

Computing beyond analyticity.

Matrix Algorithms in Inexact and Uncertain Computing

Françoise Chaitin-Chatelin *

CERFACS Technical Report TR/PA/03/110

Abstract

Given the matrices A and E in $\mathbb{C}^{n \times n}$, we consider, for the family $A(t) = A + tE$, $t \in \mathbb{C}$, questions such as i) existence and analyticity of $t \rightarrow R(t, z) = (A(t) - zI)^{-1}$, and ii) limit as $|t| \rightarrow \infty$ of $\sigma(A(t))$, the spectrum of $A(t)$. The answer depends on the Jordan structure of $0 \in \sigma(E)$, more precisely on the existence of trivial Jordan blocks (of size=1), and on Lidskii's perturbation theory.

Homotopic Deviation theory is then used to perform a comparative backward analysis for Inexact and Uncertain Computing.

Keywords : Sherman-Morrison formula, Jordan structure, Frobenius structure, frontier point, critical point, limit point, eigenprojection, analyticity, singularity, Schur complement, backward analysis, Inexact Computing, Uncertain Computing.

1 Introduction

Homotopic Deviation theory is the framework in which we study Inexact Computing. Given the matrices $A, E \in \mathbb{C}^{n \times n}$, we consider the family $A(t) = A + tE$, for $t \in \mathbb{C}$. We address such questions as :

- a) existence and analyticity of

$$t \in \mathbb{C} \rightarrow R(t, z) = (A(t) - zI)^{-1}, z \in \mathbb{C} ,$$

*Université Toulouse 1 and CERFACS, 42 avenue G. Coriolis 31057 Toulouse Cedex 1, France.
E-mail: chatelin@cerfacs.fr

- b) limit as $|t| \rightarrow \infty$ of $\sigma(A(t))$, the spectrum of $A(t)$, that is the set of singularities for $R(t, z)$.

Let $r = \text{rank } E$. The case $r = n$ is well known and offers no salient feature. The interesting case corresponds to $r < n$, and the Jordan structure of $0 \in \sigma(E)$ plays a major role. The case where $0 \in \sigma(E)$ is semi-simple has been studied in [4, 9, 10]. A particular case where E is nilpotent (hence defective) is treated in [5]. In this report, we treat the general Jordan structure for $0 \in \sigma(E)$ by means of Lidskii's perturbation theory [13].

2 The Jordan and Frobenius structures of $0 \in \sigma(E)$, $\text{rank } E = r < n$.

2.1 $E = UV^H$, with $U, V \in \mathbb{C}^{n \times r}$, $\text{rank } U = \text{rank } V = r$

Any matrix $E \in \mathbb{C}^{n \times n}$ of rank $r \leq n$ can be written under the form $E = UV^H$ where $U, V \in \mathbb{C}^{n \times r}$ of rank r represent bases for $\text{Im } E$, $\text{Im } E^H$ respectively. Such U and V can be computed from the Singular Value Decomposition of E , for example. We assume now on that $r < n$.

We define $K = \text{Ker } E$ of dimension $g = n - r$, $W = \text{Im } E$ of dimension r , and $G = V^H U \in \mathbb{C}^{r \times r}$.

Lemma 2.1 $\mathbb{C}^n = \text{Ker } E \oplus \text{Im } E \iff \text{rank } G = V^H U = r \iff 0 \in \sigma(E)$ is semi-simple.

Proof. Clear. □

When not only $\text{Ker } E \cap \text{Im } E = \{0\}$, but also $(\text{Ker } E)^\perp = \text{Im } E = \text{Im } E^H$, the direct sum of lemma 2.1 becomes orthogonal. And the bases U and V in $\text{Im } E$ satisfy $U = VB$, $B \in \mathbb{C}^{r \times r}$ of rank r . It is possible to choose $B = I_r$, then $E = UU^H$ is hermitian, positive definite.

2.2 $\text{Ker } E \cap \text{Im } E = S \neq \{0\}$

Notation associated with $\sigma(E)$.

m = algebraic multiplicity of $0 \in \sigma(E)$,

g = geometric multiplicity of $0 = \dim \text{Ker } E$, $g = n - r$

r' = rank G

\mathcal{M} = invariant subspace associated with 0 , $\dim \mathcal{M} = m$,

$\underline{\mathcal{M}}$ = invariant subspace associated with all nonzero eigenvalues,

$\dim \underline{\mathcal{M}} = n - m$

f = number of Jordan blocks of size ≥ 2 ,

K' = eigenspace generated by the $g' = g - f$ eigenvectors belonging to a (trivial) Jordan chain of length 1, $g' = \dim K'$,

S = subspace generated by the f eigenvectors *starting* a non trivial Jordan chain,

T = subspace generated by the f vectors *ending* a non trivial Jordan chain,

N = subspace of \mathcal{M} consisting of vectors in $\text{Im } E$ not in $\text{Ker } E$.

The overlap between $\text{Ker } E$ and $\text{Im } E$ introduces the following decomposition for \mathbb{C}^n :

Lemma 2.2 $\mathbb{C}^n = \mathcal{M} \oplus \underline{\mathcal{M}} = K' \oplus S \oplus N \oplus T \oplus \underline{\mathcal{M}}$ into five subspaces intersecting at $\{0\}$.

Proof. Clearly $K = \text{Ker } E = K' \oplus S$. And consider a Jordan chain of length $l > 1$ starting at $x_1 : Ex_1 = 0, Ex_2 = x_1, \dots, Ex_l = x_{l-1}$.

We see that $x_1 \in \text{Ker } E$ and $\text{Im } E, x_2$ to $x_{l-1} \in \text{Im } E$ and not to $\text{Ker } E$ and x_l belongs neither to $\text{Ker } E$ nor to $\text{Im } E$.

We conclude easily that $\mathcal{M} = K' \oplus S \oplus N \oplus T$.

Note that $S \oplus N \oplus \underline{\mathcal{M}}$ spans $\text{Im } E$, and that T , which is the complementary subspace in \mathbb{C}^n of $(\text{Ker } E \cup \text{Im } E)$, represents the f vectors non described by E . \square

Lemma 2.3 $\dim N = h = m - g - f$.

Proof. The decomposition $\text{Im } E = S \oplus N \oplus \underline{\mathcal{M}}$ implies that $f + h + n - m = r = n - g$. We remark that $f > 0 \iff m > g \iff r' < r, f < g \iff g' > 0$ and $n > m \iff r' > 0$. By construction, $h = 0$ if all non trivial Jordan blocks are of size 2. Observe that, when E is nilpotent, $n = m = g' + 2f + h$. \square

The decomposition $\mathbb{C}^n = \text{Ker } E \oplus (N \oplus T \oplus \underline{\mathcal{M}})$ confirms that $n = \dim \text{Ker } E + \dim \text{Im } E = g + (h + f + n - m) = g + (m - g - f) + f + n - m$.

2.3 The Frobenius structure of $0 \in \sigma(E)$ defective

There are two canonical normal forms for matrix polynomials in z of the form $E - zI, z \in \mathbb{C}$. The second one, the Jordan canonical form, is, by far, the better known of the two among Numerical Analysts. The first one is the Frobenius normal form which is a block diagonal form where each block is the companion matrix associated with the successive invariant polynomials for $E - zI$ of non zero degree [12,p. 262-265]. The Frobenius form (1878) extends to matrix polynomials on an arbitrary field the Smith form (1861) originally proved for matrices of integers. Unlike the Frobenius form, the Jordan form depends on the basis field, which is taken to be \mathbb{C} in Numerical Analysis.

For future reference, we introduce the following condition on $0 \in \sigma(E)$ of algebraic multiplicity m , which may or may not be satisfied.

$(\Sigma F) \iff$ the n^{th} invariant polynomial for $E(z) = E - zI$ is divisible by $z^{m-l}, 1 \leq l < g' < g$, and the l preceding invariant polynomials have each z in factor.

The condition (ΣF) means that l successive simplifications by z are possible in the minors of $E(z)$ of order $n-1, n-2, \dots, n-l$.

3 Existence of $R(t, z)$, for $z \in \text{re}(A)$

We define $M_z = -V^H(A - zI)^{-1}U \in \mathbb{C}^{r \times r}$, for $z \in \text{re}(A) = \mathbb{C} \setminus \sigma(A)$, the resolvent set for A .

By the Sherman-Morrison formula

$$R(t, z) = R(0, z)[I_n - tU(I_r - tM_z)^{-1}V^H R(0, z)] \quad (1)$$

exists for $t \neq \frac{1}{\mu_z}$, $\mu_z \in \sigma(M_z)$. We order by decreasing magnitude the r eigenvalues for M_z :

$$|\mu_{1z}| \geq |\mu_{2z}| \geq \dots \geq |\mu_{rz}| \geq 0$$

Proposition 3.1 For $1 \leq r < n$, z given in $\text{re}(A)$ such that $\text{rank } M_z = r$, the map $t \rightarrow R(t, z)$ has the following properties:

- i) it exists for any $t \neq \frac{1}{\mu_{iz}}$, $i = 1, \dots, r$,
- ii) it is analytic for $|t| < \frac{1}{|\mu_{1z}|} = \frac{1}{\rho(M_z)}$
- iii) $R(t, z)$ is analytic in $s = 1/t$ for $|t| > \frac{1}{|\mu_{rz}|} = \rho(M_z^{-1})$, and $\lim_{|t| \rightarrow \infty} R(t, z) = R(\infty, z) \neq 0$ is given by $R(\infty, z) = R(0, z)[I + UM_z^{-1}V^H R(0, z)]$.

Proof. Clear by (1). See [10]. When M_z is singular, $R(\infty, z)$ does not exist. \square

The identity (1) also shows that z in $\text{re}(A)$ is an eigenvalue for the r matrices $A(t_{iz})$, with $t_{iz} = \frac{1}{\mu_{iz}}$, $\mu_{iz} \in \sigma(M_z)$, $i = 1, \dots, r$.

The connection $t_z \mu_z = 1$ between t and z such that z is an eigenvalue for $A(t)$ shows that $|t_z| \rightarrow \infty$ as $|\mu_z| \rightarrow 0$. This is the subject of the next section.

4 $\lim_{|t| \rightarrow \infty} \sigma(A(t))$

This question was addressed in [4, 10] under the assumption (Σ) that $0 \in \sigma(E)$ is semi-simple. It was proved that g eigenvalues of $A(t)$ converge to $\sigma(\Pi)$, as $|t| \rightarrow \infty$, where $\Pi \in \mathbb{C}^{g \times g}$ is the matrix which represents $PAP_{\text{Ker } E}$, where P is the eigenprojection for E associated with $0 \in \sigma(E)$.

What happens when we relax (Σ) which says that $g' = g = m$? The condition (Σ) guarantees the continuity around $s = 0$ of the eigenprojection

$P(s)$ for $E(s) = E + sA$, associated with the g eigenvalues $\nu(s)$ converging to 0 in $O(s)$.

How much of this can be retained when $g' < g$?

Computationally, the answer is determined by the **asymptotic spectrum** of $A + tE$:

$$\sigma_\infty(A, E) = \lim_{|t| \rightarrow \infty} \sigma(A(t)) = \{\infty, \text{Lim}\}$$

which represents the possible limits for $\lambda(t) \in \sigma(A(t))$ as $|t| \rightarrow \infty$. Either $|\lambda(t)| \rightarrow \infty$ or $\lambda(t) \rightarrow z \in \text{Lim} \subset \mathbb{C}$. Lim consists of the limits as $|t| \rightarrow \infty$ of eigenvalues for $A + tE$ which are at finite distance in \mathbb{C} . We define $l_\star = \text{card Lim}$ (counting multiplicities). Clearly $l_\star \geq 0$ and $l_\star = 0$ when $\text{Lim} = \emptyset$.

4.1 Characterisation of Lim when $0 \in \sigma(E)$ is defective

The relation $A + tE = t(E + \frac{1}{t}) = \frac{1}{s}(E + sA)$, with $s = \frac{1}{t}$, shows that $A(t) = A + tE$ and $E(s) = E + sA$ share the same eigen/Jordan vector structure for $st = 1$, s and $t \neq 0$. Their eigenvalues are related by $\lambda(t) = \frac{\nu(s)}{s}$, $\nu(s) \in \sigma(E(s))$, and $\lim_{|t| \rightarrow \infty} \lambda(t) = \lim_{|s| \rightarrow 0} \frac{\nu(s)}{s}$. Given the Jordan structure for E , the differences $\nu_i(s) - \nu_i$, $\nu_i \in \sigma(E)$, are fractional powers of s , for s small enough : $\nu_i(s) - \nu_i = \alpha_i s^{1/l_i} + o(s^{1/l_i})$ with $1/l_i \leq 1$, $i = 1, \dots, n$.

The theory of V.B. Lidskii (1965), refined by H. Baumgärtel (1985) characterises the coefficients α_i , under specific non degeneracy assumptions. A clear presentation of this complex theory is given by Moro, Burke, Overton (1997) in matrix notation. The reader is referred to [13] for a complete account of Lidskii's theory.

An easy corollary of this theory is

Lemma 4.1 $g' = 0$, implies $l_\star = 0$ under the non degeneracy condition (\hat{G}) .

Proof. See[13]. The condition (\hat{G}) is stated below. □

The condition $g' \geq 1$ is not always sufficient to get $l_\star > 0$. Let us look at

Example 4.1 Let $E = e_1 e_n^T$: E has rank 1 and is defective, $E^2 = 0$ because $e_n^T e_1 = 0$. $\sigma(E) = \{0\}$ is decomposed into $(0^1)^{n-2}$ with $K' = \text{lin}(e_2, \dots, e_{n-1})$, $g' = n - 2$, and (0^2) with Jordan chain (e_1, e_n) .

Now we consider

$$A = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & & 1 & 0 \end{pmatrix}, \text{ a Jordan block of order } n. \text{ We define}$$

$$E(s) = E + sA = sA(t) \text{ with } s=1/t \text{ and } A(t) = \begin{pmatrix} 0 & & & & t \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & & 1 & 0 \end{pmatrix}$$

which is the companion matrix of the polynomial $p(z) = z^n - t = 0$. $A(t)$ has n simple eigenvalues $\lambda = t^{1/n}$ for $t \neq 0$. As $|t| \rightarrow \infty$, all eigenvalues escape to ∞ : $l_* = 0$. How is this reflected on eigenvectors?

The right eigenvector is $x(t) = (\lambda^{n-1}, \dots, 1)^T$, colinear with $(1, \dots, \frac{1}{\lambda^{n-1}})^T$, and the left eigenvector is $y(t) = (1, \bar{\lambda}, \dots, \bar{\lambda}^{n-1})^T$, colinear with $(\frac{1}{\lambda}, \dots, 1)^T$.

The normalised scalar product is of the order of $y(t)^H x(t) = \frac{n}{\lambda^{n-1}}$, and $y(t)^H A x(t) = \frac{n}{\lambda^n} = \frac{n}{t}$.

When $|t| \rightarrow \infty$, $y^H x \rightarrow 0$ for n fixed, and $x \rightarrow e_1$, $y \rightarrow e_n$: each eigenpair (x, y) converges to (e_1, e_n) , the Jordan chain for (0^2) . No eigenvector in K' is obtained in the limit $|t| \rightarrow \infty$. \triangle

This example illustrates the fact that when $g' < g$, one can have $0 = l_* < g'$, a possibility which is ruled out by (Σ) : $g' = g = l_*$.

We, now on, restrict our attention to the case of interest for Homotopic Deviation, that is $g' \geq 1$, or equivalently $f = g - g' < g$. And we look, for $\nu = 0$, for possible convergence of order ≥ 1 in s , that is $\nu(s) = \xi s + o(s)$, and $\xi \in \text{Lim}$.

The g Jordan blocks for $\nu = 0 \in \sigma(E)$ are ordered by non increasing size. Doing the same for non zero eigenvalues induces a complete Jordan basis X such that $E = X J X^{-1}$, therefore

$$E + sA = X(J + sX^{-1}AX)X^{-1} = X(J + sB)X^{-1}$$

where J is a Jordan form for E , and $X^{-1}AX = B$. We assume that $\nu = 0$ is placed first : $\bar{X} = [e_1, \dots, e_m]$ (resp. $\bar{Y}^T = \begin{bmatrix} e_1^T \\ \vdots \\ e_m^T \end{bmatrix}$) represents the right (resp. left) Jordan basis for 0 of algebraic multiplicity m .

In these bases, we select the *eigenvectors*. This defines the $n \times g$ matrix $\tilde{X} = [Z, X']$ (resp. $\tilde{Y} = [W, Y']$) such that Z, X' (resp. W, Y') are the eigenvectors starting the non trivial and trivial Jordan blocks for J (resp. J^T) respectively.

Therefore Z is a basis for S , W a basis for T and X' a basis for K' . It is easy to check that $W^T Z = 0_f$, $Y'^T X' = I_{g'}$, and $\tilde{Y}^T \tilde{X} = \begin{pmatrix} 0_f & 0 \\ 0 & I_{g'} \end{pmatrix}$.

Therefore $\tilde{P} = \tilde{X} \tilde{Y}^T$ is *not* a projection on $\text{Ker } E$. It satisfies $\tilde{P}^2 = P' = X' Y'^T$, the eigenprojection on K' . We define $\tilde{\Pi} = \tilde{Y}^T B \tilde{X} = \begin{pmatrix} \Gamma & R \\ L & \Pi' \end{pmatrix}$, with $\Gamma = W^T B Z$, $L = Y'^T B Z$, $R = W^T B X'$ and $\Pi' = Y'^T B X'$. Π' represents the Galerkin approximation $P' B P'$ restricted to K' , whereas $\tilde{\Pi}$ does not correspond to a Galerkin approximation (\tilde{P} is not a projection).

Assumption (G) : $\Gamma = W^T B Z$ has rank f

Proposition 4.2 Under (G) $l_* \geq g'$ and $\text{Lim} \supset \sigma(\Omega)$ with $\Omega = \Pi' - L \Gamma^{-1} R$.

Proof. See [13], for the particular case $j = q$, and $l_q = 1$ (notation in [13]). Ω is the *Schur complement* of Γ in $\tilde{\Pi}$. We have $l_\star \geq g'$. \square

The condition (G) on Γ guarantees that at least g' eigenvalues stay at finite distance. But it does not guarantee that $n - g'$ eigenvalues diverge to ∞ . The stronger result $\text{Lim} = \sigma(\Omega)$ requires a stronger assumption (\hat{G}) which can be described as follows.

Let $l = l_1 > \dots > l_q = 1$ be the strictly decreasing sequence of *different* sizes for the g Jordan blocks (assuming that $g' \geq 1$). For $j = 1, \dots, q$, there are r_j blocks of size l_j . We have $r_q = g'$ and $g = f + g' = \sum_{i=1}^q r_i$.

The matrix $\tilde{\Pi}$ results of an inductive process for $j = 1, \dots, q$ which builds Γ_j of order $f_j = \sum_{i=1}^j r_i$ with the eigenvectors starting the Jordan blocks of size l_1 to l_j . Γ_j is the 2×2 block matrix $\Gamma_j = \begin{pmatrix} \Gamma_{j-1} & R_j \\ L_j & \Delta_j \end{pmatrix}$. The process ends with $j = q$, $\Gamma_q = \begin{pmatrix} \Gamma_{q-1} & R_{q-1} \\ L_{q-1} & \Delta_{q-1} \end{pmatrix} = \tilde{\Pi} = \begin{pmatrix} \Gamma & R \\ L & \Pi' \end{pmatrix}$ of order g , where $\Gamma_{q-1} = \Gamma$ has order $f < g$.

Assumption (\hat{G}) : **The matrices Γ_j , $j = 1, \dots, q - 1$ are regular.**

This stronger assumption (which implies (G)) makes it sure that there is no interaction between Jordan blocks of different sizes [13]. Therefore exactly $n - g'$ eigenvalues diverge to ∞ , and exactly g' converge to $\text{Lim} : l_\star = g'$.

Note that (\hat{G}) reduces to (G) when $q = 2$.

When Γ is not invertible, there exist l_\star eigenvalues $\nu(s)$ converging to 0 in order ≥ 1 of s , where $l_\star \geq 0$ can be larger or smaller than g' .

For an example illustrating $l_\star = 0 < g' = n - 2$, we look back at Example 4.1. We reorder the Jordan basis for E as $(e_1, e_n, e_2, \dots, e_{n-1})$. Define the permutation matrix P_σ associated with $\sigma : \{1, 2, \dots, n\} \rightarrow \{1, n, 2, \dots, n - 1\}$.

Therefore $J = P_\sigma E P_\sigma = \text{diag} \left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, 0, \dots, 0 \right)$ and $B = P_\sigma A P_\sigma$ gives

$$B = \left(\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & & & 0 & 1 \\ 1 & 0 & & & & 0 \end{array} \right)$$

$b_{21} = e_2^T B e_1 = 0$. The condition (G) = (\hat{G}), with $f = 1$, is not satisfied. Indeed no eigenvalue $\lambda(t)$ remains at finite distance : $l_\star = 0$.

For an example illustrating $l_* > g'$, we turn to

Example 4.2 [10] *Let A, E be matrices of order 6 in diagonal form of 3 blocks of order 2.*

$$A = \text{diag} \left[\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ 0 & 4 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \right],$$

$$E = \text{diag} \left[\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right].$$

E has rank 2, $0 \in \sigma(E)$ is defective with $m = 5$, $g = 4$, $g' = 3$. Hence $g' < g < m < n$. $\sigma(A) = \{-1, 1, (2^2), 3, 4\}$ and $\text{Lim} = \{-1, (2^2), 2, 4\}$ yields $l_ = 5 > g' = 3$.*

If we reorder the basis we get $J = \text{diag} \left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, 0 \ 0 \ 0 \ 2 \right)$

and $B = \text{diag} \left(\begin{pmatrix} -1 & 1 \\ 0 & 4 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right)$. Condition $(G) = (\hat{G})$ is not satisfied because $b_{21} = 0$; this explains why $l_ > g'$.*

Consider the eigenvector $x(t) = \left(-1, \frac{1}{t} + \sqrt{1 + \frac{1}{t^2}} \right)^T$ associated with $\lambda(t) = t + 2 - \sqrt{t^2 + 1}$ which satisfies $\lambda(0) = 1$ and $\lambda(\infty) = 2$. As $|t| \rightarrow \infty$, it converges to $(-1, 1)^T \in K' \subset \text{Ker } E$. No other eigenvector associated with $\lambda(t) \in \sigma(A(t))$ converges to an eigenvector in K' . \triangle

Example 4.3 *When $0 \in \sigma(E)$ is semi simple, i.e. when (Σ) holds, Proposition 4.2 is valid with no condition ($f = 0$) and $\Omega = \Pi' \equiv \Pi$, the Galerkin approximation for A defined by the eigenprojection $P' \equiv P$ on $K' = \text{Ker } E$. See [4, 10]. \triangle*

We observe that $\Omega = \Pi'$ when R or $L = 0$. When E is nilpotent ($m = n$), an algorithmically important case corresponds to a *unique* non trivial Jordan block of size $l > 1$, hence $g' = g - 1 < g < n = g' + l$ and $f = 1$. Note that $(\hat{G}) = (G)$.

Proposition 4.3 *When $f = 1$ and $\gamma = e_l^T B e_1 \neq 0$, then $\text{Lim} = \sigma(\Omega)$ with $\Omega = \Pi' - \frac{1}{\gamma} u v^T$ with $u = (b_{1 \ l+1} \cdots b_{1n})^T$ and $v^T = (b_{l+1 \ l} \cdots b_{nl})$.*

Proof. Apply Proposition 4.2 with $f = 1$, $Z = [e_1]$, $W = [e_l]$ and $X' = Y' = [e_{l+1} \cdots e_n]$. $\gamma \neq 0$ implies that exactly 2 eigenvalues tend to ∞ . \square

An application in Numerical Software is the analysis of the Arnoldi method where E is nilpotent of rank 1 ($E^2 = 0$). [5]

4.2 Genericity of the deviation process (A, E) under (\hat{G})

The following

Proposition 4.4 *Let E be given in $\mathbb{C}^{n \times n}$ (resp. $\mathbb{R}^{n \times n}$). The set of matrices A such that (A, E) satisfies (\hat{G}) is a dense open subset in $\mathbb{C}^{n \times n}$ (resp. $\mathbb{R}^{n \times n}$).*

is proved in the PhD thesis of A. Ilahi, Cerfacs Rep. TH/PA/98/31 (1998). Proposition 4.4 leads naturally to the

Definition 4.1 *When $0 \in \sigma(E)$ is defective with at least one trivial Jordan block, the deviation process (A, E) is generic iff (\hat{G}) is satisfied.*

When (A, E) is generic, there are $g' \geq 1$ eigenvalues converging to $\sigma(\Omega)$ and $n - g'$ eigenvalues diverging to ∞ . We leave for a future report [11] the application of Liski's theory to the convergence of eigenvectors.

We define $F = Z\Gamma^{-1}W^T \in \mathbb{C}^{n \times n}$ under (G) .

Lemma 4.5 *F satisfies $F^2 = 0$ and $\text{rank } F = 1$*

Proof. $F^2 = Z\Gamma^{-1}(W^T Z)\Gamma^{-1}W^T = 0$ and $F \neq 0$. Therefore $\text{rank } F = 1$, and F is nilpotent. \square

Corollary 4.6 *There exist $u, v \in \mathbb{C}^n$ such that $v^T u = 0$, $v^T B u \neq 0$ and $F = \frac{1}{v^T B u} u v^T$.*

Proof. Use the characterization of a rank 1-matrix as $u v^T$, $u, v \in \mathbb{C}^n$. \square

Lemma 4.7 *$Q = FB$ is a rank 1-projection.*

Proof.

- 1) $Q = Z\Gamma^{-1}W^T B$ satisfies $Q^2 = Z\Gamma^{-1} [W^T B Z \Gamma^{-1}] W^T B = Q$. Since $Q = FB$, $1 \leq \text{rank } Q \leq \min(\text{rank } F, \text{rank } B)$.
- 2) $Q = \frac{1}{v^T B u} u v^T B$ by Corollary 4.6. \square

Lemma 4.8 *Ω represents the map $P'B(I - Q)P'$ restricted to $K' = \text{Im } X'$.*

Proof.: $P'B(I - Q)P' = X'(Y'^T B X' - Y'^T B Z \Gamma^{-1} W^T B X') Y'^T = X'(\Pi' - L \Gamma^{-1} R) Y'^T = X' \Omega Y'^T$. \square

Theorem 4.9 *When (Σ) does not hold, but (G) is valid, the matrix Ω replaces Π . This amounts to replace PAP by $P'B(I - Q)P'$, where $I - Q$ is a projection with rank $n - 1$.*

Proof. Clear by Lemma 4.7 and 4.8. □

We observe that when 0 is semi-simple, $Q = 0_n (f = 0)$.

When $f \geq 1$, the role of $I - Q$ is to express in B , in a prescribed fashion, the complexity introduced by the presence of f non trivial Jordan blocks.

The matrix B is modified into $B - BFB = B - C$ where C has rank 1. The rank 1-modification $C = BZ(W^T BZ)^{-1}W^T B$ entangles B with J into a highly *non linear* coupling. This high complexity disappears when $0 \in \sigma(E)$ is semi-simple.

4.3 Convergence of the mean $\hat{\lambda}(t) = \frac{1}{m} \sum_{i=1}^m \lambda_i(t)$

In the previous paragraphs 4.1 and 4.2, we have looked at the possible finite limit for *individual* eigenvalues $\lambda(t)$.

We now consider collectively the m eigenvalues $\lambda_i(t)$, $i = 1, \dots, m$ which are such that $\lambda_i(t) = \frac{\nu_i(s)}{s}$, and $\nu_i(s) \rightarrow 0$ as $s \rightarrow \infty$.

Let P represent the spectral projection associated with $0 \in \sigma(E)$ on the invariant subspace $M = \text{Im } P$ of dimension m . Let Γ be a closed Jordan curve around 0, isolating 0 from the other eigenvalues of E . Then $P(s) = \frac{1}{2i\pi} \int_{\Gamma} (E(s) - zI)^{-1} dz$ is defined for s small enough; it is the spectral projection associated with the m eigenvalues $\nu_i(s)$ converging to 0. And $\|P(s) - P\| \rightarrow 0$ as $s \rightarrow 0$ [1, 2, 3]. We define the arithmetic mean $\hat{\lambda}(t) = \frac{1}{m} \sum_1^m \lambda_i(t) = \frac{1}{m} \sum_1^m \frac{1}{s} \nu_i(s)$ and $\Pi = PAP|_M$.

Proposition 4.10 *There are m eigenvalues $\lambda_i(t)$, $i = 1, \dots, m$ such that $\hat{\lambda}(t) \rightarrow \frac{1}{m} \text{tr } \Pi$ as $|t| \rightarrow \infty$. Among these eigenvalues, $m - l_*$ escape to ∞ in such a way that their limit sum remains at finite distance in \mathbb{C} .*

Proof. This is a classical result. We can apply the theory of [2, chapter 4, Section 4.2] to $E(s) = E + sA$ for $s \rightarrow 0$. When $0 \in \sigma(E)$ is defective $PEP = D$ is a nilpotent matrix such that $D^m = 0$ by Theorem 1.7.1, [2], p.37. The matrix B in Theorem 4.2.2, [2], p.106, corresponds to $PEP|_M$ such that $\sigma(B) = \{0\}$. Under the perturbation sA on E , $\sigma(B)$ is transformed into $\sigma(B')$ where $\|B' - (B + s\Pi)\| = O(s^2)$ for s small enough. Now $\text{tr}(B' - B - s\Pi) = \text{tr } B' - s \text{tr } \Pi$, and $\sigma(B')$ consists of the m eigenvalues $\nu_i(s)$ of $E(s)$ converging to 0.

For any $M \in \mathbb{C}^{m \times m}$, $\frac{1}{m} |\text{tr } M| \leq \rho(M) \leq \|M\|$. Therefore

$$|\frac{1}{ms} \text{tr } B' - \frac{1}{m} \text{tr } \Pi| = O(s).$$

We conclude that $\frac{1}{m} \sum \frac{\nu_i(s)}{s} = \frac{1}{m} \sum \lambda_i(t) = \hat{\lambda}(t) \rightarrow \frac{1}{m} \text{tr } \Pi$.

If 0 is semi-simple, $D = 0$ and $\frac{1}{s}\sigma(B') \rightarrow \sigma(\Pi)$, that is $\text{Lim} = \sigma(\Pi)$ [4].

When 0 is defective, l_* eigenvalues $\lambda_j(t)$, $j = 1, \dots, l_*$ converge to Lim , with $0 \leq l_* \leq m$. Therefore $m - l_*$ escape to ∞ , with $m - l_* \geq 0$. They satisfy

$$\sum_{l_*+1}^m \lambda_i(t) \rightarrow \text{tr } \Pi - \lim_{|t| \rightarrow \infty} \sum_1^{l_*} \lambda_j(t).$$

There always exists a correlation between the m eigenvalues $\lambda_i(t)$: their sum converges to $\text{tr } \Pi$. \square

5 The kernel, limit and frontier points in $re(A)$

Observing the evolution $t \mapsto \lambda(t)$ leads to the distinction $\sigma(A) = \sigma^i \cup \sigma^e$ between *invariant* eigenvalues $\lambda \in \sigma^i$ ($\lambda(t) = \lambda$ for any $t \in \mathbb{C}$) and *evolving* eigenvalues $\lambda \in \sigma^e$ ($\lambda(t) \neq \lambda$ for almost all $t \in \mathbb{C}$).

Clearly, $\text{Lim} = \sigma^i \cup \text{Lim}^e$, where Lim^e is the set of limits of evolving eigenvalues at finite distance.

We introduce the following definitions for points in $re(A)$:

Definition 5.1 We call *limit points* the elements in $\Lambda(A, E) = \text{Lim} \cap re(A)$.

Definition 5.2 We call *kernel points* for (A, E) the values in $\sigma(\Omega)$ which are in the resolvent set $re(A)$. Points in $re(A)$ which are not kernel points are said to be *generic*. We denote $K(A, E)$ the set of kernel points, with $K(A, E) = \sigma(\Omega) \cap re(A)$.

We know that, under (G) (resp. (\hat{G})) $\text{Lim} \supset \sigma(\Omega)$ (resp. $\text{Lim} = \sigma(\Omega)$), therefore $K(A, E) \subset \Lambda(A, E)$ (resp. $K(A, E) = \Lambda(A, E)$). And, as we also know, kernel and limit points are points ξ in $re(A)$ where M_ξ is rank deficient. In general, if for $z \in re(A)$, M_z is singular, then $R(\infty, z)$ does not exist. Such a z signals the frontier of existence for $R(\infty, z)$.

Definition 5.3 We call *frontier points* the elements in $F(A, E) = \{z \in re(A); \text{rank } M_z < r\}$. A point z in $re(A)$ which is not in $F(A, E)$ is *generic*.

Definition 5.4 A point z in $F(A, E)$ is *critical* when $\rho(M_z) = 0$. A frontier point which is not critical is *regular*. The set of critical points is $C(A, E)$.

At a critical point, M_z is nilpotent ($M_z^\delta = 0$ with $M_z^{\delta-1} \neq 0$, $1 \leq \delta \leq r$). Therefore $(I_r - tM_z)^{-1} = \sum_{k=0}^{\delta-1} (tM_z)^k$, and $t \mapsto R(t, z)$ is a polynomial in t of degree δ at a critical point, $1 \leq \delta \leq r$.

When $r = 1$, the frontier points are critical and $R(t, z)$ is a polynomial of degree 1 in t for $z \in F(A, E) = C(A, E)$.

For $r > 1$, the existence of critical points is an open theoretical question which can be studied computationally by means of the visualization tools described in [7, 8, 4, 9, 10].

5.1 $F(A, E)$ and $C(A, E)$ are either finite or equal to $re(A)$

We first suppose that there exists at least one eigenvalue $\lambda \in \sigma(A)$ such that $\lim_{z \rightarrow \lambda} M_z$ does not exist. Such an eigenvalue is *normwise-observable* by the Definition 6.1.

Proposition 5.1 *If $F(A, E)$ or $C(A, E)$ are open sets in $re(A)$, they are identical to $re(A)$.*

Proof. The point z in $F(A, E)$ (resp. $C(A, E)$) are characterised as roots of algebraic equations in z of the form $\det M_z = 0$ (resp. $M_z^T = 0$). There are finitely many such roots, unless the equations are satisfied for all z in $re(A)$. For $F(A, E)$ (resp. $C(A, E)$) this yields $0 \in \sigma(M_z)$ (resp. $\rho(M_z) = 0$) for all z in $re(A)$. \square

The inclusion $C(A, E) \subseteq F(A, E)$ implies that when $C = re(A)$, the same is true for F . However, when C is discrete, F can be discrete or continuous when $r > 1$.

Example 5.1 *We consider the matrices A, E defined in Example 4.2. where (A, E) is non generic. For $z \in re(A)$, M_z is the 2×2 matrix defined by*

$$M_z = \begin{pmatrix} \frac{2z-4}{(1-z)(3-z)} & 0 \\ 0 & 0 \end{pmatrix}$$

M_z is of rank $1 < 2$. Therefore $F(A, E) = re(A)$. Observe that M_λ is defined for $\lambda = -1, 2$ and 4 , due to algebraic simplification. $F(A, E)$ can be extended to $\mathbb{C} - \{1, 3\}$ to contain the eigenvalues $\{-1, 2, 4\}$ which are normwise unobservable. Similarly $C(A, E) = \emptyset$ can be extended to be equal to $\{2\}$. \triangle

Example 5.2 *When $F(A, E)$ is discrete, it can be empty. Consider $A = \lambda I$, this yields $M_z = \frac{1}{z-\lambda} V^H U$ for $z \neq \lambda$. When $0 \in \sigma(E)$ is defective (resp. semi-simple), then M_z is singular (resp. regular) for any $z \neq \lambda$. Therefore $F(A, E) = re(A)$ (resp. \emptyset). \triangle*

Corollary 5.2 *The following statements hold for $C(A, E)$:*

1. $C(A, E) = re(A) \Leftrightarrow \sigma(A)$ is normwise unobservable

2. $C(A, E) = re(A) \Leftrightarrow \sigma(A(t)) = \sigma(A)$ for $t \in \mathbb{C}$
3. $C(A, E) = re(A) \Rightarrow E$ is nilpotent and $Lim = \sigma(A)$

Therefore, under (Σ) , $C(A, E)$ is a finite set.

Proof.

1. The first equivalence is clear by Proposition 5.1 : all eigenvalues of $M_z = 0$ for $z \in re(A)$. By continuity, M_λ exists and is nilpotent for $\lambda \in \sigma(A) : \rho(M_\lambda) = 0$.
2. Now, $\rho(M_z) = 0$ for all z in $re(A)$ implies that $R(t, z)$ is a polynomial in t , for all z in $re(A)$. Equivalently $\sigma(A(t)) = \sigma(A)$ for all $t \in \mathbb{C}$, that we assume to hold now for the reciprocal.
 $A(t) - zI$ is invertible for any $t \in \mathbb{C}$ and any $z \in re(A)$. Then, for any $\mu_z \in \sigma(F_z)$, $z \in re(A)$, $1 - t\mu_z \neq 0$ for any $t \in \mathbb{C}$.
The condition cannot hold for any $\mu_z \neq 0$, since $t = 1/\mu_z \in \mathbb{C}$. Therefore the condition holds iff $\sigma(F_z) = \{0\}$, that is $\sigma(M_z) = \{0\}$ for any $z \in re(A)$.
3. We have in particular $\rho(A + tE) = \rho(A) = |t|\rho(\frac{1}{t}A + E)$, which implies $\rho(E) = 0$ by the upper semicontinuity of $\rho(\cdot)$. Thus 3.
Under (Σ) , E cannot be nilpotent because 0 is semi-simple and $E \neq 0$. Therefore C cannot be continuous. \square

Proposition 5.3 Under (Σ) , then : $C(A, E) \subset \bigwedge(A, E) = K(A, E) = \sigma(\Pi) \cap re(A) \subset F(A, E)$, and $card F \leq (n - 1)r$.

Proof. Clear from [4,9,10], because $g' = g = m = l_*$. We recall that the geometric multiplicities at ξ in $\sigma(\Pi) \cap re(A)$ are preserved under (Σ) :

$$\dim \text{Ker} (\Pi - \xi I) = \dim \text{Ker} (M_\xi). \quad (2)$$

Now we show that, under (Σ) , $F(A, E)$ is necessarily discrete. $M_z = \frac{1}{\pi(z)}Q(z)$ where $\pi(z)$ is the characteristic polynomial of A and $Q(z) = V^H \text{adj}(zI_n - A)U$ is a matrix polynomial of order r , with leading matrix coefficient $V^H U$ for z^{n-1} . For z in $re(A)$, the values z for which at least one $\mu_z \in \sigma(M_z)$ is zero are the latent roots of $\det Q(z) = 0$. This is a scalar polynomial equation of degree $(n - 1)r$ under $(\Sigma) \Leftrightarrow V^H U$ has rank r . $\det Q = 0$ has at most $(n - 1)r$ roots in $re(A)$, which are the elements of $F(A, E)$. \square

5.2 $0 \in \sigma(E)$ is defective

When $0 \in \sigma(E)$ is defective, the 3 sets Λ , K , and F need not be identical, and they may be empty. We can illustrate the difference between Λ and F by the Example 5.1 As $|t| \rightarrow \infty$, $\text{Lim} = \{-1, (2^2), 2, 4\} \subset \sigma(A)$, consists of the invariant eigenvalues $\{-1, (2^2), 4\}$. The third copy of 2 is the limit of $\lambda(t)$ originating in $1 \in \sigma^e$. Therefore $\Lambda(A, E) = \emptyset$, whereas $F(A, E) = \text{re}(A)$, [10].

We illustrate now that one can still have $\Lambda(A, E) = F(A, E)$ when $g < m$.

Example 5.3 Let $A = J_n = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}$ be the Jordan block of order $n > 2$, $\sigma(J_n) = (0^n)$. We consider a family of n deviation matrices of the type $E_k = e_k e_n^T$, for $k = 1, \dots, n$. E_k has rank 1, it is defective ($E_k^2 = 0$) for $k < n$ with $g' = n - 2$, $\sigma(E_k) = \{(0^2), (0)^{n-2}\}$. For $k = n$, $E_n = e_n e_n^T$ is the orthogonal projection on e_n , $\sigma(E_n) = \{(0)^{n-1}, 1\}$: $g = n - 1$.

We define $A_k(t) = A + tE_k$, the eigenvalues $\lambda^{(k)}(t)$ are the n roots of the characteristic polynomial $q_k(t, z) = z^n - tz^{k-1} = z^{k-1}(z^{n-k+1} - t) = 0$.

The case $k = n$ is easy : $g = n - 1 = l_*$, $\text{Lim} = \sigma(\Pi')$ with $\Pi' = J_{n-1}$: $\text{Lim} = \{0 \text{ with algebraic multiplicity } n - 1\}$. And $0 \in \sigma(A)$: $n - 1$ eigenvalues stay invariant at the value 0, only $\lambda = t$ escapes at ∞ .

For $k < n$, (A, E_k) is non generic. $g' = n - 2$ and l_* varies with k : $l_*(k) = k - 1$ with $\text{Lim} = \{0\}$ for $1 < k < n$, and $\text{Lim} = \emptyset$ for $k = 1$ as shown in Example 4.1.

For $z \neq 0$, that is z in $\text{re}(A)$, M_z is the scalar $-z^{-(n-k+1)} \neq 0$ for all $z \neq 0$ in \mathbb{C} . M_z has rank 1, equal to r and $F(A, E_k) = \Lambda(A, E_k) = \emptyset$ for $1 \leq k \leq n$. \triangle

We show now that in the absence of (Σ) , $F(A, E)$ cannot be considered empty in finite precision.

Lemma 5.4 Let $0 \in \sigma(E)$ be defective. For any z in $\text{re}(A)$ such that $|z| > n\|A\| > \rho(A)$ with $n > 1$, the distance of M_z to a singular matrix of rank $r' < r$ is less than $\frac{\|U\|\|V\|}{\|A\|} \frac{1}{n(n-1)}$.

Proof. For $|z| > \rho(A)$, $(A - zI)^{-1} = -\sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}}$ is converging [2 page 55].

Therefore M_z can be written as $M_z = \frac{1}{z}G + \frac{1}{z}V^H \left(\sum_{k=1}^{\infty} \frac{A^k}{z^k} \right) U$, where $G = V^H U$ has rank $r' < r$.

We suppose that $|z| > n\|A\| > n\rho(A) > \rho(A)$ for $n > 1$, then $\|M_z - \frac{1}{z}G\| \leq \frac{\|U\|\|V\|}{\|A\|} \frac{1}{n(n-1)}$ because $\|A^k\| \leq \|A\|^k$ for $k \geq 1$. \square

This shows that for z far enough from 0, the matrix M_z is close to singularity. When $\frac{\|A\|}{|z|} < \frac{1}{n}$, then M_z is in a neighbourhood of size $O(\frac{1}{n^2})$ of the

singular matrix $\frac{1}{z}G$, where n can be chosen as one wishes.

It has been shown [15] that, when $t \mapsto A(t)$ is holomorphic, there are at most $2n-1$ distinct points z which cannot be eigenvalues of $A(t)$ for any $t \in \mathbb{C}$, where the equality is attainable, unless the map $t \mapsto \sigma(A(t))$ is invariant. The bound $2n-1$ has been reduced to $n-1$ when $A(t) = A + tE$ [14]. More generally, the following is true:

Theorem 5.5 *i) When the critical set is finite, then $\text{card } C \leq l_* \leq m \leq n$ and*

$$C(A, E) \subset \text{Lim} \cap \text{re}(A) = \bigwedge(A, E) \subset F(A, E)$$

with equalities when $r = 1$. When F is finite, $\text{card } F \leq (n-1)r$.

ii) When the critical set is continuous, $F(A, E) = C(A, E) = \text{re}(A)$, and $\text{Lim} = \sigma(A)$.

Proof. Part i) We have to show that $\bigwedge(A, E) \subset F(A, E)$: if $z = \lim_{|t| \rightarrow \infty} \lambda(t)$, $z \in \text{re}(A)$, $\lambda(t) = z' \in \text{re}(A)$ is an eigenvalue of $A + tE$ with $t\mu_{z'} = 1$. Therefore $|t| \rightarrow \infty$ implies $|\mu_{z'}| \rightarrow 0$ and $z \in F(A, E)$. $F(A, E)$ can be finite or continuous.

Under (Σ) , we know that $\text{card } \text{Lim} = n-r = g$, and under (\hat{G}) , $\text{card } \text{Lim} = g' < g = n-r$.

Now we show that $C(A, E) \subset \bigwedge(A, E)$. If $z \in \text{re}(A)$ such that $\rho(M_z) = 0$, by upper semicontinuity of $\rho(M_z)$ then $\forall \varepsilon, \exists \alpha : |z' - z| < \alpha \Rightarrow \rho(M_{z'}) < \varepsilon$.

By assumption, z is an isolated critical point, therefore $\rho(M_{z'}) > 0$ and $z' \in \text{re}(A)$ is an eigenvalue $\lambda(t)$ for $A + tE$ with $|t| > \frac{1}{\varepsilon}$. This shows that $\text{card } C \leq l_* \leq m \leq n$

The case $r = 1$ is straightforward: M_z is reduced to a scalar.

Part ii). By Corollary 5.2, $C(A, E)$ continuous is equivalent to $\sigma(A)$ invariant under t , hence $\text{Lim} = \sigma(A)$ and $l_* = n = m$. Thus $F(A, E) = C(A, E) = \text{re}(A) \neq \text{Lim} = \sigma(A)$. However $F = C$ can be extended to \mathbb{C} by continuity to include $\sigma(A)$. \square

6 Unobservable eigenvalues by (A, E)

In this Section, we consider the question (P) :

Can $\lambda \in \sigma(A)$ be an eigenvalue of $A + tE$ for $t \neq 0$?

This question (P) is the analogue for $\lambda \in \sigma(A)$ of the general question (Q) for $z \in \mathbb{C}$: in how many ways can $z \in \mathbb{C}$ be an eigenvalue of $A + tE$, $t \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$?

The answer is given by the correspondence $t\mu_z = 1$, for $\mu_z \in \sigma(M_z)$. For $z \in \text{re}(A)$, there are r such ways, defined by $t_i = \frac{1}{\mu_{iz}} \in \hat{\mathbb{C}}$, $i = 1, \dots, r$; $t_i \in \mathbb{C}$ iff $z \notin F(A, E)$.

For $\lambda \in \sigma(A)$, we distinguish whether λ in $\sigma(A)$ belongs to σ^i , σ^e or σ^f according to the following definitions :

- $\lambda \in \sigma^e$ (e for *evolving*) $\iff \lambda$ is an evolving value, that is $\lambda(t) \neq \lambda$ for almost all $t \in \mathbb{C}$,
- $\lambda \in \sigma^i$ (i for *invariant*) $\iff \lambda$ is an invariant value, that is $\lambda(t) = \lambda$ for all $t \in \mathbb{C}$,
- $\lambda \in \sigma^f$ (f for *final*) $\iff \lambda$ is a final value, that is $\lambda = \lim_{|t| \rightarrow \infty} \lambda(t)$, where $\lambda(t)$ originates in an evolving value $\lambda' \in \sigma^e$; $\lambda' \neq \lambda$ in general but $\lambda' = \lambda$ is possible.

Any multiple $\lambda \in \sigma(A)$ may belong to more than one of the sets σ^i , σ^e and σ^f which may overlap.

Any $\lambda \in \sigma^i$ belongs to $\sigma(A(t))$ for $t \in \mathbb{C}$, therefore the number of ways by which (P) is satisfied is noncountable. For $\lambda \in \sigma^e$, as we shall see in Proposition 6.2, $\lim_{z \rightarrow \lambda} M_z$ does not exist and the correspondence $t\mu_z = 1$ cannot be defined at $z = \lambda$ for all $\mu_z \in \sigma(M_z)$. $\lambda \in \sigma^e$ is not necessarily an eigenvalue of $A + tE$, $t \neq 0$.

The discussion above, leads to the study of the deviation E such that $\lim_{z \rightarrow \lambda} M_z$ exists, when $\lambda \in \sigma(A)$.

6.1 Three limits of observability for $\lambda \in \sigma(A)$

Using the Laurent expansion of $R(0, z)$ around $\lambda \in \sigma(A)$, we get for $z - \lambda$ small enough,

$$M_z = -V^H \left(\frac{-P_\lambda}{z - \lambda} - \sum_{k=1}^{a-1} \frac{D_\lambda^k}{(z - \lambda)^{k+1}} + \sum_{k=0}^{\infty} (z - \lambda)^k S_\lambda^{k+1} \right) U,$$

where P_λ , D_λ , S_λ denote respectively the spectral projection, nilpotent, reduced resolvent for A associated with λ , an eigenvalue of ascent (index) $a \geq 1$, [2], chapter 2, p. 59.

Under the conditions $V^H P_\lambda U = 0$, $V^H D_\lambda^k U = 0$, $k = 1, \dots, a - 1$ the matrix M_z is such that

$$\lim_{z \rightarrow \lambda} M_z = -V^H S_\lambda U = M_\lambda \text{ exists.}$$

We remark that, in this definition, λ is necessarily treated as a group, with its *algebraic* multiplicity m_λ (when multiple).

We introduce the :

Definition 6.1 $\lambda \in \sigma(A)$ is normwise unobservable (in short $\|\cdot\|$ -unobservable) by (A, E) iff $M_\lambda = \lim_{z \rightarrow \lambda} M_z$ exists.

This definition is justified by the fact that the map $z \mapsto \|M(z)\|$ has a peak at $z = \lambda$.

Proposition 6.1 A normwise-unobservable eigenvalue λ is seen by the deviation process (A, E) as a regular point at an homotopic distance to singularity equal to $1/\rho(M_\lambda)$.

Proof. The normwise observation is done by means of the map $z \mapsto \|M_z\|$. When $\lim_{z \rightarrow \lambda} M_z = M_\lambda$ happens to exist, λ cannot be seen as an eigenvalue (singularity) for the map $z \mapsto \|M_z\|$: it is seen as a regular point. If $\mu_{i\lambda}$, $i = 1, \dots, r$ are the eigenvalues of M_λ , then λ is an eigenvalue of the r matrices $A + t_i E$ with $t_i = \frac{1}{\mu_{i\lambda}} \in \hat{\mathbb{C}}$. The r values $t_i \in \mathbb{C}$ iff M_λ is regular. If M_λ has at least one multiple eigenvalue, there are less than r distinct t_i . \square

The notion of "normwise observability" was introduced in [10] under the generic name of "observability" based on the non existence of the map $z \mapsto M_z$ at $z = \lambda$.

The inequalities

$$|\mu_z| \leq \rho(M_z) \leq \|M_z\| \text{ for } z \in \text{re}(A)$$

indicate that other types of non observability are possible. For example, it is possible that λ is $\|\cdot\|$ -observable but $\lim_{z \rightarrow \lambda} \rho(M_z)$ exists. This can occur when M_z is non normal [1, 2]. Therefore we introduce the

Definition 6.2 $\lambda \in \sigma(A)$ is spectrally unobservable (in short σ -unobservable) by (A, E) iff $\sigma_\lambda = \lim_{z \rightarrow \lambda} \sigma(M_z)$ exists.

Proposition 6.2 $\lambda \in \sigma^e$ is necessarily spectrally-observable.

Proof. An evolving eigenvalue λ is such that $z \in \text{re}(A)$, arbitrarily close to λ , is an eigenvalue of $A + tE$ for t small. Because $t\mu_z = 1$, this implies that some $|\mu_z|$ can become arbitrarily large as $z \rightarrow \lambda$. Therefore $\lim_{z \rightarrow \lambda} \sigma(M_z)$ does not exist, hence $\|M_z\| \rightarrow \infty$: λ is observable, spectrally and normwise. \square

On the other hand, $\lambda \in \sigma^i$ can be observable or not, as we already know [10]. And when λ multiple belongs to σ^i and σ^e , then necessarily λ is σ -observable. See $\lambda = 1$ in Example 2.1 of [10].

If $u \in \text{Ker}(A - \lambda I) \cap \text{Ker} E$, then $Au = \lambda u$ and $(A + tE)u = \lambda u$ for any $t \neq 0$. Therefore such a λ belongs to σ^i , and M_λ exists or not.

In Example 5.1, $\sigma^i = \{-1, (2^2), 4\}$ and $\sigma^f = \{2\}$. The eigenvalue 2 (of algebraic multiplicity 2) in A is the limit of spectral rays originating in 1 :

therefore 2 is necessarily unobservable. Now -1 and 4 are $\|\cdot\|$ -unobservable, due to algebraic simplification in M_z . And $\sigma^e = \{1, 3\}$ is σ -observable.

Finally we consider a third type of nonobservability. Let the set $\sigma_{\min} \subset \sigma(M_z)$ consists of the eigenvalues of M_z with *minimum* modulus : $\mu_{rz} \in \sigma_{\min}$.

Definition 6.3 $\lambda \in \sigma(A)$ is *individually unobservable* by (A, E) iff $\lim_{z \rightarrow \lambda} \mu_z = \mu_\lambda$ exists for at least one $\mu_z \in \sigma_{\min}$.

One has the original chain : $\|\cdot\|$ -nonobservability $\implies \sigma$ -nonobservability \implies individual nonobservability.

Proposition 6.3 $\lambda \in \sigma^f$ is necessarily *individually nonobservable* $\mu_\lambda = 0$.

Proof. $\lambda \in \sigma^f$ is such that $\lambda = \lim_{|t| \rightarrow \infty} \lambda(t)$. Because of the relation $t\mu_z = 1$, it follows that when $|t| \rightarrow \infty$, $|\mu_{rz}| \rightarrow 0$ for μ_{rz} in σ_{\min} . Therefore $\mu_\lambda = 0$. \square

6.2 The question (Q)

Definition 6.4 The nodal set for $z \in \mathbb{C}$ relative to (A, E) is the set $\mathcal{N}_z(A, E) = \mathcal{N}_z = \{0 \neq t \in \hat{\mathbb{C}}, z \in \sigma(A + tE)\}$

The answer to the question (Q) can be easily formulated in terms of $\text{card } \mathcal{N}_z$, where here the point $t_i = \frac{1}{\mu_{iz}}$ in \mathcal{N}_z , whenever \mathcal{N}_z is finite, are counted according to the algebraic multiplicity of $\mu_{iz} \in \sigma(M_z)$.

The answer to (Q) is fourfold :

- 1) for $z \in \text{re}(A)$, $\text{card } \mathcal{N}_z = r$,
- 2) for $\lambda \in \sigma^e$, $0 \leq \text{card } \mathcal{N}_\lambda < r$,
- 3) for $\lambda \in \sigma^i$, $\text{card } \mathcal{N}_\lambda = \infty$ (with $\mathcal{N}_\lambda = \mathbb{C}$ or $\hat{\mathbb{C}}$).
- 4) for $\lambda \in \sigma^f$, $1 \leq \text{card } \mathcal{N}_\lambda \leq r$.

6.3 Bibliographical comments

The study of the maps $t \mapsto R(t, z)$ and $t \mapsto \sigma(A(t))$ that I have presented, with its strong emphasis on backward analysis, is original, to the best of my knowledge.

Related works include the papers by Simoncini and Alam-Bora cited in [10], as well as my book [1], where is defined the *radial convergence* for a sequence of closed linear operators in a Banach space. Alam and Bora [14], following Simoncini and my early works [1, 4, 7, 8], study the map $z \mapsto \rho(F_z)$ with $F_z = -E(A - zI)^{-1}$ of order n , in a completely different way (they do not use M_z of order $r \leq n$, which satisfies $\rho(M_z) = \rho(F_z)$). They use algebraic

geometry and independently arrive at the notion of spectral unobservability (under the form $\lim_{z \rightarrow \lambda} \rho(F_z)$ exists), which they call "stability" [14]. I find this term ambiguous, since it conflicts with the traditional notion of (numerical) stability for an evolving (σ -observable or "unstable" in the parlance of [14]) eigenvalue, which is a *forward* analysis notion : such an eigenvalue may or may not be numerically unstable, depending on its condition number [2, 3, 6].

This is the reason why I propose the term " spectral non observability" which, from my point of view, conveys satisfactorily the notion of a *backward* analysis (using M_z of order $r \leq n$ rather than $F_z = E(zI - A)^{-1}$ of order n).

Finally, we remark that, because the analysis of [14] is based on F_z and not on M_z , it cannot capture important aspects of the deviation theory, such as the analyticity of $R(t, z)$ around ∞ , and the associated frontier set $F(A, E)$ when the deviation E is singular.

7 Uncertain Computing

In the next three sections, we perform a comparative study of Inexact and Uncertain Computing. Because the latter is more familiar to Applied Numerical Analysts, we review this topic first.

7.1 What is Uncertain Computing ?

In a nutshell, Uncertain Computing is the art of Computing with *uncertain data*. It serves as a theoretical model for the two most frequent situations in Scientific Computing :

- i) in Experimental Sciences, there is uncertainty on the data, and a backward analysis is required to assess the validity of a computed result, usually using finite precision. However, assessment is necessary even in the ideal case where the result is computed exactly because of the uncertainty on the data.
- ii) in Computational Mathematics, when exact arithmetic is replaced by an arithmetic with finite precision, often the floating point arithmetic of the computer.

In the world of Uncertain Computing, the computational reality is fuzzy. One does not know A exactly, but a ball $\{A + \Delta A, \|\Delta A\| \leq \alpha\}$ centered at A , where α is the (absolute) level of uncertainty of the data A : two matrices A and B such that $\|A - B\| \leq \alpha$ are indiscernable. Zero and infinity have no meaning, no more than convergence ($x \rightarrow 0$).

The problem has been known to physicists and engineers from the beginnings of Computation, more than 4 millenia ago. However, the problem of

reliability became really pressing only in the second half of the 20th Century with the advent of the fast modern computers.

The original ideas of Turing and Von Neumann were systematically developed by Wilkinson for Numerical Linear Algebra and Software. We assume that the reader is familiar with the general normwise backward analysis for $Ax = b$ [6], and we proceed toward the eigenvalue problem.

7.2 The normwise distance of $A - zI$ from singularity, for $z \in \text{re}(A)$

Let z be given in $\text{re}(A)$: z is an eigenvalue of *infinitely many* matrices $A + \Delta A$ satisfying

$$\|\Delta A\| \geq \frac{1}{\|(A - zI)^{-1}\|}$$

Here, we have assumed that $\|\cdot\|$ is a subordinate norm of matrix. The quantity : $\delta_z = \frac{1}{\|(A - zI)^{-1}\|}$ represents the normwise distance of $A - zI$ to singularity [6].

The normwise backward analysis is a *metric* analysis where the parameter is the positive number $\alpha = \|\Delta A\| \in \mathbb{R}^+$.

For z given in $\text{re}(A)$, the positive number δ_z is a separator on the α -axis \mathbb{R}^+ into the two disjoint intervals $[0, \delta_z[$ and $[\delta_z, +\infty[$:

- a) z is a regular point for $A + \Delta A - zI$, with $\|\Delta A\| < \delta_z$,
- b) z is an eigenvalue for $A + \Delta A$, with $\|\Delta A\| \geq \delta_z$

δ_z , the normwise distance of $A - zI$ to singularity, represents therefore the normwise *backward error* for z as an eigenvalue of A : δ_z measures, in a backward way, the *smallest* $\|\Delta A\|$ such that $A + \Delta A - zI$ is singular.

Under the assumption that $\|\cdot\|$ is subordinate, the map $z \rightarrow \|(A - zI)^{-1}\|$ is subharmonic, with values $+\infty$ for $z \in \sigma(A)$, and local positive minima [7]. Its representation as a surface in 3D is the normwise *spectral portrait* of A [6].

The level sets are *pseudospectra* in $\text{re}(A)$:

$$\Lambda_\varepsilon = \{z; z \in \sigma(A + \Delta A), \|\Delta A\| \leq \varepsilon\} \cap \text{re}(A) = \{z \in \text{re}(A), \|(A - zI)\|^{-1} \geq \frac{1}{\varepsilon}\}$$

which are never empty for $\varepsilon > 0$, [6].

8 Inexact Computing

In Uncertain Computing, as we saw, one has only access to $\alpha = \|\Delta A\|$, the norm of the modification ΔA of the A . In contradistinction, in Inexact Computing, we assume that the deviation matrix E is known, and that the

modification of A takes the form $\Delta A = tE$, $t \in \mathbb{C}$. $A + tE$, $t \in \mathbb{C}$ is **inexact** data, whereas $A + \Delta A$, $\|\Delta A\| = \alpha \in \mathbb{R}^+$ is **uncertain** data.

We shall see later that the generalization from $\alpha = \|\Delta A\| \in \mathbb{R}^+$ to $\Delta A = tE$, $t \in \mathbb{C}$ implies, for the backward analysis, a significant evolution from *metric bounds* to *algebraic identities*.

8.1 What is Inexact Computing ?

Inexact Computing is the coupling of A and E by the complex parameter t into $A(t) = A + tE$, $t \in \mathbb{C}$.

It offers a computational approach for the study of the parameter dependence :

$$(t, z) \in \mathbb{C}^2 \rightarrow (A + tE - zI)^{-1} = R(t, z)$$

based on the factorization

$$A(t, z) = A + tE - zI = (I + tE(A - zI)^{-1})(A - zI)$$

We observe that in $A(t, z)$, the parameter t multiplies E (possibly rank deficient) whereas $-z$ multiplies I (non singular and semi-simple).

Homotopic perturbation theory with $t \in [0, 1]$, or $|t| \leq 1$ has been intensively used to relate A to $B = A + E$ by means of $A(t) = A + t(B - A)$, with $A(0) = A$ and $A(1) = B$ [1, 2, 3].

Such an approach uses "local" information, and may miss "global" effects in Computation which appear for $|t|$ large enough and in particular for $|t| = \infty$. This was the reason for developing the theory of Homotopic Deviation [4], where looking at what happens when $|t| \rightarrow \infty$ is an essential part of the analysis. This theory is the backbone of Inexact Computing.

8.2 The (homotopic) distances of $A - zI$ from singularity for $z \in re(A)$, z generic.

We suppose that $\text{rank } E = r \leq n$.

Let z be given in $re(A)$. z is an eigenvalue of r matrices $A(t_{iz}) = A + t_{iz}E$, with $t_{iz} = \frac{1}{\mu_{iz}}$, $\mu_{iz} \in \sigma(M_z)$, $i = 1, \dots, r$. There are exactly r distances τ_{iz} to singularity for $A(t) - zI$: $\tau_{iz} = |t_{iz}| = \frac{1}{\mu_{iz}}$, $i = 1, \dots, r$. These distances are *finite* when z is generic in $re(A)$ (for $r < n$) or for all z in $re(A)$ (for $r = n$) : that is when M_z is of rank r .

When $r \geq 2$, among the r distances τ_{iz} , two are special : the smallest distance τ_{1z} and the largest distance τ_{rz} . They signal two *limits of analyticity* for the resolvent $R(t, z)$ for z generic and fixed in $re(A)$,

- a) $t \rightarrow R(t, z)$ is analytic for $|t| < \tau_{1z}$,
- b) $R(t, z)$ is analytic in $s = 1/t$ for $|t| > \tau_{rz}$.

The homotopic *spectral* (resp. *frontier*, for $r \geq 2$) portrait is defined by the map :

$$z \in \text{re}(A) \mapsto \rho(M_z) = \frac{1}{\tau_{1z}} \quad (\text{resp. } \rho(M_z^{-1}) = \tau_{rz}, \text{ for } r \geq 2).$$

For $\varepsilon > 0$, the $\frac{1}{\varepsilon}$ level sets for $z \mapsto \rho(M_z)$ are

$$R_\varepsilon(A, E) = \{z \in \text{re}(A); \rho(M_z) \geq \frac{1}{\varepsilon}\}.$$

For $r \geq 2$, they are related to the sets :

$$S_\varepsilon(A, E) = \{z \in \sigma(A + tE), |t| \leq \varepsilon\} \cap \text{re}(A)$$

by the inclusion relation

$$R_\varepsilon(A, E) \supset S_\varepsilon(A, E)$$

which becomes an equality when the two sets are empty. For $r = 1$, the equality always holds, (even when the sets are not empty) [4,7].

The other finite distances, τ_{2z} to τ_{r-1z} correspond to matrices $A(t_{iz})$, $|t_{iz}| = \tau_{iz}$, for which z is an eigenvalue.

For a given z generic in $\text{re}(A)$, the plane $t \in \mathbb{C}$ is divided by r circles centered at z into $r + 1$ regions inside which $R(t, z)$ exists.

8.3 z is nongeneric in $\text{re}(A)$

At a *frontier* point z , M_z is rank deficient and there are at most $r - 1$ matrices at finite distance for which z is an eigenvalue. Moreover $R(\infty, z)$ does not exist. The t -plane is divided into r regions at most. If z is a frontier point such that M_z is nilpotent, z is *critical* and $\tau_{1z} = \infty$. Then $R(t, z)$ is a polynomial in t of degree δ (the degree of nilpotency of M_z).

8.4 The analyticity coefficient $an(z)$

For any z generic in $\text{re}(A)$, M_z has rank r and the product

$$an(z) = \rho(M_z)\rho(M_z^{-1}) = \frac{|\mu_{1z}|}{|\mu_{rz}|} \geq 1$$

is well defined. Observe that $an(z) = 1$ for $r = 1$ at any z in $\text{re}(A)$.

Definition 8.1 *The real number $an(z) = \rho(M_z)\rho(M_z^{-1})$ is the analyticity coefficient for (A, E) at $z \in \mathbb{C}$.*

We assume, as before, that $\|A\|$ represents a *subordinate* norm for A . We set $\text{cond}(A) = \|A\| \|A^{-1}\|$ for A regular.

Lemma 8.1 For $r = n$, $an(z)$ is defined on $re(A)$ and

$$1 \leq an(z) \leq cond(E)cond(A - zI)$$

Proof. When $r = n$, E is regular, as well as $F_z = E(A - zI)^{-1}$ for $z \in re(A)$, $F_z^{-1} = (A - zI)E^{-1}$. Therefore $\rho(F_z) \leq \|E\| \|(A - zI)^{-1}\|$ and $\rho(F_z^{-1}) \leq \|E^{-1}\| \|A - zI\|$. \square

Lemma 8.2 For $1 < r < n$, $an(z)$ is not defined at a regular frontier point. At a critical frontier point ξ , it is defined iff δ , the ascent of $0 \in \sigma(M_\xi)$, is a divisor of r . Moreover $an(\xi) = 1$ if $\delta = r$.

Proof. When z is a regular frontier point in $\Delta(A, E)$, $\mu_{rz} = 0$ but $\mu_{1z} \neq 0$.

When ξ is critical, $\mu_{1\xi} = \mu_{r\xi} = 0$, and $an(\xi)$ is of the form $\frac{0}{0}$.

We consider z generic approaching ξ such that $\|M_z - M_\xi\| = \varepsilon$, with $\varepsilon \rightarrow 0$ as $z \rightarrow \xi$. And we apply Proposition 4.3.8 of [2], p. 113. M_ξ is nilpotent ($M_\xi^\delta = 0$) and has \hat{g} Jordan blocks of size at most δ associated with $0 \in \sigma(M_\xi)$.

- 1) If the \hat{g} blocks are of equal size, $r = \hat{g}\delta$. There are r eigenvalues μ_{iz} converging to 0 at the same rate $O(\varepsilon^{1/\delta})$. And $an(\xi) \sim 1$.

When $\hat{g} = 1$ and $r = \delta$, the r eigenvalues μ_{iz} have *exactly* the same modulus. Therefore in the limit $z \rightarrow \xi$, $an(\xi) = 1$.

- 2) If the \hat{g} blocks are of unequal size, $\frac{\hat{g}}{r} > \frac{1}{\delta}$. The μ_{iz} converge to 0 at an unequal rate. The largest rate is $|\mu_{1z}| = O(\varepsilon^{1/\delta})$ and the smallest rate is $|\mu_{rz}| = \eta \leq O(\varepsilon^{\hat{g}/r})$.

Therefore $\frac{\varepsilon^{1/\delta}}{\eta} \geq \frac{\varepsilon^{1/\delta}}{\varepsilon^{\hat{g}/r}} \rightarrow \infty$ as $\varepsilon \rightarrow 0$. In this case, $an(\xi)$ is not defined. \square

The asymptotic behavior of $an(z)$ when $|z| \rightarrow \infty$ can be studied from the series expansion

$$(A - zI)^{-1} = - \sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}}$$

valid for $|z| > \rho(A)$ ([2] p. 55]).

Therefore $M_z = \sum_{k=0}^{\infty} \frac{V^H A^k U}{z^{k+1}} = \frac{G}{z} + \frac{V^H A U}{z^2} + \dots$ We define $w = \frac{1}{z}$ and

rewrite

$$M_{1/w} = O + wG + w^2 V^H A U + \dots$$

where O is the zero matrix of order r . As $w \rightarrow 0$, $\mu_{1/w} = 0 + w\tilde{g} + O(w^2)$, where \tilde{g} is an eigenvalue of G . The r eigenvalues in G are ordered by decreasing modulus.

We conclude that, when $|z| \rightarrow \infty$, $an(z)$ has a finite limit $\frac{|\tilde{g}_1|}{|\tilde{g}_r|} \geq 1$ when $r' = r$ (under (Σ)), and is not defined for $r' < r$.

We now perform a geometric study of each of the two factors $\rho(M_z)$ and $\rho(M_z^{-1})$ which define $an(z)$.

We know that the map $z \in \mathbb{C} \mapsto \rho(M_z)$ is subharmonic in $z \in re(A)$, with $\lim_{|z| \rightarrow \infty} |\mu_{1z}| = 0$ [4, 7, 8]. When $\text{card } F(A, E) < \infty$, the map $z \in re(A) \setminus F(A, E) \mapsto \rho(M_z^{-1})$ is also subharmonic, with the significant difference that now $\lim_{|z| \rightarrow \infty} |\mu_{rz}| = 0 \iff \lim_{|z| \rightarrow \infty} \rho(M_z^{-1}) = \infty$.

The first map defines a surface in \mathbb{R}^3 which is the *spectral portrait* of (A, E) , it has peaks at $+\infty$ at the σ -observable eigenvalues of A , and possibly zero minima at critical points.

The second surface in \mathbb{R}^3 is the *frontier portrait* of (A, E) : it has finitely many peaks at $+\infty$ at the regular frontier points and possibly at (some of) the critical ones.

Because the two surfaces have an opposite behavior at $|z| = \infty$, their intersection \mathcal{H} cannot be empty. The subharmonicity implies that this intersection consists of a finite number of closed curves.

Definition 8.2 *When $F(A, E)$ is discrete, we call H the set of closed curves defined by $H = \{z \in re(A), z \text{ regular such that } |\mu_{1z}| = |\mu_{rz}|^{-1}\}$. Equivalently $|\mu_{1z}||\mu_{rz}| = 1$ for z on H .*

One has the alternate characterization for the intersection $\mathcal{H} \in \mathbb{R}^3$ as $\{(z, \rho(M_z)) = (z, \rho(M_z^{-1})) \in \mathbb{R}^3, z \in H\}$.

For $r = 1$, $z \in H \iff |\mu_z| = 1$: $|\mu_z|$ is on the unit circle. The curve H is identical to the curve Γ introduced in [7].

For $r = 2$, $z \in H \iff |\det M_z| = |\mu_{1z}||\mu_{2z}| = 1$.

For $r \geq 3$, all we can say in general is that $an(z) = |\mu_{1z}|^2 = \frac{1}{|\mu_{rz}|^2}$ for $z \in H$.

For any z generic in $re(A)$, we know by Proposition 3.1 that the resolvent matrix $R(t, z)$ has two possible series expansions :

- (i) one around 0, for $|t| < \frac{1}{\rho(M_z)}$,
- (ii) a second one around ∞ , for $|s| < \frac{1}{\rho(M_z^{-1})}$, $s = \frac{1}{t}$.

For z on H , the radii of convergence for s and t are equal by definition. The curve H divides the complex plane into two regions :

- a) its *exterior* where $\rho(M_z) < \rho(M_z^{-1})$ and $\frac{1}{\rho(M_z)} > \frac{1}{\rho(M_z^{-1})}$: the convergence in t has a large radius of convergence (this region contains the frontier points $F(A, E)$),

b) its *interior* where $\rho(M_z) > \rho(M_z^{-1})$ and $\frac{1}{\rho(M_z)} < \frac{1}{\rho(M_z^{-1})}$: now the convergence in s has a larger radius (the region contains the σ -observable eigenvalues).

The curves H in \mathbb{C} and \mathcal{H} in \mathbb{R}^3 play the role of *equilibrium curves* between analyticity towards 0 or ∞ for the resolvent matrix.

We observe that for $r = 1$, the curve \mathcal{H} lies in the horizontal plane $\rho = 1$, whereas for $r > 1$, this curve is not planar in general ($\rho(M_z)$) is not constant for $z \in H$).

8.5 Analyticity of $R(t, z)$ for almost all $t \in \mathbb{C}$

Another set of interest for $an(z)$ consists of $An = \{z \in re(A); an(z) = 1\}$.

For $z \in An$, the gap in the analytic representation for $R(t, z)$ has Lebesgue measure zero in the complex plane, since $|\mu_{1z}| = |\mu_{rz}|$.

For $r = 1$, $An = re(A)$. The conditions on A and E under which $An \neq \emptyset$ for $r > 1$ are not known yet. Below we assume that $An \neq \emptyset$. We introduce the

Definition 8.3 *A matrix M is said to be circular iff its eigenvalues lie on a circle of radius $\rho(M) \geq 0$, centered at 0.*

For any $z \in An$, the corresponding matrix M_z is circular : all eigenvalues have equal modulus.

Examples.

- 1) Any unitary matrix is circular with radius = 1.
- 2) The Jordan matrix $J = \mu I + L$, where L (such that $l_{ij} = 0$ for $j \neq i - 1$, $l_{ii-1} = 1$) is nilpotent, is itself circular, with its spectrum reduced to the point μ on the circle of radius $|\mu|$.

When $\mu = 0$, $J = L$ is nilpotent and the circle is reduced to the origin point 0.

The basic QR algorithm may fail to converge on circular matrices, see Parlett [16] and [2, 3]. Convergence can be restored by appropriate shifts on the eigenvalues, see Wilkinson.

We have seen that M_z nilpotent $\iff \rho(M_z) = 0$ is a sufficient condition for the infinite series $(I_r - tM_z)^{-1} = \sum_0^{\infty} (tM_z)^k$, converging for $|t| < 1/\rho(M_z)$, to be reduced to the finite representation as a polynomial in t :

$$(I_r - tM_z)^{-1} = \sum_0^{\delta-1} (tM_z)^k, \quad M_z^\delta = 0$$

valid for any t in \mathbb{C} .

This is well known, and raises the obvious question : are there other cases of *finite* representation with $\rho(M_z) > 0$ which could be of computational interest for $R(t, z)$?

This is the subject of the next paragraph, assuming that $z \in A_n$, $A_n \neq \emptyset$.

8.6 Resolution nodes in A_n

Definition 8.4 *A point $z \in A_n$ such that the circular matrix M_z satisfies $\rho(M_z) \geq 0$ and $(I_r - tM_z)^{-1}$ has a finite representation for almost all t in \mathbb{C} is called a resolution node.*

Until further notice, we drop the subscript z . Let $M \in \mathbb{C}^{r \times r}$ be circular such that $|\mu_i| = \rho \geq 0$.

Lemma 8.3 *If $\sigma(M) = \{\mu\}$ for M circular, that is $M = \mu I_r + D$, $D^\delta = 0$ with $1 \leq \delta \leq r$, then for $t \neq \mu^{-1}$ and $\omega = \frac{t}{1 - t\mu}$*

$$(I_r - tM)^{-1} = \frac{1}{1 - t\mu} \sum_{k=0}^{\delta-1} (\omega D)^k \quad (3)$$

Proof. $I_r - tM = (1 - t\mu)I_r - tD = (1 - t\mu)[I_r - \omega D]$. The result follows. Note that $\mu = 0 \iff \omega = t$ covers the case where M is nilpotent. \square

Lemma 8.4 *Let Q be unitary with $\sigma(Q) = \{e^{i\gamma_j}\}$, $j = 1, \dots, r$ and $0 \leq \varphi_j = \frac{\gamma_j}{2\pi} < 1$. If we suppose that $\varphi_j = \frac{R'_j}{R_j}$ are rational in $[0, 1[$ with $R_j \in \mathbb{N}^*$, $R'_j \in \mathbb{N}$, then the sequence I, Q, \dots, Q^{R-1} is R -cyclic : $Q^R = I$, where R is the smallest common multiple for the r denominators R_j , $j = 1, \dots, r$.*

Proof. This follows easily from the Euler formula $e^{2ik\pi} = 1$, $k \in \mathbb{Z}$. Let $\varphi_1 = \frac{\gamma_1}{2\pi} = \frac{R'_1}{R_1}$, where $R'_1, R_1 \neq 0 \in \mathbb{N}$ are mutually prime. Then $q_1 = e^{i\gamma_1}$ is such that $R_1\gamma_1 = 2\pi R'_1$ and $q_1^{R_1} = e^{i2\pi R'_1} = 1$. Clearly $R \geq 2$ for $Q \neq I$. \square

Because $Q^R = I$, the eigenvalues q of Q satisfy $q^R = 1$. Therefore $\sigma(Q)$ consists of r numbers chosen among the R roots of 1 of order R .

Lemma 8.5 *We set $M = \rho Q$, $\rho > 0$, where Q satisfies the assumption of lemma 8.4. Then for $|t| < 1/\rho$, $\alpha = t\rho$, $\pi = \alpha^R$ with $R \in \mathbb{N}$, $R \geq 2$,*

$$(I_r - tM)^{-1} = \frac{1}{1 - \pi} (I + \alpha Q + \dots + \alpha^{R-1} Q^{R-1}) \quad (4)$$

Proof. $I - tM = I - t\rho Q$ with $\alpha = t\rho$. By assumption, the powers of Q form an R -cyclic sequence. For $|\alpha| < 1$, $(I - \alpha Q)^{-1} = \sum_0^\infty (\alpha Q)^k$ is converging since $\|\alpha Q\|_2 = |\alpha| < 1$. Moreover $\pi = \alpha^R$ is such that $|\pi| < |\alpha| < 1$ and $\sum_0^\infty \pi^k = \frac{1}{1 - \pi}$. \square

Because $M^{-1} = \frac{1}{\rho}Q^H$ exists for $\rho > 0$, there is an alternative finite representation in terms of Q^H .

$I - t\rho Q = -t\rho Q(I - \frac{1}{t\rho}Q^H)$ and $(I - tM)^{-1} = -\frac{1}{t\rho}Q^H(I - \frac{1}{t\rho}Q^H)^{-1}$ formally.

Lemma 8.6 *Under the assumptions of lemma 8.5 on M , for $|t| > 1/\rho$, $\beta = 1/t\rho$,*

$$(I_r - tM)^{-1} = \frac{1}{1 - \pi} \left[\frac{1}{\beta^{R-1}}Q^H + \dots + \frac{1}{\beta}(Q^H)^{R-1} + I \right] \quad (5)$$

Proof. We set $\beta = \frac{1}{\alpha} = \frac{1}{t\rho}$. For $|\beta| < 1$, $(I - tM)^{-1} = -\beta Q^H \sum_0^\infty (\beta Q^H)^k$ is converging. $\sigma(Q^H) = \{e^{-i\gamma_j}\}$, and the sequence of powers of Q^H is equally R -cyclic. $\beta^R = \frac{1}{\pi}$ and $|\pi| > 1$ yields $\sum_0^\infty (\beta^R)^k = \frac{1}{1 - 1/\pi} = \frac{\pi}{\pi - 1}$; and $(I - tM)^{-1} = \frac{\pi\beta Q^H}{1 - \pi} [I + \beta Q^H + \dots + \beta^{R-1}(Q^H)^{R-1}]$. (5) follows. \square

Definition 8.5 *An eigenvalue $\mu = |\mu|e^{i\gamma}$ of M such that the ratio $\varphi = \frac{\gamma}{2\pi}$ is rational is said to have a rational spectral phase φ in \mathbb{Q} .*

Any real (resp. pure imaginary) eigenvalue has a rational phase :

φ	1	1/2	1/4	3/4
μ	$ \mu $	$- \mu $	$i \mu $	$-i \mu $

Results from Lemma 8.3 to Lemma 8.6 are gathered (with the subscript z) into the

Theorem 8.7 *$z \in An$ is a resolution node if M_z falls into any of the two categories :*

- i) $\sigma(M_z) = \{\mu_z\}$ with $|\mu_z| \geq 0$, there existe δ , $1 \leq \delta \leq r$, such that $(M_z - \mu_z I)^\delta = 0$,
- ii) $\rho_z = |\mu_z| > 0$ and $M_z = \rho_z Q_z$ where Q_z is a unitary matrix with only rational spectral phases.

In case i) (resp. ii)) $(I - tM_z)^{-1}$ has a unique finite representation in $\omega_z = \frac{t}{1 - t\mu_z}$, for any $t \neq \mu_z^{-1}$ (resp. two distinct finite representations in t for $|t| < \frac{1}{\rho_z}$, and in $1/t$ for $|t| > 1/\rho_z$).

We have left aside two important questions :

- i) is $An \neq \emptyset$ for $r > 1$?
- ii) if $An \neq \emptyset$, do resolution nodes exist ?

The answer to ii) is easy for $r = 1$ ($A_n = \text{re}(A)$).

The matrix M_z reduces to the scalar μ_z . And $(1 - t\mu_z)^{-1} = \frac{1}{1 - t\mu_z}$ for any $t \neq \frac{1}{\mu_z}$, $\mu_z \neq 0$ and $= 1$ for $\mu_z = 0$, t arbitrary. This corresponds to (3) in lemma 8.3.

However, for $\mu_z \neq 0$, one can equally well consider $e^{i\gamma_z} = \frac{\mu_z}{|\mu_z|}$. When z is such that μ_z has a rational spectral phase, one has two alternative representations of type (4) (resp. (5)) which are rational fractions in t , for $|t| < \frac{1}{|\mu_z|}$ (resp. in $1/t$, $|t| > |\mu_z|$).

We leave the study of the two questions i) and ii) in the case $r > 1$ for future work.

So far, we have performed a backward analysis by looking at the correspondence $z \in \text{re}(A) \rightarrow t_z$ such that z is an eigenvalue of $A(t_z) = A + t_z E$. By comparison with the normwise approach of Section 7, the novelty is that, for z given in $\text{re}(A)$, there is usually a finite number of such t_z given by $\sigma(M_z^{-1})$. Moreover, because of the structure $\Delta A = tE$ where E is fixed, computation with t (or $1/t$) can be performed, using M_z (or M_z^{-1}). As a result, a *forward analysis* $t \mapsto \lambda(t)$ is possible by computation. This, of course, is **impossible** when one knows $\alpha = \|\Delta A\|$ only.

8.7 A forward spectral analysis $t \mapsto \lambda(t)$

We set $t = \tau e^{i\theta}$, with $\tau = |t| \in \mathbb{R}^+$ and $\theta \in [0, 2\pi[$. The matrix family $t \mapsto A(t) = A + tE$ can be divided into

- i) the matrix orbits $\mathcal{A}(\tau) = \{A + \tau e^{i\theta} E, \theta \in [0, 2\pi[\}$ where $\|tE\| = \tau\|E\|$ is fixed,
- ii) the matrix rays $\mathcal{A}(\theta) = \{A + \tau e^{i\theta} E, \tau \in \mathbb{R}^+ \}$ where θ is fixed.

This decomposition induces a similar one for the eigenvalue map $t \mapsto \lambda(t)$. One defines

- i) the spectral orbits $\Sigma(\tau) = \{\lambda(\tau e^{i\theta}), \theta \in [0, 2\pi[\}$,
- ii) the spectral rays $\Lambda(\theta) = \{\lambda(\tau e^{i\theta}), \tau \in \mathbb{R}^+ \}$.

The fate of $\lambda(t)$ as $\tau \rightarrow \infty$ is one of the following two :

- 1) $\lambda(t)$ escape to $\infty \iff |\lambda(t)| \rightarrow +\infty$,
- 2) $\lambda(t)$ has a limit at finite distance in Lim when $\text{Lim} \neq \emptyset$.

9 Normwise backward analysis for Inexact Computing

Let be given the family $A(t) = A + tE$. We compare the complete forward/backward analysis which can be performed, to the normwise backward (only) analysis that we get from using $\|\Delta A\| = \tau\|E\|$, and pretending to ignore that "in reality" $\Delta A = tE$. All the information involving the argument θ for $t = \tau e^{i\theta}$ is lost. The norm filter put on the matrix data filters out the difference between E and $e^{i\theta}E$ for $\theta \neq 0 \pmod{2\pi}$.

The normwise analysis says, at the global level of A , that $A + \Delta A - zI = (I + \Delta A(A - zI)^{-1})(A - zI)$ for $z \in \text{re}(A)$ is regular for $\|\Delta A\| < \frac{1}{\|(A - zI)^{-1}\|}$.

The wealth of results stemming from the spectral analysis of the $r \times r$ matrix $M_z = -V^H(A - zI)^{-1}U$ associated with $E = UV^H$ is concealed. Therefore the conclusions of the normwise analysis are deceptively poor when compared with the homotopic analysis. For example, the popular normwise spectral portrait of A , that is the map $z \mapsto \frac{1}{\|(A - zI)^{-1}\|}$, has nothing to say about critical points. The corresponding pseudospectrum remains unstructured.

This is a spectacular, if underestimated, consequence of considering the *norm* $\|\Delta A(A - zI)^{-1}\| \leq \|\Delta A\| \|(A - zI)^{-1}\|$ rather than the *spectral radius* $\rho(E(A - zI)^{-1}) = \rho(V^H(A - zI)^{-1}U)$. Considering the coupling matrix M_z is all the more appealing that its order is r which can be significantly smaller than n (as small as $r = 1$). We shall develop elsewhere the connection between M_z and the transfer function in Linear Systems Theory.

References

- [1] F. Chatelin. **Spectral approximation of linear operators**. Academic Press, New York, 1983.
- [2] F. Chatelin. **Valeurs propres de matrices**. Masson, Paris, 1988.
- [3] F. Chatelin. **Eigenvalues of matrices**. Wiley, Chichester, 1993. Enlarged Translation of the French Publication with Masson.

- [4] F. Chaitin-Chatelin. About Singularities in Inexact Computing. Technical Report TR/PA/02/106, CERFACS, Toulouse, France, 2002.
- [5] F. Chaitin-Chatelin. The Arnoldi method in the light of Homotopic Deviation theory. Technical Report TR/PA/03/15, CERFACS, Toulouse, France, 2003.
- [6] F. Chaitin-Chatelin and V. Frayssé. **Lectures on Finite Precision Computation.** SIAM Publ., Philadelphia, 1996.
- [7] F. Chaitin-Chatelin and E. Traviésas. Homotopic perturbation - Unfolding the field of singularities of a matrix by a complex parameter: a global geometric approach. Technical Report TR/PA/01/84, CERFACS, 2001.
- [8] F. Chaitin-Chatelin and E. Traviésas. Qualitative Computing. Technical Report TR/PA/02/58, CERFACS, Toulouse, France, 2002. To appear in **Handbook of Computation**, B. Einarsson ed. , SIAM Philadelphia.
- [9] F. Chaitin-Chatelin and M.B. van Gijzen. Homotopic Deviation: an example in acoustics. Technical Report TR/PA/03/04, CERFACS, Toulouse, France, 2003.
- [10] F. Chaitin-Chatelin and M.B. van Gijzen. Homotopic Deviation theory with an application to computational acoustics. Technical Report TR/PA/04/05, CERFACS, Toulouse, France, 2004.
- [11] F. Chaitin-Chatelin. The dynamics of matrix coupling with an application to Krylov methods. Technical Report TR/PA/04/29, CERFACS, Toulouse, France, 2004.
- [12] P. Lancaster, M. Tismenetsky. **Theory of Matrices.** Academic Press, New York, 1987.
- [13] J. Moro, J. Burke, M. 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.
- [14] R. Alam, S. Bora, *Effect of linear perturbation on spectra of matrices.* Linear Algebra and its Applications **368**, 329-342, (2003)
- [15] B. Aupetit, **A primer on Spectral Theory.** Springer-Verlag, New York (1991)
- [16] B.N. Parlett. *Global convergence of the basic QR algorithm on Hessenberg matrices.* Math. Comp. 22,803-817, 1968.

All Cerfacs Reports are available from:
<http://www.cerfacs.fr/algor/reports/index.html>