Dynamic Longitudinal Survival Analysis

Gurjinder S. Mohan

St. Peter's College

Department of Statistics
University of Oxford
Trinity, 2017

A thesis submitted for the degree of

Doctor of Philosophy

Dedication

I would like to dedicate this thesis to my parents and family and to everyone I was fortunate enough to learn from. I also dedicate this to the memory of Sabar Koti.

Acknowledgements

I would like to acknowledge and thank Jouni Helske from whom I learnt immensely; Andrew Gelman, Chris Holmes, Robin Evans, Gesine Reinert and Arnaud Doucet for invaluable feedback; Charlotte Deane for kindly allowing use of her server; Branko Ristic for providing a complimentary copy of his book and advice; Martin Brown at St. Peter's college for assistance with administrative affairs and James Carey and his team for assistance and providing data motivating this study. I would also like to acknowledge and thank my examiners Geoff Nicholls and Jane Hutton.

I'm extremely grateful to Peter Clifford and Gesine Reinert for granting financial support through a Departmental Studentship and to Gesine Reinert and Colin McDiarmid for their faith and support as DGS: am also indebted to my irrepressible friend and source of inspiration Luke Kelly, who supervised me through till completion and supported me, both entertainingly and inimitably, throughout. Special thanks to my friends Jessie, Maria, Pavinder, Jaspreet and Jeevan and to Mark, Simon, Susan and Stuart from the IT team and Emma, Jan and Beverley from the Admin team for their help throughout.

Finally, a huge thank you to my supervisor David Steinsaltz for his encouragement and support throughout, without whom none of this would have been possible.

Abstract

In recent times, a greater emphasis is being placed on understanding the underlying processes and mechanisms that drive and generate survival data. As Aalen and Gjessing (2001) note, events don't happen out of the blue, there is a development preceding each event. When observing the time till occurrence of a death, ageing is one such underlying process or 'development'. This project outlines potential methods for the analysis of longitudinal survival data, from an ageing perspective, and develops a novel dynamic model which incorporates understanding of the ageing process, illustrated through application to lifetime behavioural data available for a sample of fruit-flies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An ageing study currently under way in Mexico records the daily behavioural and movement patterns over the lifetime of a sample of Mexican fruit flies and motivates the research conducted as part of this thesis. Since the flies are observed over their lifetimes, survival times are available alongside longitudinal data. Hence, the goal set by the team of researchers is to enable modelling of the ageing process using the longitudinal data with a view to incorporating survival analysis.

Survival analysis, also known as life-time data analysis or failure-time analysis in engineering and reliability contexts, analyses the time until occurence of an event of interest and examples include the time until failure in reliability studies or the time until death or remission following disease in medical studies. Originating from the Bills of Mortality published by John Graunt in 1662, it is one of the oldest statistical disciplines with key applications in the social and biomedical sciences, economics and engineering.

If X is a non-negative random variable representing the time until event, the survival function $S(x)$ is defined as $S(x) = P(X > x)$, where $S(x)$ denotes the probability of surviving beyond x. The survival function is connected to the probability density function (pdf), $f(x)$, and the cumulative distribution function (cdf), $F(x)$, of X through the relation $S(x) = 1 - F(x) = \int_x^\infty f(u)\,du$. The hazard function $h(x)$, referred to as the instantaneous risk of event or force of mortality, is defined as:

$$h(x) = \lim_{\partial x \to 0} \left[\frac{P(x \leq X < x + \partial x | X \geq x)}{\partial x} \right] = \frac{f(x)}{S(x)}$$

where for small ∂x, $h(x)\partial x$ is the probability of dying in the interval $[x, x + \partial x)$, given survival up till x. The cumulative hazard function $H(x)$ is then $H(x) = \int_0^x h(u)\,du$, which allows the survival function and the hazard function to be written entirely in terms of one another since:

$$h(x) = \frac{f(x)}{S(x)} = \frac{\frac{d}{dx}F(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{d}{dx}\left[\log S(x)\right]$$

and so $S(x) = \exp\{-\int_0^x h(u)\,du\} = \exp\{-H(x)\}$.

Landmark papers outlining the Kaplan-Meier estimator for a survival curve (Kaplan and Meier, 1958), the inclusion of covariates in a regression model for survival data (Cox, 1972) and the introduction of counting processes as a basis for inference in survival analysis (Aalen, 1978), have set the landscape for research in this field over the last half century or so (Andersen, 2010).

In recent times, a greater emphasis is being placed on understanding the underlying processes and mechanisms that drive and generate survival data (Aalen and Gjessing, 2001). When observing the time till occurrence of a death, ageing is one such underlying

process or development.

Ageing is one of the most complex and enigmatic biological processes: Peter Medawar (1952) called it 'an unsolved problem of biology' and more than 60 years on, it continues to retain that mystique. Medvedev (1990), as part of a comprehensive review to rationally classify the different theories on ageing wrote, "there are now more than 300 theories of ageing and the number continues to grow", highlighting the complexity and myriad of views. The underlying theory and understanding of ageing nevertheless has consolidated considerably since the seminal ideas of Charles Darwin (1859) inspired early research (Weismann, 1882). With several varying views and a lack of consensus, it can be confusing at times to understand what ageing is.

1.1 What is Ageing ?

Ageing or senescence is often defined as a deteriorative process, which with time, increases the probability that an individual will die (Arking, 2006, Bonsall, 2006). A more prevalent biological definition sees ageing characterized by an accumulation of damage to an organism, leading to loss of function and increased vulnerability to ageing-related diseases and mortality (Jacobson et al., 2010). Ageing typically is considered to be a latent process, only observable through manifestations such as physiological decline.

1.2 Ageing Studies

Several ageing studies—most notably focusing on organisms under laboratory conditions—have taken place in the last twenty years or so, producing contradicting evidence for the various theories and motivating alternative explanations of ageing. Le Bourg (2001), Arking (2006), Ljubuncic and Reznick (2009), Goldsmith (2011) and Bengtson and Settersten Jr (2016) provide useful reviews.

Ageing studies are typically based on the belief that ageing manifests itself through a loss of vigour or vitality—the ability to survive. As a result, increased mortality risk, timing of death or the age-related decline in functional properties is used to track the deterioration associated with ageing (Arking, 2006). However, recognising and recording relevant deterioration in animals poses formidable difficulties (Monaghan et al., 2008) and consequently model systems or surrogates, such as fruit flies, worms and mice are often employed due to their relative convenience.

1.2.1 Behavioural studies using fruit flies

Advances in computing, coupled with novel experimental techniques, have allowed the possibility to study and analyse the ageing process in detail for living organisms.

The fruit fly in particular has proved to be a popular choice for studying ageing and age-related physiological and behavioural changes. Fruit flies develop to adulthood quickly, have a well-documented genome sequence and relatively short mean life span compared to mice or humans, and are also inexpensive to handle (Grotewiel et al., 2005). Piper

and Partridge (2016) highlight further how well-defined dietary requirements and a well-understood biology make them particularly advantageous.

Studies such as Grotewiel et al. (2005), Carey et al. (2006) and Iliadi and Boulianne (2010) provide valuable insights into behavioural changes with age. The further characterization of lifetime behavioural changes specifically are considered essential for understanding ageing in more detail. However as Zou et al. (2011) note, there remains a remarkable paucity of studies on age-specific and entire lifetime patterns, due largely to the lack of an automated system to record lifetime behavioural information until quite recently.

An ageing study currently underway in Mexico records the daily behavioural and movement patterns over the lifetime of a sample of Mexican fruit flies and motivates the research conducted as part of this thesis. Further details on the data are presented in Chapter 3.

1.2.2 Longitudinal survival data

In order to understand the ageing process in greater detail, inevitably observations and measurements are taken at specific points in time across the lifetime of the organism. Hence alongside availability of measurements on a covariate, the time until occurrence of death is also present. Data such as this is referred to as longitudinal survival data and Yashin et al. (2007) acknowledge its potential and how it remains underused due to a lack of suitable models.

Quite often there is simultaneous interest in both the time till death and longitudinal process. Several joint longitudinal survival models have emerged in the last 20 years or so (Pawitan and Self, 1993, Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997, Hogan and Laird, 1997, Xu and Zeger, 2001, Wang and Taylor, 2001, Cowling et al., 2006). In ageing studies however, ageing is thought of as a latent process since it's not considered possible to observe it directly. Joint models incorporating a latent process (Henderson et al., 2000, Hashemi et al., 2003, Lee et al., 2010) which can be used to model the ageing process, are more scarce and are evaluated in Chapter 2.

1.3 Thesis Outline

This thesis is motivated by fly lifetime behavioural data and a goal set by the team of researchers responsible for the data to model ageing and survival jointly and if possible, dynamically. With respect to this, a novel Dynamic Longitudinal Survival (DLS) model is outlined in Chapter 6 which utilises Kalman filtering techniques and Sequential Monte-Carlo (SMC) methods. The relevant Kalman filtering techniques and SMC methods are outlined in chapters 4 and 5 respectively and referred to in Chapter 6: hence, a reader familiar with these methods can go straight to Chapter 6.

Chapter 2 reviews existing ageing models and establishes the relevance of joint longitudinal survival analysis from an ageing and gerontology perspective: the work from this chapter partly comprises my contribution to Steinsaltz, Mohan and Kolb (2012), along with other results which are not presented. A dynamic survival modelling approach

which forms the basis of a DLS model is identified in this chapter.

Chapter 3 outlines the novel lifetime behavioural dataset motivating this study and provides a summary of exploratory survival on the data;

Chapter 4 outlines relevant material for state space models in order to define a survival state space modelling approach towards latter part of chapter.

Chapter 5 outlines relevant Sequential Monte-Carlo (SMC) methods and is referred to in Chapters 6 and 7. Chapter 6 outlines a novel Dynamic Longitudinal Survival (DLS) model in detail.

In Chapter 7 the DLS modelling approach is applied to the fly lifetime behavioural data from Chapter 3 and model diagnostics are presented.

In Chapter 8 the work as part of this research is reviewed, highlighting key results and open problems for further research.

Chapter 2

Literature Review

Overview

This chapter reviews existing mathematical ageing models and highlights the role of frailty models in ageing and survival analysis to capture heterogeneity. A dynamic survival modelling approach capable of modelling ageing in conjunction, is identified towards the end of the chapter in section 2.3 and forms the basis for a novel model developed in Chapter 6.

2.1 Ageing Models

Mathematical models together with the Gompertz distribution have played a key role in the understanding of biological ageing in the last half century or so. One of the

earliest models by Strehler and Mildvan (1960) postulated that an organism consisted of a number of subsystems, each of which had a specific maximum ability to restore following a 'challenge' and that death occurred when an organism could no longer keep up. The hazard rate under such a model became $h(x) = c \cdot \varphi(x)$, where c denoted the total number of challenges per unit of time and $\varphi(x)$ represented the frequency of environmental challenges capable of causing death at age x.

The frequency of challenges were considered to depend on vitality $V(x)$ – an organism's capacity to stay alive at age x – with $\varphi(x) = ke^{\frac{V(x)}{\epsilon D}}$ where k was taken to be a constant, D the relative deleteriousness of the environment and ϵD a measure of the average demand for energy (Yashin et al., 2000). As Finkelstein (2012) notes, Strehler and Mildvan (1960) through such an approach in application to male mortality data were able to throw up support for the Gompertz law of mortality.

Inspired by Strehler and Mildvan (1960)'s biodemographic analysis of mortality rates, Sacher and Trucco (1962) considered homeostatic mechanisms alongside stochastic physiological variables. Individual ageing was postulated as a stochastic differential equation satisfying:

$$d\alpha_t = -aY_t dt + bdW_t, \quad \alpha_0 = \alpha,$$

where $a > 0$ reflected homeostatic feedback, W_t represented random fluctuations through a Wiener process, b denoted a diffusion coefficient describing intensity of random fluctuations and α denoted a random variable with a pdf which satisfied the Fokker-Planck equation (Yashin et al., 2000).

Sacher and Trucco (1962), Yashin, Iachine and Begun (2000), Steinsaltz, Mohan and

Kolb (2012) and Yashin et al. (2016) together provide a fairly thorough review of developments to date of which some are presented herein.

Simpler ageing models inspired by Gompertz (1825) ignored heterogeneity and individual characteristics. A famous example is the cascading failures multistate model by Le Bras (1976) where senescence was considered as a positive integer state and the rate of moving from current state to the next, or an absorbing state, was proportional to the current state as depicted in Figure 2.1, where λ and γ denote the respective rates.

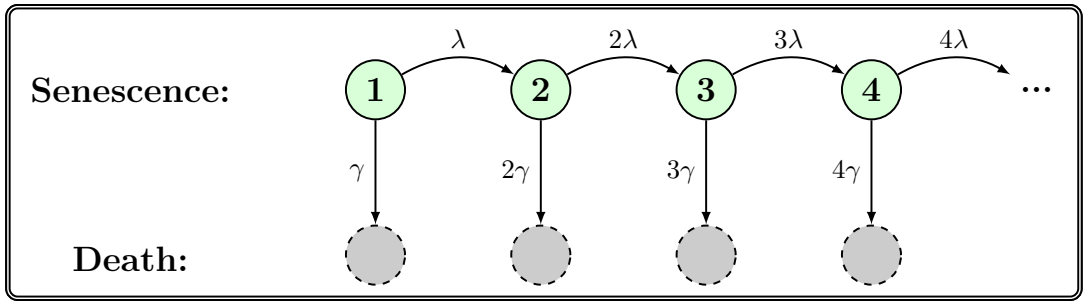

Figure 2.1: Le Bras cascading failures multistate model

Although exponentially increasing mortality (Gompertz, 1825) was observed for much of the lifespan in several biological species, in the words of Curtsinger et al. (2006):

> ...at extreme old ages, a 'mortality deceleration' occurs – the pace of mortality growth decelerates from an expected exponential curve. Sometimes this mortality deceleration progresses to the extent that mortality 'leveling-off' is observed, leading to a 'mortality plateau'. Thus, at extreme old ages, a paradoxical situation is observed when one of the major manifestations of ageing – increasing death rate – apparently fades away or even disappears.

Such a phenomenon was confirmed for Le Bras (1976)'s model. Although he had

purported exponentially increasing mortality, Gavrilov and Gavrilova (1991) showed theoretically using data that mortality plateaus would arise.

Several empirical studies around the same time provide further evidence against Gompertz-like mortality: Carey et al. (1992) notably showed mortality deceleration for a population of medflies, Kannisto et al. (1994) showed the same for humans gathering data from 27 countries and Vaupel et al. (1998) showed similar results for both humans and medflies. Curtsinger et al. (2006) provide a thorough review.

Steinsaltz and Evans (2004) providing a comprehensive analysis of mortality plateaus, showed mortality deceleration is in fact convergence to a quasi-stationary distribution and that such late behaviour of models is to be expected. Quasi-stationarity refers to a limiting state for the hazard distribution which consequently results in a constant population hazard rate (Aalen and Gjessing, 2004).

2.1.1 Frailty

Beard (1959) and later Vaupel et al. (1979) and Lancaster (1979) independently introduced the notion of frailty or random effects in survival analysis to capture the effects of unobserved heterogeneity, where frail individuals were considered to possess a greater risk of death than the overall population and similarly healthier individuals a lower risk of death. A latent frailty Z was assumed to act multiplicatively on the baseline hazard function $h_0(x)$ such that $h(x|Z) = Z \cdot h_0(x)$. Vaupel et al. (1979) assumed Z to follow a Gamma distribution.

Hougaard (1986) considered the power variance function distribution which includes the gamma, inverse Gaussian and positive stable distributions as potential candidates for frailty. Vaupel and Yashin (1985), Aalen (1994) and Hougaard (1995) summarise much of the early work and Duchateau and Janssen (2007), Wienke (2010) and Hanagal (2011) provide more comprehensive recent reviews, which include ideas such as additive frailty and correlated frailty.

Seminal papers, which include Vaupel et al. (1998), Yashin et al. (2000) and Rockwood et al. (2002), through empirical support offered frailty as an explanation for mortality plateaus. Further empirical support from studies such as Sastry (1997), Price and Manatunga (2001), Mitnitski et al. (2002), Yashin et al. (2008) and Richards (2008), where frailty was shown to capture heterogeneity and improve model fit, has ensured it remains a popular part of survival research and especially in the construction of ageing models (Yashin et al., 2016). Steinsaltz and Wachter (2006) provide classical mathematical results to aid better understanding of mortality rate deceleration and heterogeneity in connection with the choice of frailty distribution.

Li and Anderson (2009) extends these ideas by introducing a four-parameter model describing mortality as the first passage time of an abstract measure of survival capacity, where ageing itself was modelled by a stochastic process, a Wiener process with drift.

2.1.2 Damage Accumulation

A precursor to Li and Anderson (2009) was Woodbury and Manton (1977)'s random-walk model for human mortality, where an ageing process α was modelled using a

stochastic differential equation:

$$d\alpha_w(t) = u(\alpha_w, t)dt + d\xi(\alpha_w, t),$$

$$dP(\alpha_w) = \mu(\alpha_w, t)P(\alpha_w)dt,$$

and $dP(\alpha_w)$ denoted the change in survival probability. Models such as these are based on an underlying damage accumulation assumption, referred to as wear and tear in reliability and engineering contexts, whereby the accumulation of damage in an organism is postulated to result in mortality.

Gavrilov and Gavrilova (1991, 2004) promote reliability type biological ageing models based entirely on this assumption. Degradation models provide similar modelling capabilities and Singpurwalla (1995) provides early expository work outlining useful stochastic processes for modelling degradation: stochastic wear and tear.

From the perspective of ageing research, Whitmore et al. (1998) introduced a failure time model based on a bivariate Wiener process, where one process represents an observed marker process and the second represents a latent process. The failure time is defined by the latent process crossing a threshold. Lee et al. (2000) extended this to model a joint marker process and latent health status. The notion of a latent health status is appealing, since it represents the unobservable aging process.

Building on earlier work, Lee and Whitmore (2006) introduced threshold regression for survival data by postulating the deterioration of a subject's health as following a stochastic process and thereby utilising the approach of first-hitting time models: Aalen and Gjessing (2001) provide a comprehensive review of first-hitting-time models.

Following the notation in Lee and Whitmore (2006), let $\{X(t), t \geq 0\}$ denote a Wiener process with mean μ and variance σ^2 , with initial value $X(0) = x_0 > 0$. With $\mu < 0$ the process will tend to drift towards 0 and hence the first-hitting time for the zero level follows an inverse Gaussian distribution. Lee et al. (2010) further extend these ideas to introduce a threshold survival regression model with time-varying covariates in a health status setting. On the other hand, Li and Lee (2011)'s model allows for more flexible representation of the varying coefficients. Bagdonavičius and Nikulin (2001, 2009) and Bagdonavičius et al. (2004) using stochastic wear and tear ideas extend the traditional survival accelerated life model, using a gamma process.

2.1.3 Disposable Soma

One of the most influential modern ageing theories extends the ideas of damage accumulation and is the idea of 'disposable soma' proposed by Kirkwood (1977), which centres on the idea that ageing arises due to a trade-off between allocating resources to self-preservation and other bodily activities such as reproduction. A decline in function is considered to result from unrepaired damage to the organism as a result of life processes such as normal metabolism and reproduction, with damage accumulating over age (Monaghan et al., 2008).

Abrams and Ludwig (1995), Cichoń (1997) and Chu and Lee (2006) provide some of the further extensive studies on disposable soma theory. Some of the applications to ageing through mathematical modelling include Johnson and Mangel (2006), Evans and Steinsaltz (2007) and Bansaye et al. (2011).

Yashin and Manton (1997) and Yashin et al. (2007, 2008, 2011, 2012) extend disposable soma framework, proposing stochastic differential equations for capturing resistance to stresses, adaptive capacities, allostatic[1] adaptations and allostatic load in a stochastic process model for ageing (Yashin et al., 2016).

2.2 Joint Longitudinal Survival Models

Relevant data recorded at various timepoints across an individual's life is considered necessary in order to understand ageing in more detail (Yashin et al., 2016). When the data also comprises survival information as in Zou et al. (2011), joint longitudinal survival models come to the fore, providing a promising foundation from which to construct an ageing model.

Longitudinal measurements collected in such studies are typically incomplete and prone to measurement error and simply using the raw measurements as recorded in survival analysis, can lead to biases (Prentice, 1982, Raboud et al., 1993, Ibrahim et al., 2005). Lange et al. (1992) and Hoover et al. (1992) further allude to the high within subject variability when making biological and physiological observations, endorsing special attention for such covariates through the application of a longitudinal model. Diggle et al. (2002) and Fitzmaurice et al. (2008) are some of the comprehensive overall references available on longitudinal data: Newsom et al. (2013) and Arbeev et al. (2016) provide more specialised references for ageing researchers.

[1] "...allostasis is the process that keeps the organism alive and functioning, i.e. maintaining homeostasis or maintaining stability through change and promoting adaptation and coping."(McEwen, 2002)

Two approaches exist for handling longitudinal survival data in general:

1. The first approach separates the longitudinal and survival analyses into a 2-stage analysis where in the 1st stage, a linear mixed effects model is estimated for the longitudinal covariates ignoring survival information: at the 2nd stage a survival model is fitted using the subject-specific estimated values from the longitudinal model in 1st stage as time-dependent covariates (Tsiatis et al., 1995, Bycott and Taylor, 1998, Dafni and Tsiatis, 1998).

2. The second approach eliminates the isolation of the longitudinal and survival analyses by making the inference using a joint likelihood for the linear mixed-effects and Cox proportional hazards model, motivated by limitations with the 2-stage approach as summarised in Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997) and Yu et al. (2004). The 2-stage approach fails to employ survival information when modelling the longitudinal process and subsequently eliminates uncertainty with the estimated values from longitudinal model when fitting as time-dependent covariates in a Cox proportional hazards model.

Thorough reviews of joint longitudinal survival modelling literature till date are provided by Hogan and Laird (1997), Troxel (2002), Tsiatis and Davidian (2004), Yu et al. (2004), Ibrahim et al. (2005), Verbeke et al. (2010), Sousa (2011), Rizopoulos (2012), Gould et al. (2015) and Elashoff et al. (2017). However, they all omit degradation models as well as the accelerated life joint models outlined by Bagdonavičius and Nikulin (2001), Bagdonavičius et al. (2004) and Bagdonavičius and Nikulin (2009), which are also examples of joint models albeit relatively unknown.

Arbeev et al. (2016) review the various longitudinal modelling approaches (e.g splines, change-points) joint models have hitherto taken from an ageing perspective.

2.3 Dynamic Models

Dynamic survival models offer an alternative modelling approach with some of the earliest examples provided by Gamerman (1991) and Fahrmeir (1994). Gamerman (1991) considered the piecewise exponential class of distributions, through a temporal factorization of the likelihood accounting for covariate effects such as treatment groups. Assuming a random walk evolution of the parameters, the model was estimated as a Bayesian time series (Harrison and Stevens, 1976, West et al., 1985) and alternatively using linear Bayesian methods of the form presented for Dynamic Generalized Linear Models by West et al. (1985). Hemming and Shaw (2005) built on Gamerman (1991), introducing a dynamic model which fitted the log-hazard as an additive function of the baseline hazard and covariates.

Fahrmeir (1994) adopting a different perspective on dynamic survival, postulated survival to evolve in discrete time as a sequence of binary outcomes y_t, where $y_t = 1$ denoted mortality at time t: estimation was based on posterior modes (Fahrmeir and Kaufmann, 1991, Fahrmeir, 1992).

Pavlov (2010) extended the work of Fahrmeir (1994), utilising the inherent Hidden Markov Model foundation of state-space models to model a latent ageing process in addition. With a strong recent focus on dynamic predictions within joint models (Ri-

zopoulos, 2011, Barrett and Su, 2017) together with its ability to model the ageing process, Pavlov's approach becomes immediately appealing and hence is explored in greater detail herein.

Chapter 3

Fly lifetime behavioural data

Overview

Before Pavlov (2010) and Fahrmeir (1994)'s modelling approach is formalised in Chapter 4, the data under consideration as part of this thesis and motivating the research is introduced.

3.1 Data

An electronic behavioural monitoring system (BMS), developed jointly by James R. Carey at the University of California, Davis, and a team of engineers at the Instituto Nacional de Astrofisica, Optica y Electronica in Puebla, Mexico, monitors 9 flies simultaneously over the course of their lives. Barring computational problems, each fly is

Table 3.1: Excerpt from data for one fly

Fly ID	Date	Time	Behaviour	X	Y	Z
A5_10_0	03/12/2009	12:54:51.00	2	490.898	529.876	375.516
A5_10_0	03/12/2009	12:54:51.20	2	491.371	529.355	374.979
A5_10_0	03/12/2009	12:54:51.40	2	491.647	529.648	375.834
A5_10_0	03/12/2009	12:54:51.60	2	494.336	529.38	375.444
A5_10_0	03/12/2009	12:54:51.80	2	495.452	528.655	375.658
A5_10_0	03/12/2009	12:54:52.00	2	497.296	528.285	375.648
A5_10_0	03/12/2009	12:54:52.20	2	500.123	527.179	375.65

photographed 5 times per second for one minute, three times an hour, for 24 hours a day over its lifetime. Using image processing, the behavior exhibited in each photo is then automatically identified and recorded in a database together with X,Y,Z co-ordinates. Zou et al. (2011) provide further details on the data collection and behavioural classification procedure.

3.2 Data description

Table 3.1 provides an excerpt from a dataset generated on day 0 - the day of birth - for one particular fly:

From Table 3.1, the *Fly ID* is made up of the fly reference number A5, batch number 10 and day the data is recorded on, day 0 in that instance.

The datasets generated comprise the lifetime behavioural and movement patterns for 36 tephritid - Anastrepha Ludens - fruit flies, bred from the same genetic strand. The equilibration of the genetic background allows focus on age-related behavioural changes and survival information, without the effects of genetic differences (Iliadi and Boulianne,

2010).

To provide a glimpse into the depth of data available, the average age for the 36 flies is 123 days, which equates to almost 100 million behavioural and movement measurements across the flies. With monitoring from birth until the time they die, survival times are also available.

3.2.1 Experimental Design

Curiously, as seen in Table 3.2, the study design is severely imbalanced since only 4 male flies are included overall of which none are assigned to a sugar diet. The reason for such a lopsided design is that it was part of an exploratory run to test equipment for a larger study, which never transpired due to a lack of funding.

Table 3.2: Fly gender and dietary breakdown by sample size

	Diet	
	Full	**Sugar**
Female	16	16
Male	4	-

Nevertheless, 20 of the 36 flies are on a full diet consisting of protein and sugar, while the remaining 16 female flies are on a sugar only diet. This information will be used to assess whether there is a difference in survival for flies on the 2 diets. Dietary restriction (the reduction of dietary intake without causing malnutrition) can extend lifespan (Fontana et al., 2010, Poirier et al., 2010). Given the lack of data available for male flies, it's considered reasonable to exclude them from further analysis.

3.3 Lifetime Behaviour Analysis

Biomarkers are defined as measurements or observations on a biological organism, that provide information on the physiological status of the organism: biomarkers of ageing in addition are expected to be able to predict differences in mortality risk and are therefore particularly sought after but have proved elusive (Warner, 2004, Jacobson et al., 2010). A key aim of this study hence is to establish a potential biomarker of ageing for the fruit flies.

All 36 flies have data files for each of the days they lived. Typically each file has upwards of 10,000 behavioural observations which alone provide little ageing or survival information. A fly transitioning from resting to walking over a fifth of a second reveals little but as Iliadi and Boulianne (2010) note, fruit fly behaviour undergoes a progressive decline with age. Hence, it makes sense to focus on a longer behavioural observation period such as a day and then to compare changes over days.

3.3.1 Exploratory Data Analysis

Day-time for the flies is between 7am and 7pm when the cage lighting is on, with night-time outside these hours. In order to assess any age-related changes for flies on the 2 diets graphically, it's important to get an idea of the proportion of time spent in each behaviour across the entire lifetime of the flies as a whole firstly, as in Carey et al. (2010) and Zou et al. (2011).

Bringing in the dietary information through Figure 3.1, shows female flies on a full

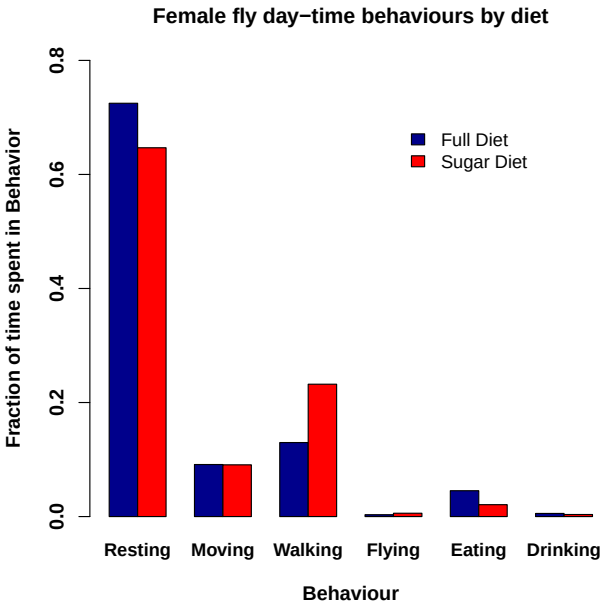

Figure 3.1: Female fly day-time behaviours by diet

Figure 3.2: Female fly night-time behaviours by diet

diet resting between 73-75% during the day, which is almost 10% more than the 65% observed for female flies on a sugar diet. By resting less, the female sugar diet flies spend more time walking, nearly 10% more than the females on a full diet. The flies on a full diet also spend more time eating than their sugar diet counterparts by nearly 2.5%. Night-time behaviours for the flies are fairly similar, with them spending between 92%-95% resting. Standard errors for the estimated proportions are almost negligible given sample sizes in excess of 100 million observations.

Age-related behavioural changes for flies are likely to be increases in rest with age and decreases in periods of intense activity such as flying and walking: to assess this visually, the daily behavioural proportions by age for each fly can be plotted. A plot for one of the flies selected at random is presented in Figure 3.3: the fly selected is a female on a full diet, for which the daily fraction of rest appears to hover around a certain level with a slight increase later on in life; walking levels appear to increase sharply early in life peaking around 50% into the fly's life and then gradually appear to decline with age. Daily behavioural fractions fluctuate quite sharply between days close-by in some cases and upon closer inspection, potential outliers are largely due to the observation period comprising only a few hours for those particular days and hence not reflecting what would transpire over a full day's monitoring. In order to take this into account, a loess curve is fitted as in Figures 3.4 and 3.5 to help provide an initial insight into how the daily behavioural fractions evolve with age. This ties up with a greater interest in long term behavioural changes rather than from one day to the next,

Interestingly, the female flies on a sugar diet largely exhibit similar shapes for the

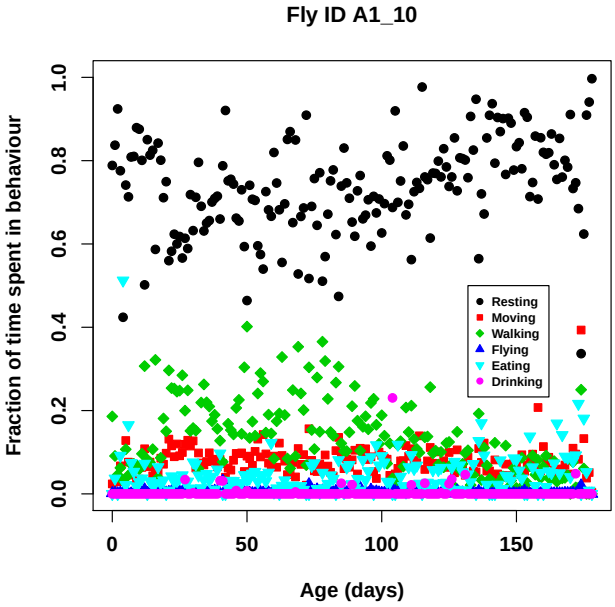

Figure 3.3: Behaviours observed over lifetime of one fly

walking behaviour: walking levels gradually increase with age from birth till about 40-50% into the fly's life, and then gradually decline till death occurs. The flies largely spend little time each day eating, drinking and flying, with little change over age. The sugar-diet flies from Figure 3.5 interestingly exhibit an increase in resting levels towards latter life perhaps as should be expected.

The raw datasets for each fly consist of behavioural observations taken every fifth of a second. Transitions from one behaviour to another as in the raw data alone provide no discernible information on ageing. Considering the transitions that take place over a day however can help estimate the probability with which a fly will continue to rest or transition to walking and so forth each day over a fifth of a second, which may change with age.

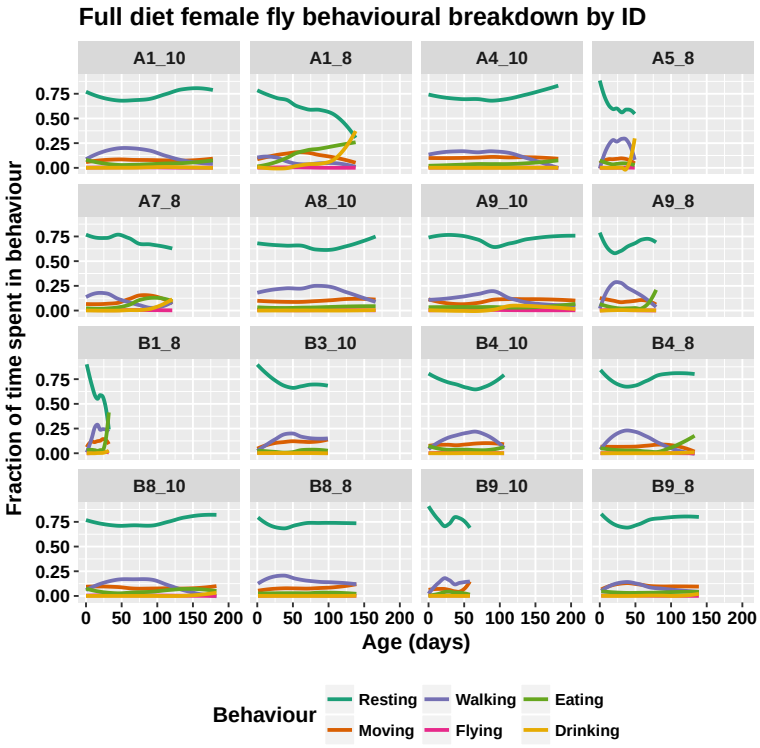

Figure 3.4: Female full diet lifetime behavioural loess fit plot

Hence, the daily empirical behavioural transition probabilities are calculated to assess how likely a fly is to retain any particular behaviour and to transition to other behaviours. Figure 3.6 provides an aggregate of this information across all the flies. As would be expected, the behaviour state space is irreducible with all the behaviours able to 'communicate' with one another, albeit with close to 0 probabilities in some cases. Standard errors are negligible since the transitions are over a million for each case.

Interestingly over $\frac{1}{5}$ of a second across the sample as a whole, flies tend to retain the behaviour they are 'in' more often than not, in fact 86% of the time when moving and upwards of 97% of the time when walking, resting, moving, eating or drinking. Flying however is only retained 15% of the time which is understandable considering

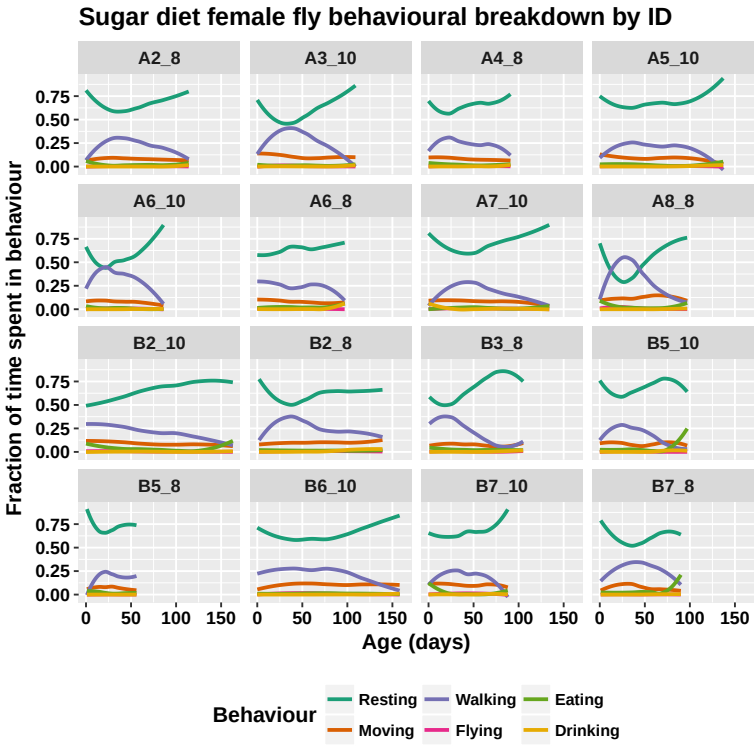

Figure 3.5: Female sugar diet lifetime behavioural loess fit plot

it is perhaps more energy intensive than the other behaviours.

3.3.2 Lifetime Activity

Pavlov (2010) focused on empirical transition probabilites and behavioural fractions for similar fly data. To establish other biomarkers for ageing, the XYZ co-ordinate information available for flies is explored. Using the distance formula, the total Euclidean distance that a fly travels each day can be calculated. As Carey et al. (2010) endorse, lifetime activity levels are likely to be a better biomarker for ageing since they reflect how physically active a fly is each day.

Taking the XYZ co-ordinates, the distance a fly covers each second can be calculated

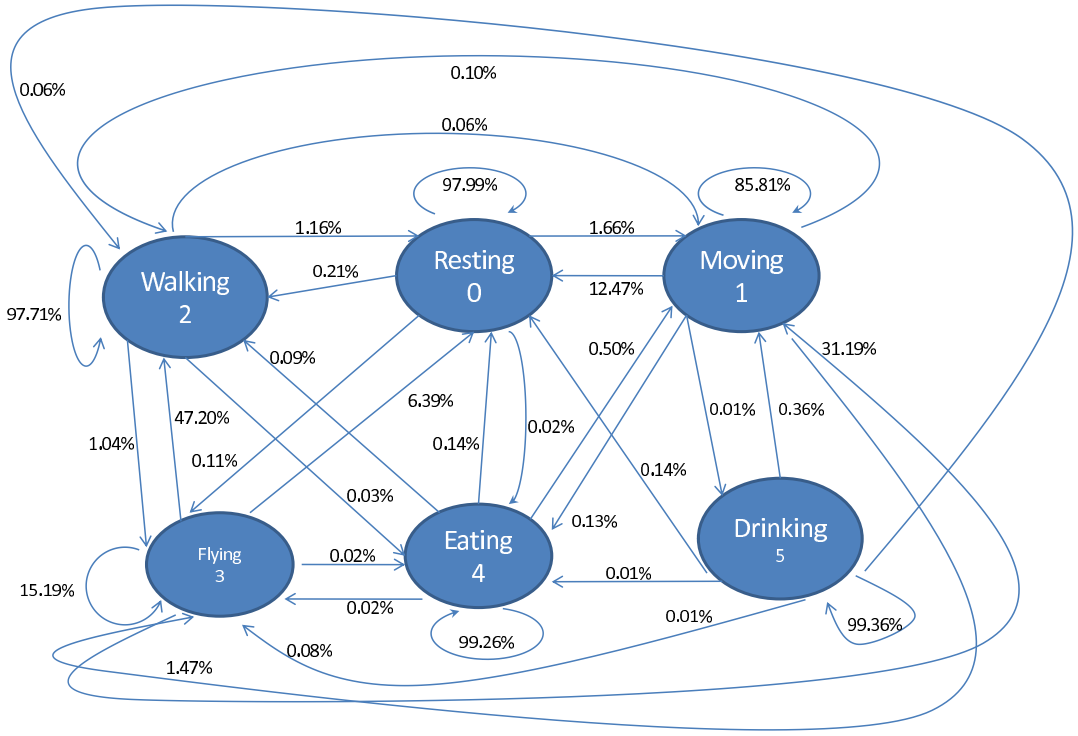

Figure 3.6: Empirical Markov chain transition rates between behaviour states

and then aggregated over each hour and used to calculate an average hourly speed. The average speeds for each hour are then used to estimate the distance a fly travels over day-time hours each day. Upon plotting, to allow a fairer comparison between days and over age, only the days with a complete set of hourly speeds (an average speed for every hour between 7am to 7pm) are plotted – see Figure 3.7.

There's a noticeable difference in the shape of the plots for sugar and full diet flies, as apparent in Figure 3.7. Sugar diet flies appear to generally be more active and the ascent to peak life activity appears more pronounced than full diet flies as a result. Upon reaching peak life activity, both sets of flies exhibit a decline in life activity till

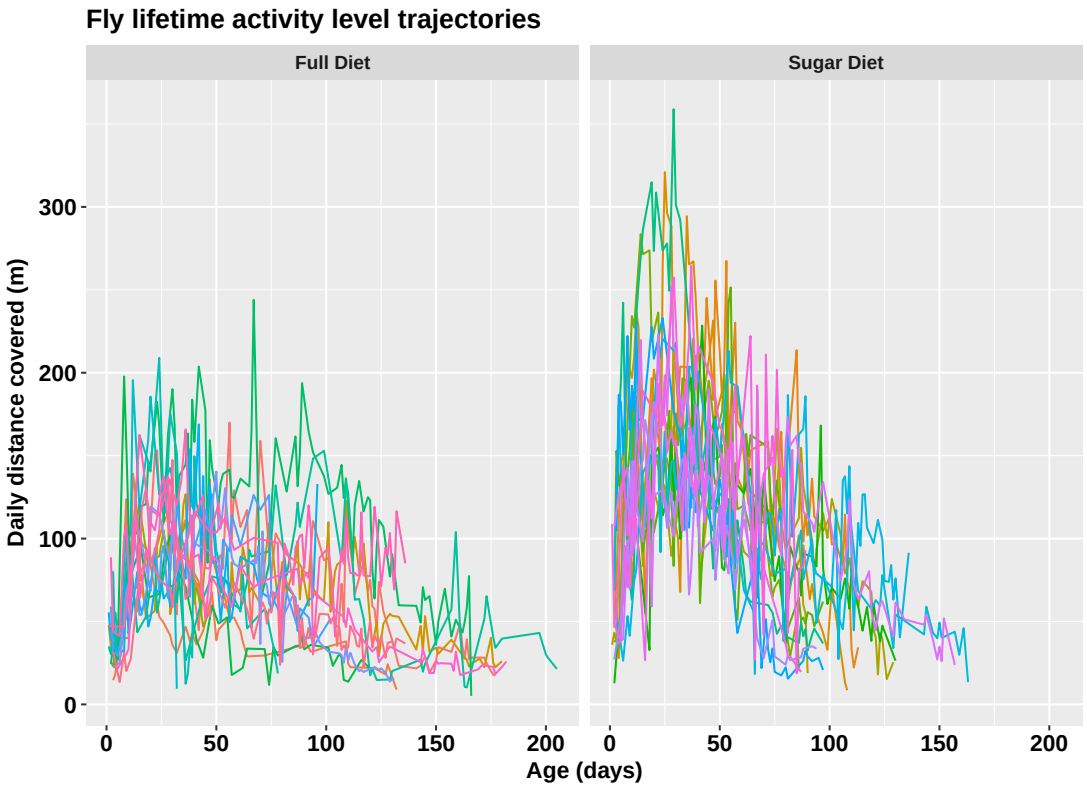

Figure 3.7: Lifetime activity level trajectories by diet

death. Focusing on the distances covered per day at this stage in Figure 3.7, shows sugar diet flies generally cover more distance per day over their lifetimes than the full diet flies, almost 200m or so more in some cases.

To assess statistically whether the lifetime activity level data across both samples arises from the same distribution, a nonparametric Kruskal Wallis Rank Sum test is conducted which returns a p-value close to 0 suggesting a significant difference.

Upon discussion with James Carey and his team of researchers at the University of California, Davis – the biologists responsible for the data – a transformation of the activity level data is considered whereby the peak activity level for each fly is tracked over time, updating when a new peak is recorded and taking the absolute value drop

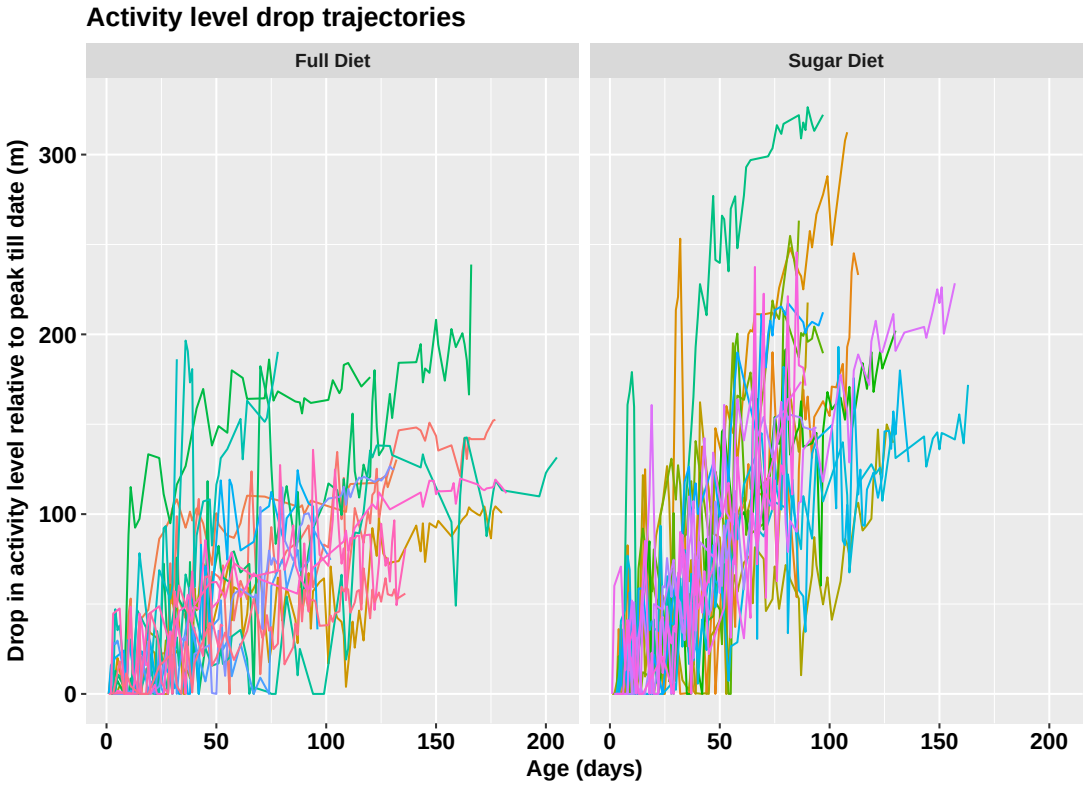

Figure 3.8: Activity level drop trajectories by diet

in activity, relative to the peak.

Defining this more mathematically, let y_t^* denote the raw activity level on day t for an individual fly for $t = 1, \ldots, n^*$, where n^* denotes the survival time and let p^* denote the peak activity level which will be tracked over time. Initially, $p^* = y_1^*$ because that's the peak activity level observed for the fly to date and hence the absolute value drop from peak, denoted by y_t, becomes $y_1 = |y_1^* - p^*| = 0$. When a new peak activity level is observed, p^* is updated and y_t from that time point on are calculated using the updated peak level till a new peak is observed as illustrated in Figure 3.8. The biologists believe this particular transformation will help track ageing more closely since the peak activity level will reflect physical maturity and drops from it potential ageing

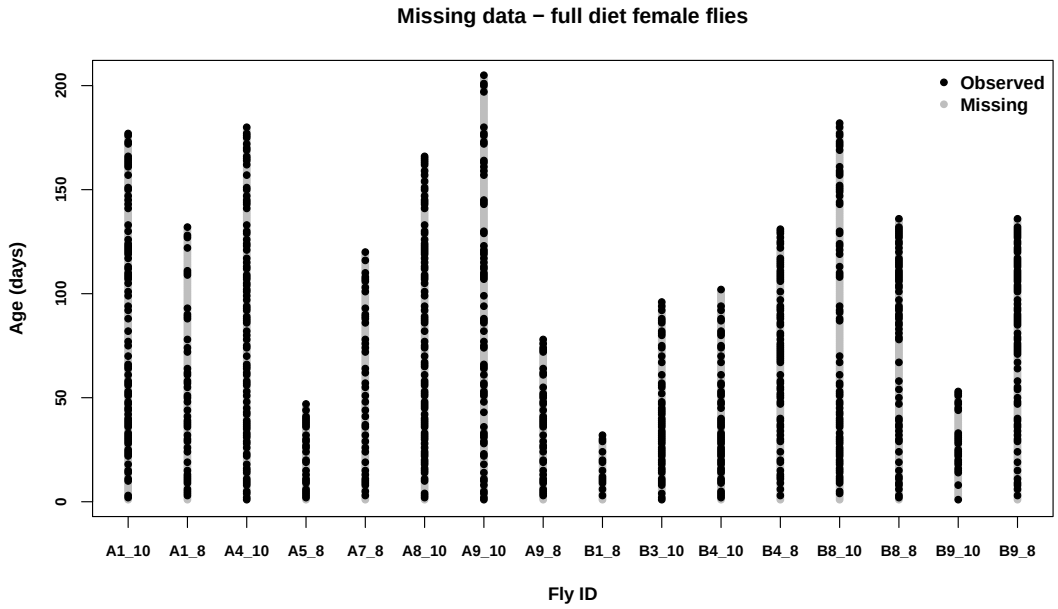

Figure 3.9: Missing data for full diet flies

effects.

Missing Data

Since the data collected was part of a run to test equipment, technical glitches along the way resulted in missing data for flies on both diets as summarised in Figures 3.9 and 3.10. The missing data will be handled (when moving onto the modelling phase) using an approach commonly used when working with state space models, as outlined in sections 4.2.3 and 5.5.

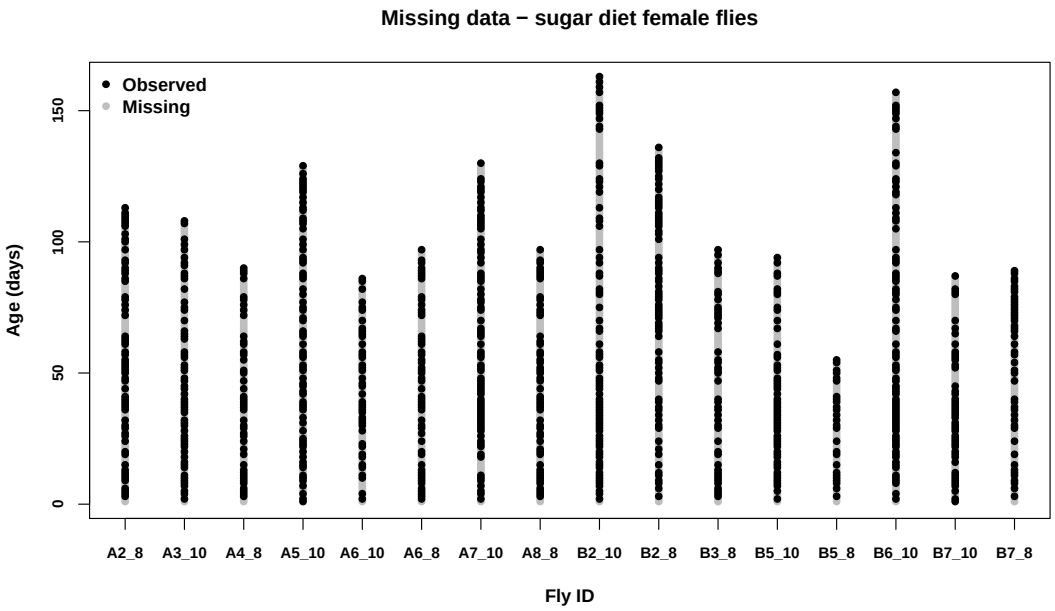

Figure 3.10: Missing data for sugar diet flies

3.4 Survival Analysis

Fly survival times can be deduced from cessation of daily behavioural data. The behavioural observations on the day of mortality mark a prolonged spell of rest so death is known to have occurred sometime on day X_i for fly i. Adopting a naive approach and letting $X_1, ..., X_n$ be the survival times as in Figure 3.11 shows a greater variance in survival times for the full diet flies as compared to the sugar diet flies. Interestingly, but perhaps not surprisingly given the greater variance, the full diet flies have both the highest and lowest survival times as summarised in Table 3.3.

Table 3.3: Summary of fly survival times

	Sample Size	**Min.**	**1st Quartile**	**Median**	**Mean**	**SD**	**3rd Quartile**	**Max.**
Female Full Diet	16	32	91.5	131.5	123.3	52.2	168.8	205
Female Sugar Diet	16	55	89.8	97.0	108.0	28.4	129.3	163
All Female Flies	32	32	89.8	110.1	115.7	42.1	136.0	205

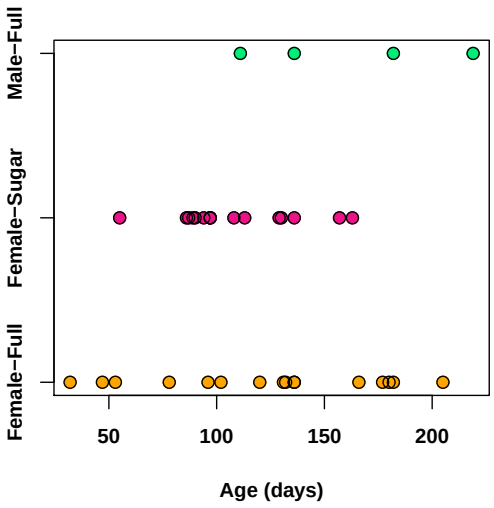

Figure 3.11: Fly survival times by gender and diet

A more detailed graphical analysis of the survival data involves estimation of the survival curve and since survival times are available for all the flies, it can be estimated

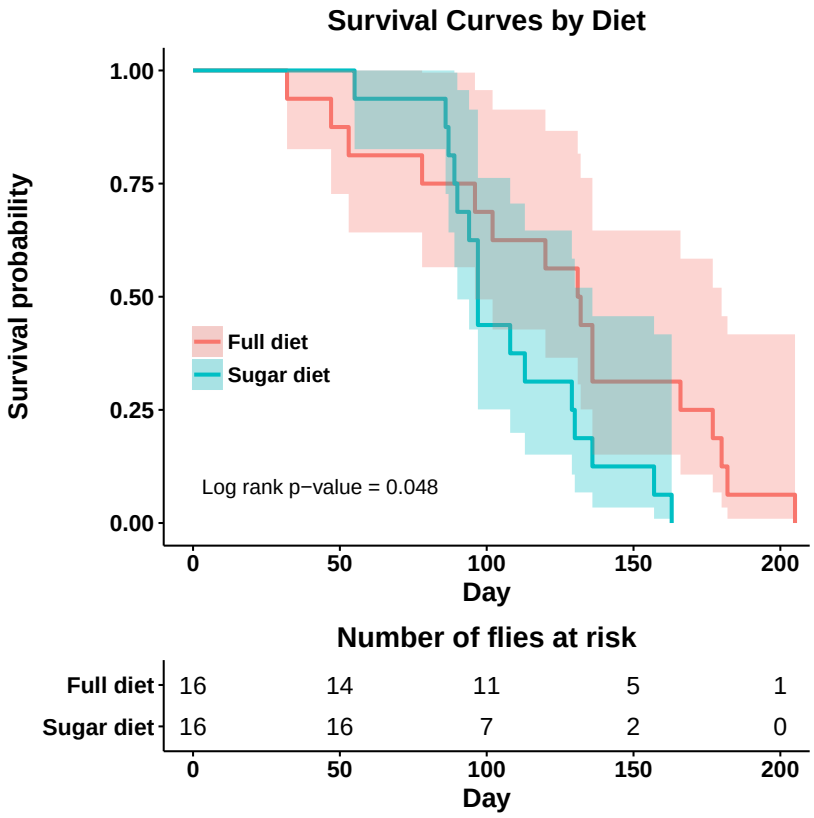

Figure 3.12: Empirical survival curve by diet

through the empirical survivor function, $\hat{S}(x)$, where:

$$\hat{S}(x) = \frac{\text{no. of flies alive} > x \text{ days}}{\text{no. of flies comprising sample}} .$$

The estimated variance of $\hat{S}(x)$ is then $\hat{S}(x)(1 - \hat{S}(x))/n$, where n denotes the sample size. Plotting the empirical survival curve by diet excluding males in Figure 3.12 suggests full diet flies live slightly longer than their sugar diet counterparts and that the hazard function is not proportional since the curves cross.

Testing formally for a difference in survival across the 2 diets via the log-rank test returns a p-value of 0.048, suggesting a significant difference. However, with the log-rank test being based on an underlying assumption of proportional hazards, the Wilcoxon

test is conducted since it's considered more appropriate when proportionality doesn't hold (Collett, 2003) as in Figure 3.12. Peforming a Wilcoxon test returns a p-value of 0.06 suggesting moderate support in favour of a dietary difference in survival.

3.4.1 Interval Censoring

Mortality for each fly is only known to have occurred between the day observations ceased and the previous day. Hence data can be written in the form $\{I_i\}_{i=1}^n$ where $I_i = (L_i, R_i)$, n is the sample size and L_i and R_i refer to the left and right endpoints respectively for individual i. Taking F to be the cumulative distribution function and letting θ denote a vector of unknown parameters, the likelihood can be expressed as:

$$L(\theta) = \prod_{i=1}^{n} [F(R_i) - F(L_i)] \tag{3.1}$$

$$= \prod_{i=1}^{n} [S(L_i) - S(R_i)] . \tag{3.2}$$

Turnbull's Algorithm

Lindsey and Ryan (1998), Sun (2007), Gómez et al. (2009) and Chen et al. (2013) provide a throrough review of methods for interval-censored survival data. One of the most popular ways to obtain a non-parametric estimator for the survival curve accounting for interval censoring is Turnbull's algorithm (Turnbull, 1976).

Following the notation in Gentleman and Geyer (1994), let $\{s_j\}_{j=0}^m$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n\}$. Then as Peto (1973) and Turnbull (1976) first showed, the likelihood for one individual depends on the cdf F only through

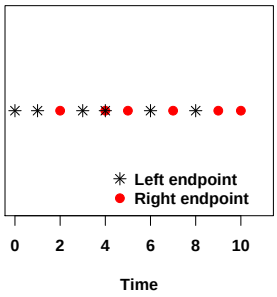

Figure 3.13: Turnbull interval plot for example data

$\{F(s_j)\}_{j=1}^{m}$ and not on the changes in F between s_j (Gentleman and Geyer, 1994) as seen from equation 3.2.

Lindsey and Ryan (1998) provide a graphical approach for finding s_j whereby all the left and right endpoints for a sample are plotted. Then visually, left and right endpoints are identified such that the left point picked is immediately followed by a right endpoint.

Illustrating this for artificial data presented in Gómez et al. (2009) gives Figure 3.13 for a set of 6 individuals with censoring intervals $\{(4,7],(3,5],(0,2],(1,4],(6,9],(8,10]\}$. The Turnbull intervals s_j can then be deduced visually as $\{(1,2],(3,4],(4,5],(6,7],(8,9]\}$.

Letting α_{ij} indicate whether $(s_{j-1}, s_j) \subseteq I_i$ and $p_j = F(s_j-) - F(s_{j-1})$ the likelihood can then be written as (Gentleman and Geyer, 1994):

$$L(\theta) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_{ij} p_j \right], \tag{3.3}$$

with the constraints $\sum_{j=1}^{m} p_j = 1$ and $p_j \geq 0$.

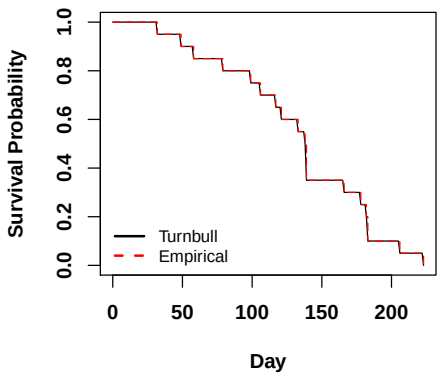

Figure 3.14: Turnbull survival curve for full diet flies

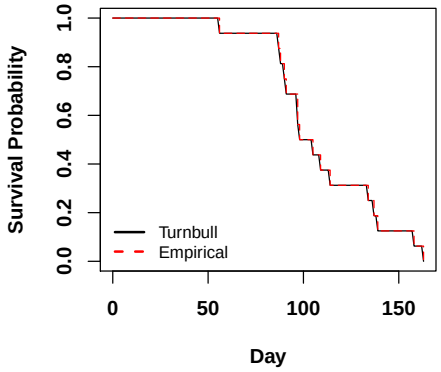

Figure 3.15: Turnbull survival curve for sugar diet flies

3.4.2 Application to fly data

Applying Turnbull's algorithm in order to construct survival curves for the fly data broken down by diet yields Figures 3.14 and 3.15 which show very negligible difference between the empirical survival curves and ones accounting for interval censoring.

Since the interval censoring for the fly data is always the very same small window for each of the 36 flies, it is sufficiently small enough to the point there's hardly any overlapping between intervals for the flies, and therefore the Turnbull algorithm takes s_j as the interval censoring endpoints for the entire sample and the empirical survival

curve provides almost the same curve as one which would be estimated using Turnbull's algorithm.

This shows that since the interval censoring is a very small window, it's reasonable to assume the survival time to be the actual day fly was observed to be dead.

Chapter 4

Survival State Space Model

Overview

This chapter outlines some key general results for state space models in order to define a survival state space model in section 4.4. A reader familiar with Kalman filtering techniques can skip straight to section 4.4.

4.1 State Space Models

In a nutshell, state space models enable insight into the development of a latent process $\alpha_1, ..., \alpha_n$ associated with a series of observations $y_1, ..., y_n$: interest centres on inference of the α_t's having observed $y_1, ..., y_n$. A state space model can be a manifestation of a Hidden Markov Model (see Figure 4.1), which in the general case comprises (Cappé

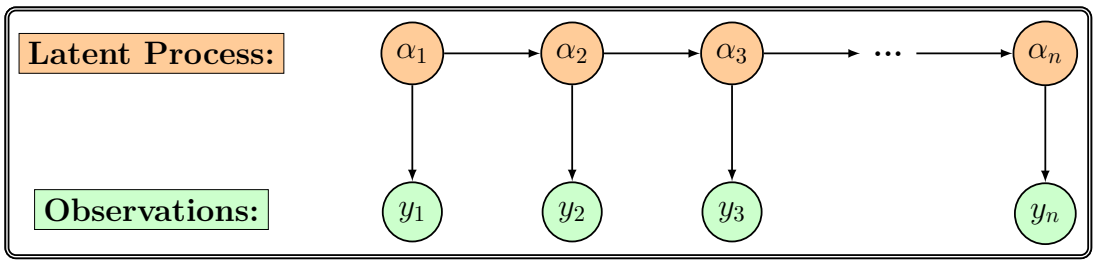

Figure 4.1: Hidden Markov model

et al., 2005, Fernhead, 2012):

i.) a hidden (latent) state, α_t, which is a Markov chain:

$p(\alpha_t|\alpha_{1:t-1}) = p(\alpha_t|\alpha_{t-1})$, where $\alpha_{1:t-1} = \{\alpha_1, \ldots, \alpha_{t-1}\}$,

ii.) and a series of observations y_t, where y_t is conditionally independent of α_s and y_s for all $s \neq t$ given α_t.

4.2 Linear Gaussian State Space Model

State space models can not be solved analytically except in the linear Gaussian case where:

$$y_t = Z_t\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t), \tag{4.1}$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad \eta_t \sim N(0, Q_t), \quad t = 1, ..., n, \tag{4.2}$$

$$\alpha_1 \sim N(a_1, P_1). \tag{4.3}$$

The matrices Z_t, T_t, R_t, H_t, Q_t and the initial distribution of α_1 are usually assumed to be known but can also be estimated as outlined in section 4.2.4. The error terms ϵ_t and

η_t are assumed to be serially independent and independent of each other at all time points (Durbin and Koopman, 2012). Equations (4.1) and (4.2) are referred to as the observation equation and state equation respectively. R_t is usually the identity matrix and can be absorbed into the parameter Q_t in any case, but authors such as Durbin and Koopman (2012) tend to separate out.

Interest centres on estimates of the filtering and smoothing distributions where filtering implies estimates for $E(\alpha_t|Y_t)$ and $Var(\alpha_t|Y_t)$ and smoothing implies estimates for $E(\alpha_t|Y_n)$ and $Var(\alpha_t|Y_n)$, where $Y_t = \{y_1, ..., y_t\}$, for $t = 1, .., n$.

Kalman (1960) famously solved the linear Gaussian case through the Kalman filter and smoother.

4.2.1 Kalman Filter

Let:

$$a_{t|t} = E(\alpha_t|Y_t), \qquad P_{t|t} = Var(\alpha_t|Y_t), \tag{4.4}$$

$$a_t = E(\alpha_t|Y_{t-1}), \qquad P_t = Var(\alpha_t|Y_{t-1}), \tag{4.5}$$

$$\hat{\alpha}_t = E(\alpha_t|Y_n), \qquad V_t = Var(\alpha_t|Y_n). \tag{4.6}$$

Then together with the Markovian structure of the model and results from multivariate normal regression theory (see Appendix A) the Kalman filter for (4.4) and (4.5) can be

derived as (Kalman, 1960, Durbin and Koopman, 2012):

$$v_t = y_t - Z_t a_t, \qquad F_t = Z_t P_t Z_t^T + H_t,$$

$$a_{t|t} = a_t + P_t Z_t^T F_t^{-1} v_t, \qquad P_{t|t} = P_t - P_t Z_t^T F_t^{-1} Z_t P_t,$$

$$a_{t+1} = T_t a_t + K_t v_t, \qquad P_{t+1} = T_t P_t (T_t - K_t Z_t)^T + R_t Q_t R_t^T,$$

for $t = 1, .., n$ where $K_t = T_t P_t Z_t^T F_t^{-1}$. The recursion is initialised using (4.3) and the one step ahead estimates can also be written as:

$$a_{t+1} = T_t a_{t|t}, \qquad P_{t+1} = T_t P_{t|t} T_t^T + R_t Q_t R_t^T,$$

4.2.2 Kalman Smoother

The Kalman smoother, using output from the Kalman filter, returns estimates for (4.6) and is given by:

$$L_t = T_t - K_t Z_t,$$

$$r_{t-1} = Z_t^T F_t^{-1} v_t + L_t^T r_t, \qquad N_{t-1} = Z_t^T F_t^{-1} Z_t + L_t^T N_t L_t,$$

$$\hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad V_t = P_t - P_t N_{t-1} P_t,$$

for $t = n, .., 1$, initialised using $r_n = 0$ and $N_n = 0$.

4.2.3 Missing Data

Missing observations in the Kalman filter and smoother are handled intuitively. Since there's no new information at the time points where observations are missing, one-step

ahead predictions are utilised (Durbin and Koopman, 2012):

$$\begin{aligned} a_{t|t} &= E(\alpha_t|Y_t) = E(\alpha|Y_{t-1}) \\ &= a_t, \end{aligned} \tag{4.7}$$

$$\begin{aligned} P_{t|t} &= Var(\alpha_t|Y_t) = Var(\alpha_t|Y_{t-1}) \\ &= P_t, \end{aligned} \tag{4.8}$$

$$\begin{aligned} a_{t+1} &= E(\alpha_{t+1}|Y_t) = E(T_t\alpha_t + R_t\eta_t|Y_{t-1}) \\ &= T_t a_t, \end{aligned} \tag{4.9}$$

$$\begin{aligned} P_{t+1} &= Var(\alpha_t|Y_{t-1}) = Var(T_t\alpha_t + R_t\eta_t|Y_{t-1}) \\ &= T_t P_t T_t^T + R_t Q_t R_t^T. \end{aligned} \tag{4.10}$$

Hence, the Kalman filter can be updated using equations (4.7) to (4.10) for the case when there's missing data. The backwards smoothing recursions of the Kalman smoother similarly become:

$$r_{t-1} = T_t^T r_t, \qquad N_{t-1} = T_t^T N_t T_t. \tag{4.11}$$

4.2.4 Parameter Estimation

Given an unknown parameter vector ψ and $\alpha_1 \sim N(a_1, P_1)$, with a_1 and P_1 known, the log-likelihood for the linear Gaussian state space model can be written as (Harvey, 1993, Durbin and Koopman, 2012):

$$\log L(Y_n; \psi) = \sum_{t=1}^{n} p(y_t|Y_{t-1}), \tag{4.12}$$

where $p(y_1|Y_0) = p(y_1)$. Since $E(y_t|Y_{t-1}) = E(Z_t\alpha_t + \epsilon_t|Y_{t-1}) = Z_t a_t$ and $v_t = y_t - Z_t a_t$, $F_t = Var(y_t|Y_{t-1})$ and hence the log-likelihood becomes:

$$\log L(Y_n; \psi) = -\frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\left(\log|F_t| + v_t^T F_t^{-1} v_t\right), \tag{4.13}$$

where p is the dimension of the observations and v_t and F_t are output from the Kalman Filter.

4.3 Simulation Smoothing

An algorithm for drawing samples from the latent process given the data is referred to as simulation smoothing and proves useful for investigating performance of techniques and in the analysis of non-Gaussian models using approximation methods such as the 'Approximating Linear Gaussian' state space model - see (Durbin and Koopman, 2000, Section 4). The goal is to generate draws from $P(\alpha_1, ..., \alpha_n|Y_n)$, which is multivariate normal since underlying model is linear and Gaussian.

One widely known early approach is the forwards filtering, backward sampling (FFBS) algorithm, developed independently by Frühwirth-Schnatter (1994), Carter and Kohn (1994) and Shephard (1994), where the identity:

$$P(\alpha_1, ..., \alpha_n|Y_n) = P(\alpha_n|Y_n)\, P(\alpha_{n-1}|Y_n, \alpha_n)...P(\alpha_1|Y_n, \alpha_2, ..., \alpha_n) \tag{4.14}$$

is employed using Bayes' theorem whereby:

$$P(\alpha_1, ..., \alpha_n|Y_n) \propto P(\alpha_n|Y_n)\, P(\alpha_{n-1}|Y_{n-1}, \alpha_n)...P(\alpha_1|Y_1, \alpha_2), \tag{4.15}$$

$$P(\alpha_t|Y_t, \alpha_{t+1}) \propto P(\alpha_{t+1}|\alpha_t)P(\alpha_t|Y_t). \tag{4.16}$$

For the linear Gaussian state space model, these quantities are readily computed by the Kalman Filter.

Building on the work of De Jong and Shephard (1995), Durbin and Koopman (2002) proposed a simulation smoother based on mean-corrections, drawing samples of the disturbances $\epsilon_1, ..., \epsilon_n$ and $\eta_1, ..., \eta_n$ given the observations Y_n.

Following Durbin and Koopman (2012), let $w = (\epsilon_1, \eta_1, ..., \epsilon_n, \eta_n)$, $\hat{w} = E(w|Y_n)$ and W =var$(w|Y_n)$. Then for the linear Gaussian state space model on page 38, the conditional density of w given Y_n is $N(\hat{w}, W)$ and $P(w) \sim N(0, \phi)$ where ϕ =diag$(H_1, Q_1, ..., H_n, Q_n)$ (Durbin and Koopman, 2012). Let w^+ denote a draw from the unconditional $P(w)$ and y^+ denote subsequent values of y generated running the Kalman filter recursions, initialised with a draw α_1^+ from α_1. Then $\hat{w}^+ = E(w|y^+)$ can be computed using the Kalman filter and smoother. Durbin and Koopman (2012) show using A.2, where the conditional variance of x given y in a multivariate normal distribution does not depend on the value of y, var$(w|y^+) = W$. Therefore given y^+, $w^+ - \hat{w}^+ \sim N(0, W)$ unconditionally, irrespective of the value of y^+ and $\tilde{w} = w^+ - \hat{w}^+ + \hat{w}$ is a random draw from $N(\hat{w}, W)$.

Based on this principle, the simulation smoother by Durbin and Koopman (2002) (summarised in Algorithm 1) (Jarociński, 2015) is often the preferred choice (Helske et al.,

Algorithm 1 Simulation Smoother (Durbin and Koopman, 2002)

1. Obtain random draws ϵ_t^+ and η_t^+ from $N(0, H_t)$ and $N(0, Q_t)$ respectively for $t = 1, .., n$.
2. Draw $\alpha_1^+ \sim N(a_1, P_1)$.
3. Using Step 2 to initialise, generate α_t^+ and y_t^+ taking output from Step 1 in place of ϵ_t and η_t in (4.1) and (4.2).
4. Compute $\hat{\alpha}_t = E(\alpha_t|Y_n)$ and $\hat{\alpha}_t^+ = E(\alpha_t|Y_n^+)$ using the Kalman filter and smoother.
5. Return $\tilde{\alpha}_t = \hat{\alpha}_t + \alpha_t^+ - \hat{\alpha}_t^+$

2013, Scott and Varian, 2014, Mikkonen et al., 2015, Zhu et al., 2017) owing to its simplicity and ease of implementation.

The estimation error $\alpha_t^+ - \hat{\alpha}_t^+$ is Normal with zero mean (Bertsekas and Tsitsiklis, 2002, Section 4.7) and can be used to establish the validity of Algorithm 1, whereby $\tilde{\alpha}_t \sim N(\hat{\alpha}_t, V_t)$ should hold. Doucet (2010) provides a succinct summary of the validity. Although Durbin and Koopman (2002) works well in practice, Jungbacker and Koopman (2007) show there may be cases where due to 'imposed ill-defined variance matrices' the method can not be implemented properly (Durbin and Koopman, 2012). The simulation smoother by De Jong and Shephard (1995), a precursor to Durbin and Koopman (2002) where sampling of the disturbances and then states was first considered, works generally.

4.4 Killed State Space Model

Ageing is typically considered to be a latent process only observable through manifestations such as physiological changes: the inherent Hidden Markov Model structure of a state space model becomes naturally appealing as a result.

Pavlov (2010) building on these ideas outlined a 'Killed' state space model of the form below, where ageing is postulated to be an unobserved process driving longitudinal measurements and mortality:

$$\begin{cases} \alpha_{t+1} &= T\alpha_t + \eta_t, \qquad \eta_t \sim N(0, Q_t), \\ y_t &= Z\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t) \\ s_t &\sim \text{Bernoulli}(p_t), \end{cases}$$

where $p_t = P(s_t = 1|\alpha_t)$, α_t refers to the latent ageing process, y_t comprise longitudinal behavioural measurements and s_t denotes a Bernoulli survival indicator of the form presented in Fahrmeir (1994); T, Z, Q_t, H_t are considered to be parameters defining the respective evolution of the longitudinal measurements and ageing over time. Aside from the Bernoulli term, the model is noticeably a linear Gaussian state space model.

4.4.1 Changing Frailty

Frailty was briefly outlined in section 2.1.1 and is the notion that frail individuals possess a greater risk of death than the overall population and similarly healthier individuals a lower risk of death. This can be reflected in the Bernoulli survival probability

p_t utilising a frailty modelling framework where a hazard function:

$$h(t|\theta_t) = h_0(t) \cdot \theta_t, \tag{4.17}$$

where θ_t denotes a random changing frailty term which is a function of the latent process α_t, assumed log-normal for ease of inference by Pavlov (2010) and $h_0(t)$ is a baseline hazard taken to follow a Gompertz distribution. Hence:

$$\begin{aligned} h(t|\theta_t) &= (Ke^{Ct})(e^{\gamma\alpha_t}), \\ h(t|\theta_t) &= Ke^{Ct+\gamma\alpha_t}, \end{aligned} \tag{4.18}$$

where K, C are Gompertz parameters and γ is the co-efficient for effect of hidden ageing process α_t. Pavlov (2010) like Fahrmeir (1994), considers analysis in discrete time and uses the survival function approximation:

$$S(t) = \exp\left\{-\int_0^t h(s)\,ds\right\} \approx \exp\left\{-\sum_{x=1}^{t} h(x)\right\}, \tag{4.19}$$

where t denotes units of days. The dataset considered comprises fly behavioural data and the longitudinal measurements considered are behavioural fractions available from day 1 when fly is born through to when it dies on day n^*: n^* doubles up as a survival time as a result and is therefore considered random. Considering each day in the life of a fly as a Bernoulli trial, Pavlov (2010) defines s_t as the binary indicator on day t, for $t = 1, ..., n^*$, where $s_t = 1$, if fly survives through till the end of the day and 0 otherwise: then, $s_1 = 1, S_2 = 1, ..., S_{n^*-1} = 1, S_{n^*} = 0$.

A survival probability $p_t = p_t(\alpha_t)$ conditional on the unobserved ageing process is then

derived by letting:

$$\begin{aligned} p_t &= p_t(\alpha_t) = P(T > t | T > t-1; \alpha_{\leq t}), \\ &= \exp\{-Ke^{Ct+\gamma\alpha_t}\}, \\ &= P(s_t = 1 | \alpha_t), \end{aligned} \tag{4.20}$$

where $\alpha_{\leq t}$ denotes all available information on ageing process up till time t.

A key underlying assumption of Pavlov's Killed state space model is the conditional independence of the longitudinal measurements y_t and the Bernoulli indicators s_t given ageing: this is considered reasonable given biological theory since it suggests ageing is responsible for the behavioural and survival information and since it impacts the two differently, creates conditional independence.

4.4.2 Likelihood

Given the conditional independence, the likelihood for a paramater vector $\psi = (Z, T, Q_t, H_t, K, c, \gamma)$ can be written as:

$$\begin{aligned} L(\psi; y, s) &= \int P(y, s|\alpha)\, P(\alpha)\, d\alpha \\ &= \int P(y|\alpha)\, P(s|\alpha)\, P(\alpha)\, d\alpha. \end{aligned} \tag{4.21}$$

Since:

$$P(y|\alpha) = \frac{P(\alpha, y)}{P(\alpha)} = \frac{P(\alpha|y)\, P(y)}{P(\alpha)}\,, \tag{4.22}$$

the likelihood simplifies to:

$$\begin{aligned} L(\psi; y, s) &= P(y) \int P(s|\alpha)\, P(\alpha|y)\, d\alpha \\ &= P(y)\, E_{\alpha|y}\{P(s|\alpha)\}. \end{aligned} \tag{4.23}$$

(4.23) allows the likelihood to be evaluated exploiting the joint linear Gaussian structure of α and y through the Kalman filter and smoother since $P(y)$ is the likelihood from Kalman filter omitting s and $E_{\alpha|y}\{P(s|\alpha)\}$ can be evaluated using the simulation smoother outlined in Algorithm 1.

The mortality time is considered as the number of measurements n^* and hence:

$$\begin{aligned} P(s|\alpha) &= \left(\prod_{t=1}^{n^*-1} p_t \right) (1 - p_{n^*}), \\ &= \left(\prod_{t=1}^{n^*-1} \exp\{-Ke^{Ct+\gamma\alpha_t}\} \right) \left(1 - \exp\{-Ke^{Cn^*+\gamma\alpha_{n^*}}\}\right). \end{aligned} \tag{4.24}$$

$E_{\alpha|y}\{P(s|\alpha)\}$ is then evaluated using Monte-Carlo integration:

$$E_{\alpha|y}\{P(s|\alpha)\} \approx \frac{1}{M} \sum_{i=1}^{M} \left[\left(\prod_{t=1}^{n^*-1} \exp\{-Ke^{Ct+\gamma\alpha_t^{(i)}}\} \right) \left(1 - \exp\{-Ke^{Cn^*+\gamma\alpha_{n^*}^{(i)}}\}\right) \right], \tag{4.25}$$

where $\alpha_t^{(i)}$ for $i = 1, .., M$ are draws from $P(\alpha|y)$ returned by Algorithm 1.

Letting:

$$L_i = \sum_{t=1}^{n^*} \exp\{-Ke^{Ct+\gamma\alpha_t^{(i)}}\}, \tag{4.26}$$

where for $t = n^*$, the term summed is:

$$\exp\left\{ \log\left[-\frac{1}{K} \log\left(1 - \exp\{-Ke^{Cn^*+\gamma\alpha_{n^*}^{(i)}}\}\right) \right] \right\}, \tag{4.27}$$

a more computationally stable form for (4.25) is deduced taking the log:

$$\log E_{\alpha|y}\{P(s|\alpha)\} \approx -KL + \log\left[\frac{1}{M}\sum_{i=1}^{M}\exp\left(-K(L_i - L)\right)\right], \tag{4.28}$$

where $L = \min\{L_1, ..., L_M\}$.

4.4.3 Approximating Linear Gaussian Model

For non-Gaussian state space models, approximate filtering techniques need to be utilised. Pavlov's model comprises a linear Gaussian component plus a Bernoulli survival indicator from the Exponential family of distributions.

Durbin and Koopman (2000) outline an 'Approximating Linear Gaussian' model which converges to the same conditional mode for the unobserved state process given the observed measurements as non-Gaussian model. Given an ageing process of the form:

$$\alpha_t = T\alpha_{t-1} + \eta_t, \quad \eta \sim N(0, Q_t), \tag{4.29}$$

where T and Q_t are assumed known. Let $\phi_t = Z\alpha_t$ for known Z and assume observations to arise from exponential family distributions with form:

$$p(y_t|\phi_t) = \exp\{y_t\phi_t - b_t(\phi_t) + c_t(y_t)\}. \tag{4.30}$$

Let:

$$\dot{b}_t = \frac{\delta b_t(\phi_t)}{\delta\phi_t}\Big|_{\phi_t=\tilde{\phi}_t}, \tag{4.31}$$

$$\ddot{b}_t = \frac{\delta^2 b_t(\phi_t)}{\delta\phi_t\delta\phi_t'}\Big|_{\phi_t=\tilde{\phi}_t}, \tag{4.32}$$

where $\tilde{\phi}_t$ is a trial value of ϕ_t. Then taking:

$$A_t = \ddot{b}_t^{-1}, \tag{4.33}$$

$$\tilde{y}_t = \tilde{\phi}_t - \ddot{b}_t^{-1}(\dot{b}_t - y_t), \tag{4.34}$$

an approximating linear Gaussian model is given by:

$$\begin{cases} \alpha_t = T\alpha_{t-1} + \eta_t, \quad \eta \sim N(0, Q_t), \\ \tilde{y}_t = Z\alpha_t + \epsilon_t, \quad \epsilon \sim N(0, A_t). \end{cases} \tag{4.35}$$

The filter is initialised with a trial value of $\tilde{\phi}_t$ with each iteration of the filter returning a new trial value until convergence is achieved: initial trial values don't need to be accurate and convergence is typically achieved between 5 and 10 iterations.

Pavlov (2010) attempts to create an approximating model following Durbin and Koopman (2001) and appears to follow steps (4.29) to (4.33) but for step (4.34) he creates a pseudo-observation for S_t which isn't elaborated but is different to what (4.34) outlines: notation is an issue and it seems to have confused the construction of model. In an attempt to make Pavlov (2010) easier to follow, and to show where he differs, a complete overhaul of his notation and a more reader friendly approach outlining his attempt gives:

$$\phi_t = \log\left(\frac{p_t}{1-p_t}\right) = -Ke^{Ct+\gamma\alpha_t} - \log\left[1 - \exp\left(-Ke^{Ct+\gamma\alpha_t}\right)\right], \tag{4.36}$$

and hence $b_t(\phi_t) = \log\left(1 + \exp^{\phi_t}\right)$.

Durbin and Koopman (2001) outline $\ddot{b}_t$ and $\dot{b}_t$ for a Bernoulli observation as:

$$\dot{b}_t = \exp \phi_t \left(1 + \exp \phi_t\right)^{-1} \tag{4.37}$$

$$\ddot{b}_t = \exp \phi_t \left(1 + \exp \phi_t\right)^{-2}. \tag{4.38}$$

Following (4.34) the pseudo-observation should be of the form:

$$\tilde{y}_t = \tilde{\phi}_t + \ddot{b}_t^{-1} y_t - \ddot{b}_t^{-1} \dot{b}_t, \tag{4.39}$$

which Pavlov (2010) agrees with for all the terms bar $\tilde{\phi}_t$ for which he suggests the first derivative of a fixed ϕ_t (rather than a trial value) times a $\tilde{\alpha}_t$, both of which are unsubstantiated. Given the iterative procedure by which subsequent trial values are generated from the Kalman filter and smoother and plugged in for mode estimation, Pavlov (2010) doesn't follow Durbin and Koopman (2001) through nor explain what's being attempted: generation of the trial values is naturally the most crucial part of the iterative procedure.

Nevertheless the framework of his model offers significant potential for an ageing model capable of modelling the fly data.

Chapter 5

Sequential Monte Carlo Methods

Sequential Monte Carlo (SMC) methods – also known as particle filtering or particle methods – offer a simulation based approach to estimating posterior distributions for state-space models and consequently come to the fore in a non-linear, non-Gaussian setting where approximate techniques are employed such as in section 4.4.3. SMC methods don't rely on local linearisation techniques or crude approximations (Doucet and Johansen, 2009) which often become problematic as in the case of Pavlov (2010). Hence, this section outlines the relevant features of SMC methods for a novel approach to the estimation of a survival state space model in Chapter 6. A reader familiar with SMC methods can skip straight to Chapter 6.

Doucet et al. (2001), Arulampalam et al. (2002), Ristic et al. (2004), Cappé et al. (2007), Fearnhead (2008), Doucet and Johansen (2009) and Creal (2012) provide an overview of methods, from which a short review is provided herein for reference using

a state-space model of the form presented on page 38. Let the transition density for the unobserved process be $P(\alpha_t|\alpha_{t-1})$ and the observation density for observations be $P(y_t|\alpha_t)$, where together they define a Hidden Markov Model of the form presented on page 38. The joint distribution for the unobserved process and observations can then be written as:

$$P(\alpha_{1:n}, y_{1:n}) = P(\alpha_1)P(y_1|\alpha_1)\prod_{t=2}^{n} P(\alpha_t|\alpha_{t-1})P(y_t|\alpha_t) \tag{5.1}$$

The filtering distribution – an estimate of the unobserved process at time t, given information from all the observations up till t – relies upon the posterior distribution. Following Cappé et al. (2007), this can be obtained using the *prediction-correction* recursion (Ho and Lee, 1964):

- **Prediction:**

$$P(\alpha_{1:n}|y_{1:n-1}) = P(\alpha_{1:n-1}, y_{1:n-1})\, P(\alpha_n|\alpha_{n-1}), \tag{5.2}$$

- **Correction:**

$$P(\alpha_{1:n}|y_{1:n}) = \frac{P(y_n|\alpha_n)P(\alpha_{1:n}|y_{1:n-1})}{P(y_n|y_{1:n-1})}, \tag{5.3}$$

where $P(y_n|y_{1:n-1}) = \int P(\alpha_n|y_{1:n-1})P(\alpha_n|\alpha_{n-1})P(y_n|\alpha_n)\, d\alpha_{n-1:n}$ is the predictive distribution of y_n, given the past observations $y_{1:n-1}$. Since it's not possible to sample directly from $P(\alpha_{1:n}|y_{1:n})$, N particle paths $\tilde{\alpha}_{1:n}^{(i)}$, $i = 1, .., N$ are sampled from a convenient importance distribution $q(\alpha_{1:n}|y_{1:n})$.

5.1 Particle Filter

Sequential importance sampling involves choosing an importance sampling distribution such that it allows a recursive procedure: a proposal can be constructed considering the factorization (Doucet et al., 2001, Cappé et al., 2007),

$$q(\alpha_{1:n}|y_{1:n}) = q(\alpha_{1:n-1}|y_{1:n-1})\, q(\alpha_n|\alpha_{n-1}, y_n). \tag{5.4}$$

The unnormalized importance weights are then (Doucet et al., 2000, Cappé et al., 2007):

$$\tilde{w}_n^{(i)} = \frac{P(\tilde{\alpha}_{1:n}^{(i)}|y_{1:n})}{q(\tilde{\alpha}_{1:n}^{(i)}|y_{1:n})}, \tag{5.5}$$

$$\propto w_{n-1}^{(i)} \times \frac{P(\tilde{\alpha}_n^{(i)}|\tilde{\alpha}_{n-1}^{(i)})P(y_n|\tilde{\alpha}_n^{(i)})}{q(\tilde{\alpha}_n^{(i)}|\tilde{\alpha}_{n-1}^{(i)}, y_n)P(y_n|y_{1:n-1})}, \tag{5.6}$$

where $w_n^{(i)} = \tilde{w}_n^{(i)} / \sum_{j=1}^{N} \tilde{w}_n^{(j)}$ and is the normalized weight. The scaling factor $P(y_n|y_{1:n-1})$ does not depend on the state sequence and since the weights are normalized, need not be computed (Cappé et al., 2007). An approximation of the expectation of any function h can be computed using the particles propagated over time where,

$$\bar{h} = \int h(\alpha_{1:n})\, P(\alpha_{1:n}|y_{1:n})\, d\alpha_{1:n}, \tag{5.7}$$

$$\approx \sum_{i=1}^{N} w_n^{(i)}\, h(\tilde{\alpha}_{1:n}^{(i)}). \tag{5.8}$$

The importance density $q(\alpha_n|\alpha_{n-1}, y_n)$ is often referred to as the optimal importance density since it minimises the Monte Carlo variance of the importance weights: for the specific case where $q(\alpha_n|\alpha_{n-1}, y_n) = P(\alpha_n|\alpha_{n-1}, y_n)$, the optimal importance density, the variance of the importance weights is zero (Doucet et al., 2000, Section D).

In practice the optimal importance density $q(\alpha_n|\alpha_{n-1}, y_n)$ proves intractable and alter-

Algorithm 2 Bootstrap Filter

For time t=1

for $i = 1, ..., N$

- Sample $\tilde{\alpha}_1^{(i)} \sim P(\alpha_1)$.
- Compute weights $\tilde{w}_1^{(i)} = g(y_1|\tilde{\alpha}_1^{(i)})$.
- Normalize weights: $w_1^{(i)} = \tilde{w}_1^{(i)} / \sum_{i=1}^{N} \tilde{w}_1^{(j)}$
- Compute filtering estimate: $\bar{h}_1 = \sum_{i=1}^{N} w_1^{(i)} \tilde{\alpha}_1^{(i)}$

For time t=2, ..., n:

- Resample: select N particle indices $j_i \in \{1, ..., N\}$ using normalized weights: $w_{t-1}^{(i)} = \frac{\tilde{w}_{t-1}^{(i)}}{\sum_{i=1}^{N} \tilde{w}_{t-1}^{(i)}}$
- Set $\alpha_{t-1}^{(i)} = \tilde{\alpha}_{t-1}^{(j_i)}$ and $w_{t-1}^{(i)} = 1/N$ for $i = 1, ..., N$.

for $i = 1, ..., N$

- Sample $\tilde{\alpha}_t^{(i)} \sim P(\tilde{\alpha}_t^{(i)}|\alpha_{t-1}^{(i)})$
- Compute weights: $\tilde{w}_t^{(i)} = g(y_t|\tilde{\alpha}_t^{(i)})$
- Normalize weights: $w_t^{(i)} = \tilde{w}_t^{(i)} / \sum_{i=1}^{N} \tilde{w}_t^{(j)}$
- Compute filtering estimate: $\bar{h}_t = \sum_{i=1}^{N} w_t^{(i)} \tilde{\alpha}_t^{(i)}$

natives are sought. Gordon et al. (1993) proposed use of the state transition density $P(\alpha_n|\alpha_{n-1})$ as part of the Bootstrap Filter, summarised in Algorithm 2. Since the transition density is known generally, it is possible to sample from it for a large number of cases and the subsequent simplified form of the importance weights has ensured the popularity and widespread application of the Bootstrap Filter.

5.2 Resampling

Sample degeneracy occurs when after a few iterations, almost all the particles have negligible weights (Arulampalam et al., 2002). A key advantage of the resampling present in the original Bootstrap Filter is that it removes particles with low weights with a high probability and hence helps solve sample degeneracy (Doucet and Johansen, 2009).

Resampling can be seen to provide stability in the future at the cost of a short term increase in the Monte Carlo variance (Doucet and Johansen, 2009). Consequently it's considered more sensible to resample when the variance of the importance weights is greater than a specified threshold like Liu and Chen (1995), who suggested the effective sample size (ESS) to assess variance of importance weights, where:

$$\text{ESS} = \left(\sum_{i=1}^{N} (w_t^{(i)})^2 \right)^{-1}, \tag{5.9}$$

and resampling when $\text{ESS} < k \cdot N$ with $k = 0.75$ or $k = 0.5$ typically.

Degeneracy is often the pivotal factor when determining whether an SMC algorithm works in practice and as a result the ESS proves popular in assessing peformance since it returns a value between 1 and N which is intuitive and provides a way to check if a sufficient number of particles remain. When the majority of particles have neglible weights, the ESS is small and similarly when the particles have relatively uniform weights, the ESS is close to N.

As a general rule, Ristic et al. (2004), Cappé et al. (2007), Doucet and Johansen (2009)

and Durbin and Koopman (2012) advise to compute estimates such as the filtering estimate before resampling since the accuracy can only drop due to Monte Carlo variation upon resampling. Chopin (2004) shows the estimate computed before resampling to be more efficient and therefore preferable.

5.2.1 Resampling Schemes

Although multinomial resampling is often assumed for simplicity, other popular and potentially better options exist. Douc et al. (2005) provide a comprehensive comparison and show residual and stratified resampling to have a lower conditional variance than multinomial resampling for all configurations of weights and show systematic, stratified and residual resampling to provide comparable results in practice.

Nevertheless, systematic resampling proves to be the most popular, despite lack of rigorous theoretical support, owing to its simplicity to implement and potential to outperform multinomial resampling generally. Cappé et al. (2005) provides a thorough account of each of the algorithms.

5.3 Particle Smoothing

Godsill, Doucet and West (2004) provide a natural extension to the Kalman smoother which generalizes for nonlinear, non-Gaussian state space models taking the joint

smoothing distribution factorized as:

$$P(\alpha_{1:n}|y_{1:n}) = P(\alpha_n|y_{1:n}) \prod_{t=1}^{n-1} P(\alpha_t|\alpha_{t+1:n}, y_{1:n}), \tag{5.10}$$

where using a Markovian assumption,

$$P(\alpha_t|\alpha_{t+1:n}, y_{1:n}) = P(\alpha_t|\alpha_{t+1}, y_{1:t}) \tag{5.11}$$

$$= \frac{P(\alpha_t|y_{1:t})f(\alpha_{t+1}|\alpha_t)}{P(\alpha_{t+1}|y_{1:t})} \tag{5.12}$$

$$\propto P(\alpha_t|y_{1:t})f(\alpha_{t+1}|\alpha_t). \tag{5.13}$$

Assuming filtering has already been carried out with approximate representations of $P(\alpha_t|y_{1:t})$ available for $t = 1, ..., n$, the particle smoother, summarised in Algorithm 3, follows intuitively from (5.13). As Godsill, Doucet and West (2004) note, this method is independent of the filtering algorithm employed and further independent realizations can be obtained by repeating the algorithm. The computational complexity however is $O(Nn)$ for each random realization and hence the procedure is computationally intensive for a large number of simulations (Cappé et al., 2007).

5.4 Prediction

Doucet et al. (2000) outline an intuitive approach for obtaining approximations of the prediction distribution as summarised in Algorithm 4.

Algorithm 3 Particle Smoother

For time t=1, ..., n:

- Run Particle Filter, storing the particles and weights for each time step.
- Choose $\tilde{\alpha}_n = \alpha_n^{(i)}$ with probability $w_n^{(i)}$

For time t=$n-1, ..., 1$:

- Calculate and normalize modified weights $\rho_t^{(i)} \propto w_t^{(i)} f(\tilde{\alpha}_{t+1}|\alpha_t^{(i)})$ for $i = 1, ..., N$.
- Choose $\tilde{\alpha}_t = \alpha_t$ with probability $\rho_t^{(i)}$.
- $\tilde{\boldsymbol{\alpha}}_{1:n} = (\tilde{\alpha_1}, \tilde{\alpha_2}, ..., \tilde{\alpha_n})$ is then an approximate independent realization from $P(\alpha_{1:n}|y_{1:n})$.

Algorithm 4 P-step ahead prediction

Run particle filter up till time t storing normalized weights $w_t^{(i)}$ for $i = 1, ..., N$.

For j=1, ..., p:

- For $i = 1, ..., N$ sample $\alpha_{t+j}^{(i)} \sim P(\alpha_{t+j}|\alpha_{t+j-1}^{(i)})$ and extend the trajectory $\alpha_{1:t}^{(i)} = (\alpha_1^{(i)}, \alpha_2^{(i)}, ..., \alpha_t^{(i)})$ with $\alpha_{1:t+j}^{(i)} = (\alpha_{1:t+j-1}^{(i)}, \alpha_{t+j}^{(i)})$.
- Compute filtering estimate $\bar{h}_{t+j} = \sum_{i=1}^{N} w_t^{(i)} \alpha_{t+j}^{(i)}$

5.5 Missing Data

The modifications to the Kalman filtering recursions outlined in section 4.2.3, for the case when there's missing data, are based on one-step ahead predictions: when handling missing data, SMC methods can be extended with a similar approach using Algorithm 4.

Chapter 6

Dynamic Longitudinal Survival Analysis

This chapter presents a novel application of SMC methods to survival analysis via a Dynamic Longitudinal Survival (DLS) model building on the Kalman Filtering techniques and SMC methods introduced in earlier chapters.

6.1 Outline of Model

The primary aim of the model is to extract an ageing signal from activity measurements, observed over the lifetime of a population of flies, in order to estimate survival chances dynamically.

Setting up a state-space model where a latent ageing process $\alpha_{i,t}$ denotes the ageing

process on day t for fly i, $y_{i,t}$ denotes the longitudinal observation on day t for fly i, $s_{i,t}$ denotes a Bernoulli survival indicator with probability $p_{i,t} = p_t(\alpha_{i,t})$ for fly i and $\alpha_{i,1} \sim N(a_1, P_1)$ denotes the initial distribution of the latent ageing process for fly i and T, Q, H, Z, a_1 and P_1 define the population level parameters:

$$\alpha_{i,t+1} = T\alpha_{i,t} + \eta_{i,t}, \qquad \eta_{i,t} \sim N(0, Q), \tag{6.1}$$

$$y_{i,t} = Z\alpha_{i,t} + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim N(0, H), \tag{6.2}$$

$$s_{i,t} \sim \text{Bernoulli}(p_{i,t}) \tag{6.3}$$

$$\alpha_{i,1} \sim N(a_1, P_1). \tag{6.4}$$

6.2 Ageing Latent Process

6.2.1 Random Walk

The aim of (6.1) is to postulate a latent ageing process. A reasonable first consideration is the random walk evolution for ageing like (Woodbury and Manton, 1977, Dasgupta, 1994, Penna and Stauffer, 1995) and Pavlov (2010): considering a drift term additionally allows incorporation of biological decay and damage accumulation and has been considered extensively in stochastic models of ageing – see Weitz and Fraser (2001) and Li and Anderson (2009).

A simple random walk can be modelled by letting T and Q equal 1. A model of this form incorporates 'disposable soma' biological theory since it allows the possibility of maintenance and repair alongside damage accumulation guided by the data.

6.2.2 Random Walk with Drift

To incorporate drift, the local linear trend model setup popularised by Harvey (1993) and Durbin and Koopman (2012) proves appealing whereby the latent process becomes:

$$\alpha_{i,t+1} = \alpha_{i,t} + v_{i,t} + \eta_{i,t}, \qquad \eta_{i,t} \sim N(0, Q_1) \tag{6.5}$$

$$v_{i,t+1} = v_{i,t} + \omega_{i,t}, \qquad \omega_{i,t} \sim N(0, Q_2), \tag{6.6}$$

or in matrix notation:

$$\begin{pmatrix} \alpha_{i,t+1} \\ v_{i,t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{i,t} \\ v_{i,t} \end{pmatrix} + \begin{pmatrix} \eta_{i,t} \\ \omega_{i,t} \end{pmatrix}, \tag{6.7}$$

where:

$$\begin{pmatrix} \eta_{i,t} \\ \omega_{i,t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \right), \tag{6.8}$$

and:

$$\begin{pmatrix} \alpha_{i,1} \\ v_{i,1} \end{pmatrix} \sim N\left(\begin{pmatrix} a_{1,1} \\ a_{2,1} \end{pmatrix}, \begin{pmatrix} Q_{1,1} & 0 \\ 0 & Q_{2,1} \end{pmatrix} \right). \tag{6.9}$$

Ageing is assumed to start at 0 and hence $a_{1,1}$ and $Q_{1,1}$ are set to zero for initialisation. A constant drift can be incorporated by specifying a value for $a_{2,1}$ and setting $Q_{2,1}$ and Q_2 to 0. Alternatively, setting $v_{i,1}$ to a constant d with $Q_1 = Q$ and $Q_2 = 0$ in (6.6)

achieves the same result: based on this, conditional densities can be expressed as:

$$\begin{aligned} P(\alpha_{i,t}|\alpha_{i,t-1}) &\sim N(T\alpha_{i,t-1} + d, Q), \\ P(\alpha_{i,t}) &\sim N(td, tQ), \\ P(y_{i,t}|\alpha_{i,t}) &\sim N(Z\alpha_{i,t}, Z^2Q + H), \\ P(y_{i,t}) &\sim N(Ztd, Z^2tQ + H). \end{aligned} \tag{6.10}$$

As a result when computing the likelihood based on (6.10), the parameter Z can be increased and the drift d decreased without changing overall effect of the two. Hence, to ensure identifiability it's necessary to set one of them to a fixed value since there are two trend parameters.

Similarly, Durbin and Koopman (2012) show using matrix notation for state space models, $y_{i,t} = Z\alpha_{i,t} + \epsilon_{i,t} = Z(T\alpha_{i,t-1} + d + \eta_{i,t-1}) + \epsilon_{i,t}$ and as a result, ideally only one trend component should be specified across the latent and observation process to ensure parameters are identifiable.

Since ageing is assumed to be a deteriorative process which increases mortality risk over time, it makes sense to consider a drift term which is analogous to ageing being pushed to a terminal level broadly rather than a simple random walk which is more speculative. This ties up with damage accumulation and wear and tear theories of ageing.

In order to specify a stochastic local linear trend, both Q_1 and Q_2 can be set to values greater than 0. The local linear trend however, as Durbin and Koopman (2012) note, can appear far from smooth since the trend is at a local level. Young et al. (1991) and

Kitagawa and Gersch (1996) advocate modelling trend in state space models through the 'integrated random walk' or 'smoothness prior' approach respectively, whereby Q_1 is set to 0 with $Q_2 > 0$, to achieve a smoother trend. Biological theories suggest a gradual development of the ageing process over time (Strehler, 1977, Arking, 2006) which implies a smoother development of the trend and consequently deem the approach of Young et al. (1991) and Kitagawa and Gersch (1996) preferable.

6.2.3 Ornstein-Uhlenbeck Process

An alternative consideration as a stochastic process for the latent ageing process is the Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930), which has been considered extensively in ageing studies (Woodbury and Manton, 1977, Yashin et al., 2000, 2008, 2016). Aalen and Gjessing (2004) describe it as a natural model to use for an underlying process in biological contexts, since it stabilises around some equilibrium point.

Yashin et al. (2016) further draw attention to its appeal for biological interpretation from an ageing biodemographic perspective since it's the only nontrivial process which is stationary, Gaussian and Markovian. Touching upon the ability of the OU process to drift towards a long term mean, which can be interpreted as homeostatic regulation[1], Yashin et al. (2016) single out the mean-reverting property of the OU process for endorsing it as the method of choice for modelling ageing trajectories.

[1] "...a fundamental property of living organisms that tends to maintain stability by returning or restoring a biological subsystem to an equilibrium state in case of disturbances of various kinds."(Yashin et al., 2016)

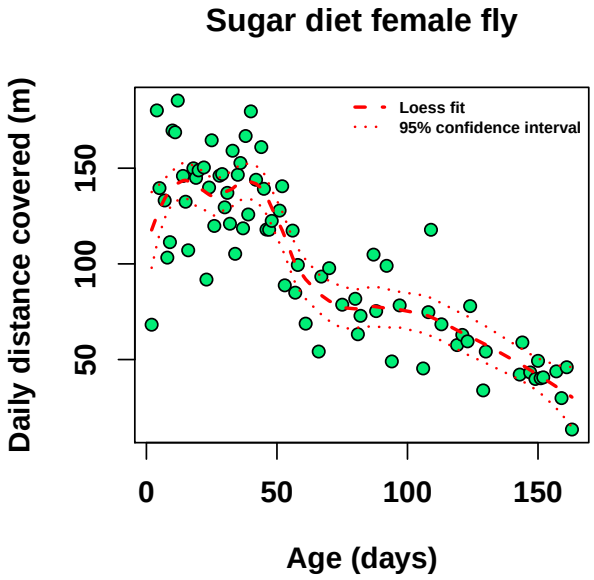

Figure 6.1: Lifetime actitivity plot

The discrete-time analog of the OU process is the stationary autoregressive order 1 (AR1) process of the form (6.1) with $|T| < 1$: restricting the marginal variance of $\alpha_{i,t}$ to be 1 gives $Q = 1 - T^2$.

6.3 Longitudinal Observations

The longitudinal data available includes daily activity levels and a daily survival indicator. Chiu et al. (2013) considering fruit-fly data, highlight how activity patterns are among the physiological processes that change most significantly as animals age and as a result provide good indicators of healthspan. Bearing this in mind, activity patterns—see Figure 6.1—considered in section 3.3.2 are re-explored. An absolute value difference in the daily activity level relative to the peak activity level observed

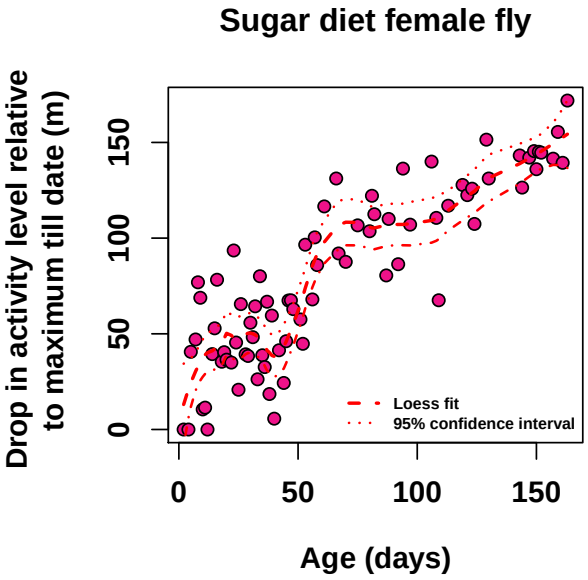

Figure 6.2: Drop in activity level over lifetime

to date, over the lifetime of a fly can be assumed to be governed by a latent ageing process: considering the absolute value drops potentially provides more insight into possible ageing effects as depicted in Figure 6.2 for the same fly from Figure 6.1. The zeroes denote when a new peak activity level is reached.

The activity level drops together with survival indicators will be used in model as longitudinal information.

6.3.1 Daily Survival Probability

The Bernoulli survival indicators are distributed with a probability $p_{i,t}$ and estimated as outlined in section 4.4. Pavlov (2010) assumed a baseline Gompertz hazard function based on its popularity in ageing studies: a baseline hazard guided by the data can

provide a better choice.

With a frailty multiplicative hazard being fitted, the baseline hazard reflects the population level of risk. Earlier in Figure 3.12 a significant difference in survival together with a non-proportional hazard was observed for the full and sugar diet flies: this motivates a separate baseline hazard and frailty group for each diet initially at least.

6.4 Baseline Hazard Estimation

The baseline hazard can be estimated assuming parametric distributions such as the Gompertz, Weibull and Log-Logistic, which collectively cover a wide range of forms the hazard can take: the parameter estimates, alongside providing potential starting values for optimisation routines later, will help evaluate whether a particular choice of baseline hazard fits but otherwise this stage is exploratory – the estimates observed here will not be used later, where full joint estimation will be carried out.

6.4.1 Gompertz Distribution

The Gompertz distribution is arguably the most popular distribution in ageing studies and biodemography, proving to be the catalyst in areas of research such as frailty modelling.

For the purposes of fitting a baseline hazard, the survival times for the flies can reasonably be assumed to be continuous since the step size of the discretization is relatively

small (Lenart, 2014) as seen in Figures 3.14 and 3.15.

Letting the survival time x follow a Gompertz distribution with density:

$$f(x) = ae^{bx}e^{-\frac{a}{b}(e^{bx}-1)}, \qquad x > 0, \tag{6.11}$$

and letting $\theta = (a, b)$ and $Q(b) = \frac{1}{n}\sum_{i=1}^{n} e^{bx_i}$, the log-likelihood is:

$$l(\theta) = \log L(\theta) = n\log(a) + b\sum_{i=1}^{n} x_i - \frac{a}{b}\sum_{i=1}^{n}(e^{bx_i} - 1) \tag{6.12}$$

$$= n\log(a) + b\sum_{i=1}^{n} x_i - \frac{na}{b}(Q(b) - 1). \tag{6.13}$$

The partial derivatives are then given by:

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \frac{nQ(b)}{b} + \frac{n}{b} \tag{6.14}$$

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n} x_i - \frac{na}{b^2} - \frac{naQ'(b)}{b} + \frac{naQ(b)}{b^2} \tag{6.15}$$

By setting 6.14 to zero, a maximum likelihood estimate (MLE) for a can be derived as:

$$\hat{a} = \frac{b}{Q(b) - 1} \tag{6.16}$$

By plugging the MLE for a into 6.15:

$$\begin{aligned}\frac{\partial l}{\partial b} &= \sum_{i=1}^{n} x_i + \frac{n(Q(b) - bQ'(b) - 1)}{b(Q(b) - 1)} \\ &= \sum_{i=1}^{n} x_i + \frac{n}{b} - \frac{nQ'(b))}{(Q(b) - 1)} \\ &= \bar{x} + \frac{1}{b} - \frac{Q'(b))}{Q(b) - 1}\end{aligned} \tag{6.17}$$

which can be solved using numerical methods. Lenart (2014) presents a similar solution

for the MLE. Asymptotic variances can be derived using:

$$\frac{\partial^2 l}{\partial a^2} = \frac{-n}{a^2} \tag{6.18}$$

$$\frac{\partial^2 l}{\partial b^2} = \frac{[Q'(b)]^2 - Q''(b)(Q(b)-1)}{(Q(b)-1)^2} - \frac{1}{b^2} \tag{6.19}$$

However, given the small sample sizes for both diets, the bootstrap algorithm for standard errors (Efron and Tibshirani, 1994) is preferable.

6.4.2 Weibull Distribution

If the survival time x is assumed to follow a Weibull distribution with pdf:

$$f(x) = ab^a x^{a-1} e^{-(bx)^a}, \qquad x > 0, \tag{6.20}$$

then the log-likelihood is given by:

$$l(\theta) = n \log a + na \log b - b^a \sum_{i=1}^{n} x_i^a + (a-1) \sum_{i=1}^{n} \log x_i, \tag{6.21}$$

where $\theta = (a, b)$ and the partial derivatives are given by:

$$\frac{\partial l}{\partial b} = \frac{na}{b} - ab^{a-1} \sum_{t=1}^{n} x_i^a \tag{6.22}$$

$$\frac{\partial l}{\partial a} = \frac{n}{a} + n \log b - b^a \sum_{i=1}^{n} (x_i^a \log x_i) - b^a \log(b) \sum_{i=1}^{n} x_i^a + \sum_{i=1}^{n} \log x_i \tag{6.23}$$

which yields a MLE for b as:

$$\hat{b} = \left(\frac{n}{\sum_{i=1}^{n} x_i^a} \right)^{\frac{1}{a}}, \tag{6.24}$$

which can be plugged into (6.23) for b to give:

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \frac{n}{\sum_{i=1}^{n} x_i^a} \sum_{i=1}^{n} (x_i^a \log x_i) + \sum_{i=1}^{n} \log x_i \tag{6.25}$$

Balakrishnan and Kateri (2008) consider maximum likelihood estimation for a Weibull pdf of the form (6.20) where their parameter $\theta = b^{-1}$ and $\beta = a$. The solution to the score equation (6.25) can be expressed as:

$$\frac{1}{a} = \frac{\sum_{i=1}^{n} (x_i^a \log x_i)}{\sum_{i=1}^{n} x_i^a} - \frac{1}{n} \sum_{i=1}^{n} \log x_i, \tag{6.26}$$

where the right hand side of (6.26) denotes $H(a; \boldsymbol{x})$. Balakrishnan and Kateri (2008) using the Cauchy-Schwartz inequality show that $H(a; \boldsymbol{x})$ is a monotone increasing function of a. Consequently a plot of the left hand side of (6.26) against the right hand side will intersect exactly once, at the MLE of a (Balakrishnan and Kateri, 2008).

Asymptotic variances can be derived using:

$$\frac{\partial^2 l}{\partial b^2} = \frac{-na}{b^2} - a(a-1)b^{a-2} \sum_{i=1}^{n} x_i^a \tag{6.27}$$

$$\frac{\partial^2 l}{\partial a^2} = -\frac{n}{a^2} - 2b^a \log b \sum_{i=1}^{n} (x_i^a \log x_i) - b^a \sum_{i=1}^{n} \left(x_i^a [\log x_i]^2 \right) - b^a [\log b]^2 \sum_{i=1}^{n} x_i^a \tag{6.28}$$

6.4.3 Log-Logistic Distribution

If the survival time x is assumed to follow a Log-Logistic distribution with pdf:

$$f(x) = \frac{a b^a x^{a-1}}{(1 + (bx)^a)^2}, \qquad x > 0, \tag{6.29}$$

then the log-likelihood which is usually maximised directly is given by:

$$l(\theta) = n\log a + na\log b + (a-1)\sum_{i=1}^{n}\log x_i - 2\sum_{i=1}^{n}\log(1+(bx_i)^a), \tag{6.30}$$

where $\theta = (a, b)$. Alternatively, the partial derivatives are:

$$\frac{\partial l}{\partial b} = \frac{na}{b} - \frac{2a}{b}\sum_{t=1}^{n}\frac{bx_i^a}{1+(bx_i)^a} \tag{6.31}$$

$$\frac{\partial l}{\partial a} = \frac{n}{a} + n\log b + \sum_{i=1}^{n}\log x_i - 2\sum_{i=1}^{n}\frac{(bx_i)^a\log(bx_i)}{1+(bx_i)^a}. \tag{6.32}$$

Al-Shomrani et al. (2016) presents a solution for the log-logistic MLE, where b is taken to be λ^{-1} in their alternative parameterisation of the distribution. Asymptotic variances can be derived using:

$$\frac{\partial^2 l}{\partial b^2} = -\frac{na}{b^2} + \frac{2a}{b^2}\sum_{i=1}^{n}\frac{(bx_i)^a(1+(bx_i)^a-a)}{(1+(bx_i)^a)^2} \tag{6.33}$$

$$\frac{\partial^2 l}{\partial a^2} = -\frac{n}{a^2} - 2\sum_{i=1}^{n}\frac{(bx_i)^a[\log(bx_i)]^2}{(1+(bx_i)^a)^2} \tag{6.34}$$

6.4.4 Estimates

Maximum likelihood estimates for the different baseline hazards along with bootstrap standard errors are calculated for the fly data as summarised in Tables 6.1 to 6.3. The observed median for female flies on a full diet is 131.5, 97.0 for female flies on a sugar diet and 110.1 for female flies overall from Table 3.3 and comparing these to the fitted medians in Tables 6.1 to 6.3 suggests the Gompertz as a good fit for the full diet female flies and the Weibull and Log-Logistic better for the sugar diet females flies and females overall. Checking graphically through Figures 6.3 and 6.4 show the Gompertz

and Weibull proving a good fit for the full diet survival times and the Weibull and Log-Logistic a better fit for the sugar diet flies. Figure 6.5 shows the Weibull proving further a good fit for the female flies overall with the Log-Logistic not fitting well towards right tail. Given the small sample the conclusions that can be drawn are limited and including frailty can potentially help improve fit.

To limit the number of potential ageing models, the Log-Logistic model is omitted hereon since it fits similarly to the Weibull and the Weibull exhibits a better fit in the right tail as seen for the sample overall in Figure 6.5.

Table 6.1: Gompertz parameter estimates

Diet	**Parameter**	**Estimate**	**Bootstrap S.E**	**Sample Size**	**Est. Median**	**Log-Lik.**
Full	a	1.12×10^{-3}	7.02×10^{-7}	16	128.6	-84.7
Full	b	0.0202	3.16×10^{-5}			
Sugar	a	5.63×10^{-4}	8.64×10^{-8}	16	111.1	-77.4
Sugar	b	0.0338	7.25×10^{-5}			
Overall	a	1.18×10^{-3}	1.93×10^{-7}	32	119.1	-165.7
Overall	b	0.0222	1.23×10^{-5}			

Table 6.2: Weibull parameter estimates

Diet	**Parameter**	**Estimate**	**Bootstrap S.E**	**Sample Size**	**Est. Median**	**Log-Lik.**
Full	a	2.71	0.504	16	121.2	-85.3
Full	b	7.21×10^{-3}	5.15×10^{-7}			
Sugar	a	4.19	0.684	16	108.8	-76.0
Sugar	b	8.43×10^{-3}	2.99×10^{-7}			
Overall	a	3.00	0.142	32	114.9	-164.3
Overall	b	7.39×10^{-3}	1.85×10^{-7}			

The suitability of the Weibull can also be explored through taking the empirical survival curve and plotting as in Figure 6.6 which should exhibit linearity broadly if the Weibull holds. However, given the extremely small samples although there's some linearity

Table 6.3: Log-Logistic parameter estimates

Diet	Parameter	Estimate	Bootstrap S.E	Sample Size	Est. Median	Log-Lik.
Full	a	3.44	1.00	16	117.4	-87.5
Full	b	8.52×10^{-3}	1.53×10^{-6}			
Sugar	a	6.85	2.84	16	104.1	-75.5
Sugar	b	9.60×10^{-3}	3.67×10^{-7}			
Overall	a	4.44	0.545	32	110.9	-166.2
Overall	b	9.02×10^{-3}	4.20×10^{-7}			

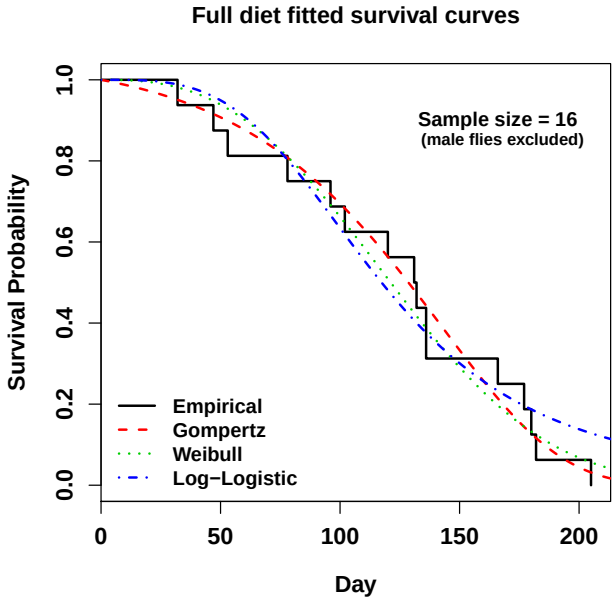

Figure 6.3: Full diet fitted survival curves

the overall relationship is not easy to ascertain without additional survival times as advocated in Pletcher (1999).

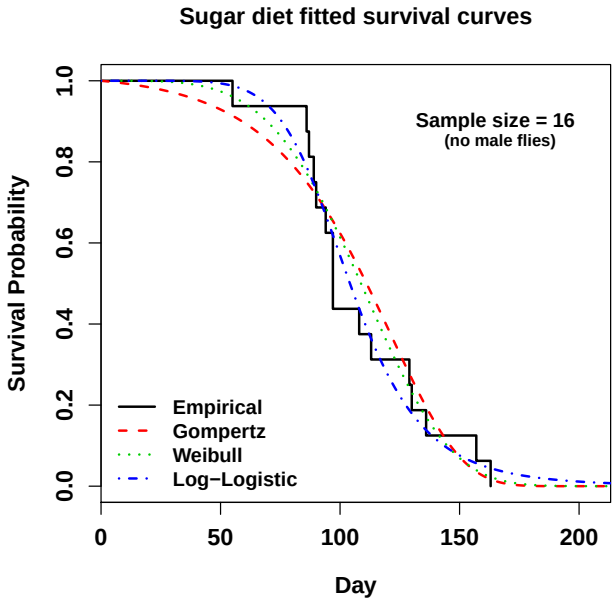

Figure 6.4: Sugar diet fitted survival curves

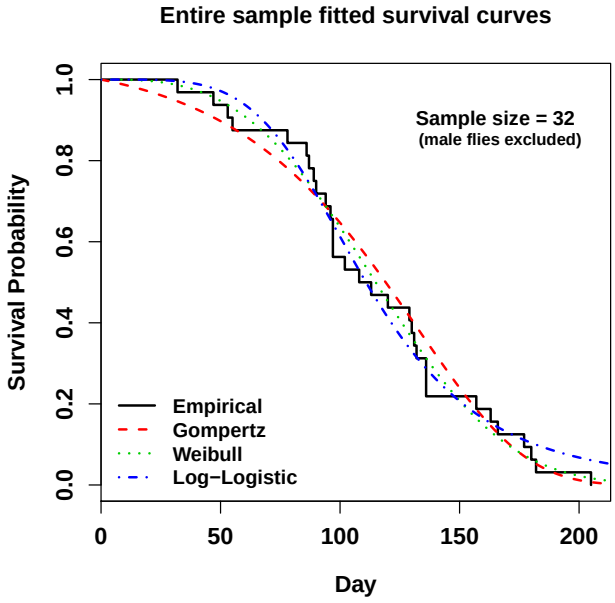

Figure 6.5: Entire sample fitted survival curves

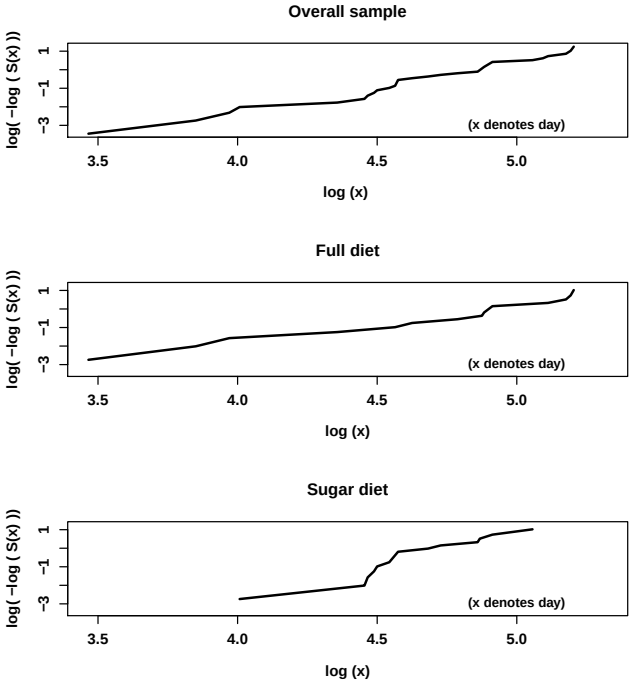

Figure 6.6: Weibull fit check

6.5 Dynamic Longitudinal Survival Model Likelihood

(6.1) to (6.4) define the structure of the DLS model. Let $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n_i})$ denote the ageing process for fly i, $y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,n_i})$ and $s_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,n_i})$ denote the observations for fly i, where n_i defines the number of observations for fly i: further, let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)$, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ and $\boldsymbol{s} = (s_1, s_2, ..., s_n)$, where n denotes the sample size (the total number of flies). Then assuming conditional independence of the observations $\boldsymbol{y}$ and $\boldsymbol{s}$ given the latent ageing process $\boldsymbol{\alpha}$, the likelihood for a parameter vector ψ can be written as:

$$\begin{aligned}
L(\psi; \boldsymbol{y}, \boldsymbol{s}) &= \prod_{i=1}^{n} \left[\int P(y_i, s_i|\alpha_i) P(\alpha_i)\, d\alpha_i \right] \\
&= \prod_{i=1}^{n} \left[\int P(y_i|\alpha_i) P(s_i|\alpha_i) P(\alpha_i)\, d\alpha_i \right] \\
&= \prod_{i=1}^{n} \left[\int \frac{P(\alpha_i|y_i)\, P(y_i)}{P(\alpha_i)} P(s_i|\alpha_i) P(\alpha_i)\, d\alpha_i \right] \\
&= \prod_{i=1}^{n} \left[P(y_i) \int P(s_i|\alpha_i)\, P(\alpha_i|y_i)\, d\alpha_i \right] \\
&= \prod_{i=1}^{n} \left[P(y_i)\, E_{\alpha_i|y_i} \left[P(s_i|\alpha_i) \right] \right]. \qquad (6.35)
\end{aligned}$$

The log-likelihood then becomes:

$$\log L(\psi; \boldsymbol{y}, \boldsymbol{s}) = \sum_{i=1}^{n} \left[\log P(y_i) + \log E_{\alpha_i|y_i} \left[P(s_i|\alpha_i) \right] \right] \qquad (6.36)$$

The flexibility of state-space models allows each individual to have his own parameters for the longitudinal observations in (6.2) and (6.35) generalises to allow that. Since α_i

and y_i are Gaussian they can be evaluated using (4.13). Assuming a log-normal frailty throughout, $P(s_i|\alpha_i)$ will depend on the choice of baseline hazard. The hazard function taking into account a multiplicative frailty, in the general case, can therefore be defined as:

$$h(t|\alpha_{i,t}) = h_0(t)\, e^{\gamma\alpha_{i,t}} \tag{6.37}$$

where $h_0(t)$ is the baseline hazard.

The fly data comprises flies following 2 different diets: a log-likelihood reflecting a diet-specific frailty can be expressed as,

$$\begin{aligned}\log L(\psi; \boldsymbol{y}, \boldsymbol{s}) = &\sum_{i=1}^{n_f} \Big[\log P(y_i) + \log E_{\alpha_i|y_i}\left[P(s_i|\alpha_i;\gamma)\right]\Big] \\ &+ \sum_{i=1}^{n_s} \Big[\log P(y_i) + \log E_{\alpha_i|y_i}\left[P(s_i|\alpha_i;\gamma_s)\right]\Big],\end{aligned} \tag{6.38}$$

where n_f and n_s denote the sample sizes for the full-diet and sugar-diet flies and γ denotes full-diet frailty coefficient and sugar-diet frailty coefficient $\gamma_s = \gamma + c$, where c is a constant factor.

6.5.1 Gompertz

For the case when a Gompertz baseline hazard is assumed:

$$\begin{aligned}h(t|\alpha_{i,t}) = h_0(t)\, e^{\gamma\alpha_{i,t}} &= ae^{bt}\, e^{\gamma\alpha_{i,t}}, \\ &= ae^{bt+\gamma\alpha_{i,t}}.\end{aligned} \tag{6.39}$$

The Bernoulli survival probability $p_{i,t}$ for $t = 1, ..., n_i$ following (4.19) and (4.20) is:

$$p_{i,t} = p_t(\alpha_{i,t}) = \exp\left\{-ae^{bt+\gamma\alpha_{i,t}}\right\}. \tag{6.40}$$

Hence:

$$\begin{aligned} P(s_i|\alpha_i) &= \left(\prod_{t=1}^{n_i-1} p_{i,t}\right)(1 - p_{i,n_i}), \\ &= \left(\prod_{t=1}^{n_i-1} \exp\left\{-ae^{bt+\gamma\alpha_{i,t}}\right\}\right)\left(1 - \exp\left\{-ae^{bn_i+\gamma\alpha_{i,n_i}}\right\}\right), \\ &= \exp\left\{-a\sum_{t=1}^{n_i-1} e^{bt+\gamma\alpha_{i,t}}\right\}\left(1 - \exp\left\{-ae^{bn_i+\gamma\alpha_{i,n_i}}\right\}\right). \end{aligned} \tag{6.41}$$

Following the approach outlined on page 48, using (6.41) the $E_{\alpha_i|y_i}[P(s_i|\alpha_i)]$ term can be evaluated using Monte-Carlo integration:

$$E_{\alpha_i|y_i}[P(s_i|\alpha_i)] \approx \frac{1}{M}\sum_{j=1}^{M}\left[\exp\left\{-a\sum_{t=1}^{n_i-1} e^{bt+\gamma\alpha_{i,t}^{(j)}}\right\}\left(1 - \exp\left\{-ae^{bn_i+\gamma\alpha_{i,n_i}^{(j)}}\right\}\right)\right], \tag{6.42}$$

where $\alpha_{i,t}^{(j)}$, for $j = 1, \ldots, M$ denote independent and identically distributed draws from $P(\alpha_i|y_i)$. A more computationally stable form can be deduced by considering (6.42) as:

$$\begin{aligned} E_{\alpha_i|y_i}[P(s_i|\alpha_i)] &\approx \frac{1}{M}\sum_{j=1}^{M}\left[e^{L_i}e^{-L_i}\exp\left\{-a\sum_{t=1}^{n_i-1} e^{bt+\gamma\alpha_{i,t}^{(j)}}\right\}\left(1 - \exp\left\{-ae^{bn_i+\gamma\alpha_{i,n_i}^{(j)}}\right\}\right)\right], \\ &= \frac{e^{-L_i}}{M}\sum_{j=1}^{M}\left[\exp\left\{L_i - a\sum_{t=1}^{n_i-1} e^{bt+\gamma\alpha_{i,t}^{(j)}}\right\}\left(1 - \exp\left\{-ae^{bn_i+\gamma\alpha_{i,n_i}^{(j)}}\right\}\right)\right], \end{aligned} \tag{6.43}$$

where $L_i = \frac{1}{M}\sum_{j=1}^{M} L_{i,j}$ and:

$$L_{i,j} = a\sum_{t=1}^{n_i-1} e^{bt+\gamma\alpha_{i,t}^{(j)}}, \tag{6.44}$$

for each iteration $j = 1, ..., M$ of the Monte-Carlo integration. Hence:

$$\log E_{\alpha_i|y_i}\{P(s_i|\alpha_i)\} \approx$$
$$-L_i - \log(M) + \log\left(\sum_{j=1}^{M}\left[e^{L_i - L_{i,j}} \times \left(1 - e^{-ae^{bn_i+\gamma\alpha_{i,n}^{(j)}}}\right)\right]\right). \tag{6.45}$$

For the likelihood (6.36), (6.45) is evaluated in turn for each individual comprising sample. The likelihoods in this section follow a discrete time cumulative hazard function approximation when calculating the survival probability as outlined in (4.19). Alternatively, the survival function for the Gompertz distribution can be used directly: if T is taken to be the survival time then $S(t) = P(T > t) = e^{-\frac{a}{b}\left(e^{bt}-1\right)}$, and

$$P(T > t \,|\, T > t-1) = \exp\left\{-\frac{a}{b}\left(e^{bt} - e^{b(t-1)}\right)\right\}. \tag{6.46}$$

In the general case, if θ denotes the frailty with $\theta > 0$, the survival function conditioned on frailty becomes $S(t|\theta) = S(t)^\theta$ (Kleinbaum and Klein, 2012). Hence assuming a log-normal frailty with the Gompertz distribution, where $S(t|\alpha_{i,t}) = e^{-\frac{a}{b}\left(e^{bt}-1\right)^{e^{\gamma\alpha_{i,t}}}} = e^{-\frac{a}{b}e^{\gamma\alpha_{i,t}}\left(e^{bt}-1\right)}$, and letting $\alpha_{i,\leq t}$ denote all the ageing information available up till time t for fly i,

$$\begin{aligned}
p_{i,t} = p_t(\alpha_{i,t}) &= P(T > t|T > t-1; \alpha_{i,\leq t}), \\
&= \frac{P(T > t; \alpha_{i,\leq t})}{P(T > t-1; \alpha_{i,\leq t})}, \\
&= \frac{\exp\left\{-\frac{a}{b}e^{\gamma\alpha_{i,t}}\left(e^{bt}-1\right)\right\}}{\exp\left\{-\frac{a}{b}e^{\gamma\alpha_{i,t-1}}\left(e^{b(t-1)}-1\right)\right\}} \\
&= \exp\left\{-\frac{a}{b}\left(e^{\gamma\alpha_{i,t}}(e^{bt}-1) - e^{\gamma\alpha_{i,t-1}}(e^{b(t-1)}-1)\right)\right\}.
\end{aligned} \tag{6.47}$$

and hence:

$$P(s_i|\alpha_i) = \exp\left\{-\frac{a}{b}\sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(j)}}(e^{bt}-1) - e^{\gamma\alpha_{i,t-1}^{(j)}}(e^{b(t-1)}-1)\right)\right\} \times \left(1-\exp\left\{-\frac{a}{b}\left(e^{\gamma\alpha_{i,n_i}^{(j)}}(e^{bn_i}-1) - e^{\gamma\alpha_{i,n_i-1}^{(j)}}(e^{b(n_i-1)}-1)\right)\right\}\right). \tag{6.48}$$

Taking,

$$L_{i,j} = \frac{a}{b}\sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(j)}}(e^{bt}-1) - e^{\gamma\alpha_{i,t-1}^{(j)}}(e^{b(t-1)}-1)\right), \tag{6.49}$$

for $j = 1, \ldots, M$ and letting L_i denote the mean as before, the log-likelihood can be expressed as:

$$\log E_{\alpha_i|y_i}\{P(s_i|\alpha_i)\} \approx -L_i - \log(M) + \log\left(\sum_{j=1}^{M}\left[e^{L_i-L_{i,j}} \times \left(1-e^{-\frac{a}{b}\left(e^{\gamma\alpha_{i,n_i}^{(j)}}(e^{bn_i}-1) - e^{\gamma\alpha_{i,n_i-1}^{(j)}}(e^{b(n_i-1)}-1)\right)}\right)\right]\right). \tag{6.50}$$

To compare the performance of a model likelihood where the survival probability is based on a cumulative hazard function approximation to one where the Gompertz distribution is used directly, parameter estimates can be calculated by setting $\gamma = 0$ in (6.45) and (6.50) to recover the marginal survival likelihood: the parameter estimates for simulated data can then be compared with the true values they were simulated from to assess performance.

1000 datasets of 500 survival times each are simulated from the Gompertz distribution using the Inversion method with parameters taken to follow each of the models from Table 6.1. The cumulative hazard approximation based approach will be referred to as 'Approach 1' and the direct Gompertz survival function approach will be referred to

as 'Approach 2'. Figure 6.7 summarises the findings observed for data simulated from the full diet Gompertz model with $a = 0.000982$ and $b = 0.0199$: a similar pattern is observed for simulations from each of the 3 models. The overall estimated mean for the parameters a and b under Approach 2 is always very close to the true parameters with neglible difference in the estimate of b between Approach 1 and 2. However for the parameter a, the Approach 1 estimator is biased with Approach 2 consistently closer to the true parameter value as seen in Figures 6.7 and 6.8.

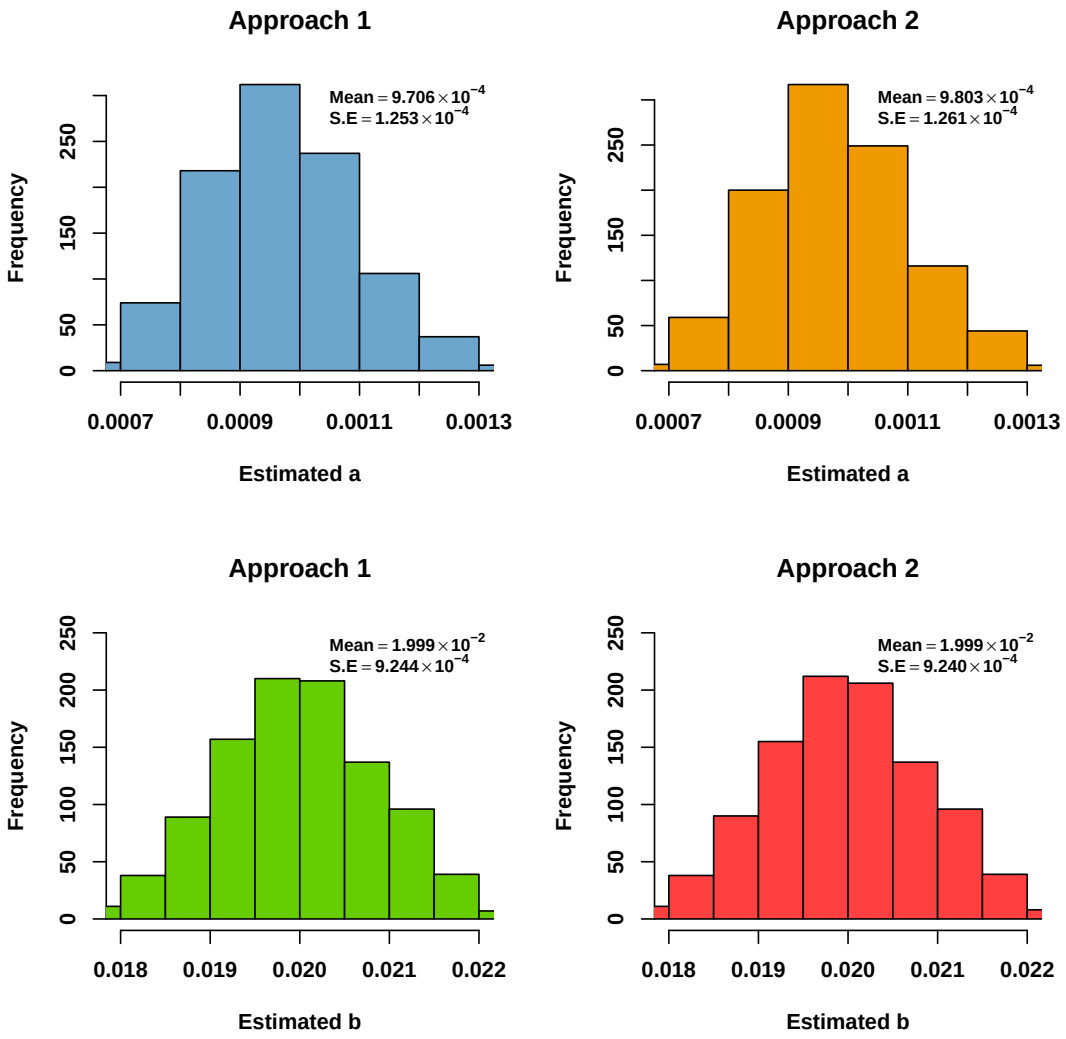

Figure 6.7: Model checking using Gompertz($a = 9.82 \times 10^{-4}$, $b = 1.99 \times 10^{-2}$)

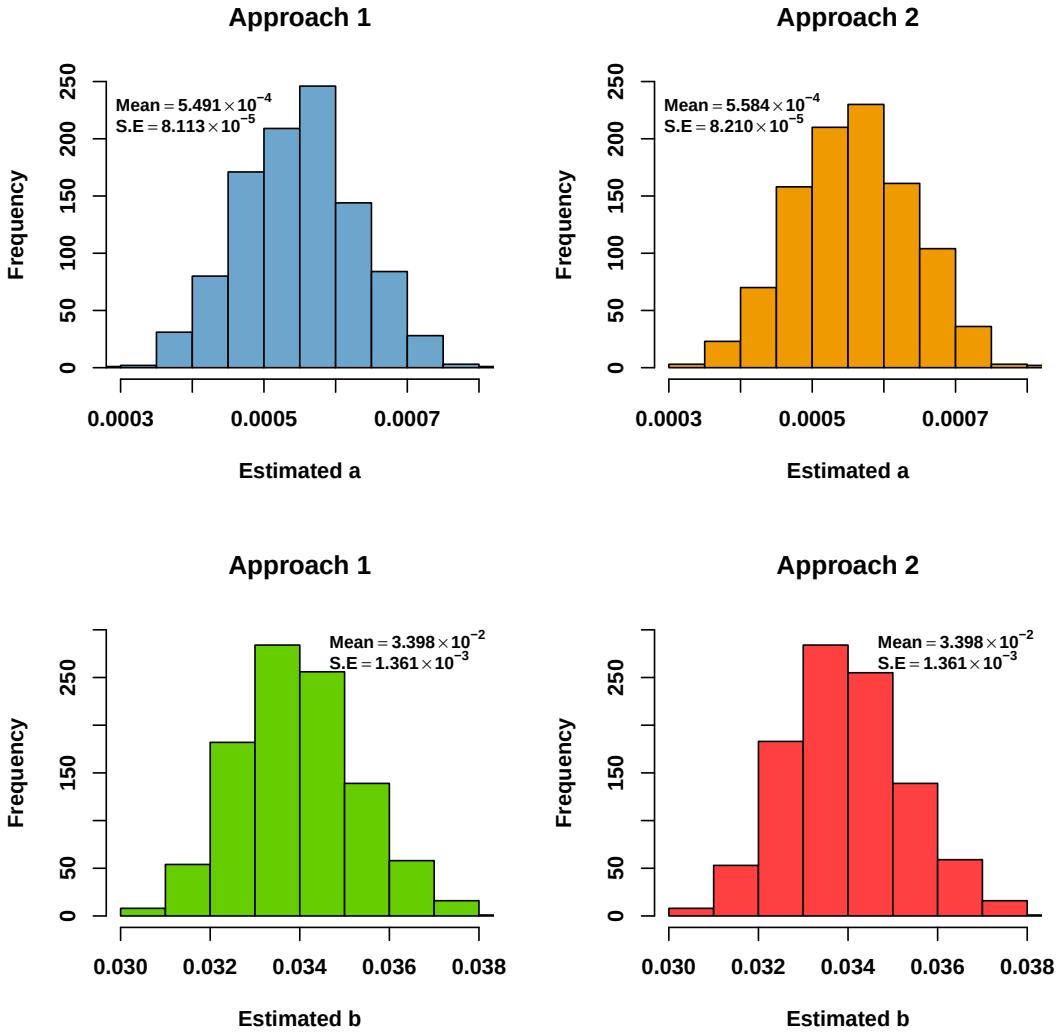

Figure 6.8: Model checking using Gompertz($a = 5.61 \times 10^{-4}$, $b = 3.38 \times 10^{-2}$)

Secondary checks to assess performance of the 2 methods versus the MLE for each of the 1000 simulated datasets shows the approximation under Approach 2 being directly comparable to the MLE as shown in Figure 6.9, with comparable standard error and percentiles. Hence the discrete time approximation taking the survival function directly is preferable and a model defined under this framework will be referred to as a dynamic

longitudinal survival (DLS) model.

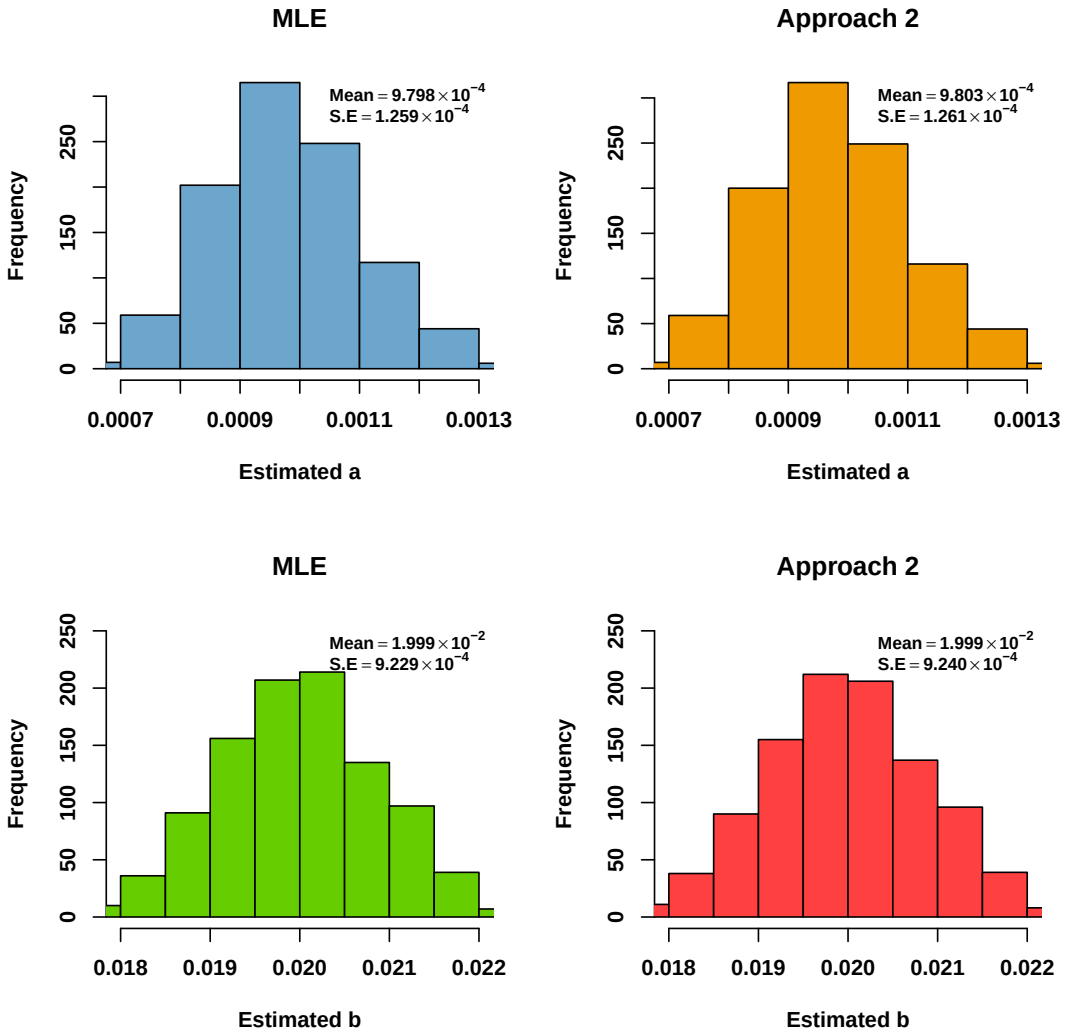

Figure 6.9: Comparison with MLE for Gompertz($a = 9.82 \times 10^{-4}$, $b = 1.99 \times 10^{-2}$)

6.5.2 Weibull

For the case when a Weibull baseline hazard is assumed, the survival function becomes:

$$S(t|\alpha_{i,t}) = \exp\left\{-(bt)^a\right\}^{e^{\gamma\alpha_{i,t}}} = \exp\left\{-e^{\gamma\alpha_{i,t}}(bt)^a\right\}. \tag{6.51}$$

The Bernoulli survival probability can then be expressed as:

$$p_{i,t} = p_t(\alpha_{i,t}) = \exp\left\{-b^a \left(e^{\gamma\alpha_{i,t}} t^a - e^{\gamma\alpha_{i,t-1}}(t-1)^a\right)\right\}, \tag{6.52}$$

for which the likelihood becomes:

$$\begin{aligned} P(s_i|\alpha_i) = \exp\left\{-b^a \sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(j)}} t^a - e^{\gamma\alpha_{i,t-1}^{(j)}}(t-1)^a\right)\right\} \\ \times \left(1 - \exp\left\{-b^a\left(e^{\gamma\alpha_{i,n_i}^{(j)}} n_i^a - e^{\gamma\alpha_{i,n_i-1}^{(j)}}(n_i-1)^a\right)\right\}\right). \end{aligned} \tag{6.53}$$

Taking,

$$L_{i,j} = b^a \sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(j)}} t^a - e^{\gamma\alpha_{i,t-1}^{(j)}}(t-1)^a\right), \tag{6.54}$$

for $j = 1, .., M$, where M denotes the number of Monte-Carlo integration draws, and letting L_i denote the mean, the log-likelihood can be expressed as:

$$\begin{aligned} \log E_{\alpha_i|y_i}\{P(s_i|\alpha_i)\} \approx -L_i - \log(M) + \\ \log\left(\sum_{j=1}^{M}\left[e^{L_i - L_{i,j}} \times \left(1 - e^{-b^a\left(e^{\gamma\alpha_{i,n_i}^{(j)}} n_i^a - e^{\gamma\alpha_{i,n_i-1}^{(j)}}(n_i-1)^a\right)}\right)\right]\right). \end{aligned} \tag{6.55}$$

6.5.3 Log-Logistic

For a Log-Logistic baseline hazard, the survival function is:

$$S(t|\alpha_{i,t}) = \left(\frac{1}{1+(bt)^a}\right)^{e^{\gamma\alpha_{i,t}}} = \exp\{-\log(1+(bt)^a)\}^{e^{\gamma\alpha_{i,t}}} \tag{6.56}$$

$$= \exp\left\{-e^{\gamma\alpha_{i,t}} \log\left(1+(bt)^a\right)\right\}, \tag{6.57}$$

The Bernoulli survival probability can then be expressed as:

$$p_{i,t} = p_t(\alpha_{i,t}) = \exp\left\{-[e^{\gamma\alpha_{i,t}}\log(1+b^a t^a) - e^{\gamma\alpha_{i,t-1}}\log(1+b^a(t-1)^a)]\right\}, \tag{6.58}$$

for which the likelihood becomes:

$$\begin{aligned} P(s_i|\alpha_i) = \exp\left\{-\sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(j)}}\log(1+b^a t^a) \;-\; e^{\gamma\alpha_{i,t-1}^{(j)}}\log(1+b^a(t-1)^a)\right)\right\} \\ \times\left(1-\exp\left(-[e^{\gamma\alpha_{i,n_i}^{(i)}}\log(1+b^a n_i^a) \;-\; e^{\gamma\alpha_{i,n_i-1}^{(j)}}\log(1+b^a(n_i-1)^a)]\right)\right). \end{aligned} \tag{6.59}$$

Taking,

$$L_{i,j} = \sum_{t=1}^{n_i-1}\left(e^{\gamma\alpha_{i,t}^{(i)}}\log(1+b^a t^a) \;-\; e^{\gamma\alpha_{i,t-1}^{(i)}}\log(1+b^a(t-1)^a)\right), \tag{6.60}$$

for $j = 1, .., M$, where M denotes the number of Monte-Carlo integration draws, and letting L_i denote the mean, the log-likelihood can be expressed as:

$$\log E_{\alpha_i|y_i}\{P(s_i|\alpha_i)\} \approx -L_i - \log(M) + \log\left(\sum_{j=1}^{M}\left[e^{L_i - L_{i,j}} \times q\right]\right), \tag{6.61}$$

where,

$$q = 1 - e^{\left(-[e^{\gamma\alpha_{i,n_i}^{(j)}}\log(1+b^a n_i^a) - e^{\gamma\alpha_{i,n_i-1}^{(j)}}\log(1+b^a(n_i-1)^a)]\right)} \tag{6.62}$$

6.5.4 Parameter Estimation

For state-space models in general, parameter estimation follows two broad approaches. As Cappé et al. (2007) outline, if the data is available beforehand, parameters are estimated prior to inference of the latent process. This first approach is consequently

referred to as 'offline' or 'batch' estimation in the literature. Alternatively, if the data is not available beforehand, the parameters are estimated sequentially as data arrives and such an approach is referred to as 'online' estimation.

Since the likelihood 6.36 factorises into a Gaussian and non-Gaussian component and the entire dataset is available, an offline approach to estimation is adopted and considered preferable: authors such as Andrieu and Doucet (2002) and Creal (2012) advocate taking into account features of the model, such as the ability to separate out a Gaussian component which has an exact solution, to enable a more efficient estimate of the parameters.

For the DLS model, the Gaussian component from 6.36 has an exact solution. The remaining part of the parameter estimation involves simulation smoothing or generating draws from the distribution $P(\alpha_i|y_i)$. Three potential simulation smoothing algorithms are outlined from a Kalman Filtering perspective in Section 4.3 and the particle smoother based on SMC methods is outlined in Section 5.3.

For Monte Carlo integration using a large number of simulations, they in principle achieve the same result. The simulation smoother by Durbin and Koopman (2002), summarised in Algorithm 1, is often preferred due to its relative simplicity; Helske (2017) provides an implementation in the software R as part of the comprehensive Kalman Filtering and Smoothing (KFAS) package based on Durbin and Koopman (2012). The particle smoother is adopted herein since it generalizes to a wider range of potential formulations for the longitudinal observations.

The Monte-Carlo draws M are taken to be $1,000$ throughout based on preliminary

analysis for the fly data which shows parameter estimates converging with negligible difference in both the estimate and precision observed past $1,000$ draws.

6.6 Model Checking

Model checking is considered a necessary precursor to parameter estimation: Gelman and Hill (2007) outline a popular general approach based on *fake-data* simulation whereby parameters of a statistical model are set to fixed true values and simulations are used to study the properties of parameter estimates and to check the method performs as expected.

6.6.1 Generating synthetic data

Fake-data or synthetic data can be generated for the DLS model using the set-up outlined on page 61. For the case when a Gompertz baseline hazard is assumed, observations can be generated according to Algorithm 5 for arbitrary a_1, P_1, Q, H, Z, T, d and survival parameters a, b and γ.

For a Weibull baseline hazard, observations can be generated using Algorithm 5, taking $p_i = \exp\left\{-b^a \left(e^{\gamma\alpha_i} i^a - e^{\gamma\alpha_{i-1}}(i-1)^a\right)\right\}$.

Algorithm 5 Simulating observations from Gompertz DLS model.

Initialization:

- Sample $\alpha_1 \sim N(a_1, P_1)$ and sample $y_1 \sim N(Z\alpha_1, H)$.
- Calculate survival probability $p_1 = \exp\left\{-\frac{ae^{\gamma\alpha_1}}{b}\left(e^b - 1\right)\right\}$.
- Sample $s \sim \text{Bernoulli}(p_1)$ and set $i = 1$.

If $s = 0$**:**

- Terminate and return y_1 and i which denotes mortality between day 0 and 1.

Else while $s = 1$**:**

- Set $i = i + 1$
- Sample $\alpha_i \sim N(T\alpha_{i-1} + d, Q)$ and sample $y_i \sim N(Z\alpha_i, H)$.
- Calculate $p_i = \exp\left\{-\frac{a}{b}\left(e^{\gamma\alpha_i}(e^{bi} - 1) - e^{\gamma\alpha_{i-1}}(e^{b(i-1)} - 1)\right)\right\}$.
- Sample $s \sim \text{Bernoulli}(p_i)$.

Return $y = y_1, ..., y_i$ and i which denotes mortality between day $i - 1$ and i.

6.6.2 Gompertz DLS Model

An initial batch of 1000 datasets of 50 survival times are generated using Algorithm 5, letting $a_1 = 0, P_1 = 0, H = 10, Q = 1, d = 0.5, Z = 1, T = 1$ and the survival parameters $a = 0.00125, b = 0.02$ and $\gamma = 0.0005$. The survival parameters are taken to follow what is observed for the fly data and similarly since drops in activity levels for the flies are always positive and largely increasing, a random walk with drift is preferred for the latent ageing process.

Parameters are estimated for each of the datasets using the log-likelihood outlined by (6.36) and (6.50). Letting $N = 1000$ and $n = 50$ denote the number of datasets and the number of survival times within each dataset respectively, estimated survival parameter estimates for the synthetic data can be summarised as in Figure 6.10. Despite the relatively small sample size n, the overall mean for a and b is very close to the true value: the overall mean for γ is approximately the true value and standard error reflects a confidence interval containing the true parameter.

The state-space model parameters are estimated for each sample letting $Q = 1, Z = 1$ and $T = 1$ throughout. The overall mean for H and the drift d is almost exactly the true value as depicted in Figure 6.11. Generating further datasets with $N = 1000$ and $n = 35, 100$ and 150 shows the accuracy of the mean γ for the synthetic data improving considerably as summarised in Table 6.4. Coverage probabilities are calculated for each of the sample sizes as summarised in Table 6.5 using standard errors from optimisation routine: they show the overall estimation of the survival parameters improving with sample size in line with Pletcher (1999), who advocates sample sizes

Table 6.4: Gompertz DLS model synthetic data mean parameter estimates

	a	**s.e**	**b**	**s.e**	**γ**	**s.e**
n=35	0.00125	0.000906	0.0200	0.0220	0.000919	0.0130
n=50	0.00125	0.000485	0.0207	0.00563	0.000460	0.00834
n=100	0.00124	0.000345	0.0204	0.00400	0.000498	0.00590
n=150	0.00125	0.000280	0.0202	0.00320	0.000502	0.00472

over 100. Computational time grows linearly with n.

Table 6.5: 95% Gompertz DLS confidence level observed coverage

Parameters	**n=35**	**n=50**	**n=100**	**n=150**
a	91.1	93.7	93.9	94.2
b	96.8	96.6	95.3	95.7
γ	96.1	96.0	95.7	95.8
H	97.1	94.2	94.2	94.2
d	97.7	95.7	95.6	95.7

Repeating the simulations with a larger observation variance H, setting it to equal $100, 200$ and 500, produces almost identical findings with similar coverage probabilities and overall means.

The observed coverage for the parameter a with $n = 35$ is below the expected level of variation for a 95% confidence interval: hence it's necessary to consider alternatives such as Bootstrap confidence intervals (Efron and Tibshirani, 1994) for small n. Davison and Hinkley (1997), Davison et al. (2003) and Carpenter and Bithell (2000) provide a thorough review.

Following Efron and Tibshirani (1994) and Carpenter and Bithell (2000), a parametric bootstrap-t confidence interval for the DLS model can be constructed as outlined in

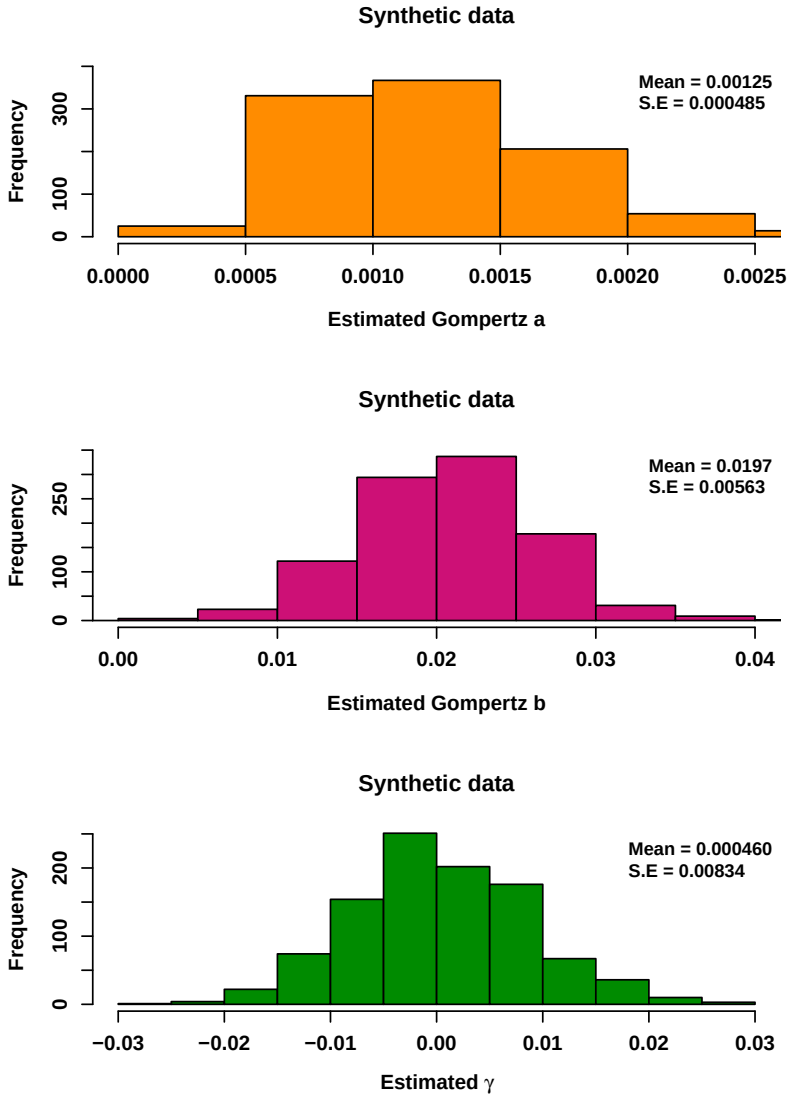

Figure 6.10: Estimated Gompertz DLS survival parameters for n=50, N=1000

Algorithm 6.

In order to apply Algorithm 6, 500 synthetic datasets are generated once parameters are estimated for an initial simulated dataset of size $n = 35$: subsequently, a coverage probability of 95.3% is calculated for the parameter a, which is a major improvement.

Considering a model of the form outlined vaguely by Pavlov (2010) and formalised

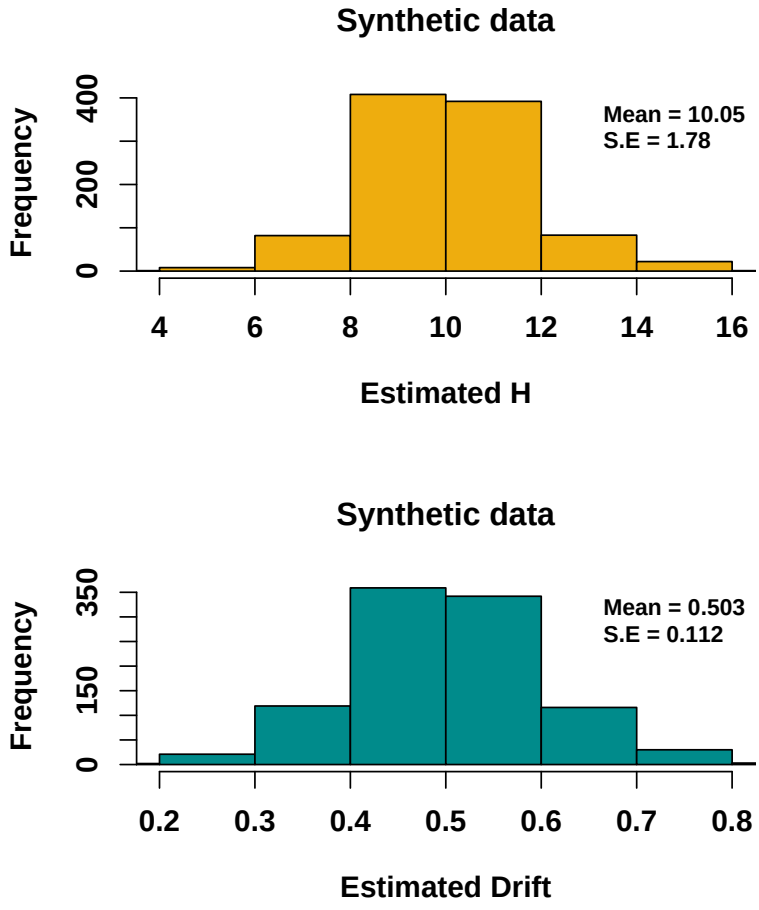

Figure 6.11: Estimated Gompertz DLS model parameters for n=50, N=1000

more thoroughly on pages 78 and 78 where a discrete time cumulative hazard function approximation is made to calculate the survival probabilities, a further $N = 1000$ datasets of size $n = 35, 50, 100$ and 150 are generated with results summarised in Tables 6.6 and 6.7. Comparing with output for the DLS model – see pages 79 and 80 where the Gompertz distribution is used directly to calculate survival probability – in Table 6.4, shows the DLS outperforming Pavlov for each of the sample sizes: the mean γ under Pavlov shows inconsistency and severe overestimation even with increases in

Algorithm 6 Constructing bootstrap-t confidence interval

1. Generate B datasets of same sample size as original data using Algorithm 5 with initial parameter estimates.
2. Estimate parameter of interest along with standard error for each of the B datasets. Let $a_{(i)}$ and $\sigma_{(i)}$ denote the parameter estimate and standard error for dataset i, where $i = 1, ..., B$.
3. Set $t^*_{(i)} = (a_{(i)} - \hat{a})/\sigma_{(i)}$, where $\hat{a}$ denotes initial parameter estimate.
4. Let t^* denote $t^*_{(1)}, ..., t^*_{(B)}$ ordered from smallest to largest. Let $t^*_{\alpha/2}$ and $t^*_{1-\alpha/2}$ denote the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of t^* respectively, where α denotes the desired level of significance.
5. The bootstrap-t confidence interval for parameter a is then, $(\hat{a} - \hat{\sigma}t^*_{1-\alpha/2}, \hat{a} - \hat{\sigma}t^*_{\alpha/2})$ where $\hat{\sigma}$ denotes initial estimate of standard error for original data.

sample size suggesting it's only weakly identifiable. The parameter a doesn't settle around its true value either, which supports findings from page 82. From Table 6.7, the coverage probability for the parameter a is also low, however Algorithm 6 can help alleviate.

Table 6.6: Pavlov model synthetic data mean parameter estimates

	a	**s.e**	**b**	**s.e**	$\boldsymbol{\gamma}$	**s.e**
n=35	0.00121	0.000603	0.0220	0.0101	6.96×10^{-5}	0.0179
n=50	0.00123	0.000457	0.0207	0.00824	0.00150	0.0153
n=100	0.00125	0.000347	0.0205	0.00505	0.000819	0.00932
n=150	0.00123	0.000283	0.0203	0.00433	0.00117	0.00799

Table 6.7: Pavlov Model 95% confidence level observed coverage

Parameters	n=35	n=50	n=100	n=150
a	86.3	91.5	92.4	92.8
b	93.8	94.3	95.6	95.3
γ	95.5	94.8	95.3	95.3
H	94.6	94.4	94.6	94.6
d	95.5	95.6	95.6	95.6

6.6.3 Weibull DLS Model

An initial batch of 1000 datasets of 50 survival times are generated using Algorithm 5 with the Weibull survival probability, letting $a_1 = 0, P_1 = 0, H = 10, Q = 1, d = 0.5, Z = 1, T = 1$ and survival parameters $a = 3, b = 0.007$ and $\gamma = 0.0005$. Parameters are estimated using the log-likelihood outlined by (6.36) and (6.55). Figure 6.12 summarises the survival parameter estimates for the 1000 datasets overall: there appears to be some bias when estimating the parameters a, p and γ. Coverage probabilities for a 95% confidence interval however are good showing each of the datasets ably capturing the true parameter values in general as seen in Table 6.9. The state space model parameter estimates are also close to the true values as summarised in Figure 6.13.

Bias when using maximum likelihood estimation with the Weibull distribution has been addressed by Cacciari et al. (1996), Hirose (1999) and Chen et al. (2017). As Shen and Yang (2015) state, "...it is widely recognized that the maximum likelihood estimator (MLE) can be quite biased, in particular when the sample size is small". Chen et al. (2017) using simulations show a significant reduction in bias once sample sizes approach 150 and beyond. Simulating a further 1000 datasets for $n = 35, 100$ and 150 confirms

these findings as seen in Table 6.8 which show the survival parameters approaching the true values as sample size increases. Coverage probabilities nevertheless are excellent across the different sample sizes as seen in Table 6.9, suggesting the standard errors are reliable. Repeating the simulations with a larger observation variance H, setting it to

Table 6.8: Weibull DLS model synthetic data mean parameter estimates

	a	**s.e**	**b**	**s.e**	$\boldsymbol{\gamma}$	**s.e**
n=35	3.08	0.827	0.00663	0.00182	0.00206	0.0108
n=50	3.05	0.569	0.00686	0.00126	0.00116	0.00708
n=100	3.03	0.395	0.00691	0.000871	0.000825	0.00482
n=150	3.04	0.340	0.00698	0.000732	0.000553	0.00406

Table 6.9: 95% Weibull DLS confidence level observed coverage

Parameters	**n=35**	**n=50**	**n=100**	**n=150**
a	96.2	97.1	96.7	95.4
b	95.3	96.1	95.8	95.1
$\boldsymbol{\gamma}$	96.8	96.7	97.0	95.5
H	96.9	94.4	94.4	94.4
d	97.2	95.5	95.5	95.5

equal $100, 200$ and 500, shows negligible difference with similar coverage probabilities and overall means.

In response to the bias issues with the Weibull distribution, Chen et al. (2017) outline a bootstrap bias-corrected estimator for a and b: $\hat{a}_{MLE}$ and $\hat{b}_{MLE}$ are initially estimated through a maximum likelihood approach. N_B bootstrap samples of the same size are then generated using $\hat{a}_{MLE}$ and $\hat{b}_{MLE}$ and parameters are re-estimated for each of the bootstrap samples. The bootstrap bias-corrected parameter estimates are then

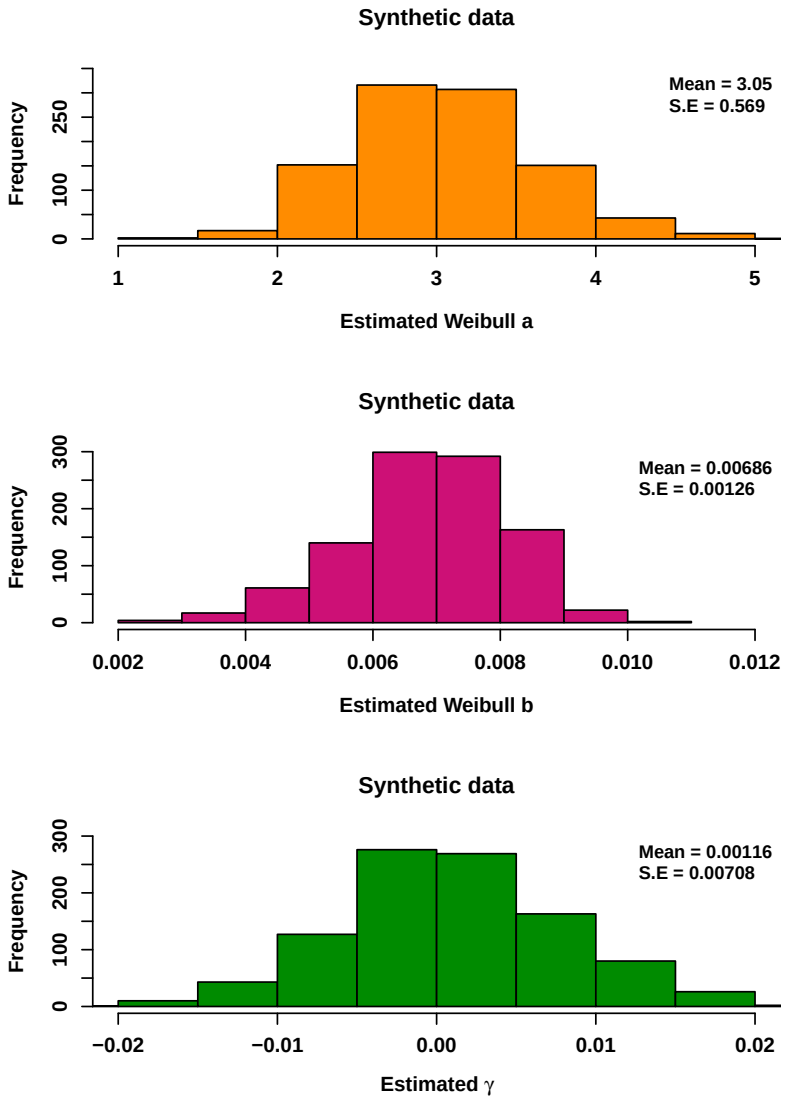

Figure 6.12: Estimated Weibull DLS survival parameters for n=50, N=1000

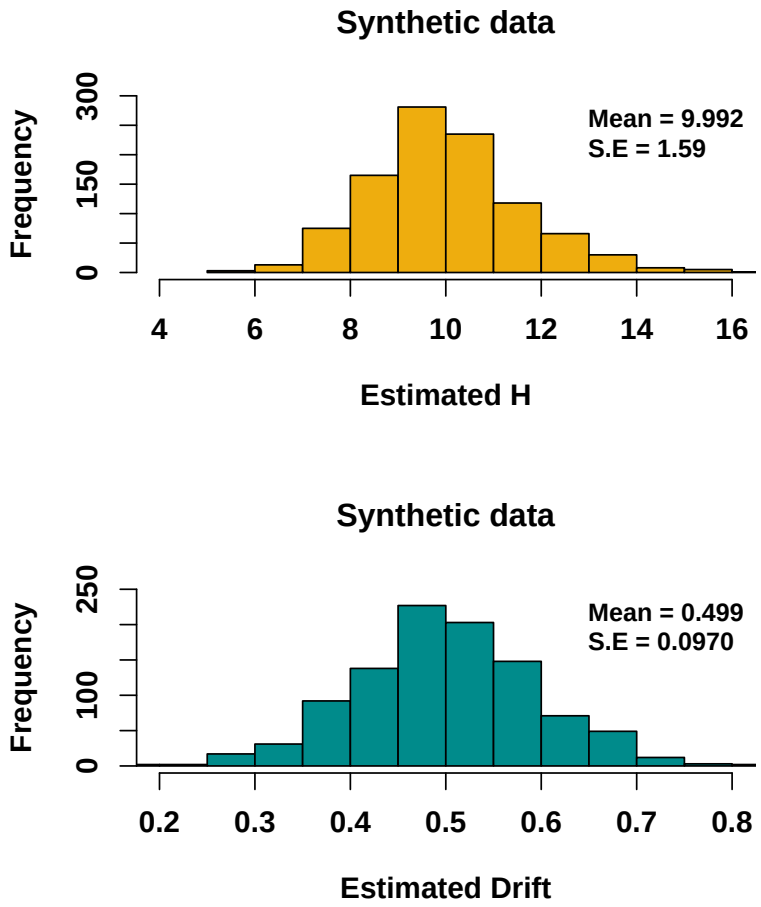

Figure 6.13: Estimated Weibull DLS model parameters for n=50, N=1000

$a = 2\hat{a}_{MLE} - (1/N_B)\sum_{i=1}^{N_B} \hat{a}^{(i)}$ and $b = 2\hat{b}_{MLE} - (1/N_B)\sum_{i=1}^{N_B} \hat{b}^{(i)}$, where $\hat{a}^{(i)}$ and $\hat{b}^{(i)}$ denote the parameter estimates for bootstrap sample i for $i = 1, ..., N_B$. This can be adopted as a parametric bootstrap for the Weibull DLS model.

Considering the 1,000 datasets simulated for each of $n = 35, 50, 100$ and 150 from earlier as parametric bootstrap samples also shows the median providing a more accurate estimator for the survival parameters than the mean since it is always closer to the true value — see Table 6.10 — with the bias reducing as the sample size grows: this supports the findings from graphical plots such as Figure 6.12, where the left skewness decreases with sample size.

Table 6.10: Parametric bootstrap estimated median

Parameters	**n=35**	**n=50**	**n=100**	**n=150**
$\boldsymbol{a = 3}$	3.03	3.01	3.01	3.01
$\boldsymbol{b = 0.007}$	0.00683	0.00694	0.00696	0.00701
$\boldsymbol{\gamma = 0.0005}$	0.00108	0.000817	0.000653	0.000507

Chapter 7

Application to Fly Data

7.1 Parameter Estimates

Taking the log-likelihood outlined by (6.36) and (6.50) together with the different postulations of ageing outlined in section 6.2 allows initial estimation at a population level across state space model and survival parameters. When estimating $|T| < 1$ for the OU ageing process, T is always close to 1 and hence setting it to $T = 0.99$ which gives $Q = 1 - 0.99^2$, allows identification of other parameters. The state space model parameters for each of the ageing processes summarised in Table 7.1, looking at all 32 female flies, remain unchanged across the Gompertz and Weibull DLS models. Tables 7.2 and 7.3 provide a summary of the corresponding estimated survival parameters: the random walk and OU ageing models for both the Gompertz and Weibull are considered significant since all the estimated survival parameters lie within twice their standard

error. Models reflecting a diet-specific frailty are found to be insignificant, which isn't surprising since samples for both the diets are made up of only 16 survival times each, which is small.

From Table 7.1, the parameter Z denotes a trend coefficient connecting the latent ageing process to the daily absolute value drop in activity level for the flies with H denoting the corresponding variance in a state space model setting. A model incorporating drift doesn't appear to fit well with H almost 2-3 times the variance under a simple random walk or OU model. The OU model fits a smoother evolution of the ageing process with it's inherent mean-reverting property – see section 6.2.3 – and perhaps not surprisingly given its popularity in ageing studies, has the lowest H of the 3 models which suggests upon a more graphical and visual interpretation later capturing the state space dynamic, we'd expect the OU to fit best.

Table 7.1: State space model parameter estimates

Model	**Z**	**S.E**	**H**	**S.E**
Random Walk	10.14	0.43	428.4	19.4
Model	**Drift**	**S.E**	**H**	**S.E**
Random Walk with Drift	1.30	0.02	1,093.5	44.4
Model	**Z**	**S.E**	**H**	**S.E**
OU	87.2	3.10	379.1	18.7

Table 7.2: Gompertz DLS survival parameter estimates

Model	**a**	**S.E**	**b**	**S.E**	γ	**S.E**
Random Walk	8.56×10^{-4}	3.93×10^{-4}	0.0226	3.81×10^{-3}	1.91×10^{-3}	9.47×10^{-4}
Random Walk with Drift	2.66×10^{-4}	1.88×10^{-4}	0.0104	5.78×10^{-3}	0.0156	3.62×10^{-3}
OU	1.03×10^{-3}	5.19×10^{-4}	0.0224	3.75×10^{-3}	8.58×10^{-4}	4.11×10^{-4}

Table 7.3: Weibull DLS survival parameter estimates

Model	a	S.E	b	S.E	γ	S.E
Random Walk	3.10	0.437	6.88×10^{-3}	5.63×10^{-4}	2.42×10^{-3}	1.17×10^{-3}
Random Walk with Drift	1.84	0.525	2.01×10^{-3}	1.27×10^{-3}	0.0149	3.74×10^{-3}
OU	3.08	0.430	7.34×10^{-3}	5.80×10^{-4}	1.13×10^{-3}	4.95×10^{-4}

The survival parameters a and b for both the Gompertz and Weibull, where significant, are close to the maximum likelihood estimates observed earlier – see Tables 6.1 and 6.2 – when fitting only a baseline hazard to the survival times in section 6.4.4. From Table 7.1, a model with drift meant the estimated variance H of the longitudinal data was high, as compared to a simple random walk or OU model: consequently, when estimating the survival parameters under a model incorporating drift, the shape parameter a for the Gompertz and scale parameter b for the Weibull are insignificant, which shows the overall model doesn't capture the data well.

The frailty coefficient γ captures the effect of a latent ageing process on survival and hence given the recursive structure of a state-space model is better evaluated graphically, taking into account its composite effect across the lifetime of a fly and for the sample overall through estimates of the survival curve and residual lifetime. The following sections will explore these ideas.

To ascertain whether the maximum likelihood estimate for γ has been obtained, the profile likelihood fixing all other parameters at their maximum likelihood values is plotted as in Figures 7.1 and 7.2: the estimates observed in Tables 7.2 and 7.3 are corroborated.

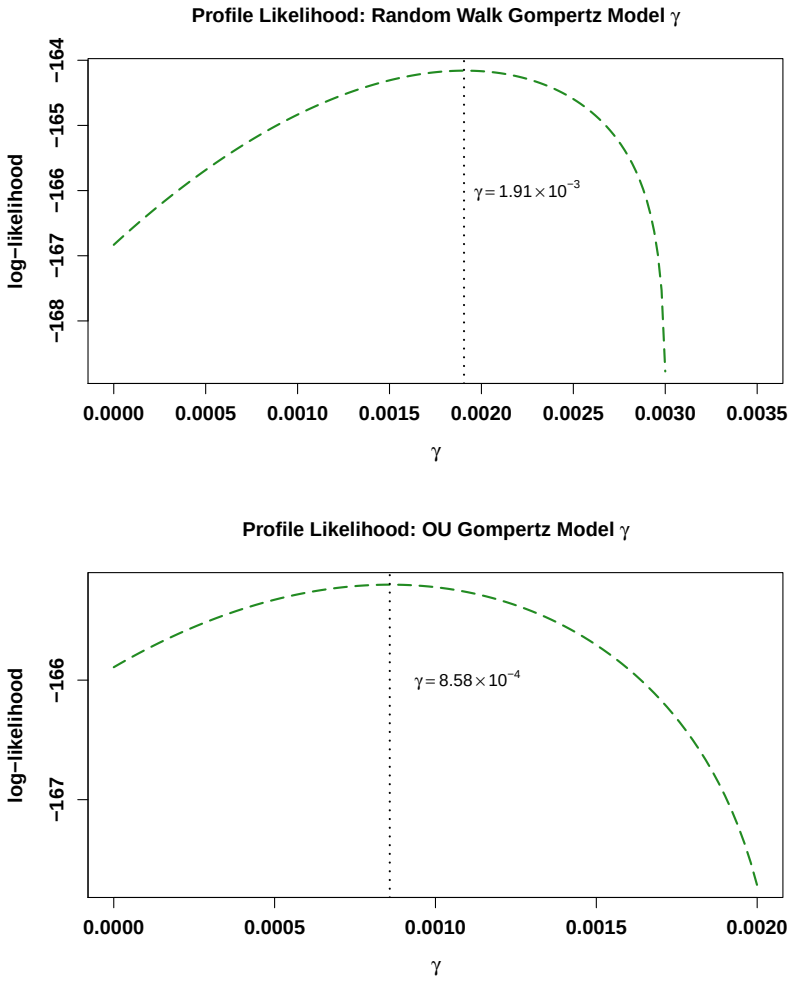

Figure 7.1: Profile likelihood plots for Gompertz model γ

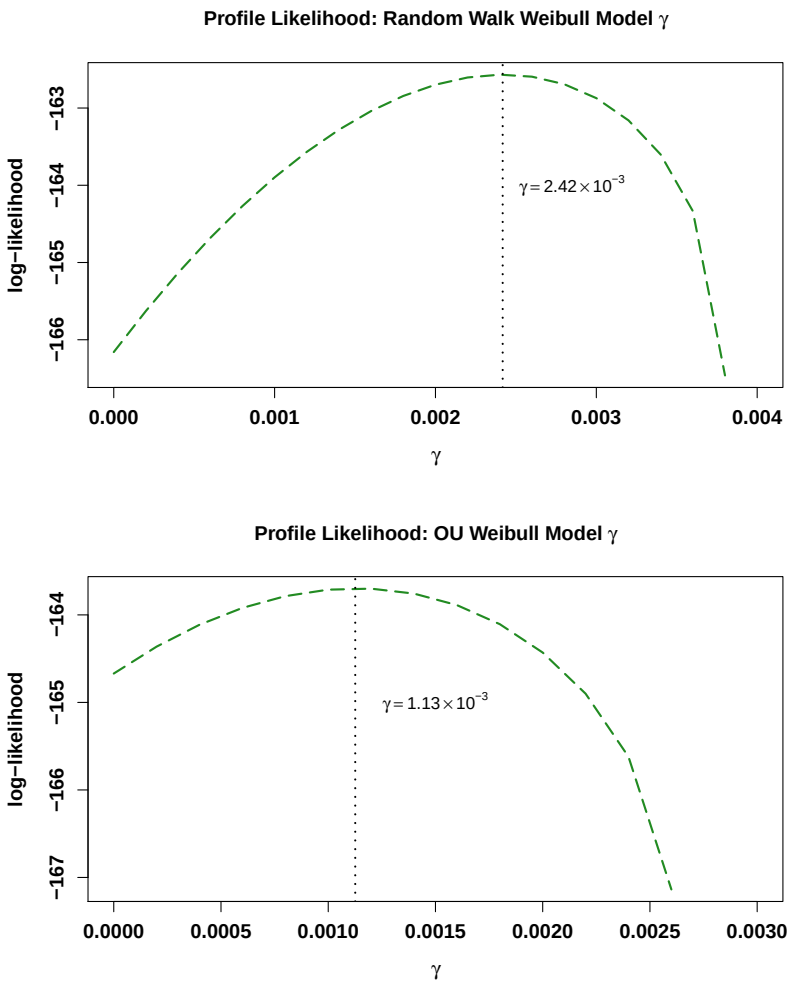

Figure 7.2: Profile likelihood plots for Weibull model γ

7.2 Filtering

The next step once a model has been fitted is to infer development of the latent process given the longitudinal information: an ageing signal based on the fly activity and survival information can be extracted through application of SMC filtering. The Bootstrap Filter from section 5.1 requires ability to draw samples from the transition density $P(\alpha_{i,t}|\alpha_{i,t-1})$. Since the latent ageing process follows (6.1), $P(\alpha_{i,t}|\alpha_{i,t-1}) \sim N(T\alpha_{i,t-1}, Q)$.

The weights require calculation of $g(y_{i,t}|\alpha_{i,t})$ where $y_{i,t}$ denotes the observation for fly i. For the DLS model, the observations are two-dimensional comprising a longitudinal observation together with a survival indicator both of which are assumed to be conditionally independent given the latent ageing process. Hence following (6.2) and (6.3) the weights become $g(y_{i,t}|\alpha_{i,t}) \times P(s_{i,t}|\alpha_{i,t})$, where $g(y_{i,t}|\alpha_{i,t}) \sim N(Z\alpha_{i,t}, H)$ and $P(s_{i,t}|\alpha_{i,t}) \sim \text{Bernoulli}(p_{i,t})$, as summarised in Algorithm 7 which can be adapted similarly for the Weibull.

7.2.1 Illustration

An application of Algorithm 7 can be illustrated considering the ageing processes outlined in Tables 7.1 and 7.2. The number of particles N is typically chosen in line with the level of accuracy required and $10,000$ is considered a sufficiently large number to that extent which can be reduced in practice (Cappé et al., 2007). Assessing this for the fly data in a preliminary run shows the filtering estimates of the latent ageing process

Algorithm 7 Gompertz DLS Bootstrap Filter

For time t=1

for $k = 1, ..., N$

- Sample $\tilde{\alpha}_{i,1}^{(k)} \sim N(a_1, P_1)$.
- Calculate $p_{i,1} = \exp\left\{-\frac{ae^{\gamma\tilde{\alpha}_{i,1}^{(i)}}}{b}\left(e^b - 1\right)\right\}$.
- Compute weights $\tilde{w}_{i,1}^{(k)} = g(y_{i,1}|\tilde{\alpha}_{i,1}^{(k)}) \times P(s_{i,1}|\tilde{\alpha}_{i,1}^{(k)})$, where $g(y_{i,1}|\tilde{\alpha}_{i,1}^{(k)}) \sim N(Z\tilde{\alpha}_{i,1}^{(k)}, H)$ and $P(s_{i,1}|\tilde{\alpha}_{i,1}^{(k)}) \sim \text{Bernoulli}(p_{i,1})$.
- Normalize weights: $w_{i,1}^{(k)} = \tilde{w}_{i,1}^{(k)} / \sum_{k=1}^{N} \tilde{w}_{i,1}^{(k)}$
- Compute filtering estimate: $\bar{h}_{i,1} = \sum_{k=1}^{N} w_{i,1}^{(k)} \tilde{\alpha}_{i,1}^{(k)}$

For time t=2, ..., n_i:

- Resample: select N particle indices $j_k \in \{1, ..., N\}$ using normalized weights: $w_{i,t-1}^{(k)} = \frac{\tilde{w}_{i,t-1}^{(k)}}{\sum_{i=1}^{N} \tilde{w}_{i,t-1}^{(k)}}$
- Set $\alpha_{i,t-1}^{(k)} = \tilde{\alpha}_{i,t-1}^{(j_k)}$ and $w_{i,t-1}^{(k)} = 1/N$ for $k = 1, ..., N$.

for $k = 1, ..., N$

- Sample $\tilde{\alpha}_{i,t}^{(k)} \sim N(T\alpha_{i,t-1}^{(k)} + d, Q)$
- Calculate $p_{i,t} = \exp\left\{-\frac{a}{b}\left(e^{\gamma\tilde{\alpha}_{i,t}^{(k)}}(e^{bt} - 1) - e^{\gamma\alpha_{i,t-1}^{(k)}}(e^{b(t-1)} - 1)\right)\right\}$.
- Compute weights: $\tilde{w}_{i,t}^{(k)} = g(y_{i,t}|\tilde{\alpha}_{i,t}^{(k)}) \times P(s_{i,t}|\tilde{\alpha}_{i,t}^{(k)})$, where $g(y_{i,t}|\tilde{\alpha}_{i,t}^{(k)}) \sim N(Z\tilde{\alpha}_{i,t}^{(k)}, H)$ and $P(s_{i,t}|\tilde{\alpha}_{i,t}^{(k)}) \sim \text{Bernoulli}(p_{i,t})$.
- Normalize weights: $w_{i,t}^{(k)} = \tilde{w}_{i,t}^{(k)} / \sum_{k=1}^{N} \tilde{w}_{i,t}^{(k)}$
- Compute filtering estimate: $\bar{h}_{i,t} = \sum_{k=1}^{N} w_{i,t}^{(k)} \tilde{\alpha}_{i,t}^{(k)}$

showing only a negligible difference past $5,000$ particles, with the ESS always above $1,500$. Hence taking this as a reasonable value for N, the random walk ageing process for an individual from the sample can be estimated as summarised in Figure 7.3, where the filtering estimate along with ESS and variance are presented. The sharp increase in variance around 190 days denotes missing data for that particular individual which is handled as outlined in section 5.5.

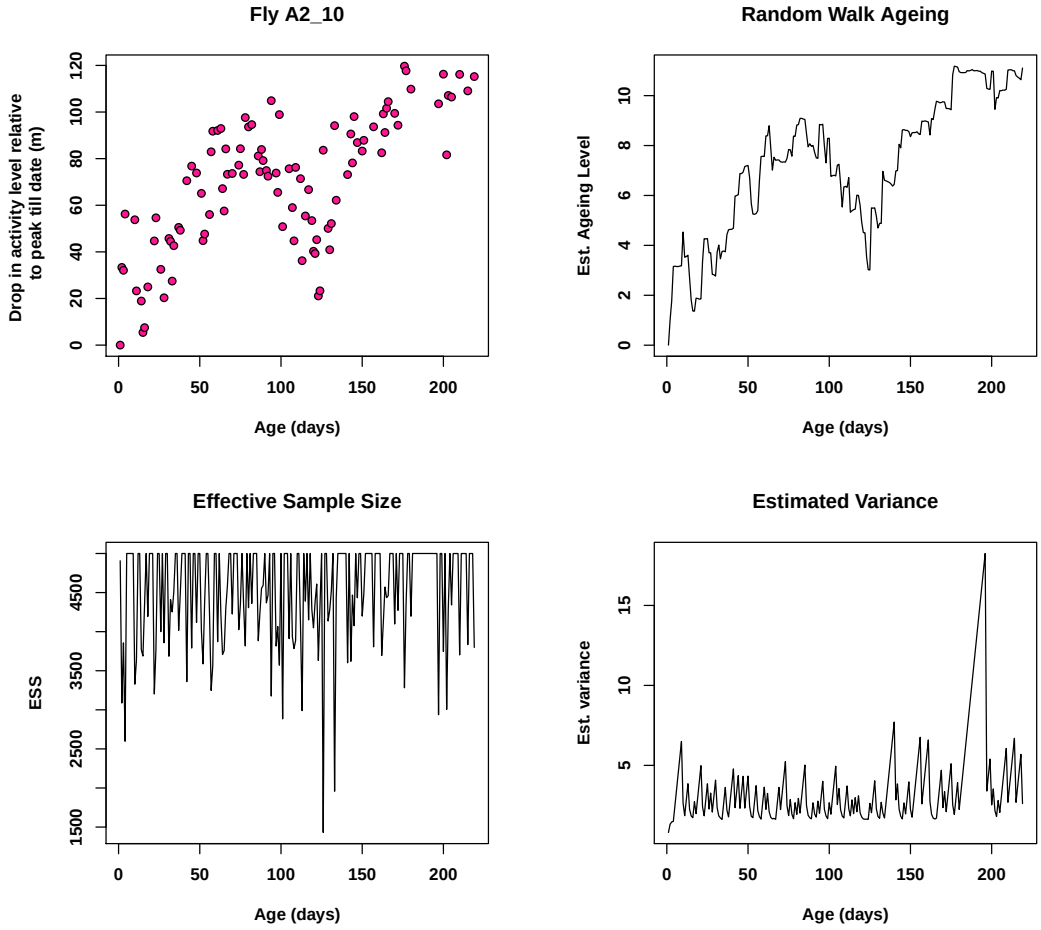

Figure 7.3: Fly A2_10: Gompertz random walk ageing

To gauge how differently the flies age modelling a random walk ageing process, 4 fe-

males flies are picked at random from both diets and their estimated ageing level is illustrated in Figures 7.4 and 7.5. In both Figures 7.4 and 7.5 the ageing level exhibits damage accumulation with repair and death appears to occur when the repair element subsides towards latter life for flies such as $A9\_8, B8\_10, B4\_8, A5\_10, A2\_8$ and $B7\_10$. The sugar diet flies also appear to age more than their full diet counterparts, with ageing levels as high as 20 observed for the flies picked at random. Considering the entire sample in Figure 7.6 where the random walk ageing levels observed at death are considered, shows flies on a sugar diet generally exhibiting higher ageing levels with an interquartile range exceeding the interquartile range observed for full diet flies. A similar pattern is observed for the OU ageing process and Weibull models.

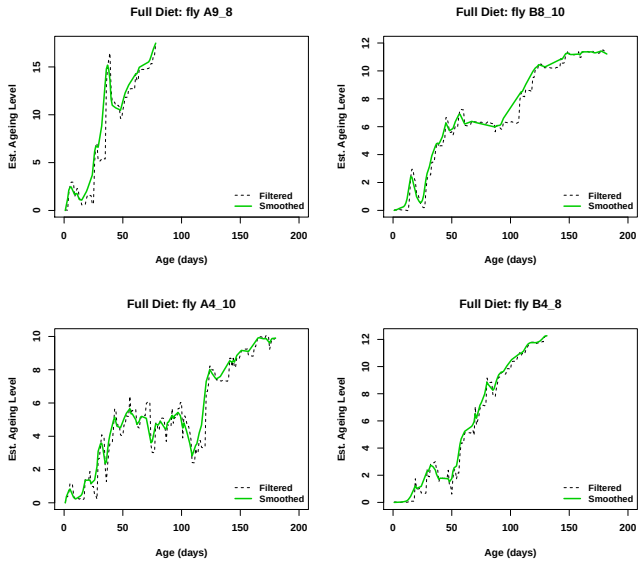

Figure 7.4: Gompertz random walk ageing for 4 full fiet flies

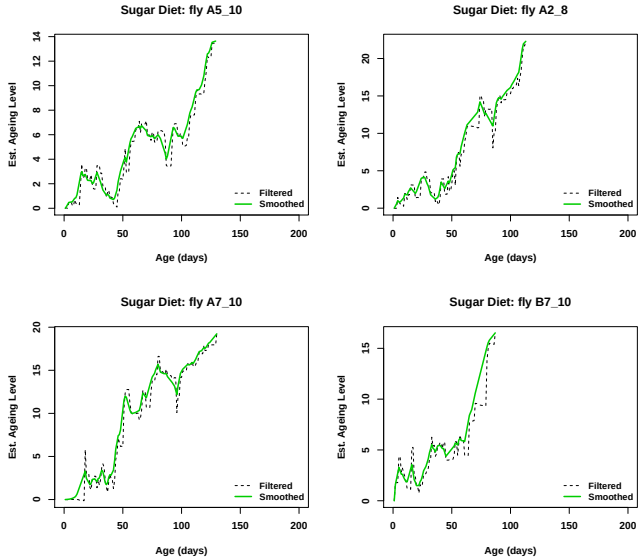

Figure 7.5: Gompertz random walk ageing for 4 sugar diet flies

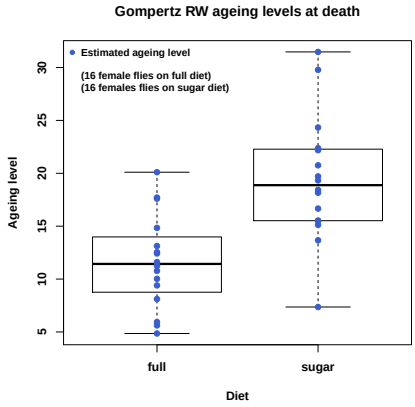

Figure 7.6: Gompertz random walk ageing level at death

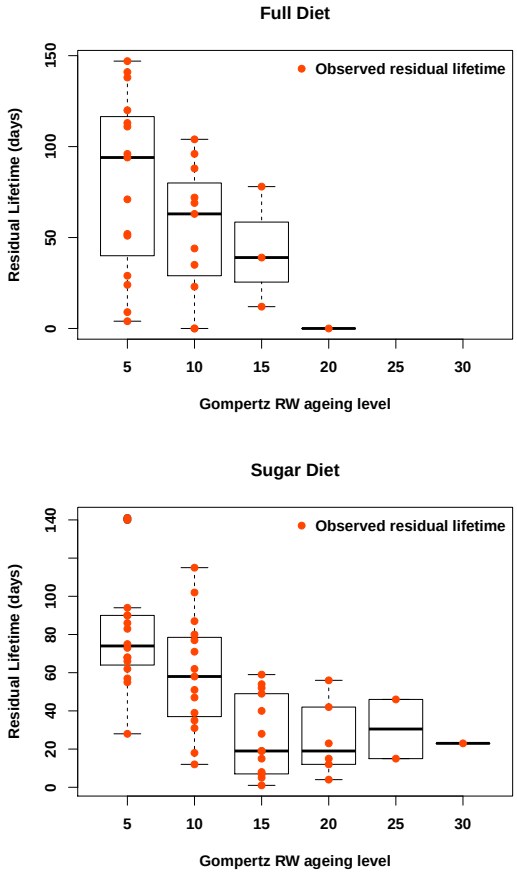

Figure 7.7: Gompertz random walk residual lifetimes by ageing level

Considering residual lifetimes from the first time each fly reaches a certain ageing level as in Figure 7.7 further highlights the difference between full and sugar diet flies, with no full diet flies living past a level of 20 and most full diet fly survival times being concentrated between an ageing level of 10-15; most of the sugar diet flies die between an ageing level of 15-25. Ageing theories focus on the ability of an organism to allocate resources between growth, repair and maintenance: the observed residual lifetimes appear to support this since even though a certain ageing level is reached, flies die some time after breaching the threshold which is in line with the organism attempting

to recover. To predict residual lifetimes taking into account the composite effect of ageing an alternative approach can be explored.

7.2.2 Predicting remaining lifetime

The ageing information deduced for each fly in its entirety can be used to estimate survival: considering an ageing level on its own ignores the evolution of the ageing process and the amalgamation of its effect on survival. Based on the filtered estimate of ageing for each time step, a dynamic probability of survival can be estimated. For the Gompertz distribution, (6.47) can be used: letting $p_t(\alpha_{i,t|t}) = P(T > t|T > t-1, \alpha_{i,\leq t})$, where $\alpha_{i,t|t}$ denotes the filtered estimate of the ageing process for fly i and $\alpha_{i,\leq t}$ denotes all information available on ageing for fly i up till t, the overall survival probability can then be estimated as:

$$P(T > t|\alpha_{i,\leq t}) = \prod_{j=1}^{t} p_j(\alpha_{i,j|j}). \tag{7.1}$$

Figure 7.8 depicts the estimated survival curve for random walk and OU ageing for fly $A2\_10$.

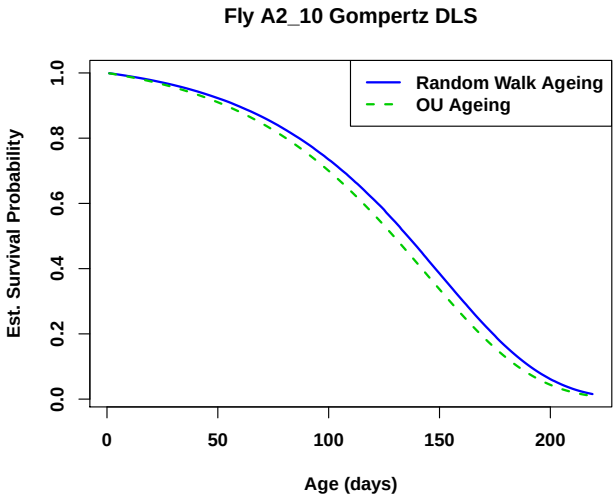

Figure 7.8: Fly A2_10 estimated survival curves

Since $p_t(\alpha_{i,t|t})$ denotes the estimated Bernoulli probability of survival through day t, given survival up till the start of day t, $1-p_t(\alpha_{i,t|t})$ can be considered as an approximate daily Geometric probability of death: based on this a crude estimate of the remaining lifetime can be calculated at each time step. Figure 7.9 provides a summary of the estimated remaining lifetime at age 100 onwards for fly $A2\_10$ under the population random walk and OU ageing model.

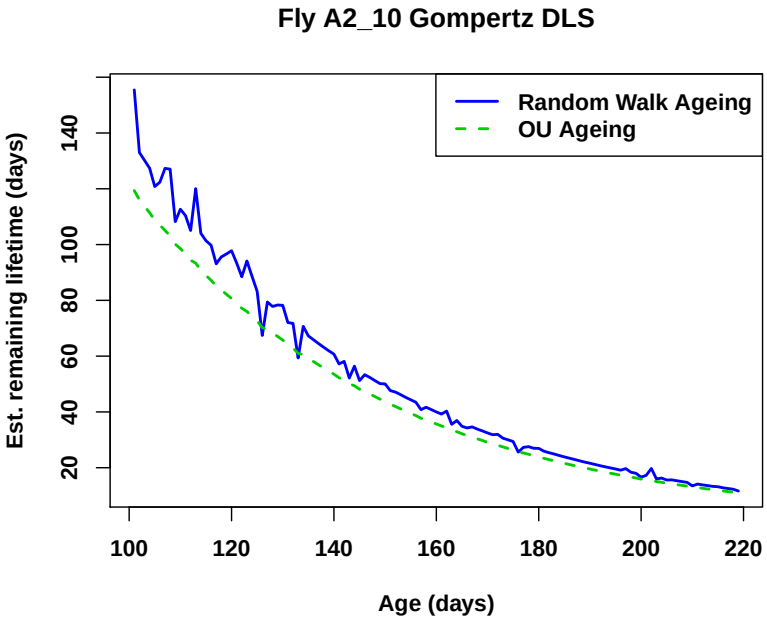

Figure 7.9: Fly A2_10 estimated remaining lifetime

Although more computationally intensive, an alternative and more informative approach to prediction of remaining lifetime can be adopted through Algorithm 8, utilising the p-step ahead SMC prediction algorithm defined in Algorithm 4. Considering fly $B1\_10$ which lived till 182 days, 500 predictions of remaining lifetime are made at ages 150 and 170 as summarised in Figure 7.10. The estimated mean remaining lifetime at age 150 is 28.8 days which is 3.2 days off the actual residual lifetime. The interquartile range provides a fairer reflection of remaining lifetime as seen from the boxplots.

Algorithm 8 Predicting remaining lifetime for Gompertz DLS

Initialization: Pick a time point t to begin prediction from. Run Algorithm 7 up till t storing $w_{i,t}^{(k)}$ and $\alpha_{i,t}^{(k)}$. Set $s^* = 1$ and a counter $t^* = 0$.

While $s^* = 1$:

for $k = 1, ..., N$

- Set $t^* = t^* + 1$.
- Sample $\tilde{\alpha}_{i,t+t^*}^{(i)} \sim N(Ta_{i,t+t^*-1} + d, Q)$.

end for.

- Compute filtering estimate: $\bar{h}_{i,t+t^*} = \sum_{k=1}^{N} w_{i,t}^{(k)} \tilde{\alpha}_{i,t+t^*}^{(k)}$
- Calculate $\hat{p}_{i,t+t^*} = \exp\left\{-\frac{a}{b}\left(e^{\gamma \bar{h}_{i,t+t^*}}(e^{b(t+t^*)} - 1) - e^{\gamma \bar{h}_{i,t+t^*-1}}(e^{b(t+t^*-1)} - 1)\right)\right\}$.
- Sample $s^* \sim \text{Bernoulli}(\hat{p}_{i,t+t^*})$.

Return: t^* which denotes predicted remaining lifetime.

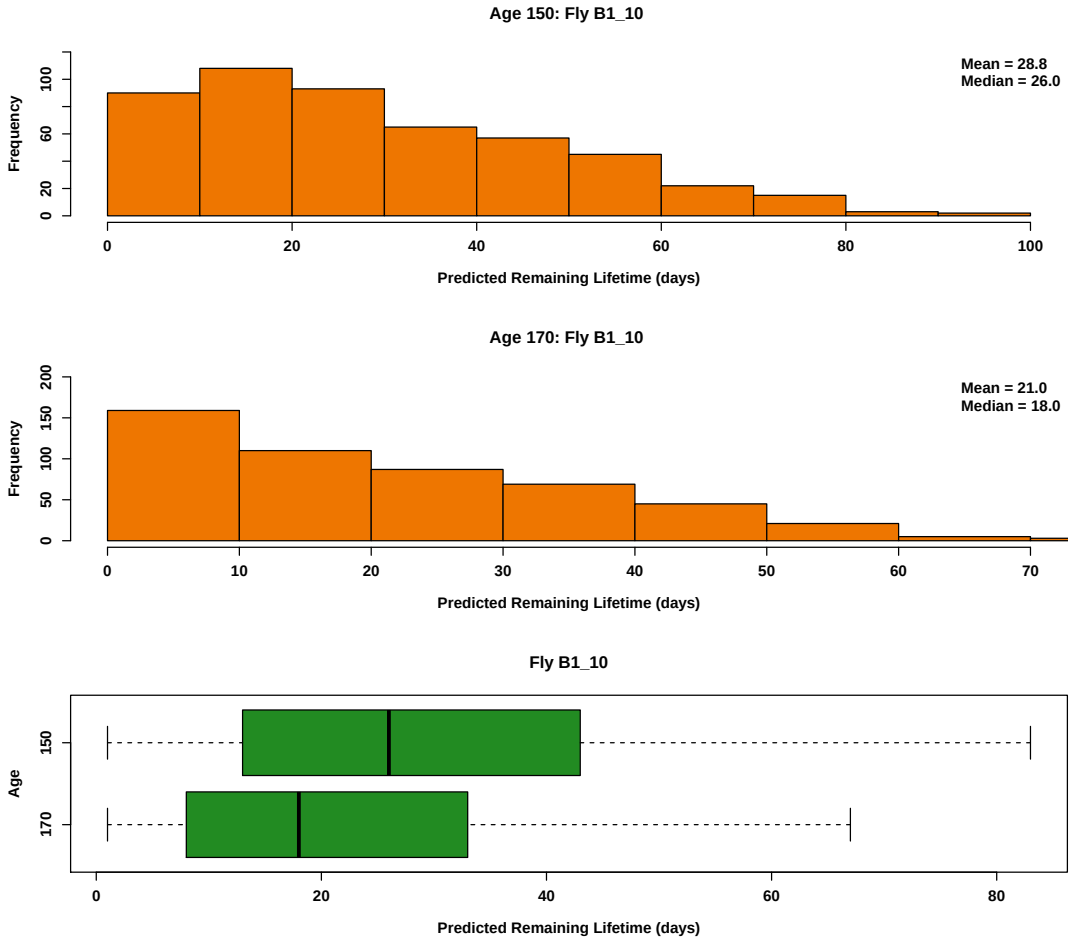

Figure 7.10: Fly B1_10: random walk ageing remaining lifetime predictions

7.3 Model Diagnostics

7.3.1 Goodness of Fit

The overall aim of the model is to provide an estimate of the population survival curve taking into account an unobserved ageing effect: hence a vital goodness of fit diagnostic comprises evaluation of fitted survival curves as compared to the empirical

survival curve.

Section 7.1 defines generative models for which empirical survival curves can be simulated as outlined in Algorithm 9: a comparison thereafter with the empirical survival curve for original data can provide a highly informative goodness of fit check graphically.

Considering the population Gompertz random walk and OU ageing models outlined in Tables 7.1 and 7.2 firstly, 500 survival curves are simulated as presented in Figure 7.11. The 500 survival curve overlay effectively creates a simulated confidence band around the observed empirical survival curve and since it covers the range of data, reveals immediately whether the model fits well or not. From Figure 7.11, the survival curves simulated under both a Gompertz random walk and OU model cover the empirical survival curve in its entirety which initially implies it's possible the empirical survival curve could have arisen as a realisation from a true population following either model. Between day 100 and 160 however, the OU model simulations provide a confidence band which covers the empirical survival curve moreso and hence appears to fit better overall.

Algorithm 9 Generating survival curve for fitted Gompertz DLS model.

Let n denote desired sample size.

For $i^* = 1, .., n$ **do:**

Initialization:

- Sample $\alpha_1 \sim N(a_1, P_1)$ and sample $y_1 \sim N(Z\alpha_1, H)$.
- Calculate survival probability $p_1 = \exp\left\{-\frac{ae^{\gamma\alpha_1}}{b}\left(e^b - 1\right)\right\}$.
- Sample $s \sim \text{Bernoulli}(p_1)$ and set $t = 1$.

If $s = 0$**:**

- Terminate and return $t_j^* = t$ which denotes mortality between day 0 and 1.

Else while $s = 1$**:**

- Set $t = t + 1$
- Sample $\alpha_t \sim N(T\alpha_{t-1} + d, Q)$ and sample $y_t \sim N(Z\alpha_t, H)$.
- Calculate $p_t = \exp\left\{-\frac{a}{b}\left(e^{\gamma\alpha_t}(e^{bt} - 1) - e^{\gamma\alpha_{t-1}}(e^{b(t-1)} - 1)\right)\right\}$.
- Sample $s \sim \text{Bernoulli}(p_t)$.

Return $t_j^* = t$ which denotes mortality between day $t - 1$ and t.

End. Return the survival times $t_1^*, ..., t_n^*$ for which an empirical survival curve can be fitted.

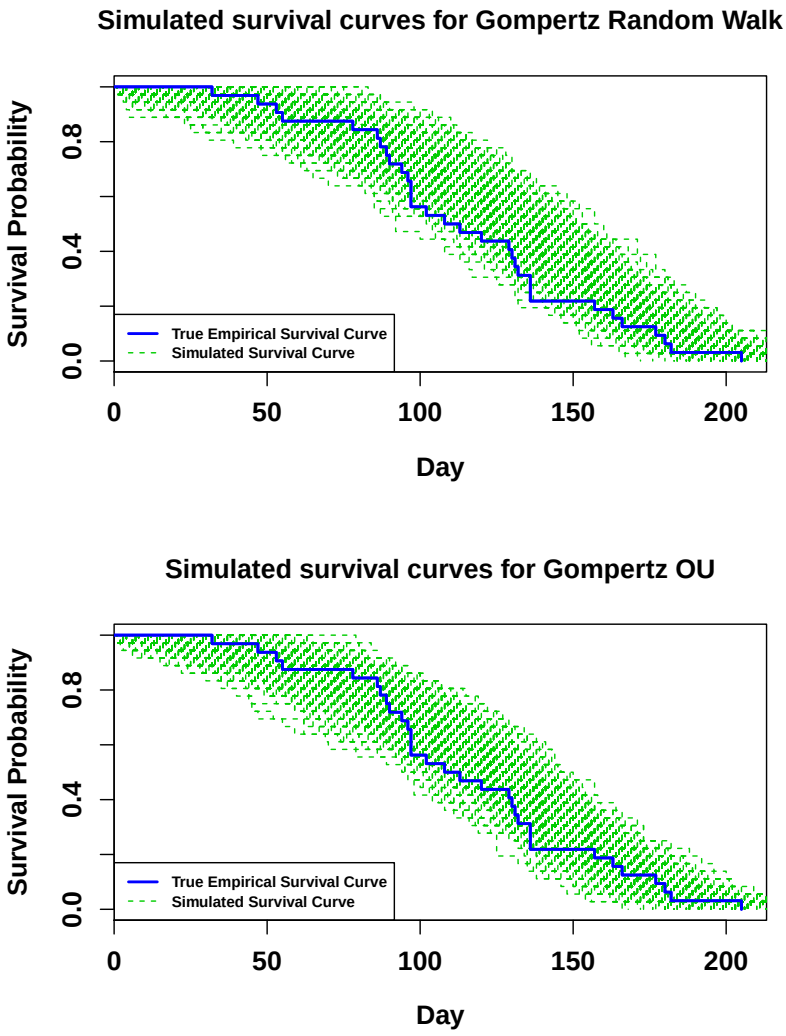

Figure 7.11: Gompertz DLS simulated survival curves

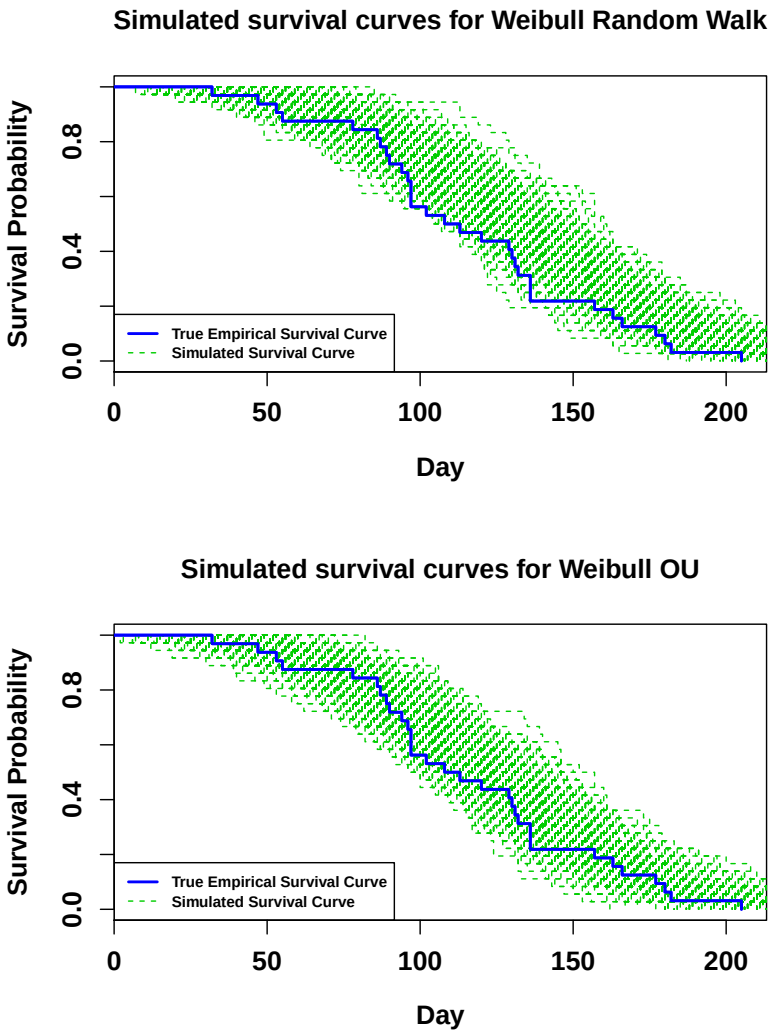

Figure 7.12: Weibull DLS simulated survival curves

Considering the Weibull model estimates in turn, taking Tables 7.1 and 7.3 firstly as for the Gompertz, shows comparative fits to before between day 100 and 160, as seen in Figure 7.12. The Weibull OU model appears to fit reasonably well overall and across the 2 distributions, the OU models provide a closer fit to observed empirical survival curve. Comparing Figures 7.11 and 7.12 shows both fitting similarly.

In order to pursue a more formal test, Algorithm 9 can be used to generate samples of survival times from which an average value of the CDF at each of the observed survival times can be calculated.

Let n denote sample size of the original data x and let x_i for $i = 1, ..., n$, refer to survival times in ascending order. Let N denote number of simulated samples and $y_{j,1}, ..., y_{j,n}$ denote survival times for sample j in ascending order, for $j = 1, ..., N$. Let $\bar{y}_i = \frac{1}{N} \sum_{j=1}^{N} y_{j,i}$, for $i = 1, ..., n$. Then an empirical CDF for the average of the simulated datasets can be calculated taking $\hat{y} = (\bar{y}_i, ..., \bar{y}_n)$.

This enables goodness of fit tests such as the Kolmogorov Smirnov and Cramér Von Mises tests which evaluate (DeGroot and Schervish, 2012):

$$H_0 : F_x(x) = F_{\hat{y}}(x), \qquad -\infty < x < \infty, \tag{7.2}$$

$$H_1 : \text{Hypothesis } H_0 \text{ is not true.} \tag{7.3}$$

The Kolmogorov-Smirnov (KS) test considers the statistic:

$$D = \sup_{-\infty<x<\infty} |F_x(x) - F_{\hat{y}}(x)|, \tag{7.4}$$

with the Cramér Von Mises (CVM) statistic considering a quadratic discrepancy measure as summarised in Anderson (1962). For the fly data $n = 36$ and taking $N = 1,000$ to calculate $\hat{y}$ for each of the models from section 7.1, enables initial evaluation through the KS and CVM test. Although computationally intensive, permutation tests are considered preferable for small sample sizes (Wasserman, 2004): repeating testing on the models, with a permutation test of size $2,500$ calcuating the KS test and CVM test

statistic each time, consolidates previous findings as summarised in Table 7.4.

Table 7.4: DLS model goodness of fit test p-values

			Permutation Test	
Model	**KS p-value**	**CVM p-value**	**KS p-value**	**CVM p-value**
Gompertz:				
RW[1]	0.35	0.34	0.10	0.15
OU	0.52	0.70	0.67	0.73
Weibull:				
RW	0.35	0.25	0.30	0.29
OU	0.78	0.78	0.65	0.84

7.3.2 Remaining Lifetime Prediction Diagnostic

One way to evaluate the accuracy of remaining lifetime predictions taking into account ageing is to pick a series of stages in each individual's life and to predict the remaining lifetime from those points against the actual residual lifetime to see which models perform better.

If P^* denotes the percentage into an individual's life from which a prediction will be made, the forecast error for individual i, $P^*\%$ into his life, can then be defined as $e_{i,P^*} = \hat{r}_{i,P^*} - T_{i,P^*}$ where $\hat{r}_{i,P^*}$ denotes predicted remaining lifetime and T_{i,P^*} denotes actual remaining lifetime for $i = 1, ..., n$. The absolute errors are then $|e_{i,P^*}|$. The mean absolute error (MAE) and the root mean squared error (RMSE) for the sample

[1]RW denotes random walk

can then be calculated as:

$$MAE_{P^*} = \frac{1}{n}\sum_{i=1}^{n}|e_{i,P^*}| = \frac{1}{n}\sum_{i=1}^{n}|\hat{r}_{i,P^*} - T_{i,P^*}|,$$
$$RMSE_{P^*} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_{i,P^*}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_{i,P^*} - T_{i,P^*})^2}.$$

Focusing on absolute errors firstly, 500 predictions of remaining lifetime are generated using Algorithm 8 for every female fly using the models from Tables 7.1 and 7.2 at a series of points as illustrated in Figures 7.13 and 7.14, where the lower quartile of the simulated lifetimes for each fly is taken as a prediction of the remaining lifetime. The lower quartile is found to be more accurate than both the mean and median because of the left skewness apparent in the distribution of simulated lifetimes as seen in Figure 7.10. When only 25% of the fly's life has been observed, the absolute errors have a large range and as more is known about the ageing process, the predictions start to improve. For both the Gompertz and Weibull, both the random walk and OU models appear to perform similarly at 25%: at 50%, the OU model for the Gompertz has a smaller interquartile range than that for the random walk whereas for the Weibull it's the other way around. At 75%, the median absolute error for both the Gompertz and Weibull models appears to be similar.

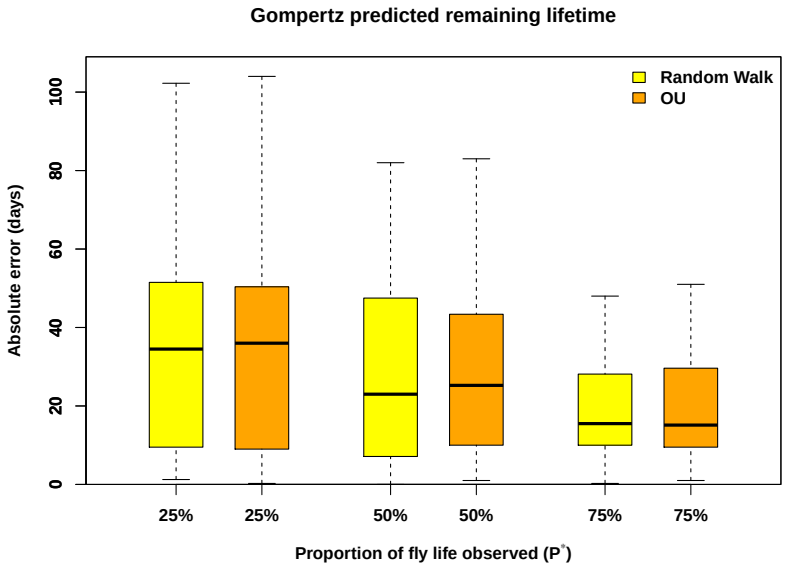

Figure 7.13: Predicted remaining lifetime absolute error boxplot for Gompertz ageing models

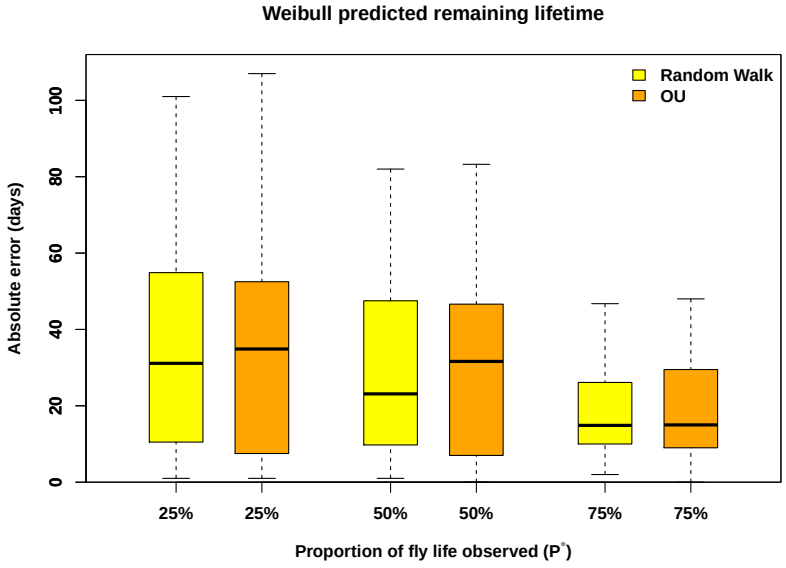

Figure 7.14: Predicted remaining lifetime absolute error boxplot for Weibull ageing models

When evaluating forecasting accuracy, the MAE and RMSE come to the fore and

Table 7.5: Predicted remaining lifetime forecasting error

P*=25%	**Gompertz**		**Weibull**	
	RW	**OU**	**RW**	**OU**
MAE	34.5	34.2	34.1	35.6
RMSE	43.9	44.1	44.3	45.6

P*=50%	**Gompertz**		**Weibull**	
	RW	**OU**	**RW**	**OU**
MAE	27.4	28.3	30.0	29.2
RMSE	35.5	35.5	37.1	37.6

P*=75%	**Gompertz**		**Weibull**	
	RW	**OU**	**RW**	**OU**
MAE	20.9	20.6	19.9	19.2
RMSE	25.8	25.6	24.7	24.1

prove a more popular and effective measure in practice than the absolute errors alone (Armstrong, 2001). From Table 7.5, considering MAE when the first 25% of a fly's life is used to predict remaining lifetime, shows little between the Gompertz and Weibull random walk and Gompertz OU models, with the Weibull OU at least a day more inaccurate on average than the others. Once 50% of the fly's life is observed and more ageing information is available, the Gompertz RW model performs better than the others and is almost 2 days more accurate on average. Once 75% of the fly's life is observed, the Weibull models start to outperform the Gompertz models, with the Weibull OU model most accurate with a predicted remaining lifetime almost a day and a half more accurate than both Gompertz models. The RMSE for the Weibull OU model is also the lowest across all the models.

Chapter 8

Conclusion

Gerontologists often conduct laboratory experiments such as outlined in Zou et al. (2011), in order to understand the ageing process in more detail: often as a result, alongside longitudinal data, survival information is also available and hence longitudinal survival models incorporating a latent process which can be modelled as the ageing process, prove appealing. Chapters 1 and 2 outline these ideas of which parts appear in Steinsaltz, Mohan and Kolb (2012).

In response to modelling requirements set out by the team of researchers responsible for the novel fly lifetime behavioural dataset outlined in Chapter 3, whereby a model capable of modelling a latent ageing process together with longitudinal survival data was desired, a DLS model outlined in state space form is explored. Pavlov (2010) building on Fahrmeir (1994) previously proposed a similar survival model: however as only one previous citation till date (by myself in Steinsaltz, Mohan and Kolb (2012)) testifies,

there are certain problems with his model such as filtering being defined vaguely in a manner contrary to the theory it cited—see section 4.4.3—and the subsequent complete lack of code to verify.

As part of this research, a novel application of SMC methods to survival analysis through a DLS model is outlined which avoids the pitfalls of linearisation in the case of non-Gaussian observations, as in the approximate Kalman filtering realm. Key results include showing the discrete time cumulative hazard approximation to be biased for the Gompertz parameter a on page 82 and showing taking the Gompertz survival function directly to be unbiased, which in follow-up results for the DLS model checking, also performs considerably better as outlined on page 95. A novel remaining lifetime prediction approach is also outlined along with model diagnostics focusing on the empirical survival information and forecasting errors.

Small sample sizes for each of the diets limited the frailty modelling that could be done at a dietary level, as all the attempted models were found to be insignificant. In future, larger sample sizes can help shed more light on ageing effects given dietary restrictions: Pletcher (1999) to this extent advocates sample sizes in excess of 100 and ideally closer to 500. The small sample sizes also created issues with bias and in future work, potential bias-correction approaches can be explored. We looked at one possible transformation (absolute value differences between peak activity recorded till date and daily observed activity) of the data and other transformations could perform better which can be explored further in future work.

Traditional joint longitudinal survival models build on the proportional hazards frame-

work and hence adopting a state space modelling approach provides an alternative which not only explicitly allows postulation of a latent process governing observations in a hidden Markov model framework but adopts a dynamic setup at the outset and models a dynamic survival probability incorporating frailty rather than the hazard function alone. There's a strong recent focus on enabling dynamic predictions using joint models (Barrett and Su, 2017) and consequently, a state space modelling framework utilising SMC methods provides both a natural and novel foundation.

All software developed as part of this research is available: the C and R code is currently being written up as a R package and in its current form is being applied to reproductive data from the 1958 Cohort Study by researchers based at the Department of Statistics in Oxford.

Future work can extend the range of possible frailties: ageing theories continue to evolve and as such the flexiblity of a state space model allows further postulations.

References

Aalen, O. (1978), 'Nonparametric inference for a family of counting processes', *The Annals of Statistics* **6**(4), 701–726.

Aalen, O. and Gjessing, H. (2001), 'Understanding the Shape of the Hazard Rate: A Process Point of View', *Statistical Science* **16**(41), 1–14.

Aalen, O. O. (1994), 'Effects of frailty in survival analysis', *Statistical Methods in Medical Research* **3**(3), 227–243.

Aalen, O. O. and Gjessing, H. K. (2004), 'Survival Models Based on the Ornstein-Uhlenbeck Process', *Lifetime Data Analysis* **10**(4), 407–423.

Abrams, P. A. and Ludwig, D. (1995), 'Optimality theory, Gompertz' law, and the disposable soma theory of senescence', *Evolution* pp. 1055–1066.

Al-Shomrani, A. A., Shawky, A., Arif, O. H. and Aslam, M. (2016), 'Log-logistic distribution for survival data analysis using MCMC', *SpringerPlus* **5**(1), 1774.

Andersen, P. (2010), 'Editorial: Recent developments in survival analysis', *Statistical Methods in Medical Research* **19**(1), 3–4.

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis, 3rd edition*, New York: John Wiley & Sons.

Anderson, T. W. (1962), 'On the distribution of the two-sample Cramer-von Mises criterion', *The Annals of Mathematical Statistics* pp. 1148–1159.

Andrieu, C. and Doucet, A. (2002), 'Particle filtering for partially observed Gaussian state space models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 827–836.

Arbeev, K. G., Akushevich, I., Kulminski, A. M., Land, K. C. and Yashin, A. I. (2016), Approaches to statistical analysis of longitudinal data on aging, health, and longevity: Biodemographic perspectives, *in* 'Biodemography of Aging', Springer, pp. 241–261.

Arking, R. (2006), *The Biology of Aging: Observations and Principles, 3rd Edition*, Oxford University Press, USA.

Armstrong, J. S. (2001), *Principles of forecasting: a handbook for researchers and practitioners*, New York: Springer Science & Business Media.

Arulampalam, M. S., Maskell, S., Gordon, N. and Clapp, T. (2002), 'A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking', *IEEE Transactions on signal processing* **50**(2), 174–188.

Bagdonavičius, V., Cheminade, O. and Nikulin, M. (2004), 'Statistical planning and inference in accelerated life testing using the chss model', *Journal of Statistical Planning and Inference* **126**(2), 535–551.

Bagdonavičius, V. and Nikulin, M. (2001), *Accelerated life models: modeling and statistical analysis*, Boca Raton, FL: CRC Press.

Bagdonavičius, V. and Nikulin, M. (2009), 'Statistical models to analyze failure, wear, fatigue, and degradation data with explanatory variables', *Communications in Statistics–Theory and Methods* **38**(16-17), 3031–3047.

Balakrishnan, N. and Kateri, M. (2008), 'On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data', *Statistics & Probability Letters* **78**(17), 2971–2975.

Bansaye, V., Delmas, J.-F., Marsalle, L. and Tran, V. C. (2011), 'Limit theorems for Markov processes indexed by continuous time Galton-Watson trees', *The Annals of Applied Probability* pp. 2263–2314.

Barrett, J. and Su, L. (2017), 'Dynamic predictions using flexible joint models of longitudinal and time-to-event data', *Statistics in Medicine* **36**(9), 1447–1460.

Beard, R. E. (1959), Appendix: Note on some mathematical mortality models, *in* 'Ciba Foundation Symposium-The Lifespan of Animals (Colloquia on Ageing), Volume 5', Wiley Online Library, pp. 302–311.

Bengtson, V. L. and Settersten Jr, R. (2016), *Handbook of Theories of Aging, Third Edition*, New York: Springer Publishing Company.

Bertsekas, D. P. and Tsitsiklis, J. N. (2002), *Introduction to probability*, Belmont, MA: Athena Scientific.

Bonsall, M. (2006), 'Longevity and ageing: appraising the evolutionary consequences of

growing old', *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**(1465), 119.

Bycott, P. and Taylor, J. (1998), 'A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model', *Statistics in Medicine* **17**(18), 2061–2077.

Cacciari, M., Mazzanti, G. and Montanari, G. (1996), 'Comparison of maximum likelihood unbiasing methods for the estimation of the Weibull parameters', *IEEE Transactions on Dielectrics and Electrical Insulation* **3**(1), 18–27.

Cappé, O., Godsill, S. J. and Moulines, E. (2007), 'An overview of existing methods and recent advances in sequential monte carlo', *Proceedings of the IEEE* **95**(5), 899–924.

Cappé, O., Moulines, E. and Rydén, T. (2005), *Inference in Hidden Markov Models*, New York: Springer.

Carey, J., Papadopoulos, N., Kouloussis, N., Katsoyannos, B., Muller, H., Wang, J. and Tseng, Y. (2006), 'Age-specific and lifetime behavior patterns in Drosophila melanogaster and the Mediterranean fruit fly, Ceratitis capitata', *Experimental Gerontology* **41**(1), 93–97.

Carey, J. R., Liedo, P., Orozco, D. and Vaupel, J. W. (1992), 'Slowing of mortality rates at older ages in large medfly cohorts', *Sci-ence* **258**, 457–91.

Carey, J. R., Zou, S., Liedo, P., Robles, L., Morice, A., Müller, H. G., Wang, J.-L., Ingram, D. and Papadopoulos, N. (2010), A high-resolution system for recording the

daily and lifetime behavioral and movement patterns of individual tephritid fruit flies, *in* 'Proceedings of Measuring Behavior', pp. 24–27.

Carpenter, J. and Bithell, J. (2000), 'Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians', *Statistics in Medicine* **19**(9), 1141–1164.

Carter, C. K. and Kohn, R. (1994), 'On Gibbs sampling for state space models', *Biometrika* pp. 541–553.

Chen, D.-G., Sun, J. and Peace, K. E. (2013), *Interval-Censored Time-to-Event Data: Methods and Applications*, Boca Raton, FL: CRC Press.

Chen, M., Zhang, Z. and Cui, C. (2017), 'On the Bias of the Maximum Likelihood Estimators of Parameters of the Weibull Distribution', *Mathematical and Computational Applications* **22**(1), 19.

Chiu, J. C., Kaub, K., Zou, S., Liedo, P., Altamirano-Robles, L., Ingram, D. and Carey, J. R. (2013), 'Deleterious effect of suboptimal diet on rest-activity cycle in Anastrepha ludens manifests itself with age', *Scientific Reports* **3**(173).

Chopin, N. (2004), 'Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference', *The Annals of Statistics* **32**(6), 2385–2411.

Chu, C. C. and Lee, R. D. (2006), 'The co-evolution of intergenerational transfers and longevity: An optimal life history approach', *Theoretical Population Biology* **69**(2), 193–201.

Cichoń, M. (1997), 'Evolution of longevity through optimal resource allocation', *Proceedings of the Royal Society of London B: Biological Sciences* **264**(1386), 1383–1388.

Collett, D. (2003), *Modelling Survival Data in Medical Research, 2nd Edition*, Boca Raton, FL: CRC press.

Cowling, B., Hutton, J. and Shaw, J. (2006), 'Joint modelling of event counts and survival times', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55**(1), 31–39.

Cox, D. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220.

Creal, D. (2012), 'A survey of sequential monte carlo methods for economics and finance', *Econometric Reviews* **31**(3), 245–296.

Curtsinger, J. W., Gavrilova, N. S. and Gavrilov, L. A. (2006), Biodemography of aging and age-specific mortality in Drosophila melanogaster, *in* 'Handbook of the Biology of Aging, 6th edition', San Diego: Academic Press, pp. 261–288.

Dafni, U. and Tsiatis, A. (1998), 'Evaluating surrogate markers of clinical outcome when measured with error', *Biometrics* pp. 1445–1462.

Darwin, C. (1859), *On the origin of the species by natural selection*, London: John Murray.

Dasgupta, S. (1994), 'A computer simulation for biological ageing', *Journal de Physique I* **4**(10), 1563–1570.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap methods and their application*, Cambridge: Cambridge University Press.

Davison, A. C., Hinkley, D. V. and Young, G. A. (2003), 'Recent developments in bootstrap methodology', *Statistical Science* pp. 141–157.

De Jong, P. and Shephard, N. (1995), 'The simulation smoother for time series models', *Biometrika* pp. 339–350.

DeGroot, M. and Schervish, M. (2012), *Probability and Statistics, 4th edition*, Boston: Addison-Wesley.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002), *Analysis of Longitudinal Data, 2nd edition*, Oxford: Oxford University Press.

Douc, R., Cappé, O. and Moulines, E. (2005), Comparison of resampling schemes for particle filtering, *in* 'Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)', pp. 64–69.

Doucet, A. (2010), 'A Note on Efficient Conditional Simulation of Gaussian Distributions) ', [Technical Report], (Updated Apr, 2010), *Available at: http://www.stats.ox.ac.uk/~doucet/doucet_simulationconditionalgaussian.pdf* . (Accessed Jun 18, 2017).

Doucet, A., De Freitas, N. and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, New York: Springer.

Doucet, A., Godsill, S. and Andrieu, C. (2000), 'On sequential Monte Carlo sampling methods for Bayesian filtering', *Statistics and computing* **10**(3), 197–208.

Doucet, A. and Johansen, A. M. (2009), 'A tutorial on particle filtering and smoothing: Fifteen years later', *Handbook of Nonlinear Filtering* **12**(656-704), 3.

Duchateau, L. and Janssen, P. (2007), *The frailty model*, Springer Science & Business Media.

Durbin, J. and Koopman, S. (2001), *Time series analysis by state space methods*, Oxford: Oxford University Press.

Durbin, J. and Koopman, S. J. (2000), 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(1), 3–56.

Durbin, J. and Koopman, S. J. (2002), 'A simple and efficient simulation smoother for state space time series analysis', *Biometrika* **89**(3), 603–616.

Durbin, J. and Koopman, S. J. (2012), *Time Series Analysis by State Space Methods, 2nd edition*, Oxford: Oxford University Press.

Efron, B. and Tibshirani, R. J. (1994), *An introduction to the bootstrap*, Boca Raton, FL: CRC press.

Elashoff, R., Li, N. et al. (2017), *Joint Modeling of Longitudinal and Time-to-Event Data*, Boca Raton, FL: CRC Press.

Evans, S. N. and Steinsaltz, D. (2007), 'Damage segregation at fissioning may increase growth rates: A superprocess model', *Theoretical Population Biology* **71**(4), 473–490.

Fahrmeir, L. (1992), 'Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models', *Journal of the American Statistical Association* **87**(418), 501–509.

Fahrmeir, L. (1994), 'Dynamic modelling and penalized likelihood estimation for discrete time survival data', *Biometrika* **81**(2), pp. 317–330.

Fahrmeir, L. and Kaufmann, H. (1991), 'On Kalman filtering, posterior mode estimation and fisher scoring in dynamic exponential family regression', *Metrika* **38**(1), 37–60.

Faucett, C. and Thomas, D. (1996), 'Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach', *Statistics in Medicine* **15**(15), 1663–1685.

Fearnhead, P. (2008), 'Computational methods for complex stochastic systems: a review of some alternatives to MCMC', *Statistics and Computing* **18**(2), 151–171.

Fernhead, P. (2012), 'Modern Computational Statistics: Alternatives to MCMC) ', [Lecture notes], (Updated Jul 27, 2012), *Available at: http://www.maths.lancs.ac.uk/~fearnhea/GTP/GTP_Slides.pdf* . (Accessed Jun 6, 2017).

Finkelstein, M. (2012), 'Discussing the Strehler-Mildvan model of mortality', *Demographic Research* **26**(9), 191–206.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008), *Longitudinal data analysis*, Boca Raton, FL: CRC Press.

Fontana, L., Partridge, L. and Longo, V. (2010), 'Extending Healthy Life Span–from Yeast to Humans', *Science* **328**(5976), 321.

Frühwirth-Schnatter, S. (1994), 'Data augmentation and dynamic linear models', *Journal of Time Series Analysis* **15**(2), 183–202.

Gamerman, D. (1991), 'Dynamic Bayesian Models for Survival Data', *Applied Statistics* pp. 63–79.

Gavrilov, L. A. and Gavrilova, N. S. (1991), 'The biology of life span: a quantitative approach.'.

Gavrilov, L. A. and Gavrilova, N. S. (2004), 'The reliability-engineering approach to the problem of biological aging', *Annals of the New York Academy of Sciences* **1019**(1), 509–512.

Gelman, A. and Hill, J. (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Gentleman, R. and Geyer, C. J. (1994), 'Maximum likelihood for interval censored data: Consistency and computation', *Biometrika* **81**(3), 618–623.

Godsill, S. J., Doucet, A. and West, M. (2004), 'Monte Carlo smoothing for nonlinear time series', *Journal of the American Statistical Association* **99**(465), 156–168.

Goldsmith, T. (2011), *An Introduction to Biological Aging Theory*, Maryland: Azinet.

Gómez, G., Calle, M. L., Oller, R. and Langohr, K. (2009), 'Tutorial on methods

for interval-censored data and their implementation in R', *Statistical Modelling* **9**(4), 259–297.

Gompertz, B. (1825), 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies', *Philosophical Transactions of the Royal Society of London* **115**, 513–583.

Gordon, N. J., Salmond, D. J. and Smith, A. F. (1993), 'Novel approach to nonlinear/non-Gaussian Bayesian state estimation', *IEE Proceedings F (Radar and Signal Processing)* **140**(2), 107–113.

Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S. and Bois, F. Y. (2015), 'Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group', *Statistics in Medicine* **34**(14), 2181–2195.

Grotewiel, M., Martin, I., Bhandari, P. and Cook-Wiens, E. (2005), 'Functional senescence in Drosophila melanogaster', *Ageing Research Reviews* **4**(3), 372–397.

Hanagal, D. D. (2011), *Modeling survival data using frailty models*, Boca Raton, FL: CRC Press.

Harrison, P. J. and Stevens, C. F. (1976), 'Bayesian forecasting', *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(3), 205–247.

Harvey, A. (1993), *Time Series Models, 2nd edition*, Cornwall: Harvester Wheatsheaf.

Hashemi, R., Jacqmin-Gadda, H. and Commenges, D. (2003), 'A latent process model for joint modeling of events and marker', *Lifetime Data Analysis* **9**(4), 331–343.

Helske, J. (2017), 'KFAS: Exponential Family State Space Models in R', *Journal of Statistical Software* **78**(10), 1–39.

Helske, J., Nyblom, J., Ekholm, P. and Meissner, K. (2013), 'Estimating aggregated nutrient fluxes in four Finnish rivers via Gaussian state space models', *Environmetrics* **24**(4), 237–247.

Hemming, K. and Shaw, J. (2005), 'A class of parametric dynamic survival models', *Lifetime data analysis* **11**(1), 81–98.

Henderson, R., Diggle, P. and Dobson, A. (2000), 'Joint modelling of longitudinal measurements and event time data', *Biostatistics* **1**(4), 465–480.

Hirose, H. (1999), 'Bias correction for the maximum likelihood estimates in the two-parameter Weibull distribution', *IEEE Transactions on Dielectrics and Electrical Insulation* **6**(1), 66–68.

Ho, Y. and Lee, R. (1964), 'A Bayesian approach to problems in stochastic estimation and control', *IEEE Transactions on Automatic Control* **9**(4), 333–339.

Hogan, J. and Laird, N. (1997), 'Mixture models for the joint distribution of repeated measures and event times', *Statistics in Medicine* **16**(3), 239–257.

Hoover, D., Graham, N., Chen, B., Taylor, J., Phair, J., Zhou, S., Muñoz, A. et al. (1992), 'Effect of CD4+ cell count measurement variability on staging HIV-1 infection', *Journal of Acquired Immune Deficiency Syndromes* **5**(8), 794–802.

Hougaard, P. (1986), 'A class of multivariate failure time distributions', *Biometrika* pp. 671–678.

Hougaard, P. (1995), 'Frailty models for survival data', *Lifetime Data Analysis* **1**(3), 255–273.

Ibrahim, J., Chen, M. and Sinha, D. (2005), *Bayesian survival analysis*, New York: Springer.

Iliadi, K. and Boulianne, G. (2010), 'Age-related behavioral changes in drosophila', *Annals of the New York Academy of Sciences* **1197**(1), 9–18.

Jacobson, J., Lambert, A., Portero-Otín, M., Pamplona, R., Magwere, T., Miwa, S., Driege, Y., Brand, M. and Partridge, L. (2010), 'Biomarkers of aging in Drosophila', *Aging Cell* **9**(4), 466–477.

Jarociński, M. (2015), 'A note on implementing the Durbin and Koopman simulation smoother', *Computational Statistics & Data Analysis* **91**, 1–3.

Johnson, L. R. and Mangel, M. (2006), 'Life histories and the evolution of aging in bacteria and other single-celled organisms', *Mechanisms of Ageing and Development* **127**(10), 786–793.

Jungbacker, B. and Koopman, S. J. (2007), 'Monte Carlo estimation for nonlinear non-Gaussian state space models', *Biometrika* **94**(4), 827–839.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering* **82**(1), 35–45.

Kannisto, V., Lauritsen, J., Thatcher, A. R. and Vaupel, J. W. (1994), 'Reductions in mortality at advanced ages: several decades of evidence from 27 countries', *Population and Development Review* pp. 793–810.

Kaplan, E. and Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481.

Kirkwood, T. (1977), 'Evolution of ageing', *Nature* **270**, 301–304.

Kitagawa, G. and Gersch, W. (1996), *Smoothness priors analysis of time series*, New York: Springer Science & Business Media.

Kleinbaum, D. G. and Klein, M. (2012), 'Survival analysis: A self-learning text, 3rd edition'.

Lancaster, T. (1979), 'Econometric methods for the duration of unemployment', *Econometrica: Journal of the Econometric Society* pp. 939–956.

Lange, N., Carlin, B. and Gelfand, A. (1992), 'Hierarchical bayes models for the progression of hiv infection using longitudinal cd4 t-cell numbers', *Journal of the American Statistical Association* pp. 615–626.

Le Bourg, É. (2001), 'A mini-review of the evolutionary theories of aging. is it the time to accept them?', *Demographic Research* **4**(1), 1–28.

Le Bras, H. (1976), 'Lois de mortalité et age limite', *Population (french edition)* pp. 655–692.

Lee, M., DeGruttola, V. and Schoenfeld, D. (2000), 'A model for markers and latent health status', *Journal of the Royal Statistical Society. Series B, Statistical Methodology* pp. 747–762.

Lee, M. and Whitmore, G. (2006), 'Threshold regression for survival analysis: Modeling

event times by a stochastic process reaching a boundary', *Statistical Science* pp. 501–513.

Lee, M., Whitmore, G. and Rosner, B. (2010), 'Threshold regression for survival data with time-varying covariates', *Statistics in Medicine* **29**(7-8), 896–905.

Lenart, A. (2014), 'The moments of the Gompertz distribution and maximum likelihood estimation of its parameters', *Scandinavian Actuarial Journal* **2014**(3), 255–277.

Li, J. and Lee, M.-L. T. (2011), 'Analysis of failure time using threshold regression with semi-parametric varying coefficients', *Statistica Neerlandica* **65**(2), 164–182.

Li, T. and Anderson, J. J. (2009), 'The vitality model: A way to understand population survival and demographic heterogeneity', *Theoretical Population Biology* **76**(2), 118–131.

Lindsey, J. C. and Ryan, L. M. (1998), 'Methods for interval-censored data', *Statistics in Medicine* **17**(2), 219–238.

Liu, J. S. and Chen, R. (1995), 'Blind deconvolution via sequential imputations', *Journal of the American Statistical Association* **90**(430), 567–576.

Ljubuncic, P. and Reznick, A. Z. (2009), 'The evolutionary theories of aging revisited–a mini-review', *Gerontology* **55**, 205–216.

McEwen, B. S. (2002), 'Sex, stress and the hippocampus: allostasis, allostatic load and the aging process', *Neurobiology of Aging* **23**(5), 921–939.

Medawar, P. (1952), *An Unsolved Problem of Biology*, London: HK Lewis.

Medvedev, Z. (1990), 'An attempt at a rational classification of theories of ageing', *Biological Reviews* **65**(3), 375–398.

Mikkonen, S., Laine, M., Mäkelä, H., Gregow, H., Tuomenvirta, H., Lahtinen, M. and Laaksonen, A. (2015), 'Trends in the average temperature in finland, 1847–2013', *Stochastic Environmental Research and Risk Assessment* **29**(6), 1521–1529.

Mitnitski, A. B., Graham, J. E., Mogilner, A. J. and Rockwood, K. (2002), 'Frailty, fitness and late-life mortality in relation to chronological and biological age', *BMC Geriatrics* **2**(1), 1.

Monaghan, P., Charmantier, A., Nussey, D. H. and Ricklefs, R. E. (2008), 'The evolutionary ecology of senescence', *Functional Ecology* **22**(3), 371–378.

Newsom, J., Jones, R. N. and Hofer, S. M. (2013), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*, New York: Routledge.

Pavlov, A. (2010), 'A new approach in survival analysis with longitudinal covariates', Ph.D. thesis, Queen's University (Canada), *URL: http://qspace.library.queensu.ca/jspui/bitstream/1974/5585/1/Pavlov_Andrey_201004_PhD.pdf* .

Pawitan, Y. and Self, S. (1993), 'Modeling Disease Market Processes in AIDS', *Journal of the American Statistical Association* pp. 719–726.

Penna, T. and Stauffer, D. (1995), 'Efficient Monte Carlo simulation of biological aging', *International Journal of Modern Physics C* **6**(02), 233–239.

Peto, R. (1973), 'Experimental survival curves for interval-censored data', *Applied Statistics* pp. 86–91.

Piper, M. D. W. and Partridge, L. (2016), *Drosophila: Methods and Protocols*, New York: Springer Publishing Company.

Pletcher, S. D. (1999), 'Model fitting and hypothesis testing for age-specific mortality data', *Journal of Evolutionary Biology* **12**(3), 430–439.

Poirier, L., Cabo, R. and Zou, S. (2010), Dietary restriction and aging in drosophila melanogaster, *in* A. V. Everitt, S. I. S. Rattan, D. G. Couteur and R. Cabo, eds, 'Calorie Restriction, Aging and Longevity', Springer Netherlands, pp. 79–95.

Prentice, R. (1982), 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika* **69**(2), 331–342.

Price, D. L. and Manatunga, A. K. (2001), 'Modelling survival data with a cured fraction using frailty models', *Statistics in Medicine* **20**(9-10), 1515–1527.

Raboud, J., Reid, N., Coates, R. and Farewell, V. (1993), 'Estimating risks of progressing to aids when covariates are measured with error', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 393–406.

Richards, S. (2008), 'Applying survival models to pensioner mortality data', *British Actuarial Journal* **14**(02), 257–303.

Ristic, B., Arulampalam, M. S. and Gordon, N. (2004), *Beyond the Kalman filter: Particle filters for tracking applications*, London: Artech House.

Rizopoulos, D. (2011), 'Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data', *Biometrics* **67**(3), 819–829.

Rizopoulos, D. (2012), *Joint Models for Longitudinal and time-to-event data: With applications in R*, Boca Raton, FL: CRC Press.

Rockwood, K., Mitnitski, A. B. and MacKnight, C. (2002), 'Some mathematical models of frailty and their clinical implications', *Reviews in Clinical Gerontology* **12**(02), 109–117.

Sacher, G. and Trucco, E. (1962), 'The stochastic theory of mortality', *Annals of the New York Academy of Sciences* **96**(1), 985–1007.

Sastry, N. (1997), 'A nested frailty model for survival data, with an application to the study of child survival in northeast brazil', *Journal of the American Statistical Association* **92**(438), 426–435.

Scott, S. L. and Varian, H. R. (2014), 'Predicting the present with Bayesian structural time series', *International Journal of Mathematical Modelling and Numerical Optimisation* **5**(1-2), 4–23.

Shen, Y. and Yang, Z. (2015), 'Bias-correction for Weibull common shape estimation', *Journal of Statistical Computation and Simulation* **85**(15), 3017–3046.

Shephard, N. (1994), 'Partial non-Gaussian state space', *Biometrika* pp. 115–131.

Singpurwalla, N. (1995), 'Survival in dynamic environments', *Statistical Science* **10**(1), 86–103.

Sousa, I. (2011), 'A review on joint modelling of longitudinal measurements and time-to-event', *Revstat–Statistical Journal* **9**, 57–81.

Steinsaltz, D. and Evans, S. N. (2004), 'Markov mortality models: implications of quasistationarity and varying initial distributions', *Theoretical Population Biology* **65**(4), 319–337.

Steinsaltz, D., Mohan, G. and Kolb, M. (2012), 'Markov models of aging: theory and practice', *Experimental Gerontology* **47**(10), 792–802.

Steinsaltz, D. R. and Wachter, K. W. (2006), 'Understanding Mortality Rate Deceleration and Heterogeneity', *Mathematical Population Studies* **13**, 19–37.

Strehler, B. L. (1977), *Time, Cells, and Aging*, New York: Academic Press.

Strehler, B. L. and Mildvan, A. S. (1960), 'General theory of mortality and aging', *Science* **132**.

Sun, J. (2007), *The Statistical Analysis of Interval-Censored Failure Time Data*, New York: Springer Science & Business Media.

Troxel, A. (2002), 'Techniques for incorporating longitudinal measurements into analyses of survival data from clinical trials', *Statistical Methods in Medical Research* **11**(3), 237–245.

Tsiatis, A. and Davidian, M. (2004), 'Joint modeling of longitudinal and time-to-event data: an overview', *Statistica Sinica* **14**(3), 809–834.

Tsiatis, A., DeGruttola, V. and Wulfsohn, M. (1995), 'Modeling the relationship of

survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS', *Journal of the American Statistical Association* pp. 27–37.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 290–295.

Uhlenbeck, G. E. and Ornstein, L. S. (1930), 'On the theory of the Brownian motion', *Physical Review* **36**(5), 823–841.

Vaupel, J., Manton, K. and Stallard, E. (1979), 'The impact of heterogeneity in individual frailty on the dynamics of mortality', *Demography* **16**(3), 439–454.

Vaupel, J. W., Carey, J. R., Christensen, K., Johnson, T. E., Yashin, A. I., Holm, N. V., Iachine, I. A., Kannisto, V., Khazaeli, A. A., Liedo, P. et al. (1998), 'Biodemographic trajectories of longevity', *Science* **280**(5365), 855–860.

Vaupel, J. W. and Yashin, A. I. (1985), 'Heterogeneity's ruses: some surprising effects of selection on population dynamics', *The American Statistician* **39**(3), 176–185.

Verbeke, G., Molenberghs, G. and Rizopoulos, D. (2010), Random effects models for longitudinal data, *in* K. van Montfort, J. H. Oud and A. Satorra, eds, 'Longitudinal Research with Latent Variables', Berlin: Springer-Verlag, pp. 37–96.

Wang, Y. and Taylor, J. (2001), 'Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome', *Journal of the American Statistical Association* **96**(455), 895–905.

Warner, H. R. (2004), 'The future of aging interventions: Current status of efforts to measure and modulate the biological rate of aging', *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **59**(7), B692–B696.

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer.

Weismann, A. (1882), *Uber die Dauer des Lebens*, Jenas.

Weitz, J. S. and Fraser, H. B. (2001), 'Explaining mortality rate plateaus', *Proceedings of the National Academy of Sciences* **98**(26), 15383–15386.

West, M., Harrison, P. J. and Migon, H. S. (1985), 'Dynamic generalized linear models and bayesian forecasting', *Journal of the American Statistical Association* **80**(389), 73–83.

Whitmore, G., Crowder, M. and Lawless, J. (1998), 'Failure inference from a marker process based on a bivariate Wiener model', *Lifetime Data Analysis* **4**(3), 229–251.

Wienke, A. (2010), *Frailty models in survival analysis*, Boca Raton, FL: CRC Press.

Woodbury, M. A. and Manton, K. G. (1977), 'A random-walk model of human mortality and aging', *Theoretical Population Biology* **11**(1), 37–48.

Wulfsohn, M. and Tsiatis, A. (1997), 'A joint model for survival and longitudinal data measured with error', *Biometrics* pp. 330–339.

Xu, J. and Zeger, S. (2001), 'Joint analysis of longitudinal data comprising repeated

measures and times to events', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(3), 375–387.

Yashin, A., Arbeev, K., Akushevich, I., Kulminski, A., Akushevich, L. and Ukraintseva, S. (2007), 'Stochastic model for analysis of longitudinal data on aging and mortality', *Mathematical Biosciences* **208**(2), 538–551.

Yashin, A. I., Akushevich, I., Arbeev, K. G., Kulminski, A. and Ukraintseva, S. (2011), 'Joint analysis of health histories, physiological state, and survival', *Mathematical Population Studies* **18**(4), 207–233.

Yashin, A. I., Arbeev, K. G., Akushevich, I., Kulminski, A., Akushevich, L. and Ukraintseva, S. V. (2008), 'Model of hidden heterogeneity in longitudinal data', *Theoretical Population Biology* **73**(1), 1–10.

Yashin, A. I., Arbeev, K. G., Akushevich, I., Kulminski, A., Ukraintseva, S. V., Stallard, E. and Land, K. C. (2012), 'The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span', *Physics of Life Reviews* **9**(2), 177–188.

Yashin, A. I., Iachine, I. A. and Begun, A. S. (2000), 'Mortality modeling: A review', *Mathematical Population Studies* **8**(4), 305–332.

Yashin, A. I. and Manton, K. G. (1997), 'Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies', *Statistical Science* pp. 20–34.

Yashin, A. I., Stallard, E. and Land, K. C. (2016), *Biodemography of Aging: Determinants of Healthy Life Span and Longevity*, Dordrecht BV: Springer.

Young, P. C., Ng, C. N., Lane, K. and Parker, D. (1991), 'Recursive forecasting, smoothing and seasonal adjustment of non-stationary environmental data', *Journal of Forecasting* **10**(1-2), 57–89.

Yu, M., Law, N., Taylor, J. and Sandler, H. (2004), 'Joint longitudinal-survival-cure models and their application to prostate cancer', *Statistica Sinica* **14**(3), 835–862.

Zhu, B., Dunson, D. B. et al. (2017), 'Bayesian functional data modeling for heterogeneous volatility', *Bayesian Analysis* **12**(2), 335–350.

Zou, S., Liedo, P., Altamirano-Robles, L., Cruz-Enriquez, J., Morice, A., Ingram, D., Kaub, K., Papadopoulos, N. and Carey, J. (2011), 'Recording Lifetime Behavior and Movement in an Invertebrate Model', *PloS ONE* **6**(4), e18151.

Appendices

Appendix A

Results for Multivariate Normal Distribution

Lemma 1

If x and y are jointly normally distributed random vectors defining a bivariate Normal distribution such that (Durbin and Koopman, 2001, Anderson, 2003):

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{pmatrix} \right]$$

where Σ_{yy} is assumed to be non-singular, then:

$$\mathrm{E}(x|y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \tag{A.1}$$

$$\mathrm{Var}(x|y) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T. \tag{A.2}$$

Lemma 2

If x, y, z are jointly normally distributed with means μ_p and covariance matrices $\Sigma_{pq} = E[(p - \mu_p)(q - \mu_q)^T]$ for $p, q = x, y$ and z with $\mu_z = 0$ and $\Sigma_{yz} = 0$, then (Durbin and Koopman, 2012):

$$\mathrm{E}(x|y,z) = \mathrm{E}(x|y) + \Sigma_{xz}\Sigma_{zz}^{-1}z, \tag{A.3}$$

$$\mathrm{Var}(x|y,z) = \mathrm{Var}(x|y) - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}^{T}. \tag{A.4}$$