UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Modeling long-term survivors: comparison of a mixture model with a flexible bimodal model

Author: **Mathilde FOULON**
Supervisor: **Catherine LEGRAND**
Reader: **Ingrid VAN KEILEGOM**
Academic year 2022–2023
Master [120] en statistique, orientation biostatistiques

Acknowledgment

I would like to thank Catherine Legrand, the supervisor of this work, for her guidance, help and insightful discussions throughout this project. I would also like to thank Ingrid Van Keilegom giving up her time to read this work.

I would also like to thank all my friends who supported and assisted me during this work: Nacim, Lara, Manon, Clémence, Clément and Stéphanie for their moral support and Alex for proofreading this master's thesis.

Futhermore, I am deeply grateful to my mother, who has supported me throughout my academic journey. I also have a thought for my dad who motivates me from the stars.

Contents

1 Introduction **4**

2 Reminder of Survival Analysis and Cure Models **6**
2.1 Survival analysis 6
2.2 Cure Models 12
2.2.1 Introduction to cure models 12
2.2.2 Mixture cure models 14
2.2.3 Model regression 14
2.2.4 Model estimation 15
2.2.5 Discussion. 15

3 3-component mixture cure model **18**
3.1 Classical Finite Mixture model 18
3.2 Survival finite mixture model 24
3.3 3-component mixture cure model 29

4 Exponentiated Log-Sinh Cauchy cure model **33**
4.1 Exponentiated log-sinh Cauchy function 33
4.2 Mixture cure ELSC model 34
4.2.1 Regression model 37
4.2.2 Likelihood maximization 38

5 Simulation **41**
5.1 3-component mixture cure model simulation 41
5.1.1 Simulation methodology 41
5.1.2 Estimation methodology 43
5.1.3 Initialization methodology 45
5.1.4 Monte Carlo Experiments 45
5.1.4.1 Simulation study 1 46
5.1.4.2 Simulation study 2 49
5.2 Mixture cure ELSC model simulation 55
5.2.1 Simulation methodology 55
5.2.2 Estimation methodology 55

5.2.3 Initialization methodology 55
5.2.4 Monte Carlo Experiments 55
5.2.4.1 Simulation study 1 56
5.2.4.2 Simulation study 2 59
5.3 Simulation with random censoring 63
5.3.1 3-component mixture cure model without covariates 63
5.3.2 ELSCcr model without covariates 66
5.3.3 ELSCcr model with covariates 68
5.4 Discussion 70

6 Application to real data **72**
6.1 Calving data 72

7 Conclusion **78**
.1 Appendix 83
.1.1 3-component mixture cure model simulation : simulation study 1 . 83
.1.2 3-component mixture cure model simulation : simulation study 2 . 86
.1.3 Mixture cure ELSC model simulation : simulation study 1 92
.1.4 Mixture cure ELSC model simulation : simulation study 2 95
.2 Simulation with random censoring 100
.2.1 3-component mixture cure model : without covariates 100
.2.2 ELSCcr model: without covariates 101
.2.3 ELSCcr model: with covariates 104

Chapter 1

Introduction

Classical survival analysis considers that all individuals in the study are likely to experience the event of interest if the follow-up is long enough. However, there are many situations in which some of these individuals will never experience the event of interest, and are statistically cured. For example, thanks to medical advances in the field of oncology, it's reasonable to believe that, if we're interested in cancer relapse, it's possible that a proportion of individuals will never experience this event. Similarly, if we are interested in cancer-related death, we can hope that for certain cancers a fraction of patients will be long term survivors who can be considered statistically cured. In another context, if we look at the age at which women become pregnant for the first time, it's possible that some of them will never have children. This part of the cured population can be taken into account in the survival analysis using cure models. One of the most popular families of models is the mixture cure model, originally discussed by Farewell [12], Boag [3] and Berkson & Gage [15]. These models see the population as a mixture of cured and uncured individuals.

It is essential to have a sufficiently long follow-up time to see that part of the population will indeed not observe the event of interest after a first wave of events, and that there is a cured part of the population. However, it is also possible that, in the very long term, a second wave of events will affect one part of the population. It's also possible that, after this second wave, one part of the population will finally be considered cured. This scenario occurs, for example, when breast cancer relapses. In fact, some patients will experience a relapse soon after their treatment, while others will experience a relapse much later, and some will simply not experience a relapse at all. Another example is when we study the risk of customers' first default on their credit. Some customers will default relatively quickly, others will default after a few years, and some will not. The data in these examples therefore undergo two waves of events and have a bimodal distribution.

Unfortunately, conventional mixture cure models are unable to model these bimodal data correctly. Two different models have been proposed in the literature to solve this particular issue. The first, proposed by Ramires and al. [30], is the exponentiated log-sinh Cauchy cure rate (ELSCcr) model, which is a mixture cure model in which the distribution for the uncured part of the population has an exponentiated log-sinh Cauchy distribution. This distribution is very flexible and can be used to model bimodal data. The second model

is proposed by Hunsberger and al [22] and Alves and Dias [5], the 3-component mixture cure model. This model considers a heterogeneous population. It is then a mixture of sub-populations, each with its own specific distribution.

The objective of this work is to analyze whether the two models can correctly model bimodal data, and then to compare them. To meet this objective, the structure of the work is as follows:

Chapter 2 provides an overview of the main features of survival analysis. Section 2.1 presents the different particularities of classical survival analysis. Section 2.2 focuses on the cure model, in particular the mixture cure model.

Chapter 3 delves into the 3-component mixture cure model. The first section provides an explanation of the classical finite mixture component model proposed by McLachlan and Peel [13]. In this section, we introduce the model and its key characteristics. Additionally, we present the EM algorithm employed for parameter estimation. The second section focuses on the finite mixture cure model in the context of survival analysis. We discuss the essential aspects of this model. Finally, the third and final section presents the 3-component mixture cure model. We explain the method used to estimate the value of the parameters of the 3-component mixture cure distribution by maximum likelihood.

Chapter 4 introduces the ELSCcr model and its various features. The first section provides a presentation of the ELSC distribution and its associated functions. Developing from this foundation, the second section, explores the integration of this distribution within the framework of the mixture cure model. To conclude this chapter, we present the methodology employed for estimating the value of the parameters of ELSCcr distribution, we also propose a way of accounting for covariates in this model.

Chapter 5 presents the simulation study conducted to evaluate the performance of the two models. The first section focuses on the simulation study for the 3-component mixture cure, considering only administrative censoring. We will present the methodology employed for the data simulation and the function developed for this purpose. Additionally, we will then explain the function created to estimate the value of the parameters of this model using the EM algorithm, along with the method of parameter initialization used in the model. Finally, we present the outcomes of the Monte Carlo experiment conducted both with and without covariates. The second part of this chapter will present the simulation study for the ELSCcr model, considering only administrative censoring. We will outline the methodology employed to simulate the data and the estimation method used to determine the value of model parameters. Following this, we will present the results obtained from Monte Carlo experiments, again with and without the presence of covariates. In the penultimate section of the chapter, we will perform a simulation study for each model, this time incorporating random censoring. To conclude the chapter, a brief discussion of the results and a summary of the models will be provided.

Chapter 6, the ELSCcr model and the 3-component mixture cure model, will be applied to real data. Firstly, to see if both models fit the data correctly, and secondly to compare the models and see if one fits better than the other.

We will conclude this work with a conclusion, providing a summary of the work and a discussion of future research directions.

Chapter 2

Reminder of Survival Analysis and Cure Models

In the first section of this chapter we will quickly recall what survival analysis is, its characteristics and the different functions related to it. Then we will talk about the likelihood function in the context of fully parametric classical survival analysis and its maximization in order to determine the value of the different parameters. Finally, we will quickly discuss a way to incorporate the effect of covariates into the different models.

The second section of this chapter gives a quick introduction to the mixture cure models. We will define this model and explain its various particularities. This family of survival models is the one on which the 3-component mixture cure model and the Exponentiated Log-Sinh Cauchy cure model are based.

This chapter is based on the work of Catherine Legrand [21].

2.1 Survival analysis

Survival data are concerned with the time to the occurrence of a so-called event of interest from the start of follow-up. Despite the name survival analysis, the event of interest is not necessarily death and can be a positive or negative event depending on the context. The objective of survival analysis is to analyze the time required to go from an "initial" state to a "final" state. The main feature of interest is therefore a positive random variable T which represents the actual time until the event of interest occurs. For example, the time it takes for a young graduate to find his or her first job, the time until a cancer recurs or the time until the death of a person with a serious illness. Survival analysis has been covered in many books such as [6, 7, 16, 17].

As usual, we will denote $F(t) = P(T \leq t)$ and $f(t) = \frac{d}{dt}F(t)$ respectively the distribution and density functions of the random variable T. However, in survival analysis, the following three functions are more commonly used:

Survival function $S(t)$: defined as the probability that an individual will survive beyond time t, it is a decreasing function that takes values in the interval $[0, 1]$. We have the

following relation with the functions defined previously,

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= P(T > t) \end{aligned} \tag{2.1.1}$$

$$f(t) = -\frac{d}{dt}S(t) \tag{2.1.2}$$

and the survival function is said to be proper when $\lim_{t\to+\infty} S(t) = 0$.

Hazard function $h(t)$: is defined as the instantaneous risk to observe an event just after time t given that the event has not yet been observed at time t ,

$$\begin{aligned} h(t) = & \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}) \\ = & \frac{f(t)}{S(t)} = \frac{-d \ \log S(t)}{dt} \end{aligned} \tag{2.1.3}$$

$h(t)$ is a positive function.

Cumulative hazard function $H(t)$: one can understand this function as being the expected number of events to be observed by time t. It is an increasing function, taking values in $[0, +\infty]$ and is defined as

$$H(t) = \int_0^\infty h(u)du \tag{2.1.4}$$

and we have

$$S(t) = \exp(-\int_0^t h(u)du) = \exp(-H(t)) \tag{2.1.5}$$

We can see that knowing only one of the functions that characterize T allows us to determine the others.

One difficulty with survival analysis is that one has to take into account the fact that part of the data may be censored and/or truncated. In this work, we will limit ourselves to taking censoring into account. Having censored data means that for some individuals, the time to the event of interest is not known exactly. We are particularly interested in right censoring, which means that only a lower bound of the actual survival time is available. This can happen when, at the last follow-up time, the individual has still not experienced the event of interest. This type of phenomenon is referred to as an administrative censoring. It is also possible that an individual stops the study before the end of this one. It is called the loss to follow-up. It is also a type of censoring on the right. In the case of administrative censoring, there will be fixed censoring, while for loss to follow-up, there will be random censoring. For the latter, the censoring is then a positive random variable C that follows the distribution function $G(.)$. It is assumed that T and C are independent and that C is uninformative. This means that, the distribution of C does not depend on the parameters of the distribution of T. Figure 2.1.1 represents the different possibilities

that can be obtained when right-censoring is observed. *Start* represents the beginning of the study. $T_{\max}$ represents the end of the study. Individual A is right-censored because loss to follow-up in the course of the study, we do not know whether he/she will suffer the event of interest (lost to follow-up). Individual B has experienced the event of interest, so he/she is not censored. Finally, individual C is also right-censored because at the time $T_{\max}$ he/she has not yet experienced the event of interest. So in reality, the observed data in the sample

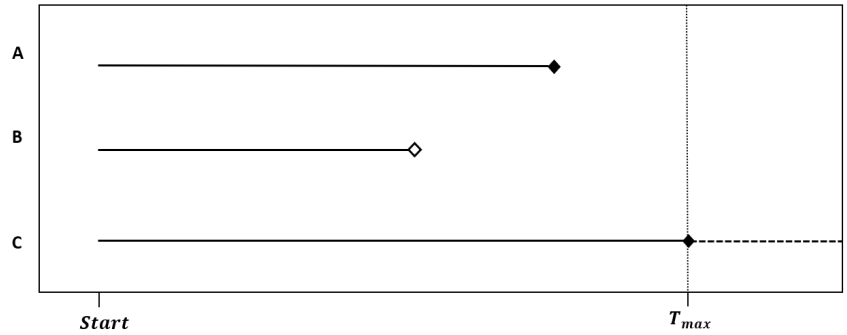

Figure 2.1.1: Representation of the time-to-event in a hypothetical study. Dark diamonds indicate censoring of the patient. The white diamond represents the observation of the event of interest .

when the right censoring is taken into account are:

- Observed times : $W = \min(T, C)$,where W represents the time either to the event or to the censoring, but not both.
- The censoring indicator δ which indicates whether an individual i is censored or not: $\delta = I(T \leq C)$

The censoring indicator therefore tells us whether (takes the value 0) or not (takes the value 1) the individual is censored. From the observed data W and δ, we would like to determine the main characteristics of the distribution of the actual survival time T. In a parametric context, the idea is to assume a distribution for T and use the available data to estimate the value of the different parameters of the distribution. We can determine the value of the parameters of the distribution by the maximum likelihood method. However, it is necessary to take into account the contribution of censored data in addition to uncensored data in the likelihood. So let us suppose a sample of n individuals with for each of them the observed data $w_i = \min(t_i, c_i)$, $i = 1, ..., n$, where w_i is the observed event time, c_i is the censoring time-to-event, t_i the real time-to-event and $\delta_i = I(t_i \leq c_i)$ the censoring indicator of the ith individual. Considering that the different observations are independent, and that the censoring is independent and not informative, the likelihood function is,

$$L \propto \prod_{i=1}^{n} (f(w_i))^{\delta_i} (S(w_i))^{1-\delta_i} \tag{2.1.6}$$

This work focuses on parametric distributions for the variable T. In the literature, we can find various proposals for the distribution of the random variable T. The distributions used in this can be for example the Exponential, Gompertz, Weibull and the Exponentiated log-sinh Cauchy (ELSC). The latter will be explained in Chapter 4. To estimate the different parameters, we utilize the appropriate density and survival functions within the likelihood function and maximize the latter with respect to the parameters of interest. Moreover, the variance of these estimators can be obtained from the Fisher matrix for the observed data. It is then possible to use the asymptotic normality property of the MLEs to obtain different hypothesis tests and confidence intervals. Additionally, the covariate effects can also be taken into account using regression models.

Although we focus on a parametric distribution of T, we also quickly address non-parametric estimation. When there is no censoring, one can use the empirical estimator of the distribution. This is given by

$$\begin{aligned}\hat{S}^{emp}(t) &= \frac{\text{Number of observations with time-to-event } \geq t}{\text{Number of observations in the data set}} \\ &= 1 - \hat{F}^{emp}(t)\end{aligned}$$

However, as already mentioned, in the context of survival analysis, censoring must be taken into account. Two estimators are proposed for this, the Nelson-Aalen estimator [1, 24] and the Kaplan-Meier estimator [9]. The latter is the most popular, and will be employed it in this work. Assume n observations with the observed times $w_1, ..., w_n$ and their associated censoring indicators $\delta_1, ...\delta_n$. Then there are r distinct event times with $r \leq n$. The ordered sample of event times is $w_{(1)}, ..., w_{(r)}$ and the number of events that occur at these different times are noted $d_{(1)}, ..., d_{(r)}$. We also have the size of the risk set $R(w_{(j)})$ at event time $w_{(j)}$. The latter can be defined as the number of observations that have not yet experienced right censoring or event of interest before $w_{(j)}$. The probability of surviving beyond time t (within the interval $[w_{(j)}, w_{(j+1)}[$) is the product of the probabilities of having survived beyond all previous time intervals:

$$S(t) = P(T > w_{(1)}|T > w_{(0)}) \times P(T > w_{(2)}|T > w_{(1)}) \times ... \times P(T > w_{(j)}|T > w_{(j+1)})$$

The probability of surviving this interval can be written as

$$1 - \frac{d_{(j)}}{R(w_{(j)})} = \frac{R(w_{(j)}) - d_{(j)}}{R(w_{(j)})}$$

The Kaplan-Meier estimator of the survival function is given by

$$\hat{S}^{KM}(t) = \prod_{j:w_{(j)} \leq t} \frac{R(w_{(j)}) - d_{(j)}}{R(w_{(j)})} \tag{2.1.7}$$

with $\hat{S}^{KM}(t) = 1$ for $t < w_{(1)}$. This estimator is a decreasing step function, with jumps at each event realization. If the last observation is censored, then it is impossible for

the estimator to reach zero. It is not possible to estimate $S(t)$ consistently beyond this last observation. The variance of the Kaplan-Meier estimator can be estimated by the Greenwood formula, given by

$$\hat{V}_{asymptotic}(\hat{S}^{KM}(t)) = (\hat{S}(t))^2 \sum_{j:w_{(j)} \leq t} \frac{d_{(j)}}{R(w_{(j)})(R(w_{(j)}) - d_{(j)})} \tag{2.1.8}$$

$\hat{S}^{KM}(t)$ is asymptotically normal:

$$\frac{\hat{S}^{KM}(t) - S(t)}{\sqrt{\hat{V}(\hat{S}^{KM}(t))}} \xrightarrow{d} N(0,1) \tag{2.1.9}$$

From the asymptotic normality we can obtain the $(100-\alpha)\%$ confidence interval of $\hat{S}^{KM}$. This is given by

$$\hat{S}^{KM}(t) \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{S}^{KM}(t))} \tag{2.1.10}$$

The log-log transformation is often preferred in order to have only points in $[0,1]$

$$\hat{S}^{KM}(t)^{\exp[\pm z_{\alpha/2}\sqrt{\hat{V}(\log(-\log \hat{S}^{KM}(t)))}\ ]} \tag{2.1.11}$$

Let us now discuss the regression model for the survival data. It is indeed important to be able to take covariates into account in the survival model. Two families are mainly used: the accelerated failure time (AFT) and the proportional hazards model. First, let's talk about the latter. This model considers that the risk function for a given individual and an associated set of covariates can be written as the product of a risk function common to all individuals $h_0(t)$ and a function dependent on the covariates. Let $\mathbf{X}$ be a set of covariates and n individuals who have as associated covariates $\mathbf{x_i}^\top = (x_{i1}, ..., x_{ip})$ with $i = 1, ..., n$ and p the number of covariates, then the proportional hazard model is given by

$$h_i(t) = h(t|\mathbf{x_i}) = h_0(t)\exp(\boldsymbol{\beta}^\top \mathbf{x_i}) \tag{2.1.12}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unobserved parameters associated to the covariates. Taking the logarithm of (2.1.12), we get

$$\log\ h_i(t) = \log\ h_0(t) + \boldsymbol{\beta}^\top \mathbf{x_1} \tag{2.1.13}$$

We can then interpret the effect of covariates on the log-hazard as for a classical linear model. The proportional hazards (PH) model takes its name directly from the hypothesis that the hazards are proportional, i.e. if we have two individuals with respective the covariates $\mathbf{x_i}$ and $\mathbf{x_j}$ the ratio of their hazards is constant over time and we have

$$\frac{h(t|\mathbf{x_i})}{h(t|\mathbf{x_j})} = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x_i})}{\exp(\boldsymbol{\beta}^\top \mathbf{x_j})} \tag{2.1.14}$$

This relationship is independent of time. The time dependence is captured in the baseline hazard function which is common to all individuals. The PH model can also be written as a survival function by the relation (2.1.5) ,

$$S_i(t) = S(t|\mathbf{x_i}) = S_0(t)^{\exp(\boldsymbol{\beta}^\top \mathbf{x_i})} \tag{2.1.15}$$

where $S_0(t) = \exp(-\int_0^t h_0(u)du)$ the baseline survival function. The baseline hazard and survival function can be chosen such that T follows a parametric distribution, the model is then a fully parametric PH model. The estimation of the parameters is done by plugging the survival function (2.1.5) and the corresponding density function in the likelihood function (2.1.6) and maximizing. However, the baseline hazard function can be left unspecified, which corresponds to the so-called "semi-parametric Cox PH model". In this last case we will maximize the partial likelihood (2.1.16) function to obtain the estimation of $\boldsymbol{\beta}$. Let $\mathbf{x_{(i)}}$ and $R(w_{(i)})$, respectively the covariates vector and the risk set for the subject with i^{th} event time,

$$L_{partial}(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x_{(i)}})}{\sum_{l \in R(w_{(i)})} \exp(\boldsymbol{\beta}^\top \mathbf{x_l})} \tag{2.1.16}$$

Let's talk briefly about the second model taking into account covariates: the AFT model. Under its survival representation, this one is defined as

$$S_i(t) = S_0(\exp(\boldsymbol{\beta}^\top \mathbf{x_i})t) \tag{2.1.17}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is the vector of regression coefficients, S_0 is a baseline survival function belonging to a parametric distribution. $\exp(\boldsymbol{\beta}^\top \mathbf{x_i}) > 0$ is an acceleration factor. We observe that the covariates have a multiplicative effect on the time which implies an acceleration (or deceleration) the time scale. We can also write this model in its log-linear form

$$\log\ T = \mu + \boldsymbol{\alpha}^\top \mathbf{X} + \sigma\epsilon \tag{2.1.18}$$

where μ is the location parameter and σ is the scale parameter, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)$ is the vector of covariates parameters and ϵ is the error term. In this work, we will use the parametric Weibull distribution with

$$f(t) = \frac{\gamma}{\lambda}(\frac{t}{\lambda})^{\gamma-1} e^{-(t/\lambda)^\gamma} \tag{2.1.19}$$

$$S(t) = e^{-(t/\lambda)^\gamma} \tag{2.1.20}$$

with $t \geq 0$, $\gamma > 0$ the shape parameter and $\lambda > 0$ the scale parameter. This distribution is a special case; indeed its an AFT model when ϵ follows a Gumbel distribution, is equivalent to the Weibull PH model, but they use a different parameterization :

$$S_{i,Weibull\ PH\ model} = \exp(-\exp(\boldsymbol{\beta}^\top \mathbf{x_i})(\frac{t}{\lambda})^\gamma) \tag{2.1.21}$$

$$S_{i,AFT} = \exp(-\exp(\frac{\log t - \mu - \boldsymbol{\alpha}^\top \mathbf{x_i}}{\sigma})) \tag{2.1.22}$$

and we have

$$\begin{aligned} \lambda &= \exp(\mu) \\ \gamma &= 1/\sigma \\ \boldsymbol{\beta} &= -\boldsymbol{\alpha}/\sigma \end{aligned} \tag{2.1.23}$$

We will use all these results later when estimating the parameters of the 3-component mixture cure model.

2.2 Cure Models

In this section, we will present the main aspects of cure models. This model is explained in detail in the book written by Maller and Zhou [27]. We will first define what a cure model is, when to use it and explain its particularities. In a second step, we will present one of the possible families of cure models: the mixture cure models. The latter was originally discussed by Farewell [12], Boag [3] and Berkson & Gage [15]. This model family is currently the most popular model for the analysis of data with a cured fraction. It will be the basis for the future development of the models we will introduce in the following chapters.

2.2.1 Introduction to cure models

In classical survival analysis, it is assumed that each individual will experience the event of interest at some point. However, in reality, this assumption is not necessarily true. Indeed, a fraction of the population may never experience the event of interest. For example, we may be interested in the time until cancer recurrence in patients who have already had cancer. A part of the population will experience recurrence, but fortunately a second part of the population will not experience cancer anymore and will be considered cured. The proportion of people who are cured is in itself interesting, in addition to the survival time of those who are not. In this case, we consider that this part of the population is cured or not susceptible, and for this fraction of individuals we have $T = \infty$. These are called long-term surviving individuals.

There exists different possible situations, which we have represented on Figure 2.2.1. We can see that individual A is censored during the study, but will actually not experience the event of interest after the censoring. Individual B is censored at the end of the study, but will in fact experience the event of interest shortly afterwards. Individual C experiences the event of interest during the study. Finally, individual D is censored at the end of the study and will in fact not experience the event of interest.

The fact of having a part of the population that will never experience the event of interest will have an impact on the survival function. This means that when t goes to infinity, the survival function will no longer be zero but will stabilize at a level which

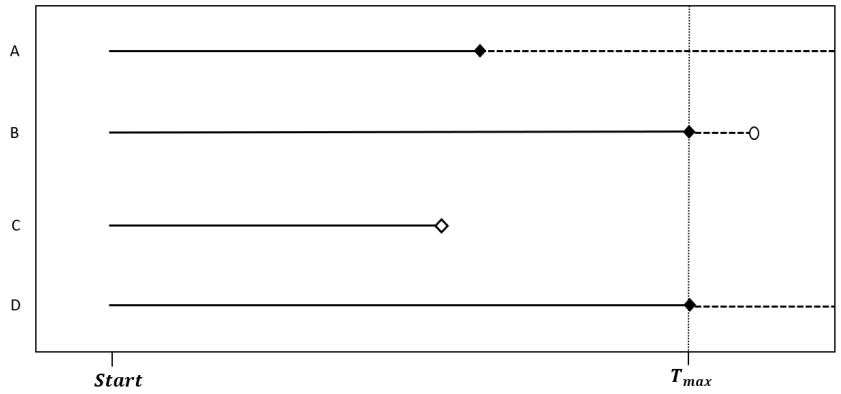

Figure 2.2.1: Representation of the time-to-event in a hypothetical study. Dark diamonds indicate censoring of the patient. The white diamond represents the observation of the event of interest. The white circle represents the observation of the event of interest after the end of the study.

corresponds to the cure rate, denoted $1 - \pi$. We therefore have

$$\lim_{t \to \infty} S(t) = 1 - \pi > 0$$

The survival function is therefore said to be improper and the value towards which the survival function tends is $1 - \pi$ corresponds to the proportion of cured persons.

At the end of the study, the cured individuals are all censored (administrative censoring). However, in the previous section we considered that the censored individuals maintained the same behavior as the uncensored individuals. It can be understood that this assumption is no longer correct in the context of the cure model. Indeed, some censored individuals can indeed observe the event of interest after censoring (like individual B on Figure 2.2.1). But others will have a time $T = \infty$ and will never experience the event, thus are cured (like individual A and D on Figure 2.2.1).

An example of a survival curve for a population with a cure fraction is shown on Figure 2.2.2 for $T \sim Exp(\lambda)$ where $\lambda = 0.4$ and $\pi = 0.6$. It can be seen that the survival function does not reach 0, but reaches what we will call a plateau. This visualization of the plateau, when we look at an estimated survival curve and when the follow-up time is long enough, will allow us to say that we are in the presence of a cure fraction. For a given data set, we can therefore observe if there is a part of the population that can be considered as cured by computing the Kaplan-Meier estimator and observing this long plateau with a large amount of right censored data. However, we will see later that if the follow-up time is not long enough, we might miss the observation that some individuals might experience, after this plateau, the event of interest. We will call this a second wave of events. It is in this situation that bimodality in the survivor density function will be observed.

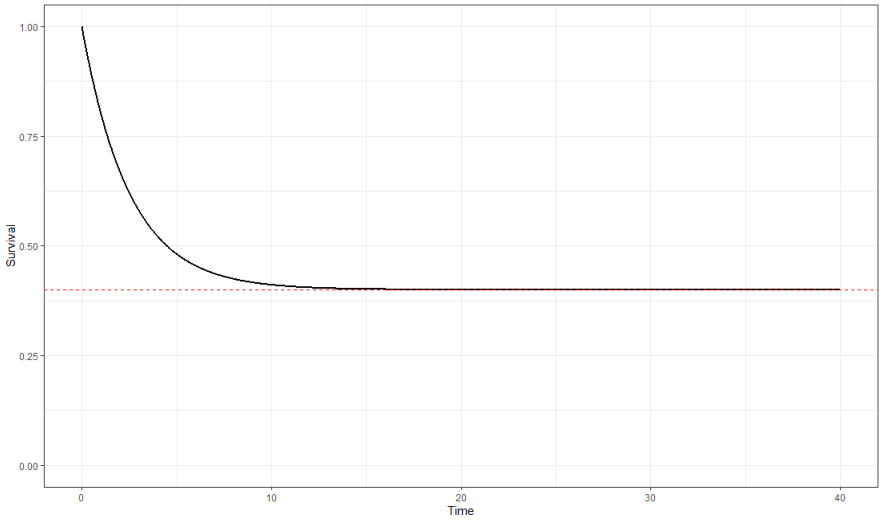

Figure 2.2.2: Survival function for $T \sim Exp(0.4)$ and $\pi = 0.6$.

2.2.2 Mixture cure models

The proportion of cured people is very important in the cure model and is a parameter of interest in the mixture cure model. In the presence of a cure fraction, it can be considered that the population can be divided into a cured (non-susceptible to experience the event of interest) and an uncured (susceptible to experience the event of interest) part. The population is then said to be heterogeneous. In fact, we have two groups with different risks. To manage the heterogeneity of the population, the mixture cure model was introduced by [12], [3], and [15].

The mixture cure model uses two sub-models. A first model for the incidence which aims to determine the group (cured or not) in which the individual is located. The second model is the latency model which models the survival function of uncured individuals. We consider a variable B which represents the cure status, with $B = 1$ if the individual is not cured, and $B = 0$ if the individual is cured. B is only partially observed due to the censored observations. The probability of being susceptible can be defined as $\pi = P(B = 1)$ and we have the survival function of susceptible population $S_u(t) = P(T > t|B = 1)$ and for the cured population we have $S_c(t) = P(T > t|B = 0) = 1, \ \forall t$. The survival function of T for the whole population is given by

$$S_{pop}(t) = P(t > t) = (1 - \pi) + \pi S_u(t) \tag{2.2.1}$$

It can be seen that despite the fact that the survival function for the population is improper, the survival function for uncured individuals is proper.

2.2.3 Model regression

The mixture cure model can be adapted to consider the impact of the covariates on the latency model and on the incidence model . For example, in the medical context, the use of

a treatment, or not, may have an impact on the survival of the patient. But, this treatment may also impact on the likelihood of being susceptible. Covariates may be common to both incidence and latency models, some may be specific to one of the sub-models. The survival function in the presence of covariate is noted:

$$S_{pop}(t|\mathbf{X}, \mathbf{Z}) = (1 - \pi(\mathbf{X})) + \pi(\mathbf{X})S_u(t|\mathbf{Z}) \tag{2.2.2}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the covariate vectors of the incidence model and the latency model respectively. There are various model choices in the literature for the two sub-models. For example, for the incidence model, we can use the logistic regression model. For the latency model, we can take fully, semi, or non-parametric models. A common choice is to use the parametric PH model. It should be noted, however, that the assumption of proportional hazards does not apply at the population level, but at the level of the susceptible part of the population.

2.2.4 Model estimation

The estimation of the mixture cure model in a fully parametric context is performed following the same idea as the classical survival model. The likelihood function defined in equation (2.1.6) is maximized while taking into account the two subpopulations (cured and uncured). We have the likelihood function

$$L = \prod_{i=1}^{n} [\pi(\mathbf{x_i}) f_u(w_i|\mathbf{z_i})]^{\delta_i} \times \prod_{i=1}^{n} [1 - \pi(\mathbf{x_i}) + \pi(\mathbf{x_i}) S_u(w_i|\mathbf{z_i})]^{1-\delta_i} \tag{2.2.3}$$

for $i = 1, ..., n$ individuals and we have i.i.d observed data $\mathbf{y_i} = (w_i, \delta_i, \mathbf{x_i}, \mathbf{z_i})$ where $\mathbf{x_i}$ and $\mathbf{z_i}$ are respectively the set of covariates that will influence the latency model and the model for individual i. As we consider a fully parametric model, we can simply maximize the likelihood function. For this we can use numerical optimization using for example Newton-Raphson algorithm.

2.2.5 Discussion.

As we have seen in this section, the cure models are to be applied when we observe a long plateau on the graph of the estimated survival curve. However, this long plateau can be misleading. Indeed, it could be that a part of the population is cured, but it could also be that after a certain time the individuals will finally experience the event of interest. We could call this a second wave of events. We can visualize this kind of situation on Figures 2.2.3 and 2.2.4, for the same data set. Indeed, when considering a follow-up time of $t_{follow-up} = 60$ (Figure 2.2.3), one could believe that the population contains a proportion of cured people. However, if one chooses a follow-up time of $t_{follow-up} = 150$ (Figure 2.2.4), one can see that there is, in fact, a second wave of realization of the event of interest. This kind of situation underlines the importance of the choice of the follow-up time. For example, when looking at the time until recurrence of breast cancer, it is well known that

some women may experience a recurrence on a relatively short time scale, while others may experience the event years later. By selecting a follow-up time of 5 years, we risk missing events of women who could experience a recurrence 10 or 15 years later.

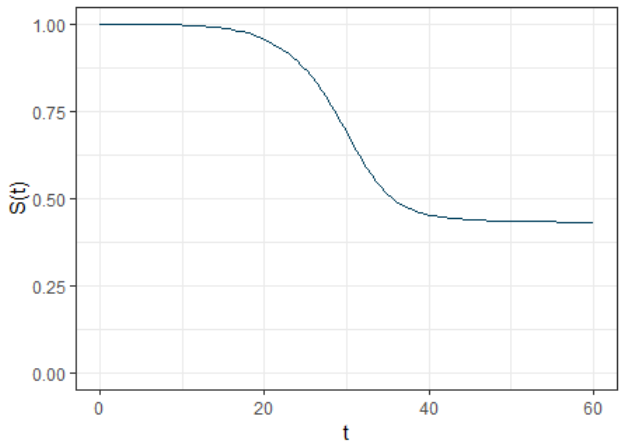

Figure 2.2.3: Survival function with $t_{followup} = 60$.

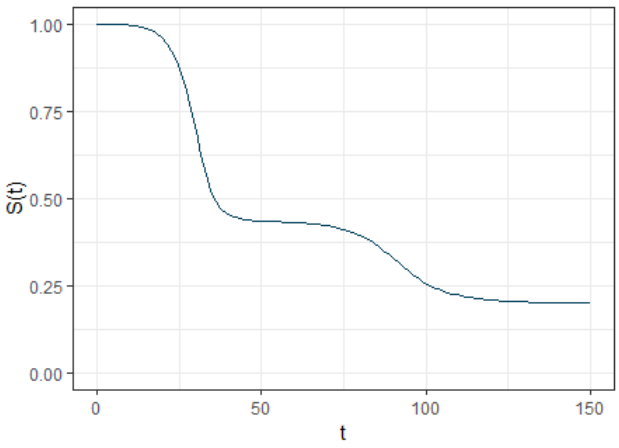

Figure 2.2.4: Survival function with $t_{followup} = 150$.

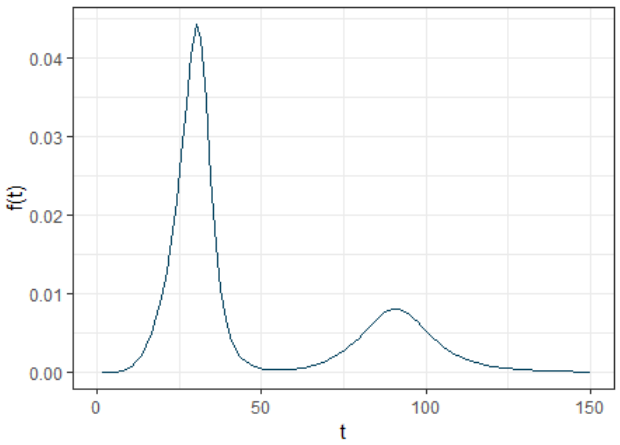

Figure 2.2.5: Density function of data with two event waves.

The density and also the hazard function of this type of data will have a bimodal form, as shown on Figure 2.2.5. In such situation, it is necessary to have an adapted model that will take into account this bimodality, which the classical cure models do not consider. It would therefore be necessary to determine a model that takes into account:

- Individuals who would experience an event relatively early during the follow-up period (first wave).
- Individuals who would experience an event relatively late during the follow-up period (second wave).

- Individuals who would not experience an event (cured) over the follow-up period

In the following chapters of this work, we will propose two distinct models able of taking these different elements into account. We will define them, explain how to estimate these two models and then compare them first on simulated data sets and then on real data.

Chapter 3

3-component mixture cure model

In this chapter we will introduce a first model allowing to take into account a heterogeneous population experiencing the event of interest at different time scales and taking into account a part of cured individuals. It is proposed to use the framework of finite mixture models [13, 5, 14] which considers a heterogeneous population. The mixture cure model, presented in Section 2.2.2, is a special case of the mixture survival model. Indeed, in this model, we had a heterogeneous population composed of 2 groups: cured and uncured. In order to capture the bimodality of the data and the fraction of non-susceptible people, we extend this model to a population composed of 3 groups: cured or not susceptible, likely to observe the event in the short term and likely to observe the event in the long term. The risk of experiencing the event of interest is specific to each population group.

In the first part of this chapter we will recall the classical finite mixture models, for more details on these models we refer to McLachlan and Peel [13]. The second section of this chapter is dedicated to the use of mixture models in survival analysis. This section is based on the work of McLachlan, McGiffin [14] and Alves, Dias [5]. The last section of this chapter focuses on the cure model context with bimodality of data. To do this we will use a 3-component mixture cure model. The specificity of this model is the consideration that one of the components of the population is considered as non susceptible which previous models did not do.

3.1 Classical Finite Mixture model

Mixture models were introduced by McLachlan and Peel [13]. We rely on their work, for all explanations and calculations in this section. The mixture models take into account a mixture of different distributions, which makes these models very flexible. Mixture models can be used when one has a heterogeneous population whose distribution cannot a priori be easily modelled in a parametric way. For example, when a population is composed of different age groups or when a population is taking different treatments and these different groups will not have the same distribution function for the event. We can then see the distribution of the population as composed of these different distributions specific to each

of the groups. This type of model can be useful when one needs to model data that have a skewed or/and asymmetric distribution.

Let $\mathbf{Y} = (\mathbf{Y_1^\top}, ..., \mathbf{Y_n^\top})^\top$ be the n i.i.d. random sample where $\mathbf{Y_i}$ for $i = 1, ..., n$ is a p-dimensional random vector and $\mathbf{y_i}$ is the observed value of $\mathbf{Y_i}$. The probability function associated with $\mathbf{Y_i}$ is $f(\mathbf{y_i})$ on $\mathbb{R}^p$. The latter function can be written as a weighted sum of the contribution of different densities. We have

$$f(\mathbf{y_i}) = \sum_{g=1}^{G} \pi_G f_g(\mathbf{y_i}) \tag{3.1.1}$$

where $f_g(\mathbf{y_i})$ are called the component densities of the mixture and $\pi_1, ..., \pi_G$ are the mixing proportions or weights such as

$$0 \leq \pi_g \leq 1 \qquad (g = 1, ..., G)$$

and

$$\sum_{g=1}^{G} \pi_g = 1$$

$f(\mathbf{y_i})$ is the mixture density and G is the number of component. We have

$$\mathbf{Y_1}, ..., \mathbf{Y_n} \overset{\text{i.i.d}}{\sim} F$$

where $F(\mathbf{y_i})$ is the distribution corresponding to the mixture density $f(\mathbf{y_i})$.

Associated with random vector $\mathbf{Y_i}$, there exists a random vector $\mathbf{B_i}$ of size G such that the gth element $(g = 1, ..., G)$ of $\mathbf{B_i}$ is denoted $B_{gi} = (\mathbf{B_i})_g$ and is equal to one if $\mathbf{Y_i}$, comes from the g-component distribution and zero if not. It's assume for independent data that $\mathbf{B_i}$ are distributed unconditionally as

$$\mathbf{B_i} \sim Mult_G(1, \boldsymbol{\pi}) \tag{3.1.2}$$

a multinomial distribution with G categories and the probabilities associated $\boldsymbol{\pi} = (\pi_1, ..., \pi_G)$ such as

$$P(\mathbf{B_i} = \mathbf{b_i}) = \pi_1^{b_{1i}} \pi_2^{b_{2i}} ... \pi_G^{b_{Gi}}$$

This random vector $\mathbf{B}$ can be observed. For example in the case where the distributions are different according to the age of the individuals. If the age of the individual is available, it is then possible to know in which group the individual is. In this first case, we consider that the data are complete

$$\mathbf{y_c} = (\mathbf{y}^\top, \mathbf{b}^\top)^\top$$

where $\mathbf{y}$ is the observed-data and $\mathbf{b} = (\mathbf{b_1^\top}, ..., \mathbf{b_n^\top})$ is the group indicator. But in many cases, the information of the membership of an individual to a group is not observed. The data is then said to be incomplete and we only have $\mathbf{y}$. The vector of probabilities $\boldsymbol{\pi}$ which is also the parameter vector of the multinomial is then not known. McLachlan and Peel [13] consider the vector of probabilities $\boldsymbol{\pi}$ as a prior probabilities vector belonging to the

different groups. The posterior probability of belonging to the gth group/component given the observation $\mathbf{y_i}$ is given by

$$\begin{aligned}\tau_g(\mathbf{y_i}) &= P(observation \ \in \ gth \ component|\ \mathbf{y_i}) \\ &= \frac{\pi_g f_g(\mathbf{y_i})}{f(\mathbf{y_i})} \qquad for\ g = 1, ..., G\ and\ i = 1, ..., n\end{aligned} \tag{3.1.3}$$

The use of the prior and posterior will be understood later when estimating the parameters of the model. In the parametric framework, the mixture density function is written

$$f(\mathbf{y_i}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y_i}; \boldsymbol{\theta_g}) \tag{3.1.4}$$

where the vector $\boldsymbol{\Psi} = (\pi_1, ..., \pi_{G-1}, \boldsymbol{\xi}^\top)^\top$ contains the unknown parameters $\pi_1, ..., \pi_{G-1}$, and $\boldsymbol{\xi} = (\boldsymbol{\theta_1^\top, ..., \theta_G^\top})^\top$ where $\boldsymbol{\theta_g}$ is the vector of parameters of the parametric distribution. The posterior probability of belonging to the gth component of the mixture becomes

$$\tau_g(\mathbf{y_i}; \boldsymbol{\Psi}) = \frac{\pi_g f_g(\mathbf{y_i}; \boldsymbol{\theta_g})}{\sum_{h=1}^{G} \pi_h f_h(\mathbf{y_i}; \boldsymbol{\theta_h})} \tag{3.1.5}$$

Let's look at the fitting of the mixture distribution.The maximum likelihood method will be used to determine the different parameters of the mixture distribution. Due to incomplete data, the estimation of the mixture distribution is done on the available information, i.e. the marginal distribution of $\mathbf{Y_i}$ and not on the joint distribution of $\mathbf{Y_i}$ and $\mathbf{B_i}$. This will be done using the EM algorithm presented by Dempster and al [2]. As a reminder, we have the complete data vector $\mathbf{y_c} = (\mathbf{y}^\top, \mathbf{b}^\top)$ with label vectors $\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_n}$. Recall that $\mathbf{b_i}$ $(i = 1, ..., n)$ is a zero-one vector that determines whether $\mathbf{y_i}$ comes from the gth group $(g = 1, ..., G)$. We will therefore make the difference between the likelihood of the observed (incomplete) data $\mathbf{y}$ noted $L(\boldsymbol{\Psi})$,

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(\mathbf{y_i}; \boldsymbol{\theta_g}) \tag{3.1.6}$$

and the likelihood of the complete data $\mathbf{y_c}$ noted $L_c(\boldsymbol{\Psi})$,

$$L_c(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{g=1}^{G} (\pi_g f_g(y_i; \boldsymbol{\theta_g}))^{b_{gi}} \tag{3.1.7}$$

The log-likelihoods are given by

$$\log\ L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log(\sum_{g=1}^{G}\ \pi_g\ f_g(\mathbf{y_i}; \boldsymbol{\theta_g})) \tag{3.1.8}$$

$$\log\ L_c(\mathbf{\Psi}) = \sum_{g=1}^{G}\sum_{i=1}^{n} b_{gi}(\log \pi_g + \log f_g(\mathbf{y_i};\boldsymbol{\theta_g})) \tag{3.1.9}$$

with $b_{gi} = (\mathbf{b_i})_g = 0$ or 1 and π_g is the mixing proportion. By the assumption made on the distribution of $\mathbf{B_i}$ (3.1.2), the distribution of the complete-data vector $\mathbf{Y_c}$ implies the appropriate distribution for the incomplete-data vector $\mathbf{Y}$ [13].

In order to determine the maximum likelihood estimator in the context where b_{gi} are unobserved, we will use the EM algorithm. This consists in iteratively performing an E-step for the expectation and a M-step for the maximization [13], [21]. These two steps are explained in more detail below:

E-step : this step uses the conditional expectation of the complete log-likelihood given the observed data $\mathbf{y}$ to handle the unobserved data (b_{gi}). For the $(k+1)$th iteration this can be written as

$$Q(\mathbf{\Psi};\mathbf{\Psi^{(k)}}) = E_{\mathbf{\Psi^{(k)}}}(\log L_c(\mathbf{\Psi})|\mathbf{y}) \tag{3.1.10}$$

where $\mathbf{\Psi^{(k)}}$ is the current value of $\mathbf{\Psi}$ at the kth EM iteration. For the first iteration we need to choose an initial value $\mathbf{\Psi^{(0)}}$. The previous equation by linearity can be written as

$$Q(\mathbf{\Psi};\mathbf{\Psi^{(k)}}) = \sum_{g=1}^{G}\sum_{i=1}^{n} \tau_g(\mathbf{y_i};\mathbf{\Psi^{(k)}})(\log\ \pi_g + \log\ f_g(\mathbf{y_i};\boldsymbol{\theta_g})) \tag{3.1.11}$$

with $E_{\mathbf{\Psi^{(k)}}}(B_{gi}|\mathbf{y}) = \tau_g(\mathbf{y_i};\mathbf{\Psi^{(k)}})$ and we have

$$\tau_g(\mathbf{y_i};\mathbf{\Psi^{(k)}}) = \frac{\pi_g^{(k)} f_g(\mathbf{y_i};\boldsymbol{\theta}_\mathbf{g}^{(\mathbf{k})})}{\sum_{h=1}^{G}\pi_h^{(k)} f_h(\mathbf{y_i};\boldsymbol{\theta}_\mathbf{h}^{(\mathbf{k})})} \tag{3.1.12}$$

the posterior probability already mentioned earlier for $g = 1, ..., G$ and $i = 1, ..., n$.

M-step : this step aims to determine at the $(k+1)$th iteration the global maximum of $Q(\mathbf{\Psi};\mathbf{\Psi^{(k)}})$ with respect to $\mathbf{\Psi} = (\pi_1, ..., \pi_{G-1}, \boldsymbol{\xi}^\top)^\top$. This step thus makes it possible to update the estimate of $\mathbf{\Psi^{(k+1)}}$ which one will use in the next E-step. Given the form of $Q(\mathbf{\Psi};\mathbf{\Psi^{(k)}})$ it is possible to determine the updated estimations $\pi_g^{(k+1)}$ independently of the $\boldsymbol{\xi}^{(k+1)}$.

The maximum likelihood estimator of π_g if the b_{gi} were observable would be given by

$$\hat{\pi}_g = \sum_{i=1}^{n} \frac{b_{gi}}{n} \qquad (g = 1, ..., G) \tag{3.1.13}$$

Since we replace b_{gi} by $\tau_g(\mathbf{y_i};\mathbf{\Psi^{(k)}})$ in the E-step, we will follow the same approach in (3.1.13), we obtain

$$\pi^{(k+1)} = \sum_{i=1}^{n} \frac{\tau_g(\mathbf{y_i};\mathbf{\Psi^{(k)}})}{n} \qquad (g = 1, ..., G) \tag{3.1.14}$$

The vector of parameters $\boldsymbol{\xi}$ is updated by maximizing the part of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ depending on this last vector of parameters. To obtain $\boldsymbol{\xi}^{(k+1)}$, one seeks the solution of

$$\sum_{g=1}^{G}\sum_{i=1}^{n}\tau_g(\mathbf{y_i}; \boldsymbol{\Psi}^{(k)})\frac{\partial \log\ f_g(\mathbf{y_i}; \boldsymbol{\theta_g})}{\partial \boldsymbol{\xi}} = 0 \tag{3.1.15}$$

The E-step and M-step are iterated until convergence, i.e. the difference

$$L(\boldsymbol{\Psi}^{(\mathbf{k+1})}) - L(\boldsymbol{\Psi}^{(\mathbf{k})}) \leq \epsilon$$

reaches a pre-specified threshold ϵ. The diagram on Figure 3.1.1 shows how the algorithm works.

A few remarks about the estimation of the parameters. Firstly, the estimation of parameters $\boldsymbol{\Psi}$ from observations only makes sense if this vector of parameters is identifiable. The identifiability for finite mixture models is not defined in the same way as for classical models. Indeed, in the context of mixture models, as defined by [13] we consider that for $\boldsymbol{\Psi} \in \Omega$, where Ω is the space of the parameters, the class of finite mixtures is identifiable if there are two members of parametric family mixture densities with G components densities belonging to the same parametric family

$$f(\mathbf{y_i}; \boldsymbol{\Psi}) = \sum_{g=1}^{G}\pi_g f_g(\mathbf{y_i}; \boldsymbol{\theta_g})$$

and

$$f(\mathbf{y_i}; \boldsymbol{\Psi}^*) = \sum_{g=1}^{G}\pi_g^* f_g(\mathbf{y_i}; \boldsymbol{\theta_g^*})$$

then

$$f(\mathbf{y_i}; \boldsymbol{\Psi}) \equiv f(\mathbf{y_i}; \boldsymbol{\Psi}^*)$$

if and only if $g = g^*$ and the component labels are interchangeable: $f_g(\mathbf{y_i}; \boldsymbol{\theta_g}) = f_g(\mathbf{y_i}; \boldsymbol{\theta_g^*})$ and $\pi_g = \pi_g^*$ for $g = 1, ..., G$. This is label switching.Label switching means that the labels of two components can be swapped. However, even if the class of finite mixtures are identifiable, $\boldsymbol{\Psi}$ is not. To remedy this problem, it is possible to impose constraints on $\boldsymbol{\Psi}$ [13]. The lack of identifiability doesn't seem to be a concern in general when estimating MLEs using the EM algorithm, since it has no impact on the qualitative interpretation of the results [20]. It may, however, have an impact during simulation, and we'll need to pay close attention to this. When covariates are added, the references [11] and [14] indicate that we may still observe estimation problems linked to identifiability for certain data, and that we need to remain cautious.

The second comment we would like to make is about the starting values $\boldsymbol{\Psi}^{(\mathbf{0})}$ of the EM algorithm. It is important to know that a bad choice of initial values can impact on the speed of convergence of the EM algorithm. And a choice of initial parameters too close to the boundary can lead to divergent results.

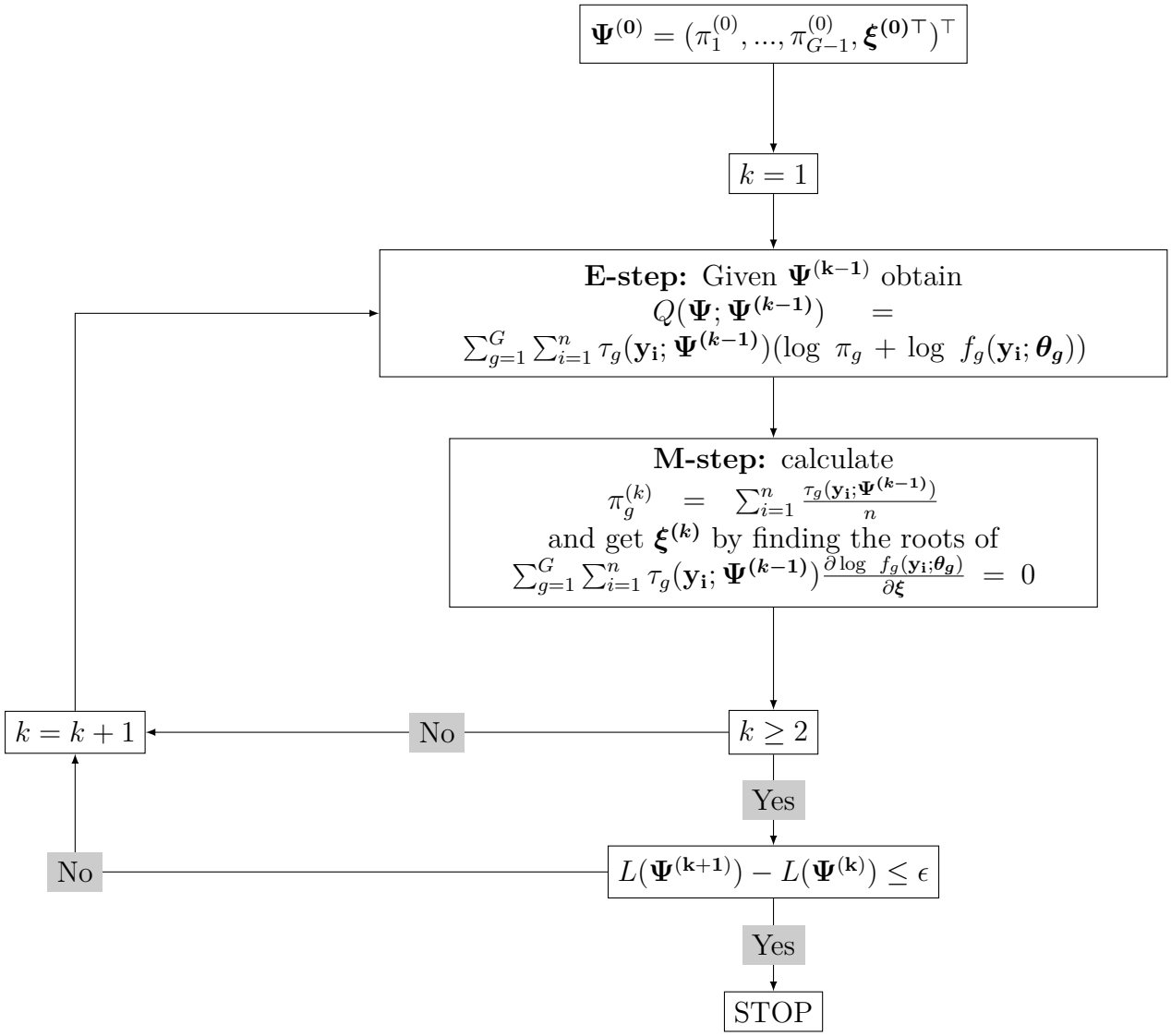

Figure 3.1.1: Diagram of the EM algorithm for the finite mixture model.

Finally, we would like to point out that it is also possible to use a direct approach by directly maximizing the likelihood of incomplete data. However, in the context of finite mixture models, the EM algorithm is favored in the literature.

To conclude this section, let's discuss the method used to obtain the asymptotic covariance matrix of maximum likelihood estimators. Under regularity conditions and with a large sample size, the parameter estimators asymptotically follow a multivariate normal distribution, with the inverse of the Fisher information matrix as the variance-covariance matrix. In practice, we use the observed information matrix [13] given by

$$(\mathbf{I}^{-1}(\hat{\mathbf{\Psi}};\mathbf{y}))^{1/2} \tag{3.1.16}$$

where

$$\mathbf{I}^{-1}(\mathbf{\Psi};\mathbf{y}) = -\frac{\partial^2 \log L(\mathbf{\Psi})}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^\top} \tag{3.1.17}$$

We can directly estimate (3.1.16), but for the finite mixture model this may be analytically difficult to evaluate. If the data are i.i.d., it is possible to use an approximation of the observed information matrix of the complete data [13] :

$$\mathbf{I_e}(\hat{\mathbf{\Psi}};\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y_i};\hat{\mathbf{\Psi}})\mathbf{s}^\top(\mathbf{y_i};\hat{\mathbf{\Psi}}) \tag{3.1.18}$$

$\mathbf{I_e}(\hat{\mathbf{\Psi}};\mathbf{y})$ is the empirical observed information matrix and where

$$\mathbf{s}(\mathbf{y_i};\mathbf{\Psi}) = E_{\mathbf{\Psi}}\left\{\frac{\partial L_{c,i}(\mathbf{\Psi})}{\partial \mathbf{\Psi}}|y\right\} \tag{3.1.19}$$

the complete-data log-likelihood is given by (3.1.9), we then obtain

$$\mathbf{s}(\mathbf{y_i};\mathbf{\Psi}) = \sum_{g=1}^{G} \tau_g(\mathbf{y_i};\mathbf{\Psi})\frac{\partial\left\{\log \pi_g + \log f_g(\mathbf{y_i};\boldsymbol{\theta_g})\right\}}{\partial \mathbf{\Psi}} \tag{3.1.20}$$

We can therefore easily calculate the variance-covariance matrix of estimators and the resulting confidence intervals.

3.2 Survival finite mixture model

The mixture models can also be used in the context of survival analysis. This has been studied by McLachlan and Peel [13], McLachlan and McGiffin [14] and Alves and Dias [5]. Indeed, mixture models are used to model heterogeneous data, which can correspond to different situations in survival analysis. For example, when analyzing events that may occur at different stages of the follow-up.

As already explained, survival analysis is concerned with the random variable T which represents the time until the event of interest occurs. This random variable is characterized by the equations (2.1.1), (2.1.3), (2.1.4). One of the difficulties of the survival analysis

as explained in Chapter 2.1 is to take into account censoring of the data. As already explained, censoring can be fixed (administrative) or can be random variable C with a specific distribution. In this work, we will focus on right censoring. In the context of survival analysis, the observed data for the failure time $\mathbf{y} = (\mathbf{y_1}^\top, ..., \mathbf{y_n}^\top)$ where $\mathbf{y_i} = (w_i, \mathbf{x_i}^\top, \delta_i)^\top$ for $i = 1, ..., n$. Here w_i is the observed time for the ith individuals with $W = min(T, C)$. The vector $\mathbf{x_i}^\top$ corresponds to the covariates of the ith individuals, and $\delta_i = I(T \leq C)$ is the censoring indicator for this ith individual. The value n represents the number of individuals in the study. We consider that the population we are studying in this mixture context is heterogeneous and comes from G possible groups. The survival time density function can be written as a mixture model and is written as in equation (3.1.1). Consequently, the resulting survival function is as follows,

$$S(t) = \sum_{g=1}^{G} \pi_g S_g(t) \tag{3.2.1}$$

where

$$S_g(t) = \int_t^\infty f_g(u)du \qquad (g = 1, ..., G)$$

and $f_1(t), ..., f_G(t)$ are the g component densities and $\pi_1, ..., \pi_G$ with $0 \leq \pi_g, \leq 1$ and $\sum_{g=1}^{G} \pi_g = 1$ are the mixing proportions or weights. The finite mixture survival models use the usual parametric survival model for the component distributions. For example, in the following, we will use the Weibull distribution. We can also take covariates into account. We thus have $S_g(t; \mathbf{x}, \boldsymbol{\theta_g})$ where $\mathbf{x}$ is the covariate vector and $\boldsymbol{\theta_g}$ contains the vectors of the model parameters and the coefficients of the covariates. One way of doing this is to use the PH model presented in Section 2.1. The mixing proportions may also depend on the covariates $\mathbf{x}$. π_g can be specified using, for example, a logistic function of $\mathbf{x}$,

$$\begin{aligned} \boldsymbol{\pi_g} &= \boldsymbol{\pi_g}(\mathbf{x}; \boldsymbol{\alpha}) \\ &= \frac{\exp(\beta_{0g} + \boldsymbol{\beta_g}^\top \mathbf{x})}{1 + \sum_{h=1}^{G-1} \exp(\beta_{0h} + \boldsymbol{\beta_h}^\top \mathbf{x})}, \qquad (g = 1, ..., G-1) \end{aligned} \tag{3.2.2}$$

and

$$\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$$

where $\boldsymbol{\alpha_g} = (\beta_{0g}, \boldsymbol{\beta_g}^\top)^\top$ and where $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}^\top, ..., \boldsymbol{\alpha_{G-1}}^\top)$ contains the logistic regression coefficients. McLachlan and McGiffin [14], and Farewell [11] warn that there may be concerns about identifiability when covariates are taken into account.

Let us now discuss the fitting by maximum likelihood adapted to the context of survival analysis. As in Section 3.1, it is necessary to estimate the unknown parameter vector $\boldsymbol{\Psi}$. When mixing proportions do not depend on covariates we have

$$\boldsymbol{\Psi} = (\pi_1, ..., \pi_{G-1}, \boldsymbol{\xi}^\top)^\top \qquad where\ \boldsymbol{\xi} = (\boldsymbol{\theta_1}^\top, ...\boldsymbol{\theta_G}^\top)^\top$$

When the mixing proportions depend on the covariates then $\mathbf{\Psi}$ is given by

$$\mathbf{\Psi} = (\boldsymbol{\alpha}^\top, \boldsymbol{\xi}^\top)$$

In the following development, we'll use the π_g notation for mixing proportions, even though these depend on covariates.

As mentioned above, censoring must be taken into account in the survival analysis and thus also in the context of model fitting. Given the observed data $\mathbf{y_i} = (w_i, \mathbf{x_i}^\top, \delta_i)^\top$ for $i = 1, ..., n$ with n individuals , the log-likelihood is obtained by taking the logarithm of the expression (2.1.6),

$$\log(L(\mathbf{\Psi})) = \sum_{i=1}^{n} \delta_i \log\ f(w_i; \mathbf{x_i}, \mathbf{\Psi}) + \sum_{i=1}^{n} (1 - \delta_i) \log\ S(w_i; \mathbf{x_i}, \mathbf{\Psi}) \tag{3.2.3}$$

where $f(w_i; \mathbf{x}, \mathbf{\Psi})$ and $S(w_i; \mathbf{x}, \mathbf{\Psi})$ are respectively the mixture density function and the mixture survival function. The first part of the log-likelihood function is the contribution of the uncensored individuals, and the second part is the contribution of the censored individuals. As in the previous section, the log-likelihood must be maximized. Maximizing the equation (3.2.3) can be complicated. Indeed, as in the classical finite mixture model part, the components from which the data come are not known and moreover one must take into account the censoring. McLachlan and McGiffin [14] propose as in Section 3.1 to consider the EM framework. To do so, we consider that the observed data vector $\mathbf{y}$ is incomplete and that there exists a non observed label vector $\mathbf{b} = (\mathbf{b_1}^\top, ..., \mathbf{b_n}^\top)^\top$ such that the complete data are $\mathbf{y_c} = (\mathbf{y}^\top, \mathbf{b}^\top)^\top$. Recall that $\mathbf{b_i}$ $(i = 1, ..., n)$ is a zero-one vector that determines whether $\mathbf{y_i}$ comes from the gth group $(g = 1, ..., G)$ with $b_{gi} = (\mathbf{b_i})_g = 0$ or 1 . As for the classical mixture models, the determination of the maximisation of the likelihood is done with the EM algorithm where the E-step and M-step are adapted to the survival framework.

E-step : The equation (3.1.11) is adapted to take into account the censoring and the covariates, we obtain

$$Q(\mathbf{\Psi}; \mathbf{\Psi^{(k)}}) = \sum_{g=1}^{G} \sum_{i=1}^{n} \tau_g(w_i; \mathbf{x_i}, \mathbf{\Psi^{(k)}}) \left\{ \log \pi_g + \delta_{ig} (w_i; \mathbf{x_i}, \boldsymbol{\theta_g}) + (1 - \delta_i)_g (w_i; \mathbf{x_i}, \boldsymbol{\theta_g}) \right\} \tag{3.2.4}$$

and $\tau_g(w_i; \mathbf{x_i}, \mathbf{\Psi^{(k)}})$

$$\tau_g(w_i; \mathbf{x_i}, \Psi^{(k)}) = \frac{\pi_g f_g(w_i; \mathbf{x_i}, \boldsymbol{\theta_g^{(k)}})^{\delta_i} S_g(w_i; \mathbf{x_i}, \boldsymbol{\theta_g^{(k)}})^{1-\delta_i}}{(\sum_{h=1}^{G} \pi_h f_h(w_i; \mathbf{x_i}, \boldsymbol{\theta_h^{(k)}}))^{\delta_i} (\sum_{h=1}^{g} \pi_h S_h(w_i; \mathbf{x_i}, \boldsymbol{\theta_h^{(k)}}))^{1-\delta_i}} \tag{3.2.5}$$

the posterior probability already mentioned earlier for $g = 1, ..., G$ and $i = 1, ..., n$. The mixing proportions and component distributions can either depend on covariates or not.

M-step : this step aims to determine at $(k+1)$th iteration the global maximum of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ according to $\boldsymbol{\Psi} = (\pi_1, ..., \pi_{G-1}, \boldsymbol{\xi}^{T})^T$ if π_g do not depend on covariates or $\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\xi}^{T})^T$ if π_g depend on covariates. The aim is to update the estimate of $\boldsymbol{\Psi}^{(k+1)}$. Given the form of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ it is possible to determine $\pi_g^{(k+1)}$ or $\boldsymbol{\alpha}^{(k+1)}$ independently of the $\boldsymbol{\xi}^{(k+1)}$. As for the regression model for the different components survival function, we can use the parametric PH model as in Section 2.1 [14]. We therefore want to determine the parameters of the baseline hazard and the parameters of the coefficients related to the covariates. It should be noted that the proportional hazard assumption holds only within the group, not in the population [14]. Maximization can be done by the Newton Raphson algorithm [26].

The algorithm is carried out from initial values iteratively until the convergence threshold is reached, as in the classical mixture model approach. The computation of the Fisher information matrix is done as in the classical finite mixture context by adapting the log-likelihood of the complete data (3.1.9) to take censoring into account. Figure 3.2.1 shows the diagram of the EM algorithm adapted to the survival framework and with the mixture proportion depending on the covariates.

Here we have focused on the survival function derived from a mixture model (3.2.1). A remark can be made about the corresponding hazard function given by

$$h(t) = \frac{f(t)}{S(t)} \tag{3.2.6}$$

$$= \frac{\sum_{g=1}^{G} \pi_g h_g(t) S_g(t)}{S(t)} \tag{3.2.7}$$

$$= \sum_{g=1}^{G} \pi_g \lambda_g(t) \tag{3.2.8}$$

$h_g(t)$ is the gth component hazard function and where

$$\lambda_g(t) = \frac{h_g(t) S_g(t)}{S(t)} \tag{3.2.9}$$

$$= \frac{h_g(t) \exp[-\int_0^t h_g(u)du]}{\sum_{j=1}^{G} \pi_j \exp[-\int_0^t h_j(u)du]} \qquad (g = 1, ..., G) \tag{3.2.10}$$

$$\tag{3.2.11}$$

Let $\mathbf{x}$ be a set of covariates and assuming that $h_g(t)$ has the proportional hazards form such that

$$h_g(t; \mathbf{x}) = h_{0g}(t) \exp(\theta_g^T \mathbf{x})$$

where $h_{0g}(t)$ is the baseline hazard function. Although $h_g(t)$ has the proportional hazard form, $h(t)$ does not necessarily have it. Therefore, proportional hazard is not observed at the population level.

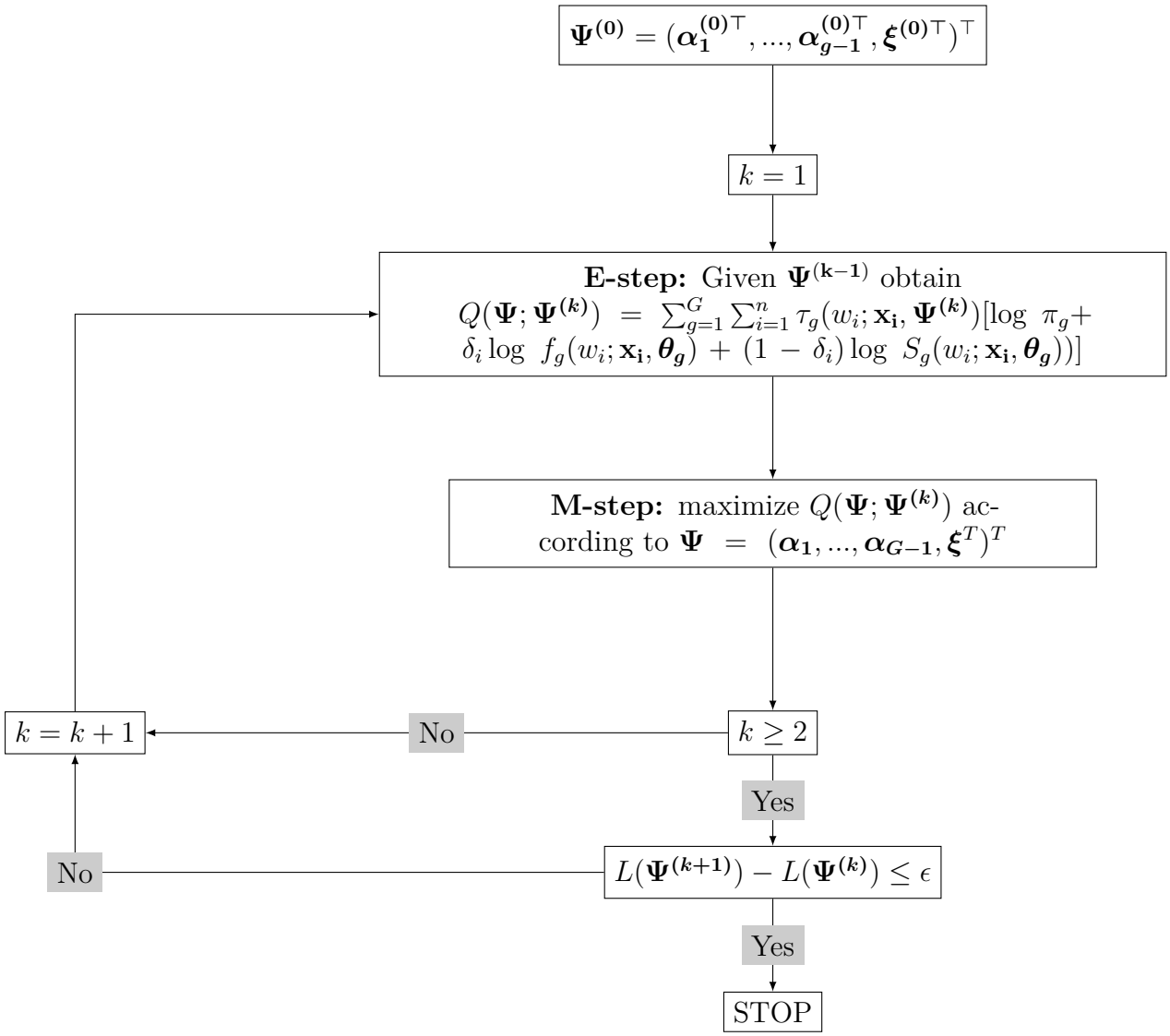

Figure 3.2.1: Diagram of the EM algorithm for the survival mixture model.

While above it was proposed to specify the mixture model by the survival function, Blackstone and al [10] propose to specify the mixture model by the hazard function

$$H(t) = \sum_{g=1}^{G} \mu_g(\mathbf{x}, \boldsymbol{\beta_g}) G_g(t, \boldsymbol{\theta_g})$$

where $H(t)$ is the mixture cumulative hazard function , $\mu_g(\mathbf{x}; \boldsymbol{\beta_g})$ is the mixing proportion, called the scaling function and $G_g(t, \boldsymbol{\theta_g})$ is the component cumulative hazard function, called the shaping function which does not depend on $\mathbf{x}$.

3.3 3-component mixture cure model

Now that we have discussed the mixture model in the classical case and in the survival framework, we can adapt the latter model to the context that interests us in this work. That is to say, we can define a survival model which takes into account a phase of realization of the event of interest on the short term, a phase of realization on the long term and a part of the population which will never experiment the event of interest. It would therefore be a 3-component mixture cure model. This model becomes a more complex version than the classical mixture cure model presented in Section 2.2.2 (which was a 2-component mixture cure model) and thus allows us to capture the bimodality of data. This model was also proposed by Hunsberger and al [22] but our methodology is different.

The model will therefore be composed of a mixture of three component distributions:

- $F_1(t; \boldsymbol{\theta_1})$, the distribution of time until the event of interest that would occur in the short-term phase. $S_1(t; \boldsymbol{\theta_1})$ and $f_1(t; \boldsymbol{\theta_1})$ respectively are the survival and density function associated with $F_1(t; \boldsymbol{\theta_1})$. π_1 corresponds to the probability of belonging to this first phase.

- $F_2(t; \boldsymbol{\theta_2})$, the distribution of time until the event of interest that would occur in the long-term phase. $S_2(t; \boldsymbol{\theta_2})$ and $f_2(t; \boldsymbol{\theta_2})$ respectively are the survival and density function associated with $F_2(t; \boldsymbol{\theta_2})$. π_2 is the mixing proportion corresponding to this second phase.

- The last component corresponds to the fraction of the population that in the very long term is still not affected by the event of interest. This population is considered to be cured. The survival function is given by $S_3(t) = 1$. The corresponding distribution function is null. The proportion of people cured is noted as $\pi_3 = 1 - \pi_1 - \pi_2$.

As for the two-component cure model, the choice of models for the component distribution is quite free and is made among the different classical parametric families of survival analysis. It could be considered to take two different parametric families for the two distributions $F_1(t; \boldsymbol{\theta_1})$ and $F_2(t; \boldsymbol{\theta_2})$ but in this work we decided to focus on a model with 2 Weibull distributions as in the article [22]. The different mixing proportions are either fixed or dependent on the covariates.

In the first case, the 3-component survival function is given by

$$S(t;\boldsymbol{\Psi}) = \pi_1 S_1(t;\boldsymbol{\theta_1}) + \pi_2 S_2(t,\boldsymbol{\theta_2}) + \pi_3 \tag{3.3.1}$$

where $\sum_{g=1}^{3} \pi_g = 1$ and $0 \leq \pi_g \leq 1$. It can also be written $\pi_3 = 1 - \pi_1 - \pi_2$. The parameter vector is given by

$$\boldsymbol{\Psi} = (\pi_1, \pi_2, \boldsymbol{\xi}^\top)^\top \qquad \textit{where } \boldsymbol{\xi} = (\boldsymbol{\theta_1^\top}, \boldsymbol{\theta_2^\top})^\top$$

or π_g depends on covariates by the relation (3.2.2), and the parameter vector is then given by

$$\boldsymbol{\Psi} = (\boldsymbol{\alpha}^\top, \boldsymbol{\xi}^\top)$$

Assuming the i.i.d. data $\mathbf{y_i} = (w_i, \mathbf{x_i}^\top, \delta_\top)^\top$ for $i = 1, ..., n$, the parameters are estimable using the (log)likelihood function. Using the equation (3.2.3) from the previous section and adapting it to the 3-component mixture cure model, we obtain

$$\begin{aligned}\log(L(\boldsymbol{\Psi})) = &\sum_{i=1}^{n} \delta_i \ \log[\pi_1(\mathbf{x_i};\boldsymbol{\alpha}) f_1(w_i;\mathbf{x_i};\boldsymbol{\theta_1}) + \pi_2(\mathbf{x_i};\boldsymbol{\alpha}) f_2(w_i;\mathbf{x_i};\boldsymbol{\theta_2})] \\ &+ \sum_{i=1}^{n} (1-\delta_i) \ \log[\pi_1(\mathbf{x_i};\boldsymbol{\alpha}) S_1(w_i;\mathbf{x_i};\boldsymbol{\theta_1}) + \pi_2(\mathbf{x_i};\boldsymbol{\alpha}) S_2(w_i;\mathbf{x_i};\boldsymbol{\theta_2}) \\ &+ (1 - \pi_1(\mathbf{x_i};\boldsymbol{\alpha}) - \pi_2(\mathbf{x_i};\boldsymbol{\alpha}))]\end{aligned} \tag{3.3.2}$$

The part of the equation that corresponds to the contribution of the uncensored individuals, corresponds to the individuals who experience the event of interest and therefore will not be cured. These individuals may experience the event in the short or long term. The second part of the equation corresponds to the contribution of the censored observations. When the data are censored it is not known whether individuals are cured or will experience the event in the short term or in the long term. As proposed in Sections 3.1 and 3.2, to maximize the log likelihood function, we use the EM algorithm. Below is explained how the E-step and M-step work for the 3-component mixture cure model.

E-step : This step was completed using equation (3.2.4) for G = 3.

$$\begin{aligned}
Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(\boldsymbol{k})}) =& \sum_{i=1}^{n} \tau_1(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log \pi_1 + \delta_i \log f_1(w_i;\mathbf{x_i},\boldsymbol{\theta_1}) + (1-\delta_i)\log S_1(w_i;\mathbf{x_i},\boldsymbol{\theta_1})\right\} \\
&+ \sum_{i=1}^{n} \tau_2(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log \pi_2 + \delta_i \log f_2(w_i;\mathbf{x_i},\boldsymbol{\theta_2}) + (1-\delta_i)\log S_2(w_i;\mathbf{x_i},\boldsymbol{\theta_2})\right\} \\
&+ \sum_{i=1}^{n} \tau_3(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log \pi_3 + (1-\delta_i)\log S_3(w_i;\mathbf{x_i},\boldsymbol{\theta_2})\right\} \\
=& \sum_{i=1}^{n} \tau_1(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log \pi_1 + \delta_i \log f_1(w_i;\mathbf{x_i},\boldsymbol{\theta_1}) + (1-\delta_i)\log S_1(w_i;\mathbf{x_i},\boldsymbol{\theta_1})\right\} \\
&+ \sum_{i=1}^{n} \tau_2(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log \pi_2 + \delta_i \log f_2(w_i;\mathbf{x_i},\boldsymbol{\theta_2}) + (1-\delta_i)\log S_2(w_i;\mathbf{x_i},\boldsymbol{\theta_2})\right\} \\
&+ \sum_{i=1}^{n} \tau_3(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})}) \left\{\log(1-\pi_1-\pi_2)\right\}
\end{aligned}$$

and $\tau_i(w_i;\mathbf{x_i},\boldsymbol{\Psi}^{(\boldsymbol{k})})$

$$\tau_1(w_i;\mathbf{x_i},\Psi^{(k)}) = \frac{\pi_1 f_1(w_i;\mathbf{x_i},\boldsymbol{\theta_i^{(k)}})^{\delta_i} S_1(w_i;\mathbf{x_i},\boldsymbol{\theta_i^{(k)}})^{1-\delta_i}}{(\sum_{l=1}^{2}\pi_l f_l(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{\delta_i}(\sum_{h=1}^{3}\pi_h S_h(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{1-\delta_i}} \tag{3.3.3}$$

$$\tau_2(w_i;\mathbf{x_i},\Psi^{(k)}) = \frac{\pi_2 f_2(w_i;\mathbf{x_i},\boldsymbol{\theta_i^{(k)}})^{\delta_i} S_2(w_i;\mathbf{x_i},\boldsymbol{\theta_i^{(k)}})^{1-\delta_i}}{(\sum_{l=1}^{2}\pi_l f_l(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{\delta_i}(\sum_{h=1}^{3}\pi_h S_h(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{1-\delta_i}} \tag{3.3.4}$$

$$\tau_3(w_i;\mathbf{x_i},\Psi^{(k)}) = \frac{[1-\pi_1-\pi_2]^{1-\delta_i}}{(\sum_{l=1}^{2}\pi_l f_l(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{\delta_i}(\sum_{h=1}^{3}\pi_h S_h(w_i;\mathbf{x_i},\boldsymbol{\theta_h^{(k)}}))^{1-\delta_i}} \tag{3.3.5}$$

the posterior probabilities for the 3 possible components. It should be noted that the survival function for group 3, representing cured individuals, is equal to 1. Moreover, only censored individuals can be classified as belonging to group 3. The mixing proportions above can either depend on covariates or not. We have considered that the distributions depend on covariates, but they may not necessarily.

M-step : the realization of the M-step is done as in the previous section. The implementation of the algorithm is done as in the previous sections.

The 3-component mixture cure model can be seen as consisting of two models: a latency model for the short-term and long-term phases which is the time to observe the event and an incidence model which corresponds to the probability of being in one of the phase types or to be cured. The parameters of the model components are interpreted conditionally on being in a certain phase. The incidence part is interpreted as usual.

Figures 3.3.1, 3.3.2 show the survival and density functions, respectively, for different Weibull 3-component mixture cure models. As expected, the survival curve is composed of two plateaus and the density is bimodal.

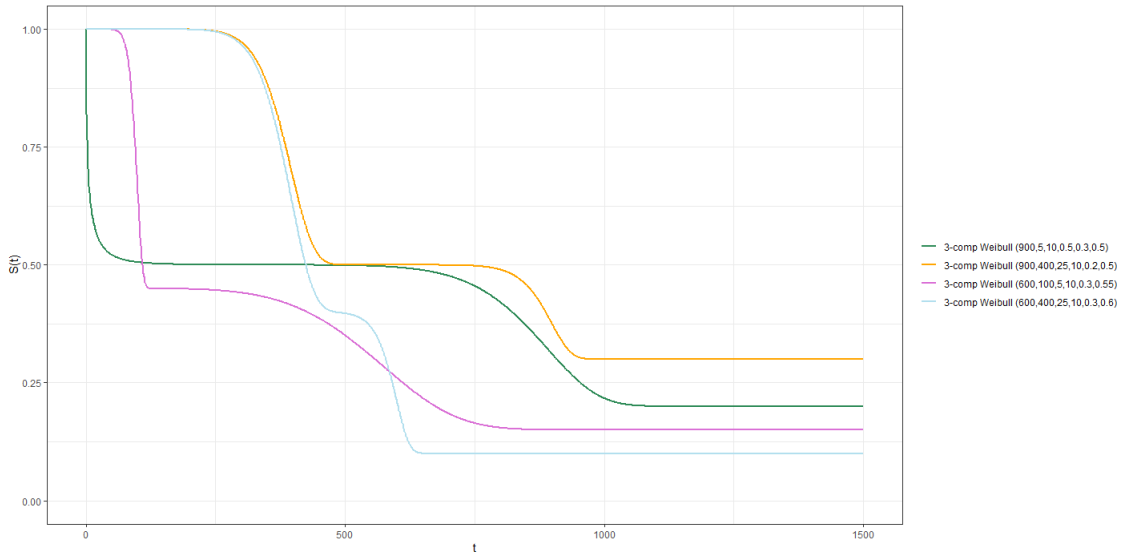

Figure 3.3.1: T $\sim$ 3-comp Weibull$(\lambda_1, \lambda_2, \gamma_1, \gamma_2, \pi_1, \pi_2)$ survival function

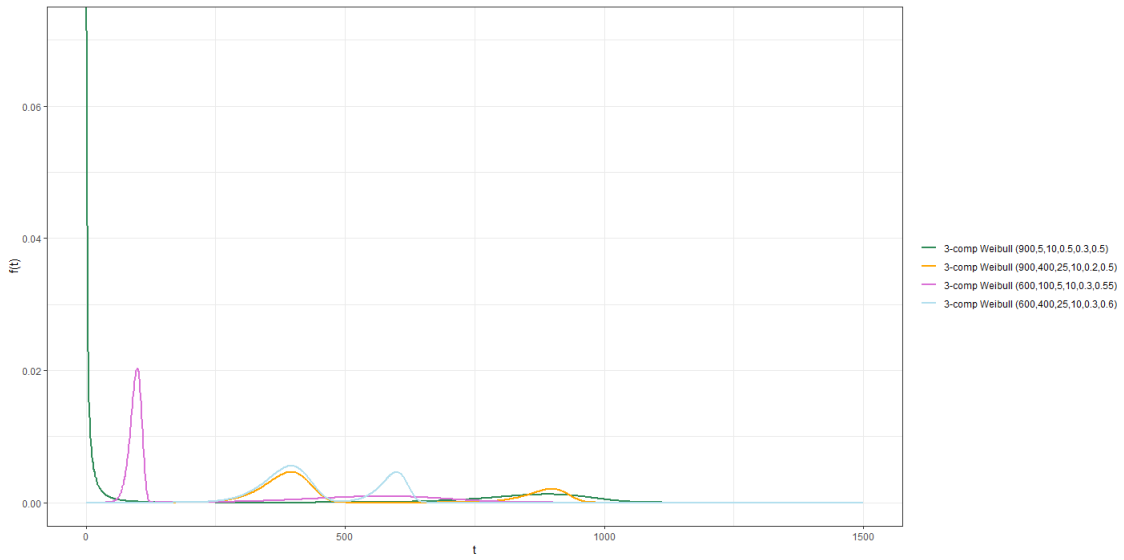

Figure 3.3.2: T $\sim$ 3-comp Weibull$(\lambda_1, \lambda_2, \gamma_1, \gamma_2, \pi_1, \pi_2)$ density function

Chapter 4

Exponentiated Log-Sinh Cauchy cure model

The second model proposed in this work to take into account the bimodal structure of the distribution is the Exponentiated Log Sinh Cauchy (ELSC) model proposed by Ramires and al.[31]. As in Section 2.2.2, the mixture cure models will be used, the novelty here will be the use of the ELSC distribution function as distribution of the uncured individuals [30]. Indeed, the ELSC distribution, contrary to distributions most often used in the literature, has not only a location and a scale parameter but also two shape parameters which characterize the bimodality, the skewness [1] and the kurtosis [2] . These two additional parameters make this distribution more flexible than the usual distributions such as Weibull, log normal, etc. This model thus offers an alternative to the 3-component mixture cure distribution proposed in the Chapter 3. In Section 4.1, we will present the ELSC distribution and its different characteristics. The Section 4.2 will present the mixture cure model associated with this distribution. The last Section of this chapter will discuss the estimation of the model.

4.1 Exponentiated log-sinh Cauchy function

The Exponentiated Sinh Cauchy (ESC) density function was introduced by Cooray [19] in order to have a distribution characterised by scale, location parameters and in addition two shape parameters. This distribution makes it possible to take into account certain data behaviors that more traditional distributions could not capture. This can include skewed, bimodal, bathtub data. The Exponentiated Sinh Cauchy function distribution can be written as follows:

$$F(u;\mu,\sigma,\nu,\tau) = \left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu \sinh\left(\frac{u-\mu}{\sigma}\right)]\right\}^{\tau} \tag{4.1.1}$$

[1] Measures the asymmetry of the distribution. This is the third standardized moment.

[2] Measures the tailedness/sharpness of a distribution. This is the fourth standardized moment.

where $u \in \mathbf{R}$, $\mu \in \mathbf{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, $\nu > 0$ is a shape parameter known as symmetry parameter of the distribution, which characterizes the bimodality of the distribution and $\tau > 0$ is also a shape parameter known as asymmetry parameter of the distribution. The last two parameters characterise the skewness, kurtosis and bimodality of the distribution.

Unfortunately, the distribution (4.1.1) has as support the domain of the reals. This is not desirable in the context of survival analysis, as the random variable T is positive. This is why [31] proposes to use the Exponentiated Log-Sinh Cauchy model by applying the exponential function to the random variable U wich follows the distribution function (4.1.1). We have $T = e^U$. We then obtain the following distribution function:

$$F_{ELSC}(t; \mu, \sigma, \nu, \tau) = \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh\left(\frac{\log(t) - \mu}{\sigma}\right)] \right\}^{\tau} \tag{4.1.2}$$

Let $\rho = [\log(t) - \mu]/\sigma$ and $t > 0$, $\mu \in \mathbf{R}$, $\sigma > 0$, $\nu > 0$, $\tau > 0$. The density function is obtained by deriving the function (4.1.2) according to t, we have

$$f_{ELSC}(t; \mu, \sigma, \nu, \tau) = \frac{\tau\nu}{t\sigma\pi} \frac{\cosh(\rho)}{[\nu^2 \sinh^2(\rho) + 1]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh(\rho)] \right\}^{\tau - 1} \tag{4.1.3}$$

As in classical survival analysis, the survival function is $S_{ELSC}(t) = 1 - F_{ELSC}(t)$ and the hazard function is $h_{ELSC}(t) = f_{ELSC}(t)/S_{ELSC}(t)$. Figures 4.1.1 are the plots of the survival and the density functions, where the parameters ν and τ vary. These plots show that it is possible to observe a first and a second wave of events and to have a bimodal structure for the distribution of the random variable T.

The quantile function is the inverse of the distribution function $F(t) = p$ where $p \in [0, 1]$. This quantile function is also a random variable. This function will later be used to generate the data during the simulation. The quantile function is given by

$$t = Q(p) = \exp\left(\mu + \sigma \operatorname{arcsinh}\left\{ \frac{1}{\nu} \tan[\pi(p^{1/\tau} - 0.5] \right\}\right) \tag{4.1.4}$$

4.2 Mixture cure ELSC model

In this section we will use the ELSC distribution as the latency distribution of the mixture cure model [30] defined in Section 2.2.2. Recall that B is an indicator of recovery status and $S_u(t) = P(T > t|B = 1)$ is the conditional survival function of uncured (susceptible) individuals and π is the probability of being susceptible. Be careful not to confuse the notation π and π which correspond respectively to the probability of being susceptible and to the value of the constant π. We thus obtain the survival function for the population by

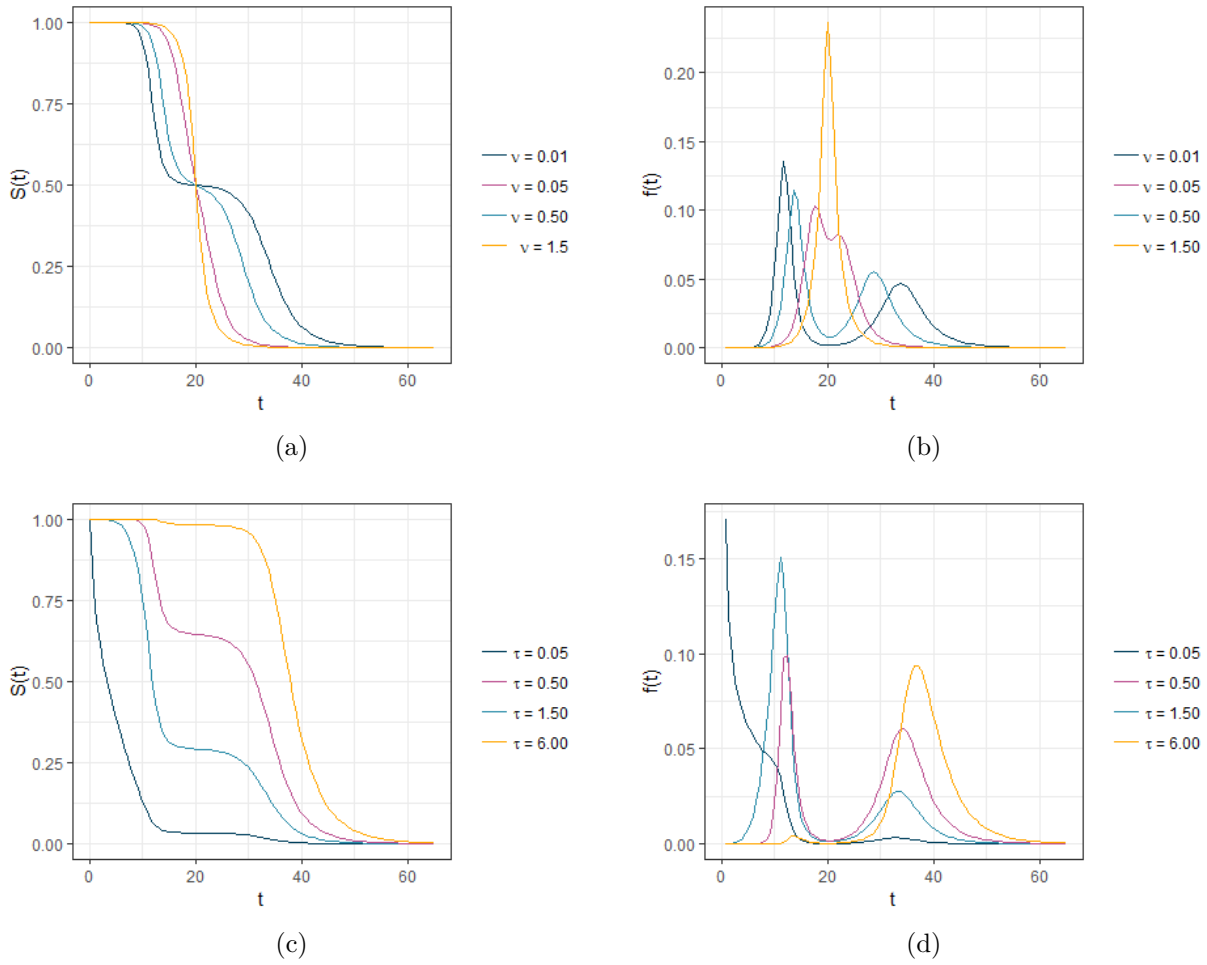

Figure 4.1.1: Plots of the ELSC survival function on the right and the ELSC density function on the left for fixed values : (a) & (b) $\sigma = 0.1$, $\mu = 3$, $\tau = 1$ and (c) & (d) $\sigma = 0.1$, $\mu = 3$, $\nu = 0.01$

inserting $S_u(t) = S_{ELSC}(\mathrm{t})$ in equation (2.2.2):

$$\begin{aligned} S_{pop} &= (1-\pi) + \pi S_{ELSC}(t) \\ &= (1-\pi) + \pi[1 - \left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu \sinh(\rho)\right\}^{\tau}] \\ &= 1 - \pi\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu \sinh(\rho)\right\}^{\tau} \end{aligned} \tag{4.2.1}$$

where $\rho = [\log(t) - \mu]/\sigma$, $t > 0$, $\mu \in \mathbf{R}$, $\sigma > 0$, $\nu > 0$, $\tau > 0$, $\pi \in [0, 1]$ is the uncure probability. The density of the population is given by

$$f_{pop}(t; \mu, \sigma, \nu, \tau) = \frac{\pi\tau\nu}{t\sigma\pi} \frac{\cosh(\rho)}{[\nu^2 \sinh^2(\rho) + 1]} \left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu \sinh(\rho)]\right\}^{\tau-1} \tag{4.2.2}$$

If the random variable T follows (4.2.2) one uses the notation $T \sim ELSCcr(\mu, \sigma, \nu, \tau, \pi)$ where cr refers to the cure proportion. Figures 4.2.1, 4.2.2 show respectively the survival function and the hazard function for different values of τ and ν

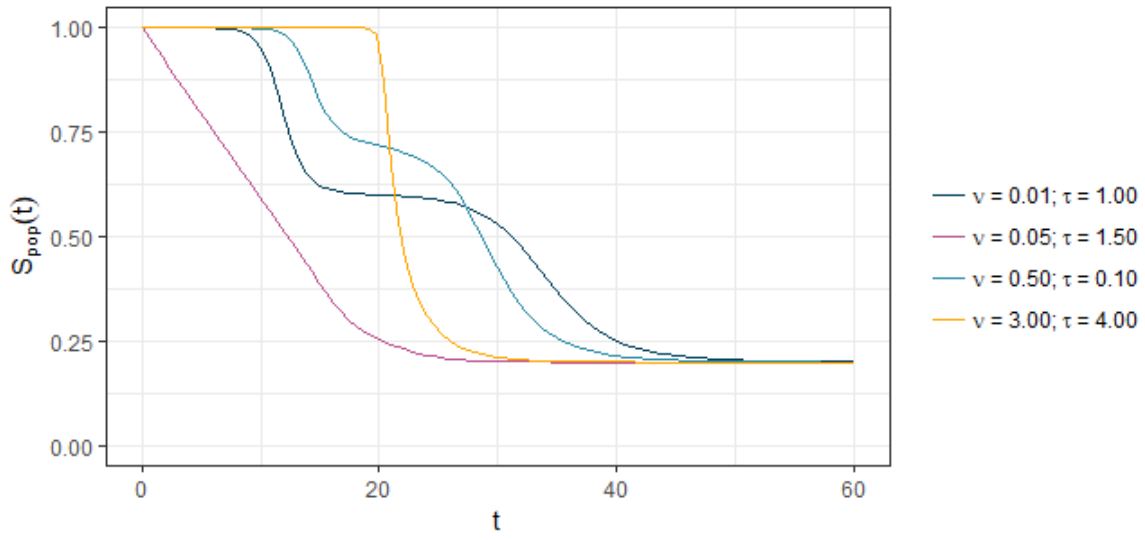

Figure 4.2.1: T $\sim$ ELSCcr$(\mu, \sigma, \nu, \tau, \Pi)$ survival function : $\mu = 4$, $\sigma = 0.1$, $\pi = 0.8$, ν and τ varies

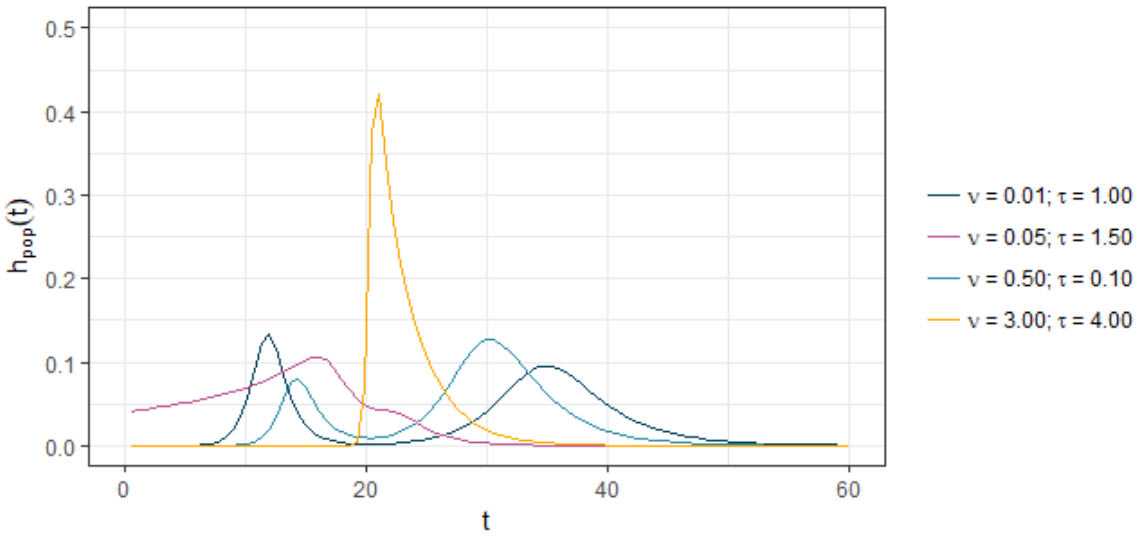

Figure 4.2.2: T $\sim$ ELSCcr$(\mu, \sigma, \nu, \tau, \Pi)$ hazard function : $\mu = 4$, $\sigma = 0.1$, $\pi = 0.8$, ν and τ varies

It can be seen that, as intended, it is possible to represent a bimodal hazard function. There is a first plateau on Figure 4.2.1 between the realization of events of interest in the short term and the realization of events of interest in the long term. A second plateau is observed at $S(t) = 1 - \pi = 0.2$, representing the proportion of individuals who can be considered cured.

4.2.1 Regression model

As mentioned in Section 2.2.2, the incidence model can depend on the covariates, as can the latency model. In order to take the covariates into account, it is proposed to use link functions to connect the parameters of the conditional distribution of T to the covariates . Let t_i the observed response variables , for $i = 1, ..., n$, n independent positive random variables, with density function $f(t_i|\boldsymbol{\theta_i})$ conditional on $\boldsymbol{\theta_i} = (\mu_i, \sigma_i, \nu_i, \tau_i, \pi_i)$ the vector of parameters of the distribution (4.2.2) for the ith individual, and each parameter is a function of explanatory variables. For each parameter of the ELSCcr distribution, we can employ $g_k(.)$, for $k = 1, \ldots, 5$ a monotonic injective and twice differentiable link function. This function relates the parameters to the covariates:

$$g_k(\boldsymbol{\theta_k}) = \mathbf{X_k}\boldsymbol{\beta_k}$$

i.e.

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \mathbf{X_1}\boldsymbol{\beta_1}, \\ g_2(\boldsymbol{\sigma}) &= \mathbf{X_2}\boldsymbol{\beta_2}, \\ g_3(\boldsymbol{\nu}) &= \mathbf{X_3}\boldsymbol{\beta_3}, \\ g_4(\boldsymbol{\tau}) &= \mathbf{X_4}\boldsymbol{\beta_4}, \\ g_5(\boldsymbol{\pi}) &= \mathbf{X_5}\boldsymbol{\beta_5}, \end{aligned}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$, $\boldsymbol{\pi}$ are the different parameter vector of length n, $\boldsymbol{\beta_k} = (\beta_{0k}, \beta_{1k}, ..., \beta_{m_k k})^\top$ is a parameter vector of length $m_k + 1$ and m_k is the number of explanatory variables associated with the kth parameter. $\mathbf{X_k} = (\mathbf{x_{1,k}}, ..., \mathbf{x_{n,k}})^\top$ is a known design matrix of explanatory variables of dimension $n \times (m_k + 1)$. There are $m = 5 + \sum_{k=1}^{5} m_k$ parameters to estimate and we have to ensure that $m < n$. The $\boldsymbol{\beta_k}$ are functionally independant. Since $\mu_i \in \mathbf{R}$, we can choose the identity function as the link function $g_1()$. For the parameters σ_i, ν_i, τ_i we choose the logarithmic link function since they must be strictly positive. For the proportion of uncured observations π_i we choose the logit link function. We have

$$\begin{aligned} \mu_i &= \mathbf{x_{i,1}}^\top \boldsymbol{\beta_1}, \\ log(\sigma_i) &= \mathbf{x_{i,2}}^\top \boldsymbol{\beta_2}, \\ log(\nu_i) &= \mathbf{x_{i,3}}^\top \boldsymbol{\beta_3}, \\ log(\tau_i) &= \mathbf{x_{i,4}}^\top \boldsymbol{\beta_4}, \\ \pi_i &= \frac{\exp(\mathbf{x_{i,5}}^\top \boldsymbol{\beta_5})}{1 + \exp(\mathbf{x_{i,5}}^\top \beta_5)} \end{aligned} \tag{4.2.3}$$

where $\mathbf{x_{i,k}}^\top = (1, x_{i1,k}, ..., x_{im_k,k})$ is the vector of explanatory variables for the ith individuals and the kth parameter. To determine the vector of parameters $\boldsymbol{\beta_k}$, since the model is fully parametric, we can rely on the maximization of the likelihood function. The regression parameters can be interpreted in the usual way.

4.2.2 Likelihood maximization

As before, we consider a sample of n individuals, with $t_1, ..., t_n$ the real time-to-event of interest and $c_1, ..., c_n$ the censoring time-to-event of each individual. The observed response variables for the ith individual is $w_i = min(t_i, c_i)$, the censoring indicator is given by $\delta_i = I(t_i \leq c_i)$. We assume that the censoring time C and the real time-to-event T are independent and that the censoring time is uninformative. The likelihood function is obtained using the likelihood of the mixture cure model (2.2.3) and using $f_u(t) = f_{ELSC}(t)$

and $S_u(t) = S_{ELSC}(t)$. Applying the logarithm we get

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \times (\log(\pi_i) + \log(\tau_i) + \log(\nu_i) + \log(\cosh(\rho_i)) - \log(\pi w_i) - \log(\sigma_i) \\ - \log(\nu^2 \sinh^2(\rho_i) + 1) + (\tau_i - 1)\log\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu \sinh(\rho_i)]\}) \\ + \sum_{i=1}^{n}(1 - \delta_i) \times \log(1 - \pi_i\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu_i \sinh(\rho_i)\}^{\tau_i}) \quad (4.2.4)$$

where $\rho_i = [\log(w_i) - \mu_i]/\sigma_i$. Given that the parameter vector $\boldsymbol{\theta}$ is assumed to follow the relations (4.2.3) , we try to determine the parameter vector $\boldsymbol{\theta'} = (\boldsymbol{\beta_1^\top}, \boldsymbol{\beta_2^\top}, \boldsymbol{\beta_3^\top}, \boldsymbol{\beta_4^\top}, \boldsymbol{\beta_5^\top})$. The maximisation of the likelihood when taking the covariates into account is done by maximising the equation 4.2.4 with respect to $\boldsymbol{\theta'} = (\boldsymbol{\beta_1^\top}, \boldsymbol{\beta_2^\top}, \boldsymbol{\beta_3^\top}, \boldsymbol{\beta_4^\top}, \boldsymbol{\beta_5^\top})$. It is assumed that the parameters are identifiable [30]. The first derivatives according to the different parameters of the vector $\boldsymbol{\theta'}$ are given by [30] ,

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta_1}} = \mathbf{X_1^\top \Omega_1 s_1},\ \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta_2}} = \mathbf{X_2^\top \Omega_2 s_2},\ \frac{\partial l(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta_3}} = \mathbf{X_3^\top \Omega_3 s_3},$$
$$\frac{\partial l(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta_4}} = \mathbf{X_4^\top \Omega_4 s_4},\ \frac{\partial l(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta_5}} = \mathbf{X_5^\top \Omega_5 s_5},$$

where

$$\boldsymbol{\Omega_1} = I_n,$$
$$\boldsymbol{\Omega_2} = (\sigma_1, ..., \sigma_n) \times I_n,$$
$$\boldsymbol{\Omega_3} = (\nu_1, ..., \nu_n) \times I_n,$$
$$\boldsymbol{\Omega_4} = (\tau_1, ..., \tau_n) \times I_n,$$
$$\boldsymbol{\Omega_5} = (\Pi_1(1 - \Pi_1), ..., \Pi_n(1 - \Pi_n)) \times I_n,$$

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

and $\mathbf{s_k} = (s_{1,k}, ..., s_{n,k})$ with $k = 1, ..., 5$, and

$$s_{i,1} = \begin{cases} -\dfrac{\tanh(\rho_i)}{\sigma_i} + \dfrac{\nu_i^2 \sinh(2\rho_i)}{\sigma_i K_i} - (\tau_i - 1)\dfrac{\nu_i \cosh(\rho_i)}{\pi \sigma_i J_i K_i} & \text{if } \delta_i = 1 \\ \dfrac{\pi_i \tau_i J_i^{\tau_i - 1} \nu_i \cosh(\rho_i)}{\pi(1 - \pi_i J_i^{\tau_i}) K_i \sigma_i} & \text{if } \delta_i = 0 \end{cases}$$

$$s_{i,2} = \begin{cases} -\dfrac{\rho_i \tanh(\rho_i)}{\sigma_i} - \dfrac{1}{\sigma_i} + \dfrac{\nu_i^2 \rho_i \sinh(2\rho_i)}{K_i \sigma_i} - \dfrac{(\tau_i - 1)\rho_i \nu_i \cosh(\rho_i)}{\pi J_i K_i \sigma_i} & \text{if } \ \delta_i = 1 \\ \dfrac{\pi_i \tau_i J_i^{\tau_i - 1} \nu_i \rho_i \cosh(\rho_i)}{\pi(1 - \pi_i J_i^{\tau_i})\sigma_i K_i} & \text{if } \delta_i = 0 \end{cases}$$

$$s_{i,3} = \begin{cases} \dfrac{1}{\nu_i} - \dfrac{2\nu_i \sinh^2(\rho_i)}{K_i} + (\tau_i - 1)\dfrac{\sinh(\rho_i)}{\pi J_i K_i} & \text{if } \ \delta_i = 1 \\ \dfrac{-\pi_i \tau_i J_i^{\tau_i - 1} \sinh(\rho_i)}{\pi(1 - \pi_i J_i^{\tau_i}) K_i} & \text{if } \delta_i = 0 \end{cases}$$

$$s_{i,4} = \begin{cases} \dfrac{1}{\tau_i} + \log(J_i) & \text{if } \ \delta_i = 1 \\ \dfrac{-\pi_i J_i^{\tau_i} \log(J_i)}{(1 - \pi_i J_i^{\tau_i})} & \text{if } \delta_i = 0 \end{cases}$$

$$s_{i,5} = \begin{cases} \dfrac{1}{\pi_i} & \text{if } \ \delta_i = 1 \\ \dfrac{J_i^{\tau_i}}{\pi_i J_i^{\tau_i} - 1} & \text{if } \delta_i = 0 \end{cases}$$

where $J_i = 1/2 + \pi^{-1} \arctan[\nu_i \sinh(\rho_i)]$ and $K_i = \nu_i^2 \sinh^2(\rho_i) + 1$.

By solving the equations below it is possible to obtain the maximum likelihood estimate of θ'

$$\left.\frac{\partial l(\boldsymbol{\theta'})}{\partial \boldsymbol{\beta_1}}\right|_{(\boldsymbol{\theta'} = \hat{\boldsymbol{\theta'}})} = 0, \ \left.\frac{\partial l((\boldsymbol{\theta'}))}{\partial \boldsymbol{\beta_2}}\right|_{(\boldsymbol{\theta'} = (\hat{\boldsymbol{\theta'}})} = 0, \ \left.\frac{\partial l(\boldsymbol{\theta'}))}{\partial \boldsymbol{\beta_3}}\right|_{(\boldsymbol{\theta'} = \hat{\boldsymbol{\theta'}})'} = 0,$$
$$\left.\frac{\partial l(\boldsymbol{\theta'}))}{\partial \boldsymbol{\beta_4}}\right|_{(\boldsymbol{\theta'} = \hat{\boldsymbol{\theta'}})} = 0, \ \left.\frac{\partial l(\boldsymbol{\theta'}))}{\partial \boldsymbol{\beta_5}}\right|_{(\boldsymbol{\theta'} = \hat{\boldsymbol{\theta'}})} = 0,$$

and $\hat{\boldsymbol{\beta}}_k = \hat{\beta}_{0k}, \hat{\beta}_{1k}, ..., \hat{\beta}_{m_k k})^\top$ is the MLE of $\boldsymbol{\beta_k} = (\beta_{0k}, \beta_{1k}, ..., \beta_{m_k k})^\top$ with $k = 1, ..., 5$. There is no closed-form for these expressions, we then maximise the likelihood numerically using the Newton-Raphson algorithm [26]. When the sample of individuals is large enough and since the estimator $\hat{\boldsymbol{\theta'}}$ is consistent, then $\hat{\boldsymbol{\theta'}}$ approximately follows a multivariate normal distribution with an asymptotic mean $\boldsymbol{\theta'_0}$ and an asymptotic variance $I_n^{-1}(\boldsymbol{\theta'_0})$ which is the inverse of the Fisher information matrix. By the large sample properties of MLE, we can write

$$\hat{\boldsymbol{\theta'}} \sim_a N_5(\boldsymbol{\theta'}_0, \mathbf{I_n^{-1}}(\boldsymbol{\theta'_0})) \tag{4.2.5}$$

When the expression $\mathbf{I_n}(\boldsymbol{\theta'_0})$ is not available it can be replaced by the matrix $\mathbf{I_n}(\hat{\boldsymbol{\theta}}'_\mathbf{n})$. Here it is the case because we have no closed-form due to the censored observation. Thanks to this asymptotic property, confidence intervals can be calculated and hypothesis tests performed.

Chapter 5

Simulation

In this chapter, we perform a simulation study to evaluate the models presented in the previous chapters and measure if they produce accurate results. Data simulation is an approach that involves generating synthetic data set using controlled parameters. This allows us to evaluate whether the models are able to correctly estimate these known parameters. The various methods were implemented using **R** software.

The objective of this chapter is to first present the data simulation methods for the two models. For the 3-component mixture cure model we used the methodology proposed by [28] and we adapted the **simsurv** function of the **simsurv** package. For the second model, we used the methodology of [30] and implemented the code provided by the authors.

For each model, a Monte Carlo simulation experiment has been performed, for different sample sizes. We first performed the simulations without considering covariates and then repeated them by taking into account a binary covariate. For each of the simulated samples, we applied the maximum likelihood parameter estimation methods discussed in the previous two chapters. To estimate the parameters of the 3-component-mixture model, we have modified the **weibullRMM_SEM** function of the **mixtools** package. The method behind this function has been proposed by Bordes and Chauveau [20]. For the ELSCcr model, we have reused the codes provided in the article [30]. For every Monte Carlo experience, we calculated the mean of the estimates, variance, bias and mean square errors. A graphical visual analysis was carried out to examine the results.

5.1 3-component mixture cure model simulation

5.1.1 Simulation methodology

For the simulation of the data we have modified the **simurv** function from the package **simsurv** [28] in order to produce data corresponding to the 3-component mixture cure model. The **simsurv** function originally simulated survival data from standard parametric distributions (Weibull, exponential and Gompertz) and the two-component mixture model. It can also take into account the effects of covariates. The methodology used to simulate

the data is described below.

In probability theory, according to the probability integral transform, let X be a random variable with continuous F the cumulative distribution function, then $F(F^{-1}(u)) = u$ for all $u \in [0,1]$, and as consequence, the distribution of $U = F(X)$ is uniform on $[0,1]$ [18]. By the relationship $S(t) = 1 - F(t)$ and if the event time T_i occurs in continuous time, the survival function is a uniform random [0,1] . For the ith individual, we have

$$S_i(T_i) = U_i \sim U(0,1) \tag{5.1.1}$$

It is then proposed to find the solutions of the equation

$$S_i(T_i) - U_i = 0 \tag{5.1.2}$$

by considering T_i as unknown. In order to determine the root of this equation, we then use Brent's univariate root finder [4]. This can be done using **uniroot** from the **stats** package [25]. The proposed alternative, that we will use only in case of problems when solving the equation, is to use the function **dfsane** from the **BB** package [32]. It is possible to improve the numerical stability by applying a transformation, we have chosen in this work to keep the one proposed in the **simsurv** package which is the log transformation and we then have

$$\log(S(T_i)) - \log(U_i) = 0 \tag{5.1.3}$$
$$-H_i(T_i) - \log(U_i) = 0 \tag{5.1.4}$$

By modifying the **simsurv** function to incorporate the Weibull 3-component mixture cure model, we have the following equation to solve

$$\log(\pi_1 \exp(-(\frac{T_i}{\lambda_1})^{\gamma_1}) + \pi_2 \exp(-(\frac{t}{\lambda_2})^{\gamma_2}) + (1 - \pi_1 - \pi_2)) - \log(U_i) = 0 \tag{5.1.5}$$

in which π_1, π_2 are the mixing proportions, λ_1, λ_2 are the scale parameters and γ_1, γ_2 are the shape parameters. Only T_i is not fixed in (5.1.5). It is also possible to take covariates into account. We have decided to use the PH model presented in Chapter 2.1. We need to solve

$$\log(\pi_1 \exp(-(\frac{T_i}{\lambda_1})_1^{\gamma})^{\exp(\boldsymbol{\theta_1 x_i})} + \pi_2 \exp(-(\frac{t}{\lambda_2})_2^{\gamma})^{\exp(\boldsymbol{\theta_2 x_i})} + (1 - \pi_1 - \pi_2)) - \log(U_i) = 0 \tag{5.1.6}$$

The implementation also allows the effect of covariates on the incidence model to be taken into account. Indeed, the proportions of the different components can be given directly or a computation of the latter according to the covariates can be performed using a multilogit function as presented in the equation (3.2.2). π_1 and π_2 then become respectively $\pi_1(\mathbf{x_i}; \boldsymbol{\beta_1})$ and $\pi_2(\mathbf{x_i}; \boldsymbol{\beta_2})$.

So we created the **Mixsimsurv** function which is a modified and simplified version of the **simsurv** function.

```
Mixsimsurv( scale, shape, x, z, betas1, betas2, b1, b2,
pmix, maxt = NULL, interval = c(1E-8, 500),
rootsolver = c("uniroot", "dfsane"),
seed = sample.int(.Machine$integer.max, 1), ...)
```

The arguments, *scale* and *shape*, are respectively the vectors of size 2 of the scale (λ_g) and shape (γ_g) parameters of the Weibull latent distributions. x and z are both data frames. The first one is required and is composed of at least one variable including a vector containing the id of individuals for whom a time must be simulated. For example x = *data.frame(id = seq(1:n))* for n individuals. The other variables of the data frame are the covariates that have an influence on the latency part. The second data frame z contains the covariates having an influence on the incidence part. It is only asked if you want to use the multilogit function to obtain the proportions. When this is the case, an intercept variable must also be created. *betas1, betas2* are data frames where each column contains a vector of repeated true coefficient values for the two different latency models. The name and length of these variables match those of the data frame x. This corresponds to the $\boldsymbol{\theta_g}$ coefficients for g = 1,2 in the equation (3.3.1). *b1, b2* are data frames where each column contains a vector of repeated true coefficient values of the incidence model. The name and the length of these variables match those of the data frame z. This corresponds to the $\boldsymbol{\alpha_g} = (\beta_{0g}, \boldsymbol{\beta_g}^\top)^\top$ coefficients for g = 1,2 in the equation (3.2.2). *pmix* is a vector of size two containing the two uncured proportions only in the absence of covariates in the incidence sub-model. *maxt* is the maximum follow up time, individuals with a simulated event time larger than this argument will be right censored. *intervals*, the data are simulated to be in the given interval. *rootsolver* is the function chosen to find the roots of the equation (5.1.6). *rootfun* is the function applied for numerical stability (5.1.3). The function returns a data frame containing the id , event time and status. However, we would like to point out that sometimes the function returns errors due to the rootsolver part when using certain scale and shape parameters that we believe do not allow the root search technique to work.

5.1.2 Estimation methodology

As explained in Section 3.3, to estimate the parameters we need to perform an EM algorithm. We have based our implementation on several papers that also proposed the use of a multi-component mixture survival model [22, 29], [20]. The first one does not use the EM algorithm to determine the parameters, it uses instead directly the maximum likelihood of the incomplete data equation. We also tried to implement this method, however the test was not conclusive and did not return consistent results for our data simulations. We would like to recall that the EM algorithm is specially designed to process incomplete data, moreover, according to the literature, it allows efficient and stable estimation. However the disadvantages are that it is sensitive to the initial value (local maximum) and can be very slow. We will explain later how to determine these initial values. The second paper uses the EM algorithm to estimate the parameters of a multi-component mixture model. However, they do not take into account the covariates nor a possible part of the population that

could be cured. Our work is mainly based on the work of Bordes and Chauveau [20]. We extend the **weibullRMM_SEM** function of the **mixtools** package to take into account covariates and the cured part of the population. We decided to remove the stochastic part and focus on the basic EM algorithm. The stochastic part is a step added between E-step and M-step, whose purpose is to generate the B component label in (3.1.9). The data belonging to the same component are then grouped together and the parameters estimated on the basis of these sub-samples.

Based on the explanation of Section 3.3 illustrated in Figure 3.2.1, we implemented the EM algorithm. In the E-step , we calculate the posterior probabilities given by equations (3.3.3), (3.3.4), (3.3.5). The M-step maximizes the incidence submodel using the **optim** function from the **stats** package. This function allows to determine the parameters that maximize the multilogit log-likelihood function taking into account the weight associated to the posterior probability. For the estimation of the parameters of the latency part, we use the **survreg** function of the **survival** package for the appropriate model. In order to appropriately weight each data according to its posterior probability we use the option *weight* of **survreg**.

The function created is

EM (**t**, d,X = NULL,Z = NULL , pmix = NULL, shape = NULL, **scale** = NULL, maxit = 200, maxit.survreg = 200, epsilon = 1e−03, averaged = TRUE)

The argument t is the vector of lifetime time such that $t = min(t_{true\ eventtime}, c)$ if c is the censoring time. d is the vector of censoring indicator. X and Z are covariate matrices. Z must be provided with at least one column equivalent to *rep(1,n)* for the intercept where n is the number of observation. *pmix, shape , scale* are respectively the vectors of initial values for the mixture probabilities, the scale and shape parameters of the Weibull distribution. *maxit* is the maximum number of iterations performed between the E-step and the M-step. *maxitsurvreg* is the maximum number of iterations allowed for the **survreg** function. *epsilon* is the threshold to reach. *averaged = true* means that the parameters are updated by taking the average of the parameter estimates obtained in the previous iterations along with the new value obtained at the M-step.

The function returns a list with:

- t : event time
- d : status
- *coef1*: coefficient vector for the PH model of the first component. If there is no covariate, the value will be 0.
- *coef2* : coefficient vector for the PH model of the second component. When there is no covariate, the value will be 0.
- *coef_incidence1* : coefficients of the incidence model associated with the first component $\boldsymbol{\alpha_1}$. When there are no covariates, returns β_{01}.

- *coef_incidence2* : coefficients of the incidence model associated with the first component $\boldsymbol{\alpha_2}$. When there are no covariates, returns β_{02}.
- *scale* : vector of length 2 with the scale parameters of the two Weibull components.
- *shape* : vector of length 2 with the shape parameters of the two Weibull components.
- *loglik*: likelihood of the model for the last iteration.
- *posterior* : matrix of posterior probabilities for the last iteration.
- *all.loglik* : matrix listing the likelihoods of each iteration.
- *all.scale* : matrix listing the scale vector of each iteration.
- *all.shape* : matrix listing the shape vector of each iteration.

To obtain the mixing proportions, apply the covariate-free multilogit function 3.2.2 using *coef_incidence1* and *coef_incidence2*. We also wanted to point out that the survreg and rweibull functions, both used in the EM function implementation, don't have the same parameterization. To deal with this situation, we use the relations (2.1.23).

5.1.3 Initialization methodology

As already mentioned, the EM algorithm is sensitive to initial values. On one hand, this may have an impact on the speed of the process, and on the other, it may converge towards a local maximum, rather than a global maximum. We have therefore set up a procedure to determine these values: the uniform binning approach [20]. Given the context, the data are separable in three groups: first wave of event, second and the cured part. In order to achieve this, we use **geom_histogram** from the **ggplot2** package [33]. This function allows to divide the timescale into three intervals and to retrieve information about interval sizes. To calculate the mixture proportions, we count the number of data belonging to each interval and divided by the total number of data. To estimate the parameters of each component Weibull survival, we use the **survreg** function on the data contained in each bin. This method has been automated and implemented in our simulations.

5.1.4 Monte Carlo Experiments

We conduct Monte Carlo simulation experiments with 1000 iterations to validate our methodology. We realized it for different sample sizes: 50, 200, 500. In this section, we will only take into account the administrative censoring due to the cured proportion in the samples. In Section 5.3, we'll take random censoring into account. At the end of each experiment, we calculated the mean, the empirical variance, the bias and the mean squared error of all result estimated parameters. In a first simulation study, we did not take into account the covariates, in a second one we included a binary covariate.

5.1.4.1 Simulation study 1

For this first experiment, we simulated 1000 samples from a 3-component mixture cure Weibull. The first component follows a Weibull distribution with parameters scale $\lambda_1 = 5$ and shape $\gamma_1 = 0.5$ while the second one follows a Weibull distribution with parameters scale $\lambda_2 = 900$ and shape $\gamma_2 = 10$. The mixing proportions are respectively for each of the components $\pi_1 = 0.5$, $\pi_2 = 0.3$ and for the cured part $\pi_3 = 1 - \pi_1 - \pi_2 = 0.2$. This means that 20% of data is administratively censored. For one simulation, when n = 500, we obtained 1054 as largest event time and 1500 as largest censored time.

For $n = 50$ and $n = 200$, the estimation of the scale parameter of component 1 failed respectively 7 and 1 times and returned NA. We decided to exclude these results from the analysis. There were no failures for $n = 500$. It can be understood that the sample size has an impact on the estimation performance of the EM algorithm. Looking more closely at the simulated samples providing these failures of the EM algorithm, we see that the failure is due to a problem in the posterior probability computation. A possible alternative that we have tested is to use the stochastic EM (SEM) algorithm [20]. This method consists in adding a stochastic step called S-step between the M-step and the E-step. This step aims at creating the missing variable **B** (3.1.2) which is the group indicator. The M-step does not maximize the likelihood of incomplete data anymore but the likelihood of complete data. We can then obtain the estimation of the parameters of the incidence sub-model by using the **multinom()** function of the **nnet** package. The parameters of the latent sub-model are obtained by using the function **survreg()** for each group. This method gives rather good estimates. Table .1.1 in Appendix shows the results obtained for the problematic samples.

Let's continue the analysis without considering the samples where the EM algorithm failed. Table 5.1.1 shows the means, biases, variances and MSEs of the estimates for each Monte Carlo experiment. It can be seen that the averages' estimates for each sample size appears to be relatively close to the true parameters. Consistent with asymptotic theory, the biases decrease with increasing sample size. The same observation can be drawn for the MSEs and the variances. The variance of the parameter λ_2 is quite high, especially for sample size n = 50.

For each sample size, we randomly selected 150 results. For each of them we have plotted the survival curve associated with the estimated parameters on Figures 5.1.1, 5.1.2, 5.1.3. When n = 50, the survival function for the estimated parameters seems to be very different from the survival function for the true parameters, and they are also very different from each other. This reflects the variability of estimates when $n = 50$. We notice that the larger the sample size, the more the light green estimated survival curves tend towards the darker line that represents the true survival curve for the true parameters. We can therefore conclude that the estimation improves with the sample size. We can notice that the estimated curves for $n = 200$ and $n = 500$ follow well the behavior of the true parameters curve. The figures also show the estimated Kaplan-Meier survival curves. We can see that the Kaplan-Meier curve tends to underestimate the cure rate. We observe that the levels of the plateaus for the estimated survival function when $n = 500$ (where the survival probability is constant) correspond to the mixing proportions.

In Appendix .1.1, we find different graphs that represent the boxplot and histogram of each estimates of parameter for each sample size. We notice that among the thousand iterations there are some outliers. Outliers are visually more numerous and far from the true value when the sample size is 50. The estimates become more precise when the sample size increases as expected by the theory. Then, the different graphs seem to show that the estimates of the parameters seem asymptotically normally distributed. Figure .1.13 in Appendix, shows the estimated Kaplan-Meier survival curve and the true survival curve associated with a sample that provides outliers. It can be seen that it is the sample that does not correspond to the true parameter curve. In fact, for this sample, the events of the second wave arrive earlier. The samples providing outliers appear pathological. There doesn't seem to be label switching (explained in Section 3.1).

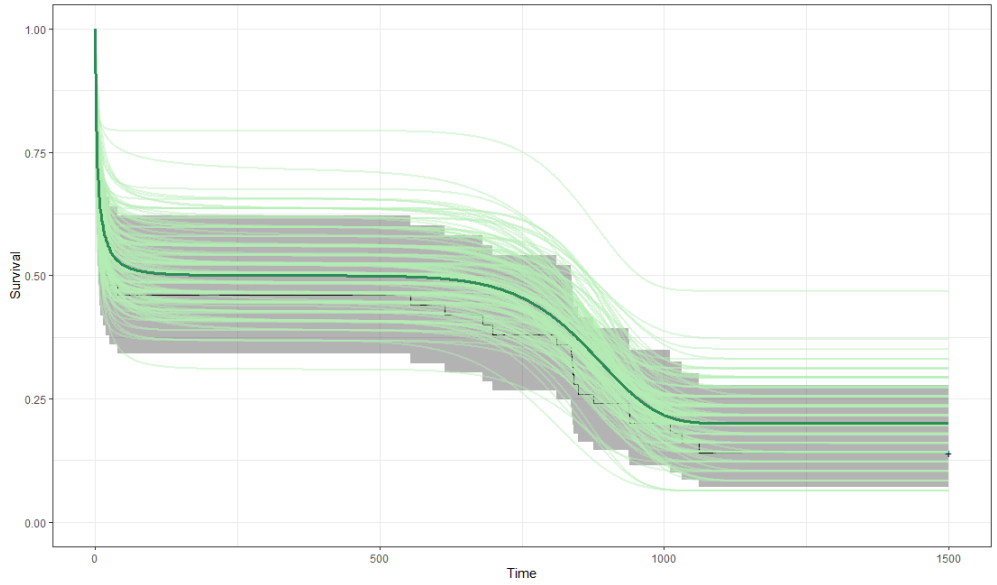

Figure 5.1.1: Sample survival curve estimated using the EM algorithm for $n = 50$ for a Weibull 3-component mixture cure model. In dark green, the survival curve for the true parameters: $\pi_1 = 0.5, \pi_2 = 0.3, \pi_3 = 0.2, \lambda_1 = 5, \lambda_2 = 900, \gamma_1 = 0.5, \gamma_2 = 10$. In black the estimated Kaplan-Meier survival curve for a random sample.

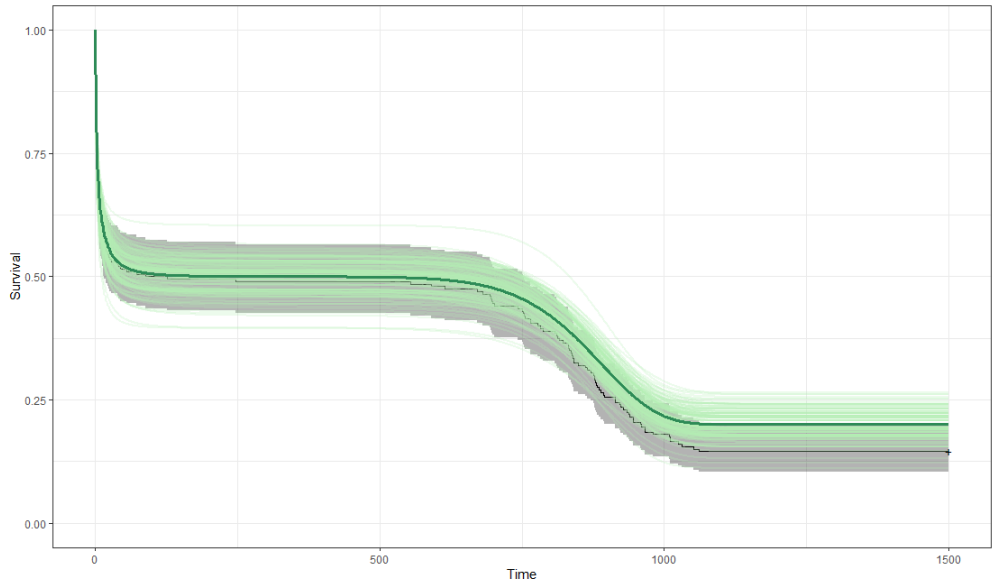

Figure 5.1.2: Sample survival curve estimated using the EM algorithm for $n = 200$ for a Weibull 3-component mixture cure model. In dark, the survival curve for the true parameters: $\pi_1 = 0.5, \pi_2 = 0.3, \pi_3 = 0.2, \lambda_1 = 5, \lambda_2 = 900, \gamma_1 = 0.5, \gamma_2 = 10$. In black the estimated Kaplan-Meier survival curve for a random sample.

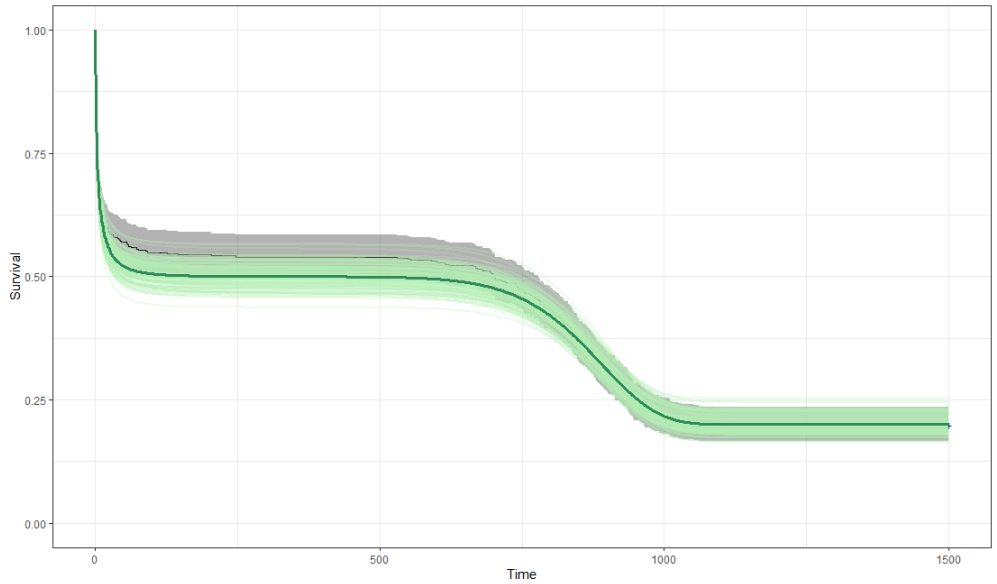

Figure 5.1.3: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model. In dark green, the survival curve for the true parameters: $\pi_1 = 0.5, \pi_2 = 0.3, \pi_3 = 0.2, \lambda_1 = 5, \lambda_2 = 900, \gamma_1 = 0.5, \gamma_2 = 10$. In black the estimated Kaplan-Meier survival curve for a random sample.

n = 50				
	Mean	Bias	Variance	MSE
λ_1	5.33	0.33	5.36	5.47
λ_2	898.05	-1.95	620.75	624.57
γ_1	0.53	0.03	0.01	0.01
γ_2	11.06	1.06	7.36	8.50
π_1	0.47	-0.03	0.00	0.00
π_2	0.34	0.04	0.00	0.00
n = 200				
	Mean	Bias	Variance	MSE
λ_1	5.06	0.06	1.13	1.13
λ_2	899.37	-0.63	157.31	157.71
γ_1	0.51	0.01	0.00	0.00
γ_2	10.24	0.24	1.15	1.21
π_1	0.48	-0.02	0.00	0.00
π_2	0.34	0.04	0.00	0.00
n = 500				
	Mean	Bias	Variance	MSE
λ_1	5.04	0.04	0.43	0.43
λ_2	899.38	-0.62	58.59	58.98
γ_1	0.50	0.00	0.00	0.00
γ_2	10.10	0.10	0.44	0.45
π_1	0.48	-0.02	0.00	0.00
π_2	0.34	0.04	0.00	0.00

Table 5.1.1: The means, biases, variances and MSE for the 3-component mixture cure Weilbull model. The true parameters are : $\lambda_1 = 5$, $\lambda_2 = 900$, $\gamma_1 = 0.5$, $\gamma_2 = 10$, $\pi_1 = 0.5$, $\pi_2 = 0.3$

5.1.4.2 Simulation study 2

In this second Monte Carlo experiment, we will take into account a covariate. Indeed, we will consider that the simulated event times come from two distinct groups. To do this we will create a covariate coming from a binomial distribution such as $X \sim Bin(1, 0.5)$. This kind of situation occurs for example when a sample is separated into two equal groups, one following a treatment and the other having a placebo. We simulated 1000 samples from a 3-component mixture cure Weibull distribution where the latency model and the incidence model take into account the covariate X. The first component follows a Weibull distribution with parameters scale $\lambda_1 = 200$ and shape $\gamma_1 = 10$, while the second one follows a Weibull distribution with parameters scale $\lambda_2 = 900$ and shape $\gamma_2 = 25$. The mixture proportions are obtained by the multilogit function. For the first mixture proportion we have the coefficient $\beta_{01} = 1$, $\beta_1 = 0.5$ and for the second we have $\beta_{02} = 0.2$, $\beta_2 = 0.5$. When X=1 this corresponds to $\pi_1 = 0.598$ for the first mixing proportion 1 to $\pi_2 = 0.269$ for the second

mixing proportion and $\pi_3 = 0.133$ for the cured proportion. When X=0 this corresponds to $\pi_1 = 0.55$ for the first mixing proportion 1 to $\pi_2 = 0.247$ for the second mixing proportion and $\pi_3 = 0.203$ for the cured proportion. For X = 1, there is 13.3% administrative censoring and for X = 0 , there is 20.3% administrative censoring. For the coefficients of the Weibull PH model we have $\theta_1 = 0.5$ for the first component and $\theta_2 = 1.5$ for the second. For one simulation, when n = 500, we obtained 965.27 as largest event time and 1500 as largest censored time.

In the Monte Carlo experiment, for $n = 50$ the EM algorithm failed 95 times, for $n = 200$ 5 times and for $n = 500$ it never failed. The sample size seems to have an influence on the ability of the EM algorithm to estimate parameters. The sample size $n = 50$ is very small when covariates are taken into account, as the EM has to estimate parameters for groups of size $n = 25$, which would explain the estimation failures. In the following analysis, those samples for which the EM algorithm failed will not be included.

Table 5.1.2 shows the mean, bias, variance and MSE of the estimates for each sample size. We observe that the means of the estimates for sample size $n = 50$ are slightly less close to the true value of the model parameters than for larger sample sizes. This implies a greater bias. The bias tends to be zero as the sample size increases. When we look at the variance, it's quite high for sample $n = 50$, especially for the shape parameter γ_2. This result was expected because, as we said earlier, this sample size seems very small to be able to estimate the parameters correctly. The variance decreases with increasing sample size. The same observation can be made for MSE. The results are consistent with asymptotic theory.

For each sample size, 150 parameter estimates were randomly selected. The corresponding survival functions have been plotted on Figures 5.1.4, 5.1.5, 5.1.6. The survival function for the true parameters is also shown in darker color. The estimated Kaplan-Meier survival curve for a randomly selected sample have also been plotted on these same figures. For greater clarity, the graphs of the survival curves for the estimates and the estimated Kaplan-Meier survival curve are available separately on the Figures .1.14, .1.15, .1.16,.1.17 , .1.18 ,.1.19, in Appendix .1.2 For the sample size $n = 50$, we observe considerable variability between the different survival functions associated with estimation. For this sample size, the estimates fail to fit the true parameter survival function correctly. When $n = 200$, the fit of these different estimated curves to the true curve improves. When the sample size is $n = 500$, the estimated survival functions are really close to the true survival function. The different plateaus for $n = 500$ correspond to the different mixing proportions for the two covariates.

Appendix .1.2 displays the different boxplots and histograms for each parameter and sample size. The boxplots show the desired results: as the sample size increases, the range decreases and the median is closer to the true value of the parameters. There are some outliers. There are the most for parameters when $n = 50$. This result was to be expected, given that we consider this sample size to be too small for a correct estimate. On the boxplot of the shape parameter γ_2, we observe more data far from the median for sample $n = 50$, which would also explain the greater variability in the estimates of this parameter. For the estimates of λ_1, θ_1 θ_2 and β_1 for samples of size $n = 200$ and $n = 500$, we have

a value very far from the median which surprises us. For each parameter, this estimate comes from the same samples. It's these samples that seem pathological. By examining the estimated Kaplan-Meier survival curve for the samples concerned on Figures .1.40 and .1.41, we realize that this does not correspond to what the model is capable of modeling. As for the histograms, we can say that the estimated parameters follow a normal distribution

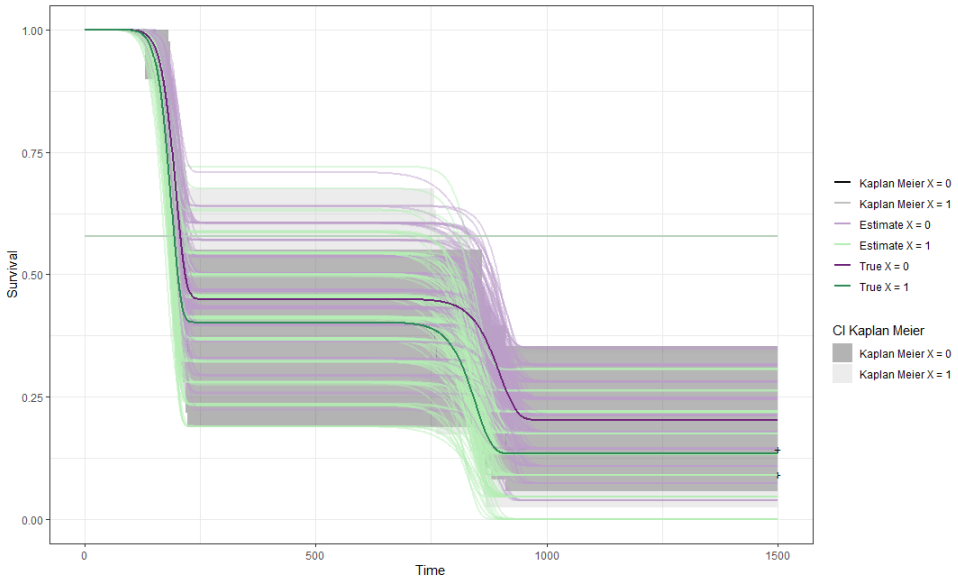

Figure 5.1.4: Sample survival curve estimated using the EM algorithm for $n = 50$ for a Weibull 3-component mixture cure model. In darker colors, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$, $\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm estimate correctly.

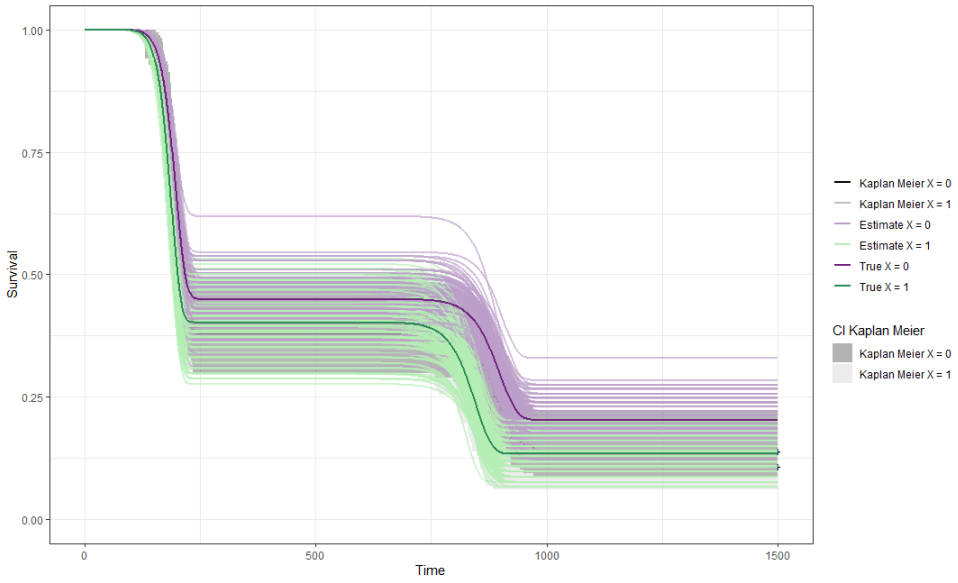

Figure 5.1.5: Sample survival curve estimated using the EM algorithm for $n = 200$ for a Weibull 3-component mixture cure model. In darker colors, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$,$\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm estimate correctly.

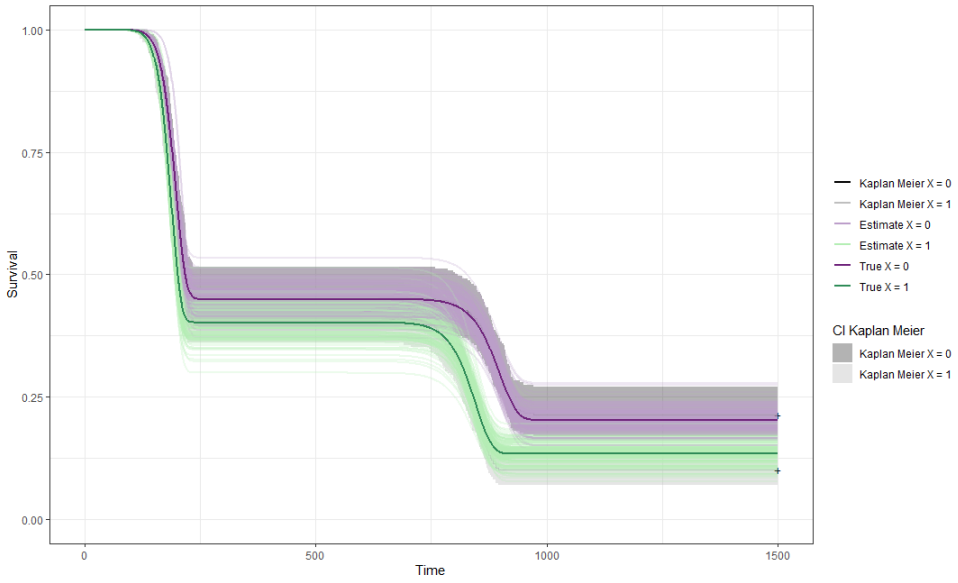

Figure 5.1.6: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model. In darker colors, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$,$\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm estimate correctly.

n = 50				
	Mean	Bias	Variance	MSE
λ_1	200.370	0.370	26.830	26.967
λ_2	898.397	-1.603	241.792	244.361
γ_1	10.822	0.822	3.566	4.241
γ_2	30.687	5.687	212.558	244.900
θ_1	0.623	0.123	0.156	0.171
θ_2	1.814	0.314	2.603	2.701
β_{01}	1.062	0.062	0.361	0.365
β_1	0.923	0.423	4.069	4.248
β_{02}	0.252	0.052	0.484	0.486
β_2	0.906	0.406	4.040	4.205
n = 200				
	Mean	Bias	Variance	MSE
λ_1	199.889	-0.111	8.693	8.706
λ_2	899.794	-0.206	56.131	56.173
γ_1	10.222	0.222	0.700	0.750
γ_2	26.035	1.035	9.381	10.452
θ_1	0.513	0.013	0.044	0.044
θ_2	1.595	0.095	1.226	1.235
β_{01}	1.004	0.004	0.059	0.059
β_1	0.549	0.049	0.308	0.311
β_{02}	0.211	0.011	0.082	0.082
β_2	0.531	0.031	0.219	0.220
n = 500				
	Mean	Bias	Variance	MSE
λ_1	200.076	0.076	4.473	4.479
λ_2	899.805	-0.195	25.153	25.191
γ_1	10.127	0.127	0.432	0.448
γ_2	25.445	0.445	3.660	3.858
θ_1	0.516	0.016	0.020	0.020
θ_2	1.560	0.060	1.377	1.381
β_{01}	0.995	-0.004	0.031	0.031
β_1	0.536	0.036	0.235	0.236
β_{02}	0.204	0.004	0.037	0.037
β_2	0.515	0.015	0.086	0.086

Table 5.1.2: The means, biases, variances and MSE for the 3-component mixture cure Weibull model. The true parameters are : $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$,$\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$.

5.2 Mixture cure ELSC model simulation

5.2.1 Simulation methodology

For the simulation of the mixture cure ELSC model data, we have used the method described in the article [30], we will also use the codes provided in that article. Their method is explained below.

Let F be the cumulative distribution $ELSCcr(\mu, \sigma, \nu, \tau)$ and U a random variable uniformly distributed on $[0, 1]$, then $T = F^{-1}(U)$ has as cdf F. The function F^{-1} is the quantile function. Therefore from the quantile function (4.1.4) and by setting the parameters $\mu,\ \sigma,\ \nu,\ \tau$, it is then possible to generate the data of the part of the population not cured and not censored. The parameters of the ELSCcr distribution can be given directly or are obtained by taking into account the covariates using the relations (4.2.3). The cured (and censored) part of the population can be generated by the quantile function of another distribution. In the following we will use data generated from a uniform distribution $T_{cure} \sim U(200, 250)$.

The number of data generated for the uncensored part is $n \times \pi$ and for the censored part $n \times (1 - \pi)$, where n is the sample size desired. The parameter π is fixed, or is obtained using the relation (4.2.3). The associated censoring indicator is such that only the data generated for the cure part is censored.

5.2.2 Estimation methodology

The estimates of the model parameters are obtained by maximizing the log-likelihood function 4.2.4. The authors of the article have performed the maximization using the **optim** function of the package **stats**.

5.2.3 Initialization methodology

The optim function allows the user to specify initial values. The authors of the article [30] have decided to use the parameters that were used to generate the data. This choice implies that the estimates are likely be close to the true parameters thanks to this initialization.

5.2.4 Monte Carlo Experiments

As proposed in the article [30] and as done for the 3-component mixture cure model , we performed a 1000 Monte Carlo experiments for different sample size: 50, 200, 500. In this section, we will consider only administrative censoring. For each experiment we calculated the mean, the bias, the variance and the mean squared error of the parameter estimates. This was done a first time without taking into account covariates and a second time taking into account a binary covariate.

5.2.4.1 Simulation study 1

We generated 1000 samples of size $n = 50,\ 200,\ 500$ from the ELSC mixture cure distribution with parameters $\mu = 4,\ \sigma = 0.2,\ \nu = 0.1,\ \tau = 1,\ \pi = 0.8$. There is 20% administrative censoring linked to the cure part of the model. For one simulation, when $n = 500$, we obtained 188.87 as largest event time and 249.72 as largest censored time.

The table 5.2.1 shows the mean, the bias, the variance and the MSE of the estimates obtained for the different sample sizes. We see that the results even for the smallest sample size are good and close to the true value of the estimated parameters. The bias decreases globally with the sample size, as well as the variance and thus the MSE. This meets the expectations of the asymptotic theory.

For each sample size, we selected 150 estimates randomly and plotted the corresponding survival function. The result can be found on Figures 5.2.1, 5.2.2 and 5.1.3. In dark green, we find the survival function for the true parameters. These figures also show the estimated Kaplan-Meier survival curve . For $n = 50$, it is observed that when the true survival curve shows a change of behavior, the estimated curves do not reproduce it very well. When the sample size increases, the estimated survival functions are closer to those of the true parameters and adopt the same behavior.

Finally, in Appendix .1.3, we present the boxplot and histogram of the estimates for each simulation, illustrating how they vary with the sample size. On the boxplot, we observe that there are some outliers. In particular, there are more outliers when the sample size is n = 50. However, we can notice that when the sample size increases the dispersion of the estimates decreases and the median gets closer to the true value of the parameters. By observing the different histograms, we can see that the estimates seem to follow a normal distribution.

n = 50				
	Mean	Bias	Variance	MSE
μ	3.996	-0.004	0.004	0.004
σ	0.183	-0.017	0.001	0.001
ν	0.084	-0.016	0.002	0.002
τ	1.056	0.056	0.057	0.060
π	0.821	0.021	0.000	0.000
n = 200				
	Mean	Bias	Variance	MSE
μ	3.995	-0.005	0.001	0.001
ν	0.190	-0.010	0.000	0.000
σ	0.087	-0.013	0.001	0.001
τ	1.046	0.046	0.012	0.014
π	0.808	0.008	0.000	0.000
n = 500				
	Mean	Bias	Variance	MSE
μ	3.995	-0.005	0.001	0.001
ν	0.191	-0.009	0 .000	0.000
σ	0.089	-0.011	0.000	0.000
τ	1.045	0.045	0.006	0.008
π	0.806	0.006	0.008	0.000

Table 5.2.1: The means, biases, variances and MSE for the mixture cure ELSC model. The true parameters are : $\mu = 4,\ \sigma = 0.2,\ \nu = 0.1,\ \tau = 1,\ \pi = 0.8$

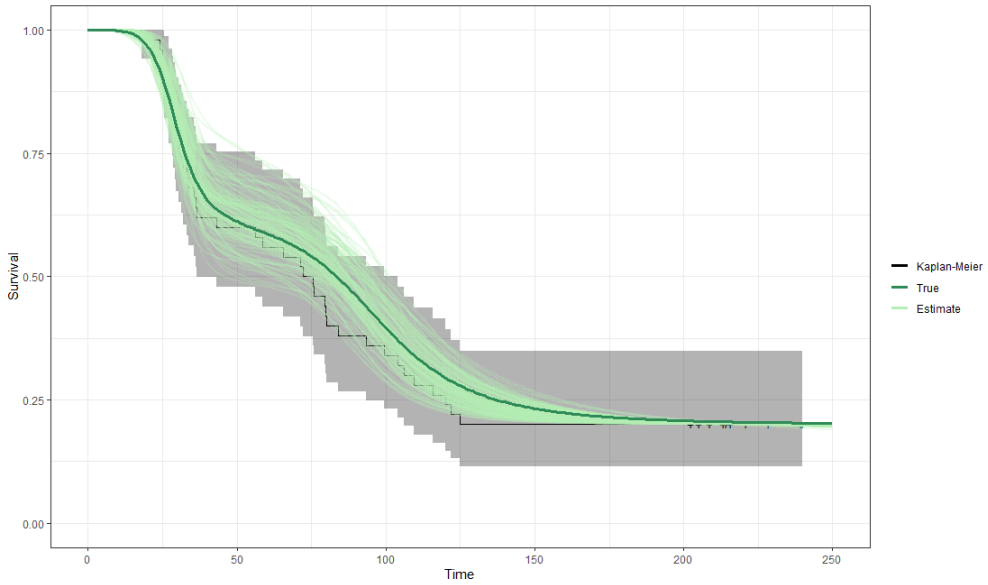

Figure 5.2.1: Sample survival curve estimated for $n = 50$ for Mixture cure ELSC model. In dark green, the survival curve for the true parameters: $\mu = 4, \nu = 0.2, \tau = 0.1, \sigma = 1, \pi = 0.8$. In black the estimated Kaplan-Meier survival curve for a random sample.

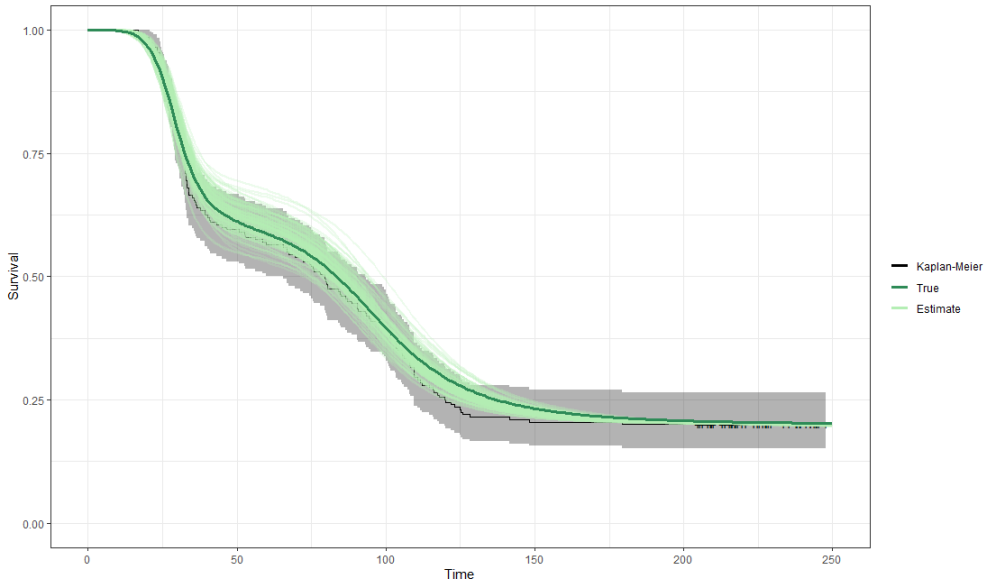

Figure 5.2.2: Sample survival curve estimated for $n = 200$ for Mixture cure ELSC model. In dark green, the survival curve for the true parameters: $\mu = 4, \nu = 0.2, \tau = 0.1, \sigma = 1, \pi = 0.8$. In black the estimated Kaplan-Meier survival curve for a random sample.

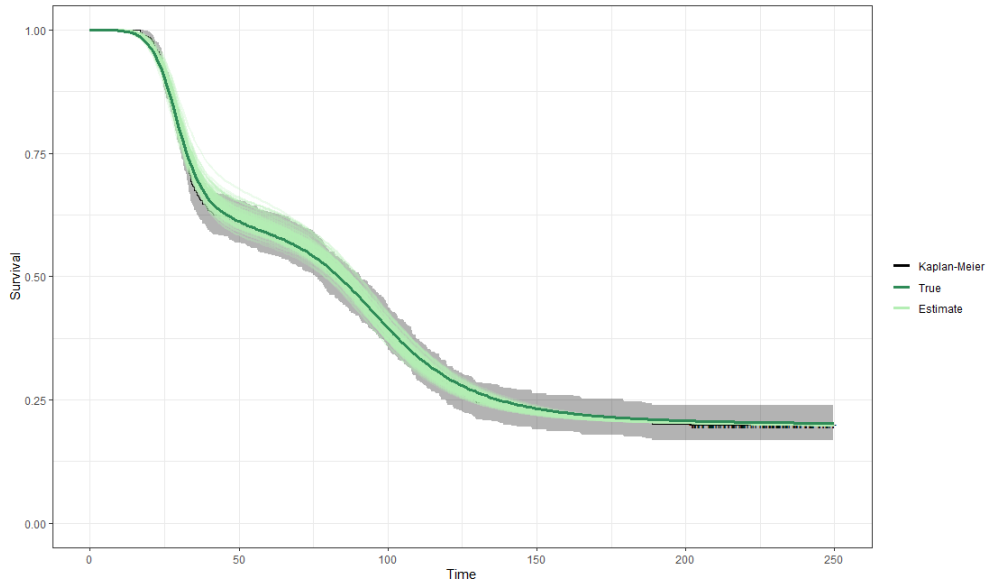

Figure 5.2.3: Sample survival curve estimated for $n = 500$ for Mixture cure ELSC model. In dark green, the survival curve for the true parameters: $\mu = 4, \nu = 0.2, \tau = 0.1, \sigma = 1, \pi = 0.8$. In black the estimated Kaplan-Meier survival curve for a random sample.

5.2.4.2 Simulation study 2

For this second Monte Carlo experiment, we introduced a binary covariate into the data simulation. When $x = 0$, the data follow the $T_0 \sim ELSCcr(3, 0.08, 0.01, 0.60, 0.88)$ distribution and when $x = 1$ the data follow the $T_1 \sim ELSCcr(3.5, 0.22, 0.22, 4.48, 0.73)$ distribution. For $X = 1$ and $X = 0$ there are respectively 27% and 12% administrative censoring. We can obtain this result when $\beta_{0,1} = 3$, $\beta_{1,1} = 0.5$, $\beta_{0,2} = -2.5$, $\beta_{1,2} = 1$, $\beta_{0,3} = -4.5$, $\beta_{1,3} = 3$, $\beta_{04} = -0.5$, $\beta_{1,4} = 2$, $\beta_{0,5} = 2$, $\beta_{1,5} = -1$ and the relations 4.2.3 become :

$$\begin{aligned} \mu_i &= 3 + 0.5x_i, \\ \sigma_i &= \exp(-2.5 + 1x_i), \\ \nu_i &= \exp(-4.5 + 3x_i), \\ \tau_i &= \exp(-0.5 + 2x_i), \\ \pi_i &= \frac{\exp(2 - 1x_i)}{1 + \exp(2 - 1x_i)} \end{aligned} \tag{5.2.1}$$

For one simulation, when $n = 500$, we obtained 138.39 as largest event time and 208.23 as largest censored time.

The results are in Table 5.2.2. As for the first experiment, we observe results close to the true parameters, a rather small bias which decreases with the sample size. In the same way, the variance and the MSE are relatively small and decrease with the size of the sample. So, as in the previous section, we have results in agreement with the asymptotic theory.

To analyze these results visually, we have randomly selected 150 estimates for each sample size and we have plotted the corresponding survival functions. These results can be seen on the figures 5.2.4, 5.2.5, 5.2.6. In darker color, we find the survival function for the true parameters. The estimated Kaplan-Meier survival curve is also found on these figures. When $n = 50$, we can see that when the true curve changes behavior, as for the experiment without covariates, the survival curves for the estimates do not quite match. But this difficulty to fit correctly disappears when the sample size increases. We can see that for $n = 500$, the curves for the estimates are close to the true survival curve.

In Appendix .1.4 are displayed the boxplot and histograms of the estimates, we observe again that when $n = 50$ the range of the estimates is greater, we also observe some outliers. The range decreases with increasing sample size. By analyzing the histograms, we see that the estimates of the parameters are normally distributed.

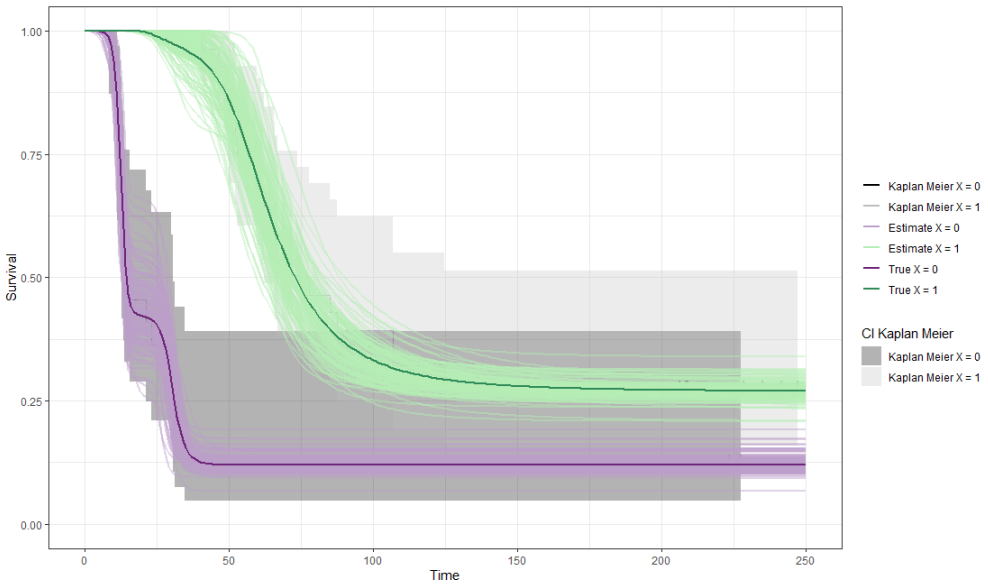

Figure 5.2.4: Sample survival curve estimated for $n = 50$ for Mixture cure ELSC model. In darker colors, the survival curve for the true parameters: $\beta_{0,1} = 3, \beta_{1,1} = 0.5, \beta_{0,2} = -2.5, \beta_{1,2} = 1, \beta_{0,3} = -4.5, \beta_{1,3} = 3, \beta_{0,4} = -0.5, \beta_{1,4} = 2, \beta_{0,5} = 2, \beta_{1,5} = -1$

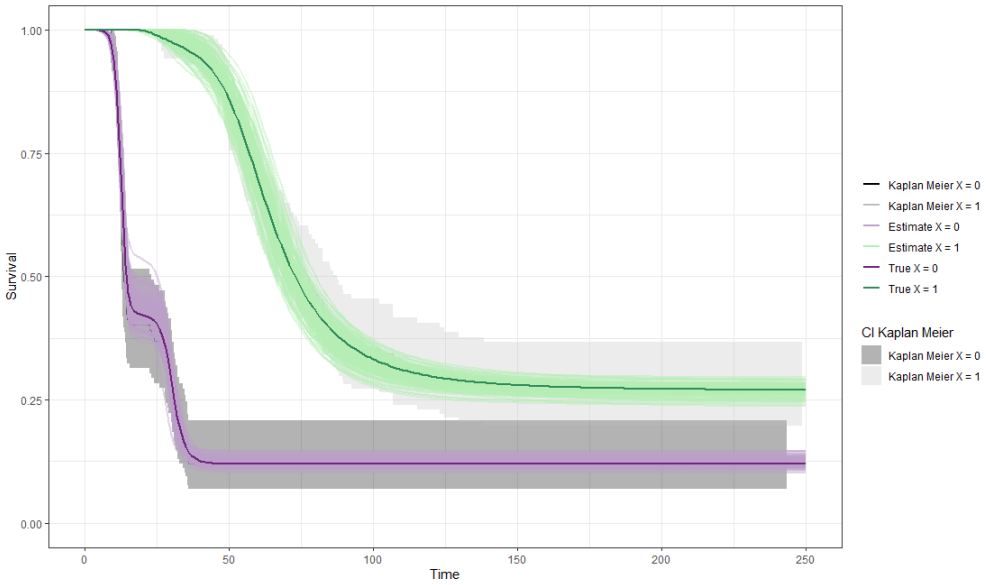

Figure 5.2.5: Sample survival curve estimated for $n = 200$ for Mixture cure ELSC model. In darker colors, the survival curve for the true parameters: $\beta_{0,1} = 3, \beta_{1,1} = 0.5, \beta_{0,2} = -2.5, \beta_{1,2} = 1, \beta_{0,3} = -4.5, \beta_{1,3} = 3, \beta_{0,4} = -0.5, \beta_{1,4} = 2, \beta_{0,5} = 2, \beta_{1,5} = -1$

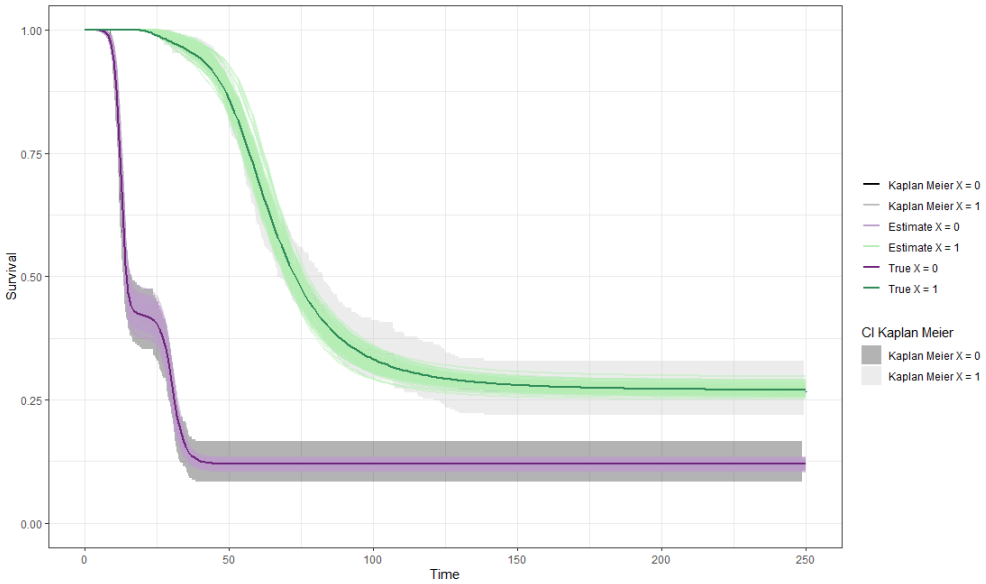

Figure 5.2.6: Sample survival curve estimated for $n = 500$ for Mixture cure ELSC model. In darker colors, the survival curve for the true parameters: $\beta_{0,1} = 3, \beta_{1,1} = 0.5, \beta_{0,2} = -2.5, \beta_{1,2} = 1, \beta_{0,3} = -4.5, \beta_{1,3} = 3, \beta_{0,4} = -0.5, \beta_{1,4} = 2, \beta_{0,5} = 2, \beta_{1,5} = -1$

n = 50				
	Mean	Bias	Variance	MSE
$\beta_{0,1}$	2.995	-0.005	0.001	0.001
$\beta_{1,1}$	0.494	-0.006	0.167	0.167
$\beta_{0,2}$	-2.603	-0.103	0.030	0.040
$\beta_{1,2}$	0.991	-0.009	0.096	0.096
$\beta_{0,3}$	-5.066	-0.566	1.012	1.332
$\beta_{1,3}$	3.599	0.599	1.237	1.595
$\beta_{0,4}$	-0.486	0.014	0.084	0.084
$\beta_{1,4}$	2.278	0.278	1.003	1.080
$\beta_{0,5}$	1.945	- 0.055	0.025	0.028
$\beta_{1,5}$	-0.948	0.052	0.045	0.048
n = 200				
	Mean	Bias	Variance	MSE
$\beta_{0,1}$	2.997	-0.003	0.000	0.000
$\beta_{1,1}$	0.572	0.072	0.022	0.027
$\beta_{0,2}$	-2.541	-0.041	0.006	0.008
$\beta_{1,2}$	0.976	-0.024	0.022	0.023
$\beta_{0,3}$	-4.681	-0.181	0.177	0.210
$\beta_{1,3}$	3.355	0.355	0.263	0.389
$\beta_{0,4}$	-0.475	0.025	0.015	0.016
$\beta_{1,4}$	1.994	-0.006	0.114	0.114
$\beta_{0,5}$	1.970	- 0.030	0.004	0.005
$\beta_{1,5}$	-0.967	0.033	0.008	0.009
n = 500				
	Mean	Bias	Variance	MSE
$\beta_{0,1}$	2.997	-0.003	0.000	0.000
$\beta_{1,1}$	0.574	0.074	0.005	0.010
$\beta_{0,2}$	-2.532	-0.032	0.002	0.003
$\beta_{1,2}$	0.981	-0.019	0.007	0.008
$\beta_{0,3}$	-4.626	-0.126	0.066	0.082
$\beta_{1,3}$	3.313	0.313	0.103	0.201
$\beta_{0,4}$	-0.475	0.025	0.007	0.007
$\beta_{1,4}$	1.975	-0.025	0.032	0.033
$\beta_{0,5}$	2.009	0.009	0.003	0.003
$\beta_{1,5}$	-1.001	-0.001	0.004	0.004

Table 5.2.2: The means, biases, variances and MSE for the mixture cure ELSC model. The true parameters are : $\beta_{0,1} = 3$, $\beta_{1,1} = 0.5$, $\beta_{0,2} = -2.5$, $\beta_{1,2} = 1$, $\beta_{0,3} = -4.5$, $\beta_{1,3} = 3$, $\beta_{0,4} = -0.5$, $\beta_{1,4} = 2$, $\beta_{0,5} = 2$, $\beta_{1,5} = -1$.

5.3 Simulation with random censoring

In this section, we examine the impact of adding random censoring to simulated samples of size $n = 500$. For each model, censored data were simulated according to an exponential distribution adapted to achieve around 10% additional censoring compared with the administrative censoring linked to the cure proportion.

5.3.1 3-component mixture cure model without covariates

For the 3-component mixture cure model without covariates, we simulated 1000 data sets where the censoring followed the exponential distribution $C \sim Exp(1/2000)$ to reach an average of 30 percent censored data versus 20 percent for samples without random censoring. Both mixture components follow a Weibull distribution, with scale parameters $\lambda_1 = 5$ and $\lambda_2 = 900$ respectively, and shape parameters $\gamma_1 = 0.5$, $\gamma_2 = 10$. The mixing proportions are $\pi_1 = 0.5$, $\pi_2 = 0.3$ and $\pi_3 = 1 - \pi_1 - \pi_2$. For one simulation, when $n = 500$, we obtained 1054.99 as largest event time and 1500 as largest censored time.

When estimating the parameters for the simulated data sets, the EM algorithm failed 293 times, returning missing values. This represents 29.3% of simulated set. In each of these cases, the problem seems to lie in the fact that the **survreg** function suddenly returns missing values, leading to error propagation. Similarly, for 72 simulated samples, the algorithm's estimation of the second component's scale parameter returned extremely large results. When estimating the other parameters of the distribution for 65 of these 72 samples, we obtained aberrant results for the parameter estimates. This result can be seen graphically on the boxplot (30%) on Figures .2.1, .2.2, .2.3, .2.4,.2.5 and .2.6 in Appendix .2.1. The figures also show the results for estimates with 20% administrative censoring. There's a big difference between the two boxplots due to outliers, especially for the parameters of the Weibull distribution of the second component. As with the missing values, it would seem that suddenly during the iterations, the function **survreg()** returns an extremely large value for the second component's scale parameter. NA's and outlier results account for 36.5% of the Monte Carlo results. The number of times the EM algorithm failed was quite high. The model parameter estimation method seems sensitive to censoring and may return outliers. However, when we focus on the median, we observe that the results for each parameter are close to the true values. These results are listed in Table 5.3.1. We can see that these results are similar to the case where there is only administrative censoring. The interquartile range is relatively small. Since the median of the estimates is close to the true value of the parameters, this suggests that the majority of estimates are close to the true value. We have opted for these statistics, as they are not affected by outliers.

We randomly selected 150 estimates and plotted the corresponding survival curves. We created 3 graphs: one displaying the selected samples superimposed on the estimated Kaplan-Meier survival curve for a simulated sample for which the EM algorithm failed Figure 5.3.1, another for which the EM algorithm returned aberrant values Figure 5.3.2, and finally for which the EM algorithm returned values close to the true values Figure 5.3.3. We have carried out this process for various samples. It seems that when the estimated

30%						
	π_1	π_2	λ_1	λ_2	γ_1	γ_2
1st quartile	0.486	0.281	4.76	894.904	0.475	9.57
Median	0.504	0.300	5.255	901.321	0.495	10.007
3rd quartile	0.521	0.317	5.847	908.149	0.514	10.734
20%						
	π_1	π_2	λ_1	λ_2	γ_1	γ_2
1st quartile	0.482	0.289	4.589	894.426	0.485	9.565
Median	0.498	0.301	5.030	899.312	0.502	9.984
3rd quartile	0.513	0.316	5.439	904.381	0.521	10.451

Table 5.3.1: Median and interquartile for the 3-components mixture cure Weibull model with 20% and 30% censoring rate. The true parameters are : $\lambda_1 = 5$, $\lambda_2 = 900$, $\gamma_1 = 0.5$, $\gamma_2 = 10$, $\pi_1 = 0.5$, $\pi_2 = 0.3$

Kaplan-Meier survival curve correctly follows the survival function of the true parameters, or overestimates the mixing proportion of the first 2 components, the EM algorithm returns good results. However, when the estimated Kaplan-Meier survival curve underestimates the first two mixing proportions, the EM algorithm fails. We can't explain why we observe this.

By analyzing the estimated survival function, we can observe that they follow the behavior of the survival curve of the true parameters. However, it can be seen that the estimated parameters for samples that returned extremely large second component parameter estimations, led to a survival curve tending to fit only the cured part.

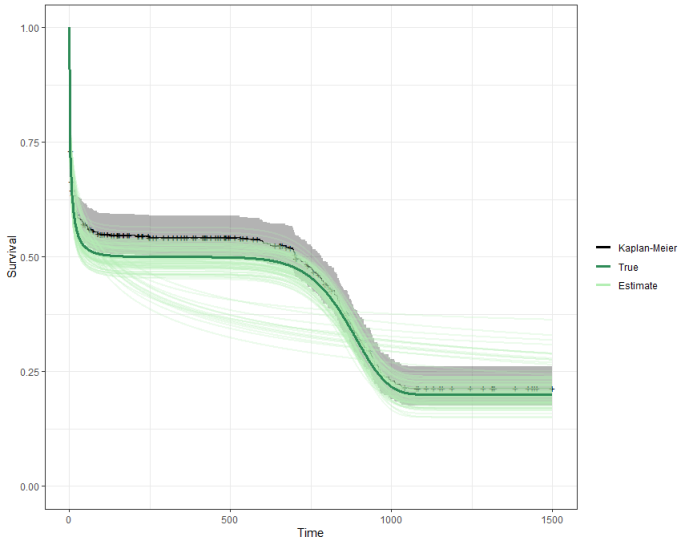

Figure 5.3.1: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model with 30% censoring. In dark green, the survival curve for the true parameters: $\pi_1 = 0.3, \pi_2 = 0.5, \pi_3 = 0.2, \lambda_1 = 900, \lambda_2 = 5, \gamma_1 = 10, \gamma_2 = 0.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm fails.

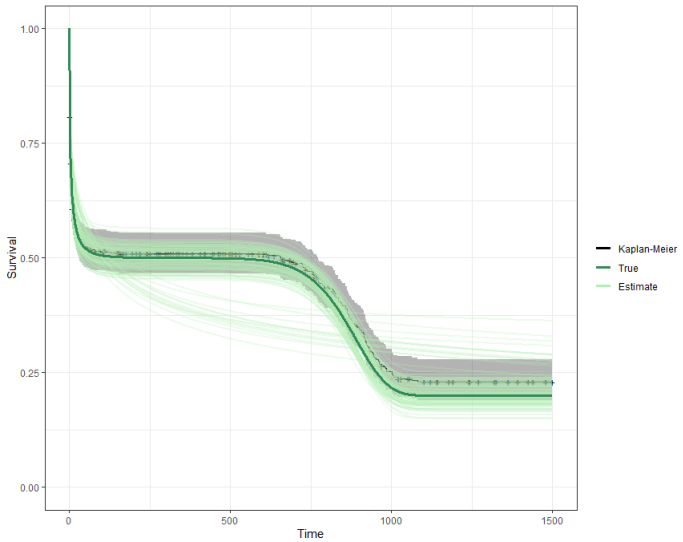

Figure 5.3.2: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model with 30% censoring. In dark green, the survival curve for the true parameters: $\pi_1 = 0.3, \pi_2 = 0.5, \pi_3 = 0.2, \lambda_1 = 900, \lambda_2 = 5, \gamma_1 = 10, \gamma_2 = 0.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm doesn't estimate correctly.

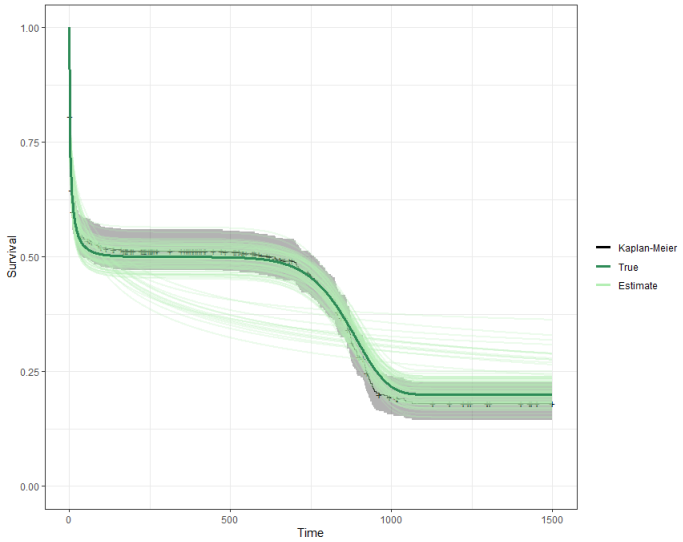

Figure 5.3.3: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model with 30% censoring. In dark green, the survival curve for the true parameters: $\pi_1 = 0.3, \pi_2 = 0.5, \pi_3 = 0.2, \lambda_1 = 900, \lambda_2 = 5, \gamma_1 = 10, \gamma_2 = 0.5$. In black the estimated Kaplan-Meier survival curve for a sample where the EM algorithm estimated correctly.

We also carried out simulation study for samples of size n = 3000, in order to assess whether sample size had an impact on estimation in the presence of random censoring and improved the algorithm's performance or not. Despite our expectations, the results obtained were not conclusive, even showing a deterioration in performance. Because of these disappointing results, we decided not to present them here.

The estimation method of the 3-component mixture cure model is not robust to random censoring. This may be due to a lack of identifiability, as information is missing to distinguish the different sub-models correctly. The first difficulty seems to be distinguishing between the component linked to the second wave and the part of the model linked to recovery. We have not yet been able to solve this problem, nor have we found any further explanations in the literature.

Given the results of this section, and despite our unsuccessful attempts to incorporate random censoring into the estimation of model parameters taking covariates into account, we have decided not to address this point in the work. However, it would be interesting to investigate the reasons why the estimation methodology doesn't seem to work in a future research

5.3.2 ELSCcr model without covariates

For the covariate-free ELSCcr model with parameters $\mu = 4$, $\sigma = 0.2$, $\nu = 0.1$, $\tau = 1$, $\pi = 0.8$, we simulated data with random censoring from the exponential distribution

$C \sim Exp(1/300)$. We simulated 1000 data sets. We achieved an average censoring rate of 29.86%.

For each dataset, we applied the parameter estimation method explained in section 5.2.1. On 60 occasions, the optimization failed to produce any values. Execution time is much faster than for the EM algorithm. ELSCcr parameter estimation without covariates looks robust to random censoring unlike the 3-component mixture cure model. Indeed, the median is close to the true value of the estimator. The same applies to the mean. Given the similarity between the median and the mean, there appears to be no outlier. Bias, variance and MSE are relatively low. These results are shown in Table 5.3.2. The results are similar to those obtained when only administrative censoring was taken into account. The same conclusions can be drawn from looking at the boxplot of parameter estimates for the model with only administrative censoring (20%) and that with random censoring (30%) on Figures .2.7, .2.8, .2.9,.2.10 and .2.11.

20%					
	Mean	Median	Bias	Variance	MSE
μ	3.994	3.994	-0.006	0.000	0.001
ν	0.192	0.191	-0.008	0 .000	0.000
σ	0.089	0.088	-0.011	0.000	0.000
τ	1.051	1.044	0.051	0.005	0.008
π	0.804	0.804	0.004	0.008	0 .000
30%					
	Mean	Median	Bias	Variance	MSE
μ	3.995	3.995	-0.005	0.001	0.001
ν	0.191	0.191	-0.009	0 .000	0.000
σ	0.089	0.088	-0.011	0.000	0.000
τ	1.045	1.043	0.045	0.006	0.008
π	0.806	0.196	0.006	0.008	0 .000

Table 5.3.2: The means,medians, biases, variances and MSE for the mixture cure ELSC model with 20% and 30% censoring rate. The true parameters are : $\mu = 4,\ \sigma = 0.2,\ \nu = 0.1,\ \tau = 1,\ \Pi = 0.8$

Once again, we randomly selected 150 of the parameter estimates and drew the corresponding survival function. This result can be seen in Figure 5.3.4. This figure also shows the estimated Kaplan-Meier survival curve for a randomly selected simulated sample. We can see that all the survival curves for the estimated parameters are close to the survival curve of the true parameters. This also corresponds to the estimated Kaplan-Meier survival curve .

All these observations lead to the conclusion that the method of estimation of the ELSCcr model seems robust to data containing random censoring. In the next section, we'll see whether this is still the case when covariates are taken into account.

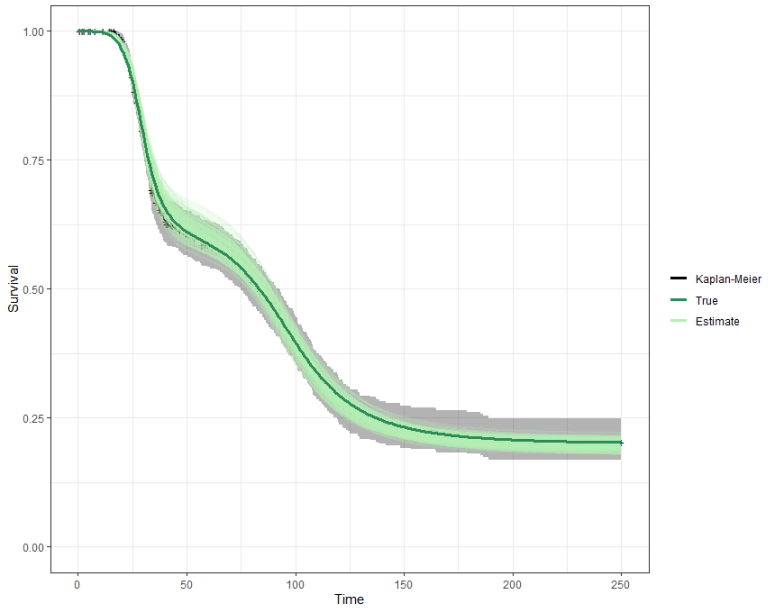

Figure 5.3.4: Sample survival curve estimated for $n = 500$ for Mixture cure ELSC model with 30% censoring. In dark green, the survival curve for the true parameters: $\pi_1 = 0.3, \pi_2 = 0.5, \pi_3 = 0.2, \lambda_1 = 900, \lambda_2 = 5, \gamma_1 = 10, \gamma_2 = 0.5$. In black the estimated Kaplan-Meier survival curve for a random sample.

5.3.3 ELSCcr model with covariates

Let's look at the addition of random censoring when the ELSCcr model takes into account the influence of covariates. The parameters chosen for the model are $\beta_{0,1} = 3$, $\beta_{1,1} = 0.5$, $\beta_{0,2} = -2.5$, $\beta_{1,2} = 1$, $\beta_{0,3} = -4.5$, $\beta_{1,3} = 3$, $\beta_{0,4} = -0.5$, $\beta_{1,4} = 2$, $\beta_{0,5} = 2$, $\beta_{1,5} = -1$, and the simulated covariates follows $X \sim Bin(1, 0.5)$. Consider random censoring $C \sim Exp(1/300)$. Samples have an average censoring rate of 29.38%. For one simulation, when $n = 500$, we obtained 138 as largest event time and 248 as largest censored time.

We have estimated the parameters for each simulated dataset. As with the model without covariates, execution time is again much faster than for the EM algorithm. First, let's take a look at the table of statistics 5.3.3 . The mean of the samples is close to the true coefficients, as is their median. Since the median and mean are close to each other, there appear to be no outliers. Looking at the bias, the variance and MSE are relatively low. However, for the two coefficients that relate the covariates to the cure rate parameter, we see that the estimate moves away from the true coefficients. Moreover, the bias, variance and MSE are higher than for estimates of the same coefficient when there is only administrative censoring.

We can observe the boxplot estimates of the various coefficients in Appendix .2.3. We note that overall the estimates are similar to those found in the context where only administrative censoring is taken into account (20%). The same observation as above can

be made, however, for the two coefficients linking the covariates to the cure rate parameter: the estimate for the data with random censoring moves away from the estimate for the data without random censoring and the true value.

30%					
	Mean	Median	Bias	Variance	MSE
$\beta_{0,1}$	2.999	2.998	-0.001	0.000	0.000
$\beta_{1,1}$	0.582	0.585	0.082	0.005	0.012
$\beta_{0,2}$	-2.529	-2.524	-0.029	0.003	0.003
$\beta_{1,2}$	0.974	0.973	-0.026	0.009	0.009
$\beta_{0,3}$	-4.627	-4.586	-0.127	0.071	0.087
$\beta_{1,3}$	3.306	3.267	0.306	0.108	0.202
$\beta_{0,4}$	-0.471	-0.474	0.029	0.007	0.008
$\beta_{1,4}$	1.945	1.943	-0.055	0.032	0.033
$\beta_{0,5}$	1.832	1.832	-0.168	0.004	0.033
$\beta_{1,5}$	-1.240	-1.238	- 0.240	0.010	0.067
20%					
	Mean	Median	Bias	Variance	MSE
$\beta_{0,1}$	2.997	2.997	-0.003	0.000	0.000
$\beta_{1,1}$	0.576	0.574	0.076	0.005	0.011
$\beta_{0,2}$	-2.532	-2.528	-0.032	0.002	0.003
$\beta_{1,2}$	0.980	0.979	-0.020	0.007	0.007
$\beta_{0,3}$	-4.628	-4.582	-0.128	0.070	0.0826
$\beta_{1,3}$	3.317	3.262	0.317	0.100	0.201
$\beta_{0,4}$	-0.474	-0.473	0.026	0.007	0.007
$\beta_{1,4}$	1.971	1.972	-0.029	0.032	0.032
$\beta_{0,5}$	2.008	1.990	0.008	0.002	0.002
$\beta_{1,5}$	-0.985	0.994	0.015	0.004	0.004

Table 5.3.3: The means, biases, variances and MSE for the mixture cure ELSC model with 20% and 30% censoring rate. The true parameters are : $\beta_{0,1} = 3$, $\beta_{1,1} = 0.5$, $\beta_{0,2} = -2.5$, $\beta_{1,2} = 1$, $\beta_{0,3} = -4.5$, $\beta_{1,3} = 3$, $\beta_{0,4} = -0.5$, $\beta_{1,4} = 2$, $\beta_{0,5} = 2$, $\beta_{1,5} = -1$.

Figure 5.3.5 displays the survival curves for the parameters estimated from 150 simulated samples. It also shows the estimated Kaplan-Meier survival curve for a randomly selected simulated sample, and the survival curve for the true parameters. It can be seen that the estimated curves fit relatively well when the covariate is 0. When the covariate takes the value 1, the estimated model overestimates the cure rate. This reflects what we had previously observed when analyzing the boxplot and the mean. It therefore seems that model estimation method in the presence of covariates is less robust to random censoring than without covariates.

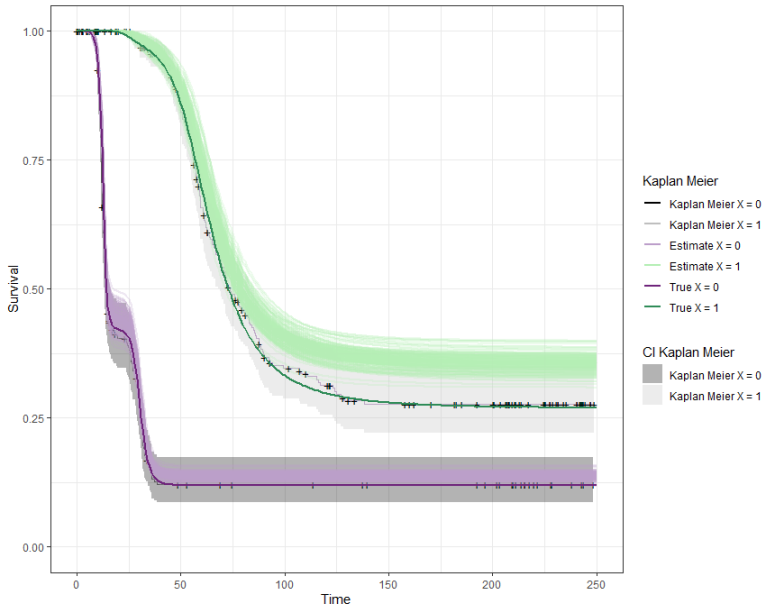

Figure 5.3.5: Sample survival curve estimated for $n = 500$ for mixture cure ELSC model with 30% censoring. In dark green, the survival curve for the true parameters: $\beta_{0,1} = 3, \beta_{1,1} = 0.5, \beta_{0,2} = -2.5, \beta_{1,2} = 1, \beta_{0,3} = -4.5, \beta_{1,3} = 3, \beta_{0,4} = -0.5, \beta_{1,4} = 2, \beta_{0,5} = 2, \beta_{1,5} = -1$. In dark and light gray respectively the estimated Kaplan-Meier survival curves for a sample when $X = 0$ and $X = 1$.

5.4 Discussion

To conclude this chapter, let us quickly discuss the results obtained. For two models, the estimation methodology allows us to estimate the parameter fairly accurately when there is only administrative censoring. When the sample size increases, each of them shows the expected asymptotic behavior. The survival curves for the estimated parameters are close to those of the true parameters for large sample size. However, the methodology employed to estimate the parameters of the ELSCcr model seems to perform better than the EM algorithm for the 3-component mixture cure model, perhaps due to the choice of the value of the initialization. The estimation seems to be closer to reality and presents less variability for this model. It should also be highlighted that estimating the parameters of the distribution ELSCcr is faster than the EM-algorithm for 3-component mixture cure distribution. Indeed, the EM algorithm is rather slow, especially when covariates are taken into account. This may be due to the less optimal choice of value initialization when covariates are taken into account, which implies a longer convergence time.

When random censoring was taken into account in parameter estimation, the methodology for the ELSCcr model performed well overall, except for the estimation of the cured proportion when covariates were taken into account. Parameter estimation of the 3-component mixture cure model by the EM algorithm is not robust in the presence of random censoring. This limitation may be due to a lack of identifiability of the model due

to the lack of information in the data. This aspect of identifiability of the 3-component mixture cure model in the presence of random censoring needs to be investigated.

In the next chapter, we will compare the two models using real data.

Chapter 6

Application to real data

In this final chapter, the ELSCcr model and the 3-component mixture cure model will be fitted to real data. Firstly, to see if both models fit the data correctly, and secondly to compare the models and see if one fits better than the other. To achieve this second objective, we will use the Bayesian Information Criterion (BIC). This criterion allows us to compare different models and determine which is the most appropriate. This criterion takes into account both the accuracy of the model and its complexity, and therefore favors simpler models. We use this criterion not because we consider it to be the most suitable, but simply because it is the most common in the literature. In this chapter we will use data without random censoring and without covariates.

6.1 Calving data

We're going to study the calving data provided in the article [30]. This data concerns the age at which Nelore heifers make their first calving. They come from the zootechnical records of a Brazilian beef cattle breeding company, located in the states of Bahia and Sao Paulo. Age at first calving is an important characteristic in the beef industry, as it has an economic impact. Indeed, when a cow has a calf earlier, it increases the profitability of cow-calf operations, and the early reproduction of Nelore heifers also increases calf weight. In literature the average age of Nelore heifers at first calving is 38 months (1157 days) [23]. However, in the industry, the age of first calving can be much earlier, depending on the genetic and nutritional strategy chosen. The event of interest is age at first calving. There will therefore be a series of heifers that will have their first calving at early puberty and a series of heifers that will have their first calving at normal puberty. The event of interest will therefore show two waves.

The sample size is $n = 1326$. The data are as follows:

- t: age in days of Nelore heifers until first calving.
- *cens*: censoring indicator (0 = no calving and therefore censored, 1 = not censored).

The study was carried out over a period of 1455 days. 32.35 % of the data are censored, of which 1.66 % is censored at time $t = 1455$ days and 30.70 % is censored at time $t = 1453$. Censoring is purely administrative. Table 6.1.1 shows the different results of the descriptive analysis of variable t and for t without the censored data. For the non censored data, the

	t (days)	$t_{no\ cens}$(days)
N	1326	897
Minimum	722	722
1st quartile	1024	872
Median	1086.5	1053
Mean	1149.514	1004.32
3rd quartile	1453	1089
Maximum	1455	1453
sd	231.24	117.64

Table 6.1.1: Descriptive analysis of variable t with and without censored data.

mean age at first calving is 1004.319 days (33 months). The median age is 1053 days (35.1 months). The standard deviation is 117.64 days. These statistics should be treated with caution, however, given the nature of the data.

The minimum age at which a heifer has had her first calving is 722 days. The highest age at which a heifer has had her first calf is 1453 days. The largest censored time is 1455 days.

Using the Kaplan-Meier estimate, the median age at which a heifer has its first calving is 1086 days with 95% confidence interval [1082, 1094]. The figure 6.1.1 shows the survival curve using the Kaplan-Meier estimator. We explained in Section 2.1 how to calculate the latter. Firstly, we can see that there are no events until 722 days. This is because heifers have to reach puberty before they can reproduce. There are two plateaus, the first between around 900 days and 1000 days, and the second after 1200 days. Visually, you'd think there is a cured fraction, but looking at the data, there are 5 heifers that had a calf at a later age. These ages are 1424, 1450, 1451, 1452 and 1453 days. We are therefore not in the cure model. We will nevertheless continue our analysis. Later, we will repeat the analysis without these 5 cows in order to have data that corresponds to a cure model.

To fit ELSCcr model to data, we reused the codes from the work by Thiago G. Ramires & al [30] . The initial estimation of the model parameters was carried out using the **gamlss()** function from the package of the same name [8]. This function provides parameter estimates for the Generalized Additive Models for Location Scale and Shape. Parameter estimation for the 3-component mixture cure model was carried out using the **EM()** function described in section 5.1.2. Initial parameter estimation was carried out in the same way as in this section 5.1.3. Table 6.1.2 shows the estimates of the different model parameters, the associated 95% confidence interval and the BIC value. For the 3-component mixture cure model, the table presents the parameters for the incident part, estimated using a logit model without covariates. By applying the logit function 3.2.2, we obtain the first mixing proportion

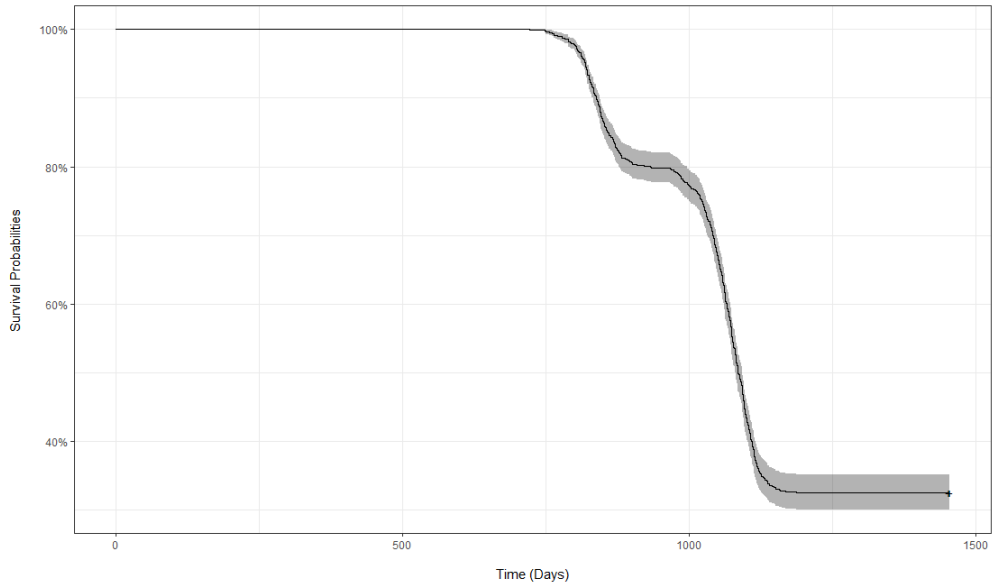

Figure 6.1.1: The estimated Kaplan-Meier survival curve for calving data

$\pi_1 = 0.15$ and the second mixing proportion is $\pi_2 = 0.52$ and hence $\pi_3 = 0.33$. It can be seen that the ELSCcr model minimizes the BIC and therefore seems to be the model that best fits the data. Figure 6.1.2 shows the estimated survival function fitted to the data. We note that the 3-component mixture cure model seems to have difficulty in fitting correctly to the first plateau of the data. When we look at the Kaplan-Meier estimate, we see that the cure rate estimate seems correct and corresponds to the percentage of censored data. The ELSCcr model, on the other hand, fits the curve relatively well and the cure rate estimate seems correct.

However, as mentioned above, some heifers had their first calf very late. This implies that we are not in a cure model. This would explain why the 3-componement mixture cure model doesn't seem to fit the data correctly.

Now let's redo the same analysis without the 5 heifers that had their calf at a later age. In this situation, the sample size is $n = 1321$, 32.48% data are censored due to administrative censoring with 1.66% at time $t = 1455$ and 30.81% at time $t = 1453$. Table 6.1.3 shows the results of the descriptive analysis.

For the non censored data, the average age at first calving is 1001 days, the median age is 1052 days. The minimum age at first calving is 772 days and the maximum is 1187 days. The largest censored time is 1455 days. Using the Kaplan-Meier estimate, the median age at which a heifer has its first calving is 1086 days with 95% confidence interval [1081, 1094]. To obtain the parameter estimates for the different models, we follow the same methodology as explained above for the complete dataset. The results of the parameter estimates and their 95% confidence intervals are shown in table 6.1.4 . By applying the logit function 3.2.2, the first mixing proportion $\pi_1 = 0.20$ and the second mixing proportion is $\pi_2 = 0.48$, we can also find $\pi_3 = 0.32$. When we look at the Kaplan-Meier estimate, we see that the cure

3 component mixture cure model			
	estimate	lower	upper
λ_1	846.168	841.550	850.785
λ_2	1089.024	1073.187	1104.861
γ_1	35.116	34.583	35.649
γ_2	11.583	10.979	12.186
β_{01}	-0.761	-1.022	-0.499
β_{02}	0.486	0.279	0.692
BIC : 12582.75			
ELSCcr			
	estimate	lower	upper
μ	6.844	6.841	6.847
σ	0.029	0.027	0.030
ν	0.022	0.016	0.003
τ	1.646	1.510	1.782
π	0.680	0.655	0.705
BIC : 11981.87			

Table 6.1.2: MLE , confidence intervals and BIC statistics : calving data

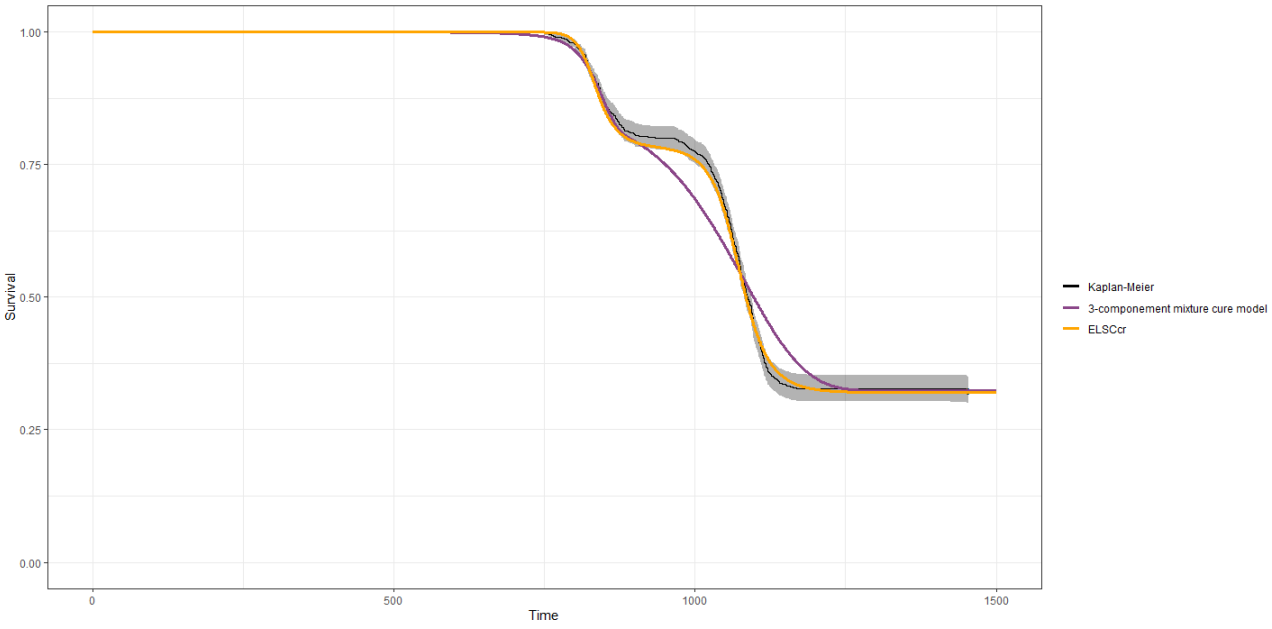

Figure 6.1.2: The estimated Kaplan-Meier survival curve , ELSCcr model, 3-component mixture cure model for calving data

rate estimate seems correct and corresponds to the percentage of censored data. Similarly, the other mixing proportions also correspond to what can be observed visually. The BIC

	t (days)	$t_{no\ cens}$(days)
N	1321	892
Minimum	722	722
1st quartile	1023	872
Median	1086	1052.5
Mean	1148.391	1001.843
3rd quartile	1453	1089
Maximum	1455	1187
sd	230.952	113.203

Table 6.1.3: Descriptive analysis of variable t with and without censored data for the reduced calving data set.

for each model is also shown. This time, the 3-component mixture cure model minimizes the BIC and seems to be the best model despite having more parameters. We can confirm this idea by observing figure 6.1.3, where it is indeed the curve of the 3-component mixture cure model that best fits the data.

3 component mixture cure model			
	estimate	lower	upper
λ_1	850.568	846.592	854.544
λ_2	1088.352	1085.078	1091.626
γ_1	29.329	26.256	32.402
γ_2	29.316	27.743	30.888
β_{01}	-0.487	-0.718	-0.257
β_{02}	0.410	0.201	0.619
BIC : 11806.58			
ELSCcr			
	estimate	lower	upper
μ	6.844	6.841	6.848
σ	0.027	0.026	0.029
ν	0.018	0.013	0.023
τ	1.617	1.482	1.751
π	0.680	0.655	0.705
BIC : 11826.36			

Table 6.1.4: MLE , confidence intervals and BIC statistics : calving data

We can conclude this section by saying that the 3-component mixture cure model is sensitive to late events. This seems a logical conclusion, given that it was designed specifically to model data with a fraction of the population cured. In the case of the population with a cured fraction without covariates, it seems to outperform the ELSCcr model. But the ELSCcr model seems more flexible and fits both datasets well.

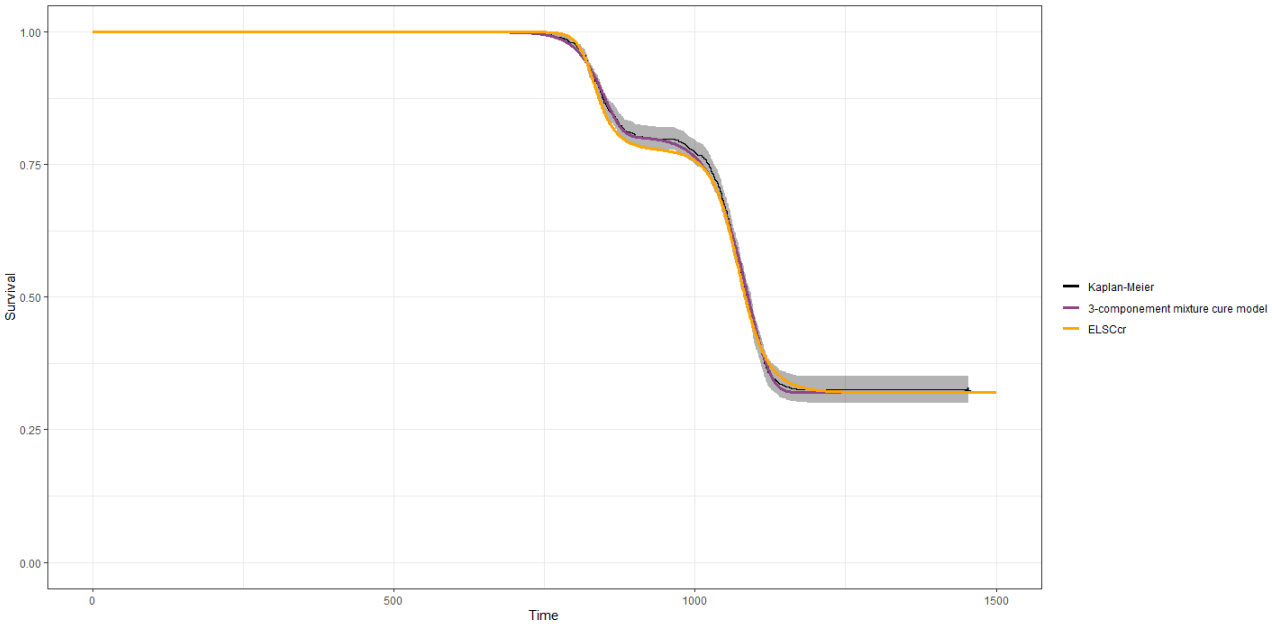

Figure 6.1.3: The estimated Kaplan-Meier survival curve, ELSCcr model, 3-component mixture cure model for reduced calving data

Chapter 7

Conclusion

In conclusion, this study on long-term survivors with a bimodal structure in the data highlights the scarcity of models that effectively incorporate this characteristic. Within the scope of our research, we thoroughly examined and compared two existing models found in the literature: the Exponentiated Log-Sinh Cauchy (ELSC) cure model and the 3-component mixture cure model.

We introduced these two models in Chapters 3 and 4. The intuition underlying these models is not quite the same. The Exponentiated Log-Sinh Cauchy cure model uses a flexible distribution capable of capturing the bimodality of the data, which is then used as the distribution for the uncured part of the mixture cure model. One of its particular features is that the integration of covariates into the model is achieved through link functions, establishing a relationship between the model parameters and the covariates. The second model takes a different approach by considering that the population is not simply divided into cured and uncured population, it assumes the presence of three distinct subgroups: the first for individuals experiencing a short-term event, the second for individuals experiencing a long-term event, and the last for cured individuals. The difficulty with this last model was that group membership was not an observed variable, so we proposed a solution using the EM framework.

One of the main results of this work was obtained in simulation experiments in the presence of administrative censoring. The methodologies used for maximum likelihood estimation proved effective for estimating the parameters of both models. However, for smaller sample sizes, the estimation seemed to perform better for the mixture cure ELSC model.

The second result found in the simulation experiments concerns the robustness to random censoring for parameter estimation of the ELSC mixture cure model. Unfortunately, we were unable to draw the same conclusions for the 3-component mixture cure model, which seems to be sensitive to the introduction of random censoring. We hypothesize this is due to the difficulty of the EM algorithm to correctly assign the censored data to the right sub-population when there is a cure part. This could be due to a lack of identifiability. It would be interesting in future research to look at these last results and try to find explanations or improve the algorithm we have implemented in order to better handle

random censoring scenarios.

The final results of this work were obtained when we wanted to fit the models to real data. When we had a cure model with only administrative censoring, the 3-component mixture cure model proved to fit the data best. This model successfully captured the underlying patterns and characteristics of the observed data. However, the ELSCcr model also fitted the data relatively well, and appeared to be less sensitive to the events that occured over an extended period.

In conclusion, it would appear that the ELSC model is a very flexible solution and outperforms the 3-component mixture cure model when modeling bimodal data with a cure proportion. However, in the case of purely administrative censoring, we still recommend adjusting the 3-component mixture cure model, as it may perform better.

For future research, it would be interesting to study the identifiability of the 3-component mixture cure model. It could also be interesting to allow the different component distributions to come from different families, or to integrate sub-models that would be semi-parametric.

Bibliography

[1] O.O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 07 1978.

[2] D.B. Rubin A.P. Dempster, N.M. Laird. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 1–22, 1977.

[3] J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, pages 15–53, 1949.

[4] R.P. Brent. *Algorithms for Minimization Without Derivatives*. Courier Corporation, 2013.

[5] José G. Dias Bruno Cardoso Alves. Survival mixture models in behavioral scoring. *Expert Systems with Applications*, pages 3902–3910, 05 2015.

[6] D. Collett. *Modelling Survival Data in Medical Research.* Chapman and Hall, New York, 2015.

[7] D. Cox, D. R. Oakes. *Analysis of Survival Data.* Chapman and Hall, New York, 1984.

[8] Robert A. Rigby D. Mikis. Stasinopoulos. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 2007.

[9] Paul Meier E. L. Kaplan. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, pages 457–481, 1958.

[10] Malcolm E. Turner Jr. Eugene H. Blackstone, David C. Naftel. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association*, pages 615–624, 1984.

[11] V.T. Farewell. Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, pages 257–262, 1986.

[12] V.T. Farwell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

[13] David Peel Geoffrey McLachlan. *Finite Mixture Models*. John Wiley Sons, Inc., 2000.

[14] Dc. McGiffin Gj. McLachlan. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research*, pages 211–226, 1994.

[15] R.P. Gage J. Berkson. Survival curve for cancer patients following treatments. *Journal of the American Statistical Association*, pages 501–515, 1952.

[16] Ross L. Prentice John D. Kalbfleisch. *The Statistical Analysis of Failure Time Data: Kalbfleisch/The Statistical.* John Wiley Sons, Inc., Hoboken, NJ, USA, 2002.

[17] Melvin L. Moeschberger John P. Klein. *Survival Analysis : Techniques for Censored and Truncated Data*. Springer-Verlag, New York, 2003.

[18] J.Segers. Lstat2410 : Copulas: models and inference. *UCLouvain*, 2021-2022.

[19] Cooray Kahadawala. Exponentiated sinh cauchy distribution with applications. *Communications in Statistics - Theory and Methods*, pages 3838–3852, 2013.

[20] D. Chauveau L. Bordes. Stochastic em algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Computational Statistics*, pages 1513–1538, 2016.

[21] C. Legrand. Lstat2230: Advanced survival models. *UCLouvain*, 2022-2023.

[22] S. Hunsberger P.S. Albert W.B. London. A finite mixture survival model to characterize risk groups of neuroblastoma. *Statistics in Medicine*, pages 1301–1314, 2009.

[23] M.H. Santos R.G. Silva G.B. Oliveira M.V.C. Ferraz, A.V. Pires. A combination of nutrition and genetics is able to reduce age at puberty in nelore heifers to below 18 months. *Animal*, pages 569–574, 2018.

[24] Waine Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, pages 945–966, 1972.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[26] S. M. Robinson R. Jennrich. A newton-raphson algorithm for maximum likelihood factor analysis. *Psychometrika*, pages 111–123, 1969.

[27] Xian Zhou Ross A. Maller. *Survival Analysis with Long-Term Survivors | Wiley*. Wiley, 1972.

[28] M. Moreno-Betancur M.J. Crowther S.L. Brilleman, R. Wolfe. Simulating survival data using the **simsurv** *R* package. *Journal of Statistical Software*, pages 1–27, 2020.

[29] M.R. Karim S.Ruhi, S. Sarker. Mixture models for analyzing product reliability data: a case study. *SpringerPlus*, page 634, 2015.

[30] Artur J. Lemonte Niel Hens Gauss M. Cordeiro Thiago G. Ramires, Edwin M.M. Ortega. A flexible bimodal model with long-term survivors and different regression structures. *Communications in Statistics - Simulation and Computation*, pages 2639–2660, 10 2020.

[31] Gauss M. Cordeiro Niel Hens Thiago G.Ramires, Edwin M. M.Ortega. A bimodal flexible distribution for lifetime data. *Journal of Statistical Computation and Simulation*, pages 2450–2470, 2016.

[32] Ravi Varadhan and Paul Gilbert. BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(4):1–26, 2009.

[33] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

.1 Appendix

.1.1 3-component mixture cure model simulation : simulation study 1

n = 50						
	λ_1	λ_2	γ_1	γ_2	π_1	π_2
simulation 90	2.88	929.67	0.604	13.376	0.512	0.294
simulation 231	8.89	887.7	0.482	12.558	0.355	0.405
simulation 511	3.29	896.25	0.37	12.77	0.458	0.317
simulation 624	7.61	960.69	0.534	14.839	0.483	0.274
simulation 812	5.1	936.5	0.675	14.028	0.391	0.284
simulation 884	4.11	874.68	0.479	12.418	0.445	0.330
simulation 924	4.05	859.57	0.492	11.579	0.469	0.258
n = 200						
	λ_1	λ_2	γ_1	γ_2	π_1	π_2
simulation 585	4.53	907.88	0.491	0.202	0.424	0.317

Table .1.1: Results of parameter estimation of simulated data with the SEM algorithm. The true parameters are : $\lambda_1 = 5$, $\lambda_2 = 900$, $\gamma_1 = 0.5$, $\gamma_2 = 10$, $\pi_1 = 0.5$, $\pi_2 0.3$.

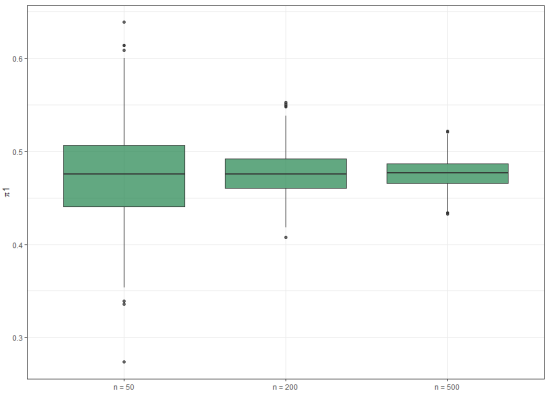

Figure .1.1: Boxplots of estimates of π_1 for 1000 replications of the EM algorithm

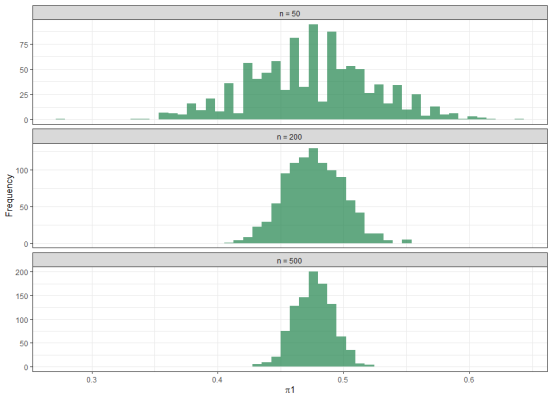

Figure .1.2: Histogram of estimates of π_1 for 1000 replications of the EM algorithm

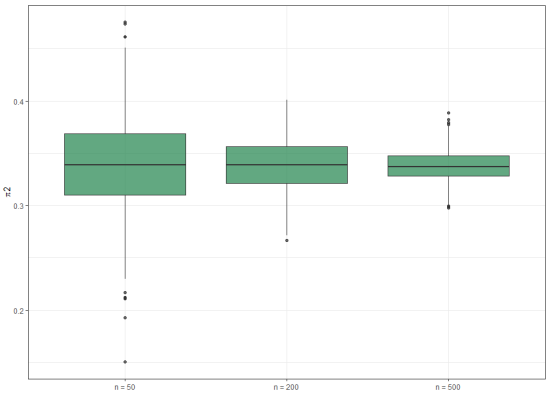

Figure .1.3: Boxplots of estimates of π_2 for 1000 replications of the EM algorithm

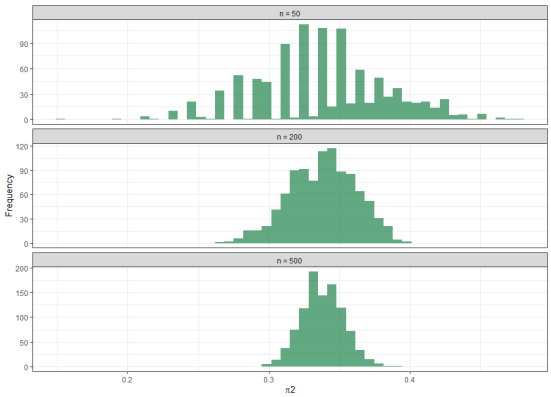

Figure .1.4: Histogram of estimates of π_2 for 1000 replications of the EM algorithm

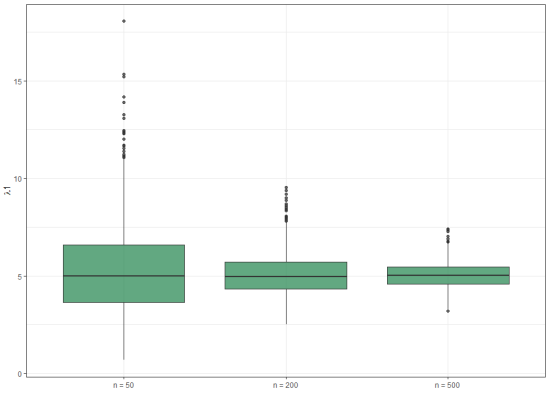

Figure .1.5: Boxplots of estimates of λ_1 for 1000 replications of the EM algorithm

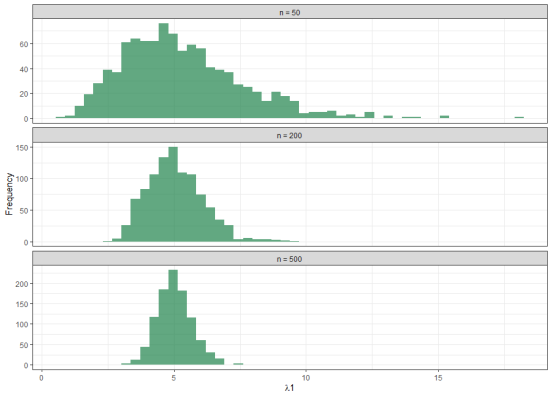

Figure .1.6: Histogram of estimates of λ_1 for 1000 replications of the EM algorithm

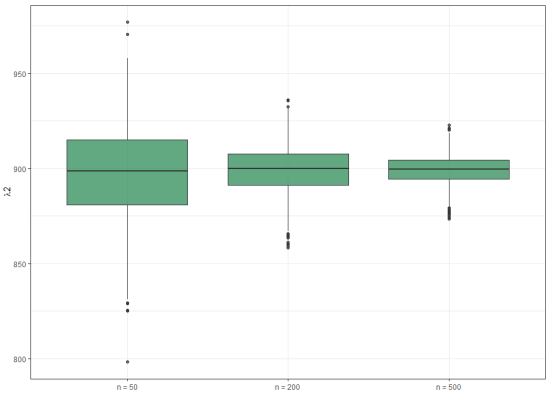

Figure .1.7: Boxplots of estimates of λ_2 for 1000 replications of the EM algorithm

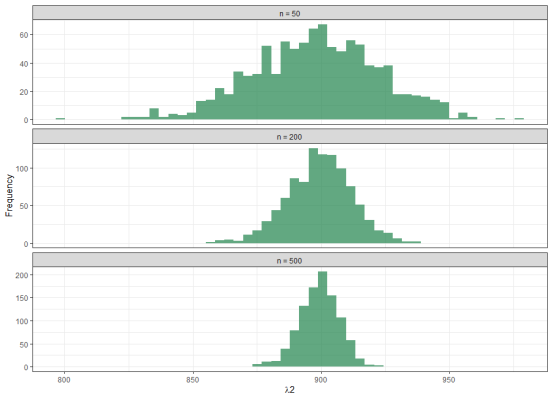

Figure .1.8: Histogram of estimates of λ_2 for 1000 replications of the EM algorithm

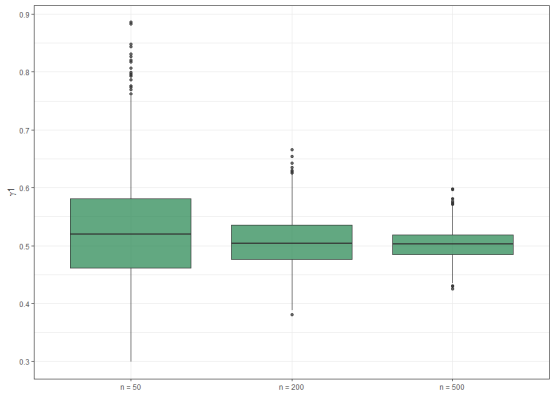

Figure .1.9: Boxplots of estimates of γ_1 for 1000 replications of the EM algorithm

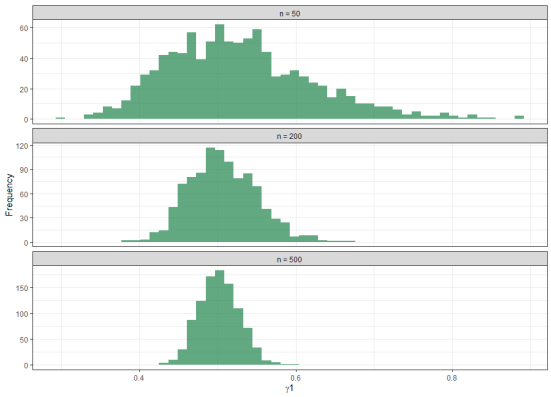

Figure .1.10: Histogram of estimates of γ_1 for 1000 replications of the EM algorithm

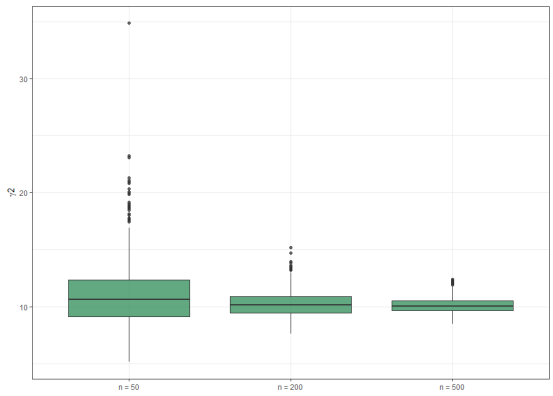

Figure .1.11: Boxplots of estimates of γ_2 for 1000 replications of the EM algorithm

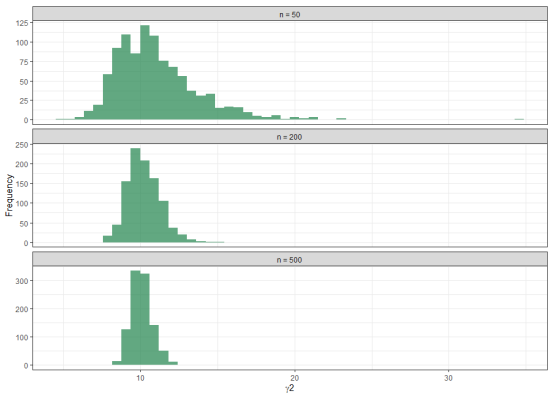

Figure .1.12: Histogram of estimates of γ_2 for 1000 replications of the EM algorithm

.1.2 3-component mixture cure model simulation : simulation study 2

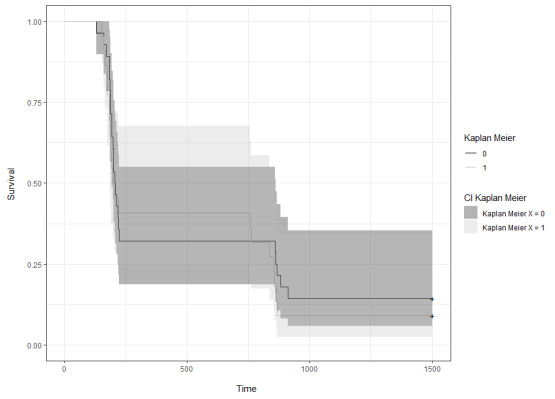

Figure .1.14: The estimated Kaplan-Meier survival curve for a sample when $n = 50$

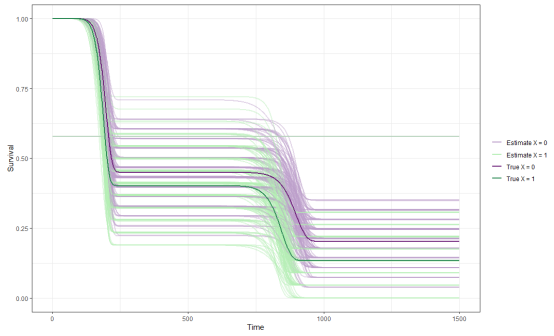

Figure .1.15: Sample survival curve estimated using the EM algorithm for $n = 50$ for a Weibull 3-component mixture cure model. In darker, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$, $\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$.

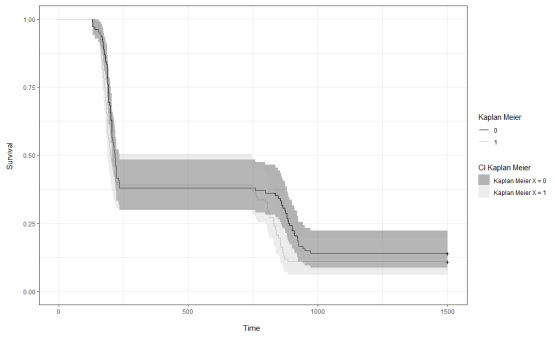

Figure .1.16: The estimated Kaplan-Meier survival curve for a sample when $n = 200$

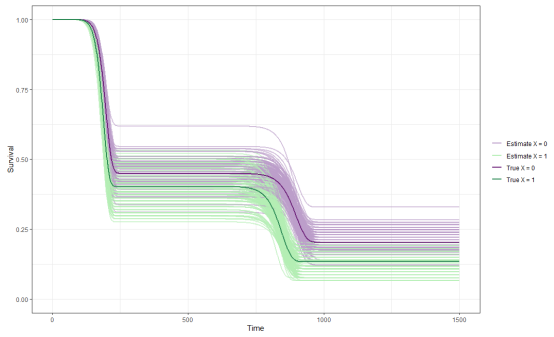

Figure .1.17: Sample survival curve estimated using the EM algorithm for $n = 200$ for a Weibull 3-component mixture cure model. In darker, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$, $\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$.

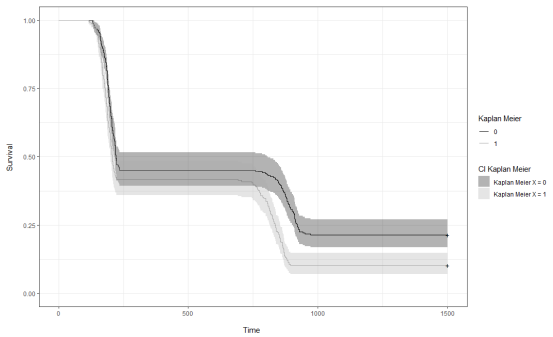

Figure .1.18: The estimated Kaplan-Meier survival curve for a sample when $n = 500$

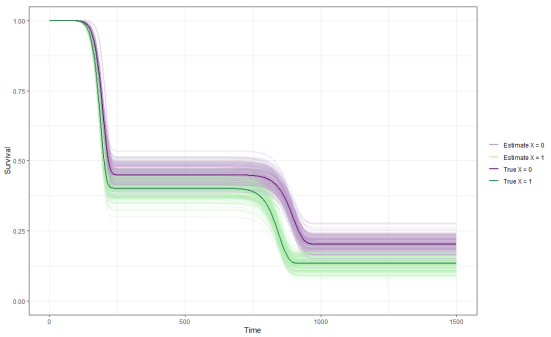

Figure .1.19: Sample survival curve estimated using the EM algorithm for $n = 500$ for a Weibull 3-component mixture cure model. In darker, the survival curve for the true parameters: $\lambda_1 = 200$, $\lambda_2 = 900$, $\gamma_1 = 10$, $\gamma_2 = 25$, $\beta_{01} = 1$ $\beta_1 = 0.5$, $\beta_{02} = 0.2$, $\beta_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 1.5$.

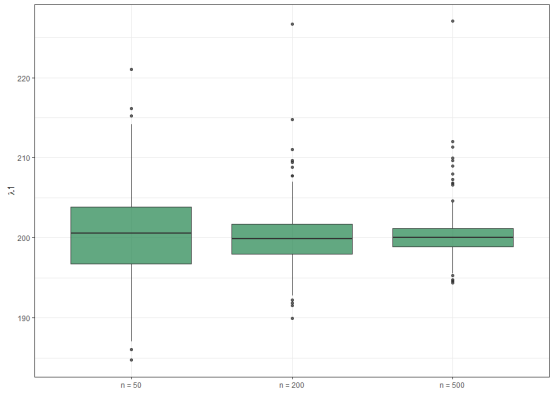

Figure .1.20: Boxplots of estimates of λ_1 for 1000 replications of the EM algorithm

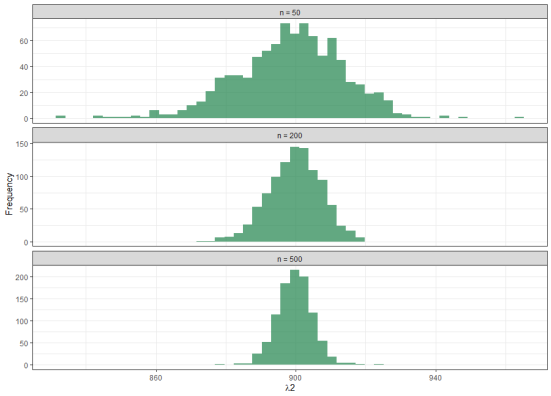

Figure .1.21: Histogram of estimates of λ_1 for 1000 replications of the EM algorithm

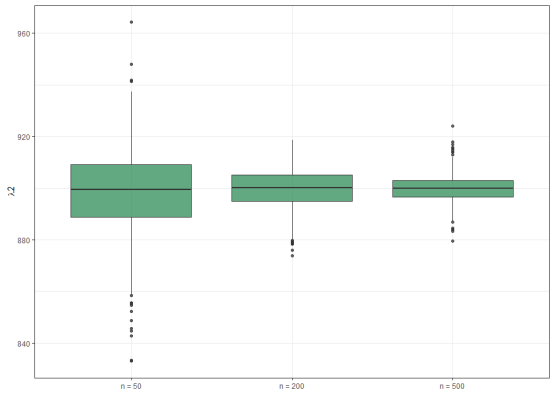

Figure .1.22: Boxplots of estimates of λ_2 for 1000 replications of the EM algorithm

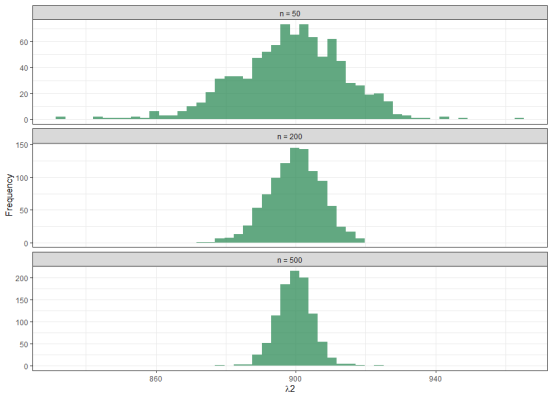

Figure .1.23: Histogram of estimates of λ_2 for 1000 replications of the EM algorithm

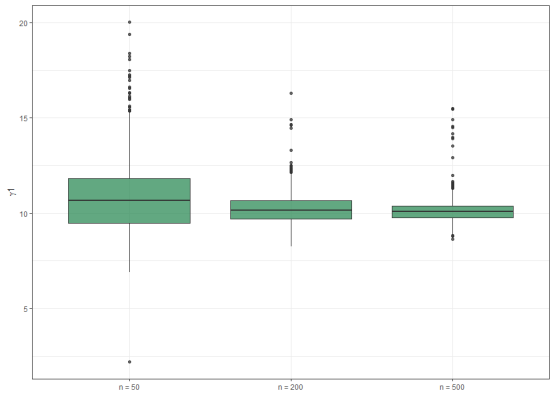

Figure .1.24: Boxplots of estimates of γ_1 for 1000 replications of the EM algorithm

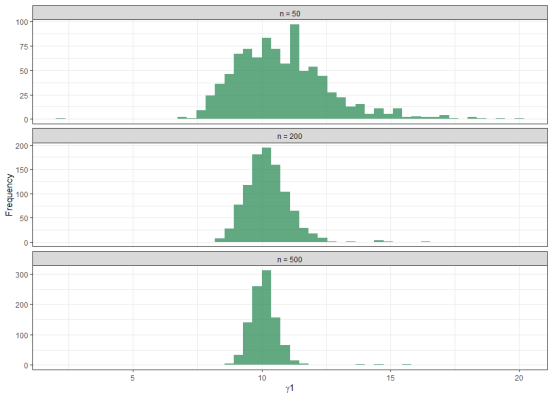

Figure .1.25: Histogram of estimates of γ_1 for 1000 replications of the EM algorithm

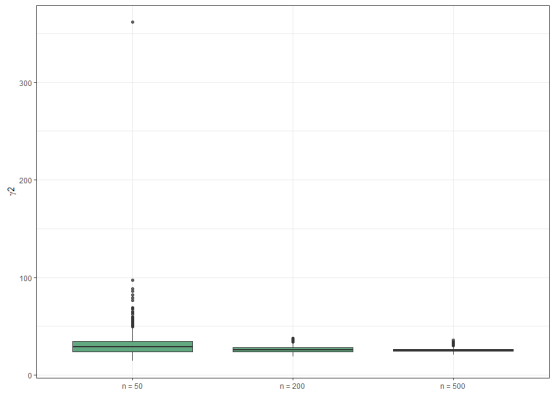

Figure .1.26: Boxplots of estimates of γ_2 for 1000 replications of the EM algorithm

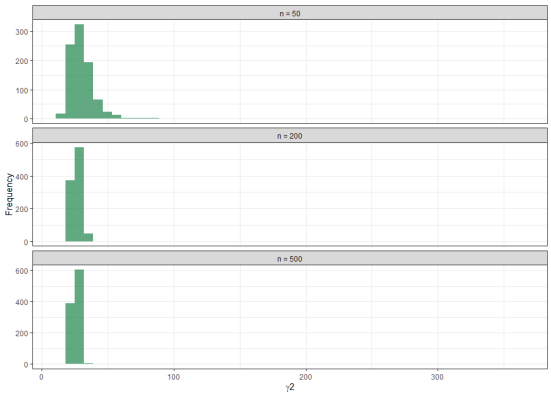

Figure .1.27: Histogram of estimates of γ_2 for 1000 replications of the EM algorithm

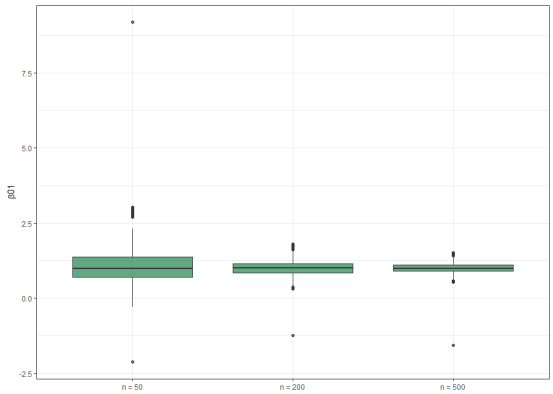

Figure .1.28: Boxplots of estimates of β_{01} for 1000 replications of the EM algorithm

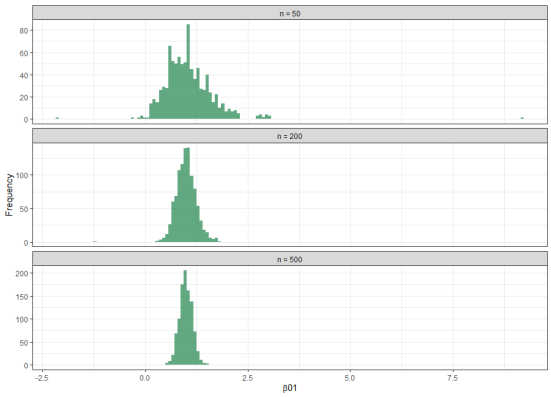

Figure .1.29: Histogram of estimates of β_{01} for 1000 replications of the EM algorithm

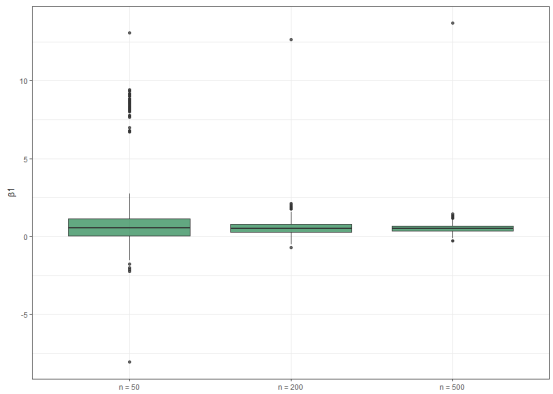

Figure .1.30: Boxplots of estimates of β_1 for 1000 replications of the EM algorithm

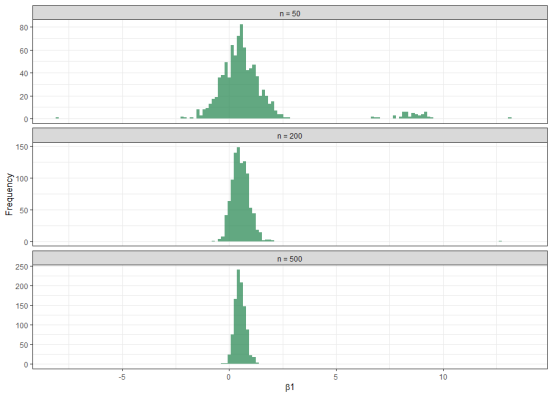

Figure .1.31: Histogram of estimates of β_1 for 1000 replications of the EM algorithm

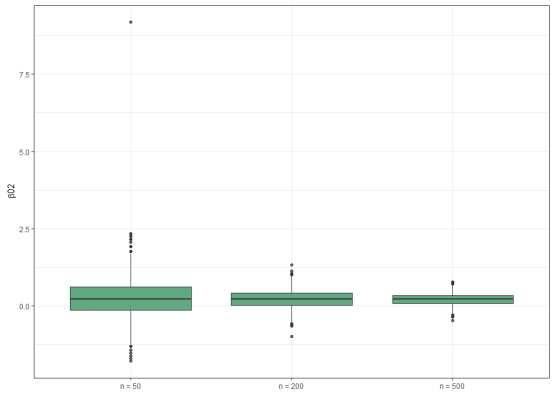

Figure .1.32: Boxplots of estimates of β_{02} for 1000 replications of the EM algorithm

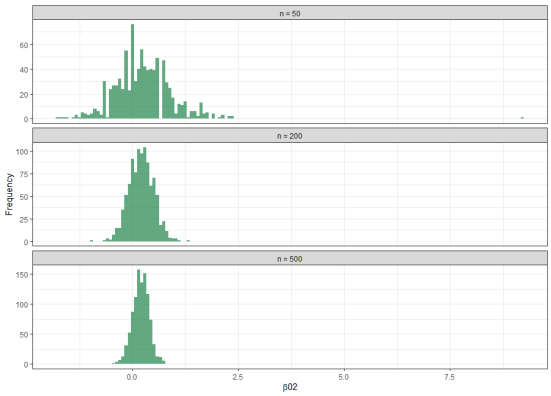

Figure .1.33: Histogram of estimates of β_{02} for 1000 replications of the EM algorithm

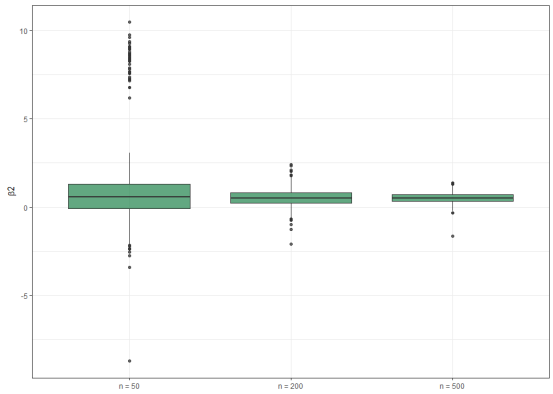

Figure .1.34: Boxplots of estimates of β_2 for 1000 replications of the EM algorithm

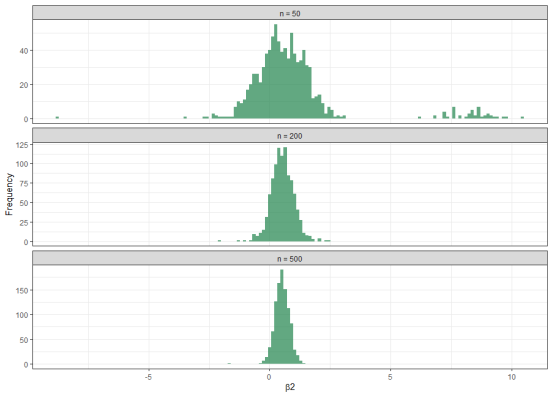

Figure .1.35: Histogram of estimates of β_2 for 1000 replications of the EM algorithm

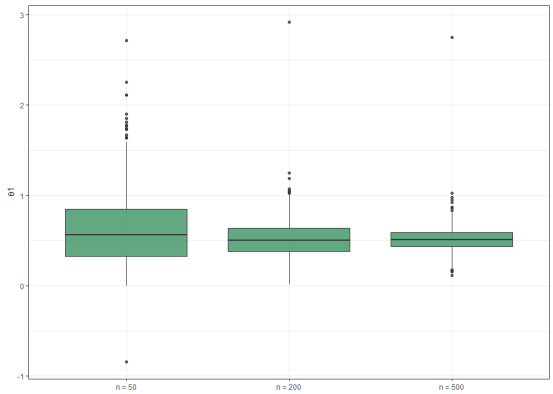

Figure .1.36: Boxplots of estimates of θ_1 for 1000 replications of the EM algorithm

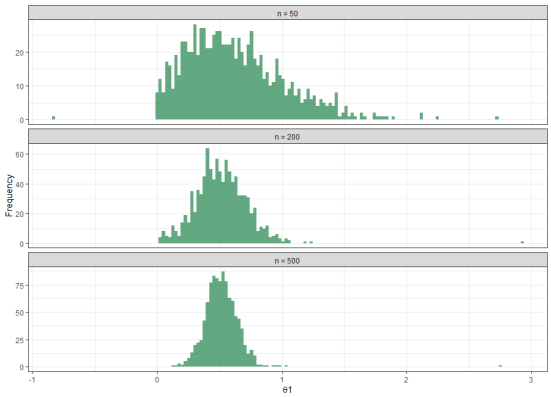

Figure .1.37: Histogram of estimates of θ_1 for 1000 replications of the EM algorithm

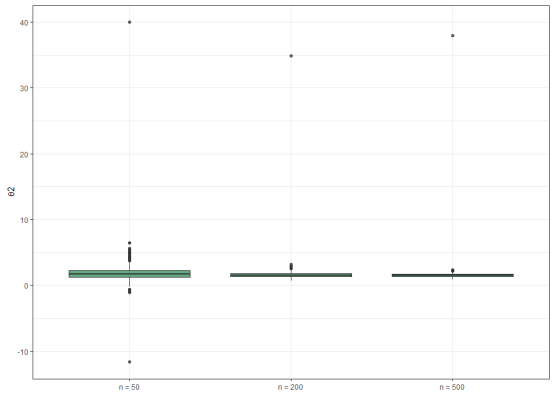

Figure .1.38: Boxplots of estimates of θ_2 for 1000 replications of the EM algorithm

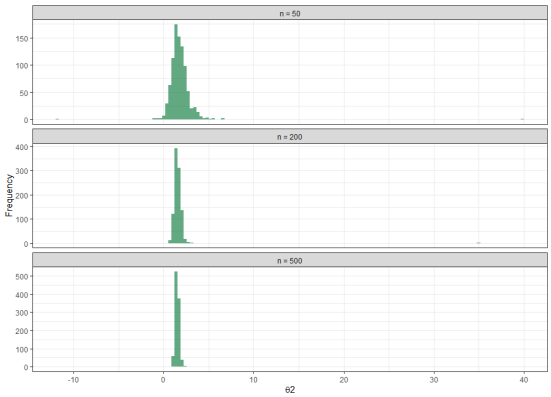

Figure .1.39: Histogram of estimates of θ_2 for 1000 replications of the EM algorithm

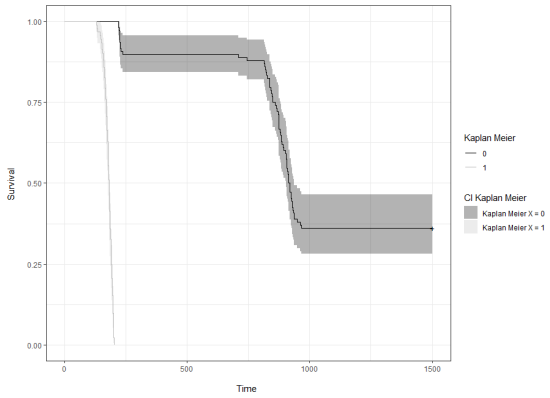

Figure .1.40: The estimated Kaplan-Meier survival curve for the pathological sample when $n = 200$

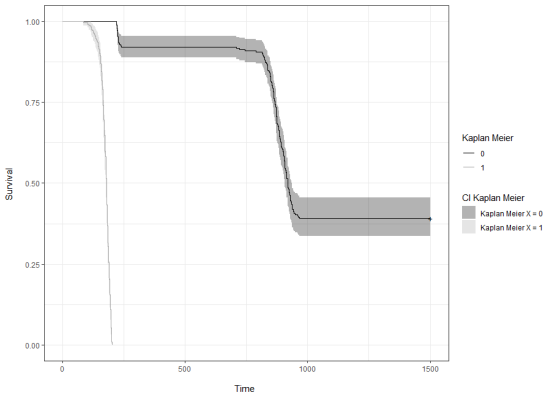

Figure .1.41: The estimated Kaplan-Meier survival curve for pathological sample when $n = 500$

.1.3 Mixture cure ELSC model simulation : simulation study 1

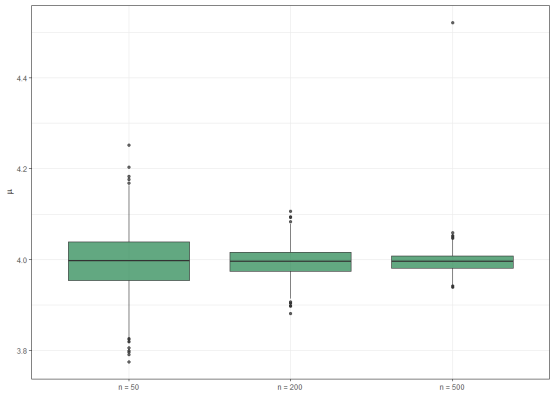

Figure .1.42: Boxplots of estimates of μ for 1000 replications of the Mixture cure ELSC model.

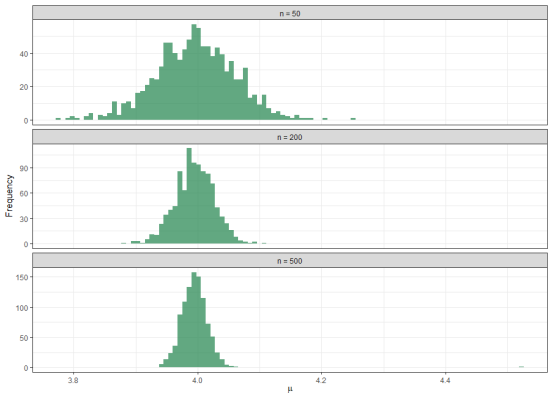

Figure .1.43: Histogram of estimates of μ for 1000 replications of the Mixture cure ELSC model.

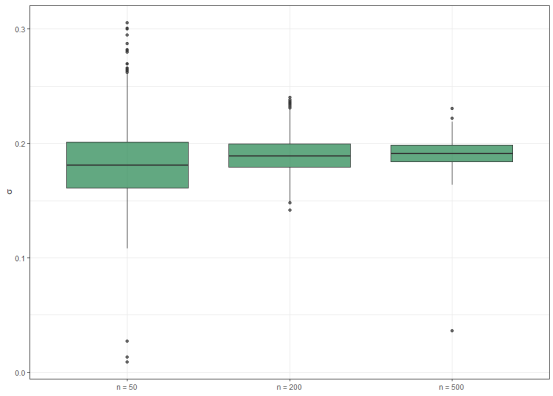

Figure .1.44: Boxplots of estimates of σ for 1000 replications of the Mixture cure ELSC model.

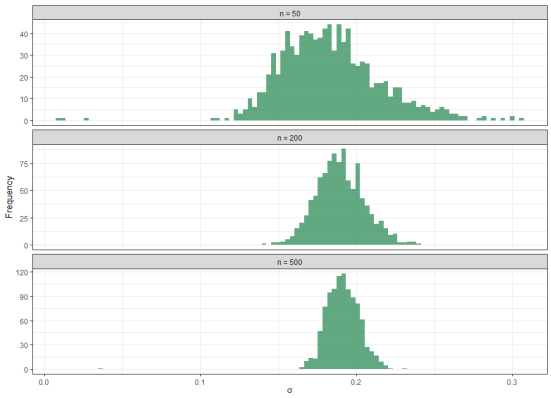

Figure .1.45: Histogram of estimates of σ for 1000 replications of the Mixture cure ELSC model.

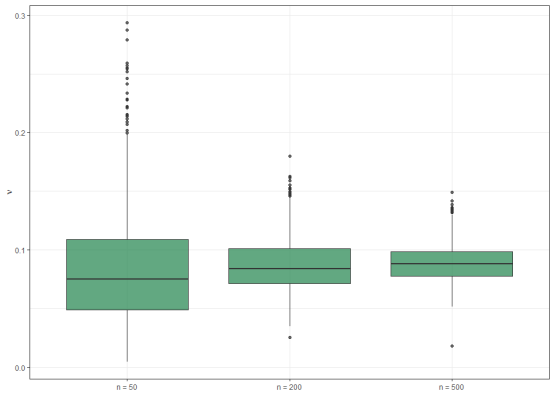

Figure .1.46: Boxplots of estimates of ν for 1000 replications of the Mixture cure ELSC model.

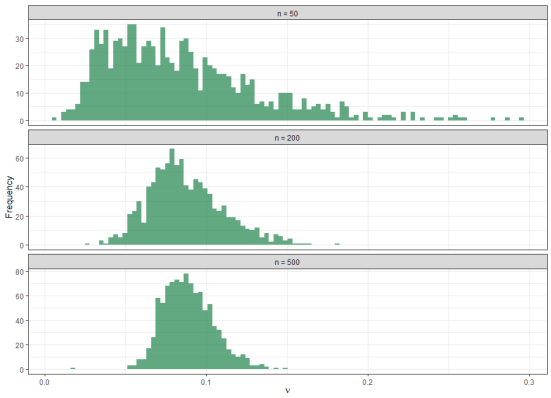

Figure .1.47: Histogram of estimates of ν for 1000 replications of the Mixture cure ELSC model.

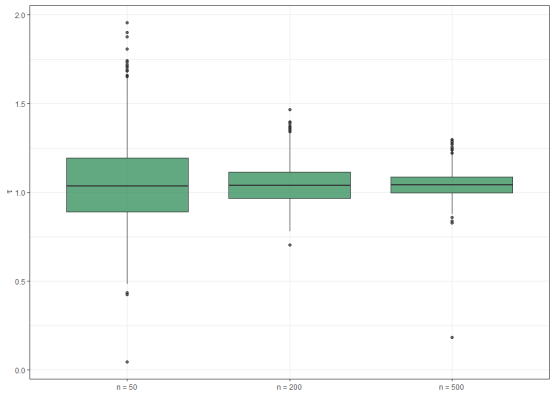

Figure .1.48: Boxplots of estimates of τ for 1000 replications of the Mixture cure ELSC model.

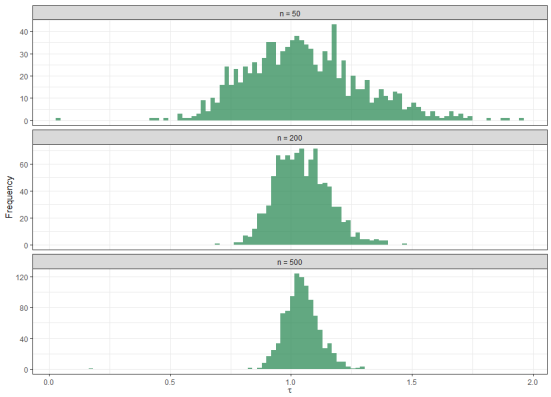

Figure .1.49: Histogram of estimates of τ for 1000 replications of the Mixture cure ELSC model.

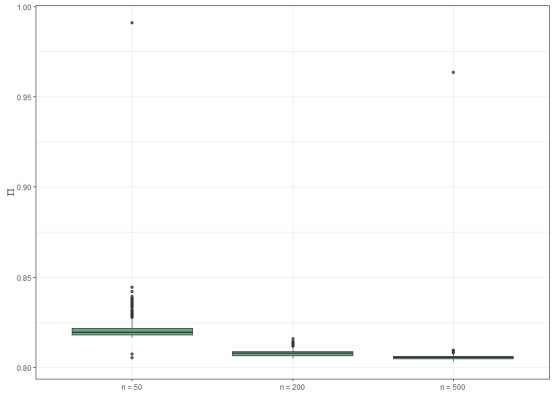

Figure .1.50: Boxplots of estimates of π for 1000 replications of the Mixture cure ELSC model.

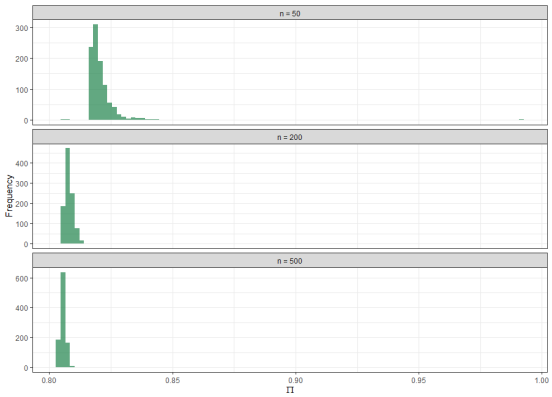

Figure .1.51: Histogram of estimates of π for 1000 replications of the Mixture cure ELSC model.

.1.4 Mixture cure ELSC model simulation : simulation study 2

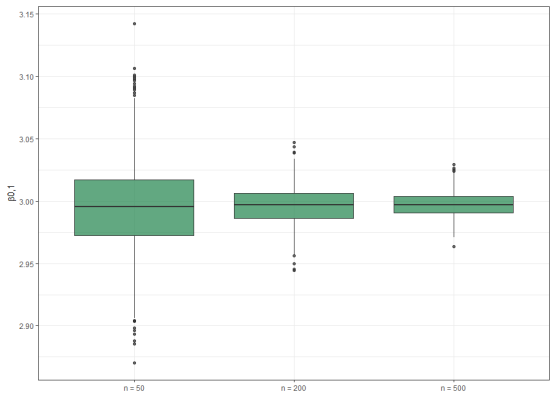

Figure .1.52: Boxplots of estimates of $\beta_{0,1}$ for 1000 replications of the Mixture cure ELSC model.

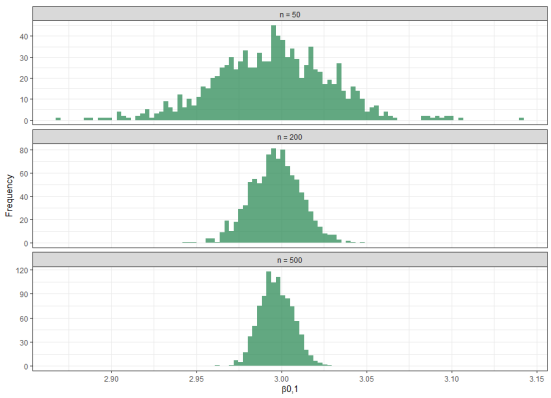

Figure .1.53: Histogram of estimates of $\beta_{0,1}$ for 1000 replications of the Mixture cure ELSC model.

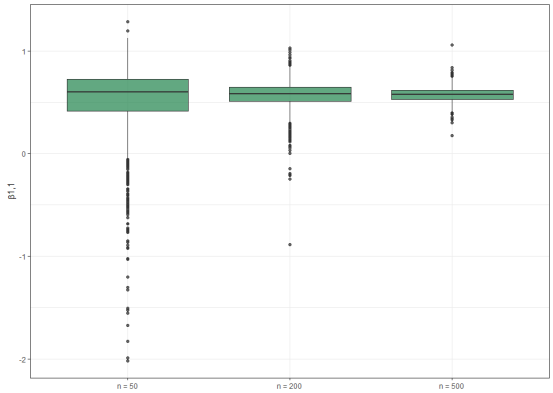

Figure .1.54: Boxplots of estimates of $\beta_{1,1}$ for 1000 replications of the Mixture cure ELSC model.

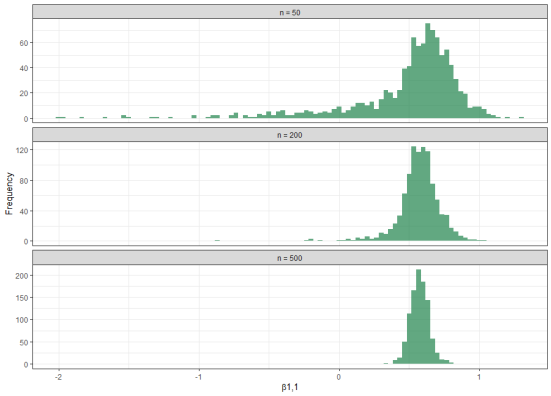

Figure .1.55: Histogram of estimates of $\beta_{1,1}$ for 1000 replications of the Mixture cure ELSC model.

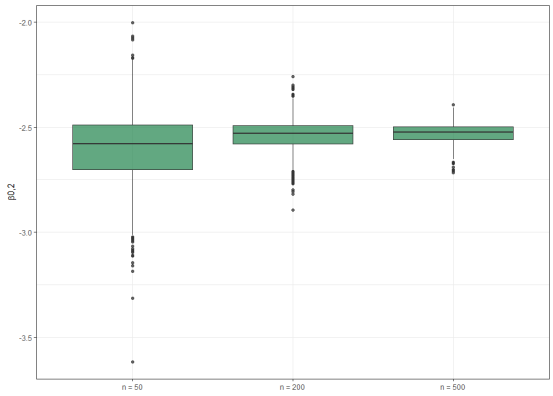

Figure .1.56: Boxplots of estimates of $\beta_{0,2}$ for 1000 replications of the Mixture cure ELSC model.

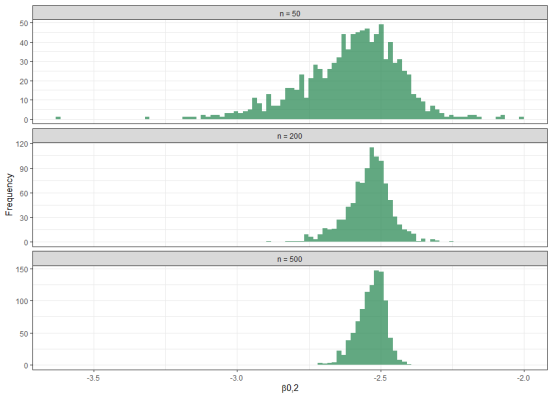

Figure .1.57: Histogram of estimates of $\beta_{0,2}$ for 1000 replications of the Mixture cure ELSC model.

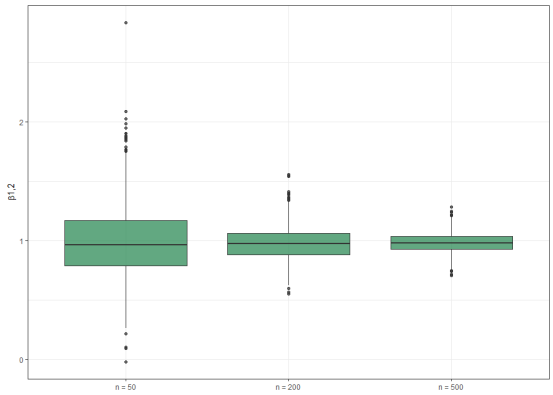

Figure .1.58: Boxplots of estimates of $\beta_{1,2}$ for 1000 replications of the Mixture cure ELSC model.

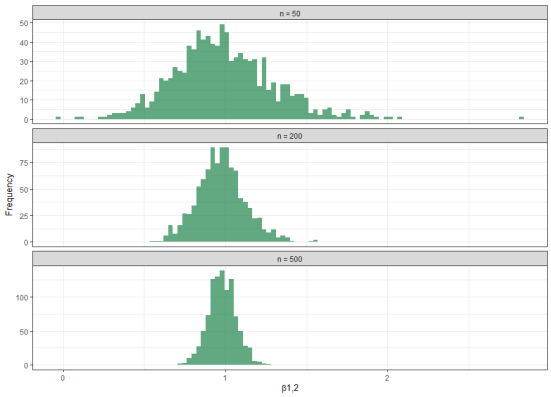

Figure .1.59: Histogram of estimates of $\beta_{1,2}$ for 1000 replications of the Mixture cure ELSC model.

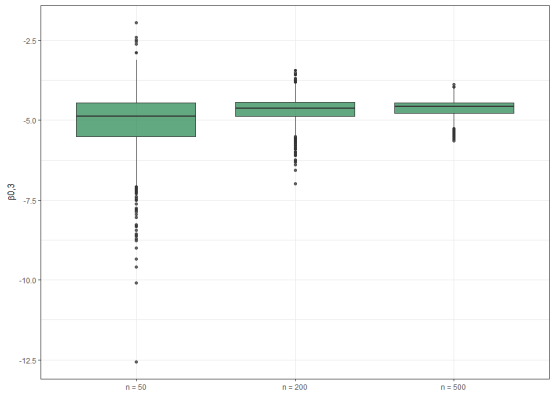

Figure .1.60: Boxplots of estimates of $\beta_{0,3}$ for 1000 replications of the Mixture cure ELSC model.

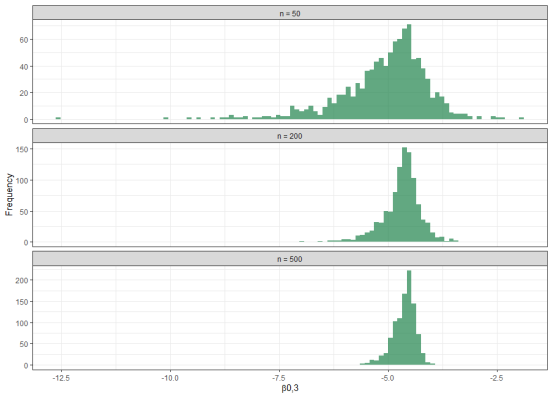

Figure .1.61: Histogram of estimates of $\beta_{0,3}$ for 1000 replications of the Mixture cure ELSC model.

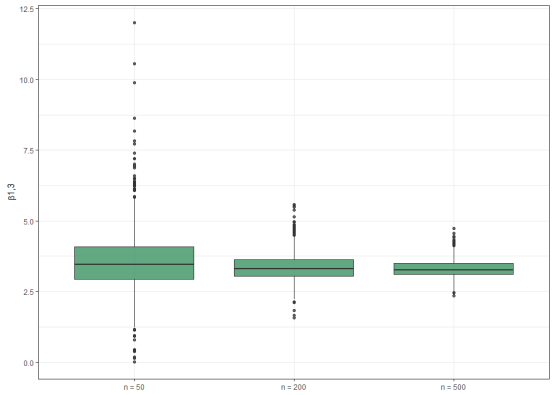

Figure .1.62: Boxplots of estimates of $\beta_{1,3}$ for 1000 replications of the Mixture cure ELSC model.

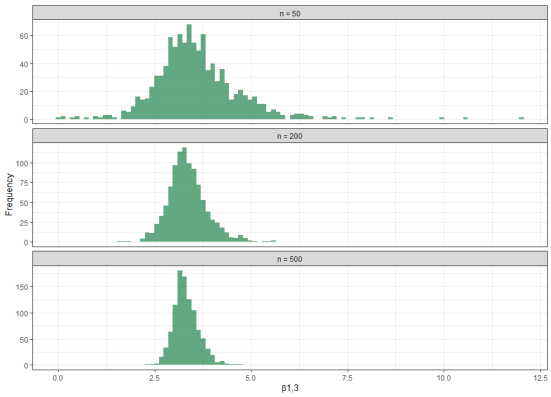

Figure .1.63: Histogram of estimates of $\beta_{1,3}$ for 1000 replications of the Mixture cure ELSC model.

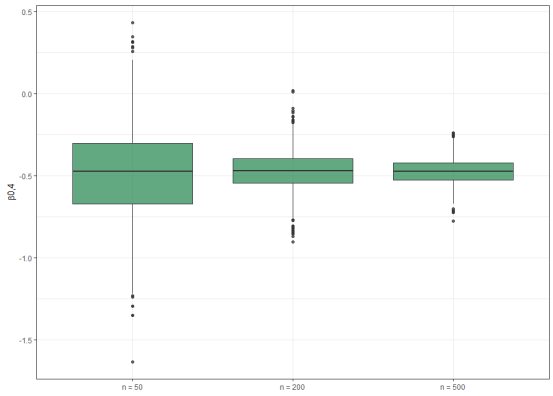

Figure .1.64: Boxplots of estimates of $\beta_{0,4}$ for 1000 replications of the Mixture cure ELSC model.

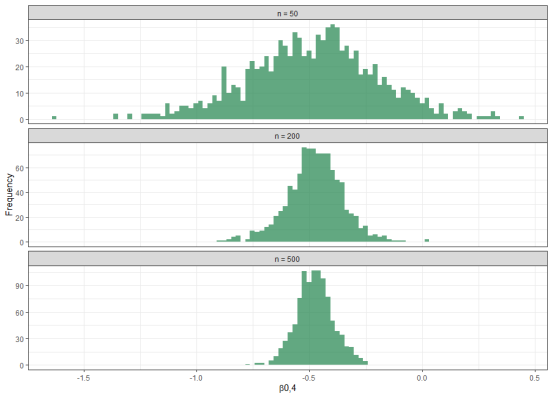

Figure .1.65: Histogram of estimates of $\beta_{0,4}$ for 1000 replications of the Mixture cure ELSC model.

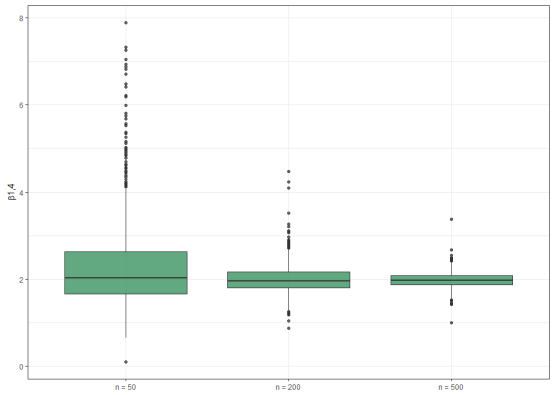

Figure .1.66: Boxplots of estimates of $\beta_{1,4}$ for 1000 replications of the Mixture cure ELSC model.

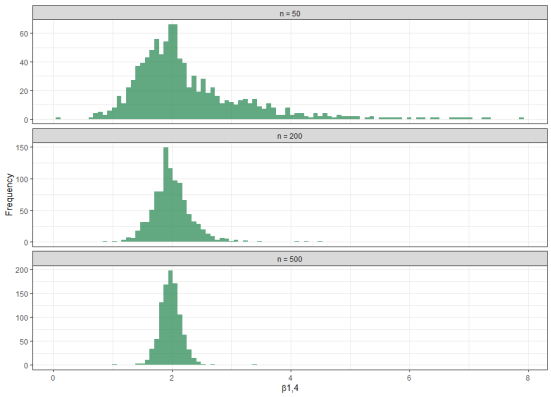

Figure .1.67: Histogram of estimates of $\beta_{1,4}$ for 1000 replications of the Mixture cure ELSC model.

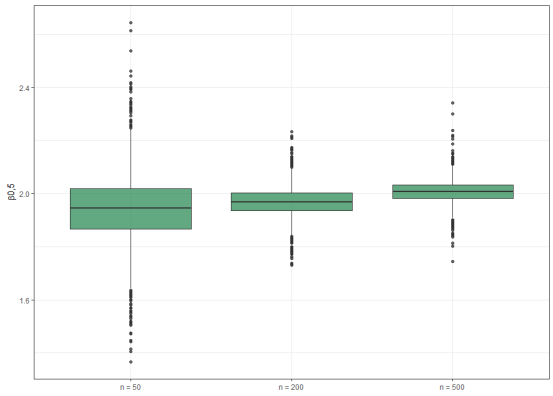

Figure .1.68: Boxplots of estimates of $\beta_{0,5}$ for 1000 replications of the Mixture cure ELSC model.

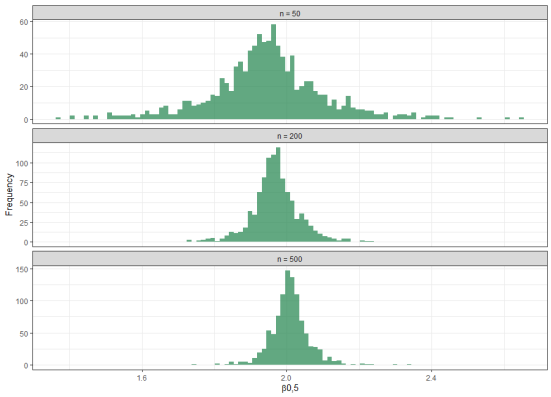

Figure .1.69: Histogram of estimates of $\beta_{0,5}$ for 1000 replications of the Mixture cure ELSC model.

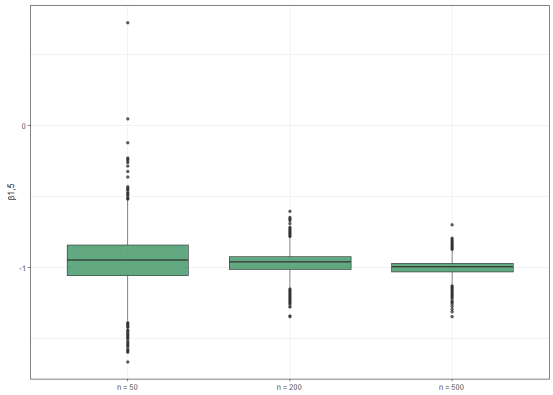

Figure .1.70: Boxplots of estimates of $\beta_{1,5}$ for 1000 replications of the Mixture cure ELSC model.

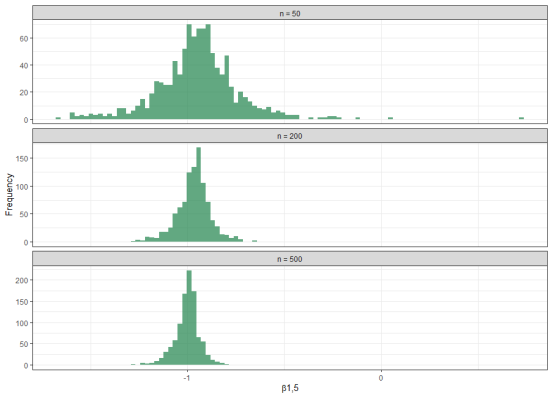

Figure .1.71: Histogram of estimates of $\beta_{1,5}$ for 1000 replications of the Mixture cure ELSC model.

.2 Simulation with random censoring

.2.1 3-component mixture cure model : without covariates

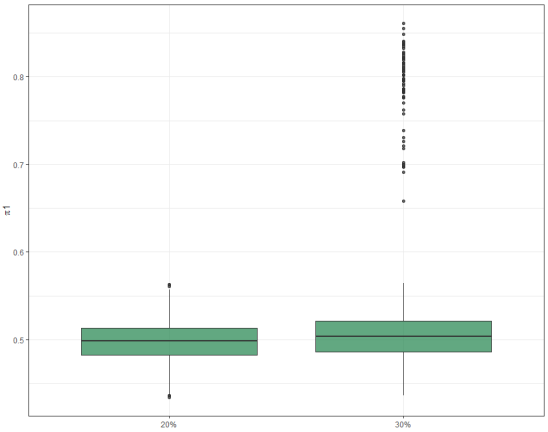

Figure .2.1: Boxplots of estimates of π_1 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

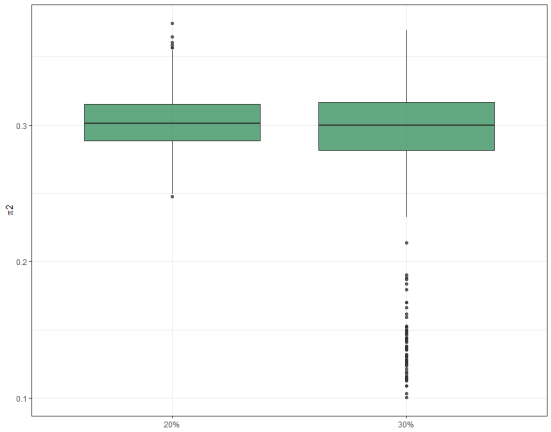

Figure .2.2: Boxplots of estimates of π_2 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

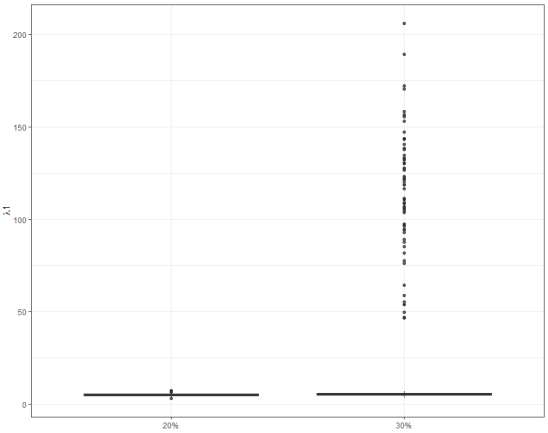

Figure .2.3: Boxplots of estimates of λ_1 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

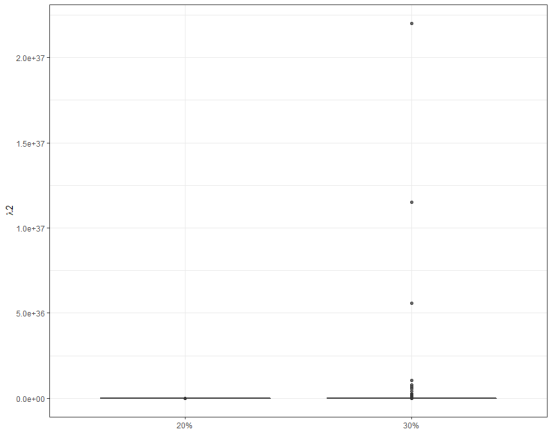

Figure .2.4: Boxplots of estimates of λ_2 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

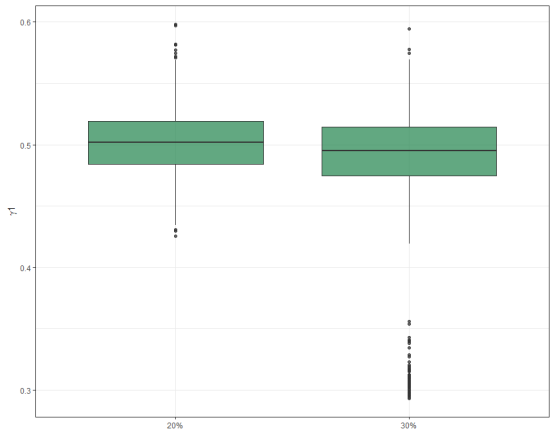

Figure .2.5: Boxplots of estimates of γ_1 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

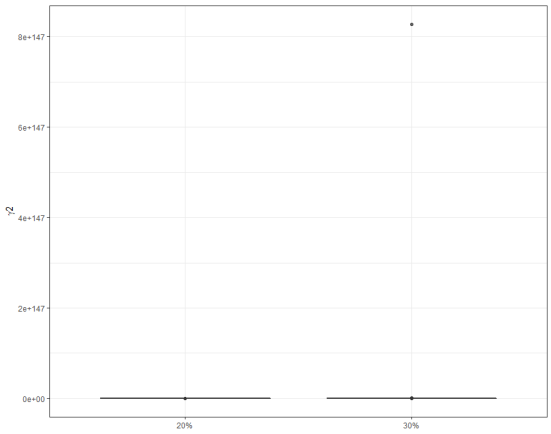

Figure .2.6: Boxplots of estimates of γ_2 for 1000 replications of the EM algorithm for samples with 20% and 30% censoring rate.

.2.2 ELSCcr model: without covariates

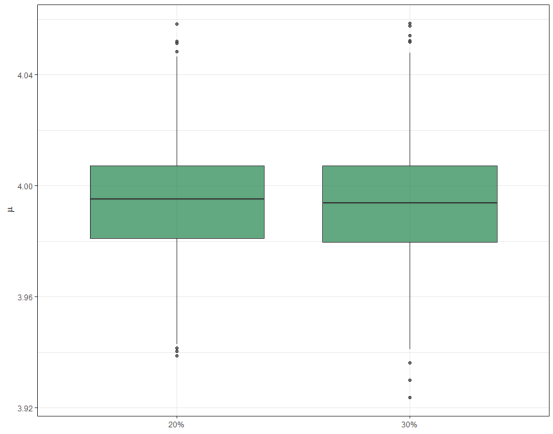

Figure .2.7: Boxplots of estimates of μ for 1000 replications of the Mixture cure ELSC modelfor samples with 20% and 30% censoring rate.

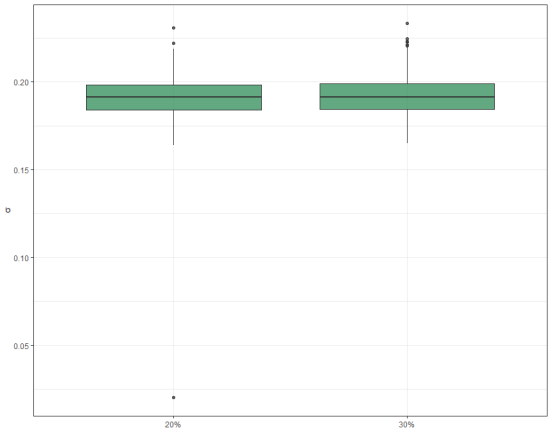

Figure .2.8: Boxplots of estimates of σ for 1000 replications of the Mixture cure ELSC modelfor samples with 20% and 30% censoring rate.

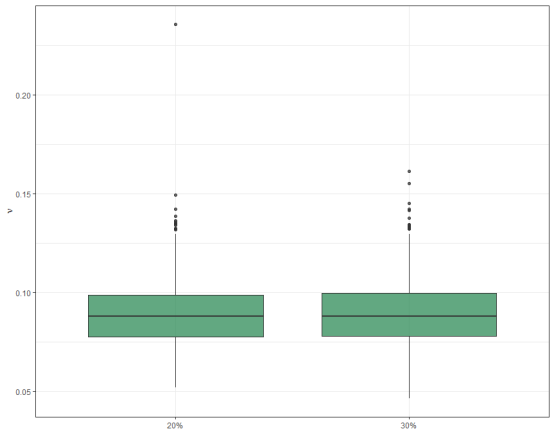

Figure .2.9: Boxplots of estimates of ν for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

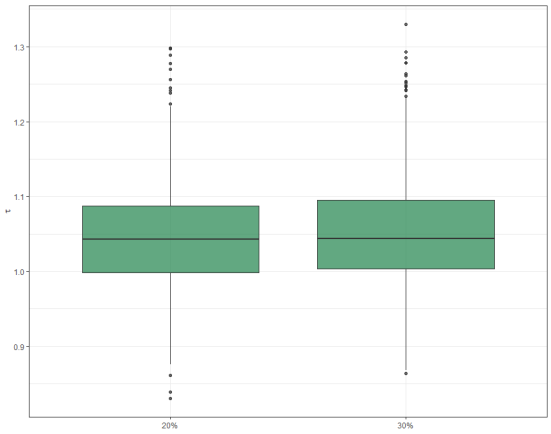

Figure .2.10: Boxplots of estimates of τ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

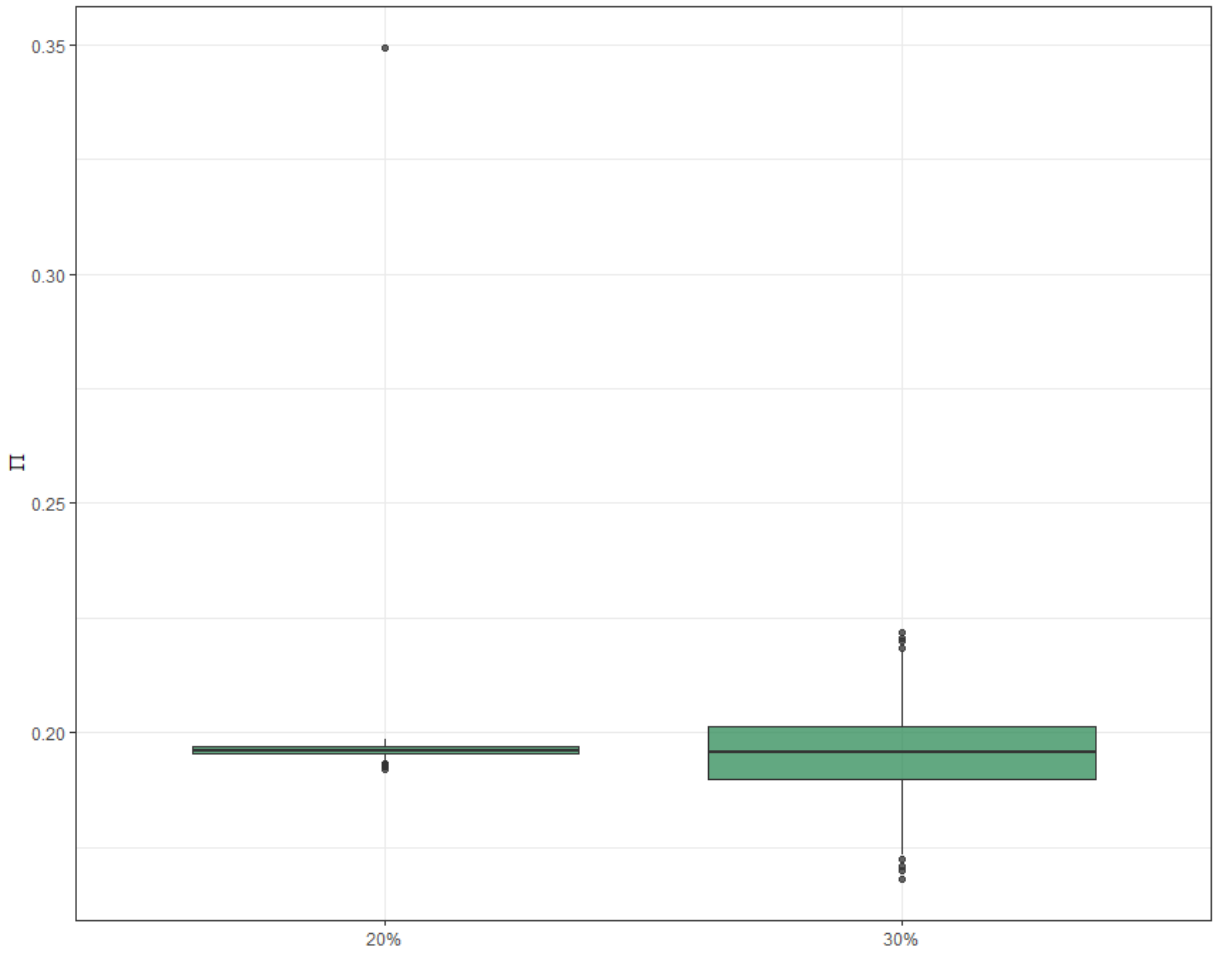

Figure .2.11: Boxplots of estimates of π for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

.2.3 ELSCcr model: with covariates

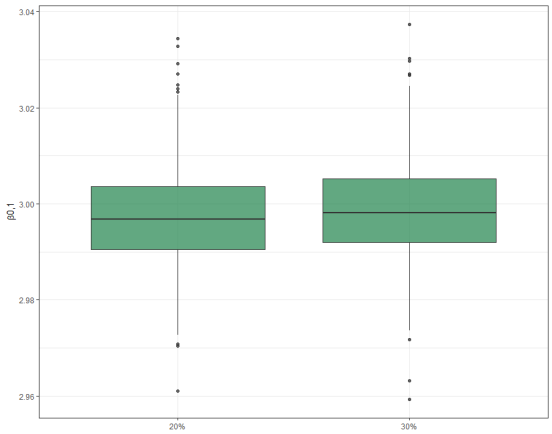

Figure .2.12: Boxplots of estimates of $\beta_{0,1}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

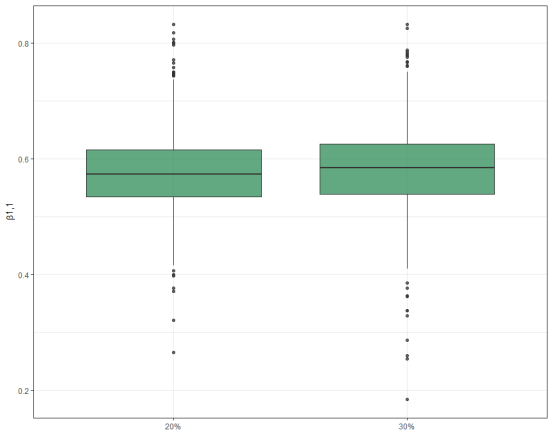

Figure .2.13: boxplotcenss of estimates of $\beta_{1,1}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

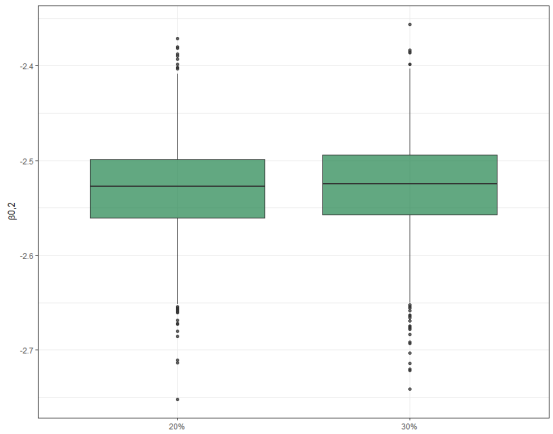

Figure .2.14: Boxplots of estimates of $\beta_{0,2}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

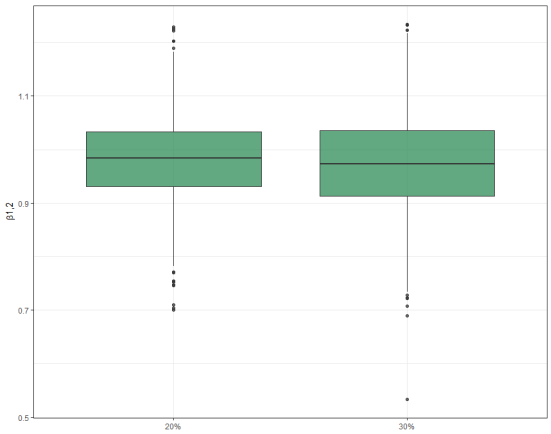

Figure .2.15: Boxplots of estimates of $\beta_{1,2}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

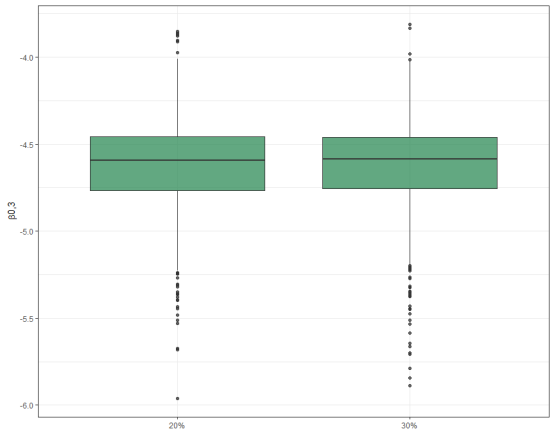

Figure .2.16: Boxplots of estimates of $\beta_{0,3}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

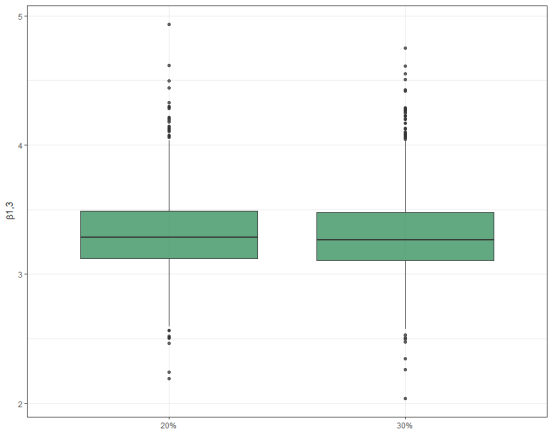

Figure .2.17: Boxplots of estimates of $\beta_{1,3}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

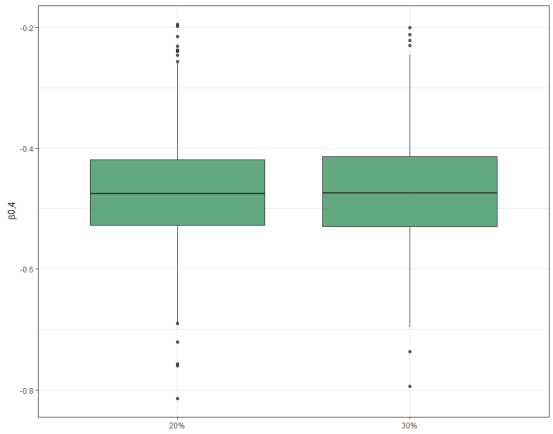

Figure .2.18: Boxplots of estimates of $\beta_{0,4}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate..

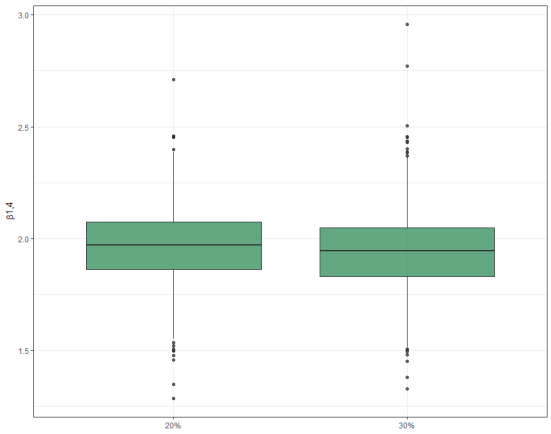

Figure .2.19: Boxplots of estimates of $\beta_{1,4}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

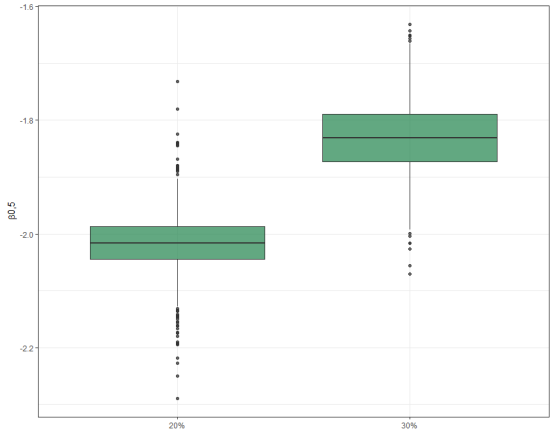

Figure .2.20: Boxplots of estimates of $\beta_{0,5}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

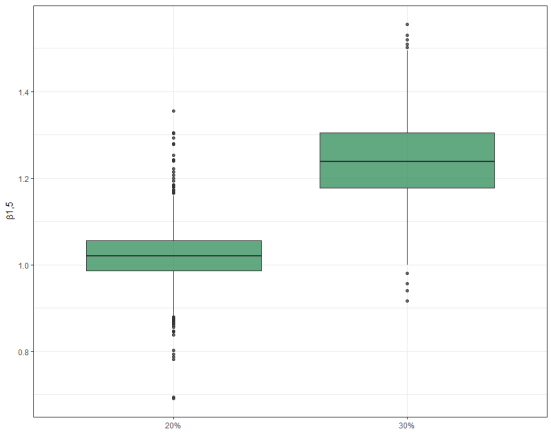

Figure .2.21: Boxplots of estimates of $\beta_{1,5}$ for 1000 replications of the Mixture cure ELSC model for samples with 20% and 30% censoring rate.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc