well documented in the literature that, under an *arbitrary* perturbation ΔA on the matrix A, a defective eigenvalue of A generically disperses into a cluster of eigenvalues with an error bound proportional to $\|\Delta A\|_2^{\frac{1}{l}}$ where l is the size of the largest Jordan block associated with the eigenvalue [2, p. 58][3, 13]. Similar and related sensitivity results can be found in the works such as [1, 14, 15]. This error bound implies that the asymptotic sensitivity of a defective eigenvalue is infinity, and only a fraction $\frac{1}{l}$ of the data accuracy passes on to the accuracy of the eigenvalue. For instance, if the largest Jordan block is 5×5, only three correct digits can be expected from the computed eigenvalues regarding the defective eigenvalue since one fifth the hardware precision (about 16 digits) remains in the forward accuracy.

It is also known that the mean of the cluster emanating from the defective eigenvalue under perturbations is not hypersensitive [11, 17]. Kahan is the first to discover the finite sensitivity $\frac{1}{m}\|P\|_2$ of a multiple eigenvalue under constrained perturbations that preserve the algebraic multiplicity m, where P is the spectral projector associated with the eigenvalue. This spectral projector norm is large only if a small perturbation on the matrix can increase the multiplicity [10]. As pointed out by Kahan, the seemingly infinite sensitivity of a multiple eigenvalue may not be a conceptually meaningful measurement for the condition of a *multiple eigenvalue* since arbitrary perturbations do not maintain the characteristics of the eigenvalue as being multiple. Theorem 4.2 sheds light on another intriguing and pleasant property of a defective eigenvalue: Its algebraic multiplicity does *not* need to be maintained under data perturbations for its sensitivity to be under control, as long as the geometric multiplicity *and* the Segre anchor are preserved. As a result, the condition number $\tau_{A,m\times k}(\lambda_*)$ provides a new and different measurement on the sensitivity of a *defective* eigenvalue λ_* when its multiplicity support is preserved.

The same eigenvalue can be ill-conditioned in the spectral projector norm while being well conditioned in multiplicity support condition number and vice versa (c.f. Example 4 in §10) with no contradiction whatsoever.

More importantly, the finite sensitivity enables accurate numerical computation of a defective eigenvalue from imposing the constraints on the multiplicity support, as we shall demonstrate in later sections. Even if perturbations are unconstrained, the problem of computing a defective eigenvalue may not have to be hypersensitive at all if the problem is properly generalized, i.e. regularized. We shall prove in Theorem 5.2 that the $m \times k$ condition number still provides the finitely bounded sensitivity of λ_* as what we call the $m \times k$ pseudo-eigenvalue of A, and this condition number is large only if m or k can be increased by small perturbations.

There are further subtleties on the condition of a defective eigenvalue. The sensitivity is finitely bounded if the multiplicity or the multiplicity support of *the* eigenvalue is preserved. Denote the collection of $n \times n$ complex matrices having an eigenvalue that shares the same multiplicity support $m \times k$ as $\mathcal{E}^n_{m\times k}$. Every $A \in \mathcal{E}^n_{m\times k}$ has an eigenvalue λ_* along with an X_* such that (A, λ_*, X_*) belongs to an algebraic variety defined by the solution set of the polynomial system $\mathbf{g}(G, \lambda, X) = \mathbf{0}$. The set $\mathcal{E}^n_{m\times k}$ is not a manifold in general so the Tubular Neighborhood Theorem does not apply. As a result, maintaining a multiplicity support is not enough to dampen the sensitivity of a particular defective eigenvalue with that multiplicity support. The matrix staying on $\mathcal{E}^n_{m\times k}$ does not guarantee the finite sensitivity of a defective eigenvalue. If a matrix $A \in \mathcal{E}^n_{m\times k}$ has two eigenvalues of the same multiplicity support $m \times k$, then A is in the intersection of images of two holomorphic mappings described in Lemma 4.1. When A drifts on $\mathcal{E}^n_{m\times k}$, the multiplicity support $m \times k$ may be maintained for one eigenvalue but lost on the other. Consequently, the *other* defective eigenvalue still disperses into a cluster.

5 A well-posed defective eigenvalue problem

A mathematical problem is said to be well-posed if its solution satisfies three crucial properties: Existence, uniqueness and Lipschitz continuity. The problem of finding an eigenvalue of a matrix in its conventional meaning is ill-posed when the eigenvalue is defective because the sensitivity of the eigenvalue is infinite with respect to arbitrary perturbations on the matrix. Lacking Lipschitz continuity with respect to data, such a problem is not suitable for numerical computation unless the problem is properly modified, or better known as being *regularized.*

We can alter the problem of

$$\textit{finding an eigenvalue of a matrix } A$$

to

$$\textit{finding a } \lambda_* \textit{ so that } (\lambda_*, X_*) \textit{ is local least squares solution to } \mathbf{g}(A,\lambda,X) = \mathbf{0}$$

where $\mathbf{g}$ is the mapping defined in (2) with proper parameters. We shall show that the latter problem is a regularization of the former.

For any fixed matrix A, a local least squares solution $(\hat{\lambda}, \hat{X})$ to the equation $\mathbf{g}(A,\lambda,X) = \mathbf{0}$ is the minimum point to $\|\mathbf{g}(A,\lambda,X)\|_2$ in an open subset of $\mathbb{C} \times \mathbb{C}^{n\times k}$ where

$$\mathbf{g}_{\lambda X}(A,\hat{\lambda},\hat{X})^\dagger\, \mathbf{g}(A,\hat{\lambda},\hat{X}) \quad = \quad \mathbf{0}$$

if $\mathbf{g}_{\lambda X}(A,\hat{\lambda},\hat{X})$ is injective. The least squares solution (λ_*, X_*) of $\mathbf{g}(A,\lambda,X) = \mathbf{0}$ can be solved by the Gauss-Newton iteration

$$(\lambda_{j+1}, X_{j+1}) \quad = \quad (\lambda_j, X_j) - \mathbf{g}_{\lambda X}(A,\lambda_j,X_j)^\dagger\, \mathbf{g}(A,\lambda_j,X_j), \qquad j = 0,1,\ldots \tag{13}$$

based on the following local convergence lemma that is adapted from [24, Lemma 2].

Lemma 5.1 *[24] Let* $\mathbf{g}$ *be the mapping in* (2)*. For a fixed* $A \in \mathbb{C}^{n\times n}$*, assume* (λ_*, X_*) *is a local least squares solution to* $\mathbf{g}(A,\lambda,X) = \mathbf{0}$ *with an injective* $\mathbf{g}_{\lambda X}(A,\lambda_*,X_*)$*. There is an open convex neighborhood* D *of* (λ_*, X_*) *and constants* $\zeta, \gamma > 0$ *such that*

$$\left\|\mathbf{g}_{\lambda X}(A,\lambda,X)^\dagger\right\|_2 \;\leq\; \zeta, \tag{14}$$

$$\left\|\mathbf{g}(A,\lambda,X) - \mathbf{g}(A,\tilde{\lambda},\tilde{X}) - \mathbf{g}_{\lambda X}(A,\tilde{\lambda},\tilde{X})\left((\lambda,X) - (\tilde{\lambda},\tilde{X})\right)\right\|_2 \;\leq\; \gamma \left\|(\lambda,X) - (\tilde{\lambda},\tilde{X})\right\|_2^2 \tag{15}$$

for all $(\lambda,X), (\tilde{\lambda},\tilde{X}) \in \overline{D}$*. Assume there is a* $\sigma < 1$ *such that, for all* $(\lambda,X) \in D$*,*

$$\left\|\left(\mathbf{g}_{\lambda X}(A,\lambda,X)^\dagger - \mathbf{g}_{\lambda X}(A,\lambda_*,X_*)^\dagger\right) \mathbf{g}(A,\lambda_*,X_*)\right\|_2 \;\leq\; \sigma \left\|(\lambda,X) - (\lambda_*,X_*)\right\|_2. \tag{16}$$

Then, from all $(\lambda_0, X_0) \in D$ *such that* $\left\|(\lambda_0,X_0) - (\lambda_*,X_*)\right\|_2 < \frac{1-\sigma}{\zeta\gamma}$ *and*

$$\left\{(\lambda,X) \in \mathbb{C}\times\mathbb{C}^{n\times k} \;\middle|\; \|(\lambda,X) - (\lambda_*,X_*)\|_2 < \|(\lambda_0,X_0) - (\lambda_*,X_*)\|_2\right\} \;\subset\; D, \tag{17}$$

the Gauss-Newton iteration (13) *is well defined in* D*, converges to* (λ_*, X_*) *and satisfies*

$$\left\|(\lambda_{j+1},X_{j+1}) - (\lambda_*,X_*)\right\|_2 \;\leq\; \mu \left\|(\lambda_j,X_j) - (\lambda_*,X_*)\right\|_2$$

for $j = 0,1,\ldots$ *with* $\mu = \sigma + \zeta\gamma\left\|(\lambda_0,X_0) - (\lambda_*,X_*)\right\|_2 < 1$. □

When the matrix A has an eigenvalue λ_* of multiplicity support $m\times k$, there is an X_* such that (λ_*, X_*) is an exact solution to $\mathbf{g}(A,\lambda,X) = \mathbf{0}$. However, when A is known through its empirical data in $\tilde{A}$, a local least squares solution $(\tilde{\lambda}, \tilde{X})$ to the equation $\mathbf{g}(\tilde{A},\lambda,X) = \mathbf{0}$ generally has a residual $\left\|\mathbf{g}(\tilde{A},\tilde{\lambda},\tilde{X})\right\|_2 > 0$, and $\tilde{\lambda}$ may not be an eigenvalue of either A or $\tilde{A}$. For the convenience of elaboration, we call such a $\tilde{\lambda}$ an $m\times k$ *pseudo-eigenvalue* of $\tilde{A}$. By changing the conventional problem of computing an eigenvalue to a modified problem of finding a pseudo-eigenvalue, the defective eigenproblem is regularized as a well-posed problem as asserted in the main theorem of this paper.

Theorem 5.2 (Pseudo-Eigenvalue Theorem) *Let* λ_* *be an eigenvalue of* $A \in \mathbb{C}^{n\times n}$ *with a multiplicity support* $m \times k$ *along with* $X_* \in \mathbb{C}^{n\times k}$ *satisfying* $\mathbf{g}(A, \lambda_*, X_*) = \mathbf{0}$ *where* $\mathbf{g}$ *is as in* (2) *with proper parameters* C *and* S. *The following assertions hold.*

(i) *The exact eigenvalue* λ_* *of* A *is an* $m \times k$ *pseudo-eigenvalue of* A.

(ii) *There are neighborhoods* Φ *of* A *in* $\mathbb{C}^{n\times n}$ *and* Λ *of* λ_* *in* $\mathbb{C}$ *such that every matrix* $\tilde{A} \in \Phi$ *has a unique* $m \times k$ *pseudo-eigenvalue* $\tilde{\lambda} \in \Lambda$ *that is Lipschitz continuous with respect to* $\tilde{A}$.

(iii) *For every matrix* $\check{A} \in \Phi$ *serving as empirical data of* A, *there is a unique* $m \times k$ *pseudo-eigenvalue* $\check{\lambda} \in \Lambda$ *of* $\check{A}$ *such that*

$$\big|\check{\lambda} - \lambda_*\big| \quad \le \quad \tau_{A,m\times k}(\lambda_*)\,\big\|\check{A} - A\big\|_2 + O\big(\big\|\check{A} - A\big\|_2^2\big). \tag{18}$$

(iv) *The* $\check{\lambda}$ *in* (iii) *is an exact eigenvalue of* $\check{A} + E\,\check{X}^\dagger$ *with a Jordan block of size at least* k *where* $\check{X}$ *is the least squares solution of* $\mathbf{g}(\check{A}, \check{\lambda}, X) = \mathbf{0}$ *and* $E = (\check{A} - \check{\lambda}\,I)\,\check{X} - \check{X}\,S$. *When* $\check{X}^{\mathsf{H}}\check{X} = I$, *the backward error* $\|E\,\check{X}^\dagger\|_F$ *is bounded by* $\big\|\mathbf{g}(\check{A}, \check{\lambda}, \check{X})\big\|_2$.

Proof. The assertion (i) is a result of Lemma 3.2 (i). For any $r > 0$, denote $\Psi_r = \big\{(\lambda, X) \in \mathbb{C}\times\mathbb{C}^{n\times k} \;\big|\; \|(\lambda, X) - (\lambda_*, X_*)\|_2 < r\big\}$ and let $r_0 > 0$ such that $\{A\}\times\overline{\Psi_{r_0}}$ is a subset of Σ in Lemma 4.1. Let $r \in (0, r_0)$. Assume there is a matrix $\tilde{A}$ with $\|\tilde{A} - A\|_2 < \varepsilon$ for any $\varepsilon > 0$ such that $\min_{(\lambda,X)\in\overline{\Psi_r}} \big\|\mathbf{g}(\tilde{A}, \lambda, X)\big\|_2$ is not attainable in Ψ_r. Let $\varepsilon \to 0$. Then $\tilde{A} \to A$ and there exists an $(\hat{\lambda}, \hat{X}) \in \overline{\Psi_r}\setminus\Psi_r$ such that $\|\mathbf{g}(A, \hat{\lambda}, \hat{X})\|_2$ is the minimum 0 of $\big\|\mathbf{g}(A, \lambda, X)\big\|_2$ for $(\lambda, X) \in \overline{\Psi_r}$ and $(\hat{\lambda}, \hat{X}) \ne (\lambda_*, X_*)$. This is a contradiction to Lemma 4.1. As a result, there is a neighborhood Φ_r of A for every $r \in (0, r_0)$ such that $\min_{(\lambda,X)\in\Psi_r} \|\mathbf{g}(\tilde{A}, \lambda, X)\|_2$ is attainable at certain $(\tilde{\lambda}, \tilde{X}) \in \Psi_r$ for every $\tilde{A} \in \Phi_r$, implying the existence of the pseudo-eigenvalue $\tilde{\lambda}$.

By Lemma 5.1, we can assume $r_1 \in (0, r_0)$ is small so that the inequalities (14), (15) and (16) hold for $\sigma = 0$ and $\|(\lambda, X) - (\tilde{\lambda}, \tilde{X})\|_2 < \frac{1}{2\,(2\zeta)\,(2\gamma)}$ for all $(\lambda, X), (\tilde{\lambda}, \tilde{X}) \in \overline{\Psi_{r_1}}$. By the continuity of $\mathbf{g}$, the corresponding Φ_{r_1} can be chosen so that, for every $\hat{A} \in \Phi_{r_1}$ with a local minimum point $(\hat{\lambda}, \hat{X}) \in \Psi_{r_1}$ for $\|\mathbf{g}(\hat{A}, \lambda, X)\|_2$, we have $\big\|\mathbf{g}_{\lambda X}(\hat{A}, \lambda, X)^\dagger\big\|_2 < 2\zeta$,

$$\begin{aligned}
\big\|\mathbf{g}(\hat{A}, \lambda, X) - \mathbf{g}(\hat{A}, \tilde{\lambda}, \tilde{X}) - \mathbf{g}_{\lambda X}(\hat{A}, \tilde{\lambda}, \tilde{X})\,((\lambda, X) - (\tilde{\lambda}, \tilde{X}))\big\|_2 \quad &< \quad 2\gamma\,\big\|(\lambda, X) - (\tilde{\lambda}, \tilde{X})\big\|_2^2, \\
\big\|\big(\mathbf{g}_{\lambda X}(\hat{A}, \lambda, X)^\dagger - \mathbf{g}_{\lambda X}(\hat{A}, \hat{\lambda}, \hat{X})^\dagger\big)\,\mathbf{g}(\hat{A}, \hat{\lambda}, \hat{X})\big\|_2 \quad &\le \quad \tfrac{1}{2}\,\big\|(\lambda, X) - (\hat{\lambda}, \hat{X})\big\|_2
\end{aligned}$$

for all $(\lambda, X), (\tilde{\lambda}, \tilde{X}) \in \Psi_{r_1}$. Let $r_2 = \frac{1}{3}r_1$, $\Psi = \Psi_{r_2}$ and $\Phi = \Phi_{r_1} \cap \Phi_{r_2}$. For every $\hat{A} \in \Phi$, the minimum of $\|\mathbf{g}(\hat{A}, \lambda, X)\|_2$ is attainable at $(\hat{\lambda}, \hat{X}) \in \Psi$ and, for any initial iterate $(\lambda_0, X_0) \in \Psi$, we have $\|(\lambda_0, X_0) - (\hat{\lambda}, \hat{X})\|_2 < \frac{1}{2\,(2\zeta)\,(2\gamma)} = \frac{1-\frac{1}{2}}{(2\zeta)\,(2\gamma)}$ and the set

$$\Omega \quad = \quad \big\{(\lambda, X) \in \mathbb{C}\times\mathbb{C}^{n\times k} \;\big|\; \|(\lambda, X) - (\hat{\lambda}, \hat{X})\|_2 < \|(\lambda_0, X_0) - (\hat{\lambda}, \hat{X})\|_2\big\} \quad \subset \quad \Psi_{r_1}$$

since, for every $(\lambda, X) \in \Omega$, we have

$$\begin{aligned}
\|(\lambda, X) - (\lambda_*, X_*)\|_2 \quad &\le \quad \|(\lambda, X) - (\hat{\lambda}, \hat{X})\|_2 + \|(\hat{\lambda}, \hat{X}) - (\lambda_*, X_*)\|_2 \\
&< \quad \|(\lambda_0, X_0) - (\hat{\lambda}, \hat{X})\|_2 + r_2 \\
&\le \quad \|(\lambda_0, X_0) - (\lambda_*, X_*)\|_2 + \|(\lambda_*, X_*) - (\hat{\lambda}, \hat{X})\|_2 + r_2 \\
&< \quad r_2 + r_2 + r_2 \quad = \quad r_1
\end{aligned}$$

By Lemma 5.1, for every $(\lambda_0, X_0) \in \Psi$, the Gauss-Newton iteration on the equation $\mathbf{g}(\hat{A}, \lambda, X) = \mathbf{0}$ converges to $(\hat{\lambda}, \hat{X})$. This local minimum point $(\hat{\lambda}, \hat{X})$ is unique in Ψ because, assuming there is another minimum point $(\check{\lambda}, \check{X}) \in \Psi$ of $\|\mathbf{g}(\hat{A}, \lambda, X)\|_2$, the Gauss-Newton iteration converges to $(\hat{\lambda}, \hat{X})$ from the initial point $(\check{\lambda}, \check{X})$. On the other hand, the Gauss-Newton iteration from the local minimum point $(\check{\lambda}, \check{X})$ must stay at $(\check{\lambda}, \check{X})$, implying $(\check{\lambda}, \check{X}) = (\hat{\lambda}, \hat{X})$.

On the Lipschitz continuity of the pseudo-eigenvalue, let $\tilde{A}, \check{A} \in \Phi$ with minimum points $(\tilde{\lambda}, \tilde{X})$ and $(\check{\lambda}, \check{X})$ of $\|\mathbf{g}(\tilde{A}, \lambda, X)\|_2$ and $\|\mathbf{g}(\check{A}, \lambda, X)\|_2$ respectively in Ψ. The one-step Gauss-Newton iterate from $(\tilde{\lambda}, \tilde{X})$ on the equation $\mathbf{g}(\check{A}, \lambda, X) = \mathbf{0}$ toward $(\check{\lambda}, \check{X})$

$$(\lambda_1, X_1) \;=\; (\tilde{\lambda}, \tilde{X}) - \mathbf{g}_{\lambda X}(\check{A}, \tilde{\lambda}, \tilde{X})^{\dagger}\, \mathbf{g}(\check{A}, \tilde{\lambda}, \tilde{X})$$

yields $\big\|(\lambda_1, X_1) - (\check{\lambda}, \check{X})\big\|_2 \le \mu \big\|(\tilde{\lambda}, \tilde{X}) - (\check{\lambda}, \check{X})\big\|_2$ with $0 \le \mu < 1$ by Lemma 5.1. Thus

$$\begin{aligned}
\big\|(\check{\lambda}, \check{X}) - (\tilde{\lambda}, \tilde{X})\big\|_2 \;&\le\; \big\|(\check{\lambda}, \check{X}) - (\lambda_1, X_1)\big\|_2 + \big\|(\lambda_1, X_1) - (\tilde{\lambda}, \tilde{X})\big\|_2 \\
&\le\; \mu \big\|(\check{\lambda}, \check{X}) - (\tilde{\lambda}, \tilde{X})\big\|_2 + \big\|(\lambda_1, X_1) - (\tilde{\lambda}, \tilde{X})\big\|_2
\end{aligned}$$

Using the identity $\mathbf{g}_{\lambda X}(\tilde{A}, \tilde{\lambda}, \tilde{X})^{\dagger}\, \mathbf{g}(\tilde{A}, \tilde{\lambda}, \tilde{X}) = \mathbf{0}$ and the Lipschitz continuity of $\mathbf{g}$ and $\mathbf{g}_{\lambda X}$, there is a constant γ such that

$$\begin{aligned}
\big\|(\check{\lambda}, \check{X}) - (\tilde{\lambda}, \tilde{X})\big\|_2 \;&\le\; \frac{1}{1-\mu} \big\|(\lambda_1, X_1) - (\tilde{\lambda}, \tilde{X})\big\|_2 \\
&=\; \frac{1}{1-\mu} \big\|\mathbf{g}_{\lambda X}(\check{A}, \tilde{\lambda}, \tilde{X})^{\dagger} \mathbf{g}(\check{A}, \tilde{\lambda}, \tilde{X}) - \mathbf{g}_{\lambda X}(\tilde{A}, \tilde{\lambda}, \tilde{X})^{\dagger} \mathbf{g}(\tilde{A}, \tilde{\lambda}, \tilde{X})\big\|_2 \\
&\le\; \frac{1}{1-\mu} \Big(\big\|\mathbf{g}_{\lambda X}(\check{A}, \tilde{\lambda}, \tilde{X})^{\dagger}\big\|_2 \, \big\|\mathbf{g}(\check{A}, \tilde{\lambda}, \tilde{X}) - \mathbf{g}(\tilde{A}, \tilde{\lambda}, \tilde{X})\big\|_2 \\
&\qquad + \big\|\mathbf{g}_{\lambda X}(\check{A}, \tilde{\lambda}, \tilde{X})^{\dagger} - \mathbf{g}_{\lambda X}(\tilde{A}, \tilde{\lambda}, \tilde{X})^{\dagger}\big\|_2 \big\|\mathbf{g}(\tilde{A}, \tilde{\lambda}, \tilde{X})\big\|_2 \Big) \\
&\le\; \gamma \big\|\tilde{A} - \check{A}\big\|_2
\end{aligned}$$

for all $\tilde{A}, \check{A} \in \Phi$. Namely, the $m \times k$ pseudo-eigenvalue is Lipschitz continuous with respect to the matrix. Furthermore, by setting $(\tilde{A}, \tilde{\lambda}, \tilde{X}) = (A, \lambda_*, X_*)$ in the above inequalities we have (18) because the residual $\big\|\mathbf{g}(A, \lambda_*, X_*)\big\|_2 = 0$ Thus $\mu = 0$ and (iii) is proved.

For the assertion (iv), the matrix $\check{X}$ is of full column rank since X_* is and the least squares solution of $\mathbf{g}(G, \lambda, X) = \mathbf{0}$ is continuous, implying $\check{X}^{\dagger} \check{X} = I$ and thus $E = E\, \check{X}^{\dagger} \check{X}$ leading to $(\check{A} - E\, \check{X}^{\dagger} - \check{\lambda} I)\check{X} = \check{X}\, S$. The eigenvalue $\check{\lambda}$ of $\check{A} + E\, \check{X}^{\dagger}$ corresponds to a Jordan block of size at least k since S in (3) is nilpotent of rank $k-1$. □

The Pseudo-Eigenvalue Theorem establishes a rigorous and thorough regularization of the ill-posed problem in computing a defective eigenvalue so that the problem of computing a pseudo-eigenvalue enjoys unique existence and Lipschitz continuity of the solution that approximates the underlying defective eigenvalue with an error bound proportional to the data error, reaffirming the $m \times k$ condition number as a bona fide sensitivity measure of an eigenvalue whether it is defective or not. This regularization makes it possible to compute defective eigenvalues accurately using floating point arithmetic even if the matrix is perturbed, and we shall present such an algorithm in next section.

6 An algorithm for computing a defective eigenvalue

The Pseudo-Eigenvalue Theorem sets the foundation for accurate computation of a defective eigenvalue even if the matrix is represented with empirical data, provided that the multiplicity support can be obtained (more to that later in §8). The computation is under the assumptions that the given matrix A is the data representation of an underlying matrix possessing a defective eigenvalue and an initial estimate λ_0 is close to that eigenvalue. Assuming the multiplicity support $m \times k$ is known, identified or estimated, we also need to set up the matrix parameters $C \in \mathbb{C}^{n\times m}$ and $S \in \mathbb{C}^{k\times k}$, while using $T \in \mathbb{C}^{m\times k}$ in (1). By Lemma 3.2, the proper C is in an open dense subset of $\mathbb{C}^{n\times m}$ so that we can set C at random. With C available, we can then set up

$$\mathbf{x}_1^{(0)} = \begin{bmatrix} A-\lambda_0 I \\ C^{\mathsf{H}} \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{0} \\ T_{1:m,1} \end{bmatrix} \tag{19}$$

$$\mathbf{x}_{j+1}^{(0)} = \alpha_j \begin{bmatrix} A-\lambda_0 I \\ C^{\mathsf{H}} \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{x}_j^{(0)} \\ \mathbf{0} \end{bmatrix} \quad \text{for } j = 1, \ldots, k-1 \tag{20}$$

$$S = \left[\begin{array}{c|ccc} 0 & \alpha_1 & & \\ \vdots & & \ddots & \\ 0 & & & \alpha_{k-1} \\ \hline 0 & 0 & \cdots & 0 \end{array}\right] \tag{21}$$

where, for $j = 1, \ldots, k-1$, the scalar α_j scales $\mathbf{x}_{j+1}^{(0)}$ to a unit vector. Denote $X_0 = \left[\mathbf{x}_1^{(0)}, \cdots, \mathbf{x}_k^{(0)}\right]$. Then $\mathbf{g}(A, \lambda_0, X_0) \approx \mathbf{0}$ and we apply the Gauss-Newton iteration (13) that converges to (λ_*, X_*) assuming the initial estimate λ_0 is sufficiently close to λ_*. When the iteration stops at the j-th step, a QR decomposition of the matrix representing $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)$ is available and thus an estimate $\left\|\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)^{\dagger}\right\|_2$ of the $m \times k$ condition number can be computed by a couple of steps of inverse iteration [12] with a negligible cost. A pseudo-code of Algorithm PseudoEig is given in Fig. 2.

7 Taking advantage of the Jacobian structure

The main cost of Algorithm PseudoEig occurs at solving for the least squares solution $(\sigma, Y) \in \mathbb{C} \times \mathbb{C}^{n\times k}$ on the linear equation $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)(\sigma, Y) = \mathbf{g}(A, \lambda_j, X_j)$ where the partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)$ corresponds to an $(n\,k + m\,k) \times (n\,k + 1)$ matrix whose QR decomposition may be needed. This matrix is pleasantly structured with a proper arrangement so that the cost of QR decomposition can be reduced substantially.

Let $X = \left[\mathbf{x}_1, \cdots, \mathbf{x}_k\right]$, the image $\mathbf{g}(A, \lambda, X) \in \mathbb{C}^{n\times k} \times \mathbb{C}^{m\times k}$ can be arranged as

$$\mathbf{g}(A,\lambda,X) = \begin{bmatrix} C^{\mathsf{H}}\mathbf{x}_k & & & & & -T_{1:m,k} \\ (A-\lambda I)\,\mathbf{x}_k & -s_{k-1,k}\,\mathbf{x}_{k-1} & -s_{k-2,k}\,\mathbf{x}_{k-2} & -\cdots & -s_{1k}\,\mathbf{x}_1 & \\ & C^{\mathsf{H}}\mathbf{x}_{k-1} & & & & -T_{1:m,k-1} \\ & (A-\lambda I)\,\mathbf{x}_{k-1} & -s_{k-2,k-1}\,\mathbf{x}_{k-2} & -\cdots & -s_{1,k-1}\,\mathbf{x}_1 & \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ & & & & -s_{12}\,\mathbf{x}_1 & \vdots \\ & & & \ddots & C^{\mathsf{H}}\mathbf{x}_1 & -T_{1:m,1} \\ & & & & (A-\lambda I)\,\mathbf{x}_1 & \end{bmatrix}.$$

Algorithm Pseudoeig

Input: matrix A, initial eigenvalue estimate λ_0, multiplicity support m, k

- set C as a random $n \times m$ matrix and $\mathbf{x}_1^{(0)}$ as in (19)
- set $\mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_k^{(0)}$ by (20)
- set $X_0 = \left[\mathbf{x}_1^{(0)}, \cdots, \mathbf{x}_k^{(0)}\right]$, S as in (21) and $\mathbf{g}$ as in (2)
- for $j = 0, 1, \ldots$ do
  - * solve the linear system $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)(\sigma, Y) = \mathbf{g}(A, \lambda_j, X_j)$ for the least squares solution (σ, Y)
  - * set $\lambda_{j+1} = \lambda_j - \sigma$, $X_{j+1} = X_j - Y$.
  - * if $\left\|\mathbf{g}(A, \lambda_j, X_j)\right\|_2 < \left\|\mathbf{g}(A, \lambda_{j+1}, X_{j+1})\right\|_2$ then

    set $(\hat{\lambda}, \hat{X}) = (\lambda_j, X_j)$, break the loop. end if

  end do

Output: eigenvalue $\hat{\lambda}$, backward error bound $\left\|\mathbf{g}(A, \hat{\lambda}, \hat{X})\right\|_2 \left\|X^\dagger\right\|_2$, $m \times k$ condition number $\left\|\mathbf{g}_{\lambda X}(A, \hat{\lambda}, \hat{X})^\dagger\right\|_2$

Figure 2: Algorithm Pseudoeig

As a result, the partial Jacobian matrix in a blockwise upper-triangular form[2].

$$\frac{\partial \mathbf{g}(A, \lambda, X)}{\partial(\mathbf{x}_k, \ldots, \mathbf{x}_1, \lambda)} = \left[\begin{array}{cccccc} C^{\mathsf{H}} & O & O & \cdots & O & \mathbf{0} \\ A - \lambda I & -s_{k-1,k}\, I & -s_{k-2,k}\, I & \cdots & -s_{1k}\, I & -\mathbf{x}_k \\ & C^{\mathsf{H}} & O & \cdots & O & \mathbf{0} \\ & A - \lambda I & -s_{k-2,k-1}\, I & \cdots & -s_{1,k-1}\, I & -\mathbf{x}_{k-1} \\ & \ddots & \ddots & & \vdots & \vdots \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & C^{\mathsf{H}} & O & \mathbf{0} \\ & & & A - \lambda I & -s_{12}\, I & -\mathbf{x}_2 \\ & & & & C^{\mathsf{H}} & \mathbf{0} \\ & & & & A - \lambda I & -\mathbf{x}_1 \end{array}\right].$$

We can further assume the matrix A is already reduced to a Hessenberg form or even Schur form. Then

$$\left[\begin{array}{c} C^{\mathsf{H}} \\ A - \lambda I \end{array}\right] = \left[\begin{array}{cccc} * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \\ & * & \cdots & * \\ & & \ddots & \vdots \\ & & & * \end{array}\right]$$

is nearly upper-triangular with $m+1$ subdiagonal lines of nonzero entries. The QR decomposition of the partial Jacobian $\mathbf{g}_{\mathbf{x}_k \cdots \mathbf{x}_1 \lambda}(A, \lambda, X)$ can then be carried out by a sequence of standard textbook Householder transformations. It is also suitable to apply an iterative method for large sparse matrices particularly if A is sparse.

[2]Matlab code is available at `homepages.neiu.edu/~zzeng/pseudoeig.html`.

8 Identifying the multiplicity support

The geometric multiplicity can be identified with numerical rank-revealing. Let λ_0 be an initial estimate of $\lambda_* \in \mathit{eig}(A)$ in Lemma 3.2 and assume

$$|\lambda_0 - \lambda_*| \;<\; \theta \;<\; \min_{\lambda \in \mathit{eig}(A)\setminus\{\lambda_*\}} \big|\lambda - \lambda_0\big|.$$

The geometric multiplicity of λ_* can be computed as the numerical nullity of $A - \lambda_0 I$ within the error tolerance θ defined as

$$m \;=\; \max\big\{ j \;\big|\; \sigma_{n-j+1}(A - \lambda_0 I) < \theta \big\} \qquad (22)$$

where $\sigma_i(\cdot)$ is the i-th largest singular value of $(\cdot)$. A misidentification of the geometric multiplicity can be detected. Underestimating m results in an undersized C in (2) so that both $\left[\begin{smallmatrix} A - \lambda_* I \\ C^{\mathsf{H}} \end{smallmatrix}\right]$ and the partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_*, X_*)$ are rank-deficient. Overestimating m renders the system $\left[\begin{smallmatrix} A - \lambda_* I \\ C^{\mathsf{H}} \end{smallmatrix}\right] \mathbf{u} = \left[\begin{smallmatrix} \mathbf{0} \\ T_{1:m,1} \end{smallmatrix}\right]$ inconsistent with a large residual norm. During an iteration in which (λ_j, X_j) approaches (λ_*, X_*), a large condition number of the partial Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)$ indicates a likely underestimated geometric multiplicity and a large residual $\big\|\mathbf{g}(A, \lambda_j, X_j)\big\|_2$ suggests a possible overestimation.

If the geometric multiplicity is identified, it is possible to find the Segre anchor by a searching scheme based on the condition number of the Jacobian $\mathbf{g}_{\lambda X}(A, \lambda_j, X_j)$ as shown in the following example.

Example 1 Let

$$A \;=\; \begin{bmatrix}
0 & 4 & 0 & -4 & 0 & -2 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 2 & 1 & 0 & 0 & -1 & 0 & 0 \\
0 & 3 & 3 & -4 & 1 & 0 & 4 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & -1 & -1 \\
1 & -4 & 2 & 10 & -3 & 1 & -7 & -2 & -1 & 3 & 2 & 0 & 0 & -1 & 0 & 3 & -1 & 1 & 1 & 2 \\
-1 & -1 & 2 & 5 & -2 & -1 & -5 & -1 & -1 & 2 & 0 & -1 & 0 & 1 & 0 & 2 & -1 & -1 & -1 & 1 \\
-1 & -2 & 2 & 1 & 1 & 1 & -1 & 0 & -2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\
1 & 4 & 1 & -12 & 4 & 2 & 13 & 3 & 0 & -4 & 0 & 0 & -2 & -1 & 1 & -6 & 1 & 1 & 0 & -3 \\
-1 & -1 & 1 & 5 & -2 & 0 & -4 & -2 & 0 & 1 & -1 & 0 & 0 & 1 & 0 & 4 & 0 & -1 & -1 & 2 \\
1 & 2 & -4 & 0 & 1 & 0 & 1 & 1 & 4 & -2 & -1 & 1 & 0 & -1 & 0 & 1 & 2 & 1 & 0 & 1 \\
0 & -5 & 2 & 10 & -5 & -1 & -10 & -2 & -1 & 6 & 3 & -2 & 0 & 0 & 0 & 3 & -3 & 0 & 1 & 2 \\
1 & 1 & 1 & -1 & 2 & 2 & 4 & 0 & 1 & -1 & -1 & 2 & 0 & -1 & 0 & 0 & 2 & 1 & -1 & 0 \\
1 & -1 & 0 & 2 & 1 & 2 & 1 & 0 & 1 & -1 & 3 & 2 & 0 & -1 & 0 & 0 & 1 & 1 & -1 & 0 \\
-1 & -3 & 0 & 5 & -1 & 2 & -4 & -1 & 0 & 1 & -1 & 4 & 4 & 1 & -2 & 2 & 0 & -1 & 0 & 1 \\
-2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 3 & 2 & 0 & 0 & 0 & -1 & 0 & 0 \\
-3 & 4 & -1 & -4 & 0 & -2 & 1 & 0 & 0 & -1 & -1 & -1 & -2 & 5 & 2 & 0 & 0 & -1 & 1 & 0 \\
-2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 & 3 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & -2 & 3 & 1 & 2 & -1 & -1 & 2 & -2 & -2 & 2 & 0 & 0 & 0 & 5 & 2 & 0 & 0 & 1 \\
6 & 3 & -6 & 3 & 6 & 4 & 7 & 0 & 7 & -7 & 1 & 5 & -2 & -6 & 1 & 0 & 8 & 6 & -1 & 0 \\
0 & 2 & -4 & -4 & 1 & -1 & 4 & 1 & 0 & -1 & 0 & -1 & -1 & 0 & 1 & -2 & 0 & 3 & 4 & -1 \\
1 & -4 & -1 & 11 & -4 & 1 & -8 & -3 & -1 & 3 & 2 & 0 & 0 & -1 & 0 & 4 & -1 & 1 & 4 & 3 \\
0 & 0 & -1 & 1 & -2 & 0 & -1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4
\end{bmatrix}$$

with $\mathit{eig}(A) = \{2, 3\}$ of nonzero Segre characteristics $\{4, 3, 3\}$ and $\{5, 5\}$ respectively.

Applying the Francis QR algorithm implemented in Matlab yields computed eigenvalues scattered around $\lambda_1 = 2.0$ and $\lambda_2 = 3.0$:

$2.000118556521482 + 0.000118397929590i$	$3.000398490901253 + 0.001224915665189i$
$2.000118556521482 - 0.000118397929590i$	$3.000398490901253 - 0.001224915665189i$
$1.999881443477439 + 0.000118714860725i$	$3.000646066870935 + 0.000469627646058i$
$1.999881443477439 - 0.000118714860725i$	$3.000646066870935 - 0.000469627646058i$
$2.000013778528383 + 0.000018105742295i$	$3.001287762162967 + 0.000000000000000i$
$2.000013778528383 - 0.000018105742295i$	$2.999753002133234 + 0.000759191914332i$
$2.000008786021464 + 0.000020979720849i$	$2.999753002133234 - 0.000759191914332i$
$2.000008786021464 - 0.000020979720849i$	$2.998957628017279 + 0.000757681758834i$
$1.999977435451235 + 0.000002873978888i$	$2.998957628017279 - 0.000757681758834i$
$1.999977435451235 - 0.000002873978888i$	$2.999201861991639 + 0.000000000000000i$

Using two computed eigenvalues above, say

$$\tilde{\lambda}_0 = \texttt{1.999881443477439 - 0.000118714860725i} \quad \text{and} \quad \hat{\lambda}_0 = \texttt{3.001287762162967 + 0.000000000000000i}$$

as initial estimates of the defective eigenvalues, smallest singular values of $A - \tilde{\lambda}_0 I$ and $A - \hat{\lambda}_0 I$ can be computed using a rank-revealing method as

$\sigma_j(A - \tilde{\lambda}_0 I)$:	$\sigma_j(A - \hat{\lambda}_0 I)$:
...	...
0.084065699924186	0.070046163725993
0.049368630759014	0.054661269198836
0.000000000003635	0.036234932328447
0.000000000000280	**0.000000000000001**
0.000000000000001	**0.000000000000000**

indicating the geometric multiplicities 3 and 2 respectively.

Set the geometric multiplicities for the initial eigenvalue estimate $\tilde{\lambda}_0$ and $\hat{\lambda}_0$ as 3 and 2 respectively. Applying Algorithm PseudoEig with increasing input $k = 1, 2, \ldots,$ as estimated Segre anchors, we list the computed eigenvalues, $m \times k$ condition numbers and residual norms in Table 1. At λ_1, for instance, underestimated values $k = 1, 2$ render the $m \times k$ condition numbers as large as 10^8 and the residuals to be tiny, while the overestimated value $k = 4$ leads to a drastic increase of residual from 10^{-16} to 10^{-3} but maintains the moderate $m \times k$ condition number, as shown in Table 1. Similar effect of increasing estimated values of the Segre anchor at λ_2 can be observed consistently. □

test k value	at $\lambda_1 = 2$, Segre anchor $k = 3$		
	computed eigenvalue	condition number	residual norm
$k = 1$	1.999881443477439 - 0.000118714860725i	560995239.6	0.000000000000001
$k = 2$	1.999999993438010 - 0.000000011324234i	147603979.2	0.000000000000001
$\rightarrow$ $k = 3$	2.000000000000000 - 0.000000000000000i	58.7	0.0000000000000006 $\leftarrow$
$k = 4$	2.109885640097783 - 0.004348977611146i	24.1	0.007
test k value	**at $\lambda_2 = 3$, Segre anchor $k = 5$**		
	computed eigenvalue	condition number	residual norm
$k = 1$	3.001287762162967	2161090332264.6	0.000000000000003
$k = 2$	3.001287762162967	7962600062.8	0.0000000000005
$k = 3$	3.001287762162967	4556940.4	0.000000003
$k = 4$	3.000000013572103	687859583.9	0.0000000000000007
$\rightarrow$ $k = 5$	3.000000000000000	33.9	0.0000000000000007 $\leftarrow$
$k = 6$	3.002451613695432	34.1	0.007

Table 1: Effect of increasing estimated Segre anchors: Underestimated values yield large condition numbers of the Jacobian and overestimated values lead to large residual norms. The results using the correct anchors are pointed out with arrows.

Identify multiplicity support in practical computation can be challenging. It is certainly a subject that is worth further studies.

9 Improving accuracy with orthonormalization

Algorithm PseudoEig uses a simple nilpotent matrix S with only one superdiagonal line of nonzero entries. By Lemma 3.2 (iii), we can modify C and S as parameters of $\mathbf{g}$

so that the matrix component $\check{X}$ of the solution to $\mathbf{g}(A, \lambda_*, \check{X}) = \mathbf{0}$ has orthonormal columns. The orthonormalization can be carried out by the following process:

- Execute Algorithm PSEUDOEIG and obtain output $\hat{\lambda}, \hat{X}, C, S$.
- Normalize $\hat{X}_{1:n,1}$ and adjust s_{12} so that $(A - \hat{\lambda} I)\, \hat{X} \approx \hat{X}\, S$ still holds.
- Reset $C_{1:n,1}$ as $\hat{X}_{1:n,1}$.
- Reset $\hat{X}_{1:n,2:k}$ as $\hat{X}_{1:n,2:k} - \hat{X}_{1:n,1}\, (\hat{X}_{1:n,1})^{\mathsf{H}} \hat{X}_{1:n,2:k}$.
- Reset $S_{1,1:k}$ as $S_{1,1:k} + (\hat{X}_{1:n,1})^{\mathsf{H}} \hat{X}_{1:n,2:k}\, S_{2:k,1:k}$.
- Obtain the thin QR decomposition $\hat{X} = Q\, R$.
- Reset S as $R\, S\, R^{-1}$ in the mapping $\mathbf{g}$.
- Set the initial iterate $(\lambda_0, X_0) = (\hat{\lambda}, Q)$ for the Gauss-Newton iteration (13).

The advantage of such an orthonormalization is intuitively clear. When we solve for the least squares solution $(\tilde{\lambda}, \tilde{X})$ of the equation $\mathbf{g}(A, \lambda, X) = \mathbf{0}$ minimizing the magnitude of the residual $(A - \tilde{\lambda}\, I)\tilde{X} - \tilde{X}\, S = E$, the backward error given in Theorem 5.2 (iv) is $\|E\|_2 \, \big\|\tilde{X}^\dagger\big\|_2$. When the norm $\|\tilde{X}^\dagger\|_2$ is large, minimizing the residual norm $\|E\|_2$ may not achieve the highest attainable backward accuracy. If the columns of $\tilde{X}$ are orthonormal, however, the norm $\|\tilde{X}^\dagger\|_2 = 1$ and the least squares solution that minimizing the residual norm $\|E\|_2$ directly minimizes the backward error bound.

Example 2 Consider the matrix

$$A \;=\; \begin{bmatrix} 2 & 1 & & & \\ & -8 & 1 & & \\ & & 2 & 1 & \\ & & & 2 & 1 \\ & -10000 & 1000 & -100 & 12 \end{bmatrix} \tag{23}$$

with an exact eigenvalue $\lambda_* = 2$ and the multiplicity support 1×5. A straightforward application of Algorithm PSEUDOEIG in Matlab yields

$$\begin{aligned}
\tilde{\lambda} \;&=\; 1.999999999999748 \\
S \;&=\; \begin{bmatrix} 0 & 0.100686223197184 & 0 & 0 & 0 \\ 0 & 0 & 0.680272615629152 & 0 & 0 \\ 0 & 0 & 0 & 0.786924421181882 & 0 \\ 0 & 0 & 0 & 0 & 0.922632632948520 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
\tilde{X} \;&=\; \begin{bmatrix} 1.00502786434024 & 0.10210319200724 & .07627239342106 & .06542640851275 & .06584192219606 \\ -0.00000000000025 & 0.10119245986833 & .06945800549083 & .06002060904501 & .06036453955047 \\ -0.00000000000253 & 1.01192459868319 & .76341851426484 & .65486429121738 & .65902236805904 \\ -0.00000000000000 & -0.00000000000051 & .68838459356550 & .60075267245722 & .60419916522969 \\ -0.00000000000000 & -0.00000000000000 & -.00000000000052 & .54170664784191 & .55427401993992 \end{bmatrix}
\end{aligned}$$

The residual norm

$$\big\|(A - \tilde{\lambda} I)\, \tilde{X} - \tilde{X}\, S\big\|_F \;\approx\; 4.5 \times 10^{-14}$$

can not be minimized further with the unit round-off about 10^{-16} considering $\|A\|_2 \approx 10^4$. The backward error

$$\big\|(A - \tilde{\lambda} I)\, \tilde{X} - \tilde{X}\, S\big\|_F \, \big\|\tilde{X}^\dagger\big\|_2 \;\approx\; 1.3 \times 10^{-9}$$

is not small enough. After orthonormalization and resetting the resulting parameter C and S in $\mathbf{g}$ in (2), we apply the Gauss-Newton iteration again and obtain

$$\begin{aligned}
\hat{\lambda} &= 2.000000000000000 \\
S &= \left[\begin{array}{rrrrr} 0 & 0.09950371902 & -0.00990049999 & 0.00099000050 & -0.99498744208 \\ 0 & 0 & 1.00493781395 & -0.00098508732 & 0.99004950866 \\ 0 & 0 & 0 & 1.00004900870 & -0.09850873917 \\ 0 & 0 & 0 & 0 & 10050.38307728113 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \\
\hat{X} &= \left[\begin{array}{rrrrr} -1.0 & -0.000000000002531 & -0.000000000000000 & -0.000000000000000 & -0.000000000000000 \\ 0.0 & -0.099503719021067 & 0.009900499987341 & -0.000990000499944 & 0.994987442082474 \\ 0.0 & -0.995037190210673 & -0.000990049999192 & 0.000099000049994 & -0.099498744209908 \\ 0.0 & -0.000000000000000 & -0.999950498725976 & -0.000009900005712 & 0.009949874421338 \\ 0.0 & -0.000000000000000 & -0.000000000000000 & -0.999999505000536 & -0.000994987010945 \end{array}\right]
\end{aligned}$$

The residual practically stays about the same magnitude

$$\left\|(A - \hat{\lambda} I)\, \hat{X} - \hat{X}\, S\right\|_F \;\approx\; 1.25 \times 10^{-14}$$

but the backward error improves substantially to

$$\left\|(A - \hat{\lambda} I)\, \hat{X} - \hat{X}\, S\right\|_F \left\|\hat{X}^{\dagger}\right\|_2 \;\approx\; 1.25 \times 10^{-14}$$

as $\|\hat{X}^{\dagger}\|_2 \approx 1$. More importantly, the forward accuracy of the computed eigenvalue improves by 3 additional accurate digits. □

When the given matrix represents perturbed data, the orthonormalization seems to be more significant in improving the accuracy, as shown in the example below.

Example 3 Using a random perturbation of magnitude about 10^{-5}, let

$$\tilde{A} \;=\; A + 10^{-5} \left[\begin{array}{rrrrr} -0.092 & -0.653 & -0.201 & -0.416 & -0.787 \\ -0.135 & -0.218 & 0.054 & -0.136 & -0.255 \\ 0.651 & 0.663 & -0.166 & -0.969 & -0.603 \\ -0.833 & 0.607 & 0.314 & 0.969 & -0.020 \\ -0.733 & -0.879 & 0.256 & -0.665 & -0.321 \end{array}\right] \tag{24}$$

be the data representation of the matrix A in (23). Table 2 lists the computed eigenvalues, residual norms, backward errors and forward errors before and after orthonormalization. The results show a substantial improvement on the both forward and backward errors even though the residual magnitudes roughly stay the same. □

	before orthonormalization	after orthonormalization
computed eigenvalue	**2.00**4413315474177	**2.000000**343999377
residual norm	2.3×10^{-6}	2.9×10^{-6}
backward error	6.7×10^{-2}	2.9×10^{-6}
forward error	4.4×10^{-3}	3.4×10^{-7}

Table 2: Comparison between computing results with or without orthonormalization of the X component of the least squares solution to $\mathbf{g}(\tilde{A}, \lambda, X) = \mathbf{0}$ for the matrix $\tilde{A}$ in (24) at the eigenvalue $\lambda = 2$. Correct digits of computed eigenvalues are highlighted in boldface.

10 What kind of eigenvalues are ill-conditioned, and in what sense?

The well documented claim that a defective eigenvalue is infinitely sensitive to perturbations requires an oft-missing clarification: Its unbounded sensitivity is with respect to *arbitrary* perturbations. The sensitivity of a defective eigenvalue is finitely bounded by the spectral projector norm divided by the multiplicity if the perturbation is constrained to maintain the multiplicity, or by the multiplicity support condition number if the multiplicity support remains unchanged.

Furthermore, the above sensitivity assertions and clarifications are applicable on the problem of finding eigenvalues in its strictly narrow sense. In the sense of computing a multiple eigenvalue via a cluster mean provided that the cluster can be grouped correctly, the sensitivity is still bounded by spectral projector norm divided by the multiplicity. The problem of finding a defective eigenvalue in the sense of computing a pseudo-eigenvalue elaborated in this paper also enjoys a finitely bounded sensitivity in terms of the multiplicity support condition number.

Of course, the problem can still be ill-conditioned even if the sensitivity is bounded. In the following example, the matrix A has an eigenvalue of multiplicity 7 and the spectral projector norm is large, so the eigenvalue is ill-conditioned in this sense. On the other hand, the same eigenvalue is well-conditioned in multiplicity support sensitivity. Interestingly, this is not a contradiction at all. The conflicting sensitivity measures imply that the cluster mean is not accurate for approximating the eigenvalue but the pseudo-eigenvalue is, and Algorithm PSEDOEIG converges to the defective eigenvalue with all the digits correct.

Example 4 A simple eigenvalue $\lambda_1 = 2.001$ and a defective eigenvalue $\lambda_2 = 2$ with the Segre characteristic $\{5, 2, 0, \ldots\}$, i.e. multiplicity support 2×2, exist for

$$A \;=\; \begin{bmatrix} 3.006 & 2 & 1.005 & -1.001 & -0.002 & -0.001 & -0.001 & -1 \\ 5 & 2 & 5 & -1 & -2 & -1 & -1 & 0 \\ -5.006 & -3 & -3.005 & 2.001 & 3.002 & 2.001 & 0.001 & 2 \\ -6 & -1 & -6 & 3 & 5 & 3 & 0 & 1 \\ -5 & -1 & -5 & 1 & 6 & 3 & 0 & 1 \\ 1 & 0 & 1 & 0 & -1 & 1 & 0 & 0 \\ -4 & -2 & -4 & 1 & 3 & 2 & 2 & 2 \\ 5 & 0 & 5 & -1 & -2 & -1 & -1 & 2 \end{bmatrix}.$$

Let P_2 be the spectral projector associated with $\lambda_2 = 2$. The defective eigenvalue λ_2 is *both* highly ill-conditioned in spectral projector norm and almost perfectly conditioned measured by its 2×2 condition number with a sharp contrast:

$$\tfrac{1}{m}\|P_2\|_2 \;\approx\; 4.05 \times 10^{14} \quad \text{while} \quad \tau_{A,2\times 2}(\lambda_2) \leq 19.95.$$

This may seem to be a contradiction except it is not. Both conditions accurately measure the sensitivities of same end (finding the defective eigenvalue) through different means (cluster mean versas pseudo-eigenvalue). The Francis QR algorithm implemented in Matlab produces computed eigenvalues

```
2.003667055821394,        2.001912473859015 + 0.002992156370408i,
1.996674198110247,        2.001912473859015 - 0.002992156370408i,
2.000000046670435,        1.998416899175164 + 0.002994143122392i,
1.999999953329568,        1.998416899175164 - 0.002994143122392i.
```

There is no apparent way to group 7 computed eigenvalues to use the cluster mean for the defective eigenvalue even if we know the multiplicity is 7. Out of all 8 possible groups of 7 eigenvalues, the best approximation to $\lambda_2 = 2.0$ by the average is 2.000142850475652 with a substantial error 1.4×10^{-4} predicted by the spectral projector norm. In contrast, Algorithm PseudoEig accurately converges to $\lambda_2 = 2.0$ with an error below the unit round off 2.2×10^{-16} using the correct multiplicity support 2×2 that can easily be identified using the method in §8, as accurately predicted by the 2×2 condition number.

This seemingly contradicting sensitivities can be explained by the fact that there are infinitely many matrices nearby possessing a single eigenvalue of nonzero Segre characteristic $\{6, 2\}$ within 2-norm distances of 5.2×10^{-5}. Namely, such a small perturbation increases the multiplicity from 7 to 8 but can not increase the multiplicity support 2×2. Using the publicly available Matlab functionality NumericalJordanForm on the matrix A with error tolerance 10^{-5} in the software package NAClab[3] for numerical algebraic computation, we obtain approximately nearest matrix B with a single eigenvalue associated with Jordan blocks sizes 6 and 2 with first 14 digits of its entries given as

3.0059955942886	1.9999978851470	1.0049959180573	-1.0010020728471	-0.0020046893569	-0.0010002300301	-0.0010132897111	-0.9999977586058
4.9999998736434	1.9999937661529	5.0000001193777	-1.0000000065301	-2.0000000129428	-0.9999999934070	-0.9999999926252	-0.0000169379637
-5.0060008381014	-3.0000021146845	-3.0050076499797	2.0009979267360	3.0019953102172	2.0009997699688	0.0009867094117	2.0000022421372
-5.9999927405774	-1.0000021677309	-6.0000074892946	2.9999962015789	4.9999997701478	2.9999999999775	-0.0000002324249	0.9999978331627
-4.9999995930006	-0.9999961877596	-5.0000095377366	0.9999880178349	5.9999870716625	3.0000002295144	-0.0000150335883	0.9999994536709
0.9999971940837	-0.0000010574036	1.0000006545259	-0.0000047736356	-1.0000023807994	0.9999966612522	-0.0000036987224	-0.0000054161918
-4.0000092166543	-1.9999997569827	-3.9999765043908	1.0000142841290	3.0000142782479	2.0000000005386	2.0000249853630	2.0000002431865
4.9999998983338	0.0000026655939	5.0000001062663	-0.9999999958672	-1.9999999894366	-1.0000000065916	-1.0000000120790	2.0000133696815

The spectrum of B consists of a single eigenvalue $\lambda = 2.00125$. This lurking nearby matrix indicates that the multiplicity 7 of $\lambda_2 = 2.0 \in \mathit{eig}(A)$ can be increased to 8 with a small perturbation $\|A - B\|_2$, which is exactly the kind of cases where spectral projectors have large norms as elaborated by Kahan [10] and grouping method fails. However, those nearby defective matrices have the same multiplicity support 2×2, implying a small perturbation does not increase either the geometric multiplicity or the Segre anchor. As a result, the multiplicity support condition number is benign, and computing the defective eigenvalue via pseudo-eigenvalue is stable.

Interestingly, even though the matrix B is only known via the above empirical data, the spectral projector associated with its eigenvalue 2.00125 is known to be identity since there is only one distinct eigenvalue. Consequently, the mean of all approximate eigenvalues computed by Francis QR algorithm is 2.000124999999987 with 14 digits accuracy, same as the empirical data. Algorithm PseudoEig produces the 2×2 pseudo-eigenvalue 2.000125000000078 with the same number of correct digits due to a small 2×2 condition number about 14.47. The software NumericalJordanForm accurately produces the Jordan Canonical Forms of both matrices A and B. □

References

[1] J. V. Burke and M. L. Overton. Stable perturbations of nonsymmetric matrices. *Linear Algebra and Its Applications*, 171:249–273, 1992.

[3] http://homepages.neiu.edu/∼naclab

[2] F. Chaitin-Chatelin and V. Frayssé. *Lectures on Finite Precision Computations*. SIAM, Philadelphia, 1996.

[3] F. Chatelin. Ill conditioned eigenproblems. In J. Cullum and R. A. Willoughby, editors, *Large Scale Eigenvalue Problems*, North-Holland, Amsterdam, 1986. Elsevier Science Publishers B. V.

[4] J. W. Demmel. A Numerical Analyst's Jordan Canonical Form. Ph.D. Dissertation, Computer Science Department, University of California, 1983.

[5] J. W. Demmel. Computing stable eigendecompositions of matrices. *Lin. Alg. and Appl.*, 79:163–193, 1986.

[6] J. W. Demmel and A. Edelman. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. *Linear Algebra and its Applications*, 230:61–87, 1995.

[7] G. H. Golub and J. H. Wilkinson. Ill-conditioned eigensystems and the computation of the Jordan canonical form. *SIAM Review*, 18:578–619, 1976.

[8] B. Kågström and A. Ruhe. Algorithm 560: JNF, an algorithm for numerical computation of the Jordan Normal Form of a complex matrix. *ACM Trans. Math. Software*, 6:437–443, 1980.

[9] B. Kågström and A. Ruhe. An algorithm for numerical computation of the Jordan normal form of a complex matrix. *ACM Trans. Math. Software*, 6:398–419, 1980.

[10] W. Kahan. Conserving confluence curbs ill-condition. Technical Report 6, Computer Science, University of California, Berkeley, 1972.

[11] T. Kato. *Perturbation Theory for Linear Operators*. Springer, Berline, Heidelberg, New York, 1966 & 1980.

[12] T.-Y. Li and Z. Zeng. A rank-revealing method with updating, downdating and applications. *SIAM J. Matrix Anal. Appl.*, 26:918–946, 2005.

[13] V. Lidskii. Perturbation theory of non-conjugate operators. *U.S.S.R. Comput. Math. and Math. Phys.*, 6:73–85, 1966.

[14] R. A. Lippert and A. Edelman. The computation and sensitivity of double eigenvalues. In *Advances in computational mathematics, Lecture Notes in Pure and Appl. Math. 202*, pages 353–393, New York, 1999. Dekker.

[15] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.*, 18:793–817, 1997.

[16] A. Ruhe. An algorithm for numerical determination of the structure of a general matrix. *BIT*, 10:196–216, 1970.

[17] A. Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT*, 10:343–354, 1970.

[18] S. Rump. Computational error bounds for multiple or nearly multiple eigenvalues. *Linear Algebra and its Applications*, 324:209–226, 2001.

[19] S. Rump. Eigenvalues pseudospectrum and structured perturbations. *Linear Algebra and its Applications*, 413:567–593, 2006.

[20] B. Sridhar and D. Jordan. An algorithm for calculation of the Jordan Canonical Form of a matrix. *Comput. & Elect. Engng.*, 1:239–254, 1973.

[21] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra*. Princeton University Press, Princeton and Oxford, 2005.

[22] J. H. Wilkinson. Sensitivity of eigenvalues. *Utilitas Mathematica*, 25:5–76, 1984.

[23] J. H. Wilkinson. Sensitivity of eigenvalues, II. *Utilitas Mathematica*, 30:243–286, 1986.

[24] Z. Zeng. The approximate irreducible factorization of a univariate polynomial. Revisited. Proceedings of ISSAC '09, ACM Press, pp. 367–374, 2009.