Lund University

Master's Thesis

Classification of Stiffness and Oscillations in Initial Value Problems

Author:
Marcus Valtonen Örnhag

Supervisor:
Professor Gustaf Söderlind

A thesis submitted in fulfilment of the requirements for the degree of Graduate student

in the

Faculty of Engineering
Centre for Mathematical Sciences
Numerical Analysis

Examinator:
Associate Professor Eskil Hansen

March 2015

Abstract

The spectrum of a linear, constant coefficient operator A can help us characterize the system $\dot{x} = Ax$ satisfactorily, but in the case of a nonlinear dynamical system such methods are not suitable. In this thesis we discuss the insufficiency of only studying the eigenvalues along the Jacobian of the solution trajectory and discuss possible indicators to better characterize such systems.

In Söderlind et al. [17] a stiffness indicator was derived, and we seek to investigate the possibility to define an *oscillation indicator*, i.e. an indicator that accurately captures the phenomenon known as oscillations. This is of scientific interest since a rigorous and computationally relevant characterization of oscillations is still missing. Furthermore, we discuss problems that are not due to nonlinearity, but non-normality, and derive a *normality indicator*.

For applications one would need computationally inexpensive indicators, and we make suggestions of such, called estimators, mimicking the behavior of the stiffness indicator and the proposed oscillation indicator. In order to demonstrate the theory sixteen computational experiments serve to illustrate a variety of different phenomena.

Populärvetenskaplig sammanfattning

Lösningarna till linjära system med konstanta koefficienter, $\dot{x} = Ax$, har varit kända i över hundra år och deras karaktär bestäms med hjälp av egenvärdena till matrisen A. Det är inte konstigt att konceptet *egenvärde* förbryllar matematik- och ingenjörsstudenter: Vad är den korrekta tolkningen? Vad betyder det? Det är inte ett enkelt begrepp att förstå, eftersom det finns många olika perspektiv på hur det ska tolkas. Inom numerisk analys talar egenvärdena om huruvida din metod kommer att konvergera och hur snabbt detta sker, men i populationsekologi förutspår egenvärdena de långsiktiga förhållandena mellan olika arter i ett ekosystem. I kvantmekanik kan egenvärdena vara energitillstånd för en partikel i en kvantbrunn och i hållfasthetsläran talar de om för dig hur du ska designa en bro för att motstå starka vindar och jordbävningar.

När man väl förstår vilket omfattande begrepp egenvärden utgör kommer ett ännu större problem – för icke-linjära dynamiska system räcker det inte att studera egenvärden. Det är dessa problem som är av intresse ute i näringslivet. För att kunna lösa sådana problem behöver vi kunna karakterisera dem, eftersom olika problem kräver olika lösningsmetoder. Detta är inte problem som du kan skriva ner för hand, utan består av miljontals ekvationer som behöver lösas, vilket även är svårt för datorer att göra om inte rätt metoder används.

I denna avhandling presenteras några förslag på hur *styvhet* och *oscillation* kan karakteriseras. Idag finns det inga vedertagna definitioner för dessa fenomen, men begreppen har existerat i forskningsvärden i över sextio år. En anledning till att det inte finns är för att det rör sig om komplexa fenomen som inte bara har en specifik egenskap. Styva ekvationer är en typ av differentialekvationer där en del numeriska metoder (explicita) är numeriskt instabila om inte steglängden är väldigt liten. Även för moderna datorer kan detta innebära långa simuleringstider och ofta överväger man att istället använda implicita metoder för att bli av med steglängdskravet. Oscillationer är inte endast lösningar till periodiska system utan även till kvasiperiodiska system och kaotiska system. Dessa kan ha olika egenskaper, t.ex. kan de vara invarianta längs med en lösningstrajektorie eller ha stabila självsvängningar.

För att demonstrera teorin analyseras sexton välkända problem med olika egenskaper och ursprung.

Acknowledgements

I would like to express my sincere appreciation to my advisor Professor Gustaf Söderlind. Without his guidance and encouragement over the past months this thesis would not have been possible.

My special thanks are extended to the NA group and other Master's thesis students at the Department of Numerical Analysis at Lund University for inspring discussions and seminars.

Nomenclature

A^{T}	Transpose of the matrix A
A^*	Complex conjugate transpose of the matrix A
$\operatorname{Tr} A$	Trace of the matrix A
$[A, B]$	Commutator of A and B
∇f	Gradient of f
$\operatorname{div} f$	Divergence of f
$\operatorname{He} A$	Hermitian part of the matrix A
$\mathrm{i}\operatorname{She} A$	Skew-Hermitian part of the matrix A
$\operatorname{No} A$	Normal part of the matrix A
$\operatorname{Ano} A$	Non-normal part of the matrix A
μ	Logarithmic norm
m_2	Upper logarithmic norm w.r.t. the spectral norm
M_2	Lower logarithmic norm w.r.t. the spectral norm
λ	Eigenvalue
σ	Singular value or s-number
$\Lambda(A)$	Spectrum of A
$\rho(A)$	Spectral radius of A
$\Lambda_\varepsilon(A)$	ε-pseudopectrum of A
$W(A)$	Numerical range of A
$\tilde{\omega}(A)$	Numerical abscissa of A
s	Stiffness indicator
ω	Oscillation indicator
ε	Complementary oscillation indicator
κ	Normality indicator
τ	Stiffness estimator
χ	Oscillation estimator

Contents

Chapter 1

Introduction

The solution to linear, constant coefficient systems such as $\dot{x} = Ax$ has been known for at least a century and the characterization of such systems are almost solely determined by the eigenvalues of A. The importance of eigenvalues, or more generally, spectra, as a tool for the mathematical sciences is unquestionably great, and throughout the evolution of computers in the 20th century, the study of spectra has been a standard tool in scientific computing. If A is diagonalizable the problem can be transformed into an eigenfunction basis and speed up the computation time. But eigenvalues are of great importance for many reasons – not only algorithmic. Lloyd N. Trefethen argues in [18] that:

> *There is a* psychological *reason for the usefulness of eigenvalues. Much of the human brain is specialized for the processing of visual information, and eigenvalues take advantage of this biological trait, supplementing the abstract notion of a matrix or operator by a picture in the complex plane. They give an operator a personality.*

In addition, researchers from different fields have their own interpretation of eigenvalues. In quantum mechanics they correspond to energy levels for a particle in a well and in population ecology they describe the long term relationship between present species. More important, however, is that the spectrum does not hold *all* information to *all* problems. For nonlinear dynamical systems such as

$$\dot{x} = f(x), \qquad x(0) = x_0,\ t \in [0, T]$$

one typically cannot expect a satisfying characterization by the spectrum of the corresponding operator. Nevertheless, it is important to be able to characterize these systems accurately, since different systems need to be treated differently.

One must keep in mind that not all complex behavior is due to nonlinearities and even linear problems may cause behavior hard to capture by analyzing the spectra. Non-normality is such a phenomenon. A *normal* matrix A satisfies the condition $A^*A = AA^*$ where A^* denotes the complex conjugate transpose. These matrices have nice features,

such as an orthogonal eigenvector basis. Non-normal matrices may have the opposite, causing distorsions when trying to transform the problem to an eigenvector basis. Since theoretically such distorsions are bounded they do not influence the system characteristics as $t \to \infty$, however, in a world with finite computing time they may cause difficulties making them almost impossible to treat with standard methods.

1.1 What is stiffness?

Although the concept of stiff problems is "known" to mathematicians and engineers alike, the absence of a rigorous definition is apparent. As Higham and Trefethen [6] expresses it:

> *What makes a stiff problem? No single answer seems right for all problems.*

Despite this, a stiff problem is commonly thought of as a problem for which explicit methods do not work. Typically the step sizes needed would have to be unreasonably small to guarantee convergence of the numerical method and one would need an implicit method to (greatly) speed up the performance.

In [17] Söderlind et al. propose a *stiffness indicator* that captures most of the commonly known characteristics of a stiff problem in a computationally relevant way. This is done by using the concept of logarithmic norms and suitable matrix decompositions to extract more information from the Jacobian along the solution trajectory than the eigenvalues alone will provide.

1.2 What are oscillations?

There are numerous examples of oscillatory systems in every field of science – from a simple pendulum in elementary physics to current research topics in complex biological models. The term *oscillatory*, however, is very broad and contains more than recurring motions. In fact, oscillatory systems include the obvious, periodic systems but also quasi-periodic systems and chaotic systems exhibiting complex dynamics. This diversity of solutions is one of the main problems in characterizing oscillatory systems. In [13] Petzold et al. writes about highly oscillatory systems:

> *What is a highly oscillatory system, and what constitutes a solution of such a system? As we will see, this question is application-dependent, to the extent that it does not seem possible to give a precise mathematical definition which would include most of the problems that scientists, engineers and numerical analysts have described as highly oscillatory.*

1.3 Aim of this thesis

Despite the obvious doubts from previous authors the aim of this thesis is to investigate the possibility to construct an *oscillation indicator* that captures most of the phenomena that commonly fall under the classification as oscillatory. By extending the analysis in [17] we seek to answer the following questions:

- Is it possible to define a rigorous and computationally relevant characterization of oscillations?
- Are stiffness and oscillations two independent phenomena?
- How does non-normality influence stiffness and oscillations?

An important question, that naturally arises, is whether or not there is a need for an oscillation indicator. Does science suffer without it? The answer is: Probably not. Ekeland et al. [4] writes similarly about stiffness:

> *It is perhaps true that a precise definition of stiffness is not crucial for practical purposes.*

I would like to think the same way about oscillations; however, it does not make the subject any less interesting, but rather captures the complexity of the phenomenon.

Chapter 2

Theory

2.1 Matrix theory

2.1.1 Cartesian decomposition

Let $A \in \mathbb{C}^{n\times n}$ and define its Hermitian and skew-Hermitian parts as

$$\operatorname{He} A = \frac{1}{2}(A + A^*) \qquad \text{and} \qquad \operatorname{She} A = \frac{1}{2\mathrm{i}}(A - A^*),$$

where A^* denotes the complex conjugate transpose of A. Clearly

$$A = \operatorname{He} A + \mathrm{i}\operatorname{She} A \qquad \text{and} \qquad A^* = \operatorname{He} A - \mathrm{i}\operatorname{She} A.$$

We will refer to this as the *Cartesian decomposition* of the matrix A, since in the scalar case $z \in \mathbb{C}$ it is simply $\operatorname{He} z = \operatorname{Re} z$ and $\operatorname{She} z = \operatorname{Im} z$. Thus, the Cartesian decomposition can be seen as a generalization of the real and imaginary parts of a complex number. Some elementary properties are listed in Table 2.1.

Table 2.1: Elementary properties of Cartesian decomposition.

Operation	A^*	$\mathrm{i}A$	A^*A	$\operatorname{He} A$	$\mathrm{i}\operatorname{She} A$	$\operatorname{She} A$
$\operatorname{He}(\cdot)$	$\operatorname{He} A$	$\operatorname{She} A^*$	A^*A	$\operatorname{He} A$	0	$\operatorname{She} A$
$\operatorname{She}(\cdot)$	$-\operatorname{She} A$	$\operatorname{He} A^*$	0	0	$\operatorname{She} A$	0

As $\operatorname{He} A$ and $\operatorname{She} A$ are Hermitian, they are unitarily similar to diagonal matrices, i.e. there exists matrices U and V such that

$$U^* \operatorname{He} AU = M = \operatorname{diag} \mu_k,$$
$$V^* \operatorname{She} AV = \Omega = \operatorname{diag} \omega_k,$$

where μ_k are called *logarithmic values* and ω_k the *angular values* of the matrix A. Furthermore,

$$\operatorname{Re}(\operatorname{Tr}[A]) = \operatorname{Tr}[\operatorname{He} A] \qquad \text{and} \qquad \operatorname{Im}(\operatorname{Tr}[A]) = \operatorname{Tr}[\operatorname{She} A].$$

Also, for real matrices, $\operatorname{Tr}[A] = \operatorname{Tr}[\operatorname{He} A]$, since $\operatorname{She} A$ has zero trace.

2.1.2 On normality and non-normality of matrices and operators

Let $A \in \mathbb{C}^{n\times n}$ and define its normal and non-normal parts as

$$\operatorname{No} A = \frac{A^*A + AA^*}{2} \qquad \text{and} \qquad \operatorname{Ano} A = \frac{A^*A - AA^*}{2},$$

giving $A^*A = \operatorname{No} A + \operatorname{Ano} A$. If A is normal, then $\operatorname{No} A = A^*A = AA^*$ and $\operatorname{Ano} A = 0$. For all matrices, $\operatorname{No} A^* = \operatorname{No} A$ and $\operatorname{Ano} A^* = -\operatorname{Ano} A$ giving

$$\|\operatorname{No} A\| = \|\operatorname{No} A^*\| \qquad \text{and} \qquad \|\operatorname{Ano} A\| = \|\operatorname{Ano} A^*\|.$$

The *commutator* is defined by $[A, B] = AB - BA$, giving the relationship

$$A^*A = (\operatorname{He} A)^2 + (\operatorname{She} A)^2 + \frac{1}{2}[A^*, A],$$

or equivalently

$$A^*A = (\operatorname{He} A)^2 + (\operatorname{She} A)^2 + \mathrm{i}[\operatorname{He} A, \operatorname{She} A]. \tag{2.1}$$

Hence A is normal if and only if $\operatorname{He} A$ and $\operatorname{She} A$ commute.

Theorem 2.1. *For every matrix $A \in \mathbb{C}^{n\times n}$ the following norm bounds hold*

$$0 \le \|\operatorname{Ano} A\|_2 \le \frac{\|A^*A\|_2}{2} \le \|\operatorname{No} A\|_2 \le \|A^*A\|_2.$$

Not all bounds hold for an arbitrary norm; however, the right-hand side can be replaced by $\max\{\|A^*A\|, \|AA^*\|\}$.

Proof. First note that $\|A^*A\|_2 = \|\Sigma\|_2^2 = \|AA^*\|_2$, with $\Sigma = \operatorname{diag}\sigma_k$, where σ_k are the singular values. The triangle inequality yields

$$\|\operatorname{No} A\|_2 = \frac{\|A^*A + AA^*\|_2}{2} \le \frac{\|A^*A\| + \|AA^*\|_2}{2} = \|A^*A\|_2.$$

This holds for any norm, except the last equality, which can be replaced by the inequality $\max\{\|A^*A\|, \|AA^*\|\}$. Consider the quadratic forms

$$0 < x^*AA^*x = x^* \operatorname{No} Ax - x^* \operatorname{Ano} Ax,$$
$$0 < x^*A^*Ax = x^* \operatorname{No} Ax - x^* \operatorname{Ano} A^*x,$$

for $x \neq 0$. Since $\|\operatorname{Ano} A\| = \|\operatorname{Ano} A^*\|$ we may pick a vector x such that the quadratic form $x^* \operatorname{Ano} Ax$ is positive and $x^* \operatorname{Ano} Ax = \|\operatorname{Ano} A\|_2$ (if not, we work with $\operatorname{Ano} A^*$ and use the second relation instead of the first). Then

$$0 < x^*AA^*x = x^* \operatorname{No} Ax - \|\operatorname{Ano} A\|_2,$$

giving $\|\text{No}\, A\|_2 \geq \|\text{Ano}\, A\|_2$. It immediately follows that

$$\|A^*A\|_2 = \|\text{No}\, A + \text{Ano}\, A\|_2 \leq \|\text{No}\, A\|_2 + \|\text{Ano}\, A\|_2 \leq 2\|\text{No}\, A\|_2 \leq 2\|A^*A\|_2,$$

establishing

$$\frac{\|A^*A\|_2}{2} \leq \|\text{No}\, A\|_2 \leq \|A^*A\|_2.$$

Finally, in order to prove that $\|\text{Ano}\, A\|_2 \leq \|A^*A\|_2/2$ we again consider the quadratic form

$$2x^*\,\text{Ano}\, Ax^* = x^*A^*Ax - x^*AA^*x \leq x^*A^*Ax \leq \|A^*A\|_2,$$

where the first inequality is due to A^*A and AA^* being positive semi-definite. Hence $2\|\text{Ano}\|_2 \leq \|A^*A\|$, and

$$0 \leq \|\text{Ano}\, A\|_2 \leq \frac{\|A^*A\|_2}{2} \leq \|\text{No}\, A\|_2 \leq \|A^*A\|_2.$$

□

2.2 Logarithmic norms and the stiffness indicator

Classically the *logarithmic norm* μ of a matrix A is defined

$$\mu[A] = \lim_{h\to 0^+} \frac{\|I + hA\| - 1}{h},$$

for some norm $\|\cdot\|$, see e.g. [16]. When deriving the *stiffness indicator* the upper (l.u.b.) and lower (g.l.b.) logarithmic norms are used. In the spectral norm, i.e. the operator norm induced by the Euclidean vector norm, this simplifies to

$$M_2[A] = \max \Lambda(\text{He}\, A) \qquad \text{and} \qquad m_2[A] = \min \Lambda(\text{He}\, A).$$

Definition 2.1 (Söderlind et al. [17])**.** The stiffness indicator s for a matrix A is defined as

$$s[A] = \frac{m_2[A] + M_2[A]}{2}.$$

From the definition it immediately follows that:

Theorem 2.2 (Söderlind et al. [17])**.** *The stiffness indicator has the following elementary properties*

1. $s[0] = 0$
2. $s[I] = 1$
3. $s[\lambda I + A] = \lambda + s[A]; \qquad \lambda \in \mathbb{R}$
4. $s[\alpha A] = \alpha s[A]; \qquad \alpha \in \mathbb{R}$
5. $m[A] \leq s[A] \leq M[A]$

2.3 Pseudospectra and numerical range

2.3.1 Definitions and elementary properties

Define the spectrum $\Lambda(A)$ of a complex-valued square matrix A as the numbers λ in the complex plane where $A - \lambda I$ is not invertible, i.e. the spectrum for an $n \times n$ matrix consists of at most n points, namely, the eigenvalues. This concept can be generalized to operators. Denote by $\mathcal{B}(\mathcal{H})$ the Banach algebra of all bounded linear operators on a complex Hilbert space $\mathcal{H}$.[1] If $\dim \mathcal{H} = n < \infty$ we have the finite dimensional case discussed above. Let $A \in \mathcal{B}(\mathcal{H})$. Then

$$\Lambda(A) = \{z \in \mathbb{C} \,:\, zI - A \text{ is not invertible in } \mathcal{B}(\mathcal{H})\}.$$

The spectrum of a matrix or operator is directly related to quantities such as $\|A^n\|$ and $\|\exp(tA)\|$ if it is normal; however, for non-normal matrices they might have nothing in common. Therefore, researchers propose a related quantity, the *pseudospectrum* (see e.g. [1, 2, 3, 5, 6, 7, 18]). The ε-pseudospectrum is defined by

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \,:\, \|(zI - A)^{-1}\| \geq \varepsilon^{-1}\},$$

where $(zI - A)^{-1}$ is known as the *resolvent*. In the spectral norm the following definition is equivalent

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \,:\, \sigma_{\min}(zI - A) \leq \varepsilon\},$$

with $\sigma_{\min}$ denoting the smallest singular value in the matrix case and the smallest s-number for an operator [18]. There are other equivalent definitions of the pseudospectrum, see e.g. [3]. The importance of studying the pseudospectrum is that it describes how the spectrum $\Lambda(A)$ transforms under small ε-perturbations. If A is normal then $\Lambda_\varepsilon(A)$ is the set of points at distance less than or equal to ε from $\Lambda(A)$. This is not the case if A is non-normal. Note also, that no matter what A is $\Lambda(A) \subseteq \Lambda_\varepsilon(A)$.

Let the numerical range W of an operator T in a Hilbert space $\mathcal{H}$ be the subset in the complex plane

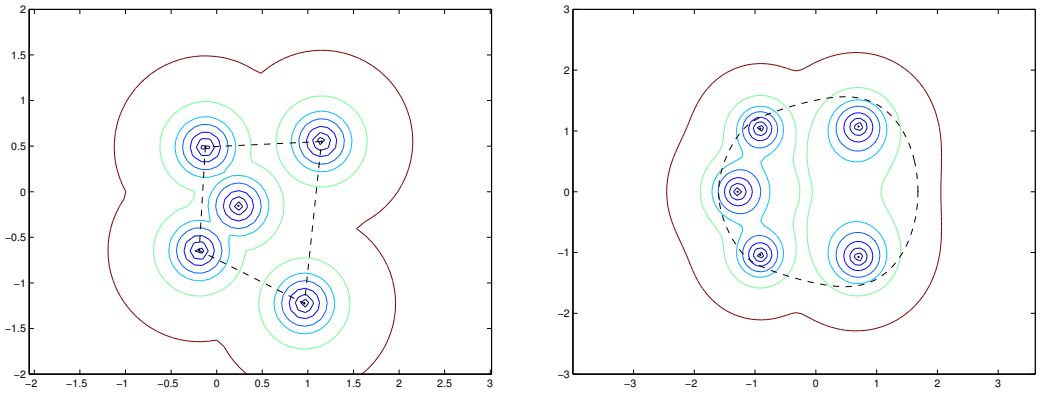
$$W(T) = \{(Tx, x) \,:\, x \in \mathcal{H},\ \|x\| = 1\}.$$

In the finite dimensional case, where $A \in \mathbb{C}^{n\times n}$, this becomes

$$W(A) = \{x^* A x \,:\, x \in \mathbb{C}^n,\ x^* x = 1\},$$

in the spectral norm. It is a direct consequence that the numerical range is the range of the *Rayleigh quotient*. Note that $\Lambda(A) \subseteq W(A)$ and if A is normal then $W(A)$ is the convex hull of its eigenvalues. In fact, by the Toeplitz-Hausdorff theorem $W(A)$ is always convex. To illustrate this, the pseudospectra and numerical range for a normal and a non-normal random 5×5 matrix are presented in Figure 2.1.

[1] This can be generalized to Banach spaces, see e.g. [1, 15].



(a) Normal matrix

(b) Non-normal matrix

Figure 2.1: The ε-pseudospectra of a normal and a non-normal matrix. The dashed line is the boundary of the numerical range. Note that the numerical range of the normal matrix (a) is the convex hull of the spectrum, whereas this is not the case for the non-normal matrix in (b).

When characterizing stability, the numerical range is a useful tool. This is perhaps not surprising, since it is a crude estimate of the spectrum. Notably the Lax-Wendroff condition involves the numerical range (cf. [9], Theorem 3).

The *numerical abscissa* of A is defined by

$$\tilde{\omega}(A) = \sup_{z \in W(A)} \operatorname{Re} z.$$

It follows from the Hille–Yosida theorem [12] that

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\exp(tA)\| \Big|_{t=0} = \tilde{\omega}(A),$$

hence describes the behavior of $\|\exp(tA)\|$ as $t \to 0$. In case of the spectral norm we have $\tilde{\omega}(A) = \mu[A] = M_2[A]$.

2.3.2 Stability and ε-pseudospectrum

Consider the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t) = Au(t), \qquad u(0) = u_0,\ t \in \mathbb{R}_0^+.$$

Applying any Runge-Kutta method to this ODE, with a fixed step size h, yields a scheme

$$U_n = p(hA)^n U_0,$$

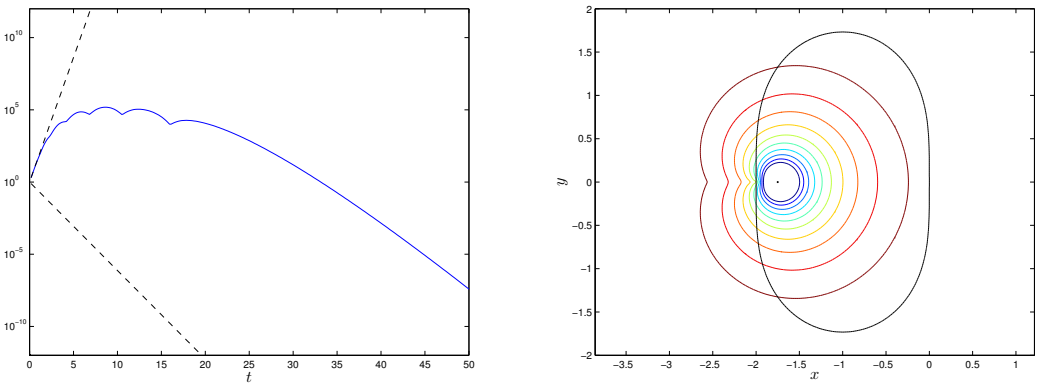
where U_n is the numerical approximations computed successively starting from given initial data $U_0 = u_0$, and p is a polynomial. For the classic RK2-method this is simply $p(x) = 1 + x + x^2/2$, i.e., the second degree Taylor expansion of e^x. The above discretization may come from a semi-discretization of a PDE, thus not limited to ODEs. The error propagation is directly linked to the growth $\|p(hA)\|^n$. Such a quantity satisfies

$$\rho(p(hA))^n \leq \|p(hA)^n\| \leq \|p(hA)\|^n,$$

where ρ denotes the spectral radius. The main reason to introduce the pseudospectrum is that these inequalities are not sharp for non-normal matrices. In [6] Higham & Trefethen illustrates this by the matrix

$$A = \begin{pmatrix} -10 & 5 & 5 & & & \\ & -10 & 5 & 5 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -10 & 5 & 5 \\ & & & & -10 & 5 \\ & & & & & -10 \end{pmatrix} \in \mathbb{R}^{16\times 16},$$

for which $\rho(p(hA)) = 0.78125$ and $\|p(hA)\|_2 \approx 2.003$, hence the bounds diverge as $n \to \infty$. In Figure 2.2 the result is replicated.



(a) **(b)**

Figure 2.2: (a) The blue line is the function $\|p(hA)^n\|_2$ and the bounds are the spectral radius and the norm respectively. Note that after $t \approx 10$ the bounds are off by more than ten orders of magnitude in both directions. In (b) the stability region (black line) of the RK2 method and the pspeudospectrum of hA, for $\varepsilon = 10^{-10}, \ldots, 10^{-1}$ is shown. The dot at $x = -1.75$ is the spectrum.

If zoomed in, the ε-pseudospectrum for $\varepsilon = 10^{-6}$ is contained in the stability region of the RK2-method, and for $\varepsilon = 10^{-7}$ it is not. Higham & Trefethen argue that this explains the hump of the blue line in Figure 2.2a being of magnitude 10^6. The following result is of importance.

Theorem 2.3 (Higham & Trefethen [6]). *Let $\dot{u} = Au$ be modeled as described above by an explicit Runge-Kutta formula with stability region S that satisfies certain technical assumptions. Then there exist positive constants C_1 and C_2, depending only on the Runge-Kutta formula and on N, such that*

$$C_1\mathcal{K} \leq \sup_{n\geq 0} \|p(hA)^n\| \leq C_2\mathcal{K}, \qquad \mathcal{K} = \sup_{z\notin S} \operatorname{dist}(z,S)\|(zI - hA)^{-1}\|,$$

Here $\operatorname{dist}(z,S)$ *denotes the usual distance of z to the set S. The constant C_1 is of modest size, depending only on the Runge-Kutta formula, while C_2 depends on the Runge-Kutta formula and also linearly on N.*

For the previous example $\mathcal{K} \approx 1.6 \cdot 10^4$ and $\sup_{n\geq 0}\|p(hA)^n\| \approx 1.5 \cdot 10^5$, yielding a more satisfactory bound. Furthermore, the authors conclude that

> *A problem is stiff for $t \approx t_0$ if the pseudospectra of this linear approximation extend far into the left half-plane as compared with the time scale of the solution for $t \approx t_0$.*

The stiffness criterion from Higham & Trefethen is summarized in Table 2.2.

Table 2.2: Summary of stiffness criterion in [6].

"Linear" theory (based on eigenvalues) ($t \to \infty$)	"Nonlinear" theory (based on norms) ($t \to 0$)	"Intermediate" theory (finite t)
A has a large spectral radius but a small spectral abscissa	A has a large norm but a small logarithmic norm	A has large pseudospectral radii but small pseudospectral abscissae

2.3.3 Relation to the logarithmic norm and the stiffness indicator

After some comparison, there are several similarities in the different approaches presented by Söderlind et al. and Trefethen & Higham. In fact, the numerical abscissa is equivalent to the logarithmic norm in the Euclidean topology.

In [14] Ransford & Rostand present a pair of 4×4 matrices having identical pseudospectra but whose squares have different norms. Furthermore, these matrices only have simple eigenvalues. This demonstrates that pseudospectra do not determine norm behavior. By our hypothesis stiffness and oscillations are topological phenomena, and the topology is induced by the norm; hence, using the pseudospectrum as the main tool to quantify these might not capture all system properties.

Chapter 3

Indicators

In this section we will present the indicators and discuss why we choose to work with them from our a priori knowledge.

3.1 Previous attempts

There has been several attempts to characterize stiffness in the past. Lambert [8] was one of the first to propose a qualitative way, by introducing the *stiffness ratio* defined as $\max|\operatorname{Re}\lambda[A]|/\min|\operatorname{Im}\lambda[A]|$, for $\lambda \in \mathbb{C}^-$. This definition, however, has obvious flaws; Söderlind et al. [17] write: *Although such a span in negative real parts of eigenvalues is often observed, it is neither necessary nor sufficient for stiffness*.

Since stability is a topological phenomenon, and eigenvalues are not, one needs to introduce a norm. Higham and Trefethen [6] propose to characterize stiffness by observing the pseudospectrum. This approach is similar to the one used by Söderlind et al. in [17]; however, the latter introduces the logarithmic norm in the definition. I believe introducing a norm is a key component to analyze stiffness; however, it does not exclude the possibility of other definitions. The benefit of the stiffness indicator is that the norm is already in the definition, whereas the pseudospectral method introduces it via the resolvent, which makes the connection to the norm-dependence indirect.

There has not been any attempts to characterize oscillations as rigorously as stiffness. Although a lot of research has been done on normal and non-normal operators the relationship to oscillations seems not to have been investigated thoroughly.

3.2 Definitions and elementary properties

The motivation behind the stiffness indicator is in [17]. The reason why we choose to include the stiffness indicator in this thesis is to investigate possible relationships between stiffness and oscillations.

The idea of characterizing oscillations is to expand the theory of the stiffness indicator. Since the Hermitian part of the operator carries information about the stiffness of the corresponding system, perhaps the oscillation indicator can be defined by using the same idea for the skew-Hermitian part. By observing that the linear system $\dot{x} = Ax$, where A is skew-Hermitian (thus having purely imaginary eigenvalues) is oscillatory, strengthens this hypothesis. In order to get a scalar quantity the norm of $\operatorname{She} A$ is the easiest but perhaps also the most efficient way to do so.

Further analysis show that this approach still bears a strong resemblance to the one used in the derivation of the stiffness indicator. Using the g.l.b. and the l.u.b. logarithmic norms respectively, will not work for real-valued matrices, as these have complex conjugate eigenvalues. Adding the g.l.b. and the l.u.b. logarithmic norms, like for the stiffness indicator, results in such an indicator being identically zero for problems with a real-valued vector field. This cannot reflect oscillations even in a linear constant coefficient system. A possibility is to choose the g.l.b. alone, but since $\operatorname{He}(\operatorname{She} A) = \operatorname{She} A$, this is equivalent to the maximum eigenvalue of $\operatorname{She} A$, which in turn is equivalent to $\|\operatorname{She} A\|$, since $\operatorname{She} A$ is Hermitian

These observations lead to the following definitions:

Definition 3.1. For a given matrix $A \in \mathbb{C}^{n\times n}$ the *oscillation indicator* is defined by

$$\omega[A] = \|\operatorname{She} A\|_2.$$

Furthermore, the *complementary oscillation indicator* is defined by

$$\varepsilon[A] = \max |\operatorname{Im} \lambda[A]|.$$

With these definitions it immediately follows that:

Theorem 3.1. *The oscillation indicator has the following elementary properties*

1. $\omega[0] = 0$
2. $\omega[I] = 0$
3. $\omega[sI + A] = \omega[A]; \qquad s \in \mathbb{R}$
4. $\omega[\alpha A] = |\alpha|\, \omega[A]; \qquad \alpha \in \mathbb{R}$

Proof. The first two properties are trivial. Assume $s \in \mathbb{R}$, then

$$\omega[sI + A] = \|\operatorname{She}(sI + A)\|_2 = \|\operatorname{She} A\|_2 = \omega[A],$$

proving property 3. The last property follows directly from $\operatorname{She}(\alpha A) = \alpha \operatorname{She} A$ for $\alpha \in \mathbb{R}$. □

Since the elementary properties in Theorem 3.1 are desirable features of an indicator measuring oscillations we expect the same results for ε.

Theorem 3.2. *The complementary oscillation indicator has the following elementary properties*

1. $\varepsilon[0] = 0$
2. $\varepsilon[I] = 0$
3. $\varepsilon[sI + A] = \varepsilon[A]; \qquad s \in \mathbb{R}$
4. $\varepsilon[\alpha A] = |\alpha|\varepsilon[A]; \qquad \alpha \in \mathbb{R}$

Proof. The first two properties are trivial. Let λ_k denote the eigenvalues of A. Then $sI + A$ has eigenvalues $s + \lambda_k$, and if $s \in \mathbb{R}$, $\operatorname{Im}(s + \lambda_k) = \operatorname{Im} \lambda_k$, and property 3 follows. Property 4 is due to the eigenvalues of αA are $\alpha\lambda_k$. □

In theory $\varepsilon[A]$ should only be markedly different from $\omega[A]$ if A is highly non-normal, due to the first being norm-dependent, and the latter norm-independent. However, one would expect $\omega[A] \approx \varepsilon[A]$ for normal matrices, due to the nice features of such matrices. Theorem 3.3 shows that this is true – in fact equality holds in the normal case.

Theorem 3.3. *Let $A \in \mathbb{C}^{n\times n}$.*

1. *If A is Hermitian, then $\omega[A] = \varepsilon[A] = 0$.*
2. *If A is normal, then $\omega[A] = \varepsilon[A]$.*

Proof. The first property follows from Hermitian matrices having real eigenvalues and that $\operatorname{She} A = 0$ for such matrices. The second follows from the fact that if A is normal it is unitarily diagonalizable, i.e. $A = UDU^*$, where U is a unitary matrix and $D = \operatorname{diag} \lambda_k$. Then $A^* = U\overline{D}U^*$, where $\overline{D} = \operatorname{diag} \overline{\lambda_k}$ and

$$\omega[A] = \frac{1}{2}\|A - A^*\|_2 = \frac{1}{2}\|D - \overline{D}\|_2,$$

where we use that $\|AU\|_2 = \|UA\|_2 = \|A\|_2$ for any square matrix A and unitary matrix U of the same size. Let $\lambda_k = \alpha_k + \mathrm{i}\beta_k$, then

$$D - \overline{D} = \operatorname{diag}(\lambda_k - \overline{\lambda_k}) = 2\mathrm{i} \operatorname{diag} \beta_k,$$

giving $\omega[A] = \max\{|\beta_k|\} = \max\{\operatorname{Im} |\lambda_k|\} = \varepsilon[A]$.

□

Since A is normal only if $\operatorname{She} A$ and $\operatorname{He} A$ commute (c.f. (2.1)) one could expect $\omega[A] \approx \varepsilon[A]$ if the quantity $\|[\operatorname{He} A, \operatorname{She} A]\|$ is small. However, we wish to handle this more rigorously. So far we have only discussed whether a matrix is normal or not. It is interesting to ask *how close* to a normal matrix a non-normal matrix is, i.e. we would like an indicator for normality. Such an indicator is of interest, since we then may compare possible correlations between non-normality and oscillations.

Recall that $\mathrm{No}\, A = (A^*A + AA^*)/2$ and $\mathrm{Ano}\, A = (A^*A - AA^*)/2$. By the decomposition of $A^*A = \mathrm{No}\, A + \mathrm{Ano}\, A$ one could proceed as with the Cartesian decomposition, $A = \mathrm{He}\, A + \mathrm{She}\, A$, to produce indicators. Similar to the oscillation indicator we choose the norm to measure normality and non-normality, i.e. $\|\mathrm{No}\, A\|$ and $\|\mathrm{Ano}\, A\|$. Rearranging the terms in Theorem 2.1 we get upper and lower bounds on the norms, i.e.

$$0 \leq \frac{\|\mathrm{Ano}\, A\|_2}{\|A^*A\|_2} \leq \frac{1}{2} \leq \frac{\|\mathrm{No}\, A\|_2}{\|A^*A\|_2} \leq 1.$$

Due to these bounds we may scale both of the quantities to be in the interval $[0, 1]$ and by subtracting them we get an indicator in the interval $[-1, 1]$.

Definition 3.2. For a given non-zero matrix $A \in \mathbb{R}^{n\times n}$ the *normality indicator* is defined by

$$\kappa[A] = \kappa_a[A] - \kappa_n[A],$$

where

$$\kappa_a[A] = \sqrt{\frac{2\|\mathrm{Ano}\, A\|}{\|A^*A\|}},$$
$$\kappa_n[A] = \sqrt{\frac{2\|\mathrm{No}\, A\|}{\|A^*A\|} - 1}\,.$$

One notable difference to the previous case is that we are only interested in quantifying normality (instead of two quantities, in the case of the Cartesian decomposition, i.e. stiffness *and* oscillations). Therefore, it seems superfluous to have two indicators; however, we cannot pick only $\kappa_n[A]$ or $\kappa_a[A]$ – we need both – which is motivated by Theorem 3.4 and Theorem 3.5 below. The main reason is that one can find a matrix A such that $\kappa_a[A] = \kappa_n[A] = 1$ can occur simultaneously, showing the insufficiency of choosing to work with either $\kappa_n[A]$ or $\kappa_a[A]$.

Lastly, since we are comparing normality against non-normality, dividing by the norm results in a relative indicator, giving a more intuitive understanding of *how normal* a matrix is.

Theorem 3.4. *The following statements are true*

$$\kappa_a[A] = 0 \quad \Rightarrow \quad \kappa_n[A] = 1, \tag{3.1}$$
$$\kappa_n[A] = 0 \quad \Rightarrow \quad \kappa_a[A] = 1, \tag{3.2}$$

but the converse does not hold, i.e.

$$\kappa_n[A] = 1 \quad \not\Rightarrow \quad \kappa_a[A] = 0, \tag{3.3}$$
$$\kappa_a[A] = 1 \quad \not\Rightarrow \quad \kappa_n[A] = 0. \tag{3.4}$$

Proof. Consider (3.1) and assume $\kappa_a[A] = 0$. By definition

$$\kappa_a[A] = 0 \quad \Rightarrow \quad \|\mathrm{Ano}\, A\|_2 = 0 \quad \Leftrightarrow \quad \mathrm{Ano}\, A = 0,$$

and by the decomposition $A^*A = \mathrm{No}\, A + \mathrm{Ano}\, A = \mathrm{No}\, A$, it follows that

$$\kappa_n[A] = \sqrt{\frac{2\|\mathrm{No}\, A\|}{\|A^*A\|} - 1} = \sqrt{\frac{2\|A^*A\|}{\|A^*A\|} - 1} = 1.$$

Consider (3.2) and assume $\kappa_n[A] = 0$. Then

$$\kappa_n[A] = 0 \quad \Leftrightarrow \quad 2\|\mathrm{No}\, A\|_2 = \|A^*A\|_2,$$

We shall prove that in this case $\|\mathrm{Ano}\, A\|_2 \geq \|\mathrm{No}\, A\|_2$. Choose a unit vector x such that $x^*A^*Ax = \|A^*A\|_2$. Then

$$0 < x^*(A^*A + AA^*)x \leq \|A^*A + AA^*\|_2 = \|A^*A\|_2 = x^*A^*Ax,$$

but since AA^* is positive semi-definite, $AA^*x = 0$. Therefore, for this specific x, we have

$$x^*(A^*A - AA^*)x = \|A^*A\|_2,$$

hence $\|A^*A - AA^*\|_2 \geq \|A^*A\|_2$, or equivalently $2\|\mathrm{Ano}\, A\|_2 \geq \|A^*A\|_2 = 2\|\mathrm{No}\, A\|_2$. By Theorem 2.1 $\|\mathrm{Ano}\, A\|_2 \leq \|\mathrm{No}\, A\|_2$ holds for all matrices A, and consequently $\|\mathrm{Ano}\, A\|_2 = \|\mathrm{No}\, A\|_2 = \|A^*A\|_2/2$ giving

$$\kappa_a[A] = \sqrt{\frac{2\|\mathrm{Ano}\, A\|_2}{\|A^*A\|_2}} = \sqrt{\frac{\|A^*A\|_2}{\|A^*A\|_2}} = 1.$$

We present a counter-example to (3.3) and (3.4) simultaneously. Consider

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

then

$$A^*A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathrm{No}\, A = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathrm{Ano}\, A = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

hence $\|A^*A\|_2 = \|\mathrm{No}\, J\|_2 = 1$, and $\|\mathrm{Ano}\, A\|_2 = 1/2$, giving $\kappa_a[A] = \kappa_n[A] = 1$.

□

Theorem 3.5. *A matrix A is normal if and only if $\kappa[A] = -1$.*

Proof. Assume A is normal. Then $\operatorname{No} A = A^*A$ and $\operatorname{Ano} A = 0$, giving $\kappa_a[A] = 0$ and $\kappa_n[A] = 1$, thus $\kappa[A] = -1$. Now, assume that $\kappa[A] = -1$. Then $\kappa_a[A] = 0$ implying $\operatorname{Ano} A = 0$, hence

$$A^*A = \operatorname{No} A \quad \Leftrightarrow \quad A^*A = AA^*,$$

proving that A is normal. □

We are now ready to list some elementary properties:

Theorem 3.6. *The normality indicator has the following elementary properties*

1. $-1 \leq \kappa[A] \leq 1$
2. $\kappa[I] = -1$
3. $\kappa[\alpha A] = \kappa[A]; \qquad \alpha \in \mathbb{C}$

Proof. The only non-trivial property is 3, which follows from the norm axioms and $\operatorname{No}(\alpha A) = |\alpha|^2 \operatorname{No} A$ and $\operatorname{Ano}(\alpha A) = |\alpha|^2 \operatorname{Ano} A$. □

Since we have seen a matrice A having $\kappa_a[A] = \kappa_n[A] = 1$ (c.f. the proof of Theorem 3.4) thus giving a normality indicator $\kappa[A] = 0$, we want to emphasize the property $\kappa[A] = 1$.

Definition 3.3. A matrix $A \in \mathbb{C}^{n\times n}$ is called *maximally non-normal* if $\kappa[A] = 1$.

3.3 Computationally inexpensive estimators

The idea in this section is to propose *estimators*, computationally inexpensive indicators, that replicate most of the properties of the stiffness indicator and the oscillation indicator. These estimators should be of low complexity and ideally Jacobian-free, i.e. one should not have to compute the complete Jacobian matrix along the solution trajectory in order to be able to compute the estimators. Such estimators could, e.g. be implemented as a step size regulator, or possibly a smooth switch between Implicit-Explicit Hybrid Methods.

Given a matrix $A \in \mathbb{C}^{n\times n}$, the stiffness indicator is the mean value of the largest and smallest eigenvalue of $\operatorname{He} A$, whereas for a real matrix $\operatorname{Tr}[A]$ is the sum of all eigenvalues. Moreover, due to the linearity of the trace,

$$\operatorname{Tr}[A] = \operatorname{Tr}[\operatorname{He} A] + \mathrm{i}\operatorname{Tr}[\operatorname{She} A],$$

and if A is real $\operatorname{Tr}[A] = \operatorname{Tr}[\operatorname{He} A]$, i.e. the trace is the sum of all eigenvalues to $\operatorname{He} A$. By normalizing the trace with the dimension of the matrix, this quantity and the stiffness indicator coincide in the case where $n = 2$. It is feasible that these quantities are related for larger n as well, if the eigenvalues are not clustered. But, there are further reasons

why we should consider the trace. The flow of a system can be analyzed in terms of how the phase volume evolves. Note that in a system $\dot{x} = f(x)$, the divergence of the vector field is $\operatorname{div}(f) = \operatorname{Tr}[\nabla f]$. For a linear system $\dot{x} = Ax$, this means that the divergence is governed by $\operatorname{Tr}[A]$. Indeed, for the flow $\exp(tA)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \det[\exp(tA)] = \operatorname{Tr}[A] \det[\exp(tA)],$$

implying that $\det[\exp(tA)] = \exp(t \operatorname{Tr}[A])$ describes the evolution of the phase volume. Phase volume preserving systems have zero trace, meaning that $\det[\exp(tA)]$ remains constant. Such systems are not stiff, but "oscillatory" in character, as they have the same behavior in forward and reverse time. Stiff problems, on the other hand, typically dissipate phase volume very fast in forward time. In these problems, the vector field has a large negative divergence, and consequently the matrix A (or the Jacobian) also has a large, negative trace. This has the further benefit, that one can obtain a rough estimate of stiffness by inspecting the diagonal of the Jacobian alone. In addition, such an estimate will directly distinguish phase volume preserving systems from stiff systems, by the direct observation of a zero trace.

With these observations in mind we make the following definition.

Definition 3.4. For a given matrix $A \in \mathbb{C}^{n \times n}$ the *stiffness estimator* is defined by

$$\tau[A] = \frac{1}{n} \operatorname{Tr}[A].$$

Theorem 3.7. *The stiffness estimator has the following elementary properties*

1. $\tau[0] = 0$
2. $\tau[I] = 1$
3. $\tau[zI + A] = z + \tau[A]; \qquad z \in \mathbb{C}$
4. $\tau[\alpha A] = \alpha \tau[A]; \qquad \alpha \in \mathbb{C}$
5. $\tau[A + B] = \tau[A] + \tau[B]$

Proof. The first two properties are trivial, and the remaining are due to the trace being linear. □

Note that the stiffness estimator does not necessarily need the complete matrix, but only the main-diagonal. Consequently, the computation time for computing the trace of a matrix is significantly less than for computing the eigenvalues ($O(n)$ compared to $O(n^3)$).

It seems harder to capture oscillations with only the information from the main-diagonal, as the off-diagonal elements describe the interaction between the different components, which are likely to cause oscillations. Since $\tau[A]$, as defined above, can be

thought of as the "mean of the eigenvalues", our hypothesis is that oscillations could be characterized as the "standard deviation of the eigenvalues", which should be large if the eigenvalues are unevenly positioned in the complex plane. For a moment, assume that the hypothesis is true. Then, the following definition is an attempt to to still use the trace, and possibly reuse $\tau[A]$ if computed.

Definition 3.5. For a given matrix $A \in \mathbb{C}^{n\times n}$ the *oscillation estimator* is defined by

$$\chi[A] = \sqrt{\tau[A^*A] - |\tau[A]|^2}\,.$$

Theorem 3.8 below strengthens our hypothesis, as the oscillation estimator has all elementary properties of the oscillation indicators (c.f. Theorem 3.1 and Theorem 3.2).

Theorem 3.8. *The oscillation estimator has the following elementary properties*

1. $\chi[A] \geq 0$
2. $\chi[0] = 0$
3. $\chi[I] = 0$
4. $\chi[sI + A] = \chi[A]; \qquad s \in \mathbb{R},\ A \in \mathbb{R}^{n\times n}$
5. $\chi[\alpha A] = |\alpha|\chi[A]; \qquad \alpha \in \mathbb{R}$

Proof. The first property follows directly from Cauchy-Schwarz inequality

$$n^2|\tau[A]|^2 = \left|\sum_{i=1}^{n}\lambda_i\right|^2 \leq n\sum_{i=1}^{n}|\lambda_i|^2 = n^2\tau[A^*A] \quad\Leftrightarrow\quad \tau[A^*A] - |\tau[A]|^2 \geq 0.$$

Properties 2 and 3 are trivial. Let $s \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ then

$$\begin{aligned}\chi^2[sI + A] &= \tau\left[(sI + A)^*(sI + A)\right] - (\tau[sI + A])^2 \\ &= \tau\left[s^2I + sA + sA^* + A^*A\right] - (s + \tau[A])^2 \\ &= s^2 + s\tau[A] + s\tau[A^*] + \tau[A^*A] - s^2 - 2s\tau[A] - (\tau[A])^2 \\ &= \tau[A^*A] - (\tau[A])^2 = \chi^2[A],\end{aligned}$$

since $\tau[A] = \tau[A^*]$ if A is real. The last property is due to the linearity of $\tau[A]$. □

In the case $n = 2$ we have shown that $s[A] = \tau[A]$, and we wish to investigate any connections between $\omega[A]$ and $\chi[A]$. In Theorem 3.9 and Theorem 3.10 we present bounds relating $\omega[A]$, $\chi[A]$ and $\tau[A]$. For this we need the following lemma.

Lemma 3.1. *Let $A \in \mathbb{R}^{n\times n}$ with real eigenvalues. Then*

$$\tau[A] + \chi[A]/(n-1)^{1/2} \leq \lambda_{\max}[A] \leq \tau[A] + \chi[A](n-1)^{1/2}\,.$$

Proof. Let $\tau[A] = \tau$, etc. Since the trace is the sum of the eigenvalues, we get

$$n\chi^2 = \sum_{i=1}^{n} \lambda_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} \lambda_i\right]^2 = \sum_{i=1}^{n} (\lambda_i - \tau)^2$$

thus giving

$$\begin{aligned} n^2(\lambda_{\max} - \tau)^2 &= \left[\sum_{i=1}^{n} \lambda_{\max} - \lambda_i\right]^2 \geq \sum_{i=1}^{n} (\lambda_{\max} - \lambda_i)^2 = \sum_{i=1}^{n} \big((\lambda_{\max} - \tau) - (\lambda_i - \tau)\big)^2 \\ &= \sum_{i=1}^{n} (\lambda_{\max} - \tau)^2 - 2(\lambda_i - \tau)(\lambda_{\max} - \tau) + (\lambda_i - \tau)^2 \\ &= \sum_{i=1}^{n} (\lambda_{\max} - \tau)^2 + (\lambda_i - \tau)^2 = n[(\lambda_{\max} - \tau)^2 + \chi^2], \end{aligned}$$

giving

$$\lambda_{\max} \geq \tau + \chi/(n-1)^{1/2}.$$

Let $\mathbf{1}$ denote the one-vector, $\boldsymbol{e}_j$ the vector with 1 at at the j:th entry and 0 elsewhere, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. Then the lower bound follows from a Cauchy-Schwarz type inequality, noting that the matrix $X = I - \mathbf{1}\mathbf{1}^{\mathsf{T}}/n$ is symmetric idempotent, hence positive semidefinite. Thus

$$(\lambda_j - \tau[A])^2 = (\boldsymbol{e}_j X \boldsymbol{\lambda})^2 \leq \boldsymbol{e}_j^{\mathsf{T}} X \boldsymbol{e}_j \cdot \boldsymbol{\lambda}^{\mathsf{T}} X \boldsymbol{\lambda} = (n-1)\chi^2,$$

giving

$$-\chi(n-1)^{1/2} \leq \lambda_j - \tau[A] \leq \chi(n-1)^{1/2},$$

and rearranging the terms, picking $\lambda_j = \lambda_{\max}$, finishes the proof. A similar proof can be found in [20]. □

From Lemma 3.1 it immediately follows that:

Theorem 3.9. *The following bounds hold*

$$\tau[(\operatorname{She} A)^2] + \chi[(\operatorname{She} A)^2]/(n-1)^{1/2} \leq (\omega[A])^2 \leq \tau[(\operatorname{She} A)^2] + \chi[(\operatorname{She} A)^2](n-1)^{1/2}.$$

Proof. Just note that

$$(\omega[A])^2 = \|\operatorname{She} A\|^2 = \lambda_{\max}[(\operatorname{She} A)^* \operatorname{She} A] = \lambda_{\max}[(\operatorname{She} A)^2].$$

□

If $n = 2$ we have the equality

$$\omega[A] = \sqrt{\tau[(\operatorname{She} A)^2] + \chi[(\operatorname{She} A)^2]}.$$

The inequality implies that when $\tau[(\operatorname{She} A)^2]$ is small, and n reasonably small, that $\omega[A] \approx \chi[(\operatorname{She} A)^2]$. The remaining question is how $\chi[(\operatorname{She} A)^2]$ is related to $\chi[A]$. This is partly answered by Theorem 3.10, for which we need the following lemma.

Lemma 3.2. *The assignment $\langle A, B\rangle = \operatorname{Tr}(B^{\mathsf{T}}A)$ yields an inner product. The norm induced by this inner product is the Frobenius norm $\|\cdot\|_F$ and is submultiplicative (although it is not an operator norm). In this inner product space the class of symmetric matrices are orthogonal to the class of skew-symmetric matrices.*

Proof. It is well-known that the trace induces the Frobenius norm and that it is submultiplicative. Moreover, assume $A^{\mathsf{T}} = A$ and $B^{\mathsf{T}} = -B$, i.e. A is symmetric and B is skew-symmetrric. Then

$$\operatorname{Tr}(B^{\mathsf{T}}A) = \operatorname{Tr}(-BA^{\mathsf{T}}) = -\operatorname{Tr}(A^{\mathsf{T}}B) = -\operatorname{Tr}((A^{\mathsf{T}}B)^{\mathsf{T}}) = -\operatorname{Tr}(B^{\mathsf{T}}A),$$

hence $\langle A, B\rangle = 0$, proving that A and B are orthogonal. □

We are now ready to prove the last theorem, which strengthens the connection between the oscillation indicator and the oscillation estimator.

Theorem 3.10. *If $A \in \mathbb{R}^{n\times n}$ then*

$$\chi[(\operatorname{She} A)^2] \leq \sqrt{n}(\chi[A])^2.$$

Proof. Let $\langle A, B\rangle = \operatorname{Tr}(AB^{\mathsf{T}})$ and $\|\cdot\|_F$ the Frobenius norm. Then, for any matrix $A \in \mathbb{R}^{n\times n}$, we have

$$n^2(\chi(A))^2 = \|A\|_F^2\|I\|_F^2 - \langle A, I\rangle^2.$$

Since $\chi((\operatorname{She} A)^2) = \chi((\mathrm{i}\operatorname{She} A)^2)$, we may work with symmetric and skew-symmetric parts instead, i.e. $\operatorname{He} A = (A + A^{\mathsf{T}})/2$ and $\mathrm{i}\operatorname{She} A = (A - A^{\mathsf{T}})/2$. Furthermore, $\|\mathrm{i}\operatorname{She} A\|_F^4 = \langle(\mathrm{i}\operatorname{She} A)^2, I\rangle^2$ hence

$$\begin{aligned}
n^4\chi(A)^4 &= \left(\|A\|_F^2\|I\|_F^2 - \langle A, I\rangle^2\right)^2 = \left(\|\mathrm{i}\operatorname{She} A\|_F^2\|I\|_F^2 + \|\operatorname{He} A\|_F^2\|I\|_F^2 - \langle \operatorname{He} A, I\rangle^2\right)^2 \\
&\geq \left(\|\mathrm{i}\operatorname{She} A\|_F^2\|I\|_F^2\right)^2 = n^2\|\mathrm{i}\operatorname{She} A\|_F^4 \geq n^2\|(\mathrm{i}\operatorname{She} A)^2\|_F^2 \\
&\geq n\left(n\|(\mathrm{i}\operatorname{She} A)^2\|_F^2 - \|\mathrm{i}\operatorname{She} A\|_F^4\right) = n\left(\|(\mathrm{i}\operatorname{She} A)^2\|_F^2\|I\|_F^2 - \langle(\mathrm{i}\operatorname{She} A)^2, I\rangle^2\right) \\
&= n^3\chi((\mathrm{i}\operatorname{She} A)^2)^2
\end{aligned}$$

and rearranging the terms yields the desired inequality. □

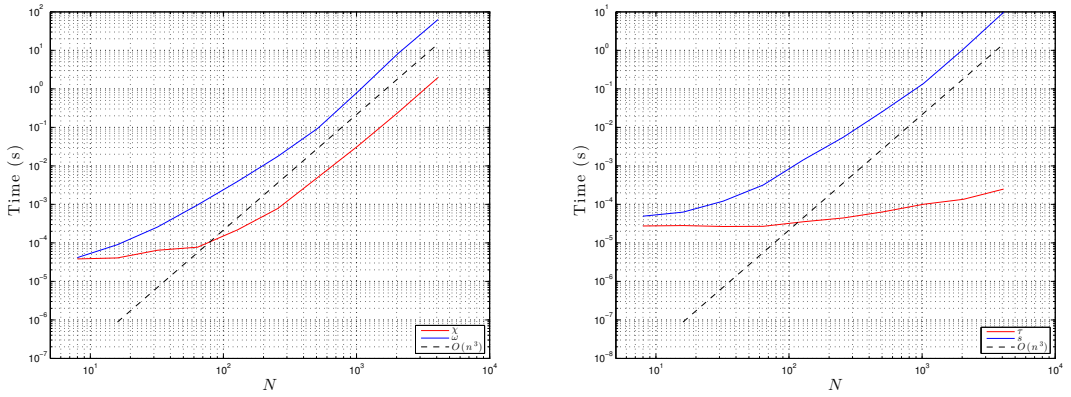
Theorem 3.9 and Theorem 3.10 suggest that if $\tau[(\operatorname{She} A)^2]$ is relatively small (compared to $\chi[A]$) then $\omega[A]$ and $\chi[A]$ are closely related. Since

$$\tau[(\operatorname{She} A)^2] = \frac{1}{2}\left(\tau[A^*A] - \tau[A^2]\right),$$

we see that if A is not maximally non-normal, or close to, it is plausible that ω and χ are in the same magnitude.

The oscillation estimator is, unfortunately, not as ideal as the stiffness estimator, since the complete Jacobian is needed in order to determine $\tau[A^*A]$. This also requires

matrix multiplication, which in the general (naive) case is of the order $O(n^3)$, giving the same complexity as the oscillation indicator. A benefit is that one does not need to call any subfunctions, hence computing $\chi[A]$ is more time efficient than $\omega[A]$, see Figure 3.1.



(a) **(b)**

Figure 3.1: Computation time for (a) ω and χ, (b) s and τ.

Chapter 4

Numerical examples

In order to demonstrate the theory proposed in previous chapters, a series of computational experiments were made from well-known problems. Several of the problems are gathered from the the Bari test set for IVPs [11]. In addition, a couple of highly oscillatory problems from different fields of applied sciences are tested and analyzed. High precision solutions were computed with MATLAB's `ode15s` solver.

4.1 Lorenz equations

Lorenz equations is well-known for exhibiting chaotic behavior for certain parameter values and initial conditions. The Lorenz equations was originally a model for atmospheric convection, and is a system of three ordinary differential equations

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y-x)\\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x(\rho-z)-y\\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy-\beta z\end{aligned}$$

where σ, ρ, and β are constants. Most commonly studied values are $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$, for which the system exhibits chaotic behavior, see Figure 4.1. We will use these values in the experiment, and $(x_0, y_0, z_0) = (10, 14, 10)$ over $t \in [0, 10]$.

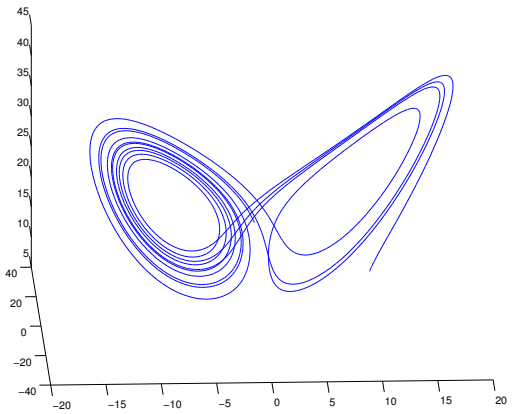


Figure 4.1: Phase portrait of Lorenz equation for $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$, with $(x_0, y_0, z_0) = (10, 14, 10)$ over $t \in [0, 10]$.

The Jacobian depends only on state, implying that the indicators also will. In Figure 4.2 the stiffness indicator is negative and mimics the behavior or the first solution component well – the maxima and minima in stiffness seems to be correlating well with the solution. The equations, however, are not what would be considered a stiff problem, but this does not necessarily mean that the stiffness indicator should be constantly zero – such problems do exist, e.g. Hamiltonian systems, and have very special structure. This shows that stiffness is something that can be present, but not necessarily dominant in a system.

The oscillation indicators ω and ε have almost identical maxima but deviate slightly from each other as the normality indicator reaches a minima. Note also that as the minima of the oscillation indicators correlate well with the maxima of the stiffness indicator.

The normality indicator has four distinct dips where $\varepsilon = 0$ due to the Jacobian having only real eigenvalues. However, note that ω is not identically zero, suggesting that the skew-Hermitian part still has influence on the solution trajectory, despite the lack of complex eigenvalues.

The estimators correlate well with the corresponding indicators, i.e. s and τ as well as ω and χ are in the same magnitude. It is not a defect that τ is constant; on the contrary, this is a nice feature for a possible step size regulator. Also, note that ω and χ have approximately the same extrema.

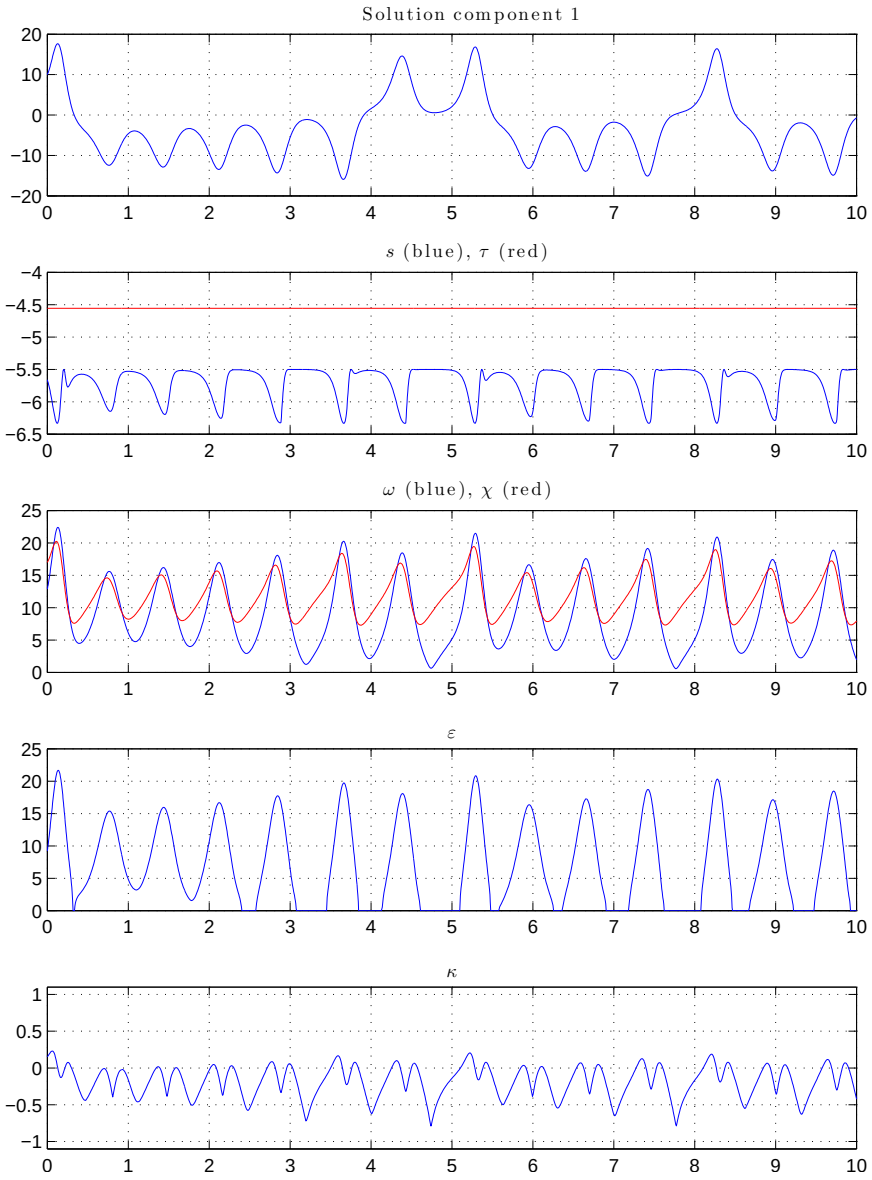


Figure 4.2: Indicators. Lorenz equation.

4.2 Chen's equation

Chen's equation is a modification of Lorenz equation and is given by

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= a(y-x)\\ \frac{\mathrm{d}y}{\mathrm{d}t} &= (c-a)x - xz + cy\\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - bz\end{aligned}$$

where a, b, and c are constants. This system is interesting as it is in the chaotic regime when the Lorenz equation is not [19]. See phase portrait in Figure 4.3. In the experiment we use $a = 35$, $b = 8/3$ and $c = 28$, with $(x_0, y_0, z_0) = (10, 14, 10)$ over $t \in [0, 10]$.

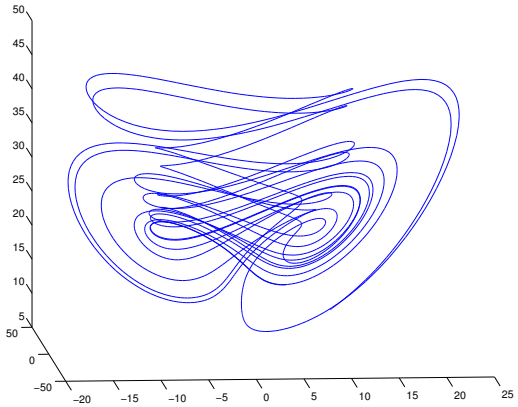


Figure 4.3: Phase portrait of Chen's equation for $a = 35$, $b = 8/3$ and $c = 28$, with $(x_0, y_0, z_0) = (10, 14, 10)$ over $t \in [0, 10]$.

Although Chen's equation is a modification of Lorenz equations and resembles the visual behavior in phase-space, the indicators, see Figure 4.4, suggests that the properties of Chen's equation are quite different from Lorenz equations. The most remarkable difference is the non-normality, which, in comparison to the Lorenz equations, are above 0.5 in magnitude along most parts of the trajectory, but has small dips where the behavior of the oscillation indicators are similar to that of the Lorenz equations as ε tends quickly to zero. Note also that the $\omega \gtrsim \varepsilon$ which should be expected since the system exhibits strong non-normality; a property that is not prevalent in the Lorenz equations.

Note also that the stiffness indicator has not been significantly altered, thus the change in non-normality in this case seems to have a larger impact on oscillations.

Again, the estimators approximate the corresponding indicators well. The stiffness estimator is constant, as in Lorenz equations, but is still in the same magnitude. The extrema of ω and χ correlate well.

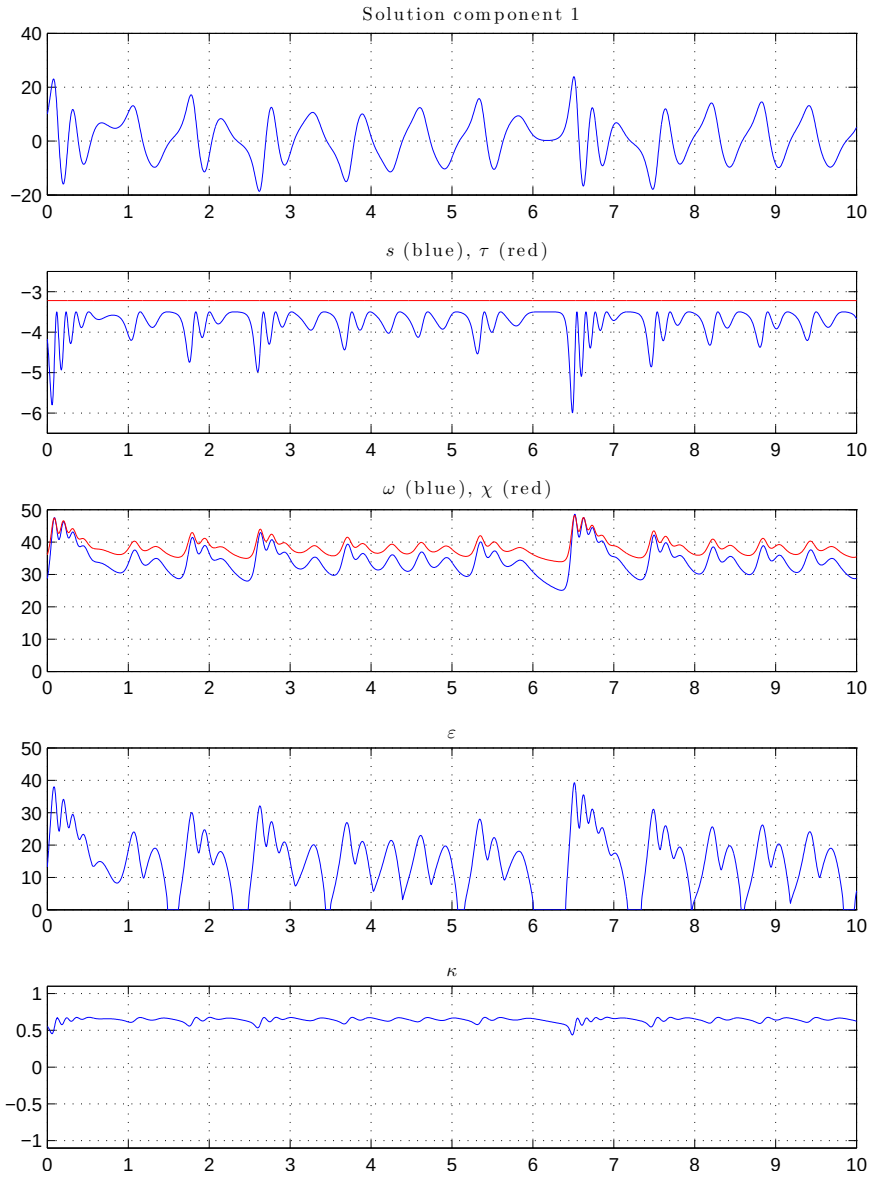


Figure 4.4: Indicators. Chen's equation.

4.3 Duffing oscillator

The Duffing oscillator is a periodically forced oscillator with a nonlinear elasticity, given by

$$\ddot{x} + \delta\dot{x} + \beta x + \alpha x^3 = \gamma \cos \omega t$$

which is rewritten

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= x_2 \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= \gamma \cos \omega t - \delta x_2 - \beta x_1 - \alpha x_1^3 \end{aligned}$$

where α, β, γ, δ and ω are constants. For $\alpha > 0$, $\beta < 0$ and $\delta > 0$ some interesting dynamics occur, see phase-plot in Figure 4.5. In the experiment we use $x_0 = (1, 1)$ over $t \in [0, 40]$ with $\alpha = 1$, $\beta = -1$, $\gamma = 0.3$, $\delta = 0.2$ and $\omega = 1$.

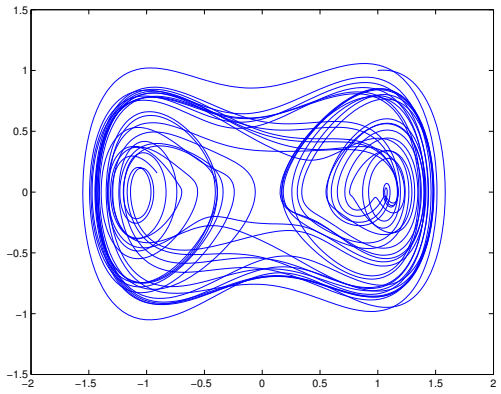


Figure 4.5: Phase portrait of the Duffing oscillator for $\alpha = 1$, $\beta = -1$, $\gamma = 0.3$, $\delta = 0.2$ and $\omega = 1$, with $x_0 = (1, 1)$ over $t \in [0, 200]$.

The Jacobian, and therefore the indicators, depend only on state. The system shows different characteristics from the previous problems as the stiffness indicator is constant. The maxima of the oscillation indicators again coincide with the maxima along the solution trajectory. Since the normality indicator is $\gtrsim 0.5$ along parts of the trajectory the deviation between ω and ε is not surprising.

There is, however, an interesting recurring phenomenon; the normality indicator drops to -1 at certain intervals. This is when the solution along the trajectory has a fixed directional derivative, and thus the Jacobian matrix is a good approximation of a linear model to the otherwise complex dynamics.

Analytical computations verify that the stiffness is only dependent on the parameter δ, more precisely, the stiffness indicator (and the estimator) $s[A] = -\frac{\delta}{2}$ and the oscillation indicator $\omega[A] = \frac{1}{2}|3\alpha x_1^2 + \beta + 1|$.

Interestingly, ω and χ seems to correlate well when $|\kappa| \lesssim \frac{1}{2}$, but deviates otherwise, i.e. for highly non-normal matrices. It is, however, unclear whether the deviation between ω and χ when the normality indicator drops to -1 is due to normality. Perhaps this phenomenon is due to the rapid change in normality.

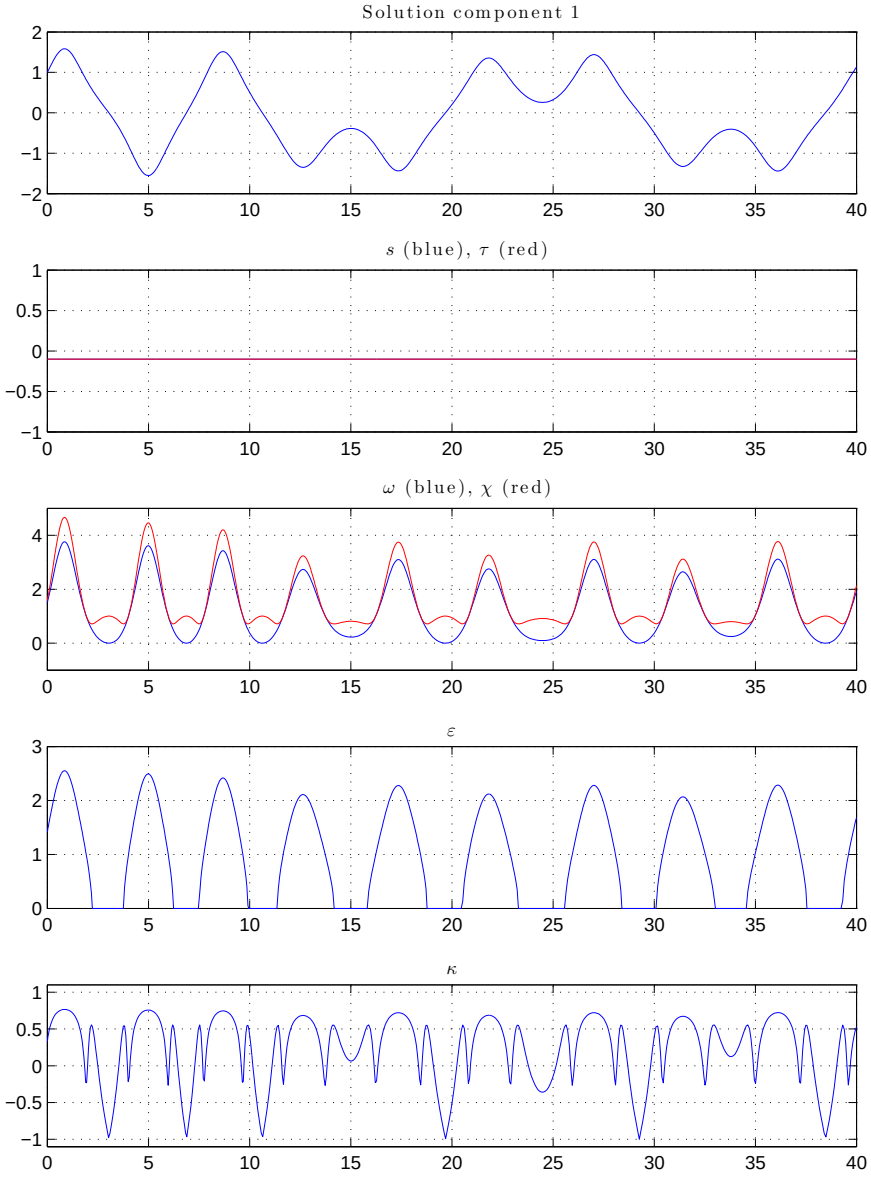


Figure 4.6: Indicators. Duffing oscillator.

4.4 Stiff spring pendulum

The following system from multibody dynamics is described in Petzold et al. [13]

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= u \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= v \\ \frac{\mathrm{d}u}{\mathrm{d}t} &= -\lambda x \\ \frac{\mathrm{d}v}{\mathrm{d}t} &= 1 - \lambda y\end{aligned}$$

where $\varepsilon^2\lambda = (r-1)/r$ and $r = \sqrt{x^2+y^2}$. The solution consist of a low frequency oscillation and a superimposed high-frequency oscillation, see Figure 4.7. This system is claimed to be highly oscillatory by the authors. In the experiment we use $(x_0, y_0, u_0, v_0) = (0.9, 0.1, 0, 0)$ and $\varepsilon^2 = 10^{-3}$.

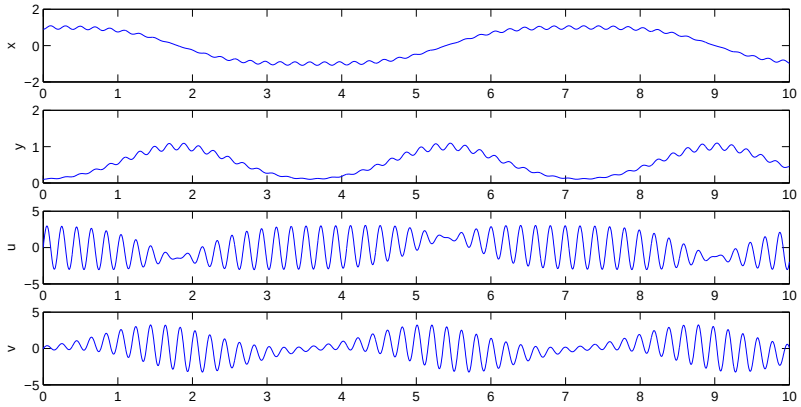


Figure 4.7: Solutions to the stiff spring pendulum for $\varepsilon^2 = 10^{-3}$ and initial values $(0.9, 0.1, 0, 0)$ over $t \in [0, 10]$.

The stiffness indicator is constantly zero due to the non-normality but what is even more interesting is that the rest of the indicators are as well. This is to be expected, since the motions are periodic and do not change drastically during a period; however, the spectrum is constantly changing along the solution trajectory. This is not reflected by the indicators, as the large purely imaginary eigenvalues (the high frequency) remain the same. This is a good feature of the oscillation indicator, since it shows that it can filter out the essential phenomenon that we want to analyze.

Note that the normality indicator is constantly 1, which is to be expected from a periodic system. The observations from earlier problems suggest that high non-normality reflects in prevalent oscillations; however, this example shows a constant behavior. Perhaps what causes oscillations in a system, stiff or nonstiff, is the *change in normality* rather than a high magnitude of non-normality.

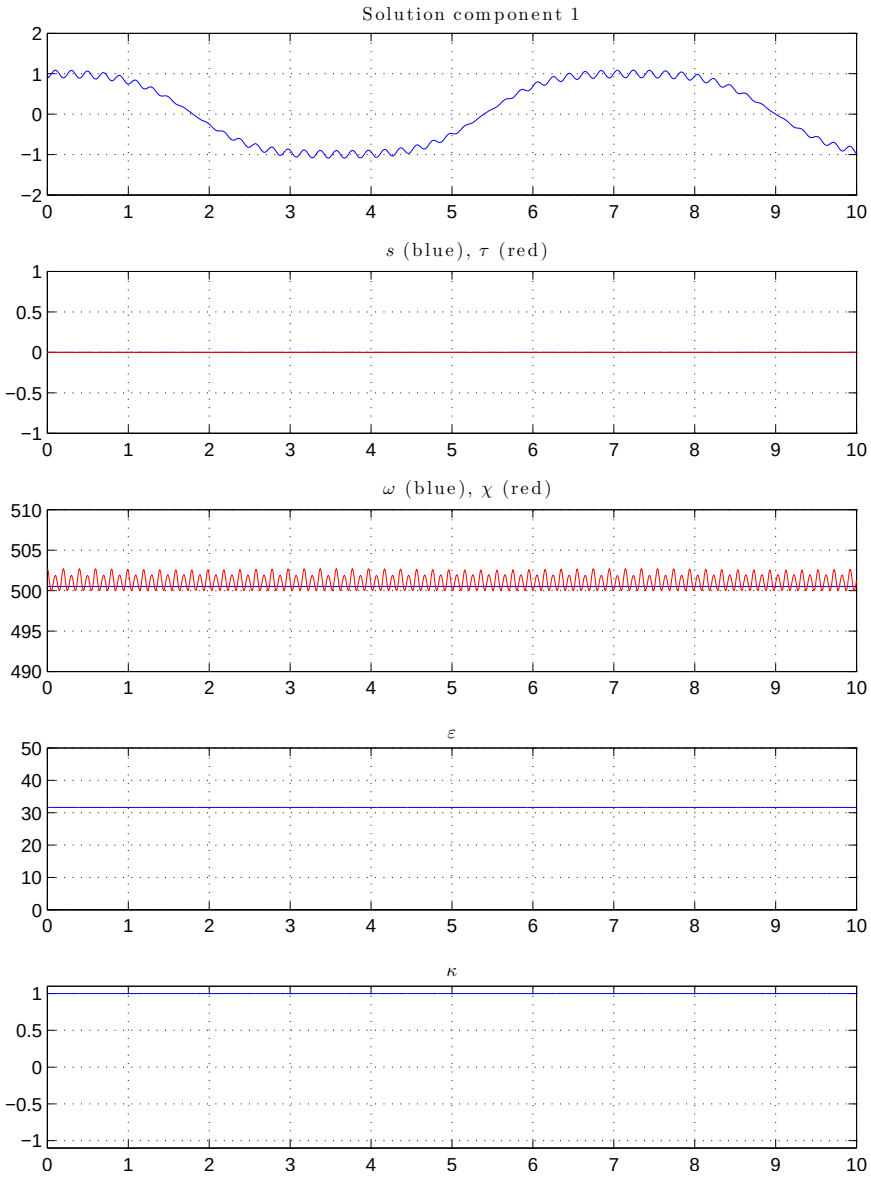


Figure 4.8: Indicators. Stiff spring pendulum.

4.5 Stellar Orbit Problem with Resonance

In [10] Lee & Engquist study a highly oscillatory system from the theory of stellar orbits in a galaxy,

$$\dot{x} = \frac{1}{\varepsilon^2}\begin{pmatrix} 0 & a & 0 & 0 \\ -a & 0 & 0 & 0 \\ 0 & 0 & 0 & b \\ 0 & 0 & -b & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ x_3^2/a \\ 0 \\ 2x_1x_2/b \end{pmatrix}$$

where $a = 2$ and $b = 1$, $\varepsilon^2 = 10^{-4}$ over $t \in [0, 30]$. The components originates from a reference circular orbit and a secondary term measuring the deviation of the orbit from the galactic plane. In Figure 4.9, the high frequency solutions are impossible to distinguish but the low-frequency resonance modes are visible.

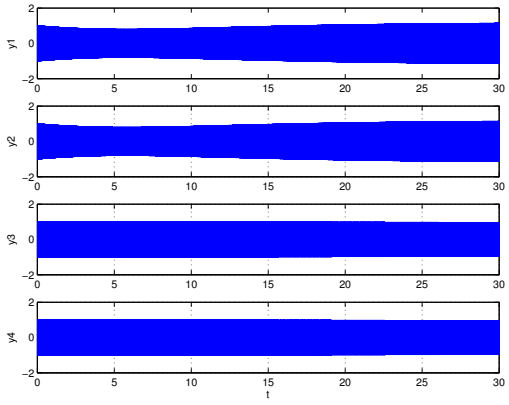


Figure 4.9: Solution components with initial conditions $(1, 0, 1, 0)$.

The indicators show that the Stellar Orbit Problem exhibits similar behavior as in the stiff spring pendulum in the sense that they are constant; however, the Jacobian along the solution trajectory is normal. This is reflected in the oscillation indicators ω and ε being identical, in accordance with Theorem 3.3. This is the opposite of what was observed in the stiff spring pendulum, although the physical models are similar – both are highly oscillatory. The reason they differ substantially is because in the stiff spring pendulum the low-amplitude, high frequency is superimposed on the low frequency mode, whereas in this problem the low frequency is modulating a high frequency. Clearly, the systems have totally different characteristics, despite both being nonstiff.

The estimators are again in the same magnitude as the corresponding indicators.

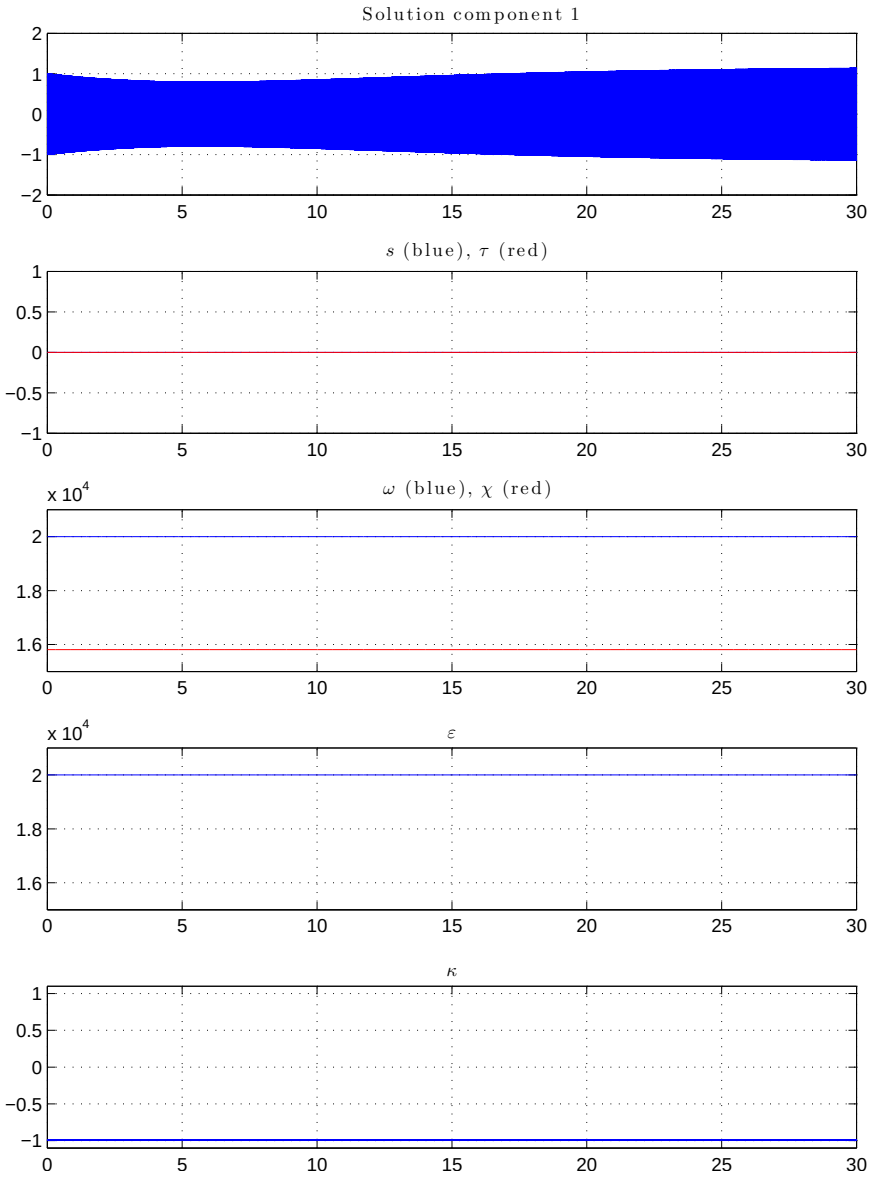


Figure 4.10: Indicators. Stellar Orbit Problem with Resonance.

4.6 Double pendulum

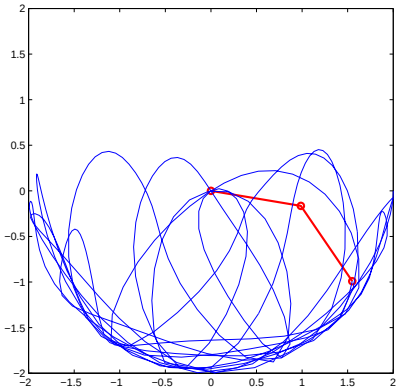
The double pendulum is a simple physics experiment known to exhibit chaotic behavior. The equations are given by

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= u \\ \frac{\mathrm{d}u}{\mathrm{d}t} &= \frac{ed - bf}{ad - cb} \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= v \\ \frac{\mathrm{d}v}{\mathrm{d}t} &= \frac{af - ce}{ad - cb}\end{aligned}$$

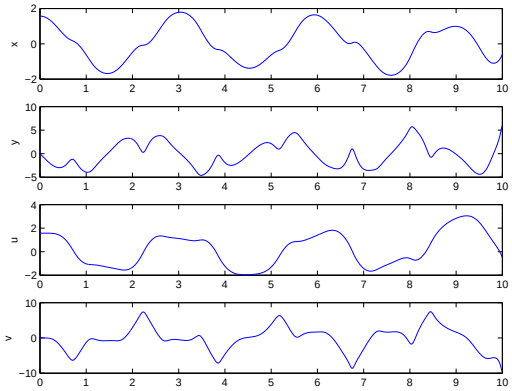
where

$$\begin{aligned} a &= (m_1 + m_2)\ell_1 & b &= m_2\ell_2 \cos(x - y) \\ c &= m_2\ell_1 \cos(x - y) & d &= m_2\ell_2 \\ e &= -m_2\ell_2 v^2 \sin(x - y) - g(m_1 + m_2)\sin x & f &= m_2\ell_1 y^2 \sin(x - y) - m_2 g \sin y \end{aligned}$$

for some constants m_1, m_2, ℓ_1, ℓ_2 and g. In the experiment we use $m_1 = m_2 = \ell_1 = \ell_2 = 1$ and $g = 9.81$ with $(x_0, u_0, y_0, v_0) = (\pi/2, 0, \pi/2, 0)$ over $t \in [0, 10]$.



(a) Solutions to the double pendulum and initial values $(\pi/2, 0, \pi/2, 0)$ over $t \in [0, 20]$, with $m_1 = m_2 = \ell_1 = \ell_2 = 1$ and $g = 9.81$.



(b) Solutions to the double pendulum and initial values $(\pi/2, 0, \pi/2, 0)$ over $t \in [0, 10]$, with $m_1 = m_2 = \ell_1 = \ell_2 = 1$ and $g = 9.81$.

The stiffness indicator and the corresponding estimator correlate well, but the estimator is smoother. The oscillation indicator and the corresponding estimator are almost identical. Due to the non-normality the difference between ω and ε is not surprising.

The stiffness indicator is changing signs, indicating that the system is indefinite, or possibly changing between dissipative and accretive (such characterizations are due to the sign of the g.l.b. and l.u.b. logarithmic norms).

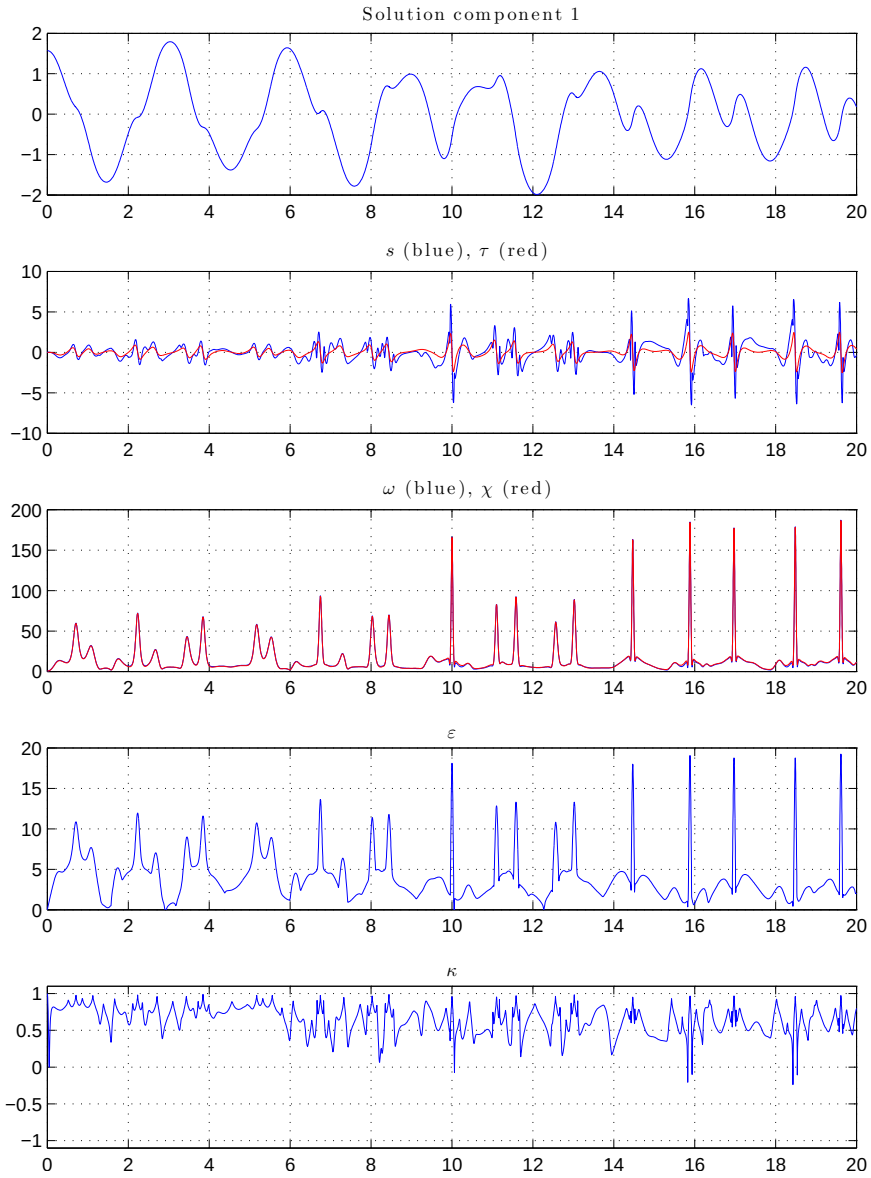


Figure 4.12: Indicators. Double pendulum.

4.7 Lotka-Volterra

The Lotka-Volterra equations is a simple predator-prey model given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(a - by)$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = -y(c - dx)$$

for constants a, b, c and d. The solutions to the systems are limit cycles, depending on the choice of constants, see Figure 4.13. In the experiment we use $(x_0, y_0) = (1, 1)$ over $t \in [0, 2.5]$ with $a = 3$, $b = 9$ and $c = d = 15$.

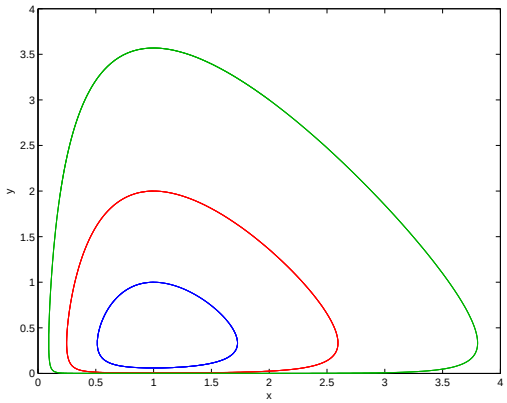


Figure 4.13: Phase portrait suggesting periodic solutions of the Lotka-Volterra equations for different initial conditions.

Since it is a 2×2 system we can compute some of the indicators analytically

$$J = \begin{pmatrix} a - by & -bx \\ dy & c - dx \end{pmatrix}$$

thus

$$\text{He}\, J = \begin{pmatrix} a - by & \frac{1}{2}dy - \frac{1}{2}bx \\ \frac{1}{2}dy - \frac{1}{2}bx & c - dx \end{pmatrix} \quad \text{and} \quad \text{She}\, J = \begin{pmatrix} 0 & -(\frac{1}{2}bx + \frac{1}{2}dy) \\ \frac{1}{2}bx + \frac{1}{2}dy & 0 \end{pmatrix}$$

which gives the stiffness indicator $s[J] = \frac{1}{2}(a - c - by + dx)$ and oscillation indicator $\omega[J] = \frac{1}{2}|bx + dy|$.

Interestingly stiffness is caused by all parameters, whereas oscillations are caused only by b and d, which are the constants modeling the interaction between the species.

The peaks of the normality indicator does not seem to correlate with the other indicators. This indicates that non-normality is not sufficient to cause stiffness nor oscillations in the system.

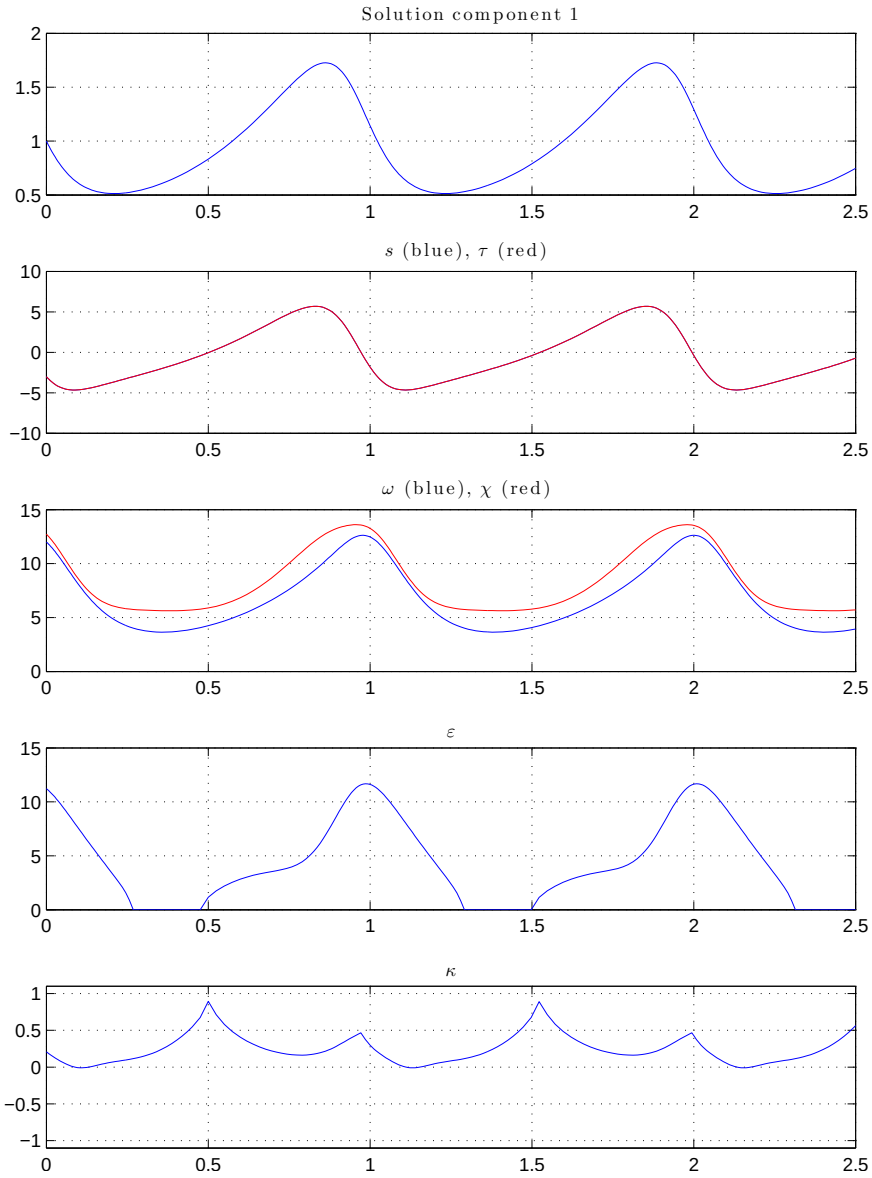


Figure 4.14: Indicators. Lotka-Volterra, for $a = 3$, $b = 9$, $c = d = 15$ over $t \in [0, 2.5]$

4.8 Oregonator

The Oregonator is a theoretical model for a chemical reaction and is given in a normalized form by

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}\theta} &= 320 \cdot s(x_1 - x_1x_2 + x_2 - qx_1^2)\\ \frac{\mathrm{d}x_2}{\mathrm{d}\theta} &= 320 \cdot (x_3 - x_2 - x_1x_2)/s\\ \frac{\mathrm{d}x_3}{\mathrm{d}\theta} &= 320 \cdot w(x_1 - x_3)\end{aligned}$$

where $\theta = t/320$, with t being the original time scaling. The solutions to the Oregonator are limit cycles, as shown in Figure 4.15. The Oregonator is a well-known stiff problem. In the experiment we use $s = 77.27$, $q = 8.375 \cdot 10^{-6}$ and $w = 0.161$ with $\boldsymbol{x}_0 = (1, 1, 2)$ over $t \in [0, 1]$.

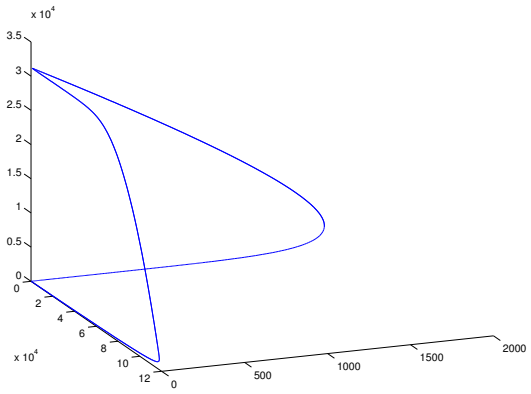


Figure 4.15: Phase portrait suggesting periodic solutions of the Oregonator.

In Figure 4.16 the indicators show that the problem is indeed stiff, but that stiffness and oscillations are not two separate phenomena. Rather, they seem to be coupled if the Jacobian matrix along the solution trajectory is highly non-normal; however, oscillations can be present with or without stiffness or non-normality, as observed in many problems before (e.g. Stiff spring pendulum).

With recent observations and the rigorous mathematical definition of stiffness and oscillations proposed in this thesis, one may criticize the expression *stiff problem*, in the sense that most such problems, as with the Oregonator, is not always stiff, but rather exhibits certain phases in which it is stiff.

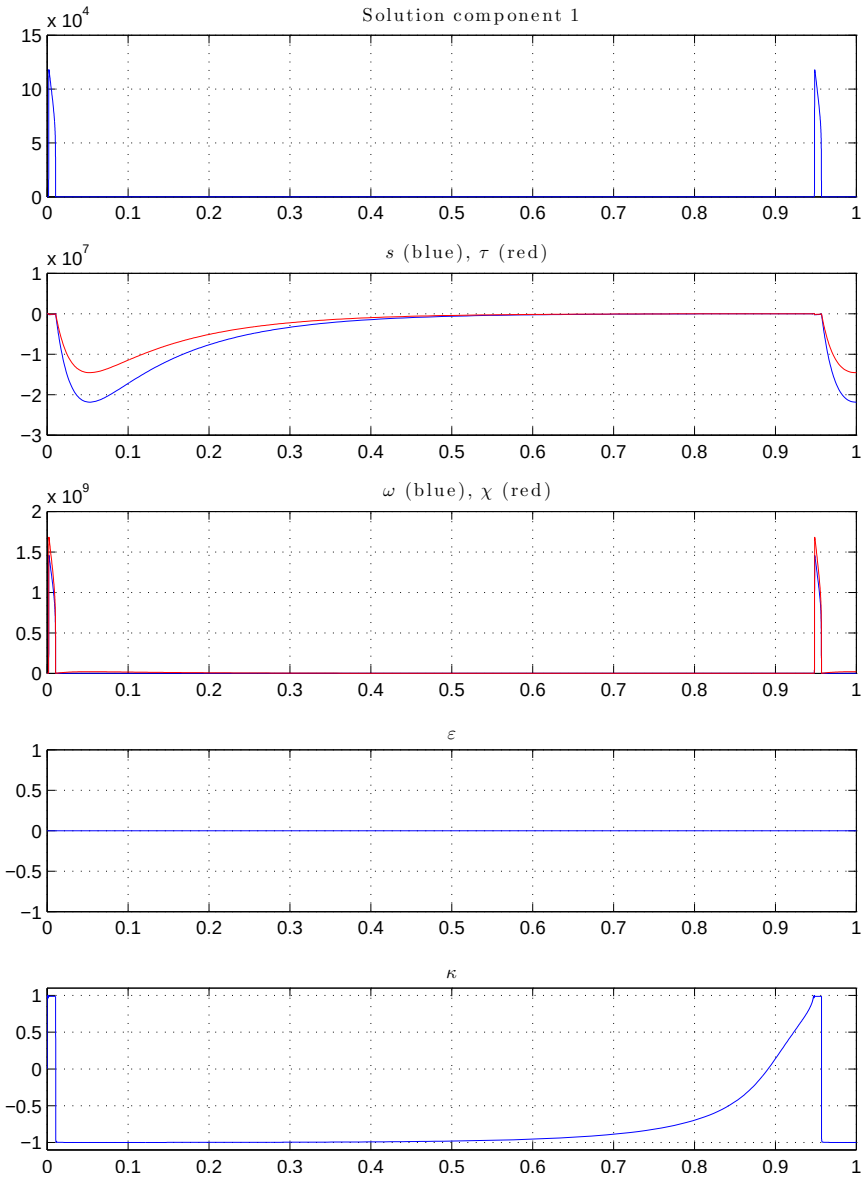


Figure 4.16: Indicators. Oregonator, $s = 77.27$, $q = 8.37 \cdot 10^{-6}$ and $w = 0.161$.

4.9 Van der Pol oscillator

The Van der Pol oscillator is an oscillator with nonlinear damping designed to model an electric circuit. The normalized van der Pol equation is given by

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= 2\mu x_2 \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= 2\mu^2(1-x_1^2)x_2 - 2\mu x_1\end{aligned}$$

where μ is a positive parameter. For large values of μ the system is known to be stiff. The solutions are limit cycles, see Figure 4.17. In the experiment we use $\mu = 200$, with $\boldsymbol{x}_0 = (2, 0)$ over $t \in [0, 1]$.

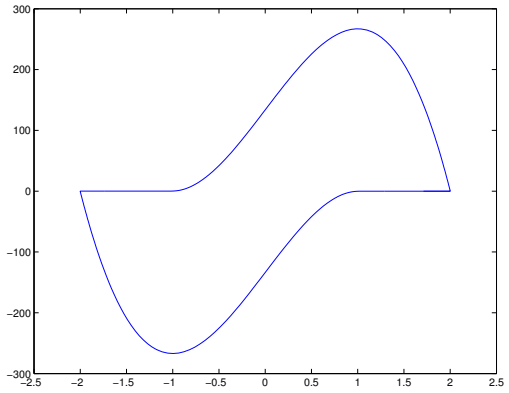


Figure 4.17: Phase portrait suggesting periodic solutions of the normalized van der Pol equation.

In Figure 4.18 we note that the stiffness indicator is positive when $t \approx 0.42$ where the transition occurs, but otherwise negative. Interestingly all indicators peak around this point, which suggests that even though the problem is considered stiff perhaps the terminology should be that it is stiff at certain parts along the solution trajectory, rather than considering the whole system as stiff.

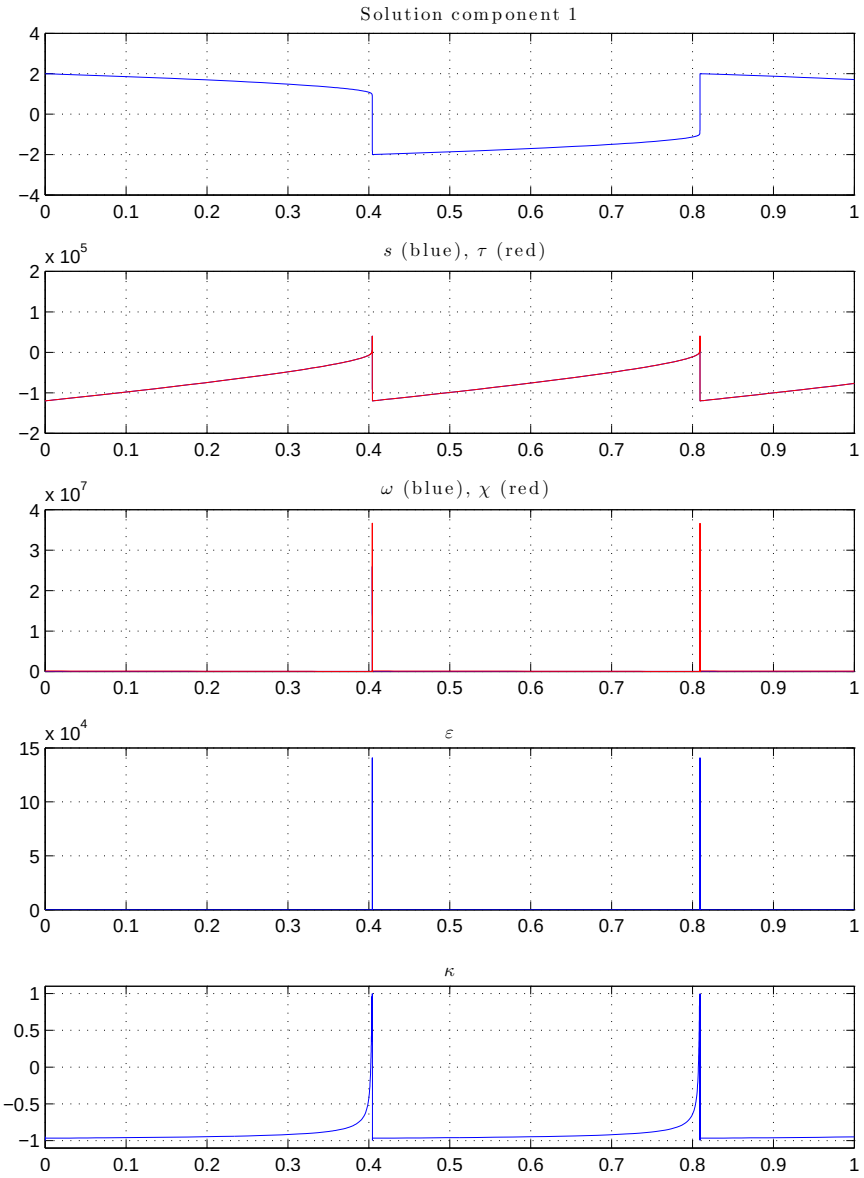


Figure 4.18: Indicators. Normalized van der Pol, with $\mu = 200$.

4.10 Verwer's Pollution model

Verwer's Pollution model is a stiff system of 20 non-linear ODEs and describes chemical reaction between a number of compounds. The solutions are shown in Figure 4.19.

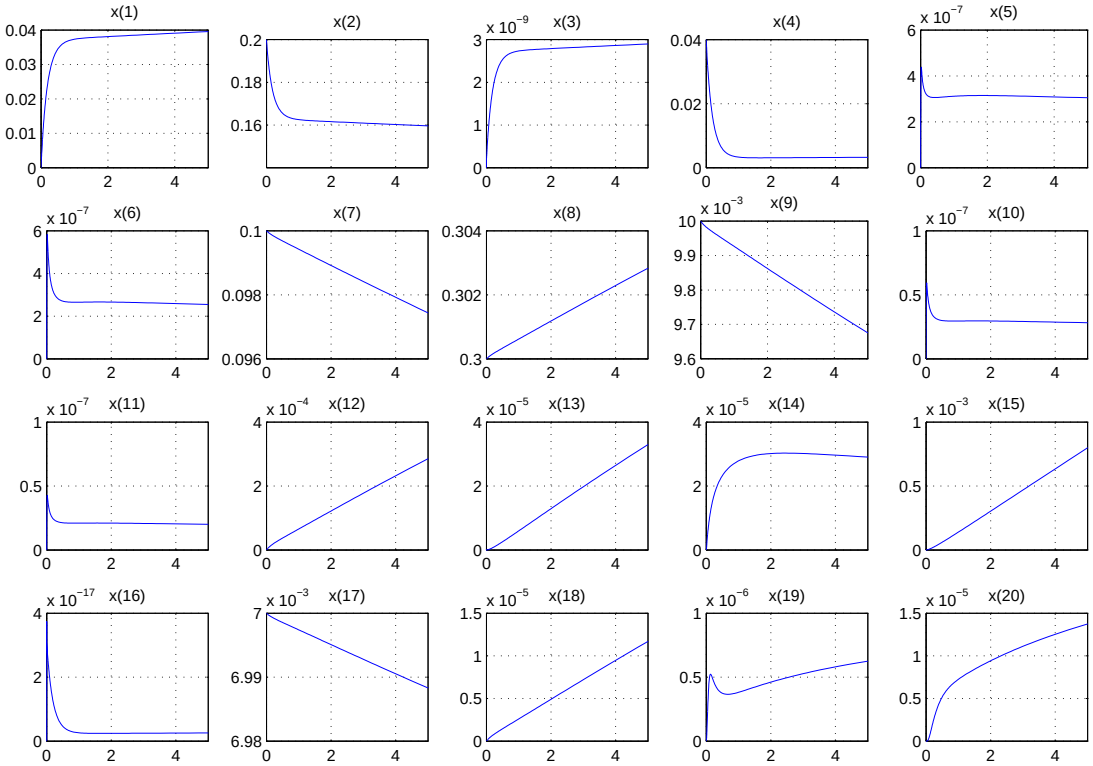


Figure 4.19: Solutions to Verwer's Pollution model.

This is the third and final problem that exhibits the property that all indicators are (almost) constant. Clearly, the stiffness indicator suggest that the problem is indeed stiff, see Figure 4.20, and unlike the Van der Pol oscillator it is constantly stiff, meaning that any attempt of using an explicit method on any part along the solutions trajectory will be in vain.

In the two previous problems where the indicators were constant the normality indicator was either 1 or -1. In this case it is constantly zero and so is the complementary oscillation indicator ε, meaning that the imaginary parts of the eigenvalues are zero. Although, as seen in Figure 4.19, the solutions are not what one would intuitively call oscillatory, the directional derivative is changing and some of the solution component have local extremas – this is where the oscillation indicator ω reveals the influence of the skew-Hermitian part.

For this system $s/\tau \approx 10$ which might underestimate the stiffness, and could cause problems with step size regulation.

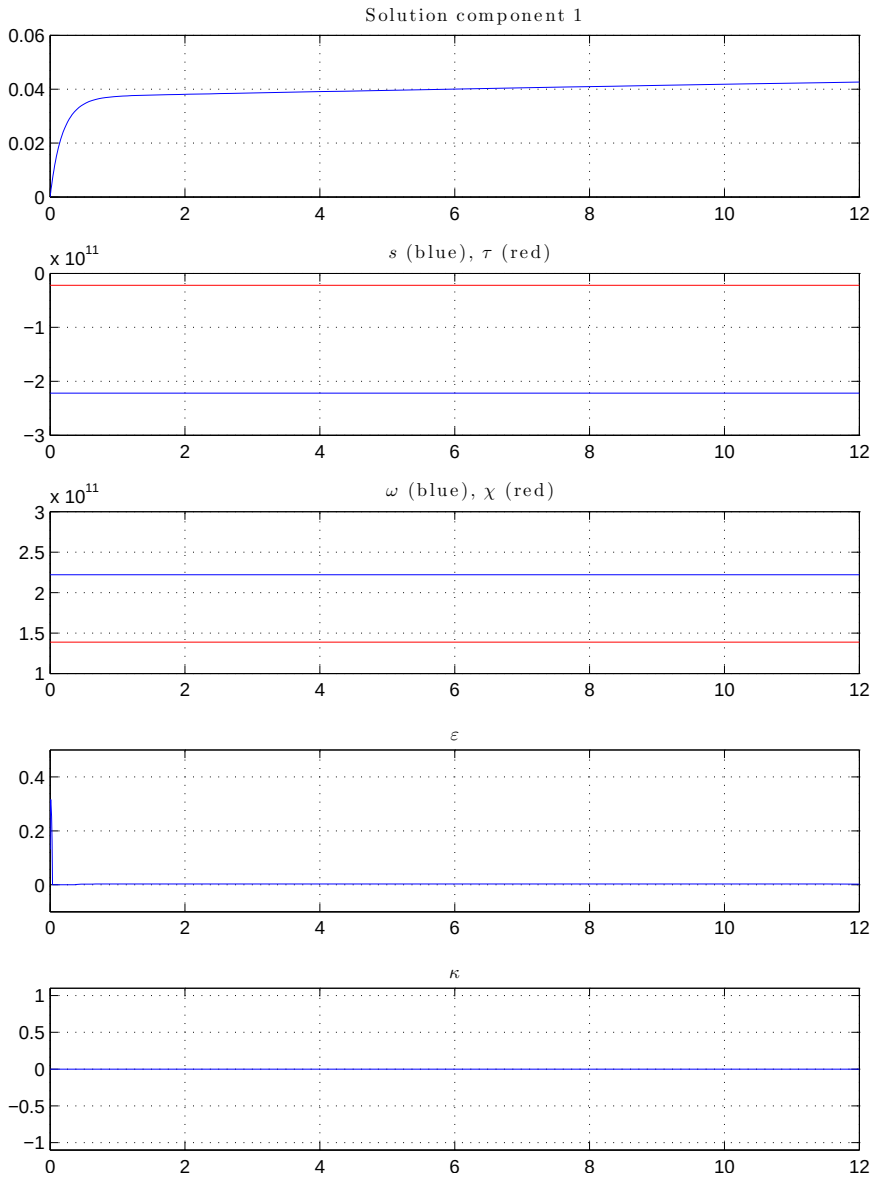


Figure 4.20: Indicators. Verwer's Pollution model.

4.11 Airy's equation

Airy's equation originates from early studies in optics and is given by

$$\ddot{x} - tx = 0,$$

which we may rewrite

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = tx_1$$

The solutions are shown in Figure 4.21. The problem is interesting because we may compute the eigenvalues analytically, and thus have a good a priori knowledge of the system. In the experiment we use $\boldsymbol{x}(-10) = (-0.25, 0)$ over $t \in [-10, 5]$.

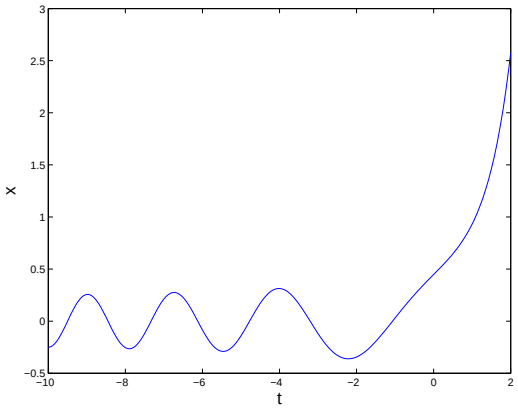


Figure 4.21: First solution component to Airy's equation over $t \in [-10, 2]$.

The Jacobian is

$$A = \begin{pmatrix} 0 & 1 \\ t & 0 \end{pmatrix}$$

giving

$$\mathrm{He}\, A = \begin{pmatrix} 0 & \nu \\ \nu & 0 \end{pmatrix} \quad \text{and} \quad \mathrm{She}\, A = \begin{pmatrix} 0 & \rho \\ -\rho & 0 \end{pmatrix}$$

with $\nu = \frac{1}{2} + \frac{1}{2}t$ and $\rho = \frac{1}{2} - \frac{1}{2}t$, giving $s[A] = 0$ and $\omega[A] = \frac{1}{2}|t - 1|$.

The eigenvalues are purely imaginary for $t < 0$ and real for $t > 0$, which is seen in complementary oscillation indicator ε, see Figure 4.20. What is more interesting though is the transition phase $-1 \leq t \leq 1$, where the problem changes character. The oscillation indicator is interesting to observe here since it is zero at $x = 1$ (instead of $x = 0$ where the spectrum is degenerate). The information coming from only observing the eigenvalues does not capture this transition phenomenon.

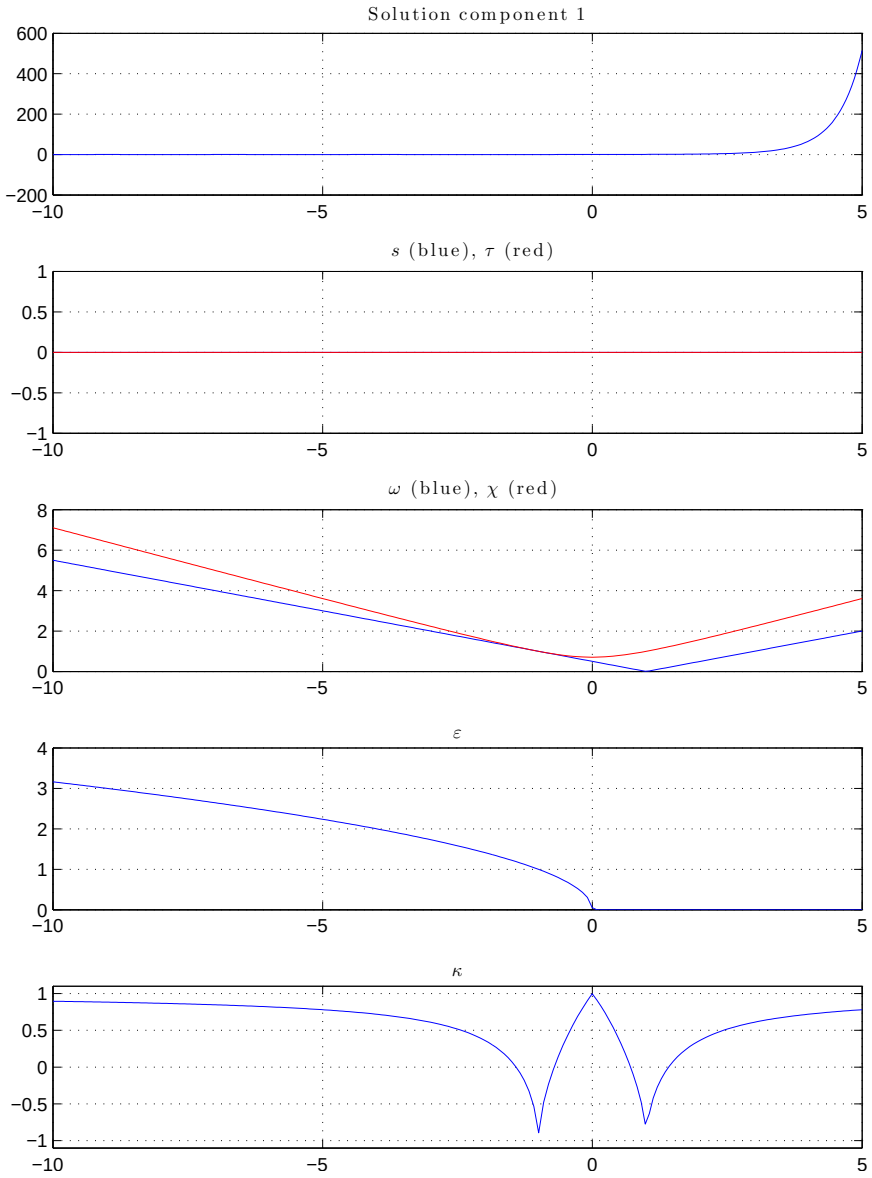


Figure 4.22: Indicators. Airy's equation.

4.12 Brusselator

The Brusselator models an autocatalytic, oscillating chemical reaction. In such reactions a species acts to increase the rate of its producing reaction. The equations are given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a - (b+1)x + ax^2y$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = bx - ax^2y$$

where $a, b > 0$. The solutions approach a limit cycle, see Figure 4.23. In the experiment we use $a = 1$, $b = 2$ with $(x_0, y_0) = (1.5, 3)$ over $t \in [0, 14]$.

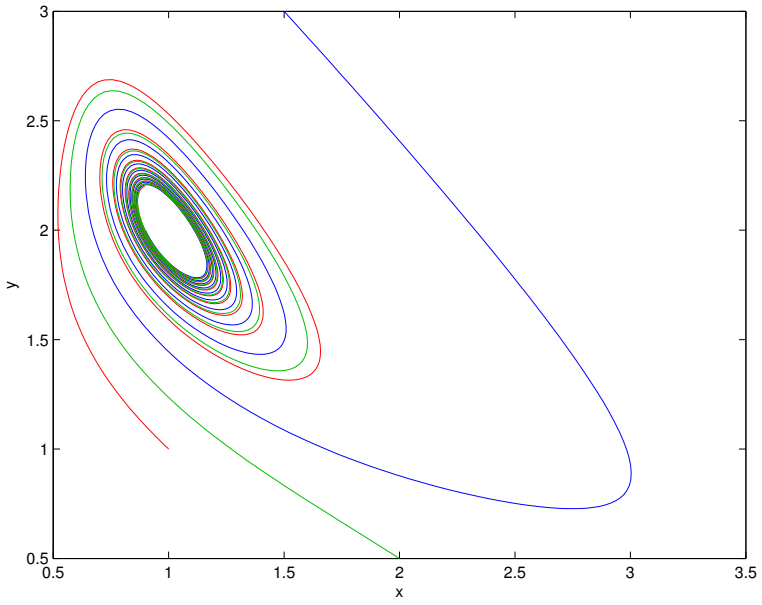


Figure 4.23: Phase portrait of the Brusselator for $a = 1$, $b = 2$ over $t \in [0, 60]$ for some different initial conditions.

In the initial phase, in the interval $0.4 \leq x \leq 0.9$ the complementary oscillation indicator is constantly zero, meaning that the eigenvalues are real-valued; however, judging by the phase portrait and intuition, this does not mean that the problem is non-oscillatory during this part of the solution trajectory. The oscillation indicator ω is more robust in the sense that it does not rapidly change, since it does not depend on the eigenvalues.

Also, the normality indicator has certain peaks, e.g. at $x = 4.80$ and $x = 10.95$ which does not seem to affect any of the other indicators nor cause any irregularities along the solution trajectory. This indicates that non-normality is not sufficient for oscillations nor for stiffness.

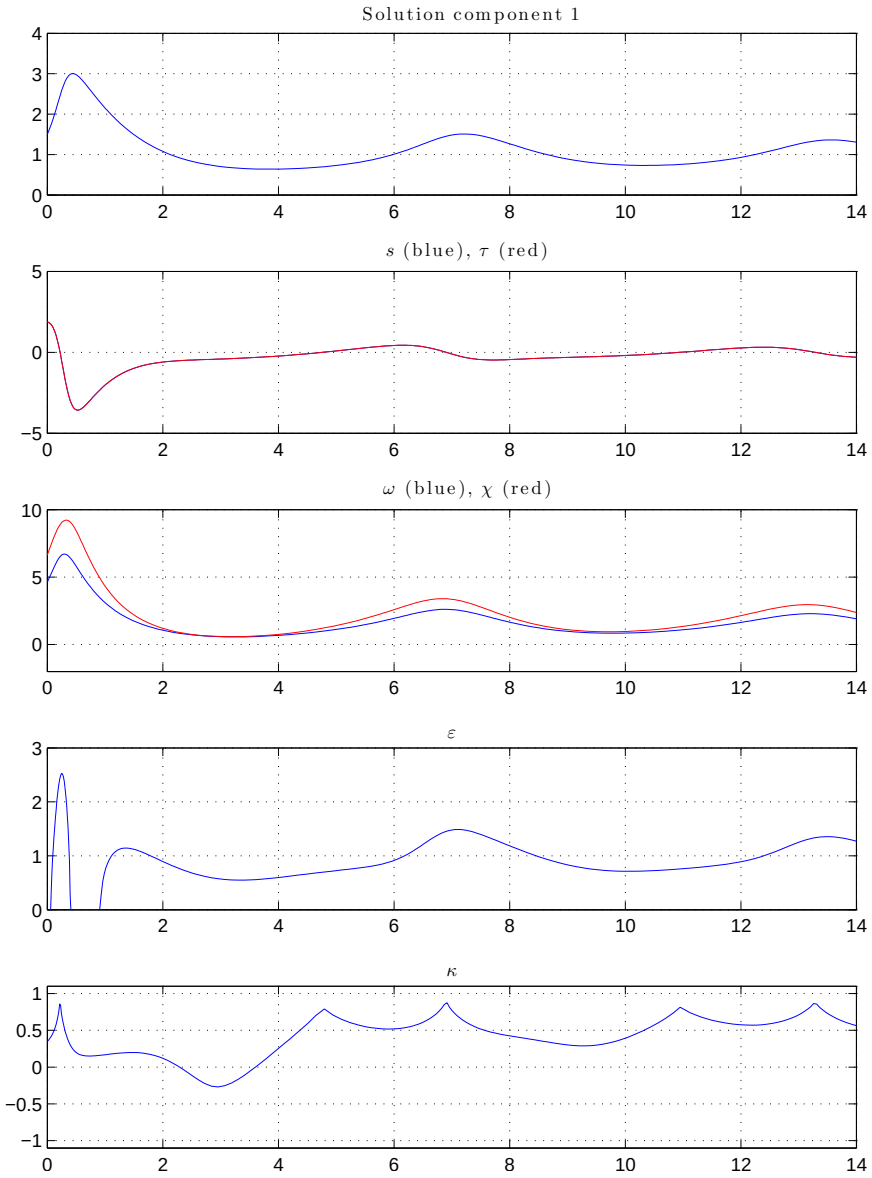


Figure 4.24: Indicators. Brusselator.

4.13 Pleiades problem

The Pleiades problem is a celestial mechanics problem modeling the orbits of seven stars in the plane. The system is nonstiff and consists of 28 equations (7 bodies each one having two spatial components and two velocity components) During the movement of the seven bodies several quasi-collisions occur, which creates some interesting dynamics, see Figure 4.25.

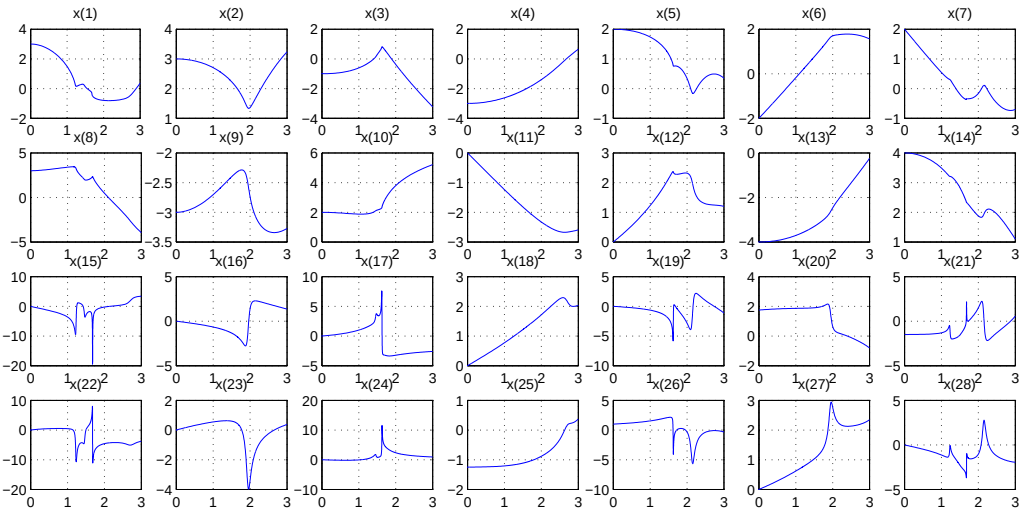
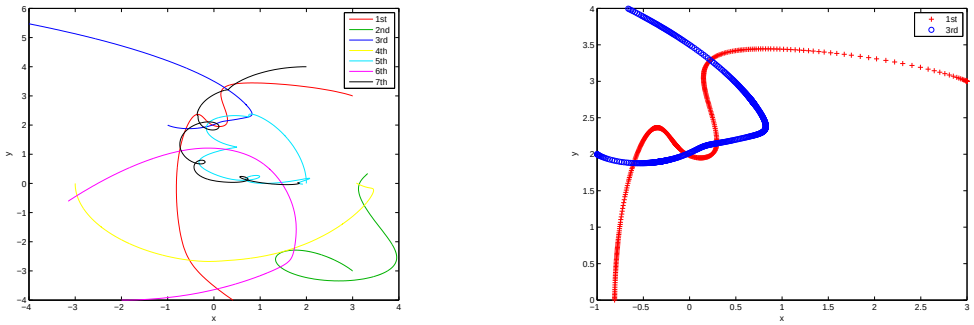


Figure 4.25: Solution components to Pleiades problem.

The star trajectories are shown in Figure 4.26a, and a zoomed in version of a quasi-collision in Figure 4.26b. The notable peaks in the oscillation indicators ω and ε seen in Figure 4.27 are directly correlated to the quasi-collisions.



(a) **(b)**

Figure 4.26: (a) All star trajectories for $t \in [0, 5]$, and (b) zoomed-in quasi-collisions.

It is interesting to see that the normality indicator is not constant, yet suggests that the problem is highly non-normal. Hence maximal non-normality is not necessarily a global property as seen in the stiff spring pendulum, but can also be a local.

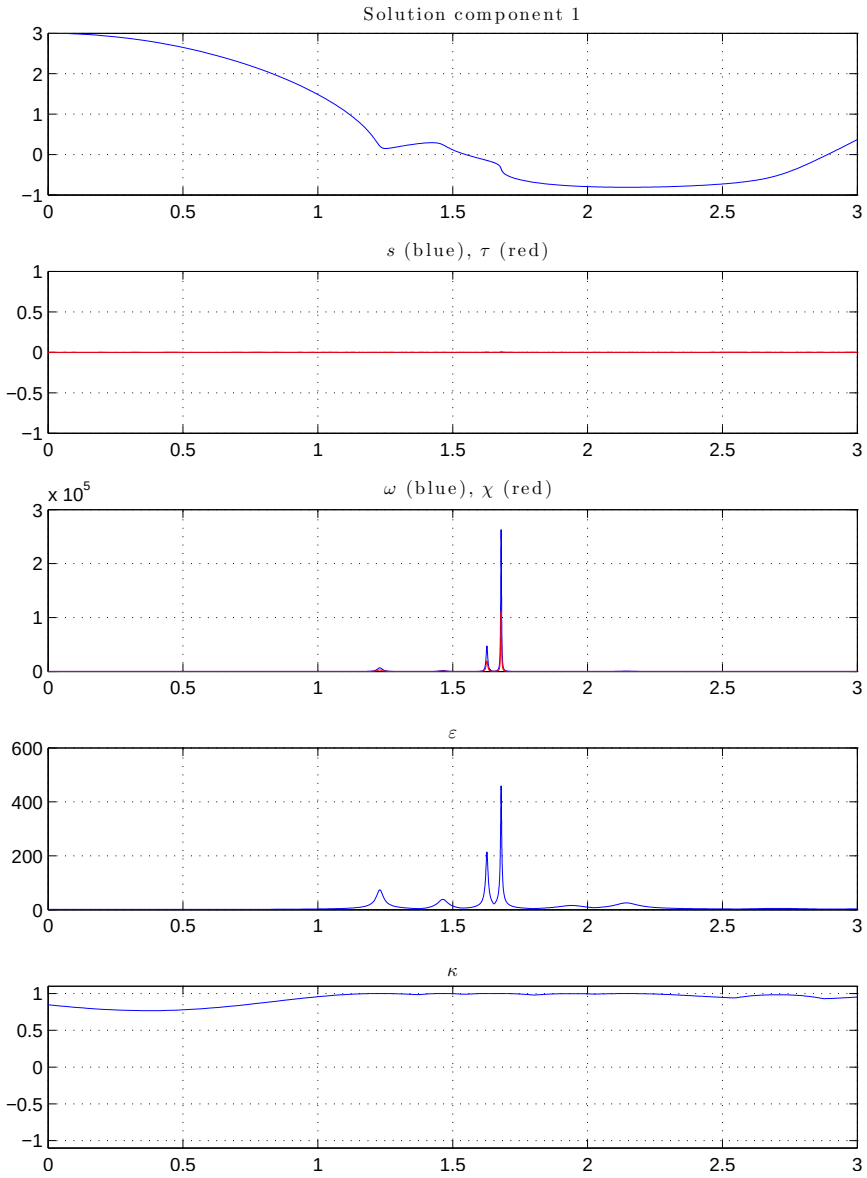


Figure 4.27: Indicators. Pleiades problem.

4.14 Robertson

The Robertson problem, like the Brusselator, is a model of an autocatalytic reaction, but differs since the reaction rate constants differ significantly from each other, leading to completely different dynamics, see Figure 4.28. The equations are

$$\begin{aligned}\frac{\mathrm{d}y_1}{\mathrm{d}t} &= -0.04y_1 + 10^4 y_2 y_3 \\ \frac{\mathrm{d}y_2}{\mathrm{d}t} &= 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2 \\ \frac{\mathrm{d}y_3}{\mathrm{d}t} &= 3 \cdot 10^7 y_2^2\end{aligned}$$

with the initial condition $\boldsymbol{y}_0 = (1, 0, 0)$ over $t \in [0, 10^{10}]$.

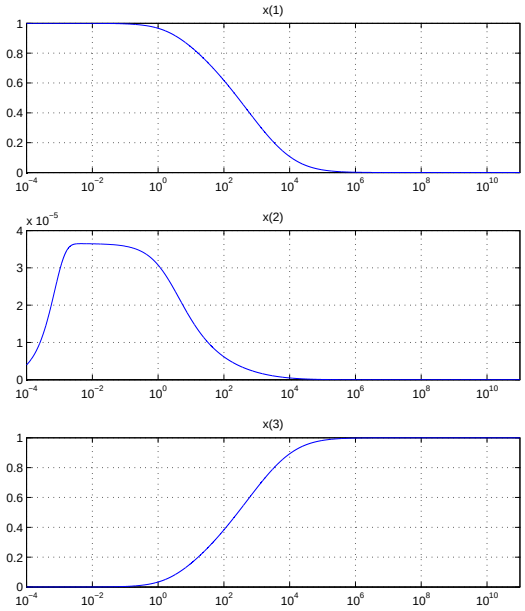


Figure 4.28: Solution components to Robertsons problem.

In Figure 4.29 we note that the problem gradually becomes stiffer, which is not the case in the Brusselator. In the interval $x > 10^5$ the oscillation indicator is large although the solution components have found an equilibrium, which is not a good property of the indicator. This tells us that a large skew-Hermitian part does not affect a system's properties of reaching an equilibrium. We also note that the normality indicator is almost constantly zero.

Note that the estimators deviate from the corresponding indicators over time, despite $\kappa \approx 0$ along the trajectory.

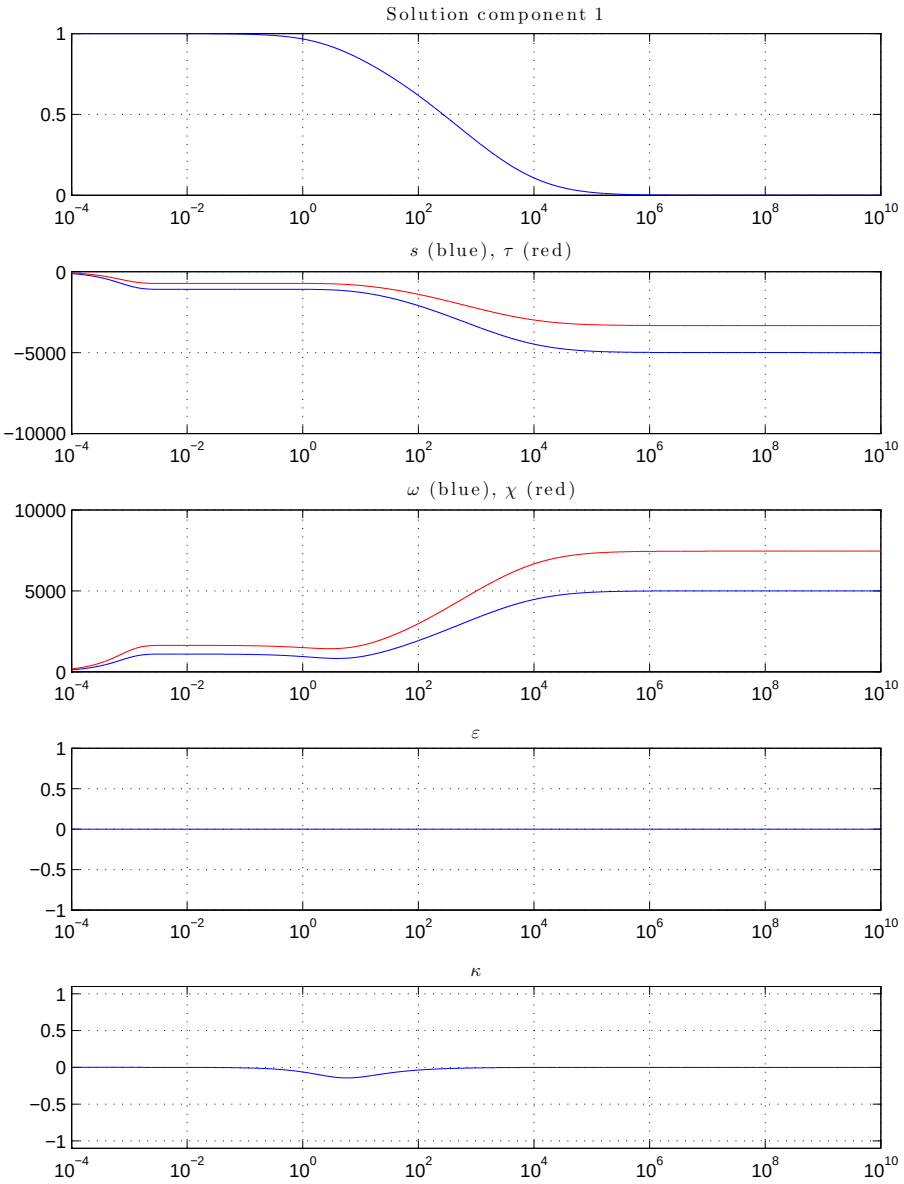


Figure 4.29: Indicators. The Robertson problem.

4.15 E5

The E5 problem is a chemical model for pyrolysis (thermochemical decomposition of organic materials) where six chemical compounds react, where two reactants are decoupled from the other (hence the dynamics for these are discarded). The system was originally posed on a smaller interval but several interesting properties occur for larger times, making it an interesting test problem. Also, the system is considered stiff. Solutions are presented in Figure 4.30.

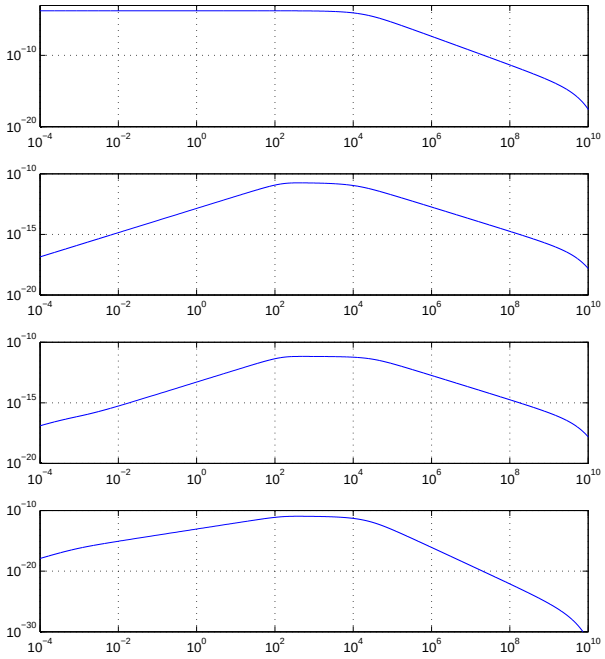


Figure 4.30: Solution components to E5 problem.

In Figure 4.31 we note an interesting change in the system behavior around $x = 10^4$, where all indicators except the complementary oscillation indicator ε change. This, again, suggests that changes along the solution trajectory is not only due to the eigenvalues. The non-normality indicator is correlating well with the oscillation indicator at this point. Note that the normality indicator increases after this point but the oscillation indicator is lower than before.

In comparison to the previous problem, The Robertson problem, where we observed that the estimators deviated from the indicators with time, the E5 problem exhibits the opposite characteristics; the estimators approach the indiactors with time. Again, this happens when $\kappa \approx 0$.

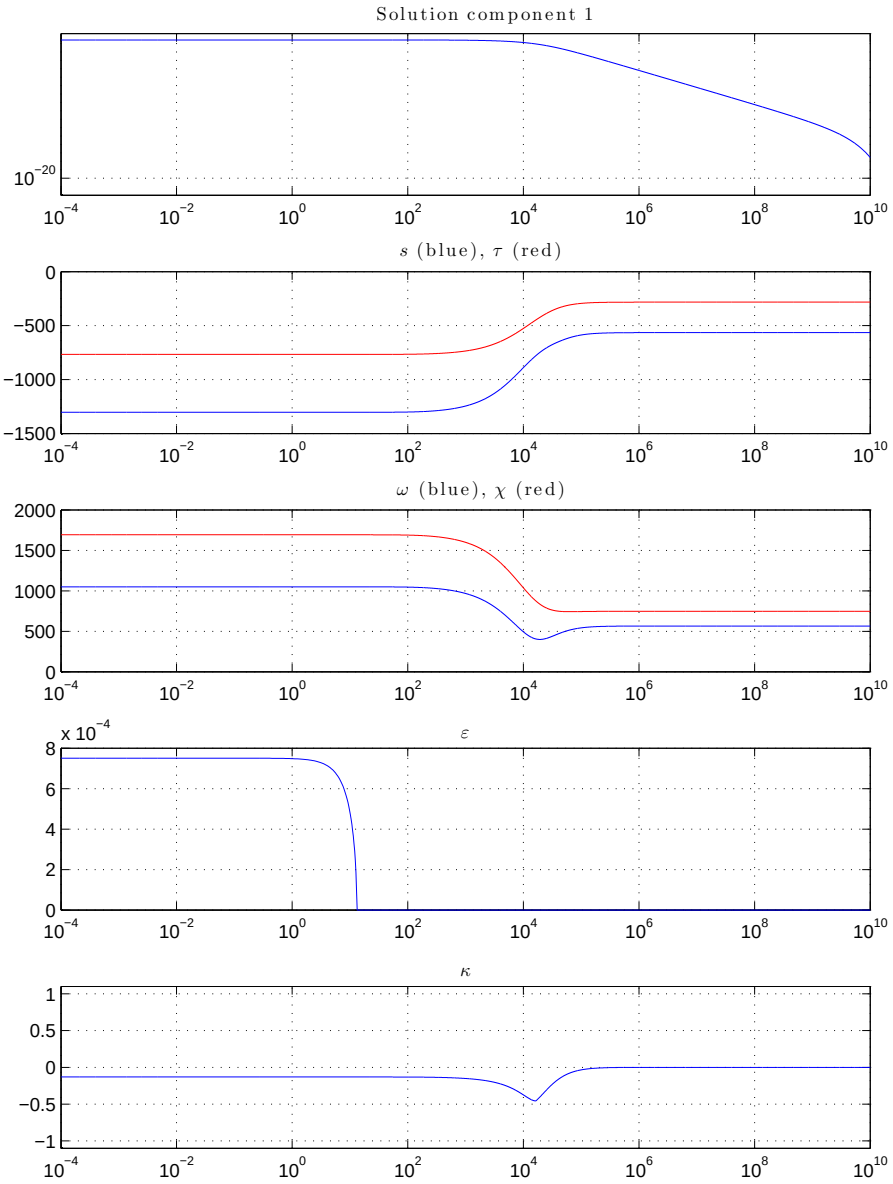


Figure 4.31: Indicators. E5 problem.

4.16 HIRES

The HIRES problem originates from plant physiology and models High Irradiance Responses (HIRES) of photomorphogenesis on the basis of phytochrome. The system contains eight reactants leading to a stiff system of eight ODEs. The solution components can be seen in Figure 4.32.

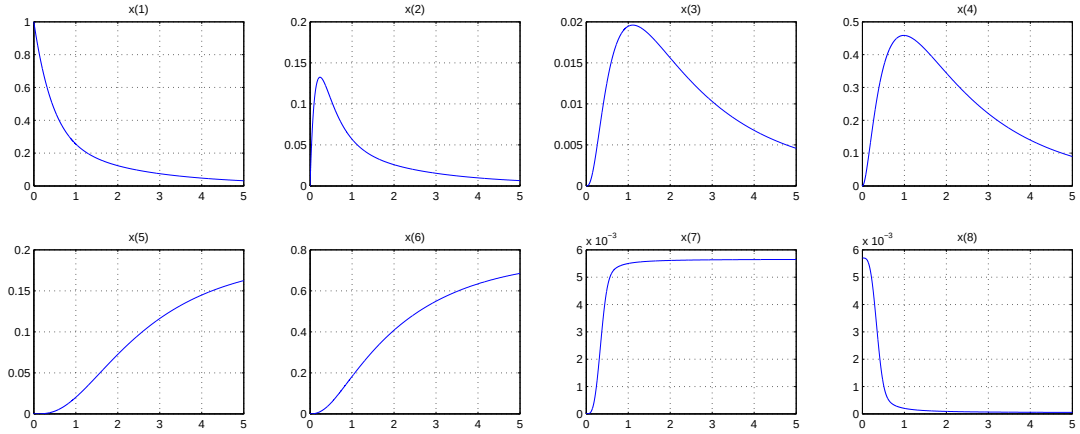


Figure 4.32: Solution components to HIRES problem.

As in the Robertson problem the eigenvalues are real-valued except in the initial phase, $0 < t < 0.1$, but again the skew-Hermitian part and the non-normality have a big impact on the solution trajectory. Note that the skew-Hermitian part becomes larger, but the normality indicator remains constant for $t \gtrsim 1$.

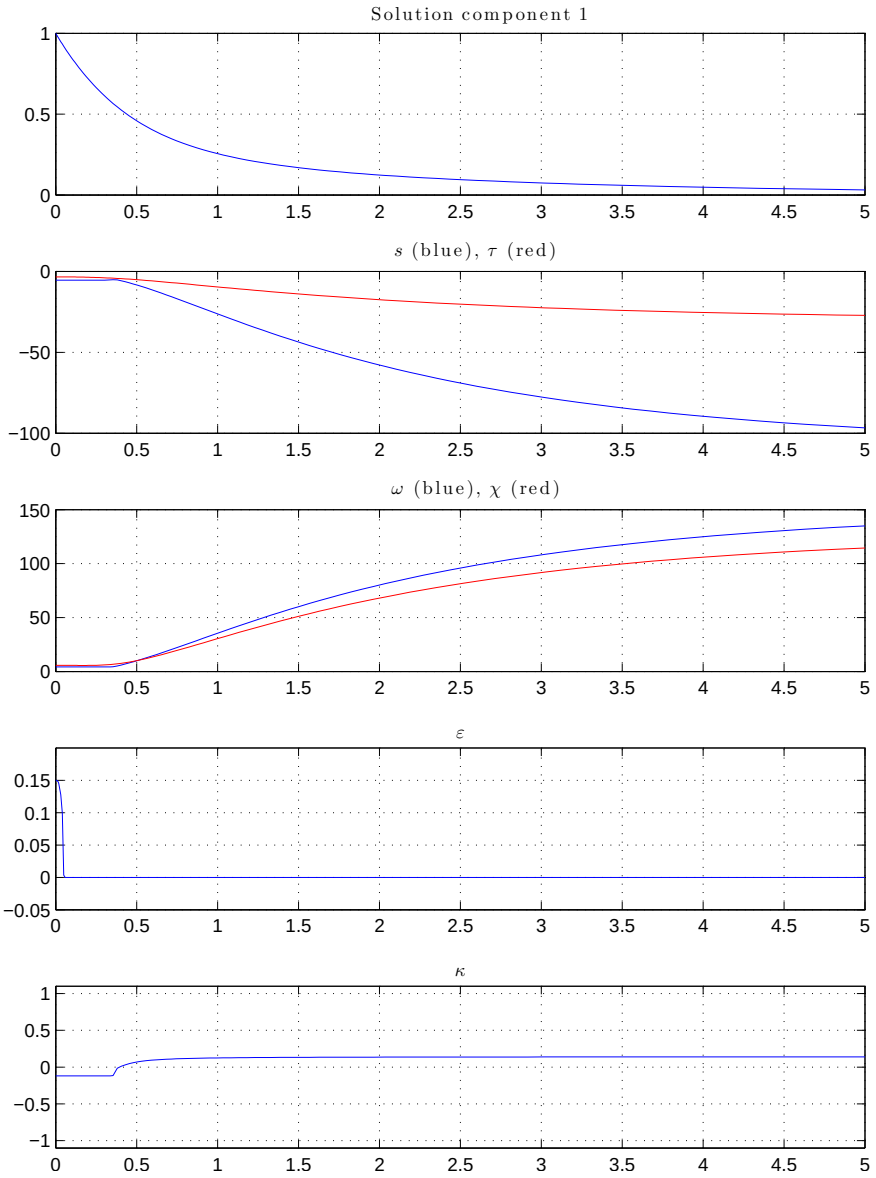


Figure 4.33: Indicators. HIRES problem.

Chapter 5

Discussion

5.1 Theory and numerical experiments

The numerical experiments show that lots of interesting dynamics, that are intuitively related to oscillations, may be observed without the eigenvalues having a large imaginary part, and that the eigenvalues alone give a poor understanding of these dynamics. In fact, the imaginary parts can be constantly zero, as in the Oregonator, Verwer's Pollution model, the Robertson problem and HIRES.

The correlation between the oscillation indicators ω and ε is good when the problem is normal; however, as the system becomes non-normal, or the change in normality is large, the correlation weakens. In such cases the oscillation indicator ω and the normality indicator are better correlated. Also, there are cases, e.g. the double pendulum, were all three correlate well.

Non-normality is not a sufficient condition for stiffness nor oscillations, as seen in Lotka-Volterra, the stiff spring pendulum and Airy's problem; however, it is prevalent in many cases such as the Oregonator and the Van der Pol oscillator. Furthermore, maximal non-normality can be a local phenomenon as well as a global phenomenon. The stiff spring pendulum is an example where it is global, whereas the double pendulum, the Oregonator, the Van der Pol oscillator and Pleiades problem are examples where it is only local.

The theory developed in Section 3.3 about the estimators did not give any strict bounds relating s and τ or ω and χ; however, special cases were treated, that suggested that they, most likely, would be in the same magnitude. The numerical experiments verify this, at least for practical problems as in Chapter 4, which is the target group of the indicators. The only exception is Verwer's Polltuion model, where τ underestimated s by a factor of ten.

It is interesting to notice that for $|\kappa| \lesssim \frac{1}{2}$ the estimators correlate well with the corresponding indicators, and deviate otherwise, as seen in the Duffing oscillatior; however, this is not a sufficient conditions, since there are cases where $\kappa \approx 0$ such

as the Robertson problem, the E5 problem and HIRES, where the distance between them grows (or shrinks) with time. Furthermore, many systems have the property $\kappa = 0$, locally, or even globally, along the solution trajectory. We discuss this in property further in Appendix A.

5.2 Stiffness and oscillations: A rigorous definition?

We have sought to extend the analysis of the stiffness indicator by looking at the skew-Hermitian part of a matrix A, instead of the Hermitian part. Indeed the oscillation indicator ω has some interesting properties; however, the question remains whether or not it actually measures *oscillations*. It is clear that the complementary oscillation indicator, based on the eigenvalues of A, is insufficient, and in many of the cases where it does not provide sufficient information the oscillation indicator does. For non-normal problems this is clear since ε is smaller, sometimes much smaller, than ω. There are also problems, such as Verwer's Pollution model, where ω is very large and $\varepsilon = 0$ along most of the solution trajectory. Judging by the solution components, is this problem really oscillatory? Perhaps not. But this leads us to a second interesting question: Are stiffness and oscillations two independent phenomena? The answer is no. Surely one can have a nonstiff highly oscillatory system, such as the stiff spring pendulum; however, judging by the analyses in this thesis, it is hard (impossible?) to construct a problem having the opposite characteristics, i.e. in non-normal systems stiffness comes oscillations, at least in the sense oscillations are defined here. Perhaps one may want to discuss quantities such as ω/s for systems (where $s \neq 0$) to quantify the conceptual, "visual" properties of oscillations.

5.3 Alternative definitions

Can one make an analogous definition of oscillations using the pseudospectrum instead? From Section 2.3.3 we know that there are many similarities between using the stiffness indicator and a pseudospectral method to characterize stiffness; however, it is not clear how the skew-Hermitian part translates into the latter framework. Since both methods are norm dependent, and the psuedospectrum for non-normal matrices (and operators) have nice properties, which could translate into a similar behaviour as the normality indicator, it is feasible that a similar approach can be made using this methodology. It is important to keep in mind that pseudospectra are computationally expensive and such methods probably would not be efficient for practical use, whereas the proposed estimators arise naturally in the framework of this thesis.

5.4 On the estimators

The estimators correlate very well with the corresponding indicators in the sense that they are always in the same magnitude in the test problems, with the exception of Verwer's Pollution model, where τ is off by a factor 10. Since τ is a lot cheaper to compute than s it is a strong candidate for a step size regulator. For the same reason it is unclear whether χ is a good regulator, since it is in the same order, $O(n^3)$, as ω.

In Lorenz equations and Chen's equation we saw that the estimator has a smoothing effect – they are constant when the corresponding indicators are not. Such an effect, although not constant, can be seen in the double pendulum as well. This is a nice feature for a step size regulator; however, in the stiff spring pendulum χ varied when ω did not, suggesting that it could cause the opposite effect. Step size regulation; however, is usually desired in stiff problems, and the stiffness estimator did not show such behavior – this is perhaps only a property of the oscillation estimator.

Chapter 6

Conclusions & future work

6.1 Conclusions

The aim of this thesis has been to establish a mathematically rigorous formalism for stiffness and oscillation, which can be applied to general systems of ordinary differential equations. Chapter 1 provided motivation of the topic, and in Chapter 2 the essential theory was presented. In Chapter 3 an overview of previous attempts was given, followed by the proposed indicators; the properties and the motivation behind them were thoroughly investigated analytically. The benefits and shortcomings of the indicators were evaluated in detail in Chapter 4, where they were tested on well-known problems. The overall discussion was presented in Chapter 5.

The key contribution of this thesis is the development of a typology for the classification of initial value problems. Furthermore, we problematize the commonly known concept of oscillations, and discuss the complexity of the terminology. We propose a rigorous mathematical definition that captures most of the phenomena known as oscillations. Analogously, a normality indicator is derived. The interactions and correlations between stiffness, oscillation and normality are discussed.

Lastly, we propose two estimators that replicate the behavior of the stiffness indicator and the oscillation indicator. Such estimators could be implemented to support step size regulators, which is of interest in many applications, e.g. stiff problems.

6.2 Future work

The typology proposed in this thesis is a strong fundament for future work. A natural step, after working with matrices, is to extend the theory to operators. A benefit of working with pseudospectral methods is that such extensions already exist, and is a current research topic. The transition between the finite dimensional case and the infinite dimensional case is straight-forward; however, as discussed in Section 2.3.3 psuedospectra do not always determine norm behavior accurately. Since stiffness and

oscillations depend on the topology induced by the norm, one might be better off avoiding these pseudospectal methods in characterizing these phenomena.

One of the benefits of working with the spectral norm, as in this thesis, is that it is a natural next step to extend to Hilbert spaces, which have nice properties. Such a transition to infinite dimensional cases is perhaps not as straight-forward as in working with psuedospectra. Note, also that some of the theorems involving upper and lower bounds rely on constants involving the dimension n, which will not hold in the infinite dimensional case.

Lastly, it would be of interest to investigate further applications, not restricting the theory to step size control, nor ODEs. Consider, e.g. the convection-diffusion equation $u_t = u_{xx} + u_x$, discretized in space using symmetric FDM. We then obtain the MOL ODE $\dot{u} = (T_{\Delta x} + S_{\Delta x})u$, where the symmetric Toeplitz matrix $T_{\Delta x}$ is negative definite, and the Toeplitz matrix $S_{\Delta x}$ is skew-symmetric. Hence

$$\mathrm{He}(T_{\Delta x} + S_{\Delta x}) = T_{\Delta x} \qquad \text{and} \qquad \mathrm{i}\,\mathrm{She}(T_{\Delta x} + S_{\Delta x}) = S_{\Delta x},$$

suggesting that applications can be found in splitting methods of PDEs.

Chapter 7

Bibliography

[1] A. Ammar and A. Jeribi. Measures of noncompactness and essential pseudospectra on Banach space. *Mathematical Methods in the Applied Sciences*, 37(3):447–452, 2014.

[2] C. Costara. Maps on matrices that preserve the spectrum. *Linear Algebra and Its Applications*, 435(11):2674–2680, 2011.

[3] J. Cui, C.-K. Li, and Y.-T. Poon. Pseudospectra of special operators and pseudospectrum preservers. *Journal of Mathematical Analysis and Applications*, 419(2):1261–1273, 2014.

[4] K. Ekeland, B. Owren, and E. Øines. Stiffness Detection and Estimation of Dominant Spectra with Explicit Runge-Kutta Methods. *ACM Transactions on Mathematical Software*, 24(4):368–382, 1998.

[5] H. Guebbai and A. Largillier. Spectra and pseudospectra of convection-diffusion operator. *Lobachevskii Journal of Mathematics*, 33(3):274–283, 2012.

[6] D. Higham and L. Trefethen. Stiffness of ODEs. *BIT*, 33(2):285–303, 1993.

[7] N. Higham and F. Tisseur. More on pseudospectra for polynomial eigenvalue problems and applications in control theory. *Linear Algebra and Its Applications*, 351:435 – 453, 2002.

[8] J. D. Lambert. *Computational Methods in Ordinary Differential Equations*. John Wiley & Sons, London, 1973.

[9] P. D. Lax and B. Wendroff. Difference schemes for hyperbolic equations with high order of accuracy. *Communications on Pure and Applied Mathematics*, 17(3):381–398, 1964.

[10] Y. Lee and B. Engquist. Variable Step Size Multiscale Methods for Stiff and Highly Oscillatory Dynamical Systems. 2013.

[11] F. Mazzia and C. Magherini. Test Set for Initial Value Problem Solvers, 2008.

[12] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Applied mathematical sciences: 44. New York ; Berlin : Springer-Vlg, cop. 1983, 1983.

[13] L. R. Petzold, L. O. Jay, and J. Yen. Numerical solution of highly oscillatory ordinary differential equations. *Acta Numerica*, pages 437–483, 1997.

[14] T. Ransford and J. Rostand. Pseudospectra do not determine norm behavior, even for matrices with only simple eigenvalues. *Linear Algebra and Its Applications*, 435(12):3024–3028, 2011.

[15] M. Seidel. On (N, ϵ)-pseudospectra of operators on Banach spaces. *Journal of Functional Analysis*, 262(11):4916 – 4927, 2012.

[16] G. Söderlind. The logarithmic norm. History and modern theory. *BIT Numerical Mathematics*, 46(3):631–652, 2006.

[17] G. Söderlind, L. Jay, and M. Calvo. Stiffness 1952–2012: Sixty years in search of a definition., 2014.

[18] L. N. Trefethen. Pseudospectra of Linear Operators. *SIAM Review*, (3):383, 1997.

[19] T. Ueta and G. Chen. Bifurcation analysis of Chen's equation. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 10(8):1917–1931, 2000.

[20] H. Wolkowicz and G. Styan. Bounds for eigenvalues using traces. *Linear Algebra and Its Applications*, 29(C):471–506, 1980.

Appendix A

On matrices with $\kappa = 0$

How come so many systems have $\kappa[A] = 0$. What does it mean?

$$\kappa_a[A] = \kappa_n[A] \quad \Rightarrow \quad \|A^*A + AA^*\|_2 - \|A^*A - AA^*\|_2 = \|A^*A\|_2.$$

Note that if A is normal then $A = 0$ (but then $\kappa[A]$ is not defined). By randomly generating matrices one easily finds this property among matrices having $\det A = 0$, e.g.

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \kappa[A] = 0,$$

but after further study this is neither necessary nor sufficient. Consider, e.g.

$$A = \begin{pmatrix} -2 & 2 & 0 \\ 0 & 0 & 4 \\ -2 & -2 & 0 \end{pmatrix},$$

where $\det A = -32$, also having this property. Note that the structure of A is quite special. It has both rows and column mutually orthogonal, so both A^*A and AA^* are diagonal. This is however, not the case for all matrices. In the 4×4 case

$$A_1 = \begin{pmatrix} 0 & -1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad A_2 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & -1 & -1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & -1 \end{pmatrix},$$

has the properties $\kappa[A_i] = 0$, $\det A_1 = 2$, $\det A_2 = 4$ but with $A_i^*A_i$ and $A_iA_i^*$ non-diagonal. To investigate this phenomenon further we make the following definition:

Definition A.1. Consider the class of matrices

$$\mathcal{A}_\mathcal{K} = \{A \in \mathbb{C}^{n\times n};\ \kappa[A] = 0\}\ .$$

If $A \in \mathcal{A}_\mathcal{K}$ we say that A is *semi-normal*.

We immediately get the following basic property:

Lemma A.1. *If $A \in \mathcal{A}_{\mathcal{K}}$, then $\lambda A \in \mathcal{A}_{\mathcal{K}}$, for every nonzero $\lambda \in \mathbb{C}$.*

Proof. Follows directly from the properties of the normality indicator. □

It is not in the scope of this thesis to explore the properties of the elements in $\mathcal{A}_{\mathcal{K}}$; however, we show some basic results for rank 1 matrices, inspired by the previous observations.

Theorem A.1. *If A is a rank 1 matrix, $A = u \otimes v$, with $\|u\|_2 = \|v\|_2 = 1$ and $|\langle u, v\rangle| = 1/\sqrt{2}$ then $A \in \mathcal{A}_{\mathcal{K}}$.*

Proof. Consider $A = u \otimes v$, where $u, v \in \mathbb{C}^n$ are nonzero vectors, such that $\|u\|_2 = \|v\|_2 = 1$ (otherwise we normalize them, as in Lemma A.1). If $u \parallel v$ the matrix A would be normal, and by Theorem 3.5 it follows that $\kappa[A] = -1$, hence $A \notin \mathcal{A}_{\mathcal{K}}$. Furthermore,

$$A^* = v \otimes u, \quad AA^* = u \otimes u, \qquad A^*A = v \otimes v,$$

hence $A^*Au = u$, $A^*Av = v$, $A^*Au = \langle v, u\rangle u$ and $AA^*v = \langle u, v\rangle v$. Let $\lambda > 0$ be an eigenvalue of $2\,\mathrm{No}\,A$ and $x = \alpha u + \beta v$ corresponding eigenvector, then

$$2\,\mathrm{No}\,Ax = \lambda x \quad \Leftrightarrow \quad [u \otimes u + v \otimes v](\alpha u + \beta v) = \lambda(\alpha u + \beta v),$$

giving

$$\alpha + \beta\overline{\langle u, v\rangle} - \lambda\alpha = 0 \quad \text{and} \qquad \alpha\langle u, v\rangle + \beta - \lambda\beta = 0,$$

since u and v are linearly independent. For $\alpha \neq 0$ or $\beta \neq 0$ we deduce that $\lambda = 1 \pm |\langle u, v\rangle|$ and $\|\mathrm{No}\,A\|_2 = 1 + |\langle u, v\rangle|$. Similarly, $\|\mathrm{Ano}\,A\|_2 = \sqrt{1 - |\langle u, v\rangle|^2}$ and

$$\kappa[A] = 0 \quad \Leftrightarrow \quad 1 = 1 + |\langle u, v\rangle| - \sqrt{1 - |\langle u, v\rangle|^2},$$

giving $|\langle u, v\rangle| = 1/\sqrt{2}$.

□

Theorem A.2. *If $A \in \mathcal{A}_{\mathcal{K}}$ and is of rank 1, with $\|u\|_2 = \|v\|_2 = 1$ and $|\langle u, v\rangle| = 1/\sqrt{2}$ then $A \oplus B \in \mathcal{A}_{\mathcal{K}}$ if $\|B\|_2 \leq 1$, $\|\mathrm{No}\,A\|_2 \leq (2 + \sqrt{2})/4$ and $\|\mathrm{Ano}\,A\|_2 \leq \sqrt{2}/4$.*

Proof. Properties of direct sum. □

From Theorem A.2 we may generate as large matrices as we wish in $\mathcal{A}_{\mathcal{K}}$, and it explains the special structure observed in some of the matrices; however, far from all matrices, have this property. We conclude this section with the following open problem.

Problem A.1. Determine necessary and sufficient conditions for a matrix $A \in \mathcal{A}_{\mathcal{K}}$.