Bayesian Evidence Synthesis for the Analysis of Biomedical Data

by

Anastasios Apsemidis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in fulfilment of the requirements for

the PhD in Statistics

Athens, Greece

April 2024

Μπεϋζιανή Σύνθεση Πληροφορίας για την Ανάλυση Βιοϊατρικών Δεδομένων

Αναστάσιος Αψεμίδης

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα

Απρίλιος 2024

ACKNOWLEDGEMENTS

I would like to thank my supervisor Nikolaos Demiris for his guidance throughout my PhD.

I am grateful to Dimitrios Karlis, Athanasios Yannacopoulos, Michael Zazanis and Nikolaos Demiris for their teaching courses during the first PhD year.

I am also grateful to Ioannis D. Vrontos for our cooperation in the project "Modeling the Economic and Financial Impact of COVID-19" and to Alexandra Livada for her help in the Excel seminars of the project "Reinforcing Educational and Research Activity of the Department of Statistics". Lastly, I thank Stavros Vakeroudis and Nikolaos Demiris for our cooperation in the project of assignments evaluation of the "Master of Science in Statistics" program. The above mentioned projects were funded by the Hellenic Foundation for Research and Innovation. I also have to mention Ioannis Ntzoufras who was always very helpful and I am deeply thankful to.

Many thanks to the Hellenic Statistical Authority that let me be part of the team for the "European Big Data Hackathon 2023" in Brussels/Belgium.

Finally, I am thankful to my family for their support.

Abstract

Anastasios Apsemidis

Bayesian Evidence Synthesis for the Analysis of Biomedical Data

April, 2024

In the era of Statistics and Data Science, the analysis of biomedical data has received increasing attention by researchers and practitioners who seek to harness the plethora of information for decision making processes. The Bayesian methodology which has also gained great popularity in the last few decades, due to computational and statistical advances in Monte Carlo simulation methods, provides a coherent framework for synthesizing evidence from multiple sources. To this end, we aim at utilizing Bayesian models for estimating critical quantities in the fields of communicable and non-communicable disease analysis. Regarding the former, we are concerned with the Covid-19 pandemic and, specifically we build discrete-time stochastic compartmental models based on the latent level of both registered and unregistered cases to infer the reproduction number and the proportion of cases observed. Further, we face the problem through the lens of dynamical systems to gain insight, but also construct quantities suitable for decision support. In the non-communicable diseases context, we propose different extrapolation methods of the survival curve taking into account projections of mortality with the aim of estimating the life years gained when a treatment is selected in place of another. The methodology is demonstrated on three studies the medical community is concerned of, regarding breast cancer, advanced melanoma and cardiac arrhythmia.

Περίληψη

Αναστάσιος Αψεμίδης

Μπεϋζιανή Σύνθεση Πληροφορίας για την Ανάλυση Βιοϊατρικών Δεδομένων

Απρίλιος, 2024

Στην εποχή άνθησης της Στατιστικής και της Επιστήμης των Δεδομένων, η ανάλυση βιοϊατρικών δεδομένων κερδίζει συνεχώς την προσοχή ερευνητών και επαγγελματιών, οι οποίοι προσπαθούν να χρησιμοποιήσουν την πληθώρα πληροφορίας σε διαδικασίες λήψης αποφάσεων. Η Μπεϋζιανή μεθοδολογία, της οποίας η δημοτικότητα έχει επίσης αυξηθεί τις τελευταίες δεκαετίες λόγω της υπολογιστικής και στατιστικής προόδου σε μεθόδους προσομοίωσης Μόντε Κάρλο, παρέχει ένα συνεκτικό πλαίσιο σύνθεσης πληροφορίας από διαφορετικές πηγές. Έτσι, στοχεύουμε στη χρήση Μπεϋζιανών μοντέλων, για να εκτιμήσουμε σημαντικές ποσότητες στα πεδία τόσο των λοιμωδών όσο και των μη λοιμωδών ασθενειών. Όσον αφορά τις λοιμώδεις ασθένειες, ασχολούμαστε με την πανδημία Covid-19 και, συγκεκριμένα, κατασκευάζουμε στοχαστικά διαμερισματικά μοντέλα διακριτού χρόνου βασισμένα στο λανθάνον επίπεδο των καταγεγραμμένων και μη κρουσμάτων, ώστε να εκτιμήσουμε το ρυθμό αναπαραγωγής και το ποσοστό των παρατηρούμενων κρουσμάτων. Επίσης, αντιμετωπίζουμε το πρόβλημα υπό το πρίσμα των δυναμικών συστημάτων με στόχο την ανάπτυξη διορατικότητας, αλλά και την κατασκευή ποσοτήτων κατάλληλων για υποστήριξη λήψης αποφάσεων. Στο πλαίσιο των μη λοιμωδών ασθενειών, προτείνουμε μεθόδους παρεκβολής της καμπύλης επιβίωσης, λαμβάνοντας υπόψη προβολές της θνησιμότητας, με στόχο να εκτιμήσουμε τα χρόνια ζωής που κερδίζονται, όταν εφαρμόζεται μία θεραπεία αντί κάποιας άλλης. Η μεθοδολογία παρουσιάζεται μέσα από τρία παραδείγματα που απασχολούν την ιατρική κοινότητα και αφορούν τον καρκίνο του μαστού, το μεταστατικό μελάνωμα και την καρδιακή αρρυθμία.

So far as the theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality.

- Albert Einstein

Contents

List of Tables

List of Figures

Chapter 1

Introduction

To ask the right question is harder than to answer it.
- Georg Cantor

The analysis of biomedical data has been increasingly popular during the last 50 years, due to the development of Statistics, which provides tools for inference and prediction that can be used on important decisions on the effectiveness of therapies/drugs, planning strategies on pharmaceutical and non-pharmaceutical measures or even combining results of different studies.

The advantage of using statistical methods in all these scenarios, as opposed to deterministic mathematical models, is that every estimate is followed by a probabilistic measure of uncertainty. Thus, the plausibility of complex models can be quantified and simulations checking specific assumptions can be conducted.

In this Thesis, we are concerned with the analysis of *communicable* and *non-communicable* diseases. The former are diseases that are caused by bacteria and viruses and are spread from one individual to another via means like air droplets or body fluids. Examples of such diseases are the influenza and the Covid-19. On the contrary, Non-Communicable Diseases (NCD) are not caused by infection and are the leading cause of deaths worldwide (74% of all deaths[79]) according to the World Health Organization

(WHO). Examples of NCD are cancers, cardiovascular diseases and diabetes.

From a statistical point of view, the analysis framework is quite different between the two types of diseases. The infectious disease data are always incomplete in the sense that the exact time of infection and the period of infectiousness are not observed for every individual. Moreover, for large-scale outbreaks like Covid-19, we typically have only daily information for the total numbers of deaths or registered cases (not individual-based numbers). Finally, in infectious diseases, the individuals participate in a chain of infection, which leads to much correlation in the data. All in all, inference and predictions based on infectious disease data requires special treatment and can easily fool someone regarding their integrity (see Lazer et al., 2014[65] for the 2013 Google Flu Trends). These problems of outbreak data do not disappear in the Big Data era, where there is plenty of information to mine, as a recent article regarding Covid-19 vaccines underlines (Bradley et al., 2021[12]) referring to the big data paradox "the bigger the data, the surer we fool ourselves" (Meng, 2018[72]). On the other hand, with NCD we are interested in the survival of patients under different treatments and the major complication in the data is the presence of censoring (defined below), which is also the reason that simple logistic regression models cannot work in such situations and more elaborate methods are needed.

Throughout the Thesis we adopt the Bayesian school of thinking, which stems more naturally from our intuition and tend of interpretation. Bayesian models fitted using Markov Chain Monte Carlo (MCMC) simulation methods can be more complex than ones fitted by frequentist methods, like Maximum Likelihood, while they also provide the theoretical guarantee of reaching to the true parameters after infinite iterations (although in practice we use only a few thousand iterations and convergence diagnostics, which in reality are more like non-convergence diagnostics).

The present Chapter is organized as follows. In Section 1.1, we are concerned with NCD and therefore introduce the reader to common concepts in survival analysis and provide necessary definitions, terminology and bibliography. We discuss the current ways of fitting and extrapolating a model as well as some basic concepts of mortality analysis. In Section 1.2, we discuss the case of communicable diseases and provide background thereon. All the analyses in the Thesis are performed using R[85], while Bayesian models

are fitted using Hamiltonian Monte Carlo (HMC)[75, 9] via the probabilistic programming language Stan[95], which has the option of implementing a minimum-tuning algorithm for HMC called No U-Turn Sampler (NUTS)[49].

1.1 Non-Communicable diseases

Regarding NCD, we are mostly interested in modeling survival curves, summarizing the course of a typical individual suffering from the specific type of NCD and belonging to a specific demographic and treatment group. Thus, survival analysis is a useful tool in health industry, since it provides insight in the effectiveness of drugs and therapies. Specifically, the analysis of time-to-event data (which extends beyond the range of biomedical applications) serves as an inference means when there arise questions like "What is the probability of a patient suffering from breast cancer to live more than 3 years after diagnosis, when receiving a certain type of drug?". Furthermore, survival analysis is utilized in prediction purposes, when the question of interest is something like "How long do we expect the aforementioned patient to live on average?". This type of questions is not actually a typical prediction, but what is referred to as extrapolation, that is prediction out of the range of given data. Extrapolation is dangerous and, although statisticians advise against it, it has become a routine in health economic evaluations.

Health economics is a field where different drugs/therapies are compared in terms of their long-term survival, as obtained by an extrapolated model, in order to decide on the best. Except of their clinical effectiveness, another factor for the decision making process is their cost effectiveness, since there are limited national resources for funding. For instance, in United Kingdom (UK) the organization responsible for such decision making procedures is the National Institute for Health and Care Excellence (NICE).

Although extrapolation may be a difficult task, its performance can be improved by using information on the survival of an external population. By doing this, it is possible to make predictions of survival probabilities way beyond the range of the data, i.e. the range of the follow-up period, by anchoring the predictions on the data (or the fitted

model) of the external population, which is more general and includes the disease group under study. However, this general population is built using past information of survival probabilities and does not correspond to survival at the current year (more on this below).

1.1.1 Survival analysis

Survival analysis is a term used for the analysis of time-to-event data. A *time-to-event* data point is the recorded time passed until a subject has experienced an event. Applications of survival analysis are found in medicine, economics, industry and other areas (so time-to-event is also referred to as failure time), but the subject of the Thesis is the analysis of time-to-event data regarding survival of patients suffering from a disease. However, the methods discussed are applicable in any survival analysis, like the completely different problem of determining time until a mechanical part of a machine is broken (in a statistical process control setting), or determining time until a goal is achieved in a football match (in a sports analytics setting). For all the following, we assume that time is a continuous variable t (although similar theory can describe a discrete time variable), therefore $t \in \mathbb{R}_+$.

One characteristic that time-to-event data have is the presence of *censoring*, i.e. some of the observations are incomplete in the sense that we are aware of survival of the subject until a time point, when the subject stopped being observed, but the event did not occur. The period during which an individual's health is being monitored as part of a study is called *follow-up* period. There are different types of censoring, but we are concerned with what is called *right-censoring*, which means that some of the subjects (patients) either dropped out of the study without experiencing the event (death) and we call them *lost to follow-up*, or the study ended without them experiencing the event. Thus, we know that such a patient is alive until the last available time, but we do not know the exact time that they experienced the event. However, such partial information of knowing that censored individuals are alive until a time point can be exploited in a survival analysis. Figure 1.1 depicts schematically the situation that two out of five patients are right-censored. Specifically, Patients 1, 2 and 4 are known to have died at time 2, 1 and 2 respectively, while Patient 3 is lost to follow-up at time 3 (we do not

know that he died a little later) and Patient 5 has not died until the time the study has ended at $t^* = 4$. Patient 5 actually died at time $t = 5$, but this is not recorded in the study. Thus, Patients 3 and 5 are right-censored.

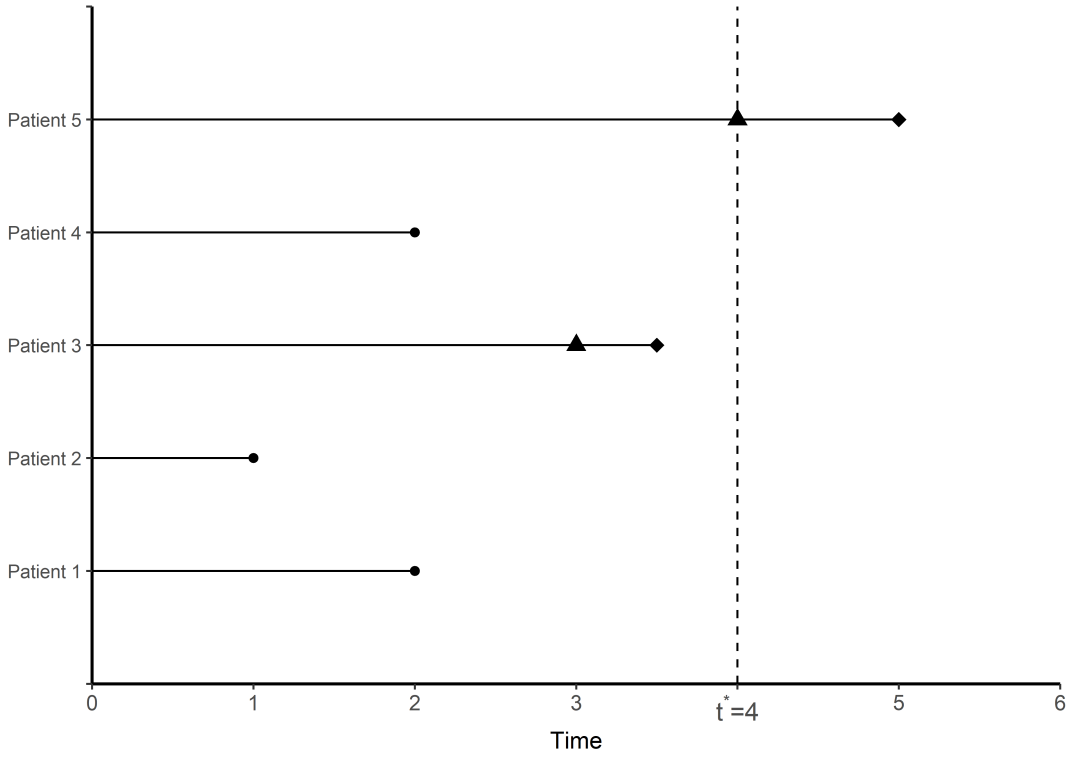

Figure 1.1: Illustration of typical survival data. In this case, the data are (2, 1, 3*, 2, 4*), where we note the censored observations with an asterisk. The hypothetical study ended at $t^* = 4$. Uncensored observations are depicted with circles, censored observations are depicted as triangles and the rhombuses are the actual event times of the censored observations (which are not known).

Let X_i for $i = 1, ..., n$ be n non-negative random variables measuring the time until an event with Cumulative Distribution Function (CDF) M and C_i for $i = 1, ..., n$ be n non-negative random variables of censoring times with CDF N. We assume that $\mathbf{X}$ and $\mathbf{C}$ are independent (a scenario called *random censoring*) and we consider the bivariate variables of the form $\mathbf{Y}_i = (T_i, \Delta_i)$, where $T_i := \min\{X_i, C_i\}$ and $\Delta_i := I(X_i \leq C_i)$. Thus, in the example given in Figure 1.1, the realized $\mathbf{X}$ is the vector of times $\mathbf{x} = (2, 1, 3.5, 2, 5)$ and the realized $\mathbf{C}$ is the vector of censoring times $\mathbf{c} = (2, 1, 3, 2, 4)$. The first component

of the sample we observe is the vector $\mathbf{t} = (2, 1, 3^*, 2, 4^*)$ and the second component is $\boldsymbol{\delta} = (1, 1, 0, 1, 0)$. We denote censored observations with an asterisk.

In an analysis of time-to-event data, we are interested in the so-called *survival function* or *reliability function* at a time point t defined as the probability that the failure will occur at a point greater than t, i.e.

$$S_T(t) := P(T \geq t) = 1 - P(T < t) = 1 - F_T(t) \tag{1.1}$$

Therefore, the survival function of T is just the complementary CDF of T. Another useful function in survival analysis is the *hazard rate* at time t defined as

$$\begin{aligned} h(t) &:= \lim_{h \to 0} \frac{P(t \leq T < t + h \,|\, T \geq t)}{h} = \lim_{h \to 0} \frac{P(t \leq T < t + h \,,\, T \geq t)}{h \cdot P(T \geq t)} \\ &= \lim_{h \to 0} \frac{P(t \leq T < t + h)}{h \cdot P(T \geq t)} = \frac{f(t)}{S(t)} \end{aligned} \tag{1.2}$$

where $f(t)$ is the Probability Density Function (PDF) of T. The last equality holds because of the definition

$$f(t) := \lim_{h \to 0} \frac{P(t \leq T < t + h)}{h}$$

Note that the hazard rate is not a PDF, but it describes the plausibility of failure at the next instant given survival up to this point.

From equation (1.1), we get that $f(t) = \dfrac{dF(t)}{dt} = \dfrac{d\big(1 - S(t)\big)}{dt} = -\dfrac{dS(t)}{dt}$ and, plugging this into (1.2) we have

$$\begin{aligned} h(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)/dt}{S(t)} &\Rightarrow h(t)dt = -dS(t)/S(t) \Rightarrow \int_0^u h(t)dt = \int_0^u -dS(t)/S(t) \\ &\Rightarrow \int_0^t h(u)du = -[\log S(u)]_0^t \\ &\Rightarrow \int_0^t h(u)du = -\log S(t) + \log S(0) \\ &\Rightarrow S(t) = exp\Big(- \int_0^t h(u)du \Big) \end{aligned}$$

because $S(0) = P(T \geq 0) = 1$. This last equation leads to the last special quantity used in survival analysis, which is the *cumulative hazard function* or *integrated hazard* and is defined by

$$H_T(t) := \int_0^t h(u)du = -\log S(t) \tag{1.3}$$

This means that the integrated hazard is just the log transform of the survival with a minus sign to ensure positivity, since the survival takes values inside $[0, 1]$. A large value of survival leads to a small value of the cumulative hazard and vice versa. Lastly, knowledge of one of $f(t)$, $h(t)$ and $S(t)$ leads to knowledge of the other three, using the equations $h(t) = f(t)/S(t)$, $f(t) = -dS(t)/dt$ and $S(t) = exp\Big(- \int_0^t h(u)du \Big)$ and can thus define the distribution of time until the event.

Let us assume that a dataset of $\mathbf{y}_i$ values are observed, where $\mathbf{y}_i = (t_i, \delta_i)$, t_i are the (independent) observed event times and δ_i indicates whether the event was actually an event or a censored time, i.e.

$$\delta_i = \begin{cases} 1 & \text{if } x_i \leq c_i \\ 0 & \text{if } x_i > c_i \end{cases}$$

Note that our data is just the $\mathbf{y}$ vector and we do not have $\mathbf{x}$ and $\mathbf{c}$. Good news is we can build the likelihood function only using $\mathbf{y}$ as follows.

We decide on a continuous distribution with domain the positive real numbers. The contribution of an observation which experiences the event, i.e. $(t_i, 1)$, is just its density $f(t_i)$ as it is in the usual settings. However, the contribution of partially observed observations, i.e. $(t_i, 0)$, is the survival $S(t_i)$. We can think that the probability of the sample to be observed is the product of the individual probabilities for every time point, but the probability of a censored time equals the probability of the event happening after that censored time. Of course the likelihood does not correspond to probabilities when using continuous variables, but this is just an intuition for the form of the likelihood, which reads

$$\begin{aligned} L(\mathbf{y} \,|\, \boldsymbol{\theta}) &= \prod_{i=1}^{n} f(t_i)_i^{\delta} S(t_i)^{1-\delta_i} && (1.4) \\ &= \prod_{i=1}^{n} h(t_i)_i^{\delta} S(t_i)_i^{\delta} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} S(t_i) h(t_i)_i^{\delta} && (1.5) \end{aligned}$$

where $\boldsymbol{\theta}$ is the vector of parameters of the family $f(\cdot)$. We can use either (1.4) or (1.5) for maximization or in a Bayesian context.

For visual inspection of the goodness of fit, we compare nonparametric estimations of the survival, hazard and cumulative hazard functions (which are treated as "data") with the fitted model. The most famous estimator of the survival function is the Kaplan-Meier defined as

$$\hat{S}(t) := \prod_{t=1}^{n}(1 - \frac{d_t}{n_t})$$

where d_t is the number of people experiencing the event at time t and n_t is the number of people at risk of experiencing the event at time t. For instance, if the event is death, then d_t is the number of people that died at time t and n_t is the number of people left in the study at time t. The quantity d_t/n_t is actually an estimate of the hazard rate $\hat{h}(t)$, but it is not used without smoothing (see for example Müller and Wang, 1994[74] and Gefeller and Dette, 1992[37]), because of its large variance. Finally, the cumulative hazard function is estimated using the Nelson-Aalen formula as

$$\hat{H}(t) = \sum_{i=1}^{t} \frac{d_i}{n_i}$$

Regarding the distributions typically used in survival analysis, we give in Table 1.1 the hazard and survival functions of some common distributions. Note that Exponential is the only distribution with constant hazard function. Although any distribution with support on $\mathbb{R}_+$ can be used, since the data regard time, the typical competitors are the Exponential, Weibull, Gamma, Log-Normal, Log-Logistic and Gompertz. Two more complex families also used are the Generalized Gamma (which can be reduced to the Exponential, Weibull and Gamma) and the Generalized F (which can be reduced to the Log-Logistic and Generalized Gamma). Other more flexible families exist in the literature which allow for hazards of various shapes (see for instance Shama et al., 2023[92] and references therein).

One quantity of interest when conducting a survival analysis is the mean survival time, which under the scope of health economic evaluations can be used to estimate the Life Years Gained (LYG) when switching from one group/drug/therapy to another. The mean survival time can be calculated as follows. Let T be the variable that counts time

until the event, with PDF $f(t)$. Then, the mean of T from time 0 until n is

$$\begin{aligned}\mathbb{E}[T] &= \int_0^n t f(t)dt = \int_0^n \int_0^t ds f(t)dt = \int_0^n \int_0^t f(t)dsdt \\ &= \int_0^n \int_s^n f(t)dtds = \int_0^n \big(F(n) - F(s)\big)ds\end{aligned}$$

Then, letting $n \to \infty$, in which case $F(n) \to 1$, we have

$$\mathbb{E}[T] = \int_0^\infty \big(1 - F(s)\big)ds = \int_0^\infty S(s)ds$$

which means that the mean survival equals the area under the survival curve. Then, the LYG can be calculated for two different groups by subtracting the two areas, i.e. finding the area between the two survival curves. When, the full survival curve is not known, we refer to the *restricted mean survival*, which is simply the area under the observed survival curve.

Distribution	Hazard	Survival
Exponential	λ	$exp\big[-\lambda \cdot t\big]$
Weibull	$(\alpha/\sigma) \cdot (t/\sigma)^{\alpha-1}$	$exp\big[-(t/\sigma)^{\alpha}\big]$
Log-Normal	$\dfrac{\phi\Big(\dfrac{\log(t)-\mu}{\sigma}\Big)}{t \cdot \sigma\Phi\Big(\dfrac{\mu-\log(t)}{\sigma}\Big)}$	$\Phi\Big(\dfrac{\mu-\log(t)}{\sigma}\Big)$
Log-Logistic	$\dfrac{(\alpha/\sigma)(t/\sigma)^{\alpha-1}}{1+(t/\sigma)^{\alpha}}$	$\dfrac{1}{1+(t/\sigma)^{\alpha}}$
Gompertz	$\lambda \cdot exp\big[\alpha \cdot t\big]$	$exp\big[-(\lambda/\alpha)e^{\alpha \cdot t-1}\big]$

Table 1.1: The hazard and survival functions of the most common distributions. Multiplying the two functions, produces the corresponding PDF. $\sigma \in \mathbb{R}_+$ indicates scale parameters, $\lambda \in \mathbb{R}_+$ indicates rate parameters, $\alpha \in \mathbb{R}_+$ indicates shape parameters, $\mu \in \mathbb{R}$ indicates location parameters, $\phi(\cdot)$ and $\Phi(\cdot)$ indicate the PDF and CDF of the standard normal distribution respectively. In the case of Gompertz, the shape can take on negative values.

1.1.2 Extrapolation

Extrapolation is referred to prediction outside the range of the data and, particularly in our case, way beyond the range of the data. Specifically, we want to fit a model on the follow-up data and use it to predict long-term survival, which can be many years after the end of the follow-up period. Thus, comparisons can be made among survival curves of different groups of people (e.g. age, sex, or treatment groups) and use these estimates in decision making processes. This task is encountered by incorporating information from a general population, which acts as anchor for the extrapolation to be more robust, since the long-term data of the general population have a "complete" survival curve. The general survival curve is defined over a lifetime horizon, so it eventually reaches zero, in contrast with typical follow-up periods of 1-5 years, which may lead to a survival of only 50%.

There exist many articles that try to tackle the survival extrapolation problem. Demiris and Sharples (2006)[26] fit a joint model on the disease and general population data and extrapolate according to the fitted model, while Benaglia et al. (2015)[8] extrapolate in the presence of cause-specific hazards. Che et al. (2023)[21] use a method that gradually blends the predicted survival of the disease population with the survival of the general population. For a review on this subject see Jackson et al. (2017)[52] and Bullement et al. (2023)[15]. Gallacher et al. (2021)[36] discuss problems regarding the goodness-of-fit statistics to select a model for extrapolation. The general idea can also be pictured in Latimer and Adler (2022)[63]. Guyot et al. (2017)[43] use splines to extrapolate survival of cancer patients. Royston and Parmar (2002)[89] do not extrapolate, but also use splines to model the hazard or odds function demonstrating on cancer datasets. The odds function is defined as the ratio of the CDF and the survival, i.e. $O(t) = [1 - S(t)]/S(t)$. Sweeting et al. (2023) [97] provide a tutorial on extrapolation using excess hazard and cure models, while Jackson (2023)[53] provides an R package for flexible extrapolation using M splines.

1.1.3 Mortality analysis

The analysis of mortality data is crucial for demographic agencies as well as insurance companies and aims at describing the rate of death of a population. The two most well known models are the Heligman-Pollard model (see Heligman and Pollard, 1980[47]) and the Lee-Carter model(see Lee and Carter, 1992[66]).

The Heligman-Pollard models the odds of death and assumes that the underlying function is composed of three additive terms: a decreasing function capturing the mortality rate of newborns, a concave "hump" function capturing the mortality due to accidents and an increasing function capturing the mortality rate of adulthood. Specifically, if we define the mortality rate as $m_x = d_x/e_x$, where d_x is the number of deaths at age x and e_x is the number of people exposed to death at age x, then the probability of dying within one year is $q_x = 1 - exp(-m_x)$ for a person aged x. The Heligman-Pollard model reads

$$\frac{q_x}{1-q_x} = A^{(x+B)^C} + Dexp\big[-E\big(\log\frac{x}{F}\big)^2\big] + GH^x$$

where A, B, C, D, E, F, G and H are parameters to be estimated.

The Lee-Carter model on the other hand, aims at describing the dynamic nature of mortality rates through time and it is expressed as

$$y_{x,t} = \alpha_x + \beta_x\kappa_t + \epsilon_{x,t}$$

where $y_{x,t} = \log(d_{x,t}/e_{x,t})$ and α_x, β_x and κ_t are parameters to be estimated, while $\epsilon_{x,t}$ is an error term. The advantage of this model is that we can use it to predict log-mortality for each age x at time t and this is a crucial step of the proposed methodology discussed in Chapter 2.

1.1.4 Data extraction from a function graph

In many cases in the literature the data used in an analysis are not available, for instance due to limited licence reasons. However, publishing the 2-dimensional graph of a function utilizing these data can be very informative for others to extract. This mining procedure

is very useful in survival analysis, since it allows for scientists to figure the data out using a published Kaplan-Meier curve.

The idea behind digitization of an image is very simple, since at its core lies only an affine transformation of the coordinate system used by the Graphical User Interface (GUI), which in our case is the default R (or RStudio) environment. We begin by inserting the image with the published Kaplan-Meier into R. This image is plotted on the R canvas, which represents the $\mathbb{R}^2$ space. We want to click on the image to select K points that help us determine the unknown underlying function and let R return their coordinates. However, these coordinates will be based on the Cartesian system that R uses and, not on the system that is shown in the image. Thus, we need to define the system shown in the image and transform every selected point based on that. We use the notation $(x, y)^R$ and $(x, y)^U$ for a point (x, y) expressed in R's and user's language (i.e. image's language) respectively.

The transformation we need is an affine transformation, since we only need to stretch the plane and also move the origin, since $(0, 0)^R$ will not be at the same as $(0, 0)^U$. Thus, we search for the coefficients of the function $y = \alpha + \beta x$, where x is a selected point expressed in R's language (obtained by clicking on the plot) and y is the same point expressed in user's language, i.e. expressed in the system we see in the image. Since, the coordinate system used in a plot is not always orthonormal, due to different units in the x- and y-axis, we need to define two different transformations (one for each coordinate) and then just use these transformations to set the coordinates of every point.

We describe the procedure for the x-coordinate and the y-coordinate is obtained similarly. We want to specify the underlying transformation $y = \alpha + \beta x$ that translates the x-coordinate of a point in R's language into the x-coordinate in user's language. Since we search for a straight line, we only need two different points that we know their translation, i.e. we know both x and y of the transformation. The x-axis shown in the image provides plenty of such points, since we can click on a point and obtain the R's expression and also read the image to get the user's expression. Thus, we select two points with named ticks from the x-axis of the image and obtain $(x_1, y_1)_R$ and $(x_2, y_2)_R$, but also their translation $(x_1, y_1)_U$ and $(x_2, y_2)_U$. Their y-coordinates are not needed, so

we proceed with just x_1^R and x_2^R and their corresponding x_1^U and x_2^U. It must hold that $x_1^U = \alpha + \beta x_1^R$ and $x_2^U = \alpha + \beta x_2^R$, so the slope β is given by the ratio of the differences $y_2^R - y_1^R$ and $x_2^R - x_1^R$, i.e.

$$\beta = \frac{y_2^R - y_1^R}{x_2^R - x_1^R}$$

For α we have to solve either of the two equations with respect to α so, using the first one and the slope we just calculated, we have

$$\alpha = x_1^U - \beta x_1^R = x_1^U - \frac{y_2^R - y_1^R}{x_2^R - x_1^R} x_1^R$$

Now, if we use the obtained transformation on the x-coordinates of all the K collected points x_k^R, $k = 1, ..., K$, we have their corresponding x_k^U values used in the image. We follow the same procedure for the y-coordinates of two points $(z_1, w_1)_R$ and $(z_2, w_2)_R$ (this time clicked on the y-axis of the image), i.e. w_1^R and w_2^R to obtain the coefficients of the y-axis transformation and then we translate the y-values of the K collected points.

The data extracted by this procedure may be treated as fixed, but in reality they are approximations of the real data and, hence there is inherent variance. If one wishes to eliminate some portion of this variance, they need to repeat the procedure N times, collecting $K_1, ..., K_N$ datasets and averaging the final estimates of the analysis conducted. So the final estimate of a variable θ would be

$$\hat{\theta} = \frac{\sum_{i=1}^{N} K_i \theta_i}{\sum_{i=1}^{N} K_i}$$

This is more useful when the function needs many points to be determined (actually estimated). A Kaplan-Meier plot with a few points is relatively easy to be digitized, because the times of steps of the function are visible. On the other hand, when the Kaplan-Meier plot visualizes survival of hundreds of patients, the steps are not easily distinguishable.

Finally, we let us use the procedure on the bladder dataset of the R package *survival*[101]. The data are times until the first recurrence of 85 patients with bladder cancer. The dataset is much more detailed, but we keep it simple for illustration purposes. We extract 21 event times and their corresponding survival probability from the plot (the x- and y-coordinates) and calculate the restricted mean survival time. The restricted mean

survival using the actual Kaplan-Meier curve is 28.617, while averaging the estimates by the digitized data we get 28.708 (see Figure 1.2).

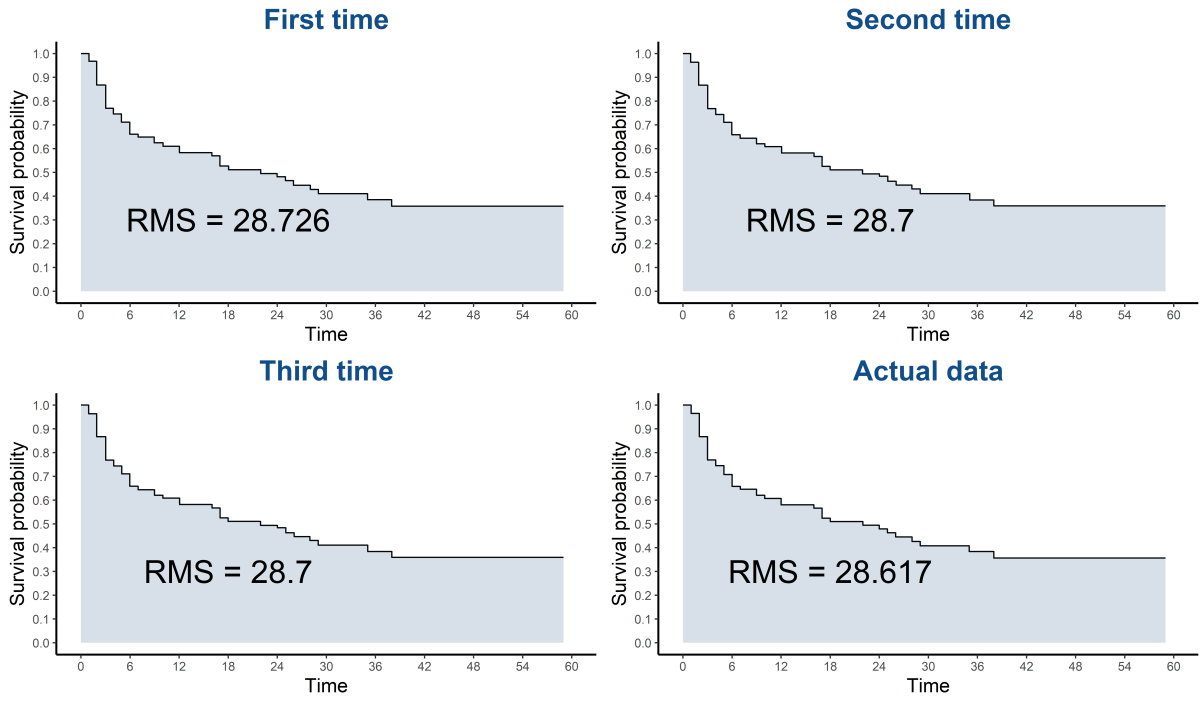

Figure 1.2: Restricted mean survival time using digitized data (upper left, upper right and lower left) and using the actual data (lower right) from the bladder dataset.

1.2 Communicable diseases

The word "epidemic" comes from ancient Greek, where "epi-" is a preposition for "upon" and "-demic" stems from the word δῆμος (/'ði.mos/), which means "people". So epidemic is something that happens on people. In the modern sense, by the term *epidemic* we mean a disease outbreak that affects a population in a short period of time, while if the epidemic is spread in the entire world we call it *pandemic*, again from ancient Greek where πᾶν (/pan/) means "whole", "everyone". John Snow is considered the father of epidemiology thanks to his studies on cholera outbreaks in London in the mid 1800. After him, epidemiology became a popular tool in dealing with a wide range of outbreaks

and an important part of public health systems in decision making. Thus, when a new epidemic occurs, researchers resort to mathematical models (see for instance Bjørnstad, 2022[11]) for the estimation of the epidemic potential and the corresponding mitigation and control measures since infectious disease outbreaks can have devastating consequences on both the society and the economy.

Pathogens co-exist with humans in this world and as long as we share this planet, epidemics will continue to occur. Some examples of epidemics that have hit the world in the past are the Measles (which exists since ancient times), Ebola (which was first identified in the 1970's) and HIV (which was first identified in the 1980's). Some recent ones are the foot-and-mouth disease in the UK (see Ferguson et al., 2001[34] and Keeling et al., 2001[56]) and the SARS (see Lipsitch et al., 2003[67] and Riley et al., 2003[88]), but the one we are concerned with in this Thesis is the latest and perhaps most well known Covid-19 pandemic, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).

Although there have been another three major coronavirus-related outbreaks, namely the SARS in 2003, the MERS in 2012 and the MERS in 2015 (see Kwok et al., 2019[62] for a review), the statistical community showed great interest in estimating quantities related to the novel virus and its spread and epidemic modeling became very popular, because the world suffered from its consequences for three years (see the WHO dashboard[107]). The virus, which first appeared in Wuhan, China in December 2019, causes many pneumonia-related symptoms and complications (chronic or not) and even death. The characteristic that made Covid-19 even more dangerous was the large proportion of asymptomatic cases that made it easier for the disease to spread (see for instance Oran and Topol, 2021[78]). It is transmitted through direct contact and droplets (see for instance Rahman et al., 2020[86]). The first case is when a person touches a contaminated object and then their mouth, eyes or nose, while the latter is when droplets after a cough or sneeze of a patient enter the mouth, eyes or nose of another person. Covid-19 is even more dangerous for certain groups of people, such as the elderly, those with cardiovascular and kidney diseases and diabetes. The World Health Organization declared Covid-19 a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 and a pandemic on

11 March 2020. Governments throughout the world either recommended or enforced protective measures against the new virus, which included lock-downs, social distancing and Personal Protective Equipment (PPE) such as wearing masks. Covid-19 costed many human lives (either directly by leading to death, or indirectly if the infected person belonged to a high-risk group), had strong social and psychological impact on people and created problems regarding the economy (see for example Dong et al., 2020[30], Kaye et al., 2021[55], Bashir et al., 2020[7]). The literature on Covid-19 includes thousands of articles that discuss different aspects of the pandemic, like modeling the transmission dynamics, assessing Non-Pharmaceutical Interventions (NPIs) effectiveness, evaluating the socio-economic impact, developing drugs and more, under the scope of Statistics, Medicine, Computer Science, Finance, or Sociology and utilizing tools such as continuous time processes, state-space models, time series, hierarchical models, compartmental models, neural networks, support vector machines, aggregated trees and more.

1.2.1 Epidemiological terminology

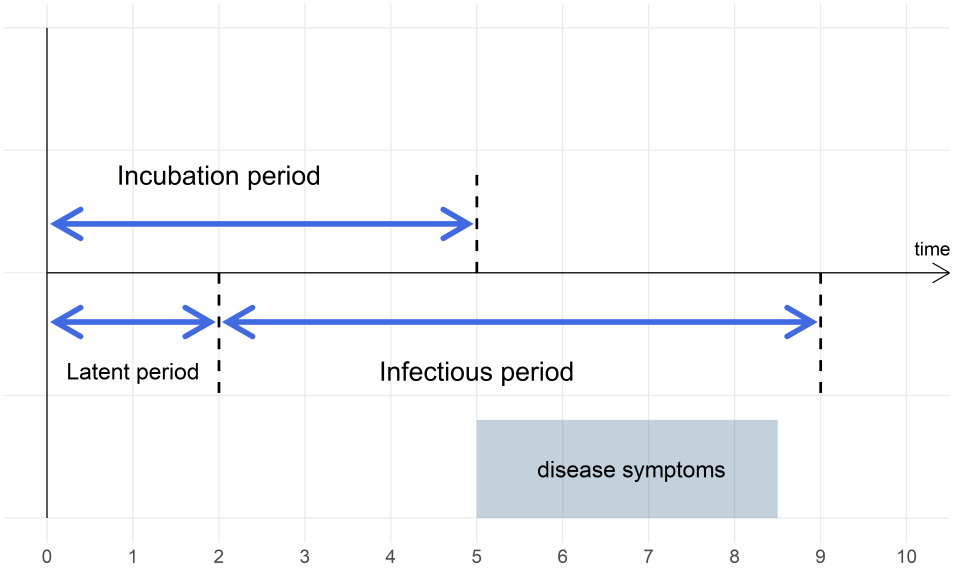

Figure 1.3: The difference among the latent, incubation and infectious periods in a hypothetical scenario. The horizontal axis measures time from infection in days. The patient experiences the disease for 3.5 days, but remains infectious for 0.5 days more.

Throughout Chapters 3 and 4 we use terms originated in Infectious Disease Epidemiology, which we shall discuss next. An *agent* (a pathogen or an infectious microorganism), which in our case is the Coronavirus, causes a disease to a host, which in our case are humans. When a human (from now we will use the word individual) carries the disease, but does not transmit it to others yet, then they are called *infected* or *exposed*. The period during which an individual is exposed is called *exposed period* or *latent period*. After that period, if it exists, the individual becomes *infectious*, meaning that they are infected and also transmit the disease. The set of all infectious individuals is called the *active set*. An individual with the potential to be infected in the future is called *susceptible*, meaning that they have not been infected yet but they can be in the future, or they have been infected, recovered and now they can be infected again in the future. There also exists a term for the time from infection until symptom onset and it is called *incubation period*. In Figure 1.3 we depict schematically a hypothetical scenario with the latent, incubation and infectious period having values 2, 5 and 7 days respectively. Finally, the *serial interval* is the time interval from infection of an individual until they transmit it to someone else.

1.2.2 The reproduction number

One most important quantity in epidemiology is the *basic reproduction number* R_0, which takes place in every decision making process about an epidemic. This number determines whether an epidemic has the potential to become large, or it is destined to faint quickly. Specifically, according to the threshold limit theorem, only if $R_0 > 1$ will the epidemic become a major outbreak assuming a large population (see Bailey, 1953[3], Whittle, 1995[106] and Ball, 1983[4]).

Let us assume that a population is *closed* (its size does not change), *homogeneous* (every individual is equally prone to the disease) and *homogeneously mixing* (every individual can get in contact with everyone else). If the infectious individuals contact on average k susceptible individuals and each contact has a probability p of transmitting the disease, then the *infection rate* is defined as $\lambda = k \cdot p$, or in informal mathematical description $\lambda = \#\, contacts \cdot P(transmission \,|\, contact)$. Note that an infectious individual

infects others with rate λ, so he infects each specific individual at rate λ/N, where N is the size of the population. The NPIs that aim at deteriorating the spread of the epidemic are such that reduce the two product terms. In the Covid-19 case for example, the first term can be reduced by a lock-down, while the second can be reduced by wearing face masks. The incubation period for the early variants of Covid-19 was approximately 5 days, but it could be as long as 14 days. Thus, when someone was suspected to had been infected, they had to stay at a 14-day quarantine, so that they have time to see if they are actually infected (of course they could just be asymptomatic).

The basic reproduction number can be written as $R_0 = \lambda \cdot \tau = k \cdot p \cdot \tau$, where τ is the mean *infectious period* (the duration of the infectiousness). For instance, suppose that someone infects 2 people per 5 days (so $\lambda = 2/5$) and he stays infectious for $\tau = 10$ days at total. Then, there will be $R_0 = 4$ people infected by the initial infected person. Sometimes, we refer to the reciprocal of the infectious period, which is the *recovery rate* γ, i.e. $\gamma = 1/\tau$. One could think of why this corresponds to the rate of recovery as follows: since individuals infect others for τ days, each day they are $1/\tau \cdot 100\%$ closer to recovery.

According to the aforementioned assumptions, the interpretation of R_0 is the following: in a completely susceptible population, each infectious individual transmits the disease to R_0 other individuals (on average) or, in other words R_0 is the expected number of secondary infections in a completely susceptible population. Therefore, it is clear that if each infectious individual infects more than one individuals, the epidemic will grow ($R_0 > 1$ case). At the beginning of Covid-19, its R_0 was estimated to be approximately 3.8 (95% CrI (2.4, 5.6)) (Flaxman et al., 2020[35]), but this value could vary depending on the country under study (see for instance Sanche et al., 2020[90]). In contrast with R_0 which refers to expected cases in a fully susceptible population, R_e represents the expected number of secondary cases in a partially susceptible to the disease population. Thus, $R_e(t)$ is the *effective reproduction number* at each time t, while R_0 refers to the beginning of the epidemic. In Chapter 3, where we are concerned with the effective reproduction number of Covid-19, we simply refer to it as reproduction number R_t (with time t as a subscript since there we deal with discrete-time models).

1.2.3 Compartmental models

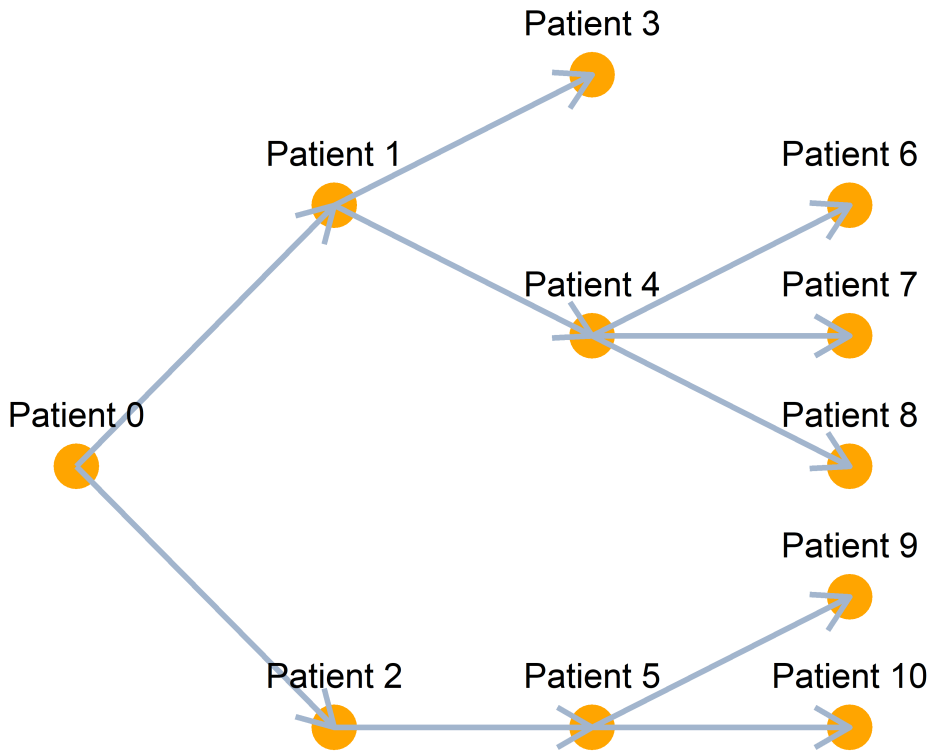

Figure 1.4: A graph that describes the chain of transmission for a case when $R_0 = 2$. Patient 0 (leftmost orange dot) transmits the disease to 2 other people, who each infect others. Each patient infects $R_0 = 2$ other individuals on average.

Scientists have developed many ways that an epidemic can be explained using mathematical and statistical models and at the moment there exist a wide spectrum of different methods one can use for a specific disease according to its type and characteristics. The most common way of thinking in epidemic modeling is the compartmental type. *Compartmental models* assume that each individual belongs to one out of many distinct states (also called compartments) that characterize their situation relatively to the epidemic. The whole population is part of a state-changing progress, which defines the epidemic and is named by the states used (see for example Iboi et al., 2020[51], who develop a compartmental model for Covid-19 in the most populous country in Africa, Nigeria, and assess spread mitigating scenarios via simulations).

For instance, if we assume that individuals are either susceptible to the disease, infectious due to the disease or immune for any reason and thus removed from the previous disease-related states, then we can define an S-state, an I-state and an R-state that stand for the labels Susceptible, Infectious and Removed that each individual can have. The R-state usually represents recovery or death. The former corresponds to a removal if we assume that individuals who get infected once cannot be re-infected. Because of these three states used, the model is called SIR. If we are interested in a disease that does not leave the host immune after infection, we can make use of an SIS (Susceptible-Infectious-Susceptible) model. Or, if the host experiences an initial period of non-infectiousness while infected and then becomes infectious, we can use a SEIR model (Susceptible-Exposed-Infectious-Removed), where the Exposed state means that, after getting exposed to the disease, individuals go through a period during which they do not transmit it (this is the latent period we have mentioned previously). Finally, note that we refer to all these models using epidemic terminology, but they can easily be applied to any scenario of transmission. For instance, the spread of some news or rumour can be modelled by an SI model, because an individual is initially not aware of the news (so susceptible) and then they become "infectious" in the sense that they have heard it and also can spread it to others. Or, we can imagine an economic crisis that is being spread throughout nations (so each individual is now a whole country) that pass the crisis to others (they "infect" them) and after a while they recover; thus we are describing an SIR situation. Liu and Cao (2022)[68] develop a Susceptible - Undocumented infected - Documented infected - Recovered model to capture the characteristics of the Covid-19 epidemic, breaking the Infected state at two parts, while Giordano et al. (2020)[39] use eight states, namely Susceptible, Infected, Diagnosed, Ailing, Recognized, Threatened, Healed and Extinct. Khan and Atangana (2022)[58] investigate the case of the Omicron variant of the Coronavirus (initially observed in South Africa in 2021) by using the states: Susceptible, Exposed, Asymptomatic infected, Symptomatic infected, Infected with the Omicron variant and Recovered.

There exist many options for epidemic modeling like the stochastic SIR-type Chain-Binomial, the Reed-Frost and the Greenwood models. For spatial epidemics, see the

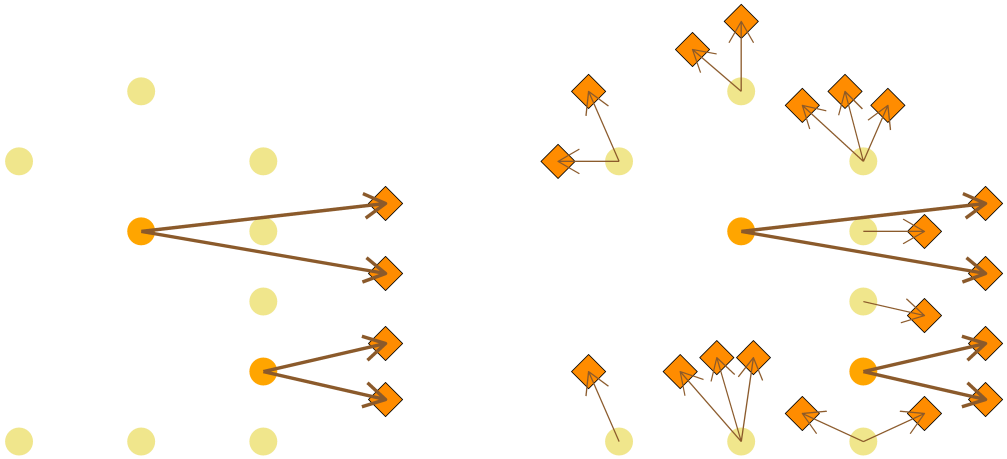

Figure 1.5: Scenario where $R_0 = 2$ with 10 initial infections (indicated as dots). Left: Only 2 patients are observed, so we expect 4 new cases (indicated as rhombuses). Right: In reality, 19 new cases are generated.

work of Mollison (1977)[73]. Another modeling procedure is through graphs (see Figure 1.4), where Patient 0 initiates the spread and transmits it to the next generation of patients, who in turn transmit it to the next generation and so on. Apart from modeling purposes, Figure 1.4 can be used to visualize an epidemic and how quickly it can be spread depending on R_0, since every dot leads to R_0 other dots on average. Now suppose that we are in a situation where $R_0 = 2$ and the epidemic starts with 10 individuals, but only 2 are observed, because the rest are asymptomatic. Then, we expect about 4 new cases for the next generation, while in reality we will have approximately 20 (see Figure 1.5). Observing only 20% of the total cases is not an extreme scenario if we think of the Covid-19 case when it first appeared and the Figure demonstrates that the total (rather than the observed) cases drive the epidemic and, so it is vital that we take into account the unregistered cases.

Basic ODE models

Describing an epidemic mathematically requires that we deal with initial value problems of Ordinary Differential Equations (ODE) using numerical integration. We do not focus on the details of these procedures as they are out of the scope of the Thesis, but

the method used is the Livermore Solver for Ordinary Differential Equations (LSODA) implemented in the R package `deSolve`[93]. Here we describe the simple SIR and the SIR with demography models, since the former sets the basis for our (more complex) stochastic models, while the latter adds a feature that completely changes the behaviour of the epidemic. This different behaviour is also accounted in the proposed models discussed in Chapter 3. This introduction helps the reader develop intuition, but more dynamics-related details can be found in Chapter 4.

The most basic yet useful epidemic model is the SIR. Its deterministic nature is described by the following system of ODE

$$\begin{aligned} \frac{dS}{dt} &= -\lambda SI/N \\ \frac{dI}{dt} &= \lambda SI/N - I/\tau \\ \frac{dR}{dt} &= I/\tau \end{aligned} \tag{1.6}$$

where $S(t)$ (abbreviated as S), $I(t)$ (abbreviated as I) and $R(t)$ (abbreviated as R) are the number of susceptible, infectious and removed individuals respectively, λ is the infection rate, τ is the infectious period and N is the population size. Although our interest lies in stochastic models, the ODE description is more useful in providing the intuition of its parameters. Moreover, the above formulation is equivalent to the stochastic case when $N \to \infty$ (more details in Andersson and Britton, 2012[1]). Note that $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$, i.e. $\frac{d}{dt}\big(S(t) + I(t) + R(t)\big) = 0 \Leftrightarrow \frac{d}{dt}N = 0 \Leftrightarrow N = constant$. Therefore, we assume that the total population size N does not change in time. The reproduction number is the one described above as $R_0 = \lambda\tau$.

Interpreting the model, we can say that the number of individuals who leave the S-state (i.e. those that have just been infected) should depend on both how many susceptible and how many infected there are at a given time, so the change in susceptible individuals should be proportional to the product of the susceptible and infected individuals at the same time (here we can use the term "infected" and "infectious" interchangeably since there is no exposed period) yielding $\frac{dS}{dt} \propto -SI$. We put a minus sign because the individuals *leave* the S-state. This happens at rate λ/N, where the denominator N assures that the infection rate does not depend on the population size. This happens

because $\lambda = k \cdot p$, where the number of contacts k is bounded by the population size N. So, having p fixed, a larger population has the potential of a larger λ. Thus, the first equation regarding the change in the susceptible individuals becomes $\frac{dS}{dt} = -\lambda SI/N$.

Regarding the second equation, it says that the change of $I(t)$ is due to the quantity $\lambda SI/N$ which was subtracted from the S-state and is now added to the I-state. Furthermore, the number of infectious individuals that escape the I-state at each instant is proportional to how many of them exist in this state with proportionality constant the recovery rate γ $(= 1/\tau)$, thus yielding the $-I/\tau$ term (again the minus sign indicates reduction). Finally, the individuals that leave the I-state at each instant (which are I/τ) enter the R-state, so the last equation is a valid update for the number of removed individuals.

In Chapter 4 we analyse the dynamics of the SIR model omitting the denominator N for λ, but keeping the variables as proportion of the population rather than absolute numbers, thus inherently dividing by N and having the same results, only expressed as percentages (so $S(t) + I(t) + R(t) = 1$). Using this idea, in Figure 1.6 we depict the simulated time series of $S(t)$, $I(t)$ and $R(t)$ for $\lambda = 0.5$ and $\tau = 6$ (so $R_0 = 3$, something like the Coronavirus case). We can see that $S(t)$ is a decreasing function, since susceptible individuals can only be reduced according to the first equation of the ODE system, $R(t)$ is an increasing function, since removed individuals have a positive derivative, but $I(t)$ has a global maximum. Setting the second equation equal to zero, we have $\lambda SI = I/\tau \Leftrightarrow S = 1/(\lambda\tau)$, i.e. the maximum corresponds to the time when $S(t) = 1/R_0$, or when the effective reproduction number at time t, $R_e(t)$, becomes equal to 1, because then it will be $R_e(t) = R_0 \cdot S(t) = R_0 \cdot 1/R_0 = 1$. Remember that $R_e(t)$ refers to the susceptible population at time t, so it scales R_0 with the proportion of individuals that are susceptible, that is $S(t)/N$, or simply $S(t)$ when we use the $N = 1$ formulation like we do here. Finally, note that the reproduction number does not tend to zero for the epidemic to end. Its usefulness and interpretation lie only to the comparison of it with the threshold 1. These results can be seen in Figure 1.6. When the infected individuals reach 0, we can see that approximately 94% of the whole population has been infected and that 6% has escaped infection.

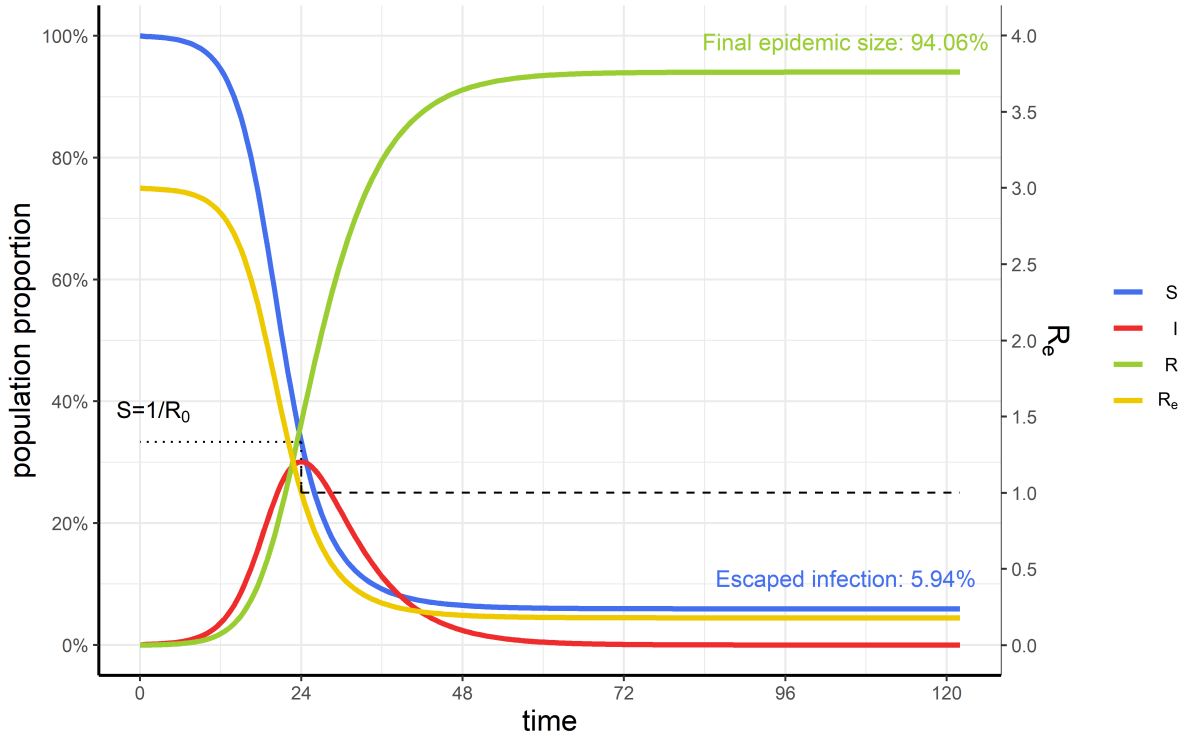

Figure 1.6: The series of $S(t)$, $I(t)$, $R(t)$ and effective reproduction number $R_e(t)$. The right axis is referred to the R_e scale. When $R_e = 1$, the I-state reaches its maximum (indicated by dashed lines), while at the same time the S-state takes the value of $1/R_0$ (indicated by dotted lines). We integrated the system with initial values $S = 0.999$, $I = 0.001$ and $R = 0$ using time steps of size 0.5 for as many steps as needed for $I(t) < 10^{-6}$ (the criterion was satisfied at $t = 122$, i.e. after 243 steps).

The second model we examine is the SIR with demography, which changes the behaviour of the epidemic by allowing it to become *endemic*, meaning that there remains a constant infectious portion of individuals that does not cease to exist as with the SIR case. The system of differential equations of the SIR with demography model is the following:

$$\begin{aligned} \frac{dS}{dt} &= -\lambda SI/N + A(N - S) \\ \frac{dI}{dt} &= \lambda SI/N - I/\tau - AI \\ \frac{dR}{dt} &= I/\tau - AR \end{aligned} \tag{1.7}$$

where A is the birth rate, which we assume equals the death rate. The idea is that people enter the susceptible state at rate AN due to births and exit the susceptible state at rate $\lambda SI/N$ due to infections and also at rate AS due to deaths (not related to the disease). Also, people enter the infectious state at rate $\lambda SI/N$ due to infections and exit that state at rate I/τ due to recovery or death because of the infection and also at rate AI due to deaths not related with the disease. Then, the removed individuals are increased with rate I/τ (due to recoveries or deaths from the previous state) and also decreased with rate AR due to deaths in the R-state not related with the disease. Note that the population size remains fixed since the added births are AN and the added deaths are $AS + AI + AR = A(S + I + R) = AN$. Lastly, the reproduction number is now given by $R_0 = \dfrac{\lambda}{\gamma + A}$, instead of $\dfrac{\lambda}{\gamma}$ as in the simple SIR, since individuals "recover" with the added rate of physical deaths.

We conduct the same simulation as we did with the simple SIR, in order to demonstrate the behaviour of the epidemic when demography is present in the model. We choose as $\lambda = 0.56$, $\tau = 1/6$ and $A = 1/50$ in order to have the same $R_0 = 3$, but now the image we get in Figure 1.7 is very different. We can see that a proportion of approximately 7% infected individuals remains in the population as endemic infections. The reader can refer to Van den Driessche and Watmough (2002)[103] for a discussion on equilibria for compartmental models.

Before closing the introduction to epidemic models, we give the SIR formulation expressed as a stochastic (rather than deterministic) process. Let $S(t)$ and $I(t)$ be the number of susceptible and infectious individuals respectively at time t. The epidemic begins at $S(0) = N$ (so the initial population is N) and $I(0) = m$, when each of the m initially infectious individuals make contact with susceptible individuals as "arrivals" in a homogeneous Poisson process with intensity λ/N. Then, these susceptible individuals become themselves infectious with independent and identically distributed (iid) infectious periods. After that period the individuals become removed. This is the description of the stochastic SIR model. Assuming further that the infectious periods are exponentially distributed with intensity $1/\tau$, then we are led to a bivariate Markov process, where the transition between state (s, i) and state $(s-1, i+1)$ of the process $(S, I) = \{(S(t), I(t)) :$

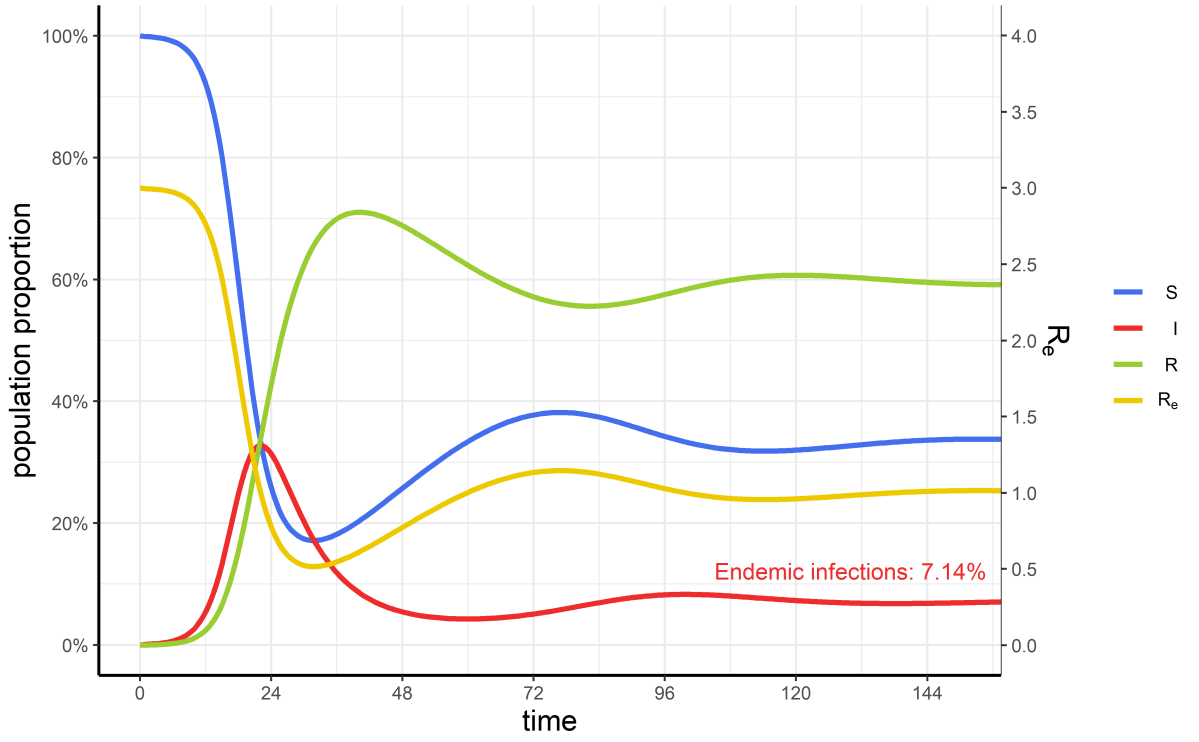

Figure 1.7: The series of $S(t)$, $I(t)$, $R(t)$ and effective reproduction number $R_e(t)$ under the SIR with demography model. The right axis is referred to the R_e scale. We integrated the system with initial values $S = 0.999$, $I = 0.001$ and $R = 0$ using time steps of size 0.5 for as many steps as needed for $\sqrt{S(t)^2 + I(t)^2} < 10^{-6}$ (the criterion was satisfied at $t = 410$, i.e. after 819 steps). The x-axis is cut to $t = 150$.

$t \in [0, \infty)\}$ happens at rate $\lambda si/N$, while the transition from (s, i) to $(s, i-1)$ happens at rate i/τ. This model is referred to as the General Stochastic Epidemic (GSE) model (see Kermack and McKendrick, 1927[57], Bartlett, 1949[6] and Britton, 2010[13]).

Assuming further that lifetime is exponentially distributed with intensity A, the Markov chain transitions inside the homogeneous and homogeneously mixing population are given in Table 1.2. Since the population size is fixed and the individuals do not move directly into the R-state, there is no need in describing transitions to the last state.

The GSE model has set the basis for many types of models that change its assumptions (see for instance Reinert, 1995[87]). A more general framework than the GSE is

From	To	At rate
(s, i)	$(s+1, i)$	AN
(s, i)	$(s-1, i)$	As
(s, i)	$(s-1, i+1)$	$\lambda si/N$
(s, i)	$(s, i-1)$	$(\gamma + A)i$

Table 1.2: Markov chain transitions between states S and I for the SIR with demography model.

the two-level mixing model (see Ball et al., 1997[5]) where we assume both a global and a local mechanism of epidemic spread. These models are more natural when someone wishes to accommodate a different rate of infection in groups like households. See Britton and Becker (2000)[14] for estimation of the critical vaccination coverage in populations consisted of households.

Chapter 2

Stable Survival Extrapolation via Transfer Learning

If I have been able to see further, it was only because I stood on the shoulders of giants.
- Isaac Newton

2.1 Introduction

Scientists of medical research are increasingly interested in survival extrapolation due to the role of the estimated Life Years Gained (LYG) when switching from one treatment to another. This estimate is used in health economics evaluations where only the best treatment/drug is to be endorsed given a finite budget. Without extrapolating the survival curve, one is limited in calculating the Restricted Mean Survival (RMS) time, which can be very misleading in the long run. Thus, although statisticians generally advise against inference/prediction beyond the range of the data, extrapolation is necessary in such applications.

However, the fact that extrapolation is prone to errors remains still and many authors have tried to overcome its difficulties. One problem with current approaches is that

the survival of the external population is based on past information. For instance, life expectancy of a 20-year-old person refers to a 20-year-old person 20 years ago and not to a 20-year-old today. In this sense, the external population curve that has been used numerous times as an anchor for extrapolating the survival of the group of interest is subject to improvement. Therefore, it is advised that extrapolation of survival of a group of patients be anchored to the survival of an external population with known long run survival. This external population can be either a similar group (which has a complete survival curve) or the general population that demographic agencies keep track of.

The contribution of the Chapter is a method to construct an external population survival curve as expressed in today's terms and not on outdated information. We elaborate on training richly parametric models of the poly-hazard type and argue on their statistical efficiency when learning of external information is transferred to provide robust extrapolation. Our methodology is motivated from and tested on three case studies regarding breast cancer, metastatic skin melanoma and cardiac arrhythmia. The breast cancer dataset due to the completeness of its survival curve also serves as a demonstrative tool using the ideas of interim analyses. Using the melanoma dataset, we estimate the LYG when the mRNA vaccine is received along with the pembrolizumab drug versus only receiving the current state-of-the-art pembrolizumab. Finally, regarding the arrhythmia dataset we estimate the LYG for patients that receive implantable cardioverter defibrilator (ICD) instead of anti arrhythmia drugs (AAD).

The remaining of the Chapter is organized as follows. Section 2.2 presents three case-studies that motivated our work and that survival extrapolation can be of crucial importance. Section 2.3 describes the methodology proposed, while Section 2.4 provides the results obtained. Finally, Section 2.5 concludes with a discussion of the results and ideas for future research.

2.2 Motivation and real case studies

Breast cancer is one of the most common types in women and the most common cancer in women in 157 countries out of 185 in 2022, according to the WHO[108]. Therefore, it is natural to ask about survival probability of women with breast cancer relatively to the general population.

Advanced melanoma is referred to a type of skin cancer at a metastatic stage. One therapy that has been proven to be effective includes pembrolizumab doses (see Villani et al., 2022[104] for a discussion on the currently available treatment options). Khattak et al. (2023)[59] published a hazard ratio of 0.561 with 95% C.I. (0.309,1.017) between the recurrence-free survival between groups of patients receiving "mRNA-4157 (V940) + pembrolizumab" versus "pembrolizumab". Thus, it is of interest the LYG when a patient shifts from receiving pembrolizumab alone to receiving the mRNA vaccine combined with pembrolizumab.

Cardiac arrhythmia refers to problems regarding the rate of the heartbeat. Survivors of ventricular fibrillation receive AAD treatment to control the disease, but ICD treatment has proven to be significantly more efficient (see Conolly et al., 2000[22]). We aim at specifying the LYG among patients receiving AAD or ICD treatment, through the lens of a cause-specific hazards scenario, where there exists an identifiable cause of hazard (cardiac arrhythmia) and also other causes of death (non-identifiable and grouped together).

2.2.1 Breast Cancer

We use the METABRIC breast cancer dataset which can be downloaded from cBioPortal[18] and include genomic and clinical data of women with breast cancer (see Curtis et al., 2012[24]) from Canada and the United Kingdom (UK). Keeping only the variables of interest and removing instances with unknown time and type of event, the final data are 1980 women with recorded time-to-event, type of event (death or censoring) and ages between 21 and 96 years old with median at 61 years. This dataset serves as a starting point, since the survival curve is observed for its most part and our estimates can be

validated.

2.2.2 Advanced Melanoma

The second dataset regards patients with advanced melanoma. We digitize the overall survival curve published in Hamid et al. (2019)[44] to obtain failure times of patients receiving pembrolizumab. Utilizing the estimated hazard ratio 0.561 of the recurrence-free survival between groups of "mRNA+pembrolizumab" versus "pembrolizumab", we derive the survival curve of patients receiving the mRNA vaccine (and then its extrapolation) assuming that the hazard ratio remains constant in time (this is a basic assumption that underlies the following results). The patients (62% males) were at the ages of 24-89 years with mean 59.4 and standard deviation 14.25 years.

2.2.3 Cardiac Arrhythmia

We can work in a cause-specific hazards scenario on the cardiac arrhythmia data digitised from the published Kaplan-Meier (KM) in Demiris and Sharples (2006)[26]. In this study, patients with cardiac arrhythmia receive an ICD and we derive its efficacy compared with AAD through the estimates of the hazard ratio in the meta-analysis of Conolly et al. (2000)[22], who report the ratio $h_{ICD}/h_{AAD} = 0.5$ with 95% C.I. $(0.37, 0.67)$ for arrhythmia-related deaths. Thus, we leave the "other causes" component to be the same in both groups and the external population and work only on the "cause-of-interest" component. In the ICD dataset, the average age at implant was 60 years and men represented 81% of the cohort (more information can be found in Buxton et al., 2006[16]).

2.3 Extrapolation using mortality projections

Let T be a random variable measuring time until an event with cumulative distribution function F. In survival analysis, there are two functions that play a critical role, namely the survival $S(t) := 1 - F(t)$ and the hazard $h(t) := \lim_{h \to 0} \dfrac{P(t \leq T < t + h \,|\, T \geq t)}{h}$. Given

a dataset, we build upon estimating these functions and we utilize non-parametric estimators to criticize the fit. For the survival function, we make use of the Kaplan-Meier estimator (see the original paper in Kaplan and Meier, 1958[54]) and, for the hazard function we use the ideas of Muller and Wang (1994)[74] and Gefeller and Dette (1992)[37] regarding smoothing of the hazard rate. In Appendix A2, we also give results on the cumulative hazard function $H(t) := -\int_0^t h(u)du$ (along with the non-parametric Nelson-Aalen estimator) and the odds function $O(t) := F(t)/S(t)$. Using the definitions of the survival and cumulative hazard functions and the facts that $S(t) = \exp(-H(t))$ and $h(t) = f(t)/S(t)$ (where f is the probability density function), it suffices to know only one to obtain the rest. Thus, when we estimate the hazard of a group of patients $h_1(t)$ and know it is half that of another group, we can simply write $h_2(t) = h_1(t)/2$ and, using Gauss-Kronrod quadrature, estimate the integral of the cumulative hazard and then the survival.

Let t_i be the observed survival time and δ_i its corresponding indicator of either event or censoring in a sample of size n. The poly-hazard survival models assume that the hazard function $h(\cdot)$ can be decomposed into a sum of M hazards, i.e. $h(t) = \sum_{m=1}^{M} h_m(t)$. Then, under the assumption of independence of the hazards, the survival function can be written as $S(t) = \prod_{m=1}^{M} S_m(t)$ and the likelihood of the survival model reads $L = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)$. This richly parametric family of models provides flexibility that usual survival models do not, since its hazard is not restricted to have at most one mode.

Our methodology assumes a joint model for the external and disease population, both of which adopt a poly-hazard form. Although such a model can capture the characteristics of a typical dataset quite well, its flexibility becomes its weakness when it comes to extrapolating beyond the range of given data. Thus, even during the follow-up times we imply dependence of the disease with the external times through a constraint on their corresponding components. To this end, let the disease group data have hazard $h_d(t) = \sum_{m=1}^{M} h_d^m(t)$ and the external data have hazard $h_p(t) = \sum_{m=1}^{M} d_p^m(t)$, where some of their components are either the same or proportional to each other. For instance, it could be the case that $h_d^1(t) = C \cdot h_p^1(t)$ and $h_d^3(t) = h_p^3(t)$ for $M = 3$, thus only the second

components are free. Then, the likelihood contributions from the two groups are

$$L_d = \prod_{i=1}^{n_d} h_d(t_i)^{\delta_i} S_d(t_i) \quad \text{and} \tag{2.1}$$

$$L_p = \prod_{j=1}^{n_p} h_p(t_j)^{\delta_j} S_p(t_j) \tag{2.2}$$

for samples of sizes n_d and n_p respectively, composing the total likelihood

$$L = L_d \cdot L_p \tag{2.3}$$

2.3.1 Construction of the external population

One problematic aspect of the survival extrapolation $S_{t>t^*}(t)$ after the last observed time t^* that takes place in the literature so far is the use of past information for the external population. We wish to aid those issues by projections of the mortality rate, which is estimated via a Lee-Carter model (Lee and Carter, 1992[66]). To this end, if $y_{x,t}$ is the log-mortality rates for age x and time t, then we write

$$y_{x,t} = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}$$

$$\kappa_t = u + \kappa_{t-1} + v_t$$

where $\epsilon_{x,t} \sim N(0, \sigma_\epsilon^2)$ and $v_t \sim N(0, \sigma_v^2)$. Following the Bayesian approach of Pedroza[82], we can project the mortality for age x used in the dataset for a future time t', call it ${}_{t'}m_x$. Then, the probability of reaching age $x+j$ from age x is estimated as

$${}_{t'}\pi_x^{x+j} = \exp[-\sum_{i=0}^{j-1} {}_{t'}m_{x+i}]$$

and, doing this for every age provides an estimate of the survival probability for the external population at the desired time. For instance, the probability a 20-year-old survives until the age of 22 in 2024 is estimated as ${}_{2024}\pi_{20}^{22} = \exp[-{}_{2024}m_{20} - {}_{2024}m_{21}]$. Finally, we use synthetic times until death from the external survival curve after it has been age-sex matched with the disease group demographic characteristics. The mortality data can be obtained from the Human Mortality Database[50], where we downloaded UK mortality rates of years 2000-2020 and projected until 2023. Regarding the arrhythmia

failure times in the external population, we also utilize the published proportions of arrhythmia-related deaths at every age to generate the synthesized data as in Benaglia et al. (2015)[8].

2.3.2 Extrapolation methods

Training the joint poly-hazard model, we obtain an estimate for the two total hazards $h_d(t;\boldsymbol{\theta}_d)$ and $h_p(t;\boldsymbol{\theta}_p)$. A plain extrapolation of the disease group hazard involves using the estimated parameters $\hat{\boldsymbol{\theta}}_d$ for new times $\mathbf{t}'$ as $h_d(\mathbf{t}';\hat{\boldsymbol{\theta}}_d)$. Note that even with this "vanilla" method of extrapolation, the predicted hazard implicitly uses external information via its dependence on the components of h_p that we have imposed making it more robust than an extrapolation of an independent (from the external population) poly-hazard model.

Another approach for extrapolating h_d is making the assumption that at the end of the follow-up period, the difference between h_d and h_p will determine their future (constant) difference. Thus, calculating the last k differences between h_d and h_p and taking the average, say D, provides another method of extrapolation, since we can write $h_d(\mathbf{t}') = D + h_p(\mathbf{t}')$, where $D = \frac{1}{k}\sum_{i=1}^{k} h_d(t_{n_d-i+1}) - h_p(t_{n_d-i+1})$. We can also try the same idea using a constant hazard ratio R, so that $h_d(\mathbf{t}') = R \cdot h_p(\mathbf{t}')$, where $R = \frac{1}{k}\sum_{i=1}^{k} h_p(t_{n_d-i+1})/h_p(t_{n_d-i+1})$. Note here that $h_p(\mathbf{t}')$ has already been estimated during the training procedure of the joint model.

The components of the poly-hazard models do not have a clear interpretation on their own. However, we can think of the component that induces the highest hazard at the beginning of follow-up as the one responsible for the disease of interest and the one that is increasing most slowly as the one capturing the effect of ageing. Thus, we can assume that the effect of the disease-related component remains the same between the disease and external groups and apply the previous constant difference and constant ratio methods only for the component of interest. Then, we add the other extrapolated components to obtain the total hazard in this pseudo cause-specific hazards scenario. In Appendix A2 we demonstrate each method of "vanilla", "constant difference", "constant

ratio", "pseudo cause specific constant difference" and "pseudo cause specific constant ratio".

2.3.3 Cause-specific hazards

When the hazard is decomposed into components, whose cause can be identified, then we work in a cause-specific hazards modeling procedure as follows (also see Benaglia et al., 2015[8]). Assume that the external population has two components of the same family of distributions, one of which relates with the cause of interest and the other refers to all other causes. For component $k = 1, 2$ we have failure times t_i^k and their corresponding censoring indicator δ_i^k for $i = 1, ..., n_k$, therefore the likelihood contribution of each of the components "cause of interest" and "other causes" is $L_p^k = \prod_{i=1}^{n_k} h_p^k(t_i^k)^{\delta_i^k} S_p^k(t_i^k)$. Then, the total hazard from all causes is

$$h_p = h_p^1 + h_p^2$$

where each component is trained using its own cause-specific data.

On the other hand, the disease group includes subjects that died either by the cause of interest or some other cause, but we do not have this piece of information for each patient like in the external population case. Thus, we assume a poly-hazard form with components in the same family as those in the external population. The constraint we use is that the second component is the one responsible for the "other causes" and it is thus the same as that of the external population. Regarding the first component, we assume that it is proportional of that in the external population and it is responsible for the cause of interest. Thus, we write

$$h_d(t) = C \cdot h_p^1(t) + h_p^2(t)$$

and the likelihood reads $L_s = \prod_{i=1}^{n} h_d(t_i)^{\delta_i} S_d(t_i)$ for a sample of failure times t and censoring indicators δ of size n. Finally, we combine the external population and disease group into a joint model with likelihood

$$L = L_s \cdot L_p^1 \cdot L_p^2$$

2.3.4 Mean survival and life years gained

In order to estimate the mean survival time of a group of patients, we need to calculate the area under the survival curve. RMS is given by the area under the survival curve from time $t = 0$ until the end of follow-up period and this can be very far from the true mean survival given a short follow-up. Therefore we need estimate the area under the whole survival function (including the extrapolation times) and this can be performed by many methods if extrapolation has been performed. We use the simple trapezium rule for numerical integration given by

$$\int_0^{t_{max}} S(t)dt \approx \sum_{z=1}^{N} \frac{S(t_{z-1}) + (t_z)}{2}(t_z - t_{z-1}) \tag{2.4}$$

where t_{max} is the maximum time point (when survival probability is considered to equal zero) and the $\{t_z \,|\, z = 1, ..., N\}$ is a partition of the interval $(0, t_{max})$ into N subintervals. Thus, the LYG between the survival functions $S_1(\cdot)$ and $S_2(\cdot)$ reaching probability zero at times t^1_{max} and t^2_{max} respectively can easily be estimated as

$$LYG_{S_1,S_2} = \int_0^{t_{max}} S_1(t)dt - \int_0^{t_{max}} S_2(t)dt \tag{2.5}$$

where $t_{max} = max\{t^1_{max}, t^2_{max}\}$.

2.4 Results

Due to the analytical form, easy interpretation and preferable results of the Weibull hazards, we mainly focus on poly-Weibull models, but we have also tried poly-LogNormal and poly-LogLogistic models, as well as hazard components not belonging to the same family. Also note that usual parametric families outside the poly-hazard type proved inadequate to the data considered in this article. In practise, two to three components are enough when considering poly-hazard models to provide the desired flexibility of an increasing, a decreasing and a roughly constant component, whose summation will produce a U-shaped hazard, which corresponds to early high hazard due to the disease, a drop due to overcoming the initial danger and a final increase due to age-related compli-

cations. More results for the external population mortality projections and for the breast cancer, melanoma and arrhythmia data can be found in Appendix A2.

2.4.1 Breast Cancer

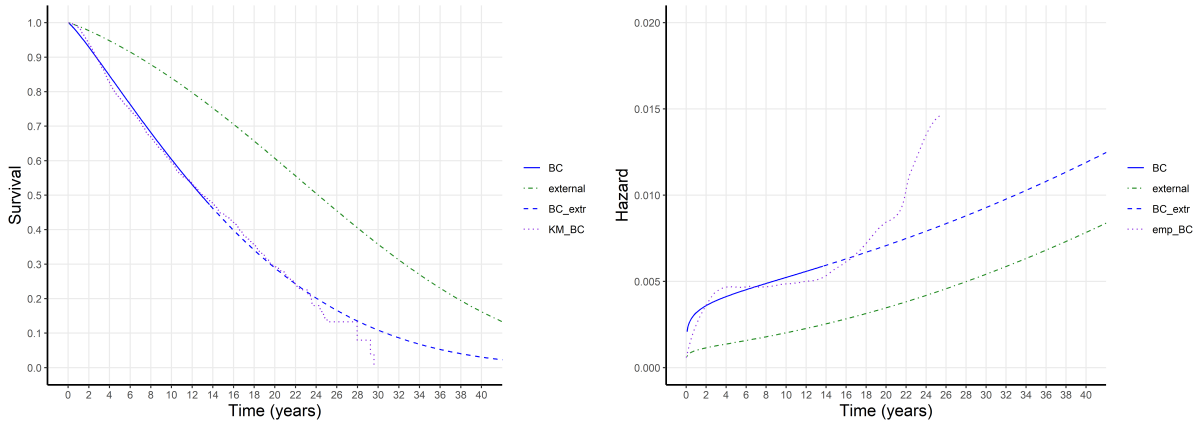

Figure 2.1: Extrapolated survival function (left) and hazard function (right) when keeping only 80% of fractional information for the breast cancer data. The breast cancer follow-up is indicated as "BC" in solid line and its extrapolation ("BC_{extr}") in dashed line. The external population is indicated with a dot-dashed line and the empirical estimation (the KM in the survival case) as a dotted line.

The breast cancer dataset has a complete at its most part survival curve, so training a model on it would only serve as a noise reduction procedure of the plain Kaplan-Meier estimate. Thus, in order to demonstrate the extrapolation methodology, we use the ideas in Demets and Lan (1994)Demets and Lan [25] regarding fractional information in interim analyses and we keep the part of the survival curve capturing 80% of fractional information, which corresponds to 49% survival probability.

Our best model for the specific dataset was a joint Bi-Weibull for the disease group and the external population. We impose the constraint that the first components of each Bi-Weibull model will be proportional to each other and the second components will be equal. Thus, we generate age-sex matched data from the external population $y_{1:n_p}$ and we assume each sample $i = 1, ..., n_p$ has hazard $h_p(y_i; \alpha_1, \alpha_2, \lambda_1, \lambda_2) = h_p^1(y_i; \alpha_1, \lambda_1) +$

$h_p^2(y_i;\alpha_2,\lambda_2)$. The hazard form of each component m is a Weibull hazard with shape and rate parameters α_m and λ_m respectively, i.e. $h_p^m(y_i;\alpha_m,\lambda_m) = \lambda_m \cdot \alpha_m \cdot y_i^{\alpha_m - 1}$. The survival function is $S(y_i;\alpha_1,\alpha_2,\lambda_1,\lambda_2) = S_p^1(y_i;\alpha_1,\lambda_1) \cdot S_p^2(y_i;\alpha_1,\lambda_1)$, where each component takes on the form $S_p^m(y_i;\alpha_m,\lambda_m) = \exp(-\lambda_m \cdot y_i^{\alpha_m})$. The likelihood is formed by equation (2.2).

For the disease data $t_{1:n_d}$, we also assume a Bi-Weibull model, but every sample $i = 1, ..., n_d$ has hazard $h_d(t_i;\alpha_1,\alpha_2,\lambda_1,\lambda_2) = C \cdot h_p^1(t_i;\alpha_1,\lambda_1) + h_p^2(t_i;\alpha_2,\lambda_2)$ and survival $S(t_i;\alpha_1,\alpha_2,\lambda_1,\lambda_2) = S_p^1(t_i;\alpha_1,\lambda_1)^C \cdot S_p^2(t_i;\alpha_1,\lambda_1)$, where $C > 0$. The likelihood is formed by equation (2.1) and, so the total likelihood is given by (2.3).

In this example, although the five extrapolation methods are similar, we prefer the vanilla method, which is the same as the pseudo cause-specific constant ratio method by construction. Thus, we use $h_d(t_i';\hat{\alpha_1},\hat{\lambda_1},\hat{\alpha_2},\hat{\lambda_2})$ for each new point t_i'. Using equations (2.4) and (2.4), we estimate a mean survival of approximately 180 months in comparison to the external mean survival of approximately 302 months, i.e. 122 life months lost due to cancer. The other extrapolation methods give similar results, except of the constant ratio method estimating approximately 173 months of mean survival, which is a bit closer to the 168 implied by the Kaplan-Meier curve. However, due to a small sample at the the of the follow-up we tend to believe the aforementioned model more than the non-parametric estimate of the KM. In Figure 2.1 we depict the extrapolated survival and hazard functions.

2.4.2 Advanced Melanoma

The best fitted model for the advanced melanoma dataset was a joint Tri-Weibull, with the constraint that the first components of the disease and external groups are proportional and the third ones are equal. Thus, we generate population data as in the breast cancer example, only here we assume $m = 3$ components, i.e. $h_p(y_i;\alpha_1,\alpha_2,\alpha_3,\lambda_1,\lambda_2,\lambda_3) = h_p^1(y_i;\alpha_1,\lambda_1) + h_p^2(y_i;\alpha_2,\lambda_2) + h_p^3(y_i;\alpha_3,\lambda_3)$. The hazard of the pembrolizumab group of patients is given the form $h_{pembro}(t_i;\alpha_1,\alpha_d,\alpha_3,\lambda_1,\lambda_d,\lambda_3) = C \cdot h_p^1(t_i;\alpha_1,\lambda_1) + h_d^2(t_i;\alpha_d,\lambda_d) + h_p(t_i;\alpha_3,\lambda_3)$. We proceed as in the previous example writing the joint likelihood.

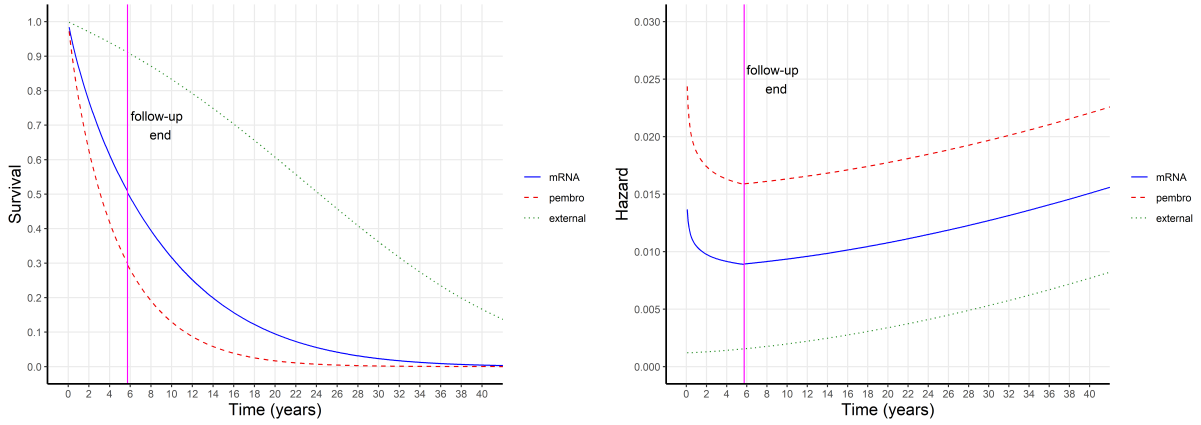

Figure 2.2: Extrapolated survival function (left) and hazard function (right) for the melanoma data. We indicate the only-pembrolizumab treated group as "pembro" (dashed line) and the pembrolizumab plus mRNA vaccine as "mRNA" (solid line). The external population is indicated as dotted line and the end time of follow-up is indicated by a vertical solid line.

In this example we suggest the results obtained using the "constant difference" method. In order to determine the number of points to use for estimating the constant difference, we need to think that the last points are more informative to estimate D_{pembro}, but using only 1 or 2 points could increase variance in D_{pembro}. We proceed with a small number of 5 points and write $h_{pembro}(t') = D_{pembro} + h_p(t')$ for a future time t'. The mRNA group of patients is implicitly derived by the pembrolizumab group by $h_{mRNA}(t) = 0.561 \cdot h_{pembro}(t)$ and extrapolation follows by its similarly estimated constant difference D_{mRNA} and writing $h_{mRNA}(t') = D_{mRNA} + h_p(t')$ for a future time t'.

Figure 2.2 depicts the extrapolated survival and hazard functions. Calculating the area between the mRNA and pembrolizumab survival functions estimates the LYG to be 43.69 months on average (see right plot of Figure 2.3), which means that patients with advanced melanoma gain 3.64 life years if they follow the treatment including the vaccine. This example also illustrates the importance of extrapolation, since the follow-up period is very short to use RMS as an estimate of mean survival (see left plot of Figure 2.3).

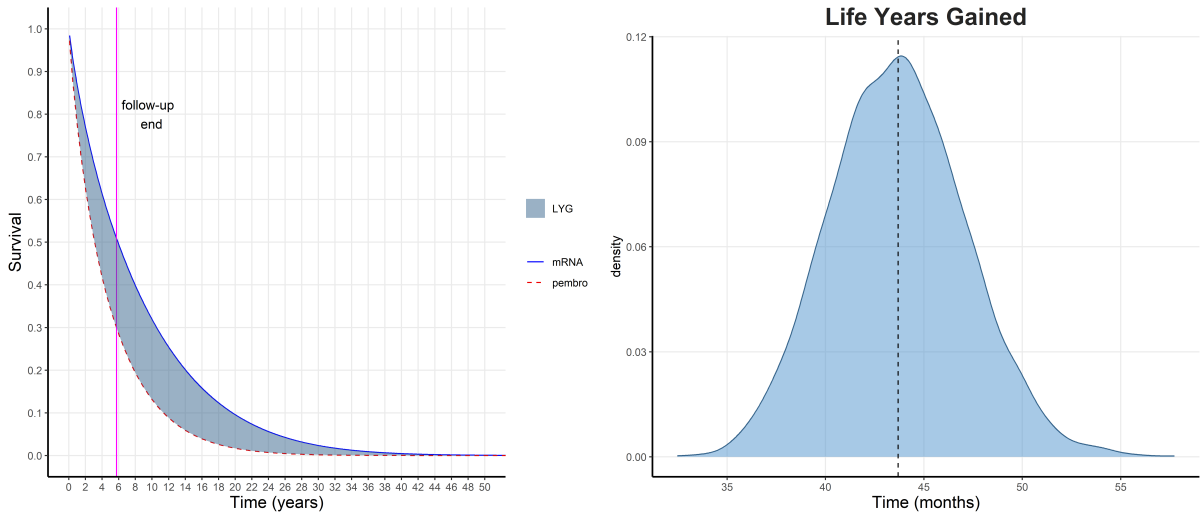

Figure 2.3: LYG for the melanoma data. Left: LYG illustrated as the area between the two survival curves with the end time of follow-up noted by a vertical line. Right: Density of the LYG with its mean indicated by a vertical line.

2.4.3 Cardiac Arrhythmia

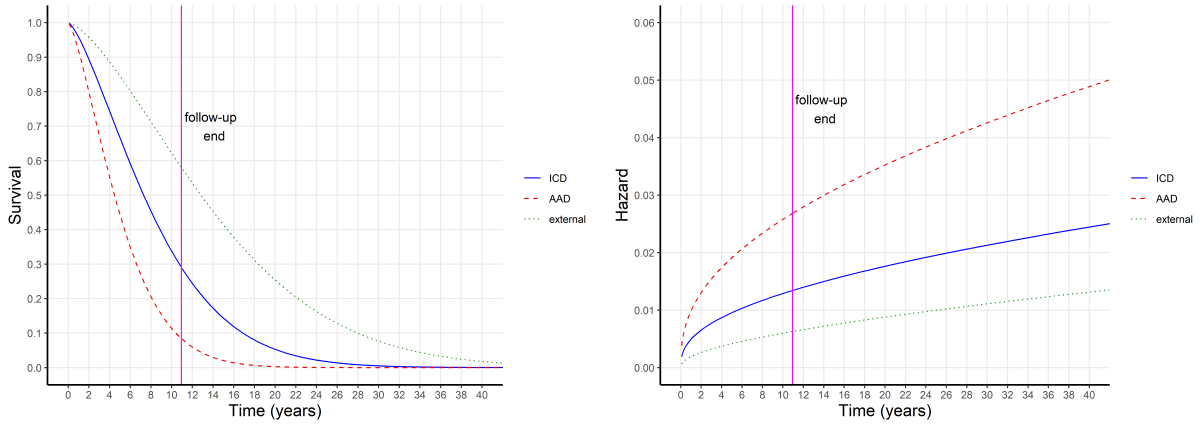

Figure 2.4: Extrapolated survival function (left) and hazard function (right) for the arrhythmia data. The external population is indicated as dotted line and the end time of follow-up is indicated by a vertical solid line.

Since Benaglia et al. (2015)[8] try a Bi-Gompertz distribution on the data and find it inferior to the Bi-Weibull, we focus on the Bi-Weibull model. Let the arrhythmia-related times until death in the external population $y_{1:n_{p_1}}$ have a Weibull hazard $h_p^1(y_i; \alpha_1, \lambda_1) =$

$\lambda_1 \cdot \alpha_1 \cdot y_i^{\alpha_1 - 1}$ and the times until death from other causes in the external population $z_{1:n_{p_2}}$ have another Weibull hazard $h_p^2(y_i; \alpha_2, \lambda_2) = \lambda_2 \cdot \alpha_2 \cdot z_i^{\alpha_2 - 1}$. Then, the joint likelihood is given by:

$$L_p(\alpha_1, \alpha_2, \lambda_1, \lambda_2) = \prod_{i=1}^{n_{p_1}} h_p^1(y_i)^{\delta_i^1} S_p^1(y_i) \prod_{j=1}^{n_{p_2}} h_p^2(z_j)^{\delta_j^2} S_p^2(z_j)$$

On the other hand, the causes of death are not known in the disease group data $t_{1:n_d}$, so we use a Bi-Weibull model with likelihood

$$L_d(C, \alpha_1, \alpha_2, \lambda_1, \lambda_2) = \prod_{i=1}^{n_d} h_d(t_i)^{\delta_i} S_d(t_i)$$

where $h_d = C \cdot h_p^1 + h_p^2$ and $S_d = (S_p^1)^C \cdot S_p^2$.

In this example, we show the vanilla extrapolation method, which is the same as the cause specific constant ratio method by construction. The AAD group is constructed by using the published hazard ratio of 0.5 (see Section 2.2), like in the melanoma case, so $h_{ICD}(t) = 0.5 \cdot h_{AAD}(t)$. In Figure 2.4 we show the extrapolated survival and hazard functions. Calculating the area between the two survival functions estimates the LYG to be approximately 39.7 months, or 3.31 years on average (see right plot of Figure 2.3). AAD patients have a mean survival of approximately 62.27 months and ICD patients have a mean survival of approximately 101.97 months.

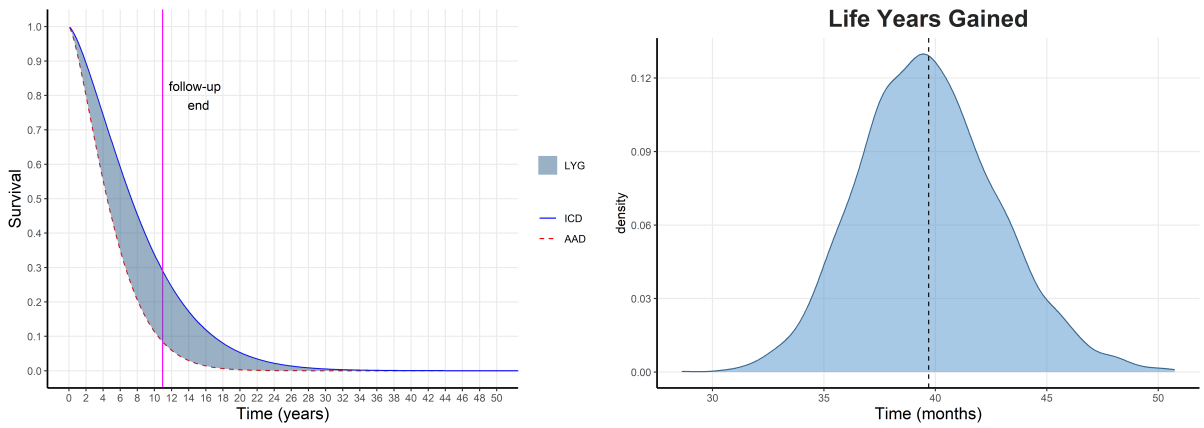

Figure 2.5: LYG for the arrhythmia data. Left: LYG illustrated as the area between the two survival curves with the end time of follow-up noted by a vertical line. Right: Density of the LYG with its mean indicated by a vertical line.

2.5 Discussion

Fitting independent poly-hazard models on the external population and the disease groups can give better results during the follow-up period, but extrapolation becomes unstable and unreliable, which illustrates the purpose of this article. The erratic behaviour can be avoided by assumptions (even during the follow-up) that transfer information from a low-variance external model. In our methods, this external model follows from projections of mortality (in contrast with existing methods) in a Bayesian evidence synthesis context. Moreover, we note the importance of working on the more sensitive hazard scale, instead of the survival, cumulative hazard or odds scale, since these three cumulative quantities tend to mask true effects. After modeling in terms of the hazard rate, we make conversions to the survival for calculation of the LYG and for demonstration purposes.

Regarding the specific case studies examined in the article, we have seen that women with breast cancer lose on average 10.17 years compared with the general population. We have demonstrated the long-term effectiveness of the mRNA-4157 (V940) vaccine in combination with pembrolizumab for treatment of advanced melanoma and concluded that 3.64 extra years of life can be gained contrasted with pembrolizumab alone. Finally, cardiac arrhythmia deaths can be lessened when switching from AAD to ICD treatment by approximately 63.8% on average.

The proposed methodology has limitations on the availability of the data, since we needed to digitize published Kaplan-Meier curves instead of working with the real datasets in the advanced melanoma and cardiac arrhythmia cases. Thus, inconsistencies related with the quality of extraction are expected. Further, the generation of synthesized data from the external population can be made more precise when expert opinion is utilized, like in the arrhythmia proportions per age in Benaglia et al. (2015)[8].

Survival modeling can be performed in other areas except of medical applications, like for instance industrial or assurance reliability functions and the idea of transferring information from an external source can be fruitful. Moreover, we have only trained poly-

hazard models with their components being independent to each other, but this need not be always the case since the introduction of copulas in this framework (see for instance Tsai and Hotta, 2013[102]).

Chapter 3

Epidemic Models for SARS-CoV-2 Transmission

All models are wrong, but some are useful.

- George Box

3.1 Introduction

The Covid-19 pandemic was initiated at Wuhan, China in late 2019 and is caused by the spread of the SARS-CoV-2 virus. The exact burden of the pandemic remains unknown and disease severity is highly variable with symptoms ranging from none or low fever to more serious, even chronic or death (see for instance Guan et al., 2020[42]). The high transmissibility of the disease (see Flaxman et al., 2020[35] and Tang et al., 2020[100] for early estimates) led to preventive measures being adopted, initially in the form of NPIs. The need for monitoring the epidemic led to the development of a vast range of models that could, in principle, assist the decision making process on the effect of NPIs against the virus. Hellewell et al. (2020)[48] assess the effect of contact-tracing and isolation, while Kucharski et al. (2020)[60] discuss the effect of travel restrictions. Eikenberry et al.

(2020)[32] investigate the use of face masks for protection. For a review over the Covid-19 modeling literature the reader can refer to Cao and Liu (2022)[17].

Traditionally, epidemic models are fitted on the recorded number of cases for inference or prediction but in the case of Covid-19 it has become apparent that only an unknown proportion of the total cases is observed. This phenomenon occurs due to the large number of asymptomatic cases as well as infected individuals not being tested for a number of reasons. Thus, statisticians should build upon these partially observed data to estimate quantities that, by definition, depend on the total number of cases.

The standard deterministic Susceptible-Infectious-Removed (SIR) model is often the basis of the analysis of infectious disease outbreak data and is typically written as

$$\begin{aligned} \frac{dS}{dt} &= -\lambda \cdot S \cdot I/N \\ \frac{dI}{dt} &= \lambda \cdot S \cdot I/N - I/\tau \\ \frac{dR}{dt} &= I/\tau \end{aligned} \tag{3.1}$$

where S, I and R denote the number of susceptible, infectious and removed individuals respectively, λ is the infection rate, τ is the infectious period and N is the size of the population. We shall return to this model in Chapter 4. The current Chapter is based upon a suitably tailored stochastic discrete-time model presented in Section 3.1.

The main contributions of this Chapter are the following. A new discrete-time stochastic epidemic framework is presented and fitted to publicly available data. The unobserved number of infectious and susceptible individuals are estimated and independently validated against external data. The marginal likelihood is analysed in detail while a distinct type of variable selection procedure is used for predicting the infection rate using mobility information. A form of Principal Components Analysis (PCA) is presented, linking the un-supervised nature of PCA to the supervised prediction task at hand.

The remainder of the Chapter is organised as follows. In Section 3.2 the models considered in the article are presented theoretically along with ideas regarding the prediction of the infection rate and the total cases and an extension allowing for re-infection. Then, in Section 3.3 we present the results of training the models on data from Greece, the United Kingdom (UK) and the United States of America (USA), the model determina-

tion procedure and predictive performance. In Section 3.4 we give an application of the Covid-19 model on explaining returns of the Greek stock market. Finally, in Section 3.5 we provide an analysis on the mobility data regarding the Principal Components (PC) and a different lens to decide on the number of PC to discard and we conclude with a discussion in Section 3.6.

3.2 Modeling framework

In this Section we describe the proposed methodology for epidemic modeling at the country level, a suitably tailored stochastic SEIR in discrete time. The model comprises of four states, namely: Susceptible, Exposed, Infectious and Removed, abbreviated as S-state, E-state, I-state and R-state respectively. The framework of the stochastic epidemic model can be applied on other communicable diseases that end up in death with some probability as Covid-19 does by suitably changing the parameters discussed next, but the focus here is on Covid-19.

3.2.1 Stochastic discrete-time transmission model

Let $\mathbf{D}$ be a random vector of daily deaths in a given country and d its recorded realization. Let also $\mathcal{F}_t$ be the history of the process $\{D_t, \theta_t \,|\, t \in \mathbb{N} \setminus \{1\}\}$ up to time t, where $\boldsymbol{\theta}$ is the vector of the conditional means of the process, i.e. $\theta_t = \mathbb{E}[D_t \,|\, \mathcal{F}_{t-1}]$. We model the mean number of daily deaths using a Negative Binomial distribution with parametrization for a random variable X and realizations x

$$NB(x; \mu, \psi) = \frac{\Gamma(x+\psi)}{x!\,\Gamma(\psi)} \left(\frac{\mu}{\mu+\psi}\right)^x \left(\frac{\psi}{\mu+\psi}\right)^\psi$$

where $\mathbb{E}[X] = \mu$ and $Var[X] = \mu + \dfrac{\mu^2}{\psi}$ with $(\mu, \psi) \in \mathbb{R} \times \mathbb{R}^+$. Thus, the likelihood reads

$$P(D_t = d_t \,|\, \mathcal{F}_{t-1}) = NB(d_t; \theta_t, \psi) \tag{3.2}$$

We wish to make inferences about the disease transmission capability based upon the total number of cases, as opposed to the observed ones. Thus, we assume that the

mean number of deaths is a proportion of the total cases that were generated during the previous days. We initiate the model assuming an equal number of cases, C, for the first $\tau + h$ days, i.e. $C_t = C_1$, for $t = 1, ..., \tau + h$. Then, for a sample of n days the model may be written as

$$\theta_t = p_t \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k \,, \quad \text{for} \quad t = 2, ..., n \tag{3.3}$$

$$C_t = \lambda_{t-1-h} \frac{S_{t-1-h} I_{t-1-h}}{N} \,, \quad \text{for} \quad t = \tau + h + 1, ..., n - 1 \tag{3.4}$$

where $p_t \in (0, 1)$ is the *Infection Fatality Ratio* (IFR) at time t, i.e. the probability of death for a given case that occurred at time $s < t$, $C_t \in \mathbb{R}^+$ are the total cases at time t, $\lambda_t \in \mathbb{R}^+$ denotes the infection rate at time t, $S_t \in \mathbb{R}^+$ and $I_t \in \mathbb{R}^+$ are the numbers of susceptible and infectious individuals respectively at time t, while $N \in \mathbb{N}$ is the (fixed) population size of the country under study. The length of the Exposed period, during which an individual is infected but not infectious, is represented by $h \in \mathbb{N}_0$, while $\tau \in \mathbb{N}$ is the infectious period. Both of those time periods are assumed to be known from previous studies.

For the infection fatality ratio, we assume a piecewise constant form with $B - 1$ change-points at times l_b, i.e. $p_t = p_{(b+1)} \cdot I\big(t \in [l_b, l_{b+1} - 1]\big)$, for $b = 0, ..., B - 1$ with $l_0 = 2$, $l_B = n$ and $p_n = p_{(B-1)}$ (note that l_0 and l_B are not actual change-points, but they are used to define the intervals of constant IFR's). For the π_{t-k} term that posits different weights to each past day, we use the sum of two Gamma distributions $\pi(t)$ as an estimate of the time from infection until death and discretize it as

$$\pi_s = \int_{s-0.5}^{s+0.5} \pi(t) dt \quad \text{, for} \quad s = 2, ..., n - 1$$

and $\pi_1 = \int_0^{1.5} \pi(t) dt$. We use these weights π_s in the model as known proportions. The infection rate is written as a piecewise constant function of time with $J - 1$ change-points at times u_j, i.e. $\lambda_t = \lambda_{(j+1)} \cdot I\big(t \in [u_j, u_{j+1} - 1]\big)$, for $j = 0, ..., J - 1$ with $u_0 = 1$ and $u_J = n - h - 1$ (like with the IFR case, u_0 and u_J are not change-points). We update at every time step the number of susceptible individuals S_t and the active set of infectious

individuals I_t as

$$S_t = S_{t-1} - C_t \tag{3.5}$$

$$I_t = \sum_{k=0}^{\tau-1} C_{t-k} \tag{3.6}$$

for $t = \tau, ..., n-h-2$. An important quantity in epidemiology of infectious diseases is the basic reproduction number R_0, typically interpreted as the number of secondary infections generated by an infectious individual in a large susceptible population. If the reproduction number R_t at day t is greater than 1 (less than) the epidemic is increasing (decreasing). This quantity is estimated using $R_t = \lambda_t \cdot \tau \cdot S_t/N$.

The description of the basic model is completed by the R-state, estimating the number of individuals that died or recovered by summing those leaving the Infectious state, i.e.

$$R_t^{(s)} = \sum_{i=1}^{t-\tau} C_i \tag{3.7}$$

for $t = \tau, ..., n-h-2$. The superscript "(s)" denotes that this is the Removed *state* at time t, as opposed to the reproduction number R_t.

Equations (3.2) through (3.7) define the dynamic stochastic discrete-time SEIR model. Note also that setting $h = 0$, the model collapses to the simple SIR type. The model represents a natural stochastic discrete time analog of the deterministic SIR system given by the ordinary differential equations (ODE) at (3.1).

Incorporating vaccination and demography

A turning point of the pandemic was the introduction of the vaccine, which played a vital role in mitigating its influence. The vaccinations are added in our model by assuming that after the first dose each individual remained susceptible for two weeks and then they became immune with probability a_1. After three more weeks (corresponding to roughly the time the second dose was done in most countries), we assume that this probability of immunity is being raised to $a_1 + a_2$. Thus, we incorporate the vaccine data by directly moving individuals to the R-state with a fixed probability. The updates of the susceptible

equation (3.5) are now performed by

$$S_t = S_{t-1} - C_t - V_t \tag{3.8}$$

where

$$V_t = a_1 \cdot \rho_{t-14} \cdot I\big(t \in [15, n-h-2]\big) + a_2 \cdot \rho_{t-35} \cdot I\big(t \in [36, n-h-2]\big)$$

and $\rho_t \in \mathbb{N}_0$ is the number of vaccinations at time t. The removed individuals at time t are now $R_t^{(s)} = \sum_{i=1}^{t-\tau} C_i + V_t$, while the infectious remain the same as in (3.6). We add further realism to the model by the inclusion of demography. This is not necessary when modeling acute outbreaks of short duration but over a period of three years one may wish to consider including births and deaths into the population dynamics. We assume that the number of births $A \in \mathbb{R}^+$ equal the one of deaths (due to reasons other than Coronavirus) and update equation (3.8) as follows:

$$S_t = S_{t-1} - C_t - V_t + A \cdot (1 - S_{t-1}/N) \tag{3.9}$$

The A term accounts for new births while the term $A \cdot S_{t-1}/N$ accounts for deaths. Furthermore, deaths from natural causes can occur inside the active set, so I_t is updated by

$$I_t = \sum_{k=0}^{\tau-1} C_{t-k} - A \cdot I_{t-1}/N \tag{3.10}$$

while the removed population is given by

$$R_t^{(s)} = \sum_{i=1}^{t-\tau} C_i + V_t - A \cdot R_{t-1}^{(s)}/N \tag{3.11}$$

Note that newborn individuals are assumed to be susceptible and so we do not add births to I_t or $R_t^{(s)}$. Thus, equations (3.2), (3.3), (3.4), (3.9), (3.10) and (3.11) synthesize the SEIR model with vaccination and demography.

The A term is assumed to be a known constant informed by the data while V_t is simply a transformation of data. The fixed parameter A can be estimated using the population distribution across the published age groups as follows. Suppose that there exist K age groups of length g_i years, for $i = 1, ..., K$, and in each group belong A_i people. Then, we assume that

$$A = \frac{A_1}{365 \cdot g_1} \tag{3.12}$$

are the newly born people set equal to the number of daily demographic deaths.

When adding vaccinations and demography to the model, $\mathcal{F}_t$ is the σ-algebra generated by $\{D_t, \theta_t, \mathbf{X}_t \,|\, t \in \mathbb{N} \setminus \{1\}\}$, where $\mathbf{X}_t$ is the "covariate" information of the weights π_t and vaccinations V_t. Note that equation (3.2) refers to the distribution of deaths conditional upon the parameters ψ, λ_t, p_t, C_1 and θ_1, so formally it should read

$$d_t \,|\, \psi, \lambda_t, p_t, C_1, \theta_1, \mathcal{F}_{t-1} \sim NB(\theta_t, \psi) \tag{3.13}$$

Prediction of the infection rate

The SEIR model we have introduced assumes a piecewise constant infection rate λ, which (in theory) can become more flexible by the introduction of covariates. The variable that we believe plays an important role on determining λ_t is the mobility of the population, since the rate of infection is analysed as the product of the probability of infection given there is a contact with the number of contacts made by an individual. The mobility data provided by Google[40] and Apple[2] are daily percent changes of mobility in certain places compared to a baseline mobility level. Google records six variables, while Apple records two, all of which are correlated to each other. Therefore, it is necessary that some form of de-correlation be done. To this end, we performed a Principal Components Analysis (PCA), which also allows for dimension reduction. In Section 3.4 we provide a more thorough analysis regarding the PCA procedure and mobility data insights.

We initially embedded this process within the wider model and used the first Principal Component (PC) of the these data as the mobility variable m_t at day t. We modelled the infection rate as a piecewise constant Autoregressive Moving Average (ARMA) process, i.e.

$$\log(\lambda_t) = \left\{ \beta_{0j} + \beta_{1j} m_t + \sum_{m=1}^{M} \phi_{mj} \log(\lambda_{t-m}) + \sum_{k=1}^{K} \delta_{kj} d_{t-k} \right\} \cdot I\big(t \in [u_j, u_{j+1} - 1]\big)$$

for $j = 0, ..., J-1$. Overall, it appears that the information in the data was insufficient to estimate the parameters that are several levels away from the data. In particular, this phenomenon has manifested itself by convergence problems of the algorithm.

Since mobility could not be inserted into the log-linear equation of λ_t, we investigated

the post-processing option of first fitting the SEIR model with a piecewise constant infection rate and then treating the produced mean λ_t estimates as a response variable in a regression scenario, i.e. ignoring the uncertainty of the λ_t estimate. In particular, we assumed that $\mathbb{E}[\log(\hat{\lambda}_t)|m_t] = g(m_t)$, where $g(\cdot)$ is a suitable function. We investigated three types of predictive models, namely Linear Regression (LR), Generalized Additive Models (GAM) and Extreme Gradient Boosted Regression Trees (XGB) and compared them in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) estimated by r-repeated k-fold Cross Validation (CV). We consider three scenarios of constructing the covariates of mobility: working with the first PC of mobility, the first two PCs, or using a specific variable chosen as described next. For each of these scenarios we test whether to smooth the variable(s) m_t or not via the discretized serial interval distribution f_t of Covid-19. In the case of smoothing we work with $m_t^* = \sum_{i=1}^{t-1} f_{t-i} m_i$, for $t = 2, ..., n-2-h$ and $m_1^* = m_1$. The discretization is performed the same way as with the π_s case. Finally, when working with the PCs we further considered two scenarios: including in the PCA procedure all of the mobility data or excluding the one referring to the places of residence. The intuition is that when this variable is excluded m_t has the natural mobility meaning, in the sense that increasing m_t leads to increased actual mobility. Otherwise, the linear combination of the mobility variables would include a variable that does not correspond to mobility, thus perplexing the interpretation.

Variable selection is performed in a way that allows the mobility variable used to contain information for both the omitted covariates and the response. Adopting matrix notation we denote m_{ik} the value of the variable k at sample i, i.e. $\mathbf{m}_k$ (for $k = 1, ..., 8$) is one of the variables to be selected. Let $\log(\lambda) = \mathbf{y}$ and $\hat{\mathbf{y}}^{(k)}$ be the fitted values of the linear regression of $\mathbf{y}$ versus $\mathbf{m}_k$, while $\hat{\mathbf{m}}^{(k)}$ be the fitted values of the multiple linear regression of $\mathbf{m}_k$ versus all the other mobility variables except $\mathbf{m}_k$. Then, we seek to solve the following problem

$$\arg\max_k \left\{ \frac{\sum_{i=1}^{n-2-h} (\hat{m}_i^{(k)} - \bar{m}_k)^2}{\sum_{i=1}^{n-2-h} (m_{ik} - \bar{m}_k)^2} + \frac{\sum_{i=1}^{n-2-h} (\hat{y}_i^{(k)} - \bar{y})^2}{\sum_{i=1}^{n-2-h} (y_i - \bar{y})^2} \right\} \tag{3.14}$$

where the bar over a variable indicates the mean over the sample data.

Prediction of the total cases

The number of tests performed every day for sure determines the number of recorded Covid-19 cases. However, this effect cannot be estimated by our models and, thus we once again used the post-processing option we also took for the mobility effect. This time, we assess the effect as follows: we aim to predict the total number of cases C_t using the number of observed cases c_t, deaths d_t, tests T_t, intensive care unit (ICU) entries H_t and the first two PC's of the 7 mobility variables smoothed by the serial interval distribution (the places of residence variable is not included). The three competing predictors are again LR, GAM and XGB.

3.2.2 Waning immunity and the SEIRS model

The SEIR model presented thus far does not allow a transition from the R-state to the S-state. While a reasonable approximation at the early phase of the pandemic, re-infections cannot be ignored for longer time-horizons. Thus, we extend the SEIR model with vaccination and demography to the SEIRS setting where recovered individuals move to susceptibility after t^* days using $r_t = (1 - p_t) \cdot \sum_{k=1}^{t-1} \pi^*_{t-k} \cdot C_k$, where π^*_{t-k} is the discretized Gamma distribution of time from infection until recovery estimated in Paul and Lorin (2021)[81]. Adding the lagged r_t to S_t in equation 3.9 via

$$S_t = S_{t-1} - C_t - V_t + A \cdot (1 - S_{t-1}/N) + r_{t-t^*} \tag{3.15}$$

allows recovered individuals to lose their immunity t^* days after infection.

3.2.3 Bayesian inference

A key modelling decision relates to the selection of the death data as the basis for inference on the total cases and disease transmisibility. While this makes sense in terms of data quality, it implicitly precludes the use of case data for learning those parameters,

effectively "cutting feedback" from the observed case data. This approach was introduced in Spiegelhalter et al. (2007)[94] and has been adopted in numerous studies in order to prevent data of low quality contaminating inference (see Plummer, 2015[83] for a review). We adopt this approach here, thus leading to a two-stage inference procedure, see for example Figure 3.2 where the observed cases are used retrospectively for estimating quantities like the observed proportion.

Computation

The SEIR model described above can be fitted by the Bayesian methodology of incorporating prior information to the unknown parameters and updating the uncertainty of the likelihood. The complexity of this non-linear stochastic model implies that no analytic likelihood calculations exist, thus we resort to drawing samples from the posterior using Hamiltonian Monte Carlo (HMC).

To this end, we define a Hamiltonian function $\mathbb{H}(\mathbf{q}, \boldsymbol{\rho})$ as the sum of the potential energy function $U(\mathbf{q}) = -\log\Big(p(d_t|\mathbf{q}) \cdot \prod_{i=1}^{M} p(q_i)\Big)$ and the kinetic energy function $K(\boldsymbol{\rho}) = \sum_{i=1}^{M} \frac{\rho_i^2}{2m_i}$, where $\mathbf{q}$ and $\boldsymbol{\rho}$ are the vectors of model variables and auxiliary variables respectively, each of dimension $\dim(\mathbf{q}) = \dim(\boldsymbol{\rho}) = M$, $p(d_t|\mathbf{q}) = P(D_t = d_t \,|\, \mathcal{F}_{t-1})$ is the likelihood function, $p(\mathbf{q}) = \prod_{i=1}^{M} p(q_i)$ is the prior and m_i's are the mass elements. Hamilton's equations become

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\rho_i}{m_i} \\ \frac{d\rho_i}{dt} &= \frac{\partial}{\partial q_i} \log\Big(p(d_t|\mathbf{q}) \cdot \prod_{i=1}^{M} p(q_i)\Big) \end{aligned}$$

for every $i = 1, ..., M$. The simulated dynamics propose at each iteration a solution $(\mathbf{q}^*, \boldsymbol{\rho}^*)$ which is accepted with probability

$$\min\left\{1, \exp\left[U(\mathbf{q}) - U(\mathbf{q}^*) + \sum_{i=1}^{M} \frac{\rho_i^2 - \rho_i^{*2}}{2m_i}\right]\right\}$$

In this research, we use the NUTS algorithm of Hoffman and Gelman (2014)[49] to perform the procedure with minimum need of tuning.

Prior specification and elicitation

We assume *a priori* independent distributions on the B infection fatality ratios $p_{(b)}$, the J infection rates $\lambda_{(j)}$, the dispersion parameter ψ and the initial value of total cases C_1 and fitted deaths d_1. The first $\tau + h$ values of C_t are assumed equal and are given a $Gamma(2, 0.0625)$ prior. The initial value of the susceptible population is then set to $S_1 = N - C_1$. For each lambda between change-points u_j and $u_{j+1} - 1$ we assign a Log-Normal $LN(0, 1)$ prior. The dispersion parameter ψ of the Negative Binomial is allocated a $Gamma(2, 0.125)$ prior. While these prior distributions are weakly informative, we also conducted sensitivity analyses to assess their effect as typically done in Bayesian robustness settings. We use a fixed $\tau = 6$ infectious period (Cereda et al., 2020[20]; Flaxman et al., 2020[35]), and a fixed exposed period of $h = 2$, since the mean incubation period is approximately 5 days (Lauer et al., 2020[64]) and infectiousness starts approximately 2 days before the symptom onset (He et al., 2020[45]).

The prior on the IFR parameters $p_{(b)}$ requires particular caution for two reasons. First, it may not be informed by the outbreak data and, second it largely drives the scale of the estimated total cases and functions thereof. For each interval of constant IFR, the prior was set to a strongly informative Gaussian distribution with standard deviation of 10^{-4} and mean computed as follows. First, we scaled the IFR published by the Centers for Disease Control and Prevention (CDC) for the 4 age groups 0-17, 18-39, 40-64 and 65+, $p_k^{(0)}$ $(k = 1, ..., 4)$, according to the age distribution of the recorded cases in the country under study. Then, we locate the change-points by inspecting the observed country-specific IFR and use the mean IFR inside each time interval. Thus, if we denote by $c_{t,k}$ the cases at time t for age group k in a given country, the IFR is computed as $p_t^* = \sum_{k=1}^{4} p_k^{(0)} \frac{c_{t,k}}{\sum_{i=1}^{4} c_{t,i}}$ and we set the mean of the Gaussian prior to $\mathbb{E}[p_{(b)}] = \frac{1}{l_{b+1} - l_b} \sum_{t=l_b}^{l_{b+1}-1} p_t^*$. We also used this mean as a point mass prior for IFR but this approach was too rigid in some cases and we therefore opted for this strongly informative Gaussian prior.

3.3 Results

In this Section, we analyse the pandemic data from the UK, Greece and the USA. Modeling the entire USA as a single entity may not be entirely appropriate due to the inherent heterogeneity stemming from both the large geographic area and population size so these results should be interpreted with caution. Other models considered for Covid-19 are given in Appendix A3, where we build upon the final methodology presented testing assumptions and different forms of the suitable models.

We present the results from the NUTS implementation of HMC. We also tried automatic differentiation variational inference (Kucukelbir et al., 2017[61]) but, while this was faster in all case, it rarely gave reliable results in all but the simplest of models for short time horizons. In addition, we used Simulated Annealing (SA) in order to maximise the un-normalized posterior. Extensive searches by SA required less than 10 minutes when sampling took more than two days. However, the SA results were very sensitive to initial values while HMC appeared stable and robust. We therefore report HMC-based results due to statistical but not computational efficiency and retained HMC as our preferred algorithm for inference.

3.3.1 Sources of evidence

The publicly available data we leverage include the recorded cases and deaths, the number of vaccinated individuals, the age distribution of cases, the demographic births and deaths, the country population, the number of tests performed and quantities estimated at the beginning of the pandemic, which are the time from infection until death, the infectious period, the exposed period and the serial interval. Below we outline the values we use for these quantities (also see Table 3.1).

For the UK, the cases and death data as well as the vaccination coverage and age distribution of cases are provided in the Public Health England[33] website. For the parameter A we use demographic evidence[27] and calculate it using $g_1 = 5$ and $A_1 = \frac{6.2 \cdot N_{UK}}{100}$. We use $N_{UK} = 67886011$ as the 2020 UK population size[109].

The Greek death data are obtained from the COVID-19 Data Repository of the Center for Systems Science and Engineering at Johns Hopkins University[29]. The vaccine doses received are taken from Our World in Data[70], while the age distribution of cases and the daily tests performed were taken from the GitHub page of Sandbird[91]. The demographic births and deaths are obtained from the Hellenic Statistical Authority (HSA[46]) which provides the population age groups and so we use equation (3.12) to calculate A with $g_1 = 10$ and $A_1 = 1049839$. The population of Greece in 2020 is set to $N_{GR} = 10816286$ following the HSA estimate.

The USA death data are taken from the Johns Hopkins University while the cases and the vaccinations are obtained from Our World in Data. The population is set to be $N_{USA} = 331230000$ and we calculate A using $g_1 = 15$ and $A_1 = \frac{18.51 \cdot N_{USA}}{100}$. For these we use the `statista.com` page, specifically the published total population[84] and age distribution[28] in 2020.

For the infection-to-death distribution we use the sum of the infection-to-onset of symptoms and onset-to-death distributions, which are Gamma with shapes 1.35 and 4.94 and rates 0.27 and 0.26 respectively, as in Flaxman et al. (2020)[35]. The same source was used for the serial interval distribution as well. The $p_k^{(0)}$ IFR can be found in the CDC[19] page. The probabilities of moving to the R-state are set to be $a_1 = 0.4$ and $a_2 = 0.1$.

	Greece		UK	
	Value	Source	Value	Source
c_t		Johns Hopkins University		Public Health England
d_t		Johns Hopkins University		Public Health England
ρ_t		Our World in Data		Public Health England
$c_{t,k}$		Sandbird Github		Public Health England
A	287.63	Hellenic Statistical Authority	2306.26	Public Health England
N	10816286	Hellenic Statistical Authority	67886011	Worldometer
T_t		Sandbird Github		
$\pi(t)$		Flaxman et al. (2020)[35]		Flaxman et al. (2020)[35]
τ	6	Cereda et al. (2020)[20], Flaxman et al. (2020)[35]	6	Cereda et al. (2020)[20], Flaxman et al. (2020)[35]
h	2	Lauer et al. (2020)[64], He et al. (2020)[45]	2	Lauer et al. (2020)[64], He et al. (2020)[45]
f_t		Flaxman et al. (2020)[35]		Flaxman et al. (2020)[35]

Table 3.1: Sources we use for the data and fixed quantities for the country-specific models of Greece and UK.

3.3.2 Epidemic parameters and functions thereof

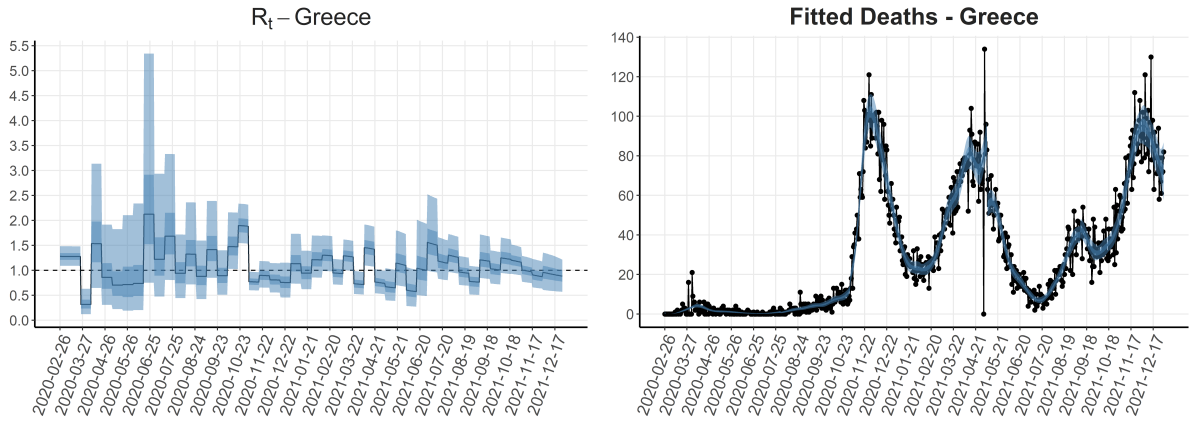

Figure 3.1: Posterior estimates of R_t (left) and θ_t (right) for Greece during the first two years of the epidemic. The median is depicted with a solid black line, along with 50% and 95% credible intervals.

We focus on the acute phase of the pandemic and fit the model to data from Greece covering the 26/2/2020 to 31/12/2021 period. We independently validate our estimates based on a large UK seroprevalence study.

In Figure 3.1 the estimated reproduction number, R_t, and mean deaths θ_t are depicted. The reasonable fit to the death data is reassuring. The piecewise constant R_t is scaled by the proportion of susceptible individuals as the theory suggests. The total number of cases at day t, C_t, i.e. is the sum of the recorded and unrecorded ones are depicted in the left plot of Figure 3.2 while the right panel contains the smoothed estimated proportion of observed cases, c_t/C_t (see Appendix A4 for the results of the SEIRS model for Greece.). We estimate that the first million infections (10% of the total population) was reached on April 2021 but observed eight months later on December 2021. The probability of recording a case initially was around 1/4 but then increased as tests and self-tests became widely available, closer to 3/4 of the total cases.

The external validity of our model was assessed by training the model on the UK death data and superimposing on Figure 3.3 the estimated number of infections against the REACT study (Ward et al., 2021[105]) which estimated an approximate 6% prevalence

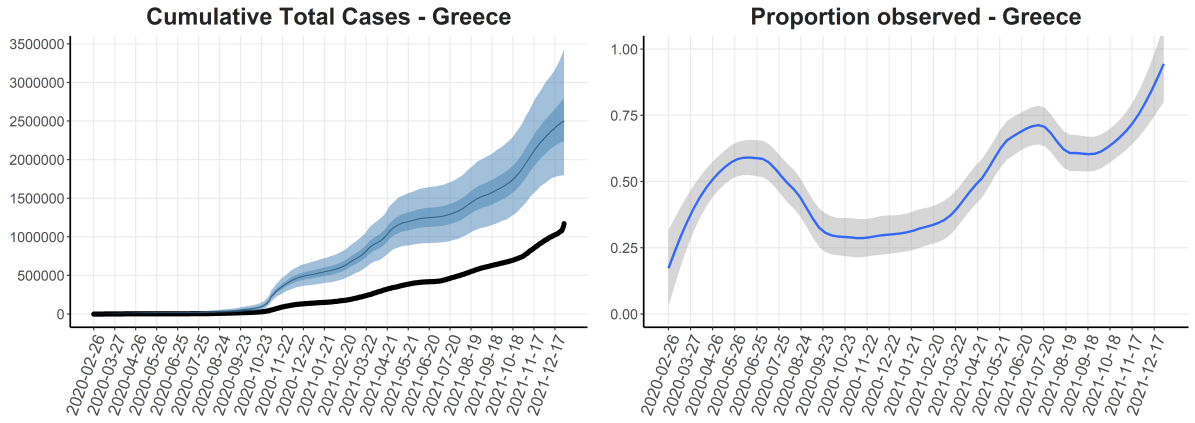

Figure 3.2: Left: Cumulative total cases: posterior median (solid black line) with 50% and 95% credible intervals. Right: Proportion of observed cases smoothed by local regression.

on July 2020. It is apparent that our model estimate agrees well with this independent data source.

3.3.3 Sensitivity analysis and model comparison

As part of the model determination process we tested the initial values, the prior distributions, we added vaccination and demography to the model and tested seven scenarios where we remove the exposed period, vaccination or demography. In Appendix A4 we provide in Table 5.16 a comparison of the eight models: four SIR and four SEIR each with zero to two of the extensions "vaccination" and "demography". The comparison is made using the information criteria AIC, BIC, DIC, DIC using half the variance of the deviance (referred to as DIC_2) and WAIC (for their definition, the reader can refer to Gelman et al., 2014[38]). The time needed to fit each model is also noted. Also in the Appendix A4, we compare in Table 5.17 the logarithm of the evidence $p(d_t)$ for each model using Bridge sampling and pairwise Bayes factors. Overall, the SIR with demography is selected by AIC and BIC, SEIR with vaccination or demography are preferred by DIC, the simple SIR is selected by DIC_2, while the SEIR with vaccination and demography is preferred by WAIC.

All the training times are similar and approximately equal to 2.5 days. The SEIR

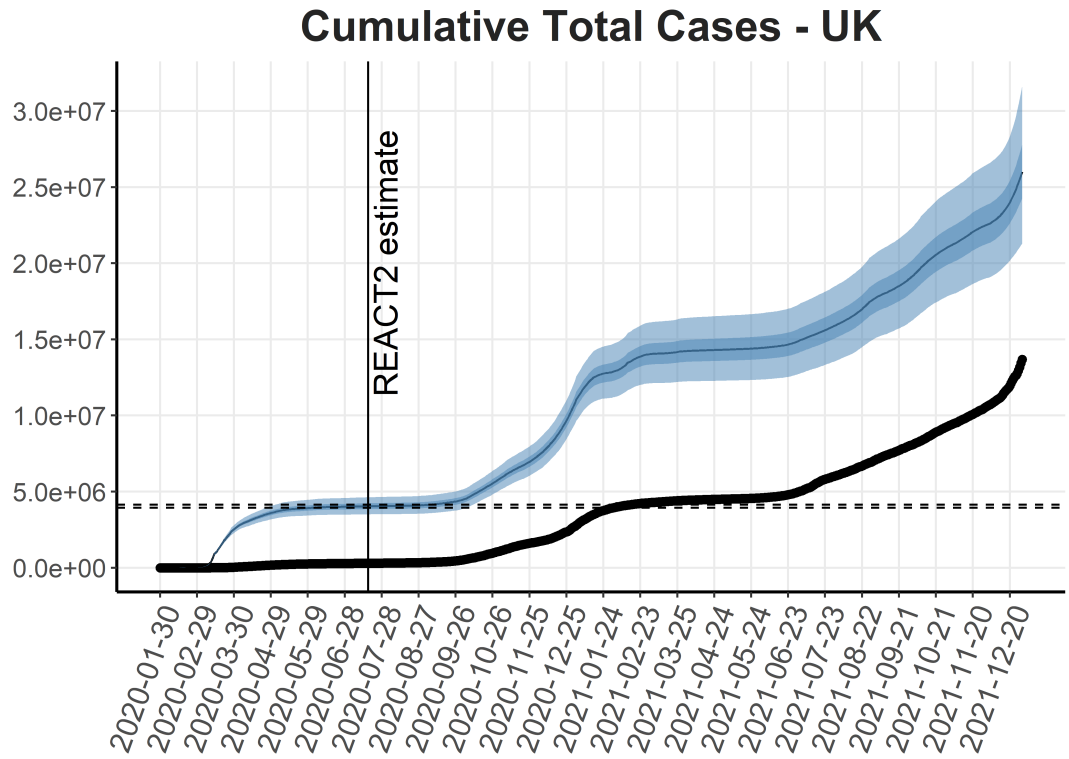

Figure 3.3: Cumulative total cases for the UK. The time when the REACT-2 study estimated the total cases is noted as a vertical line, while the reported 95% confidence interval is shown as horizontal dashed lines.

model with vaccination has the largest marginal likelihood, slightly higher than that of the SEIR with vaccination and demography while the interval estimates of the corresponding $\log p(d_t)$ show substantial overlap. Fitting the SEIRS model to the Greek data, allowing for re-infection 3 months after recovery ($3 \cdot 4 \cdot 7$ days) did not make any material difference to the SEIR version except for the last month, when the characteristics of the epidemic also change.

The distributional form of the likelihood of the death data was tested as follows. The Negative Binomial is a mixture distribution of a Poisson with a Gamma prior on the rate, so we can change this prior to any distribution with support on the positive real line. To this end, we examined the behaviour of fitting Exponential and LogNormal prior distributions. The variance for the LogNormal prior is given the same prior as the dispersion of the Negative Binomial and the means for both the Exponential and the

LogNormal are such that correspond to a mean of the previously used θ_t. Thus, the first level of the model now is $D_t \sim P(\theta_t)$ with $\theta_t \sim Exp(1/\mu_t)$ or $\theta_t \sim LNorm(m_t, s_t^2)$, where $m_t = \dfrac{\log(\mu_t^2)}{\sqrt{(\mu_t^2 + \sigma^2)}}$ and $s_t^2 = \log(1 + \sigma^2/\mu_t^2)$. Then, $\mu_t = p_t \sum \pi_{t-k} C_k$ as we have with the Negative Binomial case. The Poisson-Exponential (which is a special case of the Negative Binomial) fits the data well and also reduces training time to about a half. On the other hand, the Poisson-LogNormal struggled to converge and was unreliable. The one-point estimates of training time for the Negative Binomial, Poisson-Exponential and Poisson-LogNormal are 2.39, 0.99 and 19.82 days using the same settings for the NUTS algorithm. The simple Poisson is similar to the Negative Binomial, but needed 4.83 days. The dispersion parameter is estimated to have a median of 42.1 (95% CrI: $(31.9, 56.2)$), thus not very far from 0, as the Poisson suggests.

3.3.4 Transition to endemicity

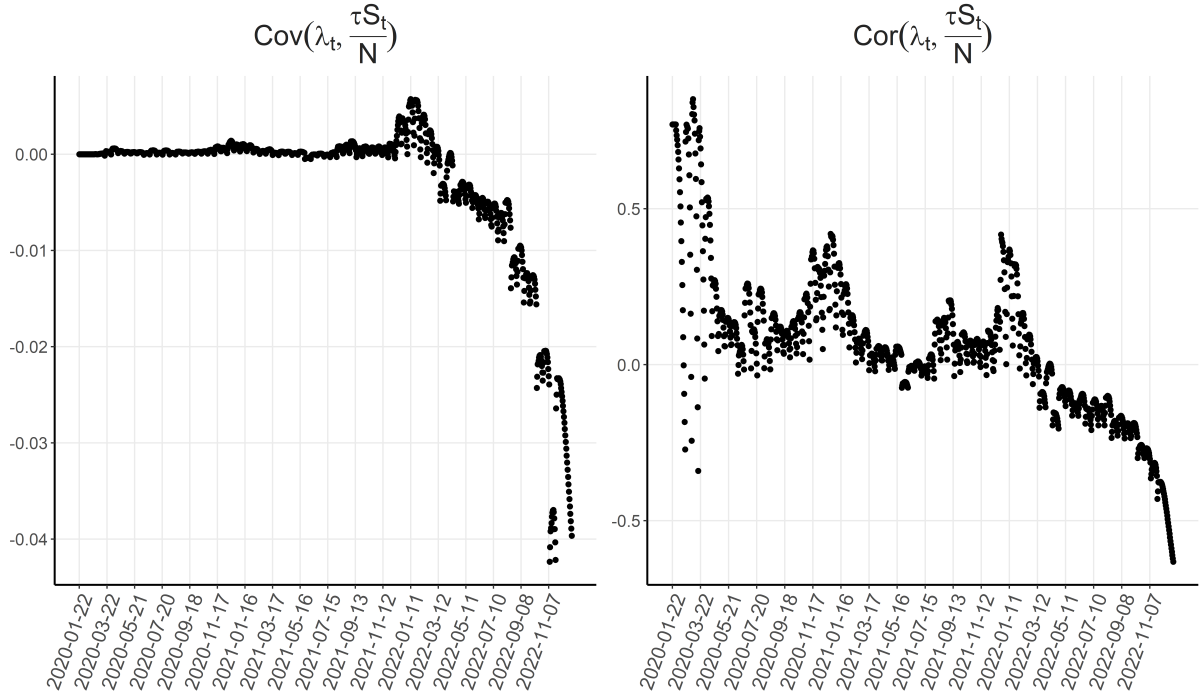

Figure 3.4: Daily estimates of the covariance (left) and correlation (right) between λ_t and $\tau S_t/N$ for USA.

Defining the end of the acute phase of the epidemic is a non-standard problem. The

definition of the reproduction number $R_t = \lambda_t \cdot \tau \cdot S_t/N$ implies that at the early stage of the epidemic R_t is approximately proportional to the infection rate. The covariance of the infection rate and the scaling factor $\tau S_t/N$ offers an insight to when those two diverge (see Figure 3.4) whence the covariance becomes negative. Another insight of the divergence between λ_t and R_t can be given through the estimated means of the infection rate λ_t and reproduction number R_t in Figure 3.5, where λ_t seems not to follow the same trend as R_t.

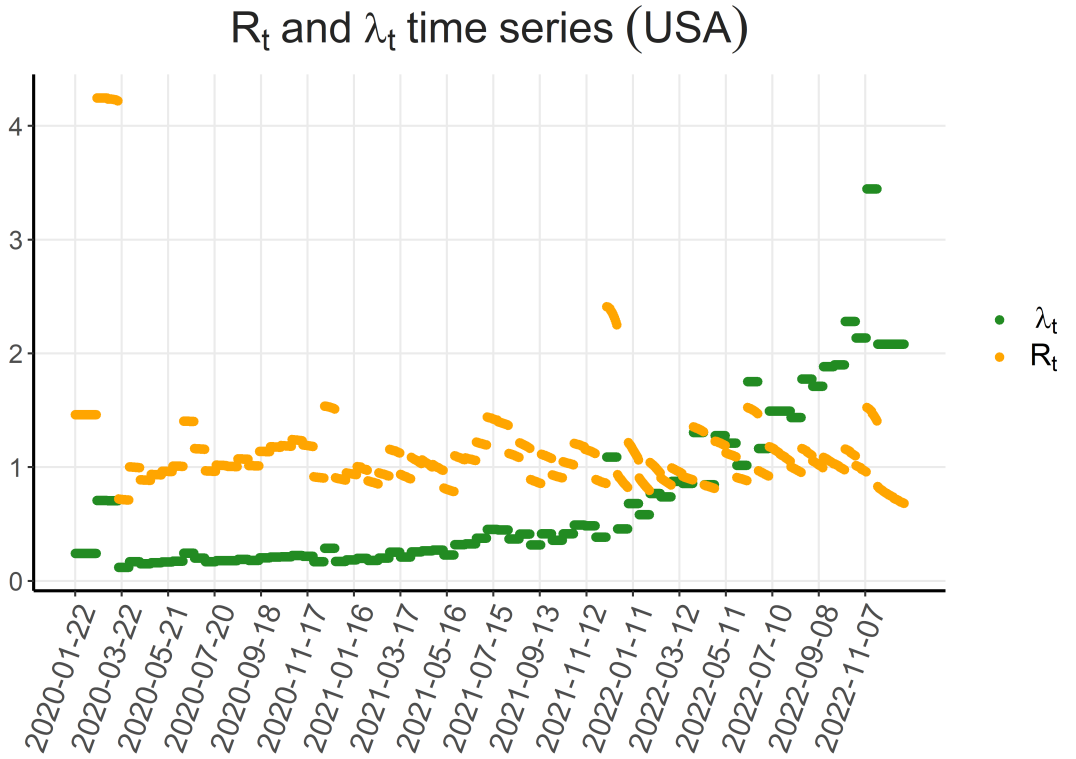

Figure 3.5: Daily estimates of the mean infection rate and reproduction number as time passes in USA.

3.3.5 Prediction of λ_t and C_t

We use PCA to reduce the 8 dependent mobility variables to 1 or 2 keeping 78.09% or 89.25% of the initial variance. The prediction of the infection rate using mobility data is summarised in Table (3.2). Both RMSE and MAE suggest that the best prediction is achieved by the XGB model using the 2 smoothed PC's where the places of residence

			LR	GAM	XGB
1 PC	Ser Int	All in	0.3267 [0.2620]	0.3120 [0.2484]	0.3043 [0.2410]
		No resid	0.3284 [0.2631]	0.3140 [0.2498]	0.2998 [0.2372]
	No Ser Int	All in	0.3336 [0.2723]	0.3312 [0.2698]	0.3197 [0.2624]
		No resid	0.3346 [0.2734]	0.3297 [0.2685]	0.3178 [0.2607]
2 PC	Ser Int	All in	0.3153 [0.2500]	0.2900 [0.2220]	0.2800 [0.2142]
		No resid	0.3128 [0.2479]	0.2927 [0.2260]	0.2699 [0.2064]
	No Ser Int	All in	0.3333 [0.2712]	0.3267 [0.2662]	0.3162 [0.2595]
		No resid	0.3337 [0.2713]	0.3247 [0.2646]	0.3137 [0.2563]
Variable Selection		Ser Int	0.3231 [0.2608]	0.3072 [0.2365]	0.3018 [0.2381]
		No Ser Int	0.3299 [0.2699]	0.3202 [0.2654]	0.3128 [0.2597]

Table 3.2: Prediction of the infection rate: 5-repeated 10-fold CV prediction error estimates for each model and scenario as captured by RMSE and MAE (in square brackets).

	RMSE	MAE
LR	1.0741	0.8309
GAM	0.5295	0.3888
XGB	0.5040	0.3669

Table 3.3: Prediction of the total cases: 5-repeated 10-fold CV prediction error estimates for each model and scenario as captured by RMSE and MAE.

are excluded. The overall prediction ability is generally low. The mobility variable that describes best all the others and simultaneously serves as a good predictor of the infection rate is the one associated with transit stations, which is rather intuitive since public transport relates well to contacts. The XGB model outperforms its competitors in predicting total cases too (see Table 3.3).

Broadly, the first PC summarizes the actual mobility in the sense that it takes positive values when mobility increases, while negative values correspond to increased stay in places of residence. The second PC is used only for prediction purposes and does not have a clear interpretation. Thus, fitting a mixture of two Gaussian distributions

using the Expectation-Maximization algorithm to the scores of the first PC, we find that "mobility" mainly stems from $0.31 \cdot N(-2.86, 1.2^2) + 0.69 \cdot N(1.26, 1.78^2)$. In Figure 3.6, we visualize the two Gaussian components that prevailed at each day during Covid-19 waves in Greece and we can see for instance that during the two lock-downs in spring 2020 and winter of 2020 to 2021, mobility was mostly generated by the first component of the mixture, which corresponds to the prevalence of the places of residence. For more details on the mobility data and the PCA procedure, as well as the EM algorithm, see Appendix A4.

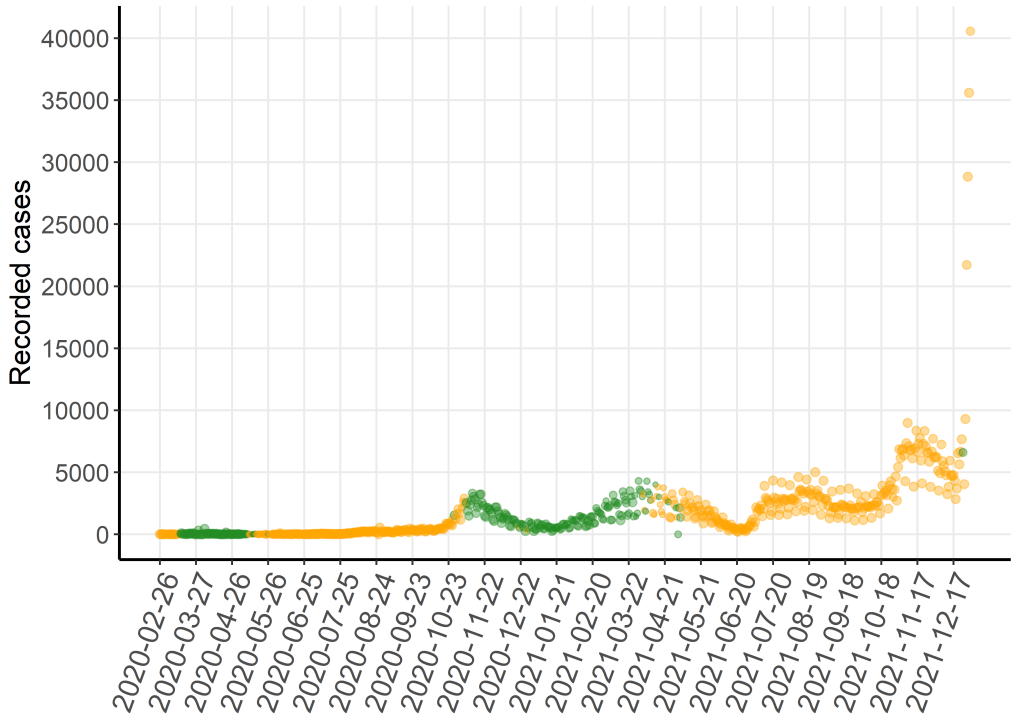

Figure 3.6: Recorded cases in Greece with green color indicating a point belonging to the first of the two Gaussian components and orange indicating a point from the second component. The classes are determined by the probability of belonging to each component estimated by Expectation-Maximization. Also, the larger the probability, the larger the size of the point.

3.4 Modeling Greek stock market returns during Covid-19

Financial markets play a critical role in modern economies, since they allocate resources and provide liquidity in businesses. However, asset prices present high volatility in presence of random factors and events such as the shock of the Covid-19 pandemic (see Pagano et al., 2020[80] for the resilience of companies during Covid-19), thus the effect of the pandemia is still under investigation (see for instance Szczygielski et al., 2023[99] and Cox and Woods, 2023[23]).

To this end, we utilize monthly return series of the Greek stock market ranging from February 2020 to October 2022 to train linear and quantile regression models, as well as break-point models, i.e. models of the following forms. For the standard linear regression model, we have

$$y_t = \alpha + \sum_{k=1}^{K} \beta_k x_{k,t} + \epsilon_t$$

where y_t is the Greek financial stock market index at time t, $x_{k,t}$ are K explanatory variables and ϵ_t are iid innovation terms with mean zero and variance σ^2. In the quantile regression, instead of the conditional mean, we model the quantiles of the response variable, so that

$$y_t = \alpha^{(\tau)} + \sum_{k=1}^{K} \beta_k^{(\tau)} x_{k,t} + \epsilon_t$$

where the parameters are associated with the τ'th quantile. The errors are assumed to be independent from a distribution with τ'th quantile equal to zero. Finally, for the break-point models we write

$$y_t = \alpha_j + \sum_{k=1}^{K} \beta_{j,k} x_{k,t} + \epsilon_t \quad , \tau_{j-1} < t < \tau_j$$

where $\epsilon_t \sim N(0, \sigma_j^2)$ and $j = 0, ..., m+1$, i.e. the parameters are allowed to change inside each interval (τ_{j-1}, τ_j). For estimation we use the method proposed by Meligkotsidou and Vrontos (2008)[71], which detects structural breaks in the series. The explanatory variables used are summarized in Table 3.4 with the corresponding transformation of each in order to achieve stationarity.

Explanatory Factors	Transformation
Market index - lag one	Δln, month-on-month % change
European Market index	Δln, month-on-month % change
Economic Policy Uncertainty index - Europe	Δlv, month-on-month differences
Economic Policy Uncertainty index - Greece	Δlv, month-on-month differences
Implied Volatility index	Δlv, month-on-month differences
Crude Oil Brent price - Europe	Δln, month-on-month % change
Long term interest rates	Δlv, month-on-month differences
Infectious disease tracker	Δlv, month-on-month differences
Reproduction number	-

Table 3.4: The set of explanatory variables and the corresponding transformation used in the analysis; Δln denotes first differences of logarithms and Δlv denotes first differences.

Regarding the mean and quantile regression models, taking into account posterior model probabilities for every possible model using the variables in Table 3.4, we have found that the European market index is the most important explanatory variable (inclusion probability 90%-96%) of Greek market returns, while the reproduction number also has a profound impact (inclusion probability slightly above 50%). We have to note that the low posterior model probabilities, as well as different performance on different quantiles suggest that there is no single model capable of explaining market returns. However, the European market index and the reproduction number appeared in the most probable models. The former has a statistically significant positive effect in all tested models, while the latter has significant negative effect on the financial return series using the conditional mean and conditional 25'th and 75'th quantile models. Regarding the break-point models, the data suggest that a factor model with one break on the European market index is best among those considered.

The current Section, which is a joint work with Dr Ioannis D. Vrontos (who leaded the research), is undertaken as a research project supported by the Hellenic Foundation for Research and Innovation under the 4th Call for Action "Science and Society" - Emblematic Action - "Interventions to address the economic and social effects of the

COVID-19 pandemic" (Project Number: 4887).

3.5 Distortion after PCA dimension reduction

The present Section aims at determining the number of useful mobility principal components. The mobility variables are described in Tables 3.5 and 3.6; in Figure 3.7 the series of the eight variables are displayed. For convenience in displaying the Figures the names of the variables are concatenated to "rere", "gropha", "par", "trast", "worl", "resid", "driving" and "walking" respectively for "Retail & recreation", "Grocery & pharmacy", "Parks", "Transit stations", "Workplaces", "Residential", "Driving" and "Walking". In this analysis, we focus only on Greece.

We return to the problem of the number of useful PC, but now seen through a different lens. We propose three new measures of the induced distortion of discarding k out of the original L variables measured on n observations. Let $\mathbf{X}$ be the $n \times L$ matrix of the centered and scaled data. PCA applies a transformation $\mathbf{A}$ on $\mathbf{X}$ (the eigen-decomposition of the covariance) resulting in a new data matrix $\tilde{\mathbf{X}}$. In the case of no dimension reduction $\mathbf{X} \cdot \mathbf{A} = \tilde{\mathbf{X}}$ is simply a rotation of the original data, so applying the inverse results again in $\mathbf{X}$. Otherwise $\tilde{\mathbf{X}} \cdot \mathbf{A}_k^{-1}$ represents a distorted image of $\mathbf{X}$. Let $\tilde{\mathbf{X}}_k$ be the distorted dataset with k the number of retained basis vectors. Note that $SS_{k,l} = (\mathbf{X}_l - \tilde{\mathbf{X}}_{k,l})^2 \cdot \mathbf{1}'$ is a similarity measure between the original variable l of $\mathbf{X}$, and the new variable l, $\tilde{\mathbf{X}}_{k,l}$ quantifying the information lost from variable l when k out of L variables are retained ($\mathbf{1}$ is a vector of n 1's). Thus, the decision of keeping k or $k-1$ components can be assisted by the information lost from a specific variable calculating $(SS_{k,l} - SS_{k-1,l})/SS_{k-1,l}$. Generalizing, we can find the percent change on the distortion over all variables by $\sum_{l=1}^{L}(SS_{k,l} - SS_{k-1,l})/\sum_{l=1}^{L} SS_{k-1,l}$, where $SS_{k,l}$ is the distortion of variable l. This kind of information is already given by the proportion of variance explained, although in a different form.

However, the distortion of each variable also helps in scenarios where the number of components should be decided based on adequate participation of a specific variable in

Google Mobility variables	
Variable	Description
Retail & recreation	Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries and movie theaters
Grocery & pharmacy	Mobility trends for places like grocery markets, specialty food shops, drug stores and pharmacies
Parks	Mobility trends for places like national parks, public beaches, marinas, dog parks, plazas and public gardens
Transit stations	Mobility trends for places like public transport hubs such as subway, bus and train stations
Workplaces	Mobility trends for places of work
Residential	Mobility trends for palces of residence

Table 3.5: Mobility variables published by Google

Apple Mobility variables	
Variable	Description
Driving	Reports describing direction requests on Apple maps
Walking	Reports describing direction requests on Apple maps

Table 3.6: Mobility variables published by Apple

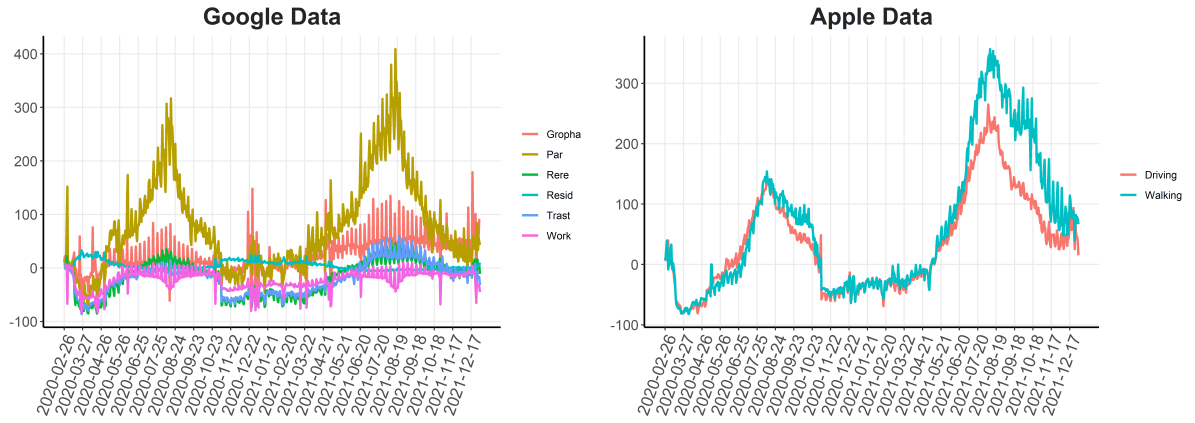

Figure 3.7: Left: The six recorded variables from Google. Right: The two recorded variables from Apple. The y-axis represents percentage change from baseline.

the resulted PC's, for instance when another analysis is to be performed based on the reduced dataset. It moves PCA from a purely unsupervised technique towards a more supervised scheme without changing the algorithm itself like for instance the partial least squares idea. Lastly, in $SS_{k,l}$ we sum the squared differences of the old and new data for variable l, but any other similarity measure can be used like in a clustering scenario (maximum distance, mean distance, minimum distance etc). For instance, instead of the sum of squares, we can measure the distortion of each variable via $W_l = \max_i |(\mathbf{X}_l - \tilde{\mathbf{X}}_{k,l})|$ for $i = 1, ..., n$.

Another natural way to quantify the distortion of each variable is by the magnitude of the scatter induced by the transformation $\mathbf{A}_k^{-1}$. Since $\mathbf{A}_L^{-1}$ returns the original data, plotting $\mathbf{X}_l$ against $\tilde{\mathbf{X}}_{k,l}$ will result in points on the $y = x$ line only when $k = L$. Else, there will be a deviation which can be measured either by comparing the sum of squares or by the angle between the fitted line of their regression and $y = x$. Thus, two measures of distortion for a specific variable l are

$$W_1 = SS_{k,l}^{-1} \cdot (\tilde{\mathbf{X}}_{k,l} - \hat{\tilde{\mathbf{X}}}_{k,l})^2 \cdot \mathbf{1}'$$

and

$$W_2 = |\text{Arctan}(1) - \text{Arctan}(\hat{b})|$$

where $\mathbf{1}$ is a vector of 1's with length n and $\hat{b}$ is the slope estimated from the aforemen-

rere	gropha	par	trast	work	resid	driving	walking
-6.96%	-32.94%	-61.61%	-4.11%	-81.37%	-27.9%	-73.05%	-41.04%

Table 3.7: Percent change in distortion through sum of squares when moving from one to two PC's.

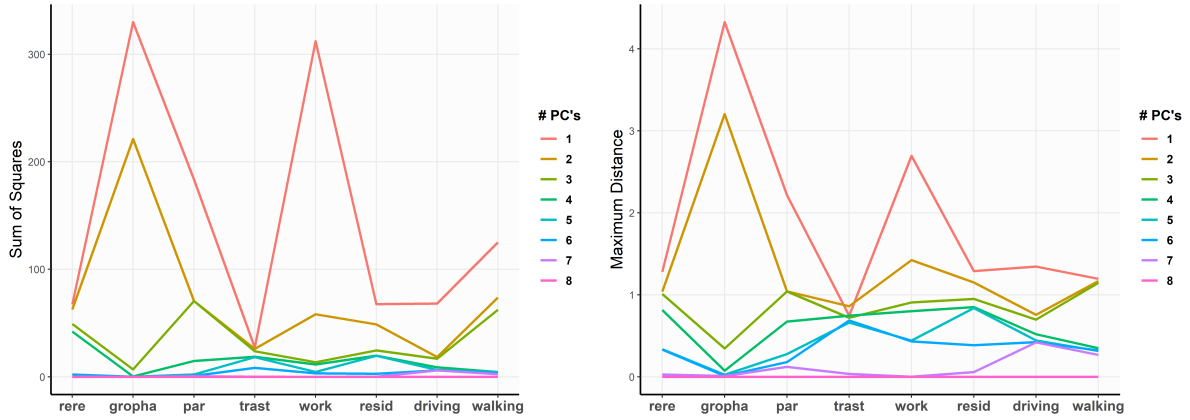

Figure 3.8: Distortion of each variable after dimension reduction. Left: through sum of squares. Right: through maximum distances.

tioned regression. Which of the above should be used is not only a matter of personal preference, but also of their sensitivity to the variables under study, in the sense that one measure may indicate a larger distortion relatively to the others.

Returning to the specific problem of mobility, we want to decide if we need one or two PC's, so we can check the first two lines from Figure 3.8. The left plot displays the distortion of each variable when discarding k out of 8 PC's through the sum of squares, while the right plot uses maximum distances. By adding the second PC, we gain a little more information for all the variables. Specifically, the percent change of the sum of squares for every variable when we move from one to two components is shown in Table 3.7 below with a total of 50.93% decrease.

Regarding the separation of the regression line from the straight line $y = x$, Figure 3.9 displays their difference in terms of the sum of squares ratio, while Figure 3.10 depicts their acute angle in a pie completion visualization. Small angles correspond to little deviation, thus more complete pie (deviation by 0 degrees is 0% and deviation by 90

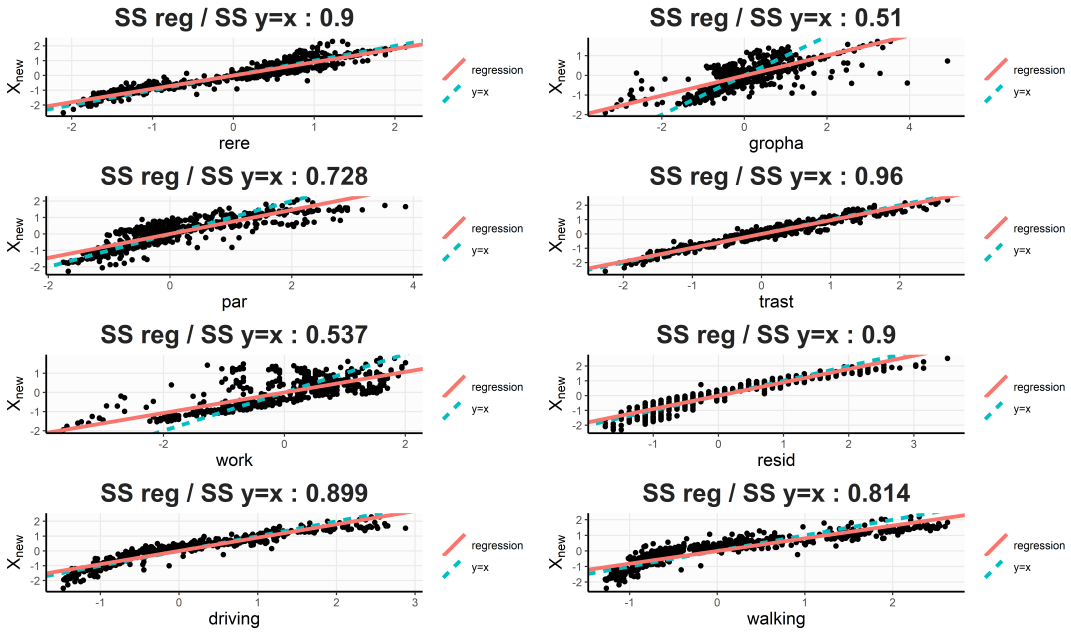

Figure 3.9: Distortion when moving from one to two PC's through comparison of the regression line to $y = x$ via the sum of squares ratio.

degrees is 100%). All of the above measures of distortion suggest that the "Grocery & pharmacy" variable is the one benefited more when adding a second PC.

On the whole, the mobility dataset can be summarized by the first principal component sufficiently and there is no exploratory reason to add a second PC. However, for predictive purposes we test our models using either one or two PC, since the second adds a little more explained variability, while it does not increase the dimensionality problem much.

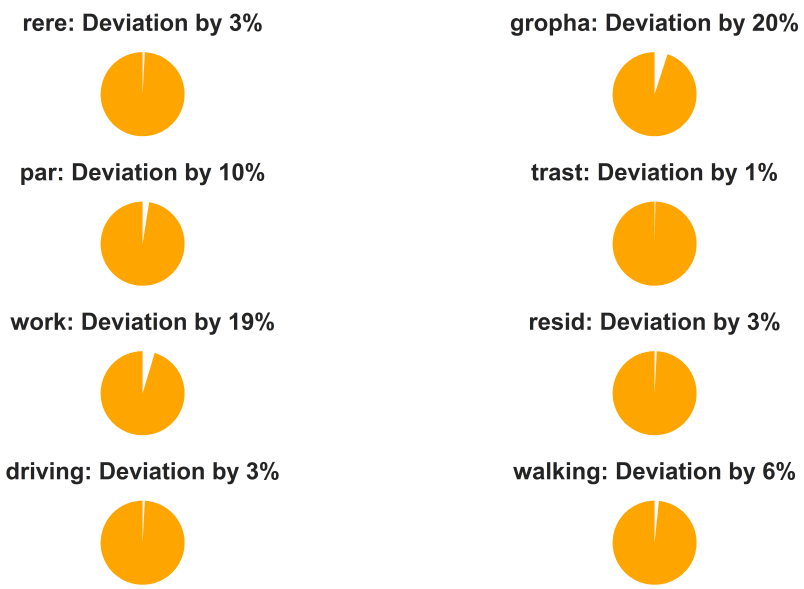

Figure 3.10: Distortion when moving from one to two PC's through comparison of the regression line to $y = x$ via angle difference.

3.6 Discussion

This paper suggests that the first two years of the Covid-19 epidemic may reasonably be described by SEIR-type models which include information on demographic births and deaths as well as vaccination. Our inference is based upon the estimated total cases and the external validity of the results is inspected through the comparison with an independent dataset from the REACT-2 study in UK. The full SEIR model with vaccination and demography is tested removing some of its structural components in turn and comparisons are made using information criteria, the marginal likelihood and Bayes factors.

The proposed framework is developed based on publicly available data which are central to this work. Estimation is based on the NUTS variant of HMC which appeared to be the most reliable compared to variational Bayes and maximization of the un-normalized posterior via Simulated Annealing. The predictive ability of the mobility data is examined within a dimension reduction framework and is shown to be relatively limited for accurate predictions and therefore for informing public health decisions.

A potential weakness of the proposed model is the reliance on an IFR estimate which affects some of the model parameters. This is due to our focus upon publicly available data (in contrast with approaches like Birrell et al., 2021[10] which use continuous time models informed by additional data sources which may or may not be publicly available) but our framework can accommodate additional evidence, like seroprevalence surveys (see Nieminen et al., 2023[76]), in a straightforward manner. Such additional data would give direct evidence on the IFR but they are not typically available for many countries and therefore are beyond the scope of this work. The computational burden for training our models poses limits to certain short-term prediction exercises. However, one-week-ahead predictions are certainly feasible and often the most realistic if one wishes to avoid strong assumptions on the population behaviour.

Chapter 4

Non-linear Dynamics of Communicable Diseases

Since Newton, mankind has come to realize that the laws of physics are always expressed in the language of differential equations.
- Steven Strogatz

In this Chapter, we investigate the Covid-19 pandemic as a dynamic non-linear system of equations and give geometric intuitions of it as a vector field in the $\mathbb{R}^2_+$ space. We derive some results about the seriousness of a disease and the effectiveness of interventions. We introduce new epidemic terminology and formalize definitions used in the area, while we also include parallelisms to physical quantities. We provide examples on simulated scenarios to build intuition and then we provide applications to models based on Covid-19 data of Greece. In Section 4.1 we refer to functions $f : \mathbb{R}_+ \to \mathbb{R}_+$, i.e. functions of time in the continuous domain, so we use the notation $f(t)$. In Section 4.2 we refer to functions $f : \mathbb{N} \to \mathbb{R}_+$, i.e. functions of time in the discrete domain, so we use the notation f_t. All the analyses are performed in a way that promotes the intuition and interpretation, rather than mathematical details and abstract results, but the unfamiliar with dynamics reader can refer to the book of Strogatz (2018)[96] and refresh their multivariate calculus with the book of Marsden and Tromba (2012)[69] which focuses on building mathematical

intuition.

4.1 The continuous deterministic case

To begin with, let $t \in \mathbb{R}_+$ be a variable that measures time since the beginning of an epidemic and $S : \mathbb{R}_+ \to \mathbb{R}_+$ and $I : \mathbb{R}_+ \to \mathbb{R}_+$ be two functions of t that measure the number susceptible and infected individuals respectively at each time t in a fixed population N. The $S(t)$ and $I(t)$ (which we write as S and I sometimes for simplicity) are the solutions of a system of ordinary differential equations (ODE), like (3.1) which we saw in Section 1.2. Actually, the third equation regarding the removed individuals plays no role in the dynamics of S and I (R does not appear in any of the other two equations), thus we omit it in this Chapter. The goal is to analyse the epidemic considering the dynamic change of S and I as time passes.

4.1.1 Definitions and terminology

Let us consider a set of ODE of the form

$$\begin{aligned} \dot{S} &= f_1(S, I) \\ \dot{I} &= f_2(S, I) \end{aligned} \tag{4.1}$$

i.e. a system of $S(t)$ and $I(t)$ involving their derivatives, where the dot symbol indicates differentiation with respect to t. This is the form of (3.1), where $f_1(S, I) = -\lambda SI/N$ and $f_2(S, I) = \lambda SI/N - \gamma I$ (in this Chapter we use the recovery rate γ in place of the infectious period τ).

Let us explain first that thinking of the susceptible and infectious population one at a time is wrong. We give a simpler example of a system to make clear what our goal in this Chapter is, as well as a little reminder of dynamics basics. Suppose we deal with the first-order system

$$\dot{x} = \cos(x)$$

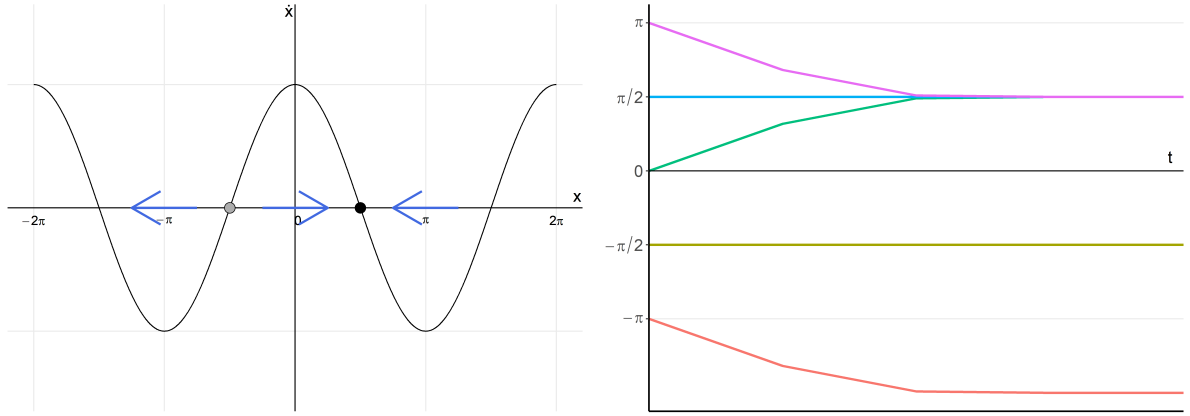

Figure 4.1: Left: The vector field of the system $\dot{x} = \cos(x)$ on the real line. At each value of x, the velocity $\dot{x}$ on the y-axis indicates the direction of motion, which takes place on the x-axis. The grey dot is an unstable fixed point, while the black one is a stable fixed point. Right: The x values as functions of time t with initial positions $-\pi$, $-\pi/2$, 0, $\pi/2$ and π.

where x is a function of time t and we are interested in its dynamic behaviour. Specifically, we would like to explain what will happen to $x(t)$ as $t \to \infty$. So, if we initially place an imaginary particle on $x(0)$, is there a specific path it will follow as it keeps moving on the real line? Will it rest somewhere? What is its qualitative behaviour based on the vector field that $\dot{x}$ dictates?

We can examine this vector field on the real line, by plotting x against $\dot{x}$, so that $\dot{x} > 0$ indicates rightward motion, $\dot{x} < 0$ indicates leftward motion and $\dot{x} = 0$ indicates a *fixed point* (i.e. a position where the particle rests), which is either a *stable fixed point*, where the particle is attracted to, or an *unstable fixed point*, where the particle is repelled away from. In Figure 4.1 we focus on the interval $(-3\pi/2, 3\pi/2)$ and in the left plot (such a plot where we can visualize the trajectories of a dynamical system is called a *phase portrait*) we can see that there exist both of these two types of points indicated with black and grey dots. For instance, suppose the initial position of the particle is $x(0) = 5\pi/4$. At first, it will accelerate until $x = \pi$, then it will decelerate until it reaches $\pi/2$, where it will be absorbed. In the right hand side of Figure 4.1 we have plotted the solutions $x(t)$ for initial positions at $-\pi$, $-\pi/2$, 0, $\pi/2$ and π.

Coming back to the epidemic example, one could say that we should consider the equations one at a time following the last procedure. We can study the dynamics on the real line by plotting S against $\dot{S}$ and observe the flow of an imaginary particle moving on this line (the x-axis) with velocity given by the corresponding values of $\dot{S}$ at each point. Thus, $\dot{S} > 0$ indicates rightward motion, $\dot{S} < 0$ indicates leftward motion and $\dot{S} = 0$ indicates a fixed point. The same reasoning holds for I against $\dot{I}$ and the whole dynamics discussion seems to have come to an end, since these tools visualize the motion of the imaginary particle (which represents the state of the epidemic) and we can answer questions like "If N_1 individuals are initially infectious, how many are they going to be after a long time?".

However, we do not really have two first-order systems, but one second-order system, that is the two variables that we are interested in are coupled together. Moreover, f_1 and f_2 are non-linear functions of S and I, so we cannot even have a closed form solution. Thus, the analysis above cannot be done to construct independent graphs $\{(S, \dot{S}) : S \in [0, N]\}$ and $\{(I, \dot{I}) : I \in [0, N]\}$ and obtain a general picture from them. We would like to have a singular plot that summarizes our thoughts.

The epidemic models stem from a system of equations (see equation (4.1)), not individual ones. Thus, we can only examine the course of the disease using both of these two pieces of information simultaneously. To this end we shall move to a representation on the bivariate space $S \times I$ and observe its dynamic behaviour on this so-called phase space (see Nolte 2010[77]), which we denote as a set $V \subseteq \mathbb{R}^2_+$. In our case the phase space is a plane, which we will refer to as the SI-plane. For all the following analysis we put the S variable on the x-axis and the I variable on the y-axis.

The dynamical system of non-linear differential equations like (4.1) can be examined as a vector field $\mathbf{F}$ on V. We define the *natural epidemic flow* to be a curve $\boldsymbol{\sigma}(t) = \big(S(t), I(t)\big)$ that satisfies

$$\mathbf{F}\big(\boldsymbol{\sigma}(\mathbf{t})\big) = \frac{d}{dt}\boldsymbol{\sigma}(t) = \frac{dS}{dt} \cdot \hat{\mathbf{i}} + \frac{dI}{dt} \cdot \hat{\mathbf{j}} \tag{4.2}$$

i.e. its velocity is dictated by $\mathbf{F} : V \to V$. Given an initial condition $\boldsymbol{\sigma}(0)$ the epidemic flow defines a trajectory on V, which we call the *natural course of the epidemic*. The flow

of a vector field can be met in many areas of physics too, e.g. the magnetic flow. We can think of the SI-plane like water that fluids in the direction the vector field defines, so if we throw a particle in there, it will move on the specific trajectory the vectors point to. In our case, the way the water fluids is the natural epidemic flow, the particle we throw is the current state of the epidemic at every time point, the position we first throw it is the initial state $\big(S(0), I(0)\big)$ where the epidemic begins, while its trajectory is the natural course of the epidemic.

The image of the realized C^1 transformation of time t to the $\mathbb{R}^2_+$ plane $\boldsymbol{\sigma} : \mathbb{R}_+ \to \mathbb{R}^2_+$ (that is the realized trajectory) is a useful plane curve, since it defines a path whose x-coordinate represents the number of susceptible individuals and whose y-coordinate represents the number of infected individuals at each time t. Thus, monitoring the road our particle follows starting from the position we threw it in the epidemic flow (the initial value $\big(S(0), I(0)\big)$) can be very informative.

The speed of the particle, which is $v(t) = ||\mathbf{v}(t)|| = \left|\left|\frac{d\boldsymbol{\sigma}(t)}{dt}\right|\right|$ (i.e. the magnitude of the velocity vector $\mathbf{v}(t)$) translates to a dangerous period when it takes on high values and vice versa. This is true, because it equals

$$v(t) = \sqrt{\left(\frac{dS}{dt}\right)^2 + \left(\frac{dI}{dt}\right)^2} \tag{4.3}$$

so when susceptible people drop quickly, the infectious rise and $v(t)$ is large (thanks to the squares of the rates). Also, when susceptible do not drop quickly, but infectious do drop, then again $v(t)$ is large. So, we can say that high speed is a characteristic we observe during an epidemic wave. Then it is natural to study the rate of change of speed, i.e. the acceleration $a(t) = \sqrt{\left(\frac{d^2S}{dt^2}\right)^2 + \left(\frac{d^2I}{dt^2}\right)^2}$. Note that, in endemic situations there exist "inverted" waves where susceptible individuals rise, so speed should be carefully examined.

Another definition we introduce is that of the *epidemic work*, which is analogous to the work in physics. Thus, we define the work of the epidemic during the time interval $[a, b]$ to be the line integral of the vector field $\mathbf{F}$ along the natural course of the epidemic

between times a and b, i.e.

$$W_{a,b} = \int_{\boldsymbol{\sigma}} \mathbf{F} ds = \int_a^b \mathbf{F}(\boldsymbol{\sigma}(t)) \cdot \boldsymbol{\sigma}'(t) dt = \int_a^b \left(\frac{dS}{dt}\right)^2 + \left(\frac{dI}{dt}\right)^2 dt \tag{4.4}$$

If this was a physics problem, we would say that the work is the energy transferred to the particle when it moves on the SI-plane, but now the word energy does not actually have a physical meaning. However, it still summarizes the seriousness of the course of the epidemic: a large number indicates bad news, while a smaller one indicates a better course.

Something else that could be of interest is the length of the epidemic course inside a time interval. However, most of the times this information is only partially exploited, since the reported numbers only regard length on the x-direction even though we deal with a two-dimensional system. The length of the course between times a and b is equal to

$$l_{a,b} = \int_a^b \left|\left|\frac{d\boldsymbol{\sigma}(t)}{dt}\right|\right| dt \tag{4.5}$$

while we can calculate only the x- or y-direction contribution by only considering the x or y coordinates of the curve $\boldsymbol{\sigma}$ respectively.

The x-direction is the one used widely, since it expresses the number of infections between times a and b,

$$l_{a,b}^{(x)} = \int_a^b \left|\frac{dS}{dt}\right| dt \tag{4.6}$$

The larger the x-direction length, the more individuals "escape" the S-state, the more cases occur. Note that we have a problem if we let the model assume that the susceptible population can also grow inside (a, b) except of being reduced, or if individuals can move to another state without first being infectious. Then, this length would not correspond to the cases occurred between a and b. If not, the new total cases are just $S(a) - S(b)$ (when S only decreases), or $S(b) - S(a)$ (when S only increases). Also, note that, if we allow for immediate removals using a model with vaccination and demography (like in 3.2), then the difference of the S-state at times a and b could only return the cases occurred during that interval if we also add the terms that account for physical deaths and immunity due to vaccination and subtract the births occurred.

If $\frac{dS}{dt}$ does not change sign inside the time interval $(t'-\tau+1, t')$, where t' is the current time, then the length $l^{(x)}_{t'-\tau+1,t'}$ is the number of individuals in the active set, which shows the infections the system has to deal with at time t'.

The y-direction between times a and b, which is not examined by practitioners is given by

$$l^{(y)}_{a,b} = \int_a^b \Big|\frac{dI}{dt}\Big| dt \tag{4.7}$$

If $\frac{dI}{dt}$ does not change sign, then the y-direction length is just $I(a) - I(b)$ (when I only decreases), or $I(b) - I(a)$ (when I only increases). This length describes how big the difference is between the number of infected individuals at the end and at the beginning of the examined period. If for example the length is given by $I(b) - I(a)$, a large and positive number means that the active set is getting larger, thus the health system is under pressure. A large and negative number means that the active set is being shrunk and the health system is being relieved (actually then we should have calculated $I(a) - I(b)$). Again, like in the S case, when I is not monotonic inside (a, b), then we lose that interpretation and any practical use of $l^{(y)}_{a,b}$.

Until now, we are based on the deterministic nature of a system of ODE to derive the natural epidemic course. However, the actual behaviour of the epidemic is far from deterministic and it is never left to act on the population without interventions. Thus, we do not experience the natural course suggested by the model, but we follow the actual course of the epidemic. The actual course of the epidemic is a curve $\boldsymbol{\gamma}(t)$ on the SI-plane whose tangent vector at some points is different from the vector field at those points (in contrast with the natural course).

Thus, we are led to one more measure that could be helpful based on the area between the positions under the actual and the natural course during the time interval (a, b). We define

$$L_{a,b} = \int_a^b \frac{||\boldsymbol{\sigma}(t) - \boldsymbol{\gamma}(t)||}{||\boldsymbol{\sigma}(t)||} dt \tag{4.8}$$

where $\boldsymbol{\sigma}$ is the natural and $\boldsymbol{\gamma}$ is the actual course, to be a measure of intervention effectiveness. The intuition is that when we intervene on a bad situation using restrictive measures, the more we change the natural course that we would follow into a less danger-

ous actual course, the better the measures. Furthermore, we would like that to happen as quickly as possible, for restrictive measures impose many problems related to the economy, psychology and more. Thus, we can calculate the Euclidean distance between positions that would occur had we followed the natural course and the corresponding positions that occur following the actual course. Then, we normalize dividing with the L_2-norm of the natural course. Large values of $L_{a,b}$ suggest large deviation of the course under the intervention from the initial course. Practically, since the natural course is not known but can be simulated assuming an underlying model, $L_{a,b}$ can also be used to assess the intervention effectiveness between measures A and B. Intervention A will be better than intervention B if $L_{a,b}^{(A)} > L_{a,b}^{(B)}$.

Another way we can calculate the intervention effectiveness is using the epidemic work between the natural and actual course during the time interval (a, b), i.e.

$$M_{a,b} = \frac{\int_{\boldsymbol{\sigma}} \mathbf{F} ds - \int_{\boldsymbol{\gamma}} \mathbf{F} dg}{\int_{\boldsymbol{\sigma}} \mathbf{F} ds} \tag{4.9}$$

where once again the denominator exists for normalization purposes. Lower speed (thus lower work) is indicative of a better course, thus the larger the difference between the actual course and the one suggested by the intervention, the better. Thus, once again $M_{a,b}^{(A)} > M_{a,b}^{(B)}$ means intervention A is to be preferred over intervention B.

4.1.2 Phase space

Let us now examine the behaviour of a system describing an epidemic using the SI plane V, a process called phase analysis, since it interprets the phases the system goes through as $t \to \infty$. One can extract useful insights for a system of ODE by plotting the phase space (the phase plane in our case) and searching for areas where the imaginary particle (the *phase point*) has some special behaviour.

SIR model

We begin with the SIR model we have already seen in Section 1.2, only this time we describe its dynamics through the lens we have introduced in the current Chapter. The system of ODE for the susceptible and infectious individuals according to the simple SIR model is

$$\begin{aligned} \dot{S} &= -\lambda SI \\ \dot{I} &= \lambda SI - \gamma I \end{aligned} \tag{4.10}$$

where λ is the infection rate and γ is the recovery rate. By "infection" we mean that an individual ends their susceptible to the epidemic property and enters a state that carries and spreads the disease, while by "recovery" we mean that they exit that state (either by actual recovery or by death). This model assumes that "infected" has the same meaning as "infectious", since there is no exposed period. Thus, we call λ the infection rate, because it expresses how quickly SI individuals lose their "susceptible" property and we call γ the recovery rate, because it expresses how quickly I individuals lose their "infected" property (details about compartmental epidemic models and basic terminology can be found in Chapter 1.2). We also remind the reader that the reciprocal of γ is the infectious period τ, which is the time interval during which an individual remains infectious.

Note that we have used a slightly different formulation than that in previous Chapters, since we do not divide λ by N. We do this because we can assume that $N = 1$ and S and I measure the proportion of susceptible and infected individuals respectively. In this Chapter, we make this assumption in order to simplify the equations. Using this formulation, V is a subset of the unit cube, i.e. $V \subseteq [0,1]^2$, rather than the whole $\mathbb{R}^2_+$. If someone wants to use the previous formulation, they just have to substitute λ/N in place of λ and let N be the actual population size.

If we throw the imaginary particle in the flow of the SI-plane, then areas of no movement are of particular interest. To find those areas, we set the two derivative equations for S and I equal to zero and obtain

$$\begin{cases} \dot{S} = 0 \\ \dot{I} = 0 \end{cases} \iff \begin{cases} \lambda SI = 0 \\ \lambda SI = \gamma I \end{cases} \iff \begin{cases} S = 0 & \text{or} \quad I = 0 \\ \begin{cases} S = \dfrac{\gamma}{\lambda} \\ I \neq 0 \end{cases} & \text{or} \quad I = 0 \end{cases}$$

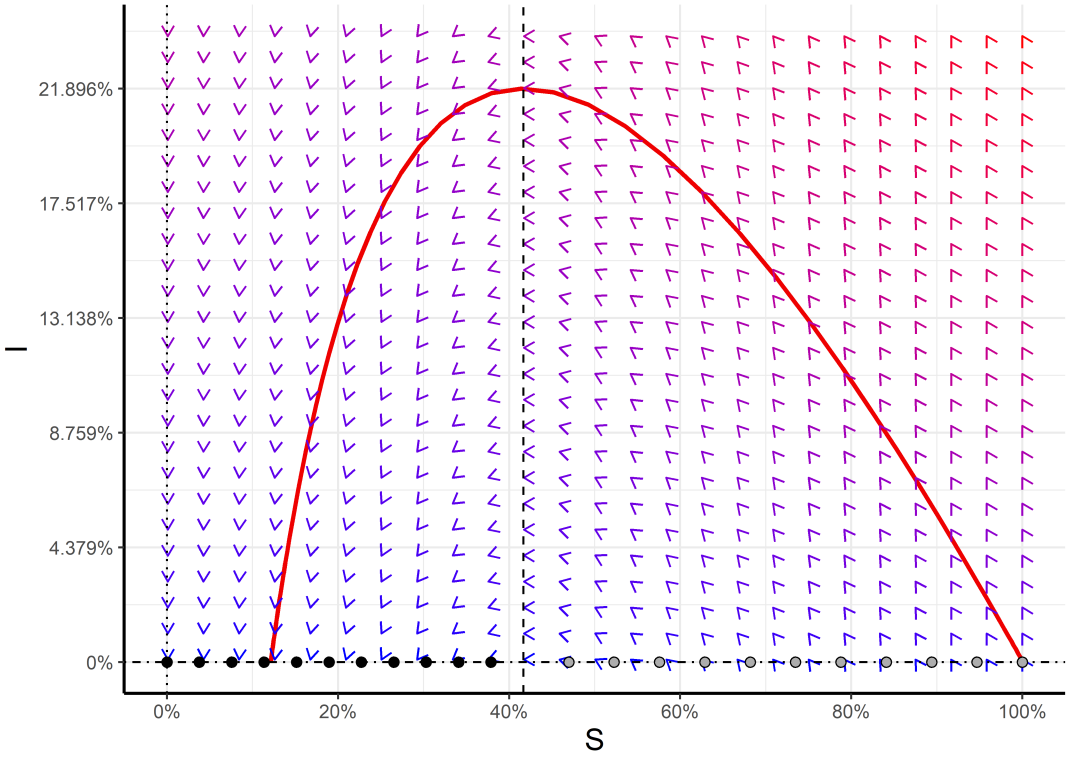

Figure 4.2: The SI phase plane of the SIR system (4.10). The vector field is shown with the color of the arrows indicating the magnitude of each vector: the more red, the larger. The vertical dashed line corresponds to $S = \gamma/\lambda$, the vertical dotted line corresponds to $S = 0$, while the horizontal dashed-dotted line corresponds to $I = 0$. The red curve corresponds to the path the phase point follows at the simulation conducted. The black and grey dots represent stable and unstable fixed points respectively.

The first equation above describes when there is no horizontal movement on the plane and the second when there is no vertical movement. Thus, on the x- and y-axis (where $I = 0$ and $S = 0$ respectively) there is no horizontal movement, so when the particle passes from there it will only move up or down the plane. This is intuitive, because zero susceptible individuals means the whole population has already been infected and

so, the people left in the I-state are being reduced continuously. Also, if the infectious individuals are set to zero, the epidemic reaches to an end, because there is no one else to transmit the disease. Regarding the second equation, there will be no vertical movement on the straight line $S = \frac{\gamma}{\lambda}$ neither on the x-axis. These three lines are called *nullclines* or *isoclines*; they are the curves where the flow is parallel to an axis and they are important because they form regions where the system has some special behaviour.

Moreover, the first equation $\dot{S} = -\lambda SI$ of the system (4.10) states that the derivative of S is always negative, since $\lambda, S, I \geq 0$ (with equality only when the epidemic is ended, when $S = 0$ or $I = 0$). This means that we only have movement from right to left, meaning that susceptible people only decrease. On the other hand, for $I \neq 0$ we have

$$\dot{I} < 0 \Leftrightarrow S < \frac{\gamma}{\lambda} \tag{4.11}$$

Thus, on the left of the vertical straight line $S = \frac{\gamma}{\lambda}$, we have downward movement (so infectious individuals decrease), in contrast with the right side of this line where we have upward movement (so the infectious individuals increase). Finally, in order for the particle to settle on a fixed point, both derivatives need to be zero and the calculations above show that there exists a whole line (the x-axis) of fixed points rather than an isolated one, specifically the line $I = 0$.

To summarize, in Figure 4.2 we have plotted the vector field that corresponds to the SIR system for $\lambda = 0.4$ and $\gamma = 1/6$, so that the reproduction number is $R_0 = 2.4$ (a plausible value for that of Covid-19 at the beginning of the pandemic). We integrated the system numerically for time steps of size 1 (they are sufficiently small to approximate the trajectory) starting with 99.9% susceptible and 0.1% infectious population until $I < 10^{-6}$ (which took 144 steps). The conclusions we derived through math are all confirmed by the directions of the arrows and the Figure provides a good visualization for the dynamics of the SIR model.

Figure 4.2 is very useful for gaining insights that are not easy to reach analytically (although in reality without a rigorous proof, they remain simple hypotheses). For instance, if a fixed point lies to the left of $S = \gamma/\lambda$, we can recognize that if a particle starts near it, it will not reach to it necessarily, meaning that the fixed point is not *attracting*.

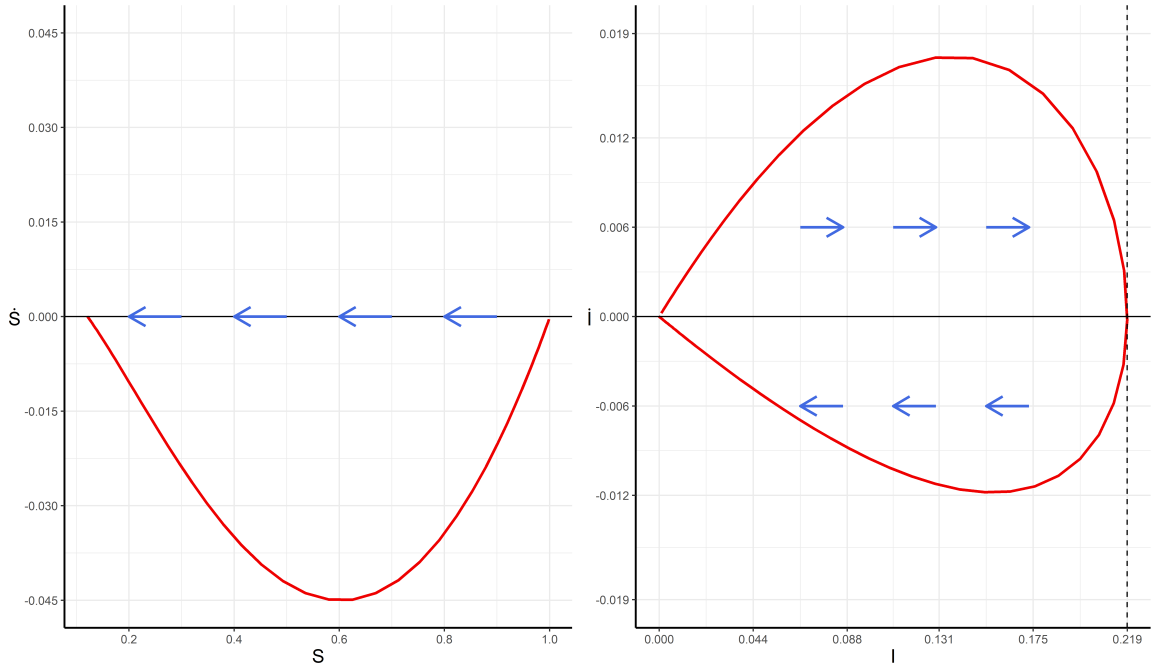

Figure 4.3: Decomposition of the SI phase plane into two parts regarding susceptible and infectious individuals. Blue arrows indicate the direction of motion, which takes place on the x-axis. The velocity at each point is given by the y-coordinate of the red curve, where a negative value means movement to the left. Left: The values $\{(S, \dot{S}) : S \in [0, 1])\}$. Right: The values $\{(I, \dot{I}) : I \in [0, 1]\}$. Note that the x-axis ticks are added below each plot for clarity, instead of the actual x-axis in the middle of the plots.

Moreover, trajectories starting close to a fixed point remain close to it meaning that they are *Liapunov stable*. Thus, the fixed points on the left of $S = \gamma/\lambda$ are *neutrally stable*. On the other hand, when a point starts on the x-axis on the right of $S = \gamma/\lambda$, then a perturbation $(S(0), I(0)) + (0, i)$ for small and positive i would initiate the epidemic (a perturbation purely on the x-axis does not initiate the epidemic). Thus, the fixed points on the right of $S = \gamma/\lambda$ are unstable. The same results can be confirmed by linearization of the system around those points (see Appendix A5).

Regarding our previous simulation, we can also break down the phase plane into two parts and re-confirm our results based on the sets $\{(S, \dot{S}) : S \in [0, 1])\}$ and $\{(I, \dot{I}) : I \in [0, 1]\}$. So in the left plot of Figure 4.3 we see that $\dot{S} < 0$, thus the particle (which is placed at the rightmost part on the x-axis) only moves to the left with velocity given by

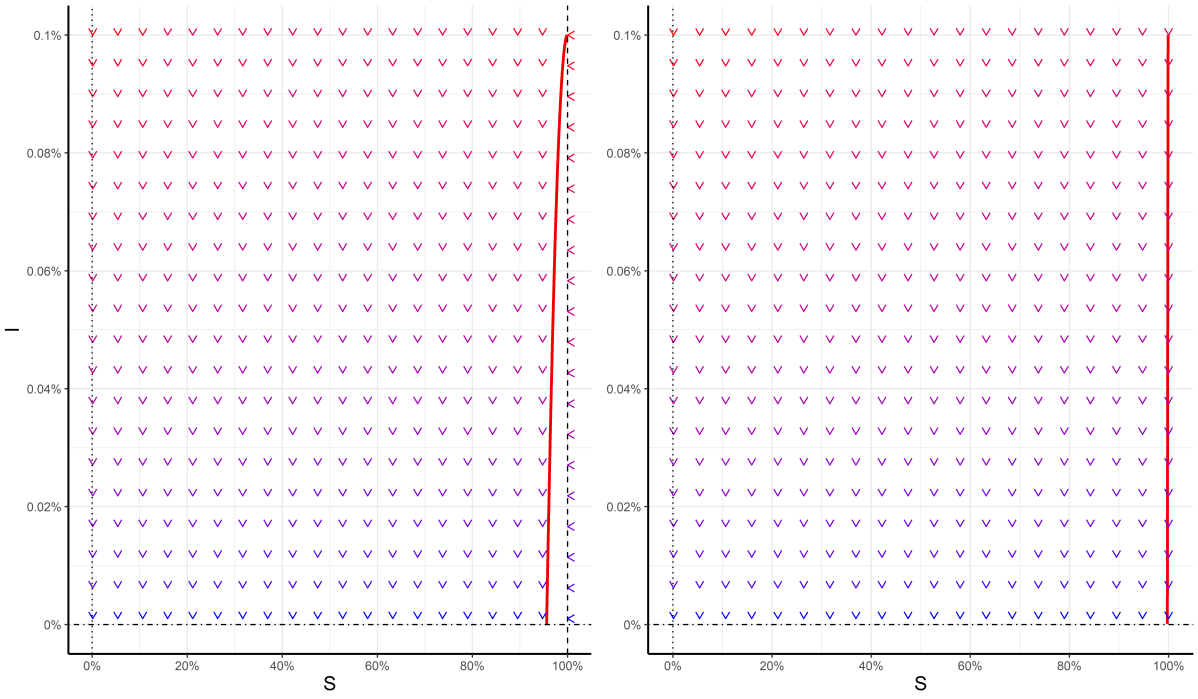

Figure 4.4: The SI phase plane of the SIR system for the case when $R_0 = 1$ (left) and $R_0 < 1$ (right). It seems that the phase point moves straight down in the second case, but this is just because of the scale of the plot; it actually moves a little to the left (as it should according to the analytical results). Also, the line $S = \gamma/\lambda$ is not present in the right plot, since $R_0 = \lambda/\gamma < 1$.

the height of the graph at each time point. In the right plot of the same Figure, we see that the particle starts at 0, moves to the right until $S = \gamma/\lambda$ (indicated as a vertical dashed line) and then it returns back to 0. This plot could have never been arisen by a first-order system of course (because of the change in direction), but it underlines once again the importance of treating the system as a second order one (where a change in direction is allowed).

The practical interpretation of the above is that if the Covid-19 epidemic can be described sufficiently well by an SIR model, then the vertical straight line $S = \gamma/\lambda$ is of big importance, because if it is close to $S = 0$, we will experience many infections until the epidemic starts to settle down to $I = 0$. Note that we can also write this line as

$S = 1/R_0$, since the reproduction number is $R_0 = \lambda/\gamma$. Therefore, a large R_0 creates a small γ/λ fraction (i.e. a vertical line close to $S = 0$), while a small R_0 creates a large γ/λ fraction (i.e. a vertical line close to $S = 1$). This is why keeping a low R_0 is critical during an epidemic: as R_0 gets smaller it moves the vertical line closer to $S = 1$ and then beyond that. The two scenarios are also displayed in Figure 4.4, where R_0 is getting smaller from the left to the right plot.

The whole analysis above assumes that $R_0 = \dfrac{\lambda}{\gamma} > 1$. The case when this does not happen is of no statistical interest since there will not be a major epidemic and probably it will not be observed at all. In the left-hand side of Figure 4.4 we show the same scenario as before, but we set $\lambda = \gamma = 1/6$ (so $R_0 = 1$), while in the right-hand side it holds that $\lambda = 0.1$ and $\gamma = 1/6$ (so $R_0 = 0.6$). In both scenarios we see that every initial position for the particle leads it directly to the interval $\mathcal{I} = \{(S, I) \in [0, 1]^2 : 0 \leq S \leq \dfrac{\gamma}{\lambda}, I = 0\}$, which is now the whole $(0, 1)$ interval of the x-axis. Also, the integration time needed to finally reach to it is different than before: it needed 1122 steps for the $R_0 = 1$ case and 105 for the $R_0 < 1$ case. Finally, observe that the arrows of the vector field on the plane are more red close to the top of each plot and more blue at the bottom confirming that large speed corresponds to dangerous period.

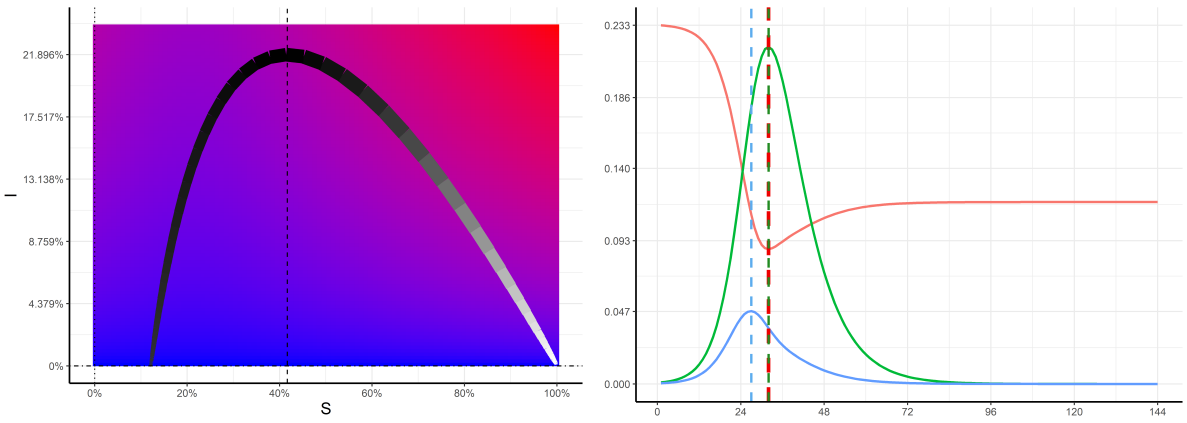

Figure 4.5: Left: Instead of the vector field, we use color to represent the speed (red=high, blue=low). The thickness of the curve is proportional to the speed. Also, darker points represent lower acceleration. Right: Series of $v(t)$ (in blue), $a(t)$ (in red) and $I(t)$ (in green). The vertical dashed lines indicate a maximum or minimum.

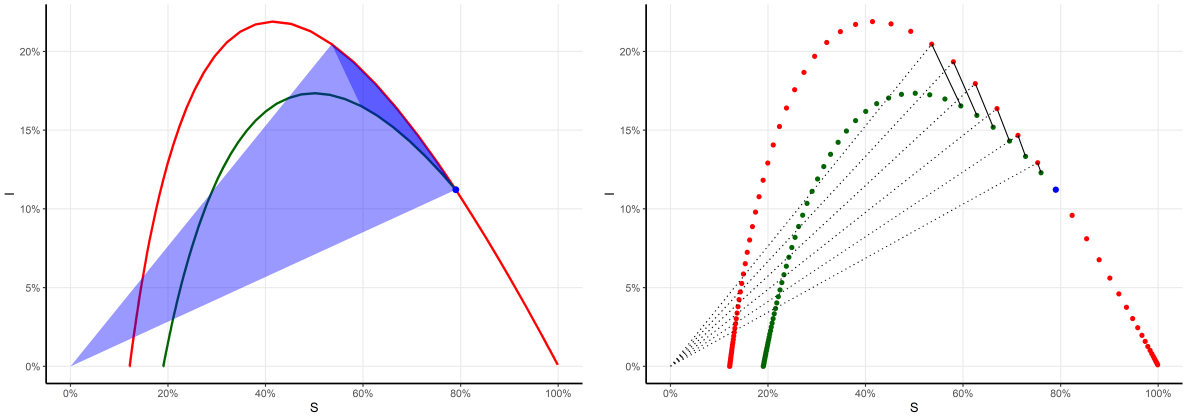

Figure 4.6: Left: Intervention effectiveness for the simulated scenario as expressed by (4.8). Right: Approximation of $L_{23,29}$ by (4.26).

Regarding the speed and acceleration for the SIR system, we have that

$$\mathbf{v}(t) = (\dot{S}, \dot{I}) = (-\lambda SI, \lambda SI - \gamma I) \tag{4.12}$$

and

$$\mathbf{a}(t) = (\ddot{S}, \ddot{I}) = (-\lambda I, \lambda S - \gamma) \tag{4.13}$$

so in the left plot of Figure 4.5 we show the vector field for the previous $R_0 > 1$ case along with the acceleration at each point of the simulation, while in the right plot of the same Figure we show the series of the speed, acceleration and the infectious individuals $I(t)$ during the epidemic course. Acceleration reaches its minimum when $I(t)$ is maximized, while the speed reaches its maximum before $I(t)$ is maximized.

Lastly, let us demonstrate the intervention effectiveness tools discussed for the general case. To this end, let us assume that the SIR model is the one used for a particular case with $\lambda_1 = 0.4$ and $\gamma_1 = 1/6$ and let $\boldsymbol{\sigma}_1(t)$ be the epidemic course we follow according to that model. After 23 days, an intervention is decided which leads to $\lambda_2 = 1/3$ and $\gamma_2 = 1/6$. So, we assume that the infectious period $\tau_1 = 1/\gamma_1 = 1/\gamma_2 = 6$ stays the same, but the new measure imposed lowers the infectious rate. The first set of parameters are those we have used before to demonstrate the phase plane and the second value of λ is chosen only for demonstration purposes.

We want to assess the effectiveness of the new measure for a time period equal to

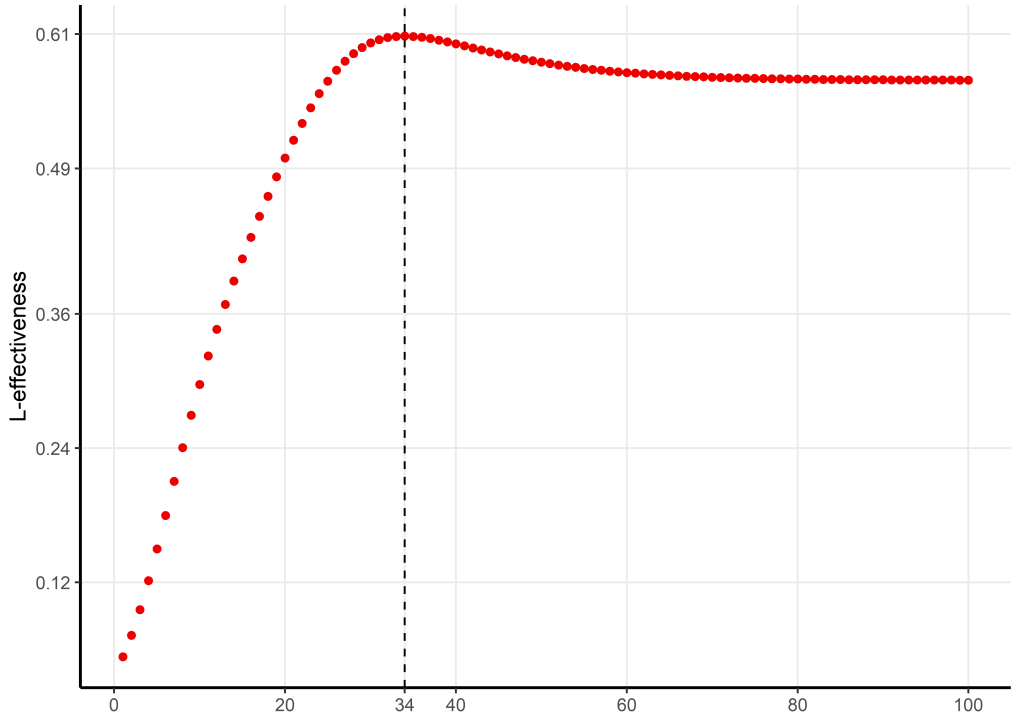

Figure 4.7: The first 100 summands taken from (4.26).

one infectious period, i.e. from day 23 up to day 29. After having quantifying the gain, one can compare it with the loss the intervention measures are connected to and decide if they are worth it. For instance, a lock-down is followed by serious socio-economic problems, so a small gain can be argued that is not worth it. One problem is how to define a "small" or "large" gain and the other is to use it in a decision-making process. The variables discussed so far regard the former problem. Moreover, different measures can be compared, since the best intervention will lead to a larger value of the L or M index.

In the left hand side of Figure 4.6, the red line is the $\boldsymbol{\sigma}_1(t)$ curve. After 23 days, indicated with the green point, the new course $\boldsymbol{\sigma}_2(t)$ begins. The L variable in (4.8) for days 23 up to 29 quantifies the effectiveness to be $L_{23,29} = 0.6733$. This can be visualized as the area between the two curves (in blue) divided by the area between $\boldsymbol{\sigma}_1(t)$ and the point $(0, 0)$ (in light and dark blue) for $t \in (23, 29)$. In order to use the M variable (4.9), we need the work for the two courses. For each curve we calculate the work and substitute in (4.9), so we get $M_{23,29} = 0.4635$.

Actually, the two intervention effectiveness measures we have calculated are not computable by (4.8) and (4.9), since there are no closed form solutions for the system of ODE to obtain the S and I functions. Instead, we use the discrete-time approximations given later in the text in (4.26) and (4.27) and visualize their components in the right plot of Figure 4.6. Lastly, the specific summands from (4.26) are plotted in Figure 4.7 for the first 100 days after the new measure has been taken. Until day 34, there is an increase in gain of the measure while, after that point, the daily gain is constant. Using an optimization scheme one can compare the utility of the measure and the loss induced by it.

SIR with demography model

It is obvious from the above that according to the SIR model (with $R_0 > 1$), it is only a matter of time until the epidemic dies out, either due to lack of susceptible or lack of infectious individuals. However, there are also cases that a disease does not cease to exist, but it becomes endemic. This can occur by an epidemic that is described by an SIR which also allows for physical births and deaths, which we refer to as SIR with demography. Such a model is sensible when the epidemic persists for a long time.

The system of ODE for this specific model is given in (1.7), but let us rephrase it using the $N = 1$ formulation and using the recovery rate, instead of the infectious period:

$$\begin{aligned} \dot{S} &= -\lambda SI + A(1 - S) \\ \dot{I} &= \lambda SI - \gamma I - AI \end{aligned} \tag{4.14}$$

We now examine the behaviour of the system (4.14) on the phase plane by first finding the isoclines that define areas of no movement.

$$\begin{cases} \dot{S} = 0 \\ \dot{I} = 0 \end{cases} \iff \begin{cases} \lambda SI = A(1 - S) \\ \lambda SI = (\gamma + A)I \end{cases} \iff \begin{cases} \begin{cases} I = \dfrac{A(1 - S)}{\lambda S} \\ S \neq 0 \end{cases} \\ \begin{cases} S = \dfrac{\gamma + A}{\lambda} \\ I \neq 0 \end{cases} \end{cases} \quad \text{or} \quad I = 0$$

Now for the isoclines, we see that no vertical motion exists on the x-axis, as well as

on the vertical straight line $S = (\gamma + A)/\lambda$. The reproduction number in this type of models is given by $R_0 = \lambda/(\gamma + A)$, since the current "recovery" can also be imposed by death (not due to the disease). Thus, the last isocline can be written as $S = 1/R_0$ (just as in the SIR case). To find an equilibrium, we need both of the two derivatives to be zero simultaneously, so substituting $S = (\gamma + A)/\lambda$ into $I = \dfrac{A(1-S)}{\lambda S}$, we obtain

$$I = \frac{A}{\gamma + A}\left(1 - \frac{\gamma + A}{\lambda}\right)$$

Together these two equations for S and I give the equilibrium endemic point

$$(S^*, I^*) = \left(\frac{\gamma + A}{\lambda}, \frac{A}{\gamma + A}\left(1 - \frac{\gamma + A}{\lambda}\right)\right)$$

If we let $a = \dfrac{\gamma + A}{A}$ and $R_0 = \dfrac{\lambda}{\gamma + A}$, then $S^* = 1/R_0$ and $I^* = \dfrac{1}{a}\left(1 - \dfrac{1}{R_0}\right) = \dfrac{1}{a}\dfrac{R_0 - 1}{R_0} = \dfrac{R_0 - 1}{aR_0}$, so

$$(S^*, I^*) = \left(\frac{1}{R_0}, \frac{R_0 - 1}{aR_0}\right)$$

The endemic equilibrium and the disease free equilibrium $(S, I) = (1, 0)$ are the only two fixed points in contrast with the previous model. We can see that the larger a, the fewer infectious individuals in the endemic state. By definition, a equals the rate of death incurred by both infections and physical deaths divided by the rate of physical deaths. Note also that the quantity a is the average lifetime $1/A$ over the average duration of infection $1/(\gamma + A)$, which means that the longer the lifetime is compared with the infection duration, the fewer endemic infections.

Finally, we can examine the type of these fixed points by linearization of the system (see Appendix A5) and conclude that, in the scientifically interesting case when $R_0 > 1$, the disease free fixed point is unstable and the endemic fixed point is stable. In Figure 4.8, we plot the phase plane for the system (4.14) when $\lambda = 0.4$, $\gamma = 1/6$, $A = 1/50$ (so $R_0 \approx 2.14$). Then, $a \approx 9.3$ and the endemic point is $(0.467, 0.057)$. For the simulation we started with an initial value of $(S, I) = (0.999, 0.001)$, performed steps of size 1 and stopped when the particle reached inside a circle of radius 10^{-6} and center (S^*, I^*) (this took 586 steps). Everything the theory suggested can be visualized using this plot. The value we chose for A does not correspond to any real scenario, but it helps in visualization purposes.

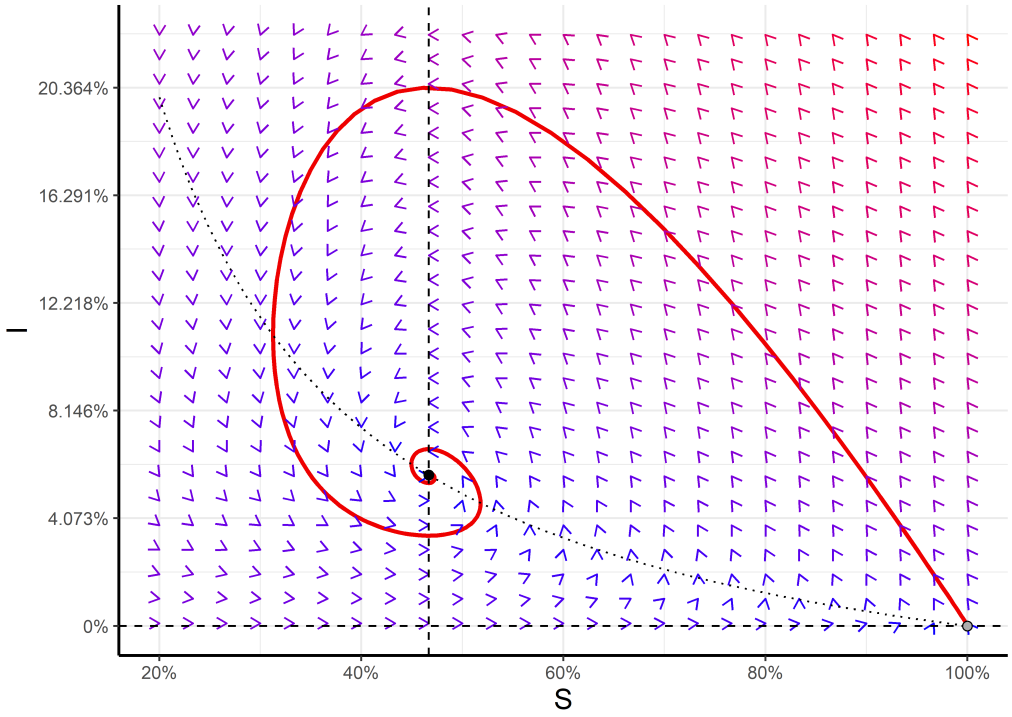

Figure 4.8: The SI phase plane of the system (4.14). The vector field is shown with the color of the arrows indicating the magnitude of each vector: the more red, the larger. Also, the isoclines are shown: the vertical dashed line corresponds to $S = 1/R_0$, the horizontal dashed line corresponds to $I = 0$ and the dotted line corresponds to $I = \dfrac{A(1-S)}{\lambda S}$. The first two indicate no vertical motion, while the last one indicates no horizontal motion. The red line corresponds to the epidemic course at the simulation conducted. The black and grey dots are the stable and unstable fixed points respectively.

Decomposing Figure 4.8, as we did in the SIR case, we can see in Figure 4.9 that both S and I constantly change direction (whenever the derivative changes its sign) and they oscillate around the x- and y-coordinate of the endemic point respectively (recall that in these plots the particle moves on the x-axis). In the case of S, this happens when the particle hits the curve $I = \dfrac{A(1-S)}{\lambda S}$, while in the case of I, this happens when the particle hits the curve $S = 1/R_0$. We have noted on the plots every time that this happens by a dashed line. Also, the motion of the particle in the upper right corner of Figure 4.8, where the red arrows indicate larger velocities, can also be seen in Figure 4.9 because at the beginning of the motion of the particle, the derivatives take on large

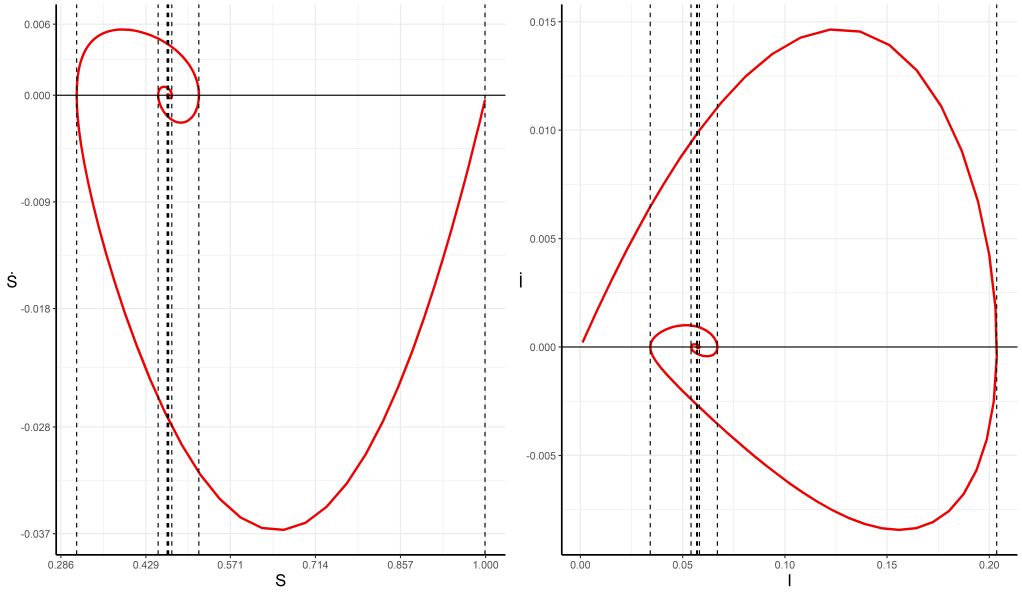

Figure 4.9: Decomposition of the SI phase plane into two parts regarding susceptible and infectious individuals. The motion takes place on the x-axis. The velocity at each point is given by the y-coordinate of the red lines, where a negative value means movement to the left. Left: The values $\{(S, \dot{S}) : S \in [0, 1])\}$. Right: The values $\{(I, \dot{I}) : I \in [0, 1])\}$. Note that the x-axis ticks are added below each plot for clarity, instead of the actual x-axis in the middle of the plots.

values (either positive or negative).

Another way we can break Figure 4.8 into separate susceptible and infectious information is Figure 4.10. This is just the two simulated time series $S(t)$ and $I(t)$ plotted together. We can see that $S(t)$ oscillates around $S^* = 0.467$ and $I(t)$ oscillates around $I^* = 0.057$. This is a very different image from the classical SIR case, where $S(t)$ is monotonically decreasing and $I(t)$ returns to zero after its wave (see Figure 1.6).

SIR through a different lens

As we have noted, the third equation regarding the creation of removed individuals plays no role in the dynamics of S and I and we exclude it from our analysis. However, before closing this Section, we provide some insights using the third-order SIR systems – with

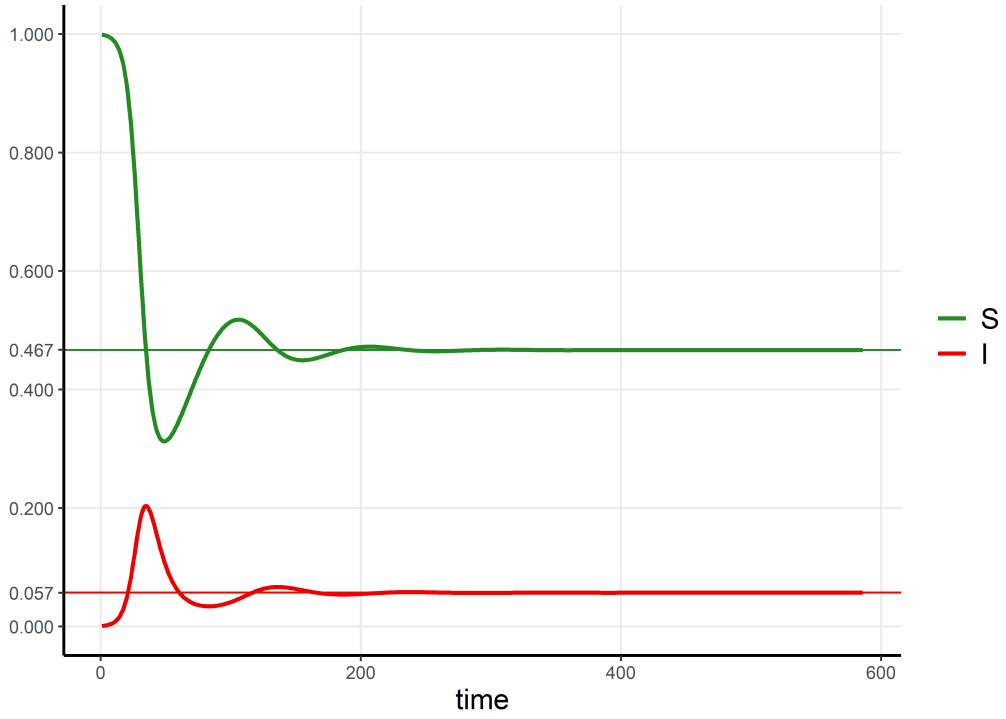

Figure 4.10: Decomposition of the SI phase plane into the time series of $S(t)$ and $I(t)$.

or without demography – which can be helpful and stimulate further research regarding the dynamics analysis or the differential geometry. To this end, let us write the full SIR model:

$$\begin{aligned} \dot{S} &= -\lambda SI \\ \dot{I} &= \lambda SI - \gamma I \\ \dot{R} &= \gamma I \end{aligned} \tag{4.15}$$

as well as the SIR with demography model:

$$\begin{aligned} \dot{S} &= -\lambda SI + A - AS \\ \dot{I} &= \lambda SI - \gamma I - AI \\ \dot{R} &= \gamma I - AR \end{aligned} \tag{4.16}$$

The term AR in the last equation expresses that removed individuals die at rate A. Also, we do not allow any newborns to be immune (thus belonging to the R-state immediately), so we do not have a term for births in the third equation.

The epidemic course in the three-dimensional space for each model takes place at a

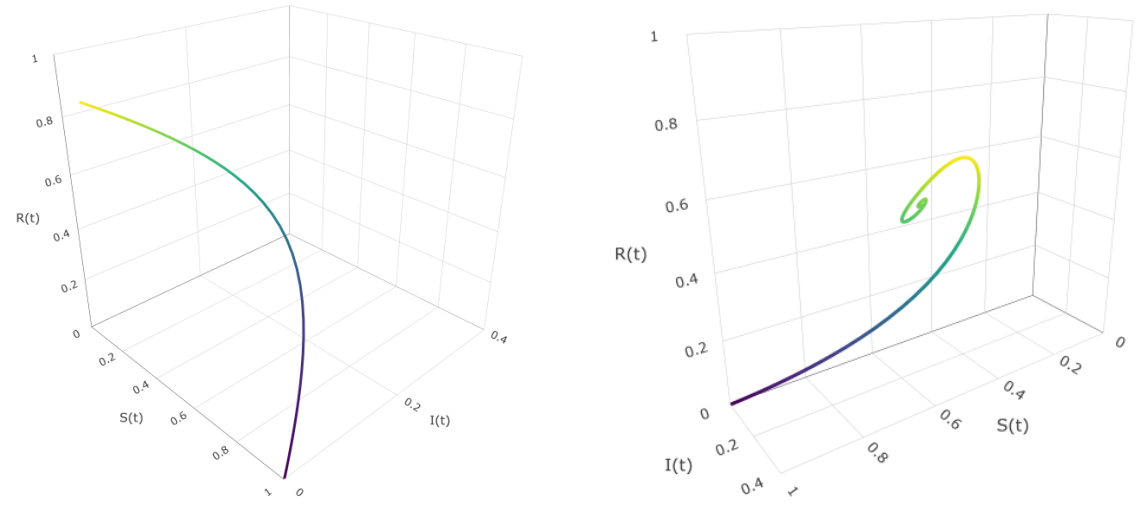

Figure 4.11: The three-dimensional epidemic course with $S(t)$, $I(t)$ and $R(t)$ in the x-, y- and z-axis respectively. Left: The simulated epidemic course for system (4.15). Right: The simulated epidemic course for system (4.16).

height different from zero as the R-state equation defines. So, a simulation using $\lambda = 0.4$, $\gamma = 1/6$ and $A = 1/50$ (like the ones we used for each model before) yields the curves shown in Figure 4.11.

Furthermore, let us try to find a conserved quantity for the case when no demography is assumed. A conserved quantity $f(\mathbf{x}(t))$ is a function of the variables $\mathbf{x}$ of a system that does not change as time passes, i.e. $\dfrac{df}{dt} = 0$. To this end, we form the quantity $\dfrac{dI}{dS}$ using the first two equations of system (4.15) and we have

$$\frac{dI}{dS} = \frac{\lambda SI - \gamma I}{-\lambda SI} \Rightarrow \frac{dI}{dS} = -1 + \frac{\gamma}{\lambda S} \tag{4.17}$$

and, by separating variables and integrating we get:

$$\begin{aligned} dI = \big(-1 + \frac{\gamma}{\lambda S}\big)dS \Rightarrow \int dI &= \frac{\gamma}{\lambda}\int \frac{dS}{S} - \int dS \\ &\Rightarrow I + S = \frac{\gamma}{\lambda}\log(S) + const \\ &\Rightarrow S(t) + I(t) - \frac{\gamma}{\lambda}\log\big(S(t)\big) = const \\ &\Rightarrow \frac{d}{dt}\Big(S(t) + I(t) - \frac{\gamma}{\lambda}\log\big(S(t)\big)\Big) = 0 \end{aligned} \tag{4.18}$$

Thus, $Q(t) = S(t) + I(t) - \dfrac{\gamma}{\lambda}\log\big(S(t)\big)$ is a conserved quantity, which is visualized

in Figure 4.12. On the other hand, Q is not conserved in the SIR with demography model and it only characterizes the "energy" function of the simple SIR, something that serves as a criterion for determining whether the SIR formulation adequately describes an epidemic. A better insight can be obtained by another simulation like that in Figure 4.2, only now we use more initial values: we use as $S(0)$ a sequence of 50 evenly spaced numbers from 0.999 to 0.45. For $I(0)$ and $R(0)$ we use $1 - S(0)$ and 0 respectively or 0.001 and $1 - S(0) - 0.001$ respectively. In the middle left and right plots of Figure 4.12 the simulated courses projected onto the SI-plane are shown for systems (4.15) and (4.16) respectively using the aforementioned initial values. In the upper left and right plots, we show the $R(t)$ values for these paths on the z-direction, while in the lower left and right plots we calculate Q over these paths. The first row of plots just generalizes a single epidemic curve in three dimensions creating a surface which might be interesting in studying as a geometric object. Regarding the last row of plots, in the simple SIR case, each course corresponds to a curve of constant height in the z-direction, while in the SIR with demography case each course creates a curve of varying height.

Based on the previous observation about the conserved quantity $Q(t)$, we can construct a variable that measures the deviation of a model from the simple SIR. Since in the latter case, $Q_0(t)$ is constant for initial values (S_0, I_0, R_0), then $(Q_0^*(t) - Q_0(t))^2$ (where $Q_0^*(t)$ is obtained by a different model) measures the deviation from the SIR for the specific set of initial values (S_0, I_0, R_0) for a specific time t (the absolute difference can also be used instead of the squared one). Thus, $\int_a^b (Q_0^*(t) - Q_0(t))^2 dt$ is the deviation of the non-SIR from the SIR model during the time interval (a, b). For instance, using the initial values $(S_0, I_0, R_0) = (0.999, 0.001, 0)$ for both the simple SIR and the SIR with demography, the energy level of the former is constant at 1.000417, in contrast with the latter which varies and creates a total deviation of 13.69964 from the beginning of the epidemic until the SIR with demography equations have converged using the same criterion as we did with previous simulations. For the total deviation we use the sum of the time-specific deviations using the values S and I produced by the ODE solver. If the correct SIR model was identified, then the individual deviations as well as the total would be zero. Thus, results about the distribution of these quantities would help in

constructing statistical tests checking the SIR hypothesis.

Finally, we can reduce the system (4.15), which is a third-order system of ODE (since it involves three different variables and their derivatives with respect to time) to a first-order system, so that it can be analysed like our example of $\dot{x} = \cos(x)$. Following some steps described in Appendix A5, we get the first-order system

$$\frac{du}{dt^*} = a - bu - e^{-u}$$

where $a = 1 - \log S_0 \frac{\gamma}{\lambda}$ and $b = \frac{\gamma}{\lambda}$. For the new parameters, it holds that $a > 1$ since we start with a small initial S value inside $(0, 1)$ and, so $\log S_0 < 0$, while $b > 0$ since it is the ratio of two positive parameters. Equation (5.28) allows us to study the SIR system of equations as a first-order system.

In order to have a picture of the vector field on the real line, we can plot the function $f(u) = a - bu - e^{-u}$ corresponding to the velocity of a phase point, as we have done in the $\dot{x} = \cos(x)$ example at the beginning of the Chapter, but we can also break the function into the graphs $f_1(u) = a - bu$ and $f_2(u) = e^{-u}$. Now, since $f(u) = 0$ corresponds to $f_1(u) = f_2(u)$, we plot the two functions and look for points where they intersect. When $f_1(u)$ is above $f_2(u)$, then movement of the phase point is to the right, since then $f(u) = f_1(u) - f_2(u) > 0$, while $f_1(u) < f_2(u)$ indicates movement to the left. Of course, $f_1(u)$ depends on the two parameters a and b, so for our simulation scenario we use $b = 1/2.4$ (so that R_0 is 2.4 like our previous simulations for the simple SIR) and $S_0 = 0.999$, i.e. $a = 1.000417$. Then, we can integrate the system using $R(0) = 0$, corresponding to $u_0 = 0/b - \log(S_0)$. In Figure 4.13 the vector field is shown with arrows indicating the way of motion. The black dot indicates a stable fixed point and the grey dot indicates a unstable fixed point. The phase point is absorbed in $u^* = 2.110041$, which means that the final size of the epidemic is $R^* = (u^* + \log S_0)/R_0 = (2.110041 + \log(0.999))/2.4 = 87.88\%$ of the total population. The left plot of Figure 4.13 depicts the vector field with the two functions $f_1(u)$ and $f_2(u)$ determining the dynamics, while in the right plot we show the same vector field in terms of the original velocity function $f(u)$.

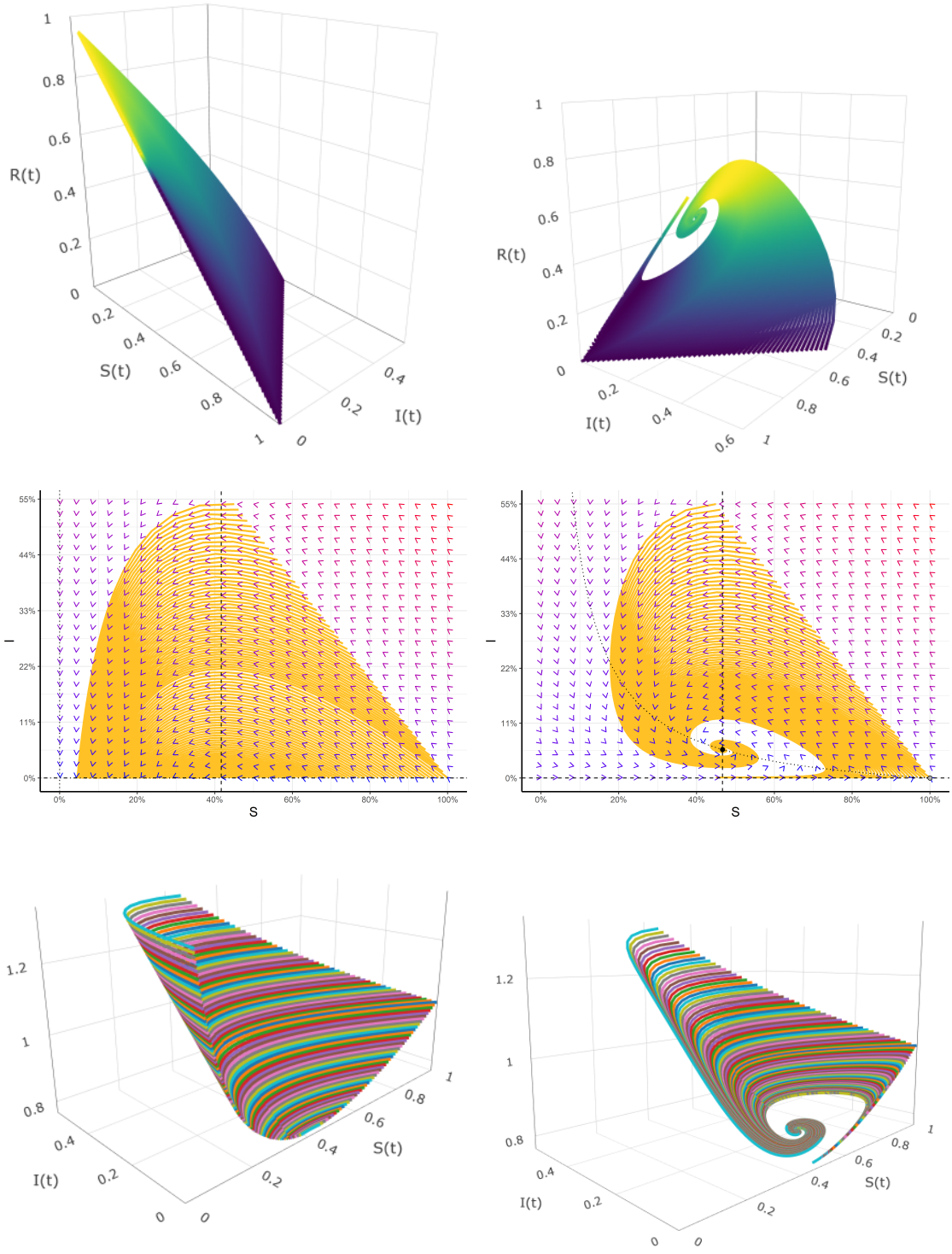

Figure 4.12: Upper left and right: The simulated three-dimensional courses for models for systems (4.15) and (4.16) respectively. Middle left and right: The simulated courses for models for systems (4.15) and (4.16) respectively projected on the SI plane. Lower left and right: The simulated variables $Q(t)$ for systems (4.15) and (4.16) respectively over the SI plane.

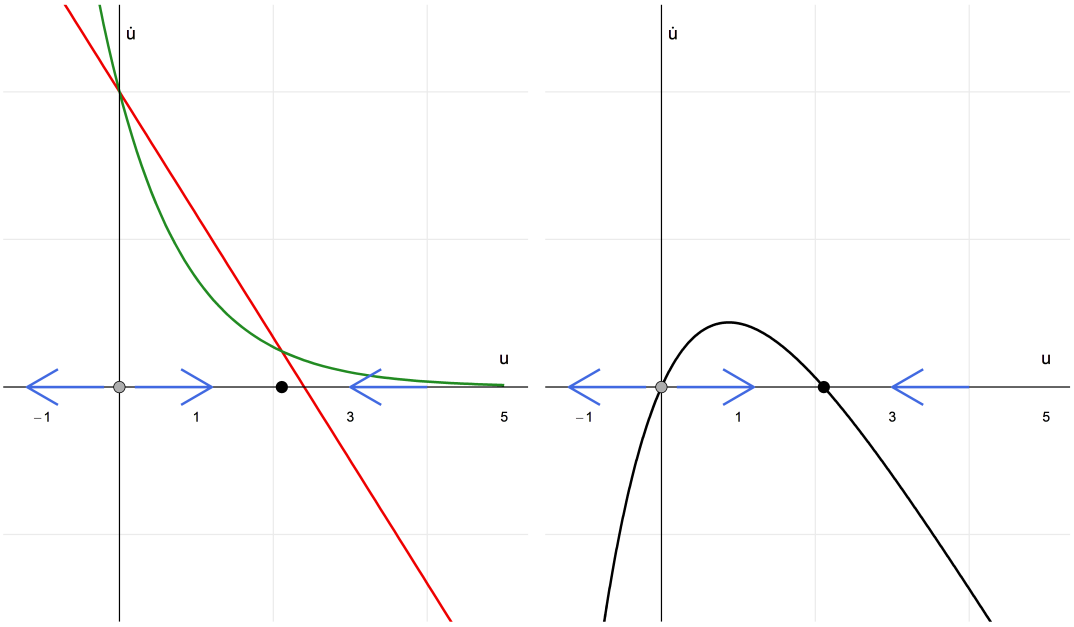

Figure 4.13: The vector field corresponding to $\dot{u} = a - bu - e^{-u}$ on the real line. Left: Visualization using the functions $f_1(u) = a - bu$ and $f_2(u) = e^{-u}$. Right: Visualization using the function $f(u) = a - bu - e^{-u}$.

4.2 The discrete stochastic case

The dynamics described in Section 4.1 regard systems on the continuous-time scale, while our models are based on discrete time. Thus, in the present Section, we first present how one can calculate the quantities mentioned so far when thinking of discrete values of time and apply them on the phase plane of a simple SIR model. Then, we continue examining more complex models from Chapter 3.

4.2.1 Translation and the SIR model

To begin with, since our models are built on the discrete time scale (on daily data), every infinitesimal quantity dt mentioned in the previous Section is translated into a time difference that always equals 1. Furthermore, our stochastic models place variability to the number of susceptible and infectious individuals – due to variability in the total cases

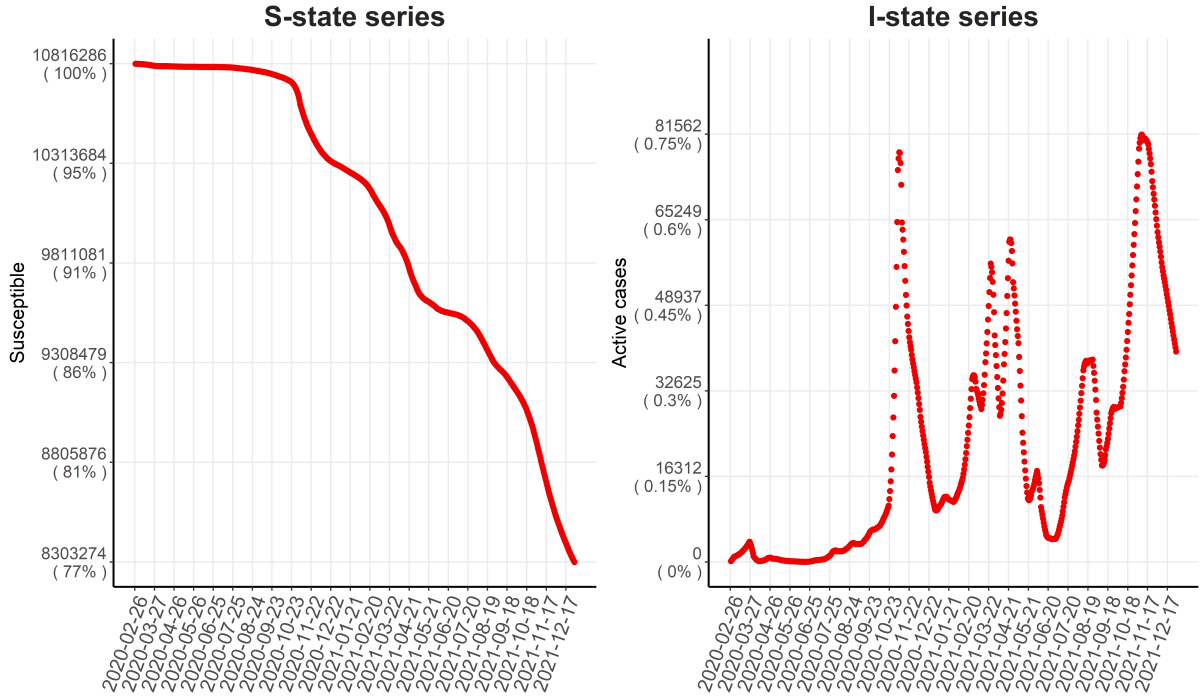

Figure 4.14: $\hat{S}_t$ and $\hat{I}_t$ series produced by equations (4.19) and (4.20) respectively of the SIR model. The median values are used as estimates.

– so we use the median value of the simulations as an estimate.

Let us first discuss the case of the simple SIR model given by the following equations:

$$
\begin{aligned}
d_t &\sim NB(\theta_t, \psi) \\
\theta_t &= p_t \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k \\
C_t &= \lambda_{t-1} S_{t-1} I_{t-1} / N \\
S_t &= S_{t-1} - C_t \qquad (4.19) \\
I_t &= \sum_{k=0}^{\tau-1} C_{t-k} \qquad (4.20)
\end{aligned}
$$

where the meaning, interpretation, exact form and prior distributions for the parameters are described in Chapter 3. What is of concern to us are the equations (4.19) and (4.20), which describe the estimation of the susceptible and infected individuals respectively.

Fitting the model on Covid-19 data from Greece from 26/02/2020 to 31/12/2021

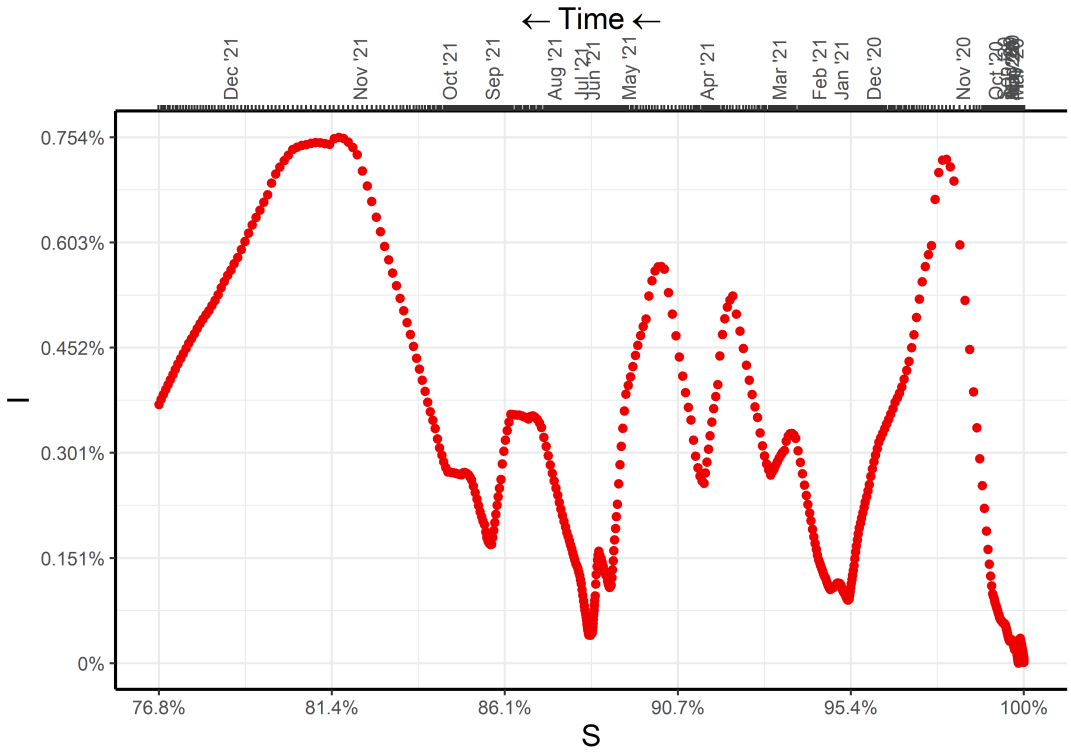

Figure 4.15: The SI-plane produced by the SIR model. The red points are median values of the susceptible and infectious individuals. The S and I states are presented as proportions of the total population, while we also included a top axis to indicate the time at which each corresponding pair occurs.

yields the estimates $\hat{S}_t$ and $\hat{I}_t$ whose median values are shown in Figure 4.14. We have argued that it is not right to treat them separately or one at a time, so combining the y-axis values of the two plots, we obtain the estimated discretized actual course of the epidemic $\hat{\boldsymbol{\gamma}}_t$ on the SI plane shown in Figure 4.15 (based on this specific model). The movement is from the bottom right corner to the left, since the susceptible population decreases. Compare this form of an actual course to that of Figure 4.2, which also refers to an SIR model. The real case is much more complicated, since the assumptions are not exactly met in the real world, the parameters change, intervention measures interrupt the theoretical dynamics and different time horizons are displayed.

We included a top axis whose ticks correspond to each $(\hat{S}_t, \hat{I}_t)$ point, because it makes obvious our point of high and low velocity: We can see that the points are far from each

other during big waves of the epidemic or, in other words, during a dangerous period the imaginary particle travels a larger distance to the left than during less dangerous periods (for example November 2020 to December 2020 compared with December 2020 to January 2021). Of course, visualizing time passing horizontally from right to left, only shows our point with respect to changes in S, i.e. with respect to new total cases. It does not capture the magnitude of changes in I, meaning that moving from (S_1, I_1) to (S_2, I_2) is the same visually as moving from (S_1, I_1) to (S_2, I_3) (also see the x-component of the length of the course below). The speed is not computable by formula (4.3), because we only have an estimate per day, so we can calculate the speed by $||\boldsymbol{\gamma}_{t+1} - \boldsymbol{\gamma}_t||$ according to the definition of the velocity $\lim_{h\to 0} \dfrac{\boldsymbol{\gamma}(t+h) - \boldsymbol{\gamma}(t)}{h}$ where $h = 1$ day (we cannot take a limit as we have discrete values at our disposal). Therefore, the discrete version of speed at time t is the root of the squared difference of the susceptible individuals plus the squared difference of the infectious ones, i.e.

$$v_t = \sqrt{(S_{t+1} - S_t)^2 + (I_{t+1} - I_t)^2} \tag{4.21}$$

and similarly $a_t = v_{t+1} - v_t$.

For the epidemic work between times a and b, we can use the discrete analogue

$$W_{a,b} = \sum_{t=a}^{b-1} (S_{t+1} - S_t)^2 + (I_{t+1} - I_t)^2 = \sum_{t=a}^{b-1} v_t^2 \tag{4.22}$$

while for the length of the course, a direct analogue is

$$l_{a,b} = \sum_{t=a}^{b-1} \sqrt{(S_{t+1} - S_t)^2 + (I_{t+1} - I_t)^2} = \sum_{t=a}^{b-1} v_t \tag{4.23}$$

thus, considering only the x- or y-component gives

$$l_{a,b}^{(x)} = \sum_{t=a}^{b-1} \sqrt{(S_{t+1} - S_t)^2} = \sum_{t=a}^{b-1} |S_{t+1} - S_t| \tag{4.24}$$

and

$$l_{a,b}^{(y)} = \sum_{t=a}^{b-1} \sqrt{(I_{t+1} - I_t)^2} = \sum_{t=a}^{b-1} |I_{t+1} - I_t| \tag{4.25}$$

Let us assume that the susceptible and infectious functions are monotonic inside (a, b). This is actually the case for the susceptible population under the SIR model with

the equation (4.19), where susceptible individuals only decrease and so, for times $i < j$ we have $S_i > S_j$. If the equation (4.19) is written as $S_t = N - \sum_{k=1}^{t} C_k$ (which may be useful if one wishes for a general rather than a recursive formula), it produces once again the interpretation of the x-axis of the SI plane: distances on the x-axis of two times i and j with $i < j$, $l_{i,j}^{(x)}$, equals the number of new infections that occurred between the two days whose distances are computed, i.e.

$$l_{i,j}^{(x)} = S_i - S_j = N - \sum_{k=1}^{i} C_k - \left(N - \sum_{k=1}^{j} C_k\right) = \sum_{k=1}^{j} C_k - \sum_{k=1}^{i} C_k$$
$$= \sum_{k=1}^{i} C_k + \sum_{k=i+1}^{j} C_k - \sum_{k=1}^{i} C_k = \sum_{k=i+1}^{j} C_k$$

Thus, on the SI plane, each point S_i on the x-axis lies $l_{i-1,i}^{(x)} = C_i$ units apart from its previous S_{i-1}. If $i = t' - \tau$ and $j = t'$, then we get

$$l_{t'-\tau,t'}^{(x)} = \sum_{k=t'-\tau+1}^{t'} C_k = \sum_{k=0}^{\tau-1} C_{t'-k} = I_{t'}$$

The last equality is just equation (4.20). So, the susceptible values just before the beginning and at the end of one infectious period, which ends at time t', lie $I_{t'}$ units apart.

On the other hand, when I_t is increasing, it means that active cases on day t are more than active cases on day $t-1$, that is the new cases are more than the removals (deaths or recoveries) and, decreasing I_t shows that the removals are fewer than the new infections. Thus, the larger the difference $I_t - I_{t-1}$, the larger severity of the epidemic at that period. We can spot such periods on the SI-plane on the "right part" of each wave. For instance, in Figure 4.15 we observe large differences when we go uphill the wave of the period mid-October to early-November 2020. When we go downhill the waves, the differences $I_t - I_{t-1}$ are negative; the larger this absolute difference, the more new removals compared with new infections. If we compute the y-distance of consecutive I_t values, we have $I_t = \sum_{i=0}^{\tau-1} C_{t-i}$ and $I_{t-1} = \sum_{i=0}^{\tau-1} C_{t-1-i}$, so

$$l_{t-1,t}^{(y)} = |I_t - I_{t-1}| = |C_t - C_{t-\tau}|$$

In conclusion, the x- distance between two consecutive points on the SI plane is the difference of the cases, while the y-distance is the difference of the cases after one infectious period. These two combined give the Euclidean distance between the points.

Lastly, the two intervention effectiveness measures in (4.26) and (4.27) can be approximated by

$$L_{a,b} = \sum_{t=a}^{b} \sqrt{\frac{(S_t^{(n)} - S_t^{(a)})^2 + (I_t^{(n)} - I_t^{(a)})^2}{S_t^{(n)2} + I_t^{(n)2}}} \tag{4.26}$$

and

$$M_{a,b} = \frac{W_{a,b}^{(n)} - W_{a,b}^{(a)}}{W_{a,b}^{(n)}} \tag{4.27}$$

during the time interval (a, b). Again, the superscripts (n) and (a) indicate natural and actual course respectively, but they can also refer to two different courses under different measures. In the right plot of Figure 4.6, we can visualize what $L_{a,b}$ does. It divides the sum of the lengths of the black lines connecting the simulated points of the two courses with the sum of the lengths of the dotted lines.

4.2.2 More complex models

Let us now shift the focus to the complete SEIR model described in Chapter 3, which includes the effect of demography and vaccination and is described by

$$\begin{aligned}
d_t &\sim NB(\theta_t, \psi) \\
\theta_t &= p_t \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k \\
C_t &= \lambda_{t-1-h} S_{t-1-h} I_{t-1-h} / N \\
S_t &= S_{t-1} - C_t - V_t + A \cdot (1 - S_{t-1}/N) \qquad (4.28) \\
I_t &= \sum_{k=0}^{\tau-1} C_{t-k} - A \cdot I_{t-1}/N \qquad (4.29)
\end{aligned}$$

Again, we are concerned with the equations producing S and I, i.e. equations (4.28) and 4.29).

The SI phase plane using the median values of our estimates is shown in Figure 4.16. Once again, this image is much different than that in Figure 4.8. Of course, Figure 4.8 refers to a slightly different model, but we use it for comparison, since they both assume an endemic point and to keep the deterministic continuous-time dynamics less complex. From this Figure, we can see that the epidemic at 31/12/2021 had led to the

point $(S, I) = (0.4388, 0.0039)$, so approximately 43.88% of the Greek population escaped infection, while 0.39% was at that time infectious.

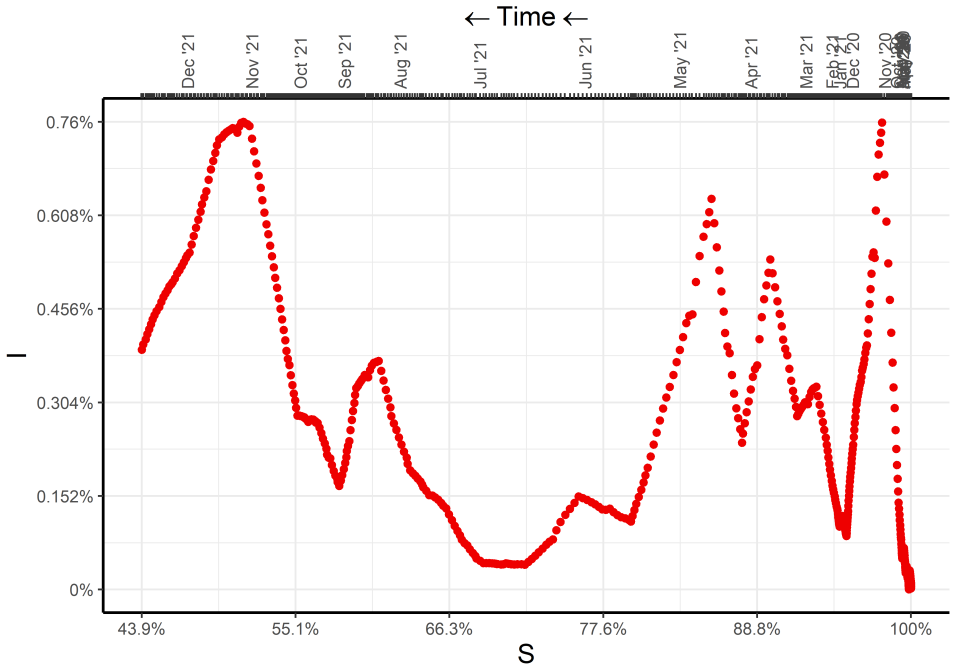

Figure 4.16: The SI plane produced by the SEIR with demography and vaccination model.

The difference between the SEIR model accounting for vaccination and the one which does not can be seen more clearly in Figure 4.17. The left plot shows the simulated course based on the SEIR with vaccination and demography model ($\boldsymbol{\sigma}_2(t)$) in green color and the same model without vaccination ($\boldsymbol{\sigma}_1(t)$) in red color, so instead of (4.28) we have $S_t = S_{t-1} - C_t + A \cdot (1 - S_{t-1}/N)$. As initial point we consider the time when vaccination began. We observe that vaccination stretches the course, in the sense that at a given time, the full model estimates the same infectious individuals, but less susceptible ones, since vaccination moves more people to the R-state (the right plot zooms in the first six weeks for better visualization).

Finally, let us examine whether the estimated (S, I) quantities of Greece suggest that an SIR model would be suitable by checking the quantity $Q_t = S_t + I_t - \frac{\gamma}{\lambda_t} \log S_t$. If the SIR assumption is correct, than Q_t has to be constant, or nearly constant due to randomness in the estimates. In the left plot of Figure 4.18, we can see a downward trend

in the supposedly conserved quantity, so the SIR assumption does not seem to hold true, except of roughly the first 224 days where it seems to be approximately constant. After that, the variance seems to get larger and then the values are obviously deviating from a fixed value. The zoomed in picture for this time interval is depicted in the right plot of the same Figure. The random fluctuations are due to errors in S and I, so the ergodic mean is also displayed in red color. We can say that Q_t values are around 1.000349 and, if the pattern continued, we would not reject the SIR assumption. The value 1.000349 is just the mean value of the ergodic mean. In Figure 4.19, a 3-dimensional plot is presented with the (S, I) values on the xy plane and Q_t on the z-coordinate.

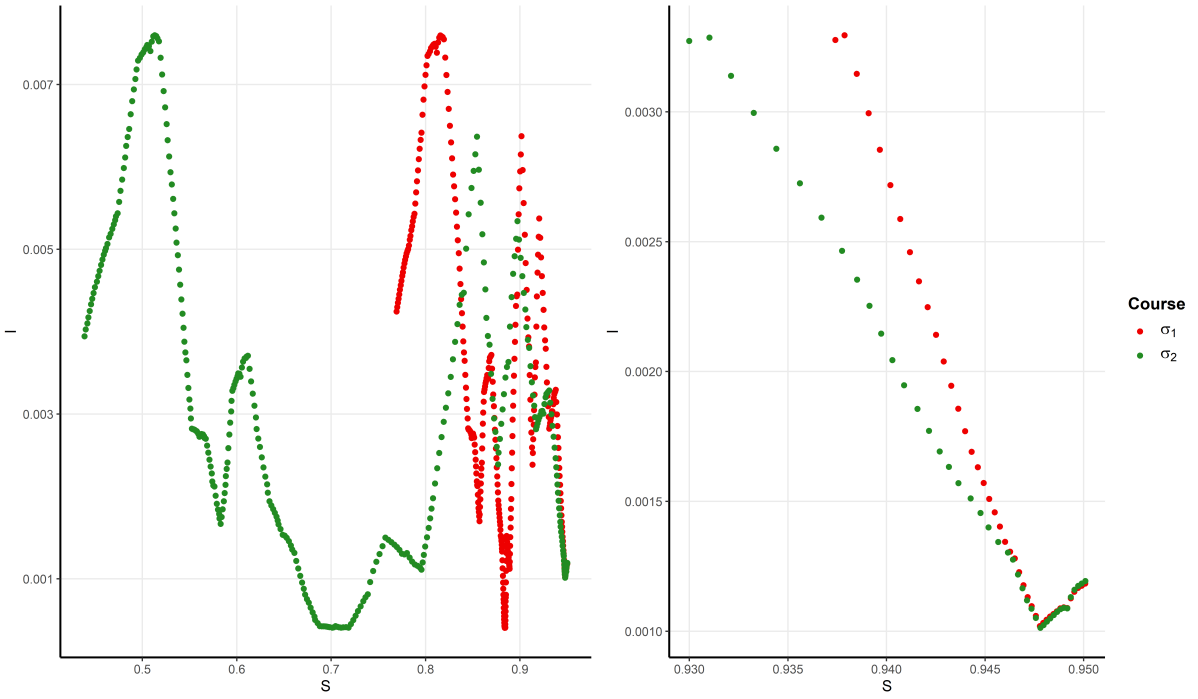

Figure 4.17: Left: The course without vaccination in red and the course with vaccination in green starting from the time vaccination was introduced. Right: Focus on the first six weeks after vaccination began.

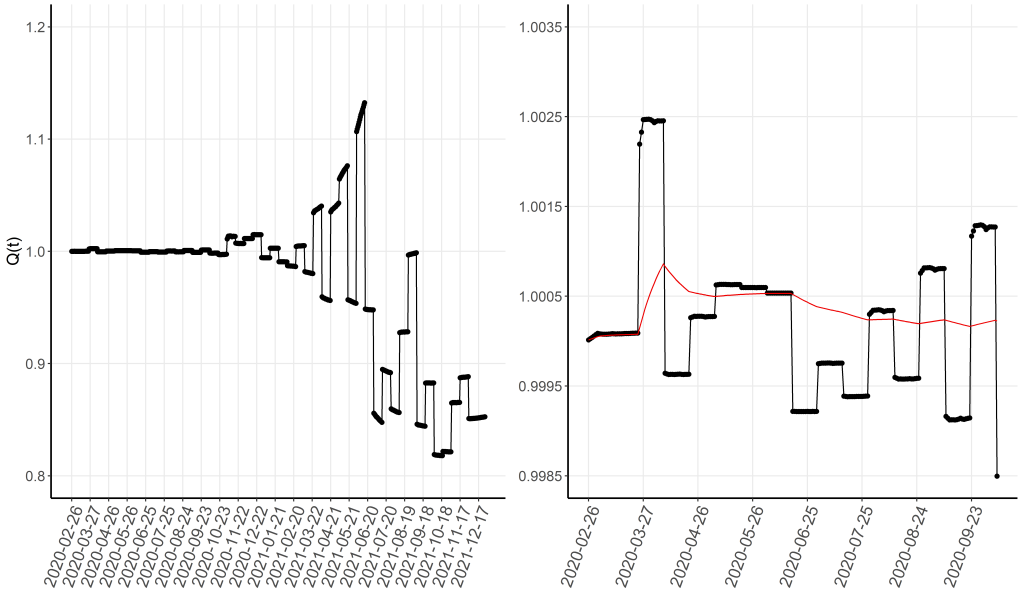

Figure 4.18: Left: Q_t for the (S, I) values of Greece. Right: Zoomed in picture of the first 225 days for better visualization of the random fluctuations of Q_t, albeit approximately constant in that interval. Red line is used for the ergodic mean of the points.

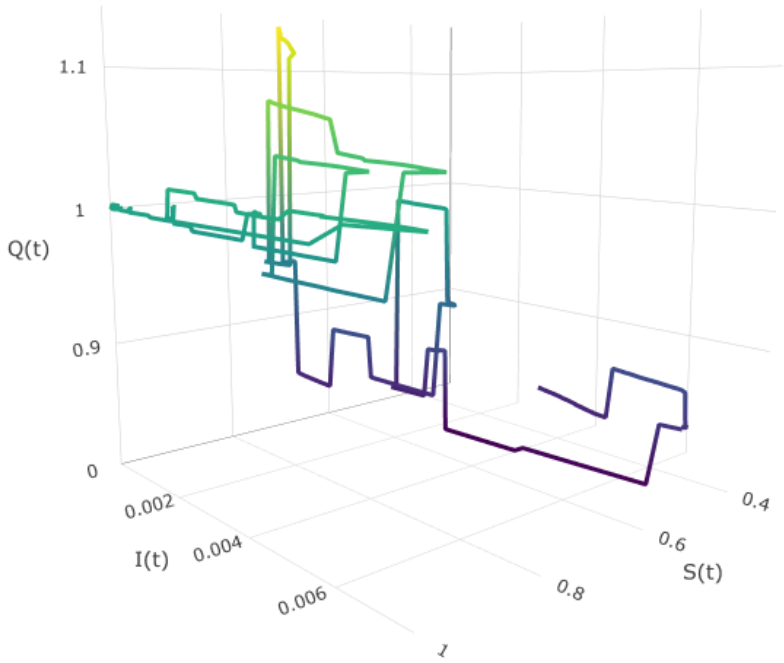

Figure 4.19: The epidemic course uplifted to Q_t height on the z-axis.

4.3 Discussion

We devoted this Chapter in analysing the dynamics of a ODE system describing an epidemic. We argued that susceptible and infectious individuals should be considered simultaneously in decision making/support processes. We introduced some concepts already existed in other fields in the epidemic framework using two basic models, namely the SIR and the SIR with demography. These two models are very different, since the latter induces an endemic situation and creates very different dynamics. Moving our focus from the series of S and I to the SI phase plane, we constructed two measures of intervention effectiveness, since an epidemic is never left to spread under no restrictions in the real world, but non-pharmaceutical interventions interrupt the theoretical dynamics. Finally, we propose the use of the conserved quantity Q as a test for the SIR model assumption, which at the moment only has a visual aspect and lacks the classical formalism for testing a statistical hypothesis, but this is something that can be fruitful for future research.

Chapter 5

Concluding Remarks

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
- John Tukey

In this Thesis we are concerned with the analysis of biomedical data under two distinct scenarios, namely communicable and non-communicable diseases. We examined the latter under the scope of survival analysis and extrapolation and provided applications on which our methodology can be useful. Specifically, we demonstrated a coherent way of survival extrapolation utilizing projections of mortality on three datasets regarding breast cancer, metastatic melanoma and cardiac arrhythmia. The first dataset also acts as demonstrative of our methods, since its survival curve is mostly known. The second estimates the life years gained when combining the usual pembrolizumab therapy with an mRNA vaccine and provides evidence of promising results. The third dataset shows the applicability of our method under the presence of cause-specific hazards.

Regarding the communicable diseases, we examined public Covid-19 data from European countries focusing more on Greece, the UK and the USA. We constructed models of the SEIRS type that can infer important epidemiological quantities such as the reproduction number and the proportion of observed cases, through the latent total cases. To this end, we synthesized different data sources using the Bayesian methodology, which

was questioned but arose as the most efficient among the tested competitors. Finally, we examined the dynamics of an epidemic using ordinary differential equations, produced some quantities regarding the course of an epidemic and applied these ideas to the discrete stochastic nature of our Covid-19 models gaining intuition about the current state of the epidemic and theoretical potential of its future state.

The two types of biomedical problems tackled in the Thesis are of major concern to scientists. Specifically, after the large Covid-19 pandemic, it has become apparent that efficient methods that describe and predict the behaviour of infectious diseases are needed. On the other hand, because of the increasing deaths caused by cancer combined with the promising results of mRNA vaccines, the survival extrapolation tools will be key to future research. However, much more work needs to be done, as explained via the difficulties and remarks throughout the Thesis in the corresponding Chapters.

Future work will be based on both epidemic and survival analysis trying to overcome the barriers imposed by the public data of Covid-19, making more in-depth analysis of the dynamics of infectious diseases, conducting simulation studies using the ideas developed in Chapter 4 in Covid-19 oriented scenarios scenarios and, finally, applying the survival extrapolation ideas on more real data, extending our methodology to cost-effectiveness scenarios, thus linking it directly to industry and aiding its use by practitioners. Finally, the two worlds meet when using the survival curve to model the proportion of susceptible individuals, so combining the ideas developed in this Thesis may lead to promising future results.

Appendix

A1. Parametrizations of models used

Distributions in Chapter 2

Here we clarify the parametrization of the Weibull models found in Chapter 2. The Weibull distribution has a PDF of the form

$$W(x;\alpha,\sigma) = \frac{\alpha}{\sigma}\Big(\frac{x}{\sigma}\Big)^{\alpha-1}\exp\Big(-\big(\frac{x}{\sigma}\big)^{\alpha}\Big) \tag{5.1}$$

for a continuous variable X that takes values $x \in \mathbb{R}_+$ with shape parameter $\alpha > 0$ and scale parameter $\sigma > 0$. Another parametrization relies on the rate parameter $\lambda > 0$ (which is the reciprocal of the scale) instead of σ and the PDF then reads

$$W(x;\alpha,\lambda) = \lambda\alpha x^{\alpha-1}\exp\big(-\lambda x^{\alpha}\big) \tag{5.2}$$

In both parametrizations, the exponential part is the survival distribution and the other is the hazard function. Moreover, in both parametrizations, the shape parameter determines whether the hazard function increases ($\alpha > 1$), decreases ($\alpha < 1$) or stays constant ($\alpha = 1$). In the last case, the Weibull reduces to the Exponential distribution. This result regarding the monotonicity of the hazard can easily be checked and we derive it for the rate parametrization as follows.

Let $h(x) = \lambda\alpha x^{\alpha-1}$ be the hazard function. Then,

$$\frac{d\,h(x)}{dx} = \lambda\alpha(\alpha-1)x^{\alpha-2}$$

The terms λ, α and $x^{\alpha-2}$ are positive, while $\alpha - 1 > 0 \Leftrightarrow \alpha > 1$. Thus, the hazard increases when $\alpha > 1$ and decreases when $\alpha - 1 < 0 \Leftrightarrow \alpha < 1$. If $\alpha = 1$, then $\frac{d\,h(x)}{dx} = 0$, therefore $h(x) = constant$ (a unique property of the Exponential distribution).

Distributions in Chapter 3

For modeling the SARS-CoV-2 transmission, we assume a likelihood of cases or deaths of the Poisson or Negative Binomial family. Thus, we clarify here the specific parametrizations we use. Let X be a discrete random variable that takes values $x \in \mathbb{N}_0$ and follows the $Poisson(\mu)$ distribution with rate $\mu \in \mathbb{R}_+$. Then, its probability mass function is

$$Poisson(x;\mu) = \frac{e^{-\mu}\mu^x}{x!} \tag{5.3}$$

with mean $\mathbb{E}[X] = \mu$ and variance $Var[X] = \mu$. Here, X is the number of events occurred in a fixed time interval, which occur with rate μ independently of the time since the last event. Therefore, the $Poisson(\mu)$ is a natural choice of modeling count events, like the number of Covid-19 cases or deaths at each day.

Now, let X be the discrete random variable that counts the number of failures in a sequence of Bernoulli trials until the ψth success. Then X follows the Negative Binomial $NB(\psi, p)$ distribution, with probability mass function

$$NB(x;\psi,p) = \binom{x+\psi-1}{\psi-1} p^{\psi}(1-p)^x \tag{5.4}$$

where $\psi \in \mathbb{N}$ is the number of successes, $x \in \mathbb{N}_0$ is the number of failures, and $p \in (0,1)$ is the probability of success. The binomial coefficient can be written as

$$\begin{aligned}\binom{x+\psi-1}{\psi-1} &= \frac{(x+\psi-1)!}{(x+\psi-1-\psi+1)!(\psi-1)!} = \frac{(x+\psi-1)!}{x!(\psi-1)!} = \frac{(x+\psi-1)!}{x!(x+\psi-1-x)!} \\ &= \binom{x+\psi-1}{x}\end{aligned}$$

and, by substituting with its real-valued analogous form $\frac{\Gamma(x+\psi)}{x!\,\Gamma(\psi)}$ using the Gamma function, we now allow ψ to take on positive real values. If we let $\mu \in \mathbb{R}_+$ to be the mean of X, $\mathbb{E}[X]$, then we have $\mu = \psi\frac{1-p}{p}$, so $p = \frac{\psi}{\mu+\psi}$ and $1-p = \frac{\mu}{\mu+\psi}$. Substituting

the equivalent form of the binomial coefficient and the probabilities p and $1-p$ in the "standard" form (5.4) gives the parametrization

$$NB(x;\mu,\psi) = \frac{\Gamma(x+\psi)}{x!\Gamma(\psi)}\left(\frac{\mu}{\mu+\psi}\right)^{x}\left(\frac{\psi}{\mu+\psi}\right)^{\psi} \tag{5.5}$$

which we use when we refer to the Negative Binomial. The variance of X is $Var[X] = \psi\frac{1-p}{p^2}$, which can be written as

$$\psi\frac{1-p}{p^2} = \psi\frac{1-p}{p}\frac{1}{p} = \mu\frac{1}{p}$$

We notice that the mean μ is part of the variance, so we decompose the above expression to a sum using a variable D:

$$\mu\frac{1}{p} = \mu + D \Leftrightarrow D = \frac{\mu}{p} - \mu = \frac{\mu-\mu p}{p} = \mu\frac{1-p}{p} = \mu\frac{\mu}{\psi}$$

Thus, $Var[X] = \mu + \frac{\mu^2}{\psi}$. In contrast to the Poisson distribution, where $\mathbb{E}[X] = Var[X] = \mu$, here the variance is larger than the mean and the parameter that controls this overdispersion is ψ (this is why it is called the *dispersion* parameter). Letting $\psi \to \infty$ leads to the Poisson case. Therefore, we can relax the restrictive assumption of the Poisson using the Negative Binomial, but still modeling just the mean parameter μ.

A2. Results on Chapter 2

UK mortality projections

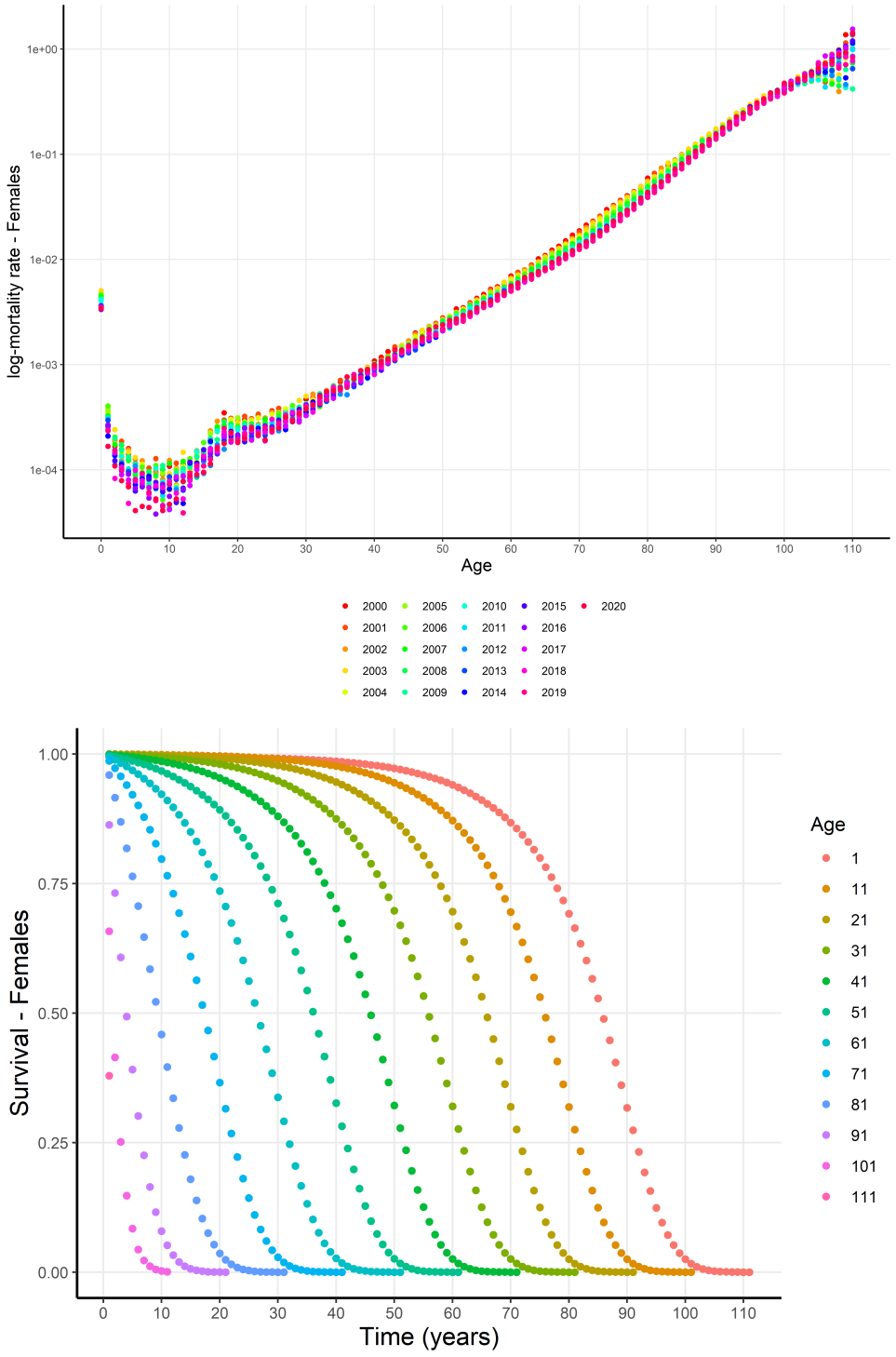

Figure 5.1: Up: log-mortality rate per age in UK for years 2000-2020 for the general female population. Down: Projected survival functions for the general female population in various ages in 2023.

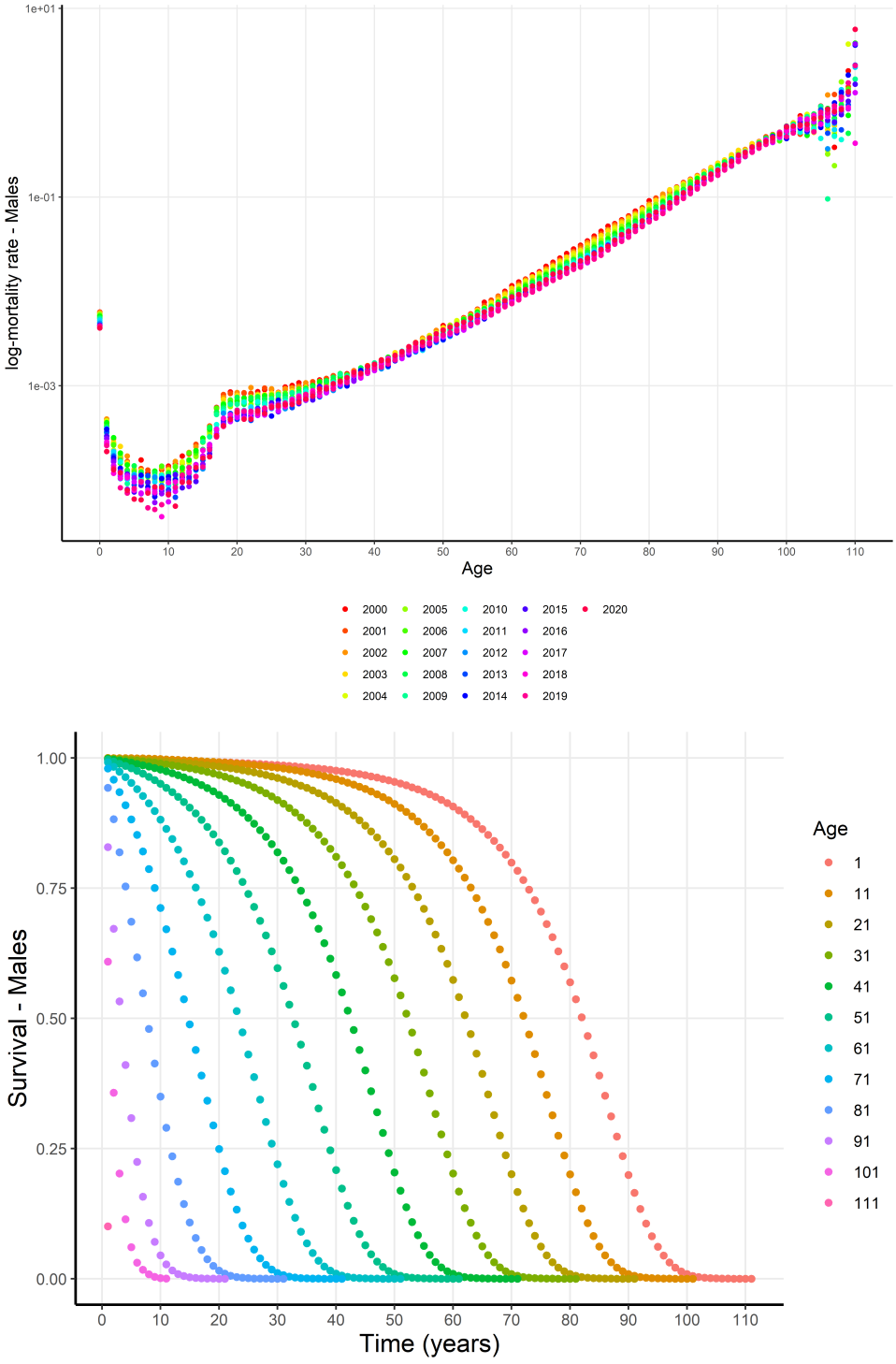

Figure 5.2: Up: log-mortality rate per age in UK for years 2000-2020 for the general male population. Down: Projected survival functions for the general male population in various ages in 2023.

Breast cancer results

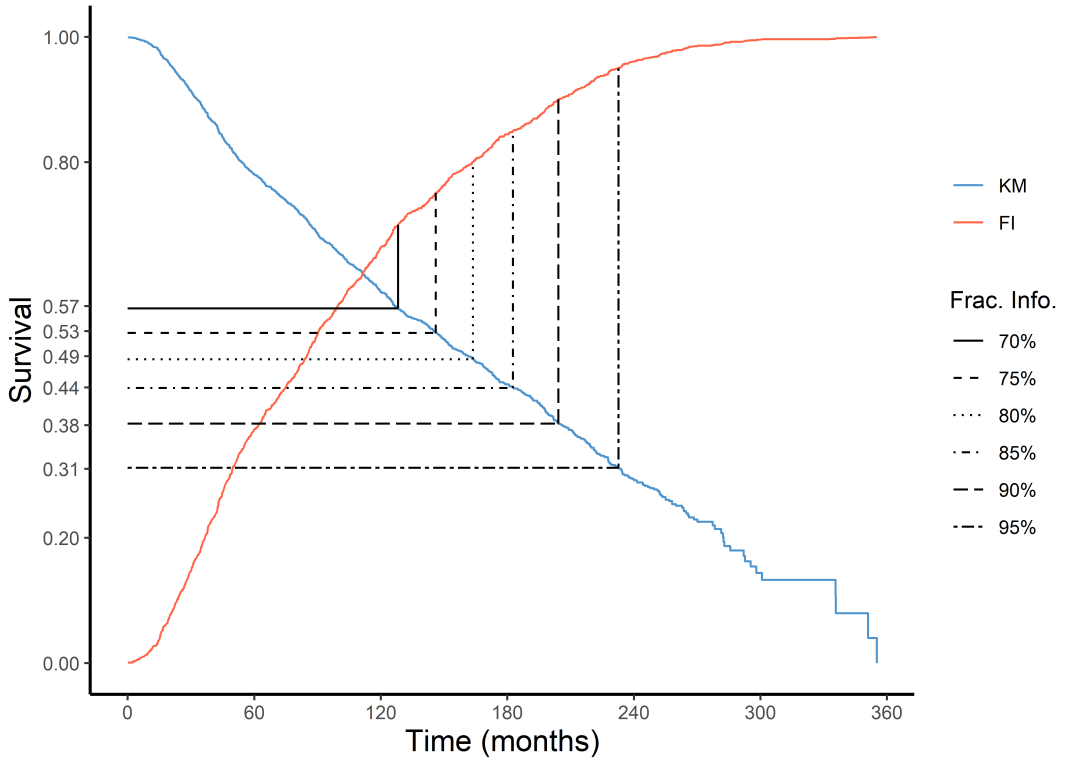

Figure 5.3: Fractional information (FI) and its corresponding Kaplan-Meier (KM) estimate of survival probability at different levels. We apply our method at the 80% level, or 49% survival probability.

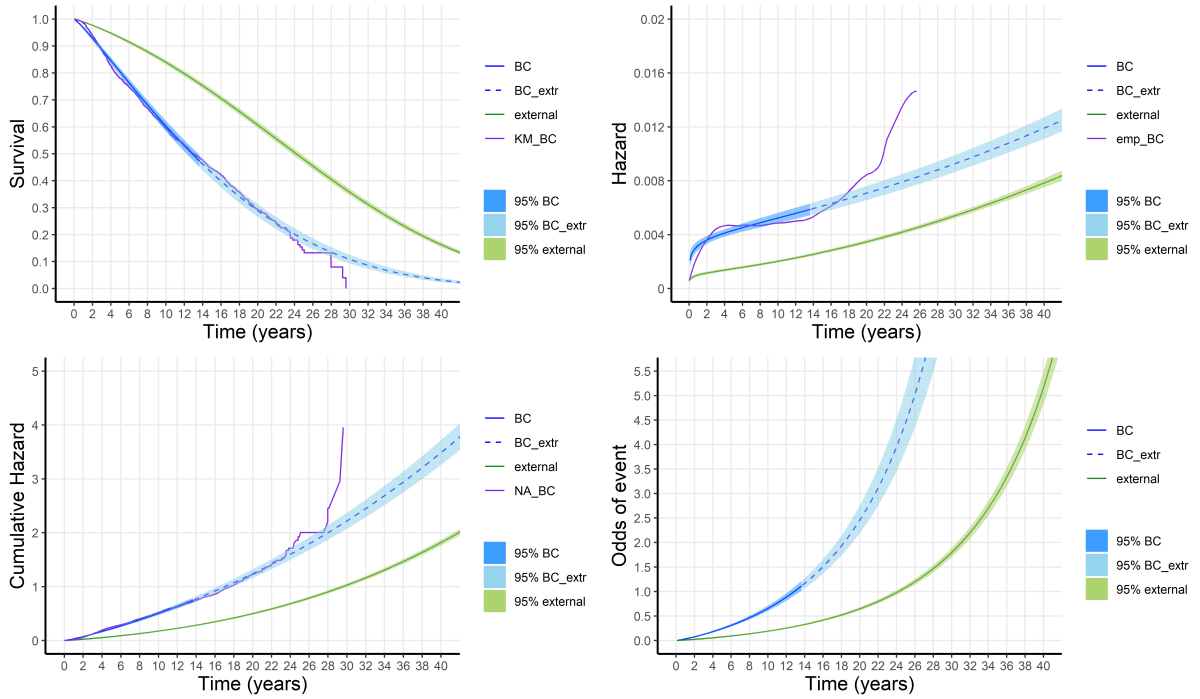

Figure 5.4: Up left and going clockwise: survival, hazard, cumulative hazard and odds functions for the breast cancer data with 95% CrI, when the vanilla method is selected for extrapolation. Non-parametric estimates are added for the follow-up period indicated as "KM" (Kaplan Meier), "emp" (hazard smoothed empirical estimate) and "NA" (Nelson-Aalen estimate).

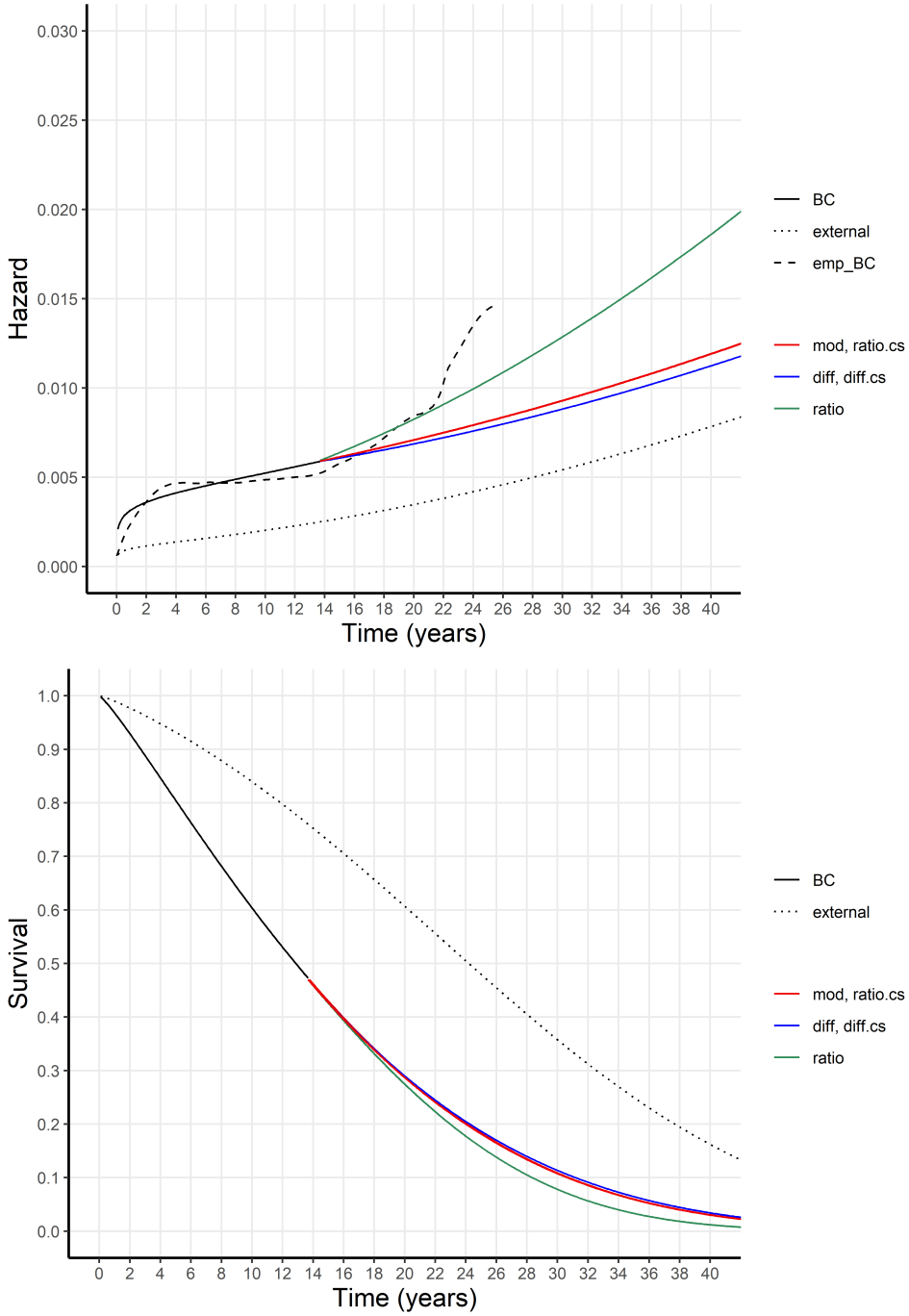

Figure 5.5: Extrapolated survival function (up) and hazard function (down) when keeping only 80% of fractional information for the breast cancer dataset using the five extrapolation methods. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

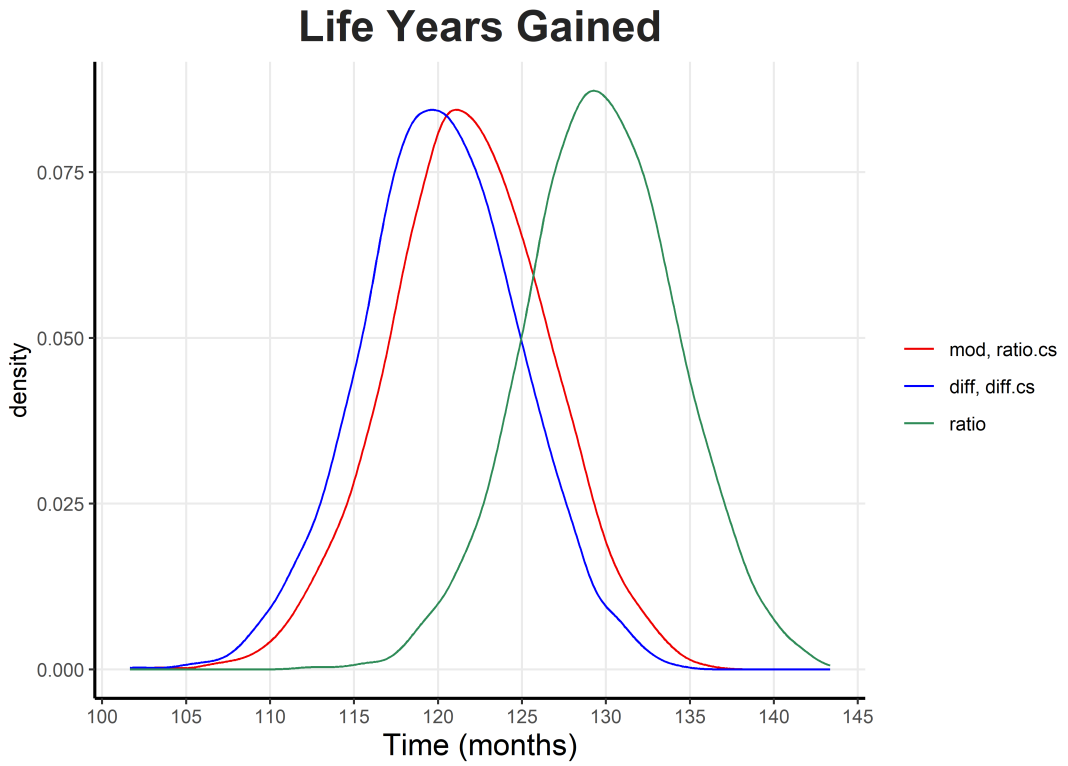

Figure 5.6: Life years lost due to breast cancer. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

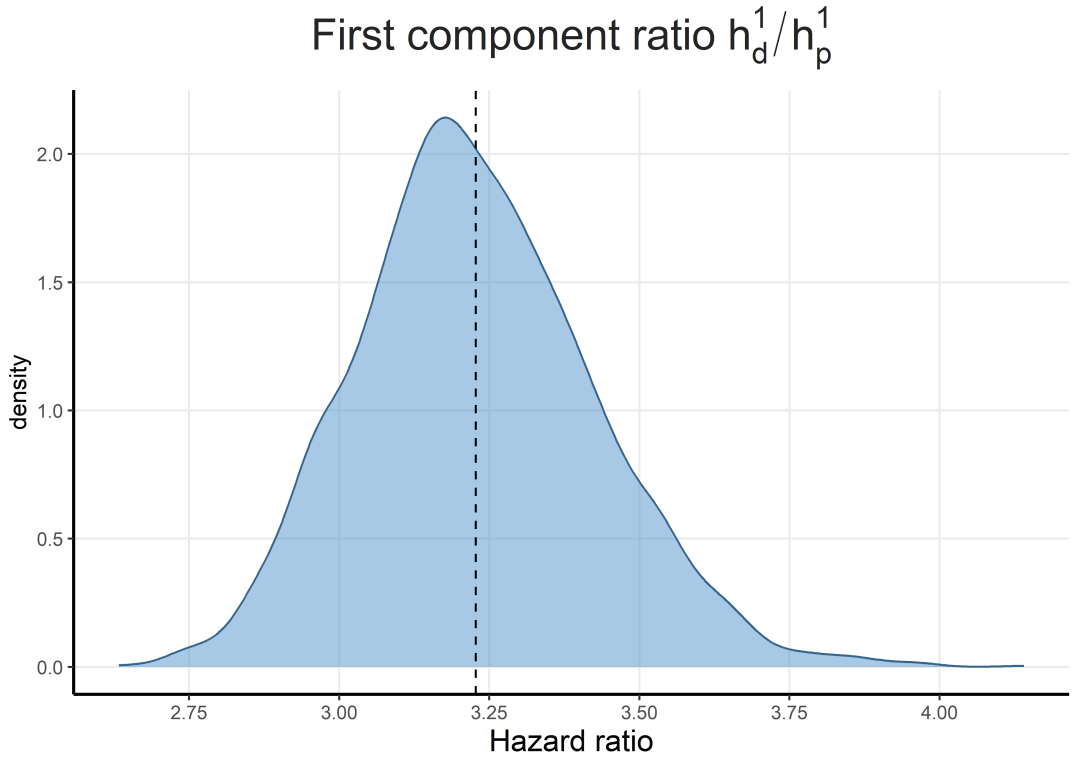

Figure 5.7: Hazard ratio density for the first components of the selected Bi-Weibull model, when it is trained using the whole dataset. Assuming it refers to the breast cancer hazard (a pseudo cause-specific scenario), the hazard ratio is 3.23 on average (indicated by a dashed line).

Advanced melanoma results

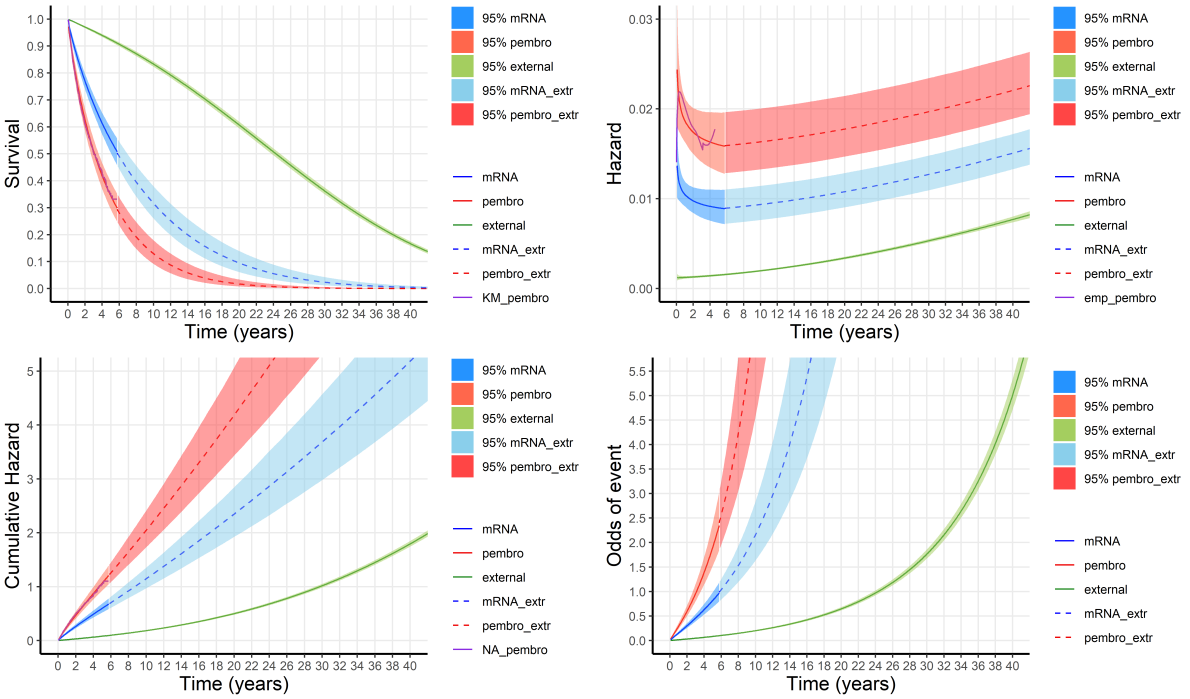

Figure 5.8: Up left and going clockwise: survival, hazard, cumulative hazard and odds functions for the advanced melanoma data with 95% CrI, when the difference method is selected for extrapolation. Non-parametric estimates are added for the follow-up period indicated as "KM" (Kaplan Meier), "emp" (hazard smoothed empirical estimate) and "NA" (Nelson-Aalen estimate).

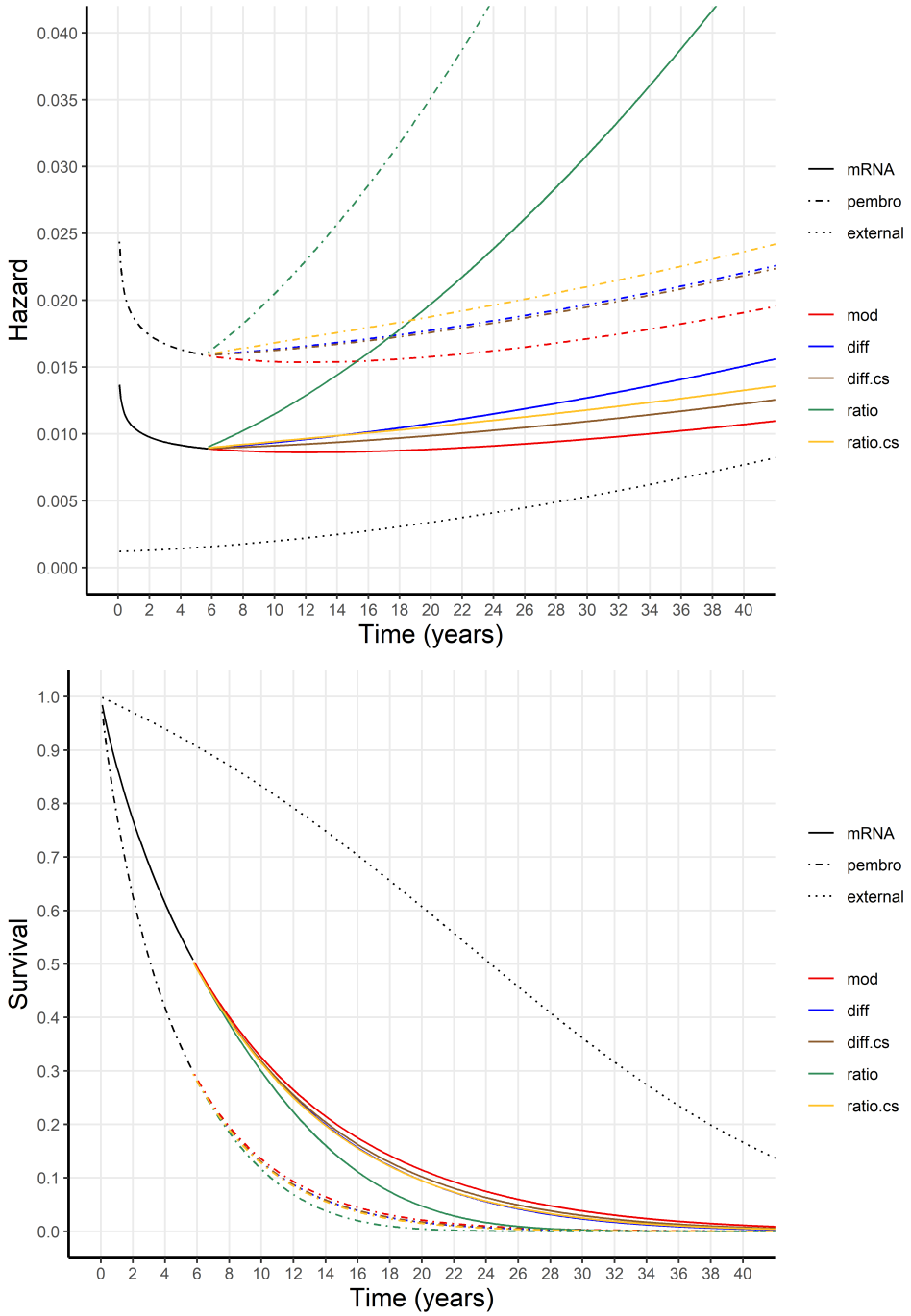

Figure 5.9: Extrapolated survival function (up) and hazard function (down) for the advanced melanoma dataset using the five extrapolation methods. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

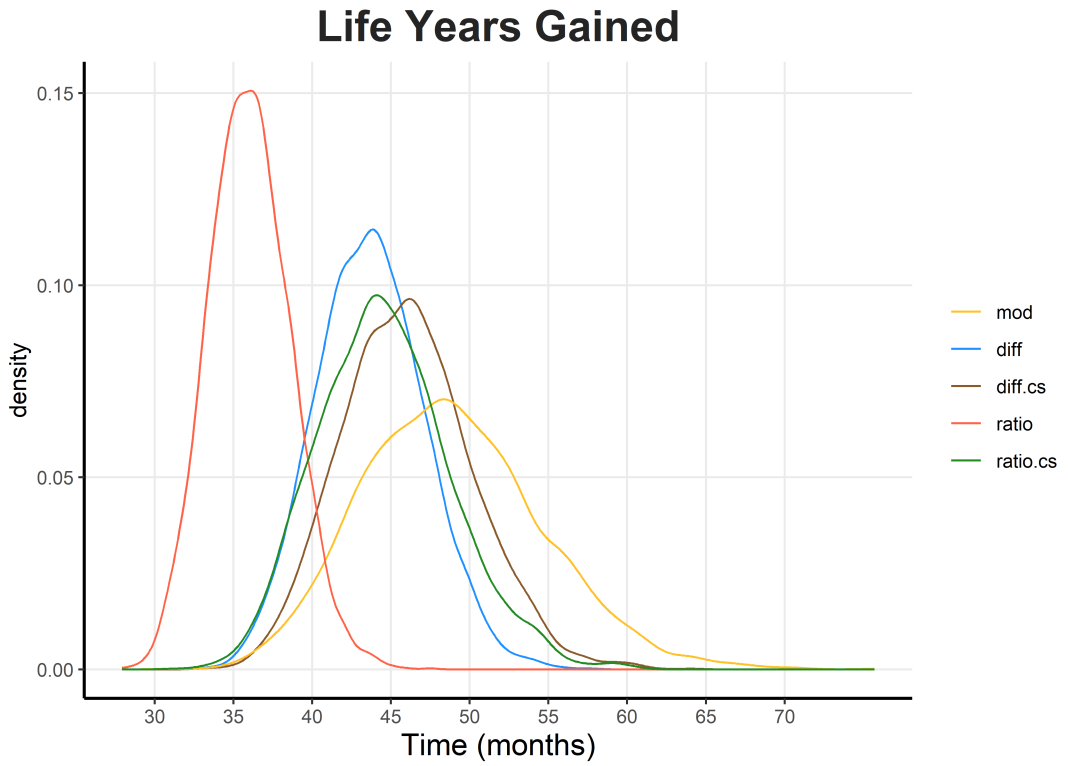

Figure 5.10: Life years gained by the mRNA vaccine compared to pembrolizumab alone. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

Cardiac arrhythmia results

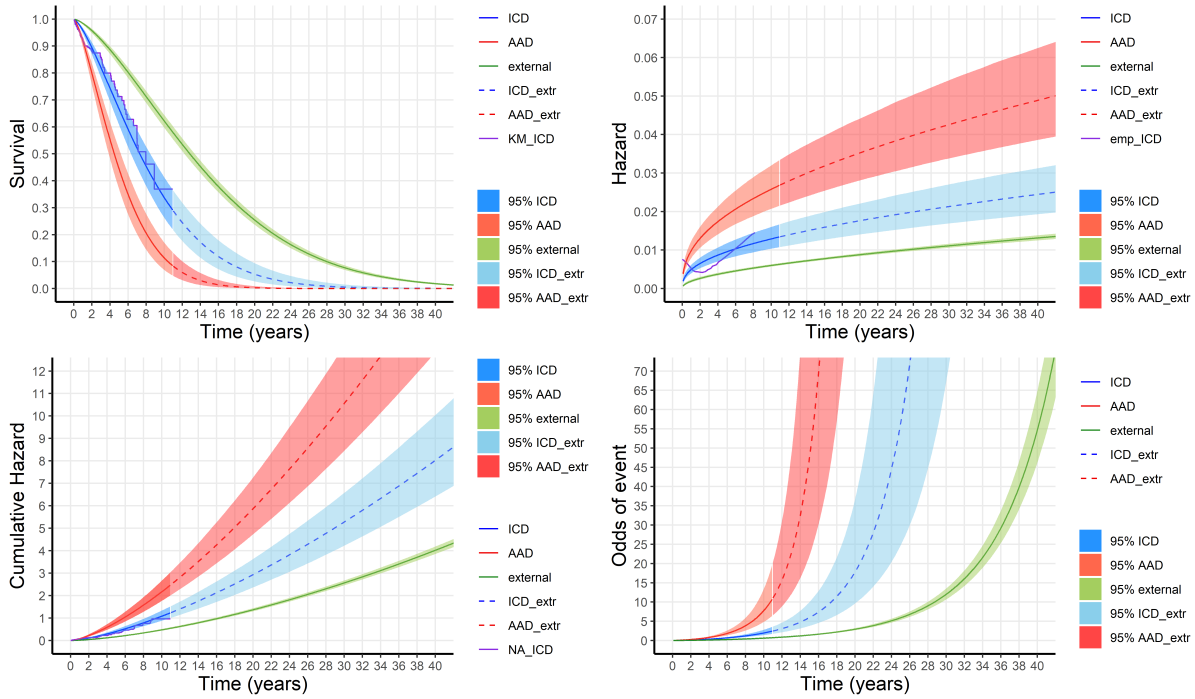

Figure 5.11: Up left and going clockwise: survival, hazard, cumulative hazard and odds functions for the advanced melanoma data with 95% CrI, when the vanilla method is selected for extrapolation. Non-parametric estimates are added for the follow-up period indicated as "KM" (Kaplan Meier), "emp" (hazard smoothed empirical estimate) and "NA" (Nelson-Aalen estimate).

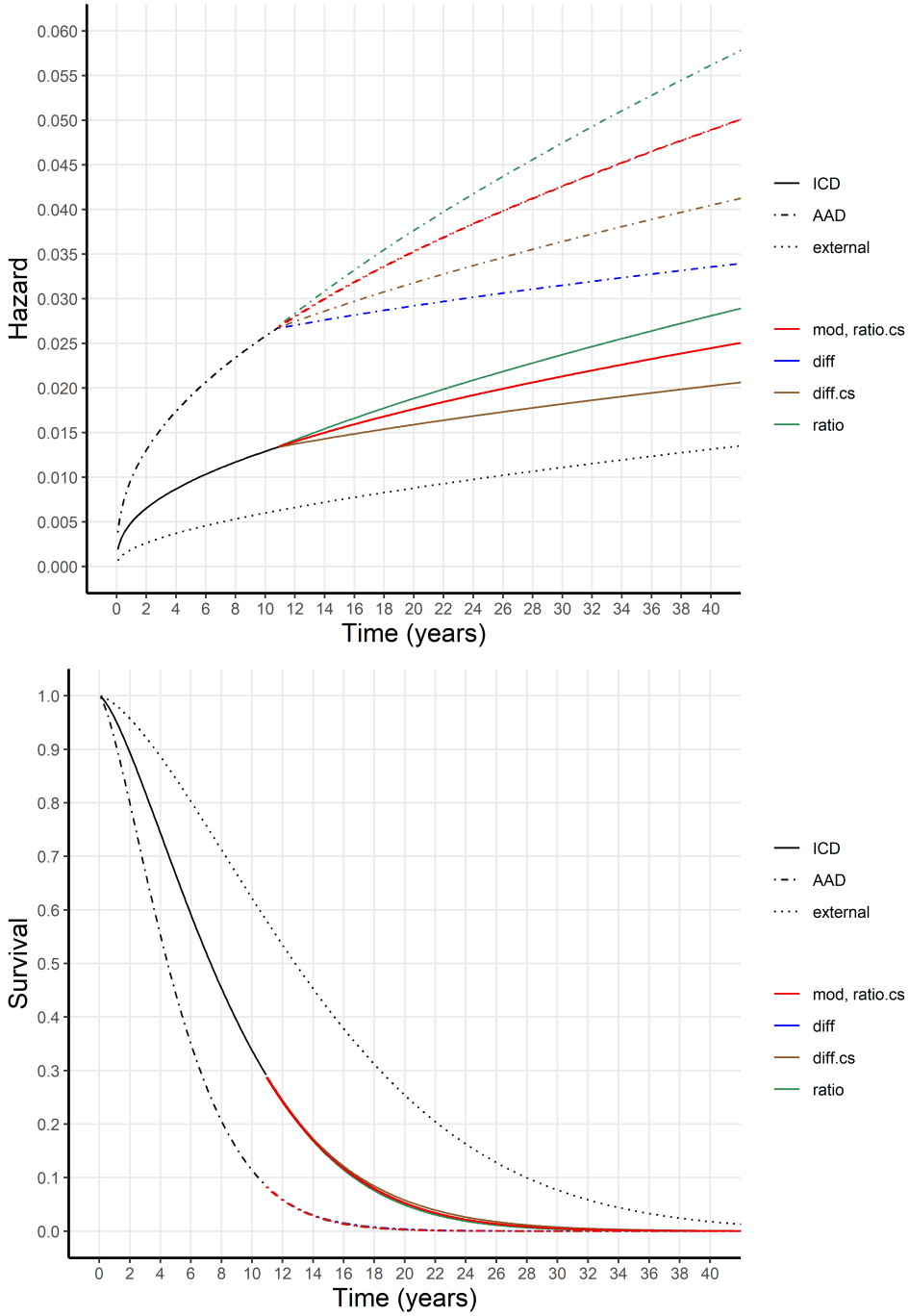

Figure 5.12: Extrapolated survival function (up) and hazard function (down) for the cardiac arrhythmia dataset using the five extrapolation methods. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

Figure 5.13: Life years gained by ICD compared to AAD. The vanilla method is indicated as "mod". The pseudo cause specific methods are indicated as ".cs".

A3. Different models considered for Covid-19

Here we describe the procedure we followed until reaching to the final Covid-19 models presented in Chapter 3 and also less competitive models that test assumptions we used in Chapter 3. We initiate the modeling procedure and the way we should monitor the course of Covid-19 when the virus reached the door of Greece in March 2020 (the first confirmed case was actually on 26/2/2020). We present many different models we tried out for a large period of time. For accurate Covid-19 inference we need a model that both fits the data well and it is based on some meaningful and interpretable assumptions. To this end we are based on the compartmental models we described in Section 1.2, but we adjust them to a non-deterministic and non-continuous framework in contrast to the original versions. Finally, we add and test many different models not only on data of Greece,

but of other countries too. First, we set the primal ideas and slowly move towards more complicated models based on the recorded number of Covid-19 cases. Then, we work with models that are based on the total number of cases, which include both the recorded and the unrecorded ones.

Models based on observed cases

First, we present some stochastic discrete-time epidemic models suitable for Covid-19 which are based on the recorded number of cases. Throughout the whole Section, we mention the training time for each model, which is just a one-point estimate, since it refers to the time we needed for fitting the particular model on the data we had at that time. For clarity, we note the last time point we have data for. As information criteria we use AIC, BIC, DIC, DIC_2 (which is the version of DIC using the variance of the log-posterior) and WAIC. For their definition, the reader can refer to Gelman et al. (2014)[38] who summarize all the above. As far as the data are concerned, we use the ones published by Johns Hopkins University[29] unless noted otherwise, which are time series of the confirmed cumulative cases c_t^0, the cumulative deaths d_t^0 and cumulative recoveries r_t^0 each day, where the superscript "0" indicates the "cumulative" quantity. From these three series, we can calculate the daily quantities by applying first order differences on each of them, i.e. the daily Covid-19 cases are $c_t = \nabla c_t^0$, the daily deaths from Covid-19 are $d_t = \nabla d_t^0$ and the daily recoveries from Covid-19 are $r_t = \nabla r_t^0$, where ∇ is the backward difference operator $\nabla Q_t = Q_t - Q_{t-1}$. Other data sources are Google[40] and Apple[2], which provide six and two time series respectively that represent the daily percentage change of mobility in different kinds of places (more information and analysis on this subject in Chapter 3). All the models are fitted by Hamiltonian Monte Carlo with the NUTS algorithm.

The likelihood is posited on count data of Covid-19 recorded cases, thus we use either a Poisson or a Negative Binomial distribution, with parametrizations given in equations (5.3) and (5.5) in Appendix A1 for clarity. The first model we tested (which included data up to March 2020) was really simplistic and did not catch any of the epidemic characteristics one wishes to explain, but set the basis for modeling an infectious disease.

It contained just a Poisson, or Negative Binomial likelihood, whose mean was written in a log-linear form with the mobility variables of Google as covariates and an autoregressive of order 1 term (from now on autoregressive of order d terms are indicated as $AR(d)$) to smooth the series. This simplistic model was tested on data of Greece, Italy, Spain and China. Thus, we set

$$c_t \sim Poisson(\theta_t) \;\; or \;\; c_t \sim NB(\theta_t, \psi)$$
$$\log(\theta_t) = \beta_0 + \boldsymbol{\beta}_1 \mathbf{M}_t + \phi \log(\theta_{t-1})$$

where beta's, ϕ and ψ have Gaussian $N(0, 5^2)$, Uniform $U(0, 1)$ and half-Cauchy $C(0, 10)$ priors respectively (which will be the default priors for these parameters in future models unless stated otherwise), c_t is the number of observed cases at day t and $\mathbf{M}_t$ is a vector of the six mobility variables. So the log-effect of each was captured by the corresponding component of $\boldsymbol{\beta}_1$. We used informative priors for beta's as a start. After all, HMC works better with such priors and, since we did not restrict their support, we could come back to making it more vague in the sensitivity analysis part after reaching to a final model. Google did not provide mobility variables for China, so we did not include the $\boldsymbol{\beta}_1 \mathbf{M}_t$ term for its models.

The Google variables were of course correlated and could not be added into the model without any pre-processing. This is why in the next version of the model we considered Principal Components Analysis (PCA) in order to reduce the dimensionality and keep only the useful information of our covariates. So we tested again our models using only 1 (the first), 2, or 3 Principal Components (PC's) and compute information criteria (AIC, BIC, DIC and DIC_2) in order to compare them (information criteria for Greece are shown in Table 5.1). Although the models fitted the data well, the epidemic aspect was actually absent, since we could not infer, for example, the most basic epidemiological quantity R_t. However, before dealing with that problem, one question that had arisen was: Is PCA the only way to include the mobility variables into the model?

As alternative candidates in place of PCA, we considered Independent Components Analysis (ICA) and Factor Analysis to be the pre-processing tools. Furthermore, since the variables were much correlated to each other, we thought that someone could use just one of them as a covariate and ignore all the others instead of transforming the

Likelihood	PC	AIC	BIC	DIC	DIC_2	Time (min)
Poisson	-	791.747	805.837	1054.210	10189.980	1.872
	1	1036.422	1041.706	1034.811	1035.232	1.004
	2	931.204	938.249	1016.380	1082.474	0.798
	3	865.259	874.065	977.605	1707.380	1.027
Neg Binomial	-	305.544	321.394	304.992	349.804	3.015
	1	301.099	308.144	303.286	324.637	1.023
	2	300.876	309.682	302.609	323.086	1.226
	3	300.176	310.743	303.258	345.319	1.666

Table 5.1: Information criteria and training time for Greece models. "-" indicates that PCA is not performed and the log-linear equation uses the original six variables. Time corresponds to 10^5 iterations after a 10^5 warm-up for 4 chains.

whole dataset. So, first we compared a few models with ICA and PCA and found out that ICA did not make any difference (in terms of information criteria) so it did not worth the trouble (after all, ICA involves one more step after PCA to make the variables independent). After that, we tried comparing PCA with 1 PC versus keeping only one of the variables (the one that is more correlated to the others, namely the "Residential" variable). Once again, the behaviour was similar (the information criteria were less than 1 unit different), but we preferred keeping the PCA method for the moment and come back to this problem in the future if needed.

Next, we tried to insert the epidemic aspect of the problem into our models by using the information of the number of Susceptible, Infected and Removed individuals for a simple SIR type model and estimating the infection rate λ. The number of Susceptible individuals can be calculated by starting at day 1 from the whole population (of Greece), which is $N = 10.72$ million and for each day subtract the new cases from the susceptible people of the previous day, i.e. $S_t = S_{t-1} - c_t$. The number of Infected individuals can be obtained by subtracting the cumulative deaths and cumulative recoveries at time t from the cumulative cases at the same time, i.e. $I_t = c_t^0 - d_t^0 - r_t^0$, while the Removed individuals can be found by adding the cumulative deaths and cumulative recoveries at each day, i.e.

$R_t = d_t^0 + r_t^0$ (not to be confused with the reproduction number at day t, R_t; from now on we analytically state "Removed individuals" instead of using the symbol). Thus, we can build the term $\frac{\lambda \cdot S \cdot I}{N}$ of the simple SIR model and have it as the mean of the chosen distribution (Poisson or Negative Binomial) for the observed cases. The best of our tries was a Negative Binomial model with mean $\frac{\boldsymbol{\theta}}{N}$, where $\theta_t = \exp(\beta_0 + \beta_1 m_t) S_{t-1} I_{t-1} + \phi \cdot \theta_{t-1}$ and m_t is the first PC of the mobility variables, which gave very plausible results for the infection rate. In other words, with this model we tried to describe the daily new cases to be the ones determined by the SIR model, but we also wrote λ in a log-linear form, which uses mobility information and we also included an AR(1) term to account for autocorrelation in the series. This model allows for estimation of the reproduction number each day t, by multiplying the estimated infection rate with the infectious period, i.e. $R_t = \lambda_t \cdot \tau$. This resulted in nice estimates of R_t with the median of the first day (that is R_0) being approximately 2.4. For τ we use random draws from a half-$Normal(0, 10^2)$ distribution. The problem is that the model is based on the confirmed Covid-19 cases, which are way less than the total ones and also the number of infected people at day t, I_t, is calculated using the recoveries data of John Hopkins, which turned out to be unreliable. At that time, neither of these issues was clear and the model seemed nice and simple.

Since, the aforementioned model seemed to work, we set our focus on whether the AR(1) term should be added on the infection rate λ, i.e.

$$\begin{aligned} c_t &\sim Poisson(\theta_t) \;\; or \;\; c_t \sim NB(\theta_t, \psi) \\ \theta_t &= \lambda_t S_{t-1} I_{t-1} / N \\ \log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1}) \end{aligned} \tag{5.6}$$

or on the mean of the distribution θ, i.e.

$$\begin{aligned} c_t &\sim Poisson(\theta_t) \;\; or \;\; c_t \sim NB(\theta_t, \psi) \\ \theta_t &= \eta_t / N \\ \eta_t &= \lambda_t S_{t-1} I_{t-1} + \phi \eta_{t-1} \\ \lambda_t &= \exp(\beta_0 + \beta_1 m_t) \end{aligned} \tag{5.7}$$

or on both the mean and λ, i.e.

$$\begin{aligned} c_t &\sim Poisson(\theta_t) \;\; or \;\; c_t \sim NB(\theta_t, \psi) \\ \theta_t &= \eta_t / N \\ \eta_t &= \lambda_t S_{t-1} I_{t-1} + \phi_1 \eta_{t-1} \\ \log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi_2 \log(\lambda_{t-1}) \end{aligned} \tag{5.8}$$

or not be added at all, i.e.

$$\begin{aligned} c_t &\sim Poisson(\theta_t) \;\; or \;\; c_t \sim NB(\theta_t, \psi) \\ \theta_t &= \lambda_t S_{t-1} I_{t-1} / N \\ \lambda_t &= \exp(\beta_0 + \beta_1 m_t) \end{aligned} \tag{5.9}$$

In all of the above, m_t is the first PC of the mobility variables. These cases were tested on both Negative Binomial and Poisson models, as well as using PCA on the mobility variables of either Google or Apple. Generally, using the Google or Apple variables produced similar results in terms of the fit, but the λ estimates when Google was used were nicer and smoother. Thus, although the information criteria tended to prefer the Apple models most of the times (even when the difference was small), we decided to continue (at least for the moment) the modeling procedure with Google. Furthermore, since Poisson and Negative Binomial models were similar in fit, but the Poisson credible intervals were narrower because of the equality of the mean and variance, we continued only with Negative Binomial likelihoods. The information criteria for the aforementioned models, as well as the time each one needed to be trained is depicted in Table 5.2 for data up to April 2020. Moreover, from now on when we experiment on the form of the infection rate λ and we do not mention the mean new cases (i.e. the θ_t), we set the mean of the Negative Binomial to be $\lambda_t S_{t-1} I_{t-1}/N$. Finally, the first values of $\log(\lambda_t)$ when AR terms are present in our models are taken from the same equation without the AR terms, for instance the first of the four models above had $\log(\lambda_1) = \beta_0 + \beta_1 m_1$.

Mobility	Likelihood	AR(1)	AIC	BIC	DIC	DIC_2	WAIC	Time (min)
Google	Poisson	λ	1623.895	1632.205	1617.280	1619.418	1703.674	2.881
		θ	2998.077	3006.388	2991.287	2993.765	3211.406	1.140
		λ & θ	1571.219	1581.607	1562.338	1565.575	1663.838	7.273
	Neg Binomial	-	3003.047	3009.280	2998.308	2999.853	3121.772	0.561
		λ	568.012	578.400	562.617	581.567	566.462	15.298
		θ	577.978	588.366	570.716	578.753	574.595	3.386
		λ & θ	570.922	583.388	561.064	573.155	565.816	58.964
		-	584.472	592.782	577.766	580.092	580.540	1.264
Apple	Poisson	λ	1618.048	1626.358	1611.321	1613.873	1720.592	2.531
		θ	2256.032	2264.343	2249.472	2253.429	2316.582	1.733
		λ & θ	1570.446	1580.833	1561.753	1565.115	1677.328	4.585
	Neg Binomial	-	2596.140	2602.373	2591.360	2592.861	2663.701	0.831
		λ	567.1767	577.564	558.721	563.804	564.042	3.9200
		θ	608.917	619.304	599.890	603.268	606.953	10.634
		λ & θ	568.245	580.711	557.523	561.733	563.784	10.727
		-	608.639	616.949	601.861	603.622	606.318	1.168

Table 5.2: Information criteria and training time for models (5.6)-(5.9). "-" indicates that we have not included any AR(1) term at all. Time corresponds to 10^5 iterations after a 10^5 warm-up for 4 chains.

The next models we tried out were designed so that infected individuals move with some probability p to either the Death state or the Recovery state (we split the Removal state in half), but we faced some problems with the model fitting procedure. Therefore, the likelihood was a mixture of two Negative Binomials, which had for state j ($j = 1$ for Death and $j = 2$ for Recovery) at time t a mean of $\frac{\boldsymbol{\theta}_j}{N}$, where

$$\theta_{jt} = \lambda_{jt} \cdot S_{t-1} I_{t-1} + \phi_{1j} \theta_{jt-1}$$
$$\log(\lambda_{jt}) = \beta_{0j} + \beta_{1j} m_t + \beta_{2j} x_{jt} + \phi_{2j} \log(\lambda_{jt-1})$$

m_t is the first PC of Google mobility at day t and x_{jt} is the number of deaths (when $j = 1$), or the number of recoveries (when $j = 2$) at day t. The AR(1) term was tested in different positions: only on λ, only on θ, on both λ and θ, or no AR(1) at all. The model above is the one with both lambda and theta smoothed, i.e. there is an AR(1) term on both the infection rate (the second equation) and on the mean new cases (the first equation). The mixing component p of the likelihood was given a $U(0, 1)$ prior. The best of the four models according to all of the information criteria and also the quickest to train (1.3 hours) was the one with no AR(1) term. The above formulation corresponds to an SIRD model and assumes that individuals move from state I to either R (Recovery) or D (Death), as opposed to a unified Removed state.

Something else we wanted to change was the infectious period τ we used for the calculation of $R_t = \lambda_t \cdot \tau$. Until now, we used random draws from a half-$Normal(0, 10^2)$ distribution to scale the infection rate λ, but this method did not correspond to any actual characteristic of the disease, so a preferred way to obtain R_t was to estimate it by data. The problem was that the available (public) data are the cases, deaths and recoveries and there is no information which person died or recovered and at which day. Thus, we assumed that those who get infected first, recover and die first, i.e. the recoveries and deaths of a day are caused by the first infected that are available in the dataset. When those recoveries and deaths have been distributed, we move on to the next ones. First we allocate the recoveries and then the deaths. For example, suppose on day 1 we had 5 cases and until day 10 we had 0 recoveries and 0 deaths. If on day 11, the first recovery arrives we assign it to one of the 5 cases of day 1. Then, if on day 12 we had 2 recoveries and 1 death, first we assume that 2 (out of the 9 left) individuals recovered and then that

1 person (out of the 7 left) died. Therefore, we had at our disposal data on time-to-death, time-to-recovery and time-to-removal. The last ones were obtained by simply not making the distinction of death and recovery and following the same procedure. We noted the days after which each individual recovers or dies and we used the mean of the days until removal for the infectious period τ when the simple SIR model was under study, or the mean number of days until recovery and the mean number of days until death when we made the distinction between deaths and recoveries. Lastly, we started to include a fixed Exposed period of 4 days by adding this lag to the Susceptible data and using these new " lagged susceptible" data for the $S_{t-1}I_{t-1}/N$ part. For instance, if on day 5 the number of susceptible individuals is 10, we assume that this was the number of the susceptible on day 1 and we had not observed it back then because of the Exposed period, i.e. in informal notation $lagged\_Susceptibles_t = Susceptibles_{t+4}$. For these models, we assumed that there is a single λ that informs us about the mean cases at day t and it depends on mobility, while the SIRD type models do not affect the likelihood (as before), but they just help us estimate R_t more accurately. Thus, the models were the same as in equations (5.6)-(5.9) with the only difference when we included an Exposed period, which affected the $S_{t-1}I_{t-1}$ term. Unfortunately, we did not find any sensible results for R_t after the procedure of estimating τ in that way and the models with the Exposed period were similar to those without it, so we decided that the best candidate is the SEIR type, since we knew this was also closer to the truth. However, some more tuning and investigation was needed, so we continued for a while with the simple SIR type models (this will be the default unless stated otherwise).

Some other aspect of these models was the way we estimate the infection rate λ. Until then, we only considered if an AR term would be necessary or not, so the next step we took was to leave the part

$$c_t \sim NB(\theta_t, \psi)$$
$$\theta_t = \lambda_t S_{t-1} I_{t-1}/N$$

as it is and investigate whether we should write λ as the exponent of the random effect of mobility, i.e.

$$\log(\lambda_t) = \beta_t^m + \phi \log(\lambda_{t-1}) \tag{5.10}$$

or a fixed intercept plus the random effect of mobility, i.e.

$$\log(\lambda_t) = \beta_0 + \beta_t^m + \phi \log(\lambda_{t-1}) \tag{5.11}$$

or a random intercept plus a fixed effect of mobility, i.e.

$$\log(\lambda_t) = \beta_t^0 + \beta_1 m_t + \phi \log(\lambda_{t-1}) \tag{5.12}$$

or a fixed intercept plus the fixed effect of mobility, i.e.

$$\log(\lambda_t) = \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1}) \tag{5.13}$$

or a random intercept plus the random effect of mobility, i.e.

$$\log(\lambda_t) = \beta_t^0 + \beta_t^m + \phi \log(\lambda_{t-1}) \tag{5.14}$$

where m_t is the first PC score of mobility at time t, while β_t^m is the random effect of mobility at time t. For this random effect, we used Normal distributions centered at the first PC scores of the mobility variables with standard deviations the standard errors obtained by Bootstrap, i.e. we run moving-blocks Bootstrap (see Efron and Tibshirani, 1994[31]) with a fixed length of blocks to estimate many times the principal component scores so that we calculate their standard deviations. The fixed and random intercepts β_0 and β_t^0 as well as the fixed effect coefficient β_1 received a $N(0, 5^2)$ prior. Moreover, we followed a different approach to calculate R_t. We assumed that τ is generated by a $Gamma(a, b)$ distribution, which we fitted by Maximum Likelihood on the recovery, deaths and removal data we constructed as described earlier, while the optimization search over the (a, b) was performed using the Broyden – Fletcher – Goldfarb – Shanno (BFGS) algorithm. These Gamma distributions were then used to generate a random τ for each iteration of λ_t. Thus, we added extra variation due to randomness of the infectious period, as opposed to the fixed one we had before. For all these models the fitted cases were fine, but the R_t values were extremely large at the beginning of the epidemic. Another model we tried had both a time-varying intercept and slope inside the log-lambda equation, i.e. $\log(\lambda_t) = \beta_t^0 + \beta_t m_t + \phi \log(\lambda_{t-1})$, but the sampler found many problems fitting it. Information about models (5.10)-(5.14) can be found in Table 5.3 for data up to mid May 2020.

The idea of time-varying covariates was interesting in the sense that both the fit and the R_t values were sensible, so we also tried another approach in which the log-linear λ depended on a fixed or random effect of mobility and an intercept that changes in a "soft" way. That is, we wrote lambda as

$$\begin{aligned} \log(\lambda_t) &= \beta_t^0 + \beta_1 X_t + \phi \log(\lambda_{t-1}) \\ \beta_t^0 &= p \cdot c_t^1 + (1-p) \cdot c_t^2 \\ p &= \frac{1}{1+\exp(-T)} \end{aligned} \tag{5.15}$$

where T is the time of the change-point and c_t^1 and c_t^2 are Gaussian $N(0, 5^2)$ variables, but we could not get trustful results due to divergent iterations of NUTS. Even worse results we obtained by making both beta coefficients to be time-varying, i.e. having $\log(\lambda_t) = \beta_t^0 + \beta_t^1 X_t + \phi\eta_{t-1}$, or even $\log(\lambda_t) = \beta_t^0 + \beta_t^m + \phi \log(\lambda_{t-1})$. There were many divergences, as well as a high percentage of iterations that needed a larger tree depth for NUTS. Somewhat better results were obtained by assuming the same form for λ as in (5.15), but writing p as

$$\begin{aligned} p &= \frac{1}{1+\exp(-K)} \\ K &= 2 \cdot (t-T) \end{aligned} \tag{5.16}$$

which worked better as a smoothing mechanism. The fit was nice and the R_t estimates were much lower than before even at the beginning of the epidemic, although they remained too large to be true (see Table 5.3 for a comparison with previous models). After including data for the whole May 2020, we refitted this last model and compared it with one to which we added to the $\log(\lambda_t)$ expression the lagged first PC of the Google mobility variables up to order 7, i.e. we wrote $\log(\lambda_t) = \beta_t^0 + \sum_{i=1}^{7} \beta_i X_{t-i} + \phi \log(\lambda_{t-1})$ to incorporate information from the last 7 days to the infection rate parameter lambda. The results were very similar and the estimates were a little smoother before, but generally good as far as the fit is concerned, but the same problem of high initial R_t values remained still. The information criteria also preferred the previous one, so in Figure 5.14 we present the estimates of R_t and mean new cases.

After the procedure of estimating the infectious period τ from the data we constructed with the method described earlier (the first-infected, first-removed method), we

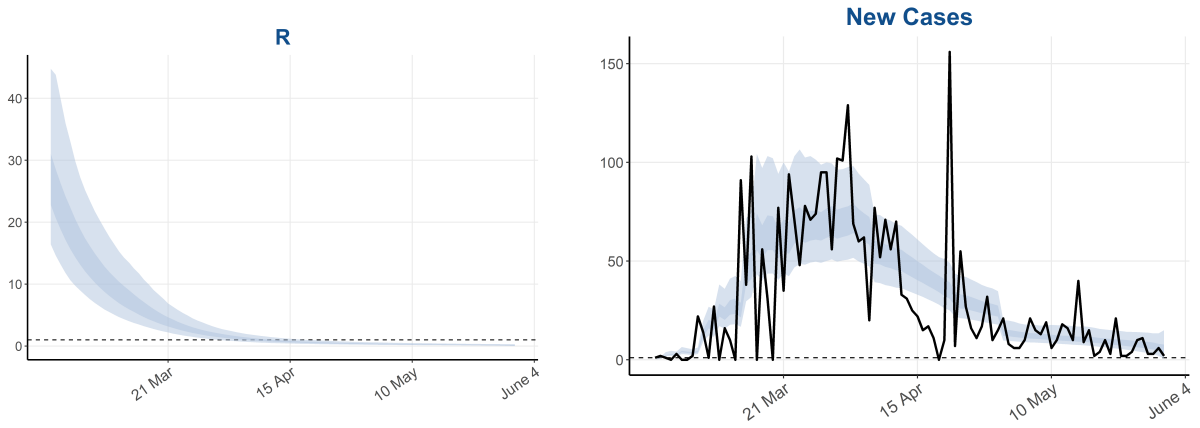

Figure 5.14: Estimates of R_t and mean new cases from the model with equation (5.16). R_t is much higher than it should be at the beginning and it decays exponentially, while the fit is fine. The dashed horizontal line on the left hand side plot is the $y = 1$ straight line below which the epidemic tends to disappear, while on the right it does not have such a meaning.

were pretty comfortable that the R_t values would be plausible and in agreement with the values that were published at that time, but this was not the case. We had many models that fitted the data well, none of which producing λ's that could be scaled by τ and reach to non-questionable R_t's. As it turned out, the recovery data have major problems and cannot be used in practice, making the statistical inference based on them impossible. However, the idea of estimating τ this way was sensible and it could work even as an approximation had the data be trustful. That was actually one of the reasons we abandoned this type of models some months later, since the recovery data played also a role in calculating I_t, the number of infected individuals at day t, which was simply $I_t = c_t^0 - d_t^0 - r_t^0$ (as described at the beginning of the Chapter).

Prediction using cases-based models

A sensible idea was to incorporate 7-days ahead predictions to our models. So, we had at our disposal a sample of size n, we fitted the model to the first $n_0 < n - 7$ days and make predictions about days $n_0 + 1, ..., n_0 + 7$. The first model we tried out was an SIR which had $\log(\lambda_t) = \beta_0 + \beta_1 X_t + \phi \log(\lambda_{t-1})$. In order to have a prediction, we assumed that $\log(\lambda_t) = \beta_0$ for $t = n_0 + 1, ..., n_0 + 7$. That is, we deleted the effect of the

Infection rate	AIC	BIC	DIC	DIC_2	WAIC	Time (min)
Equation (5.10)	562.364	569.434	585.016	3333.352	648.249	19.084
Equation (5.11)	555.883	565.310	563.463	2396.393	627.734	11.078
Equation (5.12)	523.327	532.753	526.458	2009.756	591.086	98.995
Equation (5.13)	706.103	717.887	697.488	701.035	701.802	4.117
Equation (5.14)	521.070	528.140	528.794	2145.928	591.497	169.498
Equation (5.16)	717.765	729.549	706.944	723.626	718.242	32.192

Table 5.3: Information criteria and training time for models that use as infection rate equations (5.10)-(5.16). Time corresponds to 10^5 iterations after a 10^5 warm-up for 4 chains.

covariate entirely and we also assumed that the $S_{t-1}I_{t-1}$ term of the mean daily cases stays constant for the predicted days, i.e. $S_{t-1}I_{t-1} = S_{n_0}I_{n_0}$ for $t = n_0 + 1, ..., n_0 + 7$. The predictions in that first attempt were not good. The R_t estimates remained wrong because of the bad estimates of the infectious period and the predicted daily cases were bad because of the way we dealt with $\log(\lambda_t)$ at times $n_0 + 1, ..., n_0 + 7$.

Therefore, we began experimenting on which model could produce the best predictions as measured by the absolute difference between the predicted and actual cases one week ahead, i.e. we predicted the course of the epidemic 7 days ahead, we summed the number of cases at the end of that period and we compared this number with the actual corresponding cases. This procedure was run three times. Thus, we trained the model for times $1, ..., n_1$ and predicted for $n_1 + 1, ..., n_1 + 7$. We summed the median predicted number of daily cases θ_t for $n_1+1, ..., n_1+7$, call it $\theta^\star$, we summed the observed number of daily cases c_t for $n_1+1, ..., n_1+7$, call it $c^\star$, and then we calculated the quantity $|\theta^\star - c^\star|$. Then we repeated the procedure for times $1, ..., n_2$ for training and $n_2 + 1, ..., n_2 + 7$ for prediction and times $1, ..., n_3$ for training and $n_3 + 1, ..., n_3 + 7$ for prediction, where $n_1 < n_2 < n_3 < n - 7$. For the predictions, we assumed that the $S_{t-1}I_{t-1}$ term of the mean daily cases stays constant, while the same tactic was followed for the mobility effect (when the model included it for λ), i.e. $m_t = m_{n_0}$ for $t = n_0 + 1, ..., n_0 + 7$. Thus, we constructed the predictive distribution from which we sampled the new 7 days. For the

infectious period τ, we assumed that it is generated by a Gamma distribution with shape 2.6 and rate 0.4, so that its mean is 6.5 days, so we scaled each iteration of λ_t with a random draw from this distribution to obtain R_t. The information criteria, training time and predictive ability for the models are shown in Table 5.4. The first one we tried was an SIR with an AR(1) term on λ, but no covariates (model X1 in the Table), i.e. $\log(\lambda_t) = \beta_0 + \phi \log(\lambda_{t-1})$. This model gave nice and smooth R_t estimates, only a little too high at the beginning of the epidemic (started at around 5 and decayed exponentially) and the fitted and predicted new cases were also plausible for all three periods we used as the training set. Thus, the next step was to add the mobility covariate on λ (model X2 in the Table). This resulted to a model that did not offer more in terms of the fit, the predictive ability was worse (using the criterion we mentioned before) and, furthermore depending on the number of days the fitting was done, the coefficient of mobility could be non significant. A third model was one which used as covariates the mobility variable lagged by 1 up to 7 days, instead of using the mobility of the same day (model X3 in the Table). So the only change compared with the previous one was that the new $\log(\lambda)$ depended on 7 covariates plus an intercept in order to capture dependencies due to week effects (however, the coefficients were not significant). The next model used for λ an intercept plus a random slope of mobility in the same manner as we did in previous versions, that is by using the PC score of a day as the mean of a Normal distribution with variance the one produced by Bootstrap (model X4 in the Table). The models that used the mobility effect produced too large credible intervals and not so good predictions as the simple first one. This is why we returned to the first one (with no covariates) and tried the SEIR version of it, i.e. using the lagged S_t as described earlier and the results were similar (model X5 in the Table). Lastly, we tried both an SIR (model X6 in the Table) and an SEIR (model X7 in the Table) using the "soft" change-point idea with the inverse logit proportions of two intercepts (see equation (5.16)) and no covariates on λ. Although the information criteria were close to the one without the change-point, the new fits did not seem good by visual inspection and the absolute differences of the summed cases at the end of the week we wanted the models to predict were larger than the first simple SIR. Generally, we can see from Table 5.4 that models X1 and X5 perform the best, which are the ones with just an AR(1) term on λ and no covariates. As far as

R_t is concerned, it began at around 5 and decayed exponentially, which was a progress compared to previous estimates. As a summary, in Table 5.4 we present the following. X1: SIR, AR(1) on λ, no covariates; X2: SIR, AR(1) on λ, with $\beta_0 + \beta_1 m_t$; X3: SIR, AR(1) on λ, with lagged mobility; X4: SIR, AR(1) on λ, with mobility uncertainty; X5: SEIR, AR(1) on λ, no covariates; X6: SIR, AR(1) on λ, no covariates, change-point; X7: SEIR, AR(1) on λ, no covariates, change-point.

Since we had some models that seemed to work well (either SIR or SEIR), we started experimenting on fitting the SEIR version to a number W of European countries. The way we did that was to assume the Covid-19 cases of the countries under study were generated by a Negative Binomial distribution with a mean whose quantities changed from one country to another. The likelihood of the $W = 4$ chosen countries was

$$L = \prod_{i=1}^{W} \prod_{t=1}^{n_i} NB(c_{t,i}; \theta_{t,i}, \psi_i)$$

where n_i is the number of days country i suffers from Covid-19. The mean observed cases of country i were written as $\theta_{t,i} = \lambda_{t,i} S_{t-1,i} I_{t-1,i}/N_i$, while the infection rate was $\log(\lambda_{t,i}) = \beta_{0,i} + \phi_i \log(\lambda_{t-1,i})$. The training was made only until 25 April 2020 and the predictions were performed for 7 days after that. The countries we chose were Greece, Italy, Germany and Sweden. The results were mostly satisfying, except of Germany, which had major problems with the fitted cases and the R_t estimates.

Model	Training End	AIC	BIC	DIC	DIC_2	WAIC	Time (min)	$c^\star$	$\theta^\star$	$\lvert\theta^\star - c^\star\rvert$
	25/04	541.656	550.033	535.085	537.688	538.538	4.150	103	112	9
X1	15/05	693.596	703.124	686.966	689.357	690.150	5.779	55	52	3
	02/06	801.598	811.937	794.844	797.408	798.186	4.875	121	52	69
	5/04	546.376	556.848	541.642	563.734	545.912	24.222	103	186	83
X2	15/05	693.935	705.845	685.312	688.854	689.943	5.194	55	85	30
	02/06	804.008	816.933	795.344	798.888	800.051	7.265	121	36	85
	25/04	568.112	591.150	546.072	581.751	558.584	1.219	103	585	482
X3	15/05	709.732	735.934	689.406	700.888	702.802	2.411	55	168	113
	02/06	820.157	848.592	799.463	809.139	811.876	4.324	121	51	70
	25/04	561.091	695.129	442.930	2003.706	494.600	13.176	103	134	31
X4	15/05	730.735	930.825	583.132	3441.738	648.466	21.138	55	306.5	251.5
	02/06	880.818	1144.485	772.767	9845.694	843.596	1.418	121	740.5	619.5
	25/04	541.756	550.133	535.100	537.637	538.719	3.871	103	113	10
X5	15/05	693.689	703.217	687.016	689.446	690.320	4.015	55	53	2
	02/06	801.834	811.834	794.893	797.454	798.079	4.790	121	31	90
	25/04	550.977	561.449	543.878	569.515	548.694	41.930	103	203	100
X6	15/05	694.798	706.708	686.309	690.818	690.542	11.603	55	71	16
	02/06	804.148	817.073	795.419	799.472	799.840	16.684	121	42	79
	25/04	550.764	561.235	543.795	568.837	548.823	72.147	103	203	100
X7	15/05	694.935	706.845	686.351	690.858	690.608	17.466	55	73	18
	02/06	804.088	817.013	795.512	799.507	799.796	28.336	121	42	79

Table 5.4: Results for models X1-X7. Training begins at 26/02/2020 and only the last day is indicated.

Testing period	Greece	Italy	Germany	Sweden	Cyprus	Finland	Netherlands	UK
26/05-02/05	0.068	0.439	3.979	0.075	0.064	0.223	0.849	3.156
03/05-09/05	0.452	0.502	0.027	0.055	0.200	0.197	0.424	0.329
10/05-16/05	0.563	0.243	0.088	0.066	0.188	0.306	0.284	0.604
17/05-23/05	0.238	0.108	0.376	0.049	0.364	0.100	0.043	1.064
24/05-30/05	0.162	0.020	0.354	0.281	0.222	0.348	0.265	0.474
31/05-06/06	0.492	0.053	0.244	0.587	1.000	0.990	0.283	0.598
07/06-13/06	0.713	0.323	0.221	0.420	1.000	0.368	0.443	0.511
14/06-20/06	0.644	0.363	0.682	0.377	2.500	1.368	0.111	0.306
21/06-27/06	0.410	0.620	0.560	0.185	0.125	0.509	0.175	0.235

Table 5.5: All the dates correspond to year 2020. The values correspond to the quantity $|\theta^\star - c^\star|$ we used in previous models to check their predictive ability.

After that initial try we included more countries and, also the models were fit for many different periods and each time we predicted the next 7 days. The countries were Greece, Italy, Germany, Sweden, Cyprus, Finland, Netherlands and United Kingdom (UK). The dates the model was trained until were 25/4/2020, 2/5/2020, 9/5/2020, 16/5/2020, 23/5/2020, 30/5/2020, 6/6/2020, 13/6/2020 and 20/6/2020, that is we fit it for the first time until 25/4/2020 and then we progressively add seven more days each time. Furthermore, because the absolute difference between observed and predicted cases cannot be used to make comparisons among different countries, the way we evaluated the goodness of predictions was to also divide by the number of observed cases in order for the resulted number to be on the same scale for every country. Thus, numbers close to zero suggest good predictions, while away from it suggest poor predictions. The goodness of the fit for each country was not something that depended only on the chosen model, but also on the data we used for training. For instance, when we used the data until 2/5/2020, Germany and United Kingdom had a bad fit while, when we used the data until 20/6/2020, the fit for all countries was quite good. However, in Germany, Italy and UK there was a slight shift of the fitted cases to the right, which raised our concerns. The predictive ability of this SEIR model is shown in Table 5.5. Each of the nine training procedures took between two and five hours to complete with the usual algorithmic settings of NUTS.

Introduction of deaths-based likelihoods

Deaths data are generally more trustful, since a death from Covid-19 is always registered, in contrast with just an infection. Thus, the idea of including deaths data into the likelihood seemed promising. We built a separate SEIR model for Greece, Italy and Germany with the following likelihood modification

$$L = \prod_{t=1}^{n} NB(c_t; \theta_t^{(c)}, \psi^{(c)}) \prod_{t=1}^{n} NB(d_t; \theta_t^{(d)}, \psi^{(d)})$$

where the superscript (c) refers to cases and the superscript (d) refers to deaths. The two Negative Binomial means were correlated by the fact that we wrote the mean number of deaths at day t as $\theta_t^{(d)} = \sum_{j=1}^{t-1} \pi_{t-j} \cdot \theta_j^{(c)}$. The mean new cases were once again $\theta_t^{(c)} =$

$\lambda_t S_{t-1} I_{t-1}/N$, where the infection rate λ was simply $\log(\lambda_t) = \beta_0 + \phi \log(\lambda_{t-1})$, which was shown to work well in the past. Both ψ's received the usual half-Cauchy prior while, for the first 14 days, θ_t was left to be a free half-Normal parameter. What this model expresses is our belief that daily deaths are generated from the daily cases and, specifically there is a probability π_1 that they are due to yesterday's cases, some probability π_2 that they are due to cases of two days before and so on. The proportion at day t, π_t, is calculated as $\int_{s-0.5}^{s+0.5} \pi(t)dt$, where $\pi(\cdot)$ is the density of a $Gamma(7.18, 0.3)$ with mean 23.9 days as in Flaxman et al., 2020[35]. There were underestimation problems for the fitted cases and the fitted deaths were too large. A small modification, which helped lower the number of fitted deaths, but not much for them to be sensible, was an AR(1) term on the mean estimated deaths, i.e. we wrote $\theta_t^{(d)} = \sum_{j=1}^{t-1} \pi_{t-j}\theta_j^{(c)} + \delta\theta_{t-1}^{(d)}$. R_t took generally more sensible values after stopping the estimation of τ from the data and decayed exponentially fast from around 5 to very low positive values.

As it turned out, the wrong mean deaths were due to the fact that we had not accounted for the IFR p, that is the proportion of total cases (not only the observed ones) that end up in death. Thus, our next SEIR model (the S values correspond to the lagged S) was

$$\begin{aligned} c_t &\sim NB(\theta_t^{(c)}, \psi_1) \\ d_t &\sim NB(\theta_t^{(d)}, \psi_2) \\ \theta_t^{(c)} &= \lambda_t S_{t-1} I_{t-1}/N \\ \log(\lambda_t) &= \beta_0 + \phi \log(\lambda_{t-1}) \\ \theta_t^{(d)} &= \sum_{j=1}^{t-1} p \cdot \pi_{t-j}\theta_j^{(c)} \end{aligned} \tag{5.17}$$

So after scaling $\pi(\cdot)$ by p, which was 1.12% for Greece at that time (see Sypsa et al., 2020[98], the estimates were more reasonable, but not entirely correct. Note that normally p scales the total cases, while the equation of $\theta_t^{(d)}$ posits it on the mean observed ones. There was an overestimation on the fitted cases and a shift and underestimation on the fitted deaths. Adding an AR(1) term on the mean deaths, i.e. $\theta_t^{(d)} = \sum_{j=1}^{t-1} p \cdot \pi_{t-j}\theta_j^{(c)} + \delta\theta_{t-1}^{(d)}$, the fitted cases were fine and the deaths became more plausible only shifted to the right.

R_t retained its exponentially decaying form starting at around 5. The first values for $\theta_t^{(d)}$ were explicitly written because of the AR(1) term: $\theta_1^{(d)}$ was left to be a free parameter, while $\theta_2^{(d)} = p \cdot \pi_1 \cdot \theta_1^{(c)}$. Another idea was to let p take random values from a half-$Normal(0, 5^2)$ distribution and refit the two previous models, but the problems remained unfixed. However, the infection fatality ratio was now a component of our models.

An important aspect of our final models is that they are capable of estimating the total number of cases (both observed and unobserved ones). So far, there was no such part in the models that could give those estimates. The first try was when we added to model (5.17) that

$$C_t = R_t \sum_{i=1}^{t-1} f_{t-i} C_i \tag{5.18}$$

where C_t are the total cases at day t, the first k C_t's are set to be equal and given an Exponential prior and f_{t-i} are probabilities given by the serial interval distribution, which is the distribution of time from the infection of an individual until they transmit it to someone else. The intuition is that day after day new cases are generated according to the serial interval and the cases of day t are determined by the "ease" of infection expressed by R_t. Although the fitted deaths were greatly improved, the fitted cases and R_t estimates were not satisfactory. When an AR(1) term was added to the mean new deaths, everything (even R_t, which began at approximately 2) except of the estimated total cases were plausible, but further tuning was needed. Those two models were tested with both a fixed IFR equal to 1.12% and an IFR from a $Normal(0.01, 0.1^2)$ prior. The latter version gave even worse total cases estimates (they were extremely small).

One different try of estimating the total cases was changing (5.18) with $C_t = \theta_t^{(c)} + x_t$, where x_t are the asymptomatic and infected but not tested people, i.e. those that carry the disease, but are not confirmed cases and are given a vague half-$Normal(0, 100^2)$ prior. Then, fitted deaths were calculated as $\theta_t^{(d)} = p \cdot C_t$. The fitted cases had major underestimation problems from 25th of July till the end of the training period (25/8/2020), when the confirmed cases started to grow, but this was corrected to some extent by fitting a different intercept for $\log(\lambda)$ (and the AR(1) coefficient) at that time period, i.e. we had $\log(\lambda_t) = \beta_{01} + \phi_1 \log(\lambda_{t-1})$ for the first part and $\log(\lambda_t) = \beta_{02} + \phi_2 \log(\lambda_{t-1})$ for the second one. The estimated total cases, as well as the fitted deaths were also nice. We

also tried to add a second change-point on this intercept at the beginning of the epidemic just before the first wave, but there was no benefit whatsoever; the reasoning was that there was a different level of fear for the new virus before and after the lockdown. Going back to the first SEIR with AR(1) and no change-point, we added the mobility covariate and the effect was the same as adding the first change-point on the intercept, i.e. it helped the fitted cases. Furthermore, R_t estimates of all these models start on high values (approximately equal to 15) and smoothly decrease. The last model was the only capable of producing an R_t, which after dropping below 1 during summer, it raised back over 1 when confirmed cases increased, which was a characteristic we were pleased to see.

Adding one change-point on the intercept when the mobility variable was already in the model did not make a big difference, while adding the second one ruined the fitted cases. The last try was an SEIR model that had a change-point on both λ and p on the same date, so it estimated two intercepts and two slopes for the log-linear form of λ (we included the mobility variable), two AR(1) coefficients as well as two different IFR's. It seemed that we had nicely estimated parameters and plausible results, except of the first values of R_t, which were still too high to be true. Generally, it was a subject under study whether we should add change-points on parts of λ and whether we need covariates for it or not. The reason was that as the epidemic was progressing old people were more cautious, while young people got tired of the quarantine measures, started to get out more and so, most infected people were too young to end up dead. Thus, IFR should be lower than it used to be before summer. Regarding the infectious period, we changed the $Gamma(2.6, 0.4)$ we used until now to a $Gamma(4.93, 0.26)$, which was the same as the distribution of time between onset of symptoms and death (this actually created overestimation problems of R_t, since this distribution has a larger mean).

The previously described models had a $U(0, 1)$ prior on p, which could not be estimated by the data (by construction), so a more specific prior had to be given since this would contain all the information for the final estimates. After all, we did know that the IFR for Greece at that time was approximately 1.12%, so an informative prior with almost all its mass on that point was acceptable. However, the previous models which had not incorporated this information gave an estimated mean IFR equal to 0.011. The

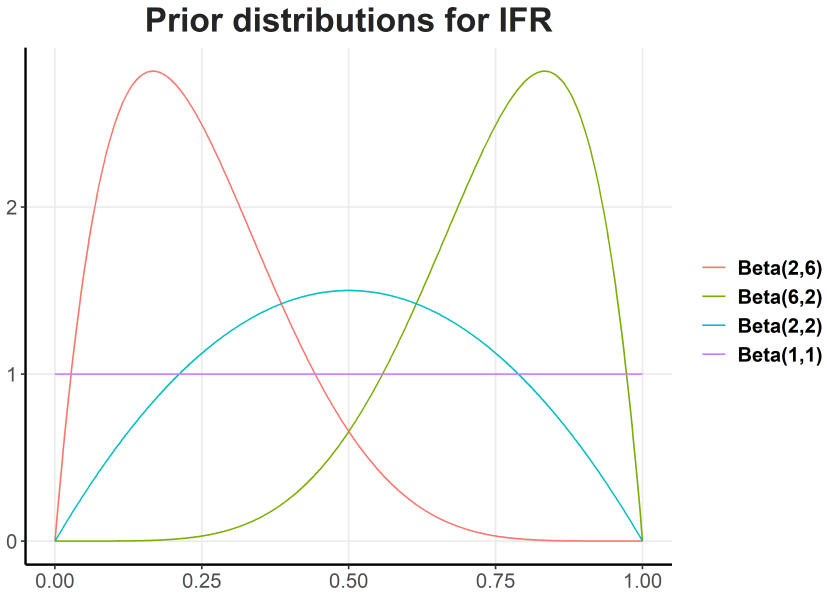

Figure 5.15: Prior distributions used for the sensitivity analysis of IFR.

last one which had two IFR's due to the change-point had a mean 0.0124 for the first and 0.0079 for the second one. This was very strange and we performed a sensitivity analysis to discover how, when no information was given for that parameter, the model did estimate it perfectly. We tested four Beta distributions, namely $Beta(2,6)$, $Beta(6,2)$, $Beta(2,2)$ and $Beta(1,1)$ (Figure 5.15) and for each case we used as initial values either the mean of the corresponding prior, or 0.0112 (Table 5.6). The results showed that indeed the estimated IFR was exactly what it should be, therefore we continued using the $U(0,1)$ prior.

Distribution, initial value	mean	mean se	sd	2.5%	25%	50%	75%	97.5%
Beta(2,6), prior mean	0.011	1.7e-05	0.0010	0.0089	0.0100	0.011	0.011	0.013
Beta(6,2), prior mean	0.011	1.8e-05	0.0011	0.0092	0.0100	0.011	0.012	0.013
Beta(2,2), prior mean	0.011	1.8e-05	0.0010	0.0088	0.0100	0.011	0.011	0.013
Uniform(0,1), prior mean	0.011	1.7e-05	0.0010	0.0088	0.0099	0.011	0.011	0.013
Beta(2,6), 1.12%	0.011	1.8e-05	0.0010	0.0089	0.0100	0.011	0.011	0.013
Beta(6,2), 1.12%	0.011	1.7e-05	0.0011	0.0092	0.0100	0.011	0.012	0.013
Beta(2,2), 1.12%	0.011	1.8e-05	0.0010	0.0089	0.0100	0.011	0.011	0.013
Uniform(0,1), 1.12%	0.011	1.9e-05	0.0010	0.0087	0.0099	0.011	0.011	0.013

Table 5.6: Sensitivity analysis performed for the estimation of IFR. The "mean" and "sd" column indicate the mean and standard deviation of the HMC estimates respectively, while "mean se" indicates the standard error of the "mean".

Something else that needed to be investigated was the way we introduce x_t (the unobserved cases at day t) into the model. Until now, we used the two equations

$$\begin{aligned} \theta_t^{(c)} &= \lambda_t S_{t-1} I_{t-1}/N \\ C_t &= \theta_t^{(c)} + x_t \end{aligned} \tag{5.19}$$

but this implied that we had deviate from the SIR/SEIR model in some sense. In SIR type models the $\lambda \cdot S \cdot I$ term is applied on the infected people, not on the observed infected people. We needed a way to put the $\lambda \cdot S \cdot I$ term on C_t. Thus, it made sense to check the behaviour of the algorithm when expressing the problem as

$$C_t = \lambda_t S_{t-1} I_{t-1}/N \qquad \theta_t^{(c)} = \lambda_t S_{t-1} I_{t-1}/N - x_t \qquad C_t = \lambda_t S_{t-1} I_{t-1}/N$$

or or

$$\theta_t^{(c)} = C_t - x_t \qquad C_t = \theta_t^{(c)} + x_t \qquad \theta_t^{(c)} = \omega \cdot C_t$$

and leaving the fitted deaths to be $\theta_t^{(d)} = p \cdot C_t$.

The first two cases were difficult to fit (the sampler got into many problems), so we used the third representation, which assumes that the observed cases are a proportion ω (with a $U(0,1)$ prior) of the total cases C_t. We only tested the SEIR type, i.e. we used the lagged susceptible population. The simple infection rate $\log(\lambda_t) = \beta_0 + \phi \log(\lambda_{t-1})$ was used (which usually provided nice estimates), but the results were not satisfactory enough. Adding a change-point at 20/07/2020 – that is before cases started to grow for the second wave in Greece – on both p and ω or only on λ did not provide any improvement. This was performed because of underestimation problems during the second wave. Furthermore, we started estimating the marginal likelihood of the models by Bridge Sampling for model comparison (see Gronau et al., 2020[41]). A comparison among the last three models with the ω representation lies in Table 5.7, where we also included three models with the previous (5.19) representation. For all these models, we did not include any covariate effect for $\log(\lambda)$ and the investigation regarded the use of change-points. It seemed that the faster (5.19) models were worse in terms of information criteria.

Change-point	AIC	BIC	DIC	DIC_2	WAIC	Evidence	Time (min)
-	2685.235	2708.665	2672.304	2677.283	2676.943	-1359.223	18.343
p & ω	2627.078	2657.202	2610.474	2618.169	2617.615	-1336.118	32.443
λ	2647.337	2677.461	2630.704	2637.775	2637.592	-1342.959	27.490
-	3094.915	3817.890	2662.048	2765.794	2708.257	-1521.742	1.885
p	3097.453	3823.776	2663.658	2771.936	2710.306	-1381.588	2.079
λ	3026.521	3756.191	2590.274	2691.081	2640.358	-1346.966	2.075

Table 5.7: Information criteria for the newly tested models. For the first three, the SEIR infection rate equation affects the total number of cases and the observed ones are just a proportion ω of them, while the last three regard the (5.19) representation. Evidence is in log-scale.

Next, we included the addition of the mobility covariate on $\log(\lambda)$, while we also kept the change-point on both p and ω, i.e.

$$\begin{aligned}
\theta_t^{(c)} &= \Big(\omega_1 \cdot I\big(t \in (1, l-1)\big) + \omega_2 \cdot I\big(t \in (l, n)\big)\Big) \cdot C_t \\
\theta_t^{(d)} &= \Big(p_1 \cdot I\big(t \in (1, l-1)\big) + p_2 \cdot I\big(t \in (l, n)\big)\Big) \cdot C_t \\
C_t &= \lambda_t S_{t-1} I_{t-1}/N \\
\log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1})
\end{aligned}$$

where l is the time of change for both p and ω (20/07/2020). All IFR and ω parameters received $U(0,1)$ prior distributions (the way we wrote IFR as $\Big(p_1 \cdot I\big(t \in (1, l-1)\big) + p_2 \cdot I\big(t \in (l, n)\big)\Big)$ is actually the way we always do when we mention a change-point on it, but here we just wrote it explicitly). Another novelty was that we started estimating samples from the predictive distribution not only to decide on the predictive ability, but to recreate the whole epidemic process and decide on the fit at least visually. The reasoning is that should the model be appropriate for the data, the new generations of samples from it will look like the original series. R_t estimates were not good initially (as they always did so far) and there was a slight underestimation of the fitted deaths for the first days. Apart from those two problems, the model was performing good.

Calculating the mean deaths by IFR times the total cases was not entirely correct, since deaths of day t were produced by cases that occurred some days before t. The distribution on those days was a sum of Gamma distributions, as was used by Flaxman et al., 2020[35]. Note that we have used this idea in previous versions with a distribution that "looked like" the one in that same article; this time we made a more specific simulation of the two $Gamma(1.35, 0.26)$ and $Gamma(4.93, 0.26)$ distributions mentioned and then discretized the result the way we have mentioned previously in the text. This way, we could use information of the infection-to-death distribution to scale each day's cases by a probability and obtain a much more accurate mean deaths estimate. Therefore, we used as mean deaths

$$\theta_t^{(d)} = p \cdot \sum_{i=1}^{t-1} \pi_{t-i} \cdot C_i$$

where π_{t-i} is given by the discretized infection-to-death distribution, i.e.

$$\pi_s = \int_{s-0.5}^{s+0.5} \pi(t)dt \quad \text{, for} \quad s \geq 2$$

and

$$\pi_1 = \int_0^{1.5} \pi(t)dt$$

We also investigated whether any improvement could be made by adding the Apple mobility data to the PCA procedure of obtaining the mobility variable m_t. The fit of all of the four models was similar and the information criteria were relatively close with a slight preference to the one using both Google and Apple without the probability adjustment (see Table 5.8). However, the difference between adjusting for deaths due to previous cases and calculating deaths only according to today's cases (without the infection-to-death probabilities) was that the former are actually closer to reality. Also, the fitted deaths had a nicer smooth form, but a slight shift to the right as well. The models were fitted for the first n_0 days and predicted the last 7 as described earlier in the text and the results are shown in Table 5.9 in terms of absolute difference and squared error loss. The numbers are not intuitive by themselves, but they can be used to compare across different models (the smaller, the better). Furthermore, we did an analogous procedure to provide a goodness of fit measure using the predictive distribution we used earlier only for a visual inspection, that is we took samples from it and generated the same $\theta^{(c)\star}$ and $\theta^{(d)\star}$ only this time these quantities corresponded to the predicted total observed cases and total deaths at the end of the training period. Thus, we can have a crude measure of the general fit: If the fitted distribution produces samples that regenerate the series closely to the original one, then the fit is good. The problem is that this technique focuses more on the final cumulative results and not on everything in between. Also, the measure of "closeness" is by itself restricted by the loss function one uses. If however the loss is used as a relative and not an absolute measure of goodness of fit/prediction, the method is fine. In Table 5.9 we use the same symbols $\theta^{(c)\star}$ and $\theta^{(d)\star}$ for the actual predictions as well. From now on when we refer to mobility effect, we mean that PCA is performed on both Google and Apple data.

Model	AIC	BIC	DIC	DIC_2	WAIC	Evidence	Time (min)
Google	2671.318	2705.024	2652.631	2659.958	2663.046	-1362.889	6.437
Google adj	2708.490	2742.197	2689.714	2697.687	2699.773	-1381.552	18.169
Google & Apple	2669.342	2703.049	2650.525	2657.471	2660.621	-1361.875	28.601
Google & Apple adj	2710.275	2743.982	2691.487	2699.709	2702.012	-1382.455	16.850

Table 5.8: Information criteria after adding the mobility variable m_t. The comparison is about using only Google mobility or Google and Apple combined, as well as using the adjustment π_{t-i} for fitted deaths. Evidence is in log-scale.

Model	$\lvert\theta^{(c)\star} - c^\star\rvert$	$\lvert\theta^{(d)\star} - d^\star\rvert$	$(\theta^{(c)\star} - c^\star)^2$	$(\theta^{(d)\star} - d^\star)^2$
Google	1.5	3	2.25	9
Google adj	116	11	13456	121
Google & Apple	181	3	32761	9
Google & Apple adj	74	11	5476	121
Google	1009.5	34	1019090	1156
Google adj	805	25	648025	625
Google & Apple	920	34	846400	1156
Google & Apple adj	729	22	531441	484

Table 5.9: Predictions and goodness of fit for models with the information criteria of Table 5.8 in terms of absolute difference and squared error loss. The quantities $\theta^{(c)\star}$ and $\theta^{(d)\star}$ correspond to $\theta^\star$ (which has been mentioned previously in the text), only the superscripts (c) and (d) denote mean observed cases and deaths respectively. $d^\star$ is the deaths analogous quantity of $c^\star$ described for previous models. The first four cases regard the actual predictions of each model (7 days ahead), while the last four regard the predictions at the end of the training period.

Next, we returned to the model with the same change-point on IFR and ω, which we present here for clarity:

$$\begin{aligned} \theta_t^{(c)} &= \omega \cdot C_t \\ \theta_t^{(d)} &= p \cdot \sum_{i=1}^{t-1} \pi_{t-i} \cdot C_i \\ C_t &= \lambda_t S_{t-1} I_{t-1}/N \\ \log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1}) \end{aligned} \tag{5.20}$$

where p and ω both change at 20/07/2020. We refitted it (including September 2020) on five periods and each time predicted the next 7 days the way we have explained with good results and also a nice fit on the observed cases. The fitted deaths presented a small shift to the right during the first wave, but this was corrected by adding two change-points on λ at 01/05/2020 and 05/06/2020; we decided on these two dates because the period in between is characterized by very few cases in Greece. However, the two change-points

on λ made R_t irrational. We also changed the way we calculated R_t from scaling it with random draws from the time distribution from symptom onset until death to draws from a $Gamma(5,1)$. The initial R_t values were now good, but it still did not rise over 1 when we knew that this was the case during the second wave. Then, we refitted model (5.20) on the same five periods but corrected the $Gamma(5,1)$ with a $Gamma(7,1)$ distribution (so that the mean infectious period is 7). Furthermore, we changed the prior of p, which was the same $U(0,1)$ distribution for each of the two periods separated from the change-point (as it was the case for ω), to a $N(0.01142, 0.001^2)$ until 04/07/2020 and $N(0.00802, 0.001^2)$ from 05/07/2020 to 11/10/2020. This was the first model that uses prior information of IFR from the age distribution of Covid-19 cases in Greece published from the National Public Health Organization (NPHO) (see Chapter 3 for its calculation). This change on the IFR priors resulted in more estimated total cases than before (with an approximate maximum median of 1400 as opposed to 700 in the previous case) and we were more confident about these estimates because of the more formal way of deriving the priors. We used this model as reference to compare with some other modified versions. The first modification was the addition of two change-points on λ, i.e.

$$\begin{aligned}\log(\lambda_t) = {} & I\big(t \in (1, u_1 - 1)\big)\beta_{01} + I\big(t \in (u_1, u_2 - 1)\big)\beta_{02} + I\big(t \in (u_2, n)\big)\beta_{03} \\ & + \Big(I\big(t \in (1, u_1 - 1)\big)\beta_{11} + I\big(t \in (u_1, u_2 - 1)\big)\beta_{12} + I\big(t \in (u_2, n)\big)\beta_{13}\Big) m_t \\ & + \Big(I\big(t \in (1, u_1 - 1)\big)\phi_1 + I\big(t \in (u_1, u_2 - 1)\big)\phi_2 + I\big(t \in (u_2, n)\big)\phi_3\Big) \log(\lambda_{t-1})\end{aligned} \tag{5.21}$$

We set u_1 to be the order of date 01/05/2020 and u_2 of date 05/06/2020. The day when p and ω changed was also set to 01/05/2020. From the above, we got some nice results, the R_t estimates were not so good (especially at the beginning of the epidemic), but the fitted deaths were nicer than before. The predictive ability and the goodness of fit results derived from the predictive distribution were similar for the two models, but the information criteria preferred the one with (5.21) over that of (5.20) (these comparisons were performed only for the fifth period of training of the models).

Discard of recorded cases

Deaths are more trustful data than cases, in the sense that we can better record them. Thus, it is reasonable to let those data guide the estimation process without the interference of the cases. Coming back to the reference SEIR model (5.20), we refit it without including the cases data (so there was no ω parameter either), but the results were really bad as far as R_t and the fitted deaths are concerned. Therefore, we continued with a comparison of two sets of models. The first set included a model of the form (5.20) with two change-points on λ (see equation (5.21)) and one change-point on p and ω (model X1 in Table 5.10), as well as another model whose only difference was that its likelihood did not take into account the cases data (model X2 in Table 5.10). The second set included the same models as those of the other, but it had two change-points on p and ω (models X3 and X4 in Table 5.10, where X4 is the only deaths data model). For p, we lowered the variance of the Normal distributions to 0.0001^2 and also the means to 0.0105 and 0.0075, because we thought the previous values were too pessimistic. When p had an extra change-point, we set its mean values to 0.0105, 0.007 and 0.008 to express our belief that the last period, which contained the second wave, the fatality of the virus was larger than that during the low-incidence second period. The times of change were 01/05/2020 and 05/06/2020 for both p, ω and λ (when one change-point was used for p and ω the time was 01/05/2020). For ω we used $U(0,1)$ priors, while τ was given a mixed prior $0.7 \cdot N(7,1) + 0.3 \cdot N(8.5,1)$ to add some skewness to the infectious period. That is, τ was approximately 7 (expressed by the $0.7 \cdot N(7,1)$ term) with a right tail (expressed by the $0.3 \cdot N(8.5,1)$ term). The resulted total cases were a little shifted to the left at the beginning of the series and the rest of the quantities of interest were similar for all these models. R_t was equally bad for all of them at the beginning of the epidemic, but the models with only deaths data had a better fit on deaths than those with both cases and deaths. Table 5.10 compares the four models in terms of their information criteria; note that we can only compare X1 with X3 and X2 with X4 due to the different likelihoods. We could conclude from these tries that introducing more change-points could not fix the inaccuracies we wished to fix.

Conducting a mini sensitivity analysis after including data until mid November 2020,

Model	AIC	BIC	DIC	DIC_2	WAIC	Time (min)
X1	3145.525	3201.149	3114.729	3126.504	3131.046	92.387
X2	759.990	805.184	735.620	744.817	742.744	14.174
X3	3148.882	211.458	3114.369	3128.201	3131.915	115.581
X4	761.945	810.615	735.574	746.223	742.871	143.629

Table 5.10: Models X1 and X3 include both cases and deaths data and they differ in that the former has one change-point on p and ω, while the latter has two. Models X2 and X4 are the same as X1 and X3, but they are based only on deaths data. Therefore, only X1-X3 and X2-X4 comparisons can be considered valid.

we lowered the IFR values at 0.9% and 0.6% (when one change-point is included), or 0.9%, 0.6% and 0.5% (when two change-points are included) and fitted three models with two change-points on λ: one with one change-point on p and ω, one with two change-points on p and ω and one which is the same as the first but not including the cases data. Also, we sampled the infectious period τ from a $Gamma(28, 4)$ (the mean is still 7) to increase the variance and check the results. Indeed, R_t did rise over 1 when it was supposed to do, but the rest of the results were similar with the previous models. One characteristic that became annoying as time passed was that during a few weeks there existed a cyclic pattern on the fitted and total cases which might indicate some seasonality, but this was actually due to false recording of the recoveries data, which we were using inside $S_{t-1}I_{t-1}$ (we did not know it then). Thus, the fitted and total cases seemed to behave somewhat irrationally. After a little experimentation on the IFR parameter, ω and λ regarding prior distributions and change-points, we present some results on Figure 5.16. We see the typical too large initial R_t, but after 27/03/2020 it seems to be plausible. The fitted cases and also the total ones (not shown here; they are just a scaled version of the fitted ones by ω) present some abrupt spikes which we did not understand at that time that are also shifted to the right of the spikes of the data. The fitted deaths are smooth and fit the trend of the series, but they only catch the general picture.

With the next set of models we tried to add flexibility to the infection rate λ by adding change-points on it, adding a second covariate, or using smoothed covariates.

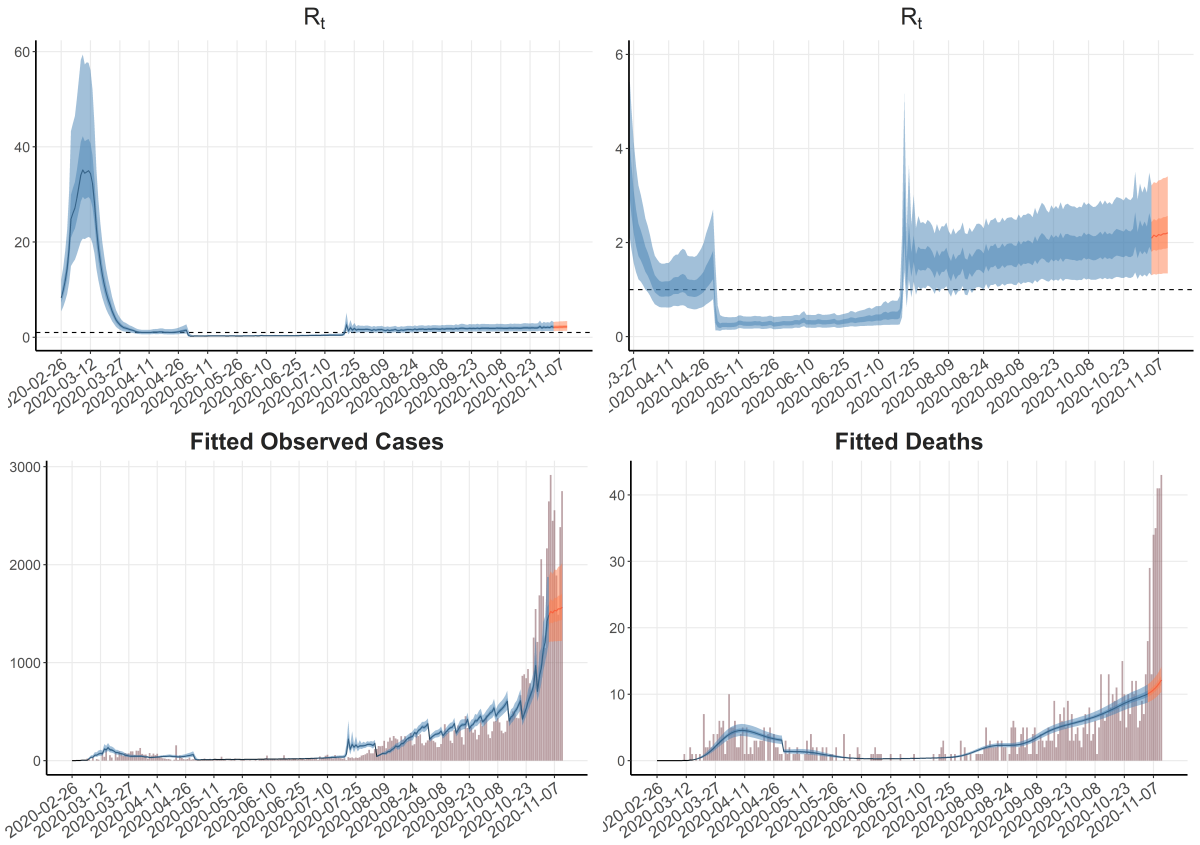

Figure 5.16: Top left: R_t estimates; Top right: R_t estimates zoomed-in after it drops for visualization purposes; Bottom left: fitted observed cases; Bottom right: fitted deaths; The orange color are predictions for the last week.

The model we were interested to was

$$\begin{aligned} \theta_t^{(c)} &= \omega \cdot C_t \\ \theta_t^{(d)} &= p \cdot \sum_{i=1}^{t-1} \pi_{t-i} \cdot C_i \\ C_t &= \lambda_t S_{t-1} I_{t-1}/N \end{aligned} \tag{5.22}$$

and λ_t was the subject under study. For these models we left the last week out of the training set to predict the course of the epidemic. The difference was that we obtained those predictions not by repeating the last value of m_t for $t = n_0 + 1, ..., n_0 + 7$, but we fitted a GAM model on the existing PC values and predicted the next 7 using this model. These predictions were then served as fixed covariates for the predictions of the epidemic model. First, we wanted to see if a third change-point on λ could help (model X3 in Table 5.11) and compare the resulted model with the 2-change-point one in equation

(5.21 (model X2 in Table 5.11) and finally compare both of them with a model with no change-points on λ at all (model X1 in Table 5.11), i.e. $\log(\lambda_t) = \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1})$. The 3-change-point model was written similarly to (5.21). The times of change for the 2-change-point model were 28/04/2020 and 25/07/2020, while for the 3-change-point one were 23/03/2020 (beginning of the first lockdown), 28/04/2020 (end of the first lockdown) and 20/07/2020. For p and ω we included one change-point at 28/04/2020 for the models with change-points on λ and one change-point at 04/07/2020 for the model with no change-points on λ. Next, we tried to incorporate information from the second PC of the mobility variables and we refitted the previous three models but with two covariates, instead of just one (models X4, X5, X6 respectively in Table 5.11). The third set of models was different from the second in that the two PC series were smoothed by the GAM model we used for the predictions, i.e. we kept the fitted values of the GAM model and used them in place of the original PC scores (models X7, X8 in Table 5.11). For this set we did not include a model with no change-points on λ (so it was comprised by two instead of three models). The fitted and total cases and fitted deaths were similar for all the aforementioned models with the difference of some spikes when a change-point was introduced. The problem with them was the underestimations that existed in general as well as the abrupt shifts that are also visible in Figure 5.16. R_t, which was the scaling of λ by a $Gamma(28, 4)$, was smooth only in the case of no change-points and also that was the case when it began from relatively low (compared to previous versions) values, approximately equal to 7 before dropping below 1 at 11/04/2020. The last set with the smoothed PC scores was the worse as far as the R_t series is concerned. The information criteria can be found in Table 5.11 for comparison and they seem to prefer the model with one PC on $\log(\lambda)$ and three change-points.

Another idea to tune model (5.22) with one change-point on p and no change-points on λ was to let ω change every day with a $U(0, 1)$ prior, which resulted in much flexibility in the fitted cases, but we still had the bad fitted deaths we also had before. For this model, we calculated the cumulative observed cases at the end of May, August, September and October 2020 and contrasted them with the estimated total cases produced by the model to get a sense of how much worse is the situation at the end of the selected dates

Model	AIC	BIC	DIC	DIC_2	WAIC	Time (h)
X1	3802.971	3838.500	3784.125	3791.504	3794.344	-
X2	3622.701	3683.101	3590.220	3602.638	3606.727	4.614
X3	3475.619	3542.753	3439.099	3458.898	3461.979	6.925
X4	3796.380	3835.462	3775.746	3784.217	3786.063	13.403
X5	3618.990	3690.049	3580.417	3596.561	3601.631	3.968
X6	3630.073	3711.792	3585.401	3613.257	3650.026	6.925
X7	3602.723	3673.782	3563.795	3581.182	3587.605	2.380
X8	3634.205	3715.923	3589.183	3607.235	3613.454	7.142

Table 5.11: Models X1 and X4 do not have any change-point on λ, models X2, X5 and X7 have 2 change-points on λ and models X3, X6 and X8 have 3 change-points on λ. The training time of model X1 was not recorded and we note it with a "-".

than that we observe (see Table 5.12). It seems that until the end of May and August we used to observe roughly one third of the total cases, while at the end of September and October (during the second wave) we used to observe approximately one forth.

The goal of the new formulation $C_t = \lambda_t S_{t-1} I_{t-1}/N$ instead of the older $c_t = \lambda_t S_{t-1} I_{t-1}/N$ was first to allow the estimation of the total cases (observed and unobserved) and, second to make the $\lambda \cdot S \cdot I$ term act on the hidden level of the total cases as the SIR-type models assume. However, the number of infected individuals at day t,

		Total Cases		
	Observed cases	Median	$95\%\, CrI$	$50\%\, CrI$
End of May	2917	9164	$(86364, 9751)$	$(7716, 10957)$
End of Aug	10317	28783	$(27238, 30483)$	$(24614, 34000)$
End of Sep	18475	74593	$(70781, 78818)$	$(63939, 87928)$
End of Oct	39251	151101	$(142523, 160462)$	$(127399, 180980)$

Table 5.12: Observed and total Covid-19 cases after May, August, September and October 2020 for Greece. The values are rounded to integers.

I_t, is being estimated by the observed cases as $I_t = c_t^0 - d_t^0 - r_t^0$, where c_t^0, d_t^0 and r_t^0 are the cumulative numbers of observed cases, deaths and recoveries respectively provided by Johns Hopkins University. Thus, in order to correct the model based on this idea, we had to get the recorded cases, as well as the recoveries out of the SI term.

Models based on a hidden level

The first successful model, which can act on the hidden total cases and estimate the number of susceptible and infected individuals (instead of calculating them by the problematic data as we have explained), was an SIR whose infection rate λ was piecewise constant and it was changing once every 30 days. The infection fatality ratio p had a change-point on 11/8/2020 (the data were updated until 10/1/2021), when the confirmed cases of younger people started to be more that those of people over 65 years old. Also, we excluded data of observed cases and the likelihood used only data of deaths. All in all, the model which was closer to the truth than any other previous version and also the basis for every other next version is the following:

$$\begin{aligned} d_t &\sim NB(\theta_t, \psi) \\ \theta_t &= \Big(p_{(1)} \cdot I\big(t \in (1, l-1)\big) + p_{(2)} \cdot I\big(t \in (l, n)\big)\Big) \cdot \sum_{i=1}^{t-1} \pi_{t-i} C_i \\ C_t &= \lambda_t S_{t-1} I_{t-1} / N \\ S_t &= S_{t-1} - C_t \\ I_t &= \sum_{i=t-6}^{t} C_i \end{aligned} \tag{5.23}$$

where for day t, $\lambda_t = \lambda_{(j+1)}$ with $j \in \{0, ..., J-1\}$ being the corresponding period between change-point u_j and $u_{j+1} - 1$ and J is the number of constant parts (note that u_0 and u_J are not an actual change-points, but the first and last time points). We set $C_t = C_1$ for $t = 1, ..., 10$, $S_1 = N - C_1$ and $\theta_1 = 1$. l corresponds to the time index of 11/8/2020 when we assume p changes. The R_t estimates (estimated by multiplying λ_t by 7, i.e. we assume a constant infectious period of seven days) and the fit on the daily deaths are shown in the left and right hand side of Figure 5.17 respectively. Both of them are plausible and satisfactory and the estimated total cases were also good. Lastly, we

set the following priors: $\lambda_j \sim LN(0,1)$, for $j = 1, ..., 10$ (at that time we had $J = 11$, i.e. λ_t had 11 distinct constant parts), $\psi \sim Gamma(2, 0.125)$, $C_1 \sim Gamma(2, 0.0625)$, $p_1 \sim Normal(0.011, 0.0001^2)$, $p_2 \sim Normal(0.009, 0.0001^2)$. The model run with no problems and the sampler needed 2.5 days to train it. The same model was also tested by assuming that λ changes every 2 weeks (instead of 30 days) and the results were also good and plausible (model X1 in Table 5.13). One new problem that came along with our new models was the training time, which was escalated from hours to days, since basing the inference on something we do not observe is a much more complicated problem. Before we continue, we clarify that IFR from now on has always the form

$$p_t = \begin{cases} p_{(1)} & \text{for } t = l_0, ..., l_1 - 1 \\ p_{(2)} & \text{for } t = l_1, ..., l_2 - 1 \\ \vdots & \\ p_{(B)} & \text{for } t = l_{B-1}, ..., l_B - 1 \end{cases}$$
$$= p_{(b+1)} \cdot I\big(t \in (l_b, l_{b+1} - 1)\big)\,,\ b = 0, ..., B-1$$

where $l_0 = 2$, $l_B = n$, $p_n = p_{n-1}$ and each $p_{(b)}$ receives a $N(\cdot, 0.0001^2)$ prior distribution centered at a value we explicitly state (this is actually the same way we model λ). We no longer experiment on other distributions on this parameter, only with the number of change-points and the mean values that should be used. We make two final notes here. First, from now on we use many change-points on these two parameters (especially for λ), so a parameter with subscript (i) indicates the i'th constant value of that parameter, while i (without the parentheses) indicates the value of the parameter at time i. Lastly, from now on when we state that $p = 0.001$ (0.001 is a random value), we mean that the Normal prior is centered at 0.001.

Setting $\log(\lambda_t) = \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1})$ (model X2 in Table 5.13) was not better in terms of information criteria and it was also an overfit to the data. Removing the AR(1) coefficient for the first days (which are more difficult to fit) also did not help (model X3 in Table 5.13). However, the coefficient β_1 was statistically significant (0 was outside the 95% CrI). Another similar idea was to exclude the mobility effect, but have a different intercept and AR(1) coefficient in the $\log(\lambda)$ equation every 4 weeks (model X4 in Table 5.13). The fit was not so bad as before (although there was one slight shift during the

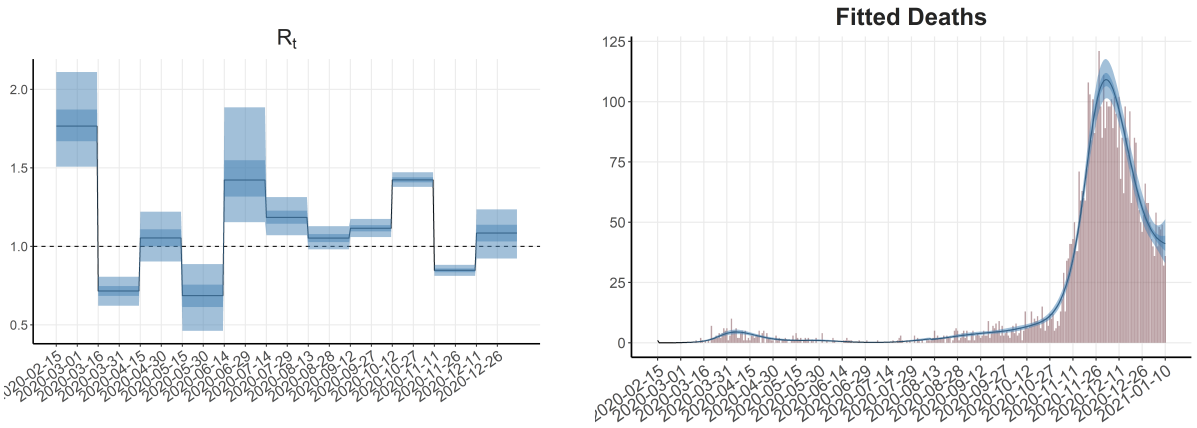

Figure 5.17: R_t and fitted deaths produced by model (5.23).

Model	AIC	BIC	DIC	DIC_2	WAIC	Evidence	Time (*days*)
X1	1586.646	1703.642	1527.862	1556.528	1549.638	−863.169	3.185
X2	1901.264	1928.563	1888.520	1893.392	1892.700	−972.704	2.602
X3	1932.200	1959.499	1919.718	1926.965	1923.522	−987.566	3.926
X4	1631.394	1701.592	1596.727	1613.369	1612.089	−863.169	4.017

Table 5.13: Information criteria for the trained models. Evidence is in log-scale.

second wave), the information criteria suggested this one over the one λ per day version, but still the model with the piecewise constant λ and no mobility effect was better. Also, R_t presented a strange oscillation of decaying magnitude inside each interval of constant β_0 and ϕ.

It seemed like there were nine different cases that we needed to investigate. The models we called type A had λ as a free piecewise constant parameter (of this type was model (5.23) and also model X1 in Table 5.13), i.e.

$$\lambda_t = \lambda_{(j+1)} \cdot I\big(t \in (u_j, u_{j+1} - 1)\big)\,,\ j = 0, ..., J-1$$

Models of type B also had a piecewise constant intercept in the $\log(\lambda)$ equation plus an AR(1) term (like model X4 in Table 5.13), i.e.

$$\log(\lambda_t) = \big(\beta_{0,j+1} + \phi_{j+1}\log(\lambda_{t-1})\big) \cdot I\big(t \in (u_j, u_{j+1} - 1)\big)\,,\ j = 0, ..., J-1$$

and models of type C were the same as B only they also included the mobility for $\log(\lambda)$

(so we had piecewise constant set of betas), i.e.

$$\log(\lambda_t) = \big(\beta_{0,j+1} + \beta_{1,j+1} m_t + \phi_{j+1} \log(\lambda_{t-1})\big) \cdot I\big(t \in (u_j, u_{j+1} - 1)\big)\,,\ j = 0, ..., J-1$$

For each model type we wanted to test three cases: piecewise constant parameters that change every 10, 20, or 30 days and indexed as 1, 2 and 3 respectively. According to that definition, model (5.23) is of type A3, model X1 in Table 5.13 is of type A2 and model X4 in Table 5.13 is of type B3. Thus, until then, the information criteria suggested the A3 model, that is the one with no $\log(\lambda)$ equation (so "A-") and λ that changes every 30 days (so "-3"), which was also the fastest one to train (it needed 2.5 days).

One small change that we made as we moved on to tuning the models was to set the total number of cases for the first seven days to be equal (as opposed to the first 10 days we had before) and, rather than having λ pieces of 10, 20, or 30 days, we switched it to 14, 21, or 28 (i.e. 2, 3, or 4 weeks). These modifications were made due to the fact that changes happen in multiples of number of weeks, not in tens of days. For instance, this way we could catch better week effects due to weekends, when people go out more. Also, we put an $Exp(1)$ prior on the mean number of deaths at day 1, θ_1, which until then was set to 1.

HMC for models of type C needed over three weeks to be completed for 4 chains, 5000 iterations per chain and 15000 warm-up (which were the usual settings), so we never got results from these models. Reducing the iterations imposed problems with the algorithm, so we decided to leave the mobility effect out of the models we investigate. After including data up to April 2021, we compared our models using the information criteria shown in Table 5.14 as well as visually. Although model B3 is preferred by the information criteria, the unexplainable noise in estimates as well as the high training time led us to tend towards type A models. Moreover, more λ pieces resulted in more training time, so models B2 and B1 would take even longer and the results had a high chance of being as bad as those of B3, thus we did not proceed with their training. Model A1 was flexible enough to give good R_t estimates and it gave the best DIC. One disadvantage was the large training time. A2 and A3 were similar and their training time was not deterrent like that of A1 or B3. Furthermore, since we included more data, one more IFR change-point was needed at 01/04/2021, since the fatality was larger than that during

	AIC	BIC	DIC	DIC_2	WAIC	Evidence	Time (*days*)
Model A1	1981.87	2126.45	1911.15	1945.79	1936.91	−1073.64	8.3
Model A2	1963.47	2067.89	1912.81	1944.26	1932.86	−1054.57	2.2
Model A3	1963.93	2048.27	1923.22	1955.28	1938.66	−1050.89	1.5
Model B3	1959.73	2048.09	1917.24	1946.14	1932.41	−1039.70	6.8
Model D1	7039.74	7140.15	6990.95	7016.40	7023.35	−3594.70	0.7
Model D2	6999.56	7120.04	6940.83	6971.53	6974.14	−3588.68	0.8
Model D3	7001.71	7162.36	6923.12	6966.47	6970.55	−3612.57	0.8

Table 5.14: Information criteria for the models trained. Models A and B are not comparable with D, since they are based on different likelihoods. Evidence is in log-scale.

the second wave (partly because of a new variant) and the prior for the last period was set to be $N(0.011, 0.0001^2)$.

One new class of models we started experimenting on was type D, which was the same as type A with the only difference of including the observed cases data in the likelihood. So models D were

$$\begin{aligned} d_t &\sim NB(\theta_t^{(d)}, \psi_1) \\ c_t &\sim NB(\theta_t^{(c)}, \psi_2) \\ \theta_t^{(d)} &= p \cdot \sum_{i=1}^{t-1} \pi_{t-i} C_i \\ \theta_t^{(c)} &= \omega \cdot C_t \\ C_t &= \lambda_t S_{t-1} I_{t-1}/N \\ S_t &= S_{t-1} - C_t \\ I_t &= \sum_{i=t-6}^{t} C_i \end{aligned}$$

where ω was a piecewise constant function of time with change-points at the same times as p and each of those three parts received a $U(0,1)$ prior distribution. The dispersion parameter ψ_2 received the same prior as ψ_1. The information criteria of models D1, D2 and D3 are shown in Table 5.14 (again 1, 2, 3 means λ changes every 2, 3 and 4 weeks respectively). They were all similar and gave plausible results, with the largest difference

	Epidemic period		Change-points	Mean values
	From	To		
Greece	26/02/2020	30/06/2021	28/07/2020, 01/11/2020, 09/02/2021, 09/05/2021	0.01142, 0.007994, 0.01142, 0.0115, 0.008
Portugal	02/03/2020	30/06/2021	01/06/2020, 16/11/2020, 01/05/2021	0.011576650, 0.008165473, 0.011576650, 0.009
Germany	27/01/2020	30/06/2021	27/5/2020, 25/11/2020, 24/2/2021	0.011039996, 0.005434144, 0.011420979, 0.007
UK	31/01/2020	30/06/2021	18/07/2020, 01/10/2020, 01/03/2021	0.01035, 0.007245, 0.0095, 0.007
Norway	26/02/2020	30/06/2021	13/05/2020, 23/12/2020, 21/04/2021	0.0091, 0.006, 0.007, 0.004
Sweden	26/02/2020	30/06/2021	16/07/2020, 02/10/2020, 01/04/2021	0.0103, 0.007, 0.009, 0.007

Table 5.15: Change-points for the models of each of the six chosen countries and the means of their Normal distributions; the standard deviations are all equal to 0.0001.

being at the beginning of R_t. Model D1 estimated it approximately equal to 5, model D2 around 2.5 and model D3 around 1.5. So D2 was closer to the truth and was also a balanced solution between the many-changes model D1 and few-changes model D3. Furthermore, in terms of AIC and BIC D2 was the best among the three and it was close to the others as far as the other criteria are concerned. Lastly, the training time of type D models was considerably smaller (approximately 0.8 days) than the models that included only deaths data.

One last modification to the aforementioned D models was to set lambda to be constant for the first four weeks and then change as we described before (every 3 or 2 weeks). The models with this modification are denoted with a ".1" at the end, so the two new models are D2.1 and D1.1 respectively. We fitted the two models with this modified λ and the R_t estimates at the beginning became more plausible as we wished to be the

case (it was around 1.5-1.6 and it dropped during the first quarantine in Greece), since at the beginning there is not much information.

Then, we tried modeling data from Portugal, Germany, UK, Norway and Sweden using models of type A and D. Model D1.1 for Portugal gave good results, but the fitted deaths were a little shifted during the first wave (spring 2020) and in January 2021. We continued with Germany and UK and we faced a problem in the fitted deaths of UK, which were shifted during the first wave. On the contrary, the model of type A.1 was fitted just fine on the UK data with no problems and plausible results. Furthermore, the estimated total cases were close to those estimated by the REACT2 study (more on this subject in Chapter 3). Norway had the characteristic that many days had recorded number of deaths equal to zero, but the next day jumped to a high value. This indicated a problematic recording system and it also created many artificial zeros in our data. However, the fit was fine. Interestingly, model A produced worse estimates, in contrast with the UK case we talked about before. Model A1.1 seemed to perform well for Sweden. We added data up to 30/6/2023 and present the change-points for the six countries in Table 5.15. In Figure 5.18 we present the R_t estimates after fitting the A1.1 model to each country.

Some changes needed to be made regarding the proportion of cases that we observe. This proportion is calculated by dividing the actual observed cases by the median total ones (as estimated by the model) and must lie inside the $(0, 1)$ interval. Actually, the smoothed version of it must be between 0 and 1, since the daily percentage is subject to too much noise. The smoothing technique we use is a local regression model with polynomials of degree 2 and a span of 0.75. For Greece and Germany there was no problem with the smoothed percentage, but the other chosen countries needed some tuning, since their percentages were above 1 (see for example Figure 5.19).

For instance, UK gave very large estimates in the last interval of IFR and this was an indication that perhaps the p parameter should be lowered. The data were a little problematic due to a period of 0's in the recorded deaths followed by a huge jump, which we corrected by evenly splitting the last value to that interval. Then, after lowering the IFR in the intervals that the observed proportion exceeded 1, the results were sensible.

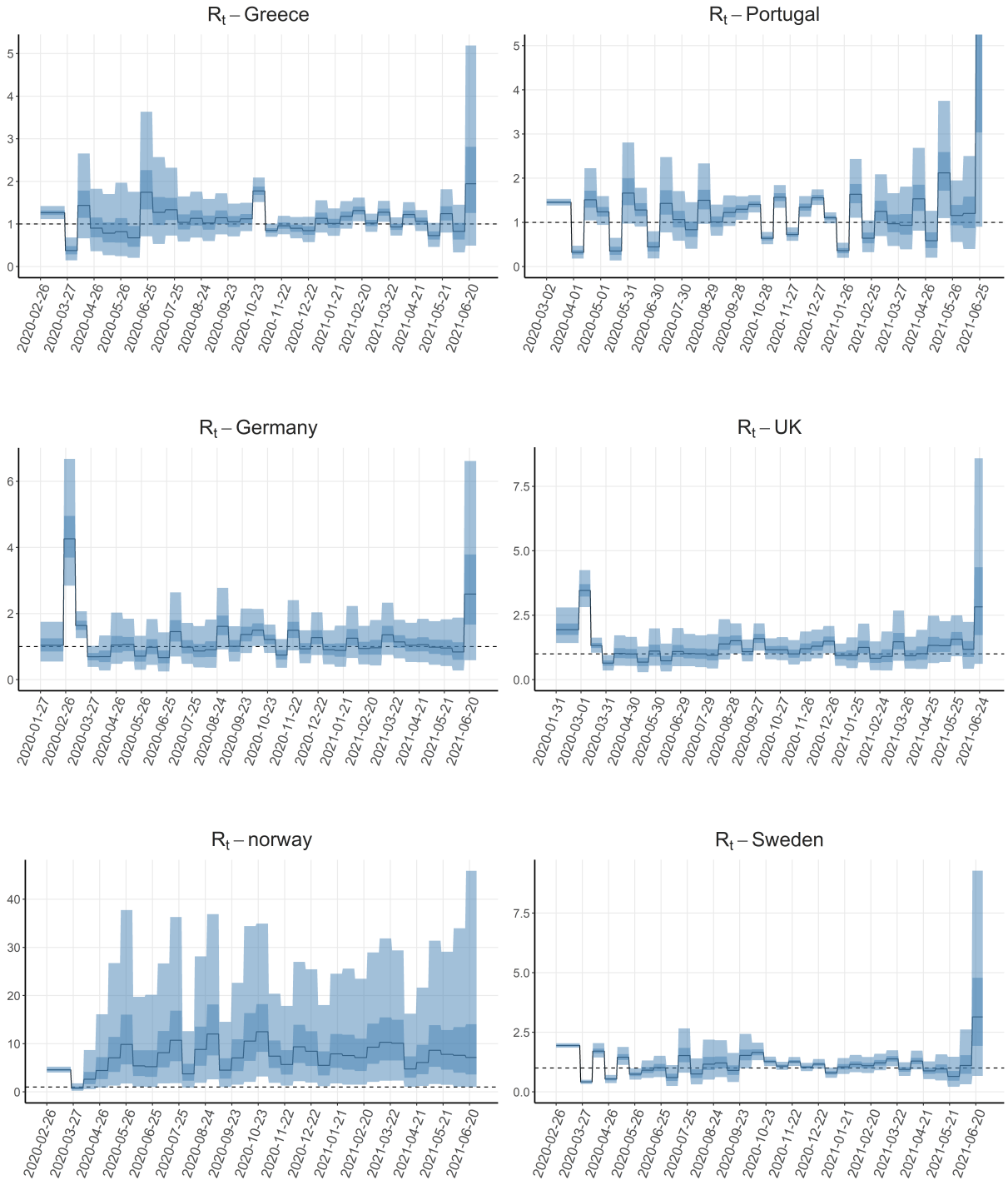

Figure 5.18: R_t estimates for each of the six chosen countries produced by model A1.1.

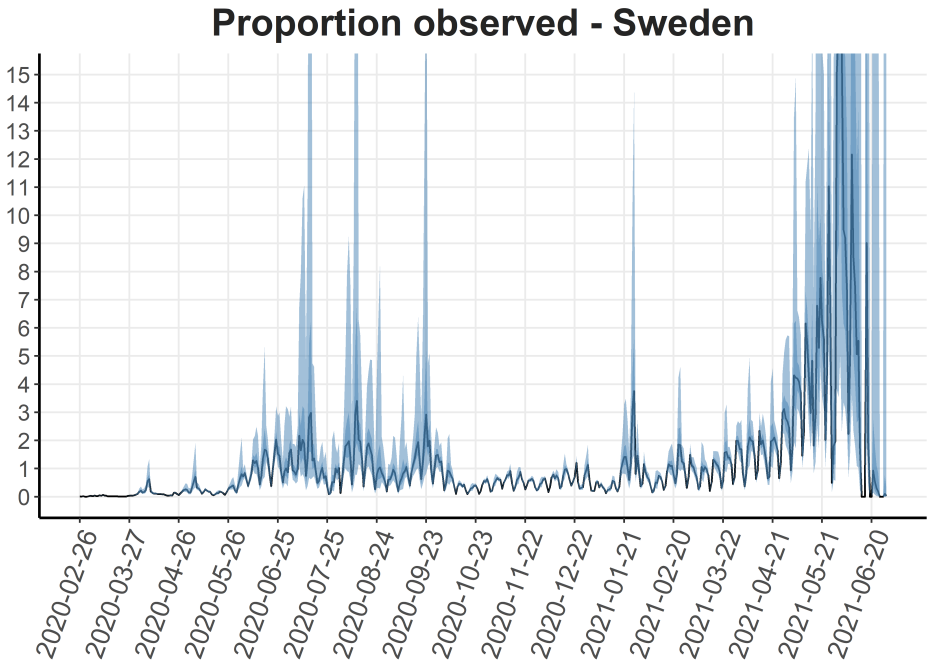

Figure 5.19: Proportion of cases observed out of the total ones for Sweden. The clusters indicate the location of the IFR change-points and, based on previous values, the adjustment of their values as well.

The procedure of IFR adjustment was repeated for the other countries as well, by locating clusters in the plotted proportion and adjusting the time of change-points and their values.

The SEIR model

Introduction of the Exposed period

After having fixed the SIR model for the chosen countries, we decided to extend our ideas to the SEIR version we had already used in the past. In order to do that, we had to include an Exposed period of length h during which the infected individuals are not infectious. Thus, we replaced equation $C_t = \lambda_{t-1} S_{t-1} I_{t-1}/N$ by $C_t = \lambda_{t-1-h} S_{t-1-h} I_{t-1-h}/N$ where we used $h = 2$. Now, the new total cases are updated according to the situation of the

epidemic $h+1$ days before instead of 1 day before. The model became

$$\begin{aligned} d_t &\sim NB(\theta_t, \psi) \\ \theta_t &= p \cdot \sum_{i=1}^{t-1} \pi_{t-i} C_i \\ C_t &= \frac{\lambda_{t-1-h} S_{t-1-h} I_{t-1-h}}{N} \\ S_t &= S_{t-1} - C_t \\ I_t &= \sum_{i=t-\tau+1}^{t} C_i \end{aligned}$$

where (from now on) we set $\tau = 6$ and $h = 2$. The country we tested the new model was Greece and we found similar results with the SIR version. R_t had a little wider credible intervals, but the general picture was the same.

Inclusion of vaccination

As time passed and data about vaccinations became available, there arose the need for inclusion of them in the model, since vaccinations play an important role in every epidemic. We had data about the number of people that had their first dose of a vaccine (for the vaccinations data see Chapter 3) and we wanted to express the following idea: 14 days after vaccination, individuals got 68% probability of immunity against Coronavirus and three weeks later this probability was raised to 95%, because that is the time when a second dose is received. Thus, the way we incorporated the vaccinations data into our SEIR model was to set the Susceptible equation as

$$S_t = S_{t-1} - C_t - (0.68 \cdot v_{t-14} + 0.27 \cdot v_{t-35})$$

where v_t is the number of vaccinations at time t. For times before 14, the number vaccinations were 0, while between 15 and 35 we used $S_t = S_{t-1} - C_t - 0.68 \cdot v_{t-14}$. The scalar 0.27 is used because we wanted the total number of individuals that get out of the Susceptible state to be 95% of those vaccinated 35 days ago, so we subtract 68% two weeks after vaccination and the rest 27% five weeks after vaccination, since $27 = 95 - 68$.

Inclusion of covariates and demography

We wanted to extend the idea of the basic SEIR model presented in the previous subsection with the inclusion of the mobility covariates that we had tried in the past, so we tried to include the first PC of mobility by writing either $\log(\lambda_t) = \beta_0 + \beta_1 m_t$, or $\log(\lambda_t) = \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1})$, both of which produced bad results with R_t being between 1 and 1.2 at all times and the credible intervals being extremely narrow. Also, neither the total cases nor the fitted deaths seemed rational.

Another way to include the mobility effect was to keep the piecewise idea of the infection rate and apply it on the coefficients of the log-linear equation, i.e. $\log(\lambda_t) = \beta_{0i} + \beta_{1i} m_t$ for every interval i we previously had just λ_i as a free parameter.Since the estimates were not trustful yet, we further added an AR(1) term to smooth the series, so lambda became $\log(\lambda_t) = \beta_{0i} + \beta_{1i} m_t + \phi \log(\lambda_{t-1})$. The information criteria were really close (but pointed to the one with the AR(1) term). Lastly, we tried to let ϕ change along with the $\boldsymbol{\beta}$ coefficients, which was also proved insufficient to produce reasonable estimates (there were too many parameters to be estimated).

Another way we wrote λ was to include the effect of past observed deaths in an Autoregressive Moving Average (ARMA)-like representation, i.e. $\log(\lambda_t) = \beta_0 + \beta_1 m_t + \phi \log(\lambda_{t-1}) + \delta d_{t-1}$. Unfortunately, these parameters could not be estimated and we stopped including the mobility effect in the λ equation.

Another new aspect for our models was the introduction of the effect of demography, assuming that the death and birth rates are both equal to A. Thus, the susceptible equation became $S_t = S_{t-1} - C_t - V_t + A \cdot (N - S_{t-1})$ where the term $A \cdot N$ accounts for the new births, while the term $A \cdot S_{t-1}$ accounts for deaths. The infectious equation became $I_t = \sum_{i=t-\tau+1}^{t} C_i + A \cdot I_{t-1}$, where we added the last term to account for physical deaths inside the active set. V_t is the individuals we remove from the susceptible due to vaccination the way we described previously.

Extension to bivariate likelihood models

An extension of the SEIR model presented above is the incorporation of the cases data in the likelihood. That is we can model the vector (c_t, d_t), instead of just d_t. This was considered in the past (we referred to it as type D model), but now we come back to it by also including vaccinations, demography and also the number of tests as a covariate of the proportion of the total cases that are eventually recorded. Therefore, we have

$$c_t = \omega_t \cdot C_t$$
$$\omega_t = \frac{1}{1 + exp(b_0 + b_1 T_{t-4})}$$

where T_t is the number of tests at time t. Although the effective sample size was low, the fitted deaths seemed good, the proportion of observed cases seemed plausible and the R_t was a little high for every interval, so generally the model seemed promising. We refer to this model as the bivariate SEIR from now on. We let the simple SEIR model to be our default and we refer to this model explicitly as the bivariate one when needed. We also tried to replace the inverse-logit representation of the proportion ω with free $U(0, 1)$ parameters with similar results. Then, we also added a lag to the generation of the observed cases based on the fact that for the first 1-2 days an infected person has no symptoms, then they suffer the symptoms for the next 1-2 days, then they get tested and the next day they obtain the positive result. So we set $\theta_t^{(c)} = \omega \cdot C_{t-6}$.

Extension to hierarchical models

In order to set a more compact structure for the total cases C_t and absorb some noise of the estimates, we inserted a new level of hierarchy into our model, by assuming that $C_t \sim Gamma(a_t, b_t)$, where $\frac{a_t}{b_t} = \frac{\lambda_{t-1-h} S_{t-1-h} I_{t-1-h}}{N}$ and $\frac{a_t}{b_t^2} = 400$, i.e. a mean calculated as C_t in the previous version and a constant variance of 400. Due to slow computation time, exceeded tree-depth algorithm problems, low effective sample size and R-hat between 1 and 2 for the estimated parameters, we stopped experimenting on a new hierarchy for the model.

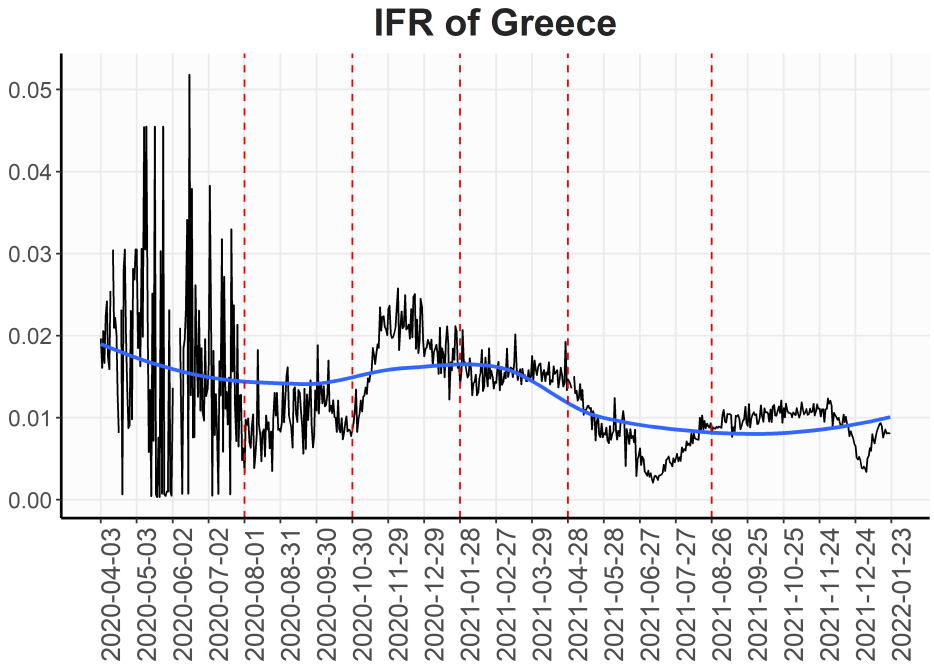

Figure 5.20: The IFR series of Greece. The black line is the estimate we obtain from the data and the blue one is a smoothed version by local regression. The vertical lines indicate the times of change.

Including delta and omicron variation data

When we decided to add more data to train our models and the new ending time point was in mid January 2022 (thus including the delta variation in Greece), we needed to revisit the IFR elicitation. Following the same procedure as we did in the past, we obtained the series in Figure 5.20 and put the IFR change-points where the vertical lines lie with values at each interval 0.0115, 0.008, 0.0115, 0.0114, 0.007 and 0.008. Another change that we made to our models was making the last period of constant λ to be four weeks long in order to reduce the credibility reduction of the estimates, due to the distribution of infection-to-death which suggests that deaths are more probably generated 20 days after infection. So for the last 2 weeks of training there was not much information for our estimates.

We made two changes for the SEIR model: first, we started removing 40% 14 days after vaccination and 10% more 35 days after vaccination, because the old 95% propor-

tion led to high λ estimates (since it decreased S_t more than it should). Furthermore, the birth rate A was not quite right using the reciprocal of the average lifetime, because this corresponded to 30.060 births per day, in contrast to the Hellenic Statistical Authority (HSA) data which suggested approximately 287.63 births per day. This estimate was calculated as follows. The data from HSA mentioned 1049839 births per year for the age group 0-9, which corresponds to 104983,9 births for the first year approximately. Therefore, dividing with 365 we get approximately 287.63 births per day, which corresponds to $A = 287.63$. Then, we added A to the susceptible group and we subtracted A (due to deaths) from both the susceptible and infectious groups with weights S_t/N and I_t/N respectively.

As a final step to the SEIR model, we started experimenting on the value of the infectious period τ and the IFR change-points, fine tuning its mean values and the times of change for Greece, Portugal, UK. We decided to let the bivariate model aside, since we had no available data to tune the proportion of observed cases and constrain the fitting procedure (there were non-identifiability issues). The final SEIR models are presented in Chapter 3, where a formal description and more details are given.

The SEIRS model

As a final step to the Covid-19 models we wanted to introduce, we constructed a model that allows individuals to be re-infected. We assumed that after seven months from the beginning of the epidemic every individual that has been infected, but not died returns to the Susceptible state. However, there was the wrong assumption that someone who survives becomes immediately susceptible.

Thus, the change we made was based on the following reasoning. The total cases that are prone to death at each time t is given by the term $\sum_{k=1}^{t-1} \pi_{t-k} C_k$ as was the case for the SEIR model. From these cases, only a proportion p_t ends up in death, so the rest $1 - p_t$ recover. Thus, we can construct the variable

$$r_t = (1 - p_t) \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k$$

and update the number of susceptible individuals by

$$S_t = S_{t-1} - C_t - V_t + A \cdot (1 - S_{t-1}/N) + r_{t-4\cdot 7\cdot 7} \tag{5.24}$$

where the lag $4 \cdot 7 \cdot 7$ on r_t brings back to susceptibility cases that occurred 28 weeks ago. Of course, r_t is zero for the first 28 weeks. Unfortunately, the results were not as promising as they were before.

The last idea for the way a SEIRS model could be constructed was to estimate the recovered individuals by $r_t = (1 - p_t) \cdot \sum_{k=1}^{t-1} \pi^*_{t-k} \cdot C_k$, where π^*_{t-k} is the discretized Gamma distribution of time from infection until recovery estimated in Paul and Lorin, 2021[81]. Then, we can move the recovered back to susceptibility using equation (5.24). Finally, we changed the estimation of R_t, multiplying with the proportion of the remaining susceptible people, i.e. $R_t = \lambda_t \tau \frac{S_t}{N}$, for both the SEIR and the SEIRS model.

Then, we tried to conduct some form of sensitivity analysis for the choice of the likelihood function, which until then was a Negative Binomial (some Poisson models had also been trained in the beginning of the research). The Negative Binomial distribution can be seen as a mixture distribution of a Poisson with rate following a Gamma distribution. Therefore, it is natural to think that other distributions can replace the prior of the Poisson rate as long as they are continuous with domain on $\mathbb{R}_+$, such as the Exponential or the LogNormal. The two distributions were trained to have mean rate equal to the mean of the previously default Negative Binomial. The former resulted in reduced training time for the model with equally plausible results, while the algorithm for the latter faced problems resulting in very long training time. A plot showing the difference between the energy transition density and the marginal energy estimated during HMC for the model with the LogNormal prior is depicted in Figure 5.21, where we can see that the target distribution is not explored as it should. Similar problems arose when we had added a prior on the total cases C_t, so it seems that the proposed models do not work well with complicated families placed in a hierarchy.

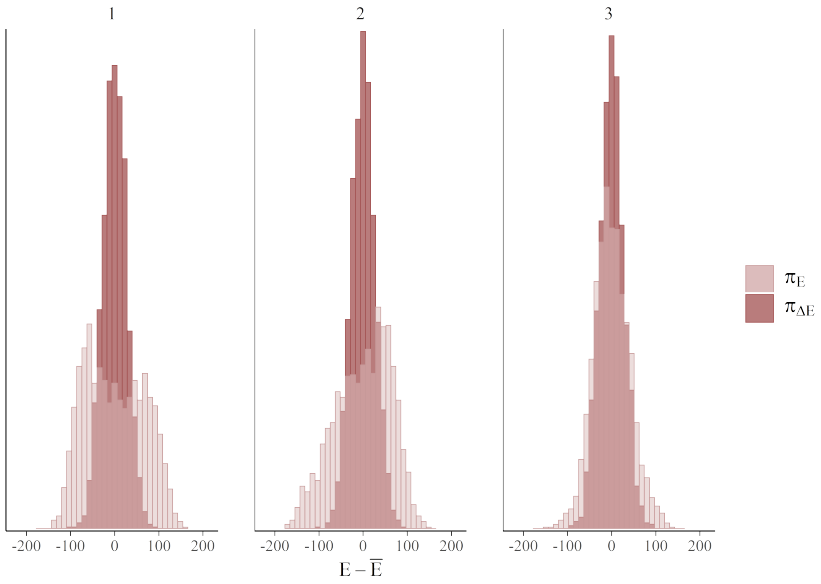

Figure 5.21: The differences between the energy transition density and the marginal energy estimated during HMC for each of three chains demonstrate that the algorithm run into problems for the Poisson with LogNormal prior model.

A4. Results on Chapter 3

Testing assumptions of the basic model

	AIC	BIC	DIC	DIC_2	WAIC	Time (d)
SIR	4034.75	4283.06	3924.75	3974.89	3980.82	2.67
SIR.vacc	4034.66	4282.97	3924.66	3976.76	3980.66	2.43
SIR.dem	4034.93	4283.24	3924.93	3978.20	3981.05	2.57
SIR.vacc.dem	4035.06	4283.37	3925.06	3977.96	3981.06	2.54
SEIR	4034.91	4283.22	3924.91	3976.59	3979.68	2.39
SEIR.vacc	4034.62	4282.93	3924.62	3974.38	3979.09	2.37
SEIR.dem	4034.52	4282.83	3924.52	3974.88	3978.61	2.50
SEIR.vacc.dem	4034.56	4282.87	3924.56	3975.72	3977.84	2.39

Table 5.16: Information criteria for the eight tested models and their training time measured in days. The results are rounded to 2 decimal digits.

	SIR	SIR.vacc	SIR.dem	SIR.vacc.dem	SEIR	SEIR.vacc	SEIR.dem	SEIR.vacc.dem
SIR	-2147.906							
SIR.vacc	2140.939	-2140.237						
SIR.dem	0.80493	0.000376	-2148.123					
SIR.vacc.dem	2190.752	1.023267	2721.667	-2140.214				
SEIR	9597373	4482.786	11923240	4380.859	-2131.829			
SEIR.vacc	35259470613	16469158	43804395361	16094691	3673.867	-2123.620		
SEIR.dem	13992052	6535.473	17382943	6386.873	1.457904	0.000397	-2131.452	
SEIR.vacc.dem	33910820866	15839225	42128908302	15479080	3533.344	0.961751	2423.577	-2123.659

Table 5.17: The log-marginal likelihood estimate using Bridge sampling for each model lies in the diagonal, while the pairwise Bayes factors lie below it with the model referred in the row in the nominator and the model referred in the column in the denominator. The evidence is rounded to 3 decimal digits and the Bayes factors to 6.

The SEIRS model for Greece

As an illustration of the SEIRS model, we present the results of fitting it on data of Greece until the end of 2022 (the data reach up to 20/12/2022). We set the IFR mean value for the whole 2022 to 0.0015 and the time of immunity after being infected to approximately 4 months ($4 \cdot 4 \cdot 7 = 112$ days). The fitted deaths, reproduction number, cumulative total cases and smoothed observed proportion of cases is shown in Figure 5.22.

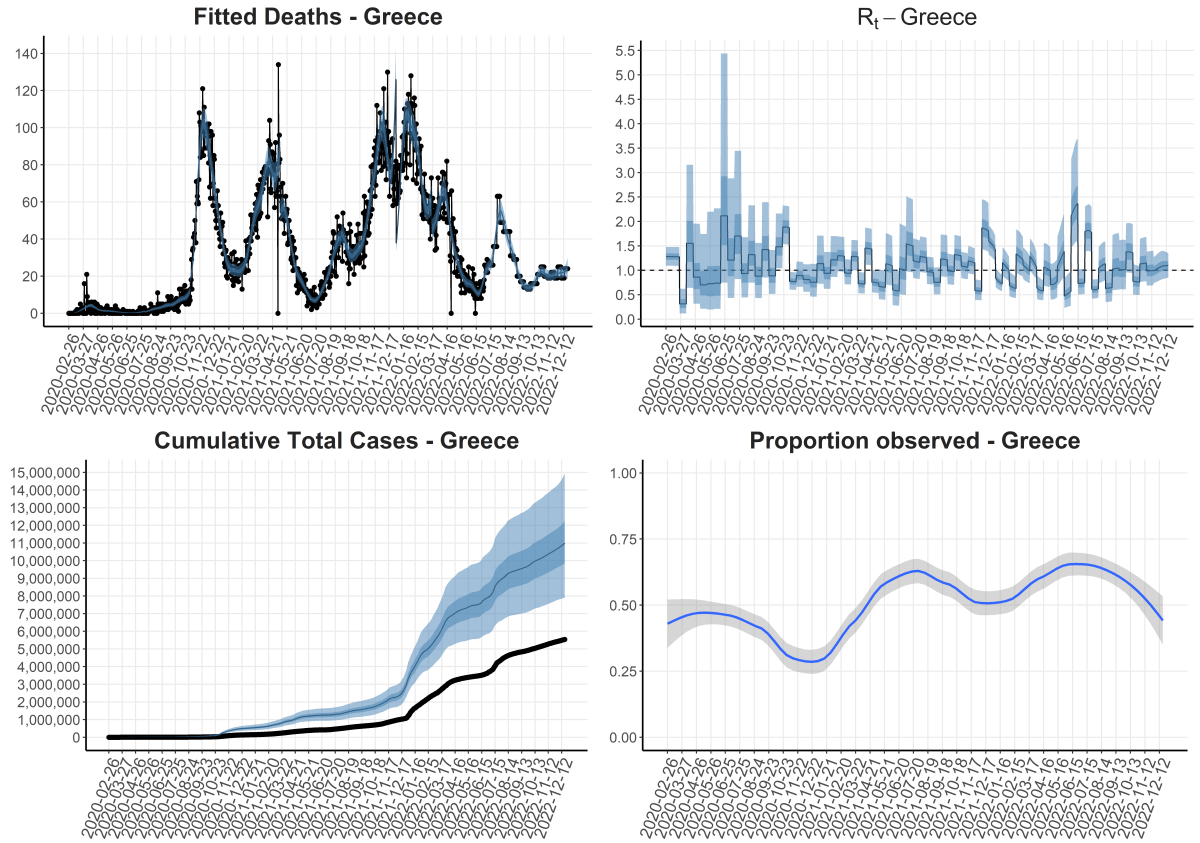

Figure 5.22: Results of the fitted SEIRS model for Greece. Up-Left: Fitted deaths. Up-Right: Reproduction number. Down-Left: Cumulative total cases. Down-Right: Proportion of observed cases.

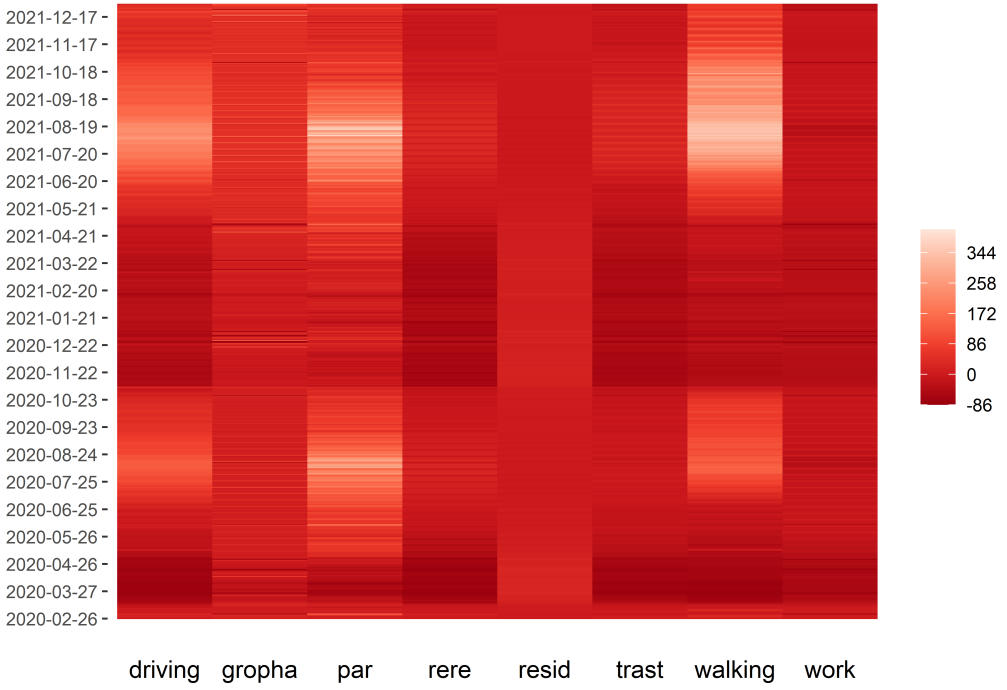

Figure 5.23: The lighter the red color, the highest increase of mobility for the specific variable, while the darker the red color, the lowest decrease of mobility. Plus/minus A means plus/minus A% change.

Mobility data

Google provides much more detailed information about the mobility of people during the epidemic, that is data about different types of places they visit, while Apple reports the mobility in terms of driving or walking, when someone uses Apple maps. The two Apple variables seem to be very correlated and it is highly unlikely that both of them can be used effectively by a model. Both Google and Apple refer to percentage change compared with a baseline. So 0 means "no change", while a value of 50 means 50% change. The Google baseline is the mean mobility from January 3 until February 6, 2020. The Apple baseline is the requests volume on January 13, 2020. A heatmap plot is shown in Figure 5.23.The lockdowns can be spotted more easily by the "driving" variable, which is darker for months 3/2020-4/2020 and 11/2020-4/2021.

Next, let us investigate the variances and correlations between the variables. The

rere	gropha	par	trast	work	resid	driving	walking
1032.041	1015.967	7377.364	857.402	281.931	67.080	6339.491	12638.350

Table 5.18: The variance of each variable is considerably different to the others.

Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8
0.7809	0.1116	0.0578	0.0274	0.0116	0.0057	0.0033	0.0017

Table 5.19: The proportion of variance explained by each component rounded at four decimal digits.

upper part of Figure 5.24 depicts the correlations between every pair of variables. As expected "Driving" and "Walking" are highly correlated ($\rho = 0.96$). Regarding the other variables, one of the largest correlations is between "Transit stations" and "Retail & recreation" ($\rho = 0.93$). Lastly, "Residential" is the only negatively correlated variable with all the others, since it refers to mobility for places of residence in contrast with the other variables. The most negatively correlated pair is between "Residential" and "Transit stations" ($\rho = -0.93$), which makes absolute sense. The lower part of Figure 5.24 depicts the autocorrelation of every mobility variable, which is significant at least for 43 days.

Next, we conduct a Principal Components Analysis, in order to reduce the dimensionality of the data and construct uncorrelated variables that can be used for prediction. Since, the variances of the data differ for every variable substantially (shown in Table 5.18), PCA is conducted on the scaled variables, i.e. we transform the correlation matrix. In order to decide how many components are useful, we examine the scree plot in Figure 5.25, which depicts the proportion of variance explained by each principal component (left-hand side) and the cumulative variance explained as more PC's are added (right-hand side). It seems that either one or two are sufficient, since just the first PC explains 78.09% of the total variance and the second just adds 11.16% (the proportions are shown in Table 5.19.). Kaiser criterion suggests only one.

Although the scree plot is way more popular on deciding on the number of useful

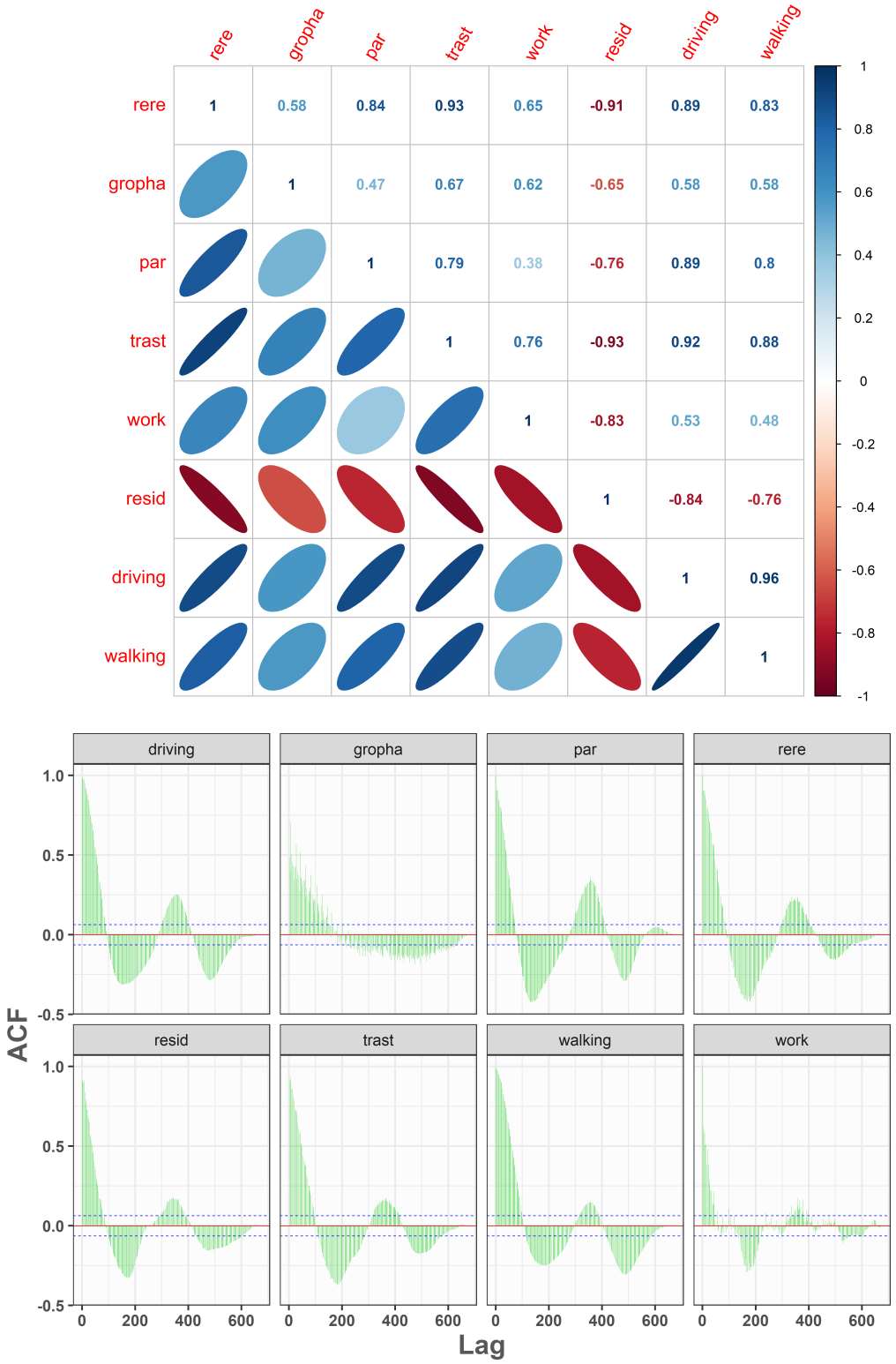

Figure 5.24: Top: Correlations between the mobility variables. Bottom: Autocorrelation of every variable.

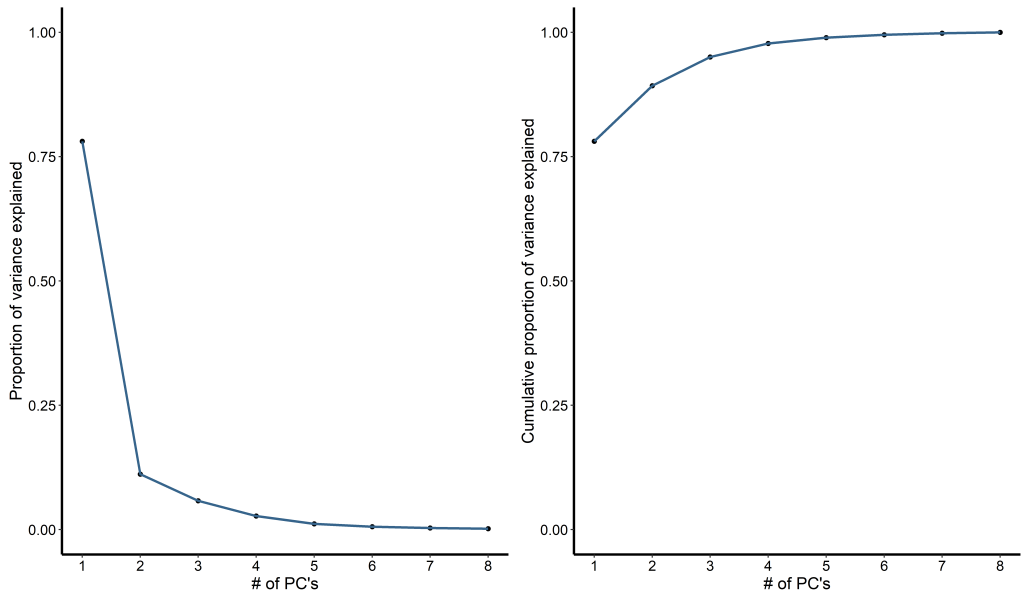

Figure 5.25: Scree plot for the PCA conducted.

PC components based on the explained variance, another visual way can be the boxplot of the PC scores for every component; a much more informative plot, since it depicts descriptive information not included in the scree plot (Figure 5.26). The decreasing variability is clearly seen and the first component has much greater variability than the rest. The difference between the second and third components is not so clearly distinct, so one can choose to keep at most two PC. Lastly, following the "notable elbow" rule of the scree plot, it is impossible for someone to keep only one PC, so boxplots are more advisable for this task.

Interpretation of the mobility components

The interpretation of the new orthogonal axes should begin with the biplot displayed in Figure 5.27. We can see that the first component has positive values on the variables that show outside mobility and negative ones for the "Residential" variable. A high positive value of the first PC means that people go out more than an average day, while a negative value means that they stay inside. The second PC has no clear interpretation, so the only reason to keep it is to use both as predictors in subsequent analysis. One observation we can make is that "Parks" are loaded on the positive y-axis, while "Work" is loaded on

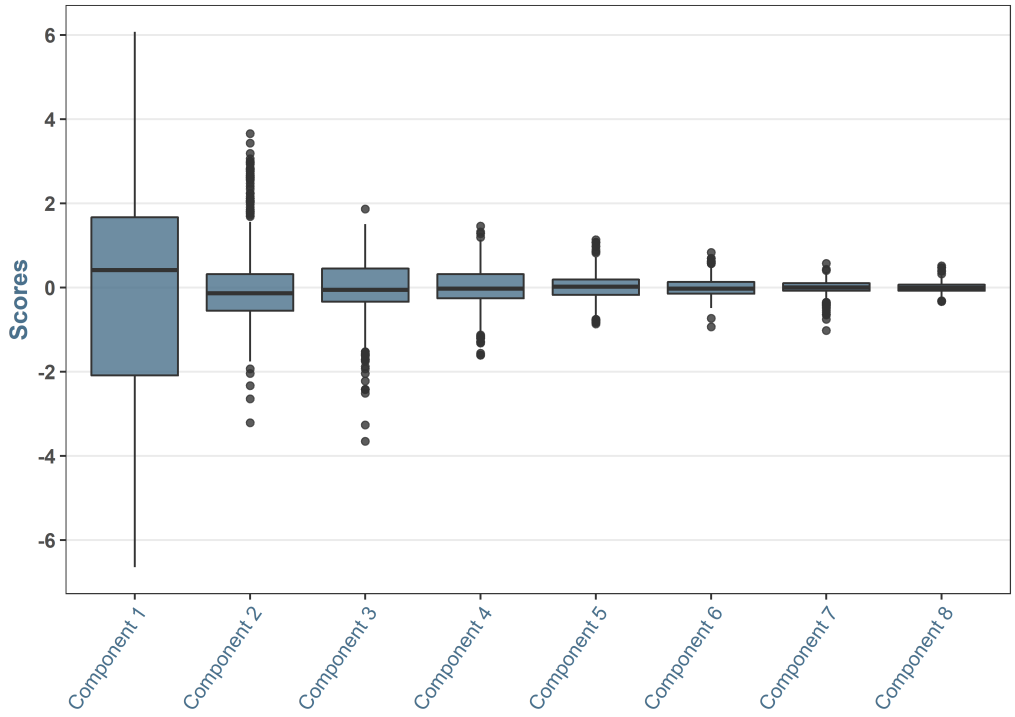

Figure 5.26: Boxplot of the PC scores for every component.

the negative y-axis. The loadings are also shown in Table 5.20.

Next, let us group the data on the plane of the first two components in waves: 26/2/2020 - 1/5/2020, 2/5/2020 - 30/9/2020, 1/10/2020 - 31/12/2020, 01/01/2021 - 20/6/2021 and 21/6/2021 - 31/12/2021. The second period is not considered a wave, since the epidemic was almost over. The grouped data are displayed in the upper part of Figure 5.28. Generally, during the periods of many cases, when restrictive measures were active, we see that the data gather on the negative x-axis (which is loaded by the "Residential" variable). The last wave is an exception, since the restrictive measures were less strict and people were more ignorant. The recorded cases per wave are displayed in the lower part of Figure 5.28.

Furthermore, if we compute the correlations of the 1st and 2nd PC's with all of the original variables (see Table 5.21), we find that the three largest for the first PC are those with "Transit stations", "Retail & recreation" and "Residential" (displayed in the left plot of Figure 5.29), while for the second PC are those with "Work", "Parks" and

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
rere	0.380	0.088	0.208	0.219	0.797	0.083	0.333	0.055
gropha	0.286	-0.425	-0.829	0.213	0.064	-0.024	0.013	0.019
par	0.34	0.434	0.009	0.614	-0.448	0.205	0.145	-0.233
trast	0.392	-0.043	0.087	-0.185	0.078	0.569	-0.684	-0.067
work	0.293	-0.650	0.378	-0.121	-0.332	0.208	0.425	0.002
resid	-0.379	0.177	-0.279	-0.182	0.035	0.739	0.393	0.112
driving	0.379	0.288	-0.070	-0.232	-0.202	-0.105	0.063	0.812
walking	0.361	0.292	-0.192	-0.627	-0.032	-0.162	0.246	-0.516

Table 5.20: PC loadings synthesizing the new orthogonal axes.

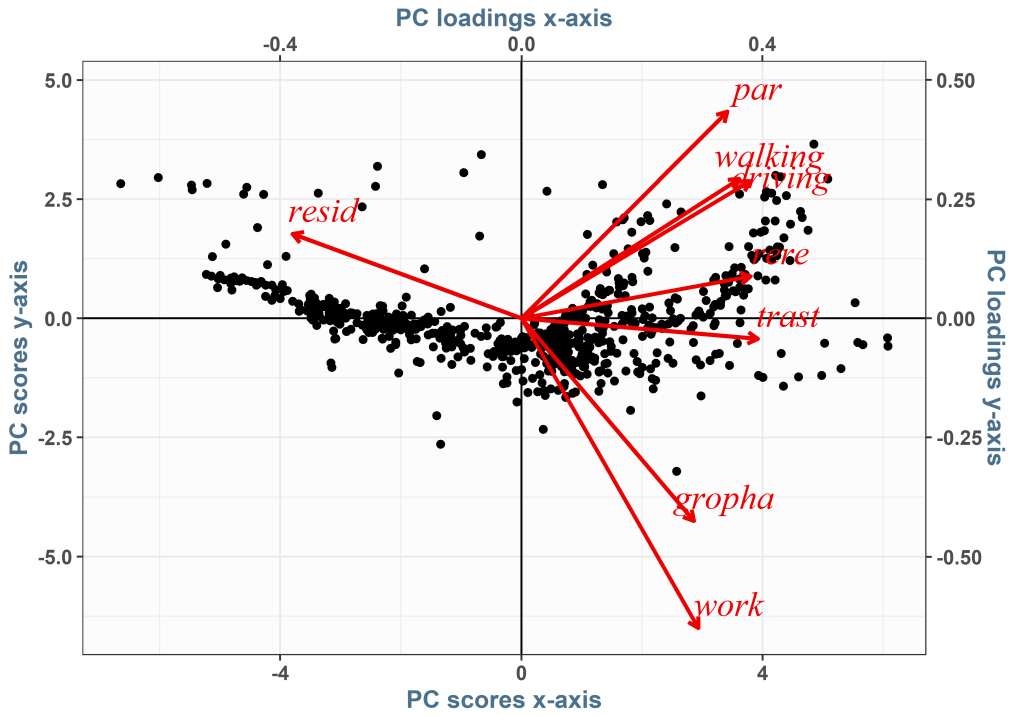

Figure 5.27: Biplot for the PCA conducted. Left and bottom axes correspond to the PC scores displayed as dots, while right and top axes correspond to the PC loadings displayed as red arrows.

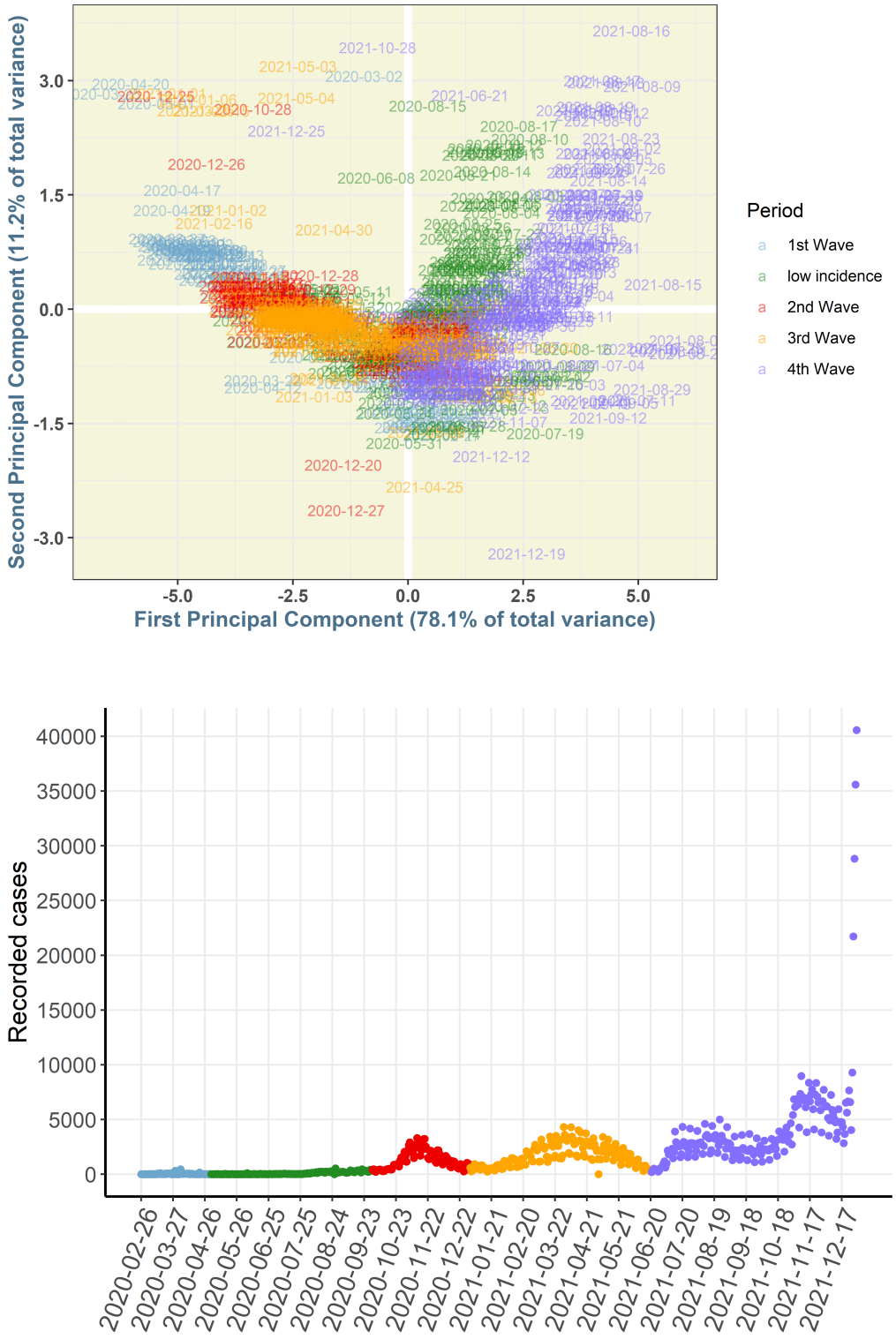

Figure 5.28: Up: The plane of the first two PC's with the grouped data according to the wave they belong to. Down: Recorded cases per wave.

	rere	gropha	par	trast	work	resid	driving	walking
1st PC	0.949	0.714	0.853	0.980	0.733	-0.948	0.948	0.902
2nd PC	0.083	-0.402	0.410	-0.041	-0.614	0.167	0.272	0.276

Table 5.21: The largest correlations between the first two PC's and the original variables.

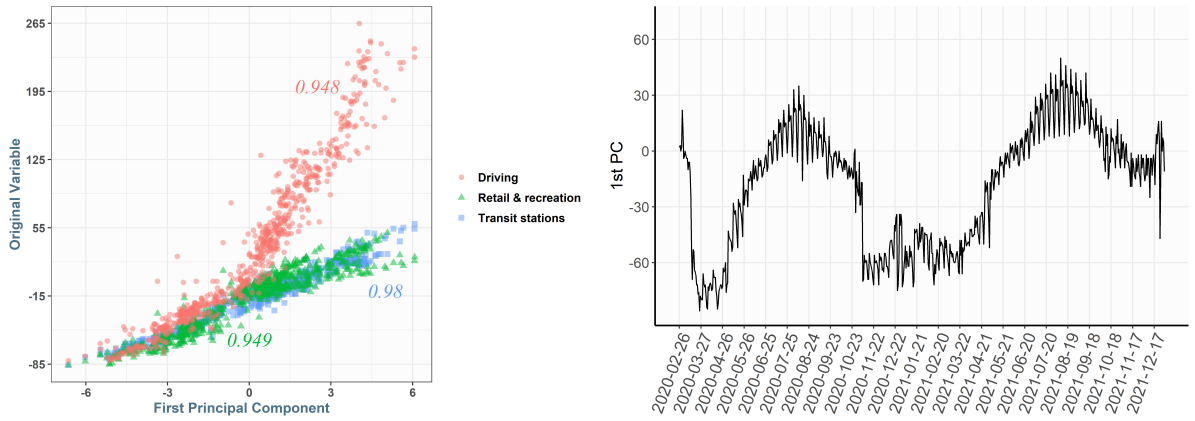

Figure 5.29: Left: The three most correlated variables with the first component. Right: The series of the first PC of mobility.

"Grocery & pharmacy".

Lastly, the series of the first PC is displayed in the right plot of Figure 5.29. Based on the previous visualization, we can see the periods of increased stay in homes when the variable takes on negative values, while more intense "mobility" is observed for larger values. For example, mobility during the first quarantine was low, it increased afterwards (during the summer season) and then there is a sudden drop when the second quarantine started (around 7/11/2020).

Fitting a mixture of two Gaussian distributions using EM

Fitting a mixture of two Gaussian distributions using the Expectation-Maximization algorithm to the scores of the first PC, we find that "mobility" stems from $0.31 \cdot N(-2.86, 1.2^2) + \ 0.69 \cdot N(1.26, 1.78^2)$, as we have described in Chapter 3. Below, we describe the maximization procedure for a mixture of two Gaussian distributions using

EM.

Suppose we have a sample of n observations X_i, which are generated by a mixture of K components. Then, the marginal distribution of X_i is

$$\mathrm{P}(X_i) = \sum_{k=1}^{K} \mathrm{P}(Z_i = k)\mathrm{P}(X_i = x \,|\, Z_i = k)$$

where $Z_i \in \{1, ..., K\}$ denotes the component that generated observation i. Let the mixture components be Gaussian densities each with its own mean and variance, thus we have

$$f(x_i) = \sum_{k=1}^{K} \pi_k f(x_i \,|\, z_i = k)$$

where $f(x_i \,|\, z_i = k) = N(x_i \,;\, \mu_k, \sigma_k^2)$. Therefore, if we let $\boldsymbol{\theta}$ to be the vector of the $3K$ parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$, then the log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\theta}) := \log f(\mathbf{x} \,;\, \boldsymbol{\theta}) &= \log \prod_{i=1}^{n} f(x_i) \\ &= \log \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i \,|\, z_i = k) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(x_i \,|\, z_i = k) \end{aligned}$$

which cannot be maximized analytically. However, if we knew which component each observation belongs to (that is if we knew z_i), the likelihood would be much easier to work with. If we let $z_{ik} = \mathrm{I}(z_i = k)$, then $f(x_i) = \prod_{k=1}^{K} \Big(\pi_k f(x_i \,|\, z_i = k) \Big)^{z_{ik}}$ and the *complete* log-likelihood is

$$\begin{aligned} \log f(\mathbf{x}, \mathbf{z} \,;\, \boldsymbol{\theta}) &= \log \prod_{i=1}^{n} f(x_i) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \log \Big(\pi_k f(x_i \,|\, z_i = k) \Big)^{z_{ik}} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \big(\log \pi_k + \log f(x_i \,|\, z_i = k) \big) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \big(\log \pi_k + \log N(x_i \,;\, \mu_k, \sigma_k^2) \big) \end{aligned}$$

Now, we take the expectation with respect to the conditional distribution of the latent variables $\mathbf{z}$ given the *incomplete* $\mathbf{x}$:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) : &= \mathbb{E}_{\mathbf{Z}\,|\,\mathbf{X}}\Big[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\big(\log\pi_k + \log N(x_i\,;\,\mu_k, \sigma_k^2)\big)\Big] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{Z}\,|\,\mathbf{X}}[z_{ik}]\big(\log\pi_k + \log N(x_i\,;\,\mu_k, \sigma_k^2)\big) \end{aligned}$$

Thus, given an initial value for the parameters, $\boldsymbol{\theta}^*$, the E-step is the computation of $\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{X}}[z_{ik}]$. It holds that

$$\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{X}}[z_{ik}] = P(z_i = k\,|\,x_i) = \frac{P(x_i\,|\,z_i = k)P(z_i = k)}{P(x_i)} = \frac{\pi_k N(x_i\,;\,\mu_k, \sigma_k^2)}{\sum_{j=1}^{K}\pi_j N(x_i\,;\,\mu_j, \sigma_j^2)}$$

so we use $\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{X}}[z_{ik}] = \dfrac{\pi_k^* N(x_i\,;\,\mu_k^*, \sigma_k^{*2})}{\sum_{j=1}^{K}\pi_j^* N(x_i\,;\,\mu_j^*, \sigma_j^{*2})}$ in the expression of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and obtain

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\pi_k^* N(x_i\,;\,\mu_k^*, \sigma_k^{*2})}{\sum_{j=1}^{K}\pi_j^* N(x_i\,;\,\mu_j^*, \sigma_j^{*2})}\big(\log\pi_k + \log N(x_i\,;\,\mu_k, \sigma_k^2)\big)$$

Finally, the M-step is to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}$ and obtain a new $\boldsymbol{\theta}^*$.

Regarding the maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, we take the derivatives with respect to μ_k, σ_k^2 and π_k, set them equal to zero and obtain the maximizers $\hat{\mu}_k$, $\hat{\sigma}_k^2$ and $\hat{\pi}_k$ as follows:

Call $w_{ik} = \dfrac{\pi_k^* N(x_i\,;\,\mu_k^*, \sigma_k^{*2})}{\sum_{j=1}^K \pi_j^* N(x_i\,;\,\mu_j^*, \sigma_j^{*2})}$. Then,

$$\begin{aligned}
\frac{dQ}{d\mu_k} = 0 &\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{d}{d\mu_k} \log N(x_i\,;\,\mu_k, \sigma_k^2) = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{d}{d\mu_k} \Big(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2} - \frac{1}{2}\log\sigma_k^2 - \frac{1}{2}\log(2\pi) \Big) = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{(x_i-\mu_k)}{\sigma_k^2} = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik} x_i - \sum_{i=1}^n w_{ik}\mu_k = 0 \\
&\Leftrightarrow \hat{\mu}_k = \frac{\sum_{i=1}^n w_{ik} x_i}{\sum_{i=1}^n w_{ik}}
\end{aligned}$$

i.e. the mean of each component is a weighted average of the data x_i with weights w_{ik}. Regarding the variances we have:

$$\begin{aligned}
\frac{dQ}{d\sigma_k^2} = 0 &\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{d}{d\sigma_k^2} \log N(x_i\,;\,\mu_k, \sigma_k^2) = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{d}{d\sigma_k^2} \Big(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2} - \frac{1}{2}\log\sigma_k^2 - \frac{1}{2}\log(2\pi) \Big) = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik} \frac{(x_i-\mu_k)^2 - \sigma_k^2}{2\sigma_k^4} = 0 \\
&\Leftrightarrow \sum_{i=1}^n w_{ik}(x_i-\mu_k)^2 - \sum_{i=1}^n w_{ik}\sigma_k^2 = 0 \\
&\Leftrightarrow \hat{\sigma}_k^2 = \frac{\sum_{i=1}^n w_{ik}(x_i-\mu_k)^2}{\sum_{i=1}^n w_{ik}}
\end{aligned}$$

For the mixing proportions we have to take into account the restriction that they sum to 1 and use Lagrange multipliers. If $L = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \lambda(\sum_{k=1}^K \pi_k - 1)$, where λ is the Lagrange multiplier, we need to solve the system

$$\begin{cases} \dfrac{\partial L}{\partial \pi_k} = 0 \\ \dfrac{\partial L}{\partial \lambda} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} \frac{w_{ik}}{\pi_k} - \lambda = 0 \\ \displaystyle\sum_{j=1}^{K} \pi_j = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_k = \dfrac{\displaystyle\sum_{i=1}^{n} w_{ik}}{\lambda} \\ \displaystyle\sum_{j=1}^{K} \frac{\displaystyle\sum_{i=1}^{n} w_{ij}}{\lambda} = 1 \end{cases} \Leftrightarrow$$

$$\begin{cases} \hat{\pi}_k = \dfrac{\displaystyle\sum_{i=1}^{n} w_{ik}}{n} \\ \lambda = \displaystyle\sum_{j=1}^{K}\sum_{i=1}^{n} \pi_j = \sum_{i=1}^{n} 1 = n \end{cases}$$

Note that the index k in the differentiations with respect to μ_k, σ_k and π_k is used to specify a specific k, while the sums over k regard the whole range of k values. This is why the sum over k disappears when we differentiate with respect to a specific μ_k, σ_k or π_k. Therefore, at each iteration of the algorithm the following steps are performed.

Estimate $w_{ik} = \dfrac{\pi_k^* N(x_i\,;\,\mu_k^*, \sigma_k^{*2})}{\sum_{j=1}^{K} \pi_j^* N(x_i\,;\,\mu_j^*, \sigma_j^{*2})}$ using the parameter values of the previous iteration and form $\hat{\mu}_k = \dfrac{\displaystyle\sum_{i=1}^{n} w_{ik} x_i}{\displaystyle\sum_{i=1}^{n} w_{ik}}$, $\hat{\sigma}_k^2 = \dfrac{\displaystyle\sum_{i=1}^{n} w_{ik}(x_i - \mu_k)^2}{\displaystyle\sum_{i=1}^{n} w_{ik}}$ and $\hat{\pi}_k = \dfrac{\displaystyle\sum_{i=1}^{n} w_{ik}}{n}$. Plug the estimated parameters into $\log f(\mathbf{x}; \boldsymbol{\theta}) = \displaystyle\sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{\pi}_k N(x_i\,;\,\hat{\mu}_k, \hat{\sigma}_k^2)$. If the change between two successive values of the log-likelihood is more than 10^{-8}, then re-estimate w_{ik} using the new parameters and repeat. In the scenario, which we study with the mobility variables, we set $K = 2$, since we seek to find two different clusters.

A5. Results on Chapter 4

Fixed points of an SIR model

We can gain insights on the type of fixed points of an SIR system through linearization, although such a method is not always valid for non-isolated fixed points. Following this procedure, we find that the determinant of the Jacobian matrix

$$J = \begin{bmatrix} \frac{\partial \dot{S}}{\partial S} & \frac{\partial \dot{S}}{\partial I} \\ \frac{\partial \dot{I}}{\partial S} & \frac{\partial \dot{I}}{\partial I} \end{bmatrix} = \begin{bmatrix} -\lambda I & -\lambda S \\ \lambda I & \lambda S - \gamma \end{bmatrix}$$

is zero at the fixed points on $I = 0$ and its trace $\lambda S - \gamma$ is negative for $S < \gamma/\lambda$ and positive for $S > \gamma/\lambda$. The eigenvalues are 0 and $\lambda S - \gamma$.

Another way we can think about the geometry of the fixed points is the following. The particle tends to reach to the interval $\mathcal{I} = \{(S, I) \in [0, 1]^2 : 0 \leq S \leq \frac{\gamma}{\lambda}, I = 0\}$ and settle in there, when starting from anywhere on V, except of the interval $\mathcal{J} = \{(S, I) \in [0, 1]^2 : \frac{\gamma}{\lambda} \leq S \leq 1, I = 0\}$ of the x-axis (then it will not move at all and the epidemic will not even occur). So $\mathcal{I}$ attracts an open set of initial conditions in V and $V \setminus \mathcal{J}$ is the basin of attraction of $\mathcal{I}$. This interval is also invariant, because if the particle starts in there, it will not escape from it. Finally, there is no subinterval of $\mathcal{I}$ that exhibits the two aforementioned properties, so $\mathcal{I}$ is minimal and, thus $\mathcal{I}$ is an attractor.

Fixed points of an SIR with demography model

We study the type of fixed points in the SIR with demography model as we did with the SIR model. The determinant of the Jacobian matrix

$$J = \begin{bmatrix} \frac{\partial \dot{S}}{\partial S} & \frac{\partial \dot{S}}{\partial I} \\ \frac{\partial \dot{I}}{\partial S} & \frac{\partial \dot{I}}{\partial I} \end{bmatrix} = \begin{bmatrix} -\lambda I - A & -\lambda S \\ \lambda I & \lambda S - \gamma - A \end{bmatrix}$$

equals $\lambda I(\gamma + A) - A\lambda S + A(\gamma + A)$ and its trace equals $\lambda(S - I) - \gamma - 2A$. Thus, on the disease free fixed point $(1, 0)$, we have $\det J = A^2 + A(\gamma - \lambda)$ and $\mathrm{tr} J = \lambda - \gamma - 2A$,

while on the endemic fixed point (S^*, I^*) we have $\det J = (\gamma + A)\left(A + \lambda\frac{R_0 - 1}{aR_0}\right) - \frac{A\lambda}{R_0}$ and $\mathrm{tr}J = \lambda\frac{a - R_0 + 1}{aR_0} - \gamma - 2A$.

Regarding the disease free state,

$$\det J > 0 \Leftrightarrow A^2 + A(\gamma - \lambda) > 0 \Leftrightarrow A(\gamma - \lambda) > -A^2 \Leftrightarrow \gamma - \lambda > -A$$
$$\Leftrightarrow \gamma + A > \lambda \Leftrightarrow \frac{\lambda}{\gamma + A} < 1 \Leftrightarrow R_0 < 1$$

Now if $R_0 < 1 \Leftrightarrow \lambda < \gamma + A$, it also holds that $\lambda < \gamma + A + A$ since A is positive and, equivalently $\lambda - \gamma - 2A < 0 \Leftrightarrow \mathrm{tr}J < 0$. Finally, it holds that $(\mathrm{tr}J)^2 - 4\det J = (\lambda - \gamma)^2 > 0$. In conclusion, when $R_0 < 1$, then $\det J > 0$, $\mathrm{tr}J < 0$ and $(\mathrm{tr}J)^2 - 4\det J > 0$, so the disease free fixed point is a stable node. On the other hand, when $R_0 > 1$, we have that $\det J < 0$, so the disease free fixed point is a saddle point (with the stable manifold being the x-axis).

Regarding the endemic state,

$$\det J > 0 \Leftrightarrow (\gamma + A)\left(A + \lambda\frac{R_0 - 1}{aR_0}\right) - \frac{A\lambda}{R_0} > 0$$

The term $(\gamma + A)A - \frac{A\lambda}{R_0}$ is zero, because

$$(\gamma + A)A - \frac{A\lambda}{R_0} = 0 \Leftrightarrow A\gamma + A^2 - \frac{A\lambda}{R_0} = 0$$
$$\Leftrightarrow \gamma R_0 + AR_0 - \lambda = 0$$
$$\Leftrightarrow R_0 = \frac{\lambda}{A + \gamma}$$

which holds true, so we have

$$\det J > 0 \Leftrightarrow (\gamma + A)\left(\lambda\frac{R_0 - 1}{aR_0}\right) > 0$$
$$\Leftrightarrow (\gamma + A)\lambda(R_0 - 1) > 0$$
$$\Leftrightarrow R_0 - 1 > 0$$
$$\Leftrightarrow R_0 > 1$$

We can also check that

$$\mathrm{tr}J < 0 \Leftrightarrow \frac{\lambda a - \lambda R_0 + \lambda}{aR_0} - \gamma - 2A < 0$$
$$\Leftrightarrow \lambda > \gamma + A - aA$$

which is true when $R_0 > 1$, i.e. $\lambda > \gamma + A$ (we have omitted the calculations). Therefore, the endemic fixed point is stable when $R_0 > 1$. If $R_0 < 1$, then the y-coordinate of the fixed point becomes negative, which is impossible in practice and thus uninteresting.

Reducing the SIR ODE system

We use the fact that the total population size N remains constant. In our case, we set $N = 1$ and work with S, I and R as proportions of the total population, but the calculations below can all be done with a specific N only by changing λ to a $\lambda^* = \lambda/N$.

We can write the susceptible population at time t as a function of the removed at time t. From the third equation of (4.15) we get $I = \dot{R}/\gamma$ and substituting into the first equation, we have

$$\begin{aligned}
\dot{S} = -\lambda SI = -\lambda S\frac{\dot{R}}{\gamma} &\Rightarrow \frac{dS}{dt} = -\frac{\lambda S}{\gamma}\frac{dR}{dt} \\
&\Rightarrow \frac{dS}{S} = -\frac{\lambda}{\gamma}dR \\
&\Rightarrow \log S = -\frac{\lambda}{\gamma}R + c \quad , \quad c = \text{constant} \\
&\Rightarrow S(t) = exp\Big(\frac{-\lambda R(t)}{\gamma}\Big)e^c
\end{aligned}$$

Then, for $t = 0$ we have $R(0) = 0$, so $S(0) = e^c \Rightarrow c = \log(S(0))$. Thus,

$$S(t) = S(0)exp\Big(\frac{-\lambda R(t)}{\gamma}\Big) \tag{5.25}$$

Now, since $S + I + R = 1$, the third equation of (4.15) becomes $\dot{R} = \gamma I = \gamma(1 - R - S)$ and substituting (5.25) into S, we get

$$\dot{R} = \gamma\Big[1 - R - S_0 exp\Big(\frac{-\lambda R}{\gamma}\Big)\Big] \tag{5.26}$$

where $S_0 = S(0)$. Now, we need to nondimensionalize equation (5.26). Let $u = \dfrac{\lambda R}{\gamma} - \log S_0 \Leftrightarrow R = (u + \log S_0)\dfrac{\gamma}{\lambda}$. Then, $\dfrac{dR}{dt} = \dfrac{\gamma}{\lambda}\dfrac{du}{dt}$ and, by equation (5.26), we have

$$\frac{\gamma}{\lambda}\frac{du}{dt} = \gamma\Big[1 - (u + \log S_0)\frac{\gamma}{\lambda} - e^{-u})\Big] \tag{5.27}$$

Now, let $t^* = \lambda t$, so $dt^* = \lambda dt$. Then, $\frac{du}{dt} = \frac{du}{dt^*}\frac{dt^*}{dt} = \lambda\frac{du}{dt^*}$. Therefore, turning back to equation (5.27) we have

$$\frac{\gamma}{\lambda}\lambda\frac{du}{dt^*} = \gamma\Big[1 - \log S_0\frac{\gamma}{\lambda} - \frac{\gamma}{\lambda}u - e^{-u}\Big] \Rightarrow \frac{du}{dt^*} = a - bu - e^{-u} \tag{5.28}$$

where $a = 1 - \log S_0\frac{\gamma}{\lambda}$ and $b = \frac{\gamma}{\lambda}$.

Glossary

AAD	Anti Arrhythmia Drugs
AIC	Akaike Information Criterion
AR(d)	Autoregressive of Order d
ARMA	Autoregressive Moving Average
BFGS	Broyden – Fletcher – Goldfarb – Shanno
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
CI	Confidence Interval
CrI	Credible Interval
CV	Cross Validation
DIC (or DIC_2)	Deviance Information Criterion
EM	Expectation - Maximization
GAM	Generalized Additive Model
GSE	General Stochastic Epidemic
GUI	Graphical User Interface
HMC	Hamiltonian Monte Carlo
HSA	Hellenic Statistical Authority
IFR	Infection Fatality Ratio
iid	Independent and Identically Distributed
ICA	Independent Components Analysis
ICD	Implantable Cardioverter Defibrilator
ICU	Intensive Care Unit
KM	Kaplan-Meier
LR	Linear Regression
LSODA	Livermore Solver for Ordinary Differential Equations
LYG	Life Years Gained
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo

NCD	Non-Communicable Diseases
NICE	National Institute for Health and Care Excellence
NPIs	Non-Pharmaceutical Interventions
NUTS	No U-Turn Sampler
ODE	Ordinary Differential Equations
PC	Principal Component
PCA	Principal Components Analysis
PDF	Probability Density Function
PPE	Personal Protective Equipment
PHEIC	Public Health Emergency of International Concern
RMS	Restricted Mean Survival
RMSE	Root Mean Square Error
SA	Simulated Annealing
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SEIR	Susceptible - Exposed - Infectious - Removed
SEIRS	Susceptible - Exposed - Infectious - Recovered - Susceptible
SIR	Susceptible - Infectious - Removed
SIRD	Susceptible - Infectious - Recovered or Dead
UK	United Kingdom
USA	United States of America
WAIC	Watanabe – Akaike Information Criterion
WHO	World Health Organization
XGB	Extreme Gradient Boosted Regression Trees

Symbols / Notation

$\mathbb{R}_+$	The set $\{x \in \mathbb{R} : x \geq 0\}$
$\mathbb{N}_0$	The set $\mathbb{N} \cup \{0\}$
X'	Transpose of matrix X
$\lvert \cdot \rvert$	Determinant of a matrix or absolute value of a number
det$X$	Determinant of matrix X
tr$X$	Trace of matrix X
$\lvert\lvert \cdot \rvert\rvert$	L_2 Norm ($\lvert\lvert \cdot \rvert\rvert_2$) of a vector
$I(A)$	Indicator function taking the value 1 if A is true and 0 otherwise

Distributions

Continuous

$U(a, b)$	Uniform with lower bound a and upper bound b
$N(\mu, \sigma^2)$	Gaussian (Normal) with mean μ and variance σ^2
$Exp(\theta)$	Exponential with rate θ
$Gamma(a, b)$	Gamma with shape a and rate b
$Beta(a, b)$	Beta with parameters a and b
$LN(a, b)$	LogNormal with location a and scale b
$C(a, b)$	Cauchy with location a and scale b
$W(a, \sigma)$	Weibull with shape a and scale σ
$W(a, \lambda)$	Weibull with shape a and rate λ

Discrete

$Poisson(\theta)$	Poisson with rate θ
$NB(\theta, \psi)$	Negative Binomial with mean θ and dispersion ψ

Index

Bibliography

[1] Andersson, H. and Britton, T. (2012). *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media.

[2] Apple mobility trends reports.

[3] Bailey, N. T. (1953). The total size of a general stochastic epidemic. *Biometrika*, pages 177–185.

[4] Ball, F. (1983). The threshold behaviour of epidemic models. *Journal of Applied Probability*, 20(2):227–241.

[5] Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability*, pages 46–89.

[6] Bartlett, M. (1949). Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):211–229.

[7] Bashir, M. F., Ma, B., and Shahzad, L. (2020). A brief review of socio-economic and environmental impact of Covid-19. *Air Quality, Atmosphere & Health*, 13(12):1403–1409.

[8] Benaglia, T., Jackson, C. H., and Sharples, L. D. (2015). Survival extrapolation in the presence of cause specific hazards. *Statistics in Medicine*, 34(5):796–811.

[9] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

[10] Birrell, P., Blake, J., Van Leeuwen, E., Gent, N., and De Angelis, D. (2021). Real-time nowcasting and forecasting of Covid-19 dynamics in England: the first wave. *Philosophical Transactions of the Royal Society B*, 376(1829):20200279.

[11] Bjørnstad, O. N. (2022). *Epidemics: models and data using R*. Springer Nature.

[12] Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700.

[13] Britton, T. (2010). Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35.

[14] Britton, T. and Becker, N. G. (2000). Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics*, 1(4):389–402.

[15] Bullement, A., Stevenson, M. D., Baio, G., Shields, G. E., and Latimer, N. R. (2023). A systematic review of methods to incorporate external evidence into trial-based survival extrapolations for health technology assessment. *Medical Decision Making*.

[16] Buxton, M., Caine, N., Chase, D., Connelly, D., Grace, A., Jackson, C., Parkes, J., and Sharples, L. (2006). A review of the evidence on the effects and costs of implantable cardioverter defibrillator therapy in different patient groups, and modelling of cost-effectiveness and cost-utility for these groups in a UK context. *Health Technology Assessment (Winchester, England)*, 10(27):iii–iv.

[17] Cao, L. and Liu, Q. (2022). Covid-19 modeling: A review. *medRxiv*, pages 2022–08.

[18] cBioPortal. Available at `https://www.cbioportal.org/study/summary?id=brca_metabric`. Data downloaded on 28/6/2023.

[19] CDC. Available at `https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html`. Accessed in 2021.

[20] Cereda, D., Tirani, M., Rovida, F., Demicheli, V., Ajelli, M., Poletti, P., Trentini, F., Guzzetta, G., Marziano, V., Barone, A., Magoni, M., Deandrea, S., Diurno, G.,

Lombardo, M., Faccini, M., Pan, A., Bruno, R., Pariani, E., Grasselli, G., Piatti, A., Gramegna, M., Baldanti, F., Melegaro, A. and Merler S. (2020). The early phase of the COVID-19 outbreak in Lombardy, Italy. *arXiv preprint arXiv:2003.09320*.

[21] Che, Z., Green, N., and Baio G. (2023). Blended survival curves: A new approach to extrapolation for time-to-event outcomes from clinical trials in health technology assessment. *Medical Decision Making 43*(3), 299–310.

[22] Connolly, S. J., Hallstrom, A., Cappato, R., Schron, E. B., Kuck, K.-H., Zipes, D. P., Greene, H. L., Boczor, S., Domanski, M., Follmann, D., Gent, M. and Roberts R. S. on behalf of the investigators of the AVID, CASH and CIDS studies (2000). Meta-analysis of the implantable cardioverter defibrillator secondary prevention trials. *European heart journal 21*(24), 2071–2078.

[23] Cox, J. and Woods, D. (2023). Covid-19 and market structure dynamics. *Journal of Banking & Finance*, 147:106362.

[24] Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson., P, Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486*(7403), 346–352.

[25] Demets, D. L. and Lan, K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352.

[26] Demiris, N. and Sharples, L. (2006). Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Statistics in Medicine*, 25(11):1960–1975.

[27] Available at `https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest`. Accessed in 2022.

[28] Available at `https://www.statista.com/statistics/270000/age-distribution-in-the-united-states/`. Accessed in 2023.

[29] Dong, E., Du, H., and Gardner, L. (2020a). An interactive web-based dashboard to track Covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

[30] Dong, Z.-Q., Ma, J., Hao, Y.-N., Shen, X.-L., Liu, F., Gao, Y., and Zhang, L. (2020). The social psychological impact of the covid-19 pandemic on medical staff in china: A cross-sectional study. *European Psychiatry*, 63(1).

[31] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[32] Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., and Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the Covid-19 pandemic. *Infectious disease modelling*, 5:293–308.

[33] `https://coronavirus.data.gov.uk/`. Accessed in 2022.

[34] Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in great Britain. *Nature*, 413(6855):542–548.

[35] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H, Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Imperial College COVID-19 Response Team, Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C. and Bhatt S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820): 257–261.

[36] Gallacher, D., Kimani, P., and Stallard, N. (2021). Extrapolating parametric survival models in health technology assessment: a simulation study. *Medical Decision Making*, 41(1):37–50.

[37] Gefeller, O. and Dette, H. (1992). Nearest neighbour kernel estimation of the hazard function from censored data. *Journal of statistical computation and simulation*, 43(1-2):93–101.

[38] Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016.

[39] Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., and Colaneri, M. (2020). Modelling the Covid-19 epidemic and implementation of population-wide interventions in Italy. *Nature medicine*, 26(6):855–860.

[40] Google COVID-19 Community Mobility Reports.

[41] Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29.

[42] Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., Li, S., Wang, J., Liang, L., Peng, Y., Wei, L., Liu, Y., Hu, Y., Peng, P., Wang, J., Liu, J., Chen, Z., Li, G., Zheng, Z., Qiu, S., Luo, J., Ye, C., Zhu, S., and Zhong N. (2020). Clinical characteristics of Coronavirus disease 2019 in China. *New England journal of medicine*, 382(18):1708–1720.

[43] Guyot, P., Ades, A. E., Beasley, M., Lueza, B., Pignon, J.-P., and Welton, N. J. (2017). Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making*, 37(4):353–366.

[44] Hamid, O., Robert, C., Daud, A., Hodi, F., Hwu, W., Kefford, R., Wolchok, J., Hersey, P., Joseph, R., Weber, J., Dronca, R., Mitchell, T. C., Patnaik, A., Zarour, H. M., Joshua, A. M., Zhao, Q., Jensen, E., Ahsan, S., Ibrahim, N., and Ribas., A. (2019). Five-year survival outcomes for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001. *Annals of Oncology 30*(4), 582–588.

[45] He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. X., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F. and Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*, 26(5): 672–675.

[46] Available at `https://www.statistics.gr/el/statistics/-/publication/SAM03/-`. Accessed in 2022.

[47] Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107(1):49–80.

[48] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Funk, S., and Eggo, R. M. (2020) Feasibility of controlling Covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496.

[49] Hoffman, M. D., and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15 15*(1), 1593–1623.

[50] Human Mortality Database. Max Planck Institute for demographic research (Germany), University of California, Berkeley (USA), and French Institute for demographic studies (France). Available at `https://www.mortality.org`. Data downloaded on 01/08/2023.

[51] Iboi, E., Sharomi, O. O., Ngonghala, C., and Gumel, A. B. (2020). Mathematical modeling and analysis of Covid-19 pandemic in nigeria. *MedRxiv*, pages 2020–05.

[52] Jackson, C., Stevens, J., Ren, S., Latimer, N., Bojke, L., Manca, A., and Sharples, L. (2017). Extrapolating survival from randomized trials using external data: a review of methods. *Medical decision making*, 37(4):377–390.

[53] Jackson, C. H. (2023). survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology*, 23(1):282.

[54] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[55] Kaye, A. D., Okeagu, C. N., Pham, A. D., Silva, R. A., Hurley, J. J., Arron, B. L., Sarfraz, N., Lee, H. N., Ghali, G. E., Gamble, J. W., Liu, H., Urman, R. D.,

and Cornett, E. M. (2021). Economic impact of covid-19 pandemic on healthcare facilities and systems: International perspectives. *Best Practice & Research Clinical Anaesthesiology*, 35(3):293–306.

[56] Keeling, M. J., Woolhouse, M. E., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817.

[57] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

[58] Khan, M. A. and Atangana, A. (2022). Mathematical modeling and analysis of Covid-19: A study of new variant omicron. *Physica A: Statistical Mechanics and its Applications*, 599:127452.

[59] Khattak, A., Carlino, M., Meniawy, T., Ansstas, G., Medina, T., Taylor, M., Kim, K., McKean, M., Long, G., Sullivan, R., Faries, M., Tran, T., Cowey, C., Pecora, A., Segar, J., Atkinson, V., Gibney, G. T., Luke, J., Thomas, S., Buchbinder, E., Hou, P., Zhu, L., Zaks, T., Brown, M., Aanur, P., Meehan, R. S., and Weber, J. S. (2023). A personalized cancer vaccine, mRNA-4157 (V940), combined with pembrolizumab versus pembrolizumab alone in patients with resected high-risk melanoma: Efficacy and safety results from the randomized, open-label phase 2 mRNA-4157-P201/KEYNOTE-942 trial. *CancerRes.*, 83:CT001.

[60] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., and Eggo, R. M. on behalf of the Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group (2020). Early dynamics of transmission and control of Covid-19: a mathematical modelling study. *The lancet infectious diseases*, 20(5):553–558.

[61] Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research.*

[62] Kwok, K. O., Tang, A., Wei, V. W., Park, W. H., Yeoh, E. K., and Riley, S. (2019). Epidemic models of contact tracing: systematic review of transmission studies of severe acute respiratory syndrome and middle east respiratory syndrome. *Computational and structural biotechnology journal*, 17:186–194.

[63] Latimer, N. R. and Adler, A. I. (2022). Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ medicine*, 1(1).

[64] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of Coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.

[65] Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *science*, 343(6176):1203–1205.

[66] Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

[67] Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D., and Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *science*, 300(5627):1966–1970.

[68] Liu, Q. and Cao, L. (2022). Modeling time evolving Covid-19 uncertainties with density dependent asymptomatic infections and social reinforcement. *Scientific Reports*, 12(1):5891.

[69] Marsden, J. and Tromba, A. (2012). *Vector Calculus*.

[70] Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., and Rodés-Guirao, L. (2021). A global database of Covid-19 vaccinations. *Nature human behaviour*, 5(7):947–953.

[71] Meligkotsidou, L. and Vrontos, I. D. (2008). Detecting structural breaks and identifying risk factors in hedge fund returns: A Bayesian approach. *Journal of Banking & Finance*, 32(11):2471–2481.

[72] Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2):685–726.

[73] Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(3):283–313.

[74] Muller, H.-G. and Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, pages 61–76.

[75] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

[76] Nieminen, T. A., Auranen, K., Kulathinal, S., Härkänen, T., Melin, M., Palmu, A. A., and Jokinen, J. (2023). Underreporting of SARS-CoV-2 infections during the first wave of the 2020 Covid-19 epidemic in Finland - Bayesian inference based on a series of serological surveys. *Plos one*, 18(6):e0282094.

[77] Nolte, D. D. (2010). The tangled tale of phase space. *Physics today*, 63(4):33–38.

[78] Oran, D. P. and Topol, E. J. (2021). The proportion of SARS-CoV-2 infections that are asymptomatic: a systematic review. *Annals of internal medicine*, 174(5):655–662.

[79] World Health Organization. `https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases`. Accessed in December 2023.

[80] Pagano, M., Wagner, C., and Zechner, J. (2023). Disaster resilience and asset prices. *Journal of Financial Economics*, 150(2):103712.

[81] Paul, S. and Lorin, E. (2021). Estimation of Covid-19 recovery and decease periods in Canada using delay model. *Scientific Reports*, 11(1):1–15.

[82] Pedroza, C. (2006). A Bayesian forecasting model: predicting us male mortality. *Biostatistics*, 7(4):530–550.

[83] Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25:37–43.

[84] Available at `https://www.statista.com/statistics/263762/total-population-of-the-united-states/`. Accessed in 2023.

[85] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[86] Rahman, H. S., Aziz, M. S., Hussein, R. H., Othman, H. H., Omer, S. H. S., Khalid, E. S., Abdulrahman, N. A., Amin, K., and Abdullah, R. (2020). The transmission modes and sources of Covid-19: A systematic review. *International Journal of Surgery Open*, 26:125–136.

[87] Reinert, G. (1995). The asymptotic evolution of the general stochastic epidemic. *The Annals of Applied Probability*, pages 1061–1086.

[88] Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., Chau, P., Chan, K.P., Lo, S.-V., Leung, P.Y., Tsang, T., Ho, W., Lee, K.-H., Lau, E. M. C., Ferguson, N. M., and Anderson, R. M. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

[89] Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21(15):2175–2197.

[90] Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N., and Ke, R. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome Coronavirus 2. *Emerging infectious diseases*, 26(7):1470.

[91] `https://github.com/Sandbird/covid19-Greece`. Accessed in 2022.

[92] Shama, M. S., Alharthi, A. S., Almulhim, F. A., Gemeay, A. M., Meraou, M. A., Mustafa, M. S., Hussam, E., and Aljohani, H. M. (2023). Modified generalized Weibull distribution: theory and applications. *Scientific Reports*, 13(1):12828.

[93] Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Package deSolve: Solving initial value differential equations in R *Journal of Statistical Software*, 33(9):1–25.

[94] Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). Openbugs user manual. *Version*, 3(2):2007.

[95] Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.5.

[96] Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press.

[97] Sweeting, M. J., Rutherford, M. J., Jackson, D., Lee, S., Latimer, N. R., Hettle, R., and Lambert, P. C. (2023). Survival extrapolation incorporating general population mortality using excess hazard and cure models: a tutorial. *Medical Decision Making*, 43(6):737–748.

[98] Sypsa, V., Roussos, S., Paraskevis, D., Lytras, T., Sotirios, T. S., and Hatzakis, A. (2020). Modelling the SARS-CoV-2 first epidemic wave in greece: social contact patterns for impact assessment and an exit strategy from social distancing measures. *medRxiv*, pages 2020–05.

[99] Szczygielski, J. J., Charteris, A., Bwanya, P. R., and Brzeszczyński, J. (2023). Which Covid-19 information really impacts stock markets? *Journal of International Financial Markets, Institutions and Money*, 84:101592.

[100] Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y., and Wu, J. (2020). Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of clinical medicine*, 9(2):462.

[101] Therneau, T. M. (2024). *A Package for Survival Analysis in R*. R package version 3.5-8.

[102] Tsai, R. and Hotta, L. K. (2013). Polyhazard models with dependent causes.

[103] Van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1-2):29–48.

[104] Villani, A., Potestio, L., Fabbrocini, G., Troncone, G., Malapelle, U., and Scalvenzi, M. (2022). The treatment of advanced melanoma: therapeutic update. *International journal of molecular sciences*, 23(12):6388.

[105] Ward, H., Atchison, C., Whitaker, M., Ainslie, K. E., Elliott, J., Okell, L., Redd, R., Ashby, D., Donnelly, C. A., Barclay, W., Darzi, A., Cooke, G., Riley, S. and Elliott, P. (2021). SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. *Nature communications*, 12(1): 905.

[106] Whittle, P. (1955). The outcome of a stochastic epidemic - a note on Bailey's paper. *Biometrika*, 42(1-2):116–122.

[107] WHO Dashboard. Available at `https://data.who.int/dashboards/covid19/cases?n=c`.

[108] World Health Organization. `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`.

[109] Worldometer. Available at `https://www.worldometers.info/world-population/uk-population/`. Accessed in 2022.