A new approach in survival analysis with longitudinal covariates

by

Andrey Pavlov

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
April 2010

Abstract

In this study we look at the problem of analysing survival data in the presence of longitudinally collected covariates. New methodology for analysing such data has been developed through the use of hidden Markov modeling. Special attention has been given to the case of large information volume, where a preliminary data reduction is necessary. Novel graphical diagnostics have been proposed to assess goodness of fit and significance of covariates.

The methodology developed has been applied to the data collected on behaviors of Mexican fruit flies, which were monitored throughout their lives. It has been found that certain patterns in eating behavior may serve as an aging marker. In particular it has been established that the frequency of eating is positively correlated with survival times.

Acknowledgements

First of all I am grateful to my supervisors Dr. David Steinsaltz and Dr. David Thomson for their support and advice during the years of my work.

I also would like to thank Dr. Dongsheng Tu for the perfect introduction to survival analysis; his lectures helped me determine my statistical interests.

My very special thanks go to Andrew Day, without whom I simply would not be the kind of statistician I am. Most of my R and SAS skills developed during his computational analysis classes, and were essential in obtaining results for this project.

And of course I thank my wife Valerie for her incredible patience, love and support.

Table of Contents

Chapter 1

Introduction

The questions of aging and mortality have been attracting scientists' attention for centuries. To date, there is no consensus on how exactly aging occurs and how it manifests itself through physiological characteristics one can observe. Supported by growth of mathematical knowledge and advancement of computer technologies on the one side, and increasing demand from industry on the other, the research on aging became very active in the second half of the 20th century. Rich with theories and ideas, it produced a variety of mathematical models, aiming to describe mortality in a population and to understand the mechanism of aging.

The most notable model of the past, which dates back to 1825, is known as the Gompertz law, which postulates exponential mortality, [12]. A significant amount of research evolved around this law, attempting to justify it theoretically as well as to collect empirical envidence either for or against it. It was noticed that exponential growth of mortality may be remarkably consistent with the observed mortality in certain species, including ourselves. As more sizable datasets became available, the Gompertz law was often found to fail to describe observed mortality throughout the

entire lifespan, especially in late life. Much evidence was collected on this account; in particular, Vaupel et al [53] presented mortality data for large cohorts of several species. The cohorts exhibited very different patterns of mortality. A striking feature of most mortality curves was their tendency to decelerate and in some cases even decline at advanced ages.

Central to the further development of mortality research is recognition of heterogeneity in a population. Individuals may age differently, depending on their genetic background, particular environmental conditions, risk factors they are exposed to and their interactions. Heterogeneity has its impact on the overall mortality curve that one observes in a cohort. In particular, several authors ([91]) have demonstated that it causes the cohort mortality to plateau, even if individual mortalities do not. Many different approaches have been taken to model heterogeneity. It may either be an observed characteristic, such as gender or treatment, or an unobserved random effect, usually termed 'frailty'. Some of these methods, most notably the celebrated Cox proportional hazards model [22], have become classical survival analysis tools, and are offered by major statistical packages such as SAS, SPSS and S-Plus.

Although some models have been more successful than others, there is little hope to arrive at a universal theory which would provide a mathematical formula for aging. What is clear though is that aging can be observed indirectly. Numberless studies have shown association between survival and various observable subject specific characteristics, usually referred to as survival covariates. It is thus common to see descriptive or predictive survival studies, seeking to predict one's lifespan given certain background variables or characteristics obtained in a cross-sectional study. Studies trying to answer *why* certain mortality patterns arise are much more rare.

Modern survival analysis goes further in exploiting subject specific information to predict survival. Several studies have shown that effects of aging may be seen in changes of certain characteristics over time. Such an idea calls for a longitudinal study, where the variable of interest is measured several times for each subject. For instance, Akushevich et al. [43] applied their model to the Framingham Heart Study 46-year follow-up data. One's health state was represented by a 11-dimensional vector, which included the components of age, pulse, blood pressure, body mass index and several others. In recognition of the fact that there could be other dimensions of the health state, the covariates were modeled as a stochastic diffusion.

While it is common with clinical researchers to focus on predicting survival based on observed covariates, authors from the aging research community have attempted to model aging explicitly. The stochastic nature of aging was realized in several works, which did or did not consider covariate data. The idea brought a new concept to the field - appearing in many variations and interchangeably termed 'vitality', 'viability', 'senescence' or 'health state' - a random process responsible for aging and eventually killing of an individual. Technical grounds in this approach are usually easy to criticize, because it is hard to justify that the aging process has one mathematical form or another. However the idea of an unobserved random aging process has proven to produce very flexible models, thus also being attractive from the survival forecast perspective. Furthermore, it allows for hierarchical modeling of longitudinal data, providing an intuitive model interpretation.

Recent years have seen a substantial advancement in joint modeling of longitudinal and survival data. However, as Yashin et al pointed out in their paper of 2007,

statistical theory for longitudinal survival analysis is somewhat behind the ever growing amount of longitudinal data. A lack of modeling techniques in the field leaves most of the data unexplored. Thus, the approach of random senescence and random covariates seems very attractive and worth developing.

In this study we link together several ideas that appeared in aging research and survival analysis, including the Gompertz law, proportional hazards and stochastic aging. We exploit the power, flexibility and intuitiveness of state space models and suggest a very generic framework for looking at longitudinal and survival data jointly. In our model, a latent senescence process will simultaneously determine mortality and influence the distribution of observed covariates. We show how easily a researcher can manipulate the definition of aging and test his theories about the process of aging, such as aging dynamics, its effect on observable data, form of individual mortality, and population mortality plateaus. We find that modern software is just a step behind being able to handle model fitting and diagnostics, and give our own solution. We propose a concept of approximate sufficiency and develop a way to make especially large datasets manageable and computationally feasible. We then illustrate our developments on a cohort of Mexican fruit flies, followed from birth to death, with their behavior histories recorded. Several behavioral patterns will be considered and their connection to the survival will be explored. We will also consider different specifications of aging. The resulting models can be appreciated by a survival analyst as competing for better fit, or by an aging researcher, who is interested in alternative descriptions of aging. Beyond the aging and survival questions, our analysis will give a comprehensive picture of the patterns in behavior which Mexican fruit flies exhibit during the course of their lives.

We begin the study with a literature review. This will illustrate the key ideas of aging research, and some of the survival analysis, which were briefly outlined above, and will provide motivation and building blocks for our approach. In chapter 3 we present the Mexican fruit fly data set, and discuss several limitations of the techniques already available in the literature. The issues raised will be addressed in the subsequent chapters. We propose a statistical framework for analysis of the data described, and give details on the theoretical foundations and review literature on this methodology in chapter 4. Then in chapter 5 we discuss ways to summarize the data and make it suitable for analysis. We introduce the model with various realizations of dynamic aging and fit these models in chapter 6, and develop several goodness of fit and diagnostic tools in chapter 7. Our conclusions and suggestions of future work follow in chapter 8.

Chapter 2

Models for mortality

2.1 Survival distributions

Survival analysis studies the distribution properties of a random variable τ, time before a particular event occurs. The event is usually death of a population member. Survival analysis typically works with the following characteristics of a probability distribution: survival function,

$$S(t) = P\{\tau > t\},$$

and mortality,

$$m(t) = -\frac{S'(t)}{S(t)},$$

also known as hazard function or failure rate. A variety of functions have been used in different areas of application to model mortality. The simplest is a constant, which corresponds to the exponential distribution of time to event. Due to its memoryless property, it has been widely used in such areas as reliability theory ([69], [103]) and

queueing theory ([26], [99]), but it is hardly suitable for modeling survival in a population. More general distribution families allow for either polynomial or exponential growth of mortality. The latter is known as the Gompertz law, introduced by Benjamin Gompertz in 1825, [12]. This is a two parameter distribution defined by its mortality function:

$$m(t) = Ke^{Ct}, \tag{2.1}$$

where K and C are positive constants. These parameters have a convenient natural interpretation: K stands for the mortality at birth and C determines the rate of aging. The exponential mortality growth principle has received substantial support from the observed mortality data and has been in the focus of many theoretical and applied studies. To mention just a few of them, we refer to Simms [42], Finch et al [16], Johnson [104], Promislow [28], Riggs and Millecchia [52], Witten and Satzer [75], Shao et al [112]. Several theories have been developed to justify the Gompertz law. In one, Strehler and Mildvan [14] consider an organism as a number of subsystems, each of which has a certain maximum ability to restore initial conditions after a stress. The repair capacity is determined by a vitality parameter, which declines linearly with age. The Gompertz exponential mortality is then obtained assuming the Maxwell-Boltzmann distribution for stress magnitudes. Gavrilov and Gavrilova provide a different rationale, deriving the Gompertz law from the results of reliability theory, and promoting this idea in a series of papers: [61], [62]. Questions of statistical estimation have also been addressed. Derivation of the maximum likelihood estimates for the Gompertz parameters can be found in Garg et al [71]. Mueller et al [64] compare several different techniques to estimate the parameters, including linear and nonlinear regression and the maximum likelihood.

With good empirical performance and theoretical justifications, the Gompertz law stands out as a classical tool for the description of survival. However, in the 1990s experiments began to accumulate significant evidence that observed mortality often deviated from the Gompertz model. Hirsh et al [5], analysed data on medfly mortality and concluded that a less steep mortality of Weibull distribution provides a significantly better fit. Carey et al [47] studied a large cohort of Mediterranean flies and pointed out that mortality slowed down at older ages. Similar observations were made later by Vaupel et al [53], who presented mortality curves for several large cohorts. It was shown in their study that human and drosophila mortality curves flattened at late ages, while mortalities of wasps and medflies even dropped after a certain age.

This issue has been addressed by many researchers, bringing ideas from a variety of science domains. For example, the Mueller and Rose [63] hypothesis is based on mutation accumulation[1] and antagonistic pleiotropy[2] theories of senescence. These theories can reproduce both exponential mortality growth for the most part of life as well as the late life mortality plateaus. Another approach relies on the disposable soma theory[3]: dynamic models are built to determine the optimal relationship among reproduction, diversion of resources from repair and senescence mortality, leading to early exponential increases in mortality followed by a slowdown at later ages (Abrams and Ludwig [84]). Many of these theories provide valuable insights into the mechanisms of aging, and the literature on the subject is vast. Our immediate goal is to track the development of statistical models that emerged in the attempts to describe

[1]The force of natural selection is much weaker in late life. Thus mutations which have a detrimental effect in late life are not efficiently eliminated by the natural selection (Medawar [89]).

[2]Selection of genes which increase fitness in early life but are harmful in late life (Williams [38]).

[3]An organism needs to allocate energy between metabolism, reproduction and repair. Insufficient repair causes gradual deterioration of the organism (Kirkwood [105]).

observed mortality. We therefore restrict our attention to just a few of the theories, which are key to our study.

2.2 Heterogeneous populations

The survival models we have been discussing assume that all population members have the same hazard function, or in the parametric context, the same parameters of their survival distribution. However, it was recognized many decades ago that different individuals may vary substantially in their longevity endowment. This may depend on various factors which could be fully controlled, observed or completely unknown to the researcher. We say that a population is heterogeneous if we recognize that certain survival related factors may vary among members of that population. It is usually of interest to model dependence of the survival distribution on such sources of heterogeneity. In this regard we will distinguish between subject-specific and population survival distribution: the former describes the lifespan of a given population member[4], while the latter pertains to a random population member. Subject-specific models have potential to make better prediction of a subject's lifespan, and they are more flexible in describing the survival in population.

Classical survival analysis, which developed in the second half of the 20th century, is generally concerned with estimating effects of fully observed variables on the time to event distribution. Such variables are often called covariates. There are several well established methods to model survival with covariates. Accelerated failure time (AFT) models, for example, are framed as generalized linear models for the logarithm

[4]In fact, the definition applies any group of subjects having identical covariates. We prefer to use the term 'subject-specific' rather than 'covariate-specific' to emphasize conditioning on individual background or historical data.

of the survival time. The effect of a covariate is thus to multiply the expected event time by some constant (Cox and Oakes, [24]). In the simplest form, the distribution of the time to event τ is assumed to follow a log-normal distribution with the mean being a linear function of the covariates Z:

$$\log \tau = Z\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2) \tag{2.2}$$

In the more general formulation, τ follows a parametric distribution with a location parameter, the logarithm of which is a linear function of the covariates. In [23], Cox proposed a proportional hazard (PH) semi-parametric model, in which the effect of covariates is to multiply hazard by some constant. It assumes a nonparametric baseline mortality $m_0(t)$, which corresponds to the mortality of an individual having all covariates equal to zero. The mortality of any population member is then given by

$$m(t) = m_0(t) \exp\{Z\beta\} \tag{2.3}$$

Easily interpretable and rather cheap computationally, these two models have earned tremendous popularity among applied scientists, in particular in the analysis of clinical data. In some situations other approaches may be preferred. For instance, if the period of following the subjects is fixed, one may model the event indicator by the logistic regression (Cox, [22], Woodbury et al [77]).

Unobservable heterogeneity factors have also received considerable attention. Unlike the obvious necessity to use covariates in survival models, this stream of research was perhaps motivated by the need to explain marginal mortality patterns. For instance, Strehler and Mildvan [14] pointed out that variations in vitality may lead to deviations from the Gompertz straight line of mortality rate, when looked at the population level. It was in the 1990s that the heterogeneity concept was drawn upon

to explain mortality plateaus[5], which had been quite well documented by that time (Vaupel et al [53]). Vaupel, Manton and Stallard put this idea on mathematical footing in their paper of 1979, [54]. The authors introduced the term 'frailty' as a measure of one's general susceptibility to all causes of death. It was defined as an age independent mortality multiplier, and assumed to be a random variable with a gamma distribution over the population:

$$m(t|\theta) = \theta m_0(t),\ \theta\ \sim \text{Gamma}(1, \sigma) \tag{2.4}$$

The resulting population mortality $\bar{m}(t)$ was averaged over possible values of frailty. The authors argued that individuals aged faster than the cohort: frail individuals died quicker, gradually reducing their impact on the overall population mortality. They showed that heterogeneity can thus substantially influence the shape of the population mortality curve. In particular, models which take heterogeneity into account are capable of reproducing mortality plateaus. To give an example, we consider the Gompertz model for the subject-specific mortality. In fig. 2.1 we plot five Gompertz curves with the random gamma frailty, and the resulting population mortality. It clearly has a plateau, and its functional form is the logistic function.

This publication inspired much discussion of the heterogeneity explanation. Several empirical studies promoted this idea. Brooks et al [2] suggest heterogeneity as an explanation of the mortality plateaus observed in a population of nematodes. To demonstrate this, the authors first considered genetically identical population and observed a Gompertz mortality. To the contrary, mixing populations with different mean lifespans caused the mortality curve to decelerate. Kowald et al [6] re-analysed the medfly mortality data obtained by Carey et al [47]. They were able to mimic

[5]The term is usually used to speak about slowing down or leveling off mortality at advanced ages.

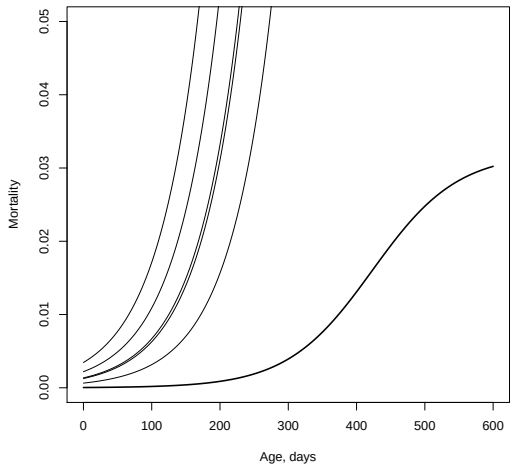

Figure 2.1: Five thin lines: subject-specific Gompertz curves with parameters K = 0.0023, C = 0.016, and a random gamma frailty with mean 1 and variance 0.5. Thick line: the population mortality, obtained from the frailty-averaged survival function.

the declining mortality curve with a mixture of Gompertz functions. However the authors warned that heterogeneity explanation should be taken with caution, because uncertainty in estimating mortality of the oldest individuals is so high that the observed plateaus could occur purely by chance. In application to a large drosophila melanogaster cohort, Drapeau et al [70] rejected the heterogeneity assumption by finding no evidence of it in the data, but later this paper was criticized by Steinsaltz [29], pointing out some flaws in the statistical analysis. In 2002 Mangel [73] illustrated the random frailty model on the Gompertz law, and pointed out that the resulting population mortality decelerated at late ages. Manton et al [57] showed that the gamma-Gompertz models fit US and Swedish mortality better than a simple Gompertz model without heterogeneity. Yashin et al [9] confirmed their result. A comprehensive theoretical treatment of frailty models can be found in Steinsaltz and

Evans [31], who fully describe the behavior of the cohort mortality given the distribution of frailties and the baseline mortality. Today the random frailty model is a methodology widely recognized and used routinely by applied statisticians to model overdispersion of survival times and to account for clustering. It has been realized in major software like STATA and R.

2.3 Randomness of aging

Heterogeneity considered so far, either observed or not, classified individuals into groups with the same risk. Thus, two individuals with the same set of covariates, or two individuals with the same frailty were assumed to have the same mortality throughout their lives. It may reasonably be argued that this assumption can easily be violated, as two subjects with similar background could still be exposed to very different risks and as a result, have discrepancy in mortality curves. In their review paper of 2000 Yashin et al. [8] stated several postulates about aging and mortality, which had received substantial support in literature. One of those statements says "there exists a stochastic component to aging: genetically identical individuals in the same environment can have different patterns of physiological aging and different lifespans". Acknowledging this, several authors considered a random aging process, referred to as 'changing frailty', 'acquired heterogeneity', 'physiological state', and other terminology. Randomness was motivated by the impossibility to account for all factors that influence mortality. A number of stochastic mechanisms have been proposed to model individual time-specific mortality. Yet, in common with the marginal population mortality models, there is no unique theory as to what causes certain mortality dynamics. Yashin et al point this out in their recent paper of 2007, [10]. In

the sequel we illuminate some of the trends in this direction.

One possible way to generalize the fixed frailty approach is to assume that there are n stages of frailty, through which an organism passes until it is killed as it reaches the n-th state. Transition between states occurs according to the law of a Markov chain with a finite state space (see chaper 4 for definition). This approach received considerable attention in the 1970s. In a simpler form, the transition from state i may occur either to the next state $i+1$, or to the absorbing state n (death). The transition rates do not depend on age, but increase monotonically with i. For development of this model, see Le Bras [40] and Gavrilov and Gavrilova [60]. More generally, transient states could have arbitrary transition probabilities and there can be several absorbing states. The time to absorption of a Markov chain model has a distribution known as the 'phase type distribution'. Aalen [82] argues that such models can generate many kinds of failure distributions. In particular, they necessarily have a leveling mortality curve, and thus can be used to model mortality plateaus. Some general results for this approach were derived by Yashin [7]. From a purely mathematical point of view, a finite state space may suffice to closely describe an observed survival distribution, even though determining the number of states is a quite complicated task (see [96] on this matter). It is not however clear why an organism should pass through a finite number of frailty states, which implies abrupt jumps in its mortality. Hougaard [88] gives a review of various survival models based on Markov chains. These are typically used when along with survival data we have fully observed covariates, so that the number of states and the transition matrix structure can be decided based upon the possible values of those covariates. For instance, the states could represent the number of children borne to a woman, presence of a disability or a stage of a

certain disease. When no particular state space is inherent to the problem, a model with continuously changing frailty is a reasonable alternative. We review some of the examples below.

One of the early stochastic aging models is due to Sacher and Trucco [37]. Aging is driven by a stochastic process V_t, satisfying the differential equation

$$dV_t = -aV_t dt + bdW_t,\ V_0 = V$$

Death occurs upon the process V_t reaching a lethal boundary, which represents the individual's vital capacity. This threshold declines linearly with the calendar age, in the flavor of Strehler and Mildvan's approach mentioned earlier. Individual mortality is thus modeled indirectly through the distribution of the stopping time. Surprisingly enough, marginal mortality turns out to be approximately the exponential Gompertz law.

Woodbury and Manton [76] proposed a diffusion model for the physiological state process V_t, which determined an organism's survival probability P_t:

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

$$dP_t(V_t) = m(V_t, t)P_t(V_t)dt$$

Notably, the process V_t is multidimensional, including both observed and unobserved variables relevant to survival. The random component ξ is due to the effects of exogenous variables. The authors consider a specification, where the error terms are normal, the drift u is a linear function, and the hazard m is a quadratic function of the state.

Rossolini and Piantanelli [36] realize the acquired heterogeneity by means of a vitality parameter, which they define as a random process influencing the probability

of survival.

Weitz and Fraiser [55] consider a different diffusion model for aging. In their approach, it is assumed that a random process represents an individual's viability to stay alive, and aging is realized by the presence of a linear negative drift:

$$V_t = -at + bW_t,\ V_0 \sim \text{Gamma}$$

The organism dies as it runs out of its viability. Once again, death is characterized by the first passage time. The authors show that their model leads to the inverse Gaussian distribution for the survival time and has a rich variety of mortality shapes.

In 2004, Steinsaltz and Wachter [30] obtained a very general result for survival time being modeled by a killed Markov process. They proved that there is a common feature of the mortality approaching a plateau as a consequence of converging to a quasistationary distribution[6]. In particular they showed that the plateau level in the model of Weitz and Fraiser is $\frac{1}{2}(a/b)^2$. The authors also consider a generalized Weitz-Fraiser model where they allow the initial viability V_0 to have an arbitrary probability distribution. They provide necessary and sufficient conditions for a particular mortality to be attainable through such construction. It turns out that the family of attainable distributions is rather rich.

2.4 Survival analysis with longitudinal data

As we have seen, introduction of a latent random aging process in one form or another allows one to model almost any survival law (Steinsaltz and Wachter [30]). The difficulty in interpreting the latent process and justifying its particular dynamics

[6]The invariant conditional distribution of a Markov chain, given that the chain has not yet reached the absorbing state (given survival).

becomes an issue. Yashin et al [8] pointed out that survival data alone "are not enough to distinguish between different mechanisms capable of generating observed mortality patterns". In order to make inference about the path of the aging process, some information about it has to be collected as it develops in time. Longitudinal measurements on certain covariates could provide such information if the distribution of these covariates is influenced by aging.

The last two decades have seen a rapid development of the statistical methods for survival analysis with longitudinal data. Interestingly though, the majority of developments are not coming from the aging research community, but mostly from clinical applications. It is generally not feasible to collect time dependent data at more than a few time points per subject. In such situations the random effects models are often used in place of the more technical and cumbersome continuous diffusions. To describe this important class of joint models, we give a brief overview of the random effects approach.

Suppose we obtain measurements Y_{ij} on subject i at time j. A linear random effects model is given by

$$Y_{ij} = X_i\beta + Z_i\beta_i + \varepsilon_{ij}$$

β are fixed effects and β_i are random effects. The latter are usually assumed to have normal distribution with zero mean. The error term ε_{ij} may be allowed to have serial correlation (that is, along the time index j). Because both fixed and random effects are present, such models are also called 'mixed models'. The role of β_i is very similar to that of frailty - to model heterogeneity across the subjects. This class of models is a very powerful tool and is widely used in applications. For development of the theory, we refer to Liard and Ware [79], Rutter and Elashoff [19] and Littell et al [95].

In joint analysis it is common to model the longitudinal component with a mixed model and the survival component with either PH (2.3) or AFT (e.g., (2.2) is a special case with the log-normal survival distribution). The two parts have to be linked together to ensure that it is possible for the longitudinal data to carry information about survival. Among the early examples of joint modeling is the study of De Gruttola and Tu [107], who considered CD4-lymphocyte count as a longitudinal measure of progression of HIV infection. In their model the random effects from the longitudinal model entered the AFT model as covariates.

A different approach considers the PH model for the survival part of data. In this case, the random intercept term can be interpreted as frailty, in the sense of Vaupel, Manton and Stallard's random mortality multiplier. For application of this model, see Faucett and Thomas [15], Wulfsohn and Tsiatis [111], Zeng and Cai [32]. Ratcliffe, Guo and Ten Have [102] considered cluster-specific rather than subject-specific frailties. Hsieh, Tseng and Wang [34] investigated properties of the maximum likelihood estimation for such models, in particular its standard error and robustness properties. Wang [20] compared several estimation strategies for this model.

There are numerous variations on this setup. For the longitudinal component, Chen, Ibrahim and Sinha [68] used a quadratic function to model the mean of the response. Xu and Zeger [56] considered a generalized linear model. Other authors relaxed the normality assumption for the random effects. Song, Davidian and Tsiatis [109] proposed to use a semiparametric realization of the model, where random effects were only assumed to have a smooth density. The density was then approximated by a specified class of functions. Brown, Ibrahim and De Gruttola [33] choose a non-parametric approach and model longitudinal data with subject specific cubic splines.

For the survival component, Chi and Ibrahim [110] considered a multidimensional endpoint, distinguishing between different causes of death. Moustaki and Steele [44] modeled survival probability as a generalized linear model with the logit link function.

Typically the role of the latent process implied by the models considered above is simply to represent the true covariate value, which we get to observe with an error. Many authors point out that ignoring the error will result in biased estimates. Among others, Prentice [97], 1982, addressed this question directly in the Cox regression set up. In particular, he demonstrated that allowing an error in measurements leads to different estimates of the covariate's effect on the relative risk, depending on the level of noise.

From the aging perspective, it would be desirable to give the latent process an interpretation similar to what we saw in the stochastic aging models, where no covariates were available. However, to the best of our knowledge, the literature on joint analysis of longitudinal and mortality data is very scarce outside clinical applications. In 1997 Yashin and Manton [11] published a review paper, where several models for failure data were discussed. They ultimately proposed a hidden diffusion model for aging, which allowed for longitudinal observations. Later this approach was developed by Yashin et al [10], where the structure of the hidden diffusion was systematically chosen based on various theories of aging.

To summarize and conclude the literature review, we point out that statistical modeling strategies are well established and more or less standard in the clinical area. These models usually serve descriptive or predictive purpose. This is why the latent process, which models population heterogeneity in the longitudinal models, is generally treated as a nuisance. In an aging study, on the contrary, this process is of

primary interest, because it may give us insight into the underlying dynamics of aging. It is thus not clear whether either of these models is going to become a standard in longitudinal aging research. In this work we develop a statistical framework which we hope can be of practical use to both biostatisticians and aging researchers.

Chapter 3

Motivation for the study

3.1 The challenge

In the preceding chapter we discussed modeling survival data using longitudinally collected information, and pointed out the benefits of modeling survival and longitudinal data jointly. This new approach has shown that longitudinal data may possess richer information on survival than a cross-sectional measurement. In view of this, statistical models to handle such data have been extensively developing. For brevity, such models will be called 'joint models'.

In spite of the diversity of data sources and models proposed for analysis, joint models have several features that are common to many studies:

1. The data are continuous measurements, or can be treated as continuous (CD4 counts in AIDS studies, for instance).

2. If measurements are taken longitudinally, there are only a few to a moderate number of time points.

3. The hidden variable or process involved in the joint model is generally not of interest to practicing statisticians.

While traditional joint modeling methods in most cases are successful, in analysis of clinical trials data in particular, we claim that certain kinds of data require a fresh look at the problem. We intend to show that the joint modeling can be adapted to handle vast amounts of longitudinal data of any type. Furthermore, it can incorporate complex aging dynamics, and thus turn into a powerful aging research tool. At the same time, joint modeling can remain a conventional routine in the applied statistician's toolkit.

It is now appropriate to introduce the dataset, which will be central to supporting the above propositions. All developments of this work will be presented in application to these data, which will be the source of our examples.

In this research we analyze life long observations of Mexican fruit flies. The data were obtained through a new technology developed by a team of researchers working at UC Davis in the US and the Instituto Nacional de Astrofisica Optica y Electronica in Mexico. Full description of the technology is in [100].

A machine vision tool able to recognize several behaviors of a fly was used. The behaviors included resting, moving, walking, flying, drinking and eating. Every 200 milliseconds a fly was photographed, the image processed, the behavior identified and the results stored in a database. The experiment was set up in such a way that the system could track 9 flies, each of which was kept in an isolated cage. Only one out of the 9 flies could be monitored at a time. The observation period lasted a fixed amount of time (10 or 60 seconds), then passing on to the next fly. This observation process was occasionally interrupted by the cage maintenance. Half of the day the

system used white light and the other half it used near infrared light at 800nm to avoid disturbing the flies while taking night measurements.

With three systems run simultaneously, we have at our disposal behavior histories of 27 fruit flies, followed from birth to death. As we have claimed, several features of the data make it unique, and challenge modern techniques for joint modeling. First, contrary to what we stated to be common, behavior data is of unordered categorical type. Clearly, this requires us to think beyond the convenience of the normal distribution for the measurements. Second, in spite of significant 'blind time' due to monitoring other flies or overall interruption of the observation process, the longitudinal data are very frequent, and the amount of output information is huge. To give the reader an idea, the longest lived fly died at the age of 202 days. During this period, 9,325,450 images of the fly were taken, analyzed, and the determined behavior recorded. From these 103,181 behavior transitions were observed.

Many of the ideas of joint modeling appear inapplicable to such data. Indeed, a fly's behavior is a bad survival covariate: one may hardly believe that, say, switching from resting to walking causes an immediate change in the fly's mortality, which is reversed back if the fly decides to rest again. It is awkward to assume that current behavior is a marker for the fly's aging state. Also fitting a joint model normally requires substantial computational resourses for handling the latent variables or latent process, which links observations and hazard. With just a few time points, the amount of longitudinal data has never been reported an issue, but with frequent observations the demand for computational resourses will quickly outgrow available capacity.

3.2 Outline of our approach

It is implausible that behaviors themselves carry information about aging: there is too much randomness in the behavior process in the short term to consider it a mortality marker. On the other hand, evidence is being accumulated that long term behavior patterns may be predictive of survival. Kowald et al [6] pointed out that age-related behavior changes may affect mortality risk. "If older flies lead quieter lives, (for example, crawling rather than flying), mortality may level off or even decline in spite of the fact that the animal gets progressively more frail". Papadopoulos, Carey et al [80] made an astonishing observation that onset of supine behavior[1] is predictive of remaining lifespan.

Such results give a clue that long term distributional changes in the behavior process may be a senescence marker, as opposed to short term patterns or single observations. We propose to track these long term changes by means of constructing appropriate summary statistics. These summaries are to be calculated over relatively short periods of time, such as a day or an hour, where aging can be reasonably assumed to stay constant. Detailed explanation of the summarizing process, motivation for particular choices of the summaries, and fruit fly data examples are presented in chapter 5. The result of this data reduction procedure is a discrete sequence of (possibly multivariate) data, which we then treat as if they were the measurements taken on the subjects.

The first stage of the analysis consists of choosing the summaries and the summarizing time window, calculating these summaries, and determining their distributions.

[1]The fly begins to lie on its back immobile for extended periods, and then resumes normal movement

This work is necessary for the fruit fly behaviors, as it transforms the data to quantities that may be meaningfully associated with hazard. Furthermore, data reduction makes the enourmous amount of behavior histories computationally manageable. In principle, this first stage can be skipped if the raw data are not as sizable and can be naturally interpreted as a senescence biomarker.

The second stage of the analysis is actually the joint model for the survival times and the discrete histories of the summary statistics. This can be accomplished with any available technique, but we advocate the use of hidden Markov models for this purpose. With such models aging can be realized as a latent (or hidden, unobserved) Markov process, and the distribution of the observed statistics depends on the state of this process. The hidden Markov framework has great flexibility in deciding on the dynamics of the aging process and on the distribution of the summary statistics. In particular, it includes all examples of aging models we discussed in chapter 2. In chapter 4 we give a brief overview of the hidden Markov theory and specialize it to the case of state space models, where model building and inference are especially transparent and simple.

Chapter 4

Hidden Markov models

In chapter 2 we discussed various ways to model survival and covariate data jointly, the approach often called joint modeling. At the core of a joint model is an unobservable entity, which has an impact on the distribution of both observations and survivorship. This entity may be conceptualized as frailty, health state, senescence and so forth. Mathemetically, such models can be built in a very general framework, which gives rise to models of very diverse structure and data distribution. This class of models is the hidden Markov models.

Hidden Markov models are used to model partially observed data. They have been extensively developing during the last few decades, and found application in many areas. They provide sufficient generality and complexity, and at the same time feasible inference techniques. Areas where hidden Markov have found application include, among others, speech recognition (Rabiner and Juang [65], Jelinek [35]), econometrics (Hamilton and Raj [51], Kim and Nelson [17]), computational biology (Durbin et al [92], Koski [106]), computer vision (Bunke and Caelli [39]), finance (Shephard [81]). We begin with general definitions and then discuss main inference

principles in hidden Markov structures. Our exposition of the theory relies on the books 'Diffusions, Markov processes and martingales', [66] and 'Inference in hidden Markov models', [83].

4.1 Definitions and facts

Markov processes As the term 'hidden Markov' suggests, the model involves a process which is not observed, or hidden. Central to the theory of hidden Markov models is the Markov property of the this hidden process, for which we reserve notation V_t. In other words, V_t is a Markov process. In this section we give main definitions and facts about Markov processes. Although greater generality is possible, we will assume that V_t develops in time, and t may be either a continuous or a discrete time index.

Definition 1. *Let V_t be a random process with values in a measurable space $(X, \mathcal{X})$. Let $V_{\leq t}$ be the sigma-algebra generated by the events $\{V_{s_1} \in A_1, ..., V_{s_n} \in A_n\}$, where $s_i \leq t$, and A_i are measurable sets. V_t is called a Markov process, if it satisfies the Markov property:*

$$P\{V_{t+s} \in A | V_{\leq t}, V_t = x\} = P\{V_{t+s} \in A | V_t = x\} \tag{4.1}$$

where $t \geq 0$, $s > 0$ and A is measurable. X is called the state space, and if $V_t = x$ we say that V is in state x at time t. The right hand side of (4.1) is called the transition kernel of the process, often written as $P_{t,t+s}(x, A)$. The kernel represents the conditional probability to find V in the set A after time s has passed from the moment of time t, when it was known to occupy state x. Thus, the transition kernel is a probability measure for any given values of s, t and x. If $P_{t,t+s}$ does not depend

on t, we say that the process is time-homogeneous. If t is discrete, the process is often called a 'Markov chain'.

Any Markov process (equivalently, its transition kernel) satisfies the Chapman-Kolmogorov equation, which expresses the transition probability using the intermediate transition:

$$P_{t,s}(x, A) = \int P_{t,u}(x, dy) P_{u,s}(y, A) \tag{4.2}$$

for any u, $t < u < s$.

Finite state space Of particular interest to us will be the case when X is a finite set, $X = \{1, ..., K\}$. Transition kernel of such Markov process can be conveniently represented by the family of transition matrices $P_{t,s}$, with the ij-th entry being

$$(P_{t,s})_{ij} = P_{t,s}(i, \{j\}), \tag{4.3}$$

where $\{j\}$ is a set consisting of one element: j. The Chapman-Kolmogorov equations translate into $P_{t,s} = P_{t,u} P_{u,s}$.

If t is continuous time, the transition kernel has a generator representation. The generator matrix is defined as

$$Q(t) = \frac{d}{ds} P_{t,t+s} \Big|_{s=0} = \lim_{s \to 0} \frac{P_{t,t+s} - I}{s} \tag{4.4}$$

Here $I = P_{t,t}$ is the identity matrix. The matrix $Q(t)$ has non-positive entries on the diagonal, and non-negative entries off-diagonal, with

$$Q_{ii} = -\sum_{k \neq i} Q_{ik} \tag{4.5}$$

For homogeneous processes, the transition matrix $P_{t,t+s}$ does not depend on t, and we write $P_{t,t+s} = P_s$. The family of the transition matrices thus forms a semigroup[1], with

[1] A semigroup assumes that an associative binary operation is defined on its elements. A semigroup is a group if in addition there exists an identity element and each element has its inverse.

the matrix multiplication operation, $P_s P_t = P_{s+t}$. The transition semigroup admits a generator, which is a constant matrix: $Q(t) = Q$. The Kolmogorov equations $P_t' = P_t Q = Q P_t$ are satisfied, and the transition matrices can be recovered through the matrix exponential: $P_t = \exp\{tQ\}$.

A Markov process V_t defined on a finite state space is a step process: it occupies some state X_1 for a random period of time D_1, then instantly jumps into another state X_2, where it spends time D_2, and so forth. Thus, V_t can be represented by the sequence of pairs (X_n, D_n), where X_n is a sequence of states the chain has visited, and D_n is the time spent in X_n. Homogeneous Markov processes have especially simple distribution of the pairs (X_n, D_n), see e.g. Rogers and Williams [66]:

- The sequence $X_1, X_2, ...$ is a homogeneous discrete time Markov chain with the transition matrix P,
$$P_{ij} = \begin{cases} -Q_{ij}/Q_{ii}, & i \neq j \\ 0, & i = j \end{cases} \tag{4.6}$$
X_i is called the embedded chain.

- $(D_1, D_2, ...)$ are independent random variables, given $(X_1, X_2, ...)$

- The conditional distribution of D_n, given that $X_n = i$, is exponential with rate $Q_i = -Q_{ii}$:
$$P\{D_n > t | X_n = i\} = e^{-Q_i t} \tag{4.7}$$

The property above allows a convenient jump-and-hold construction, which makes it straightforward to simulate the process paths:

Algorithm 1 (Simulating a Markov process)**.**

1. *Draw initial state X_0, let $t_0 = 0$*

2. *Having obtained the current state $X_n = i$, iterate:*

   (a) *Draw a holding time D_n from the exponential distribution with rate $-Q_{ii}$. Let $t_{n+1} = t_n + D_n$ and define $V_s = i$ for $s \in [t_n, t_{n+1})$.*

   (b) *Draw the next state $X_{n+1} = j$, $j \neq i$, with probability $-Q_{ij}/Q_{ii}$*

Hidden Markov models Any hidden Markov model has two components: the unobservable Markov process V_t, and the obsevations (call them Y_t). The hidden part V_t can either be the process studied but unavailable for direct observation, or it can be just an auxiliary tool to model the distribution of observations. In this research in particular, the hidden process is going to represent aging, which cannot be directly evaluated, and is only observable through mortality and perhaps the longitudinal measurements. The observable component of the model is a vector valued process Y_t. Through observing it, we can make inference about the hidden part V_t.

The hidden Markov theory is generally developed for discrete time observations, and this is the kind of hidden Markov models that will be used in the present study. We therefore assume that $t = 1, 2, ..., n$. Next we give a rigorous definition of the hidden Markov model.

Definition 2. *The process (V_t, Y_t), $t = 1, 2, ...n$ is called a hidden Markov model, if the following three conditions hold:*

- *V_t is a Markov chain*

- *For any time indices* $t_1 < ... < t_k \leq n$ *the observations* $Y_{t_1}, ..., Y_{t_k}$ *are conditionally independent, given realization of the state:*

$$E\left(\prod_{i=1}^{k} f_{t_i}(Y_{t_i})|V_{t_1}, ..., V_{t_k}\right) = \prod_{i=1}^{k} E\left(f_{t_i}(Y_{t_i})|V_{t_i}\right)$$

where f_{t_i} *are any bounded continuous functions.*

- Y_t *is observable and* V_t *is not observable*

The process V_t *is called the state of the hidden Markov model.*

Besides conditional independence, the second condition states that the conditional distribution of the observation Y_t is determined by V_t alone, regardless of the previous or subsequent states. It can be seen that the joint process (V_t, Y_t) is Markov with a special kind of transition kernel:

$$P_{t,t+s}((v, y), A) = \int_A P_{t,t+s}(v, dv')P_{t+s}(dy'|v') \tag{4.8}$$

where under the integral sign are the kernel of the process V_t and the conditional distribution of Y_{t+s} given V_{t+s}. Notice that the right hand side of (4.8) does not depend on y. Generalizations of (4.8) are possible, for instance dependence on y may be allowed, so that Y_t is a Markov chain when conditioned on $V_{\leq t}$. In discrete state spaces, such models are sometimes called 'Markov switching'.

What makes hidden Markov models an attractive tool to handle partially observed data is relative simplicity of inference. The main difficulty of making inferences about partially observed data is that the unconditional distribution of the observed component is generally not available to us in a closed form. This, in particular, complicates the use of the maximum likelihood estimation. What distinguishes hidden Markov models as models for partially observed data, is the possibility to compute

the likelihood sequentially. The algorithm is based on the procedure known as the forward recursion, originally developed by Baum et al [58], which we outline below.

Before we proceed, some notation is needed. Let $y_1, ..., y_n$ be the observed values of $Y_1, ..., Y_n$ and let $v_1, ..., v_n$ be the realization of the hidden process $V_1, ..., V_n$. Assume that the kernel of the hidden Markov process has a density with respect to some measure (normally, either the Lebesgue or the counting measure):

$$f_t(v_{t-1}, v_t) = P(V_t \in dv_t | V_{t-1} = v_{t-1})/dv_t$$

We also put $f_1(v_1) = P(V_1 \in dv_1)/dv_1$, the density of the initial distribution of the state. Likewise, let the observations have conditional densities

$$g_t(y_t|v_t) = P(Y_t \in dy_t | V_t = v_t)/dy_t$$

with respect to a (possibly another) measure.

Next we introduce the smoothing distributions. We will use the subscript of the form $t_1, ..., t_k | s_1, ..., s_m$ to denote the density of $(V_{t_1}, ..., V_{t_k})$ given $(y_{s_1}, ..., y_{s_m})$:

$$\mathcal{F}_{t_1,...,t_k|s_1,...,s_m}(v_{t_1}, ..., v_{t_k}) =$$

$$P(V_{t_1} \in dv_{t_1}, ..., V_{t_k} \in dv_{t_k} | Y_{s_1} = y_{s_1}, ..., Y_{s_m} = y_{s_m})/(dv_{t_1} \cdot ... \cdot dv_{t_k}) \tag{4.9}$$

These distributions are called smoothing distributions. The important special cases are the distributions of the single current state V_t, given the history of observations. If the history is also taken up to time t, the distribution is called filtering, and we denote it by $\mathcal{F}_{t|t} = \mathcal{F}_{1,...,t|1,...,t}$. If the entire history is taken, we put $\mathcal{F}_{t|n} = \mathcal{F}_{1,...,t|1,...,n}$.

Finally, define the conditional densities of the observations:

$$L_t = P(Y_t \in dy_t | Y_1 = y_1, ..., Y_{t-1} = y_{t-1})/dy_t \tag{4.10}$$

Clearly, the likelihood of interest is then the product of the conditional likelihoods

L_t:

$$Lik(y_1, ..., y_n) = \prod_{t=1}^{n} L_t \tag{4.11}$$

The next two equations follow from the definitions of L_t, $\mathcal{F}_{t|t}$ and the hidden Markov structure [83, p.63]:

$$L_t = \int \int f_t(v_{t-1}, v_t) g_t(y_t|v_t) \mathcal{F}_{t-1|t-1}(v_{t-1}) dv_{t-1} dv_t \tag{4.12}$$

and

$$\mathcal{F}_{t|t}(v_t) = L_t^{-1} \int f_t(v_{t-1}, v_t) g_t(y_t|v_t) \mathcal{F}_{t-1|t-1}(v_{t-1}) dv_{t-1} \tag{4.13}$$

(4.12) and (4.13) define the so called forward recursion with which one can evaluate the likelihood function, proceding forward in time to obtain the filters $\mathcal{F}_{t|t}$ and the conditional likelihoods L_t. This algorithm permits a quick update of the likelihood (4.11) if new observations arrive.

4.2 Overview of estimation techniques

The problem of estimation presumes that the observations are sampled from a distribution, which is a member of a parametric family. We will denote the parameter of the family by θ, and specify our hidden Markov model through the densities of the hidden process kernel f_t^θ and the conditional densities of the observations g_t^θ.

There are three main approaches to estimating θ in hidden Markov models: direct likelihood maximization, expectation-maximization algorithm, and the Bayesian method. The first two aim to maximize the likelihood (4.11), and the third one assumes a prior distribution on θ and computes the posterior mean. In application to the fruit fly data analysis, we have explored all three approaches. Since the direct

likelihood maximization method was eventually chosen, our exposition of the two other methods will be brief. The direct likelihood method will be given additional attention below in the state space model discussion, and further in chapter 6 in the joint model context.

The Expectation Maximization (EM) algorithm Theoretically, any hidden Markov model allows evaluation of the observations' likelihood through the forward recursion (4.12), (4.13) and (4.11). In practice, however, the quantities L_t as a rule are either tiny or huge, prohibiting their efficient calculation and storage. The usual way around this is to compute $\log L_t$ and optimize the log-likelihood. As a rule, we can take advantage of the logarithm transform if it can be evaluated analytically. This is not the case with the hidden Markov models, as can be seen from (4.12). For this reason, computation of the log-likelihood through (4.12)-(4.13) may be problematic.

The alternative method, developed by Dempster et al [4] in 1977, seeks to optimize a surrogate function $Q(\theta;\theta')$, which is defined under mild regularity conditions on the distribution family. These conditions are stated in [83, p.350]. The quantity $Q(\theta;\theta')$ is the conditional expectation of the joint log-likelihood of the state V and observations Y, taken with respect to another member of the distribution family with parameter θ'. In the case of a hidden Markov model Q reduces to

$$Q(\theta;\theta') = \sum_{t=1}^{n} E_{\theta'}[\log f_t^{\theta}(V_{t-1}, V_t)|y_1, ..., y_n] + \sum_{t=1}^{n} E_{\theta'}[\log g_t^{\theta}(y_t|V_t)|y_1, ..., y_n] \quad (4.14)$$

Here the expectation $E_{\theta'}$ involves integration with respect to the smoothing distributions $\mathcal{F}^{\theta'}_{t-1,t|n}$ in the first sum and $\mathcal{F}^{\theta'}_{t|n}$ in the second sum. The algorithm proceeds as follows.

Algorithm 2 (Expectation-maximization)**.**

1. *Pick initial parameter* θ_0

2. *Repeat until convergence achieved:*

   (a) *Expectation step: use (4.14) to compute the quantity* $Q(\theta; \theta_i)$

   (b) *Maximization step: let* θ_{i+1} *be the value that maximizes* $Q(\theta; \theta_i)$ *with respect to the first argument*

3. *Take the last computed* θ_i *to be the estimate of* θ.

It turns out that each next candidate θ_i does not decrease the likelihood. Moreover, if convergence of $Lik(\theta_i)$ takes place, then θ_i converge to a stationary point of the likelihood. For conditions that guarantee global convergence of the EM algorithm, see [83, p.389]. However, there is no way to ensure that the stationary point to which θ_i converge is the global maximum of the likelihood. Therefore, θ_0 needs to be carefully chosen.

We remark that the EM algorithm is numerically stable, because it works with the logarithms of the densities, and also because it is an ascent procedure (the value of the likelihood increases at each step). EM is attractive when both steps can be performed analytically. This is not the case in the analysis of the fruit fly data, as we will see later. In particular, the filtering and smoothing densities involved are not available in a closed form. In principle, one can perform the expectation step through Monte Carlo integration; this approach is called the Monte Carlo EM algorithm. However, with this technique one can perform direct likelihood evaluation and optimization, which is generally superior in performance.

Bayesian approach Another approach to estimation of parameters completely bypasses evaluation of the likelihood and its optimization. It is built entirely on simulations, and its validity is justified by the general principles of the Markov chain Monte Carlo method (MCMC), [83, ch. 6]. As in any Bayesian analysis, MCMC algorithms for hidden Markov models treat the parameter θ as random, with distribution density $\pi(\theta)$. This is called the prior distribution. Inference is done with the posterior distribution of θ, that is, the conditional distribution given observations. As a rule, θ is estimated as the posterior mean.

In the context of a hidden Markov model, the posterior distribution $\pi(\theta|y_1..., y_n)$ is generally not available. It turns out however that it is possible to obtain the joint posterior distribution of θ and $V_1, ..., V_n$, where we treat both θ and the hidden process path as unknown random parameters, with the prior given by

$$\pi(\theta; v_1, ..., v_n) = \pi(\theta) \prod_{t=1}^{n} f_t^{\theta}(v_{t-1}, v_t)$$

To simulate from this distribution, a variation of the MCMC method is employed, called the Gibbs sampler, which works as follows. First, the initial parameter guess $(\theta^0; V_1^0, ..., V_n^0)$ is drawn from the prior density. At each subsequent iteration the Gibbs sampler updates one or several of the coordinates of the parameter vector, by drawing it from the conditional distribution, given observations. All coordinates are thus updated one after another, and we obtain the next parameter draw $(\theta^i; V_1^i, ..., V_n^i)$. The resulting sequence forms a Markov chain, whose stationary distribution is the joint posterior required. Once we obtain the sequence and it is long enough to consider it stationary, it remains to compute the average of θ to complete the estimation:

$$\hat{\theta} = \frac{1}{M} \sum_{i=1}^{M} \theta^i$$

To illustrate, we present a variant of the Gibbs sampler for a hidden Markov model.

Algorithm 3 (Gibbs sampling for HMM)**.**

1. *Draw* θ^0 *from* $\pi(\theta)$ *and the unconditional hidden path* $V_1^0, ..., V_n^0$*, given* θ^0*.*

2. *Iterate for* $i = 1, ..., M$*:*

   (a) *Simulate* θ^i *from* $\pi(\theta|v_1^{i-1}, ..., v_n^{i-1}; y_1, ..., y_n)$

   (b) *For* $t = 1, ..., n$ *simulate* V_t^i *from* $\pi(v_t|\theta^i; v_1^i, ..., v_{t-1}^i, v_{t+1}^{i-1}, ..., v_n^{i-1}; y_1, ..., y_n)$

3. *Discard the first several iterations*

Because of the hidden Markov model structure, the density in step 2 reduces to

$$\pi(v_t|\theta^i; v_{t-1}^i, v_{t+1}^{i-1}; y_t) \propto f_t^{\theta^i}(v_{t-1}^i, v_t) f_{t+1}^{\theta^i}(v_t, v_{t+1}^{i-1}) g_t^{\theta^i}(y_t|v_t) \tag{4.15}$$

Step 3 is necessary to allow the Markov chain to enter its stationary mode. This is usually called the 'burn-in' period. There seems to be no rule of thumb regarding the number of iterations to be discarded.

Despite simplicity of estimation, the Bayesian approach suffers from a number of drawbacks. First of all, the densities (4.15) generally do not lend themselves to straightforward simulation. Algorithms such as the accept-reject procedure may be required [83, p. 166]. This takes additional computational efforts, but more importantly, there is no generic recipe as to how the simulations can be implemented. Thus, with the Bayesian approach, the simulation strategy needs to be taylored to the specific structure of the hidden Markov model. Lastly, step one of the Gibbs sampler depends on the prior distribution $\pi(\theta)$. This is usually chosen to be conjugate[2] to the joint density of the hidden state and the observations. Such a choice is often criticized

[2]A prior distribution on the set of parameters is said to be conjugate to the distribution of observations, if the posterior distribution belongs to the same distribution family as the prior.

for its subjectivity. Moreover, conjugate priors are not universally available, in which case the Bayesian approach looses its usefulness.

4.3 State space models

Any hidden Markov model $(V_t, Y_t),\ t = 1, 2, ..., n$ can be equivalently defined through a functional representation, known as the state space form:

$$\begin{cases} V_{t+1} &= a_t(V_t, \eta_t) \\ Y_t &= b_t(V_t, \varepsilon_t) \end{cases}$$

where the sequences η_t and ε_t are jointly independent, and are usually called 'disturbances'. The first equation specifies the dynamics or evolution of the state (also often called 'the signal'), whereas the second is the observation equation. We will consider a special case of this representation, where the functions a_t and b_t are linear, and the disturbances are jointly normal. Without loss of generality, we assume that the processes V_t and Y_t have zero mean:

$$\begin{cases} V_{t+1} &= A_t V_t + \eta_t,\ \eta_t \sim \mathcal{N}(0, R_t) \\ Y_t &= B_t V_t + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \Sigma_t) \end{cases} \tag{4.16}$$

This class of hidden Markov models is called a 'linear Gaussian state space model'. It appeared in the 1960s and has seen wide application in the fields of engineering and econometrics. Inference in this framework is particularly simple and elegant. Specifically, the forward recursions (4.12)-(4.13) have explicit solutions. These are delivered by the Kalman filter algorithm. The problem of filtering and smoothing is discussed below in detail. We largely rely on the results presented in Durbin and Koopman [49].

Kalman filtering and smoothing The fact that the state and observations jointly form a normal random vector ensures that all smoothing distributions $\mathcal{F}_{t_1,...,t_k|s_1,...,s_m}(v_{t_1},...,v_{t_k})$ are also normal. The required distributions are therefore fully specified by the respective means $m_{t_1,...,t_k|s_1,...,s_m}$ and variances $R_{t_1,...,t_k|s_1,...,s_m}$. The Kalman filter iteratively recovers $m_{t|t-1}$ and $R_{t|t-1}$, while the Kalman smoother finds $m_{t|n}$ and $R_{t|n}$.

The Kalman equations obtain $(m_{t+1|t}, R_{t+1|t})$ given $(m_{t|t-1}, R_{t|t-1})$ plus a new observation y_t:

Algorithm 4 (Kalman filter)**.**

1. *Specify the initial mean* $m_{1|0}$ *and variance* $R_{1|0}$.

2. *Proceeding forward in time* $t = 1, ..., n$, *compute:*

   (a) *Innovation:* $\nu_t = y_t - B_t m_{t|t-1}$

   (b) *Innovation covariance:* $\Gamma_t = B_t R_{t|t-1} B_t' + \Sigma_t$

   (c) *Kalman gain:* $K_t = A_t R_{t|t-1} B_t' \Gamma_t^{-1}$

   (d) *Prediction mean:* $m_{t+1|t} = A_t m_{t|t-1} + K_t \nu_t$

   (e) *Prediction variance:* $R_{t+1|t} = A_t R_{t|t-1} (A_t - K_t B_t)' + R_t$

In the linear space of normal random variables, $B_t m_{t|t-1}$ is the orthogonal projection of y_t on the linear subspace of the previous observations $y_1, ..., y_{t-1}$. The vector of innovations ν_t is then orthogonal to, or uncorrelated with the observation history, hence the term. The process of computing innovations is thus the Gram-Schmidt orthogonalization, and so the observations can be equivalently represented by the

orthogonal innovations $\nu_1, ..., \nu_n$. For orthogonal observations the likelihood is computed simply as the product of the likelihoods of individual innovations ν_t. The latter has normal distribution with mean zero and variance Γ_t. Thus, the Kalman filter can be used to obtain the likelihood (4.10):

$$-2 \log L_t = d_t \log(2\pi) + \log |\Gamma_t| + \nu_t' \Gamma_t^{-1} \nu_t \tag{4.17}$$

where d_t is the dimension of y_t and $|\Gamma_t|$ denotes the determinant of the matrix Γ_t.

There are various ways to initialize the Kalman filter, that is to specify $(m_{1|0}, R_{1|0})$. For discussion of this issue, see [49, Ch. 5]. Depending on the hidden process, either a degenerate distribution or the stationary distribution of the chain will be used in this study.

The following recursion performs smoothing of the states, that is it obtains the mean $m_{t|n}$ and the variance $R_{t|n}$ of the conditional distributions.

Algorithm 5 (Kalman smoother)**.**

1. *Run the Kalman filter to obtain* $m_{t|t-1}$ *and* $R_{t|t-1}$ *for* $t = 1, ..., n$.
2. *Proceeding backwards from* $t = n$ *down to* $t = 1$*, compute:*
   (a) $r_{t-1} = B_t \Gamma_t^{-1} \nu_t + (A_t - K_t B_t)' r_t$
   (b) $N_{t-1} = B_t \Gamma_t^{-1} B_t' + (A_t - K_t B_t)' N_t (A_t - K_t B_t)$
   (c) *Smoothed mean:* $m_{t|n} = m_{t|t-1} + R_{t|t-1} r_{t-1}$
   (d) *Smoothed variance:* $R_{t|n} = R_{t|t-1} - R_{t|t-1} N_{t-1} R_{t|t-1}$

Non-linear and non-Gaussian cases The Kalman recursions we have described obtain the posterior means and variances for the hidden process. Because of normality, these means are also the modes. It turns out that if the distribution of the

disturbances is non-normal, or if the functions a and b are non-linear, the Kalman algorithms are still capable of recovering the modes of the filtering or smoothing densities. This is achieved by applying the Kalman filter iteratively to the best linear normal approximation model. To be more precise, we seek to replace the original model by a linear normal one, such that the filtered state under this approximate model coincides with the conditional mode of the state density under the original model. In this study we call this procedure 'the iterated Kalman filter'.

To meet all our needs while keeping sufficient generality of exposition, we consider the case of non-normal observations Y_t, the conditional distribution of which belongs to the exponential family of distributions. These are the distributions which densities admit the following representation:

$$P(Y_t \in dy_t|\varphi_t)/dy_t = \exp\{y_t'\varphi_t - h_t(\varphi_t) + c_t(y_t)\} \tag{4.18}$$

where h is twice differentiable and c does not depend on φ_t. φ_t is called 'the natural parameter', and we assume that it is a smooth function of the state, so that $\varphi_t = \varphi_t(B_tV_t)$. In this context $\varphi_t(\cdot)$ is called 'the link function'.

The iterated Kalman filter repeats these two steps: finding a linear normal approximation model and running the Kalman filter to obtain the next proposal for the mode $\tilde{V}_t$:

Algorithm 6 (Iterated Kalman Filter)**.** *Run the Kalman filter to obtain initial proposal modes $\tilde{V}_t^0$ for $t = 1, ..., n$. Repeat until convergence achieved:*

1. *Let the proposal modes $\tilde{V}_t^i$ be obtained for $t = 1, ..., n$. Find the approximating model by computing:*

   (a) *Link: $\tilde{\varphi}_t^i = \varphi_t(\tilde{V}_t^i)$*

(b) Signal matrix: $\tilde{B}_t^i = \nabla\varphi_t(\tilde{V}_t^i)$

(c) Variance: $\tilde{\Sigma}_t^i = (\nabla^2 h_t(\tilde{\varphi}_t^i))^{-1}$

(d) Observation: $\tilde{Y}_t = \tilde{\Sigma}_t^i\,(Y_t - \nabla h_t(\tilde{\varphi}_t^i)) + \tilde{B}_t^i\tilde{V}_t^i$

2. *To obtain the next proposal* $\tilde{V}_t^{i+1}$ *run the Kalman filter on the following linear normal state space model:*

$$\begin{cases} V_{t+1} &= A_t V_t + \eta_t,\ \eta_t \sim \mathcal{N}(0, R_t) \\ \tilde{Y}_t^i &= \tilde{B}_t^i V_t + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \tilde{\Sigma}_t^i) \end{cases} \tag{4.19}$$

Justification for the algorithm presented can be found in Durbin and Koopman [49, pp.191-200]. Convergence of $\tilde{V}_t^i$ is usually attained in just a few iterations.

We remark that when only part of the observation vector is non-linear or non-normal, approximation needs not be done on the normal components. A good initial proposal $\tilde{V}_t^1$ in this case can be obtained by discarding the non-normal dimensions from the observation vector. Later in chapter 7 we will specialize (4.19) to the case when Y_t is composed of a normal vector and a binary variable. This will be the general state estimation tool for state space models with survival information.

Importance sampling We have seen that the iterated Kalman filter can be used to find the conditional mode in the case of non-linear or non-Gaussian observations. This is achieved by constructing an approximating linear Gaussian state space model, such that its filtered mean $m_{t+1|t}$ coincides with the mode of the filtering distribution $\mathcal{F}_{t+1|t}$. However, the true filtering distribution $\mathcal{F}_{t+1|t}$ is not the same as the normal distribution $\mathcal{N}(m_{t+1|t}, R_{t+1|t})$. Similarly, the other smoothing distributions differ from the respective normal distributions of the approximating model. In this paragraph

we treat the question of estimating the functional

$$\ell(f) = E_Y(f(V)) \tag{4.20}$$

where Y is the collection of available observations, and V is the collection of state vectors. f here is an arbitrary function of the state such that $Ef(V)$ exists. Straightforward Monte Carlo evaluation of (4.20) is usually not possible, because the smoothing distribution $P(V|Y)$ is not available is a closed form, and simulation from it is not feasible. On the other hand, simulation is possible from the corresponding normal distribution, obtained through the Kalman smoother.

At this point we prefer a more general treatment, and assume that simulation is possible for some distribution $Q(V|Y)$. Computation of (4.20) can then be based on the following identity:

$$E_Y(f(V)) = \int f(V)\frac{P(V|Y)}{Q(V|Y)}Q(dV|Y) \tag{4.21}$$

(4.21) admits Monte Carlo evaluation through independent random draws from the distribution $Q(V|Y)$. This distribution is called the 'importance distribution', and the procedure is called the 'importance sampling'.

Besides the availability of random draws from the importance distribution, computation of (4.20) can be enhanced through manipulations with the Radon-Nikodym derivative $\frac{P(V|Y)}{Q(V|Y)}$. It is typical with state space models that the joint distribution $P(V,Y)$ is simple, while the conditional distribution $P(V|Y)$ is intractable. Therefore, it is beneficial to compute (4.20) as

$$E_Y(f(V)) = \frac{Q(Y)}{P(Y)}\int f(V)\frac{P(V,Y)}{Q(V,Y)}Q(dV|Y) = \frac{Q(Y)}{P(Y)}\int f(V)w(V)Q(dV|Y)$$

where we put $w(V) = \frac{P(V,Y)}{Q(V,Y)}$. This quantity is called the importance weight. Letting

$f(V) = 1$, we find that

$$\frac{Q(Y)}{P(Y)} = \left(\int w(V) Q(dV|Y) \right)^{-1}$$

We have arrived at the following formula for the functional (4.20):

$$\ell(f) = \frac{\int f(V) w(V) Q(dV|Y)}{\int w(V) Q(dV|Y)} \tag{4.22}$$

The additional advantage of (4.22) is that the importance weights $w(V)$ are normalized, and therefore the formula remains valid in the case when the joint densities $P(V, Y)$ and $Q(V, Y)$ are only known up to a constant. The importance sampling algorithm is the Monte Carlo approximation of (4.22):

Algorithm 7 (Importance sampling)**.**

1. *Repeat for* $i = 1, ..., M$*:*

   (a) *Draw* V^i *from* $Q(V, Y)$*.*

   (b) *Compute the importance weight* $w(V^i) = \frac{P(V^i, Y)}{Q(V^i, Y)}$ *and evaluate* $f(V^i)$*.*

2. *Approximate* $\ell(f)$ *with*

$$\hat{\ell}(f) = \frac{\sum_{i=1}^{M} f(V^i) w(V^i)}{\sum_{i=1}^{M} w(V^i)}$$

Description of the importance sampling, including derivation of (4.22) and examples, can be found in Durbin and Koopman [49, p. 190] and Cappe et al [83, p. 210]. In chapter 7 the algorithm will be specialized to treat Bernoulli components in the observations Y.

Chapter 5

Longitudinal summary statistics

We proposed in chapter 3 that preliminary data reduction may be a reasonable, or even a necessary step in situations when the amount of data is vast. This is the case with the fruit fly behavior observations, where the raw data is unmanageable both statistically and computationally. By data reduction we mean replacing the the observations we have with a smaller collection of data, which is derived from the observations. Such derived data are called statistics. In this chapter we introduce several interesting summary statistics for the fruit fly behavior histories, and explore what we lose and gain when we perform this data reduction.

5.1 On data reduction

As a rule, the data at hand is the raw material for the data mining factory. Data may be transformed, screened for outliers, have missing values imputed, but still it is the distribution of observations that we model. If the data is reduced, and then modeled, one may argue that we might have lost some important information about the process

or phenomenon under investigation. We should keep in mind that no data can fully represent a natural phenomenon, and information loss is inevitable in the process of collecting data. In this regard, any data collection can be thought of as implicit data reduction, dictated by limited resources of the study and analytic tools available. If however we perform reduction of information already available, the information loss can be assessed. This will be the subject of the present section.

If our analysis has been carefully planned beforehand, if the models to be used in the analysis have been decided upon, then we do have a legitimate way to reduce collected data without any further loss. This is achieved by sufficient statistics. Sufficient statistics are known for many distributions, and they are especially simple in the case of the normal distribution.

However, it is desireable to relax the stringent requirement of sufficiency. First of all, this is because most statistical models used today do not possess a simple sufficient statistic in a closed form. This can sometimes be overcome using some ad hoc tricks. For instance, Kharrati-Kopaei et al [72] considered a vector ARMA(p,q) process, for which they were able to derive an 'almost sufficient' statistic by means of approximating the process' covariance structure. The quality of the statistic was assured by an asymptotic result for the covariance matrix. For a review of other attempts to use the idea of approximate sufficiency, see Abril [45].

Secondly, sufficient statistics are sufficient only within the parametric framework considered by a statistician. Relaxing the sufficiency requirement ackowledges the fact that 'all models are wrong'. Next we formalize the concept of data reduction and put manipulations with the fly behavior histories on a more solid ground.

Approximate sufficiency As we know, for an observation X, a function $f(X)$ is called a statistic, or a summary. In principle, any function can be considered, but of course some summaries are better than others. In order to assess the quality of a summary, we need to assume that the distribution law of the original data X belongs to a certain family of distributions. Good summaries preserve information about the family parameters while bad summaries lose information. Perfect summarization is achieved with sufficient statistics. For example, if $X = (X_1, ..., X_n)$ is sampled from the normal distribution with mean μ, then the sample mean statistic $\bar{X}$ is sufficient for μ. However, consider the sample median $\tilde{X} = X_{(n/2)}$. It is not sufficient and loses information about μ; precisely, the asymptotic relative efficiency[1] of the sample median is $2/\pi$ (Sherman [74]). In spite of this, the sample median is a very good estimate of μ and is sometimes preferred to $\bar{X}$ for its robustness to outliers. On the other hand, such summary as $\check{X} = (X_1 + X_n)/2$ is apparently bad, because it discards all but two observations. This example shows that insufficiency does not necessarily mean uselessness. In view of this, we will not pursue asymptotic efficiency for proposed summaries. Instead, our definition of approximate sufficiency will rely on the finite sample at hand. We say that a statistic $f(X)$ is approximately sufficient, if it is sufficient for another parametric family of distributions, which is close to the original one.

There are many ways to measure closeness of two distributions. We make our choice in favor of the Kullback-Leibler divergence, also called the relative entropy, introduced in [101]. For simplicity of exposition, we assume that the distributions involved are absolutely continuous with respect to a common measure μ, and denote the corresponding densities by the lowercase letters. With this convention the

[1] Ratio of mean squared errors of two estimators.

Kullback-Leibler divergence is defined as

$$d_{KL}(P||Q) = \int \log \frac{p(x)}{q(x)} p(x)\mu(dx) \tag{5.1}$$

$d_{KL}(P||Q)$ is nonnegative with $d_{KL}(P||Q) = 0$ if and only if $P = Q$. It is not a metric however, and in particular it is not symmetric. Distribution P can be interpreted as the true distribution of observations X, while Q is the distribution we use to model X. $d_{KL}(P||Q)$ is thus the expected log-likelihood ratio. From the information-theoretic perspective, $d_{KL}(P||Q)$ is the expected number of extra bits that must be transmitted to identify a value x drawn from X, if a code is used corresponding to the probability distribution Q, rather than the true distribution P.

Unfortunately, the magnitude of $d_{KL}(P||Q)$ has no universal meaning, and depends on the distributions involved. To have a reference point for this quantity we also consider the Shannon entropy, defined as

$$H(P) = -\int [\log p(x)] p(x)\mu(dx) \tag{5.2}$$

and introduce

$$d(P||Q) = \frac{d_{KL}(P||Q)}{H(P)} \tag{5.3}$$

This quantity measures information loss relative to the actual amount of information contained in the true distribution.

Now we move on to our formal definition of insufficiency. Let a random element X be drawn from a statistical experiment $(\mathcal{X}, \mathcal{A}, P_\theta),\ \theta \in \Theta$, and let $f(X)$ be a statistic. We say that $f(X)$ is approximately sufficient (for θ) with insufficiency $\kappa(f)$ if

$$\kappa(f) = \inf_{\{Q_\theta\}} \sup_{\theta} d(P_\theta||Q_\theta) \tag{5.4}$$

where inf is taken over all parametric families of distributions $\{Q_\theta\},\ \theta \in \Theta$, defined on the same space $(\mathcal{X}, \mathcal{A})$, such that $f(X)$ is sufficient for θ under the experiment $(\mathcal{X}, \mathcal{A}, Q_\theta)$.

First of all, notice that any sufficient statistic $f(X)$ has zero insufficiency, $\kappa(f) = 0$. This is a consequence of the fact that $d(P||P) = 0$, and we are allowed to take $Q_\theta = P_\theta$. Next, consider a parametric family consisting of just one measure: $Q_\theta = Q$. Any statistic $f(X)$ is sufficient for such family, and therefore we have

$$\kappa(f) \leq \inf_{Q} \sup_{\theta} d(P_\theta||Q) \tag{5.5}$$

The bound is attained by $f(X) = 0$ (or any other constant), because the only family Q_θ for which a constant is sufficient is $Q_\theta = Q$. Thus, constants are the worst summaries and provide the upper bound (5.5) for insufficiency of any other statistic $f(X)$. We remark that $d(P||Q)$ could alternatively be defined by normalizing the Kullback-Leibler divergence with the right hand side of (5.5). The resulting quantity would then belong to the interval $[0, 1]$. Such definition however leaves $d(P||Q)$ undefined whenever the right hand side of (5.5) is infinite, which seems undesirable.

Calculating insufficiency for a given family P_θ and a statistic $f(X)$ may be a difficult task. For instance, continuing the example with the sequence of independent observations from a normal distribution with an arbitrary location parameter and a unit variance, we may want to find insufficiency of the sample median, $\kappa(\tilde{X})$. If we take the Laplace family, for which the median is known to be sufficient, we get an upper bound on insufficiency. In Appendix A we show that the Laplace distribution yields $\kappa(\tilde{X}) \leq 0.0342$. Thus, by using the median in place of the mean, we may waste no more than 3.42% of transmitted bits. It seems to be a difficult problem to establish whether some other family provides a better bound.

The concept of insufficiency is much more usable if applied backwards, when we first decide on the family of distributions Q_θ and the associated sufficient statistic. Practically, Q_θ will be 'the working model' - a family with some unnatural features, but for which a sufficient statistic is readily available. P_θ on the other hand will be conceptually meaningful for the phenomenon under study, but too complicated, and will not possess a tractable sufficient statistic.

Application to hidden Markov models Our primary interest will be in applying the data reduction ideas to hidden Markov models. We begin with a lemma, which provides a way to construct sufficient statistics for hidden Markov models. Then we describe how a stream of observations, such as the fruit fly behavior histories, can be handled with this approach.

Lemma 1. *Let (V_t, Y_t), $t = 1, ..., n$ be a hidden Markov model, the distribution of which belongs to a parametric family indexed by θ. Suppose for each t the conditional density of Y_t given $V_t = v_t$ belongs to a parametric family g_λ, and that λ at time t is determined by $\lambda_t = \lambda_t(\theta, v_t)$. Suppose also that g_λ admits a sufficient statistic Z (for λ), and let $Z_t = Z(Y_t)$. Then $(Z_1, ..., Z_n)$ is a sufficient statistic for θ. Moreover, (V_t, Z_t) form a hidden Markov model.*

A proof is given in Appendix A. The statement of the lemma means that we can replace observations Y_t with statistics Z_t, which are sufficient for the conditional distribution. This works if this distribution is specified in a special way, namely if its parameter is a function of the state and the global model parameter θ. Such specification, as we will later see, has a very natural interpretation.

At first glance, this lemma does not give us much. Indeed, if one attempts to

model observations themselves with a hidden Markov model, there will be nothing to summarize. We are going to apply lemma 1 in a less intuitive way, and this idea will be at the core of the first step of estimation, as declared in chapter 3.

Suppose that observations B_s (the fruit fly behaviors in particular) arrive in continuous time, and the observation period is $0 \leq s \leq T$. Consider a partition of the observation period into contiguous blocks of equal length, which for simplicity will also be the unit of the time scale. For clarity of exposition and in relevance to the fruit fly data, we will call this unit a day. Now group the observations by day:

$$Y_t = (B_{s_t}, ..., B_{s^t})$$

where s_t is the first and s^t is the last occasion of measuring B_s on day t. Also let n be the last day of observation. The process B_s is thus given by the sequence Y_t, $t = 1, ..., n$. Our approach is to treat this sequence as observations, and model them with a hidden Markov model. Thus, the value of the hidden process V_t will determine the distribution of B_s during day t.

Next, lemma 1 comes into play. Assuming that sufficient statistics Z_t are available for the conditional distribution of Y_t, we perform data reduction by replacing Y_t with Z_t, and continue analysis with the hidden Markov model (V_t, Z_t), $t = 1, ..., n$. Note that this reduction loses no information, because it is done with sufficient statistics. Of course, we have to keep in mind that information is not lost only in the framework of this hidden Markov model. The model itself can't be accepted as the 'true mechanism' of B_s because of two undesired features. First, we have assumed that a single state value V_t determines the distribution of B_s during an entire day. In the fly data context, it will mean that a fly ages abruptly at midnight, while it does not age at daytime. The second unnatural assumption pertains to the distribution of Y_t, which

will be chosen so that a sufficient statistic can be constructed.

At this point we begin analysis of the fruit fly data, presented in chapter 3. Our strategy will be as follows. First, we introduce various kinds of statistics Z_t. We then describe the distribution families g_λ, for which these statistics are sufficient. The respective hidden Markov models are going to be what we called the measure Q_θ in the discussion of approximate sufficiency. Thus, the data reduction will have small insufficiency with respect to measures P_θ, which are close to Q_θ. While P_θ will be left behind the scene, we understand it to be the 'true' distribution of the behavior process, where the assumptions of stepwise aging and special forms of g_λ are relaxed. In principle, insufficiency can be assessed, but that would require specification of P_θ. For instance, we may choose a finer blocking and obtain Y_t on, say, an hourly basis, and assume that the resulting hidden Markov model is the true P_θ. Then we can evaluate $d_{TV}(P_\theta, Q_\theta)$ and get an upper bound on insufficiency. However we will not pursue this in the present study, and instead focus on building and estimating the hidden Markov model (V_t, Z_t), assuming it provides a reasonable reflection of how aging affects behaviors. In this chapter we explore statistics Z_t. Various structures for (V_t, Z_t) will be considered in chapter 6.

5.2 Summary statistics for the fruit fly data

In this section we are concerned with obtaining summary statistics Z_t for the behavior process $Y_t = (B_s,\ t < s < t+1)$, observed during day t. For convenience, we drop the subscript t from notation. The observations are periodically interrupted, so we have interval censoring of the behavior process B_s. Let A be the observation set, which is

a union of time intervals:

$$A = \bigcup_{\nu=1}^{m} A^\nu$$

Typically, A^ν are equidistant, each about 10 seconds long, and A covers approximately 10% of the day (the observing machine is either halted or monitors the other flies for the rest of the time).

Due to the fast sampling frequency and obviously much longer behavior durations, we regard our data to be continuous observations over intervals A^ν. Since we only observe $N = 6$ different behaviors, we can equivalently represent B_s, $s \in A^\nu$, with the embedded process (X_k^ν, D_k^ν), $k = 1, ..., K^\nu$, where X_k^ν is the behavior occupied and D_k^ν is the time this behavior lasted. By $\mathbf{1}\{C\}$ we denote the indicator function of an event C. Consider the following quantities:

$$N_{ij}^\nu = \sum_{k=2}^{K^\nu} \mathbf{1}\{X_{k-1}^\nu = i, X_k^\nu = j\},\ N_{ij} = \sum_{\nu=1}^{m} N_{ij}^\nu \tag{5.6}$$

$$N_i^\nu = \sum_{k=2}^{K^\nu - 1} \mathbf{1}\{X_k^\nu = i\},\ N_i = \sum_{\nu=1}^{m} N_i^\nu \tag{5.7}$$

where empty sums are understood as zero, and

$$T_i^\nu = \sum_{k=1}^{K^\nu} D_k^\nu \mathbf{1}\{X_k^\nu = i\},\ T_i = \sum_{\nu=1}^{m} T_i^\nu \tag{5.8}$$

(5.6) counts the number of transitions from behavior i to behavior j, (5.7) counts the number of times behavior i was occupied and (5.8) is the time spent in behavior i - during the interval A^ν or the entire day, depending on whether the superscript ν is present. Note that the first and the last observations are censored. We know what behaviors they are but they don't contribute to the count (5.7). We also put $T = \sum_{j=1}^{N} T_j$. This is the total observation time during the day, because at any time a fly can only do exactly one of the behaviors $1, ..., N$.

Summary I: The Markov process The family of summaries we consider in this study is a collection $Z = (T_i, N_{ij},\ i, j = 1, ..., N,\ i \neq j)$. To give it a more intuitive meaning, we represent Z equivalently with

$$\hat{Q}_{ij} = \frac{N_{ij}}{T_i} \tag{5.9}$$

for various pairs (i, j) with $i \neq j$ from $\{1, ..., N\}$, and

$$\hat{\pi}_i = \frac{T_i}{T} \tag{5.10}$$

for $i = 1, ..., N$. So we have $Z = (\hat{Q}_{ij}, \hat{\pi}_i,\ i, j = 1, ..., N,\ i \neq j)$.

The reader familiar with Markov processes will recognize (5.9) as the estimate of the off-diagonal entries of the infinitesimal generator matrix Q (4.4), and (5.10) as its stationary distribution. Indeed, if we assume that $B_s,\ s \in A^\nu$, is a homogeneous Markov process on the finite state space $\{1, ..., N\}$, then the ratio of the transition count to the total occupation time, calculated on A^ν, is the MLE of Q. For the classical result see Billingsley [85, pp. 46-47]. More details can be found in the work of Albert [1], who also obtained asymptotic distribution for this estimate. Our situaton is slightly different due to censoring on both ends of the interval A^ν. The likelihood of observations can be obtained using the embedded process representation (4.6)-(4.7). The embedded Markov chain on A^ν contributes the following term to the likelihood:

$$L_X^\nu = P\{X_1^\nu\} \prod_{k=2}^{K^\nu} P_{X_{k-1}X_k}$$

Here $P\{X_1^\nu\}$ is the likelihood of the initial observation, and P is the transition matrix of the embedded chain (4.6), so that $P_{ij} = Q_{ij}/Q_i$, $j \neq i$. The terms can be rearranged

so we obtain

$$L_X^\nu = P\{X_1^\nu\} \prod_{i,j=1}^{N} Q_{ij}^{N_{ij}^\nu} Q_i^{-N_i^\nu}$$

Note that N_i^ν is precisely the uncensored count (5.7), because we don't observe transitions into the first behavior and from the last behavior. Conditionally on the realization of X_k, the holding times are independent exponential random variables, and therefore contribute the following term to the likelihood:

$$L_{D|X}^\nu = \exp\{-Q_{X_1} D_1\} \left(\prod_{k=2}^{K^\nu - 1} Q_{X_k} \exp\{-Q_{X_k} D_k\} \right) \exp\{-Q_{X_{K^\nu}} D_{K^\nu}\}$$

The first and the last terms of this expression reflect the fact that D_1 and D_{K^ν} are censored (i.e. we don't observe the entire duration times of the first and the last behavior in A^ν), and therefore contribute the survival function rather than the density. Further notice that the terms comprising $L_{D|X}^\nu$ can be regrouped by collecting together those corresponding to the same values of the chain X_k:

$$L_{D|X}^\nu = \prod_{i=1}^{N} \left(\exp\{-Q_i T_i^\nu\} \right) Q_i^{N_i^\nu}$$

Again, we point out that N_i^ν does not count censored behaviors, and therefore the terms $Q_i^{N_i^\nu}$ from L_X^ν and $L_{D|X}^\nu$ cancel. Altogether, the likelihood of $B_s,\ s \in A^\nu$ factors as

$$L^\nu = L_{D|X}^\nu L_X^\nu = P\{X_1^\nu\} \left(\prod_{i=1}^{N} \exp\{-Q_i T_i^\nu\} \right) \left(\prod_{i,j=1}^{N} Q_{ij}^{N_{ij}^\nu} \right)$$

We find that $Z^\nu = (T_i^\nu, N_{ij}^\nu,\ i,j = 1, ..., N)$ provides a sufficient statistic for the generator (provided that $P\{X_1^\nu\}$ does not depend on the parameters of Q).

If now we assume that B_s is a homogeneous Markov process throughout the entire day, estimating the generator becomes a very tough problem because of interval censoring. In this case we do not know when the jumps occured, nor do we know

how many jumps occured between any two subsequent intervals A^ν and $A^{\nu+1}$. Metzner et al [90] give a review of generator estimation methods available as of 2007. In particular, they describe the maximum likelihood method, implemented through the EM algorithm. Direct likelihood evaluation appears to be extremely complicated. To have an idea, we may look at the formula for the likelihood obtained by Healy and DeGruttola [13, formula 2.2], who considered a special case of exactly one transition occurring in the censored interval. Apparently, the exact likelihood evaluation does not yield a manageable sufficient statistic, i.e. that of a fixed dimension, which prohibits using it in our approach.

Alternatively, we can make an additional assumption regarding the process B_s: we assume that B_s, $s \in A^\nu$ are independent realizations of the same Markov process. Then (5.9)-(5.10) is still sufficient, because in this case we have

$$L = \prod_{\nu=1}^{m} L^\nu = \left(\prod_{\nu=1}^{m} P\{X_1^\nu\} \right) \left(\prod_{i=1}^{N} \exp\{-Q_i T_i\} \right) \left(\prod_{i,j=1}^{N} Q_{ij}^{N_{ij}} \right)$$

which is of the same form as L^ν. In terms of the definitions of lemma 1, we have identified the parametric family g_λ as the distribution of m independent paths of a Markov process. The parameter λ is the generator Q, while the initial distributions are left unspecified. More generally, Z is approximately sufficient for measures close to g_λ, which is specified as

$$g_\lambda(Y) = f(\lambda, Z) h(B_s,\ s \in \bigcup_\nu A^\nu)$$

This may include relaxing the assumptions of independence and identical distribution of the paths, as well as the Markovian dependence structure and homogeneity. In this case $\hat{Q}$ can be thought of as the generator of an 'average' homogeneous Markov process, which best fits the data, just as we interpret a linear regression fit to possibly

non-linear data. For this reason, we call $\hat{Q}$ the generator summary. We note however that this interpretation is loose and should be used with caution, because $\hat{Q}$ may be biased as a generator estimate in the general case.

Likewise, (5.10) is reminiscent of the stationary distribution π of this approximating Markov process. To see this, recall that for a Markov process on a finite state space stationarity of a distribution π is equivalent to satisfying the global balance equations: $\pi Q = 0$, [86, p. 343]. For the distribution $\hat{\pi}$ we have

$$(\hat{\pi}\hat{Q})_j = \sum_{i=1}^{N} \hat{\pi}_i \hat{Q}_{ij} = \sum_{i \neq j} \frac{T_i}{T} \frac{N_{ij}}{T_i} - \frac{T_j}{T} \frac{\sum_{k \neq j} N_{jk}}{T_j} = \frac{1}{T}(\sum_{i \neq j} N_{ij} - \sum_{k \neq j} N_{jk}) \qquad (5.11)$$

In parentheses we have the difference between the number of transitions into j and from j. These differ at most by 1 if there is no interval censoring, which is asymptotically negligible compared to the growing observation time T. Convergence under no censoring also follows directly from the ergodic theorem, applied to the functions $f_j(B_s) = \mathbf{1}\{B_s = j\}$. With censoring, however, the difference between the transition count to and from j can be substantial, and convergence of $\hat{\pi}$ may fail. We therefore prefer a more cautious understanding of $\hat{\pi}_j$ as merely the observed fraction of time occupied by the behavior j during the day.

Summary II: The daily cycle A day is a natural time unit to look at. We have presented a summary obtained on a daily basis, which estimates the Markov parameters assuming homogeneity of the behavior process throughout any given day. While this may be fruitful for the purpose of joint analysis, it is of interest to look at the more local daily patterns, which can be missed entirely by daily summaries.

To accomplish this, we examine the behavior process hourly. Our first observation is that there is a consistent daily cycle, through which every fly goes every day of its

life. Fig. 5.1 shows hourly behavior fractions, as calculated in the pooled collection of all days and flies. In the subsequent analysis we restrict our attention to the 13-hour daylight period, from 7 am to 7 pm. This does not cover the night period, when the flies sleep most of the time and their activity is substantially lower.

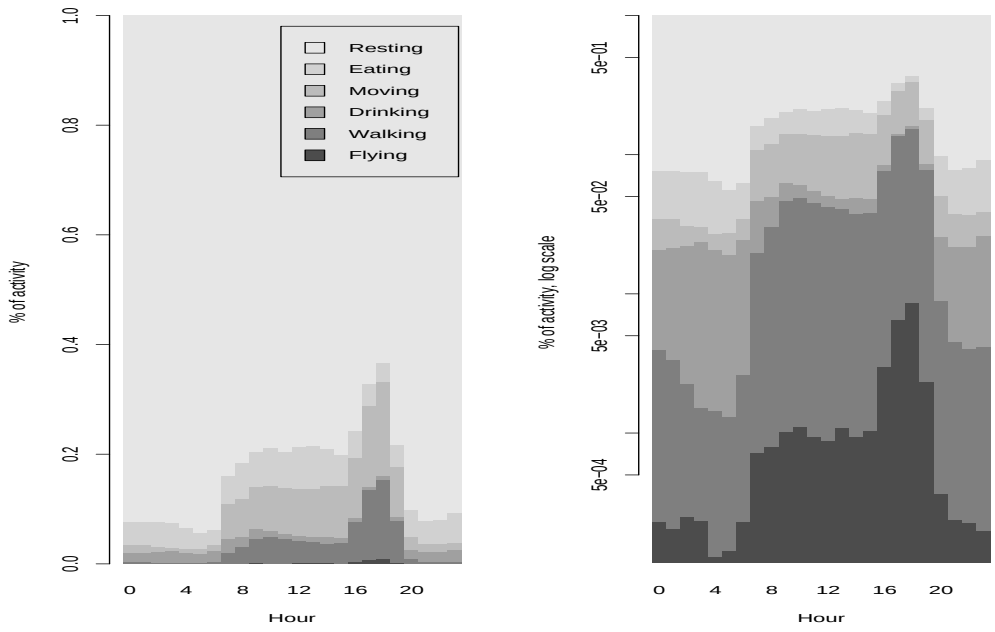

Figure 5.1: Observed proportions of time for each of the six activities. The proportions are grouped by hour and calculated using data from all flies and all days. In the chart the proportions are represented by stacked bars. Left: probability scale, right: logarithmic scale.

Within any given hour, we again derive the Markov summaries Z. These statistics are sufficient for a homogeneous Markov process, but now the assumptions apply within an hour rather than a day. To make distinction between day and hour, we use t for the days and h for the hours. Thus, the summaries Z can be grouped by time in the following way: $Z_t = (Z_{t,h})$, where h ranges from 7 am to 7 pm. So Z_t is now a collection of vectors $\hat{\pi}_i$ and $\hat{Q}_{ij}$, each of dimension 13. For example, the estimated

fraction of resting is now given by a vector of 13 such fractions within each of the 13 hour intervals.

Further analysis can be based on the statistics $(Z_{t,h})$ themselves, but we find it more meaningful, in terms of the daily cycle assessment, to first transform these vectors. For the clarity of presentation, we explain how the transformation is done on the example of just one of the components, $\hat{\pi}_1$. Let U be a square matrix formed of orthonormal basis column vectors in the 13-dimensional space, and we consider $\tilde{\pi}_1 = \hat{\pi}_1 U$. Instead of the vector of hourly fractions of behavior 1, $\tilde{\pi}_1$ contains its coefficients under basis U. The purpose of this transformation becomes clear when we specify the matrix U.

Let Π be the matrix of the hourly fractions of behavior 1, calculated from the pooled collection of all days, separately for each fly. Thus, Π_{fh} is the fraction of time behavior 1 was observed during hour h of fly f. With our dataset, Π has 27 rows corresponding to the flies and 13 columns corresponding to the hours 7 am through 7 pm. Now we run the singular value decomposition (SVD) on Π, [93, Sec. 7.3]:

$$\Pi = V \Sigma U'$$

Here V is a 13×27 matrix with orthogonal columns, Σ is a 13×13 diagonal matrix and U is a 13×13 orthogonal matrix. In addition, the numbers on the diagonal of Σ (the singular values) are ordered by descending absolute value. The columns of U can be interpreted as the principal components of Π. Thus, the first column of U is the vector that explains the most variability among the rows of Π in the least-squares sense, and Σ_{11}^2 is the amount of variance explained. It gives the principal pattern in the hourly statistics across the flies. The next principal component, or pattern, is derived as the vector which is orthogonal to the first component and explains the

most of the remaining variance. The rows of the matrix $V\Sigma$ contain the expansion coefficients for the rows of Π in the basis U.

The orthogonal matrix U from the SVD will define the transformation applied to the hourly summaries $\hat{\pi}_1$. It identifies 13 hourly patterns, ordered by their strength as measured by the variance explained. The resulting vectors $\tilde{\pi}_1$ will show how strong each of the patterns is during a given day. The transformation procedure we have just outlined is similarly applied to the other components of the summary $Z_{t,h}$, and we arrive at $\tilde{Z}_{t,h} = (\tilde{\pi}_i;\tilde{Q}_{ij})$. Note that the statistic $\tilde{Z}_{t,h}$ has the same sufficiency properties as the untransformed one, because the transformation is invertible.

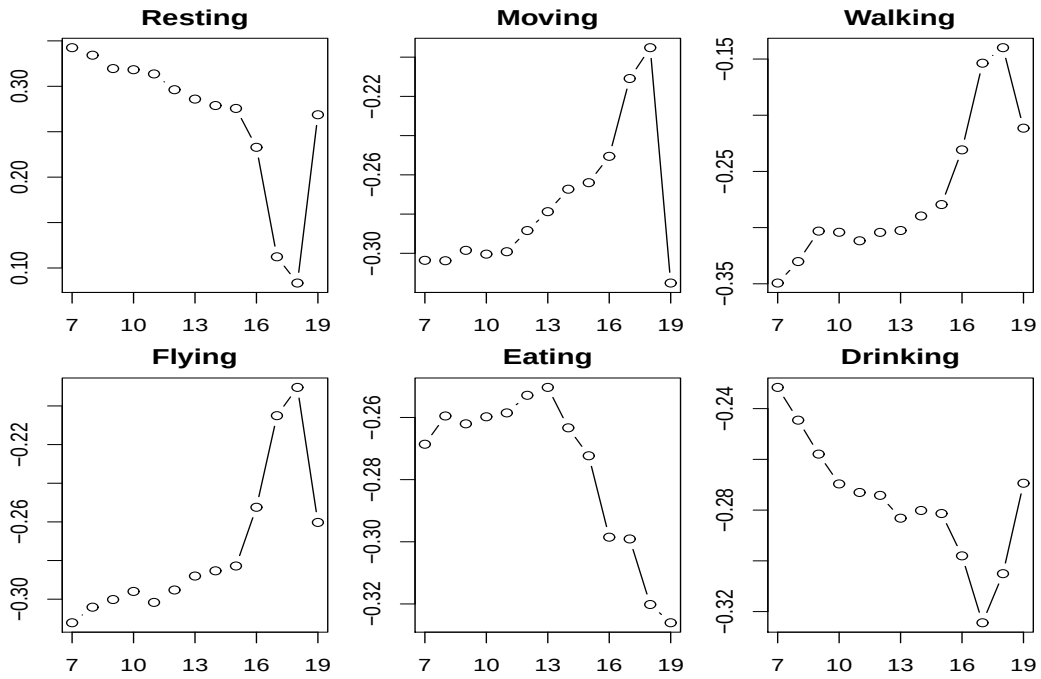

Figure 5.2: Principal components of the hourly pattern in behavior logits. The hours (x-axis) are dimensions and the y-axis are the respective coordinates of the principal vectors. All principle vectors are positively correlated with the behavior frequencies, so higher coordinate values mean the behavior is more frequent during those hours.

Fig. 5.2 illustrates the approach described for hourly patterns for the behavior

fractions (on the logit scale). We plotted the principal vectors, and observed that a fly's activity is generally not homogeneous during the day. Upon examining the singular values, we find that the daily variability is captured to a great extent by the principle singular vector. This means high consistency of the pattern across flies.

Distribution of the summaries An important question is the distribution of the derived sufficient statistic Z. In order for it to enter a hidden Markov model (V_t, Z_t) as an observation Z_t, we need to specify how Z is distributed. If Z is a maximum likelihood estimate, then its asymptotic distribution can be used, which is known to be normal. This is the case with the generator summary, and therefore $\hat{Q}$ is approximately normal with mean Q. Albert [1] obtained the asymptotic variance for this estimate under growing number of samples, $m \to \infty$, and under growing observation duration of a single path. It turns out that the estimates $\hat{Q}_{ij}$ are asymptotically independent for different pairs (i, j). If we treat $B_s,\ s \in A^\nu$ as independent realizations, and in addition assume that all A^ν are of the same length Δ, then the first scenario applies $(m \to \infty)$, and we have the following asymptotic variance for $\hat{Q}_{ij}$:

$$\mathrm{Var}(\hat{Q}_{ij}) \sim \frac{Q_{ij}}{\int_0^\Delta P\{B_s = i\}ds},\ m \to \infty$$

To estimate this variance, we replace Q_{ij} with $\hat{Q}_{ij}$, and the denominator with $\Delta\hat{\pi}_i$:

$$\hat{\mathrm{Var}}(\hat{Q}_{ij}) = \frac{\hat{Q}_{ij}}{\Delta\hat{\pi}_i} \tag{5.12}$$

There are also at least two alternative ways to estimate the variance, and those are more robust to the assumptions of the model. The first is the sandwich estimate, which is generally valid for misspecified model. The other is the jackknife method,

whereby we produce $\hat{Q}_{ij}^{-\nu}$ by excluding the ν-th observation interval A^{ν} from calculations (5.6)-(5.8). The variance is then estimated by

$$\hat{\mathrm{Var}}(\hat{Q}_{ij}) = \frac{1}{m}\sum_{\nu=1}^{m}(\hat{Q}_{ij}^{-\nu} - \bar{Q}_{ij})^2 \tag{5.13}$$

where $\bar{Q}_{ij} = \frac{1}{m}\sum_{\nu=1}^{m}\hat{Q}_{ij}^{-\nu}$.

Event history charts With the development of computer technology, graphical tools have become an essential part of any data analysis. It is hard to overestimate the human eyeball as a pattern recognition and data mining tool. For instance, to assess normality of data a statistician will look at the Q-Q plot, and he will examine the spectrum to identify periodic components.

Graphical representation of longitudinal data is challenging. Such chart has to show the overall trends in the characteristic of interest, with the possibility to track down each subject individually. It is especially difficult to plot data in the context of joint modeling, where it is also desireable to add the survival information. The problem was solved by Carey et al [48], who proposed an elegant way of presenting longitudinal survival data. In their paper of 1998, the authors examined the association between fertility data of medflies and their survival. The proposed graphical tool was named event history chart.

The idea of an event history chart is the following. Each subject is represented by a rectangle with a unit height, and the base stretching from time zero to the observed (or censored) lifespan. The rectangles are stacked one upon the other, in the decreasing order of the lifespans. This way, the right sides of the rectangles produce the pattern of the empirical survival curve. We then paint the rectangles, using color to encode the value of the observed longitudinal measure. After introduction in 1998,

event history charts were used by Carey and his collaborators to analyse behavioral patterns in cohorts of drosophila and medfly, [46]. The results were astonishing. The complex dynamics of longitudinal data and their connection to longevity can now be easily seen. Event history charts revealed that almost all behaviors of drosophila are age specific, and some behaviors of medflies are predictive of their remaining lifespan. In addition, the pictures show periodic nature of some behaviors in both cohorts, although the authors did not point that out.

To conclude, event history charts are an excellent way to display the data of the present study, and perform preliminary qualitative analysis. Unlike the original approach, the nature of our data precludes us from using the observations themselves in coloring the chart: the vast local information may easily hide interesting long-term trends. We therefore need to consider a summary statistic as a longitudinal measure, and this will be the statistic $Z = (\hat{Q}_{ij}, \hat{\pi}_i,\ i, j = 1, ..., N,\ i \neq j)$ which we introduced above. We note that these statistics were produced on a daily basis, so in fact we have a sequence in time of vector valued statistics Z_t. The event history charts will be produced for each one component of Z_t considered at a time.

First we examine the observed daily fractions of each behavior on the logit scale, each computed based on the daylight period (7am through 7pm), fig. 5.3. We used the traffic light coloring scheme as in Carey et al [46], where red represents larger fractions and green represents smaller fractions. For each behavior j, the thresholds between the colors were determined by the tertiles in the overall collection of $\hat{\pi}_i$, pooled across all days and all files. Grey areas under the survival curve represent missing statistics, which may happen if there is not sufficient information to compute the summary.

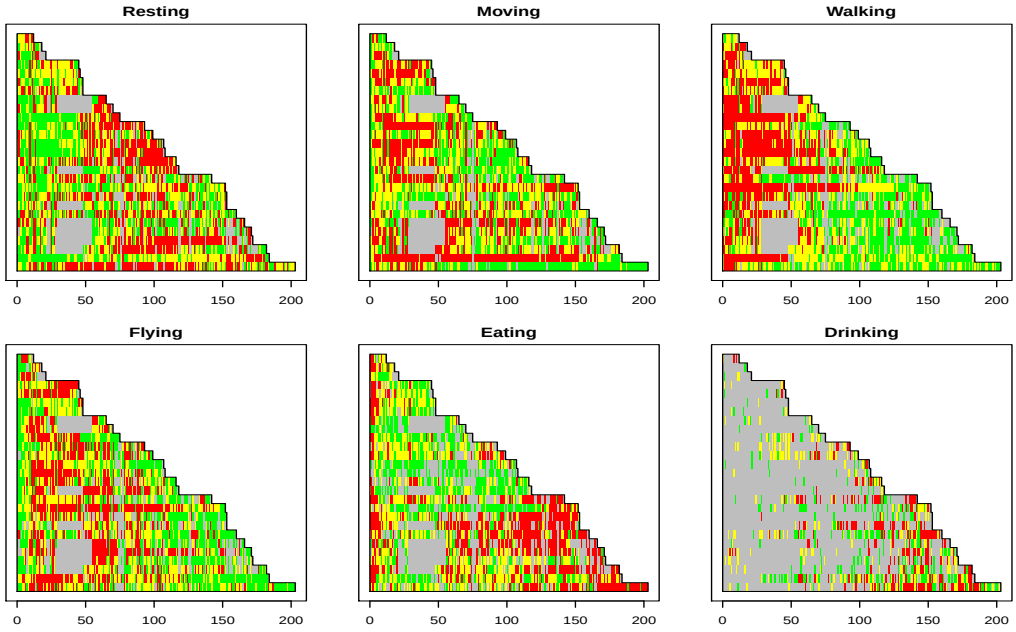

Figure 5.3: Daily fractions of time spent in each behavior. A fly is represented by a horizontal strip. Each day on the strip is colored in red, yellow or green depending on whether the behavior frequency was in the upper, middle or lower tertile that day. The strips are ordered by descending lifespan. Grey areas represent missing data.

There are a few patterns which we can identify from fig. 5.3. Some of the behaviors tend to change their frequency over time. The sharpest pattern is in the walking chart, showing that flies walk a lot more in early life (till about day 50) than they do later. Moreover, between day 50 and day 100 there appear to be more green areas in the longer-lived flies. Similarly, resting is not as frequent in early life (again till about day 50), as it becomes in the later life. Flying is not frequently observed in the first few days of life as well as in very late life, after day 100. There is no clear indication of any survival related patterns in flying. Next, eating shows an interesting picture. It is frequent during the first few days after birth, and then the green color prevails. At approximately day 50, the longer surviving flies begin to eat more, while

flies with the shorter lifespans continue to be on the green side. Finally, drinking is generally not frequent enough, and the corresponding summary is as a rule missing. In summary, we have identified two statistics, which may potentially exhibit survival related patterns: the daily fraction of walking and eating time.

Fig. 5.5 shows the rest components of the Markov summary Z, the generator matrix $\hat{Q}$. We presented them in the matrix form which follows the structure of $\hat{Q}$. Note that we have nothing on the diagonal, because these are redundant statistics and are not part of the Z vector. The rows and columns corresponding to eating and drinking were omitted due to very sparse non-missing statistics. Moreover, the walking-to-moving and moving-to-walking transitions are non-existent. The respective entries of the Q matrix may be regarded as zeroes. The data are considered on the log-scale. Again, the statistics were produced on a daily basis.

Most of these charts suggest that some changes to the generator summary occur over time. There is a weak indication in the flying-to-moving and flying-to-walking charts that these entries may have survival related patterns.

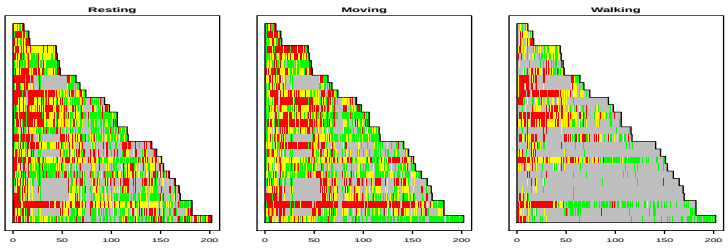

Figure 5.4: Principle hourly pattern coefficients for the behavior frequency logits.A fly is represented by a horizontal strip. Each day on the strip is colored in red, yellow or green depending on whether the coefficient was in the upper, middle or lower tertile that day. The strips are ordered by descending lifespan. Grey areas represent missing data.

Finally, fig. 5.4 contains the event history charts for a part of the summary $\tilde{Z}_{t,h}$.

We plotted the principle coefficient for each of the hourly behavior frequency logit. In our notation these statistics are $(\tilde{\pi}_i)_1$, where i indexes behaviors and 1 means the first (principle) coefficient of the resulting vector. Due to the much finer time scale, the rate of missing statistics is considerably higher than in the daily charts 5.3 and 5.5. Non-missing statistics for flying, eating and drinking are so sparse that we omitted the corresponding charts.

There is no clear pattern in these charts, except maybe for the walking behavior, where we see more red in early life. However, missing statistics prevail, and the pattern may be not reliable.

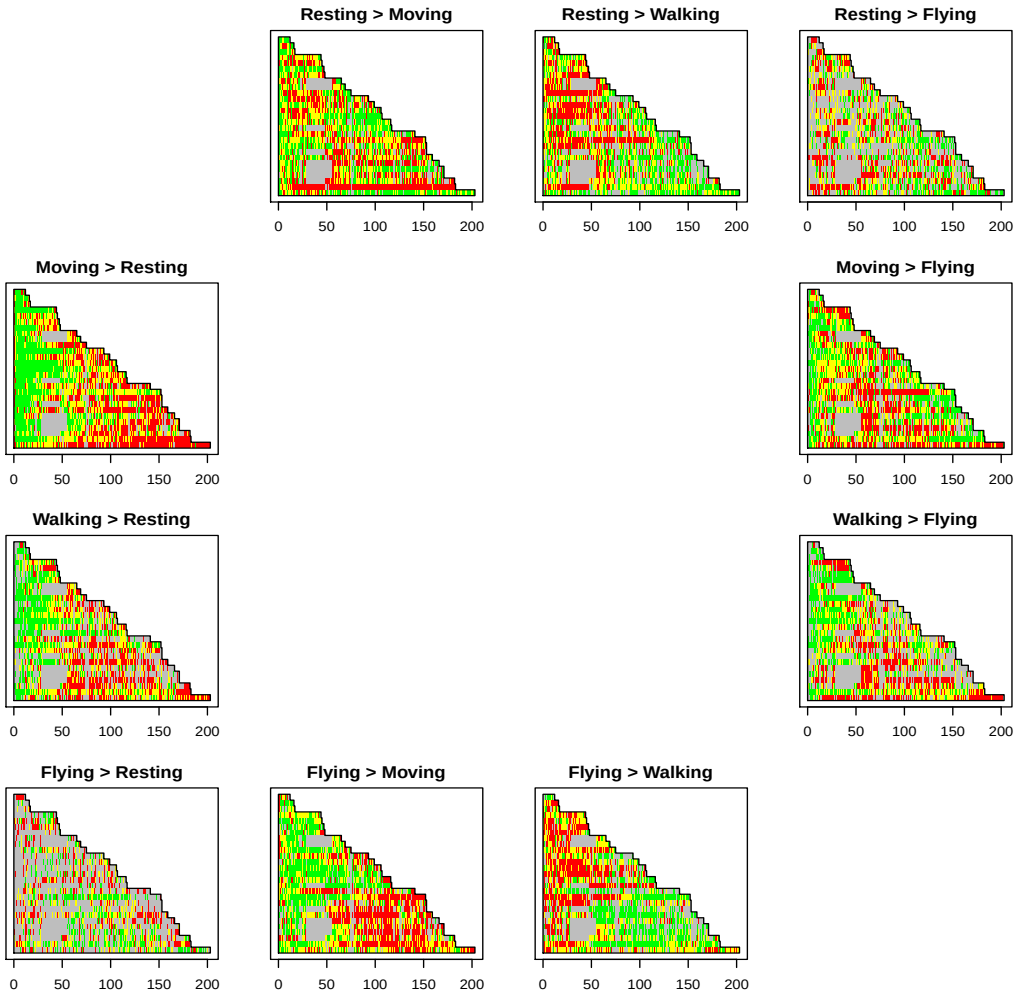

Figure 5.5: Daily Markov generator statistics. A fly is represented by a horizontal strip. Each day on the strip is colored in red, yellow or green depending on whether the generator statistic was in the upper, middle or lower tertile that day. The strips are ordered by descending lifespan. Grey areas represent missing data.

Chapter 6

The joint model

Having introduced the longitudinal summary statistics Z_t in chapter 5 we complete the first stage of analysis and pass on to the second - fitting a hidden Markov model (V_t, Z_t) in discrete time. It is at this point that we decide on the nature and dynamics of the hidden process V_t and conditional distribution of the observations Z_t. We see a close connection of this kind of analysis to both the joint models and the models of aging. In the exposition below we do not favor one context over the other, but give remarks showing the approach from both angles.

6.1 The hidden process

We begin the description of the hidden Markov model (V_t, Z_t) with specifying its hidden component V_t. We first discuss its role as the aging process in application to the fruit flies. Then we propose several choices for the dynamics of V_t for analysis of these particular data.

Understanding the hidden process In joint modeling the longitudinal and survival component are linked by unobserved random variables. These random variables may have very different meaning, depending on the application. They may simply be the true value of the covariate, while the observations available to the statistician are noisy covariate measurements. In other applications, the hidden variable may be a hypothetical construct, reflecting heterogeneity in population or individual variability over time.

In the fruit fly experiment, genetical and environmental heterogeneity is minimized: the flies were bred from the same genetic strand and kept under constant environmental conditions. The only factor that has a significant impact on mortality and, as we hypothesize, on the behaviors, is aging. Many joint analysis studies used the time variable as a covariate, and this may be thought of as a deterministic aging process. However, it has been commented many times by different authors that there is more to aging than a deterministic increase in mortality. Rossolini and Piantanelli [36], for example, wrote "Among the most serious difficulties is the evidence that aging appears as a multi-faceted process generated by interrelated functions varying at different rates not only in different individuals of a population, but also at different speed in the same individual". With these considerations, all changes that occur in mortality and in the long term behavior distribution can be attributed to aging. Statistically, we represent aging by the hidden process V_t.

Determining the dynamics of V_t is difficult, because we never get to observe it, and thus we rely on the particular context of the application. Several ideas expressed in the literature suggest that aging should be cumulative in nature. "Complex systems are made of numerous sub-systems interacting with each other; each subsystem is further

divided into interacting lower components and so on at lower levels of organization. They are highly dependent upon initial conditions. Thus in a cohort, even very small differences occurring at certain times can cause higher and higher differences at later ages in most of individual phenotypic characteristics", Rossolini and Piantanelli [36]. Thus a random walk type of model seems a reasonable choice for the fruit fly data. One of the simplest stochastic models that satisfies this requirement is a Brownian motion with a deterministic linear drift, considered by Weitz and Fraiser [55] to describe the hidden viability of an organism. A more general formulation was considered by Woodbury and Manton [76], where V_t was described by a stochastic differential equation. Wang and Taylor [111] used the integrated Ornstein-Uhlenbeck process for V_t, although it did not have the meaning of aging in their application. Yashin et al. [10] have a very elaborate discussion on the choice of the hidden aging process.

Modeling the hidden process In application to the fly data, we consider three Gaussian classes to model the process V_t, two of which are cumulative and one is stationary: Brownian motion (BM), Ornstein-Uhlenbeck (OU) and integrated Ornstein-Uhlenbeck (IOU). The discrete versions of these processes are, respectively, the random walk, the autoregression of order 1 and the cumulative autoregression of order 1. Using ARIMA notation, these processes are ARIMA(0,1,0), ARIMA(1,0,0) and ARIMA(1,1,0), and thus have simple state space representations. Indeed, these distributions can be defined through the linear normal state equation, as in (4.16):

$$V_{t+1} = A_t V_t + \eta_t,\ \eta_t \sim \mathcal{N}(0, R_t)$$

To define the ARIMA models we need is to specify the evolution matrices A_t and the variance matrices R_t.

It is straightforward to see that

$$A_t^{BM} = 1,\ R_t^{BM} = 1 \tag{6.1}$$

defines a random walk with unit variance increments, and

$$A_t^{OU} = \alpha,\ R_t^{OU} = 1 - \alpha^2 \tag{6.2}$$

defines a stationary order 1 autoregression with unit marginal variance. To obtain the state equation for the integrated autoregression, we consider the vector $(V_t, V_{t-1})'$ as the state. We have $\Delta V_t = V_t - V_{t-1}$ and $\Delta V_t = \alpha \Delta V_{t-1} + \eta_t$, where η_t are iid normal variables with variance $1 - \alpha^2$. Thus,

$$V_{t+1} = V_t + \Delta V_{t+1} = V_t + \alpha \Delta V_t + \eta_{t+1} = (1+\alpha)V_t - \alpha V_{t-1} + \eta_{t+1}$$

If follows that the integrated autoregression can be defined as the first component of the two dimentional state $(V_t, V_{t-1})'$, which has the following dynamics:

$$A_t^{IOU} = \begin{pmatrix} 1+\alpha & -\alpha \\ 1 & 0 \end{pmatrix},\ R_t^{IOU} = \begin{pmatrix} 1-\alpha^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.3}$$

Again, we chose the variance parameter so that the variance on the process increment is one.

The distribution of the initial value V_0 is the distribution of initial aging state. Since there is no particular reason to believe that the flies were born with different survival chances, we require $V_0 = 0$. In another application one might consider V_0 random with a normal distribution.

The three processes presented differ in a way we interpret aging. BM and IOU represent accumulated damage and repair, which adds up as the fly progresses into

the next day of its life. In the BM setting, the increments are independent of the current state while in the IOU case they vary smoothly from day to day. The OU setup assumes a stationary state rather than cumulative. Thus a fly is more frail in some periods of its life than in the other, and it is the newly received damage or repair that factors in the current risk of death under this model. By considering these three specifications we allow greater flexibility in modeling the data, and illustrate how easily different theoretical concepts can be realized with the state space approach. Should we decide to model the aging dynamics in a different way, all we would have to do is redesign the state equation for the model. Next we explain how V_t affects the mortality and the observations.

6.2 Subject specific mortality

Before we start the discussion about mortality, some clarification is required concerning the time variable t. We are developing the analysis in discrete time, where t indexes the days, while the definition of mortality $m(s)$ normally implies continuous time s. In this regard we consider the discrete function $\tilde{m}(t)$, defined for $t = 1, 2, ...$ as follows:

$$\tilde{m}(t) = \int_{t-1}^{t} m(s)ds$$

The survival function then becomes

$$S(n) = \exp\left\{-\int_0^n m(s)ds\right\} = \exp\left\{-\sum_{t=1}^{n} \tilde{m}(t)\right\}$$

Thus, $\tilde{m}(t)$ is the average of $m(s)$ over the time interval $[t-1, t]$. We have the approximate equality $\tilde{m}(t) \approx m(t)$, with exact equality occurring in the case of the

piecewise constant mortality. Similarly,

$$S(n) \approx \exp\left\{-\sum_{t=1}^{n} m(t)\right\}$$

In the sequel, we will not distinguish between $\tilde{m}$ and m, even though they may not be exactly the same. In particular, we will speak about the Gompertz law in discrete time, although the function $\tilde{m}$ for the Gompertz law (2.1) is slightly different from m. Having said that, we continue to work with the discrete time variable t.

In random frailty models subject specific mortality, or hazard function, is usually specified as

$$m(t|\theta) = \theta m_0(t)$$

where θ is the individual frailty and is considered to be random in the population. To make this term identifiable it is often assumed that $E\theta = 1$, while $\mathrm{Var}\theta = \sigma^2$ is a parameter which quantifies heterogeneity of the population. The most commonly used distributions for θ are gamma, log-normal and positive stable. In chapter 2 we discussed and provided some references to extensions of the frailty idea to the case where longitudinal observations are available, and frailty may be time specific. In such models, the subject specific mortality is determined by the value of a random process θ_t. Generally, we define individual mortality as

$$m(t|\theta_{\leq t}) = m(t|\theta_t) = \theta_t m_0(t) \tag{6.4}$$

The traditional interpretation of frailty is the ratio of an individual's hazard to the baseline hazard m_0. With the changing frailty model, frailty is the current hazard ratio. Another way to look at changing frailty is the effect of aging on a given individual. If no frailty is present, every individual in the population ages in the same way. For example, with the Gompertz law (2.1), log-mortality increases with

time as Ct. Including the changing frailty term will make individual rates of hazard acceleration randomly fluctuate around Ct. Thus some individuals may age faster, while others may age slower. In this regard, aging is viewed as the individual's stochastic clock (Mangel [73]).

There is an endless possibility for specifications of θ_t and $m_0(t)$. Acknowledging the profound role of the Gompertz distribution in aging studies, which has been established by the aging research community over many years, we consider the Gompertz baseline hazard (2.1):

$$m_0(t) = Ke^{Ct} \tag{6.5}$$

A variety of frailty distributions has been considered by many authors, among them gamma (Vaupel et al [54]), positive stable (Hougaard [87]) and log-normal (McGilchrist and Aisbett [18]). Extensive discussion of merits of these and other frailty distributions can be found in the recent book by Duchateau and Janssen [59]. We make our choice in favor of the log-normal distribution for purely practical reasons, resulting in the most transparent exposition of model building and estimation. Precisely, we consider

$$\theta_t = \exp\{\gamma' V_t\} \tag{6.6}$$

where V_t is the hidden process we introduced above, and γ is a vector of parameters, defining the effect of the possibly multidimensional state V_t. Altogether, (6.4)-(6.6) define the survival distribution of the subjects in the fly cohort. With the three hidden process specifications we consider, (6.1), (6.2) or (6.3), the frailty process θ_t has log-normal marginal distributions (the IOU state specification (6.3) has a second technical dimension, so in this case we take $\gamma = (\gamma_1, 0)'$). Note that any other frailty distribution could be modeled with little effort. This would lead to a non-normal

state equation and require additional steps in model fitting (see the iterated Kalman filter discussion in chapter 4).

Before we proceed, we remark that the population mortality has a very different structure than the subject specific mortality, and generally does not exhibit the baseline pattern $m_0(t)$. It is known in particular that random frailty models have plateaus in the population mortality $\bar{m}(t)$ even if the baseline $m_0(t)$ does not possess this feature (for the most general treatment, see Steinsaltz and Evans [31]). It is thus of interest to examine this phenomenon in our changing frailty model. As an illustration, we described $\bar{m}(t)$ and obtained its plateau level for the case when the log-frailty V_t is the Brownian motion. These results are formulated in continuous time and are stated in lemmas 3-5 of Appendix A.

6.3 Killed state space models

The next building block of the joint model is the observation equation, which provides the conditional distribution of observations, given the current hidden state. After the data reduction stage, we have a discrete sequence of summary statistics Z_t, $t = 1, ..., n$, where t indexes days of a fly's life, and n is the last day of observation. As was stated earlier, our intent is to treat (V_t, Z_t), $t = 1, ..., n$, as a hidden Markov model. But there is more to the observation process than the longitudinal data Z_t: having collected Z_1 through Z_t, we also know that the subject has survived up to time t. To distinguish the random survival time from its realization n, we use τ to represent the former. Thus, the entire observation history of each subject consists of $Y = (Z_1, ..., Z_n; \tau = n)$. In our data, n is always the death time, but in other applications censoring may be involved. Although censoring is not treated explicitly in the present study, it can be

accounted for by replacing $\tau = n$ with $\tau > n$. Our immediate goal is to make the observations vector Y suitable for hidden Markov modeling. To do this, we need to represent Y as a sequence $Y = (Y_1, ..., Y_n)$, where the distribution of Y_t is determined by the current state V_t.

Survival indicators We will represent the survival time τ by a sequence of Bernoulli trials S_t, $t = 1, 2, ...$, terminated right after the first failure. Definition of S_t is very similar to the construction of the geometric random variable as the time to failure in Bernoulli trials, albeit in our case the trials will have different probabilities of success. In order for the sequence $S_1, S_2, ...$ to equivalently represent τ, the events $(S_1 = 1, ..., S_{n-1} = 1, S_n = 0)$ and $\tau = n$ must have the same probability for any n. This can be achieved if, in particular, the respective conditional probabilities, given the state realization $V_{\leq n}$, are equal. We have, with the mortality specification (6.4) and (6.6),

$$P\{\tau > n | V_{\leq n}\} = \exp\left\{-\sum_{t=1}^{n} e^{\gamma' V_t} m_0(t)\right\}$$

Now define $p_t = p_t(V_t)$ as

$$p_t = P\{\tau > t | \tau > t-1; V_{\leq n}\} = \exp\left\{-e^{\gamma' V_t} m_0(t)\right\} \tag{6.7}$$

Noting that

$$P\{\tau = n | V_{\leq n}\} = P\{\tau \leq n | \tau > n-1; V_{\leq n}\} \prod_{t=1}^{n-1} P\{\tau > t | \tau > t-1; V_{\leq n}\} =$$

$$(1 - p_n(V_n)) \prod_{t=1}^{n-1} p_t(V_t), \tag{6.8}$$

we conclude that $P\{S_t = 1 | V_t\} = p_t(V_t)$ given in (6.7) provides the correct choice of the distribution of the indicators S_t. With this definition it appears natural to

interpret the event $S_t = 1$ as survival of the fly through day t, and $S_t = 0$ as death on day t. However, such a direct connection with τ prevents (V_t, S_t) from being a hidden Markov model. Indeed, in this case $S_t = 0$ would imply $S_{t+1} = 0$, while a hidden Markov model structure requires conditional independence of S_t and S_{t+1}. Instead, we postulate conditional independence of $S_1, S_2, ..., S_t$ given $V_{\leq t}$, where t is allowed to run past $\tau = n$. Thus, (V_t, S_t) is a hidden Markov model by definition, and the distribution of τ coincides with the distribution of the first time zero occurs in the sequence $S_1, S_2, ...$

The last step in converting the observation Y into a hidden Markov model is rather obvious: we need to put $Y_t = (Z_t, S_t)$. There is a catch, however. We have defined the survival indicators so that (V_t, S_t) is a hidden Markov model, and (V_t, Z_t) is a hidden Markov model by assumption. Moreover, by construction, the sequences Z and S are conditionally independent, given $V_{\leq n}$, so that (V_t, Y_t) is also a hidden Markov model. This does not mean, however, that $Y = (Z, \tau)$ and (Z, S) have the same likelihood, even though this is true of the marginal likelihoods of τ and S, due to (6.8). This problem does not occur if we assume that Z and τ are conditionally independent, given V. Considering $Y_t = (Z_t, S_t)$ is fully legitimate in this case. We believe that such assumption is quite realistic in application to the fruit flies, as it says that in the absence of aging, survival time is independent of behavior.

Longitudinal observations Regarding the distribution of Z_t, we will assume that it is normal, conditionally on the current state. This assumption is rather technical, and is hardly possible to verify because of the conditioning on an unobservable variable. Departures from normality could potentially be assessed indirectly with a goodness of fit test, but if such test showed a lack of fit, it would be impossible to

tell whether the state equation or the observation equation is inadequate. Moreover, each summary Z_t comes with an error, which is due to the estimation in stage 1. As we have seen in chapter 5, these errors can often be treated as normally distributed. This is because Z_t were derived as the MLE or approximate MLE in a local observation model, such as the Markov behavior process during a day. Thus, Z_t remains conditionally normal after the estimation error is accounted for.

The next question about Z_t we need to answer is their conditional serial correlation. A hidden Markov model assumes no such correlation present. However, it may be desired to allow serial correlation among Z_t even after conditioning on the frailty state V_t. Indeed, suppose a fly does not age, and its level of V_t stays constant from day to day. In this case we might still expect to observe correlation between the behavior pattern summaries Z_{t_1} and Z_{t_2}, calculated from two different days. For example we may want Z_t to be conditionally stationary rather than independent. This possibility can be easily accomodated by a hidden Markov model if we introduce extra technical dimensions to the state, which will not be related to frailty. We have already seen this trick when the IOU state equation was defined in (6.3). Specific examples will be given when we continue the analysis of the fruit fly data.

Computing the likelihood At this point we can collect all statements and assumptions about the observations Z and the survival times τ into the following definition.

Definition 3. *Let τ be a positive valued random variable with $P\{\tau < \infty\} = 1$. Suppose Z_t are observed for $t = 1, ..., \tau$. We say that $(Z_1, ..., Z_\tau; \tau)$ follow a (linear normal) killed state space model, if their joint likelihood coincides with that of the hidden Markov model $(V_t, (Z_t, S_t))$, observed for $t = 1, ..., \tau$ with $\tau = \min\{u : S_u = 0\}$,*

and defined by the following equations:

$$\begin{cases} V_{t+1} &= A_t V_t + \eta_t, & \eta_t \sim \mathcal{N}(0, R_t) \\ Z_t &= B_t V_t + \varepsilon_t, & \varepsilon_t \sim \mathcal{N}(0, \Sigma_t) \\ S_t, & & \text{Bernoulli}(p_t(V_t)) \end{cases} \tag{6.9}$$

We called (6.9) a state space model, because it has the form (4.16), save the binary component S_t. Conditional independence of Z and S is inherent in this state space formulation. An interesting feature of (6.9), which will be exploited below, is that it can produce observations even after $S_t = 0$ has been generated. Indeed, if $S_t = 0$ comes up, we just stop observing the model and calculate the likelihood, but this of course does not render the future observables Z and S undefined.

As has been explained in chapter 4, state space models permit a simple iterative computation of the likelihood through the Kalman filter. In (6.9) however, one of the equations is not a linear normal equation, and more complicated algorithms are required. Since the Bernoulli distribution belongs to the exponential family (4.18), the likelihood evaluation can be done with the iterated Kalman smoothing, outlined in chapter 4. A single likelihood evaluation thus requires to linearize the model, to run the Kalman filter to obtain the normal likelihood, and then to perform importance sampling to find the adjusting factor for the likehood (see Durbin and Koopman [49] for details). That is a substantial computational burden. Fortunately, our case is special, and some of these steps can be avoided. Conditional independence of S and Z allows us to write the likelihood function as follows:

$$P(Z,S) = \int P(Z,S|V)P(V)dV = \int P(Z|V)P(S|V)P(V)dV =$$

$$P(Z)\int P(S|V)P(V|Z)dV = P(Z)E_Z P(S|V) \tag{6.10}$$

$P(Z)$ is the likelihood of the normal component, easily obtained by the Kalman filter, applied to the state space model with the binary component S omitted. The term $E_Z P(S|V)$ could be thought of as a penalty or a correction to the normal likelihood, adjusting for the fact that in some states the model is more likely to survive than in others. It involves computation of the expected survival probability under the normal distribution of V (the one supplied by the Kalman smoother). This task may be straightforward in the cases where such expectation can be analytically evaluated. In other cases we may need to resort to numeric approximation methods such as the Monte Carlo integration. We note that even in this case there is no need to linearize the model at each likelihood evaluation.

Inference with Gompertz killing Under the Gompertz mortality (6.5), the probability $p_t(V_t)$ is

$$\log p_t(V_t) = -Ke^{Ct+\gamma' V_t} \tag{6.11}$$

In this paragraph we discuss possible ways of evaluation of the correction term in (6.10), and present the details for one of the methods, the Monte Carlo integration.

We have

$$E_Z P(S|V) = E_Z \left(\prod_{t=1}^{n-1} \exp\{-Ke^{Ct+\gamma' V_t}\} \times (1 - \exp\{-Ke^{Cn+\gamma' V_n}\}) \right) =$$

$$E_Z \left(\exp\left\{ -K \sum_{t=1}^{n-1} e^{Ct+\gamma' V_t} \right\} \times (1 - \exp\{-Ke^{Cn+\gamma' V_n}\}) \right)$$

Note that the terms for $t \leq n-1$ comprise a random variable which is a sum of dependent log-normally distributed random variables, and the expectation can be viewed (ignoring the term for $t = n$) as the moment generating function of that random variable. We found no indication in the literature of any significant progress

toward evaluating this expectation analytically, even in the simplest case of a single log-normal variate.

Furthermore, it is known that the moment generating function of a log-normal distribution cannot be evaluated as a series due to rapid growth of moments and alternating signs. A recent result of Asmussen and Rojas-Nandayapa [98] shows that the sums of dependent log-normal variables again have heavy tails of the log-normal type, prohibiting the moment or the cumulant series approach.

The convergence problem mentioned above was solved by Leipnik [94, formula 50 and 58] in the univariate case, who found a rapidly converging series for the characteristic function of the log-normal distribution. However, evaluation of the series given by Leipnik [94] is complicated by the nature of the series, where vanishing coefficients get multiplied by huge Hermite polynomials. The result is therefore very sensitive to the precision of calculations. Moreover, Leipnik's work does not cover the case of the sum of log-normally distributed variables. What we will do instead is Monte Carlo integration, the general technique recommended by Durbin and Koopman [49] to perform non-linear or non-gaussian Kalman filtering. Below we detail the computing aspects of the method.

The probability $E_Z P(S|V)$ is to be approximated by the following sample average, with a sufficiently large sample size M:

$$E_Z P(S|V) \approx \frac{1}{M} \sum_{i=1}^{M} \left[\prod_{t=1}^{n-1} \exp\{-Ke^{Ct+\gamma' V_t^{(i)}}\} \times \left(1 - \exp\left\{-Ke^{Cn+\gamma' V_n^{(i)}}\right\}\right) \right] \quad (6.12)$$

Here $V^{(i)},\ i = 1, ..., M$ are iid draws from the normal distribution $P(V|Z)$. Straightforward evaluation of (6.12) may quickly cause over or underflow of computer arithmetic operations due to large arguments supplied for the exponential functions. We therefore factor out the principal term in the Monte Carlo sum.

To begin with, we notice that the term corresponding to $t = n$ in (6.12), being a probability, belongs to the interval $(0, 1)$, and can therefore be rewritten in the form similar to the terms with $t < n$. This is achieved by putting

$$Cn + \gamma' \tilde{V}_n^{(i)} = \log \left[-\frac{1}{K} \log \left(1 - \exp \left\{ -K e^{Cn + \gamma' V_n^{(i)}} \right\} \right) \right] \tag{6.13}$$

In the sequel we assume that this transformation has been done for the last state value of each path i and omit the tilda from notation, so that all terms are in the same form for $t = 1, ..., n$. Now define

$$L_i = \sum_{t=1}^{n} \exp\{Ct + \gamma' V_t^{(i)}\}$$

so that

$$\prod_{t=1}^{n} \exp\{-K e^{Ct + \gamma' V_t^{(i)}}\} = \exp\{-K L_i\}$$

Finally we put $L = \min_i L_i$. This is the principal term in the Monte Carlo sum. After factoring it out and applying logarithm, we obtain the following approximation for the killing log-likelihood:

$$\log E_Z P(S|V) \approx -KL + \log \left[\frac{1}{M} \sum_{i=1}^{M} \exp\{-K(L_i - L)\} \right]$$

Note that each term in this sum is positive, and on the other hand, does not exceed 1. Furthermore, at least one term is exactly 1, so the average must belong to the interval $[1/M, 1]$. Thus, this expression is stable and can be evaluated without problems.

To draw $V^{(i)}$ from the smoothed density, we use the algorithm of Durbin and Koopman, presented in [50]. The authors refer to the procedure as 'simulation smoothing'. As we know, the smoothed density is normal with the mean $\hat{V} = E(V|Z)$, which is obtained by the Kalman smoothing routine. According to [50], the disturbance around this mean can be simulated as follows:

Algorithm 8 (Simulation smoothing)**.**

1. *Simulate an unconditional state path V^* together with observations Z^*, using the state space equations.*

2. *Put the simulated observations through the Kalman smoother to obtain $\hat{V}^* = E(V^*|Z^*)$.*

3. *$\hat{V} + (V^* - \hat{V}^*)$ and $\hat{V} - (V^* - \hat{V}^*)$ both have the same distribution as V given Z.*

We note that the algorithm gives two equiprobable draws from a single simulation. The authors refer to these as antithetic variables with the sample balanced for location. They argue that using both variables enhances the performance of the Monte Carlo average, because we obtain two draws at the cost of one simulation, and also because the two variables are negatively correlated. Further discussion of constructing antithetic variables is found in [49, p. 205], however we will limit our consideration to just this simplest antithetic pair.

6.4 Analysis of the fruit fly data

We are now ready to build the joint model for the fruit fly summary statistics Z_t. We restrict our attention to modeling one summary at a time, although we are well equipped to handle multidimensional observations. The reason for this is primarily the small population size, which prohibits any serious mortality modeling. More comments on this matter will be given in the conclusion chapter 8.

Specification of the model When designing the model, we understand that the following features could be present in the summary statistics Z_t:

- A deterministic time pattern, aside from the random aging effect
- A serial correlation

The first feature is accomodated by subtracting a trend from the observations, which in this study will be linear:

$$\tilde{Z}_t = Z_t - B_0 - B_1 t \tag{6.14}$$

The second feature is ensured by the $B_t V_t$ term in the observation equation (4.16). The question is, however, whether all observed autocorrelation is due to aging, or whether there is dependence in the data that is not related to aging. To explore this option, we propose various structures of the joint model. While there are many possibilities here, we only present four models, which cover the three aging dynamics introduced earlier, (6.1)-(6.3).

Full specification of the killed state space model consists of the state space matrices A_t, R_t, B_t and Σ_t, the killing parameters K, C and γ involved in (6.11), and the trend parameters B_0 and B_1, appearing in (6.14). If the hidden process V_t has an extra technical dimension, we adopt the convention that the first dimension will always be aging, and in this case $\gamma = (\gamma_1, 0)'$. The matrix Σ_t in the univariate case reduces to a scalar parameter, which we call σ^2. In addition to it, the observation variance contains the variance of stage 1 estimation, which is time dependent. Inclusion of this empirical variance into the model downweights the days when less data were available. We call this estimated variance $\hat{\sigma}_t^2$, so the overall observation variance is given by

$$\Sigma_t = \sigma^2 + \hat{\sigma}_t^2 \tag{6.15}$$

The matrices A_t, R_t and B_t will not be common for all models, so we present them for each of the aging definition.

Model	I	II	III	IV
A_t	$\begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$	(α)	$\begin{pmatrix} 1+\alpha & -\alpha \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1+\alpha & -\alpha \\ 1 & 0 \end{pmatrix}$
R_t	$\begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha^2 \end{pmatrix}$	$(1-\alpha^2)$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$
B_t	$(B_2 \quad B_3)$	(B_2)	$(B_2 \quad 0)$	$(B_2 \quad -B_2)$
Σ_t	$(\sigma^2+\hat{\sigma}_t^2)$	$(\sigma^2+\hat{\sigma}_t^2)$	$(\sigma^2+\hat{\sigma}_t^2)$	$(\sigma^2+\hat{\sigma}_t^2)$

Table 6.1: State space structure of the four models fit to the fruit fly data. The aging process is the random walk in model I, order 1 autoregression in model 2 and integrated order 1 autoregression in models 3 and 4.

In table 6.1 we present all matrices which define the linear normal state space equations. Next we explain how these models were constructed.

I. If V_t is the random walk (6.1), the observations become a random walk plus noise, according to the observation equation (4.16). Some degree of smoothness can be allowed by including a stationary component to the observation process. In the first model we realize this by the AR(1) process, the order 1 autoregression, which will be independent of the aging process. With this extra dimension, the state space matrices have the following form:

$$A_t^I = \begin{pmatrix} A_t^{BM} & 0 \\ 0 & \alpha \end{pmatrix}, \; R_t^I = \begin{pmatrix} R_t^{BM} & 0 \\ 0 & 1-\alpha^2 \end{pmatrix}, \; B_t^I = \begin{pmatrix} B_2 & B_3 \end{pmatrix}$$

II. This model assumes stationary aging (6.2). The state in this case is one dimensional, and we take $A_t^{II} = A_t^{OU}$, $R_t^{II} = R_t^{OU}$ and $B_t^{II} = (B_2)$.

III and IV. Lastly, we consider the integrated AR(1) process for aging, determined by $A_t^{III} = A_t^{IV} = A_t^{IOU}$ and $R_t^{III} = R_t^{IV} = R_t^{IOU}$. With the matrix B_t, we consider two variations. The first, version is given by $B_t^{III} = (B_2, -B_2)$. Such specification makes the observations depend on the stationary increments of the aging process. The other specification $B_t^{IV} = (B_2, 0)$ produces a model with non-stationary observations, which directly depend on aging.

Estimation of parameters In section 5 we introduced several summary statistics: the behavior fractions, the generator, and the hourly patterns. With the six behaviors and four models per each statistic, we have over 500 possible models to explore. This is obviously too many, and we therefore concentrated only on the few interesting summaries, detected upon examining the event history charts (see fig. 5.3 and 5.5). In what follows, we limit our presentation to one of the most informative statistic, the daily fraction of eating. Two observed paths of this statistic are shown in fig. 6.1.

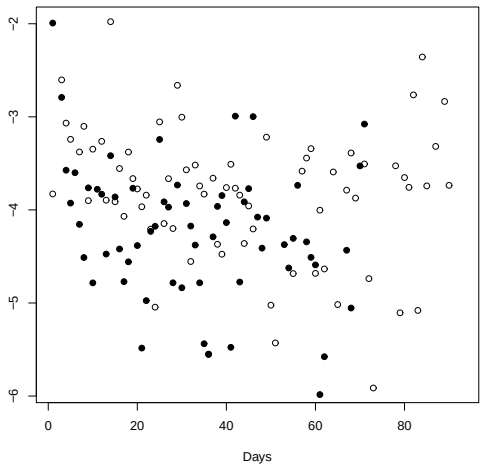

Figure 6.1: Observed daily eating frequencies (transformed with logit) for two flies. The flies lived for 75 days (filled circles) and 93 days (unfilled circles).

In chapter 5 we used $\hat{\pi}$ to denote the collection of observed fractions of time spent in each behavior, (5.10). We add the subscript t to indicate that the statistic is based on day t, and also we only keep the component of $\hat{\pi}$ which corresponds to eating. Before modeling, the logit transformation was applied to this statistic:

$$Z_t = \log \frac{\hat{\pi}_t}{1 - \hat{\pi}_t}$$

Thus, Z_t has a meaning of the log-odds of observing eating behavior at a random time between 7 am and 7 pm. Again we stress that this interpretation is not precise, unless we make certain assumtions regarding the behavior process.

Model (-2LL)	I (5796.22)		II (5813.70)		III (5826.38)		IV (5814.66)	
Param	Est	SE	Est	SE	Est	SE	Est	SE
α	0.52	0.097	0.98	0.0036	-0.21	0.13	0.98	0.0036
B_0	-3.27	0.075	-3.02	0.071	-3.01	0.070	-3.03	0.071
B_1	-0.0015	0.0020	-0.0016	0.0015	-0.0040	0.0028	-0.0012	0.0015
B_2	-0.14	0.012	-1.23	0.11	-0.25	0.035	-1.23	0.11
B_3	0.55	0.049	-	-	-	-	-	-
K	0.0013	0.00057	0.0014	0.00065	0.0012	0.00057	0.0021	0.00073
C	0.022	0.0041	0.020	0.0040	0.022	0.0045	0.016	0.0028
γ_1	0.063	0.027	0.40	0.23	0.10	0.049	0.0027	0.0027
σ^2	0.42	0.057	0.60	0.018	0.60	0.019	0.60	0.018

Table 6.2: Parameter estimates for (logit of) the daily fraction of eating: α - the autoregression coefficient in the state, B_0 and B_1 - the intercept and the slope of the observations trend, B_2 and B_3 - parameters defining the signal matrix B_t, K and C - parameters of the Gompertz law, γ_1 - the signal parameter for the frailty, σ^2 - the observation noise variance parameter

In table 6.2 we show parameter estimates for each of the four models, supplied with the standard errors. The likelihoods of the models are similar, although they are not directly comparable as the models are not nested. The deterministic trend is quite consistent across the models, and the estimate for the slope B_1 systematically turns out to be smaller than what we obtain from the weighted least squares: $B_1 = 0.0079$.

This is explained by the aging effect. Although V_t has unconditional mean zero, its mean is negative given survival up to time t, because some of the subjects with high positive values of V_t are expected to have died by t. Since in all models B_2 is negative, the mean effect of V_t on the observations is positive. This is compensated by the smaller B_1.

Of our primary interest are the parameters B_2 and γ_1. They measure the strength of the aging signal in the observations and mortality, respectively. They unfortunately are not directly comparable across the four models, because V_t has a different structure every time. It is important however that both B_2 and γ_1 are significantly different from zero, otherwise the longitudinal and the survival data have no common signal. In this regard, model IV may not be picking up this signal, as its γ_1 is within one standard error of zero. This observation will be confirmed by the graphical test we develop in chapter 7 (see fig. 7.5). The Gompertz parameters K and C in this case are very close to the maximum likelihood estimates 0.0023 and 0.016, obtained with just the survival data.

In spite of the apparent role of B_2 and γ_1 as quantities measuring the dependence between the longitudinal and survival information, it is not clear if they can be universally used to assess the predictive power of a joint model or to perform model selection. In the next chapter we develop a methodology which attempts to address both questions.

Chapter 7

Analysis of Results

In chapter 6 we described the killed state space model, which simultaneously fits the longitudinal summaries and survival data, modeling their dependence by means of a hidden Markov process. We presented estimates for one of the fly behavior summaries, obtained under several possible dynamics, and claimed that the hidden process can be interpreted as aging in the fruit fly cohort. In this chapter we formulate ideas which will help to perform the tasks a statistician usually does after fitting a model. We will address the goodness of fit assessment, testing for independence between longitudianl and survival data, and making predictions. Our approach to the goodness of fit and testing for independence will be prediction based. Therefore we begin with estimating the current hidden state and predicting the remaining lifespan.

7.1 Estimation of the state

In this section we learn how to estimate the value of the hidden process V_t, after having observed longitudinal information up to time t. In other words, we need to

evaluate the mode of the filtering distribution of V_t, given $Z_1, ..., Z_t$ and $\tau > t$.

The iterated Kalman filter When evaluating the likelihood function for the model (6.9), we exploited its special structure and managed to use the ordinary Kalman filter, in spite of the presence of the non-gaussian survival component. This saved us considerable computation time. Unfortunately, we cannot use the trick when we need the actual filtered state. Now, along with the normally distributed summaries, we have to consider the binary survival variables. These variables are non-normal, and furthermore, depend on the hidden state in a non-linear way.

To obtain the most likely state value, given the behavior summaries and the survival information, we will employ the iterated Kalman filter. Note that once the model (6.9) has been fit, we will only need to run this algorithm once for each subject to obtain the most likely state at each time point. The iterated Kalman filter has already been introduced in chapter 4. Below we specialize this algorithm for our particular case, when the observations contain the Gompertz survival indicators S_t. To do this, we need to specify the link function φ_t, observations $\tilde{Y}_t$, their signal matrix $\tilde{B}_t$ and covariance matrix $\tilde{\Sigma}_t$. The algorithm will then update the proposal mode by running the Kalman filter on the state space model (4.19).

The Bernoulli distribution belongs to the exponential family (4.18), with φ_t being the log-odds of success, $c_t = 0$, and h_t given by

$$h_t(\varphi_t) = \log(1 + e^{\varphi_t}) \tag{7.1}$$

With this h_t, we have

$$\tilde{H}_t = \nabla^2 h_t^{-1} = \frac{(1 + e^{\varphi_t})^2}{e^{\varphi_t}} \tag{7.2}$$

and

$$\tilde{H}_t \nabla h_t(\varphi_t) = 1 + e^{\varphi_t} \tag{7.3}$$

When the distribution of S_t is defined by the Gompertz law (6.11), we have

$$\varphi_t(V_t) = -Ke^{Ct+\gamma' V_t} - \log\left(1 - \exp\{-Ke^{Ct+\gamma' V_t}\}\right) \tag{7.4}$$

and thus

$$\nabla\varphi_t(V_t) = \left[-Ke^{Ct+\gamma' V_t} - \frac{K\exp\{-Ke^{Ct+\gamma' V_t}\}e^{Ct+\gamma' V_t}}{1 - \exp\{-Ke^{Ct+\gamma' V_t}\}}\right]\gamma' \tag{7.5}$$

The substitute observation for S_t then becomes

$$\tilde{S}_t = \tilde{H}_t S_t - \tilde{H}_t \nabla h_t(\varphi_t) + \nabla\varphi_t(V_t)\tilde{V}_t \tag{7.6}$$

With these calculations we can obtain the linear normal observation equation to substitute for the binary observations. There is no need to alter the equations for the summary statistics, because they are already in the required form, and are conditionally independent with the binary data. Thus, $\tilde{Y}_t$ and $\tilde{B}_t$ are obtained by stacking the corresponding quantities from the equation for Z_t and the linear equation we obtained for S_t:

$$\tilde{Y}_t = \begin{pmatrix} Z_t \\ \tilde{S}_t \end{pmatrix}, \ \tilde{B}_t = \begin{pmatrix} B_t \\ \nabla\varphi_t(V_t) \end{pmatrix} \tag{7.7}$$

Similarly, the covariance $\tilde{\Sigma}_t$ is the block matrix, formed of the corresponding covariances:

$$\tilde{\Sigma}_t = \begin{pmatrix} \Sigma_t & 0 \\ 0 & \tilde{H}_t \end{pmatrix} \tag{7.8}$$

With all these facts we can specialize the iterated Kalman filter algorithm 6 to the case of a killed state space model. The algorithm obtains the filtering distribution mode $m_{t|t-1}$ as follows:

Algorithm 9 (Iterated Kalman filter for killed state space models)**.**

1. *Obtain the initial proposal $\tilde{V}_t^0$ for the mode by running the Kalman filter on the normal component of the model (6.9), ignoring the Bernoulli trials S_t.*

2. *Iterate until convergence of $\tilde{V}_t^i$ is achieved:*

   (a) *Obtain the approximation model (4.19), using formulas (7.2)-(7.8).*

   (b) *Update the proposal mode $\tilde{V}_t^i$ by running the Kalman filter on the approximation model (4.19).*

3. *Take the last computed proposal as the estimate of the filtering distribution mode $m_{t|t-1}$.*

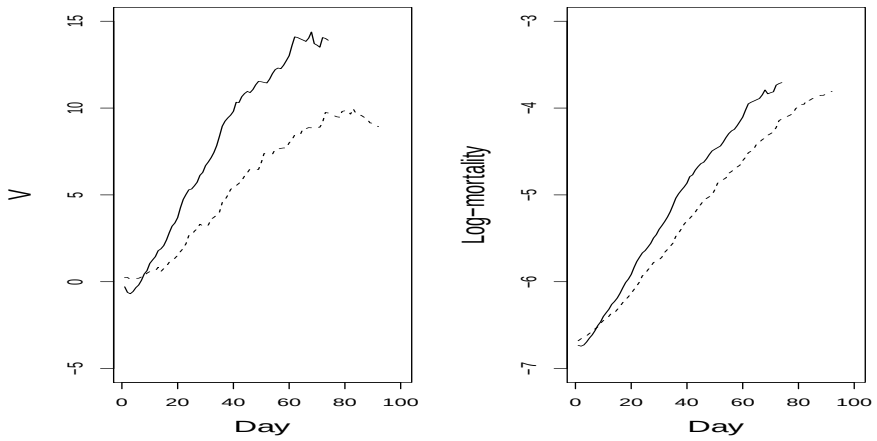

Figure 7.1: Iterated Kalman filter output for two flies. Left: the filtered mode of the aging process V_t; right: estimated logarithm of mortality. Calculations were done assuming model III.

To illustrate the work of the algorithm, we computed $m_{t|t-1}$, the estimated mode of the process V_t, given their eating history and survival up to time t. Fig. 7.1 shows $m_{t|t-1}$ (left pane) and the resulting log-mortality function (right pane) for two flies,

assuming model III. These flies died on the 75th and 93rd day; their paths are shown in solid and dashed lines, respectively. We observe that the fly which lived longer had consistently lower mortality at any fixed age t. Furthermore, their mortalities were about equal at the day of death, although one fly was 18 days older than the other. Of course not all pairs of flies exhibit this pattern perfectly. This is partly because the behavior history does not carry the pure signal, and there is uncertainty about the current state V_t. Secondly, subjects with low mortality value still have a chance to die, so occasionally we may observe flies with smaller mortality and yet shorter lifespan. However, we will establish that the pattern described is pertinent to our fruit fly cohort, and is unlikely to be due to chance.

Predicting the remaining lifespan One of the most important applications of the filtered state is inference about the subject's remaining lifespan. Our goal is to make predictions about the variable

$$\tau(t) = \tau - t \tag{7.9}$$

having observed the subject up to time t and knowing that it is still alive. This is the subject's conditional remaining lifespan. Once we know the filtering distribution $\mathcal{F}_{t+1|t}$ of the state V_t, we can easily obtain the distribution of $\tau(t)$. Indeed, we have

$$P\{\tau(t) = u | S_{\leq t}; Z_{\leq t}\} =$$

$$\int P\{S_{t+1} = 1, ..., S_{t+u-1} = 1, S_{t+u} = 0 | S_{\leq t}; Z_{\leq t}; V_{t+1}, ..., V_{t+u}\} \times$$

$$P(dV_{t+1}, ..., dV_{t+u} | S_{\leq t}; Z_{\leq t}) =$$

$$\int \left(\prod_{s=t+1}^{t+u-1} p_s(V_s) \right) (1 - p_{t+u}(V_{t+u})) P(dV_{t+1}, ..., dV_{t+u} | S_{\leq t}; Z_{\leq t}) \tag{7.10}$$

We also notice that conditioning in the state density only applies to the initial state V_{t+1}, because of the state evolution equation. So, we can write

$$P(V_{t+1}, ..., V_{t+u}|S_{\leq t}; Z_{\leq t}) = \left(\prod_{s=t+1}^{t+u} P(V_s|V_{s-1}) \right) P(V_{t+1}|S_{\leq t}; Z_{\leq t})$$

The term on the right is the filtering distribution $\mathcal{F}_{t+1|t}$, and we can see how the state equation for V_t propagates conditioning on $Z_{\leq t}$ and $S_{\leq t}$ into the future. We have shown that the conditional distribution of $\tau(t)$ is the same as has the life of the following killed state space model:

$$\begin{cases} V_{t+1} & \sim & \mathcal{F}_{t+1|t} \\ V_{t+s+1} & = & A_{t+s}V_{t+s} + \eta_{t+s}, \qquad \eta_{t+s} \sim \mathcal{N}(0, R_{t+s}) \\ S_{t+s} & \sim & \text{Bernoulli}(p_{t+s}(V_{t+s})) \end{cases} \tag{7.11}$$

where t is fixed, and the model develops in time $s = 1, 2, ...$ The parameters A_{t+s}, R_{t+s} and p_{t+s} are the same as in the original model. By 'life' of this model we mean the first time u when $S_{t+u} = 0$. Note that there is no need to consider any other observables except the survival indicators, since they don't contribute to the distribution of S_{t+s}.

While (7.11) does not provide an analytical representation of the remaining lifespan distribution, it gives a way to approximate it by a histogram. The algorithm is as follows:

Algorithm 10 (Simulating conditional remaining lifespan)**.**

1. *Draw* V_{t+1}^i *from* $\mathcal{F}_{t+1|t}$.

2. *Repeat for* $s \geq 1$ *until* $S_{t+s}^i = 0$*:*

   (a) *Draw* S_{t+s}^i *from* Bernoulli$(p_{t+s}(V_{t+s}^i))$

   (b) *Draw* V_{t+s+1}^i *given* V_{t+s}^i *using the state equation of (7.11).*

3. Let $\tau^i(t) = s - 1$

A sequence of iid $\tau^i(t)$ thus obtained represents the distribution of $\tau(t)$. It can be summarized by the mean to give a prediction of the remaining lifespan. The only difficulty with algorithm 10 is the need to draw from the filtering distribution $\mathcal{F}_{t+1|t}$. We have already remarked in chapter 4 that this task is generally not simple, and suggested the importance sampling (algorithm 7) as a possible solution. Next we specialize the algorithm 7 for the purpose of computing functionals of the remaining lifespan.

We remind that the importance sampling seeks to estimate a functional of the form

$$E_Y f(\tau(t)) = \sum_{u=0}^{\infty} f(u) P(\tau(t) = u | S_{\leq t}; Z_{\leq t}) =$$

$$\sum_{u=0}^{\infty} \int f(u) P(\tau(t) = u; dV_1, ..., dV_{t+1} | S_{\leq t}; Z_{\leq t})$$

This representation gives rise to the importance sampling estimate:

$$\hat{E}_Y \tau(t) = \frac{\sum_{i=1}^{M} \tau(t)^i w(\tau(t)^i, V^i_{\leq t+1})}{\sum_{i=1}^{M} w(\tau(t)^i, V^i_{\leq t+1})}$$

where $\tau(t)^i$ are drawn from some importance distribution Q, and $w(\cdot)$ are the importance weights:

$$w(\tau(t), V_{\leq t+1}) = \frac{P(\tau(t) = u; V_{\leq t+1}; S_{\leq t}; Z_{\leq t})}{Q(\tau(t) = u; V_{\leq t+1}; S_{\leq t}; Z_{\leq t})}$$

Because of the hidden Markov structure and conditional independence of Z and S, the numerator factors as

$$P(\tau(t) = u; V_{\leq t+1}; S_{\leq t}; Z_{\leq t}) = P(\tau(t) = u | V_{t+1}) P(Z_{\leq t} | V_{\leq t}) P(S_{\leq t} | V_{\leq t}) P(V_{\leq t+1})$$

A suitable choice of Q is the normal distribution delivered by the iterated Kalman filter, defined up to time t, and the distribution of (7.11) after time t. It factors in

the same way as the original distribution P, and is only different from it in the term $Q(S_{\leq t}|V_{\leq t})$. The importance weights thus reduce to

$$w(\tau(t), V_{\leq t+1}) = \frac{P(S_{\leq t}|V_{\leq t})}{Q(S_{\leq t}|V_{\leq t})} = \prod_{u=1}^{t} \frac{P(S_u|V_u)}{Q(S_u|V_u)} \tag{7.12}$$

$P(S_u|V_u)$ are Bernoulli probabilities defined by (6.11):

$$\log P(S_u = 1|V_u) = -Ke^{Cu+\gamma' V_u} \tag{7.13}$$

The distributions $Q(S_u|V_u)$ are univariate normal, and their parameters are obtained by the iterated Kalman filter (algorithm 9). If these parameters are $\tilde{A}_u$ and $\tilde{\Sigma}_u$, then Q has the form (up to a constant)

$$\log Q(S_u = 1|V_u) = -\frac{1}{2\tilde{\Sigma}_u}\left(1 - \tilde{A}_u V_u\right)^2 \tag{7.14}$$

We summarize these results in an algorithm for computing functionals of the remaining lifespan, given observations on the subject up to time t.

Algorithm 11 (Functionals of the conditional remaining lifespan)**.**

1. *Run the iterated Kalman filter (algorithm 9) to determine parameters of the approximating model.*

2. *For $i = 1, ..., M$ perform simulation smoothing (algorithm 8) to obtain iid paths $V^i_{\leq t+1}$, using the normal linear model obtained in step 1.*

3. *Compute the importance weight $w(V^i_{\leq t+1})$ using (7.12)-(7.14).*

4. *Simulate the remaining lifespan $\tau(t)^i$ using the state equation of (7.11) with V^i_{t+1} as the initial state (see algorithm 10).*

5. *Estimate $E_Y f(\tau(t))$ with*

$$\hat{E}_Y f(\tau(t)) = \frac{\sum_{i=1}^{M} \tau(t)^i w(V^i_{\leq t+1})}{\sum_{i=1}^{M} w(V^i_{\leq t+1})}$$

We remark that simulation of the remaining lifespan is straightforward if the state V_{t+1} is given. However, the resulting distribution needs to be averaged over all possibilities for V_{t+1}, which are distributed according to the filtering distribution $\mathcal{F}_{t+1|t}$. This averaging constitutes the major part of work in algorithm 11. Simulation of V_{t+1} from $\mathcal{F}_{t+1|t}$ is replaced by simulation from $\mathcal{N}(m_{t+1|t}, \Sigma_{t+1|t})$. The rest of the simulated path $V_{\leq t}$ is only needed to compute the importance weights.

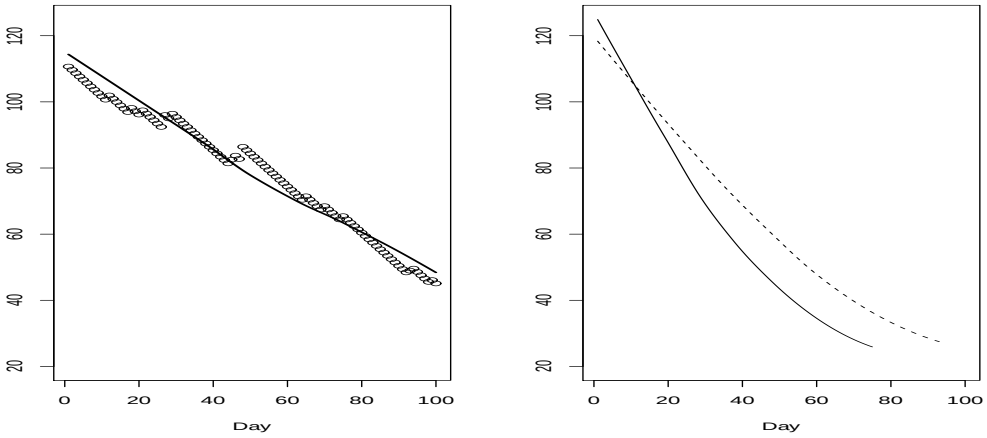

Figure 7.2: Conditional expected remaining lifespan (CERL). Left: population CERL obtained under model III (solid line), and empirical population CERL (circles); right: subject-specific CERL for two fruit flies, given their eating histories.

We conclude with an example of predicting the remaining lifespan with $E_Y\tau(t)$. We obtain the prediction for the two fruit flies, whose estimated modes $m_{t|t-1}$ were displayed in fig. 7.1. In the right pane of fig. 7.2 we show their conditional expected remaining lifespan (CERL) $\hat{E}_Y\tau(t)$. At each time t we ran algorithm 11 assuming model III. A detailed treatment of CERL will be given in the next section, where we discuss its role in the goodness of fit assessment.

7.2 Goodness of fit

In this section we address the basic question of the model diagnostics, the goodness of fit assessment. Precisely, we are interested in the capability of our models to reproduce the distribution of the survival times τ. The fit of the longitudinal data in this context plays a secondary role. Its fit may represent a separate interest to the statistician, and it can be assessed with traditional methods, but it will not be in the focus of our work. We present a new view on this topic, using conditional predictions. The ideas will be illustrated by the fruit fly example.

Conditional expected remaining lifespan Traditionally, the non-parametric approach to goodness of fit of a certain distribution is to look at the distance between the empirical and the specified proposal cumulative distribution function (CDF), or equivalently, between the respective survival curves. The Kolmogorov statistic $D = \sup \left| F - \hat{F} \right|$ is one example of such goodness of fit measure. It is not immediately clear however how D can be generalized to the case when along with survival times τ we observe covariates Z. One possibility was explored by Andrews [21], who extended the definition of the classical Kolmogorov statistic to deal with the situation when the conditional distribution $\tau|Z$ is given by the model. The joint distribution of (τ, Z) is to be tested for goodness of fit. Andrews' idea is to compare the empirical CDF of (τ, Z) against the respective semi-empirical CDF, comprised of the empirical distribution of Z and the specified distribution of $\tau|Z$. What is attractive in this approach is that the marginal distribution of Z is a nuisance parameter and does not contribute to the goodness of fit criterion. Our situation is somewhat different. Not only are our covariates time dependent, but as we collect information on Z up to time

t, we also get to know that $\tau > t$. In view of this we develop a different approach to the goodness of fit, which looks at the problem from the perspective of making conditional predictions.

We have already mentioned the concept of conditional expected remaining lifespan, or briefly CERL, and used it to predict the remaining lifespan of a subject. CERL is a function of time, where at any time t we condition τ on any information available about the subject up till this time, and compute the expectation. Depending on whether we collect the subject-specific covariates or not, CERL can be either subject specific or population specific. In the latter case conditioning is restricted to survivorship, $\tau > t$.

We begin with formulating the ideas for the population CERL. It turns out that there is more to CERL than a prediction tool: it completely determines the distribution of survival times τ, just as does the CDF or the density. Precise definition of the population CERL and statement of its properties follow.

Suppose that we observe a sequence $\tau_1, \tau_2, ..., \tau_n$ of iid survival times, or lifespans, with survival function $S(t)$. n is the cohort size. We assume in addition that the support of τ is $0 < t < t_{\max}$, where $t_{\max} \leq \infty$, and also that $S(t)$ decays fast enough so that $E\tau$ exists. CERL is the following quantity $E(t)$:

$$E(t) = E(\tau(t)|\tau > t) = -\int_0^\infty u\frac{dS(t+u)}{S(t)} = \frac{1}{S(t)}\int_0^\infty S(t+u)du \tag{7.15}$$

It is not hard to see that the function $E(t)$ defines the distribution and may be used in place of $F(t)$ or $S(t)$. Precise statement is formulated in lemma 6 of Appendix A. The empirical counterpart $\tilde{E}(t)$ is obtained by replacing $S(\cdot)$ in (7.15) with the empirical survival $\tilde{S}(\cdot)$. We note that $\tilde{E}(t)$ is defined only for $0 \leq t < \tau_{(n)}$. Consider the risk sets $R_t = \{f : \tau_f > t\}$, and let r_t be their cardinality. For $0 \leq t < \tau_{(n)}$ we

have

$$\tilde{E}(t) = \frac{\sum_f (\tau_f - t)_+}{\sum_f \mathbf{1}\{\tau_f > t\}} = \frac{\sum_{f \in R_t} (\tau_f - t)}{r_t}, \tag{7.16}$$

so $\tilde{E}(t)$ is the sample average of the remaining lifespans among the subjects still at risk at time t. If $\tau_{(n)} \leq t < t_{\max}$ then we let $\tilde{E}(t) = 0$. As one may expect, for each point t in the domain of $E(t)$ we have convergence $\tilde{E}(t) \to E(t),\ n \to \infty$ (see lemma 7 in appendix A). Population CERL is shown in the left pane of fig. 7.2. The circles are the empirical CERL for the observed survival times of the flies. The solid line gives a theoretical CERL, assuming model III (see table 6.1). Computation of the model based CERL is done with the algorithm 11, where $f(\tau(t)) = \tau(t)$, and subject-specific observations are set to missing.

The benefit of considering $E(t)$ is that it lends itself to generalization to the case of longitudinal covariate information. Suppose now that additional information is obtained on the subjects as they are observed over time: $Z_{\leq t}$. We define CERL in the same way as in (7.15), but now conditioning is going to be on both survivorship and the covariate path:

$$E_Z(t) = E(\tau - t | \tau > t; Z_{\leq t}) = \frac{1}{S(t|Z_{\leq t})} \int_0^\infty S(t+u|Z_{\leq t}) du \tag{7.17}$$

Here by $S(u|Z_{\leq t})$ we denoted the conditional probability to survive past time $u, u \geq t$, given that we have observed the covariate process up to time t.

We note that $E_Z(t)$ is a subject specific function. The right pane of fig. 7.2 presents CERL for two fruit flies, where conditioning is done on their daily fractions of eating time. Evaluation is done with the algorithm 11 under model III.

Before we proceed, we mention one property of the subject specific CERL:

$$E(E_Z(t) | \tau > t) = E(t) \tag{7.18}$$

This is a direct consequence of the law of iterated expectation, because the sigma-algebra of events $(\tau > t)$ is contained in the sigma-algebra generated by events $(\tau > t; Z_{\leq t})$. Thus, the averages of $E_Z(t)$, computed among the subjects at risk, should be close to $E(t)$.

Goodness of fit statistics As has been stated, we say that a joint model fits well if it can mimic the distribution of the survival times τ. Next we develop statistics which can be used to evaluate proximity of the empirical and model predicted distributions of τ.

The global null hypothesis we want to test is that the distribution of survival time τ is a certain given distribution ('the null distribution' or simply 'the null'). In view of lemma 6 this hypothesis can be formulated as

$$H_0 : E(t) = E_0(t),\ 0 < t < t_{\max} \tag{7.19}$$

where $E(t)$ is the CERL of τ and $E_0(t)$ defines the null distribution. Thus, goodness of fit statistics can be based on the CERL.

A crude goodness of fit assessment can be done by plotting $E(t)$ and $\tilde{E}(t)$ together against time. The left pane of fig. 7.2 shows the model based CERL (solid curve) and the empirical CERL (circles) for the 27 fruit flies, assuming model III. The model CERL was obtained for $t = 1, ..., 100$ by the algorithm 11, and smoothed by a loess smoother (Cleveland [108]) to minimize the impact of the Monte Carlo error. Visual assessment does not reveal gross deviations of the model CERL from what is observed.

The obvious disadvantage of direct comparison of $E(t)$ and $\tilde{E}(t)$ is the lack of reference for the typical magnitude of their difference $\tilde{E}(t) - E(t)$. In this regard, we

consider the following studentized difference between the empirical and model CERL:

$$\zeta(t) = \begin{cases} \frac{\tilde{E}(t)-E(t)}{(\frac{1}{r_t-1}\sum_{f\in R_t}(\tau_f - t - \tilde{E}(t))^2)^{1/2}}\sqrt{r_t}, & t < \tau_{(n-1)} \\ 0, & \tau_{(n-1)} \leq t \leq t_{\max} \end{cases} \tag{7.20}$$

If we substitute the definition of $\tilde{E}(t)$ into (7.20), we easily see that for $t < \tau_{(n-1)}$ $\zeta(t)$ is the usual one sample Student's t statistic for testing the local null hypothesis $H_0^t : E(\tau(t)) = E(t)$. It cannot however be approximated with a t-distribution with $r_t - 1$ degrees of freedom, because the sample size r_t is also a random variable. For t exceeding the second largest observed lifespan we have extended the definition of $\zeta(t)$ by zero.

The values of $\zeta(t)$ near zero indicate that the average remaining lifespan of survivors at time t is about that predicted by the model. Large positive $\zeta(t)$ speaks of the model's tendency to underestimate the longevity potential of surviving subjects, while negative $\zeta(t)$ implies overestimation. We warn however that such judgement is not reliable for all possible $0 < t < \tau_{(n)}$, because at advanced ages the risk set becomes very thin. It was shown in the proof of lemma 7 that $\tilde{E}(t)$ is biased downwards. Although the bias vanishes asymptotically for any fixed t as the number of subjects grows, it may be quite substantial for large t in a fixed population. Furthermore, it is not clear how the randomness of r_t affects the distribution of $\zeta(t)$. Simulated paths of $\zeta(t)$ display quite erratic behavior for large t (fig. 7.3). For this reason closeness of $\tilde{E}(t)$ to $E(t)$ should be evaluated on a time interval $t \in [0, t^*]$, where $t^* < \tau_{(n)}$. The choice of t^* is not obvious. If t^* is too small, we don't assess the goodness of fit in the distribution's tail. Too large t^*, on the other hand, will cover the range where $\tilde{E}(t)$ and $E(t)$ need not be close. In fig. 7.2 we restricted attention to the range from 1 to $t^* = 100$ days. With larger cohort sizes survival to advanced ages becomes more

frequent, $\tilde{E}(t)$ gains precision as an estimate of $E(t)$, and larger intervals $[0, t^*]$ can be considered. We leave it as an open question how t^* should be picked.

To this point we have not considered the covariate histories $Z_{\leq t}$. If this information is collected, the model based CERL is no longer $E(t)$. Instead, for each subject it equals $E_Z(t)$, which leads to the following amendment to the goodness of fit statistic:

$$\zeta_Z(t) = \begin{cases} \frac{\tilde{E}(t) - \frac{1}{r_t}\sum_{f \in R_t} E_Z(t)}{(\frac{1}{r_t - 1}\sum_{f \in R_t}(\tau_f - t - \tilde{E}(t))^2)^{1/2}}\sqrt{r_t}, & t < \tau_{(n-1)} \\ 0, & \tau_{(n-1)} \leq t \leq t_{\max} \end{cases} \tag{7.21}$$

Assessing the fit Traditionally, statistical tests are built on one-dimentional statistics, for which the null distribution is obtained analytically or through Monte Carlo simulations. In recent years, however, graphical assessment methods have gained popularity. Such assessments consider process-valued test statistics, which are plotted against time, and then a few paths of this process are simulated from the null distribution and plotted in the same chart. This allows the statistician to see how typical the observed process is if the null is true. It also allows spotting of particular areas or time intervals where the observed path deviates from the null. This can potentially provide a clue as to how the model might be improved. A good example of graphical testing is the series of assessment procedures developed by Lin, Wei and Ying [27] for the Cox model diagnostics. Their methodology is currently implemented in SAS proc PHREG, and offers graphical checking of the proportional hazards assumption and the functional form of covariates. Such graphical assessments can be supplemented with a global test, which is based on a certain functional, evaluated on the observed process path. Lin, Wei and Ying use the supremum functional, and

obtain its null distribution through Monte Carlo simulations. This approach will be at the core of our testing procedures. We propose two graphical goodness of fit assessments, based on the statistics $\zeta(t)$ and $\zeta_Z(t)$. For simplicity we will refer to the former as 'population zeta test' and to the latter as 'subject-specific zeta test'.

Computaton of $\zeta(t)$ is straightforward. The only quantity that requires Monte Carlo simulations is $E(t)$. $\tilde{E}(t)$ is then obtained through (7.16), and $\zeta(t)$ comes with (7.20). Simulation of the null paths requires more computation. Under the null hypothesis (7.19) our model provides the correct distribution for the survival times τ. Thus, to produce a null-distribution path, we run simulations of the killed state space model (6.9) until it generates $S_u = 0$, when we put $\tau = u$. This way we need to create the same number of simulated lifespans as we have in the original sample - this collection constitutes a single simulated sample, and then we compute $\zeta(t)$ using this artificial sample. Matching the original cohort size is necessary in order for $\zeta(t)$ to have the exact null distribution. Note that since conditioning on $Z_{\leq t}$ is not involved, simulation from the model reduces to sequentially drawing the next state V_t using the state equation, and the next survival indicator S_t given V_t.

In fig. 7.3 we show the observed statistic $\zeta(t)$ along with 10 paths, simulated under the null distribution. This has been repeated for the four aging dynamics we are considering (table 6.1). We do not observe any serious deviations of the observed curves from the typical paths generated under the null distribution. The test therefore cannot suggest which of the models should be discarded and which should be preferred. This is not surprising, given only 27 subjects in the sample. In a larger cohort, different state equations would lead to noticable differences in the lifespan, and we expect this to be visible in such plots. In addition to a few null

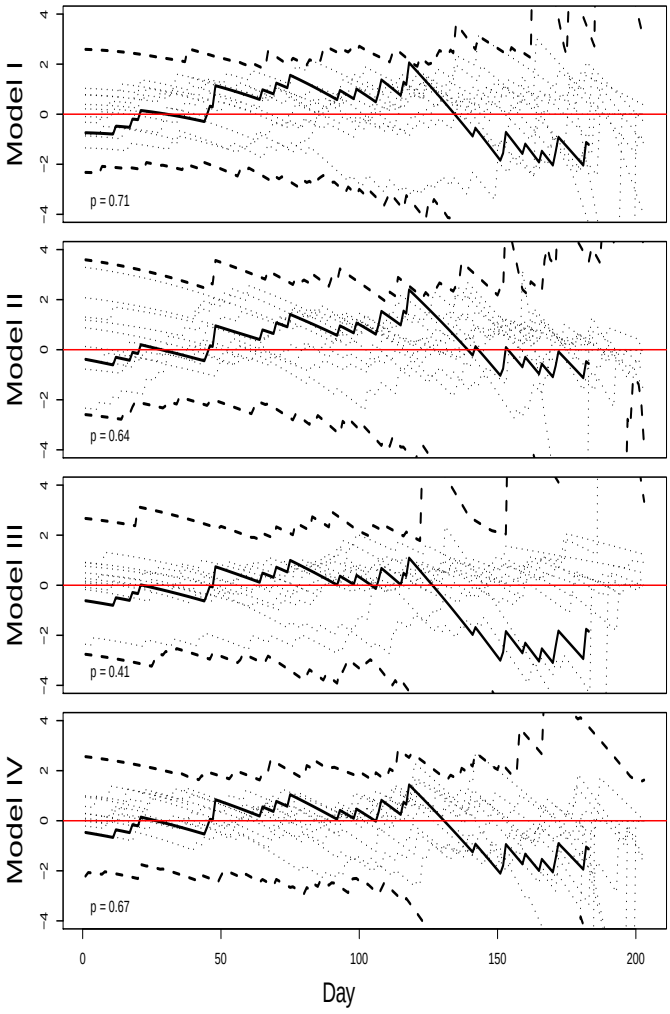

Figure 7.3: Goodness of fit based on population CERL. Solid black line is the observed statistic $\zeta(t)$. 10 paths simulated from the null are shown in dotted lines. Dashed lines provide the simulated 95% confidence band. P-values supplied are for the path supremum test. A small p-value (< 0.05) or lack of coverage of the observed path by the confidence band indicates poor fit.

paths we have computed the p-value for the supremum test:

$$P(\sup_t |\zeta(t)| > \zeta)$$

where ζ is the observed supremum. This probability was approximated by the empirical quantile of the simulated paths' suprema. Finally, we constructed a 95%

confidence band, which is expected to cover the observed path $\zeta(t)$ with 95% probabilty if the given killed state space model provides the correct survival distribution. Confidence bands were not considered in the work of Lin, Wei and Ying [27]. A confidence band can be constructed using the same Monte Carlo sample we generated to perform the supremum test. Here is the algorithm:

Algorithm 12 (Confidence band)**.**

1. *Generate M paths $\zeta^1(t), ..., \zeta^M(t)$ from the null distribution of $\zeta(t)$.*
2. *Compute $\zeta^i = \sup_t |\zeta^i(t)|$ for $i = 1, ..., M$ and order them: $\zeta^{(1)} < ... < \zeta^{(M)}$.*
3. *Discard $M\alpha$ paths with the largest supremum by considering $i \in J$,*
$$J = \{j : \ \zeta^j < \zeta^{(M\alpha)}\}$$
4. *Let $U(t) = \sup_{i \in J} \zeta^i(t)$ and $L(t) = \inf_{i \in J} \zeta^i(t)$. These are the upper and the lower boundary of a $100\alpha\%$ confidence band.*

Notice that the Kolmogorov test could have been used to test (7.19). This would still require simulation of the lifespans from the model, because we do not have the null distribution in a closed form. The advantage of our procedure is the possibility to see which particular time intervals are poorly fit. The Kolmogorov test is rather a global assessment, which only looks at the worst deviation. One could argue that the process $\tilde{S}(t) - S(t)$ (the supremum of which defines the Kolmogorov D statistic) can be considered without taking the supremum. However, this process is forced to be small for large t and might not reveal misfit in the distribution's tail. Conditioning on survival solves this problem to some extent: we can take advantage of conditioning only if there are at least a few events after t. Detailed comparison of the proposed zeta test with the Kolmogorov test is outside the scope of this study.

The more interesting benefit we get from the conditional approach is the possibility to condition on time-dependent covariates. The next goodness of fit assessment that we propose cannot be done, even globally, with the Kolmogorov type approach. The test is based on the statistic $\zeta_Z(t)$ (7.21), and includes plotting the observed statistic $\zeta_Z(t)$, a few paths of $\zeta_Z(t)$ simulated from the null, a confidence band and the supremum test.

In fig. 7.4 we show the results of the subject-specific zeta test for the four models. The time dependent covariates in this example are the daily eating frequency logits. All models except model III appear to be acceptable. For model III the observed statistic $\zeta_Z(t)$ does not lie within the simulated confidence band, and the supremum test p-value is 0.1, which is somewhat low. This suggests that model III might not be performing well. We believe the problem with model III is inappropriate distribution for the observations, because this model did not fail the population zeta test, and thus its survival distribution appears acceptable.

Next we explain the computational procedure for the subject-specific zeta test. The first difference from the population test is the quantity $E_Z(t)$, which replaces $E(t)$. Computation of $E_Z(t)$ is based on the Monte Carlo integration and is carried out according to algorithm 11. The more difficult question is how to sample from the null distribution of $\zeta_Z(t)$. To do this, we need to draw $\tau_1^i, ..., \tau_n^i$ from the appropriate distribution and use (7.21) to produce the i-th path $\zeta_Z^i(t)$. To understand what the appropriate distribution for τ is we note that

- τ is the life of the null killed state space model
- The event $\tau = t$ consists only of the following realizations of the model:

$$(S_1 = 1, Z_1), ..., (S_{t-1} = 1, Z_{t-1}), (S_t = 0, Z_t)$$

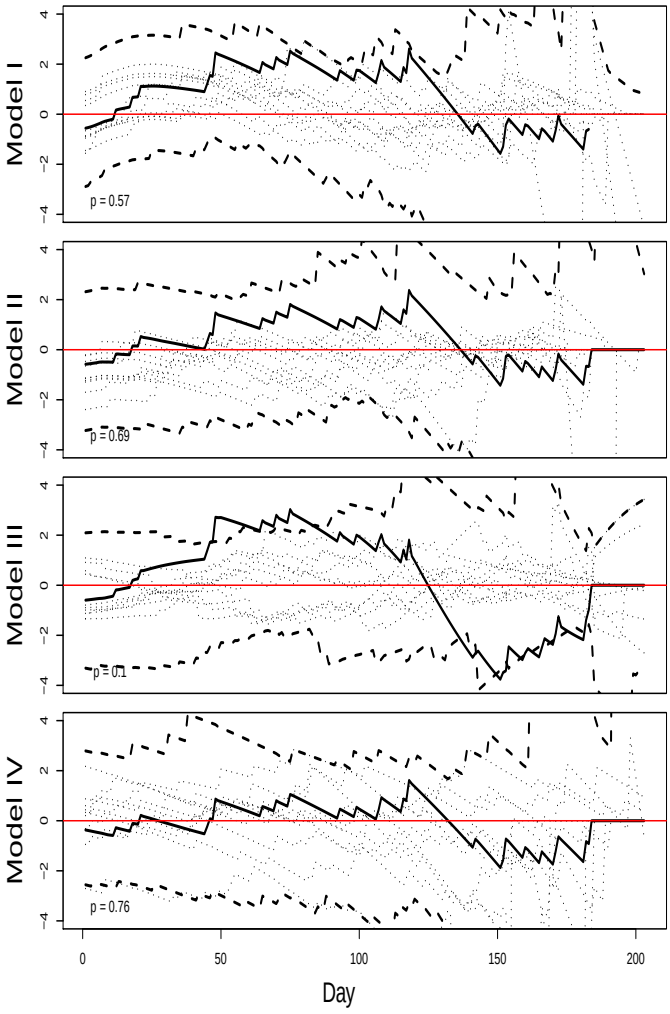

Figure 7.4: Goodness of fit based on subject-specific CERL. Solid black line is the observed statistic $\zeta_Z(t)$. 10 paths simulated from the null are shown in dotted lines. Dashed lines provide the simulated 95% confidence band. P-values supplied are for the path supremum test. A small p-value (< 0.05) or lack of coverage of the observed path by the confidence band indicates poor fit.

where $Z_1, .., Z_t$ are the fixed values of the covariate observations. If t exceeds the lifespan of the original subject, the observation sequence is extended by missing values.

In view of this drawing τ can be done sequentially. Suppose V_{t-1} and S_{t-1} have

been drawn, and $S_{t-1} = 1$. The next state should be drawn conditionally on V_{t-1} and also on the covariate history $Z_{\leq t}$ (up to time t!). Notice that since conditioning on V_{t-1} is present, conditioning on $Z_{\leq t}$ reduces to conditioning on Z_t alone. The distribution of V_t given V_{t-1} and Z_t is normal, and its parameters are obtained similarly to the Kalman filter updating (algorithm 4). Precise formulae are given in Cappe et al [83, p. 128, proposition 5.2.2], and with our notation they are:

$$E(V_t|V_{t-1}, Z_t) = A_{t-1}V_{t-1} + R_t B_t'(B_t R_t B_t' + \Sigma_t)^{-1}(Z_t - B_t A_{t-1} V_{t-1}) \tag{7.22}$$

$$\mathrm{Var}(V_t|V_{t-1}, Z_t) = R_t - R_t B_t'(B_t R_t B_t' + \Sigma_t)^{-1} B_t R_t \tag{7.23}$$

For convenience we formulate these steps in an algorithm:

Algorithm 13 (Simulating conditional survival time)**.**

1. *Repeat for* $t \geq 1$ *until* $S_t^i = 0$*:*
   (a) *Draw* V_t^i *given* V_{t-1}^i *and* Z_t *from the normal distribution with mean (7.22) and variance (7.23). If* Z_t *does not exist, draw* V_t^i *using the state equation.*
   (b) *Draw* S_t^i *from* Bernoulli$(p_t(V_t^i))$
2. *Let* $\tau^i = t$

To conclude the goodness of fit discussion we remark that the there is one more way to assess the quality of observation modeling. We can exploit the relation (7.18), which says that averaging the subject-specific CERL $E_Z(t)$ should approximate the population CERL $E(t)$. The following normalized statistic can be considered:

$$\eta_Z(t) = \frac{\frac{1}{r_t}\sum_{R_t} E_Z(t) - E(t)}{(\frac{1}{r_t - 1}\sum_{R_t}(E_Z(t) - \frac{1}{r_t}\sum_{R_t} E_Z(t))^2)^{1/2}}\sqrt{r_t} \tag{7.24}$$

The population CERL $E(t)$ does not depend on the observed values, and thus the quality of observation modeling is irrelevant to $E(t)$. On the other hand, $E_Z(t)$

depends on how the model describes the distribution of Z. If this distribution is not appropriate, the quantities $E_Z(t)$ will systematically deviate from $E(t)$, and this will be captured by the statistic $\eta_Z(t)$. We will not pursue further development of this test, but we view it as an interesting subject for future research.

7.3 Assessing independence

Finally, we address the main question of joint modeling: is there indeed any information about survival, which can be recovered by observing the longitudinal data? Does, in particular, observation of the daily eating frequency gain us knowledge about a fly's aging, and ultimately its remaining days?

A straightforward approach is to compare the fit of our model and the one with independent longitudinal and survival components. For example, for the four models considered for the fly data, we can test the null hypothesis $H_0 : (B_2 = 0 \text{ or } \gamma_1 = 0)$. Using the data in table 6.2, we conclude that model IV does not reject this hypothesis, while the rest of the models do. Such a test however compares the overall fit of both survival and longitudinal data, while our task is to improve survival fit, and the logitudinal data dynamics plays only a secondary role. We therefore will look at the fit from the survival analysis perspective.

We would like to see whether individual-specific information, accumulated during life, improves the fit to the observed survival times. In this section we look at the general problem of estimating the distribution of the survival random variable τ, given time dependent covariates.

When fitting a parametric model to data we usually perform a test whether the model provides a significantly better fit to the data than if no modeling were done.

This is often done with the likelihood ratio test, which test statistic is approximately chi-square distributed. The null hypothesis, as a rule, reduces to the statement that all (or a certain group of) parameters in the model are zero. For example, in linear regression such a test would compare the likelihood of the fit model to the likelihood of the intercept only model, other regression parameters being set to zero.

In joint analysis we may do a similar test, based on the likelihood of the model. However, it will test for the overall improvement in fitting both the survival and the covariate data. From the point of view of survival analysis, we want to focus on fitting the survival part, while fitting the longitudinal component is not of much importance. Therefore the likelihood ratio test, based on the full model likelihood, does not seem appropriate for this purpose. Indeed, restricting the hidden process to be constant over time or to be non-random would affect the likelihood of both survival and longitudinal data. The corresponding likelihood ratio test will be sensitive to misfitting the covariate data, even more so if the covariate data is as vast as in the fruit fly dataset. A more appropriate comparison would be against the model where all longitudinal data are replaced with missing, because it would preserve the mortality structure and discard any information gained from observing the covariates. However, by doing so we no longer compare nested models (i.e. the null hypothesis is not obtained by restricting parameters of the full model), so the likelihood ratio test is not valid.

To propose a meaningful alternative to the likelihood ratio approach, recall the decomposition (6.10) of the full likelihood. The survival component's contribution to the fit is the term $E_Z P(S|V)$. Its meaning is the survival probability, averaged over the filtered distribution of the state. In contrast, consider the expectation $EP(S|V)$,

where the distribution of survival S gets averaged over unfiltered hidden process paths. To assess whether filtering provides an information gain on the fit of S, it is sensible to compare $E_Z P(S|V)$ with $EP(S|V)$. Equivalently, the equality condition can be rewritten in terms of the survival time τ: for any function f such that $Ef(\tau) < \infty$,

$$E_Z f(\tau) = Ef(\tau) \tag{7.25}$$

This is the condition for no association between the covariate paths Z and survival times τ. For brevity, we call (7.25) the condition of no association. Our null hypothesis states that (7.25) is true for all f, and the alternative is that it does not hold for some function f. Of course it is impossible to embrace all functions f within a feasible testing procedure. Our strategy will be to perform testing for particular choices of f. If we reject the null under a specific choice of f, we would also have to reject (7.25). For example, take $f(x) = x$, in which case the null reduces to the equality of conditional and unconditional expectations. While such comparison may lead to significant conclusions, the test is too conservative as it replaces the equality of two distributions by equality of their expectations. We will choose f such that the test is based on more detailed information. Now we state two implications of the condition of no association, based on which our tests will be developed.

Lemma 2. *Let for all functions f such that $Ef(\tau) < \infty$, $E_Z f(\tau) = Ef(\tau)$. Then*

1. *for all $t > 0$, $E(\tau | Z_{\leq t}; \tau > t) = E(\tau | \tau > t)$*
2. *for all $t > 0$, $\mathrm{Cov}(\tau, E_{Z_{\leq t}} V_t | \tau > t) = 0$*

For proof we refer to Appendix A. Let us examine the statements of the lemma. The first one wouldn't change if we subtracted t from both sides of the equation. The resulting statement means that the population CERL is the same as the subject-specific CERL, where we also condition on the observed covariate history $Z_{\leq t}$.

The second statement of the lemma says that the filtered state V_t, which is $E_{Z_{\leq t}}V_t$, is uncorrelated with the survival time. Once again, our statements are prediction based. We will concentrate on the correlation property, and in the next paragraph explain how the test statistic is constructed. As before, we propose a graphical assessment against the null distribution.

Correlation test The test we propose to assess whether the covariates Z are informative of survival times τ will be based on the second statement of lemma 2. For each $t > 0$, define conditional correlations

$$\rho_t = \mathrm{Cor}(\tau, E_{Z_{\leq t}}V_t | \tau > t) \tag{7.26}$$

According to lemma 2, the null hypothesis should be formulated as

$$H_0 : \rho_t = 0,\ t > 0$$

Any estimate of ρ_t can serve as a test statistic. One possibility is to compute sample correlations between observed τ and filtered states V_t. However, we propose to use the Spearman rank correlations instead, and denote them $\hat{\rho}_t$. Correlations based on ranks are usually recommended for skewed data, and are robust to outliers. Essentially we will see if the order of filtered states $E_{Z_{\leq t}}V_t$ is consistent with the order of lifespans of subjects at risk. Remarkably, we can replace τ with the remaining lifespans $\tau(t)$, which is more easily interpreted. This will not change the Spearman statistics, because the order is preserved.

Now we will describe how the graphical assessment can be realized for our test statistic $\hat{\rho}_t$. Under the condition of no association, we can simulate the covariate histories Z independently of the survival times τ. For example, this could be achieved by letting Z and τ be generated by independent realizations of the state V. However,

the resulting $\hat{\rho}_t$ paths will deviate from zero not only because of randomness in τ, but also because of randomness in Z. This is undesirable, since we stated that the distribution of Z is a nuisance rather than an object of modeling. Apparently, allowing Z to vary will waste power of the testing procedure. We therefore suggest holding Z fixed and only draw survival times τ. Here is the algorithm:

Algorithm 14 (Simulation under the condition of no association)**.**

1. *Let* $V_0 = 0,\ S_0 = 1$ *and repeat until* $S_t = 0$*:*
   (a) *Draw* V_t *given* V_{t-1}*, using the fitted state evolution equation*
   (b) *Draw* S_t *given* V_t*, according to the probability* $p_t(V_t)$
2. *Set* $\tau^{(i)} = t$
3. *Define* $Z_t^{(i)} = Z_t$ *for* $t \leq \min(\tau, \tau^{(i)})$*. If* $\tau^{(i)} > \tau$ *then define* $Z_t^{(i)}$ *as missing for* $t > \tau$

Again, some clarification is necessary here. The data Z were observed only up to the time of death τ of the corresponding subject, and does not exist past this time. How can we then draw a random time $\tau^{(i)}$ and associate it with the history Z? The answer is simple, and is given by the state space model we use. There is no structural requirement preventing the process Z_t from being observed after the first occurence of $S_t = 0$. Likewise, there is no restriction for S_t to continue generating if the observations Z_t stop arriving. Most importantly, the distribution of V_t given $Z_{\leq t}^{(i)}$ coincides with the distribution of V_t given $Z_{\leq t}$. In particular, $E(V_t|Z_{\leq t}^{(i)}) = E(V_t|Z_{\leq t})$, $t > 0$. This follows from the Kalman filter algorithm.

In application to the process $\hat{\rho}_t$, the algorithm can be formulated in terms of the filtered state $E(V_t|Z_{\leq t}^{(i)})$ instead of the full histories of observations $Z^{(i)}$. Indeed, the

quantity $E(V_t|Z^{(i)}_{\leq t})$ is equal to $E(V_t|Z_{\leq \min(t,\tau)})$. These expectations are derived by the Kalman filter for $t \leq \tau$. For $t > \tau$ we simply need to put $E(V_t|Z_{\leq \tau})$ through the state evolution equation $t - \tau$ times, [49, p.92]. Thus, only one run of the Kalman filter per subject is necessary to compute all simulated paths $\hat{\rho}_t{}^i$. To illustrate, we present the graphical diagnostics described for the eating frequency logits in figure 7.5.

We plotted the observed Spearman correlation process $\hat{\rho}_t$, obtained with each of the four models. 10 paths simulated from the null distribution according to the proposed algorithm are also shown. Under models I through III we observe that the correlation is close to zero in the beginning, attesting to the fact that we haven't accumulated any information yet. Then it steadily goes into the negative side, and is strongest between day 50 and 70. We also note that in all models the observations have a negative signal coefficient B_2. This means that until about day 70, the flies which were more frequently observed eating were less frail and lived longer lives. After day 70 situation changes to the opposite: the flies with the least frequent eating have died, and among survivors the pattern is different. With the data at hand, simulated paths show that this reverting pattern might be due to chance alone. This happens because the risk set thins out considerably, and the confidence band covers the entire range $[-1, 1]$. Larger populations are necessary to make solid statements about the late-life patterns. If the correlation reverting pattern is confirmed by fresh data, the model may need refinement, because current specification implies a steady mortality increase for low eating frequency logits.

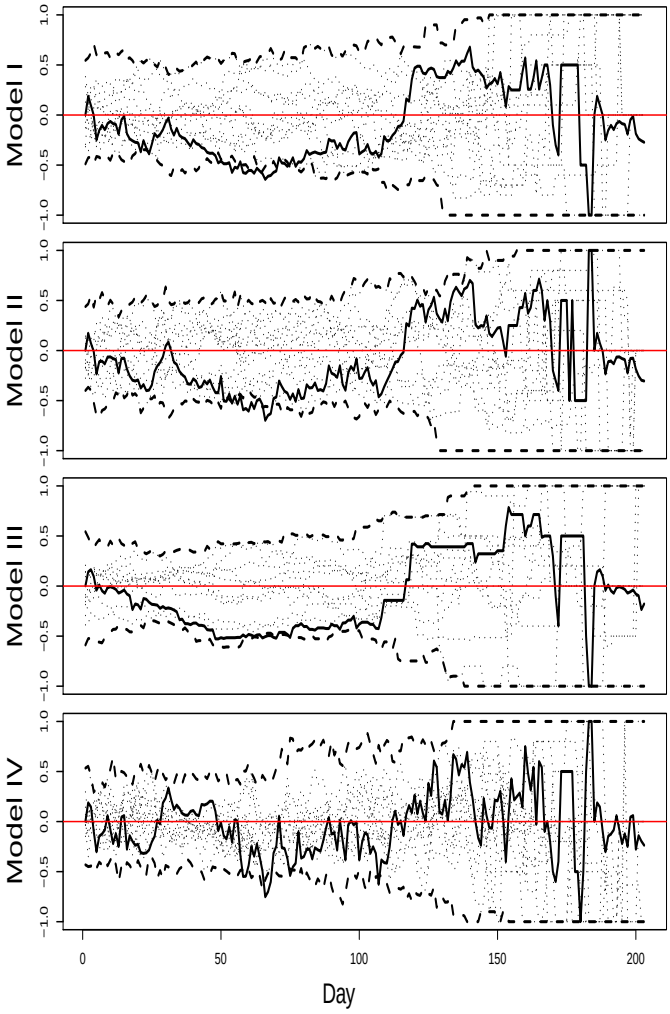

Figure 7.5: Correlation test. Solid black line is the observed statistic $\hat{\rho}_t$. 10 paths simulated under the condition of no association are shown in dotted lines. Dashed lines provide the simulated 95% confidence band. Lack of coverage of the observed path by the confidence band indicates presence of association between longitudinal and survival data.

Chapter 8

Conclusion

We conclude with a summary of the key points we made in this work. The original intent of the present research was to develop an approach that would answer the question: can longitudinal data, collected during life, help understand the aging process. In particular, the goal was to investigate this possibility with the behavioral records of a cohort of 27 Mexican fruit flies. Survival time was chosen as the observable endpoint for the subject's aging state.

Extensive literature study revealed that there are relatively few works in the field of aging, which utilized longitudinal data. On the other hand, there is a significant body of work done in applied clinical studies, where longitudinal observations were used as covariates in the survival analysis. In these studies, a number of models have been introduced to model longitudinal and survival data jointly. The term 'joint model' has become common to describe this kind of analysis. In spite of the considerable attention and growing data collection possibilities, there appears to be no well established statistical framework for the joint models. In particular, major statistical software packages do not offer tools for joint analysis. Thus, we position

the results of our research as a comprehensive methodology for joint modeling, be it in the context of survival analysis or an aging study.

The methodology developed in this work consists of two stages. The purpose of the first stage is to perform data reduction. This step may be necessary for two reasons. First, to smooth out local patterns in the longitudinal data that are not pertinent to aging. The second reason is to make the data computationally manageable. With only several time points of data collection, this stage can be skipped. In other applications with extensive data collection, such as monitoring the fruit fly behaviors, data reduction can hardly be avoided. We came to the conclusion that the summary statistics, with which summarization is done, need not necessarily estimate some meaningful characteristic (although of course it is convenient if they do). What is important is that these summaries preserve the signal of aging, if it is present in the original data. We also argued that since we possess the richer, unsummarized data, the quality of data reduction may be assessed and measured. We proposed an insufficiency measure for this purpose.

The second stage of analysis consists of state space modeling, which is very broadly used in engineering, physics and financial applications, but to the best of our knowledge has not been applied in aging research. We adapted the famous Kalman filter to handle the survival data. Such an approach provides a very general and at the same time simple means to perform joint modeling.

With our model, the researcher can focus on designing the dynamics of the state. There is a possibility to assess various model assumptions and choose between several competing models. We developed a series of graphical assessment tools for this purpose. All our tests are constructed from the prediction perspective, putting emphasis

on the survival times and giving the longitudinal data a secondary role.

For the fruit fly dataset, we considered many summary statistics, and used the event history charts to pick the most interesting patterns. The daily eating frequency statistic was chosen for further exploration. We performed the second stage of analysis with four candidate models, each having different mechanisms of aging and distribution of the longitudinal observations. We concluded that there can possibly be relationship between the fly's aging and its eating patterns, although the sample size is too small to make strong statements. Because of the same reason we found no evidence as to which of the four models should be favored. The cohort size is a big limitation of the fruit fly data. With a richer survival information it would also be of interest to model multiple summary statistics simultaneously, a task our model permits with no difficulty.

Finally, we determine several directions for the future research of joint modeling.

1. We have introduced the concept of approximate sufficiency, but have not performed actual calculations for the fruit fly data. It is of interest to see how much we lose by treating the blocks of uninterrupted observations as independent paths of a Markov process. The comparison could be done against an interval-censored Markov process or a time-inhomogeneous Markov process.

2. Diagnostics

   (a) Of primary interest is further development of the tests for goodness of fit and longitudinal-survival association. Extensive research is required to determine power of the supremum tests. The ability of the graphical tests and supremum tests to discriminate between the models should be investigated.

(b) For all graphical tests considered we have observed that the simulated paths fluctuate wildly at advanced ages. This is likely the result of inappropriate normalization of the test statistics. It is desirable to improve upon this issue and have the confidence band roughly parallel to the x axis. Such a property would justify the use of the supremum test, and avoid the necessity to cut off the region of high age (t^*).

(c) The correlation test proposed does not tell us how much of the variance in the remaining lifespan is explained by the joint model. Development of a statistic similar to the regression R^2 would be an important addition to the qualitative model assessment we have done.

3. Censoring has not been considered in this study, although almost any survival analysis includes this possibility. Likelihood evaluation and parameter estimation is not complicated by censoring. On the other hand, it is not immediately clear how the graphical tests will handle censored longitudinal paths, because they rely on the observed remaining lifespan. Should this difficulty be overcome, the tests could be performed online, while the cohort is still being followed.

4. More efforts need to be directed towards improving the software the author has developed for fitting the fly data. It is desirable to prepare an R package that would perform custom joint modeling and offer the graphical diagnostics discussed in this work.

Bibliography

[1] Albert A. Estimating the infinitesimal generator of a continuous time finite state Markov process. *Annals of mathematical statistics*, 38:727–753, 1962.

[2] Brooks A., Lithgow G., and Johnson T. Mortality rates in a genetically heterogeneous population of Caenorhabditis Elegans. *Science*, 263:668–671, 1994.

[3] Comtet A., Monthus C., and Yor M. Exponential functionals of Brownian motion and disordered systems. *Journal of applied probability*, 35:255–271, 1998.

[4] Dempster A., Laird N., and Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[5] Hirsh A., Williams R., and Mehl P. Kinetics of medfly mortality. *Experimental Gerontology*, 29(2):197–204, 1994.

[6] Kowald A., Kirkwood T., Robine J., Ritchie K., Vaupel J., Carey J., and Curtsinger J. Explaining fruit fly mortality. *Science*, 260:1664–1666, 1993.

[7] Yashin A. Filtering of jump processes. *Automation and remote control*, 31:725–730, 1970.

[8] Yashin A., Iachine I., and Begun A. Mortality modeling: a review. *Mathematical Population Studies*, 8(4):305–322, 2000.

[9] Yashin A., Vaupel J., and Iachine I. A duality of aging: the equivalence of mortality models based on radically different concepts. *Mechanisms of aging and development*, 74:1–14, 1994.

[10] Yashin A., Arbeev K., Akushevich I., Kulminski A., Akushevich L., and Ukraintseva S. Stochastic model for analysis of longitudinal data on aging and mortality. *Mathematical Biosciences*, 208:538–551, 2007.

[11] Yashin A. and Manton K. Effects of unobserved and partially observed covariate processes on system failure: A review of models and estimation strategies. *Statistical Science*, 12(1):20–34, 1997.

[12] Gompertz B. On the nature of the function expressive of the law of human mortality, and on the new model of determining the value of life contingencies. *Philosophical Transactions Royal Society, A*, 115:513–580, 1825.

[13] Healy B. and DeGruttola V. Hidden Markov models for settings with interval-censored transition times and uncertain time origin: application to HIV genetic analyses. *Biostatistics*, 8(2):438–452, 2007.

[14] Strehler B. and Mildvan A. General theory of mortality and aging. *Science*, 132(3418):14–21, 1960.

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

[16] Finch C., Pike M., and Witten M. Slow mortality rate accelerations during aging in some animals approximate that of humans. *Science*, 249:902–905, 1990.

[17] Kim C. and Nelson C. *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT Press, 1999.

[18] McGilchrist C. and Aisbett C. Regression with frailty in survival analysis. *Biometrics*, 47(2):461–466, 1991.

[19] Rutter C. and Elashoff R. Analysis of longitudinal data: Random coefficient regression modelling. *Statistics in Medicine*, 13:1211–1231, 1994.

[20] Wang C. Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*, 16:235–253, 2006.

[21] Andrews D. A conditional Kolmogorov test. *Econometrica*, 65(5):1097–1128, 1997.

[22] Cox D. *The analysis of binary data*. Methuen, London, 1970.

[23] Cox D. Regression models and life tables. *Journal of Royal Statistical Society, Series B*, 34:187–202, 1972.

[24] Cox D. and Oakes D. *Analysis of survival data*. CRC Press, 1984.

[25] Dufresne D. The integral of geometric Brownian motion. *Advances of applied probability*, 33:223–241, 2001.

[26] Gross D. and Harris C. *Fundamentals of queueing theory*. Wiley, 1998.

[27] Lin D., Wei L., and Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572, 1993.

[28] Promislow D. Senescence in natural populations of mammals: comparative study. *Evolution*, 45:1869–1887, 1991.

[29] Steinsaltz D. Reevaluating a test of the heterogeneity explanation for mortality plateaus. *Experimental Gerontology*, 40:101–113, 2005.

[30] Steinsaltz D. and Wachter K. Understanding mortality rate deceleration and heterogeneity. *Mathematical Population Studies*, 13:19–37, 2006.

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

[32] Zeng D. and Cai J. Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis*, 1:151–174, 2005.

[33] Brown E., Ibrahim J., and DeGruttola V. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61:64–73, 2005.

[34] Hsieh F., Tseng Y., and Wang J. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62:1037–1043, 2006.

[35] Jelinek F. *Statistical models for speech recognition.* MIT Press, 1997.

[36] Rossolini G. and Piantanelli L. Mathematical modeling of the aging processes and the mechanisms of mortality: paramount role of heterogeneity. *Experimental Gerontology*, 36:1277–1288, 2001.

[37] Sacher G. and Trucco E. The stochastic theory of mortality. *Annals of New York Academy of Sciences*, 96:985–1007, 1962.

[38] Williams G. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 11:398–411, 1957.

[39] Bunke H. and Caelli T. *Hidden Markov models: applications in computer vision*. World Scientific, 2001.

[40] Le Bras H. Lois de mortalité et âge limite. *Population*, 31:655–692, 1976.

[41] Matsumoto H. and Yor M. Exponential functionals of Brownian motion, I: Probability laws at fixed time. *Probability Surveys*, 2:312–347, 2005.

[42] Simms H. Logarithmic increase in mortality as a manifestation of aging. *Journal of Gerontology*, 1:13–25, 1945.

[43] Akushevich I., Kulminski A., and Manton K. Life tables with covariates: Dynamic model for nonlinear analysis of longitudinal data. *Mathematical Population Studies*, 12:51–80, 2005.

[44] Moustaki I. and Steele F. Latent variable models for mixed categorical and survival responses, with an application to fertility preferences and family planning in Bangladesh. *Statistical Modelling*, 5:327–342, 2005.

[45] Abril J. On the concept of approximate sufficiency. *Pakistan Journal of Statistics*, 10:171–177, 1994.

[46] Carey J., Papadopoulos N., Kouloussis N., Katsoyannos B., Muller H.-G., Wang J.-L., and Tseng Y.-K. Age-specific and lifetime behavior patterns in Drosophila

Melanogaster and the Mediterranean fruit fly, Ceratitis Capitata. *Experimental Gerontology*, 41:93–97, 2006.

[47] Carey J., Liedo P., Orozco D., and Vaupel J. Slowing of mortality rates at older ages in large medfly cohorts. *Science*, 258:457–461, 1992.

[48] Carey J., Liedo P., Muller H.-G., Wang J.-L., and Vaupel J. A simple graphical technique for displaying individual fertility data and cohort survival: case study of 1000 mediterranean fruit fly females. *Functional Ecology*, 12:359–363, 1998.

[49] Durbin J. and Koopman S. *Time series analysis by state space methods*. Oxford University Press, 2001.

[50] Durbin J. and Koopman S. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–615, 2002.

[51] Hamilton J. and Raj B. *Advances in Markov-switching models: applications in business cycle research and finance*. Springer, 2003.

[52] Riggs J. and Millecchia R. Mortality among the elderly in the US, 1956-1987: demonstration of the upper boundary to Gompertzian mortality. *Mechanisms of ageing and development*, 62:191–199, 1992.

[53] Vaupel J., Carey J., Christensen K., Johnson T., Yashin A., Holm N., Iachine I., Kannisto V., Khazaeli A., Leido P., Longo V., Zeng Y., Manton K., and Curtsinger J. Biodemograpic trajectories of longevity. *Science*, 280(5365):855–860, 1998.

[54] Vaupel J., Manton K., and Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

[55] Weitz J. and Fraser H. Explaining mortality rate plateaus. *Proceedings of the National Academy of Sciences*, 98:15383–15386, 2001.

[56] Xu J. and Zeger S. The evaluation of multiple surrogate endpoints. *Biometrics*, 57(1):81–87, 2001.

[57] Manton K., Stallard E., and Vaupel J. Alternative models for heterogeneity in mortality risks among the aged. *The Journal of American Statistical Association*, 81(395):635–644, 1986.

[58] Baum L., Petrie T., Soules G., and Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[59] Duchateau L. and Janssen P. *The frailty model*. Springer, 2008.

[60] Gavrilov L. and Gavrilova N. *The biology of lifespan: a quantitative approach*. Harwood Academic Publishers, 1991.

[61] Gavrilov L. and Gavrilova N. The reliability theory of aging and longevity. *Journal of Theoretical Biology*, 213(4):527–545, 2001.

[62] Gavrilov L. and Gavrilova N. The reliability-engineering approach to the problem of biological aging. *Annals of the New York Academy of Sciences*, 1019:509–512, 2004.

[63] Mueller L. and Rose M. Evolutionary theory predicts late life mortality plateaus. *Proceedings of the National Academy of Science, USA*, 93:15249–15253, 1996.

[64] Mueller L., Nusbaum T., and Rose M. The Gompertz equation as a predictive tool in demography. *Experimental Gerontology*, 30(6):553–569, 1995.

[65] Rabiner L. and Juang B. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[66] Rogers L. and Williams D. *Diffusions, Markov processes and martingales*. Cambridge University Press, 2001.

[67] Abramowitz M. and Stegun I. *Modified Bessel Functions I and K. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1972.

[68] Chen M., Ibrahim J., and Sinha D. A new joint model for longitudinal and survival data with a cure fraction. *Journal of Multivariate Analysis*, 91:18–34, 2004.

[69] Crowder M., Kimber A., Smith R., and Sweeting T. *Statistical analysis of reliability data*. Chapman and Hall, London, 1991.

[70] Drapeau M., Gass E., Simison M., Mueller L., and Rose M. Testing the heterogeneity theory of late-life mortality plateaus by using cohorts of Drosophila Melanogaster. *Experimental Gerontology*, 35:71–84, 2000.

[71] Garg M., Rao R., and Redmond C. Maximum likelihood estimation of the parameters of the Gompertz survival function. *Applied Statistics*, 19(2):152–159, 1970.

[72] Kharrati-Kopaei M., Nematollahi A., and Shishebor Z. On the sufficient statistics for multivariate ARMA models: approximate approach. *Statistical Papers*, 50:261–276, 2007.

[73] Mangel M. Environment and longevity: emergence without interaction, multiple steady states and stochastic clock. *Evolutionary ecology research*, 4:1065–1074, 2002.

[74] Sherman M. Comparing the sample mean and the sample median: an exploration in the exponential power family. *The American Statistician*, 51(1):52–54, 1997.

[75] Witten M. and Satzer W. Gompertz survival model parameters: estimation and sensitivity. *Applied mathematics letters*, 5:7–12, 1992.

[76] Woodbury M. and Manton K. A random walk model for human mortality and aging. *Theoretical Population Biology*, 11:37–48, 1977.

[77] Woodbury M., Manton K., and Stallard E. Longitudinal models for chronic disease risk: an evaluation of logistic multiple regression and alternatives. *International Journal of Epidemiology*, 10:187–197, 1981.

[78] Yor M. On some exponential functionals of Brownian motion. *Advances of applied probability*, 24:509–531, 1992.

[79] Laird N. and Ware J. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[80] Papadopoulos N., Carey J., Katsoyannos B., Kouloussis N., Muller H.-G., and Liu X. Supine behaviour predicts the time to death in male Mediterranean

fruit flies (Ceratitis Capitata). *Proceedings of Royal Society, Series B, UK*, 269(1501):1633–1637, 2002.

[81] Shephard N. *Stochastic volatility : selected readings*. Oxford University Press, 2005.

[82] Aalen O. Phase type distribution in survival analysis. *Scandinavian journal of statistics*, 22(4):447–464, 1995.

[83] Cappe O., Moulines E., and Ryden T. *Inference in hidden Markov models*. Springer Science+Business Media, Inc, 2005.

[84] Abrams P. and Ludwig D. Optimality theory, Gompertz law, and the disposable soma theory of senescence. *Evolution*, 49(6):1055–1066, 1995.

[85] Billingsley P. *Statistical inference for Markov processes*. The University of Chicago Press, 1961.

[86] Bremaud P. *Markov chains. Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[87] Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:671–678, 1986.

[88] Hougaard P. Multi-state models: a review. *Lifetime Data Analysis*, 5(3):239–264, 1999.

[89] Medawar P. *An Unsolved Problem of Biology*. H.K. Lewis, London, 1952.

[90] Metzner P., Dittmer E., Jahnke T., and Schutte C. Generator estimation of Markov jump processes. *Journal of Computational Physics*, 227:353–375, 2007.

[91] Service P. Heterogeneity in individual mortality risk and its importance for evolutionary studies of senescence. *The American Naturalist*, 156(1):1–13, 2000.

[92] Durbin R., Eddy S., Krogh A., and Mitchison G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[93] Horn R. and Johnson C. *Matrix Analysis*. Cambridge University Press, 1985.

[94] Leipnik R. On lognormal random variables: I-the characteristic function. *The Journal of the Australian Mathematical Society, Series B*, 32:327–347, 1991.

[95] Littell R., Milliken G., Stroup W., Wolfinger R., and Schabenberger O. *SAS for Mixed Models, Second Edition*. Cary, NC: SAS Institute Inc, 2006.

[96] MacKay R. Estimating the order of a hidden Markov model. *The Canadian Journal of Statistics*, 30(4):573–589, 2002.

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

[98] Asmussen S. and Rojas-Nandayapa L. Asymptotics of sums of lognormal random variables with gaussian copula. *Statistics and Probability Letters*, 78(16):2709–2714, 2008.

[99] Bose S. *An introduction to queueing systems*. Kluwer/Plenum Publishers, 2002.

[100] Hernandez-Rodriguez S., Altamirano-Robles L., Carey J., and Liedo P. Automatization of the Mexican fruit fly activity recognition process using 3d and

gray-level features. *Proceedings of the IASTED International Conference 'Advances in Computer Science and Technology'*, 2006.

[101] Kullback S. and Leibler R. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[102] Ratcliffe S., Guo W., and Ten Have T. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, 60:892–899, 2004.

[103] Aven T. and Jensen U. *Stochastic Models in Reliability*. Springer-Verlag, New York, 1999.

[104] Johnson T. Age-1 mutants of Caenorhabditis Elegans prolong life by modifying the Gompertz rate of aging. *Science*, 249:908–911, 1990.

[105] Kirkwood T. Evolution of aging. *Nature*, 270:301–304, 1977.

[106] Koski T. *Hidden Markov models for bioinformatics*. Kluwer, 2001.

[107] De Gruttola V. and Tu X. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

[108] Cleveland W. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[109] Song X., Davidian M., and Tsiatis A. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4):742–753, 2002.

[110] Chi Y. and Ibrahim J. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62:432–445, 2006.

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

[112] Shao Z., Gao W., Yao Y., and Zhuo Y.and Riggs J. The dynamics of aging and mortality in the People's Republic of China. *Applied Mathematics Letters*, 67:239–246, 1993.

Appendix A

Proofs

Insufficiency of the sample median for the location parameter of a normal distribution After defining insufficiency of a statistic in chapter 5, we gave an example of estimating the location parameter μ of a normal distribution with the sample median, based on n independent observations $X_1, ..., X_n$. We claimed that $\kappa(\tilde{X}) \leq 0.0342$. Below we prove this claim.

Proof. According to our assumption, the true distribution of $X = (X_1, ..., X_n)$ has the following density:

$$p_\mu^n(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

Let the model distribution Q_μ be the distribution of n iid Laplace random variables with location μ and scale $b > 0$. Its density is given by

$$q_\mu^n(x_1, ..., x_n) = \frac{1}{(2b)^n} \exp\left\{-\frac{1}{b}\sum_{i=1}^{n}|x_i - \mu|\right\}$$

Since the sample median is sufficient for such experiment, substitution of q_μ into (5.4) is legitimate and gives the following inequality:

$$\kappa(\tilde{X}) \leq \sup_{\mu} d(\mathcal{P}_\mu^n || \mathcal{Q}_\mu^n)$$

Because of additivity of both the entropy and the relative entropy in the case of independent observations, and of identical marginal distributions, $d(\mathcal{P}_\mu^n || \mathcal{Q}_\mu^n)$ reduces to $d(\mathcal{P}_\mu || \mathcal{Q}_\mu)$. Thus we only need to consider univariate normal and Laplace distributions. For these distributions we have

$$d_{KL}(\mathcal{P}_\mu || \mathcal{Q}_\mu) = E_{P_\mu}\left(-\log\sqrt{2\pi} - \frac{1}{2}(X_1-\mu)^2 + \log(2b) + \frac{1}{b}|X_1-\mu|\right)$$

The square term evaluates to the variance, which is 1, and the absolute value term gives the first absolute moment, which is $\sqrt{2/\pi}$. Furthermore, minimizing with respect to b we find that we should pick $b = \sqrt{2/\pi}$. Thus the relative entropy evaluates to

$$d_{KL}(\mathcal{P}_\mu || \mathcal{Q}_\mu) = \frac{1}{2} + \log(2/\pi)$$

The entropy of the normal distribution straightforwardly evaluates to

$$H(\mathcal{P}_\mu) = \log\sqrt{2\pi} + \frac{1}{2}$$

Dividing the two expressions, we obtain

$$d(\mathcal{P}_\mu || \mathcal{Q}_\mu) = \frac{\frac{1}{2} + \log(2/\pi)}{\log\sqrt{2\pi} + \frac{1}{2}} = 0.03412...$$

which is rounded upwards to 0.0342 to maintain the upper bound. Since there is no dependence on μ, we also have $\kappa(\tilde{X}) \leq 0.0342$, and the proof is completed.

□

Sufficient statistics for hidden Markov models In chapter 5 we stated lemma 1, where sufficient statistics were constructed for a hidden Markov model. Below is the proof of this result.

Proof. Because of the hidden Markov structure, the joint distribution of observations can be represented as

$$P_\theta\{Y_1 = y_1, ..., Y_n = y_n\} = \int \prod_{t=1}^{n} g_{\lambda_t}(y_t) P_\theta\{V_1 \in dv_1, ..., V_n \in dv_n\}$$

Dependence of the integrand on the state is hidden in the parameter values λ_t, which are assumed to be functions of v_t. These quantities also depend on the model's parameter θ, as does the distribution of the state.

Now we recall the factorization theorem, which gives an equivalent condition of sufficiency. Thus, for the statistic Z_t it means that

$$g_{\lambda_t}(y_t) = f(\lambda_t, Z_t) h(y_t)$$

where h is a function which is independent of λ_t, and therefore also of v_t and θ. Plugging this decomposition into the joint distribution expression, we arrive at

$$\begin{aligned} P_\theta\{Y_1 = y_1, ..., Y_n = y_n\} &= \int \prod_{t=1}^{n} f(\lambda_t, Z_t) h(y_t) P_\theta\{V_1 \in dv_1, ..., V_n \in dv_n\} \\ &= \left(\int \prod_{t=1}^{n} f(\lambda_t, Z_t) P_\theta\{V_1 \in dv_1, ..., V_n \in dv_n\} \right) \left(\prod_{t=1}^{n} h(y_t) \right) \end{aligned}$$

The first product on the last line depends on the parameter θ through the statistics $Z_1, ..., Z_n$, while the second term does not depend on the parameter. This shows that the collection $(Z_1, ..., Z_n)$ is a sufficient statistic for θ.

Notice that Markovian structure of the process V_t has not been used. The second statement of the lemma that (V_t, Z_t) is a hidden Markov model follows directly from the facts that (V_t, Y_t) is, and that each Z_t is a function of Y_t, $Z_t = Z(Y_t)$. It remains to remark that the function Z does not depend on the state v_t or the parameter θ, because it emerged as a sufficient statistic for the family g_λ, which is not related to the hidden Markov structure or parameterization. □

Population mortality plateau In compliment to the mortality plateau discussion of chapter 6, we present a description of the population mortality $\bar{m}(t)$ for a special case. We consider the Gompertz individual baseline hazard (6.5), and the log-frailty process $V_t = \sigma W_t$, which is a Brownian motion with variance σ^2. To begin with, we compute the population survival function, which is the individual survival averaged over all possible frailty paths:

$$\bar{S}(t) = \bar{S}_{K,C,\sigma}(t) = E\left(\exp\{-\int_0^t Ke^{Ct+V_s}ds\}\right) \tag{A.1}$$

Next we obtain a useful representation for the hazard $\bar{m}(t)$.

Lemma 3. *For any* $t \geq 0$

$$\bar{m}(t) = Ke^{(C+\sigma^2/2)t}\frac{\bar{S}_{K,C+\sigma^2,\sigma}(t)}{\bar{S}_{K,C,\sigma}(t)}$$

Proof. First we find that

$$\bar{S}'(t) = -KE\left(e^{Ct+\sigma W_t}\exp\{-K\int_0^t e^{Cs+\sigma W_s}\}ds\right)$$

Now we recall the Girsanov's theorem. In its simplest form, the theorem says that for any c and any continuous functional f,

$$Ef(W_s + cs,\ 0 \leq s \leq t) = e^{-c^2t/2}Ee^{cW_t}f(W_s,\ 0 \leq s \leq t)$$

Take $c = \sigma$ and apply this identity to the expression for $S'(t)$:

$$\bar{S}'(t) = -Ke^{(C+\sigma^2/2)t}E\left(\exp\{-K\int_0^t e^{Cs+\sigma^2 s+\sigma W_s}\}ds\right)$$

Notice that the last expectation represents $\bar{S}_{K,C+\sigma^2,\sigma}(t)$, while by our notation convention $\bar{S}(t)$ with no subscripts is the survival function corresponding to the original parameters K, C, σ. It remains to substitute the expression for $\bar{S}'(t)$ into the identity:

$$\bar{m}(t) = -\frac{\bar{S}'(t)}{\bar{S}(t)}$$

□

Lemma 3 shows that in early life the hazard is approximately Gompertz with the rate $C+\sigma^2/2$. This rate has an increase of $\sigma^2/2$ from the individual hazard rate C, as more frail individuals get killed quicker. Describing $\bar{m}(t)$ for high t is a more delicate matter. Note that $A(t) = \int_0^t e^{Cs+V_s} ds$ is the integrated geometric Brownian motion with drift. The distribution of $A(t)$ was extensively studied in 1990s, see for instance [78], [3] and [25]. Although various analytic expressions exist for the density of $A(t)$ and its moments, they typically involve an inverse Laplace transform and integrated modified Bessel functions, and therefore not convenient for our purpose. However, asymptotic behavior of $E\exp\{-KA(t)\}$ is available in a closed form and can be found in the review of Matsumoto and Yor, [41]. We will state this result without proof.

Lemma 4.

$$\lim_{t\to\infty} t^{3/2} e^{C^2 t/2} E\left(\exp\{-\alpha \int_0^t e^{2(Cs+W_s)} ds\}\right) = \frac{2^{C-3/2}}{\sqrt{\pi}} \Gamma\left(\frac{C}{2}\right)^2 \frac{K_0(\sqrt{2\alpha})}{(\sqrt{2\alpha})^C},\ C>0 \tag{A.2}$$

$$\lim_{t\to\infty} t^{1/2} E\left(\exp\{-\alpha \int_0^t e^{2W_s} ds\}\right) = \sqrt{\frac{2}{\pi}} K_0(\sqrt{2\alpha}),\ C=0 \tag{A.3}$$

Here K_0 is the modified Bessel function of the second kind (see [67]). These asymptotic formulas allow us to obtain the following result.

Lemma 5. *For the Gompertz baseline individual hazard ($C > 0$), the population mortality $\bar{m}(t)$ approaches a positive plateau, given by*

$$\lim_{t\to\infty} \bar{m}(t) = \frac{1}{2}\left(\frac{C}{\sigma}\right)^2 \tag{A.4}$$

If there is no mortality acceleration ($C = 0$, exponential individual survival), then

$$\lim_{t\to\infty} t\bar{m}(t) = \frac{1}{2} \tag{A.5}$$

Proof. First we need to rework the statements of lemma 4 to allow for an arbitrary variance of the log-frailty process. To do this, we use the fact that $\sqrt{a}W_t \stackrel{d}{=} W_{at}$. We have

$$A(t) = \int_0^t e^{Cs+V_s} ds \stackrel{d}{=} \nu \int_o^{t/\nu} e^{C\nu s' + \sigma\sqrt{\nu} W_{s'}} ds'$$

Now take ν such that $\sigma\sqrt{\nu} = 2$:

$$A(t) \stackrel{d}{=} \frac{4}{\sigma^2} \int_0^{t\sigma^2/4} e^{2(2C/\sigma^2 s' + W_{s'})} ds'$$

Define $t' = t\sigma^2/4$, $\alpha = 4K/\sigma^2$ and $C' = 2C/\sigma^2$, and plug them into (A.2) and (A.3). We obtain:

$$\bar{S}(t) = Ee^{-KA(t)} \sim \left(\frac{t\sigma^2}{4}\right)^{-3/2} \exp\left\{-\frac{C^2 t}{2\sigma^2}\right\} \frac{2^{2C/\sigma^2 - 3/2}}{\sqrt{\pi}} \Gamma\left(\frac{C}{\sigma^2}\right)^2 \frac{K_0(\sqrt{8K/\sigma^2})}{(\sqrt{8K/\sigma^2})^{2C/\sigma^2}}$$

The same formula holds for $\bar{S}_{K,C+\sigma^2,\sigma}$ with C replaced by $C + \sigma^2$. When we take the ratio of the two survival functions, it is easy to check that the expression simplifies to

$$\frac{\bar{S}_{K,C+\sigma^2,\sigma}(t)}{\bar{S}_{K,C,\sigma}(t)} \sim \frac{1}{2K} \left(\frac{C}{\sigma}\right)^2 e^{-(C+\sigma^2)t/2}$$

Putting this together with the result of lemma 3 yields (A.4).

In the case $C = 0$ we have

$$\bar{S}(t) \sim \left(\frac{t\sigma^2}{4}\right)^{-1/2} \sqrt{\frac{2}{\pi}} K_0(\sqrt{8K/\sigma^2})$$

and

$$\frac{\bar{S}_{K,\sigma^2,\sigma}(t)}{\bar{S}_{K,0,\sigma}(t)} \sim \frac{1}{2K} t^{-1} e^{-t/2}$$

Together with the statement of lemma 3 this gives (A.5).

□

Conditional expected remaining lifespan This paragraph contains two lemmas, which justify our claims about CERL we made in chapter 7.

Lemma 6. *Let τ be a positive valued random variable with $E\tau < \infty$. Then (7.15) defines the CERL function for $0 < t < t_{\max}$, where $t_{\max} = \sup\{t : P\{T > t\} > 0\}$.*

Conversely, any function $E(t)$ such that

- *$E(t)$ is defined for $0 < t < t_{\max}$, where $t_{\max} = \inf\{t : \lim_{s\to t} E(s) = 0\}$*
- *$E(t) > 0$ in $0 < t < t_{\max}$*
- *$E(t)$ is right differentiable with $E'(t) \geq -1$*

defines a time to event distribution with support $0 < t < t_{\max}$.

Proof. The condition $E\tau < \infty$ guarantees that the first integral in (7.15) exists. In particular, $uP\{\tau > u\} \to 0,\ u \to \infty$ must hold in order for the integral to converge. Therefore, integration by parts yields the second integral.

To show that CERL uniquely identifies the distribution, consider its derivative. From (7.15) it follows that $E(t)$ is (right-)differentiable. Indeed, $S(t)$ is differentiable, and the integral converges uniformly in t. We have by the quotient rule:

$$\begin{aligned} E'(t) &= \int_0^\infty \frac{S'(t+u)S(t) - S(t+u)S'(t)}{S(t)^2} du \\ &= \frac{S(t+u)}{S(t)}\Big|_0^\infty + \frac{-S'(t)}{S(t)} \int_0^\infty \frac{S(t+u)}{S(t)} du \\ &= -1 + m(t)E(t) \end{aligned}$$

We used the fact that $S'_t(t+u) = S'_u(t+u)$ and definition of the mortality function $m(t)$. It follows that $E(t)$ defines mortality by the relation

$$m(t) = \frac{1 + E'(t)}{E(t)}$$

The conditions $E(t) > 0$ and $E'(t) \geq -1$ ensure that $0 \leq m(t) < \infty$, and thus the relation above specifies a valid mortality function, which in turn defines a distribution uniquely.

Finally, it is clear from this representation that mortality blows up to infinity only if $E(t)$ approaches zero at some finite point $t_{\max}$. In this case τ cannot assume values exceeding $t_{\max}$, and $E(t)$ is not defined for $t \geq t_{\max}$. □

Lemma 7. *Let $\tau_1, ..., \tau_n$ be iid draws from the distribution with a finite mean and variance, defined by its CERL $E(t)$. Let $\tilde{E}(t)$ be the empirical CERL. Then $\tilde{E}(t)$ is an asymptotically unbiased, consistent and asymptotically normal estimate of $E(t)$, as $n \to \infty$.*

Proof. To compute expectation of $\tilde{E}(t)$, we use the law of total expectation by conditioning on the random variable $\delta = (\mathbf{1}\{\tau_1 > t\}, ..., \mathbf{1}\{\tau_n > t\})$. δ assumes values in the set Δ of every possible sequences of 0 and 1 of length n. Let k_δ be the sum of all components of δ. If $p_t = P\{\tau > t\}$, then $P\{\delta\}$ is given by the binomial probability $Bin(n, p_t)$. We have

$$E\tilde{E}(t) = \sum_{\delta \in \Delta : k_\delta > 0} P\{\delta\} \frac{\sum_i E[(\tau_i - t)_+ | \mathbf{1}\{\tau_i > t\}]}{k_\delta}$$

We excluded one term in this sum, having all $\tau_i < t$, because in this case $\tilde{E}(t)$ has been defined as zero. The expectations $E[(\tau_i - t)_+ | \mathbf{1}\{\tau_i > t\}]$ are zero if $\tau_i < t$ and equal $E(t)$ otherwise, and there are exactly k_δ non-zeros in the sum. Thus we have

$$E\tilde{E}(t) = (1 - (1 - p_t)^n)E(t)$$

For every fixed t in the domain of $E(t)$ $p_t > 0$, and thus the bias vanishes with $n \to \infty$.

To show consistency, we notice that $(\tau_i - t)_+$ are iid with the expectation equal to $p_t E(t)$, while $\mathbf{1}\{\tau_i > t\}$ are iid with the expectation equal to p_t. By the strong law of large numbers, the averages converge to these expectations with probability 1. Since $p_t > 0$, the ratio of the averages converges (a.s.) to the ratio of the expectations, which is $E(t)$. With Slutsky's theorem, asymptotic normality also follows.

□

The condition of no association In this paragraph we prove lemma 2, where we stated two implications of the condition (7.25).

Let for all functions f such that $Ef(\tau) < \infty$, $E_Z f(\tau) = Ef(\tau)$. Then

1. for all $t > 0$, $E(\tau | Z_{\leq t}; \tau > t) = E(\tau | \tau > t)$
2. for all $t > 0$, $\mathrm{Cov}(\tau, E_{Z_{\leq t}} V_t | \tau > t) = 0$

Proof. First, notice that (7.25) implies

$$E_{Z_{\leq t}} f(\tau) = Ef(\tau) \tag{A.6}$$

for all f and $t > 0$. Indeed, the sigma-algebra generated by paths $Z_{\leq t}$ up to time t is contained in the sigma-algebra generated by full paths Z. Now we apply expectation $E_{Z_{\leq t}}$ to both sides of (7.25). By the law of iterated expectation, $E_{Z_{\leq t}} E_Z f(\tau) = E_{Z_{\leq t}} f(\tau)$. On the right hand side we have $E_{Z_{\leq t}} Ef(\tau) = Ef(\tau)$.

Next we show that the expectations involved can be replaced by conditional expectations given survival up to time t. To do this, take $f(x) = \mathbf{1}\{x > t\}$ and use (A.6). This yields $P_{Z_{\leq t}}\{\tau > t\} = P\{\tau > t\}$. Now consider (A.6) with the following functions:

$$f_t(x) = f(x) \frac{\mathbf{1}\{x > t\}}{P\{\tau > t\}} = f(x) \frac{\mathbf{1}\{x > t\}}{P_{Z_{\leq t}}\{\tau > t\}}$$

We have

$$Ef_t(\tau) = \int_t^\infty f(x) \frac{P\{\tau \in dx\}}{P\{\tau > t\}} = E(f(\tau) | \tau > t)$$

$E_{Z_{\leq t}} f_t(\tau)$ is obtained in the same way, and so we have

$$E_{Z_{\leq t}}(f(\tau) | \tau > t) = E(f(\tau) | \tau > t)$$

Taking $f(x) = x$ yields the first statement of the lemma.

To prove the second statement, we once again employ the law of iterated expectation:

$$E(\tau E_{Z_{\leq t}} V_t | \tau > t) = E\left(E[\tau E_{Z_{\leq t}} V_t | Z_{\leq t}; \tau > t] | \tau > t\right)$$

The quantity $E_{Z_{\leq t}} V_t$ is measurable with respect to the sigma-algebra generated by events $(Z_{\leq t}; \tau > t)$, and therefore

$$E\left(E[\tau E_{Z_{\leq t}} V_t | Z_{\leq t}; \tau > t] | \tau > t\right) = E\left([E_{Z_{\leq t}} V_t] E_{Z_{\leq t}}[\tau | \tau > t] | \tau > t\right)$$

By statement 1 of the lemma, we can replace $E_{Z_{\leq t}}(\tau | \tau > t)$ with $E(\tau | \tau > t)$, which is a constant. Therefore,

$$E\left([E_{Z_{\leq t}} V_t] E_{Z_{\leq t}}[\tau | \tau > t] | \tau > t\right) = E\left([E_{Z_{\leq t}} V_t] E[\tau | \tau > t] | \tau > t\right) =$$

$$E\left(E_{Z_{\leq t}} V_t | \tau > t\right) E[\tau | \tau > t]$$

It remains to apply the formula for covariance:

$$\mathrm{Cov}(\tau, E_{Z_{\leq t}} V_t | \tau > t) = E(\tau E_{Z_{\leq t}} V_t | \tau > t) - E(E_{Z_{\leq t}} V_t | \tau > t) E(\tau | \tau > t) = 0$$

□