Modelling interval-censored event times in the presence of a cure fraction:

From building to refining

A thesis submitted in partial fulfillment of the requirements for the degree of Docteur en Sciences, Université catholique de Louvain, September 2016

Sylvie Scolas

Thesis committee:

Catherine Legrand, Supervisor (Université catholique de Louvain)
Anouar El Ghouch, Supervisor (Université catholique de Louvain)
Philippe Lambert (Université de Liège, Université catholique de Louvain)
Emmanuel Lesaffre (Katholieke Universiteit Leuven)
Abderrahim Oulhaj (United Arab Emirates University)
Ingrid Van Keilegom (Université catholique de Louvain)

Contents

1 Introduction **13**

1.1 Introduction 14

1.2 Survival analysis: a basic introduction 16

1.2.1 A brief history of survival analysis 17

1.2.2 Main survival functions 19

1.2.3 Parametric modelling 21

1.2.4 Non-parametric estimation of the survival distribution . 24

1.2.5 Regression modelling 25

1.3 Survival analysis in the presence of a cure fraction 28

1.3.1 The mixture cure model 28

1.3.2 Parametric mixture cure model 29

1.3.3 Semi-parametric mixture cure model 29

1.3.4 The promotion time cure model, an alternative to the mixture cure model 31

1.4 Diagnostic checks in survival analysis 32

1.4.1 The Cox Snell residuals 32

1.4.2 The martingale residuals 33

1.4.3 The deviance residuals 34

1.5 Variable selection methods 34

1.5.1 The best subset selection 35

1.5.2 Forward, backward and stepwise selection 35

1.5.3 Shrinkage methods 36

1.6 Real Data: Alzheimer's disease 38
1.7 Objectives and outline of the thesis 39

2 Taking into account interval-censoring and cure 43
2.1 Introduction 44
2.2 Data generation 44
2.3 Data generation for this chapter 45
2.4 The presence of a cure fraction 45
2.4.1 Survival curves comparison 46
2.4.2 Regression modelling 48
2.5 The presence of interval-censoring 51
2.5.1 Survival curves comparison 51
2.5.2 Regression modelling 52
2.6 Discussion 52

3 The adaptive LASSO 67
3.1 Introduction 69
3.2 Methodology 70
3.2.1 The adaptive LASSO: least square approximation . . . 70
3.2.2 The adaptive LASSO in the presence of cured individuals 71
3.2.3 Tuning parameter selection and variance estimation . . 72
3.3 Simulation study 73
3.3.1 Simulation setting 73
3.3.2 Simulation results 73
3.4 Real data analysis 75
3.5 Discussion 76

4 Diagnostics checks 87
4.1 Introduction 89
4.2 Methodology 91
4.2.1 Checking the survival function: Cox-Snell residuals . . . 91
4.2.2 Detecting non-linearity: Deviance residuals 93
4.3 Simulation study 96
4.3.1 Simulation setting 96
4.3.2 Simulation results 97
4.4 Real data analysis 101
4.5 Discussion 101

5 The SNP representation 113

5.1 Introduction 115
5.2 Methodology 116
5.2.1 The SNP representation as an estimation method 117
5.2.2 The SNP representation for goodness-of-fit testing . . . 119
5.3 Simulation study 121
5.3.1 Simulation setting 121
5.3.2 Simulation results 121
5.4 Real data analysis 124
5.5 Discussion 125

6 Discussion **137**
6.1 Conclusion 138
6.2 Conclusion on Alzheimer's disease database 143
6.3 Further work 144

Bibliography **147**

Appendices **157**

A Appendix to Chapter 2 **157**
A.1 R Code to generate data 157
A.2 Tables 160

B Appendix to Chapter 3 **163**
B.1 Non-convergence 163
B.2 R code 163

C Appendix to Chapter 4 **171**
C.1 Deviance with the Extended Generalized Gamma distribution . 171
C.1.1 Deviance for right-censored observations in the EGG-AFT mixture cure model. 171
C.1.2 Deviance for exact event time in the EGG-AFT mixture cure model. 172
C.1.3 Deviance for interval-censored observations in the EGG-AFT mixture cure model. 172
C.2 Additional plots 174

D Appendix to Chapter 5 **175**
D.1 Estimation 175
D.1.1 Results with the BIC criterion 175

D.1.2 Results of estimation using true and Normal distribution for Scenario B. 175
D.2 Goodness-of-fit 175
D.2.1 Testing Weibull event times 175
D.2.2 Testing Exponential event times 179
D.3 R-code 179

"If we knew what it was we were doing, it would not be called research, would it?"

- Albert Einstein (1879-1955)

"I have not failed. I've just found 10,000 ways that won't work."

- Thomas Edison (1847-1931)

Acknowledgments

When starting a PhD thesis, we may think that we will change the world, make the best thesis in the world and, why not, save some lives! Our motivation is thus at its highest level. Of course as time goes by, the "best thesis in the world" becomes a "good" thesis, and ends up by being a "ok-let's-just-make-a-thesis" thesis. This is a challenging opportunity, not always easy, and this is why people must be thanked for the support that they have brought all along these four years.

First of all, of course, my promoters, Catherine and Anouar. Without you, I would not even have started a PhD! So thank you for this, but thank you also for all the time you invested in me! Thanks for making our meetings a pleasant time, for giving me ideas when I needed them, and for understanding me (which was not always easy, I admit it!). Of course, I thank all the members of my jury and committee, Abderrahim, Philippe, Emmanuel and Ingrid. Thank you for taking the time to read my work and for makings comments and remarks that certainly improved this thesis.

Second, I must thank all my "office-mates" at the D372 office (the best in the world!) who had to bear with me for such a long time and (sometimes) laugh at my silly jokes. I hope the white blackboard will not stay empty, and I count on Nathalie and Florian to prevent this from happening. Special thanks of course to Marco and Anne: you welcomed me on the very first day, and I can still hear Marco saying that he likes to be at the office early in the morning to have the afternoon clear. Then I learned that early meant 4 or 5 am for him! You two were always there when I needed you, whether it be for statistics, fruit juices or Walking Dead discussions;). Jennifer, thank you for sharing your first year with me, for the train rides, for the tea-times and so many other things,

and I wish I will see you soon in Kangaroo's island! Adrien, thank you for the casual Thursday's, and for the 'Prince' and 'Pimm's' that you missed so often!

Another series of thanks to the stat-actu crew, for the 10, hmm.. no! The 10.30, hmm... no! The 11 am coffee break. Thank you Nadja, Tatiana, Sophie and Nancy for caring for us like you did. Thanks to our IT crew, in particular Raphaël, who help me keep my computer alive a little longer. Thanks to everybody for the lunch breaks, the tea breaks, the Friday drinks. For all the fun we had, without which I would have gone crazy a long time ago! In particular, thank you Aurélie for the nice talks we had about everything, including statistics but also Canada and Magic Mike; thanks to Benjamin for his splendid music quizzes and to Alain for making us the 'champions of the world'; thanks to Anna for the "bubbles" and for making me discover Chipolata! Sebastien, thank you for sharing your historical knowledge with me, I really enjoyed it and I will think of you whenever I will see a turtle or a snail! Vincent, I am not forgetting you: thanks for the sunscreen and the poem of Lisbon! I also thank Nicolas for the ERCIM conference in London, Maïlis for the birthday balloons, Nathan for always bringing up a special topic (such as Albert, other topics can not be mentioned in a thesis), Hélène for the funny EDT talks, Samuel for the breaks in the afternoon and for Draco (no! I will not watch this!), Mickaël, for the tea-times and for being so nice by letting my redecorate your office (and giving me nice exam surveillances!). Special thanks also to Cedric for motivating us to exercise (ok, to play badminton and eat a "healthy" diner afterwards)! Nathalie, thank you for being such a nice new colleague, and for our serious Zalando talks. And Florian, thank you for being such a nice Luke!

I am also grateful to the SMCS team: Manon[1] (and the couch I never went on), Céline (and your wonderful strawberry sorbet although wonderful is not strong enough), Catherine, Nathalie and Alain for answering all my questions and for being so nice (well at least Catherine and Nathalie!). Big thanks to the great professors at UCL, who make statistics understandable for, well, some students!

To all of my friends who, in a way or another, made me who I am today, thanks to you too: Choup', Gene, Ma Poule, Sophie, ma Grosse, Greg, Lau, Sophie, Fabien (aka Stéphane) and Dany. I really hope you will enjoy reading this thesis, and not only the part in which you appear!

Cette section de remerciements ne pourrait évidement pas être complète sans une mention hyper spéciale pour mes parents. Sans eux, je ne serais *littéralement* pas là où j'en suis actuellement. Ils m'ont non seulement donné la vie, mes pre-

[1] your office is a SMCS office!

miers repas et mes premiers mots, mais aussi tout le support nécessaire pour en arriver là! Vous êtes *vraiment* les meilleurs parents du monde! Merci!

Un autre énorme merci à "frère", avec qui j'ai passé les meilleures années de ma vie, même si c'était pour se battre la plupart du temps! Je pense que toute personne devrait avoir un frère comme toi (ou une soeur comme moi!;)). Merci à toi et à ma super belle-*sista*, pour m'avoir aidé à me changer les idées avec les films et les jeux, mais aussi pour avoir fait de moi la marraine la plus fun de tous les temps!

And of course he would hate me if I do not mention him... Thank you Matthieu! You gave me the strength the take a new path, you encouraged and helped me, even though you said it would be a (Belgian[2]) rough road. You supported me, and god knows the word 'support' is not too strong in this case.

[2]this is a private joke

CHAPTER 1

Introduction

"You can't jump straight to the end, the journey is the best part."

- Robin Scherbatsky

1.1 Introduction

This thesis was motivated by the difficulties encountered to analyze a database containing some particular features [Oulhaj et al., 2009]. The objective of this study was to detect the conversion from a healthy status to a Mild Cognitive Impairment (MCI) status, a possible precursor of Alzheimer's disease. It is a matter of the utmost importance to detect risk factors associated with the onset of this conversion. Indeed, Alzheimer's disease is one of the worst plague of this century and the most common cause of dementia. Dementia causes deterioration in cognitive skills, such as memory or thinking skills, and affects the ability to perform everyday activities. According to the World Health Organization (WHO), 47.5 million people worldwide have dementia and 7.7 million new cases occur every year. Such a medical condition has social and economic impact on society, in addition to the obvious impact on the patient himself.

In this database, 241 elderly patients were followed at scheduled interviews. This means that if a conversion to MCI occurred, the exact occurrence date was not exactly known: the only information at hand is the last visit at which the patient was cognitively healthy, and the first visit at which the patient was diagnosed with a MCI status. These visits are the left and right endpoints of an interval of time during which the event happened. In the statistical literature, this feature is called *interval-censoring*.

A large proportion of patients did not experience the event during the study period. Indeed, some patients left the study before the end of the follow-up, or had simply not converted to MCI at the end of the study period. This case is referred to as *right-censoring* in the statistical literature, and the only available information is a lower bound on the true event-time. Moreover, some patients will in fact never experience the event, no matter how long the follow-up is, because they are not susceptible to it. In the statistical literature, those patients are said to be "*cured patients*", "*long-term survivors*", "*non-susceptible patients*" or "*immune patients*" [Maller and Zhou, 1996]. In the following, we will use the term "cure" to denote an observation that will never experience the event, whatever the context and without relying on the medical interpretation of the word "cure".

The goal of this thesis is to provide appropriate tools to analyze data presenting those features: right-censoring, interval-censored event times instead of exactly observed event-times and the presence of a cure fraction. The field of statistics that consists in studying the time until an event occurs is called survival analysis. It is a domain of outstanding interest with applications not only in medical sciences, but also in social sciences, economics (duration analysis), engineering (reliability analysis), and many others. Examples other than

Alzheimer's disease include studying the time until the failure of an engine, or until the first cigarette or alcohol consumption.

One of the major characteristics of survival analysis is the presence of censoring, resulting in incomplete information. A large majority of methods in survival analysis has been developed in the context of right-censoring. A more challenging context is the presence of both interval- and right-censored data, in which case classical techniques must be adapted. Furthermore, it is usually assumed that if a patient is right-censored, the event will eventually occur at a later time point. This is not the case in our database, as a fraction of the population is cured from the event. Failing to take this cure fraction into account may produce inaccurate estimation and, consequently, erroneous conclusions [Maller and Zhou, 1996].

In the past decades, numerous authors have proposed alternatives to standard survival techniques to take a cure fraction into account. Pioneers in that field were [Boag, 1949] and [Berkson and Gage, 1952], who laid the basis of the *mixture cure model*. This model assumes that the entire population under study consists of two sub-populations: the cured sub-population, containing patients who will never experience the event and are consequently always right-censored; and the uncured sub-population, containing patients who will eventually experience the event and *may* be right-censored. The mixture cure model consists of two parts: the incidence part models the probability to experience the event (i.e. one minus the cure probability) while the latency part models the time-to-event for the uncured observations.

In many occasions, the event time and the cure probability depend on covariates. This dependence is usually modelled by a *regression* model. In the incidence part, a logistic or probit regression model is very often used. The Proportional Hazards (PH) or the Accelerated Failure Time (AFT) models, described hereinafter, are widely used in the latency part [Farewell, 1977b, Farewell, 1977a, Kuk and Chen, 1992, Yamaguchi, 1992, Taylor, 1995, Sy and Taylor, 2000, Peng and Dear, 2000, Li and Taylor, 2002, Zhang and Peng, 2007, Cai et al., 2012].

While the PH model is very popular in survival analysis, it is based on some quite strong assumptions that may not always be verified. The AFT model provides an excellent alternative, for example when the PH hypothesis is not met [Wei, 1992]. Besides, as the popular statistician Sir David Cox stated, "accelerated life models are in many ways more appealing because of their quite direct physical interpretation" [Reid, 1994].

Both the PH and AFT model can be parametric (i.e. some assumptions about the form of the survival distribution are imposed), or semi-parametric (i.e. the survival distribution is unspecified), leading to a parametric or semi-parametric

PH or AFT mixture cure model. In the case of interval-censoring, using a semi-parametric AFT model is quite complicated and the use of parametric methods is thus very appealing.

An advantage of the mixture cure model is that it allows a direct interpretation of the effects of covariates on the cure probability, and on the survival distribution for uncured individuals, separately. Interestingly, these two sets of covariates may not necessarily be the same, nor have the same impact. However, this can lead to a large number of potential covariates to be included in each component of the model, and in this case variable selection is needed for the final model to be easily interpretable and to possess good predictability.

It is generally assumed that the covariates have a linear impact. However, this may not be true and a covariate that is linear in one part of the model is not necessarily linear in the other part. It is then of the utmost interest to verify the validity of such linearity assumptions, and refine the model if necessary.

In this thesis, the primary interest is on building an appropriate mixture cure model to handle interval-censored data with a fraction of cured individuals. In particular, we will propose a parametric but flexible logistic-AFT mixture cure model for which the issues of variable selection and model checking will be addressed. Additionally, we will present an alternative to parametric models, which can also be used as a goodness-of-fit procedure to assess the fit of a parametric distribution.

The remainder of this chapter will introduce the concepts and models used in the thesis. The basis of classical survival analysis are covered in Section 1.2. The method used to handle the cure fraction is explained in Section 1.3.1. A very brief introduction to the issue of model checking and variable selection is presented in Section 1.4 and 1.5. The Alzheimer's disease database that motivated this thesis is presented in more details in Section 1.6. Lastly, Section 1.7 gives more details about the objectives and the organization of the thesis.

1.2 Survival analysis: a basic introduction

This section presents a brief introduction to survival analysis and, more specifically, to the concepts used throughout this thesis. This section as well as the following of the thesis will assume that the event times are interval-censored instead of exactly observed, unless otherwise stated. Moreover we suppose in this section that every observation will eventually experience the event of interest, i.e. there is no cure fraction.

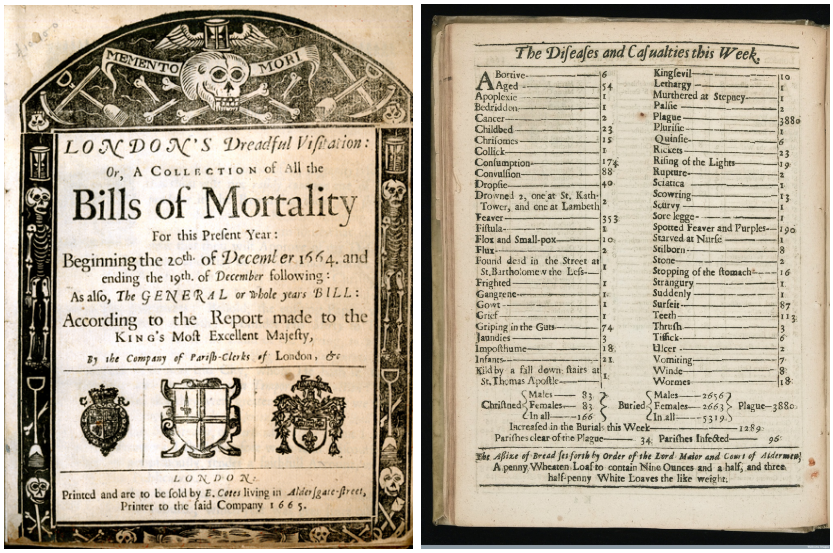


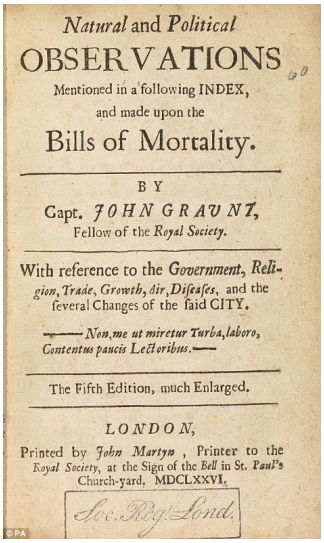
Figure 1.1 – London's Dreadful Visitation, a collection of all the *Bills of Mortality*

1.2.1 A brief history of survival analysis

While survival analysis is nowadays considered as a specific branch of statistical analysis in itself, it is interesting to note that statistic actually has its very first roots in today's survival analysis.

The great plague of 1603 led the people of London to show an increasing interest in *Bills of Mortality*, i.e. weekly reports containing information about the number of christenings and burials, and, more specifically, the number of deaths caused by the plague (Figure 1.1 gives an example of these *Bills of Mortality*). A sudden increase in deaths from the infectious disease could thus warn the population about a possible epidemic. However, a merchant named John Graunt (1620-1674) granted little credit as to the accuracy of these *Bills of Mortality*, and began a work that led to, in 1662, the publication of his only but remarkable book, *Natural and Political Observations upon the Bills of Mortality* (see Figure 1.2), which is considered as one of the very first sign of statistical analysis.

For instance, John Graunt observes that before the plague-year 1625, the reported number of deaths amounted between 7.000 and 8.000 persons per year, whereas during the plague-year, the Bills of Mortality reported 35.417 persons dead from the plague, and 18.848 persons dead from other causes. He thus assumes that 11.000 more persons than initially reported should have died


Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

Capt. JOHN GRAUNT,
Fellow of the *Royal Society*.

With reference to the *Government, Religion, Trade, Growth, Air, Diseases*, and the several Changes of the said CITY.

——*Non me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,

Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXXVI.

Soc. Reg. Lond.

Figure 1.2 – *Natural and Political Observations upon the Bills of Mortality*, 1676

from the plague. Another example is how, based on the information contained in the Bills of Mortality, he estimates the number of inhabitants in London as well as the rate of the population growth. In the same vein, he estimates a mortality rate and presents one of the first life-tables. While John Graunt is considered as the founder of statistics, it is worth mentioning that the paternity of his book is sometimes attributed to his scientific friend, William Petty, one of the founder of the Royal Society (of which John Graunt is elected as a member thanks to his book). William Petty is known as the inventor of the name *Political Arithmetic*, the former name of modern statistics [1], which was at the basis of political economy and demography.

Many other scientists followed Graunt's and Petty's footsteps. The mathematicians Christian (1629-1695) and Lodewijk Huygens (1631-1699) proposed a first calculation of the mean duration of life. The astronomer Halley (1656-1742), famous for giving his name to a comet, tried to disclaim popular superstitious beliefs, such as the influence of the phases of the moon on health, and constructed a table giving the population according to age. The famous mathematician Laplace (1749-1827) also contributed to improve methods to estimate the population of a country. He based his calculations on birth-rates in France,

[1] Actually, the word statistic comes from the Italian word *stato*, designating the *State*. Statistics were in fact, at that time, a collection of non-necessarily numerical data and observations about the states. It is the economist Knies (1821-1898) who assigned the name *Statistics* to Political Arithmetic.

which he found by *selecting* some departments in the country, and enumerating their inhabitants. This selection may be recognized as a *sample* of the population. Another famous mathematician, Lavoisier (1743-1794), also worked on population estimation, but tried to estimate the number of horses, sheep or pigs. Lavoisier based his calculations on the quantity of some necessities that he supposed to be constant in the population.

Concerning the use of statistics in medicine, James Lind (1716-1794) conducted in 1747 what is considered as the first controlled clinical trial. Halley's finding were used by Daniel Bernoulli (1700-1782) to answer questions about the small-pox, such as the frequency of the disease, or its lethality, i.e. the rate of mortality amongst the diseased.

At that time, the notion of censoring has not yet appeared, and the information about a patient who left the study was simply not considered. It is in 1897 that the word truncation appears. Francis Galton (1822-1911), a cousin of Charles Darwin (1809-1882), analyzed a truncated sample by assuming the normal distribution[2]. By then, censoring and truncation were part of a unique concept. The first apparition of the word *censoring* was marked in 1949, in a paper of Hald [Hald, 1949], who credited J.E. Kerrich for this name. Hald considers differently a truncated sample from which no observation is available, from a truncated sample for which some information is incomplete, i.e. a censored sample, as we are accustomed to.

In 1958, Edward Kaplan (1920-2006) and Paul Meier (1924-2011) proposed a solution[3] to estimate non parametrically the survival distribution of a right-censored sample [Kaplan and Meier, 1958]. In 1971, sir David Cox (1924-) lays another stone in the history of survival analysis by proposing a regression model, the proportional hazards model [Cox, 1972], one of the most cited papers in statistics.

Research in survival analysis is still very popular nowadays, with subjects including competing risks, frailty models, joint longitudinal models and many more. The reader interested in the history of statistics may consult [Westergaard, 1932].

1.2.2 Main survival functions

In this section, we briefly introduce the basic concepts of survival analysis. The literature counts a plethora of references giving a comprehensive review on the

[2]Ten years after his famous paper describing Maximum Likelihood Estimator, Ronald Fisher (1890-1962) applies his methodology to the same truncated sample, and finds results equivalent to Galton's.

[3]Actually, they both submitted their work to the *Journal of the American Statistical Association* and the editor, John Tukey (1915-2000), convinced them to combine their work into a single paper.

subject. We refer for example to [Kalbfleisch and Prentice, 2002, Klein and Moeschberger, 2003, Collett, 2003, Lawless, 2003].

Let $t_1, \cdots, t_n$ be the realizations of n independent positive continuous random variables $T_1, \cdots, T_n$, representing the event times of interest. While the cumulative distribution function, $F(t) = P(T \leq t)$, is commonly used to describe random variables in statistics in general, three quantities are more relevant in the field of survival analysis, and are all related to each other:

1. The survival distribution: $S(t) = P(T > t) = 1 - F(t)$, giving the probability to experience the event after a certain time t;
2. the hazard function or hazard rate, $h(t) = \lim_{\Delta t \to 0} \frac{P(t<T\leq t+\Delta t|T>t)}{\Delta t}$, representing the instantaneous risk of experiencing the event;
3. the cumulative hazard function, $H(t) = \int_0^t h(x)dx$, representing the accumulation of instantaneous risk over time.

Writing $f(t)$ the density function of T, it is trivial to see that

$$\begin{aligned} H(t) &= -\log(S(t)) \\ h(t) &= \frac{-\partial}{\partial t}\log(S(t)) = \frac{f(t)}{S(t)}, \end{aligned}$$

so that the knowledge of one quantity implies the knowledge of the other two quantities.

Usual statistical techniques, such as the empirical cumulative distribution function, cannot directly be used to estimate any of these quantities, due the presence of censoring. Three types of censoring exist:

- left-censoring: the event of interest occurs before some time say r_i: $t_i \in]0, r_i]$;
- right-censoring: the event of interest occurs after some time say l_i: $t_i \in [l_i, \infty[$;
- interval-censoring: the event of interest occurs between two times, say l_i and r_i: $t_i \in]l_i, r_i]$.

Note that an observation that is left-, right- or interval-censored is an incomplete observation, as the true event time t_i is not observed exactly. It can be seen with a slight abuse of notation, that interval-censoring encompasses left-censoring ($l_i = 0$), right-censoring ($r_i = \infty$), as well as exactly observed event times ($l_i = r_i = t_i$). In this thesis, we assume that every observation is either right-censored or interval-censored. Let $l_1, \cdots, l_n$ and $r_1, \cdots, r_n$ represent the

observed left and right endpoints of the time intervals such that $l_i < t_i \leq r_i$ for $i = 1, \cdots, n$. We furthermore assume non informative and independent censoring, i.e. the censoring mechanism contains no other information than t_i is bracketed by two observed values [Self and Grossman, 1986, Sun, 2006].

The censoring indicator δ_i is used to identify whether a patient experienced the event or not: $\delta_i = 1$ indicates an interval-censored observation (i.e. $r_i < +\infty$), and $\delta_i = 0$ indicates a right-censored observation (i.e. $r_i = +\infty$).

1.2.3 Parametric modelling

Some parametric distributional forms are quite popular to model survival data. In this section, some of the most commonly used distributions are briefly reviewed. Other distribution functions are given in [Klein and Moeschberger, 2003], p.38 for instance.

- The exponential distribution has parameter λ and is such that the hazard rate is constant over time. The density, survival and hazard rate are, respectively:

$$\begin{aligned} f(t) &= \lambda \exp(-\lambda t) \\ S(t) &= \exp(-\lambda t) \\ h(t) &= \lambda \end{aligned}$$

  for $0 \leq t < \infty$.

- The Weibull distribution generalizes the exponential distribution and is characterized by two positive parameters, $\lambda > 0$ and $\rho > 0$. The density, survival and hazard rate are, respectively:

$$\begin{aligned} f(t) &= \rho\lambda t^{\rho-1} \exp(\lambda t^{\rho}) \\ S(t) &= \exp(\lambda t^{\rho}) \\ h(t) &= \rho\lambda t^{\rho-1}, \end{aligned}$$

  for $0 \leq t < \infty$, where ρ controls the *scale* of the distribution and λ controls the *shape*. Remark first that the hazard function for Weibull-distributed event times is monotone; and that if $\rho = 1$, the distribution reduces to the exponential distribution.

- The log-Normal distribution is another possible distribution appropriate for survival data. The positive random variable T is log-normally distributed if $\log(T)$ is normally distributed, with mean μ and variance σ^2,

that is

$$f(t) = \frac{1}{t\sigma}\varphi\left(\frac{\log(t)-\mu}{\sigma}\right)$$
$$S(t) = \int_t^{+\infty} f(x)dx = 1 - \phi\left(\frac{\log(t)-\mu}{\sigma}\right)$$
$$h(t) = \frac{\varphi\left(\frac{\log t-\mu}{\sigma}\right)}{t\sigma\left[1-\phi\left(\frac{\log(t)-\mu}{\sigma}\right)\right]}$$

where φ is the standard Normal density function, and ϕ is the standard Normal cumulative distribution function.

- The Extended Generalized Gamma distribution (EGG) encompasses both the Weibull and the log-Normal distributions. A random variable T has an $EGG(q)$ distribution with parameters μ and σ if

$$\log(T) = \mu + \sigma\varepsilon$$

and ε has probability density function

$$f(v;q) = \begin{cases} \frac{|q|}{\Gamma(q^{-2})}(q^{-2})^{q^{-2}}\exp(q^{-2}(qv - e^{qv})) & \text{if } q \neq 0 \\ \frac{1}{(2\pi)^{1/2}}\exp(-v^2/2) & \text{if } q = 0, \end{cases}$$

and survival distribution

$$S(v;q) = \begin{cases} 1 - I(q^{-2}e^{qv}, q^{-2}) & \text{if } q > 0 \\ I(q^{-2}e^{qv}, q^{-2}) & \text{if } q < 0 \\ \int_v^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx & \text{if } q = 0, \end{cases}$$

where $\Gamma(k)$ is the gamma function, i.e. $\Gamma(k+1) = k!$, and $I(\cdot, k)$ is the incomplete gamma integral [Lawless, 1980], that is

$$I(\cdot, k) = \frac{1}{\Gamma(k)}\int_0^{\cdot} x^{k-1}e^{-x}dx.$$

The Weibull distribution is obtained when $q = 1$, i.e. ε follows an extreme value distribution; the log-normal distribution is obtained when $q = 0$, i.e. ε follows a Normal distribution.

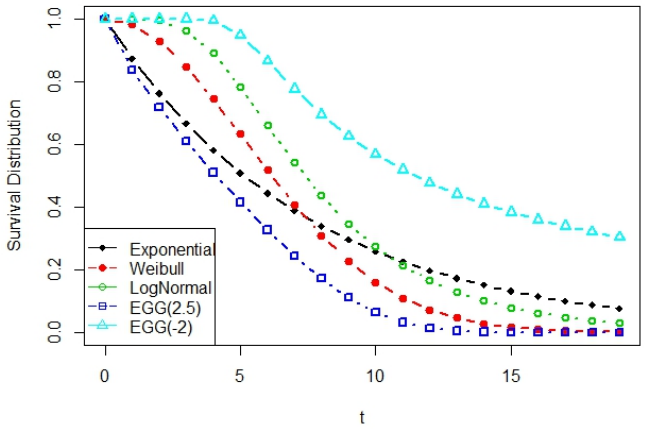


Figure 1.3 – Plot of parametric survival distributions (Exponential, Weibull, log-Normal and EGG).

A plot of the survival distribution for the Exponential, Weibull, log-Normal, $EGG(2.5)$ and $EGG(-2)$ is given in Figure 1.3, for the same mean and standard deviation (i.e. exp(2) and 0.5 respectively).

Likelihood function and estimation

Under the assumption that the event times $T_1, \cdots, T_n$ follow a survival distribution S known up to some parameter(s) $\boldsymbol{\theta}_0$, i.e. $S(t|\boldsymbol{\theta}_0)$, usual maximum likelihood theory applies. Note that in the following, when no confusion may arise, we will write $S(t)$ for $S(t|\boldsymbol{\theta}_0)$. The likelihood function for observations that are either right- or interval-censored is proportional to:

$$L(\boldsymbol{\theta}_0; l_i, r_i) = \prod_{i=1}^{n} \left[S(l_i) - S(r_i)\right].$$

We may equivalently use the censoring indicator δ_i and obtain the following notation

$$L(\boldsymbol{\theta}_0; l_i, r_i, \delta_i) = \prod_{i=1}^{n} S(l_i)^{(1-\delta_i)} \left[S(l_i) - S(r_i)\right]^{\delta_i}. \tag{1.1}$$

Standard optimization techniques such as the Newton-Raphson algorithm can be used to estimate $\boldsymbol{\theta}_0$. Theoretical large sample properties of the maximum likelihood estimator follow, such as consistency and unbiasedness; and the stan-

dard errors can easily be computed trough the Hessian matrix of the likelihood at its Maximum Likelihood Estimator (MLE).

1.2.4 Non-parametric estimation of the survival distribution

In the presence of right-censored and exact observations only, the Kaplan-Meier estimator is the most popular non-parametric estimator of the survival distribution [Kaplan and Meier, 1958]. The concept is quite simple: first, order the unique event times and obtain $0 < s_1 < \cdots < s_D < +\infty$, where D represents the number of distinct event times. Also, note d_j the number of events occurring at time s_j, and n_j the number of observations still *at risk* at time s_j, i.e. observations for which the event time or right-censoring time is larger than or equal to s_j. The Kaplan-Meier estimator at any time t is given by:

$$\hat{S}(t) = \prod_{s_j < t} (1 - \frac{d_j}{n_j}).$$

The estimated survival distribution is a discrete distribution with jumps only at the set of unique event times $s_1, \cdots, s_D$.

The Kaplan-Meier estimator involves the ordering of the event times. In the context of interval-censored event times, such an ordering is impossible due to the possibility of overlapping intervals. Suppose for example that the intervals of two observations are $]2, 4]$ and $]1, 5]$, it is then impossible to know for which observation the event occurred first. The Kaplan-Meier estimator is thus not directly applicable for interval-censored data, and needs to be adapted .

The following solution is due to [Peto, 1973] and [Turnbull, 1976]. Define $0 = s_0 < s_1 < \cdots < s_m < s_{m+1} = +\infty$, the set of the unique ordered elements of $\{l_i, r_i\}_{i=1}^n \cup \{0, +\infty\}$. For $i = 1, \cdots, n$ and $j = 1, \cdots, m+1$, define $\alpha_{ij} = 1$ if $]s_{j-1}, s_j] \subseteq]l_i, r_i]$ and 0 otherwise. For notational convenience, let $\boldsymbol{p} = (p_1, \cdots, p_{m+1})$ with $p_j = \mathbb{P}(T_i \in]s_{j-1}, s_j]) = S(s_{j-1}) - S(s_j)$ for $j = 1, \cdots, m+1$, representing the probability that the event occurred within the interval $]s_{j-1}, s_j]$. It results that, for $l_i < r_i$, $S(l_i) - S(r_i) = \sum_{j=1}^m \alpha_{ij} p_j$, and the likelihood function is proportional to

$$L(\boldsymbol{p}; \alpha_{ij}) = \prod_{i=1}^{n} \sum_{j=1}^{m+1} \alpha_{ij} p_j,$$

where $\sum p_j = 1$. Maximizing this likelihood leads to the nonparametric survival estimator. It will correspond, as for the Kaplan-Meier estimator, with a discrete distribution that changes value only at some s_j. Actually, [Turnbull, 1976]

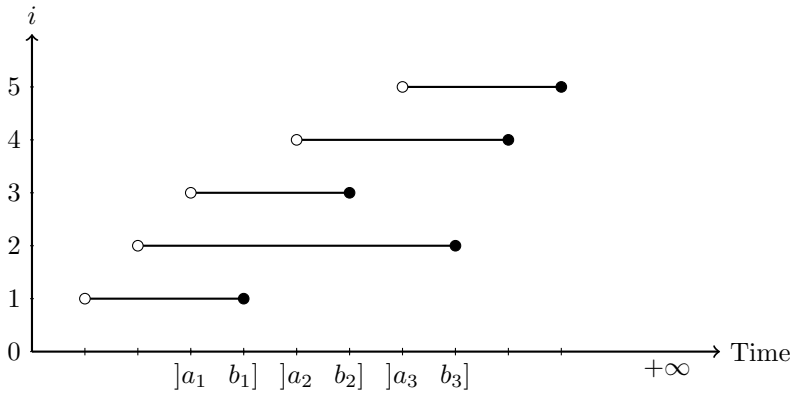


Figure 1.4 – Representation of Turnbull intervals for $n = 5$ observations. The horizontal lines represent the censoring intervals for each observation. On the x-axis, the Turnbull intervals are represented.

showed that the estimator changes value on intervals of the form $[a_j, b_j]$, with $a_1 \leq b_1 < a_2 \leq ... \leq b_{m-1} < a_m \leq b_m$, and which contain no other points of $\{l_i, r_i\}$ except at their end points (see example in Figure 1.4).

Some algorithms have been proposed to find the non-parametric MLE: the self-consistency algorithm based on the Expectation Maximization (EM) algorithm [Turnbull, 1976], which has a low convergence rate; the Iterative Convex Minorant (ICM) algorithm [Groeneboom and Wellner, 1992], which was modified by Jongbloed to insure convergence of the algorithm [Jongbloed, 1998]; or the hybrid EM-ICM algorithm [Wellner and Zhan, 1997]. We refer to [Sun, 2006] for a review of these algorithms.

1.2.5 Regression modelling

A regression model is used to model the dependence on covariates, and estimate the impact that these covariates can have on the survival distribution. We describe in this section two major regression models. The first one is the Proportional Hazards (PH) model and the second one is the Accelerated Failure Time (AFT) model.

In the following, let $\boldsymbol{X}_i' = (1, X_{i1}, \cdots, X_{im}) \in \mathbb{R}^{(m+1)}$ denote the $(m+1)$-vector of covariates that may impact T_i through the parameter μ_i, and note $\boldsymbol{\beta}' = (\beta_0, \beta_1, \cdots, \beta_m)$ the corresponding $(m+1)$-vector of unknown coefficients. Furthermore, we assume that the effect of covariates is linear, that is $\mu_i = \boldsymbol{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}$. We note $\mathcal{O}_i$ the observed dataset for an individual i such that $\mathcal{O}_i = (x_i, l_i, r_i, \delta_i)$, $i = 1, \cdots, n$.

Proportional Hazards (PH) model

The PH model assumes a direct relationship of the covariates with the hazard function, and states that

$$h(t_i) \equiv h(t_i|\boldsymbol{X_i} = \boldsymbol{x_i}) = h_0(t_i)\exp(\mu_i),$$

where h_0 is a baseline hazard function, common to all individuals, and where $\beta_0 = 0$ to avoid identifiability issues. The assumption of proportionality of hazards naturally follows: two observations with a different value of a covariate have proportional hazard functions. The model can be rewritten in terms of the conditional survival distribution, i.e. $S(t_i) \equiv S(t_i|\boldsymbol{X_i} = \boldsymbol{x_i}) = S_0(t_i)^{\exp(\mu_i)}$, where S_0 corresponds to the baseline survival distribution.

When a parametric distribution is assumed for S_0, like one of those introduced in Section 1.2.3, estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \cdots, \hat{\beta}_m)$ of $\boldsymbol{\beta}'$ are obtained by maximizing the parametric likelihood function given in Equation (1.1).

Otherwise, in the context of right-censored data, Cox proposed to leave the baseline hazard function h_0 unspecified. This results in the so-called "semi-parametric PH model", often referred to as the Cox model. As the baseline hazard can be treated as a nuisance function (parameters $\boldsymbol{\beta}$ are not involved in h_0), an estimation of the coefficients $\boldsymbol{\beta}$ can be obtained. However, maximizing the likelihood function given Equation (1.1) is not possible anymore and Cox developed a partial likelihood to estimate the parameters $\boldsymbol{\beta}$ [Cox, 1972]. As for the Kaplan-Meier estimator, the ordering of the event times is used in the estimation method. In the presence of interval-censoring, other techniques need to be used. For example, [Finkelstein, 1986] uses the score function to obtain estimates. Other references are [Kooperberg and Clarkson, 1997, Goetghebeur and Ryan, 2000, Betensky et al., 2002, Cai and Betensky, 2003].

In the Cox model with right-censoring only, an estimation of the baseline cumulative hazard function is due to Breslow [Breslow, 1972]. As in section 1.2.4., note $s_1 < \cdots < s_D$ the ordered unique event times and d_j the number of events at s_j. Then, given the estimates $\hat{\beta}$, an estimation of H_0 is obtained as

$$\hat{H}_0(t) = \sum_{j|s_j<t} \hat{h}_0(s_j)$$

where $\hat{h}_0(s_j) = \frac{d_j}{\sum_{k\in R_j} \exp(x'_k\hat{\beta})}$ and where R_j represents the set of individuals still at risk just before time s_j.

Accelerated Failure Time (AFT) model.

The Accelerated Failure Time (AFT) model states that

$$\log(T_i) = \boldsymbol{X_i'}\boldsymbol{\beta} + \sigma\varepsilon$$

where σ is a scale parameter, and ε an error term with survival distribution S_0. Note that, while S_0 represents an equivalent to the baseline survival function in the PH model, the effect of a covariate in the AFT model is quite different than in the PH model. In fact, in the AFT model, the conditional survival distribution $S(t_i) \equiv S(t_i|\boldsymbol{X_i} = \boldsymbol{x_i}) = S_0((\log(t_i) - \boldsymbol{x_i'}\boldsymbol{\beta})/\sigma)$.

In a parametric context, i.e. assuming a parametric distribution for S_0, estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m)$ and $\hat{\sigma}$ of $\boldsymbol{\beta}'$ and σ are obtained by maximizing the parametric likelihood function given in Equation (1.1).

In a semi-parametric context, i.e. leaving S_0 unspecified, a partial likelihood treating S_0 as a nuisance function is not available. Indeed, the parameters $\boldsymbol{\beta}$ and σ can not be separated from the survival distribution $S_0((\log(t) - \boldsymbol{x}'\boldsymbol{\beta})/\sigma)$. With right-censored data, estimating equations approaches based on linear-rank tests are used [Kalbfleisch and Prentice, 2002]. Their extension to interval-censored data were studied by [Rabinowitz et al., 1995] and [Betensky, 2001], but these methods are computationnaly applicable only for low-dimensional covariates. To allow for higher dimensional covariates, [Tian and Cai, 2006] proposed a Markov Chain Monte Carlo approach. More recently, a smoothing approach was developed and implemented in a R-package, smoothSurv [Komárek et al., 2005]. Another flexible method, presented in our context in Chapter 5, is the semi-nonparametric representation [Zhang and Davidian, 2001, Zhang and Davidian, 2008] .

The interpretation of the estimated coefficients in an AFT model is similar to the interpretation of coefficient in a linear regression: a covariate acts directly on the survival time T. A positive coefficient implies that a larger value of the corresponding covariate will decelerate the time until the event occurs, i.e. the event occurs later in time, while a negative coefficient shortens the time to the event, i.e the event occurs earlier.

In the following, $\boldsymbol{\eta}$ is the vector of unknown parameters, where $\boldsymbol{\eta}' = (\boldsymbol{\theta}', \boldsymbol{\beta}')$ and $\boldsymbol{\theta}$ represents the vector of other parameters, for example the parameter of the baseline distribution $\boldsymbol{\theta}_0$ and/or the scale parameter σ in the AFT model.

1.3 Survival analysis in the presence of a cure fraction

1.3.1 The mixture cure model

Let $y_1, \cdots, y_n$ be realizations of the indicator random variables $Y_1, \cdots, Y_n$, representing the status, cure or uncured, of an observation, i.e. $Y_i = \mathbb{1}(T_i < +\infty)$. Note that the uncured status y_i is unknown for a right-censored observation and that, by definition, an interval-censored observation corresponds to an uncured subject, i.e. $y_i = 1$. The status Y_i may depend on the covariate vector $\boldsymbol{Z}_i' = (1, Z_{i1}, \cdots, Z_{is}) \in \mathbb{R}^{(s+1)}$ through the parameter ν_i, and we assume the effect of covariates to be linear, that is $\nu_i = \boldsymbol{z_i'\gamma}$, where $\boldsymbol{\gamma}' = (\gamma_0, \gamma_1, \cdots, \gamma_s)$ is a $(s+1)$-vector of unknown coefficients.

The mixture cure model assumes that

$$S(t_i) = p_i S_u(t_i) + 1 - p_i,$$

where $S(t_i) = P(T_i \geq t_i | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{z}_i)$ is the improper conditional survival distribution of T_i given $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, where improper means that $\lim_{t \to +\infty} S(t) > 0$; $p_i = P(Y_i = 1 | \boldsymbol{Z}_i = \boldsymbol{z}_i)$ denotes the conditional probability to experience the event of interest given $\boldsymbol{Z}_i$; and $S_u(t_i) = P(T_i \geq t_i | \boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = 1)$ denotes the proper conditional survival distribution of the uncured sub-population given $\boldsymbol{X}_i$. Note that in the absence of cure (i.e. $p_i = 1 \, \forall i$), S coincides with S_u.

In the incidence part of the mixture cure model, the probability p_i to experience the event is usually modelled by a logistic regression, i.e.

$$p_i = \frac{\exp(\boldsymbol{z}_i'\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i'\boldsymbol{\gamma})}.$$

Note that another model can be used, such as a probit regression for instance.

In the latency part of the mixture cure model, a regression model is used to model to dependence of the time-to-event for the uncured on the covariates $\boldsymbol{X}$. In this thesis, the model is either the PH or AFT model, and the survival distribution for the uncured is then given by one of the two following models:

$$S_u(t_i) = \begin{cases} S_{u,0}(t_i)^{\exp(\mu_i)} & \text{(PH)} \quad (1.2a) \\ S_{u,0}\left(\dfrac{\log(t_i) - \mu_i}{\sigma}\right) & \text{(AFT)} \quad (1.2b) \end{cases}$$

where $\mu_i = \boldsymbol{X}_i'\boldsymbol{\beta}$ and $S_{u,0}$ represents the baseline (PH model) or error term (AFT model) distribution for the uncured individuals. Note that the covariates in $\boldsymbol{X}$ may share some or all of the components of $\boldsymbol{Z}$.

The vector of all unknown parameters to be estimated is $\boldsymbol{\eta}$, where $\boldsymbol{\eta}' = (\boldsymbol{\theta}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ and $\boldsymbol{\theta}$ represents the vector of other parameters such as the parameter of the baseline distribution $\boldsymbol{\theta}_0$ and/or the scale parameter σ in the AFT model. The observed dataset for an individual i is $\mathcal{O}_i = (x_i, z_i, l_i, r_i, \delta_i)$, $i = 1, \cdots, n$, and $\mathcal{O}$ represents the observed data for the entire sample.

1.3.2 Parametric mixture cure model

The parametric mixture cure model assumes that the distribution $S_{u,0}$ of the uncured is known up to a finite number of parameter. Maximum likelihood theory can be applied. An interval-censored observation is uncured with probability p_i and its contribution to the likelihood is therefore $p_i(S_u(l_i) - S_u(r_i))$. A right-censored observation is either uncured, with probability p_i, or cured, with probability $1 - p_i$. Its contribution to the likelihood is $p_i S_u(l_i) + (1 - p_i)$. The likelihood in a parametric mixture cure model with interval- and right-censored observations is proportional to:

$$L(\boldsymbol{\eta}, \mathcal{O}) = \prod_{i=1}^{n} \left[p_i(S_u(l_i) - S_u(r_i))\right]^{\delta_i} \left[(1 - p_i) + p_i S_u(l_i)\right]^{(1-\delta_i)}. \tag{1.3}$$

Identifiability of a mixture cure model was studied in extensive details by [Li et al., 2001] and [Hanin and Huang, 2014]. Given the parametric form of the latency and incidence parts as well as the presence of covariates in the incidence part, the model given in Equation (1.3) is identifiable if the follow-up period is long enough [Li et al., 2001, Hanin and Huang, 2014]. In Chapter 2, simulations show the impact of a follow-up that is not long enough. Maximizing the above likelihood function with respect to $\boldsymbol{\eta}$ leads to a consistent and asymptotically efficient estimate $\hat{\boldsymbol{\eta}}' = (\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ [Casella and Berger, 2001]. The corresponding asymptotic variance-covariance matrix can be obtained as usual through the Hessian matrix. The plug-in method then leads to the estimators $\hat{\boldsymbol{\theta}}$, $\hat{\mu}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, $\hat{\nu}_i = \boldsymbol{z}_i'\hat{\boldsymbol{\gamma}}$, $\hat{p}_i \equiv p(\hat{\nu}_i)$, $\hat{S}_u(t_i) \equiv S_u(t_i|\hat{\mu}_i, \hat{\boldsymbol{\theta}})$ and $\hat{S}(t_i) = \hat{p}_i \hat{S}_u(t_i) + 1 - \hat{p}_i$ of $\boldsymbol{\theta}$, μ_i, ν_i, p_i, $S_u(t_i)$ and $S(t_i)$, respectively.

1.3.3 Semi-parametric mixture cure model

The semi-parametric mixture cure model leaves the survival distribution of the uncured unspecified. In this case, $S_{u,0}$ is considered as a nuisance, and the fact that the uncured status is not known does not allow to use classical methods developed in the context without a cure fraction. Typically, in this case, the Expectation-Maximization (EM) algorithm is used for estimation purposes (see

for instance, for right-censored data: [Taylor, 1995, Sy and Taylor, 2000, Peng and Dear, 2000, Li and Taylor, 2002, Zhang and Peng, 2007, Cai et al., 2012]).

Denote the complete dataset by $\mathcal{O}_c$, which is the observed data augmented by the uncured status y_i, i.e. $\mathcal{O}_{i,c} = (\mathcal{O}_i, y_i)$, $i = 1, \cdots, n$. The likelihood based on this dataset is called the complete likelihood, and three types of observation give a different contribution to this likelihood. First, the contribution of an interval-censored observation, which is by definition uncured, i.e. $y_i = 1$, is $p_i(S_u(l_i) - S_u(r_i))$. Second, the contribution of a right-censored observation that is uncured, i.e. $y_i = 1$, is $p_i S_u(l_i)$. Third, the contribution of a right-censored observation that is cured, i.e. $y_i = 0$, is $1 - p_i$. This gives the *complete* log-likelihood:

$$\begin{aligned} l_C(\boldsymbol{\eta}; \mathcal{O}_c) = &\sum_{i=1}^{n} \delta_i y_i \log\left[p_i(S_u(l_i) - S_u(r_i))\right] \\ &+ (1-\delta_i)(1-y_i)\log\left[1-p_i\right] \\ &+ (1-\delta_i) y_i \log\left[p_i S_u(l_i)\right] \\ &= l_{inc}(\boldsymbol{\gamma}; \mathcal{O}_c) + l_{lat}(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathcal{O}_c) \end{aligned}$$

where the indexes *inc* and *lat* stand for incidence and latency respectively, and

$$\begin{aligned} l_{inc}(\boldsymbol{\gamma}; \mathcal{O}_c) &= \sum_{i=1}^{n} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \\ l_{lat}(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathcal{O}_c) &= \sum_{i=1}^{n} \left[y_i \delta_i \log((S_u(l_i) - S_u(r_i)) + y_i(1-\delta_i)\log(S_u(l_i))\right]. \end{aligned}$$

The EM algorithm is an iterative algorithm with two steps. The expectation step computes the conditional expectation of the complete likelihood l_C, with respect to the (partially observed) uncured status y_i, given the current value of $\boldsymbol{\eta}$, say $\boldsymbol{\eta}^{(k)}$, and the observed data. For this, one needs to compute the expectation of the uncured status, given $\eta^{(k)}$ and $\mathcal{O}$, that is:

$$\xi_i^{(k)} = E\left(Y_i | \boldsymbol{\eta}^{(k)}, \mathcal{O}_i\right) = \delta_i + (1-\delta_i)\frac{\hat{p}_i^{(k)} \hat{S}_u^{(k)}(l_i)}{\hat{p}_i^{(k)} \hat{S}_u^{(k)}(l_i) + 1 - \hat{p}_i^{(k)}}$$

where $\hat{p}_i^{(k)}$ and $\hat{S}_u^{(k)}$ are estimated at the current values of $\boldsymbol{\eta}^{(k)}$. Note for future reference that to compute $\hat{S}_u^{(k)}$, one also needs an estimator of $S_{u,0}$.

The second step of the algorithm, the M-step, maximizes the expected complete-data likelihood given $\xi_i^{(k)}$:

$$\begin{aligned} E(L_{inc}) &= \sum_{i=1}^{n} \left[\xi_i \log(p_i) + (1-\xi_i)\log(1-p_i)\right] \\ E(L_{lat}) &= \sum_{i=1}^{n} \left[\delta_i \xi_i \log(S_u(l_i) - S_u(r_i)) + \xi_i(1-\delta_i)\log(S_u(l_i))\right] \\ &= \sum_{i=1}^{n} \left[\delta_i \log((S_u(l_i)^{\xi_i} - S_u(r_i)^{\xi_i}) + (1-\delta_i)\log((S_u(l_i)^{\xi_i}))\right] \end{aligned}$$

where the last line results from the fact that $\delta_i = 1 \Rightarrow \xi_i = 1$. It follows that $E(l_{lat})$ is actually the likelihood of a standard AFT model without cure, except for the constant term ξ_i.

In the context of right-censored data, an estimation of $S_{u,0}$ is given by the Breslow estimator, so that parameter estimates can be obtained in the PH or AFT model [Taylor, 1995, Sy and Taylor, 2000, Peng and Dear, 2000, Li and Taylor, 2002, Zhang and Peng, 2007]. These methods are implemented in the R-package smcure [Cai et al., 2012].

The extension of these method to interval-censored data is quite challenging, as the Breslow estimator of $S_{u,0}$ is not available [Lesaffre et al., 2005].

1.3.4 The promotion time cure model, an alternative to the mixture cure model

The promotion time cure model is another class of models dealing with the presence of a cure fraction. It was motivated by biological considerations and defined in a Bayesian context by Chen [Chen et al., 1999]. Other authors have developed this model in a more general context [Yakovlev et al., 1996, Tsodikov, 1998, Tsodikov, 2001, Tsodikov et al., 2003]. To take the cure fraction into account, the promotion time cure model imposes an upper bound on the cumulative hazard function of the population, i.e.

$$\lim_{t \to +\infty} (H(t)) = \theta.$$

This model is hence also called the Bounded Cumulative Hazard (BCH) model. The survival distribution of the population is given by

$$S(t) = \exp(-\theta F(t))$$

where $F(t) = H(t)/\theta$ is the distribution function of a continuous non-negative random variable satisfying $F(\infty) = 0$, i.e. F is proper. Taking the limit of the survival distribution gives the cure fraction: $\lim_{t\to+\infty} S(t) = exp(-\theta)$. When covariates $\boldsymbol{X}$ are present, a proportional hazards model can be assumed and is obtained by letting $\boldsymbol{\theta} = exp(\boldsymbol{X\beta})$.

An advantage of this model over the mixture cure model is that it preserves the proportionality of hazards. However, it does not allow to estimate the impact of covariates on the probability to be cured and on the time-to-event separately.

1.4 Diagnostic checks in survival analysis

Building a model does not only consist in estimating the regression parameters. Indeed, the assumptions that are made about this model need to be validated. A commonly used approach is to plot the residuals of the model and is called "diagnostic checks". For instance, a QQ-plot or a histogram of the residuals can be plotted to verify the Normality assumption in classical linear regression. Several types of residuals have been defined in linear regression and in generalized linear regression [Cook and Weisberg, 1983, McCullagh and Nelder, 1989, Weisberg, 2005, Seber and Lee, 2012].

Survival analysis makes no exception, but the presence of censoring makes the definition of residuals less clear. This section aims at presenting some of the most commonly used residuals used for diagnostic checks in a right-censored context [Collett, 2003]. While an extension to interval-censored data in a PH model has been studied [Farrington, 2000], an extension for the interval-censored AFT model has not been developed. Chapter 4 of this thesis focuses on studying the definition and extension of residuals in the context of interval-censored data with a fraction of cured individuals, for the PH and the AFT model.

1.4.1 The Cox Snell residuals

The Cox-Snell residuals are defined as

$$r_{CS}(t_i) = -log(\hat{S}_i(t_i)).$$

where $\hat{S}$ is the estimated survival distribution of T. A right-censored observation results in a right-censored Cox-Snell residual: $r_{CS}(l_i)$; and an interval-censored observation results in interval-censored residuals: $]r_{CS}(l_i), r_{CS}(r_i)]$.

If the model is correctly fitted, the Cox-Snell residuals should resemble a right-censored sample from a unit exponential distribution. Indeed, $-\log(S_i(T_i))$ follows an exponential distribution of parameter one. To see this, first recall that if X is a continuous random variable with distribution function F, then $Y = F(X)$ follows a uniform distribution on $[0, 1]$. Then, remark that

$$T_i \sim S_i \Leftrightarrow S_i(T_i) \sim U[0,1] \Leftrightarrow -log(S_i(T_i)) \sim Exp(1),$$

as $P(-\log(S(T)) < c) = P(S(T) < e^{-c}) = e^{-c}$. A plot of the Kaplan-Meier (for right-censored data) or Turnbull (for interval-censored data) survival estimator of the residuals should therefore reveal points aligned on a straight line of slope one and zero intercept.

While very popular, the Cox-Snell residuals have some serious drawbacks in the Cox model. For instance, the justification of the closeness of residuals to an exponential distribution is based on true values, and not on a nonparametric estimation. This closeness is also not suitable for small sample sizes [Lagakos, 1981], and a departure from a straight line may be due to the uncertainty in estimates (rather than to a wrong model), which is larger in the right-tail and for small sample sizes [Klein and Moeschberger, 2003]. In fact, in the semi-parametric PH model, those residuals have not proved to be useful because to show departure from a straight line, the fitted model has to be really wrong [Collett, 2003]. Additionally no indication is given as regards the cause of the departure, if there is any.

1.4.2 The martingale residuals

The martingale residuals in a context of right-censored and exact observations are defined as

$$r_M(t_i) = \delta_i - r_{CS,i}$$

where $\delta_i = 1$ stands for an exact observation, and $\delta_i = 0$ stands for a right-censored observation, as usual.

For interval-censored data, [Farrington, 2000] proposes to replace the interval residuals with their expected value under the exponential distribution, i.e.

$$r_M(l_i, r_i) = \frac{\hat{S}(l_i)\log(\hat{S}(l_i)) - \hat{S}(r_i)\log(\hat{S}(r_i))}{\hat{S}(l_i) - \hat{S}(r_i)}.$$

The drawback of this extension is that the expectation is taken under an *assumed* distribution, which may be seriously wrong in some cases.

The martingale residuals play a crucial role in survival analysis, as they are viewed as the difference between the observed number of events for an obser-

vation, and its expected number of events. They can thus be used to detect departure from linearity of a covariate in a regression modelling, when plotted against the corresponding covariate [Therneau et al., 1990]. A shortcoming is that they are skewed and not symmetrically distributed, which may result in a laborious detection of non-linearity in the residuals plots. The use of deviance residuals is then advocated.

1.4.3 The deviance residuals

From a general point of view, the deviance is defined as twice the difference between the log-likelihood of the model and the corresponding saturated likelihood, which is the likelihood of the same model, but with the maximum number of parameters. The deviance residuals are defined as the contribution of each observation to the deviance. A small deviance indicates that the model fits the data correctly, whereas a large deviance indicates a problem in the fit. As with martingale residuals, a plot of the deviance residuals versus a covariate should help to detect if a covariate was erroneously entered linearly in the model.

In a PH model with only right-censored and exact observations, it has been shown that the deviance residuals are given as a transformation of martingale residuals, i.e.

$$r_{d,i} = sign(r_{M,i})[-2(r_{M,i} + \delta_i \log(\delta_i - r_{M,i}))]^{1/2}.$$

However, such an expression in terms of martingale residuals is no longer appropriate if the event times are interval-censored, or if the AFT model is assumed instead of the PH model.

The extension for the interval-censored PH model is given in [Farrington, 2000]. The AFT case is covered from the perspective of the mixture cure model in Chapter 4. Indeed, in this chapter, we propose an extension of the deviance residuals in the PH and AFT model, for the case of interval-censored data with a fraction of cure.

1.5 Variable selection methods

When building a model, one can be tempted to include as much covariates as possible. However, sparsity may be preferred for ease of interpretation and prediction, hence the need for a variable selection procedure. In the statistical literature, several variable selection methods are proposed, and the most common ones are briefly presented hereinafter. For a more comprehensive review,

see [Harrell, 2001, Collett, 2003, Hosmer et al., 2011, Bühlmann and van de Geer, 2011].

1.5.1 The best subset selection

The best subset selection computes all possible models, according to all combinations of the covariates. Note that this implies that 2^m models have to be fitted, with m the number of covariates in the model. A comparison between the models can be made with the AIC or BIC criterion:

$$AIC(k) = \hat{l}_k(\hat{\boldsymbol{\eta}}, \mathcal{O}) - k,$$

$$BIC(k) = 2\hat{l}_k(\hat{\boldsymbol{\eta}}, \mathcal{O}) - \log(n)k,$$

where $\hat{l}$ is the maximized log-likelihood of one of the fitted models, with subscript k indicating the number of covariates in the model. The model leading to the largest AIC or BIC is chosen. Obviously, this method implies a large number of models to be fitted, if the number of potential covariates, m, is large. More parsimonious methods include forward, backward and stepwise selection.

1.5.2 Forward, backward and stepwise selection

Forward selection starts from a null model, containing no covariates, and adds one covariate at a time in the model. The covariate that, when included in the model, gives the largest increase of likelihood (or the smallest p-value), is entered in the model. The process is repeated until no further covariate can increase the likelihood by a significant predetermined amount (or until no further covariate is significant at a predetermined α-level).

Backward selection starts from the full model, containing every covariates, and deletes one at a time from the model. The covariate that, when removed from the model, gives the smallest decrease in the likelihood (or largest p-value) is deleted from the model. The process is repeated until the next covariate to be removed decreases the likelihood by more than a predetermined amount (or when all covariates are significant at a predetermined α-level).

The stepwise selection procedure is very similar, except that a covariate may be entered or deleted at each stage of the process: in stepwise forward selection, a covariate that has entered the model can be removed if it is no longer significant. In stepwise backward, a deleted covariate can re-enter the model if its addition is significant.

Here are listed the major drawbacks of these methods [Harrell, 2001]:

- The R^2 values in linear regression are biased high

- The F and χ^2 test statistics do not have the claimed distribution anymore
- The method results in falsely narrow confidence intervals
- The method yields too small p-values, which do not take account of the multiple comparison problem
- The estimated parameters are biased and need shrinkage

Moreover, the method is unstable: small changes in the data can lead to large changes in the selection process [Breiman, 1996].

1.5.3 Shrinkage methods

Shrinkage methods are based on a penalized likelihood function. The idea is to shrink the estimates toward zero, therefore performing estimation and variable selection at the same time. These procedures are continuous in the sense that the coefficients are continuously shrunk towards zero. Discrete procedures, on the other hand, enter or delete one variable at a time.

The idea is that the following penalized log-likelihood must be minimized:

$$-l(\boldsymbol{\eta}, \mathcal{O}) + n\lambda \sum_{j=1}^{m} p_j(|\beta_j|), \tag{1.4}$$

where l is a log-likelihood function such as the one given in Equation (1.1). In the second term of (1.4), λ represents the penalty term (the tuning parameter), controlling for the amount of shrinkage of the estimates. If it is equal to zero, then minimizing (1.4) leads to the usual unpenalized MLE; otherwise, the coefficients are shrunk towards zero. The function $p_j(|\cdot|)$ is the penalty function and can take several forms. For example,

- the Least Absolute Shrinkage and Selection Operator (LASSO) penalty [Tibshirani, 1996]:
$$p_j(|\beta_j|) = |\beta_j|,$$
- the adaptive LASSO penalty [Zou, 2006]:
$$p_j(|\beta_j|) = |\beta_j| w_j,$$
with $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ being a known fixed weight vector.
- the ridge penalty [Hoerl and Kennard, 1970]:
$$p_j(|\beta_j|) = \beta_j^2,$$

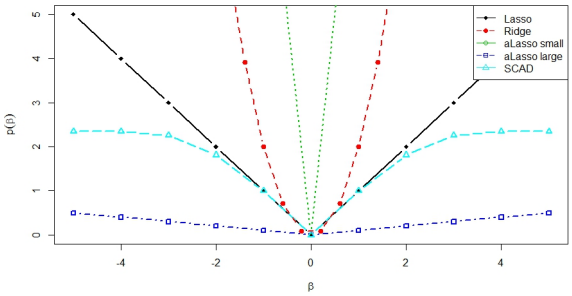


Figure 1.5 – Some penalty functions: Lasso, Ridge, Adaptive Lasso with either a small and large unpenalized coefficient estimate (i.e. w_j), and SCAD.

- the Smoothly Clipped Absolute Deviation (SCAD) penalty [Fan and Li, 2001], which depends directly on λ. This penalty is characterized by its first derivative, i.e.:

$$p'_{\lambda,j}(|\beta_j|) = \lambda(\mathbb{1}(|\beta_j| \leq \lambda)) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda}\mathbb{1}(|\beta_j| > \lambda),$$

  for $a > 2$, where $\mathbb{1}$ represents the indicator function and $x_+ = min(0, x)$.

An illustration of some penalty functions is given in Figure 1.5.

There has been extensive work about penalized likelihood and shrinkage methods. Many algorithms have been proposed to solve the optimization problem, like the LARS algorithm [Efron and Hastie, 2004], the coordinate descent algorithm [Friedman and Hastie, 2007], the unified algorithm with quadratic approximation [Fan and Li, 2001], or the recently proposed algorithm dedicated to the Cox model [Goeman, 2010]. The main disadvantage of the LASSO is that it does not possess the oracle property. The oracle property states that the variable selection method identifies the right subset of covariates and has an optimal estimation rate, i.e. $\sqrt{n}(\tilde{\beta} - \beta) \rightarrow \mathcal{N}(0, \Sigma)$, where $\tilde{\beta}$ represents the coefficients estimated with the variable selection procedure, β are the coefficients of the true model, and Σ is the covariance matrix knowing the true model [Zou, 2006]. The adaptive LASSO does possess the oracle property, as long as the weights w_j are data-dependent and wisely chosen [Zou, 2006]. The SCAD penalty also possesses the oracle property [Fan and Li, 2001].

Note that the construction of p-values or confidence intervals of the (adaptive) LASSO estimates is not straightforward and is actually mainly based on resampling techniques. It has recently been studied by Tibshirani and co-authors for the LASSO [Lockhart et al., 2014].

In the right-censored Cox model, the LASSO, adaptive LASSO and SCAD penalty have been studied by several authors [Tibshirani, 1997, Zhang and Lu, 2007, Fan and Li, 2002]. The adaptive LASSO for a right-censored AFT model was covered in [Wang and Song, 2011]. Concerning the mixture cure model, the SCAD was applied to a right-censored mixture cure model with a Cox regression in latency [Liu et al., 2012]. The extension to a parametric interval-censored AFT mixture cure model will be covered in Chapter 3.

1.6 Real Data Analysis: Oxford Project To Investigate Memory and Aging (OPTIMA)

The main motivation of this thesis is to propose methods to analyze a database concerning Alzheimer's disease. The database concerns a study for which 241 healthy elderly volunteers were followed for up to 20 years with regular assessments of their cognitive abilities using the Cambridge Cognitive Examination (CAMCOG). Among them, 91 converted to MCI (37.8%), and the other 150 (62.2%) were right-censored.

One of the objectives of the study was to identify a set of cognitive scores that predict the probability of conversion from healthy to Mild Cognitive Impairment (MCI) stage in elderly subjects. MCI often represents the pre-dementia stage of a neuro-degenerative disorder, including Alzheimer's disease, vascular dementia, or other dementia syndromes, and hence early detection of its onset is of great relevance for patients, carers and government. The CAMCOG score ranges from 0 to 107 with high scores indicating higher abilities. It is comprised of sub-tests including orientation, comprehension, expression, recent memory, remote memory, learning, abstract thinking, perception, praxis, attention, and calculation. Other recorded covariates were the age, the number of years of total education, the gender, the number of folate cells, the total Homo-cysteine (tHcy) rate, the Mini-Mental State Examination (MMSE) score, and the presence or absence of Apolipoprotein E4 (ApoE4) (a gene known to increase the risk to develop Alzheimer disease [Corder et al., 1993]). Figure 1.6 shows the distribution of the sub-test scores, whereas Table 1.1 presents the mean, standard deviation and range of the covariates Age, Education, tHcy, and folate.

Criteria for diagnosis of MCI and control were carried out according to international guidelines. For more details see ([Oulhaj et al., 2009]). To summarize, conversion to MCI was determined by a neuropsychologist at each visit, which took place in average every year and a half. The data are clearly interval-censored since conversion actually occurred between visits, so that the exact date is not known.

These data were previously analyzed by [Oulhaj et al., 2009], considering interval-censoring only and using a semi-parametric AFT model with smoothing error [Komárek et al., 2005]. In that analysis, baseline CAMCOG sub-tests along with other baseline covariates such as age, years of total education, gender, and ApoE4 were used as potential predictors of the probability of conversion from healthy to MCI. They identified three significant covariates. Two of them with a positive impact on time to MCI-conversion: expression and learning scores at baseline, and one with a negative impact: age at baseline. However, there is a biological evidence that a proportion of patients will never convert to MCI [Petersen et al., 2001]. Figure 1.7 shows the Turnbull non-parametric survival estimator, taking interval-censoring into account [Turnbull, 1976]. This curve shows a plateau with only one event after more or less 12.5 years, confirming the possibility that a fraction of the population will never experience the event. The previously fitted model did not take this feature into account. This is why we will analyze the data with a mixture cure model which considers both interval censoring and a cure proportion.

Table 1.1 – Mean, standard deviation (sd) and range for Age, Education, tHcy and folate cells in the Alzheimer's disease database.

	Mean	SD	Range
Age	71.85	8.83	[34,94]
Education	18.79	3.28	[14,28]
tHcy	12.61	3.89	[6,28]
Folate cells	11.51	3.97	[4,20]

1.7 Objectives and outline of the thesis

The main objective of this thesis is to provide appropriate tools to analyze interval-censored data with a fraction of cure.

As we have seen, a semi-parametric AFT mixture cure model is quite complicated, and a parametric approach seems convenient. In this case, the use of a flexible distribution, such as the EGG distribution is a good compromise. A continuous variable selection technique has not been developed in a parametric mixture cure model. Therefore, Chapter 3 will focus on extending the *adaptive LASSO* in the context of a parametric mixture cure model with an AFT regression in the latency. This chapter makes the assumption that the error term in the AFT model follows the EGG distribution.

To provide with diagnostic checks in parametric mixture cure models with interval-censoring, Chapter 4 proposes an adaptation of existing residuals di-

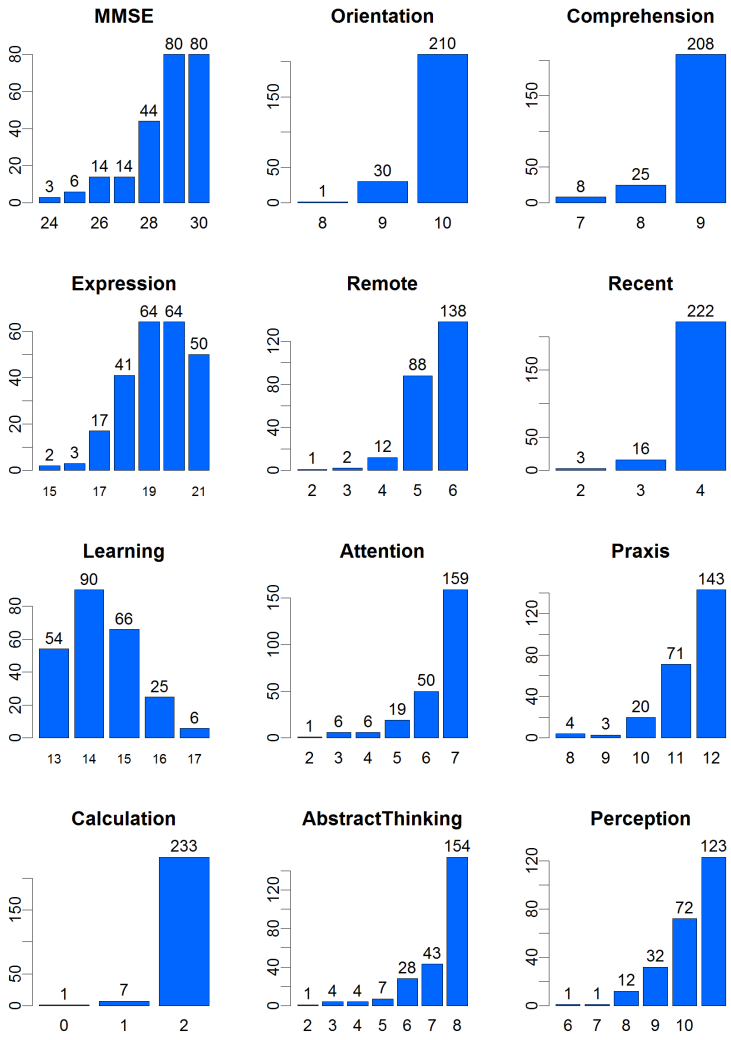


Figure 1.6 – Barplots of selected covariates from the Alzheimer's disease database.

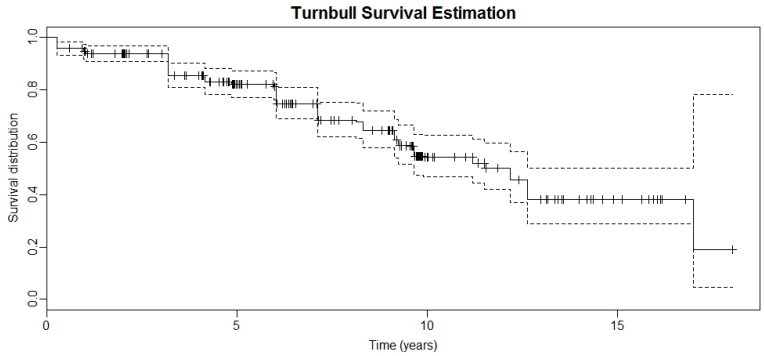


Figure 1.7 – Turnbull Survival Curve, taking interval-censoring into account

agnostic checks, presented in Section 1.4 of this introduction, with either the PH or AFT model in the latency. Firstly, we will define and study two types of Cox-Snell residuals, one type based on the survival distribution of the entire population, and the other type based on the survival distribution of the uncured sub-population. Secondly, to detect a needed transformation in the model and more specifically in the latency and in the incidence parts of the model, we will define three types of deviance residuals: global, latency, and incidence deviance residuals.

Lastly, Chapter 5 will present the semi-nonparametric (SNP) representation, with two objectives: the first objective is to propose an estimation method for a mixture cure model with interval-censoring, without imposing any parametric assumptions on the survival. This is thus an alternative to parametric models, and a more general method than using the EGG distribution. The second objective is to propose a goodness-of-fit test. Indeed, previous findings may suggest a parametric form for the survival distribution. The SNP representation naturally allows to test any parametric distribution, and may therefore confirm or inform these findings.

A discussion about our proposed methods and findings, as well as some ideas for future work are given in Chapter 6.

CHAPTER 2

Taking into account interval-censoring and a cure fraction: a simulation study

"Success is a science; if you have the conditions, you get the result."

- Oscar Wilde (1854-1900)

2.1 Introduction

In this chapter, we show by means of simulations the impact of using naive methods in the presence of interval-censoring and/or a cure fraction. In the context of interval-censoring, the naive approach is to use classical methods by creating a data-set with the mid-point of intervals treated as an exact observation. As for the presence of a cure fraction, one could simply ignore it and consider all cure subjects as right-censored. Both naive solutions may lead to incorrect conclusions and results.

This chapter is organized as follows: Section 2.2 presents the algorithm used to generated right- or interval-censored data, with or without a cure fraction. Section 2.3 present the more specific data generation process used in this chapter. Section 2.4 concerns the impact of the presence of a cure fraction on the comparison of two survival curves, and on regression modelling. The impact of the presence of interval-censoring is discussed in Section 2.5 while a conclusion is given in Section 2.6.

2.2 Data generation: Generation of cured individuals and interval-censored event times

In this chapter and throughout this thesis, we will study the finite sample size properties of our proposed approaches by means of simulations. For these simulations, the data will be generated according to the following algorithm unless otherwise stated. The algorithm can be used to generate a dataset for a mixture cure model, with or without interval-censoring.

1) Generate t_i from an AFT model, considering for example three covariates, i.e.
$$\log(T|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\varepsilon \tag{2.1}$$

2) Compute the cure probability $p_i = p(\boldsymbol{Z}_i)$ through a logistic model considering for example two covariates, i.e.
$$p(\boldsymbol{Z}) = \frac{\exp(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2)}{1 + \exp(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2)}. \tag{2.2}$$

3) Generate y_i, the uncured status, from a Bernoulli random variable of parameter p_i, i.e. $Y_i \sim Bern(p_i)$. If $y_i = 0$, set $t_i = +\infty$.

4a) In case of right-censoring only (no interval-censoring): Generate l_i, the right-censoring times through a given distribution such as the Exponential

distribution with parameter $1/\tau$. If $t_i \leq l_i$ then $\delta_i = 1$ and the observed response is t_i; if $t_i > l_i$ then $\delta_i = 0$ and the observed response is l_i.

4b) In case of interval-censoring and right-censoring, we mimic a real data study and suppose that the observations are patients followed at regular visits. For each subject i, generate a first visit time v_i. Fix a maximum number of visits, say K, and suppose that each visit is separated by a time length a. If $t_i < v_i$, set $l_i = 0$, $r_i = v_i$. Else, if $t_i > v_i + aK$, the observation is right-censored; set $l_i = v_i + aK$, $r_i = \infty$. Otherwise, there exists $k_i = 1, 2, 3, \cdots, K$ such that $v_i + a(k_i - 1) \leq t_i < v_i + ak_i$; in this case set $l_i = v_i + a(k_i - 1)$ and $r_i = v_i + ak_i$.

The values of β_0, β_1, β_2, β_3, σ, γ_0, γ_1, γ_2, a and K, as well as the distribution of X_1, X_2, X_3, ε, Z_1 and Z_2 will be given in each chapter. A R-code producing such a dataset is given in Appendix A.1.

2.3 Data generation for this chapter

The generation of the data used in this chapter may differ according to the aims of the simulations presented in the next sections. Generally, data will be generated following the algorithm described in Section 2.2, where Equations (2.1) and (2.2) are replaced by simpler expressions, respectively:

$$\log(T) = \beta_0 + \beta_1 X + \exp(\alpha)\varepsilon \tag{2.3}$$

$$p(Z) = \frac{\exp(\gamma_0 + \gamma_1 Z)}{1 + \exp(\gamma_0 + \gamma_1 Z)}. \tag{2.4}$$

The error term ε follows an EGG distribution of parameter $q = 1$, i.e. T follows a Weibull distribution, and the covariate X is binary such that $X \sim Bern(0.5)$. In the following, the covariate Z will either be equal to X or have the same distribution, i.e. $Z \sim Bern(0.5)$. Unless otherwise stated, $\beta_0 = 0$ and $\alpha = 0$. The values of β_1, γ_0, γ_1 may differ in each simulation setting discussed hereunder and are given accordingly in each paragraph. The right-censoring distribution is always exponential with mean $1/\tau$, τ taking different values for each setting and given in each following paragraph as well. Unless otherwise stated, the sample size is fixed to $n = 500$ and the number of replication is 500.

2.4 The presence of a cure fraction

Ignoring the presence of a cure fraction and making use of non-adapted tools may results in incorrect conclusions. In survival analysis one is typically in-

terested in either hypothesis testing or regression modelling. Within these contexts, we will try to give an answer to the following questions: "What if I use a cure model when in fact everyone will experience the event?", "What if I use classical methods when in fact there is a cure fraction?" We will also focus on the length of the follow-up period. Indeed, some authors argued that a cure model needs to be used only when there is evidence of a cure fraction, which is usually detected by the presence of a plateau in the estimated non-parametric survival curve [Taylor, 1995, Peng et al., 1998]. This plateau may not be detected if the follow-up period is not sufficiently long. Questions about model identifiability have also been raised when a covariate is present in both the latency and the incidence part of the model, i.e. when $X = Z$, or if only a binary covariate is present in the incidence part. This section tries to investigate this phenomenon.

2.4.1 Survival curves comparison

In the subsection, we suppose that one wants to compare the survival curve for $X = 0$ versus the survival curve for $X = 1$, where X represents one or another treatment for example. We will in fact test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. In this case, some tests like the log-rank test are widely used [Collett, 2003, Kalbfleisch and Prentice, 2002]. In the presence of a cure fraction however, these tests may not perform as expected. Indeed, the presence of cured individuals could create a difference between groups but could also hide an existing difference. It is thus interesting to study the impact of a cure fraction on the power and type-I-error of this test. We will thus first test the null hypothesis with the classical log-rank test which does not account for the presence of a cure fraction. We will then also test the same hypothesis by using the Wald test from a mixture cure model (MCM), assuming an AFT regression with EGG distribution in the latency part. By construction, this test takes the cure fraction into account. To assess the power of the test, β_1 is fixed to 0.5, and to assess the type-I-error, β_1 is set to 0. In the following, the objective is to study the behavior of these two tests.

First, we study the impact of the presence or not of cured observations in the database, i.e. both tests are applied to:

- a dataset in which all observations are uncured, but some are right-censored,
- a dataset in which some observations are cured and thus right-censored, and the cure fraction does not depend on any covariate.

Second, we study the impact of a cure fraction that depends on a covariate Z. Two cases will be covered:

- the covariate Z is **different from** the covariate X in the latency.
- the covariate Z is **equal to** the covariate X in the latency.

The presence or not of cured observations

In this paragraph, data are generated as explained in Section 2.3 to reach two different cases. In the first case, all observations are uncured and the right-censoring proportion varies from 10% to 60%. The right-censoring distribution is exponential with mean $1/\tau$, see Table 2.1 for values of τ leading to the desired right-censoring rates. In the second case, all right-censored observations are cured and the cure rate, thus the right-censoring proportion, varies from 10% to 60%. Table 2.1 also gives the values of γ_0 used to reach these cure rates.

The power of both tests is shown in Figure 2.1a as a function of the right-censoring rate. Concerning the log-rank test when there is no cure fraction in the data, the power decreases slowly as the percentage of right-censoring increases. When there are cured observations, the power decreases drastically as the cure rate increases, indicating the need for more appropriate methods. Concerning the Wald test from the MCM, there is practically no difference in power between both cases, i.e. in the presence or not of a cure fraction.

Regarding the type-I-error: using the log-rank test in both cases, i.e. with or without a cure fraction in the data, leads to a type-I-error that is close to the nominal 5%-level, see Figure 2.1b. Therefore, the presence of a cure fraction does not seem to impact the type-I-error of the log-rank test in this case. Using the Wald test from the MCM in the absence of a cure fraction tends to produce larger type-I-errors when the right-censoring proportion increases. In the presence of a cure fraction, the Wald test is conservative with type-I-errors always below the 5%-level.

In conclusion, the log-rank test lack power when there is a "significant" cure fraction in the data. On another hand, using the Wald test from the MCM in the absence of a cure fraction has a limited impact on the results.

The cure fraction depends on a covariate

This paragraph studies the impact of a cure fraction that depends on a binary covariate Z. The data are generated as in Section 2.3 and two cases are covered. First, the binary covariate Z is equal to X ($Z = X$) and second, Z is different from X ($Z \neq X$) but generated similarly. The cure rate in group Z is fixed to 20% by setting $\gamma_0 = \log(4)$, and the cure rate in group $Z = 1$ varies from 15% to

40%, by setting $\gamma_1 = \log(a/1-a)$ with $a = 0.90, 0.85, 0.80, 0.75, 0.70, 0.65, 0.60$. In this case, the right-censoring distribution is exponential of mean $1/4$.

Figure 2.2a shows the power of both tests as a function of the cure rate in the group where $Z = 1$. When $X \neq Z$, the power of the log-rank and Wald tests decreases when the cure rate increases, which is expected. When $X = Z$, the power of the log-rank test increases when the cure rate increases. Actually, as $\beta > 0$, the event times in group 1 (X=Z=1) tends to appear later than in group 0 (X=Z=0). A cure rate that is larger in group 1 than in group 0 means that more individuals in group 1 have infinite event times. This explains the fact that the difference between the two groups is accentuated when the cure rate in group 1 increases, hence a larger power in this case. A smaller cure rate in group 1 than in group 0 has the opposite effect and leads to a smaller power. Concerning the Wald test in the case $X = Z$, the power of the test is quite stable with respect to the cure rate but is quite low. This is actually explained by the length of the follow-up that may not be sufficiently long (see Section 2.4.2 hereunder for more information). We therefore performed the same analysis with a longer follow-up. For that, we generated the right-censoring distribution through an exponential of mean $1/10$. In this case the power was much larger (see gray curve in Figure 2.2a).

Figure 2.2b shows the type-I-error of both tests as a function of the cure rate in the group where $Z = 1$. When $X \neq Z$, the type-I-error is close to the nominal 5%-level for the log-rank and Wald tests, so that the fact that the cure proportion differs between two groups has no impact on the results. When $X = Z$, the fact that the cure rate differs between the two groups has an impact on the log-rank test. Indeed, it implies a difference between the global survival curves for the two groups, so that the type-I-error is too large when the cure proportion in group $X = Z = 0$ is different from the cure proportion in the other group.

In conclusion, when the cure rate depends on the covariate X, the log-rank test will too often conclude to a significant difference between treatments, leading to type-I-errors above the nominal 5%-level. This difference is however imputable to the difference of the cure rate between treatments. This information should clearly be taken into account.

2.4.2 Regression modelling

In this subsection, we aim to answer several questions and investigate different situations. In particular, we study:

1. the impact of using a mixture cure model in the absence of a cure fraction,
2. the impact of using a classical model in the presence of a cure fraction,

3. the impact of the cure and right-censoring proportion,

4. the impact of the follow-up length, in case a covariate is present in both the latency and incidence part.

In this section, the data are generated following Section 2.3, with $\beta_1 = 0.6$.

Using a cure model in the absence of a cure fraction

Firstly, we want to investigate the impact of using a mixture cure model when the cure fraction is in fact not present. To do this, we generated data as hereinabove but without a cure fraction. The parameter τ of the right-censoring distribution is set to τ =100, 13, 6.5, 4, 2.2, or 1.5 to reach 10%, 20%, 30%, 40%, 50% or 60% right-censored observations respectively.

We first fitted the data with the EGG-AFT model then with the EGG-AFT mixture cure model. The EGG-AFT model as well as the latency part of the mixture cure model give extremely similar results, see Table 2.3. For the incidence part, the estimation of the intercept is very large, leading to an estimated uncured fraction close to 100%. In conclusion, using a mixture cure model in the absence of a cure fraction does not seem to have any consequences on the latency part.

Failing to account for a proportion of cured observations

Secondly, we investigate the opposite situation: not taking the cured fraction into account when there actually are cured observations. We generated data with a fraction of cured observations and compared a model that takes cure into account with a model that does not. The right-censoring proportion is fixed at 60%, and the cure proportion increases from 10% to 60%. The values of τ, γ_0 and γ_1 used to reach these proportions are given in Table 2.2.

We first fitted the data with an AFT model with an EGG distribution for the error term (EGG-AFT model), then with the mixture cure model with an EGG-AFT model in the latency. Table 2.4 gives the parameter estimates in both cases. Concerning the EGG-AFT model, the bias and MSE are increasing when the cure proportion increases. The mixture cure model always gives smaller bias and MSE for the parameters of the latency part. Even when the cure proportion is low, such as 10%, the bias and MSE can be up to four times larger when using the model that does not take cure into account. A visual representation is given in Figure 2.3 in case there is 10% cure.

The impact of the right-censoring proportion

Thirdly, we show the behavior of parameter estimates as a function of the proportion of right-censoring. We discuss the impact of letting the right-censoring proportion vary from 60% to 10%. The data are generated following Section 2.3, and the right-censoring distribution is exponential of mean $1/\tau$. Values of τ, γ_0 and γ_1 used to reach the desired proportion of cured and right-censored observations are given in Table 2.2. We also investigate the impact of the presence of the same covariate in both the incidence and the latency part of the MCM, i.e. the impact of $X = Z$. Indeed, some authors showed that this could lead to identifiability issues [Hanin and Huang, 2014]. We thus perform two similar analyses: one where the data are generated with $X = Z$, and the other one where the data are generated with $X \neq Z$ and $Z \sim Bern(0.5)$. The results were extremely similar and only the case $X = Z$ is thus discussed here.

Table 2.5 shows the bias and MSE when there is 60% to 10% right-censoring, along with a cure proportion varying from 10% to 60%. First, for a fixed right-censoring proportion, the bias and MSE are decreasing for all parameters when the cure rate increases (i.e. from left to right). Second, for a fixed cure proportion, the bias and MSE are decreasing if the right-censoring proportion decreases. When using a mixture cure model, the more right-censored individuals are cured, the more accurate the results will be, in terms of bias and MSE. A visual representation is given in Figure 2.4 for parameters β and γ_1.

The impact of the follow-up length with the same covariate in incidence and latency.

To investigate the identifiability issues discussed in the literature [Hanin and Huang, 2014], we also investigate the impact of the follow-up length when the same covariate is present in both the incidence and latency parts of the model. We performed another simulation study, similar to those herein-above, but we generated the event times and the cure rates with X and Z being continuous covariates instead of binary. X and Z follow a standard Normal distribution. To reach respectively 10%, 20%, 30% and 40% cure rates, we set γ_0 =2.5, 1.7 ,1 or 0.5; and $\gamma_1 = 1$. The right-censoring distribution parameter is set to $\tau = 10$. To assess the impact of the length of the follow-up, we further impose that the study ends after 6, 4 or 2 units of time (i.e. years). To assess to evolution with the sample size, we generated data with n =500, 1000 or 5000.

Table 2.6 shows the results for n =500. We observe larger bias and MSE if the follow-up is short or when $X = Z$. Figure 2.5 shows the evolution with the sample size by means of boxplots representation. In this Figure are represented the estimates for β, γ_0 and γ_1 with $X \neq Z$, $X = Z$ and when the

study is designed to end after 2 units of time. We observe a clear improvement in the variance of the parameter estimates with a larger sample size.

2.5 The presence of interval-censoring

In many cases, interval-censored data are taken into account by using a naive "midpoint imputation" approach. This method creates a data-set by imputing the exact event times by the midpoint of the intervals. By doing so, classical survival techniques can be used.

The main goal of this section is to assess the impact of using this naive method. The first part concerns the comparison of two survival curves, pertaining to two treatments for instance. The second part concerns regression modelling.

2.5.1 Survival curves comparison

The comparison of two survival curves in the presence of interval-censoring has been studied by several authors [Mantel, 1967, Self and Grossman, 1986, Finkelstein, 1986, Gina R. Petroni, 1994, Fay, 1996, Pan, 2000, Fang et al., 2002]. The use of the naive approach, the midpoint imputation approach, in the context of the log-rank test has been analyzed in [Law and Brookmeyer, 1992]. In this paper, the authors conclude that if the length between two visits is the same in each arm of the treatment, then the midpoint imputation approach performs satisfactorily. Otherwise, if there is a difference of interval length, the type-I-error becomes too large when the length difference between the two groups is large. In the simulation study presented hereunder, we reached the same conclusions.

We generated right- and interval-censored data from an AFT model as in Section 2.3 with $\beta = 0$ and without a cure fraction. The length between 2 visits is generated through a uniform random variable $U[0.5, a]$, where a varies from 1 to 12. Note that by construction, the event has a time range of occurrence of approximately 0 to 40 units of time. The results of the type-I-error ($\beta = 0$) using the log-rank test are given in Table 2.7 for 2000 replications. In a first case, the length between two visits is the same for $X = 0$ and $X = 1$. In a second case, the value of a in group $X = 0$ is fixed to 1 and in group $X = 1$, the value is either 2, 4, 6, 8 or 12. Using the log-rank test with the midpoint imputation approach gives very satisfying results if the length between two visits is the same in the two treatments. However, when the difference in interval length between the two groups is large, we observe an increase in the type-I-errors. It is thus recommended to design analyses in which the visits are scheduled according to the same method in each treatment.

2.5.2 Regression modelling

In regression modelling also, a comparison with the mid-point imputation approach has been achieved for parametric modelling [Lindsey, 1998]. The conclusion was that, generally, the midpoint approach gives a satisfying result. In a semi-parametric Cox model, the imputation approach and Finkelstein's approach were compared[Finkelstein, 1986]. The authors conclude that Finkelstein's approach should be preferred [Sun and Chen, 2010]. The goal of this section is twofold. First, we want to investigate the impact of the interval length on the estimates. Second, we want to investigate the behavior of the midpoint imputation approach in a simple regression setting. To do this, we generated data and interval-censored event times following Section 2.3. The length between 2 visits is generated through a uniform random variable $U[0.5, a]$, where a is equal to 0.5, 1, 2, 3, 4 or 5. We also compare three different sample sizes: $n = 100$, $n = 200$ and $n = 500$.

Tables A.1 and A.2 in Appendix A show the results for the six different values of a, and Figure 2.6 shows the 500 estimated survival curves when using the exact likelihood or the midpoint imputation approach, with the true survival curve superimposed (for $n = 500$ and a= 1 and 4). Using the exact likelihood, the bias slightly increases if the interval length increases as well. Using the midpoint, the bias increases much faster. We nevertheless notice a larger MSE for the estimates of the parameter q of the EGG distribution when using the exact likelihood method.

When the length between two visits is large (i.e. $a > 2$), the estimated survival curves using the midpoint imputation approach are biased, whereas the estimated curves with the exact likelihood are not. As the exact likelihood method always outperforms the midpoint imputation approach, we recommend to use this method unless the length of the intervals is small compared to the usual time range of occurrence of the event.

2.6 Discussion

In the presence of cured individuals, the use of a model that does not take cure into account leads to incorrect conclusions about the time to occurrence of the event. Furthermore, the impact of using a mixture cure model rather than a classical model in the absence of a cure fraction is nearly nonexistent as regards the latency part. It is thus essential to use a model such as the mixture cure model in order to obtain unbiased parameter estimates.

On the other hand, Farewell states that those models should be used with caution, and when there is a biological evidence of a possibility of cure [Farewell,

1986]. Regarding the incidence part, the results of our simulation study support this conclusion. When the cure proportion is low compared to the right-censoring rate, the bias and MSE in the incidence part can be large. We also observe that if a covariate is present in both parts of the model, the follow-up length is of even higher importance as well as the sample size. Caution should thus be taken in cases where the follow-up is short and/or the sample size is small when drawing conclusions about the incidence part.

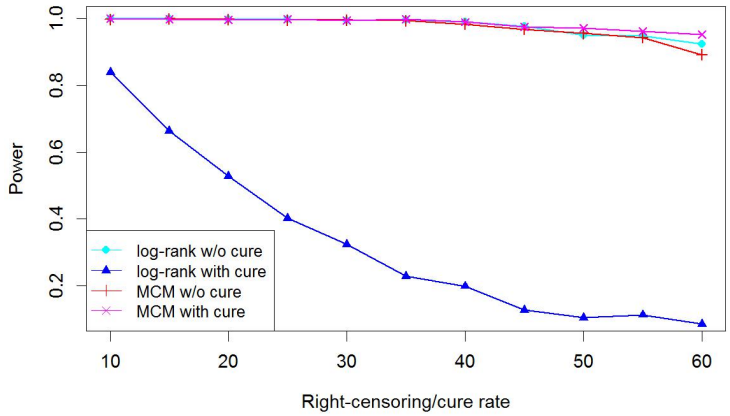
Pertaining to the presence of interval-censoring, the naive approach that consists on imputing data by the mid-point of the interval can be very satisfying in some cases. This may however be due to our particular simulation setting. Indeed, the length between two intervals is generated through a uniform distribution, giving equivalent weight to all values of the interval. It may be interesting to study the behavior of the methods with another distribution. When the length between two visits is large compared to the time of occurrence of the event, the estimated survival can be biased. It seems therefore useful to develop and use methods that take into account the fact that the event times may not be exactly observed but interval-censored. Those methods are usually also applicable in the classical right-censoring context and thus in a general framework for all types of censoring (left, interval, right).

Table 2.1 – Values of τ and γ_0 chosen to reach a given level of right-censoring (RC) or cure rate.

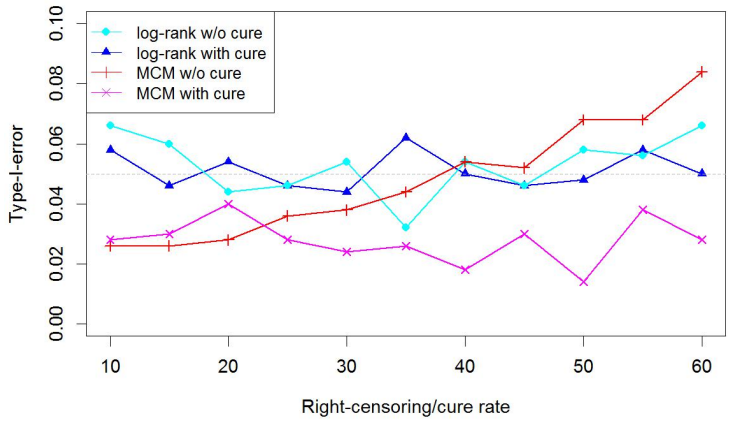
RC (Cure)	τ when $\beta = 0,5$	τ when $\beta = 0$	γ_0
10%	30	20	$\log(0.9/0.1)$
15%	20	15	$\log(0.85/0.15)$
20%	15	10	$\log(0.80/0.20)$
25%	10	8	$\log(0.75/0.25)$
30%	8	6	$\log(0.70/0.30)$
35%	6	5	$\log(0.65/0.35)$
40%	5	4	$\log(0.60/0.40)$
45%	4	3	$\log(0.55/0.45)$
50%	3	2.5	$\log(0.50/0.50)$
55%	2.5	2	$\log(0.45/0.55)$
60%	2	1.5	$\log(0.40/0.60)$

Table 2.2 – Values of τ combined with values of γ_0 and γ_1 chosen to reach a given level of right-censoring (RC) and cure rate.

Cure Rate	γ_0	γ_1	10% RC	20% RC	30% RC	40% RC	50% RC	60% RC
10% Cure	1.7	1	100	13	6.5	4	2.2	1.5
20% Cure	1	0.8		100	10	6	3	2
30% Cure	0.4	1			40	9	4.5	2.5
40% Cure	-0.5	2				50	8	3.5
50% Cure	-0.75	1.5					60	6.5
60% Cure	-1.5	2						25

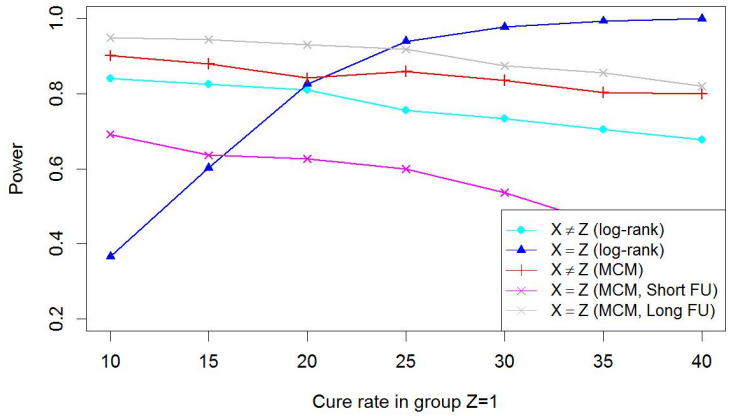


(a) Power

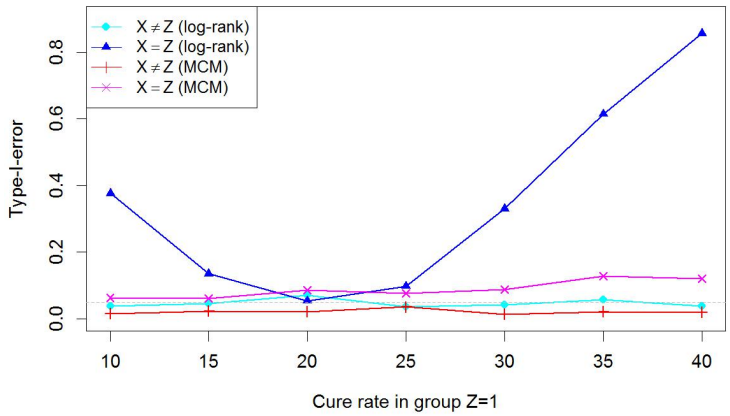


(b) Type-I-error.

Figure 2.1 – Power and Type-I-error as a function of cure or right-censoring rate. All right-censored observations are cured. log-rank: log-test rank without (w/o) or with a cure fraction in the data. MCM: Wald test from a MCM model, without (w/o) or with a cure fraction in the data.



(a) Power



(b) Type-I-error

Figure 2.2 – Power and type-I-error as a function of cure rate in group $Z = 1$. The cure rate in group $Z = 0$ if fixed to 20%. The covariate Z is either equal to X or different from X. log-rank: log-test rank test. MCM: Wald test from a MCM model. FU: Follow-up

Table 2.3 – Parameter estimates when using an EGG-AFT model (top) or a EGG-AFT mixture cure model (bottom) when there is no cured fraction. The true values are $q = 0.5$, $\beta = 0.6$ and $\alpha = 0$.

RC prop.	EGG-AFT model				
	q	β	α		
10% RC	0.469	0.586	-0.009		
20% RC	0.462	0.583	-0.009		
30% RC	0.457	0.580	-0.010		
40% RC	0.448	0.574	-0.010		
50% RC	0.428	0.561	-0.010		
	EGG-AFT mixture cure model				
	q	β	α	γ_0	γ_1
10% RC	0.470	0.586	-0.009	12.927	2.093
20% RC	0.471	0.583	-0.013	11.833	0.709
30% RC	0.474	0.578	-0.019	10.604	0.070
40% RC	0.473	0.567	-0.024	9.859	-0.222
50% RC	0.471	0.540	-0.035	8.585	-0.739

Table 2.4 – Bias and MSE of estimates using a EGG-AFT model without cure (top) or an EGG-AFT Mixture Cure model (bottom). A visual representation is given in Figure 2.3

	EGG-AFT model with 60% RC											
	10% Cure		20% Cure		30% Cure		40% Cure		50% Cure		60% Cure	
q	0.665	0.473	1.134	1.315	1.646	2.738	2.418	5.893	3.013	9.139	4.401	19.54
β	0.140	0.035	0.201	0.058	0.325	0.126	0.679	0.483	0.681	0.494	1.004	1.056
α	-0.207	0.048	-0.347	0.125	-0.455	0.212	-0.514	0.271	-0.632	0.407	-0.752	0.577
	EGG-AFT mixture cure model with 60% RC											
q	-0.167	0.120	-0.187	0.135	-0.171	0.123	-0.135	0.099	-0.085	0.054	-0.068	0.041
β	-0.077	0.030	-0.048	0.028	-0.045	0.028	-0.044	0.023	-0.037	0.019	-0.030	0.016
α	0.010	0.018	0.026	0.020	0.022	0.018	0.005	0.013	-0.012	0.007	-0.017	0.004

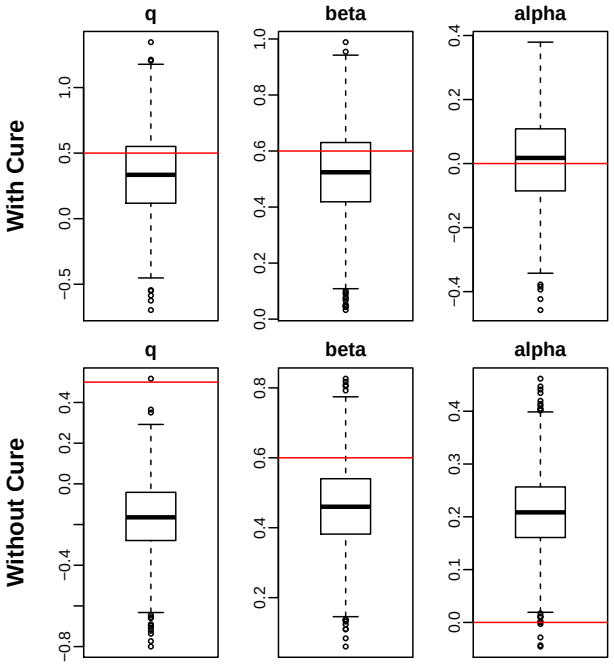


Figure 2.3 – Boxplots of q, β and α estimates for 10% cure rate along with 60% RC. Top: using the EGG-AFT mixture cure model. Bottom: using the EGG-AFT model not taking cure into account. The horizontal line represents the true value of the parameter.

Table 2.5 – Bias and MSE with the EGG-AFT Mixture Cure Model, for several right-censoring and cure rates. A visual representation from β and γ_0 is given in Figure 2.4

	10% Cure		20% Cure		30% Cure		40% Cure		50% Cure		60% Cure	
	60 % RC											
q	-0.167	0.120	-0.187	0.135	-0.171	0.123	-0.135	0.099	-0.085	0.054	-0.068	0.041
β	-0.077	0.030	-0.048	0.028	-0.045	0.028	-0.044	0.023	-0.037	0.019	-0.030	0.016
α	0.010	0.018	0.026	0.020	0.022	0.018	0.005	0.013	-0.012	0.007	-0.017	0.004
γ_0	0.724	3.291	0.209	0.362	0.087	0.093	0.036	0.031	0.010	0.024	-0.010	0.029
γ_1	1.573	12.95	0.926	5.708	0.310	1.549	0.187	0.849	-0.001	0.061	0.009	0.051
	50 % RC											
q	-0.123	0.073	-0.115	0.066	-0.092	0.047	-0.072	0.039	-0.048	0.026		
β	-0.051	0.021	-0.038	0.020	-0.035	0.018	-0.031	0.014	-0.021	0.012		
α	0.005	0.011	0.006	0.010	-0.001	0.007	-0.010	0.005	-0.016	0.003		
γ_0	0.268	0.690	0.082	0.081	0.025	0.029	0.009	0.020	-0.002	0.019		
γ_1	1.468	10.32	0.284	1.683	0.027	0.121	0.011	0.076	0.004	0.039		
	40 % RC											
q	-0.075	0.034	-0.068	0.028	-0.060	0.026	-0.044	0.022				
β	-0.030	0.013	-0.028	0.013	-0.027	0.012	-0.019	0.010				
α	-0.002	0.005	-0.005	0.004	-0.009	0.003	-0.014	0.002				
γ_0	0.080	0.091	0.026	0.033	0.007	0.021	0.001	0.017				
γ_1	0.662	4.409	0.023	0.152	0.003	0.065	0.005	0.047				
	30 % RC											
q	-0.058	0.022	-0.051	0.020	-0.042	0.017						
β	-0.024	0.010	-0.023	0.010	-0.018	0.009						
α	-0.006	0.003	-0.009	0.003	-0.011	0.002						
γ_0	0.045	0.054	0.013	0.026	0.001	0.017						
γ_1	0.209	1.327	0.008	0.087	0.002	0.046						
	20 % RC											
q	-0.042	0.015	-0.034	0.013								
β	-0.019	0.008	-0.015	0.008								
α	-0.009	0.002	-0.011	0.002								
γ_0	0.024	0.041	0.006	0.021								
γ_1	0.042	0.234	0.005	0.057								
	10 % RC											
q	-0.032	0.011										
β	-0.015	0.007										
α	-0.010	0.001										
γ_0	0.017	0.032										
γ_1	0.013	0.115										

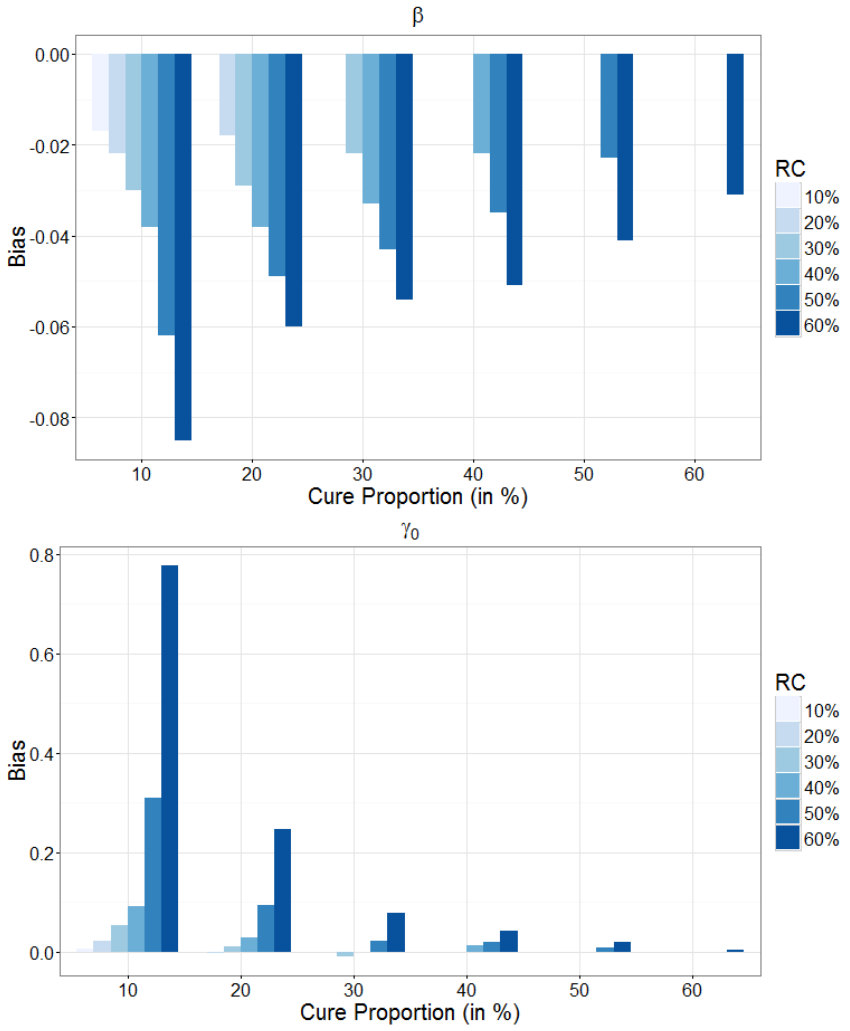
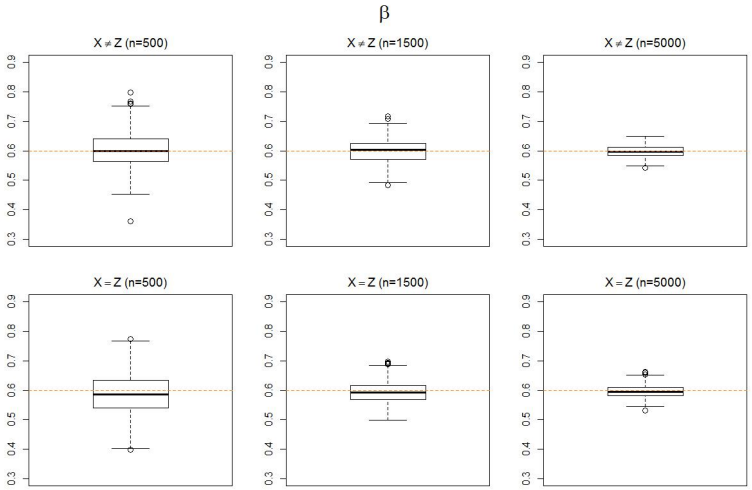


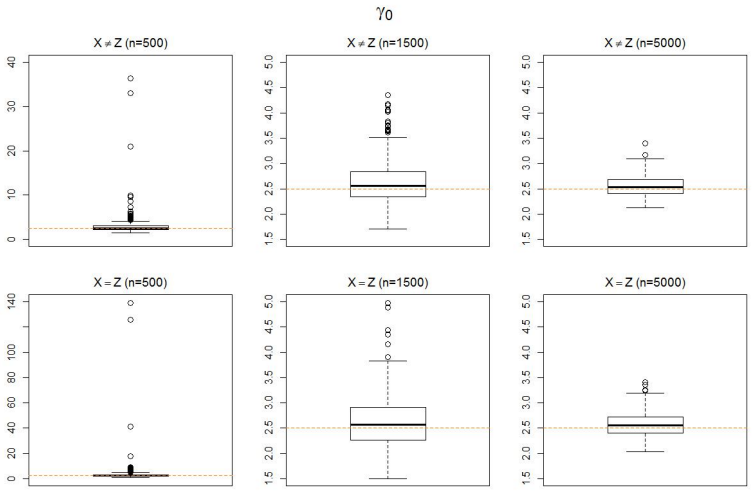
Figure 2.4 – Plots of the bias of β (top) and γ_0 (bottom) as a function of the cure rate and the right-censoring proportion.

Table 2.6 – Bias and MSE of parameter estimates. X and Z are continuous covariates and $X \neq Z$ (top) or $X = Z$ (bottom). The study is designed to end after 6, 4 or 2 units of time. A too short follow-up implies non-stable estimates, in particular when $X = Z$.

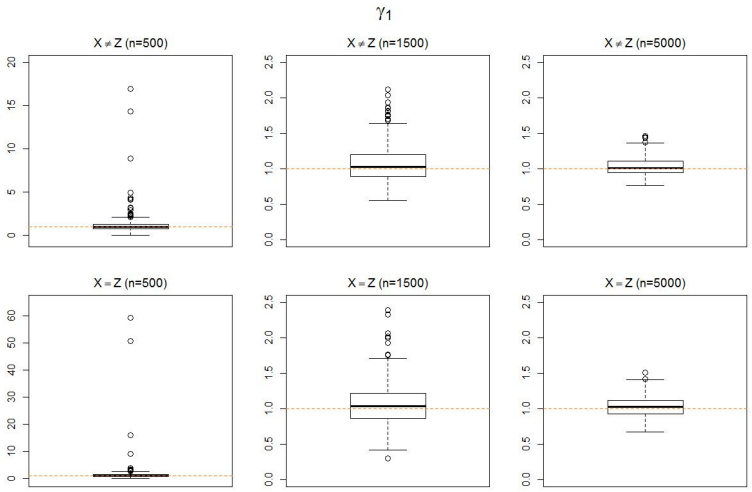
	$X \neq Z$													
			q		β_0		β_1		γ_0		γ_1		α	
	Cure rate	Right-censoring	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
End = 6	10	22	-0,006	0,046	0,001	0,011	-0,005	0,004	0,056	0,084	0,032	0,053	-0,017	0,004
	20	30	-0,015	0,049	-0,004	0,012	0,000	0,004	0,031	0,034	0,025	0,028	-0,014	0,004
	30	40	-0,028	0,054	-0,009	0,014	-0,001	0,005	0,021	0,022	0,021	0,022	-0,015	0,004
	40	47	-0,013	0,068	-0,002	0,017	0,001	0,006	0,008	0,015	0,011	0,022	-0,020	0,005
End = 4	10	25	-0,005	0,060	0,000	0,012	0,000	0,004	0,130	0,291	0,066	0,121	-0,010	0,005
	20	32	-0,013	0,069	0,000	0,013	0,005	0,005	0,053	0,076	0,045	0,050	-0,012	0,006
	30	42	-0,038	0,082	-0,007	0,016	0,000	0,005	0,039	0,035	0,037	0,032	-0,018	0,007
	40	50	-0,045	0,112	-0,009	0,019	0,005	0,007	0,036	0,027	0,033	0,029	-0,012	0,010
End = 2	10	35	-0,024	0,089	-0,008	0,015	0,001	0,004	0,407	5,785	0,195	1,289	-0,012	0,010
	20	40	-0,054	0,114	-0,011	0,019	-0,003	0,004	0,169	0,275	0,123	0,133	0,000	0,011
	30	50	-0,043	0,133	-0,007	0,019	0,004	0,006	0,071	0,117	0,060	0,073	-0,013	0,015
	40	55	-0,021	0,196	-0,007	0,026	-0,007	0,007	0,048	0,078	0,045	0,049	-0,029	0,023
	$X = Z$													
			q		β_0		β_1		γ_0		γ_1		α	
	Cure rate	Right-censoring	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
End = 6	10	22	-0,017	0,040	-0,006	0,009	0,002	0,003	0,058	0,094	0,035	0,067	-0,009	0,003
	20	30	-0,015	0,049	-0,001	0,012	-0,006	0,004	0,029	0,043	0,023	0,042	-0,017	0,004
	30	40	-0,014	0,065	-0,008	0,017	-0,001	0,005	0,016	0,026	0,007	0,030	-0,014	0,005
	40	47	-0,018	0,069	-0,002	0,015	-0,010	0,007	0,007	0,017	-0,005	0,021	-0,018	0,005
End = 4	10	25	-0,006	0,059	-0,006	0,011	-0,006	0,004	0,078	0,299	0,041	0,132	-0,017	0,006
	20	32	-0,047	0,093	-0,023	0,017	-0,002	0,005	0,049	0,118	0,033	0,070	-0,013	0,006
	30	42	-0,055	0,116	-0,018	0,018	-0,005	0,007	0,037	0,054	0,032	0,049	-0,008	0,009
	40	50	-0,066	0,165	-0,018	0,022	-0,005	0,008	0,054	0,058	0,032	0,047	-0,008	0,012
End = 2	10	35	-0,080	0,132	-0,021	0,016	-0,012	0,005	1,037	72,669	0,451	12,711	0,001	0,014
	20	40	-0,093	0,204	-0,026	0,021	-0,013	0,007	0,300	1,393	0,152	0,437	-0,006	0,020
	30	50	-0,158	0,284	-0,028	0,030	-0,015	0,009	0,238	0,471	0,142	0,196	0,012	0,025
	40	55	-0,152	0,408	-0,017	0,033	-0,022	0,012	0,206	0,406	0,130	0,177	0,003	0,043



(a) Parameter estimates for $n = 500$, $n = 1500$ or $n = 5000$ for β. With $X \neq Z$ (top) and $X = Z$ (bottom).



(b) Parameter estimates for $n = 500$, $n = 1500$ or $n = 5000$ for γ_0. With $X \neq Z$ (top) and $X = Z$ (bottom). Larger variance if $X = Z$, which decreases as the sample size increases.

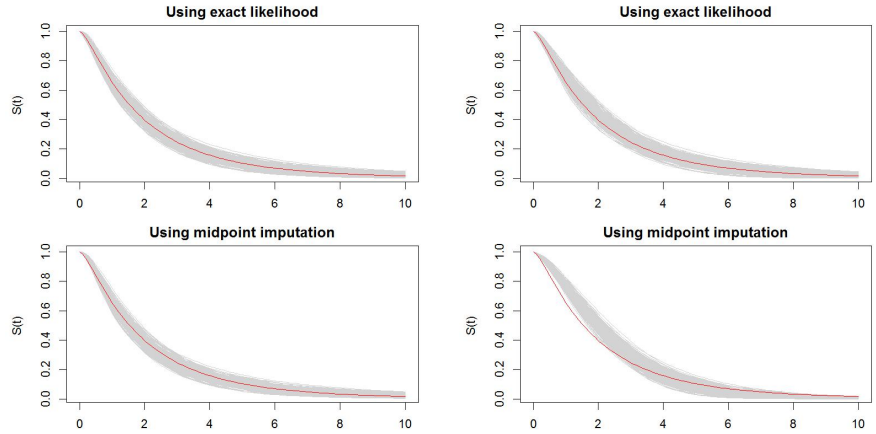


(c) Parameter estimates for $n = 500$, $n = 1500$ or $n = 5000$ for γ_1. With $X \neq Z$ (top) and $X = Z$ (bottom). Larger variance if $X = Z$, which decreases as the sample size increases.

Figure 2.5 – Parameter estimates in case of too short follow-up: evolution with sample size for β, γ_0 and γ_1. The study ends after 2 units of time.

Table 2.7 – Type-I-error (over 2000 replications). The first six rows give the type-I-error when the interval length is generated similarly in each treatment arm. The last five rows give the type-I-error when the interval length differs between the two groups. The first column pertains to the length in group $X = 0$, the second column to the length in group $X = 1$, the third column gives the type-I-error when there is light-censoring (10%) and the last column gives the type-I-error when there is medium censoring (25%). We observe a larger type-I-error when the difference of length between the two groups is large (8 or 12).

a $(X = 0)$	a $(X = 1)$	Light Censoring	Medium Censoring
1	1	0.055	0.062
2	2	0.048	0.040
4	4	0.050	0.042
6	6	0.058	0.047
8	8	0.048	0.056
12	12	0.043	0.051
1	2	0.045	0.047
1	4	0.060	0.054
1	6	0.057	0.064
1	8	0.072	0.071
1	12	0.087	0.067



(a) Length parameter: a =1 (b) Length parameter: a =4

Figure 2.6 – Survival estimation using the exact likelihood (top plots) or midpoint imputation approach (bottom plots) for 500 replications with $n = 500$. The true survival curve is superimposed. Each sub-figure corresponds to a different interval length.

CHAPTER 3

Variable selection: the adaptive LASSO

"Nothing is impossible, the word itself says, "I'm possible"! "

- Audrey Hepburn (1929-1993)

This chapter is mainly based on the paper "Variable selection in a flexible parametric mixture cure model with interval-censored data" By Scolas S., El Ghouch A., Legrand C. and Oulhaj A., Statistics in Medicine 2016; 35(7): 1210-1225.

In standard survival analysis, it is generally assumed that every individual will someday experience the event of interest. However, this is not always the case, as some individuals may not be susceptible to this event. Also, in medical studies, it is frequent that patients come to scheduled interviews and that the time to the event is only known to occur between two visits. That is, the data are interval-censored with a cure fraction. Variable selection in such a setting is of outstanding interest. Covariates impacting the survival are not necessarily the same as those impacting the probability to experience the event. The objective of this chapter is to develop a parametric but flexible statistical model to analyze data that are interval-censored and include a fraction of cured individuals when the number of potential covariates may be large. We use the parametric mixture cure model with an accelerated failure regression model for the survival, along

with the extended generalized gamma for the error term. To overcome the issue of non-stable and non-continuous variable selection procedures, we extend the adaptive LASSO to our model. By means of simulation studies, we show good performance of our method, and discuss the behavior of estimates with varying cure and censoring proportion. Lastly, our proposed method is illustrated with a real dataset studying the time until conversion to Mild Cognitive Impairment, a possible precursor of Alzheimer disease.

3.1 Introduction

The mixture cure model allows a direct interpretation of the effect of covariates on the cure probability and on the survival distribution of the uncured individuals, separately. These two sets of covariates may not necessarily be the same, and the number of potential covariates to be included in each component of the model can be large. Variable selection is thus needed so that the final model possesses good predictability and can easily be interpreted. Classical variable selection methods, like the well known best subset or stepwise selection, suffer from some serious drawbacks. For example, the computational load increases with an increasing number of variable in the model, and the process is discrete and non-stable, as it either enters or deletes a covariate from the model. Several other drawbacks are described by [Fan and Li, 2001, Harrell, 2001]. On the contrary, shrinkage methods, such as the LASSO [Tibshirani, 1996], the SCAD [Fan and Li, 2001], and adaptive LASSO [Zou, 2006] are continuous processes: the general idea is to shrink some coefficients towards zero. This allows to perform simultaneous variable selection and coefficient estimation. Moreover, newly proposed algorithms, such as the LARS algorithm [Efron and Hastie, 2004], the coordinate descent [Friedman and Hastie, 2007] and the unified algorithm with quadratic approximation [Fan and Li, 2001], ensure that the results can be obtained in an efficient way.

Dealing with right-censoring only, the SCAD procedure was extended to a Cox mixture cure model [Liu et al., 2012]. The authors use the fact that a mixture cure model, in which a Cox is assumed in the latency, can be estimated iteratively in two parts with the EM algorithm: the Cox model and the logistic regression. In this context, the use of existing techniques for the Cox model and for the logistic regression is straightforward [Fan and Li, 2001]. However, such a split in parametric models, where the EM algorithm is typically not used, is not feasible, so that existing methods can not be applied directly. Developing a method to use the adaptive LASSO in a parametric mixture cure model can therefore really be convenient.

In this Chapter, we assume an AFT regression in the latency part of the mixture cure model. To handle interval-censored event times, we will suppose a parametric distribution for the error term. This allows to use classical methods of maximum likelihood. To keep the model as flexible as possible, we assume that the error term follows an EGG distribution. This distribution is very general and covers a wide class of distributions such as the Weibull and the log-Normal. The adaptive LASSO procedure is extended to our model to perform a continuous variable selection in the latency and the incidence.

This chapter is divided as follows: Section 3.2 presents our extension of the adaptive LASSO to the presence of a cure fraction. We investigate the fi-

nite sample properties of the method via a simulation study in Section 3.3. Lastly, we present results of the application of the method to the aforementioned Alzheimer's disease database in Section 3.4, and we end with a conclusion.

3.2 Methodology

3.2.1 The adaptive LASSO: least square approximation

In this chapter, we propose the adaptive LASSO, briefly introduced in Section 3.2, as a continuous variable selection method. Consider first the case of no cure fraction, that is, a classical AFT model with parameter $\boldsymbol{\eta}' = (\boldsymbol{\theta}', \boldsymbol{\beta}')$, where $\boldsymbol{\theta}' = c(\theta_0, \sigma)$. Hereafter, we assume that the covariates are standardized. The penalized likelihood is

$$-l(\boldsymbol{\eta}; \mathcal{O}) + n\lambda \sum_{j=1}^{m} p_j(|\beta_j|),$$

where $l_n(\boldsymbol{\eta}; \mathcal{O})$ is the log-likelihood function given in Equation (1.1) and where

$$p_j(|\beta_j|) = |\beta_j| w_j,$$

with $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ being a known weight vector. We follow the proposal of [Zhang and Lu, 2007] to take $w_j = 1/|\hat{\beta}_j|$, where $\hat{\beta}_j$ is the unpenalized MLE, reflecting somehow the importance of corresponding covariates. Any consistent estimator can be used. In case of multicollinearity, a robust estimator, not necessarily consistent, can be used instead [Zou, 2006]. Alternatively, one could use the adaptive elastic-net proposed by [Zou and Zhang, 2009].

The LARS algorithm was originally aimed at solving penalized least square problems. Nevertheless, any likelihood function can be expressed in an asymptotic least square equivalent, so that use of LARS algorithm is possible. Following [Wang and Leng, 2007], using Taylor expansion, $l_n(\boldsymbol{\eta})$ can be approximated by

$$l(\hat{\boldsymbol{\eta}}) + \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})' \ddot{l}(\hat{\boldsymbol{\eta}})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}),$$

where $\hat{\boldsymbol{\eta}}$ is an unpenalized consistent estimator, and $\ddot{l}(\hat{\boldsymbol{\eta}})$ represents the matrix of second derivatives of the log-likelihood at $\hat{\boldsymbol{\eta}}$.

The following equation is the *least square approximation* (LSA) of the log-likelihood $l(\boldsymbol{\eta}; \mathcal{O})$:

$$Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) = (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})' \ddot{l}(\hat{\boldsymbol{\eta}})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}). \tag{3.1}$$

The minimizer of -$Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})$ is different from the estimates obtained by minimizing the minus log-likelihood, henceforth, the maximizer of (3.1) is called the LSA estimator [Wang and Leng, 2007].

3.2.2 The adaptive LASSO in the presence of cured individuals

In the presence of cured individuals $\boldsymbol{\eta}' = (\boldsymbol{\theta}, \boldsymbol{\beta}', \boldsymbol{\gamma}')$ and the variables impacting the probability of being cured may not necessarily be the same as those impacting the survival distribution of the uncured people. Therefore, we propose to penalize both the incidence and the latency part, allowing a different penalty term in each part. This leads to the following minimization criterion:

$$-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\beta \sum_{j=1}^{m} \frac{|\beta_j|}{|\hat{\beta}_j|} + n\lambda_\gamma \sum_{j=1}^{s} \frac{|\gamma_j|}{|\hat{\gamma}_j|},$$

where λ_β is the tuning parameter for the β's, and λ_γ is the tuning parameter for the γ's. Again, we assume that all covariates are standardized.

To solve this optimization problem with the LSA estimator and the LARS algorithm, one can proceed iteratively in several steps. We optimize first with respect to the β's, holding every other parameter fixed, then do the same for the γ's. This way, we can easily obtain adaptive LASSO solutions, with two different penalty terms. This gives the following algorithm:

Step 1. Obtain the unpenalized MLE $\hat{\boldsymbol{\eta}}' = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ by maximizing $l(\boldsymbol{\eta}; \mathcal{O})$ given in Equation (1.3).

Step 2. Set $\boldsymbol{\eta}' = (\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}', \hat{\boldsymbol{\gamma}}')$, i.e., every parameter other than $\boldsymbol{\beta}$ is fixed. Minimize $-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\beta \sum_{j=1}^{m} \frac{|\beta_j|}{|\hat{\beta}_j|}$ to get adaptive LASSO estimate $\tilde{\boldsymbol{\beta}}$.

Step 3. Set $\boldsymbol{\eta}' = (\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}', \boldsymbol{\gamma}')$, i.e., every parameter other than $\boldsymbol{\gamma}$ is fixed. Minimize $-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\gamma \sum_{j=1}^{m} \frac{|\gamma_j|}{|\hat{\gamma}_j|}$ to get adaptive LASSO estimate $\tilde{\boldsymbol{\gamma}}$.

Step 4. Set $\boldsymbol{\eta}' = (\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\gamma}}')$, i.e., every parameter other than $\boldsymbol{\theta}$ is fixed. Maximize the unpenalized likelihood $l(\boldsymbol{\eta})$ with respect to $\boldsymbol{\theta}$. We then have $\tilde{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$.

Step 5. Repeat step 2 to 4 until convergence.

The extra tuning parameter λ does not lead to any identifiability issues of the parameters of interest $\boldsymbol{\eta}$. However, during our simulation studies, some numerical instabilities (divergence of the algorithm or incoherent estimates) were observed. In such case, the algorithm was rerun, starting from different initial values, a few times (see Appendix B.1).

3.2.3 Tuning parameter selection and variance estimation

The choice of the optimal penalty $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\beta, \hat{\lambda}_\gamma)$ is of crucial importance and is made via a BIC selection criterion [Wang and Leng, 2007]. First, for fixed λ_β and λ_γ, let $\tilde{\boldsymbol{\beta}}_{\lambda_\beta}$ and $\tilde{\boldsymbol{\gamma}}_{\lambda_\gamma}$ be the adaptive LASSO estimates with λ_β and λ_γ, respectively. We minimize

$$BIC(\boldsymbol{\lambda}) = -Q(\tilde{\boldsymbol{\eta}}_{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}}) + \log(n) df_\lambda,$$

where $\tilde{\boldsymbol{\eta}}'_{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}'_{\lambda_\beta}, \tilde{\boldsymbol{\gamma}}'_{\lambda_\gamma})$ and df_λ is the number of non-zero coefficients in $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\lambda}}$. We then take

$$\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\beta, \hat{\lambda}_\gamma) = \arg \min_{(\lambda_\beta, \lambda_\gamma)} BIC((\lambda_\beta, \lambda_\gamma)).$$

The minimization can be made via a grid search amongst selected values of λ_β and λ_γ and we take the combination leading to the smallest BIC. This procedure allows $\hat{\lambda}_\gamma$ to be different from $\hat{\lambda}_\beta$, therefore a different amount of shrinkage in the latency part and in the incidence part can be reached.

Standard errors for adaptive LASSO estimates are calculated based on a ridge regression approximation and on the sandwich formula for computing the covariance matrix of the estimates [Tibshirani, 1996, Fan and Li, 2001, Zou, 2006].

Denote H the matrix of second derivatives of the log-likelihood at $\tilde{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$. Define

$$A = diag\left(1, 1, \frac{\lambda_\beta}{\tilde{\beta}_1^2}, \cdots, \frac{\lambda_\beta}{\tilde{\beta}_m^2}, 1, 1, \frac{\lambda_\gamma}{\tilde{\gamma}_1^2}, \cdots, \frac{\lambda_\gamma}{\tilde{\gamma}_s^2}\right).$$

Also, define

$$\begin{aligned} D = diag\Bigg(1, 1, \frac{\mathbb{1}(\tilde{\beta}_1 \neq 0)\lambda_\beta}{\tilde{\beta}_1^2}, \cdots, \frac{\mathbb{1}(\tilde{\beta}_m \neq 0)\lambda_\beta}{\tilde{\beta}_m^2}, 1, \\ 1, \frac{\mathbb{1}(\tilde{\gamma}_1 \neq 0)\lambda_\gamma}{\tilde{\gamma}_1^2}, \cdots, \frac{\mathbb{1}(\tilde{\gamma}_s \neq 0)\lambda_\gamma}{\tilde{\gamma}_s^2}\Bigg). \end{aligned}$$

Then the sandwich formula gives the following estimated covariance matrix:

$$cov(\hat{\boldsymbol{\eta}}) = (H + A)^{-1} (H + D) H^{-1} (H + D) (H + A)^{-1}.$$

The estimated variance of a coefficient set to zero is equal to zero. More details about this equation can be found in [Lu and Zhang, 2006]. The variance-covariance matrix provides a straightforward means to obtain standard errors of the LASSO estimates. However, it is based on several approximations and may

fail to provide satisfactory estimation. A bootstrap approach may therefore provide more accurate results.

3.3 Simulation study

In this simulation study, we aim to study the behavior of the adaptive Lasso. First of all, we compare the estimates obtained when using the model knowing the true covariates with the estimates obtained when using the adaptive Lasso selection procedure. Second, we check if the aLasso correctly detect non-zero covariates in both parts of the mixture cure model: the AFT and the logistic regression. We also study the impact of having slightly correlated covariates.

3.3.1 Simulation setting

Data are generated following the algorithm of Section 2.2 for interval-censoring, and values of parameters as well as distribution of variables are given in Table 3.1. To assess the performance of the variable selection method, we furthermore added 10 standard normal variables $X_4, \ldots, X_{13}, Z_4, \cdots, Z_{13}$, in the latency and the incidence parts, whose coefficients are truly zero. Three scenarios are covered to reach different proportion of cured and right-censored (RC) individuals: 20% cure and 40% RC in scenario 1, 30% cure and 40% RC in scenario 2, and 40% cure and 60% RC in scenario 3. All observations that are not right-censored are interval-censored. The scale parameter σ of the AFT model depend on covariates such that $\sigma = \exp(\alpha_0 + \alpha_1 X_1) = \exp(-2 + 0.5X_1)$.

3.3.2 Simulation results

The objective of this simulation study is to evaluate the adaptive LASSO procedure described above, both in term of estimation and variable selection. We use the mixture cure model with the EGG-AFT model in latency and the logistic regression in incidence; and our adaptation of the LSA R code to perform variable selection and estimation [Wang and Leng, 2007], available in Appendix B.2.

Firstly, for comparison purposes, we fitted the data without the additional 10 covariates, so without any variable selection procedure. Tables 3.2, 3.3 and 3.4 show the results for $n = 200$, $n = 300$ and $n = 500$, respectively. For any sample size, the bias and MSE for the latency part are low. However, for the smallest sample size ($n = 200$), the bias and MSE in the incidence part is large for scenario 1, where the cure proportion is low compared to the right-censoring rate. As explained in Chapter 2, we need enough information, that is, enough

cured individuals, in order to discriminate between cured and uncured, and to be able to perform accurate estimation in the incidence part. The bias and MSE are decreasing with the sample size.

Secondly, we fitted the data with the additional 10 covariates, so using the adaptive LASSO. Tables 3.5, 3.6 and 3.7 show the results for $n = 200$, $n = 300$, and $n = 500$. The upper part shows bias and MSE for the truly non-zero coefficients, and the lower part gives the average number of correct (resp., incorrect) zero's, i.e. the average number of times the adaptive LASSO sets a coefficient to zero when it truly is zero (resp., non-zero). In the simulations, the optimal tuning parameter $\boldsymbol{\lambda}$ was chosen via the BIC-type selection criterion from Section 3.2.3. Globally, those results reflect the same trend as the analysis without variable selection, i.e. low bias and MSE, except for small sample size ($n = 200$) in scenario 1. Compared to the analysis without variable selection, for non-zero coefficients, we detect larger bias and MSE in incidence. This behavior is expected as the coefficients are shrunk to zero by the adaptive LASSO.

For $n = 500$, we see that our method performs well for both coefficient estimation and variable selection. The average number of correct zero's is very close to the optimal value of 10, in both latency and incidence parts. The average number of incorrect zero's is very close to the optimal value of 0 in the latency part, and higher in the incidence part. This is explained by the fact that, in the logistic regression some covariates (here, Z_2 and Z_3) do not have an impact on the cure probability. So, the adaptive LASSO procedure interestingly sets these coefficients to zero. As a consequence, the bias for $\hat{\gamma}_2$ and $\hat{\gamma}_3$ is slightly larger. The effect of the cured proportion and right-censoring rate, concerning variable selection, follows the same trend as analyzed before: the number of correct zero slightly decreases when there is more right-censoring.

Thirdly, we assess the performances of our method on the basis of the second scenario, where there is 30% cured individuals and 40% right-censored individuals, by adding 25 covariates whose coefficients are truly zero, in each part of the model. Table 3.8 gives results of 2000 replications for $n = 500$. These results can be compared with the second column of Table 3.7. Bias and MSE are slightly larger when there are more zero covariates, but conclusions about selected variables stay the same.

Lastly, we investigate the issue of correlated variables. Indeed, in presence of strong correlation, it is well known that the LASSO, or adaptive LASSO, may choose only one of the two correlated variables. To see the effect of small correlation on adaptive LASSO, we slightly modified the second scenario by adding a correlation structure between covariates. More specifically, X_2,X_3, X_4 and X_5 are generated though a multivariate normal distribution with covariance

matrix V:

$$V = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{3.2}$$

where $\rho \in [0, 1]$ is chosen to reach a different level of correlation between the variables. The same generation process is used for Z_2,Z_3, Z_4 and Z_5. Results of the 2000 replications are given in Table 3.9 for different values of ρ, and can be compared with the second column of Table 3.7. We see that the correlation has practically no effect on the latency part. In the incidence part, we see that the number of incorrect zero's is lower for negative values of ρ, slightly improving the variable selection. In this setting, we do not observe any issues related to correlation.

Overall, the adaptive LASSO performs satisfactorily for estimation as well as for variable selection, as it includes variables that truly have an impact on the model.

3.4 Real data analysis

We apply our approach to the Alzheimer's disease dataset. In our analysis, 12 potential prognostic factors were included in the model, both in the latency and in the incidence part : MMSE, expression, remote memory, learning, attention, praxis, abstract thinking, perception, ApoE4 status, gender, age, and years of total education, resulting in a total of 26 parameters. We used the EGG-AFT mixture model to get unpenalized maximum likelihood estimates, and the adaptive LASSO procedure described in Section 3.2 to perform variable selection.

Table 3.10 shows the adaptive LASSO estimate, the standard error estimated using formula from Section 3.2.3, and the exponentiated estimates. This allows a direct interpretation of the impact of covariates, in terms of acceleration or deceleration of the time to the event in latency; and in terms of increase or decrease in odds for the incidence.

Focusing on uncured people (the latency part), there are 3 variables increasing the expected duration, thus having a positive impact on the survival, by at least 15%: Expression (38%), Perception (20%) and Education (16%). On the other hand, only the age shortens the duration by at least 15%: When age increases by 5 years, the expected time until conversion is shorten by 27%. For comparison, without considering cure, the perception score was not significant, whereas the learning score was significant, with a positive impact on the survival. However, we see that learning still has a positive impact, but

in the incidence part, reducing the risk to be uncured. Three other variables have a positive impact on the probability to be uncured: MMSE (-89%), praxis (-73%), ApoE4 Status (-43%). At the opposite, the abstract thinking (62%) and the total years of education (883%) have here a highly negative impact and significantly increases the odds ratio.

With these results, we estimated the average cure proportion in the whole sample to 20%. Analyzing these data taking a cure fraction into account lead to more information: first, the positive impact of the learning variable is now due to the fact that it reduces the probability to convert to MCI. Second, we now consider other variables that have an impact: those impacting the probability to experience the disease.

3.5 Discussion

In this chapter, we extended a continuous variable selection method to a flexible parametric AFT mixture cure model.

Although widely used in the context of dimension reduction, when the number of covariates exceeds the number of observations, shrinkage methods such as the adaptive LASSO are also useful in our context. Indeed, the number of covariates may be large, as each covariate can be included twice, i.e. once in both parts of the model. This is why we believe that such shrinkage methods should be extended to the mixture cure model.

Some shrinkage methods have been studied in a semi-parametric PH mixture cure model, but without taking interval-censoring into account [Liu et al., 2012]. In this paper, the EM algorithm is used to maximize the complete-data likelihood (see Section 1.3.3). This allows a split of the likelihood into a Cox model and a logistic regression. This means that existing methods can be applied for the Cox model, and for the logistic regression. In a parametric AFT mixture cure model, where the EM algorithm is typically not used, another methodology needs to be developed. By extending the adaptive LASSO to our model, that chapter fills a significant gap.

In the presence of interval-censoring and cure, it is very difficult to develop simple yet efficient estimation procedures without imposing parametric restrictions. We thus used the AFT mixture cure model with the EGG distribution in the latency. The EGG distribution is capable of capturing a lot of characteristics, and is an excellent compromise in this context. Based on our simulation study, it performs very adequately for estimation and variable selection.

Our simulation study highlighted different aspects. Firstly, as the adaptive LASSO is a shrinkage method, some bias may appear in the parameters estimates. This bias was however quite small and the estimation remained more

than satisfying. Secondly, the variable selection performed correctly in both parts of the model, even with a large number of supplementary covariates (in our case, up to 25 covariates with coefficients 0 were added in the model). It is worth noting that in the incidence part, covariates that did not have an impact on the response were not entered into to model. The adaptive LASSO therefore selected covariates that truly affected the response. On a final note, we also investigated the issue of correlated covariates, and could not report any problem with it, as long as the correlation remains below 75%.

In conclusion, the extension of the adaptive LASSO to our parametric AFT mixture cure model provides an excellent alternative to methods like the best subset or stepwise selection, and allows the analyst to obtain a model that is easily interpretable.

Table 3.1 – Values of parameters and generation of variables for each scenario.

	Scenario 1 20% Cure 40% RC	Scenario 2 30% Cure, 40% RC	Scenario 3 40% Cure, 60% RC
β_0	4.1	4.1	4.1
β_1	-0.2	-0.2	-0.2
β_2	0.5	0.5	0.5
β_3	-0.5	-0.5	-0.5
X_1	$Bern(0,5)$	$Bern(0,5)$	$Bern(0,5)$
X_2	$\mathcal{N}(0,0.16)$	$\mathcal{N}(0,0.16)$	$\mathcal{N}(0,0.16)$
X_3	$\mathcal{N}(0,0.25)$	$\mathcal{N}(0,0.25)$	$\mathcal{N}(0,0.25)$
ε	$EGG(0)$	$EGG(0.5)$	$EGG(1)$
γ_0	2	1	0.85
γ_1	-1	-0.2	-0.85
γ_2	0.2	0.2	0.2
γ_3	-0.4	-0.4	-0.4
Z_1	X_1	X_1	X_1
Z_2	$\mathcal{N}(0,0.4)$	$\mathcal{N}(0,0.4)$	$\mathcal{N}(0,0.4)$
Z_3	$\mathcal{N}(0,0.5)$	$\mathcal{N}(0,0.5)$	$\mathcal{N}(0,0.5)$
K	14	14	12
a	4	4	4

Table 3.2 – Results of 2000 simulations: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=200						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0.023	0.107	0.010	0.098	0.214	0.811
β_0	-0.001	0.001	-0.000	0.001	0.003	0.002
β_1	-0.000	0.001	-0.000	0.002	0.004	0.003
β_2	-0.001	0.002	0.003	0.002	0.004	0.005
β_3	0.000	0.001	-0.001	0.002	-0.001	0.003
α_0	-0.049	0.016	-0.053	0.022	-0.135	0.123
α_1	0.014	0.025	0.007	0.026	0.013	0.045
γ_0	0.320	5.286	0.051	0.105	0.063	0.158
γ_1	-0.209	1.668	-0.020	0.170	-0.047	0.226
γ_2	0.066	7.085	0.002	0.246	0.020	0.284
γ_3	-0.073	1.357	-0.016	0.171	-0.014	0.169

Table 3.3 – Results of 2000 simulations: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=300						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0.048	0.096	-0.013	0.077	0.050	0.210
β_0	-0.002	0.000	-0.001	0.001	0.002	0.001
β_1	-0.000	0.001	-0.000	0.001	0.001	0.003
β_2	0.003	0.006	0.007	0.009	0.004	0.016
β_3	-0.003	0.003	-0.003	0.004	-0.005	0.008
α_0	-0.025	0.012	-0.029	0.016	-0.058	0.054
α_1	0.011	0.019	0.006	0.019	0.008	0.034
γ_0	0.398	2.435	0.052	0.104	0.124	0.518
γ_1	-0.298	2.111	-0.031	0.148	-0.100	0.492
γ_2	0.024	0.028	0.006	0.008	0.014	0.012
γ_3	-0.054	1.007	-0.034	0.433	-0.059	0.650

Table 3.4 – Results of 2000 simulations: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=500						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0.025	0.050	-0.006	0.036	0.045	0.089
β_0	-0.001	0.000	-0.001	0.000	0.002	0.001
β_1	-0.001	0.001	0.001	0.001	0.001	0.001
β_2	0.001	0.004	0.004	0.005	0.005	0.008
β_3	0.000	0.002	-0.002	0.002	0.000	0.004
α_0	-0.016	0.007	-0.016	0.009	-0.040	0.026
α_1	0.003	0.011	0.007	0.011	0.002	0.017
γ_0	0.122	0.397	0.022	0.050	0.030	0.085
γ_1	-0.081	0.385	-0.002	0.073	-0.020	0.105
γ_2	0.009	0.007	0.004	0.004	0.006	0.004
γ_3	-0.025	0.399	-0.007	0.237	0.016	0.268

Table 3.5 – Results of 2000 simulations, with adaptive LASSO.

Sample Size : n=200						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	0.007	0.207	0.109	0.271	0.007	0.504
β_0	0.000	0.001	0.002	0.001	-0.009	0.003
β_1	0.016	0.002	0.020	0.003	0.031	0.008
β_2	-0.011	0.002	-0.012	0.003	-0.010	0.006
β_3	0.009	0.001	0.009	0.002	0.003	0.005
α_0	-0.139	0.041	-0.184	0.079	-0.217	0.156
α_1	0.071	0.047	0.086	0.056	0.129	0.127
γ_0	0.874	5.767	0.176	0.216	0.312	0.498
γ_1	-0.562	5.272	0.002	0.153	0.009	0.525
γ_2	-0.138	0.415	-0.182	0.059	-0.185	0.125
γ_3	0.274	0.472	0.332	0.173	0.325	0.189
	Average correct/incorrect number of zero's					
	Latency					
Correct	9.202		9.283		8.685	
Incorrect	0.006		0.010		0.086	
	Incidence					
Correct	9.176		9.805		9.707	
Incorrect	1.966		2.633		2.138	

Table 3.6 – Results of 2000 simulations, with adaptive LASSO.

Sample Size : n=300						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	-0.031	0.099	0.080	0.109	0.178	0.276
β_0	-0.002	0.001	0.003	0.001	0.001	0.001
β_1	0.008	0.001	0.013	0.002	0.023	0.004
β_2	-0.007	0.001	-0.007	0.002	-0.010	0.003
β_3	0.006	0.001	0.004	0.001	0.004	0.002
α_0	-0.077	0.016	-0.112	0.031	-0.190	0.096
α_1	0.034	0.020	0.046	0.025	0.076	0.049
γ_0	0.310	0.379	0.104	0.100	0.140	0.169
γ_1	-0.200	0.486	0.047	0.087	0.051	0.255
γ_2	-0.175	0.073	-0.183	0.049	-0.180	0.050
γ_3	0.304	0.180	0.324	0.157	0.334	0.158
	Average correct/incorrect number of zero's					
	Latency					
Correct	9.646		9.640		9.306	
Incorrect	0.000		0.001		0.019	
	Incidence					
Correct	9.624		9.858		9.859	
Incorrect	1.921		2.601		2.034	

Table 3.7 – Results of 2000 simulations, with adaptive LASSO.

Sample Size : n=500						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	-0.017	0.045	0.029	0.042	0.178	0.152
β_0	-0.001	0.000	0.001	0.000	0.006	0.001
β_1	0.004	0.001	0.007	0.001	0.016	0.002
β_2	-0.005	0.001	-0.004	0.001	-0.004	0.002
β_3	0.004	0.000	0.003	0.001	0.002	0.001
α_0	-0.043	0.007	-0.056	0.011	-0.133	0.048
α_1	0.019	0.011	0.025	0.012	0.038	0.020
γ_0	0.152	0.148	0.051	0.045	0.059	0.067
γ_1	-0.091	0.201	0.086	0.053	0.055	0.117
γ_2	-0.182	0.053	-0.178	0.047	-0.189	0.043
γ_3	0.306	0.152	0.309	0.140	0.315	0.142
	Average correct/incorrect number of zero's					
	Latency					
Correct	9.859		9.834		9.705	
Incorrect	0.000		0.000		0.002	
	Incidence					
Correct	9.845		9.896		9.909	
Incorrect	1.862		2.540		1.872	

Table 3.8 – Results of 2000 simulations, with adaptive LASSO, for $n = 500$ and **25 covariates with zero coefficients**.

Param.	(30% Cure, 40% RC)	
	Bias	MSE
q	0.114	0.082
β_0	0.004	0.000
β_1	0.012	0.001
β_2	-0.008	0.001
β_3	0.006	0.001
α_0	-0.132	0.030
α_1	0.066	0.018
γ_0	0.125	0.073
γ_1	0.045	0.070
γ_2	-0.187	0.043
γ_3	0.337	0.146
	Average correct/incorrect number of zero's	
	Latency	
Correct	24.540	
Incorrect	0.000	
	Incidence	
Correct	24.813	
Incorrect	2.539	

Table 3.9 – Result of 2000 simulations, with adaptive LASSO for $n = 500$. Each column represents a different value of ρ in the correlation matrix V. The case where $\rho = 0$ is given in the second column of Table 3.7.

Param.	$\rho = 0.5$		$\rho = 0.75$		$\rho = -0.75$		$\rho = -0.3$	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
q	0.036	0.036	0.033	0.033	0.052	0.057	0.042	0.050
β_0	0.001	0.000	0.001	0.000	0.002	0.001	0.001	0.000
β_1	0.008	0.001	0.007	0.001	0.010	0.001	0.009	0.001
β_2	-0.002	0.000	-0.003	0.000	-0.000	0.000	-0.001	0.000
β_3	0.002	0.000	0.003	0.000	0.001	0.000	0.001	0.000
α_0	-0.060	0.011	-0.054	0.010	-0.081	0.017	-0.071	0.014
α_1	0.028	0.011	0.027	0.011	0.038	0.015	0.034	0.014
γ_0	0.050	0.041	0.044	0.034	0.062	0.047	0.056	0.043
γ_1	0.086	0.051	0.103	0.048	0.082	0.055	0.087	0.053
γ_2	-0.177	0.038	-0.185	0.041	-0.080	0.043	-0.130	0.034
γ_3	0.237	0.090	0.284	0.110	0.095	0.073	0.116	0.052
	Average correct/incorrect number of zero's							
	Latency							
Correct	9.843		9.852		9.817		9.839	
Incorrect	0.000		0.000		0.000		0.000	
	Incidence							
Correct	9.866		9.838		9.729		9.818	
Incorrect	2.155		2.345		1.745		1.736	

Table 3.10 – MCI results: adaptive LASSO estimates, standard errors and exponential of the estimates.

Parameter	Estimates	SD	Exp (Estimate)
Latency part: EGG-AFT			
Intercept (Lat.)	2.628	0.155	
MMSE	-	-	
Expression	0.321	0.077	1.38
Remote	-	-	
Learning	-	-	
Attention	-0.057	0.043	0.94
Praxis	-	-	
Abstract Thinking	0.086	0.034	1.09
Perception	0.182	0.052	1.20
APOE E4	-0.092	0.025	0.91
Gender	-0.061	0.022	0.94
Age (5y.)	-0.321	0.144	0.73
Total Education	0.152	0.163	1.16
Incidence part: Logistic regression			
Intercept (Inc.)	2.657	0.985	
MMSE	-2.250	0.802	0.11
Expression	-	-	
Remote	-	-	
Learning	-0.969	0.425	0.38
Attention	-	-	
Praxis	-1.302	0.564	0.27
Abstract Thinking	0.483	0.456	1.62
Perception	-	-	
APOE E4	-0.556	0.264	0.57
Gender	-	-	
Age (5y.)	-	-	
Total Education	2.285	2.081	9.83

CHAPTER 4

Diagnostics checks through residuals

"We know what we are, but know not what we may be."

- William Shakespeare (1564-1616)

This chapter is mainly based on the paper "Diagnostic checks in mixture cure model with interval-censored data" by Scolas S., Legrand C., Oulhaj A. and El Ghouch A., accepted for publication in Statistical Methods in Medical Research.

Models for interval-censored survival data presenting a fraction of "cure" or "immune" patients have recently been proposed in the literature, particularly extending the mixture cure model to interval-censored data. However, little is known about the goodness-of-fit of such models. In a mixture cure model, the survival distribution of the entire population is improper and expressed in terms of the survival distribution of uncured individuals, i.e. the latency part of the model; and the probability to experience the event of interest, i.e. the incidence part. To validate a mixture cure model, assumptions made on both parts need to be checked, i.e. the survival distribution of uncured individuals, the link function used in the latency and the linearity of the covariates used in the both parts of the model. In this work, we investigate the Cox-Snell and deviance residuals and show how they can be adapted and used to perform diagnostics checks when all subjects are right- or interval-censored and some subject are cured with unknown cure status. A large simulation study investigates the ability of these residuals

to detect a departure from the assumptions of the mixture model. Developed techniques are applied to a real data set about Alzheimer's disease.

4.1 Introduction

In this chapter, we focus our attention on the use of residuals, in case where event times are either right- or interval-censored, to assess the fit of mixture cure models considering either a parametric PH or AFT model in the latency part. Mixture cure models rely on assumptions on both the incidence and latency part. These assumptions must hold to ensure the validity of the model and all of its resulting conclusions, including inferences and predictions.

The inspection of the residuals is one of the usual methods to assess the assumptions of a given model, with a long tradition in linear and generalized linear models. For instance, to detect a departure from normality in a classical linear regression model, a histogram and/or a quantile-quantile (QQ) plot of residuals are two of the main graphical tools used in practice. Although a visual inspection of a QQ-plot is not a formal test of normality, it provides practitioners with relevant and easily interpretable information. When the normality assumption is not met, it might then be better to transform the dependent variable, and graphical tools like a QQ-plot may help to find an adequate transformation. Concerning the usual assumption of linearity of covariates effects, scatter plot of the residuals versus fitted values and versus each of the predictors is one of the most commonly used diagnostic plot in practice. The presence of systematic features in these plots generally implies a missing covariate or a nonlinear effect. Although the exact nature of the problem may be hard to find, residuals plots may be very helpful to reveal a lack-of-fit with respect to a covariate and to select an adequate transformation. An alternative method would be to test whether a transformed covariate enters significantly into the model, but this implies testing many transformations and can be cumbersome, while a glance at a residuals plot might be quicker and more meaningful.

While the definition of residuals for a linear or a generalized linear model is unambiguous [McCullagh and Nelder, 1989], it becomes more complex in the context of time-to-event analysis mainly due to the presence of censoring. In the context of right-censored data, several types of residuals have been defined with different goals, see Section 1.4. The most often used are probably the Cox-Snell residuals, the martingale residuals and the deviance residuals. Cox-Snell and deviance residuals are implemented in most statistical software's and are widely used in classical survival analysis.

In mixture cure models, to the best of our knowledge, there is not such an easy graphical tool to assess the quality of the fit and make the model closer to the data at hand. The extension of the existing methods would allow a thorough data analysis when handling cure and censoring, but remains a chal-

lenging issue. In fact, the nature of mixture cure models makes diagnosing more complex since a residuals analysis must include the assessment of:

(i) the assumed distribution for the uncured sub-population (uncured survival distribution),

(ii) the assumed model for the probability to experience the event of interest, and

(iii) the resulting global survival distribution of the entire population, which is improper.

Thus, in our case, residuals not only have to allow the detection of a global lack-of-fit, but also have to pinpoint specific levels of the model. Residuals based on the global survival distribution would not allow to detect in which part of the model a transformation might be needed. On the other hand, residuals based on the uncured survival distribution can only detect a needed transformation in the latency part of the model.

The fact that the global survival distribution is improper makes the use of the classical Cox-Snell residuals questionable and it is not obvious that they will keep their good properties. An additional complication is that the cure status (cured or uncured) of a right-censored individual is unknown. As a consequence, in order to assess the assumed uncured survival distribution, one needs first to identify uncured individuals in the population sample. To this end, it is necessary to predict their status using both the data and the model under study. As will later be explained, investigating and solving these issues leads to two types of Cox-Snell residuals, intended to assess the global and the uncured survival distribution. The same difficulties are encountered with deviance residuals when it comes to check the linearity assumption. In fact, keeping in mind that the mixture cure model consists of two regression parts (the latency and the incidence), covariates entering the latency may not be the same as covariates entering the incidence. Moreover, a covariate may have a linear impact in the latency part, while having, for example, a quadratic impact, or no impact at all, in the incidence part. The fact that the deviance residuals are based on the likelihood function facilitates their adaptation and makes them more relevant for mixture cure models. Moreover, using the likelihood gives a straightforward method to take into account the fact that the event times are interval-censored. Residuals that are based on counting processes like the martingale residuals do no easily allow to deal with interval-censoring. The complete-data likelihood (defined in Section 1.3.3) induces a natural split between latency and incidence, and takes into account interval-censored event times. Based on this likelihood, we will introduce two deviance residuals: the latency deviance residuals, and

the incidence deviance residuals. Plotting these two types of deviance residuals will allow the detection of a needed transformation in each part of the model: the latency and the incidence.

[Peng and Taylor, 2016] also proposed residuals in the presence of cure, considering a Cox mixture cure model with right-censoring. The context is therefore less general than ours, since we consider the development of a diagnostic check procedure for both the Cox and AFT model with cured and interval-censored observations .

In this chapter, we aim to extend the use of residuals to perform diagnostic checks in a parametric mixture cure model with right- and interval-censoring. Our first objective is to discuss how to define the Cox-Snell residuals intended to check the survival distribution of the uncured sub-population and of the entire population. We first study the properties of the Cox-Snell residuals for the entire population, then define an approach to estimate the status, cured or uncured, of a right-censored observation. Our second objective is to investigate how to detect non-linearity in the latency and the incidence part of the model using our proposed deviance residuals.

This chapter is organized as follows: Section 4.2.1 develops the Cox-Snell residuals in mixture cure models with and without interval-censoring. Residuals aiming at detecting non-linearity are covered in Section 4.2.2. Section 4.3 shows the behavior of the proposed residuals in a simulation study. Section 4.4 presents the results of the application of our method to our real data set on Alzheimer's disease. Finally, our findings are shortly discussed in Section 4.5.

4.2 Methodology

4.2.1 Checking the survival function: Cox-Snell residuals

In this section, Cox-Snell residuals aiming at checking the marginal survival function (S) and the uncured survival function (S_u) in a mixture cure model are defined and studied.

For simplicity, let's temporarily assume that no interval-censoring is present, i.e. the event times t_i are exactly observed when $\delta_i = 1$. To assess the validity of the hypothesized survival distribution S_u of the uncured sub-population, the Cox-Snell residuals $r_{CS,u}(t_i) = -\log(\hat{S}_u(t_i))$ can be computed for all i with $Y_i = 1$, i.e. only for the uncured observations.Since the uncured status Y_i is unknown for right-censored observations, we propose to replace Y_i with its expected value given the observed data :

$$E\left(Y_i|\mathcal{O}_i\right) = \delta_i + (1-\delta_i)\frac{p_i S_u(l_i)}{p_i S_u(l_i) + 1 - p_i}.$$

Since p_i and S_u are unknown, this expectation is estimated by ξ_i:

$$\xi_i = \delta_i + (1 - \delta_i)\frac{\hat{p}_i \hat{S}_u(l_i)}{\hat{p}_i \hat{S}_u(l_i) + 1 - \hat{p}_i}. \tag{4.1}$$

As ξ_i can take any value between 0 and 1, we need a threshold to predict the uncured status. In this work, we take 0.5 as a threshold and classify an individual with $\xi_i > 0.5$ to the uncured sub-population and classify an individual with $\xi_i \leq 0.5$ to the cured sub-population. We thus define uncured Cox-Snell residuals as

$$r_{CS,u}(t_i) = -\log(\hat{S}_u(t_i)) \text{ for } i : \xi_i > 0.5.$$

A plot of the Cox-Snell residuals on the x-axis versus their Kaplan-Meier or Nelson-Aalen estimated cumulative hazard based on the right-censored sample $(\delta_i, r_{CS,u}(t_i))$ on the y-axis, should exhibit points aligned on a straight line with a unit slope and zero intercept. A departure from this line may suggest a model inadequacy but, in such a case, no indication of the cause of this inadequacy is provided by the plot. Despite this drawback, the Cox-Snell residuals plot remains useful, for example, to compare two possible distributions.

The same approach can be used to assess the hypothesized global survival distribution S of the whole population, cured and uncured. For this purpose, we define the Global Cox-Snell residuals as

$$r_{CS}(t_i) = -\log(\hat{S}(t_i)) = -\log\left(\hat{p}_i \hat{S}_u(t_i) + 1 - \hat{p}_i\right),\ i = 1, \cdots, n.$$

This definition is motivated by the fact if T has an improper cumulative distribution $F(T) = pF_u(T)$, where $F_u(T) = 1 - S_u(T)$ is a proper cumulative distribution, then for $0 \leq t < p$,

$$\begin{aligned} P(F(T) \leq t) &= pP(T \leq F_u^{-1}(t/p) | T < +\infty) + (1-p)P(F_u(T) \leq t/p | T = +\infty) \\ &= t \end{aligned}$$

and for $t \geq p$, $P(F(T) \leq t) = P(F_u(T) \leq t/p) = 1$. Consequently, if the global survival function S fitted to the data is satisfactory, the global Cox-Snell residuals r_{CS} should behave in $[0, -\log(1-p))$ like a censored sample from a mean one exponential distribution. In this case, as for the uncured observations, a plot of the global Cox-Snell residuals versus their estimated cumulative hazard should reveal points aligned on a straight line of unit slope and zero intercept.

The same argument can also be applied to the case of interval-censored data to check the validity of both S_u and S. In the following, only S_u is dis-

cussed, as the same approach can be applied to S. From the observed interval-censored data $]l_i, r_i]$, the interval-censored Cox-Snell residuals are obtained: $]-\log(\hat{S}_u(l_i), -\log(\hat{S}_u(r_i)]$, for $i = 1, \ldots, n$. Their cumulative hazard function can be estimated by the Turnbull's self-consistency algorithm [Turnbull, 1976]. If the model is adequate, a plot of the estimated function against the residuals interval endpoints should approximately resembles a straight line of unit slope and zero intercept. These residuals are not easy to handle given their interval nature. For this reason, [Farrington, 2000] suggest to replace the interval residuals with their expected values under the unit exponential distribution, leading to the following adjusted Cox-Snell residuals for S_u :

$$r_{CS,u}(l_i, r_i) = \frac{\hat{S}_u(l_i)(1 - \log(\hat{S}_u(l_i))) - \hat{S}_u(r_i)(1 - \log(\hat{S}_u(r_i)))}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}. \tag{4.2}$$

Once the hypothesis on S_u are validated, the validity of the global survival S can also be assessed via the Cox-Snell residuals $r_{CS}(l_i, r_i)$ obtained from (4.2) but with $\hat{S}$ instead of $\hat{S}_u$.

4.2.2 Detecting non-linearity: Deviance residuals

One way of assessing the adequacy of a model is to compare it with the corresponding saturated model. This is the model with the same distribution and the same structure as the model under study, but with the maximum number of parameters of interest that can be estimated ("perfect" estimable model). In our case, the saturated model allows μ_i and ν_i to be different for each observation, whereas the hypothesized regression model assumes that $\mu_i = \boldsymbol{x_i'\beta}$ and $\nu_i = \boldsymbol{z_i'\gamma}$.

In the mixture cure model, the observed data are $\mathcal{O}_i = (\delta_i, l_i, r_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The complete data set is obtained by augmenting the observed data set by the partially unobserved uncured status, y_i: $\mathcal{O}_{i,c} = (\mathcal{O}_i, y_i)$. Based on the complete data set, the complete-data log-likelihood function for the saturated model is defined as

$$l_n(\boldsymbol{\theta}, \mu_i, \nu_i; \mathcal{O}_c) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}, \mu_i, \nu_i) \tag{4.3}$$

where $l_i(\boldsymbol{\theta}, \mu_i, \nu_i)$, $i = 1, \ldots, n$, are the individual log-likelihood contributions given by

$$\begin{aligned} l_i(\boldsymbol{\theta}, \mu_i, \nu_i) = {} & y_i \log(p_i(\nu_i)) + (1 - y_i) \log(1 - p_i(\nu_i)) \\ & + \delta_i \log(S_u(l_i|\mu_i; \boldsymbol{\theta}) - S_u(r_i|\mu_i; \boldsymbol{\theta})) + (1 - \delta_i) y_i \log(S_u(l_i|\mu_i; \boldsymbol{\theta})). \end{aligned}$$

Following [Nelder and Wedderburn, 1972], we define the "complete" deviance statistic $D(\mathcal{O}_c)$ as twice the difference between the maximum achievable complete-data log-likelihoods under the saturated model and under the current regression model. Thus,

$$D(\mathcal{O}_c) = 2\sum_{i=1}^{n}(l_i(\hat{\boldsymbol{\theta}}, \tilde{\mu}_i, \tilde{\nu}_i; \mathcal{O}_c) - l_i(\hat{\boldsymbol{\theta}}, \hat{\mu}_i, \hat{\nu}_i; \mathcal{O}_c)),$$

where $\tilde{\mu}_i$ and $\tilde{\nu}_i$, $i = 1, \ldots, n$, are the points maximizing the saturated complete-data log-likelihood (4.3), and where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator under the current model of the unknown nuisance parameter $\boldsymbol{\theta}$. Some easy algebra shows that the maximum attainable value for the individual log-likelihood contribution is

$$\begin{aligned} l_i(\hat{\boldsymbol{\theta}}, \tilde{\mu}_i, \tilde{\nu}_i) =& y_i \log(p(\tilde{\nu}_i)) + (1 - y_i)\log(1 - p(\tilde{\nu}_i)) \\ &+ \delta_i \log\left(S_u(l_i|\tilde{\mu}_i, \hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i, \hat{\boldsymbol{\theta}})\right), \end{aligned}$$

where $p(\tilde{\nu}_i) = y_i$ and $\tilde{\mu}_i$ is the root of the equation

$$\frac{\partial}{\partial \mu_i} S_u(l_i|\mu_i; \hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial \mu_i} S_u(r_i|\mu_i; \hat{\boldsymbol{\theta}}),$$

for $i = 1, \ldots, n$. For the PH model, the solution is such that

$$\exp(\tilde{\mu}_i) = \frac{\log(-\log(S_{u,0}(r_i))) - \log(-\log(S_{u,0}(l_i)))}{\log(S_{u,0}(l_i)) - \log(S_{u,0}(r_i))},$$

whereas for the AFT model, the solutions depends on the assumed distribution. For example, as the log-Normal and the Weibull distribution are special cases of the EGG distribution, calculation for the EGG distribution are given in Appendix C.1.

Consequently, $D(\mathcal{O}_c) = 2\sum_{i=1}^{n} d_i(y_i)$, where $d_i(y_i)$ is the contribution of the observation i to the deviance and is given by

$$\begin{aligned} d_i(y_i) =& l_i(\hat{\boldsymbol{\theta}}, \tilde{\mu}_i, \tilde{\nu}_i) - l_i(\hat{\boldsymbol{\theta}}, \hat{\mu}_i, \hat{\nu}_i) \\ =& y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{p}_i}\right) \\ &+ \delta_i \log\left(\frac{S_u(l_i|\tilde{\mu}_i, \hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i, \hat{\boldsymbol{\theta}})}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}\right) - (1 - \delta_i) y_i \log(\hat{S}_u(l_i)). \end{aligned}$$

As the uncured status y_i is not observed for right-censored observations, we propose to replace it by its expected value, ξ_i, defined in Equation (4.1). This

leads to the "observed" deviance $D = 2\sum_{i=1}^{n} d_i$, where d_i is a shortcut for $d_i(\xi_i)$. D measures the closeness of the hypothesized model to the saturated model, but without distinguishing between the latency and the incidence part. A nice feature of this deviance is that it can be split into two additive parts: an incidence part D_{inc} and a latency part D_{lat}. More precisely, D can be written as:

$$D = D_{inc} + D_{lat} = 2\sum_{i=1}^{n} d_{i,inc} + 2\sum_{i=1}^{n} d_{i,lat},$$

where

$$d_{i,inc} = \xi_i \log\left(\frac{\xi_i}{\hat{p}_i}\right) + (1-\xi_i)\log\left(\frac{1-\xi_i}{1-\hat{p}_i}\right), \text{ and}$$
$$d_{i,lat} = \delta_i \log\left(\frac{S_u(l_i|\tilde{\mu}_i,\hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i,\hat{\boldsymbol{\theta}})}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}\right) - (1-\delta_i)\xi_i \log(\hat{S}_u(l_i)).$$

Observe that if all the individuals are uncured ($\xi_i = \hat{p}_i = 1$), then D reduces to D_{lat}, the classical deviance for interval and right-censored data. Observe also that D_{inc} is the deviance for the classical binary logistic model, with ξ_i as response variable. A "large" value of D_{inc} (D_{lat}) indicates a poorly fitted incidence (latency) part. This may be caused by a missing covariate or a covariate needing a transformation in the incidence part ("large" D_{inc}), in the latency part ("large" D_{lat}) or in both parts.

As in classical regression models (with or without censoring), we propose to measure the individual contribution to the deviance via the deviance residuals. In general, they are defined as $s_i\sqrt{2d_i}$, where d_i is the individual contribution to the deviance. In the classical logistic regression model, s_i is the sign of the difference between the observed response and its estimated expected value. In survival analysis, s_i is the sign of $\delta_i - r_i$, with r_i the Cox-Snell residuals, which can be interpreted as the difference between the observed and expected number of events for the individual i over the interval $(0, t_i)$[Collett, 2003, Farrington, 2000]. Therefore, in our case, for an individual i, we define three types of deviance residuals. One aiming at checking the global model, one aiming at checking the latency, and the last one aiming at checking the incidence: :

- Global deviance residuals: $r_D(l_i, r_i) = sgn(\delta_i - r_{CS}(l_i, r_i))\sqrt{2d_i}$,
- Latency deviance residuals: $r_{D,lat}(l_i, r_i) = sgn(\delta_i - r_{CS,u}(l_i, r_i))\sqrt{2d_{i,lat}}$,
- Incidence deviance residuals: $r_{D,inc}(l_i, r_i) = sgn(\xi_i - \hat{p}_i)\sqrt{2d_{i,inc}}$.

In the above expressions, $sgn(\cdot)$ represents the sign function and $r_{CS}(l_i, r_i)$ and $r_{CS,u}(l_i, r_i)$ are the Cox-Snell residuals for S and S_u, respectively; see

(4.2). As in classical survival analysis, deviance residuals can, for example, be plotted against the estimated linear predictors ($\hat{\mu}_i$ or $\hat{\nu}_i$) or against the explanatory variables in the linear predictors, and any unusual pattern may indicate an omission of an important covariate or a non linear effect of one or more covariates in the latency part, in the incidence part, or in both parts.

4.3 Simulation study

With this simulation study, we aim to illustrate the behavior of global and uncured Cox-Snell residuals in different settings; and to show that incidence and latency deviance residuals allow to correctly detect the need for a transformation in a covariate in the incidence and/or latency.

4.3.1 Simulation setting

Data are generated following the algorithm described in Section 2.2 for right-censoring, except that Equations (2.1) and (2.2) are replaced by the following equations:

1. Scenario A: All terms are linear, both in the latency and incidence part of the model

$$\begin{cases} \log(T) = -0.5 + X_1 - X_2 - X_3 + 0.5\varepsilon \\ p(\boldsymbol{X}) = \phi(\gamma_0 - X_2 + X_3) \end{cases}$$

2. Scenario B: A quadratic term is included in the latency part of the model

$$\begin{cases} \log(T) = 1.5 + X_1 - X_2^2 - X_3 + 0.5\varepsilon \\ p(\boldsymbol{X}) = \phi(\gamma_0 - X_2 + X_3) \end{cases}$$

3. Scenario C: A quadratic term is included in the incidence part of the model

$$\begin{cases} \log(T) = -0.5 + X_1 - X_2 - X_3 + 0.5\varepsilon \\ p(\boldsymbol{X}) = \phi(\gamma_0 - 0.5X_2^2 - X_3) \end{cases}$$

4. Scenario D: A quadratic term is included in the latency and in the incidence parts of the model

$$\begin{cases} \log(T) = -1 + X_1 - X_2^2 - X_3 + 0.5\varepsilon \\ p(\boldsymbol{X}) = \phi(\gamma_0 - 0.5X_2^2 - X_3). \end{cases}$$

In these equations, ϕ is taken to be either a logit or probit function. Whenever necessary, we use the notation A_l and A_p to distinguish scenario A using ϕ=logit from scenario A using ϕ=probit, and equivalently for scenarios B, C and D. This will allow to study the impact on residuals checking of a misspecified link function in the incidence, since we will always use logit in our fitting procedure. Regarding the time-to-event, the error term ε is generated following the extreme-value distribution, so that T_u, the time to event of the uncured observations, follows a Weibull distribution. Note that in this case, our simulation study covers both the PH and the AFT model, as both models are equivalent when the event times follow a Weibull distribution [Collett, 2003]. For each scenario, three proportions of right-censored observations (including the cured observations, which are always right-censored) are covered: light (20% cured, 30% right-censored), medium (30% cured, 40% right-censored), and heavy (50% cured, 60% right-censored).

This simulation setting leads to a total of 24 settings (four scenarios, two link functions, three right-censoring/cure proportions). To study the effect of interval-censoring on residuals checking, two other scenarios complement the simulation study: event times from scenario B_l and D_l that were exactly observed, are now interval-censored, leading to six additional settings: two scenarios and three levels of cure/right-censoring. Interval censoring was obtained following item 4b) in Section 2.2. These two additional scenarios allow a visual assessment of the impact of analyzing interval-censored (in addition of right-censored) observations instead of exactly observed event times.

Values of γ_0, τ, and generation of covariates and given in Table 4.1. The sample size was set to 500, and we replicated 500 datasets. For obvious reasons of space, only the main results are presented in this chapter; further results are given in Appendix C.2.

4.3.2 Simulation results

Cox-Snell residuals

For each setting, we have fitted a mixture cure model assuming a logistic regression (ϕ=logit) in the incidence; and an AFT model for the latency part with either a log-Normal, a Weibull or an EGG distribution. All covariates (X_1, X_2

Table 4.1 – Value of parameter and distribution of variables for each of the 4 scenarios and 3 levels of right-censoring and cure

RC level		Scenario A	Scenario B	Scenario C	Scenario D
	X_1	$\mathcal{N}(0.5, 1.5)$	$\mathcal{N}(0.5, 1.5)$	$\mathcal{N}(0.5, 1.5)$	$\mathcal{N}(0.5, 1.5)$
	X_2	$\mathcal{U}(-3, 3)$	$\mathcal{U}(-3, 3)$	$\mathcal{U}(-3, 3)$	$\mathcal{U}(-3, 3)$
	X_3	$Bern(0.5)$	$Bern(0.5)$	$Bern(0.5)$	$Bern(0.5)$
Light RC	γ_0	2	2	3.5	3.5
Medium RC	γ_0	1	1	3	3
Heavy RC	γ_0	0.5	0.5	1.5	1.5
Light RC	τ	30	20	30	20
Medium RC	τ	30	20	20	10
Heavy RC	τ	30	5	20	10

and X_3 for the latency and X_2 and X_3 for the incidence) were included in the model, assuming a linear effect. The following results are supported by Figures 4.1 and 4.2, which show the global Cox-Snell residuals for 500 datasets.

In the following, it is important to keep in mind that the Weibull and log-Normal distributions are special cases of the EGG distribution. Therefore, since the true generating distribution is Weibull, the plots of the Cox-Snell residuals for the models fitted with the EGG distribution are expected to be similar to those obtained with the Weibull distribution, or better aligned if the Weibull model is actually misspecified. Furthermore, a plot of Cox-Snell residuals based on the log-Normal distribution is expected to show more departure from the straight line than a plot of Cox-Snell residuals based on the EGG or Weibull distribution.

A first observation is that the use of the indicator random variable $\mathbb{1}(\xi_i > 0.5)$ provides a prediction of the uncured status which is satisfactory to compute the Cox-Snell residuals. Indeed, plots of the uncured Cox-Snell residuals based on the cutoff 0.5 are extremely similar to those based on the true uncured status Y_i, as shown in Figures 11 to 18 in Appendix C.2. Furthermore, although a larger cutoff could prevent misclassifying a cured individual as an uncured individual, we only observe extremely small sensitivity to the choice of the cutoff value.

Second, as global and uncured Cox-Snell residuals may be impacted in a similar way by a misspecification in the incidence part, we expect that a plot of the global Cox-Snell residuals will lead to the same conclusion as a plot of the uncured Cox-Snell residuals. We therefore use simulations to compare the global and uncured Cox-Snell residuals. In every scenarios, the conclusions are

indeed similar, and in the following, for the sake of brevity, we will only discuss results obtained with the global Cox-Snell residuals.

Third, we investigate the impact of the right-censoring and cure proportion on the ability of the Cox-Snell residuals to correctly identify the Weibull distribution as the true generating distribution. In every scenario, and for all three proportions of cured and right-censored observations, the Weibull distribution is always correctly chosen over the log-Normal. This conclusion is supported by Figure 4.1, showing Cox-Snell residuals based on the three distributions, and related to scenario C_l for the three levels of cured and right-censored proportion. However, as the proportion of right-censored observations increases, the distinction becomes less evident: compare the plots in Figure 4.1 from top to bottom, i.e. from light to heavy censoring.

We have also investigated whether the Cox-Snell residuals can still be used to identify the most appropriate underlying distribution even in the presence of a misspecification in the incidence or the latency part of the model. In this simulation study, two misspecifications in the incidence part are covered: assuming a logit link instead of a probit link, as in scenarios A_p, B_p, C_p and D_p; and assuming linearity in covariates whereas a transformation is needed, as in scenarios C and D. In the latency part, one type of misspecification is covered: assuming linearity in covariates whereas a transformation is needed, as in scenarios B and D. The Cox-Snell residuals plots support the Weibull and EGG distribution over the log-Normal distribution in every setting: Figure 4.2, giving Cox-Snell residuals for the medium level of cured/right-censored proportions of scenarios C_l, D_l and D_p, illustrates this conclusion in relation to the non-linear covariate in incidence (first row), to the non-linear covariate in both latency and incidence (second row), and to the non-linear covariate in both latency and incidence in addition of a wrong link function in incidence (third row).

Note also that a misspecification in the latency part is reflected in the Cox-Snell residuals plot for models fitted with the Weibull distribution (see second and third row of Figure 4.2 based on scenario D_l and D_p). As the EGG distribution is more flexible, the Cox-Snell residuals plots for models fitted with the EGG distribution show less departure from the straight line. We conclude that Cox-Snell residuals in a mixture cure model are effective to distinguish between two possible distributions even in the case of misspecification in incidence or latency.

The last objective of this simulation study is to discuss the impact of interval-censoring on residuals checking. Unlike the Kaplan-Meier estimator, the Turnbull estimator does not show jumps at every event time, but rather at the set of unique values of the endpoints of the interval. This yields a visual inspection

more cumbersome than if data were exactly observed instead of being contained within an interval. For the sake of visibility, we show the Cox-Snell residuals plot for one dataset only, based on Scenario C_l: the first row of Figure 4.3 (without interval-censoring) can be compared to the second row of Figure 4.3 (with interval-censored event times): interval-censored residuals show more departure from a straight line. The distinction between two possible distributions is then more complicated.

Deviance residuals

In this subsection, a mixture cure model is fitted to each generated dataset assuming a logistic regression in the incidence and a Weibull survival distribution in the latency, and assuming that covariates have a linear effect in both parts of the model. Three types of deviance residuals plots are displayed: global, latency and incidence deviance residuals, versus the covariate X_2, which needs a transformation for scenario B, C and D. For one dataset, it is common to plot the deviance residuals on the y-axis versus the covariates values on the x-axis. To help distinguishing a pattern, one generally adds a lowess smoothing curve to the plot [Cleveland, 1979].

In this simulation study, each plot concerns 500 datasets. This means 500×500 residuals points and 500 lowess smoothing curves are superimposed on each plot. To enhance visibility, points are deleted from the graphs, so that only the 500 lowess smoothing curves are shown on each plot. Latency deviance residuals are expected to show a quadratic trend for scenarios B and D; and incidence deviance residuals are expected to show a quadratic trend for scenarios C and D.

As a first point, we observe that global, latency and incidence deviance residuals show a trend when appropriate. This can be seen by looking at Figure 4.4, giving deviance residuals for scenarios C_l, D_l and D_p for a medium cured and right-censored proportion. However, when a transformation is needed in both the latency and the incidence part, a trend may be hidden in the global deviance residuals plot, as can be seen on the global deviance residuals plot for medium censoring in the second row of Figure 4.4 . Hence the importance of checking the latency and incidence deviance residuals as well.

Second, latency deviance residuals are somewhat affected by the increase of the right-censored proportion: a high censoring proportion can hide a trend in the smoothing curve (see latency deviance residuals of Figure 4.5, where the three types of deviance residuals are shown for the three levels of cured/right-censored proportion of scenario B_l: the lowess lines does not reveal a curve for heavy censoring), or suggest another transformation than the true one (see latency residuals of Figure 4.4, where some curves may suggest a cubic transformation).

When looking at a deviance residual plot for only one data set, as given in Figure 4.6 for scenario B_l for example, a trend is highly detectable when right-censored and uncensored observations are plotted in different colors. Plotting right-censored residuals with a different symbol may thus circumvent the issue, and, following [Collett, 2003], we recommend to do so.

Third, the same issue may arise with incidence deviance residuals with a too low cured proportion. Indeed, less cured observations implies less information for the modeling of the incidence part.

4.4 Real data analysis

In this section, we demonstrate how our proposed residuals may bring some further insight in the analysis of the data from the Alzheimer's disease study. We are interested in applying Cox-Snell and deviance residuals for checking the fit of a mixture cure model assuming either an AFT with a log-Normal, a Weibull, or an EGG distribution in the latency part. This analysis evaluates the association between the CAMCOG score and the time to MCI conversion, adjusting for gender, age at baseline, number of folate cells, total Homo-cysteine (tHcy) rate, Mini-Mental State Examination (MMSE) score, and expression of the APO E4 gene (APOEE4). Maximum likelihood estimates are given in Table 4.2. A visual inspection of Cox-Snell residuals given in Figure 4.7 suggests that the log-Normal distribution does not fit the data as well as the Weibull and EGG distribution, the latter two giving quite similar plots. A likelihood ratio test nevertheless rejects the null hypothesis of a Weibull distribution for the event times, and the EGG is then preferred for the remainder of the analysis. Global deviance residuals do not exhibit strong pattern when plotted against any of the five covariates Age, CAMCOG, Folate, MMSE or tHcy. The same conclusion is reached for latency and incidence deviance residuals plots; see Figure 4.8. Therefore, with data and covariates at hand, and based on Cox-Snell and deviance residuals, we can conclude that there is no need to reconsider the use of the EGG distribution in this analysis, nor to reconsider linearity of our variables in the latency and incidence.

4.5 Discussion

In this chapter, we proposed to check the assumptions of a parametric PH or AFT mixture cure model by plotting newly proposed residuals of the model. Although widely used in linear regression, and also in classical survival analysis, diagnostic checks based on residuals have not been studied in the context of possibly interval-censored data with a sub-population of cured individuals.

This chapter focused on extending both the Cox-Snell and deviance residuals to this setting, allowing to check various assumptions of the fit of parametric mixture cure models for right- and possibly interval-censored event times.

In parametric mixture cure models, assumptions are made on the survival distribution of the uncured observations, and we discussed the use of Cox-Snell residuals to detect if those assumptions are correct. A major challenge in this situation is clearly that the uncured status is unobserved for right-censored data and that the global survival distribution is improper. In this chapter, we have shown that global Cox-Snell residuals, based on the improper survival distribution of the entire population, still follows a mean one exponential distribution if the model is correctly fitted. Furthermore, we have also defined uncured Cox-Snell residuals, computed based on an estimation of the uncured status for each observation. Our simulation study demonstrates that global and uncured Cox-Snell residuals can suitably be used to assess the hypothesis about the survival distribution in the latency part, even if the incidence part is incorrectly specified.

Another important feature of these models is that covariates are typically entered in the model in a linear fashion, both in the incidence and the latency part of the model. However, covariates may have a non-linear effect in either or both parts of the model. We thus also proposed deviance-based residuals allowing to detect if a covariate needs a transformation, and in which part of the model. Additionally, we have shown that non-linearity in latency and in incidence is correctly detected by plotting the latency and incidence deviance residuals against a covariate. As expected, a heavy right-censored proportion may hide a trend in the plot and prevent detecting the need for a transformation in the latency. A low cured fraction has an identical effect in incidence. Like in other models, such residuals plots should be interpreted with care since, as stated by [Collett, 2003], no pattern in the residuals plot does not imply a correct model but rather the absence of reasons to think it is incorrect.

The subjective nature of residual checking techniques is sometimes criticized, and the use of goodness-of-fit hypothesis test may be advocated. However in practice, these graphical checks are still widely used. For example, to check the normality of a variable, one generally agrees that a Q-Q plot or a simple histogram is more meaningful than a formal normality test, which results in only one piece of information, largely impacted by the sample size: whether the hypothesis is rejected or not. In classical survival analysis, several authors stress the importance of residuals checking in the modelling process [Therneau et al., 1990, Collett, 2003, Klein and Moeschberger, 2003, Lawless, 2003].

The residuals presented in this chapter have been implemented in parametric mixture cure models, and an extension to semi-parametric mixture cure models

should be straightforward, as only the final estimates of the model are required in the definition of Cox-Snell and deviance residuals. These residuals can therefore also be used to check a semi-parametric mixture cure model, such as the ones discussed in [Sy and Taylor, 2000, Zhang and Peng, 2012]. However, Cox-Snell residuals are not as useful in a semi-parametric context, due to the fact that the baseline survival distribution has to be estimated non-parametrically [Collett, 2003], so that the approximation to the unit exponential distribution is less likely to hold. Deviance residuals, on the other hand, do not suffer the same issues and can be used in a semi-parametric context to check the linearity of the covariates.

Our results are similar to those obtained by Peng and Taylor [Peng and Taylor, 2016]. First, they show that the Cox-Snell residuals based on the global survival distribution do follow an exponential distribution in $[0, -\log(p))$. They can thus be used as usual, i.e. a plot of the estimated cumulative hazard should show points aligned on a straight line of slope one. Second, they also define an equivalent to our uncured Cox-Snell residuals, but they use a weighted Cox-Snell residuals instead of using a cutoff value (i.e. $\xi_i < 0.5$). Thirdly, as their article is based on right-censored data only, they use martingale residuals to detect a needed transformation in the latency part of the model. The idea is similar to our deviance residuals, but their method can only be used in a Cox model with right-censoring, whereas our method is more general.

Other types of residuals could be extended to the mixture cure model, with or without interval-censored data. Residuals to detect outliers and influential values have been extended to the case of cure models [Ortega et al., 2008], but the extension of residuals to check either the proportionality of hazards or the AFT assumptions within a mixture cure model is subject of future work.

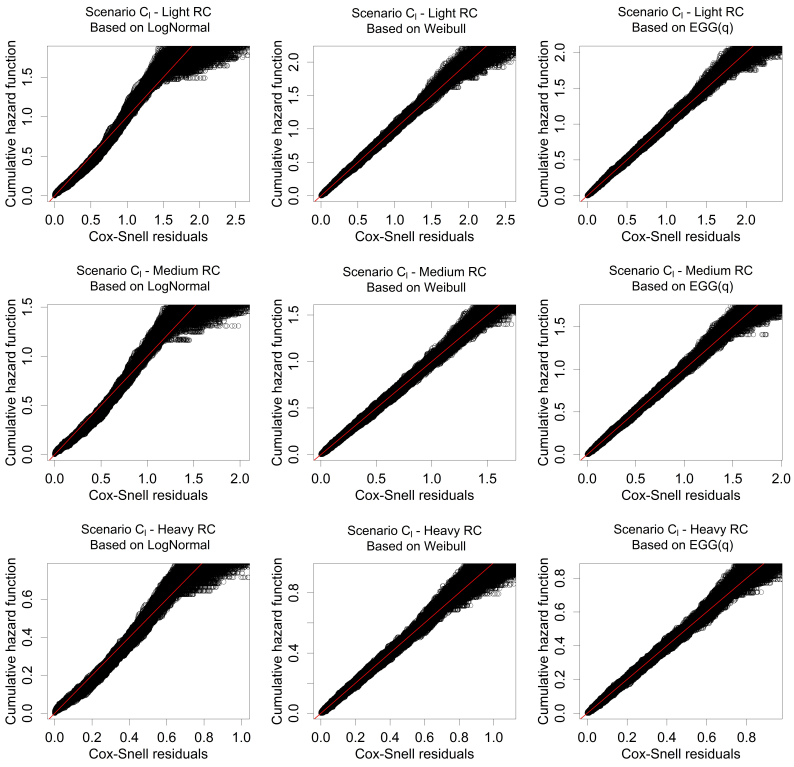


Figure 4.1 – Global Cox-Snell residuals for scenario C_l. Top to bottom: light to heavy cured/right-censored proportion; left to right: based on log-Normal, Weibull, and EGG(q). Residuals for 500 datasets are superimposed. The EGG and Weibull distributions are chosen over the log-Normal distribution.

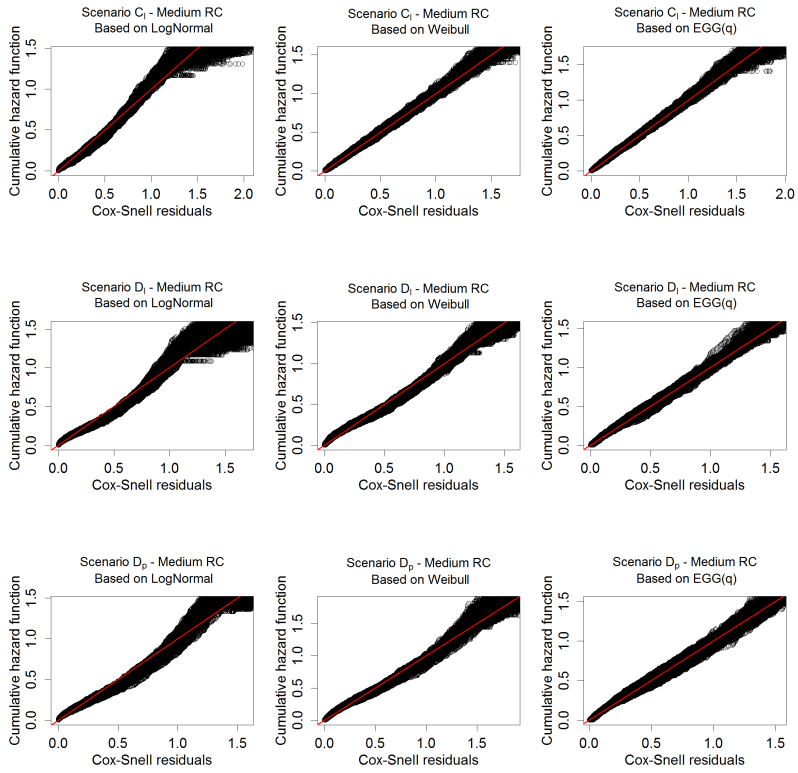


Figure 4.2 – Global Cox-Snell residuals for scenarios C_l, D_l, and D_p, for medium censoring. Left to right: based on log-Normal, Weibull, and EGG(q). Residuals for 500 datasets are superimposed. The EGG and Weibull distributions are chosen over the log-Normal distribution for each scenario.

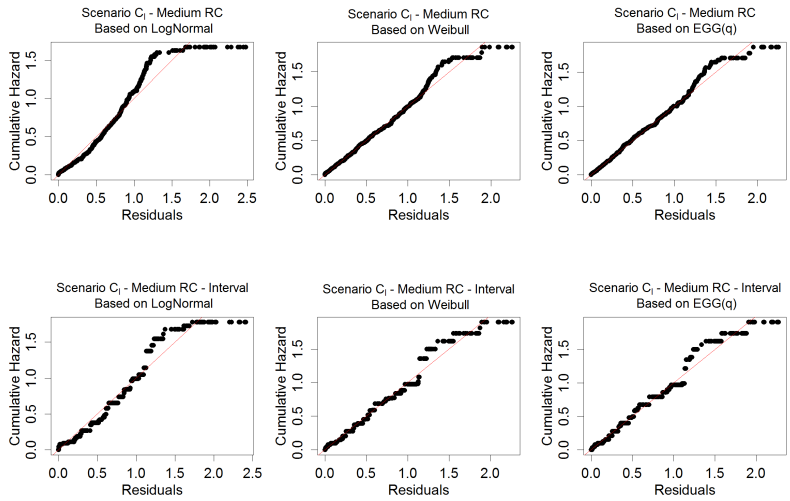


Figure 4.3 – Global Cox-Snell residuals for scenario C_l with medium cured/right-censored proportion. Above: exact event times data, bottom: interval-censored data. Left to right: based on log-Normal, Weibull, and EGG(q). Residuals for one dataset only. Interval-censoring makes a visual inspection more cumbersome, but possible.

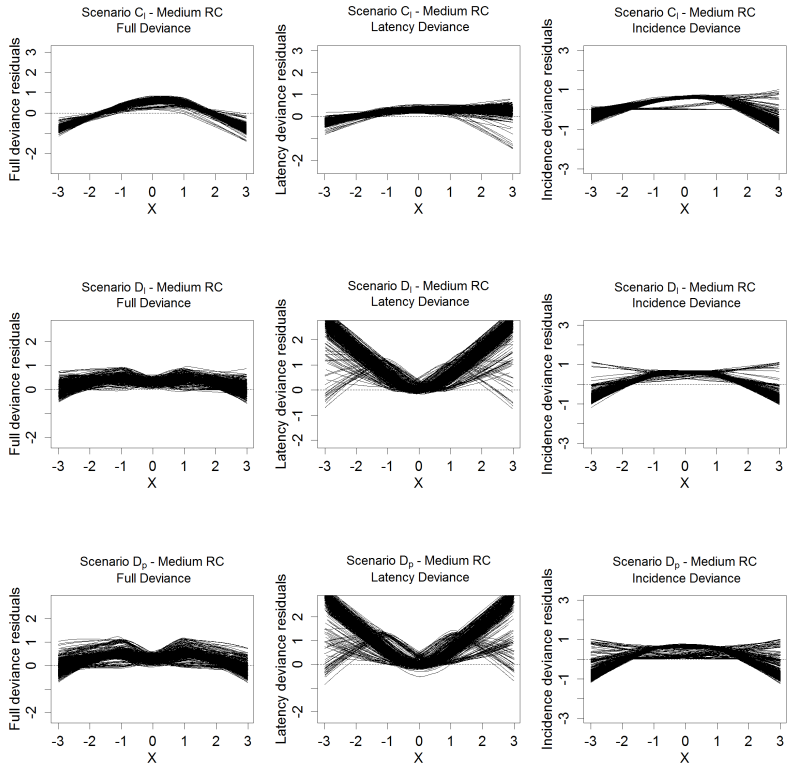


Figure 4.4 – Deviance residuals for scenario C_l (first row), D_l (second row), and D_p (third row) for medium censoring. Left to right: global, latency and incidence deviance residuals. Lowess smoothing curve for 500 datasets are superimposed. A curve is detectable when appropriate.

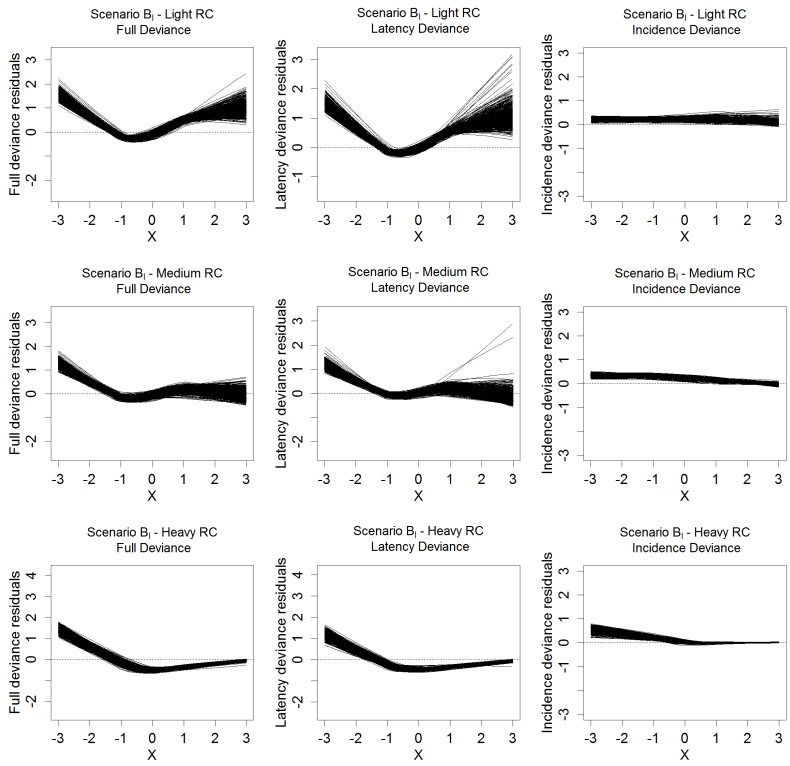


Figure 4.5 – Deviance residuals for scenario B_l. Top to bottom: for light to heavy cured/right-censored proportion; left to right: global, latency and incidence deviance residuals. Lowess smoothing curve for 500 datasets are superimposed. A heavy right-censored proportion may hide a trend in the latency deviance residuals plots.

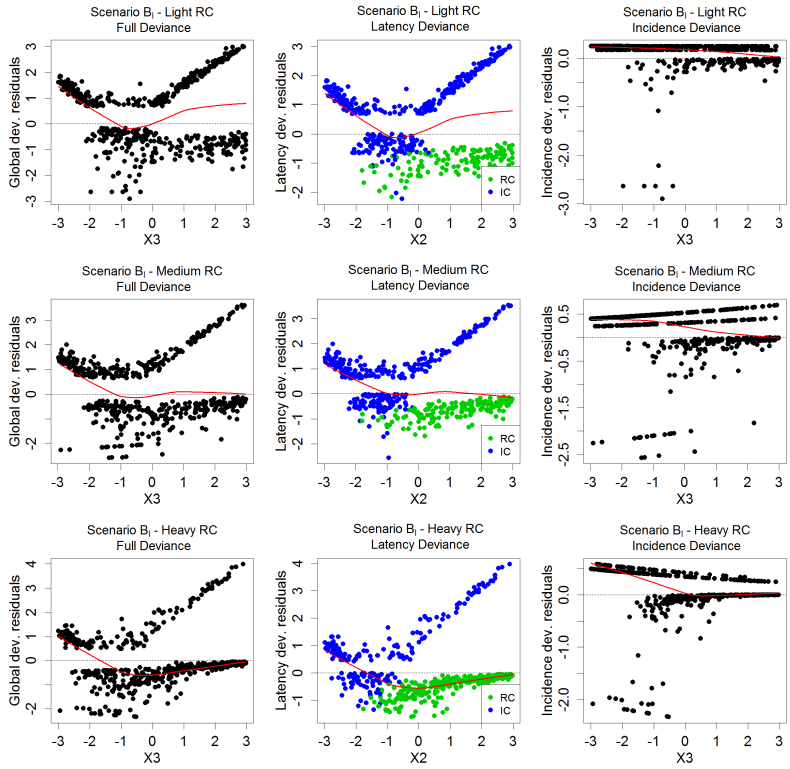


Figure 4.6 – Deviance residuals for scenario B_l. Top to bottom: for light to heavy cured/right-censored proportion; left to right: global, latency and incidence deviance residuals. Residuals and lowess smoothing curve for one dataset only. Plotting right-censored observations in a different color may help distinguishing a pattern.

Table 4.2 – Parameter estimates and standard errors of the EGG-AFT mixture cure model for the AD database, with scaled covariates

Parameter	Estimates	SD	P-value	Exp (Estimates)
Latency part: EGG-AFT model				
q	12.12	33.25	0.37	
Intercept Latency	2.75	0.12	0.00	
Age	-0.21	0.08	0.01	0.81
CAMCOG	0.31	0.07	0.00	1.36
tHcy	-0.10	0.05	0.06	0.90
Folate	-0.13	0.07	0.08	0.87
MMSE	-0.10	0.07	0.16	0.91
Gender	-0.04	0.06	0.32	0.96
APOE E4	-0.05	0.06	0.26	0.95
$\log(\sigma)$	-2.77	2.75	0.24	0.06
Incidence part: Logistic regression				
Intercept Incidence	6.06	5.88	0.23	
Age	1.19	1.26	0.26	3.28
CAMCOG	-1.24	1.28	0.25	0.29
tHcy	-0.55	1.03	0.35	0.57
Folate	-0.66	0.71	0.26	0.52
MMSE	-5.31	5.99	0.27	0.00
Gender	1.42	0.86	0.10	4.12
APOE E4	-0.05	0.56	0.40	0.95

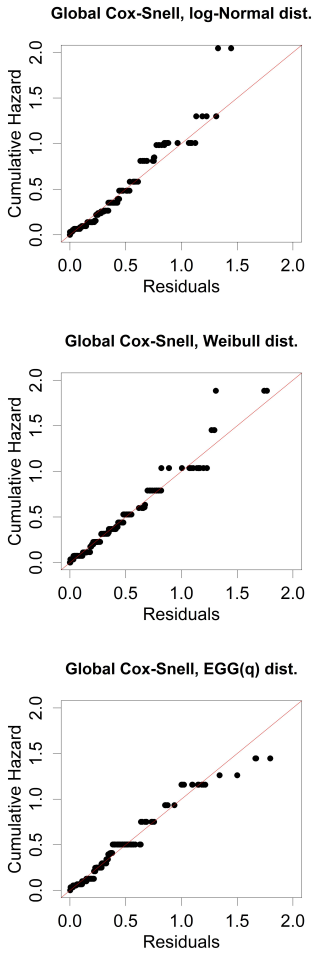


Figure 4.7 – Global Cox-Snell residuals for the MCI database, based on EGG-AFT mixture cure model. The EGG distribution is best supported by the data.

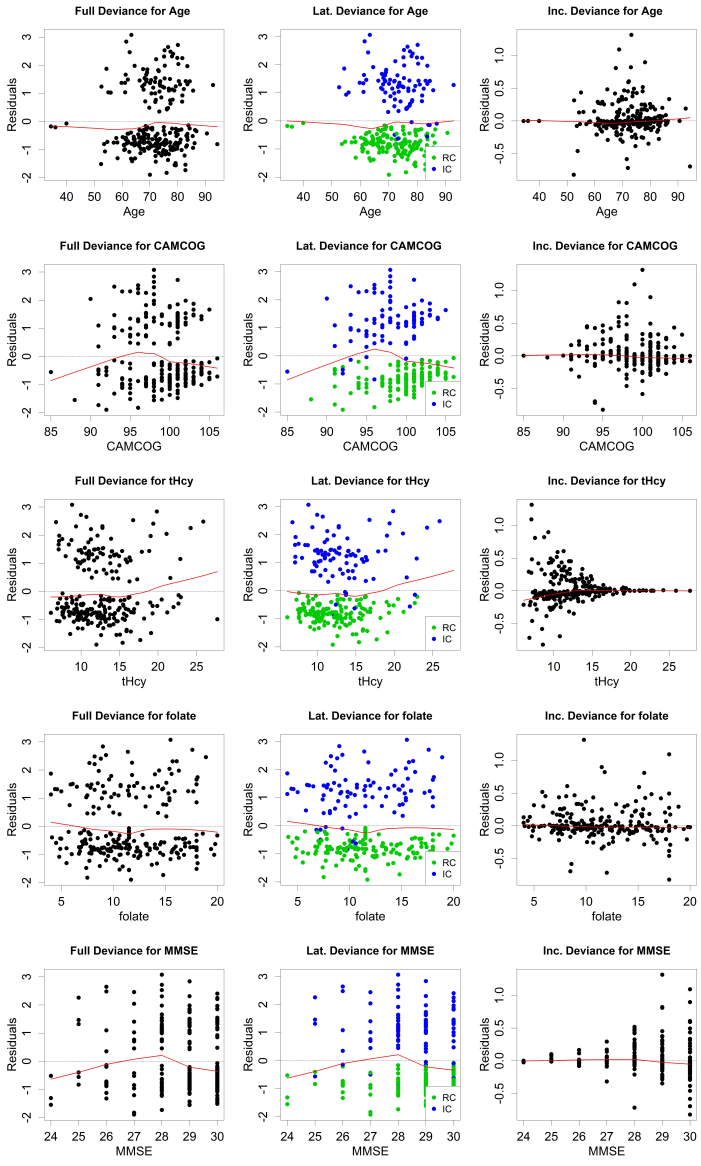


Figure 4.8 – Global, latency and incidence deviance residuals for the MCI database, based on EGG-AFT mixture cure model, for Age, CAMCOG and tHcy, folate and MMSE. No trend is detectable from the plots.

CHAPTER 5

Estimation and Goodness-of-fit with the SNP representation

"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."

- J.R.R. Tolkien (1892-1973)

This chapter is mainly based on the paper "The SNP representation in mixture cure models with interval-censoring: estimation and goodness-of-fit testing" By Scolas S., El Ghouch A., Oulhaj A. and Legrand C. to be submitted.

Contrary to what is generally assumed in survival analysis, a fraction of the population under study may not develop the event of interest. This special feature is more and more encountered, and referred to as the presence of cure fraction. It is also the case that in many other medical studies, patients come at scheduled interviews. The event times are thus not exactly observed, but interval-censored in addition to the possibility of being right-censored as usual. A mixture cure model can handle the cure fraction. This model consists of two parts: the incidence part models the probability to be cured, while the latency part models the event times for those who will eventually experience the event. When an AFT regression model is used in the latency, a parametric distribution for the error term is typically assumed in order to use classical maximum

likelihood theory. In some cases however, we do not want to use a parametric distribution, or usual parametric distributions do not fit the data correctly enough. In this paper, we present the semi-nonparametric (SNP) representation, which offers a solution to both issues: it is proposed as an alternative to parametric models; and as a goodness-of-fit test, because it can test whether an assumed parametric distribution correctly fits the data.

5.1 Introduction

In the previous chapters, a parametric distribution was assumed for the survival of the uncured individuals, leading to a parametric mixture cure model. Indeed, a semi-parametric AFT model is considerably complicated if the event times are interval-censored in addition of being right-censored. Assuming a parametric form for the distribution of the event times is thus very appealing. Furthermore, with the availability of remarkably flexible distributions (such as the Extended Generalized Gamma (EGG) [Lawless, 1980] or the Generalized F [Peng et al., 1998] distribution), criticisms about parametric techniques do not carry the same weight as in the past. An even more flexible method can be considered: the semi-non parametric (SNP) representation of [Gallant and Nychka, 1987], which has demonstrated considerable effectiveness for estimation purposes.

Based on complex mathematical developments, the SNP representation is nonetheless very popular and has been applied in a variety of domains [Gallant et al., 1990, Gabler et al., 1993, Davidian and Gallant, 1992, Chen et al., 2002, Song et al., 2002]. This popularity is certainly due to the intuitive concept and the effectiveness of the method. To approach a sufficiently 'smooth' density $h(z)$, the basic idea is to extend a base density $\varphi(z)$ (e.g. the standard Normal density) by a polynomial $P_k(z)$ of some degree k: $h(z) \approx \varphi(z)P_k^2(z)$. By doing so, estimation can be achieved as in a parametric context [Zhang and Davidian, 2008]. [Zhang and Davidian, 2008] used the SNP representation as an estimation method in a survival analysis context, but the maximum percentage of right-censored observations in their simulations is 25%, whereas the right-censoring rate is typically higher in the presence of cured individuals.

The SNP representation provide an excellent estimation method and can actually also be used for goodness-of-fit purposes. This is interesting when, for example, we want to confirm previous empirical findings that the data follow a known parametric distribution. Indeed, a parametric distribution has readily interpretable characteristics and particular behavior (for instance an increasing or decreasing hazard rate depending on the value of a parameter). Using the appropriate parametric distribution may enhance the interpretation of the results and lead to more powerful analyses. It is nonetheless undeniable that a false assumption about the survival distribution can lead to considerable consequences about the interpretation of the results. This is why it is of the utmost interest to test whether a particular assumed distribution is a correct candidate for the available data.

For the purpose of goodness-of-fit testing, the SNP representation is very intuitive as well. The polynomial $P_k(z)$ represents the extend of departure from the base density φ. By defining φ as the normal density for instance, one could test whether the data are normally distributed or not by measuring the

extend of departure from the base density. In a general paper, [Aerts et al., 1999] lay the foundations of this test, which is further developed in [Nysen et al., 2012] for the more specific purpose of testing an assumed density in an arbitrarily censored time-to-event context (with possibly left-, right-, interval-censored and exactly observed data). Their simulations, however, concern a model without covariates.

The objective of this chapter is twofold. First, we propose a very flexible AFT mixture cure model based on the SNP representation of the survival distribution of the error term in the latency part. This allows to easily take interval-censoring into account by using the likelihood function. We will show, via simulations, that this method provides good estimation in the context of a mixture cure model and with up to 60% right-censored observations; as is commonly encountered when cure are present. Second, we assess the SNP representation in the context of goodness-of-fit testing in the presence of a cure fraction, right-censoring, interval-censoring, and covariates in both parts of the mixture cure model.

This chapter is organized as follows: Section 5.2 describes the SNP representation, apply it for estimation in our model, and explains how the SNP representation can be used for goodness-of-fit testing. The results of a large scale simulation study are reported in Section 5.3. We analyze the previously mentioned Alzheimer's disease database in Section 5.4 and ends with a discussion in Section 5.5.

5.2 Methodology

In this section, we first summarize the SNP representation as described in [Zhang and Davidian, 2008]. We then specify how this representation can be used for estimation purposes in a mixture cure model, and end the section by applying the goodness-of-fit procedure proposed by [Aerts et al., 1999] and [Nysen et al., 2012] in our setting.

Assuming any density function, say f_0, to be smooth enough (see [Gallant and Nychka, 1987] for details), the SNP representation provides the following approximation to f_0:

$$f_0(z) \approx P_k^2(z)\varphi(z) = (a_0 + a_1 z + \cdots + a_k z^k)^2 \varphi(z)$$

for a fixed $k \in \{0, 1, \cdots\}$; where $\varphi(z)$ is a known kernel density with known moment generating function, and where the coefficients $\boldsymbol{a}' = (a_0, a_1, \cdots, a_k)$ are chosen to satisfy the density condition $\int_{-\infty}^{\infty} P_k^2(z)\varphi(z)dz = 1$. Extensive math-

ematical details concerning the approximation as well as convergence properties can be found in [Gallant and Nychka, 1987].

To ensure that the SNP representation is a density, [Zhang and Davidian, 2001] propose to transform the Cartesian coordinates $\boldsymbol{a}$ in spherical coordinates. The idea goes as follows. Let Z be a random variable with density φ, and $\boldsymbol{W}' = (1, Z, \cdots, Z^k)$, the condition can be rewritten as

$$\int_{-\infty}^{\infty} P_k^2(z)\varphi(z)dz = E(P_k^2(Z)) = \boldsymbol{a}'E(\boldsymbol{W}\boldsymbol{W}')\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{A}\boldsymbol{a} = 1,$$

with $\boldsymbol{A} = E(\boldsymbol{W}\boldsymbol{W}')$. The known matrix $\boldsymbol{A}$ is positive definite, so that the Cholesky decomposition $\boldsymbol{B}$ of $\boldsymbol{A}$ exists: $\boldsymbol{A} = \boldsymbol{B}'\boldsymbol{B}$. Let $\boldsymbol{c} = \boldsymbol{B}\boldsymbol{a}$, then the condition becomes $\boldsymbol{a}'\boldsymbol{A}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{a} = \boldsymbol{c}'\boldsymbol{c} = 1$. The last equality suggests that $\boldsymbol{c}' = (c_1, \cdots, c_{k+1})$ are points lying on a sphere of radius one, and thus a spherical transformation of the original coordinates $\boldsymbol{a}$, that is :

$$\begin{cases} c_1 & = \sin(\rho_1) \\ c_2 & = \cos(\rho_1)\sin(\rho_2) \\ \vdots & \\ c_k & = \cos(\rho_1)\cos(\rho_2)\cdots\cos(\rho_{k-1})\sin(\rho_k) \\ c_{k+1} & = \cos(\rho_1)\cos(\rho_2)\cdots\cos(\rho_{k-1})\cos(\rho_k), \end{cases}$$

with $-\pi/2 < \rho_j \leq \pi/2$ for $j = 1, \cdots k-1$ to achieve a unique representation, and $\rho_k \in [0, 2\pi]$. The original coordinates can thus be found with $\boldsymbol{a} = \boldsymbol{B}^{-1}\boldsymbol{c}$.

5.2.1 The SNP representation as an estimation method

To cope with interval-censored data in a mixture cure model, we will assume that the error term in the AFT model can be approximated by using the SNP representation. In the SNP representation $P_k^2(z)\varphi(z)$, the degree k of the polynomial P_k controls the extent of departure from the base density φ, so that a higher degree k brings more flexibility in approximating the true density. To avoid over-parametrization, a model selection criterion penalizing for the number of parameters, such as the AIC or BIC, is used to determine the 'optimal' degree k of the polynomial. Moreover, as k increases, the number of models to be fitted increases, as does the number of parameters to estimate in one model. It is thus computationally interesting to limit the degree of the polynomial to a maximal value, say k_{max}, so that $k \in \{0, 1, \cdots, k_{max}\}$.

According to [Zhang and Davidian, 2001] and [Fenton and Gallant, 1996], a large majority of densities can be well approximated with the SNP representa-

tion using the standard normal density as base density, along with a polynomial of degree no larger than $k_{max} = 4$ (an illustration of several SNP hazard functions is given in Figure 5.1). Some densities, that may be very skewed for example, are nonetheless better approximated when the exponential density is used as base density. Allowing the choice of either the normal or the exponential distribution in the SNP representation allows to further reduce the maximal degree of the polynomial. Based on extensive simulations, [Zhang and Davidian, 2008] showed that in this case, a degree no larger than $k_{max} = 2$ for each base density provides a satisfactory estimation.

Using the SNP representation in a survival analysis context is straightforward, and we derive the formulation of the survival distribution in an AFT model, with either the normal or exponential density as base density. In the first case, we assume that the density function of ε is approximated by $P_k^2(z)\varphi(z)$, with $\varphi(z) = (2\pi)^{-1/2}e^{-z^2/2}$. Note that in this case, when $k = 0$, the event times for uncured are log-normally distributed. For $k \in \{0, 1, \cdots\}$, the conditional survival distribution of uncured observations is then given by

$$S_{u,k}(t_i) \quad = \quad \int_{(\log(t_i)-\boldsymbol{x}_i'\boldsymbol{\beta})/\sigma}^{\infty} P_k^2(v)\varphi(v)dv.$$

By replacing S_u by $S_{u,k}$ in the likelihood function given in Equation (1.3), the maximum likelihood estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ can be found using a Newton-Raphson type algorithm. For computational purposes, it is interesting to observe that

$$\int_z^{\infty} P_k^2(v)\varphi(v)dv = \int_z^{\infty} \sum_{j=0}^{2k} d_j v^j \varphi(v)dv = \sum_{j=0}^{2k} d_j I(j,z), \tag{5.1}$$

with d_j, $j = 0, 1, \cdots, 2k$ being a function of the elements of c_i's, and $I(j,z) = \int_z^{\infty} v^j\varphi(v)dv$. In the normal basis density case, $I(j,z)$ can be defined recursively as:

$$\begin{cases} I(0,z) & = 1 - \phi(z) \\ I(1,z) & = \varphi(z) \\ I(j,z) & = z^{j-1}\varphi(z) + (j-1)I(j-2,z) \text{ for } j \geq 2 \end{cases}$$

where $\phi(\cdot)$ is the standard normal cumulative distribution function. This closed form expression offers computational advantages, as only the integration of the normal density is required.

In the exponential case, the density of $\varepsilon^* = \exp(\varepsilon)$ is approximated by $P_k^2(z)\varphi(z)$, with $\varphi(z) = \exp(-z)$. In this case, when $k = 0$, the error term ε follows an extreme-value distribution and the event times a Weibull distri-

bution. For $k \in \{0, 1, \cdots\}$, the conditional survival distribution for uncured observations is then given by:

$$S_{u,k}(t_i) = \int_{(t_i/e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^{1/\sigma}}^{\infty} P_k^2(z)\varphi(z)dz,$$

and as for the normal base density, $\hat{\boldsymbol{\eta}}$ can be obtained by maximizing the likelihood (1.3) where S_u is replaced by $S_{u,k}$. Note also that in the exponential case, $I(j, z)$ introduced in (5.1) can be defined recursively as:

$$\begin{cases} I(0,z) & = e^{-z} \\ I(j,z) & = z^j\varphi(z) + jI(j-1,z) \text{ for } j \geq 1. \end{cases}$$

Other density functions can be chosen as base density, but a recursive expression as in the normal and exponential case may not be available, so that integration is unavoidable.

For each base density, denote $\hat{l}_k(\hat{\boldsymbol{\eta}}; \mathcal{O})$ the maximized log-likelihood at its MLE $\hat{\boldsymbol{\eta}}$, corresponding to the k-th model. The choice of the best combination of density and degree k can be made with the popular AIC criterion, defined as:

$$AIC(k) = \hat{l}_k(\hat{\boldsymbol{\eta}}; \mathcal{O}) - (k + m + 2),$$

where m is the number of covariates in the latency part of the mixture cure model. The model with the density and degree k leading to the largest AIC is chosen as the best model [Zhang and Davidian, 2008]. Note that the BIC can also be used: $BIC(k) = \hat{l}_k(\hat{\boldsymbol{\eta}}; \mathcal{O}) - \frac{\log(n)}{2}(k + m + 2)$.

If willing to assume that the chosen degree and base density can be treated as predetermined, the usual parametric maximum likelihood theory applies, e.g. the use of the inverse of the Hessian matrix to obtain standard errors of the estimates (see Web Appendix A of [Zhang and Davidian, 2008] for a brief discussion on this subject).

5.2.2 The SNP representation for goodness-of-fit testing

The purpose of this subsection is to test the null hypothesis $H_0 : f_0(z) = \varphi(z)$, against the general alternative $H_1 : \exists k \in \{1, 2, \cdots\}$ such that $f_0(z) = P_k^2(z)\varphi(z)$. It is obvious how the SNP representation can be useful in testing an hypothesized distribution. The base density φ, i.e. the null model, is extended with polynomials of degree $k > 0$, and the null hypothesis is rejected if, given the data, an extended model provides a better approximation of the true density.

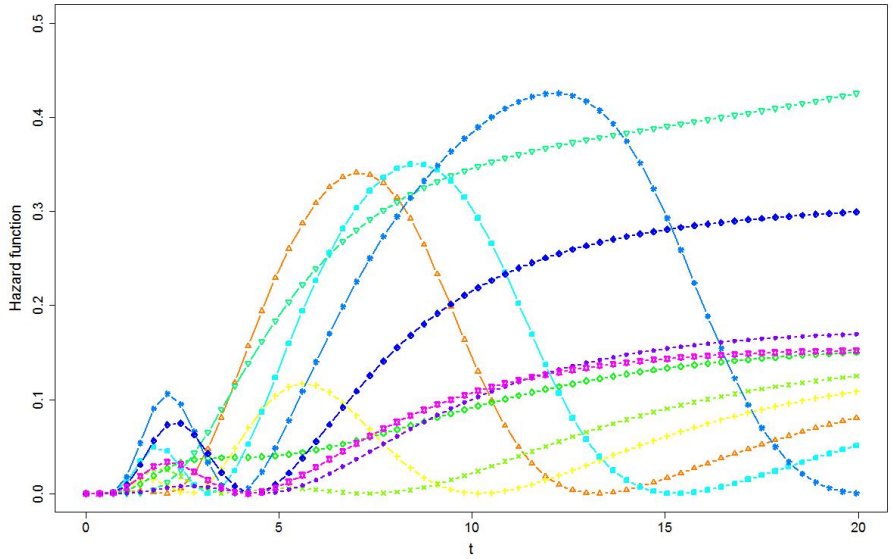


Figure 5.1 – Plot of several SNP hazard functions

The idea is to reject H_0 if an extended model (i.e. $k > 0$) is chosen over the null model (i.e. $k = 0$) by a data-driven model selection criterion [Aerts et al., 1999]. A test of level α against a general alternative is to reject H_0 if $MAIC(k, C_\alpha)$ is larger than 0 for some $k \in \{1, 2, \cdots\}$, where MAIC is the modified AIC defined as

$$MAIC(k, C_\alpha) = 2\left(\hat{l}_k(\hat{\boldsymbol{\eta}}; \mathcal{O}) - \hat{l}_0(\hat{\boldsymbol{\eta}}; \mathcal{O})\right) - C_\alpha k \quad \text{for } k > 0,$$

with C_α a constant larger than 1, controlling the level of the test. It can be shown that taking $C_\alpha = 4.18$ leads to a test of asymptotic level 5% (see [Hart, 1997], p.178 for choosing values of C_α leading to different test levels). In theory, k can be fixed or tend to infinity [Aerts et al., 1999]. In practice, only a finite number of models are needed, and simulations have shown that limiting k to $k_{max} = 7$ is generally sufficient for goodness-of-fit testing [Nysen et al., 2012].

An equivalent version of the test is to reject the null hypothesis if $Q \geq C_\alpha$, where

$$Q = \max_{1 \leq k \leq k_{max}} 2(\hat{l}_k - \hat{l}_0)/k.$$

It has been shown that, under some mild conditions, the asymptotic distribution of Q under H_0 is given by

$$P(Q \leq q) = \exp\left(-\sum_{s=1}^{\infty} \frac{P(\chi_s^2 > sq)}{s}\right),$$

where χ_s^2 is the chi-square distribution with s degrees of freedom.

Note that Q is not a single likelihood ratio test statistic, but rather the largest of a set of weighted log-likelihood ratio statistics [Aerts et al., 1999].

5.3 Simulation study

5.3.1 Simulation setting

Data are generated following the algorithm of Section 2.2 for interval-censoring, and Table 5.1 displays the parameters used to generate the data. There are three scenarios: in Scenario A, ε follows a Normal (i.e. T follows the log-Normal distribution), in scenario B, ε follows the extreme value distribution (i.e. T follows a Weibull distribution), and in scenario C, ε follows a mixture of two Normal distributions. The total percentage of cured and right-censored observations (including cured patients) is 10% and 20% (light right-censoring), 20% and 35% (medium right-censoring) or 30% and 55% (heavy right-censoring), approximately. All observations that are not right-censored are interval-censored. For each scenario (A, B and C) and each level of censoring (light, medium and heavy), we generated 500 datasets. The sample sizes considered are $n = 200$ and $n = 500$.

5.3.2 Simulation results

Estimation method

The first objective of this simulation study is to investigate the performance of the SNP representation for estimation purposes in a mixture cure model. We examine the appropriateness of the limit $k_{max} = 2$ and the AIC criterion on the basis of the relative bias and MSE of the estimates, for each level of cure and right-censoring. We apply the methodology described in subsection 5.2.1 to fit three models, i.e. $k \in \{0, 1, 2\}$, for each of the two basis densities, i.e. the Normal and Exponential density. We then choose the best model with the AIC criterion amongst these six fitted models. Tables 5.2 and 5.3 report the relative bias (i.e. $\frac{E(\hat{\beta}-\beta)}{\beta}$) and MSE for scenario A, B and C, for $n = 200$ and

$n = 500$, respectively. The estimated survival curves for scenario A, B and C (heavy censoring) are given in Figure 5.3 for the sample size $n = 500$.

We first observe that for all three scenarios, the relative bias and MSE are small for each parameter. As the right-censoring proportion increases, the MSE in latency increases; and as the cure proportion increases, the MSE in incidence decreases. We can also observe that the MSE in incidence is always larger than in the latency part, which may be explained by the low level of cured individuals (at most 30%), i.e. limited information in the logistic regression. The results when using the BIC criterion instead of the AIC are shown in Appendix D.1.1, and are very similar, except that the MSE is slightly smaller when using the BIC criterion.

The limit $k_{max} = 2$, leading to fit at most 3 models per density, does not seem too restrictive and provide a satisfactory estimation in our context. For the sake of comparison, Scenario A and B were fitted using the true density (i.e. log-Normal and Weibull respectively). Results for Scenario A are shown in Table 5.4, and in Appendix D.1.2 for scenario B. As expected, using the SNP representation leads to slightly larger bias and MSE, which remains nonetheless satisfying. In Scenario A for example, the largest observed difference in bias was -0.06 (using the SNP) versus -0.02 (using the Normal) for σ; and the largest difference in MSE was 0.36 (SNP) versus 0.11 (Normal) for β_0.

For another comparison, Scenario A was fitted supposing the event times follow the Weibull distribution and Scenario B the log-Normal distribution. We discuss the results for scenario A, given in Table 5.5 and refer to Appendix D.1.2 for Scenario B. In the latency, the relative bias using the Weibull are generally more than twice as large as with the SNP representation. In the incidence, we note much larger bias with the Weibull instead of the SNP for the light proportion of censoring, whereas the reverse trend is observable for medium and heavy censoring, to a lesser extend.

Regarding the choice of the degree k and the base density, Tables 5.6 ($n = 200$) and 5.7 ($n = 500$) show how often the AIC criterion selected each of the six fitted models. For scenario A, where the error term follows a Normal distribution, the most currently chosen model is the Normal base density associated with a polynomial of degree 0, which is expected given the data generation. The choice of higher order models (i.e. $k = 1$ or $k = 2$) or the other base density *increases* as the right-censored and cured proportion *increases*, but *decreases* as the sample size *increases* from $n = 200$ to $n = 500$.

Similar conclusions are reached for scenario B, except that the most currently chosen model is the Exponential base density associated with a polynomial of degree 0.

For scenario C, where the error term is generated via a mixture of two Normal densities, there is no combination that is systematically preferred, but the normal base density is nearly always chosen. The choice of higher order models *increases* as the sample size *increases*. Indeed, a larger sample size provides more information about the true density, hence the need for a more complex model. As the right-censored and cured proportion *increases*, the choice of higher order models *decreases*. This is explained by the fact that with heavy censoring, the mixture of normal densities become hardly distinguishable from a single normal density.

While our simulation settings were different, our conclusions are consistent with the findings previously obtained in an AFT model without cure and with relatively light censoring [Zhang and Davidian, 2008]. The SNP representation provides a satisfactory estimation in terms of bias and MSE, in our setting of mixture cure models with heavy censoring. In some cases (e.g. a large proportion right-censored individuals or a larger sample size), there is a need for a higher degree of the polynomial. However, allowing the choice of either the normal or exponential base density seems to prevent the need for polynomial of degree higher than two and thus, reduces the computational burden.

Goodness-of-fit testing

The second objective of the simulation study is to investigate the performance of the SNP representation for testing an assumed survival distribution, in the AFT mixture cure model with interval-censored event times and covariates. We present here in details the results obtained when we test the null hypothesis that the event times for uncured patients are log-normally distributed, that is, the base density is the Normal density. The results when testing if the event times followed a Weibull or an Exponential distribution are displayed in Appendix D.2.

We apply the methodology described in subsection 5.2.2. We fit at most $k_{max} = 7$ models with the Normal base density. The limiting level is 5%, with $C_\alpha = 4.18$. A plot of the error-term density for Scenario B and C is shown in Figure 5.2, with the Normal base density curve superimposed for comparison purposes. In all cases, we report the proportion of times the null hypotheses are rejected. This is the type-I-error for scenario A, and the observed power of the test for scenario B and C. Results for each scenario are given in Table 5.8.

For scenario A, the type-I-error is very close to the 5%-level. In the case of heavy censoring and small sample size, the percentage of rejection is 8%, slightly more than the expected 5%, but nearly reaches the 5%-level when the sample size increases to $n = 500$.

Concerning scenario B, the power of the test decreases as the right-censored and cured proportion increases, but is nearly 100% when censoring is light and the sample size is 500. Figure 5.2 shows the closeness of the density of the error term (i.e. the extreme value density) to the Normal density, explaining the need of a larger sample size to reject the null hypothesis.

Scenario C is based on a mixture of two Normal densities, hence the low percentage of rejection of the null hypothesis, which decreases for higher level of censoring. This behavior is expected since the departure from a single Normal density becomes smaller as the right-censoring proportion increases. We nevertheless observe that the increase of sample size from 200 to 500 improves the power of the test.

In Appendix D.2, we provide the results of further simulations with the same data generation process. We tested if the event times follow a Weibull distribution, or an Exponential distribution. The conclusions are similar: if the censoring level increases, the power of the test decreases and a larger sample size may be needed .

5.4 Real data analysis

The Alzheimer's disease database was previously analyzed with the EGG-AFT mixture cure model. Firstly, in Chapter 3, we were interested in the effect of some baseline cognitive scores on the onset of MCI, and we selected some variables with the adaptive LASSO, see results on Table 3.10. Secondly, in Chapter 4, we were interested in knowing the impact of the aggregate cognitive scores (CAMCOG) and other covariates on the onset of MCI. The results concerning this model are given in Table 4.2. In this section, we will use the SNP representation to analyze the two models corresponding to Tables 3.10 and 4.2 (i.e. the models with the same covariates).

The Weibull distribution is frequently used in survival analysis and in studies about Alzheimer's disease [Carroll, 2003, Khachaturian et al., 2004, Dysken et al., 2014]. Moreover, our analysis of the Cox-Snell residuals suggested that the event times could come from a Weibull distribution, see Section 4.4. To verify this, we test, for both models, if the event times follow a Weibull distribution by using the goodness-of-fit procedure described in Section 5.2.2. In both cases, the null hypothesis is rejected at the 5%-level, and we thus apply the SNP representation to obtain the parameter estimates.

For the first model, corresponding to Table 3.10, the combination preferred by the AIC criterion is the Normal base density with polynomial of degree 2. Table 5.9 contains the parameter estimates using the SNP representation. The results are similar to those obtained with the EGG-AFT method. The

major differences are the following: In the latency, the value of the intercept is larger when using the EGG instead of the SNP (2.628 vs 0.662); and the sign of the coefficient for "Education" is positive with the EGG, but negative with the SNP (0.152 vs -0.040). This covariate was however not significant with the EGG and the SNP. In the incidence, the coefficient for "Education" is much larger when using the EGG instead of the SNP (2.285 vs 0.128) but was not significant with the EGG and the SNP. It is worth noting that the estimates from Chapter 3 were obtained while performing a variable selection, while no variable selection was undertaken for the SNP representation in which we simply used the same variables as those selected in Chapter 3. A graphical comparison of the results is given in Figure 5.4 for the presence and absence of APOE E4 gene. There is an obvious difference between the EGG and SNP distribution. One difference is that the uncured EGG survival distribution levels off quite abruptly after a certain number of years: 10 years for patients with APOEE4 gene and 15 years for the others. However, the estimated cured proportion remains the same using the EGG or SNP method.

For the second model, corresponding to Table 4.2, the same combination was preferred by the AIC, i.e. the Normal base density with polynomial of degree 2. The parameter estimates are given in Table 5.9. In this case also, the results are similar to those previously obtained with the EGG and given in Table 4.2. The major difference here is that in the incidence using the EGG, no covariates were significant, whereas the MMSE significantly reduces to risk to convert to MCI when using the SNP representation. As the SNP representation is more general than the EGG, we will conclude that the MMSE is significant in the incidence.

5.5 Discussion

In this chapter, we investigated the SNP representation as an estimation method and goodness-of-fit procedure in the setting of mixture cure models with interval-censored data. As an estimation method, we showed that the SNP representation leads to highly satisfactory results. Lying halfway between parametric and non-parametric methods, the SNP representation is more general than many existing flexible distributions like the EGG or Generalized F distributions. It can therefore embrace more densities.

In addition to its flexibility, one advantage of the method is the simplicity of the idea behind it: a well-known base density, i.e. the normal or exponential, is extended by a polynomial, whose degree is chosen by a popular model selection criterion, such as the AIC. By imposing a limit on the degree of the polynomial, the computational burden remains acceptable. Furthermore, although

the model is chosen adaptively, experience has shown that it can be treated as a predefined parametric model, so that standard inference techniques apply [Coppejans and Gallant, 2002, Kim, 2007, Zhang and Davidian, 2008].

The chosen model informs about the departure from the base density, as a higher polynomial degree implies a higher disparity between the true density and the base density. For example, the SNP representation provides an easy way to test for (log-)normality when data are left-, right- or interval-censored while usual normality tests, such as Kolmogorov-Smirnov test are known to perform poorly in the presence of censoring [Nysen et al., 2012].

The SNP representation allows to test not only the normal distribution, but other distributions as well. Changing the basis density allows to test almost every distribution, but this may however imply to compute several integrals if a recursion formula is not available, as for the normal and exponential densities. This drawback can actually easily be circumvented with another approach: testing if the censored Cox-Snell residuals (see Chapter 4) follow an exponential distribution using the SNP representation. We have indeed shown that if the assumed model is correct, these residuals will behave like a censored sample from an exponential distribution.

Our simulations are consistent with previous findings and indicated that the rate of convergence of the test may be slow. A bootstrap procedure can be recommended for small sample sizes [Nysen et al., 2012, Geerdens et al., 2013]. However, replicating interval-censored data with a cure fraction for the purpose of goodness-of-fit testing is not a straightforward issue. We believe that studying this issue and providing a bootstrap approach would improve the performance of the test.

Lastly, the proposed goodness-of-fit test has a nice interpretation in term of likelihood ratios as it is in fact based on a set of weighted log-likelihood ratio tests. Unlike a single likelihood ratio test, it is thus consistent against any alternatives within the SNP family. In conclusion, the SNP representation provides an elegant method to analyze a mixture cure model with interval-censored data.

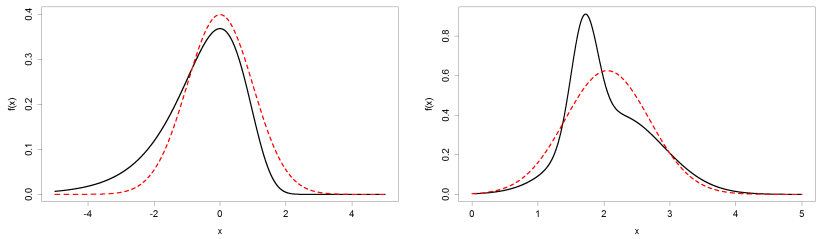


Figure 5.2 – Plot of the Weibull (left) and mixture of Normal (right) along with a standard Normal density (dotted line).

Table 5.1 – Values of parameters and distribution of variables used for data generation

RC level		Scenario A	Scenario B	Scenario C
	β_0	2	-1	see ε
	β_1	-2	2	-2
	β_2	0.5	0.5	0.5
	β_3	-0.5	-0.5	-0.5
	σ	0.75	0.14	see ε
	X_1	$Bern$	$Bern$	$Bern$
	X_2	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
	X_3	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
	ε	$\mathcal{N}(0,1)$	$exp(-exp(x))$	$0.3\mathcal{N}(0,0.7)+$ $0.7\mathcal{N}(0,0.7)$
Light	γ_0	3.2	3.2	3.2
Medium	γ_0	2.5	2.5	2.5
Heavy	γ_0	2	2	2
	γ_1	-1	-1	-1
	γ_2	-1	-1	-1
	Z_1	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$
	Z_2	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$
Light	K	50	16	50
Medium	K	28	11	30
Heavy	K	14	6	20
	a	0.5	0.5	0.5

Table 5.2 – Relative bias and MSE for scenario A, B and C; for sample size $n = 200$, using the SNP representation.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	-0.03	0.36	0.00	0.36	0.01	0.35
β_1	0.00	0.13	0.00	0.15	-0.01	0.19
β_2	0.00	0.06	0.00	0.07	-0.01	0.09
β_3	-0.01	0.06	-0.01	0.07	-0.02	0.09
σ	-0.01	0.15	-0.02	0.15	-0.06	0.15
	Incidence Part: Logistic regression					
γ_0	0.05	1.07	0.05	0.84	0.05	0.90
γ_1	0.00	1.18	-0.01	0.90	0.02	0.99
γ_2	0.08	1.18	0.11	0.97	0.03	1.19
Scenario B						
	Latency Part: AFT Model					
β_0	-0.02	0.04	-0.03	0.04	-0.03	0.04
β_1	0.00	0.01	0.00	0.02	0.00	0.03
β_2	0.00	0.01	0.00	0.01	0.00	0.02
β_3	0.00	0.01	0.00	0.01	0.00	0.02
σ	-0.03	0.01	-0.06	0.01	-0.14	0.01
	Incidence Part: Logistic regression					
γ_0	0.05	0.85	0.05	0.68	0.04	0.69
γ_1	0.09	0.94	0.09	0.79	0.08	0.77
γ_2	0.03	0.92	0.05	0.80	0.04	0.86
Scenario C						
	Latency Part: AFT Model					
β_0	-0.11	0.51	-0.07	0.44	-0.06	0.39
β_1	-0.01	0.10	-0.01	0.11	0.00	0.13
β_2	0.00	0.05	0.00	0.06	0.00	0.07
β_3	-0.01	0.05	0.00	0.06	0.01	0.07
σ	-0.04	0.13	-0.07	0.12	-0.09	0.13
	Incidence Part: Logistic regression					
γ_0	0.05	1.06	0.03	0.77	0.01	0.65
γ_1	0.08	1.16	0.03	0.89	0.01	0.80
γ_2	0.01	1.12	0.03	0.86	0.02	0.75

Table 5.3 – Relative bias and MSE for scenario A, B and C; for sample size $n = 500$, using the SNP representation.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	-0.03	0.33	-0.05	0.39	-0.02	0.42
β_1	0.00	0.08	0.00	0.09	0.00	0.12
β_2	-0.01	0.04	-0.01	0.04	-0.01	0.06
β_3	0.00	0.04	0.00	0.04	0.00	0.06
σ	0.02	0.10	0.03	0.13	0.02	0.15
	Incidence Part: Logistic regression					
γ_0	0.04	0.67	0.03	0.62	0.10	0.96
γ_1	0.03	0.78	0.01	0.62	0.09	0.74
γ_2	0.09	0.67	0.06	0.60	0.15	0.80
Scenario B						
	Latency Part: AFT Model					
β_0	-0.02	0.04	-0.02	0.04	-0.03	0.04
β_1	0.00	0.01	0.00	0.01	0.00	0.02
β_2	0.00	0.00	0.00	0.01	0.00	0.01
β_3	0.00	0.00	0.00	0.01	0.00	0.01
σ	0.05	0.01	0.04	0.01	-0.02	0.01
	Incidence Part: Logistic regression					
γ_0	0.02	0.48	0.01	0.40	0.01	0.41
γ_1	0.03	0.54	0.00	0.48	-0.02	0.52
γ_2	0.03	0.57	0.01	0.47	0.00	0.50
Scenario C						
	Latency Part: AFT Model					
β_0	-0.22	0.69	-0.18	0.64	-0.16	0.57
β_1	-0.01	0.06	-0.01	0.07	0.00	0.08
β_2	0.00	0.03	0.00	0.04	0.00	0.04
β_3	0.00	0.03	0.00	0.04	0.00	0.04
σ	0.02	0.12	-0.01	0.12	-0.02	0.11
	Incidence Part: Logistic regression					
γ_0	-0.01	0.54	0.00	0.46	-0.01	0.40
γ_1	0.00	0.61	0.01	0.50	-0.02	0.46
γ_2	-0.04	0.64	-0.01	0.52	-0.02	0.47

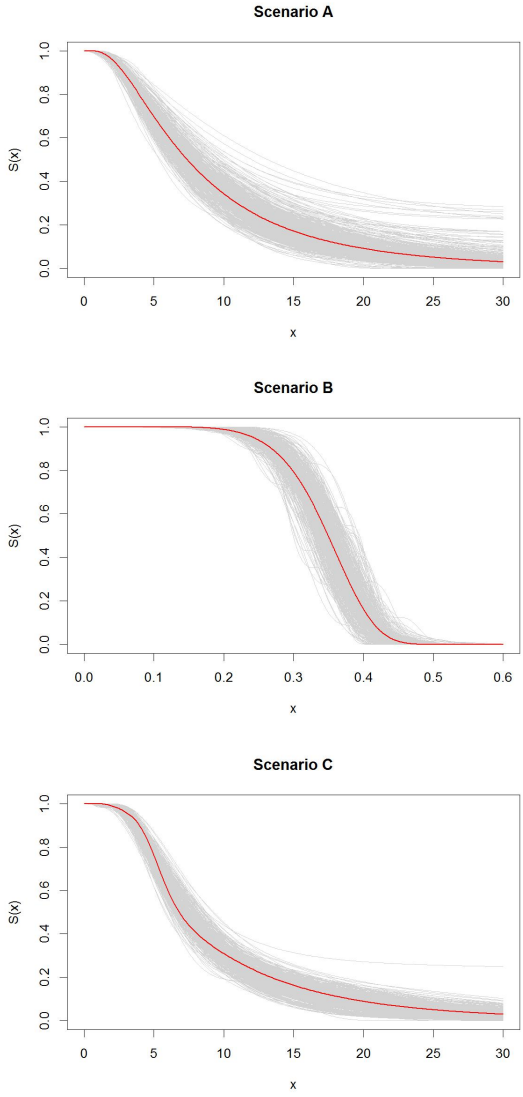


Figure 5.3 – Estimated survival curves for scenarios A, B and C using the SNP estimation with the best AIC. Sample size $n = 500$.

Table 5.4 – Relative bias and MSE for scenario A, for sample size $n = 200$, using the true model for estimation purposes.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	0.00	0.11	0.00	0.13	0.01	0.17
β_1	0.00	0.13	0.00	0.15	0.00	0.19
β_2	0.00	0.06	0.00	0.07	0.00	0.08
β_3	-0.01	0.06	-0.01	0.07	-0.01	0.09
σ	-0.01	0.05	-0.01	0.05	-0.02	0.07
	Incidence Part: Logistic regression					
γ_0	0.06	1.08	0.04	0.81	0.03	0.76
γ_1	0.01	1.19	-0.01	0.89	0.00	0.87
γ_2	0.08	1.17	0.09	0.92	0.03	0.89

Table 5.5 – Relative bias and MSE for scenario A, for sample size $n = 200$, using the Weibull distribution for estimation purposes.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	0.21	0.45	0.15	0.24	0.15	0.30
β_1	-0.12	0.30	-0.04	0.17	-0.02	0.21
β_2	-0.09	0.12	-0.03	0.08	-0.04	0.09
β_3	-0.08	0.11	-0.03	0.08	-0.04	0.10
σ	0.15	0.27	-0.12	0.07	-0.17	0.08
	Incidence Part: Logistic regression					
γ_0	1.13	18.13	0.01	0.78	-0.02	0.71
γ_1	-2.29	8.48	-0.05	0.87	-0.04	0.83
γ_2	-2.19	8.08	0.05	0.88	-0.01	0.85

Table 5.6 – Choice of degree and basis density φ for scenario A, B and C; for each level of cure and right-censoring (RC). The sample size is $n = 200$.

Cure	10%		20%		30%	
RC	20%		35%		55%	
φ	Norm	Exp	Norm	Exp	Norm	Exp
	Scenario A					
$k = 0$	75%	0.4%	75.6%	1.8%	68.6%	5.4%
$k = 1$	7.2%	3.8%	4.2%	5%	6.8%	3.4%
$k = 2$	5.8%	7.8%	6.8%	6.6%	7.2%	8.6%
	Scenario B					
$k = 0$	3%	56.2%	4.4%	46.8%	13.4%	32%
$k = 1$	13%	12.2%	14.6%	8.8%	12.8%	13.4%
$k = 2$	5.8%	9.8%	11.2%	14.2%	14.6%	13.8%
	Scenario C					
$k = 0$	27.4%	0%	37.4%	0.2%	44%	0.2%
$k = 1$	40.6%	0.2%	32.6%	0.2%	26.2%	0.8%
$k = 2$	29.2%	2.6%	24.6%	5%	20.8%	8%

Table 5.7 – Choice of degree and basis density φ for scenario A, B and C; and each level of cure and right-censoring (RC). The sample size is $n = 500$

Cure	10%		20%		30%	
RC	20%		35%		55%	
φ	Norm	Exp	Norm	Exp	Norm	Exp
	Scenario A					
$k = 0$	80.2%	0%	79.8%	0%	74.2%	0%
$k = 1$	12.4%	1.4%	10.8%	3.2%	11.4%	3.8%
$k = 2$	5%	1%	4.8%	1.4%	7%	3.6%
	Scenario B					
$k = 0$	0.2%	68%	0.8%	64.6%	6.2%	49.4%
$k = 1$	5.8%	16.4%	11.4%	10.2%	13.2%	8.4%
$k = 2$	3%	6.6%	4.8%	8.2%	9.2%	13.6%
	Scenario C					
$k = 0$	4.2%	0%	9.8%	0%	20.6%	0%
$k = 1$	55.8%	0%	47.8%	0%	46.6%	0%
$k = 2$	40%	0%	42%	0.4%	31.8%	1%

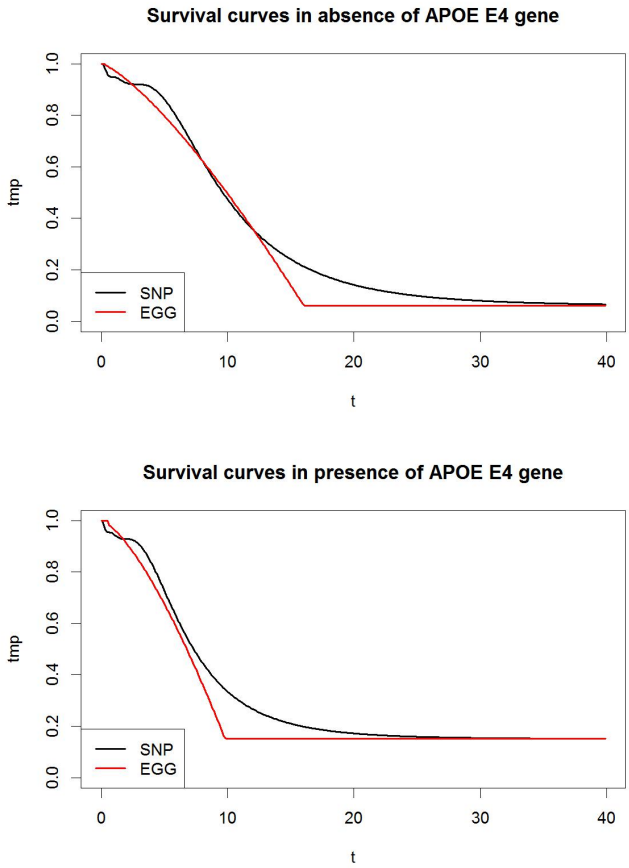


Figure 5.4 – Survival curves estimates with best SNP or EGG distribution for the Alzheimer's disease database for patients with or without the presence of the APOE E4 gene.

Table 5.8 – Goodness-of-fit: Testing the normality of the error term, for each scenario and each level of cure and right-censoring (RC). Percentage of rejection over 500 replications with the SNP representation, for sample sizes $n = 200$ and $n = 500$.

Cure	10%	20%	30%
RC	20%	35%	55%
$n = 200$			
Scenario A	4.00%	4.00%	8.00%
Scenario B	64.80%	54.60%	46.80%
Scenario C	60.20%	50.40%	38.60%
$n = 500$			
Scenario A	5.0%	4.20%	4.80%
Scenario B	98.60%	89.40%	68.60%
Scenario C	97.00%	93.80%	87.20%

Table 5.9 – Parameter estimates, standard deviation (SD), P-value (Wald test) and the exponential of the estimates, for the Alzheimer's disease database, with aggregates CAMCOG score and other covariates, using the SNP representation.

Parameter	Estimates	SD	P-value	Exp (Estimates)
Latency part: EGG-AFT model				
Intercept Latency	0.662	0.055	0.000	
Expression	0.220	0.049	0.000	1.246
Attention	-0.076	0.040	0.068	0.927
Abstract Thinking	0.074	0.049	0.129	1.077
Perception	0.092	0.049	0.071	1.096
APOE E4	-0.591	0.130	0.000	0.554
Gender	-0.095	0.090	0.230	0.909
Age	-0.211	0.054	0.000	0.810
Total Education	-0.040	0.054	0.301	0.961
σ	1.042	0.106	0.000	
Incidence part: Logistic regression				
Intercept Incidence	2.878	1.037	0.008	
MMSE	-1.855	0.820	0.031	0.156
Learning	-0.901	0.386	0.026	0.406
Praxis	-1.469	0.696	0.043	0.230
Abstract Thinking	0.572	0.447	0.176	1.772
APOE E4	-2.475	1.053	0.025	0.084
Total Education	0.128	0.584	0.390	1.137

Table 5.10 – Parameter estimates, standard deviation (SD), P-value (Wald test) and exponential of the estimates for the Alzheimer's disease database using the SNP representation.

Parameter	Estimates	SD	P-value	Exp (estimates)
Latency part: EGG-AFT model				
Intercept Latency	0.872	0.106	0.000	
Age	-0.256	0.055	0.000	0.774
CAMCOG	0.220	0.049	0.001	1.246
tHcy	0.049	0.040	0.311	1.050
Folate	-0.110	0.049	0.043	0.896
MMSE	-0.071	0.049	0.185	0.931
Gender	-0.079	0.130	0.313	0.924
APOEE4	-0.275	0.090	0.031	0.760
σ	0.725	0.054	0.000	
Incidence part: Logistic regression				
Intercept Incidence	2.743	0.054	0.063	
Age	0.189	1.037	0.382	1.208
CAMCOG	-1.699	0.820	0.134	0.183
tHcy	0.325	0.386	0.359	1.384
Folate	-0.311	0.696	0.324	0.733
MMSE	-2.454	0.447	0.046	0.086
Gender	1.417	1.053	0.261	4.125
APOEE4	-1.012	0.584	0.270	0.363

CHAPTER 6

Discussion

"All our dreams can come true, if we have the courage to pursue them."

- Walt Disney (1901-1966)

6.1 Conclusion

In survival analysis, we study the time until an event of interest occurs. In many cases, an individual is followed at scheduled visits, so that his/her time of occurrence is interval-censored. For a wide range of reasons, the event may not be observed at all before the end of the follow-up, in which case the corresponding observation is right-censored. In this case, while usual survival analysis methods assume that the event will eventually occur later in time, it may also happen that the event will in fact never occur, and the individual is said to be "cured" or "immune".

Failing to correctly take this cure fraction into account may produce inaccurate conclusions, as seen in Chapter 2. We have also seen that it is nevertheless important to use a cure model only when there is empirical evidence of the existence of cured individuals. We have also seen in Chapter 2 that using naive methods to handle the presence of interval-censored observations may lead to biased estimation. This thesis thus aimed at providing methods to analyze interval-censored data in which a fraction of the population is cured.

Two main models exist in the literature to consider a cure fraction: the mixture cure model and the promotion time cure model [Yakovlev et al., 1996, Tsodikov et al., 2003]. The latter was developed to maintain the assumption of proportional hazards, and is based on a very specific biological interpretation. It assumes an upper bound for the cumulative hazard function, and is hence also called the bounded cumulative hazard model. The probability to be cured is in fact a part of the survival distribution of the entire population. As a consequence, an interpretation of the coefficients does not allow to know the impact of a covariate on the probability to be cured, or on the time-to-event separately. On the contrary, such an interpretation is possible with the mixture cure model. This model is very intuitive, as it considers two sub-populations: the cured and the uncured. For these reasons, we decided to account for a cure fraction by using the mixture cure model.

In the incidence part of the model, a logistic regression is often assumed to model the probability to experience the event. Concerning the latency part, i.e. modelling the time-to-event of uncured individuals, two popular options are the PH and the AFT regression models. The PH model is widely used in classical survival analysis, provided that the assumption of proportionality of hazards is met. Typically, this translates into survival curves that do not cross with each other. In the presence of a cure fraction, even if the survival distribution of the uncured truly comes from a PH model, the global survival curves can cross with each other [Sy and Taylor, 2000]. To our knowledge, there is no method that distinguishes crossing hazards that are due to the presence of cure from crossing hazards that are due to a true non proportionality in the

latency. An AFT model can thus be an excellent alternative, as it circumvents this issue. Moreover, the AFT model provides a straightforward interpretation of the results: a covariate either accelerates or decelerates the time until the event.

If a parametric form is assumed for the distribution of the error term, the likelihood function for a parametric AFT model with possibly interval-censored observations is easily formed and its maximization is straightforward. Otherwise, a semi-parametric AFT model is quite challenging, with or without a cure fraction. One of the main reasons is that the error term distribution $S_{u,0}$ can not be easily "removed" or treated like a nuisance function. Indeed, in the AFT model, the parameters β_j and the survival distribution $S_{u,0}$ are in fact *bundled* together and can thus not be separated [Huang and Wellner, 1997]. A partial likelihood approach, typically used for the PH model, is thus not available.

Semi-parametric AFT models are generally analyzed using estimating equations based on linear-rank statistics [Kalbfleisch and Prentice, 2002]. Their extension to interval-censoring have been studied by [Rabinowitz et al., 1995, Betensky, 2001]. However, the methods are generally limited to a low number of covariates [Lesaffre et al., 2005]. As a solution to this issue, a Markov Chain Monte Carlo approach was proposed [Tian and Cai, 2006]. Other promising methods are flexible approaches like the semi-nonparametric representation discussed in this thesis or smoothing techniques [Komárek et al., 2005].

It is nonetheless unquestionable that a parametric modelling is very appealing in this context and a flexible distribution can be an excellent compromise. Unfortunately, parametric modelling has not received much attention in the past, and some highly attractive features have not been adapted to the parametric mixture cure model. Consequently, Chapters 3 and 4 aimed at developing useful methods in this context.

In model building, we typically have to face the issue of selecting a relatively low number of covariates that have a strong impact on the outcome. Sparsity allows an easier interpretation of the results. Besides, it may lead to more cost effective procedures in the future because the removed covariates will not need to be measured to fit these models. In a mixture cure model setting, variable selection is made more challenging by the fact that covariates in the latency part may be different from those in the incidence part. This actually leads to a large number of possible subsets to be selected.

Discrete methods entering or deleting a covariate from a model, such as the forward/backward stepwise selection procedure, are widely used and very popular, but have serious shortcomings. Amongst them, they are known to be computationally intensive if the number of covariate is large [Harrell, 2001]. Shrinkage methods, on the other hand, are superseding the popular best subset

and stepwise selection procedures. They are continuous and easy to implement methods. The adaptive LASSO belongs to this class of methods [Zou, 2006], and has already been studied in survival analysis for the Cox regression [Zhang and Lu, 2007].

Chapter 3 investigated the extension of the adaptive LASSO in our parametric AFT mixture cure model. The extension was made possible by using a least-square approximation to our particular likelihood [Wang and Leng, 2007]. Our simulations showed that results were satisfying with a large number of covariates in each part of the model. One must however pay attention to the fact that estimates are *shrunk* towards zero, therefore implying a possible bias. This bias reported in our simulation remains acceptable.

In regression models, it is commonly assumed that covariates are entered linearly in the model. This assumption may sometimes be too restrictive and a transformation of some covariates may be necessary. One approach is to add many possible transformations of a covariate in the model, and then detect which one is the best by using a selection procedure. However, a more convenient and direct approach is to acquire a visual representation of the transformation needed, and this can be achieved by plotting the residuals of the model.

Residuals checking is a very useful practical tool, allowing to explore the structure of the data. In classical survival analysis, martingale and deviance residuals are extensively used to detect non linearity in a model. The idea is to plot the residuals versus the values of the covariate. A pattern suggests that a transformation may be necessary. In a mixture cure model with interval-censoring, a straightforward extension of martingale residuals is not possible. Firstly, the martingale theory is not available with interval-censored data because it is based on counting processes. Secondly, a covariate that is present in both parts of the model can have a different impact in each part. The residuals should therefore allow to detect a needed transformation in the latency and in the incidence separately.

A first objective of Chapter 4 was to study and develop such residuals. This objective was achieved by using the complete data likelihood to define the latency, the incidence, and the global deviance residuals. Indeed, the complete-data likelihood naturally induces a separation between latency and incidence. It therefore allows to define deviance residuals based on each part of the model. We showed that the latency deviance residuals are useful to detect a non-linear trend in the latency. As heavy right-censoring may hide a pattern in the plots, the detection can further be enhanced by plotting right- and interval-censored observations with a different symbol or color [Collett, 2003]. The incidence deviance residuals correctly detect a trend in the incidence, as long as the

fraction of cured observations provides enough information, i.e. as long as there is enough cured individuals.

These deviance residuals were developed in a parametric context, making therefore assumptions about the survival distribution. Those assumptions also need to be checked, and the Cox-Snell residuals can be used in this objective. In a mixture cure model however, the probability to observe infinite event times is strictly positive. The survival distribution of the entire population is thus improper. We thus first had to study the properties of the Cox-Snell residuals based on the improper *global* survival, which we called the global Cox-Snell residuals. We proved that they can be used as in classical survival analysis. The points of the estimated nonparametric survival estimate versus these residuals should be aligned on a line of unit slope and zero intercept.

In a parametric mixture cure model, assumptions are made about the form of the survival distribution for the uncured individuals. We thus defined the uncured Cox-Snell residuals aiming at checking the *uncured* survival distribution. Only uncured observations follow this distribution, so that those residuals should only be applied on uncured individuals. However, the status of a right-censored individual is not known and we proposed a method to identify this status. Developing and studying these Cox-Snell residuals was the second objective of Chapter 4.

A visible pattern in these residuals plots should raise questions about the fit of the model. However, the absence of a pattern does not necessarily mean that the fitted model is perfect [Collett, 2003]. Diagnostic tools should give a clue about "where" and "to what extent" the model is wrong. But it is the knowledge of the subject that should refine, confirm or infirm these findings.

Residuals checking gives a general idea about the fit of a model, and can also help to find a transformation. A plot of the global or uncured Cox-Snell residuals may however seem too subjective, and it may be the case that a formal goodness-of-fit test is needed. In the presence of censoring, some popular tests such as the Kolmogorov-Smirnov test perform quite poorly [Nysen et al., 2012]. In our case, goodness-of-fit tests like log-likelihood ratio tests are better adapted as they account for both censoring and cure. However, they are usually based on an alternative hypothesis that may be too restrictive. The SNP representation presented in Chapter 5 allows a goodness-of-fit test that possesses the good sides of a log-likelihood ratio test and that is consistent against a wide range of alternatives (within the SNP family), as the procedure is based on a set of weighted log-likelihood ratios [Aerts et al., 1999].

The idea of the SNP representation is to extend a base density like the Normal density by a polynomial of degree k. The degree of the polynomial informs us about the departure from the base density, implying therefore a goodness-of-

fit test. This formulation allows to test any survival density by appropriately choosing the base density. This may however imply that some complicated integrals have to be calculated. Indeed, we have seen that with the Normal or Exponential density, only one integral calculation is required to estimate a density. This may not necessarily be the case with other densities. One approach to circumvent the issue is to test if the Cox-Snell residuals discussed in Chapter 4 indeed follow an Exponential distribution. Our simulation results confirm the results of other authors, i.e. satisfying results for the type-I-errors of the test, but the need for a bootstrap procedure to reach a larger power for small sample sizes [Nysen et al., 2012, Geerdens et al., 2013]. We have indeed seen in some cases a considerable improvement of the power when using a larger sample size.

Chapter 5 provided a method to test if an assumed parametric distribution is reasonable. In some cases however, we do not want to use a parametric distribution, or usual parametric distributions do not fit the data correctly enough. The SNP representation then offers an alternative to parametric methods. It allows to use the usual methods of maximum likelihood without imposing any other assumption than sufficient smoothness on the survival distribution. Chapter 5 therefore also developed the use of the SNP representation for estimation purposes. This method is effective, easy to understand, and can approximate nearly every suitable density in survival analysis. Our simulation results for estimation purposes are similar to previous findings where the SNP representation is applied to a PH or AFT model with interval-censored observations but without a cure fraction and with light right-censoring only[Zhang and Davidian, 2001, Zhang and Davidian, 2008].

In Chapter 3, the EGG distribution was used in the latency part of the mixture cure model. While this distribution is very flexible and encompasses both the log-Normal and Weibull distribution that are widely used in survival analysis, the SNP representation can be seen as an even more flexible distribution. It encompasses the log-Normal and the Weibull as well, but it also encompasses the EGG distribution, if the EGG is used as base density (with the drawback that one would have to compute integrals without closed form for the survival distribution).

It is worth mentioning on a final note that a method competing with the SNP could be the smoothing technique based on Gaussian densities [Komárek et al., 2005]. A comparison with the SNP was performed by means of simulations in a non-cured model [Doehler and Davidian, 2008]. No clear-cut conclusions can be drawn: both approaches have their advantages depending on the situation. It might thus be interesting to study the extension of this method to a mixture cure model.

In conclusion, this thesis proposed new methods to deal with interval-censored data and the presence of a cure fraction. The variable selection procedure used with the EGG distribution in Chapter 3 and the SNP representation of Chapter 5 may be applied in a context of right-censored data with or without a cure fraction. In this context, semi-parametric methods have been developed and may be preferred because they do not impose any assumption on the distribution of uncured observations. Parametric assumptions allow to avoid having to rely on the EM algorithm for estimation. This is not a negligible advantage as this algorithm tends to be slow and that the standard errors are not directly available, as is the case with parametric methods. The diagnostic check procedure developed in Chapter 4 is a newly proposed method applicable to a PH or AFT mixture cure model with right-censored data and/or the presence of interval-censoring. While the definition of residuals in an interval-censored PH model has been studied, no such study exist for the AFT model. Recently, a diagnostic check methodology was developed for mixture cure models, but only for the Cox mixture cure model with right-censoring [Peng and Taylor, 2016], while we cover the case of an interval-censored AFT model and a PH or AFT mixture cure model with or without interval-censoring.

The methods presented in this thesis allow to fit an AFT mixture cure model with or without imposing parametric assumptions on the survival distribution of the uncured. They allow to select important variables to be included in the model and also to refine the model by checking the assumptions made about it.

6.2 Conclusion on Alzheimer's disease database

The methodologies presented in this thesis were motivated by the need to analyze a real database concerning Alzheimer's disease. We were able to detect which cognitive scores had an impact on the probability to convert to MCI. These scores were not the same as the scores impacting the event times for uncured. The scores Learning and Praxis, as well as the covariates "APOE E4" and MMSE had a positive impact on the cure probability, as a larger value decreases the risk to convert to MCI. The scores for Abstract Thinking as well as the number of years of Education had an opposite effect.

Regarding the latency part, the scores for Expression, Abstract Thinking and Perception decelerates the time to conversion: individuals with a larger score will convert later in time. On the contrary, the older an individual is, the sooner he is expected the convert to MCI. A larger value for Attention or the presence of the APOE E4 gene has the same effect.

In Chapter 4, the interest was on the aggregate score to the cognitive tests, i.e. the CAMCOG score. The model was adjusted for other covariates: Age, tHcy, Folate, MMSE, Gender and the presence/absence of APOEE4 gender. Our diagnostic tools allowed us to be confident about the proposed model and results.

As both previous studies and the Cox-Snell residuals plot suggested that the Weibull distribution could be appropriate, we applied the goodness-of-fit procedure of Chapter 5. The test rejected the assumption of Weibull distributed event times. The estimated parameters obtained with the EGG distribution for the error term in Chapter 3 and 4 were then compared to the estimates obtained using the SNP representation. The major differences obtained were essentially values that were larger when using one method compared to the other, and conclusion discussed herein above remains similar.

6.3 Further work

While this thesis provided methods to analyze a mixture cure model with interval-censored data, some issues are still left open.

Firstly, throughout the thesis, we used the logistic regression in the incidence part. This modelling is not fault free. For instance, some of the parameters may quickly tend to infinity in configurations where the cure fraction is in fact present, but very small. Also, in the presence of high-dimensional covariates, the sample size needs to be sufficiently large to obtain accurate estimates. Parametric alternatives include the probit regression, which is in fact extremely similar to the logistic regression, or the complementary log-log, which may be more useful in the case of rare events. A non parametric alternative has recently been used in the context of mixture cure modelling [Wang et al., 2012], but for identifiability issues, they do not impose any parametric assumption on the baseline survival of the PH regression in the latency. While extensive work is still needed on that matter, we believe that using a logistic regression in the incidence, together with a diagnostic check procedure to identify possibly non-linear covariates, provides a satisfying model.

A second research topic would be a deeper study of the diagnostic checks. When using the diagnostic plots, the subjectivity of the method is often criticized. Concerning the Cox-Snell residuals, a solution to this issue would be to use the Cramer-von Mises criterion [Anderson, 1962, Peng and Taylor, 2016]. This criterion is used to quantify the distance between the assumed distribution and the exponential distribution. A small distance indicates a correct fit. By comparing the distance given by different assumed distributions, e.g. the Normal and the Weibull, one can also detect which distribution is a better fit

for the data. Regarding the deviance residuals, the ideas used in the definition of the cumulative martingale residuals present interesting features [Lin et al., 1993, Lin et al., 2002]. Cumulative sums of martingale residuals are defined as a partial-sum processes of martingale residuals, i.e. $\sum_{i=1}^{n} 1(X_{ij} < x) r_M(t_i)$, where r_M represents the martingale residual, and X_{ij} is the value of the j-th covariate for observation i. The authors use the fact that under the null hypothesis of no misspecification, the distribution of these processes can be approximated by simulating certain Gaussian processes. This will create an envelope around the martingale residuals. It would then be interesting to study how to extend this kind of envelope to our deviance residuals.

Another possible application of diagnostic checks is to check the proportionality of hazards. For this, one method is to use the Schoenfeld residuals, which are defined in terms of the derivatives of the partial likelihood. Keeping in mind that, in a semi-parametric mixture cure model, the complete data log-likelihood can be separated into a Cox and a logistic regression, an extension of the Schoenfeld residuals could seem natural. Further research is however needed on this subject. An alternative method is to include in the model, for each covariate, its corresponding time-varying covariate by creating an interaction with the survival times. If it is significant, the proportionality of hazards should seriously be questioned. It would then be interesting to see the behavior of the adaptive LASSO in this context.

This method raises the question of how to deal with time-varying covariates in a mixture cure model. This is also a subject of considerable importance. Indeed, the mixture cure model assumes that the population is divided into a cured and an uncured sub-population. The presence of time-varying covariates would imply that the two sub-populations are not fixed. It could then be more natural to use another cure model such as the promotion time [Kim et al., 2013]. However, changing one's cure status over time is not at all unnatural and the use of the mixture cure model in this case is also highly interesting.

A third research topic, mentioned in Chapter 5, is the extension of bootstrap procedures in the presence of interval-censoring and cure. Indeed, the simulations concerning goodness-of-fit testing in Chapter 5 suggested that a larger sample size could be needed to acquire a larger power. A parametric bootstrap could thus enhance the results for small sample sizes. This means that event times need to be generated from the assumed distribution. These event times need to be interval-censored and right-censored. For this, sampling left and right interval endpoints from the original intervals is not enough. In a right-censoring context one only needs to right-censor the event times and this can be achieved by estimating the right-censoring distribution. Estimating

the interval-censoring distribution is however not straightforward and require further research.

Finally, the Alzheimer's disease database was analyzed considering that individuals who died during the study were simply right-censored. Typically, in these cases, death can be seen as a competing risk, so that the use of a competing risk model could be more appropriate. To the best of our knowledge, an extension of the mixture cure model with interval-censored data to the issue of competing risk has not been studied, but deserves our entire attention.

Bibliography

[Aerts et al., 1999] Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association*, 94(447):869–879.

[Anderson, 1962] Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Statist.*, 33(3):1148–1159.

[Berkson and Gage, 1952] Berkson, J. and Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

[Betensky, 2001] Betensky, R. a. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika*, 88(3):703–711.

[Betensky et al., 2002] Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21(2):263–275.

[Boag, 1949] Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53.

[Breiman, 1996] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6):2350–2383.

[Breslow, 1972] Breslow, N. (1972). Comment on D.R. Cox (1972) paper. *Journal of the Royal Statistical Society*, (34):216–217.

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg.

[Cai et al., 2012] Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260.

[Cai and Betensky, 2003] Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59(3):570–579.

[Carroll, 2003] Carroll, K. J. (2003). On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials*, 24(6):682 – 701.

[Casella and Berger, 2001] Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

[Chen et al., 2002] Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3(3):347–360.

[Chen et al., 1999] Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

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

[Collett, 2003] Collett, D. (2003). *Modelling Survival Data in Medical Research, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

[Cook and Weisberg, 1983] Cook, R. D. and Weisberg, S. (1983). *Residuals and Influence in Regression*. Monographs on Statistics and Applied Probability, 18. Chapman and Hall/CRC, 1st edition.

[Coppejans and Gallant, 2002] Coppejans, M. and Gallant, A. (2002). Cross-validated SNP density estimates. *Journal of Econometrics*, 110(1):27 – 65.

[Corder et al., 1993] Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261:921–923.

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

[Davidian and Gallant, 1992] Davidian, M. and Gallant, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics*, 20(5):529–556.

[Doehler and Davidian, 2008] Doehler, K. and Davidian, M. (2008). 'Smooth' inference for survival functions with arbitrarily censored data. *Statistics in Medicine*, 27(26):5421–5439.

[Dysken et al., 2014] Dysken, M., Sano, M., Asthana, S., and et al (2014). Effect of vitamin E and memantine on functional decline in Alzheimer disease: The TEAM-AD VA cooperative randomized trial. *JAMA*, 311(1):33–44.

[Efron and Hastie, 2004] Efron, B. and Hastie, T. (2004). Least angle regression. *The Annals of statistics*, 32(2):407 – 499.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[Fan and Li, 2002] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

[Fang et al., 2002] Fang, H.-B., Sun, J., and Lee, M.-L. T. (2002). Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica*, 12(4):1073–1083.

[Farewell, 1977a] Farewell, V. (1977a). A model for a binary variable with time-censored observations. *Biometrika Trust*, 64(1):43–46.

[Farewell, 1986] Farewell, V. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

[Farewell, 1977b] Farewell, V. T. (1977b). The combined effect of breast cancer risk factors. *Cancer*, 40(2):931–6.

[Farrington, 2000] Farrington, C. P. (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics*, 56(2):473–482.

[Fay, 1996] Fay, M. P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics*, 52(3):811–822.

[Fenton and Gallant, 1996] Fenton, V. M. and Gallant, A. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics*, 74(1):77 – 118.

[Finkelstein, 1986] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854.

[Friedman and Hastie, 2007] Friedman, J. and Hastie, T. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

[Gabler et al., 1993] Gabler, S., Laisney, F., and Lechner, M. (1993). Seminonparametric estimation of binary-choice models with an application to labor-force participation. *Journal of Business & Economic Statistics*, 11(1):61–80.

[Gallant et al., 1990] Gallant, A., Hansen, L., and Tauchen, G. (1990). Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution. *Journal of Econometrics*, 45(1-2):141–179.

[Gallant and Nychka, 1987] Gallant, R. A. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica*, 55(2):363–390.

[Geerdens et al., 2013] Geerdens, C., Claeskens, G., and Janssen, P. (2013). Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty. *Biostatistics*, 14(3):433–446.

[Gina R. Petroni, 1994] Gina R. Petroni, R. A. W. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics*, 50(1):77–87.

[Goeman, 2010] Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal. Biometrische Zeitschrift*, 52(1):70–84.

[Goetghebeur and Ryan, 2000] Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics*, 56(4):1139–44.

[Groeneboom and Wellner, 1992] Groeneboom, P. and Wellner, J. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. DMV Seminars. Birkhäuser Basel.

[Hald, 1949] Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal*, 1949(1):119–134.

[Hanin and Huang, 2014] Hanin, L. and Huang, L. S. (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis*, 130:261–274.

[Harrell, 2001] Harrell, F. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer.

[Hart, 1997] Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Series in Statistics. Springer-Verlag New York.

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82.

[Hosmer et al., 2011] Hosmer, D., Lemeshow, S., and May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley Series in Probability and Statistics. Wiley.

[Huang and Wellner, 1997] Huang, J. and Wellner, J. A. (1997). *Interval Censored Survival Data: A Review of Recent Progress*, pages 123–169. Springer US, New York, NY.

[Jongbloed, 1998] Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7(3):310–321.

[Kalbfleisch and Prentice, 2002] Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

[Khachaturian et al., 2004] Khachaturian, A., Corcoran, C., Mayer, L., Zandi, P., Breitner, J., and Investigators, C. C. S. (2004). Apolipoprotein E E4 count affects age at onset of Alzheimer disease, but not lifetime susceptibility: The cache county study. *Archives of General Psychiatry*, 61(5):518–524.

[Kim, 2007] Kim, K. I. (2007). Uniform convergence rate of the seminonparametric density estimator and testing for similarity of two unknown densities. *The Econometrics Journal*, 10(1):1–34.

[Kim et al., 2013] Kim, S., Zeng, D., Li, Y., and Spiegelman, D. (2013). Joint modeling of longitudinal and cure-survival data. *Journal of Statistical Theory and Practice*, 7(2):324–344.

[Klein and Moeschberger, 2003] Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

[Komárek et al., 2005] Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14(3):726–745.

[Kooperberg and Clarkson, 1997] Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics*, 53(4).

[Kuk and Chen, 1992] Kuk, A. and Chen, C.-h. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

[Lagakos, 1981] Lagakos, S. W. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika*, 68(1):93–98.

[Law and Brookmeyer, 1992] Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12):1569–1578.

[Lawless, 1980] Lawless, J. F. (1980). Inference in the Generalized Gamma and Log Gamma Distributions. *Technometrics*, 22(3):409–419.

[Lawless, 2003] Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Wiley.

[Lesaffre et al., 2005] Lesaffre, E., Komárek, A., and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, 14(6):539–552.

[Li and Taylor, 2002] Li, C.-S. and Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Statistics in medicine*, 21(21):3235–47.

[Li et al., 2001] Li, C.-S., Taylor, J. M. G., and Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, 54(4):389–395.

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

[Lin et al., 2002] Lin, D. Y., Wei, L. J., and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58(1):1–12.

[Lindsey, 1998] Lindsey, J. K. (1998). A Study of Interval Censoring in Parametric Regression Models. *Lifetime data analysis*, 4:329–354.

[Liu et al., 2012] Liu, X., Peng, Y., Tu, D., and Liang, H. (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in medicine*, 31(24):2882–91.

[Lockhart et al., 2014] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

[Lu and Zhang, 2006] Lu, W. and Zhang, H. H. (2006). Variable selection for linear transformation models via penalized marginal likelihood. Institute of Statistics Mimeo Series 2580, North Carolina State University.

[Maller and Zhou, 1996] Maller, R. A. and Zhou, X. (1996). *Suvival analysis with long term survivor*. John Wiley & Sons, Inc.

[Mantel, 1967] Mantel, N. (1967). Ranking procedures for arbitrarily restricted observation. *Biometrics*, 23(1):65–78.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[Nelder and Wedderburn, 1972] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384.

[Nysen et al., 2012] Nysen, R., Aerts, M., and Faes, C. (2012). Testing goodness of fit of parametric models for censored data. *Statistics in medicine*, 31(21):2374–85.

[Ortega et al., 2008] Ortega, E. M. M., Cancho, V. G., and Lachos, V. H. (2008). Assessing influence in survival data with a cure fraction and covariates. *SORT*, 32(2):115–140.

[Oulhaj et al., 2009] Oulhaj, A., Wilcock, G. K., Smith, a. D., and de Jager, C. a. (2009). Predicting the time of conversion to MCI in the elderly: role of verbal expression and learning. *Neurology*, 73(18):1436–42.

[Pan, 2000] Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56(1):199–203.

[Peng and Dear, 2000] Peng, Y. and Dear, K. B. G. (2000). A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics*, 56(1):237–243.

[Peng et al., 1998] Peng, Y., Dear, K. B. G., and Denham, J. W. (1998). A Generalized F Mixture Model for Cure Rate Estimation. *Statistics in Medicine*, 17:813–830.

[Peng and Taylor, 2016] Peng, Y. and Taylor, J. M. G. (2016). Residual-based model diagnosis methods for mixture cure models. *Biometrics*.

[Petersen et al., 2001] Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., Ritchie, K., Rossor, M., Thal, L., and Winblad, B. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology*, 58(12):1985–1992.

[Peto, 1973] Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(1):86–91.

[Rabinowitz et al., 1995] Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, 82(3):501–513.

[Reid, 1994] Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, 9(3):439–455.

[Seber and Lee, 2012] Seber, G. and Lee, A. (2012). *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

[Self and Grossman, 1986] Self, S. G. and Grossman, E. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*, 42(3):521–30.

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

[Sun, 2006] Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer-Verlag New York.

[Sun and Chen, 2010] Sun, X. and Chen, C. (2010). Comparison of Finkelstein's method with the conventional approach for interval-censored data analysis. *Statistics in Biopharmaceutical Research*, 2(1):97–108.

[Sy and Taylor, 2000] Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

[Taylor, 1995] Taylor, J. M. G. (1995). Semiparametric estimation in Failure Time Mixture Models Estimation. *Biometrics*, 51(3):899–907.

[Therneau et al., 1990] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77:147–160.

[Tian and Cai, 2006] Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika*, 93(2):329–342.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B(Methodological)*, 58(1):267–288.

[Tibshirani, 1997] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–95.

[Tsodikov, 1998] Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–16.

[Tsodikov, 2001] Tsodikov, A. (2001). Estimation of Survival Based on Proportional Hazards When Cure is a Possibility. *Mathematical and Computer Modelling*, 33:1227–1236.

[Tsodikov et al., 2003] Tsodikov, A., Ibrahim, J., and Yakovlev, A. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, 98:1063–1078.

[Turnbull, 1976] Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped , Censored and Truncated Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

[Wang and Leng, 2007] Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

[Wang et al., 2012] Wang, L., Du, P., and Liang, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735.

[Wang and Song, 2011] Wang, X. and Song, L. (2011). Adaptive Lasso Variable Selection for the Accelerated Failure Models. *Communications in Statistics - Theory and Methods*, 40(24):4372–4386.

[Wei, 1992] Wei, L. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11:1971–1879.

[Weisberg, 2005] Weisberg, S. (2005). *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley.

[Wellner and Zhan, 1997] Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, 92(439):945–959.

[Westergaard, 1932] Westergaard, H. (1932). *Contributions to the History of Statistics*. P.S. King & Son, LTD., Westminster.

[Yakovlev et al., 1996] Yakovlev, A., Tsodikov, A., and Asselain, B. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Mathematical Biology and Medicine, Vol 1. World Scientific.

[Yamaguchi, 1992] Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan. *Journal of the American Statistical Association*, 87(418):284–292.

[Zhang and Davidian, 2001] Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802.

[Zhang and Lu, 2007] Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

[Zhang and Peng, 2007] Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26:3157–3171.

[Zhang and Peng, 2012] Zhang, J. and Peng, Y. (2012). Semiparametric estimation methods for the accelerated failure time mixture cure model. *Journal of the Korean Statistical Society*, 41:415–422.

[Zhang and Davidian, 2008] Zhang, M. and Davidian, M. (2008). "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, 64(2):567–76.

[Zou, 2006] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

[Zou and Zhang, 2009] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, 37(4):1733–1751.

APPENDIX A

Appendix to Chapter 2

A.1 R Code to generate data

Here is an example of an R code that generated interval-censored data:

```
DataGeneration = function(n,survDens, IC ="Yes",
Cure="Yes",gamma0=2, distCens='Fix', cc=1/20,mm,ll){
  #-------------------------------------------------#
  #    This function generates data whose survival
  #    distribution is either EGG, a mixture of normal,
  #    or a normal.
  #     A positive cure fraction is present by setting
  #    Cure="Yes". The cure rate depends on the value
  #    of gamma's.
  #-------------------------------------------------#
  regg =function(n, q, sigma, mu){
    if (q!=0){
      qs <- q/sigma
      lbd <- exp(-qs*mu)/q^2
      (rgamma(n,q^(-2))/lbd)^(1/qs)
    }
    else rlnorm(n,mu,sigma)
  }
  # --- Step1 : Create survival distribution for uncured people.
  #        Event times follow an AFT model.
```

```
if(survDens == "egg"){
  tbeta = c(-1,2,0.5,-0.5)
  X=cbind(1,rbinom(n,1,0.5),rnorm(n,0,1), rnorm(n,-1,1))
  tmuX=as.vector(X%*%tbeta)
  tq=1      ; talpha=-2 ; tsigma =exp(talpha)
  Tu=regg(n,tq,tsigma,tmuX)
  down=0  ;up=1  ;c=rexp(n, 1/120)
  a=0.001;b=0.1    ;
  MM=mm #max number visits
  LL=ll  # time between 2 visits
}
else if(survDens== "mixNorm"){
  tbeta = c(1.2,-2,0.5,-0.5)
  X=cbind(1,rbinom(n,1,0.5),rnorm(n,0,1), rnorm(n,-1,1))
  tmuX=as.vector(X%*%tbeta)
  components <- sample(1:2,prob=c(0.3, 0.7),size=n,replace=TRUE)
  mus = c(0.5,1)
  sds = c(0.2,0.7)
  e0 <- rnorm(n,mean=mus[components],sd=sds[components])
  Tu = exp(tmuX+e0)
  down=0  ;up=1  ;c=rexp(n,1/80)
  a=0.001;b=0.15
  MM=mm #max number visits
  LL=ll  # time between 2 visits
}
else if(survDens=="norm"){
  tbeta = c(2,-2,0.5,-0.5)
  X=cbind(1,rbinom(n,1,0.5),rnorm(n,0,1), rnorm(n,-1,1))
  tmuX=as.vector(X%*%tbeta)
  down=0;up=1
  e0=rnorm(n,0,0.75)
  Tu = exp(tmuX + e0)
  a=0.005;b=0.2
  MM=mm #max number visits
  LL=ll # time between 2 visits
}
  tbeta = c(-1,2,0.5,-0.5)
  X=cbind(1,rbinom(n,1,0.5),rnorm(n,0,1), rnorm(n,-1,1))
  tmuX=as.vector(X%*%tbeta)
  tq=1      ; talpha=0; tsigma =exp(talpha)
  Tu=regg(n,tq,tsigma,tmuX)
  down=0  ;up=1  ;c=rexp(n, 1/120)
  a=0.001;b=0.1    ;
```

```
    MM=mm #max number visits
    LL=ll  # time between 2 visits
  }
  #--- Step2 : Create cure probability#
  # tgamma= true cure parameter
  tgamma=c(gamma0,-1,-1)
  Z =cbind(1, runif(n,down,up),runif(n,0,1))
  tgammaZ= as.vector(Z%*%tgamma)
  tpcureZ = exp(tgammaZ)/(1+exp(tgammaZ))
  Y.susc=sapply(X=tpcureZ,FUN=rbinom,size=1,n=1)
  if(Cure =="No") Y.susc=rep(1,n)

  #--- Step 3 : Create the observed event time
  To=Tu+(Y.susc==0)/Y.susc   # global event time for everyone
  #--- Step4: Intervals creation if IC = "Yes",
  #           else, observed times are right censored
  if(IC=='Yes'){
    L=rep(0,n)
    R=rep(Inf,n)
    U = runif(n,a,b)
    event=rep(NA,n)
    for(i in 1:n) {
      if(To[i]<=U[i]) {R[i]=U[i];L[i]=0;event[i]=2}
      if(To[i]>U[i]+MM*LL) {R[i]=Inf;L[i]=U[i]+MM*LL;event[i]=0}
      for(k in 1:MM){
        if((U[i]+(k-1)*LL<To[i])&(To[i]<=U[i]+k*LL)) {
R[i]=U[i]+k*LL;L[i]=U[i]+(k-1)*LL;event[i]=3}
      }
    }
  }
  else{
    if(distCens == "Expo") c = rexp(n,cc)
    else if(distCens =="Unif") c = runif(n,0.01,cc)
    else if(distCens =="Fix") c = cc
    else if(distCens=="NA") c = Inf
    Toc=pmin(To,c)
    event = To==Toc
    L =Toc
    R = 0
  }
  return(list(X, Z, L, R,event,Y.susc))
}
```

A.2 Tables

Table A.1 – Bias and MSE when using exact likelihood (n=200, 300 or 500)

Using exact likelihood									
		q		β_0		β_1		α	
a	n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0,5	100	-0,247	1,675	-0,130	0,099	-0,021	0,067	-0,063	0,062
	200	-0,187	0,135	-0,071	0,035	-0,016	0,031	-0,028	0,008
	500	-0,103	0,040	-0,039	0,012	-0,005	0,010	-0,007	0,002
1	100	-0,287	2,234	-0,123	0,113	-0,021	0,069	-0,080	0,078
	200	-0,173	0,133	-0,058	0,034	-0,018	0,032	-0,037	0,009
	500	-0,095	0,040	-0,031	0,012	-0,005	0,010	-0,012	0,003
2	100	-0,040	3,578	-0,043	0,116	-0,024	0,069	-0,137	0,150
	200	-0,079	0,117	0,005	0,035	-0,024	0,034	-0,063	0,015
	500	-0,046	0,035	0,009	0,014	-0,009	0,011	-0,026	0,004
3	100	0,426	7,826	0,065	0,148	-0,027	0,073	-0,230	0,339
	200	0,036	0,128	0,085	0,049	-0,034	0,035	-0,092	0,024
	500	0,022	0,035	0,063	0,019	-0,016	0,012	-0,042	0,006
4	100	1,117	13,042	0,204	0,229	-0,045	0,080	-0,383	0,645
	200	0,250	1,156	0,184	0,091	-0,044	0,037	-0,145	0,075
	500	0,109	0,052	0,132	0,037	-0,024	0,013	-0,060	0,008
5	100	2,073	22,859	0,318	0,332	-0,041	0,093	-0,569	1,106
	200	0,438	2,029	0,280	0,149	-0,056	0,041	-0,192	0,120
	500	0,194	0,087	0,202	0,063	-0,030	0,014	-0,080	0,013

Table A.2 – Bias and MSE when using midpoint imputation approach (n=200, 300 or 500)

Using midpoint imputation approach									
		q		β_0		β_1		α	
a	n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0,5	100	-0,587	0,727	-0,226	0,129	-0,056	0,067	-0,044	0,022
	200	-0,410	0,256	-0,156	0,052	-0,041	0,031	-0,016	0,006
	500	-0,310	0,120	-0,121	0,024	-0,023	0,010	0,000	0,002
1	100	-0,483	0,542	-0,153	0,093	-0,070	0,067	-0,075	0,023
	200	-0,337	0,189	-0,091	0,031	-0,058	0,031	-0,053	0,009
	500	-0,267	0,093	-0,068	0,013	-0,042	0,011	-0,034	0,003
2	100	-0,073	0,251	0,091	0,060	-0,099	0,062	-0,188	0,063
	200	-0,024	0,055	0,125	0,032	-0,099	0,035	-0,160	0,034
	500	-0,036	0,017	0,126	0,022	-0,088	0,016	-0,124	0,018
3	100	0,360	0,356	0,336	0,148	-0,143	0,066	-0,336	0,162
	200	0,291	0,132	0,345	0,131	-0,151	0,044	-0,281	0,090
	500	0,218	0,062	0,332	0,115	-0,140	0,027	-0,223	0,053
4	100	0,848	1,918	0,571	0,354	-0,211	0,090	-0,519	0,359
	200	0,585	0,409	0,545	0,307	-0,200	0,060	-0,407	0,183
	500	0,451	0,219	0,522	0,277	-0,188	0,042	-0,319	0,106
5	100	1,215	2,919	0,754	0,595	-0,265	0,118	-0,663	0,554
	200	0,847	0,823	0,721	0,528	-0,249	0,081	-0,517	0,290
	500	0,651	0,442	0,689	0,478	-0,232	0,061	-0,406	0,170

APPENDIX B

Appendix to Chapter 3

B.1 Non-convergence

In Chapter 3, we report results of simulations concerning variable selection using the adaptive LASSO. For some scenarios, the algorithm failed to converge. In Table B.1, we report, for each scenario and each sample size, the number of failures experienced.

Table B.1 – Number of failure (percentage of failures) over 2000 replications.

Sample Size	(20% Cure) (40% RC)		(30% Cure) (40% RC)		(40% Cure) (60% RC)	
$n = 200$	54	(2,7%)	1	(0,1%)	19	(1,0%)
$n = 300$	24	(1,2%)	0	(0,0%)	3	(0,2%)
$n = 500$	0	(0,0%)	0	(0,0%)	0	(0,0%)

B.2 R code

In this section, we include the R-code we use to get estimates for our EGG-AFT mixture cure model, with adaptive LASSO.

```
regg =function(n, q, sigma, mu){
  #Generate EGG-distributed random variables
  if (q!=0){
```

```
    qs <- q/sigma
    lbd <- exp(-qs*mu)/q^2
    (rgamma(n,q^(-2))/lbd)^(1/qs)
    # E0=rgamma(n,q^(-2)); E=log(q^2*E0)/q; X=exp(mu+sigma*E)
  }
  else rlnorm(n,mu,sigma) # Y=rnorm(n); X=exp(mu+sigma*Y)
}
Segg=function (t, q, sigma, mu) {
  # Survival distribution of the EGG.
  if (q != 0) {
    qs <- q/sigma
    lbd <- exp(-qs * mu)/q^2
    if (q > 0) 1-pgamma(lbd * t^qs, q^(-2))
    else pgamma(lbd * t^qs, q^(-2))  }
  else 1-plnorm(t, mu, sigma) }
fegg=function (t, q, sigma, mu){
  #Density function of the EGG
  if (q != 0) {
    qs <- q/sigma
    lbd <- exp(-qs * mu)/q^2
    newt <- t^qs
    dgamma(lbd * newt, q^(-2)) * lbd * abs(qs) * t^(qs-1)
  }
  else dlnorm(t, mu, sigma)
}
loglik = function(param, X, Z1, Z2, L, R, event,fegg,Segg){
  # Likelihood function for interval censoring , right
  # censoring, left censoring and cure.
  # param = c(q, beta, gnamma, alpha) as in the paper
  q = as.vector(param[1])
  X = as.matrix(X); p = ncol(X)
  beta= as.vector(param[2:(p+1)])
  muX = as.vector(X%*%beta)
  n = nrow(as.matrix(L))
  if (is.null(Z1)) {s1=0; pcureZ1 = rep(1,n)} else
  {Z1 = as.matrix(Z1); s1= ncol(Z1) #if Z1=1 pcure=constant;
   gama= as.vector(param[(p+2):(p+s1+1)])
   gamaZ1= as.vector(Z1%*%gama)
   pcureZ1 = exp(gamaZ1)/(1+exp(gamaZ1))}
  if(is.null(Z2)) {s2=0; sigmaZ2=param[p+s1+2]} else
{Z2 = as.matrix(Z2); s2= ncol(Z2) #if Z2=1 var=constant;
 alpha= as.vector(param[(p+s1+2):(p+s1+s2+1)])
 sigmaZ2 =as.vector(exp(Z2%*%alpha))}
fl=fegg(L,q,sigmaZ2,muX)
```

```
Sr=Segg(R,q,sigmaZ2,muX)
Sl=Segg(L,q,sigmaZ2,muX)
L.E=(pcureZ1*fl)[event==1]
L.LC=(pcureZ1*(1-Sr))[event==2]
L.IC=(pcureZ1*(Sl-Sr))[event==3]
L.RC=((pcureZ1*Sl)+(1 - pcureZ1))[event==0]
logLik =sum(log(L.E))+sum(log(L.LC))+sum(log(L.IC))+sum(log(L.RC))
return(-logLik)
}
loglik_fixQ = function(param,q, X, Z1, Z2, L, R, event, fegg, Segg){
  # EGG-AFT mixture cure likelihood, with fixed q,
  # to test for weibull (q=1) or lognormal(q=0)
  n = nrow(as.matrix(L))
  paramQ = c(q, param)
  ll = loglik(paramQ,X, Z1, Z2, L, R, event, fegg,Segg)
  ll
}
loglik_fixQ_Alpha = function(param,q,alpha, X,
Z1, Z2, L, R, event, fegg, Segg){
  # EGG-AFT mixture cure likelihood, with fixed q AND alpha,
  # to test for exponential (q=1, alpha=0)
  n = nrow(as.matrix(L))
  paramQ = c(q, param,alpha)
  ll = loglik(paramQ,X, Z1, Z2, L, R, event, fegg,Segg)
  ll
}
initVal = function(X, Z,Xsigma, L, R, event,lasso=FALSE){
  #Gives initial values
  n=length(event)
  response =ifelse(L==0, R, L)
  status = event
  if(any(event==3) || (event==2)){
    response = ifelse(event==3, (L+R)/2, response)
    status = ifelse(event==0, 0 ,1)
  }

  inits.beta = if(!is.null(X)) {lm(log(response)~X[,-1])$coef}
             else c()
  inits.gamma = if(!is.null(Z)) {glm(status~Z[,-1],
family ="binomial")$coef} else c()

  inits.param = c(1, inits.beta,inits.gamma,
rep(1, ncol(as.matrix(Xsigma))))
  inits.param
```

```
}


#------------------- ADAPTIVE LASSO-----------------------#
require(lars)
lsaS <- function(intercept, n, Sigma, beta.ols)
{  #LSA version allowing "our" log-likelihood.
  # You NEED to run to original LARS variant for LSA from
# [Wang&Leng,2007].
  t1 <- sort(l.fit$BIC, index.return=TRUE)
  t2 <- sort(l.fit$AIC, index.return=TRUE)
  beta <- l.fit$beta
  if(intercept) {
    beta0 <- l.fit$beta0+beta.ols[1]
    beta.bic <- c(beta0[t1$ix[1]],beta[t1$ix[1],])
    beta.aic <- c(beta0[t2$ix[1]],beta[t2$ix[1],])
  }
  else {
    beta0 <- l.fit$beta0
    beta.bic <- beta[t1$ix[1],]
    beta.aic <- beta[t2$ix[1],]
    beta.bic2 = beta[t1$ix[2],]
  }
  obj <- list(beta.ols=beta.ols, beta.bic=beta.bic,
              beta.aic = beta.aic, BIC= l.fit$BIC,
              dof = l.fit$dof,all.beta =beta )
  obj
}

# --------- Adaptive Lasso for our mixture cure model ----------#
DoubleLassoDoubleLik = function(inits, X,Z1,Z2,L,R,event,fegg,
Segg,method){
  ##############################################################
  #   Input :
  #    inits : the initial value for the likelihood optimisation
  #    X  : the covariates to be included in the latency part
  #    Z1 : the covariates to be included in the incidence part
  #    Z2 : the covariates to be included in the scale part
  #    L  : the left bound of the interval for interval censored
# events
  #    R  : the right bound of the interval for interval censored
#           events
  #    event : indicator variable (0 if RC, 1 if event, 3 if IC)
  #    fegg : egg density distribution
```

```
  #    Segg : egg survival distribution
  #
  #  Output :
  #    coefficients estimations with exact zero
  #
  #  Method :
  #  This function performs variable selection and estimation
#  through the alasso for a mixture cure model and a likelihood
#  'loglik'(here: egg likelihood).
  #  It proceeds iteratively with 2 steps : first the adaptive
#  lasso on the latency part (i.e. on the columns of the
#  hessian of the likelihood for the 'X' matrix), then on
#  the incidence part (i.e. on the columns of the hessian of
#  the likelihood for the Z1 matrix). No shrinkage is
  #  made for intercepts and scale part.
  #
  ################################################################

  X = as.matrix(X); Z1 = as.matrix(Z1); Z2 = as.matrix(Z2)
  Xor =X; X = scale(X, F, T)               #Scaling X
  scaleX = attr(X,"scaled:scale")
  Z1or = Z1;Z1 = scale(Z1, FALSE, TRUE)    #Scaling Z1
  scaleZ1 = attr(Z1,"scaled:scale")
  p=ncol(X)                 # Number of covariates in latency
  s=ncol(Z1)                # Number of covariates in incidence
  v = ncol(Z2)              # Number of covariates in location
  n=nrow(X)                 # Number of observations
  #-------Iteration
  eps = 0.0001
  delta = 1
  conv = 0
  inits = inits*c(1, scaleX, scaleZ1, rep(1, ncol(Z2)))

  optim.obj =optim(inits,loglik,X=X,Z1=Z1,Z2=Z2,L=L,R=R,
event=event,fegg=fegg, Segg=Segg,hessian = TRUE,
method = method, control = list(maxit = 2000))
  p.optim = optim.obj$par
  q.optim = p.optim[1]
  beta.optim= p.optim[(2):(p+1)]
  gamma.optim = p.optim[(p+2):(p+s+1)]
  alpha.optim = p.optim[(p+s+2) : length(p.optim)]
  beta.lsa.old = beta.optim[-1]
  gamma.lsa.old = gamma.optim[-1]
  while(delta>eps ){
```

```
    #Beta-Part
    optim.Lat = optim(beta.optim,likBeta,param=c(q.optim,
            gamma.optim,alpha.optim), X=X,Z=Z1,
ZZ=Z2,L=L,R=R,event=event,fegg=fegg,Segg=Segg,
               hessian=T,method = method,
control = list(maxit = 1000))
    beta.par = optim.Lat$par;ddf.Lat = optim.Lat$hessian
    lsa.obj.Lat = lsaS(intercept=F,n=n,
          Sigma=solve(ddf.Lat[-1,-1]),beta.ols=beta.par[-1])
    beta.lsa = lsa.obj.Lat$beta.bic
    #Gamma-Part
    optim.Inc = optim(gamma.optim,likGamma,param=c(q.optim,
beta.optim,alpha.optim), X=X,Z=Z1,ZZ=Z2,L=L,
R=R,event=event,fegg=fegg,Segg=Segg,
                      hessian=T,method = method,
control = list(maxit = 1000))
    gamma.par = optim.Inc$par;ddf.Inc = optim.Inc$hessian
    lsa.obj.Inc = lsaS(intercept=F,n=n,
          Sigma=solve(ddf.Inc[-1,-1]),beta.ols=gamma.par[-1])
    gamma.lsa = lsa.obj.Inc$beta.bic

    inits =c(q.optim, beta.par[1],beta.lsa, gamma.par[1],
gamma.lsa, alpha.optim)
    optim.obj =optim(inits,loglik,X=X,Z1=Z1,Z2=Z2,L=L,R=R,
event=event,fegg=fegg,
                Segg=Segg, hessian = TRUE,method =method,
control = list(maxit = 1000))
    p.optim = optim.obj$par
    q.optim.new = p.optim[1]
    alpha.optim.new = p.optim[(p+s+2) : length(p.optim)]
    delta = max(c(abs(q.optim-q.optim.new),
abs(beta.lsa-beta.lsa.old),
            abs(gamma.lsa-gamma.lsa.old), alpha.optim.new-alpha.optim))
    beta.optim =beta.par
    beta.lsa.old= beta.lsa
    gamma.optim= gamma.par
    gamma.lsa.old = gamma.lsa
    q.optim= q.optim.new
    alpha.optim = alpha.optim.new
    inits =c(q.optim, beta.par[1],beta.lsa,
gamma.par[1],gamma.lsa, alpha.optim)
    conv = conv+1
    if(conv ==25) stop("non convergence")
  }
```

```
  return(inits/c(1, scaleX, scaleZ1, rep(1, ncol(Z2))))
}
```

APPENDIX C

Appendix to Chapter 4

C.1 Deviance with the Extended Generalized Gamma distribution

The purpose of this section to present the deviance in a mixture cure model assuming an AFT model with the EGG distribution in the latency part. Contributions to the deviance of a right-censored observation, an exact event time observation, and an interval-censored observation are given separately. The deviance can be computed by summing all contributions.

C.1.1 Deviance for right-censored observations in the EGG-AFT mixture cure model.

The contribution to the saturated complete-data log-likelihood of a right-censored observation is:

$$y_i \log(p_i(\nu_i)) + (1 - y_i) \log(1 - p_i(\nu_i)) + y_i \log(S_u(t_i|\mu_i)). \qquad \text{(C.1)}$$

Maximizing according to μ_i and ν_i leads to $S_u(t_i|\tilde{\mu}_i) = 1$ and $\tilde{\nu}_i$ such that $p(\nu_i) = y_i$. The contribution to the saturated log-likelihood of a right-censored observation is then always equal to $y_i \log(p_i(\tilde{\nu}_i)) + (1 - y_i) \log(1 - p_i(\tilde{\nu}_i))$. Consequently, the contribution to the deviance for a right-censored observation

is:

$$\xi_i \log\left(\frac{\xi_i}{\hat{p}}\right) + (1-\xi_i)\log\left(\frac{1-\xi_i}{1-\hat{p}}\right) - (1-\delta_i)\xi_i \log(\hat{S}_u(t_i)), \tag{C.2}$$

where ξ_i is the estimated value of the partially unobserved variable y_i.

C.1.2 Deviance for exact event time in the EGG-AFT mixture cure model.

The contribution to the log-likelihood of an exact observation is:

$$y_i \log(p_i(\nu_i)) + \log(f_u(t_i|\mu_i)). \tag{C.3}$$

Maximizing according to μ_i and ν_i leads to $\tilde{\mu}_i$ such that $\tilde{\mu}_i = \log(t_i)$ and $\tilde{\nu}_i$ such that $p(\tilde{\nu}_i) = y_i$. Consequently, $\frac{\log(t_i)-\tilde{\mu}_i}{\sigma} = 0$ and $f_u(t_i|\tilde{\mu}_i) = (\sigma t_i)^{-1} f_{u,\varepsilon}(0)$. Noting $v_{t_i} = \frac{\log(t_i)-\mu(\boldsymbol{x}_i)}{\sigma}$, the contribution to the saturated log-likelihood of an exact observation is:

$$\begin{aligned}
\log(f_u(t_i|\tilde{\mu}_i) - \log(\hat{f}_u(t_i)) &= \log((\sigma t_i)^{-1} f_{u,\varepsilon}(0)) - \log((\sigma t_i)^{-1} f_{u,\varepsilon}(v_{t_i})) \\
&= \log\left[(\sigma t_i)^{-1} \frac{|q|}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp(q^{-2}(0 - e^0))\right] \\
&- \log\left[(\sigma t_i)^{-1} \frac{|q|}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp(q^{-2}(q v_{t_i} - e^{q v_{t_i}}))\right] \\
&= -q^{-2} - q^{-2}(q v_{t_i} - e^{q v_{t_i}}) \\
&= -q^{-2}\left(1 + q v_{t_i} - e^{q v_{t_i}}\right).
\end{aligned}$$

Consequently, the contribution to the deviance in the EGG-AFT model for an exact observation is:

$$\xi_i \log\left(\frac{\xi_i}{\hat{p}}\right) - q^{-2}\left(1 + q v_{t_i} - e^{q v_{t_i}}\right),$$

where ξ_i is the estimated value of the partially unobserved variable y_i.

C.1.3 Deviance for interval-censored observations in the EGG-AFT mixture cure model.

The contribution to the likelihood of an interval-censored observation is

$$y_i \log(p_i(\nu_i)) + \log(S_u(l_i|\mu_i) - S_u(r_i|\mu_i)).$$

Maximizing according to ν_i leads to $p(\tilde{\nu}_i) = y_i$, and maximizing $S_u(l_i|\mu_i) - S_u(r_i|\mu_i)$ is completed by equating to zero the first derivatives with respect to μ_i. Focusing on the case $q > 0$, and noting $m_{t_i} = q^{-2}e^{q\frac{\log(t_i)-\mu_i}{\sigma}} = q^{-2}t_i^{q/\sigma}\exp(-q\mu_i/\sigma)$,

$$\begin{aligned} S_u(l_i|\mu_i) - S_u(r_i|\mu_i) &= \frac{1}{\Gamma(q^{-2})}\int_0^{m_{r_i}} x^{q^{-2}-1}e^{-x}dx - \frac{1}{\Gamma(q^{-2})}\int_0^{m_{l_i}} x^{q^{-2}-1}e^{-x}dx \\ &= \frac{1}{\Gamma(q^{-2})}\int_{m_{l_i}}^{m_{r_i}} x^{q^{-2}-1}e^{-x}dx. \end{aligned}$$

The Leibniz integral rule is applied and

$$\begin{aligned} \frac{\partial}{\partial\mu_i}\left(S_u(l_i|\mu_i) - S_u(r_i|\mu_i)\right) &= 0 \qquad \text{(C.4)} \\ m_{r_i}^{q^{-2}-1}\exp(-m_{r_i})m_{r_i}\frac{-q}{\sigma} &= m_{l_i}^{q^{-2}-1}\exp(-m_{l_i})m_{l_i}\frac{-q}{\sigma} \\ m_{r_i}^{q^{-2}}e^{-m_{r_i}} &= m_{l_i}^{q^{-2}}e^{-m_{l_i}} \\ \left(\frac{m_{l_i}}{m_{r_i}}\right)^{q^{-2}} &= e^{-m_{r_i}}e^{m_{l_i}} \\ q^{-2}\log(m_{l_i}/m_{r_i}) &= m_{l_i} - m_{r_i} \\ q^{-2}\log\left(\frac{q^{-2}l_i^{q\sigma^{-1}}\exp(-q\mu_i\sigma^{-1})}{q^{-2}r_i^{q\sigma^{-1}}\exp(-q\mu_i\sigma^{-1})}\right) &= q^{-2}l_i^{q\sigma^{-1}}\exp(-q\mu_i\sigma^{-1}) \\ &\quad - q^{-2}r_i^{q\sigma^{-1}}\exp(-q\mu_i\sigma^{-1}) \\ \exp(-q\mu_i\sigma^{-1}) &= \frac{\log\left(l_i/r_i\right)^{q\sigma^{-1}}}{l_i^{q\sigma^{-1}} - r_i^{q\sigma^{-1}}} \\ \mu_i &= -\frac{\sigma}{q}\log\left(\frac{\log\left(l_i/r_i\right)^{q\sigma^{-1}}}{l_i^{q\sigma^{-1}} - r_i^{q\sigma^{-1}}}\right). \qquad \text{(C.5)} \end{aligned}$$

The saturated log-likelihood is maximized if $\tilde{\mu}_i$ satisfies Equation (C.5). Consequently, the contribution to the deviance in the EGG-AFT model for an interval-censored observation is:

$$\xi_i\log\left(\frac{\xi_i}{\hat{p}}\right) + \log\left(\frac{S_u(l_i|\tilde{\mu}_i) - S_u(r_i|\tilde{\mu}_i)}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}\right), \qquad \text{(C.6)}$$

where ξ_i is the estimated value of the partially unobserved variable y_i.

C.2 Additional plots

Supplementary online material contains additional plots at `http://dial.uclouvain.be/`.

APPENDIX D

Appendix to Chapter 5

D.1 Estimation

D.1.1 Results with the BIC criterion

Tables D.1 and D.2 present the results corresponding to Tables 2 and 3 from the main paper, but using the BIC criterion instead of the AIC criterion to chose the best fitting model. Conclusions are similar except that the MSE is slightly lower when using the BIC criterion

D.1.2 Results of estimation using true and Normal distribution for Scenario B.

We fitted the data from Scenario B using the true distribution, i.e. the Weibull, see Table D.3. Results when data are fitted with the log-Normal distribution are given in Table D.4.

D.2 Goodness-of-fit

D.2.1 Testing Weibull event times

Table D.5 shows the results of testing if the error term of the AFT model follows an extreme-value distribution. The power is quite high for Scenario

Table D.1 – Relative bias and MSE for scenario A, B and C; for sample size $n = 200$ with BIC criterion.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	0.00	0.15	0.01	0.17	0.02	0.21
β_1	0.00	0.13	0.00	0.15	0.00	0.19
β_2	0.00	0.06	0.00	0.07	-0.01	0.08
β_3	-0.01	0.06	-0.01	0.07	-0.01	0.09
σ	-0.02	0.05	-0.03	0.07	-0.04	0.10
	Incidence Part: Logistic regression					
γ_0	0.06	1.07	0.04	0.81	0.03	0.78
γ_1	0.01	1.18	-0.01	0.89	0.00	0.93
γ_2	0.08	1.17	0.09	0.92	0.01	0.98
Scenario B						
	Latency Part: AFT Model					
β_0	-0.01	0.02	-0.01	0.02	-0.01	0.03
β_1	0.00	0.01	0.00	0.02	0.00	0.03
β_2	0.00	0.01	0.00	0.01	0.00	0.02
β_3	0.00	0.01	0.00	0.01	0.00	0.02
σ	-0.01	0.01	-0.02	0.01	-0.05	0.01
	Incidence Part: Logistic regression					
γ_0	0.05	0.85	0.05	0.68	0.04	0.69
γ_1	0.09	0.94	0.10	0.79	0.08	0.78
γ_2	0.03	0.92	0.05	0.79	0.04	0.86
Scenario C						
	Latency Part: AFT Model					
β_0	-0.02	0.26	-0.01	0.25	-0.01	0.21
β_1	-0.01	0.10	-0.01	0.11	0.00	0.13
β_2	0.00	0.05	0.00	0.06	0.00	0.07
β_3	0.00	0.05	0.00	0.06	0.00	0.07
σ	-0.07	0.09	-0.09	0.09	-0.09	0.09
	Incidence Part: Logistic regression					
γ_0	0.04	1.02	0.02	0.77	0.01	0.64
γ_1	0.07	1.14	0.03	0.89	0.01	0.80
γ_2	0.00	1.10	0.02	0.85	0.01	0.74

Table D.2 – Relative bias and MSE for scenario A, B and C; for sample size $n = 500$ with BIC criterion.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	-0.01	0.14	-0.01	0.14	0.00	0.15
β_1	0.00	0.08	0.00	0.09	0.00	0.11
β_2	-0.01	0.04	-0.01	0.04	-0.01	0.05
β_3	0.00	0.04	0.00	0.04	0.00	0.05
σ	0.00	0.05	0.00	0.05	-0.01	0.06
	Incidence Part: Logistic regression					
γ_0	0.03	0.59	0.02	0.49	0.04	0.50
γ_1	0.01	0.69	-0.01	0.55	0.03	0.56
γ_2	0.09	0.67	0.05	0.57	0.08	0.56
Scenario B						
	Latency Part: AFT Model					
β_0	0.00	0.02	0.00	0.02	-0.01	0.02
β_1	0.00	0.01	0.00	0.01	0.00	0.02
β_2	0.00	0.00	0.00	0.01	0.00	0.01
β_3	0.00	0.00	0.00	0.01	0.00	0.01
σ	-0.01	0.00	-0.01	0.01	-0.01	0.01
	Incidence Part: Logistic regression					
γ_0	0.02	0.48	0.01	0.40	0.01	0.41
γ_1	0.03	0.54	0.00	0.48	-0.02	0.51
γ_2	0.03	0.57	0.01	0.47	0.00	0.51
Scenario C						
	Latency Part: AFT Model					
β_0	-0.11	0.45	-0.06	0.38	-0.05	0.32
β_1	-0.01	0.06	-0.01	0.07	0.00	0.08
β_2	0.00	0.03	0.00	0.04	0.00	0.04
β_3	0.00	0.03	0.00	0.04	0.01	0.04
σ	-0.03	0.10	-0.06	0.10	-0.07	0.08
	Incidence Part: Logistic regression					
γ_0	-0.01	0.54	0.00	0.46	-0.01	0.40
γ_1	0.00	0.61	0.01	0.50	-0.02	0.46
γ_2	-0.04	0.63	-0.02	0.52	-0.02	0.46

Table D.3 – Relative bias and MSE for scenario B, using the true model for estimation purposes. Sample size: $n = 200$.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	-0.01	0.04	-0.01	0.04	-0.01	0.06
β_1	0.00	0.04	0.00	0.04	0.00	0.07
β_2	0.00	0.02	0.00	0.02	0.00	0.03
β_3	0.00	0.02	0.00	0.02	0.00	0.03
σ	-0.02	0.01	-0.03	0.01	-0.05	0.02
	Incidence Part: Logistic regression					
γ_0	0.05	0.87	0.05	0.67	0.04	0.65
γ_1	0.08	0.97	0.11	0.78	0.09	0.72
γ_2	0.03	0.92	0.06	0.78	0.04	0.80

Table D.4 – Relative bias and MSE for scenario B, using the log-Normal distribution for estimation purposes. Sample size: $n = 200$.

	10% Cure 20% RC		20% Cure 35% RC		30% Cure 55% RC	
	Rel. Bias	MSE	Rel. Bias	MSE	Rel. Bias	MSE
Scenario A						
	Latency Part: AFT Model					
β_0	-0.09	0.05	-0.09	0.06	-0.09	0.07
β_1	0.00	0.04	0.00	0.05	0.01	0.08
β_2	0.00	0.02	0.00	0.02	0.00	0.03
β_3	0.01	0.02	0.00	0.02	0.00	0.03
σ	0.25	0.02	0.26	0.02	0.27	0.03
	Incidence Part: Logistic regression					
γ_0	0.06	0.89	0.06	0.69	0.05	0.66
γ_1	0.09	0.99	0.12	0.79	0.10	0.73
γ_2	0.04	0.94	0.07	0.79	0.04	0.81

Table D.5 – Goodness-of-fit: Testing if the error term follows an exponential distribution, for each scenario and each level of cure and right-censoring (RC). Percentage of rejection over 500 replications with the SNP representation, for $n = 200$.

Cure	10%	20%	30%
RC	20%	35%	55%
Scenario A	98.4%	94.2%	78.2%
Scenario B	5.8%	10.2%	14.8%
Scenario C	100.0%	99.8%	99.8%

A and C. The test rejects the null hypothesis too often for smaller sample sizes: for scenario B, when the event times follows a Weibull, the type-I-error is close to the 5%-level when there is light censoring. However, the increasing censoring level increases the type-I-error as well, and reaches 15% in case of heavy censoring. Increasing the sample size from $n = 200$ to $n = 500$, the type-I-error becomes 4.4%, 5.40% 13% for light, medium and heavy censoring respectively, illustrating the need for larger sample sizes in the presence of heavy censoring.

D.2.2 Testing Exponential event times

For comparison purposes, we also tested, at the 5%-level, if the data came from an exponential distribution ($\sigma = 1$) for $n = 200$. For this small sample size, the power of this test is high: the results for scenario A were, respectively 100%, 99.80% and 94.20% rejection, and for Scenario B and C, the rejection percentage was 100% for every level of censoring. When data truly came from an exponential distribution, the type-I-error was 2%, 4% and 5% for light, medium and heavy censoring respectively, very close to the 5% level.

D.3 R-code

```
#Notations are based on the article of Zhang & Davidian

momentGen=function(density, k){
#Generate kth moment of Exponential or Normal
moment.normale = function(k){#Compute kth moment of a
                             #Standard Normal distr.
  res=gamma((2*k)+1) / ((2^k)* gamma(k+1))
  res
 }
```

```
moment.exponential = function(k){#Compute kth moment of an expo. distr.
  res = gamma(k+1)
  res
}
if(density == "normal"){
  if(k%%2==1) mom = 0
  else mom = moment.normale(k/2)
}
if(density == "exponential"){
  mom = moment.exponential(k)
}
return(mom)}

matrix.A = function(density, k){
  vecMom = sapply(c(0:(2*k)),momentGen,density=density)
  A = matrix(0, ncol =k+1 ,nrow = k+1)
for(i in 1:(k+1)){
A[i,] = vecMom[i:(k+i)]
}
A}

vector.C = function(phi, k){
  #----------------------------------------#
  #Compute vector C of spherical coordinate  #
  # s.t. t(C)%*%C = 1                       #
  # Length(phi) = degree polynom = k        #
  # Length(C) = k+1 (for the intercept)     #
  #----------------------------------------#
  if(k==0) C = 1
  else{
  C = c(sin(phi[1]), cos(phi[1]))
  if(k>1){
  for(i in 2:k){
    C = c(C, C[i]*cos(phi[i]))
    C[i] = C[i] * sin(phi[i])
  }}}
  return(C)
}

h_k = function(phi, density, k, x){
  #----------------------------------------#
  # Compute kth degree appoximation SNP     #
  # for base density = Normal, Exponential  #
  #                                         #
```

```
  #----------------------------------------#
  if(k==0) poly_k =1
  else{
A = matrix.A(density,k)
B = chol(A)
C = vector.C(phi, k)
a = solve(B)%*%C
# On the basis of phi, get back to "a"-coordinates
matrix.x=matrix(1,nrow=length(x), ncol=1)
for(i in 1:k){matrix.x= cbind(matrix.x, x^i)}
  #matrix with columns 1, X,X2,...
poly_k = matrix.x%*%a}
 if(density=="normal") {base.density = dnorm}
 else if (density =="exponential") {
base.density = function(x) {exp(-x)}}

 h_k_value = (poly_k)^2 * base.density(x)
 h_k_value
}

f0_k = function(phi, density, k, t, mu, sigma){
 #----------------------------------------#
 #  Compute SNP density                   #
 #  If density == exponential, it is the  #
 #   distribution of exp(epsilon)         #
 #  If density is normal or logistic,     #
 #   it is the distribution of epsilon    #
 #----------------------------------------#
 if(density=='exponential'){
    f_k =(sigma*exp(mu/sigma))^(-1) * t^(1/(sigma)-1) *
      h_k(phi,density,k, (t/exp(mu))^(1/sigma))}
 else {
   f_k = 1/(t*sigma) * h_k(phi, density,k, (log(t)-mu)/sigma)}
  f_k
}

integrale.I = function(density, k, x){
# Compute iterative integrals to estimate survival
# for exponential and normal densities.
  if(density=="normal"){
    if(k==0)   I = 1-pnorm(x)
      else {
    I = 1-pnorm(x)
    I = cbind(I, dnorm(x))
```

```
    for(i in 2:(2*k)){
    I= cbind(I,x^(i-1) * dnorm(x) + (i-1) *I[,i-2+1])}
  }}
  else if(density=='exponential'){
    if(k==0) I =exp(-x)
    else {
    I = as.matrix(c(exp(-x)))
    for(i in 1:(2*k)){
      I= cbind(I,x^(i) *exp(-x)  + i *I[,i-1+1])
    }}}
  else if (density=='userDefined'){
    fct_to_integrate = function(u){
      res = u^k * userDefined.density(u)
      res
    }
    I = integrate(fct_to_integrate,x,Inf)
  }
return(I)}

coeff.I = function(k, a){
  a.mat = a%*%t(a)
  vec = rep(0,(2*k+1))
  for(m in 0:(2*k)){
    for(i in 1:(k+1)){
      for(j in 1: (k+1)){
        if((i-1)+(j-1)==m)  {vec[m+1] = vec[m+1]+ a.mat[i,j]}
      }
    }
  }
  vec
}

S0_k = function(phi, density, k, t,mu,sigma){
  #linear combination of coeff.I and Integrale.I
  A = matrix.A(density,k)
  B = chol(A)
  C = vector.C(phi, k)
  a = solve(B)%*%C
 if (density == "exponential") x = (t/exp(mu))^(1/sigma)
 else x=(log(t)-mu)/sigma
  I = integrale.I(density, k, x)
 D = coeff.I(k,a)
 S0_k = as.matrix(I)%*%as.matrix(D)
 S0_k
```

```
}

expectationZ = function(k,a, density){
  #To get initial Values as in Zhang & Davidian
  coeff = coeff.I(k,a)
  if(density =="normal"){
    vectZ = c(0,1,0,3,0,15,0,105)}
  else if (density =="exponential"){
    vectZ = c(1,2,6,24,120,720,5040,40320)}
  vectZ = vectZ[1:(2*k+1)]
  EZ = coeff%*%vectZ
  EZ
}

varianceZ = function(k,a,density){
  #To get initial Values as in Zhang & Davidian
  coeff = coeff.I(k,a)
  if(density =="normal"){
    vectZ = c(1,0,3,0,15,0,105,0)}
  else if (density=="exponential"){
    vectZ = c(2,6,24,120,720,5040,40320)}
  vectZ = vectZ[1:(2*k+1)]
  VZ = coeff%*%vectZ
  VZ}

initsParam = function(X,Z,L,R,event){
  require(survival)
  time =ifelse(L==0, R, L)
  status = event
  if(any(event==3) || (event==2)){
    time = ifelse(event==3, (L+R)/2, time)
    status = ifelse(event==0, 0 ,1)}
  # TO BE IMPROVED:
  #   if(model == "AFT"){
  #     tmp = survreg(Surv(time, status)~X, dist = 'lognormal')
  #     mu0 = tmp[[1]][1]
  #     sigma0=exp(tmp[[2]][2])
  #     beta0 = tmp$coeff[-1]}
  #   else if (model == "PH"){
  #     fit = coxph(Surv(time, status)~X)
  #     baseline= summary(survfit(fit))
  #     baselineSurv = baseline$surv
  #     baselineTime=baseline$time
  #     ET0= sum( baselineSurv*diff(c(0,baselineTime)))
```

```
  #     ET02 = sum(2*baselineTime*
baselineSurv*diff(c(0,baselineTime)))
  #     mu0 = 2*log(ET0) - 0.5*log(ET02)
  #     sigma0=sqrt(log(ET02)-2*log(ET0))
  #     beta0=fit$coef}
  #   else if (model =="PO"){
  #     tmp = survreg(Surv(time, status)~X, dist = 'loglogistic')
  #     mu0 = tmp[[1]][1]
  #     sigma0=exp(tmp[[2]][2])
  #     beta0 = tmp$coeff[-1]
  #     beta0=beta0/sigma0
  #     sigma0 = pi^2 *sigma0^2 /3}
  #   sigma2 = sigma0^2 / varianceZ(k,a,density)
  #   mu = mu0 - sqrt(sigma2)*expectationZ(k,a,density)
  #
  #   inits.mu = mu
  #   inits.sigma= sqrt(sigma2)
  #   inits.beta = beta0
  inits.mu = 0
  inits.sigma = 1
  inits.beta = lm(log(time)~-1+X)$coeff
  inits.gamma = if(!is.null(Z)) {glm(status~Z[,-1], family ="binomial")$coef}
else c()
  inits.param = c(inits.mu, inits.sigma,inits.beta,inits.gamma)
  inits.param}

lik_Cure_SNP= function(param,density,k,model='AFT',L,R,X,Z,event){
  #compute full likelihood for all censoring and presence of cure.
  X = as.matrix(X)
  n = nrow(X)
  mu = param[1]
  sigma = param[2]
  nX = ncol(X);  beta = param[3:(nX+2)]
  if (is.null(Z)) {nZ=0; puncure = rep(1,n)} else
  {Z = as.matrix(Z); nZ= ncol(Z);  gamma = param[(nX+3):(nX+nZ+2)]
  #puncure
   gammaZ= as.vector(Z%*%gamma)
   puncure = exp(gammaZ)/(1+exp(gammaZ))}
  phi = param[-c(1:(nX+nZ+2))]
  if(model=='AFT'){
    fL_k = exp(-X%*%beta)* f0_k(phi, density, k,
L*exp(-X%*%beta),mu, sigma)
    SL_k = S0_k(phi,density, k, L*exp(-X%*%beta), mu,sigma)
    SR_k = S0_k(phi,density, k, R*exp(-X%*%beta), mu,sigma)
```

```
      }
  else if (model =="PH"){
   SL_k = S0_k(phi, density, k, L, mu, sigma)^(exp(X%*%beta))
   tmp = f0_k(phi, density, k,L,mu, sigma)/S0_k(phi, density, k,L,
mu, sigma)
   fL_k = exp(X%*%beta) * tmp * SL_k
    if(any(event==3)) {
   SR_k = S0_k(phi, density, k, R, mu, sigma)^(exp(X%*%beta))}
    else {
   SR_k = 0}}
  else if (model == "PO"){
    tmp = exp(X%*%beta) + S0_k(phi,density, k, L,
mu,sigma)*(1- exp(X%*%beta))
    fL_k = f0_k(phi, density, k, L,mu,sigma) *
exp(X%*%beta) * (tmp)^(-2)
      SL_k = S0_k(phi, density, k, L,mu, sigma)*tmp^(-1)
    if(any(event==3)){
      SR_k = S0_k(phi, density, k, R,mu, sigma)*tmp^(-1)
    }else SR_k=0
  }
  RC = (log((puncure*SL_k)+(1-puncure)))[event==0]
  EV = (log(puncure*fL_k))[event==1]
  LC = (log(puncure*(1-SR_k)))[event==2]
  IC = (log(puncure*(SL_k - SR_k)))[event==3]
  loglik = sum(EV)+sum(RC)+sum(LC)+sum(IC)
  -loglik
}
```

```
MaxLikSNPCure=function(X,Z,L,R,event,density="normal",
k=0,model="AFT", par_init){
#-------------------------------------------------------------------#
# Function that will compute, for a given density, k, and model,  #
# the MLE for the SNP likelihood.                    #
# The output is the value of the likelihood at its maximum, the   #
# status of convergence and the  corresponding estimates          #
#-------------------------------------------------------------------#
  #Initial Values
  nX = ncol(as.matrix(X))
  nZ = if(is.null(Z)) 0 else ncol(as.matrix(Z))
  obj1 = try(optim(par=par_init,lik_Cure_SNP,density = density,
```

```
          k=k,model=model,L=L,R=R,X=X,Z=Z, event=event,
          method = "BFGS", hessian=T,control = list(maxit=1000)),
silent = T)

  iter=0
  while(is(obj1, "try-error")& iter<60) {
    par_init = par_init+0.05*(-1)^(iter+1)/(10*iter+1)
    obj1 = try(optim(par=par_init,lik_Cure_SNP,density=density,
            k=k,model=model,L=L,R=R,X=X,Z=Z,event=event,
            method = "BFGS", hessian=T,control = list(maxit=1000)),
silent=T)
    iter = iter+1}
  if(iter==60){pr = rep(NA,(2+nX+nZ))
               val = 999999
               statusConv=99
               hessian=NULL}
  else{
    pr = obj1$par
    val = obj1$val
    statusConv=obj1$convergence
    hessian=obj1$hessian
      }
  par_est = pr[1:(2+nX+nZ)]
  phi_est = pr[-(1:(2+nX+nZ))]
  return(list(statusConv=statusConv,val=val,
              par_est= par_est,phi_est = phi_est,hessian=hessian))
}

SNP_Cure = function(X,Z,L,R,event, dens=c("normal", "exponential"),
                    kk= c(0,1,2), modl = c("AFT"),
                    phi_vect = c(-pi/4, 0,pi/4)){
  #-------------------------------------------------------------------#
  # Function that will compute, for each 'density', each 'kk' and    #
  # each 'modl', the MLE for the                                      #
  # SNP likelihood.                                                   #
  # It will compute the AIC, BIC, HQ.     #
  # Output is the density, model, and k chosen by HQ, the            #
  # corresponding AIC, BIC, HQ, et the corresponding estimates.      #
  #-------------------------------------------------------------------#
  X = as.matrix(X);nX = ncol(X)
  if(!is.null(Z)) {Z=as.matrix(Z);nZ=ncol(Z)} else {nZ=0}
  n=nrow(X)
  res_table = NULL;row_names = NULL;Model = NULL
  hessian.list = vector("list", length(dens)*length(kk)*length(modl))
```

```
hl=1
param_init=initsParam(X,Z,L, R, event)
for(i_dens in 1:length(dens)){
  for(i_modl in 1:length(modl)){
   phi_old = NULL
   for(i_kk in 1:length(kk)){
    if(kk[i_kk]==0) {
      par_init = param_init
      res = MaxLikSNPCure(X,Z,L,R,event,dens[i_dens],
                          kk[i_kk], modl[i_modl], par_init)}
    else if(kk[i_kk]>0){
      maxloglik=9999999
      for(iphi in 1:length(phi_vect)){
        par_init = c(param_init, phi_old,phi_vect[iphi])
        res.new = MaxLikSNPCure(X,Z,L,R,event,
              dens[i_dens],kk[i_kk], modl[i_modl], par_init)
        if(res.new$val<maxloglik) {
          res = res.new
          maxloglik = res$val
        }
      }
    }
    val = res$val
    par=res$par_est
    npar = 2+nX+nZ+(i_kk-1)
    AIC= round(2*val + 2*npar *1,5)
    BIC= round(2*val + 2*npar *log(n)/2,5)
    HQ = round(2*val + 2*npar *log(log(n)),5)
    res_table = rbind(res_table, c(i_dens,i_modl,
                  i_kk,res[1],val,AIC,BIC,HQ,par))
    row_names = c(row_names,paste(dens[i_dens],"-",
                        modl[i_modl],"-",kk[i_kk]))
    rownames(res_table) = row_names
    phi_old = res$phi_est
    hessian.list[[hl]] =  res$hessian
    hl=hl+1
   }
  }
  }
colnames(res_table)[1:10] = c("Density","Model","k","Conv",
                      "-loglik","AIC","BIC","HQ","mu","sigma")
return(list(res_table, hessian.list))
}
```

```
SNP_GOF = function(X,Z,L,R,event, dens="normal",
                   kk= c(0,1,2), modl = "AFT",
phi_vect = c(-pi/4, 0,pi/4), fix_index=0,fix_val=1){
  #-------------------------------------------------------------------#
  #
  #     Function that provides a goodness of fit test of 'density':
  #     X: Covariates in latency, WITHOUT intercept
  #     Z: COvariates in incidence, with intercept
  #     L: Left handside of the interval
  #     R: Right handside of the interval
  #     event: indicator of RC or IC
  #     dens: the density we wish to test
  #     kk: the degree of the polynomial to be tested with a
  #     maximum (7 for gof)
  #     modl: AFT or PH
  #     phi_vect: vector of initial value for the coefficients of P(z)
  #     fit_index: index of the parameter that has to be fixes
  #               (ex. sigma to test for exponential)
  #     fix_val: value of the parameter that has to be fixed
  #-------------------------------------------------------------------#
  X = as.matrix(X);nX = ncol(X)
  if(!is.null(Z)) {Z=as.matrix(Z);nZ=ncol(Z)} else {nZ=0}
  n=nrow(X)
  param_init=initsParam(X,Z,L, R, event)
  phi_old = NULL
  i_kk=1
  while(i_kk  < (length(kk)+1)){
      if(kk[i_kk]==0) {
          par_init = param_init
          res = MaxLikSNPCure(X,Z,L,R,event,dens,kk[i_kk],
                    modl, par_init, fix_index,fix_val)
          val0=res[2]
          }
      else if(kk[i_kk]>0){
         maxloglik=9999999
          for(iphi in 1:length(phi_vect)){
            par_init = c(param_init, phi_old,phi_vect[iphi])
            res.new = MaxLikSNPCure(X,Z,L,R,event,dens,
                         kk[i_kk], modl, par_init)
```

```
          if(res.new[2]<maxloglik) {
            res = res.new
            maxloglik = res[2]
          }
         }
        denom = ifelse(fix_index==0,kk[i_kk],kk[i_kk]+1)
         test.val = 2*(-maxloglik+val0)/denom
          if(test.val > 4.18){
          k_stop = kk[i_kk]
          val_stop = test.val
          test = 1
          i_kk= (length(kk)+1)
          estParam = res[-c(1,2)]
          }
          else {
            k_stop=tail(kk,1)
            val_stop=test.val
            test=0
            estParam = res[-c(1,2)]
            phi_old = tail(res, kk[i_kk])
          }

    }
  i_kk=i_kk+1

}
  return(list(test=test, k_stop=k_stop,
              val_test=val_stop,estParam=estParam))

}
```