UNIVERSITÀ DEGLI STUDI DI UDINE

Dipartimento di Scienze Matematiche, Informatiche e Fisiche

Corso di Laurea Magistrale in Matematica

Tesi di Laurea

Numerical computation of the basic reproduction number in population dynamics

Relatore:
Prof. Dimitri Breda

Laureando:
Francesco Florian

Correlatori:
Prof. Jordi Ripoll
Prof. Toshikazu Kuniya

ANNO ACCADEMICO 2017-2018

1 Abstract

This thesis focuses on the study and, mostly, the numerical computation, of the basic reproduction number, or R_0, a quantity defined in ecology and epidemiology as a mean to investigate what formally are the properties of stability of the zero solution of a linear system of equations.

The genesis of R_0 can be set in the early years of the twentieth century, when the concept was used without having been defined and formalized yet (as in [23]); the basic reproduction number has soon become an important tool for determining whether a (usually small) population can grow in a certain environment, or it is doomed to extinction.

In most modern models the population is structured, i.e. individuals' fertility and mortality are differentiated by some properties, like age, sex, or dimension; in those models the basic reproduction number is characterized as the spectral radius of an operator, called next generation operator.

Despite the importance of this quantity, and the number of works devoted to its applications in epidemiology, the only attempt to develop an algorithm for its numerical computation was carried out in [16].

The main contribution of this thesis is the development and implementation of an algorithm which is more general, and more accurate than the existing one at parity of computing resources.

A preliminary analysis of the convergence has also been attempted.

2 Sommario

Lo scopo di questa tesi è lo studio, e soprattutto il calcolo numerico, di R_0, o "basic reproduction number", una quantità definita in ecologia, ed epidemiologia come un mezzo per studiare quelle che formalmente sono le proprietà di stabilità della soluzione nulla di un sistema lineare di equazioni differenziali.

La nascita di R_0 può essere collocata agli inizi del ventesimo secolo, quando il concetto è stato usato senza essere definito, e prima che fosse ancora formalizzato (per esempio in [23]); R_0 è velocemente diventato un metodo importante per determinare se una popolazione (generalmente assunta piccola) può crescere in certe circostanze ambientali o è destinata all'estinzione.

Nonostante l'importanza di questo numero, e la quantità di articoli dedicati alle sue applicazioni in epidemiologia, l'unico tentativo di sviluppare un algoritmo per il calcolo numerico è quello presentato in [16].

Il contributo principale di questa tesi è lo sviluppo e l'implementazione di un algoritmo più generale e, a parità di risorse utilizzate, più accurato di quello esistente.

È anche stata tentata un'analisi di convergenza preliminare.

Contents

Chapter 1

Introduction

The traditional approach for the study of asymptotic behavior in population dynamics relies on computing the so called Malthusian parameter ([2]).

Since the work of Ross on malaria diffusion [23], a different approach has emerged, which is based on the basic reproduction number, or R_0. It is defined as the expected number of newborns generated by a single typical individual during its entire life; in epidemiology, the focus is on the disease, so "newborns" means newly infected individuals, and "lifetime" is the infectious period; that is, R_0 is the number of secondary cases produced by a "typical" infected individual during its infectious period, assuming a completely susceptible population.

This approach has particularly grown in importance with the formulation of epidemic models, and age structured populations, where computing the Malthusian parameter is often a difficult task.

For age structured population models in epidemiology R_0 is characterized as the spectral radius of a linear operator, called next generation operator, which we will define in Section 2.3.2, and this characterization is used for the explicit computation, in the few cases in which this can be done.

Despite the importance of the basic reproduction number for the study of stability in population models, and its extensive use in literature (see e.g. [6, 11, 12, 13]), the issue of the numerical computation of R_0 has only arisen in recent times, and it is a nearly unexplored area of research.

Given the importance of R_0, we investigate a method for its numerical computation, based on the approximation of the next generation operator, and we apply it to two biological models; the first is from [16]: in that article the first method for the computation of R_0 has been developed; it has been included because it shows the approximation of the next generation operator as a method for computing R_0, and this was our starting point for developing a new class of methods for the approximation of that operator; as a consequence an ongoing collaboration with the author has started; the second is an extension of the "cell population" example in [2]; it is examined because a collaboration project on that model has been proposed by prof. Ripoll.

1.1 Approximation

In both models the state space is an infinite-dimensional subspace of $L^1([\,0, l\,])$, so the next generation operator cannot be explicitly represented on a computer in order to numerically compute R_0. Moreover, an explicit representation of the next generation operator is not available in general; this means that it cannot be directly approximated using a matrix.

However, for the model to be defined, two operators B and M, called the birth and mortality operators, must be available; they define the next generation operator as $K \coloneqq BM^{-1}$.

Under some hypothesis on K, which we will make explicit, the spectral radius of the next generation operator can then be approximated by the spectral radius of the matrix $B_n M_n^{-1}$, where B_n and M_n are suitable approximations of B and M respectively; i.e., one of the following (equivalent) problems has to be solved:

$$B_n M_n^{-1} \Phi = \lambda \Phi$$
$$B_n \Psi = \lambda M_n \Psi;$$

the largest obtained λ gives then an approximation of R_0.

Two main approaches have been used in this work, which actually are based on the same method, i.e. dividing the domain $[\,0, l\,]$ in intervals and approximating continuous functions on each interval using polynomials; they are nevertheless expected to yield different results.

If a fixed (small) number of intervals are used we speak of spectral approach, where convergence is reached increasing the polynomials degree; this method is expected to yield spectral accuracy for analytic functions, i.e. the error should decrease as

$$O\left(\left(\frac{n}{q}\right)^{-n}\right),$$

where q is a constant and n is the polynomial degree [24, Chapter 4].

If a (small) polynomial degree is fixed, and the convergence is obtained increasing the number of intervals, we speak of finite elements approach; this method is expected to yield an error of order $O(n^{-k})$ for C^k functions, where k is also the degree of the polynomials, and n is the number of intervals.

The choice typically depends on the regularity of the eigenvector φ corresponding to the eigenvalue R_0: since for functions in C^k the convergence is limited at most to $O(n^{-k})$, unless the operator M and B are smoothing, fixing k as the polynomial degree gives a faster method while maximizing the order of convergence; this is mostly useful if M_n and B_n are block-diagonal, since the highest computational cost of this method comes from solving the eigenvalues problem (and inverting M_n, if solving the standard eigenvalues problem instead of the general eigenvalues problem), which has cubic complexity in the matrix dimension.

Also, note that it is not necessary that every $\varphi \in X$ is regular to obtain an high order of convergence, only the eigenvector φ relative to R_0: indeed, if it is regular,

computing the spectral radius of K or that of its restriction to the functions at least as regular as φ yield the same result.

Since in the models considered in this thesis this eigenvector is usually (piecewise) analytic, a small, fixed number of intervals yields faster convergence; we will not restrict to this choice, however, because our aim is to compare these different instances of the general method.

The code, written in `c++`, is available at https://github.com/f-florian/thesis and https://github.com/f-florian/thesis-differential[1].

1.2 Conclusions

The method developed for this thesis produced the expected results on some test cases in which the exact result was known, using the spectral approach.

It was then tested on some data for which the solution was unknown, here it exhibits the expected convergence order, and produces results that are thus most probably correct, both with the spectral and the finite elements approach, with the spectral approach being the most accurate, in accordance with the analysis in the previous section, at parity of computing time, which is of order $O(n^3)$, where n now is the polynomials degree times the number of intervals.

The convergence could also be proved for some special cases.

As a side effect, we noted that a speedup can be obtained writing the code in `c++` instead of `Matlab`.

[1] This part of the work was carried out as an activity of the Computational Dynamics Laboratory (http://cdlab.uniud.it/).

Chapter 2

An introduction to R_0 and population dynamics

The first problem regarding population dynamics is probably that of "Fibonacci rabbits", appeared in 1228; unlike most modern models, it is a discrete-time problem. For the second model, we have have to wait for Euler in 1748[8], whose problem is again a discrete-time one: "If the number of inhabitants of a certain province should increase by the thirtieth part each year, moreover at start there were 100 000 people in the province, the number of inhabitants is sought after 100 years."

2.1 Malthusian parameter

Euler model, however, is most known through the work of Malthus, who in 1978 published an essay claiming that an exponential growth of the population soon becomes non sustainable in terms of subsistence means (food) produced [19]. Malthus's analysis, like Euler's, is performed in a discrete-time settings. He assumes, from the observation of what happened in the colonization of the United States of America, that when the births of the population are not controlled, the population doubles every twenty-five years, thus growing in a geometrical ratio; he also states that the increase in food production is at most arithmetic; also, some chapters later, he concludes that the population growth cannot be estimated by the birth/death ratio, but rather by the food production increase, which is the real limiting factor.

Besides its social, political and economical implications, Malthus's essay shows it the initial part, in which he studies the same problem as Euler, the first use of what is now called "Malthusian parameter"; though Malthus himself considered that model unfeasible on long time scales, it can describe quite well the evolution on short times, and the Malthusian parameter has then become the traditional tool for the study of continuous-time dynamics.

It was soon used in continuous-time settings with reference to exponential solutions $x(t) = \exp(rt)$, where r is a constant, precisely the Malthusian parameter; since then, its meaning has extended to include non-scalar equations and general

solutions: it is now defined as the exponent of a positive dominant exponential solution.

For the simplest case, consider the system

$$p'(t) = Ap(t)$$

where $p(t) \in \mathbb{R}^n$ and A is a $n \times n$ matrix; then the Malthusian parameter is the spectral bound of A:

$$r = s(A) := \max_{\lambda \in \sigma(A)} \Re\lambda.$$

A similar definition can be given for some classes of non-constant matrices A (see [13]); in that case the malthusian parameter is the exponential parameter of a periodic times exponential dominant solution.

2.2 Malthusian parameter vs basic reproduction number

The basic reproduction number, also called R_0, is a parameter arising as an alternative to the Malthusian parameter.

A simple example is worth to show how R_0 is obtained and how these two parameters relate:

$$\begin{cases} x'(t) = \beta x(t) - \mu x(t) \\ x(0) = x_0. \end{cases} \tag{2.1}$$

Its solution is $x(t) = \mathrm{e}^{(\beta-\mu)t}x_0$; the Malthusian parameter is thus $r = \beta - \mu$.

But Eq. (2.1) can be also solved using the variation of constants formula, taking β as the non homogeneous part; the result is

$$x(t) = \mathrm{e}^{-\mu t}x_0 + \int_0^t \mathrm{e}^{-\mu(t-s)}\beta x(s)\mathrm{d}s; \tag{2.2}$$

setting $b(t) = \beta x(t)$ and $\tau = t - s$ this becomes the so called "renewal equation", introduced by Lotka in 1925 and then extensively used in epidemiology:

$$b(t) = \beta \mathrm{e}^{-\mu t}x_0 + \beta \int_0^t \mathrm{e}^{-\mu\tau} b(t-\tau)\mathrm{d}\tau. \tag{2.3}$$

Since we are only interested of in the newborns, we neglect the first term; then let $t \to +\infty$ and $b = 1$ in the rhs to get

$$R_0 = \beta \int_0^t \mathrm{e}^{-\mu\tau}\mathrm{d}\tau = \frac{\beta}{\mu},$$

which is, in this simple scalar case, the basic reproduction number.

It should be noted here that R_0 depends on the choice of β and μ, which is not unique, since only $\beta - \mu$ is determined by the model; however in all cases the extinction threshold is $R_0 = 1$, corresponding to $r = 0$.

2.3 Basic reproduction number

2.3.1 A survey on the genesis of the basic reproduction number

The basic reproduction number is now considered an important tool in demography, and it is the most important quantity in the study of epidemics and in comparing population dynamical effects of disease control strategies: as is said in [12], the concept of R_0 is very powerful in epidemiology, as it is directly related to the amount of control effort needed to eliminate an infection from a population.

It is defined as the expected number of newborns generated by a single "typical" individual during its entire life, or, in epidemiology, the number of secondary cases produced by an infected individual during its infectious period, assuming a completely susceptible population.

The basic reproduction number was first used in demography: it is first mentioned in 1886, by the Director of the Statistical Office of Berlin, Richard Böckh, who talks about the number of females born to one female during her entire reproductive period (see [11, 12]).

As already said, the first person to practically use and popularize this concept was Ross, a medical doctor and a colonel in the British Army in India [12, 23]. He discovered in 1898 that (bird) malaria was transmitted by mosquitoes and that malaria was not caused by bad air from marshes as was previously believed. He received the Nobel Prize for this discovery in 1902.

In 1911 Ross argued that local eradication of malaria was possible by decreasing the density of mosquitoes in the area [23]; this was in contrast to the general opinion at the time that fighting mosquitoes was a difficult route to eradicate malaria, because it would be practically impossible to kill all mosquitoes locally and therefore impossible to stop transmission of malaria.

Ross identified the main factors in malaria transmission and calculated the number of new infections arising per month as the product of these factors. He then showed that a critical mosquito density exists, such that malaria will eventually extinct if the density is below the threshold, therefore with no need to kill every mosquito. Ross referred to his discovery as his "Mosquito Theorem". This statement was later empirically verified in India, with the discovery of neighboring areas with malaria, where the mosquito density was above the threshold, and without malaria where the mosquito density was below the critical value.

Ross himself did not interpret his statement as the number of secondary cases arising from one infected individual being greater or smaller than 1. It was then Lotka who in 1919, replying to the work of Ross, interpreted the threshold that way (see [12]): if

$$F(a) \coloneqq exp(-\int_0^a \mu(\alpha)\mathrm{d}\alpha)$$

is the survival probability to age a (where μ is the mortality rate), Lotka wrote

$$r > 0 \iff \int_0^\infty b(a)F(a)\mathrm{d}a > 1,$$

where r is "the rate of natural increase per head", i.e. the Malthusian parameter.

The same Lotka in 1925, together with Dublin, and then Kuczynski alone in 1928, formalized the concept in the demographic context, showed how to calculate it and introduced the notation R_0, always for the case of a single scalar equation: in this case, R_0 is obtained as

$$R_0 := \int_0^\infty b(a)F(a)\mathrm{d}a,$$

where $b(a)$ is the average number of offspring that an individual will produce per unit of time at age a.

A greater generation comes from the work of Kermack and McKendrick, in 1927, which generalizes that of Ross's, since it does not assume that the infectivity of an individual is constant over time. The Kermack and McKendrick model uses the following hypotheses [14]:

- One (or more) infected person is introduced into a community of individuals, more or less susceptible to the disease in question.
- The disease spreads from the affected to the unaffected by contact infection.
- Each infected person runs through the course of his illness, and finally is removed from the number of those who are sick, from recovery or death.
- The chances of recovery or death vary from day to day during the course of the illness.
- The chances that the affected can send infection to the unaffected are likewise dependent on the stage of the disease.
- As the epidemic spreads, the number of unaffected members of the community becomes reduced.
- Since the course of an epidemic is short in relation to the life of an individual, the population can be considered as remaining constant, except in as far as it is modified by deaths due to the epidemic disease itself.

Finally, in 1952 Mcdonald publishes the paper "The analysis of equilibrium in malaria" in the Tropical Diseases Bulletin, which focuses on malaria; in one paragraph of his appendix, however, he takes a more general view of epidemic phenomena, in which he defines what he calls "basic reproduction rate" (of malaria), as "The number of infections distributed in a community as the direct result of the presence in it of a single primary non-immune case" [12].

2.3.2 Basic reproduction number as spectral radius of an operator

We have to wait for recent times and age structured population models in epidemiology, to see this concept applied to non-scalar equations (e.g. [6]); by non-scalar we here mean that the state space is either infinite-dimensional, or finite-dimensional with dimension greater that 1.

In these models R_0 maintains the same definition, but is characterized as the spectral radius of a linear operator, called "next generation operator", first introduced by Dublin and Lotka (they called it "ratio in successive generations", see [12]); this means that in the scalar case the next generation operator K is given by $Kx = R_0 x$, so we only used implicitly: since a stationary solution produces a constant birth rate, setting $b = 1$ in Eq. (2.3) means precisely taking th spectral radius of K.

For the non-scalar case, a stationary population still leads to constant birth and mortality rates; but they are expressed in terms of linear operators B and M; then, if $\{\mathrm{e}^{-Mt}\}_{t\geq 0}$ is the semigroup generated by $-M$,

$$x(t) = \mathrm{e}^{-Mt}x_0 + \int_0^t \mathrm{e}^{-M(t-s)}Bx(s)\mathrm{d}s;$$

setting $b(t) = Bx(t)$, $\tau = t - s$, neglecting the contribute of the initial population (i.e. $x_0 = 0$)

$$b(t) = B\int_0^t \mathrm{e}^{-M\tau}b(t-\tau)\mathrm{d}\tau,$$

then, considering b constant:

$$b = B\int_0^t \mathrm{e}^{-M\tau}\mathrm{d}\tau\ b = BM^{-1}\left(1 - \mathrm{e}^{-Mt}\right)b,$$

and letting $t \to \infty$ we obtain

$$b = BM^{-1}b;$$

the next generation operator is then defined as BM^{-1}.

In both the scalar and non-scalar cases, thanks to the definition of the basic reproduction number, the following criterion holds: $R_0 < 1$ leads to extinction; on the contrary $R_0 > 1$ means that population can grow (or the disease can spread through the population), as long as the linear model is a good approximation[1].

2.4 Some examples of use of R_0

The models in this section are taken from [6]; they involve various "structuring variables" such as age, disposition and sexual activity, which we collective call *state*, elements of a state space Ω.

Computing the spectral radius of a linear operator is, in general, a difficult task; nevertheless, some special cases exist, in which that task is simple.

[1] See [6] for the proof in the case of a disease.

2.4.1 Separable mixing rate

The first case is that of an operator of one dimensional range; this has an ecologic interpretation, namely that the state distribution of the newborns (or the infected) does not depend on the parent state. In epidemiology this case is called "separable mixing rate" or "separable infectivity and susceptibility", or "(separably) weighted homogeneous mixing".

The calculation is as follows: assume that the next generation operator is defined by

$$K\phi(\xi) = \int_\Omega A(\xi,\eta)\phi(\eta)\mathrm{d}\eta,$$

where

$$A(\xi,\eta) = a(\xi)b(\eta),$$

with a and b positive and ξ, η are the state of the child and the parent respectively; then in the definition of K, a can be taken outside the integral, so

$$K\phi(\xi) = S(\xi)a(\xi)\int_\Omega b(\eta)\phi(\eta)\mathrm{d}\eta.$$

Indeed, the dimension range of K is one, and there is thus only one (eigenvalue, eigenvector) pair with a nonzero eigenvalue, which is

$$(\int_\Omega b(\eta)S(\eta)a(\eta)\mathrm{d}\eta, Sa).$$

So the spectral radius is that eigenvalue,

$$R_0 = \int_\Omega b(\eta)S(\eta)a(\eta)$$

because the rhs is positive.

2.4.2 Separable mixing rate with enhanced infection within each group

In this case R_0 cannot be explicitly calculated, but the threshold related to $R_0 = 1$ can.

The individuals are characterized by a "group", they preferentially mix with other individuals in the same group, and the mixing with other groups is weighted, with the weights only depending on the groups. If the state variables are constant over time (at least in the time scale of the infection), then we can write K as

$$K\phi(\xi) = S(\xi)\left(c(\xi)\phi(\xi) + a(\xi)\int_\Omega b(\eta)\phi(\eta)\mathrm{d}\eta\right)$$

where $c(\xi)\phi(\xi)$ is the initial offspring produced within the same group; so we can state the eigenvalue problem as

$$(\lambda - S(\xi)c(\xi))\,\phi(\xi) = S(\xi)a(\xi)\int_\Omega b(\eta)\phi(\eta)\mathrm{d}\eta$$

or

$$\phi(\xi) = \frac{S(\xi)a(\xi)}{\lambda - c(\xi)S(\xi)}\int_\Omega \beta(\eta)\phi(\eta)\mathrm{d}\eta$$

multiplying both sides by $b(\xi)$ and integrating over Ω yields the characteristic equation

$$\int_\Omega \frac{S(\xi)a(\xi)}{\lambda - c(\xi)S(\xi)} = 1. \tag{2.4}$$

Thus the criterion for R_0 states: $R_0 > 1$ if $c(\xi)S(\xi) > 1$ for some $\xi \in \Omega$, that is, the disease can be maintained due to one single group; or, since the lhs in Eq. (2.4) is a decreasing function of λ, the largest value for λ is greater than one if

$$\int_\Omega \frac{S(\xi)a(\xi)}{1 - c(\xi)S(\xi)} > 1,$$

that is, the disease is maintained by all the groups collectively.

2.4.3 Multigroup separable mixing

A natural generalization of Section 2.4.1 is to assume that the dimension of the range of K is finite.

We limit to the cases in which this has a biological interpretation.

Let $\Omega = \bigcup_{0\le i\le n}\{i\}\times\Omega_i$ for some $n \in \mathbb{N}$, and $\xi := (i, \xi_i)$, where $\xi_i \in \Omega_i$.

Now assume, similarly to what we did in Section 2.4.1, that

$$\int_0^\infty A(\tau, (i,\xi_i), (j,\xi_j))\mathrm{d}\tau = a_i(\xi_i)b_{i,j}(\xi_j).$$

Then

$$K\phi(i,\xi_i) = S(i,\xi_i)a_i(\xi_i)\sum_{j=0}^{n}\int_{\Omega_j} b_{i,j}(\xi_j)\phi(j,\xi_j)\mathrm{d}\xi_j.$$

For ϕ to be an eigenvector is thus necessary that $\phi(i,\xi_i) = \sigma_i S(i,\xi_i)a_i(\xi_i)$, where σ is an eigenvector of the matrix $M = (m_{i,j})$ such that

$$m_{i,j} = \int_{\Omega_j} b_{i,j}(\xi_j)\phi(j,\xi_j)\mathrm{d}\xi_j$$

As a consequence, R_0 is the dominant eigenvalue of M.

2.4.4 Sexually transmitted diseases: heterosexual transmission only

The state is now sex; that is $\Omega = \{0\} \times \Omega_0 \cup \{1\} \times \Omega_1$, where we conventionally set 0 for males and 1 for females. Adopting the separable mixing rate assumption (i.e., the state of the infected does only depend on the sex variable, recall Section 2.4.3), and neglecting homosexual transmission we get the matrix

$$M = \begin{pmatrix} 0 & m_{1,2} \\ m_{2,1} & 0 \end{pmatrix}$$

where

$$m_{1,2} = \int_{\Omega_2} b_{1,2}(\xi_2) S_2(\xi_2) a_2(\xi_2) \mathrm{d}\xi_2,$$
$$m_{2,1} = \int_{\Omega_1} b_{2,1}(\xi_1) S_1(\xi_1) a_1(\xi_1) \mathrm{d}\xi_1.$$

thus $R_0 = \sqrt{m_{1,2} m_{2,1}}$, the spectral radius of M.

Chapter 3

Models

Here we consider two biological models, as said in the introduction, for which we want to compute R_0.

The first is taken from [16] and describes the evolution of a generic "influenza-like" disease in the Japanese population; the second is taken from [2] and describes a population of bacteria living in the intestine of an animal.

Both models can be seen as instances of a more general problem which describes a type I structured population model as defined in [2].

3.1 General framework

3.1.1 Age structured populations

We study age structured population models: at any time an individual is characterized by a state, i.e. a value $s \in Z$ where Z is a subset of $\mathbb{R}^q$ for some q. The population at time t is thus described by $p(s,t)$, which is actually a population density, unless Z is a discrete set; s is called *structuring variable*.

The key point here is that the evolution processes can also depend on s; these processes include birth and death rates as usual, plus the description of when an individual's state changes, which is called transition process.

Typical structuring variables are age (which may be related to body size) and space, which are both continuous. Examples of discrete variables would be a simplified epidemic model, in which the individuals are divided in healthy, infected and recovered (it is simplified because the properties of the infected individuals usually depend on the time since the infection began); and a population whose properties depends not on the exact position of each individual, but rather on which region they are in; since individuals can move between regions, these cannot be treated as independent problems. The famous problem of "Fibonacci rabbits" can also be seen as a structured model, even if it would make the treatment harder instead of simpler; the structuring variable depends on age and is one of 0, i.e. age is 0 and 1, i.e. age is at least 1: pairs with age greater or equal to 1 generate a new pair of age

0, while pairs of age 0 don't; the evolution is then described by

$$\begin{cases} p(0, t+1) = p(1, t) \\ p(1, t+1) = p(1, t) + p(0, t) \end{cases}$$

or also

$$p(t+1) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} p(t).$$

3.1.2 Definitions

We now define a few concepts that we will use in what follows; these definitions are taken from [7], [18] and [1]

Definition 3.1 (Banach Lattice). Let *lattice* denote a partial order where every two elements have a common supremum.

Let X be a Riesz space, i.e. a partially ordered vector space whose order is a lattice. A norm $\|\cdot\|$ on a Riesz space is a lattice norm if

$$|x| \geq |y| \implies \|x\| \leq \|y\|.$$

A complete Riesz space equipped with a lattice norm is known as a Banach lattice.

Definition 3.2 (Strongly continuous semigroup). Consider a family $\mathcal{F} \coloneqq (T(t))_{t \geq 0}$ of bounded linear operators on a Banach space X. We say that $\mathcal{F}$ is a strongly continuous semigroup (or C_0-semigroup) if

- it is a semigroup w.r.t. the composition operation, i.e.

  $$T(t+s) = T(t)T(s) \ \forall s, t \geq 0$$

  and

  $$T(0) = I$$

- the maps $\xi_x \colon t \mapsto \xi_x(t) \coloneqq T(t)x$ are continuous, as maps from $\mathbb{R}^+$ into X for any $x \in X$.

Definition 3.3 (Positive function). Let $X \coloneqq L^p(\Omega)$ a Banach space, $f \in X$; then f is said to be a positive function if $0 \leq f(s)$ a.e. on Ω.

Definition 3.4 (Positive strongly continuous semigroup). A strongly continuous semigroup $\mathcal{F} \coloneqq (T(t))_{t \geq 0}$ of bonded operators on a Banach lattice X is called positive if any operator $T(t) \in \mathcal{F}$ is positive, i.e.

$$0 \leq f \in X \implies 0 \leq T(t)f \ \forall t \geq 0.$$

3.1.3 Problem statement

We consider a single species evolving according to linear birth, death and transition processes. So let X be a Banach lattice; let the following functions be linear operators: the birth operator $B\colon X \to X$, the mortality operator $M\colon \mathcal{D}(M) \subseteq X \to X$ (it also includes the transition process); let

$$\mathcal{D}(M) = \{u \in X\colon Cu = 0\}, \tag{3.1}$$

where $C\colon X \to \mathbb{R}^p$ for some $p \in \mathbb{N}$ is linear.

Following [2] we ask that $-M$ generates a positive, strongly continuous semigroup, whose spectral bound is strictly negative, so that the solution of

$$\begin{cases} v'(t) = -Mv(t) \\ v(0) = v_0 \end{cases}$$

tends to zero for any initial condition $v_0 \in \mathcal{D}(M)$; the biological interpretation of this condition is that any population with no births should tend to extinction.

Moreover we ask that B is positive, since the population cannot decrease due to births, and bounded, because any individual must have a bounded newborns production rate.

Consider now the problem

$$U' = BU - MU \tag{3.2}$$

with the initial condition given implicitly by the domain of M.

The next generation operator, as defined in Section 2.3.2 is then

$$K = BM^{-1},$$

and R_0 is its spectral radius.

We assume that K is a compact operator; in this situation, thanks to the Krein-Rutman theorem (see [15]), the basic reproduction number is an eigenvalue; that is R_0 is the largest value λ for which a nonzero solution of one of the following (equivalent) problems exists:

$$BM^{-1}\phi = \lambda\phi \tag{3.3}$$

$$B\psi = \lambda M\psi. \tag{3.4}$$

Note that if B and M are matrices, like in the case of a discrete structuring variable, these are standard numerical problems, in the sense that some reliable tools to solve the problem already exist; this does not mean that numerically solving them always yields accurate results: in particular, Eq. (3.4) can be solved using some factorization of B and M, which is usually more accurate than computing the inverse of a matrix.

We mention for completeness that type II models, as defined in [2], differ from those of type I described so far because the codomain of B is a space different from X. In this case we cannot write Eq. (3.3); since we are only interested in models which are of type I, we will not investigate further this class of models.

In what follows we define $X \coloneqq L^1([\,0, l\,])$ for some $l > 0$.

3.2 Generic disease

We analyze the model described by the following equations, resulting from the linearization of the equations of a nonlinear model [16]. Since l is here the maximum age of individuals of the population, it is often called $a_\dagger$:

$$\begin{cases} (\partial_t + \partial_a)I(t,a) = S^0(a)\displaystyle\int_0^{a_\dagger} \beta(a,\sigma)I(t,\sigma)d\sigma - (\mu+\gamma)(a)I(t,a) \\ I(t,0) = 0 \\ I(0,a) = I_0(a) \end{cases} \tag{3.5}$$

for $t > 0, a \in]\,0,\, a_\dagger\,[$, where:

- $I(t,a)$ is the infective population of age a at time t;
- S^0 is the total population, i.e. the susceptible population in the disease free state; we assume that for each age a, $S_+ > S^0(a) > 0$, for some $S_+ > 0$ (this is coherent with the usual structure of a human population);
- β is the transmission coefficient, which we assume strictly positive, because we want nobody to be immune to the disease;
- μ is the human mortality, which is strictly positive (from a statistical viewpoint);
- γ is the recovery rate; we want to model an influenza-like disease, so we assume that anybody recovers; therefore we assume it is strictly positive;
- I_0 is the initial density of infective individuals.

The second condition in Eq. (3.5) comes from the fact that we assumed there is no vertical transmission.

Following [16] we also assume that S^0, β, γ, μ are continuous and uniformly bounded[1].

We recall that $X \coloneqq L^1([\,0,\, a_\dagger\,])$ and define the operators $M\colon \mathcal{D}(M) \subseteq X \to X$, $B\colon X \to X$ as follows:

$$M\varphi(a) \coloneqq \varphi'(a) + (\mu+\gamma)(a)\varphi(a) \tag{3.6a}$$

$$B\varphi(a) \coloneqq S^0(a)\int_0^{a_\dagger} \beta(a,\sigma)\varphi(\sigma)\mathrm{d}\sigma \tag{3.6b}$$

with

$$\mathcal{D}(M) \coloneqq \left\{\varphi \in X\colon \varphi' \in X \text{ and } \varphi(0) = 0\right\} \tag{3.7}$$

[1] A more precise model should actually require that $\lim_{x\to a_\dagger} \mu(x) = +\infty$, since otherwise some of the individual may survive the maximum age; in this work we will however not try to improve the models, since our focus is on R_0 and its computation.

Following [16] we can write the inverse of M.
If $\zeta(a) \coloneqq M\varphi(a)$, from Eq. (3.6a), which becomes

$$\zeta(a) = \varphi'(a) + (\mu + \gamma)(a)\varphi(a),$$

and the boundary condition Eq. (3.7) we obtain

$$\begin{cases} \varphi'(a) = \zeta(a) - (\mu + \gamma)(a)\varphi(a) \\ \varphi(0) = 0 \end{cases}$$

Using now the variation of constants formula we can write

$$\begin{aligned} \varphi(\sigma) &= \mathrm{e}^{-\int_0^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\varphi(0) + \int_0^\sigma \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)\mathrm{d}\rho \\ &= \int_0^\sigma \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)\mathrm{d}\rho. \end{aligned}$$

Hence

$$M^{-1}\zeta(\sigma) = \int_0^\sigma \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)\mathrm{d}\rho, \tag{3.8}$$

and the next generation operator K is thus given by

$$K\zeta(a) \coloneqq BM^{-1}\zeta(a) = S^0(a)\int_0^{a_\dagger} \beta(a,\sigma)\int_0^\sigma \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)\mathrm{d}\rho\mathrm{d}\sigma. \tag{3.9}$$

We want now to show that K is compact. We need the following theorem (see [4, Theorem 4.26]):

Theorem 3.5 (Kolmogorov-Riesz-Fréchet). *Let $\mathcal{F}$ be a bounded subset of $L^p(\mathbb{R}, \mathbb{R}^d)$, with $1 \leq p < \infty$. Assume that*

$$\lim_{h\to 0} \|\tau_h f - f\|_p = 0$$

uniformly in $f \in \mathcal{F}$, i.e., for each $\varepsilon > 0$ there exists $\delta > 0$ such that $\|\tau_h f - f\|_p < \varepsilon$ for all $f \in \mathcal{F}$ and all $h \in \mathbb{R}$ such that $|h| < \delta$. Then the closure of $\mathcal{F}|_\Omega$ in $L^p(\Omega, \mathbb{R}^d)$ is compact for any measurable set $\Omega \subset \mathbb{R}$ with finite measure. Here τ_h denotes the translation by h defined by $(\tau_h f)(t) \coloneqq f(t + h)$ and $\mathcal{F}|_\Omega$ denotes the restrictions to Ω of the functions in $\mathcal{F}$.

Theorem 3.6. *The next generation operator K defined by Eq. (3.9) is compact.*

Proof. By the definition of compact operator we need to show that K maps any bounded subset $L^p([\,0,\, a_\dagger\,], \mathbb{R})$ in a relatively compact set of $L^p([\,0,\, a_\dagger\,], \mathbb{R})$.

Since K is linear, this statement is equivalent to the following: K maps $B_{0,1}$ in a relatively compact set of $L^p([\,0,\, a_\dagger\,], \mathbb{R})$. Here $B_{0,1}$ is the unit ball in $L^1([\,0,\, a_\dagger\,], \mathbb{R})$, i.e.

$$\varphi \in B_{0,1} \iff \|\varphi\|_{L^1} = \int_0^{a_\dagger} |\varphi(a)|\mathrm{d}a \leq 1.$$

So we apply Theorem 3.5 for $\Omega \coloneqq [\,0,\, a_\dagger\,]$ and

$$\mathcal{F} \coloneqq \left\{\varphi\colon \varphi|_{[\,0,\,a_\dagger\,]} \in KB_{0,1}, \varphi(a) = 0 \text{ for } a \notin [\,0,\, a_\dagger\,]\right\},$$

i.e., we have extended the functions of $L^p([\,0,\, a_\dagger\,], \mathbb{R})$ by zero outside the domain to match the hypotheses of the theorem.

We thus need to show that

$$\lim_{h\to 0} \int_0^{a_\dagger} |f(a+h) - f(a)| \mathrm{d}a = 0$$

uniformly for $f \in \mathcal{F}$; the integral is performed only in $[\,0,\, a_\dagger\,]$ because the functions in $\mathcal{F}$ are extended by 0 outside this interval.

Let $\zeta \in B_{0,1}$; then $K\zeta(a)$ is given by Eq. (3.9). Setting

$$g(\rho) \coloneqq \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)$$

we can write

$$f(a) \coloneqq K\varphi(a) \coloneqq \int_0^{a_\dagger} S^0(a)\beta(a,\sigma) \int_0^\sigma g(\rho)\mathrm{d}\rho\mathrm{d}\sigma.$$

and

$$\begin{aligned}
&\int_0^{a_\dagger} |f(a+h) - f(a)| \mathrm{d}a = \\
&= \int_0^{a_\dagger} \left| \int_0^{a_\dagger} S^0(a+h)\beta(a+h,\sigma) \int_0^\sigma g(\rho)\mathrm{d}\rho\mathrm{d}\sigma - \int_0^{a_\dagger} S^0(a)\beta(a,\sigma) \int_0^\sigma g(\rho)\mathrm{d}\rho\mathrm{d}\sigma \right| \mathrm{d}a = \\
&\quad = \int_0^{a_\dagger} \left| \int_0^{a_\dagger} \left(S^0(a+h)\beta(a+h,\sigma) - S^0(a)\beta(a,\sigma)\right) \int_0^\sigma g(\rho)\mathrm{d}\rho\mathrm{d}\sigma \right| \mathrm{d}a \leq \\
&\quad\quad \leq \int_0^{a_\dagger} \int_0^{a_\dagger} \left|S^0(a+h)\beta(a+h,\sigma) - S^0(a)\beta(a,\sigma)\right| \left| \int_0^\sigma g(\rho)\mathrm{d}\rho \right| \mathrm{d}\sigma\mathrm{d}a
\end{aligned}$$

We now remark that since γ and μ are positive, $|g(\rho)| \leq |\zeta(\rho)|$, then

$$\left| \int_0^\sigma g(\rho)\mathrm{d}\rho \right| \leq \int_0^\sigma |g(\rho)|\mathrm{d}\rho \leq \int_0^\sigma |\zeta(\rho)|\mathrm{d}\rho \leq \int_0^{a_\dagger} |\zeta(\rho)|\mathrm{d}\rho \leq 1;$$

the last inequality holds because $\zeta \in B_{0,1}$; note that the bound we just obtained means that in what follows we do not have to check that the limitation is uniform in f, since f was defined in terms of ζ (and then g), but the part in S^0 and β is common to any $f \in \mathcal{F}$. Then, for any $f \in \mathcal{F}$

$$\begin{aligned}
\int_0^{a_\dagger} |f(a+h) - f(a)| \mathrm{d}a &\leq \int_0^{a_\dagger} \int_0^{a_\dagger} \left|S^0(a+h)\beta(a+h,\sigma) - S^0(a)\beta(a,\sigma)\right| \mathrm{d}\sigma\mathrm{d}a \leq \\
&\leq a_\dagger^2 \sup_{a,\sigma\in[\,0,\,a_\dagger\,]} \left|S^0(a+h)\beta(a+h,\sigma) - S^0(a)\beta(a,\sigma)\right|
\end{aligned}$$

and since S^0 and β are continuous function, and then so is $S^0\beta$, and they are uniformly continuous, because the domain is compact

$$\lim_{h\to 0}\|\tau_h f - f\|_1 a \le a_\dagger^2 \lim_{h\to 0}\sup_{a,\sigma\in[0,a_\dagger]}\left|S^0(a+h)\beta(a+h,\sigma) - S^0(a) - \beta(a,\sigma)\right| = 0$$

So we have obtained the hypothesis of Theorem 3.5, since, as already remarked, the second limit does not depend on f; we thus proved that the image by K of $B_{0,1}$ is relatively compact, and as a consequence K is compact, so the thesis is proved. □

3.3 Bacteria

The variable here is the spatial density at time t: $u(\cdot,t) \in X = L^1([0,\, l\,])$ of bacteria, depending on time $t \ge 0$, and position $x \in [0,\, l\,]$, where l is the intestine length; the evolution equations are:

$$\begin{cases} \partial_t u(x,t) + \partial_x(c(x)u(x,t) - D(x)\partial_x u(x,t)) + (\beta+\mu)(x)u(x,t) = \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = 2\beta(x)u(x,t) \\ c(0)u(0,t) - D(0)\partial_x u(0,t) = 0 \\ c(l)u(l,t) - D(l)\partial_x u(l,t) = 0 \end{cases} \tag{3.10}$$

where $c(x) \ge 0$ is the velocity of the flow, $D(x) \ge 0$ is the diffusion coefficient, $\beta(x) \ge 0$ and $\mu(x) > 0$ are the fertility and mortality rates.

To define the birth and mortality operators we need to state what a birth event is; when a cell divides we can assume either that the event is the birth of a cell or that it is the birth of two cell and the death of one (the parent); the former is generally preferred if the sizes of the two cells are clearly different, so that the parent and the child are clearly distinguishable, while the latter is preferred for (almost) symmetric division, when it cannot be told who the parent is, thus it is assumed that the parent dies, leaving two children.

In general it may happen that some cells divide symmetrically, and other don't; if θ is the probability of asymmetric division we may define the birth and mortality operators $B\colon X \to X$ and $M\colon \mathcal{D}(M) \subseteq X \to X$ as follows:

$$\begin{aligned} B\varphi(x) &= (\theta\beta(x) + (1-\theta)2\beta(x))\varphi(x) \\ M\varphi(x) &= (c(x)\varphi(x))' + ((1-\theta)\beta(x) + \mu(x))\varphi(x) - (D(x)\varphi'(x))' \end{aligned}$$

However, here we set $\theta = 0$, and write

$$B\varphi(x) \coloneqq 2\beta(x)\varphi(x) \tag{3.11a}$$

$$M\varphi(x) \coloneqq (c(x)\varphi(x))' + (\beta+\mu)(x)\varphi(x) - (D(x)\varphi'(x))' \tag{3.11b}$$

with

$$\mathcal{D}(M) \coloneqq \left\{\varphi \in W^{2,1}([0,\, l\,])\colon c(0)\varphi(0) = D(0)\varphi'(0), c(l)\varphi(l) = D(l)\varphi'(l)\right\}. \tag{3.12}$$

These equations explain the reason why Eq. (3.10) is written in that nonstandard way, with β in the lhs and 2β in the rhs.

We limit our work to the cases in which K is compact, since this is not true in general, as we are going to show.

3.3.1 Non-compact case

We show here a case in which K in not compact; we prove it by showing that in this case R_0 is not an eigenvalue: thanks to the Krein-Rutman theorem (see [15]) this is in contrast with K being compact.

Let $D = 0$; then Eq. (3.10) becomes

$$\begin{cases} \partial_t u(x,t) + \partial_x(c(x)u(x,t)) + (\beta+\mu)(x)u(x,t) = 2\beta(x)u(x,t) \\ c(0)u(0,t) = 0 \\ c(l)u(l,t) = 0 \end{cases} \tag{3.13}$$

and assume that c is differentiable, $c(x) > 0$ for $x \in]\,0,\, l\,[$, $c(0) > 0$, $c(l) = 0$ and

$$\int_0^l \frac{1}{c(x)} \mathrm{d}x = +\infty.$$

This means that the flow is always positive, except at position l; i.e., bacteria are transported towards the end of the intestine, but do not exit.

Also, note that the condition $c(l) = 0$ renders the boundary condition $c(l)u(l,t) = 0$ in Eq. (3.13) always automatically satisfied; this is desirable, since the first equation with $D = 0$ is of the first order, and thus only one condition is usually expected.

To explain the reason behind the last assumption, we first compute the time needed for a cell to arrive at $x = l$ starting from $x = x_0$: since the speed in x is $c(x)$, this time is

$$T_{x_0} \coloneqq \int_{x_0}^l \frac{1}{c(x)} \mathrm{d}x.$$

In any closed set $[\,0,\, x_0\,]$ the function c is continuous and strictly positive, hence it has a minimum, which we call $c_0 > 0$.

Thus

$$\int_0^{x_0} \frac{1}{c(x)} \mathrm{d}x \leq \frac{x_0}{c_0},$$

which is finite, implying $T_{x_0} = +\infty$ for any $x_0 < l$. As a consequence, requiring

$$\int_0^l \frac{1}{c(x)} \mathrm{d}x = +\infty$$

means that the time needed for a cell to get from any position to the intestine end is infinite. That is, the bacteria will not eventually accumulate at $x = l$, which is a reasonable thing to ask.

With these assumptions, we now define, like in the previous section, the operators B and M:

$$B\varphi(x) \coloneqq 2\beta(x)\varphi(x)$$
$$M\varphi(x) \coloneqq (c(x)\varphi(x))' + (\beta+\mu)(x)\varphi(x)$$
$$\mathcal{D}(M) \coloneqq \left\{\varphi \in W^{1,1}([\,0,\,l\,])\colon \varphi(0)=0\right\} = \left\{\varphi \in L^1([\,0,\,l\,])\colon \varphi \in L^1([\,0,\,l\,]), \varphi(0)=0\right\}.$$

Last, suppose by contradiction that R_0 is an eigenvalue. Substituting the operators above into Eq. (3.4) we get

$$\begin{cases} 2\beta(x)\varphi(x) = R_0(c(x)\varphi(x))' + R_0(\beta+\mu)(x)\varphi(x) \\ \varphi(0)=0 \end{cases}$$

which we can rewrite as

$$\begin{cases} \varphi'(x) = \frac{2\beta(x) - R_0c'(x) - R_0(\beta+\mu)(x)}{R_0c(x)}\varphi(x) \\ \varphi(0)=0 \end{cases}$$

which is an initial value problem, whose first equation is defined for $x \neq l$ (because $c(l)=0$); it satisfies the hypotheses of the Cauchy-Lipschitz theorem (since it is a linear problem), and thus the solution $\varphi(x)=0$ (which is a solution indeed) is the only solution.

But this means that R_0 is not an eigenvalue.

As already said, we will not investigate further this case.

3.3.2 Explicit solutions

In the special case of constant β and μ we can solve the problem explicitly: let $p(t) = \int_0^l u(x,t)\mathrm{d}x$ be the total population at time t; then integrating the first equation of Eq. (3.10), yields

$$p'(t) = c(0)u(0,t) - c(l)u(l,t) - (\beta+\mu)p(t) + 2\beta p(t) + D(l)\partial_x u(l,t) - D(0)\partial_x u(0,t).$$

By applying the boundary condition of Eq. (3.10) the terms containing D and c cancel:

$$p'(t) = (\beta+\mu)p(t) + 2\beta p(t);$$

This equation is in the same form as Eq. (2.1), which allowed a direct computation of R_0; in this case the result is

$$R_0 = \frac{2\beta}{\beta+\mu}. \tag{3.14}$$

The eigenvector of BM^{-1} relative to the eigenvalue R_0 can also be computed explicitly:

$$B\psi = R_0 M\psi$$

means

$$2\beta\psi(x) = \frac{2\beta}{\beta+\mu}\left((c(x)\psi(x))' + (\beta+\mu)\psi(x) - (D(x)\psi'(x))'\right)$$

and since all terms containing β and μ cancel, by integration

$$c(x)\psi(x) - D(x)\psi'(x) = k \tag{3.15}$$

for some constant k. Considering Eq. (3.15) in $x = 0$ it must be $k = 0$, due to the boundary condition in Eq. (3.10), or equivalently for the constraints on ψ imposed by the domain of M (Eq. (3.12)).

Then it holds:

$$\psi(x) = \mathrm{e}^{\int_0^x \frac{c(s)}{D(s)}\mathrm{d}s}$$

which is the only eigenvector associated to R_0.

Chapter 4

Numerical approximation

An explicit representation of the next generation operator is not available in general; this means that it cannot be directly approximated using a matrix. On the contrary, for the model to be defined, the operators B and M must be available.

The spectral radius of the next generation operator can then be approximated by the spectral radius of the matrix $B_n M_n^{-1}$, where B_n and M_n are suitable approximations of B and M respectively. This can be done by solving the discrete version of one of Eqs. (3.3) and (3.4), namely

$$B_n M_n^{-1} \Phi = \lambda \Phi$$

$$B_n \Psi = \lambda M_n \Psi.$$

As already remarked, these problems are equivalent in exact arithmetic, but not in machine arithmetic.

4.1 General method

Let $X = L^1([\,0,\, l\,])$, $X_n = \mathbb{C}^n$ (for any $n \in \mathbb{N}$).

Let

$$J_n \colon X_n \to X$$

and

$$P_n \colon X \to X_n$$

be bounded linear operators (we do not require, however, that the bound is uniform in n), such that

$$P_n J_n = I_{X_n}. \tag{4.1}$$

Lemma 4.1. *Under the conditions above it also holds* $J_n P_n|_{J_n X_n} = I_X$.

Proof. Since

$$(J_n P_n)\, J_n \phi = J_n\, (P_n J_n)\, \phi = J_n \phi$$

the thesis holds. □

This means that J_n is an immersion of X_n into X, and $J_n P_n$ is a projection of X onto $J_n P_n X$.

From now on, whenever it may be useful we will abuse of notation and say that $X_n \subseteq X$, and P_n projects X onto X_n; that is, we want to identify the functions of the space $J_n X_n$ with the vectors of X_n which represent them in a chosen (fixed) basis.

Definition 4.2. Given any linear operator $L\colon X \to X$, define

$$L_n \coloneqq P_n L J_n.$$

Given any linear functional $Q\colon X \to \mathbb{R}^p$, define

$$Q_n \coloneqq Q J_n$$

This mean that we want L_n to be an endomorphism of X_N which approximates L on this subspace, and Q_n a matrix approximating Q.

Lemma 4.3. *Let $K\colon X \to X$ or $K\colon X \to \mathbb{R}^p$; then the operator*

$$K \mapsto K_n$$

is linear. Moreover, suppose that $L(X_n) \subseteq X_n$; then

$$(KL)_n = K_n L_n \tag{4.2}$$

Proof. The linearity is obvious.

Let $K\colon X \to \mathbb{R}^p$; Eq. (4.2) means

$$K L J_n = K J_n P_n L J_n \tag{4.3}$$

which is true since L maps $J_n X$ into itself and because of Lemma 4.1.

If $K\colon X \to X$ Eq. (4.3) continues to hold; applying P_n to both sides concludes the proof. □

We note however that approximating KL by $K_n L_n$ may yield good results even if L does not map X_n into itself, as we will see in Lemma 5.2 and Theorem 5.3 ; also, we may not be able to write the matrix L_n, in which case we must resort to a good approximation.

Applying Definition 4.2 for $L = M$ and $L = B$, we get two square matrices, that we call $\tilde{M}_n$ and $\tilde{B}_n$.

However, we did not consider the constraints due to the domain of M, so we must also apply Definition 4.2 with $Q \coloneqq C$, and the condition $C\phi = 0$ becomes $C_n\Phi = 0$.

Then, we have two ways to force the condition deriving from the domain of M. The first method involves finding $\ker C_n$ first, and writing it in the form

$$\Phi_1 = E\Phi_2 \tag{4.4}$$

where $\Phi = (\Phi_1, \Phi_2)$ (except at most for an index reordering; row vector are used in the last expression for typographic reasons); then $\tilde{B}_n$ and $\tilde{M}_n$ have to be modified accordingly, and only applied to the space generated by Φ_2 (thus obtaining B_n and M_n, both of dimension $(n-p)\times(n-p)$), since then Φ_1 is recovered from the equation above.

The second method relies on a direct modification of the action of $\tilde{B}_n$ and $\tilde{M}_n$ in such a way that the condition on the domain is satisfied for any vector solving the eigenvalues problem

$$B_n M_n^{-1}\Phi = \lambda\Phi \tag{4.5}$$

or, equivalently

$$B_n\Phi = \lambda M_n\Phi. \tag{4.6}$$

Regarding this alternative we note that a matrix which correctly approximates the action of M on the space of polynomials cannot be square, in general (due to the conditions in $\mathcal{D}(M)$); and that in order to solve the problem the action of B is also actually needed only in the domain of M, because B acts on $M^{-1}\phi \in \mathcal{D}(M)$ in Eq. (3.3); in Eq. (3.4) it acts on ψ, like also M does, so ψ must be in $\mathcal{D}(M)$ too.

What we want to do is then stack the action on $\mathcal{D}(M)$, described by rectangular matrices of dimensions $(n-p)\times n$, and the domain conditions which use C_n, of dimensions $p\times n$ together with a zero matrix of the same dimension. We still need to know the decomposition of Φ, although we do not need the matrix E of Eq. (4.4). So we define

$$M_n = \begin{pmatrix} C_n \\ \tilde{M}_n \end{pmatrix} \tag{4.7a}$$

$$B_n = \begin{pmatrix} 0 \\ \tilde{B}_n \end{pmatrix} \tag{4.7b}$$

where 0 is here the $p\times n$ zero matrix.

The correctness (with respect to Definition 4.2) of these expressions is evident from Eq. (4.6), which decomposes in

$$\begin{cases} 0\Phi = \lambda C_n\Phi & \text{(equations describing } \Phi_1\text{)} \\ B_n\Phi = \lambda M_n\Phi & \text{(equations describing } \Phi_2\text{).} \end{cases}$$

Also, Eqs. (4.5) and (4.6) are equivalent, so we can choose any of the two.

4.2 Choice of the method

The numerical method is now determined by J_n, P_n and the choice for how to impose the condition on the domain.

Since the functions of X do not have a particular form (e.g. they are not periodic, or symmetric), the standard choice is to set

$$J_n X_n = \Pi_{N,\nu},$$

for some N, ν: $N\nu = n$; i.e., having a division of the domain $[\,0,\, l\,]$, namely $0 = a_0 < \cdots < a_N = l$, $\Pi_{N,\nu}$ is the space piecewise polynomials defined by

$$\Pi_{N,\nu} = \left\{ f \in L^1([\,0,\, l\,])\colon f \text{ is a polynomial of degree } \nu \text{ in } [\,a_{j-1},\, a_j\,], 1 \le j \le N \right\}.$$

We call *external mesh* the set of points a defined above, or equivalently the set of the intervals that these points induce.

4.2.1 Spectral method or finite elements method?

Here another choice is to be made. The first alternative is to use the space of polynomials of degree at most n, i.e. $N = 1, \nu = n$; or fix N, and then improve the approximation by increasing the polynomial degree. This is what we call called "spectral method". The second alternative is to use piecewise polynomials, fixing the degree ν and using an increasing number N of intervals to improve the approximation. We will call this a "finite elements" method.

The choice typically depends on the regularity of the eigenvector ϕ, since higher degree polynomials yield faster convergence for regular functions.

That is, if we use a single polynomial (or a piecewise polynomial on a fixed number of intervals), and the eigenvector is analytic, the error is

$$O\left(\left(\frac{\nu}{q}\right)^{-\nu}\right)$$

for some q which depend on the eigenvector; this behavior is called "spectral accuracy" [24, Chapter 4].

On the contrary for functions in C^k, since the order of convergence is limited at most to k, unless the operator M and B are smoothing, fixing $\nu = k$ as the polynomial degree gives a faster method while reaching the maximum order of convergence $O(N^{-k})$; this is mostly useful if M_n and B_n are block-diagonal, since the highest computational cost of the method comes from solving the eigenvalues problem (and inverting M_n, if solving the standard eigenvalues problem), which has cubic complexity in the matrix dimension.

Also, note that it is not necessary that every function in X is regular, only the eigenvector ϕ relative to R_0. Indeed, suppose that the spectral radius of K is an eigenvalue ρ, and ϕ is an eigenvector relative to ρ; when restricting to the functions

at least as regular as ϕ, the actions of B and M do not change, only the domain do, so ϕ is still an eigenvector, relative to the same eigenvalue; then the spectral radius is at least ρ; moreover, it cannot be greater, otherwise we would have a regular eigenvector relative to this greater eigenvalue, but it would, in particular, be an eigenvalue of K, whose spectral radius would be greater.

In the models of Chapter 3 this eigenvector is usually C^∞: let p be the greatest derivative order of ψ appearing in $M\psi$; write $M = M_1 + M_2$, where $M_2\psi$ only depend on $\psi^{(p)}$, not on ψ or the lower order derivatives themselves, and $M_1\psi$ does not depend on $\psi^{(p)}$. Consider now Eq. (3.4), which we can also write as

$$\lambda M_2\psi = (B - \lambda M_1)\psi$$

then, if $\psi \in C^k$, with $k \geq p$, and $B, M \in C^{k-p+1}$ then

$$(B - \lambda M_1)\psi \in C^{k-p+1}$$

(since $p-1$ is the maximum derivative order appearing in the rhs), so we also get $\psi^{(p)} \in C^{k-p+1}$, i.e. $\psi \in C^{k+1}$; that is, if $B, M \in C^\infty$, it must also hold $\psi \in C^\infty$.

Since ψ is regular, a single polynomial should yield faster convergence; we will not restrict to this choice, however, because our aim is to compare these different instances of the general method presented in the previous section.

4.2.2 Choice of the nodes

Having chosen the (piecewise) polynomials as the approximating space, we now choose a basis of that space. We do so by choosing a set of nodes in each interval of the external mesh, and then by using the Lagrange basis on those nodes, i.e. the $j\nu + j + m$-th element of the basis, with $0 \leq m \leq \nu$, and $0 \leq j < N$, is a polynomial of the Lagrange basis when restricted to the j-th interval, and is 0 outside that interval[1].

We can then define J_n as the isomorphism which associates the j-th element of the canonical basis of $\mathbb{C}_n$ to the j-th element of the chosen basis of Π_n and P_n as the operator mapping a function of X to the (piecewise) polynomial interpolating it on the chosen nodes.

We call *internal mesh* the set of nodes in a specific interval, and *full mesh* the whole set of nodes, i.e. the union of internal meshes of all intervals.

Since not all nodes distributions are well suited for polynomial approximation (e.g. uniform nodes are not, in general) we restrict here to two types; those "optimal" for integration, i.e. Gauss nodes, which are (for $\nu + 1$ points) the zeros of the ν-th Legendre polynomial and can be computed solving a tridiagonal eigenvalue problem (see Appendix A.1.4); those "optimal" for interpolation, i.e. Chebyshev extremal nodes, given (for $\nu + 1$ points) by

$$\cos\left(\frac{\nu - i}{\nu}\pi\right), \text{ for } 0 \leq i \leq \nu;$$

[1] See Appendix A.1.2 for the definition of the polynomial of the Lagrange basis.

Chebyshev nodes are located in $[-1, 1]$, so they must then be scaled to be used in the needed interval; on the contrary Gauss nodes can be directly computed in any interval; however computing them in a single interval and then using the same scaling as for Chebyshev nodes leads to faster computation.

Different nodes choices also have side effects, particularly when it comes to the conditions on the domain: since Chebyshev nodes include the two extremes, if domain conditions are only imposed on (one of) such points, which is a common situation, there exists a natural decomposition of Φ; on the contrary, when using Gauss nodes, it is generally more complicated to write the decomposition.

Also, using the finite elements method and the Chebyshev nodes, the points of the external mesh (excluding the interval endpoints) are used twice, representing the left and right limit of the function at that point. This is not a problem, but, since the eigenvector relative to R_0 is continuous, we can set

$$J_n X_n \coloneqq \Pi^{(0)}_{N,\nu+1} \coloneqq$$
$$\left\{ f \in C^0([\,0,\, l\,]) \colon f \text{ is a polynomial of degree } \nu+1 \text{ in } [\,a_{j-1},\, a_j\,], 1 \le j \le N \right\},$$

i.e. the space of *continuous* piecewise polynomials, were n is now given by $N\nu + 1$ (intuitively, $N - 1$ parameters are fixed by the continuity conditions); this is done including the first (resp. last) point of the internal mesh only in the first (resp. last) interval of the external mesh, which produces a smallest matrix, destroying, however, any possible block-diagonal structure, because any two consecutive blocks will share a column.

In terms of the basis of the space, this choice means that it contains elements which are nonzero polynomials on two intervals of the external mesh; these elements are those with the same index as a point which is the first or the last of an internal mesh (except for the endpoints 0 and l).

4.3 Matrices construction

Let $b_j,\ 0 \le j \le \nu$ be nodes in $[\,0,\, 1\,]$,

$$\Delta a_j \coloneqq a_{j+1} - a_j,\ 0 \le j \le N-1,$$

and $p_{j\nu+j+m} \coloneqq a_j + b_m \Delta a_j$ for $0 \le j \le N-1,\ 0 \le m \le \nu$ be the points of the full mesh.

If $\phi \in X$ we define $\Phi_k = \phi(p_k)$, for $0 \le k \le N(\nu+1)-1$, where Φ_k is the k-th component of Φ (which is coherent with our choice of P_n and J_n).

We will also, as already said, abuse of notation and identify the vector Φ and the polynomial it represents in the Lagrange basis, i.e.

$$\Phi|_{[\,a_k,\, a_{k+1}\,]} = \sum_{j=(k+1)\nu}^{(k+2)\nu} l_j \Phi_j,$$

where l_j is the j-th polynomial of the basis of $J_n X_n$ defined above.

The matrices approximating the birth and mortality operators are simpler to describe using indexes ranging from 0 to $n-1$, so we will adopt this convention.

4.3.1 Disease

Boundary condition From Eq. (3.7) we get

$$C\phi = \phi(0), \tag{4.8}$$

which in X_n translates to

$$\phi_n(0) = \sum_{j=0}^{\nu} l_j(0)\Phi_j;$$

we can thus write, using Definition 4.2

$$C_n = (l_0(0)\ l_1(0)\ \dots\ l_\nu(0)\ 0\ \dots\ 0), \tag{4.9}$$

i.e. a $1 \times n$ matrix.

Using Chebyshev nodes, its entries are all 0 except the first; on the contrary, using Gauss nodes $\phi_n(0)$ depends on the values on the whole first internal mesh, i.e. the first $\nu+1$ entries are nonzero.

Thus using the Chebyshev nodes we can only replace the first line; the Gauss nodes let us choose any of the first ν rows; we choose the first as well, since we have no reasons to do otherwise.

In what follows we thus omit the description of the first row.

Operator action We rewrite Eqs. (3.6a) and (3.6b) as

$$M \coloneqq H + \operatorname{diag}(\mu + \gamma) \tag{4.10a}$$

$$B \coloneqq \operatorname{diag} S^0 \Sigma, \tag{4.10b}$$

where for a function f we define $\operatorname{diag} f$ as the operator F such that$(F\phi)(a) = f(a)\phi(a)$; H is the derivation operator $H\phi = \phi'$; and Σ is the following integral operator:

$$(\Sigma\phi)(a) = \int_0^l \beta(a,\sigma)\phi(\sigma)\mathrm{d}\sigma.$$

We would now like to apply Lemma 4.3. We can use it for Eq. (4.10a), but not for Eq. (4.10b), because Σ does not map X_n into itself in general; and anyway we are not able to write the exact action of Σ on X_n, unless we know β, and only for a few special functions β.

Since we do not want our method to depend on β we approximate the entire integral with a quadrature formula; We will then need to prove, in Lemma 5.8, that this choice is a good one and does not lead to problems in the convergence.

Since Eqs. (4.10a) and (4.10b) involves derivatives and integrals, we choose $q_j,\ 0 \le j \le \nu$ weights of a quadrature formula and $d_{j,k},\ 0 \le j,k \le \nu$ weights of a differentiation formula on the nodes of any internal mesh; since we actually use the same nodes on any interval of the external mesh, scaled to fit the interval, we also choose a single set of quadrature and differentiation weights and scale them so that they continue to be the weights of quadrature and differentiation formulas of the desired polynomial order on the given (scaled) nodes.

So we approximate separately H, $\operatorname{diag}(\mu+\gamma)$, $\operatorname{diag} S^0$ and Σ. Since H_n is given by

$$(H_n)_{i\nu+i+h,j\nu+j+m} = \delta_{i,j}\frac{d_{h,m}}{\Delta a},$$

we obtain (still denoting, with some abuse of notation, the approximated matrices by B_n and M_n)

$$(B_n)_{i\nu+i+h,j\nu+j+m} = S^0(a_i + b_h\Delta a_i)q_m\beta(a_i + b_h\Delta a_i, a_j + b_m\Delta a_j) \tag{4.11}$$

$$(M_n)_{i\nu+i+h,j\nu+j+m} = \delta_{i,j}\left(\frac{d_{h,m}}{\Delta a} + \delta_{h,m}(\gamma+\mu)(p_t)\right), \tag{4.12}$$

for $i\nu + i + h \neq 0,\ 0 \le h,m \le \nu,\ 0 \le i,j < N$, where $\delta_{a,b}$ is the Kronecker delta, which is 1 if $a = b$, and 0 otherwise.

On the Chebyshev nodes we can also write:

$$(B_n)_{i\nu+h,j\nu+m} = S^0(a_i + b_h\Delta a_i)\Delta a_j q_m\beta(a_i + b_h\Delta a_i, a_j + b_m\Delta a_j)$$

for $1 \le h \le \nu,\ 0 \le i < N$ and $1 \le m < \nu,\ 0 \le j < N$ or $j\nu+m = 0$ or $j\nu+m = n-1$; in the remaining cases

$$(B_n)_{i\nu+h,j\nu} = S^0(a_{i\nu+h})(\Delta a_j q_m + \Delta a_{j-1}q_0)\beta(a_i + b_h\Delta a_i, a_j)$$

i.e. for the same values of i,h and $0 < j < N$.

In other words, the matrix B_n is constructed using blocks, whose last column overlaps with the first of the following block, i.e.

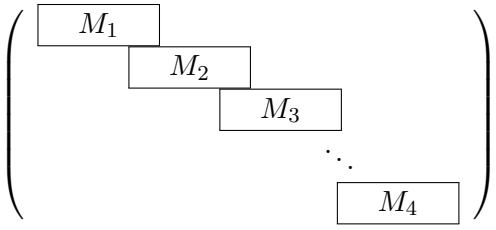
$$\left(\begin{array}{c|c|c|c|c}
B_{1,1}(0:\nu-1) & \begin{array}{c}B_{1,1}(\nu)\\+\\B_{1,2}(0)\end{array} & B_{1,2}(1:\nu-1) & \cdots & B_{1,N}(1:\nu)\\
\hline
B_{2,1}(0:\nu-1) & \begin{array}{c}B_{2,1}(\nu)\\+\\B_{2,2}(0)\end{array} & B_{2,2}(1:\nu-1) & \cdots & B_{2,N}(1:\nu)\\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots\\
\hline
B_{N,1}(0:\nu-1) & \begin{array}{c}B_{N,1}(\nu)\\+\\B_{N,2}(0)\end{array} & B_{N,2}(1:\nu-1) & \cdots & B_{N,N}(1:\nu)
\end{array}\right),$$

where an item $B_{i,j}(h:k)$ are the columns from h to k of the block $B_{i,j}$, and $B_{i,j}(k)$ is only the k-th column; $B_{i,j}$ is obtained fixing i and j in Eq. (4.11), and eliminating the first row of the obtained block.

Like for B_n, M_n also has overlapping blocks; but since it previously was block-diagonal, the structure remains simpler

$$(M_n)_{i\nu+h,j\nu+m} = \delta_{i,j}\left(\frac{d_{h,m}}{\Delta a} + \delta_{h,m}(\gamma+\mu)(p_t)\right)$$

for $1 \le h \le \nu$, $0 \le i,j < N$ and $0 \le m \le \nu$; i.e. it is now quasi-block-diagonal, having the blocks (derived from in Eq. (4.12) in the same way as those of B_n) which overlap in the same way as in B_n, and the diagonal perturbed by $\gamma+\mu$; the structure is the following:



4.3.2 Bacteria

Let H_n be the derivation matrix described above.

Boundary conditions We rewrite Eq. (3.12) in terms of C, whose image is now $\mathbb{R}^2$:

$$C\phi = \begin{pmatrix} c(0)\phi(0) - D(0)\phi'(0) \\ c(l)\phi(l) - D(l)\phi'(l) \end{pmatrix}$$

or, if $V_0 f \coloneqq f(0)$ and $V_l f \coloneqq f(l)$

$$C \coloneqq \begin{pmatrix} V_0 \\ V_l \end{pmatrix} (\operatorname{diag} c - \operatorname{diag} DH)\,.$$

Here H maps Π_n into itself, but $\operatorname{diag} c$ and $\operatorname{diag} D$ do not, in general; so again we can only use Lemma 4.3 only for the linearity. We nevertheless separately approximate H, $\operatorname{diag} c$, $\operatorname{diag} D$ and $(V_0, V_l)^T$, which will yield good results, thanks to what we will see in Lemma 5.2 and Theorem 5.3.

We can thus construct C_n: if

$$v_0 \coloneqq (l_0(0)\ l_1(0)\ \dots\ l_\nu(0)\ 0\ \dots\ 0),$$

$$v_l \coloneqq (0\ \dots\ 0\ l_{(N-1)\nu}(l)\ l_{(N-1)\nu+1}(l)\ \dots\ l_{N\nu}(l)),$$

then

$$C \coloneqq \begin{pmatrix} v_0 \\ v_l \end{pmatrix} (\operatorname{diag}_n c - \operatorname{diag}_n DH_n)$$

where for a function f we define $\operatorname{diag}_n f$ as the diagonal matrix F such that

$$F_{i,i} \coloneqq f(p_i)$$

and p_i is th i-th point of the full mesh.

This time we have to choose two rows to describe the domain conditions; they need to be one in the first block (i.e. blocks of $\nu+1$ rows) and one in the last, using both the Chebyshev and the Gauss nodes (due to the derivative); we choose the first and the last row.

Operators action Then Eq. (3.11b) can be written as

$$M = H\left(\operatorname{diag} c + \operatorname{diag}_n DH\right) + \operatorname{diag}(\beta + \mu)$$

This is very similar to the boundary condition, so the same considerations on the use of Lemma 4.3 hold.

Again, we approximate H, $\operatorname{diag} c$, $\operatorname{diag} D$ and $\operatorname{diag}(\beta+\mu)$ separately, and we delay the proof that this yields good results to Lemma 5.2 and Theorem 5.3:

$$M_n\Phi = H_n\left(\operatorname{diag}_n c + \operatorname{diag}_n DH_n\right)\Phi + \operatorname{diag}_n(\beta+\mu)\Phi.$$

Eq. (3.11a) has no problems since it contains no operator multiplications and becomes

$$B_n \coloneqq 2\operatorname{diag}_n \beta.$$

Chapter 5

Proofs of convergence

5.1 General method

In this preliminary convergence analysis we are going to assume some facts that are actually not true, namely that there are no border conditions and that M is bounded.

We now suppose that we can search for the eigenvector relative to R_0 in a space of regular functions; we thus define, for some set of points $0 = \alpha_0, \dots, \alpha_{N_0} = l$, the normed spaces

$$C^\nu_{N_0} = (\{f \in C^0([\,0,\, l\,])\colon f|_{[\,\alpha_{j-1},\,\alpha_j\,]} \in C^\nu, 1 \le j \le N_0\}, \|\cdot\|_\infty)$$
$$X_n = \left(\mathbb{C}^n, \|\cdot\|_{X_n}\right) \text{ for any } n \in \mathbb{N},$$

where

$$\|\Phi\|_{X_n} \coloneqq \max_{j=1}^{n} |\Phi_j|$$

for $\Phi \in X_n$ and $n \in \mathbb{N}$; here Φ_j denotes the j-th component of Φ. Note that $C^\nu_{N_0}$ is not a Banach space, unless $\nu = 0$.

When it does not create confusion we write C in place of $C^\nu_{N_0}$.

We also consider the same operators P_n and J_n of the previous chapter, which now have a different codomain and domain respectively; we may omit the subscript in the norms when it is clear which norm is used.

We define $\tilde{N} = \{N\nu\colon N \in \mathbb{N}\}$, and for $f\colon \tilde{N} \to \mathbb{R}^+$

$$O(f(n)) = \left\{g\colon \tilde{N} \to \mathbb{R}^+\colon \exists a, b > 0\colon g(x) \le a + bf(x) \forall x \in \tilde{N}\right\}$$

Suppose that for some functions ω, ι, κ

Assumption 5.1.

$$\|J_n\|_{C\leftarrow X_n} \in O(\iota(n))$$
$$\|P_n\|_{X_n\leftarrow C} \in O(\iota(n))$$
$$\|J_nP_n - I_C\| \in O(\omega(n)) \subseteq O(\kappa(n))$$
$$\iota^d(n)\omega(n) \in O(\kappa(n))\forall d \in \mathbb{N}.$$

Actually, we should call these objects $\omega_\nu, \iota_\nu, \kappa_\nu$, because they depend on ν, but we omit the subscript when there is no ambiguity.

For a linear operator L, let L_n be defined as in the previous chapter.

Lemma 5.1. *Fix L; if it is bounded, the following hold:*

$$\|J_nL_nP_n - L\| \in O((\iota^2\omega)(n))$$

Proof. Using the definition of L_n, the triangular inequality and the operator norms properties we can write

$$\begin{aligned}\|J_nL_nP_n - L\| = \|J_nP_nLJ_nP_n - L\| \leq \\ \leq \|J_nP_nLJ_nP_n - LJ_nP_n\| + \|LJ_nP_n - L\| \leq \\ \leq \|J_nP_n - I_C\|\|L\|(\|J_nP_n\| + 1).\end{aligned}$$

Moreover, thanks to Assumption 5.1 $\|J_nP_n - I_C\| \in O(\omega(n))$, $\|J_nP_n\| + 1 \in \iota^2(n)$, and so the whole expression is in $\|L\|O((\iota^2\omega)(n)) \coloneqq O(\|L\|(\iota^2\omega)(n))$. Since $\|L\|$ is a multiplicative constant we can absorb it into the $O((\iota^2\omega)(n))$ term, and the thesis follows. □

We already proved in Lemma 4.3 that if L maps $C_{N,\nu}$ into itself, for a linear operator K

$$(KL)_n = K_nL_n;$$

This is, however, a rather restrictive requirement; we then wish to generalize that statement.

Lemma 5.2. *Let $L\colon C \to C$, and $K\colon C \to C$ or $K\colon C \to \mathbb{R}^p$ be bounded operators; then*

$$\|(KL)_n - K_nL_n\| \in O((\iota^2\omega)(n))$$

Proof. Let $K\colon C \to C$; then thanks to the operator norms properties and Assumption 5.1

$$\|P_nKJ_nP_nLJ_n - P_nKLJ_n\| \leq \|P_n\|\|K\|\|J_nP_n - I_C\|\|L\|\|J_n\| \in \|K\|\|L\|O((\iota^2\omega)(n)) \tag{5.1}$$

and since K and L are fixed (bounded) operators their norm can be absorbed in the term $O((\iota^2\omega)(n))$ and the thesis is proved in this case.

If $K\colon C \to \mathbb{R}^p$ Eq. (5.1) holds without the factor P_n on the right; so the conclusion $\|KJ_nP_nLJ_n - KLJ_n\| \in \|K\|\|L\|O((\iota^2\omega)(n))$ is still true; as before the norm of K and L can be absorbed in the term $O((\iota^2\omega)(n))$, which completes the proof.□

As a special case, if L and L^{-1} are bounded $\left\|(L^{-1})_n L_n - I\right\| \in O((\iota^2\omega)(n))$.

We are interested in an approximation of the next generation operator $K = BM^{-1}$. We thus suppose

Assumption 5.2. The operators B, M and M^{-1} are bounded;

We now can prove

Theorem 5.3. *Under Assumptions 5.1 and 5.2, and if* $\left\|(M_n)^{-1}\right\| \in O(\iota^\tau(n))$*, for some* $\tau \in \mathbb{N}$*, then* $\left\|J_n B_n (M_n)^{-1} P_n - BM^{-1}\right\| \in O(\kappa(n))$.

Proof.

$$\begin{aligned}\left\|J_n B_n (M_n)^{-1} P_n - BM^{-1}\right\| \leq \\ \leq \left\|J_n B_n (M_n)^{-1} P_n - J_n B_n P_n M^{-1}\right\| + \left\|J_n B_n P_n M^{-1} - BM^{-1}\right\| \leq \\ \leq \|J_n\|\|B_n\|\left\|(M_n)^{-1} P_n - P_n M^{-1}\right\| + \|J_n B_n P_n - B\|\left\|M^{-1}\right\|.\end{aligned}$$

Thanks to Lemma 5.1 and the bound on M^{-1} the second term is in $O((\iota^{2+\tau}\omega)(n))$. As for the first term, we write

$$\begin{aligned}\left\|(M_n)^{-1} P_n - P_n M^{-1}\right\| \leq \\ \leq \left\|(M_n)^{-1}\right\|\|P_n\|\left\|I - MJ_n P_n M^{-1}\right\| \leq \\ \leq \left\|(M_n)^{-1}\right\|\|P_n\|\|M\|\|I - J_n P_n\|\left\|M^{-1}\right\|; \quad (5.2)\end{aligned}$$

and

$$\|B_n\| \leq \|P_n\|\|B\|\|J_n\|$$

that is,

$$\|J_n\|\|B_n\|\left\|(M_n)^{-1} P_n - P_n M^{-1}\right\| \leq \|J_n\|^2\|P_n\|^2\|B\|\|M\|\|I - J_n P_n\|\left\|M^{-1}\right\|^2;$$

so the first term is in $O((\iota^{4+2\tau}\omega)(n)) \subseteq O(\kappa(n))$, because $\|B\|, \|M\|, \left\|M^{-1}\right\|$ are constants which can be absorbed in the $O(\kappa(n))$ term.

Finally, suppose that M_n and B_n are not directly approximations of B and M, but rather, if $B = B_1 B_2$, $B_n \coloneqq P_n B_1 J_n P_n B_2 J_n$.

Then,

$$\begin{aligned}\left\|J_n P_n B_1 J_n P_n B_2 J_n (M_n)^{-1} P_n - BM^{-1}\right\| \leq \\ \leq \left\|J_n P_n B_1 J_n P_n B_2 J_n (M_n)^{-1} P_n - J_n P_n B J_n (M_n)^{-1} P_n\right\| + \\ + \left\|J_n P_n B J_n (M_n)^{-1} P_n - BM^{-1}\right\| \leq \\ \leq \|J_n\|\|P_n B_1 J_n P_n B_2 J_n - P_n B J_n\|\left\|(M_n)^{-1}\right\|\|P_n\| + \\ + \left\|J_n P_n B J_n (M_n)^{-1} P_n - BM^{-1}\right\|\end{aligned}$$

We already proved that the second term is in $O(\kappa(n))$; for the first term

$$\|P_n B_1 J_n P_n B_2 J_n - P_n B J_n\| \in O((\iota^2\omega)(n))$$

thanks to Lemma 5.2, and the remaining part is in $O(\iota^{2+\tau}(n))$, i.e. the whole product in the first term is in $O((\iota^{4+\tau}\omega)(n)) \subseteq O(\kappa(n))$, and then so is the whole

$$\left\|J_n P_n B_1 J_n P_n B_2 J_n (M_n)^{-1} P_n - BM^{-1}\right\|.$$

Similarly, if $M = M_1 M_2$, and $M_n \coloneqq P_n M_1 J_n P_n M_2 J_n$, we need to modify Eq. (5.2). We thus write

$$\begin{aligned}
&\left\|(M_n)^{-1}P_n - P_n M^{-1}\right\| \leq \\
&\qquad \leq \left\|(M_n)^{-1}\right\|\|P_n\|\left\|I - M_1 J_n P_n M_2 J_n P_n M_2^{-1} M_1^{-1}\right\| \leq \\
&\qquad \leq \left\|(M_n)^{-1}\right\|\|P_n\|\|M_1\|\left\|I - J_n P_n M_2 J_n P_n M_2^{-1}\right\|\left\|M_1^{-1}\right\| \leq \\
&\qquad \leq \left\|(M_n)^{-1}\right\|\|P_n\|\|M_1\|\left\|I - M_2 J_n P_n M_2^{-1}\right\|\left\|M_1^{-1}\right\| + \\
&\quad + \left\|(M_n)^{-1}\right\|\|P_n\|\|M_1\|\left\|M_2 J_n P_n M_2^{-1} - J_n P_n M_2 J_n P_n M_2^{-1}\right\|\left\|M_1^{-1}\right\| \leq \\
&\qquad \leq \left\|(M_n)^{-1}\right\|\|P_n\|\|M_1\|\|M_2\|\|I - J_n P_n\|\left\|M_2^{-1}\right\|\left\|M_1^{-1}\right\| + \\
&\quad + \left\|(M_n)^{-1}\right\|\|P_n\|\|M_1\|\|I - J_n P_n\|\left\|M_2 J_n P_n M_2^{-1}\right\|\left\|M_1^{-1}\right\| \in O((\iota^{3+\tau}\omega)(n))
\end{aligned}$$

So that the object we are interested in, namely

$$\left\|J_n B_n (M_n)^{-1} P_n - BM^{-1}\right\|$$

is in $O((\iota^{6+\tau}\omega)(n)) \subseteq O(\kappa(n))$, which gives the thesis. □

Following [17, Lemma 2.25], we state the following

Lemma 5.4. *Let $\mathcal{U}$ be a Banach space, A a linear and bounded operator on $\mathcal{U}$ and $\{A_N\}_{N\in\mathbb{N}}$ a sequence of linear and bounded operators on $\mathcal{U}$ such that $\|A_N - A\|_{\mathcal{U}\leftarrow\mathcal{U}} \to 0$ for $N \to +\infty$. If $\mu \in \mathbb{C}$ is an eigenvalue of A with finite algebraic multiplicity ν and ascent l, and Δ is a neighborhood of μ such that μ is the only eigenvalue of A in Δ, then there exists a positive integer $\bar{N}$ such that, for any $N \geq \bar{N}$, A_n has in Δ exactly ν eigenvalues $\mu_{N,j}, j \in 1,\dots,\nu$, counting their multiplicities. Moreover, by setting $\varepsilon_N \coloneqq \left\|(A_N - A)|_{\varepsilon_\mu}\right\|_{\mathcal{U}\leftarrow\varepsilon_\mu}$, where ε_μ is the generalized eigenspace of μ equipped with the norm $\|\cdot\|_{\mathcal{U}}$ restricted to ε_μ, the following holds:*

$$\max_{j\in\{1,\dots,\mu\}} |\mu_{N,j} - \mu| \in O(\varepsilon_N^{1/l}).$$

Proof. By [5, Example 3.8 and Theorem 5.22], the norm convergence of A_N to A implies the strongly stable convergence $A_N - \mu I_{\mathcal{U}} \overset{ss}{\to} A - \mu I_{\mathcal{U}}$ for all μ in the resolvent set of A and all isolated eigenvalues μ of finite multiplicity of A. The thesis follows then by [5, Proposition 5.6 and Theorem 6.7]. □

The ascent denotes the maximum dimension of the Jordan blocks relevant to μ; if $\mu \coloneqq R_0$, we know from the compactness of K that $l = 1$.

So the following holds:

Theorem 5.5. *Suppose that Theorem 5.3 holds for $\nu = 0$, with $\kappa_0(n) \to 0$ for $n \to +\infty$; assume also that Assumptions 5.1 and 5.2 hold, and there are no boundary conditions.*

Suppose that $\varepsilon_{R_0} \subseteq C^{\nu}_{N_0}$).

Then, $\|R_{0,n} - R_0\| \in O(\kappa(n))$.

Proof. We apply Lemma 5.4 with $\mathcal{U} := C^0([0, l])$, $A_N := K_n = B_n(M_n)^{-1}$; thanks to Theorem 5.3 $\varepsilon_N \in O(\kappa(N))$ and so the thesis is proved. □

Finally, we prove that in some cases Assumption 5.1 holds.

Lemma 5.6. *Let J_n, and P_n defined as in the previous chapter, fix ν and let $\Delta a := \max_{0\le k<N} \Delta a_k \in O(g(n))$ for some $g(n)$ such that $\lim_{n\to\infty} g(n) = 0$. Then*

- $\|P_n\|_{X_n \leftarrow C} \in O(1)$
- $\|J_n\|_{C \leftarrow X_n} \in O(1)$
- $\|J_n P_n - I_C\| \in O((g(n))^{\nu})$

Proof.

- Let $\psi \in X$, with $\|\psi\| \le 1$. Then $\|P_n\psi\| = \max_{j=0}^{n} |\psi(p_j)| \le 1$, and the first point is proved.
- $\|J_n\|$ is related to the Lebesgue constant:

$$\|J_n\|\Psi = \max_{1\le k\le N} \sup_{x\in[a_{k-1}, a_k]} \sum_{j=0}^{\nu} |l_{k\nu+k+j}(t)| = \max_{1\le k\le N} \Lambda_{\nu}(p_{(k+1)\nu}, \dots, p_{(k+2)\nu});$$

  since ν is fixed, this is bounded, and thus the second point is proved.

- Let $\psi \in C$, with $\|\psi\| \le 1$; then $J_n P_n \psi$ is the piecewise polynomial of $\Pi_{N,\nu}$ which interpolates ψ; we know that the error is in $O((\Delta a))^{\nu} = O((g(n))^{\nu})$ (see [22, Eq 7.21]). □

If $g(n) \to 0$ for $n \to +\infty$, it also holds, for $C = C^0$, $\|J_n P_n - I_{C^0}\| \to 0$ for $n \to +\infty$, and we can thus apply Theorem 5.5.

5.2 Generic disease

We will now prove that part of Assumption 5.2 holds.

Recall that β, S^0, γ, μ are continuous on a compact set, thus they are uniformly bounded; we call $\beta_+, S^0_+, \gamma_+, \mu_+$ their maximum.

We also recall that $\|\varphi\| = \|\varphi\|_{\infty} = \sup_{a\in[0, a_\dagger]} |\varphi(a)|$. We can then state

Lemma 5.7. *Let B, M be defined as in Eqs. (3.6a) and (3.6b); then B and M^{-1} are bounded .*

Proof. Let $\varphi \in C^0([\,0,\, a_\dagger\,])$; then $\|B\varphi\| \leq a_\dagger S^0_+ \beta_+ \|\varphi\|$, thus B is bounded.

Using the expression for the inverse of M (Eq. (3.8)), we write

$$M^{-1}\zeta(\sigma) = \int_0^\sigma \mathrm{e}^{-\int_\rho^\sigma (\mu+\gamma)(\eta)\mathrm{d}\eta}\zeta(\rho)\mathrm{d}\rho \leq a_\dagger \|\zeta\| \qquad \square$$

In Section 4.3.1 we also left the following lemma to prove:

Lemma 5.8. *Let $\Sigma\phi(a) \coloneqq \int_0^l \beta(a,\sigma)\phi(\sigma)\mathrm{d}\sigma$, and suppose that for each $a \in [\,0,\, a_\dagger\,]$ it holds*

$$\beta(a,\sigma)\phi(\sigma) \in C^\nu_{N_0}$$

as a function of σ.

Let a_j be the nodes of the external mesh and p_i those of the full mesh, as defined in Chapter 4.

Define a piecewise quadrature formula $I_\nu\phi(a) \coloneqq \sum_{j=0}^{N-1} I_{j,\nu}\phi(a)$, with

$$\int_{a_j}^{a_{j+1}} \beta(a,\sigma)\phi(\sigma)\mathrm{d}\sigma \simeq I_{j,\nu}\phi(a) \coloneqq \sum_{i=j(\nu+1)}^{j(\nu+1)+\nu} q_i \beta(a,p_i)\phi(p_i),$$

where $q_i \coloneqq \int_{a_j}^{a_{j+1}} l_i(a)\mathrm{d}a$ are the quadrature weights.

Then $\|\Sigma\phi(a) - I_\nu\phi(a)\| \in O((\Delta a)^\nu)$.

Proof. We start with a single interval:

$$\left|\int_{a_j}^{a_{j+1}} \beta(a,\sigma)\phi(\sigma)\mathrm{d}\sigma - \sum_{i=j(\nu+1)}^{j(\nu+1)+\nu} q_i\beta(a,p_i)\phi(p_i)\right| \leq$$

$$\leq \int_{a_j}^{a_{j+1}} \left|\beta(a,\sigma)\phi(\sigma) - \sum_{i=j(\nu+1)}^{j(\nu+1)+\nu} l_i(\sigma)\beta(a,p_i)\phi(p_i)\right|\mathrm{d}\sigma \leq$$

$$\leq \Delta a_j \sup_{\sigma\in[\,a_j,\,a_{j+1}\,]} \left|\beta(a,\sigma)\phi(\sigma) - \sum_{i=j(\nu+1)}^{j(\nu+1)+\nu} l_i(\sigma)\beta(a,p_i)\phi(p_i)\right| \in O((\Delta a_j)^{\nu+1})$$

Thanks to the properties of the interpolating polynomial ([22, Eq 7.21]).

Then, taking the sum over j, we get $\|\Sigma\phi(a) - I_\nu\phi(a)\| \in O((\Delta a)^\nu)$, and the thesis is proved. $\square$

5.3 Bacteria

Like in the previous case, we will now prove that part of Assumption 5.2 holds.

Recall that β, μ are continuous on a compact set, thus they are uniformly bounded; we call $\beta_+, S^0_+, \gamma_+, \mu_+$ their maximum.

We also recall that $\|\varphi\| = \|\varphi\|_\infty = \sup_{a\in[\,0,\,a_\dagger\,]}|\varphi(a)|$. We can then state the following

Lemma 5.9. *Let B defined as in Eqs. (3.11a) and (3.11b); then B is bounded*

Proof. Let $\varphi \in C^0([\,0,\, l\,])$; then $\|B\varphi\| \leq \beta_+ \|\varphi\|$, thus B is bounded.

5.4 Remarks on the missing hypotheses

Chapter 6

Simulation results

We tested our method on the models presented in Chapter 3, using different parameters.

All graphs in this chapter plot, in logarithmic scale on both axes, the absolute error in the computation of R_0, i.e. the difference between the current computed value and a reference value which is usually the value computed by the "spectral" method with the highest-degree polynomial used, because the results obtained suggest it is correct up to machine precision.

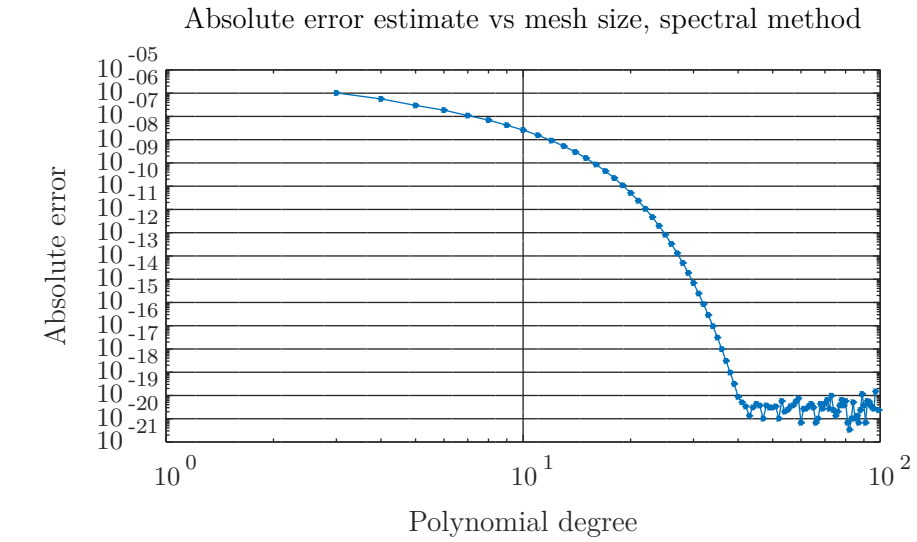
6.1 Sample functions

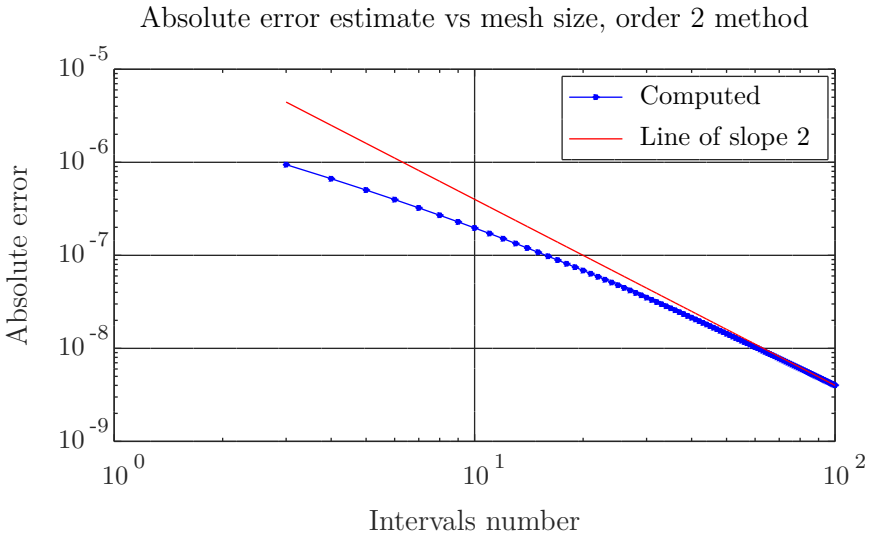
As a first test we choose the model described by Eq. (3.5), using the following parameters:

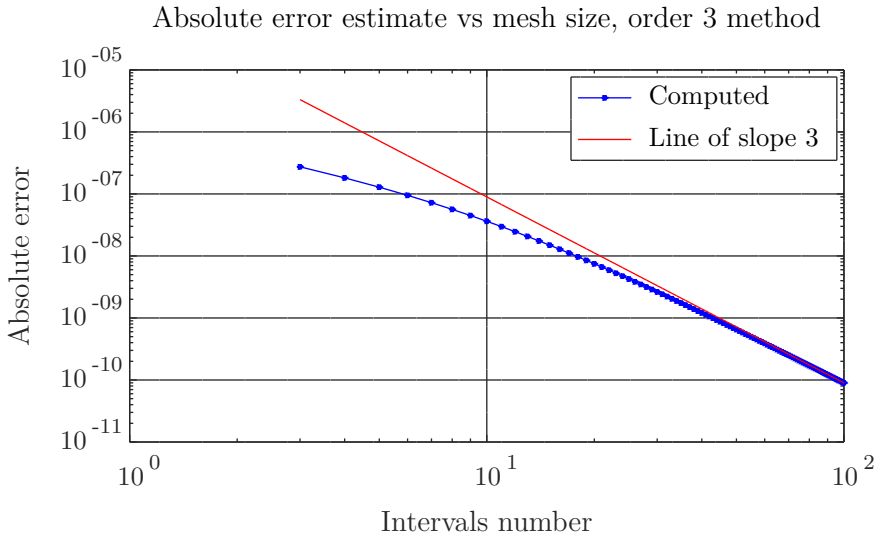
$$
\begin{aligned}
a_\dagger &= 1 \\
S^0(x) &= 5187 + 226.438x - 2.777x^2 \\
\gamma(x) &= 52 \\
\mu(x) &= \frac{8.3675}{110 - x} \\
\beta &= 1,8 \cdot 10^{-11} \left(100^2 - (x1 - x2)^2\right) + 3 \cdot 10^{-8}.
\end{aligned}
$$

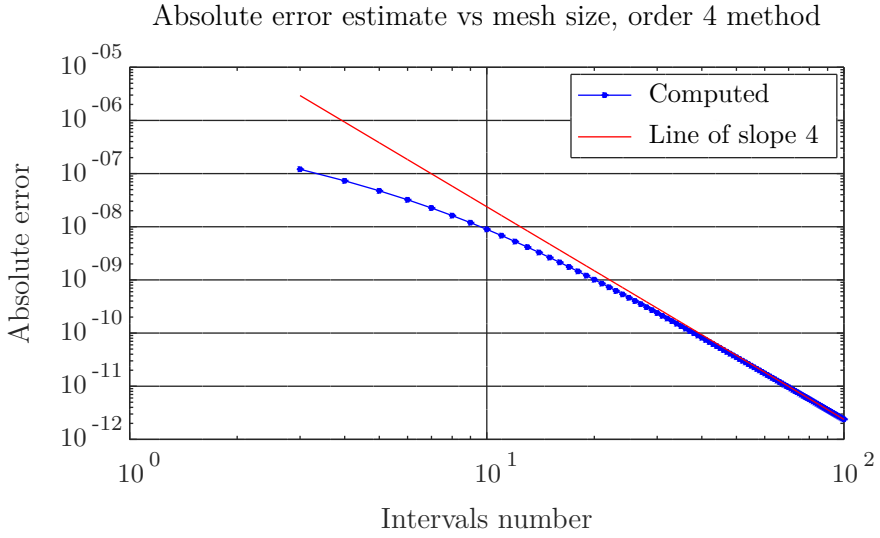
Here γ, β are taken from [16]; μ, S^0 are constructed as analytic functions which are not too far from the experimental data [20, 21]; finally, the maximum age $a_\dagger$ is set to 1, in order to get a simple problem, since the error also depends from the domain size.

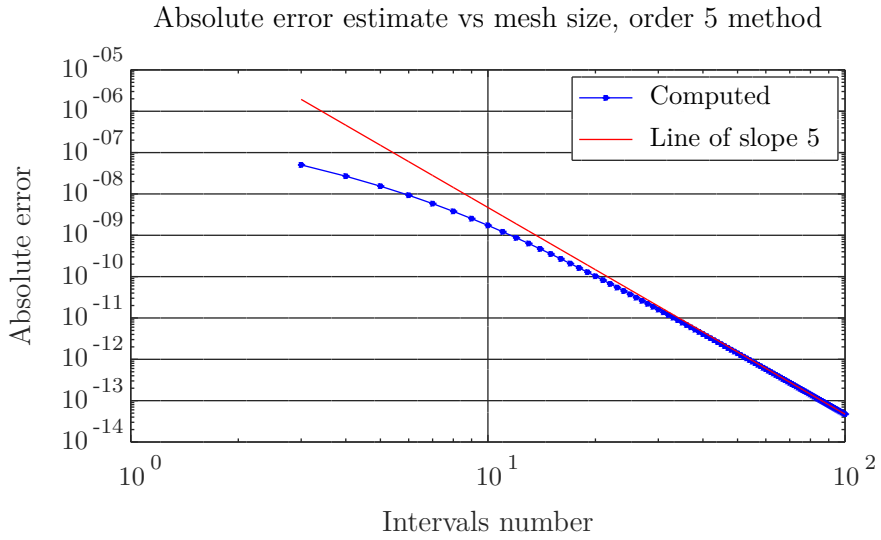
The first graph is obtained using the spectral method, i.e. a single polynomial of increasing degree is approximating the eigenvector. The obtained reference value is $R_0 \simeq 2.095 \cdot 10^{-5}$. The following graphs are obtained through a "finite elements" like method: the polynomial degree is kept fixed and the number of intervals is increased.

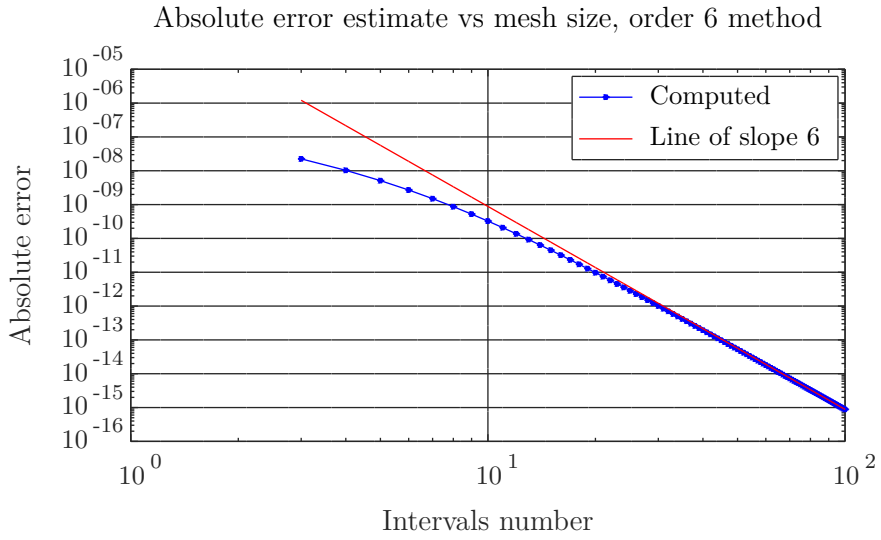

Absolute error estimate vs mesh size, spectral method
Absolute error
Polynomial degree


Absolute error estimate vs mesh size, order 2 method
Computed
Line of slope 2
Absolute error
Intervals number


Absolute error estimate vs mesh size, order 3 method
Absolute error
Intervals number
Computed
Line of slope 3
10 -05
10 -06
10 -07
10 -08
10 -09
10 -10
10 -11
10 0
10 1
10 2


Absolute error estimate vs mesh size, order 4 method
Absolute error
Intervals number
Computed
Line of slope 4
10 -05
10 -06
10 -07
10 -08
10 -09
10 -10
10 -11
10 -12
10 0
10 1
10 2





These graphs show that the algorithm produces the expected results: the spectral method reaches machine precision when the polynomial of degree is about 40; the finite elements methods shows the asymptotic expected behavior, approaching an error of $O(n^{-k})$, which in a log-log plot is a straight line of slope k.

However, they are all far from reaching the machine precision (10^{-20}, since R_0 itself is of the order of 10^{-5}), even if the results from methods of order can still be considered quite accurate results.

This confirms that the spectral method is the best choice for analytic eigenvectors, in the sense that it reaches the lowest possible error at parity of computational resources, i.e. time and memory: let n be the number of intervals times the degree of the polynomial; the computing time is mainly due to the $O(n^3)$ cost of solving

the generalized eigenvalues problem, and the memory is maily used to store the matrices M_n and B_n; then the spectral method reached machine precision with $n \simeq 40$, while for the similar values of n the order 2 method produced an error of about 10^{-7} (meaning two correct decimal digits, obtained with $N = 13, \nu = 3$), and the method of order 6 produced an error of about 10^{-9} (four correct decimal digits, with $N = 6, \nu = 7$); to better compare the resources requirement, we add that the order 6 method reaches an absolute precision of 10^{-15}, which as said in quite an accurate result, for $n = 701$, that is using about 300 times the memory and 5000 times the time needed by the spectral method to reach full precision.

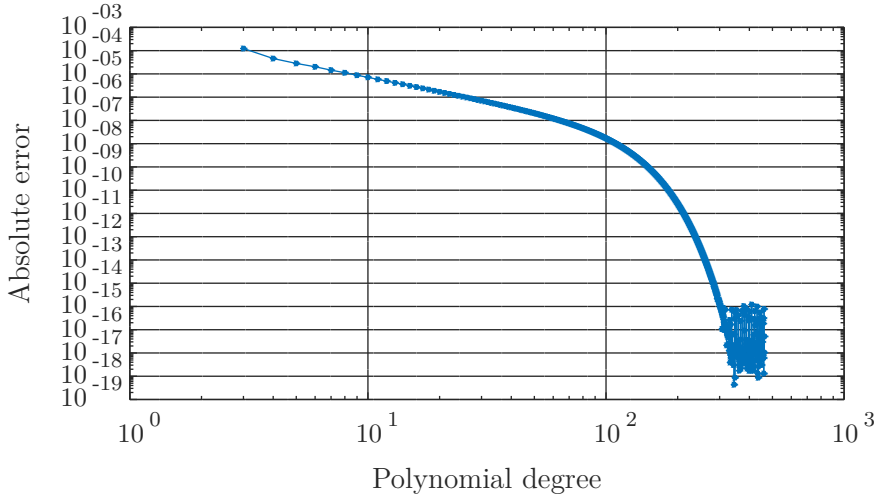
6.2 Generic disease: test using analytic functions

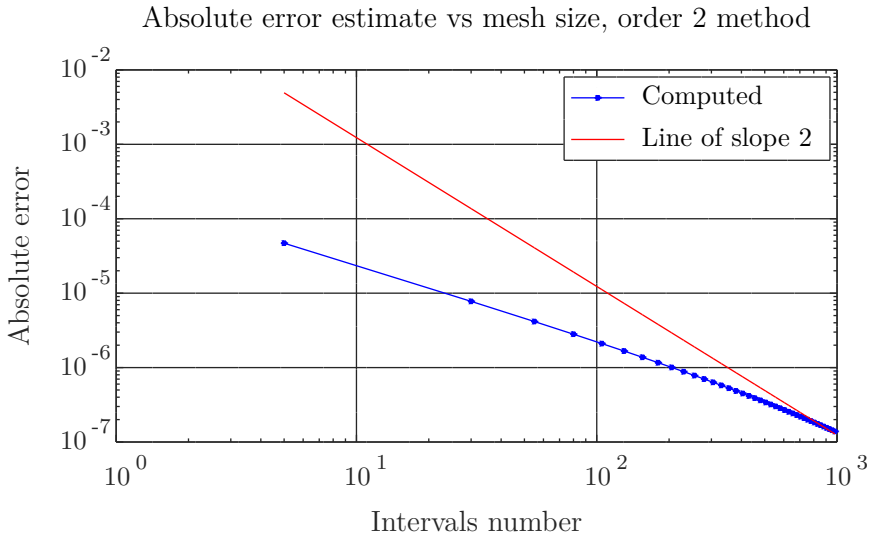
We now wan to test our method to a more realistic example: we use the same parameters as before, but now the maximum age is set to 100, which is more realistic if we are talking about a human population of susceptibles, as defined in Section 3.2

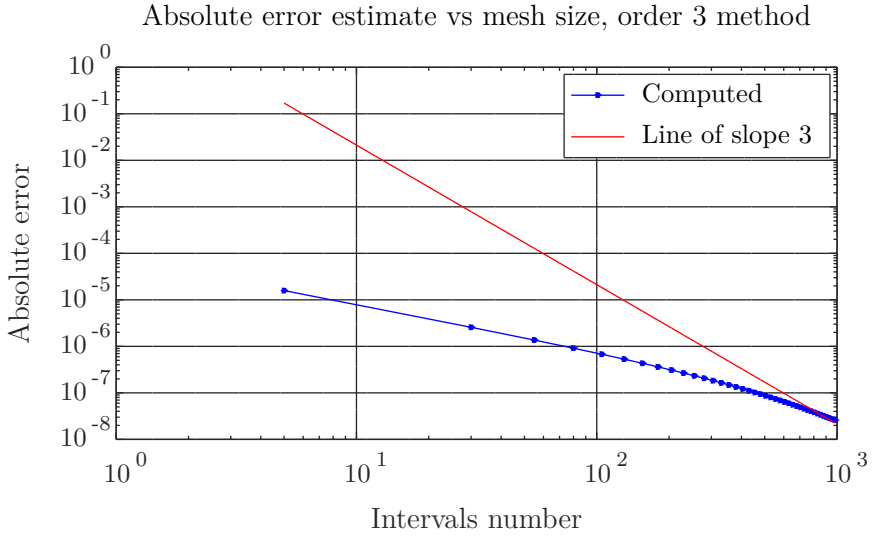
$$
\begin{aligned}
a_\dagger &= 100 \\
S^0(x) &= 5187 + 226.438x - 2.777x^2 \\
\gamma(x) &= 52 \\
\mu(x) &= \frac{8.3675}{110 - x} \\
\beta &= 1,8 \cdot 10^{-11} \left(100^2 - (x1 - x2)^2\right) + 3 \cdot 10^{-8}
\end{aligned}
$$

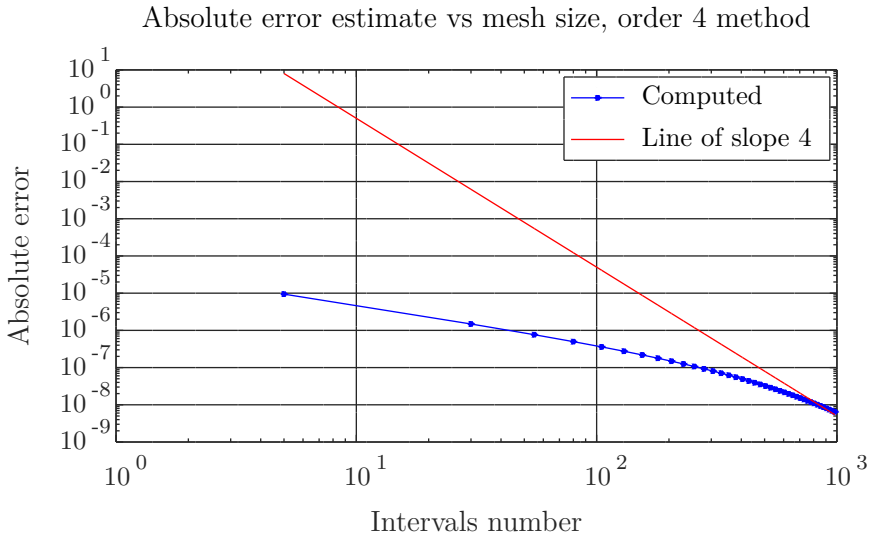
Again, the first graph is obtained using the "spectral method". The obtained reference value is $R_0 = 2.636 \cdot 10^{-3}$. The following graphs are obtained through a "finite elements"method.

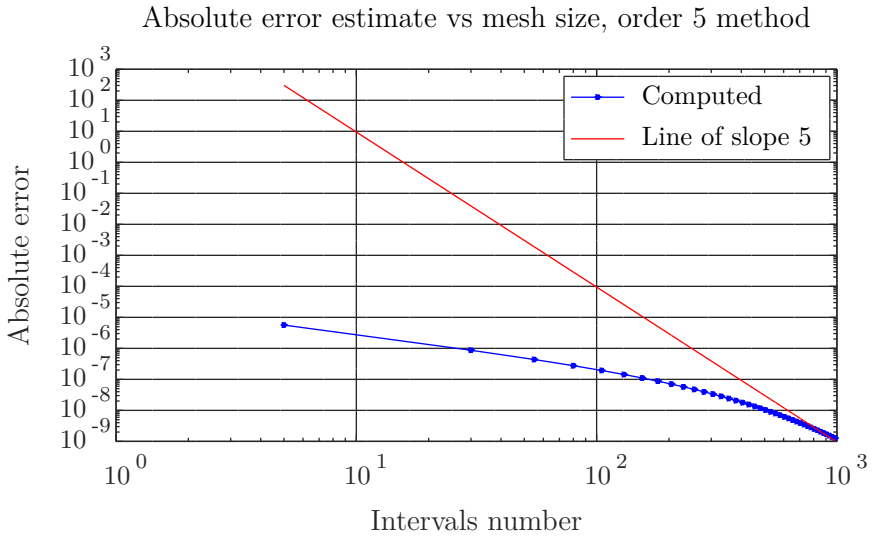
Absolute error estimate vs mesh size, spectral method




Absolute error estimate vs mesh size, order 2 method
Absolute error
Intervals number
Computed
Line of slope 2
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^0
10^1
10^2
10^3


Absolute error estimate vs mesh size, order 3 method
Absolute error
Intervals number
Computed
Line of slope 3
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^0
10^1
10^2
10^3





Again, the spectral method reaches machine precision, although this time a polynomial of degree 200 is needed; however the finite elements methods do not show the asymptotic expected behavior; this is due to the large maximum age, which gives a higher multiplicative constant in the error behavior $O(n^{-k})$ (see also the proof of Lemma 5.7, although not all the hypothesis of that lemma are verified); as we can see the slope of the blue line increases with the intervals number, which means that the asymptotic behavior will be reached with more intervals (and computing time: this is the reason why we limited to 1000 intervals, since whole order 4 graph took more than an hour on a 3.2 GHz CPU, and 6 GB of RAM, which is far more than needed).

6.3 Generic disease

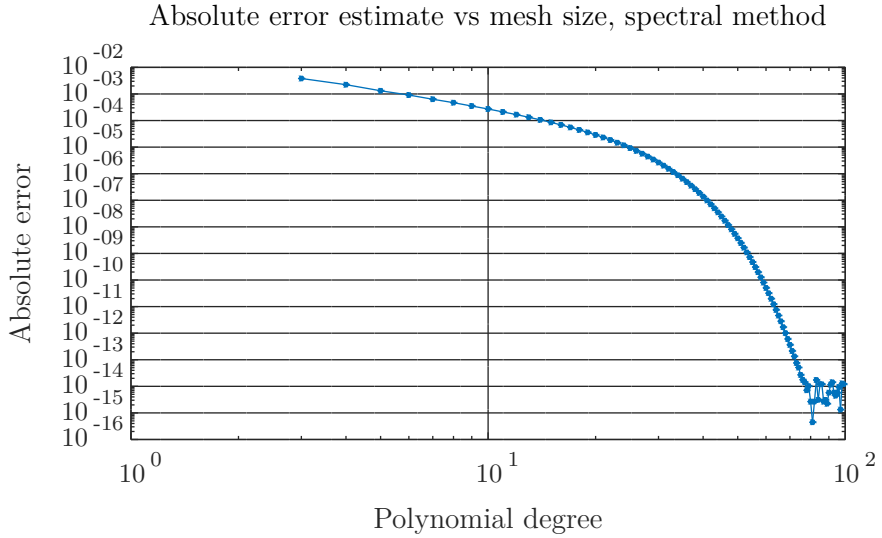
Finally, as in [16] we tested our code on a real data set: we took the values of S^0 from [21], and the values of μ from [20], taking a value every 5 years in order to match the same sampling of S^0; both these data were interpolated by cubic splines to obtain the actual functions.

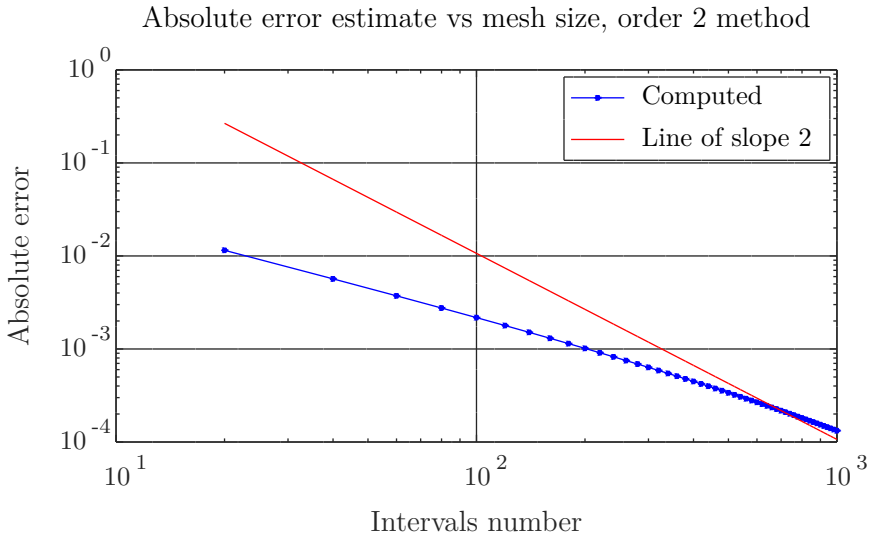
We kept the same maximum age (as suggested by [21]), and following [16] the same values for γ and β, which model the disease:

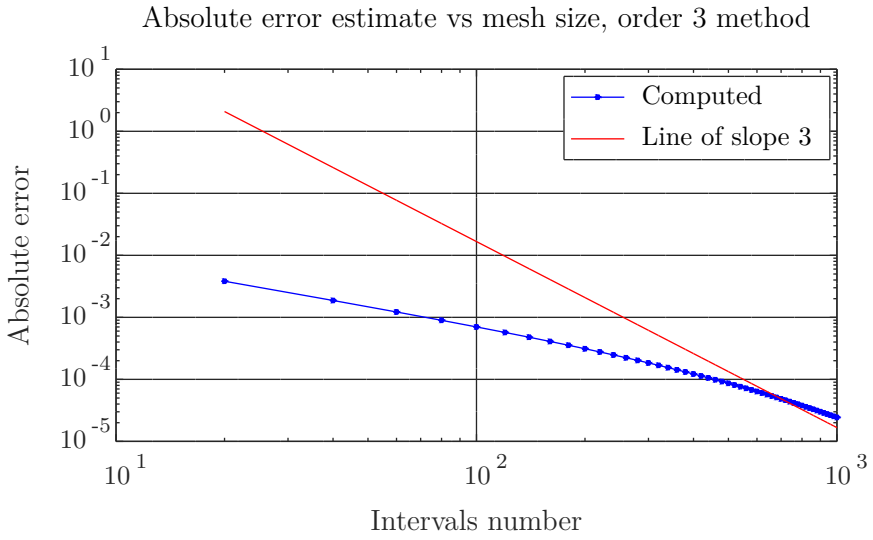
$$a_\dagger = 100$$
$$\gamma(x) = 52$$
$$\beta = 1,8 \cdot 10^{-11} \left(100^2 - (x1 - x2)^2\right) + 3 \cdot 10^{-8}.$$

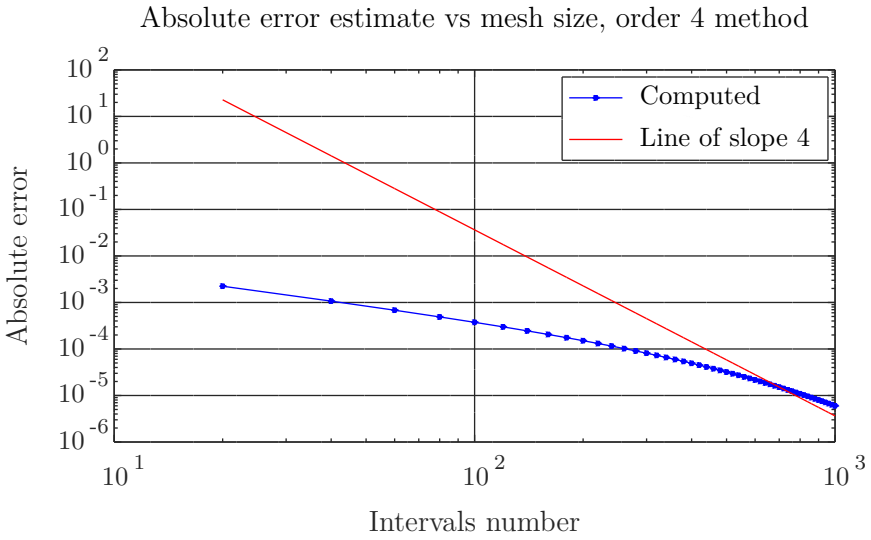
In this case (some of) the functions are not analytic; they are instead piecewise analytic, on intervals of 5 years. We thus fixed 20 intervals instead of one for the spectral method, and used a number of intervals multiple of 20 for the finite elements method.

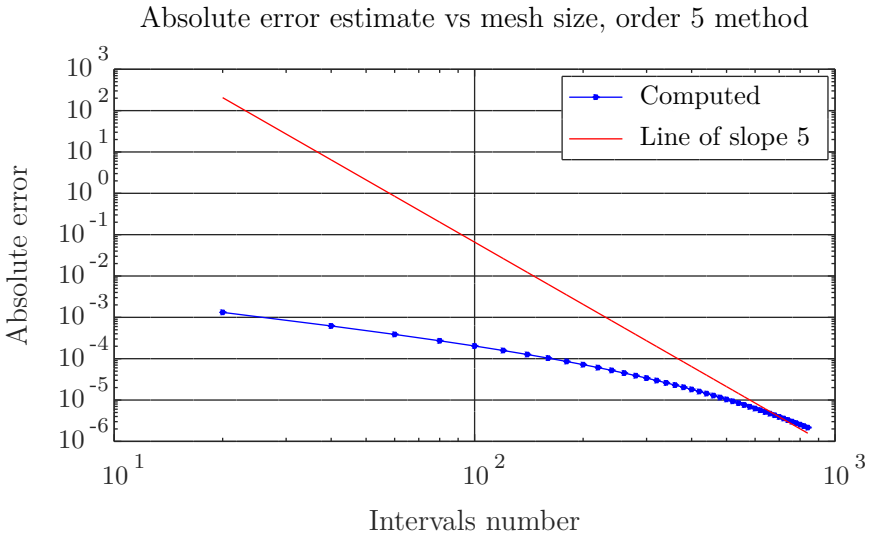
Again, the first graph is obtained using the "spectral method"; The obtained reference value is $R_0 = 2.286$. The following graphs are obtained through a "finite elements"method.




Absolute error estimate vs mesh size, order 2 method
Absolute error
Intervals number
Computed
Line of slope 2
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{1}
10^{2}
10^{3}


Absolute error estimate vs mesh size, order 3 method
Absolute error
Intervals number
Computed
Line of slope 3
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{1}
10^{2}
10^{3}





The same conclusions of the previous case are valid: the polynomial degree needed for convergence is about 80, but using 20 distinct polynomials; the finite elements methods do not show the asymptotic expected behavior yet, but, like in the previous case, the slope of the blue line is increasing, i.e. the convergence is most likely still slow because of the high maximum age, and the computing times for the finite elements are almost the same.

Nevertheless, not only the spectral method, but also high order methods still compute an acceptable approximation, given that the statistical data are not exact (we used interpolating splines), hence augmenting the precision above a given threshold does not give more accurate information.

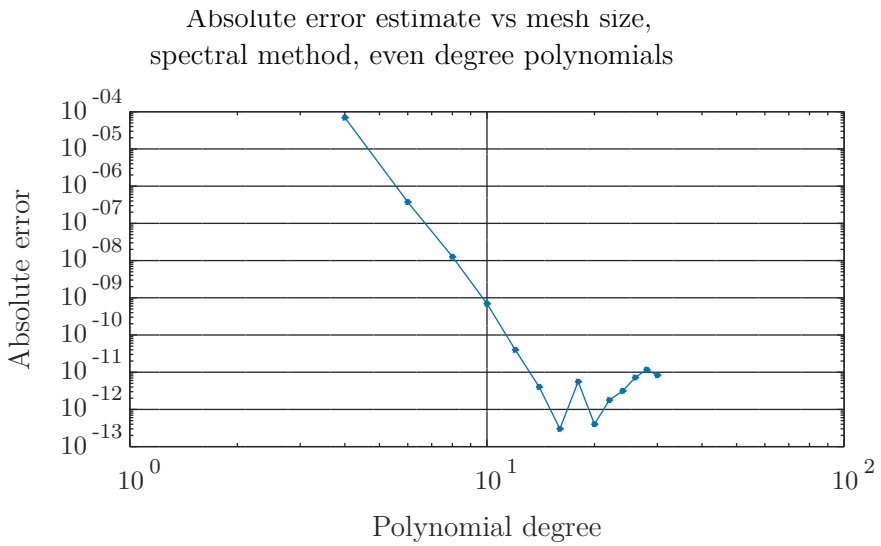
6.4 Bacteria

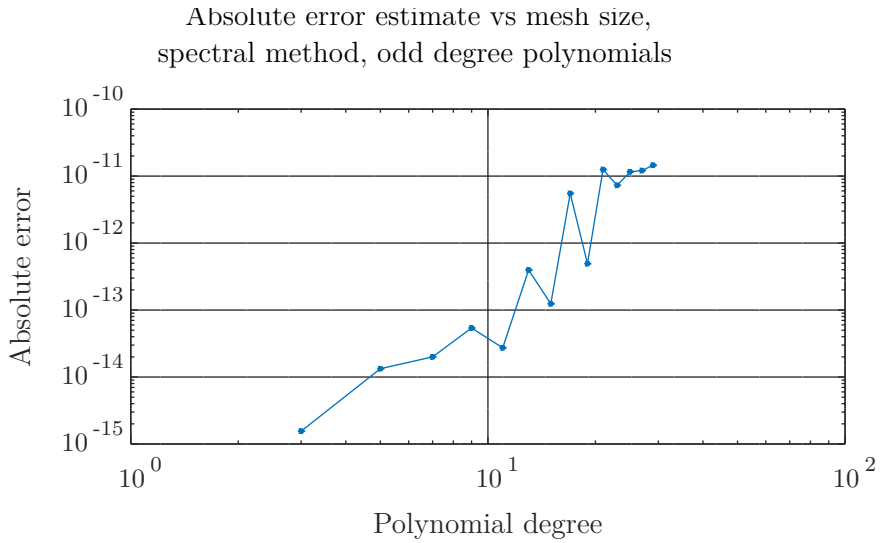
We chose these functions for testing:

$$
\begin{aligned}
l &= 1 \\
\beta(x) &= 12 \\
\mu(x) &= 1 \\
c(x) &= 2 + x \\
D(x) &= 0.5 + x.
\end{aligned}
$$

The choice of constant functions for β and μ let us compute R_0 exactly, given from Eq. (3.14); for this reason we used the exact value 24/13 as the reference value for R_0.

What follows is the plot for the spectral method; it still exhibits spectral accuracy [24, Chapter 4], but this time the approximation with odd degree polynomials is always exact to machine precision, and there seems to be more instability for high degree polynomials (for the sake of clariny, we split the plot based on he parity of the used polynomial degree).





We will not show the graphs produced by the finite elements method since those methods seem not to converge for even degree polynomials. Investigating the reason why it is so is left as a future work.

Appendix A

Implementation

The code has been written in C++, using the Gnu Scientific Library, and is available at https://github.com/f-florian/thesis and https://github.com/f-florian/thesis-differential, except that for plotting graphs which consist of trivial `Octave` scripts not deserving particular attention.

We discuss here some numerical problems which are not strictly related to the approximation of R_0 (i.e., their solutions also have other applications), but are nevertheless to be solved in order to implement the algorithm described in the previous chapters.

For the sake of clarity we limit this discussion to one sample interval; since it can be replicated (with proper scaling) in any interval of the external mesh this is not a limitation[1].

A.1 Nodes, interpolation, quadrature and differentiation

The choice has been made, of using the same nodes for interpolation, quadrature and differentiation. This is not the only possible choice: we can write a polynomial interpolating the function f on some nodes p as

$$f_n(x) \coloneqq \sum_{i=0}^{n} f(p_i) l_i(x), \tag{A.1}$$

where we used the Lagrange basis; then the integral of f could be approximated by that of f_n, using a quadrature formula on a different set of nodes b:

$$\sum_{j=0}^{n_q} \left(\tilde{q}_j \sum_{i=1}^{n} f(p_i) l_i(b_j) \right); \tag{A.2}$$

[1] In fact, also the code in the `thesis-differential` library has been written for a single interval.

for the derivative in a point y we could use yet another set of nodes c_y, obtaining a similar formula, with quadrature weights replaced by differentiation weights:

$$\sum_{j=0}^{n_y} \left(\tilde{d}_{y,j} \sum_{i=1}^{n} f(p_i) l_i(c_{y,j}) \right). \tag{A.3}$$

However, using the same nodes p also for quadrature and differentiation leads to much simpler formulas:

$$\sum_{j=0}^{n} q_j f(p_j), \tag{A.4}$$

and the same for the derivative at one of the points p, changing only the weights.

A.1.1 Polynomial basis

Also, we are not even forced to use the Lagrange basis of polynomials; on the contrary it might be thought that a Newton basis is better suited for numerical computation because the polynomial evaluations are then more accurate.

The first problem is that in order to compute the coefficients for the representation of a polynomial in the Newton basis it is necessary to solve a triangular linear system, (which is equivalent to computing the divided differences tableau); even worse, if the polynomial is the interpolation of the eigenvector, i.e. it is an unknown, we need to invert the triangular matrix

$$\begin{pmatrix} 1 & & & \\ 1 & x_1 - x_0 & & \\ \vdots & \vdots & \ddots & \\ 1 & x_n - x_0 & \dots & \prod_{j=0}^{n-1}(x_k - x_j) \end{pmatrix}$$

and perform a matrix-matrix multiplication. This last problem could be solved using the same nodes for interpolation, quadrature and differentiation, and using weights which allows using the coefficients instead of the polynomial values as unknowns.

However using barycentric representation of Lagrange polynomials is even simpler, and yields accurate results for both evaluating the polynomials and computing the differentiation weights, as shown in [3]; as for the quadrature weights, the are commonly computed so that they can be used as in Eqs. (A.2) and (A.4), so the choice of Lagrange polynomial is again simpler (and accurate enough).

A.1.2 Lagrange polynomials in barycentric form

So we write the generic polynomial of the Lagrange basis, showing how to obtain the barycentric formulation:

$$l_j(x) = \frac{\prod_{k=0,k\neq j}^{n}(x - p_k)}{\prod_{k=0,k\neq j}^{n}(p_j - p_k)}; \tag{A.5}$$

since the denominator does not depend on the evaluations point, and we can can collect

$$l(x) = \prod_{k=0}^{n} (x - p_k)$$

(which does not depend on the index j) from the numerator, we define the barycentric weights

$$w_j := \prod_{k=0, k\neq j}^{n} (p_j - p_k)$$

and write

$$l_j(x) = \begin{cases} \frac{l(x)}{(x-p_j)w_j} & x \neq p_j \\ 1 & x = p_j. \end{cases} \tag{A.6}$$

Interpolating the function 1 we get

$$1 = \sum_{i=0}^{n} l_i(x) = l(x) \sum_{i=0}^{n} \frac{1}{(x - x_i)w_i} \tag{A.7}$$

so dividing by 1 in Eq. (A.6) we get

$$l_j(x) = \frac{\frac{1}{(x - x_j)w_j}}{\sum_{j=0}^{n} \frac{1}{(x - x_j)w_j}} \tag{A.8}$$

for $x \neq p_i$, and 1 otherwise.

The main advantage of this formulation is that after the weights w_j have been computed in $O(n^2)$ operations, evaluating a function f_n by

$$f_n(x) = \frac{\sum_{j=0}^{n} \frac{f(p_j)}{(x - x_j)w_j}}{\sum_{j=0}^{n} \frac{1}{(x - x_j)w_j}}$$

requires only $O(n)$ operations (including the check that x is not one of the nodes) and the computation quite accurate.

From now on we therefore assume we have chosen the Lagrange basis and the same set of nodes for interpolation, quadrature and differentiation.

A.1.3 Differentiation weights

We want to approximate $f'(x)$, using only the evaluations of f in the nodes p; we note that, in order to write the birth and mortality operators we only need this

when $x = p_k$ for some index k; so we can write, using Eq. (A.1) and linearity of the derivative

$$f'(p_k) \simeq f'_n(p_k) = \sum_{i=0}^{n} f(p_i) l'_i(p_k).$$

We thus need to compute the derivatives of the polynomials in the Lagrange basis at those points; in order to do this we start from Eq. (A.5); since we are differentiating a product we get

$$l'_i(x) = \frac{1}{w_i} \sum_{m=0, m\neq i}^{n} \prod_{j=0, j\neq m,i}^{n} (x - p_j)$$

and so, for $k \neq i$

$$l'_i(p_k) = \frac{1}{w_i} \sum_{m=0, m\neq i}^{n} \prod_{j=0, j\neq m,i}^{n} (p_k - p_m) = \frac{1}{w_i} \prod_{j=0, j\neq k,i}^{n} (p_k - p_j) = \frac{w_k}{(p_k - p_i) w_i}.$$

For $k = i$ we differentiate Eq. (A.7) and get

$$0 = \sum_{j=0}^{n} l'_j(x)$$

and finally

$$l'_i(p_i) = - \sum_{j=0, j\neq i}^{n} l'_j(p_i)$$

A.1.4 Nodes and quadrature weights

We were able to obtain an analytic expression for the derivative of the Lagrange polynomials. For quadrature rules things get more complicated, so we limit to Gauss and Chebyshev nodes, like in Chapter 3 (this means that we only use those nodes for differentiation too).

For Gauss nodes, the algorithm given by Golub and Welsch in [10] gives nodes and weights of quadrature computing eigenvalues and eigenvectors of a tridiagonal matrix. The article by Golub and Welsch describes the tridiagonal matrix for an integral which uses a general weight function; the particular matrix we need is described in the function `gauss` in [25][2]. Once the nodes and weights have been computed, the integral evaluation is exact for all polynomials of degree at most $2n+1$, which is the greatest attainable polynomial order: in this sense Gauss quadrature is often considered the "optimal" one.

For Chebyshev nodes, the quadrature rule is called "Clenshaw-Curtis" quadrature, and has polynomial order n, i.e. it can integrate exactly polynomials up to degree n; [25] gives an algorithm for computing the quadrature weights on these nodes.

[2] We actually used the Gauss weights and nodes provided by the Gnu Scientific Library, without the need of re-implement the method.

A.2 Eigenvalues computation

Once the matrices approximating the birth and mortality operators have been written, we have to solve the eigenvalues system; as already pointed out Eqs. (3.3) and (3.4) are mathematically equivalent, but their numerical properties may differ; the first was solved by inverting a matrix and then computing the Schur decomposition $BM^{-1} = ZTZ^T$; the diagonal of T gives the eigenvalues. The second is solved using the so called QZ method to compute the generalized Schur decomposition $B = QTZ^T$, $M = QSZ^T$ [9, Sections 15.3 and 15.6].

Surprisingly, the two methods seem to produce similar results (although some simulations have been performed using only the second method); the matrix inversion was performed by computing a QR decomposition first, in order to get more accurate results.

Finally, since we are only interested in the eigenvalue of greatest magnitude, we could also use a method which only find a subset of the eigenvalues, like the power method, possibly leading to an execution time speedup; the opportunity of doing so has not been investigated and is left as a possible future improvement.

Bibliography

[1] C. D. Aliprantis and O. Burkinshaw. *Positive Operators*. Academic Press, 1985.

[2] C. Barril, À. Calsina, and J. Ripoll. "A practical approach to R0 in continuous-time ecological models". In: *Math Meth Appl Sci* (2017), pp. 1–14. DOI: https://doi.org/10.1002/mma.4673.

[3] J.-P. Berrut and L. N. Trefethen. "Barycentric Lagrange Interpolation". In: *SIAM Review* 43.3 (2004-09), pp. 501–517. DOI: 10.1137/S0036144502417715. URL: https://www.jstor.org/stable/20453536.

[4] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York: Springer, 2011. DOI: 10.1007/978-0-387-70914-7.

[5] F. Chatelin. *Spectral Approximation of Linear Operators*. Philadelphia: Society for Industrial and Applied Mathematics, 2011. DOI: 10.1137/1.9781611970678.

[6] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations". In: *Journal of Mathematical Biology* 28 (1990), pp. 365–382.

[7] K.-J. Engel and R. Nagel. *One-Parameter Semigroups for Linear Evolution Equations*. New York: Springer, 2000.

[8] L. Euler. *Introductio in analysin infinitorum*. Vol. 1. 1748. URL: http://www.17centurymaths.com/contents/introductiontoanalysisvol1.htm.

[9] M. Galassi et al. *GNU Scientific Library*. 2017. URL: https://www.gnu.org/software/gsl/doc/latex/gsl-ref.pdf.

[10] G. H. Golub and J. H. Welsch. "Calculation of Gauss Quadrature Rules". In: *Math. Comp.* 23 (1969), pp. 221–230. DOI: 10.1090/S0025-5718-69-99647-1.

[11] J. A. P. Heesterbeek and K. Dietz. "The concept of R_0 in epidemic theory". In: *Statistica Neerlandica* 50.1 (1996), pp. 89–110.

[12] J. Heesterbeek. "A brief history of R_0 and a recipe for its calculation". In: *Acta Biotheoretica* 50 (2002), pp. 189–204.

[13] H. Inaba. "The Malthusian parameter and R_0 for heterogeneous populations in periodic environments". In: *Mathematical biosciences and engineering : MBE* 9.2 (2012-04), pp. 313–346. DOI: 10.3934/mbe.2012.9.313. URL: https://www.researchgate.net/publication/230696233_The_Malthusian_parameter_and_R_0_for_heterogeneous_populations_in_periodic_environments.

[14] W. O. Kermack and A. G. McKendrick. "A Contribution to the Mathematical Theory of Epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772 (1927), pp. 700–721. ISSN: 09501207. URL: https://www.jstor.org/stable/94815.

[15] M. Krein and M. Rutman. *Linear operators leaving invariant a cone in a Banach space*. Vol. 26. 1950.

[16] T. Kuniya. "Numerical approximation of the basic reproduction number for a class of age-structured epidemic models". In: *Applied Mathematics Letters* 73.Supplement C (2017), pp. 106–112. ISSN: 0893-9659. DOI: https://doi.org/10.1016/j.aml.2017.04.031. URL: http://www.sciencedirect.com/science/article/pii/S0893965917301568.

[17] D. Liessi. "Pseudospectral methods for the stability of periodic solutions of delay models". PhD thesis. 2018.

[18] W. Luxemburg and A. Zaanen. *Riesz Spaces*. Vol. 1. London: Elsevier, 1971.

[19] T. Malthus. *An Essay on the Principle of Population*. London: J. Johnson, in St. Paul's Church-yard, 1798. URL: www.esp.org/books/malthus/population/malthus.pdf.

[20] Ministry of Health, Labour and Welfare, Japanese Government. *The 22nd Life Tables*. 2015. URL: http://www.mhlw.go.jp/english/database/db-hw/lifetb22nd/index.html.

[21] Ministry of Internal Affairs and Communications. *Population by Age (5-Year Age Group) and Sex, Monthly Estimates - Total population, Japanese population, the First Day, Each Month*. Data from December 2015. 2016-11-25. URL: https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&toukei=00200524&bunya_l=02&tstat=000000090001&cycle=7&year=20150&month=0&tclass1=000001011679&stat_infid=000031495570&result_back=1&cycle_facet=cycle&second2=1.

[22] A. Quarteroni, R. Sacco, and F. Saleri. *Matematica numerica*. Seconda edizione. Milano: Springer, 2002.

[23] R. Ross. *The prevention of malaria*. London: John Murray, 1911. URL: http://krishikosh.egranth.ac.in/handle/1/2047440.

[24] L. N. Trefethen. *Spectral Methods in Matlab*.

[25] L. N. Trefethen. "Is Gauss quadrature better than ClenshawCurtis?" In: ().