KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

UCLouvain

Faculté des sciences

Cure Models in Survival Analysis: From Modelling to Prediction Assessment of the Cure Fraction

Maïlis Amico

A thesis submitted in partial fulfillment of the requirements for the joint degree of

Doctor of Sciences
from the Université catholique de Louvain

&

Doctor in Business Economics
from the Katholieke Universiteit Leuven.

Examination committee:
Prof. Dr. Catherine Legrand, *Université catholique de Louvain*, Supervisor
Prof. Dr. Ingrid Van Keilegom, *KU Leuven*, Supervisor
Prof. Dr. Ricardo Cao, *Universidade da Coruña*
Prof. Dr. Gerda Claeskens, *KU Leuven*
Prof. Dr. Anouar El Ghouch, *Université catholique de Louvain*
Prof. Dr. Philippe Lambert, *Université de Liège & Université catholique de Louvain*

November 2018

Aknowledgements

Remerciements

Il y a un peu plus de cinq ans, en débutant ce doctorat en tant qu'assistante d'enseignement au sein de l'Université catholique de Louvain, je n'aurais jamais imaginé la direction qu'allait prendre la voie dans laquelle je venais de me lancer. Cinq ans plus tard, me voilà en fin de doctorat, non plus à l'UCLouvain mais à la KU Leuven, consacrant tout mon temps à faire de la recherche, et force est de constater que la personne que je suis devenue est bien différente de celle que j'étais.

Cette aventure n'aurait jamais débuté si, un jour de Mars 2013, la Professeure Ingrid Van Keilegom n'avait pas demandé à la mémorante que j'étais si j'avais déjà songé à mon avenir professionnel et si la possibilité de faire un doctorat en faisait partie. J'ai gardé cela au fond de moi pendant un peu plus d'un mois avant d'oser en parler, toute émue que j'étais par ce que l'on venait de me demander. Et puis j'ai dit oui, et me voilà aujourd'hui. Je dois donc à Ingrid Van Keilegom la première pierre de cet édifice qu'est ensuite venue sceller la Professeure Catherine Legrand. Ingrid, Catherine, je vous dois à toutes les deux énormément pour l'accomplissement de ce travail. Vous m'avez accompagnée sur ce chemin, chacune à votre façon, avec beaucoup de bienveillance. Je voudrais vous remercier pour la confiance que vous m'avez accordée, pour votre soutien, pour la qualité de votre encadrement mais aussi pour avoir cru en moi.

Ingrid, je suis admirative de ta vivacité d'esprit, de ta rigueur et de la passion avec laquelle tu fais ton métier. Je mesure la chance que j'ai d'avoir travaillé à tes côtés. J'ai beaucoup apprécié tous nos échanges scientifiques, mais aussi ceux un peu moins scientifiques, souvent riches et denses intellectuellement. Je voudrais également te remercier pour ta grande disponibilité, surtout lors de la deuxième partie de cette thèse, toujours avec le sourire, malgré ton agenda bien chargé, ainsi que pour m'avoir offert la possibilité de passer chercheur. Je n'y croyais plus et tu l'as fait. Je ne saurais jamais assez comment te remercier.

Catherine, tu as indéniablement apporté une touche appliquée à cette thèse. Merci pour ton expertise en biostatistiques, pour ton point de vue appliqué et pour tes commentaires et suggestions sur le dernier papier. Je voudrais également te remercier pour m'avoir fait prendre conscience que le perfection-

isme a ses limites, et qu'il est important de savoir prendre du recul quand le flot émotionnel devient trop important.

My thanks also goes to Professor Gerda Claeskens and Professor Philippe Lambert, both members of my supervisory committee, and to Professor Anouar El Ghouch, who all followed my progress during these five years and who provided me many comments and suggestions at many different occasions that contributed to improve the contents of this thesis. I would also like to thank Professor Ricardo Cao, external member of this thesis committee, for his enthusiasm and his interest for my work, and for his comments and suggestions during the private defence. Finally, thanks to Professor Robert Boute for having accepted to chair the public defence of this thesis.

J'ai passé les trois premières années de ce doctorat au sein de l'Institut de Statistiques, Biostatistiques et Sciences Actuarielles de l'Université catholique de Louvain. Je voudrais remercier tous les professeurs de l'institut pour leur sympathie.

Au cours de ces trois années, j'ai également eu l'opportunité de participer aux activités du SMCS. Je voudrais donc remercier toute l'équipe du SMCS pour m'avoir permis de garder un pied dans le monde des données réelles ainsi que pour leur convivalité. En particulier, je voudrais remercier Catherine Rasse pour m'avoir fourni toutes ces consultances qui m'ont permis d'améliorer ma connaissance et ma pratique des modèles mixtes, ainsi que Céline Bugli et Lieven Desmet avec lesquels j'ai pris beaucoup de plaisir à travailler en tant qu'assistante de leurs cours.

Un très grand merci à Nancy Guillaume pour son aide d'un point de vue administratif, mais aussi au niveau personnel, et pour sa bienveillance, ainsi qu'à toute l'équipe administrative de l'Institut de Statistiques, Biostatistiques et Sciences Actuarielles: Sophie Malali, Nadja Peiffer, Tatiana Regout and Marguerite-Marie Hanon.

Je voudrais également remercier mes collègues de Louvain-la-Neuve, les assistants qui m'ont accueillie au début de cette aventure: Nicolas, Aurélie, Benjamin, Vincent, Mathieu, Cédric, Nathan, Baptiste et Majda, ceux avec lesquels j'ai partagé un bureau: Michał, Joris, Kassu et Yuwei, ainsi qu'Héléne, Nathalie, Sylvie, Manon, Marco, Anne, François, Gildas, Samuel, Anna, Laure, Mickaël, Florian, Oswaldo, Rebecca, Stefka, Ana, Josefine, Adrien et Jennifer. Merci à vous tous pour les nombreuses discussions, pour l'aide et le soutien au quotidien et pour tous les bons moments passés ensemble à l'institut et ailleurs.

In Leuven, I would like to thank those who welcomed me: Thomas, Eugen, Ali, Jing, Daumantas, Roel, Luca, Ruben and Frits. Thanks for your enthusiasm, your open-mindedness and your efforts to integrate me. I would also like to thank Elif, Andrea, Leonard, Jeong Min, Aris, Negera, Ghassem, Vivienn, and particularly Motahareh for being such a nice officemate. Thank you all for this very nice last year.

I would finally like to thank Nicole Meesters and Annie Vercruysse for their administrative support at the Faculty of Economics and Business at KU Leuven.

D'un point de vue personnel, je voudrais tout d'abord remercier Cécile M., Jean-François D. et Véronique DN, témoins particuliers de cette aventure, bien

au délà de sa dimension professionnelle. Merci. Du fond du coeur.

J'ai passé sept ans en Belgique, loin des miens, et mes racines m'ont souvent manqué. Malgré tout, il y a tout un peloton de sudistes, de Marseille à Sahorre en passant par Lodève, Roqueredonde, Castries et Villelongue-Del-Monts sur lequel j'ai pu compter. Merci aux Amico de Marseille, aux Reverbel du mas, aux Millan, aux Soler et à Dany, pour tous les bons moments et les retours aux sources. C'est aussi grâce à vous si j'ai tenu le coup et si je suis arrivée au bout.

Enfin, je voudrais remercier mes parents et ma soeur. Merci d'être ce roc sans faille sur lequel je peux compter dans les plus grandes joies comme dans les pires tempêtes. Ma sensibilité a souvent été mise à rude épreuve pendant ces cinq années et vous en avez souvent fait les frais. Malgré tout, vous n'avez jamais vacillé. Je mesure chaque jour un peu plus la chance que j'ai de vous avoir à mes côtés. MERCI, jusqu'à la lune et les étoiles.

Pour finir, mes dernières mots seront pour Linoa et Chloé qui mieux que personne ont su me redonner le sourire quand le ciel était parfois bien gris.

Maïlis Amico, October 2018

Summary

Survival analysis examines and models the time it takes for events to occur. The typical event is death, from which the name 'survival analysis' and much of its terminology derives. Since the data can only be collected over a finite period of time, the 'time to event' may not be observed for all the individuals. This is the case for example when a patient leaves a clinical study before it ends or she/he is still alive by the end of the study. In such a case, the death time (time to event) for this individual is unknown. Such a phenomenon, named censoring, creates some unusual difficulties in the analysis of survival data that cannot be handled properly by standard statistical methods. In traditional survival analysis, all subjects in the population are assumed to be susceptible to the event of interest, that is, every subject has either already experienced the event or will experience it in the future. However, in many situations it may happen that a fraction of individuals (long-term survivors) will never experience the event, that is, they are considered to be event free. For example, a treatment is assigned to patients in order to evaluate the effect on the recurrence of a disease. Many individuals never experience recurrences and thus can be regarded as cured or immune individuals.

In the literature on cure models there are basically two types of models: the mixture cure model and the so-called promotion time cure model. In the former model one models the survival function by assuming that the underlying population is a mixture of two sub- populations: the sub-population of 'susceptibles' (i.e. those who will experience the event and have finite survival time) and the sub-population of 'non-susceptibles' (i.e. those who are event free and have an infinite survival time). On the other hand, the promotion time cure model is motivated by an underlying biological interpretation in terms of time to onset of cancer, and uses a direct modelling approach without separating susceptibles and non-susceptibles as is the case in the mixture cure model. In that sense, the two modelling approaches are quite different. Both models have been extensively studied in the literature, conditions under which the models are identifiable have been obtained and different parametric, semiparametric and nonparametric estimation procedures have been proposed and studied both asymptotically and for finite samples.

In this thesis we are interested in investigating three directions related to these models. The first contribution consists in providing a state of the art on cure models reviewing the many different points investigated in the literature and providing a formal and a numerical comparison of the two models through

an application on real data.

The second contribution of this thesis focuses on the mixture cure model and more precisely on the uncure proportion. Often, this quantity is modelled parametrically, assuming a logistic regression model. However, there is no reason to strictly constrain and limit the cure proportion to a logistic form. Our aim is then to propose a more flexible modelling approach for the cure proportion, by assuming a single-index structure, that is, a generalised linear model in which the link function is left unspecified, and by considering a Cox proportional hazards model for the conditional survival function of uncured subjects.

Finally, beside modelling and model selection, an important topic of statistical analysis is the question of model assessment, that is, the evaluation of the predictions that can be made from a given model. For cure models, predictions can be performed for two outcomes, the survival at a given time and the cure status, both of them being binary. Often, when one wants to assess the binary classification performance, the Receiver Operating Characteristic (ROC) curve is considered. However, while standard ROC curves suppose that the classes of the outcome are fully observed, building a ROC curve from cure survival data is a non-trivial problem since survival data are subject to censoring and hence, the cure status is unobserved. This last contribution concerns therefore the development of ROC curves to evaluate the performance of cure status prediction that can be made from survival data in the presence of a cure fraction.

Contents

Chapter 1

Introduction

Survival analysis is the branch of statistics dedicated to the analysis of time-to-event or survival data. By time-to-event, one means the time from a well define origin until the occurrence of a particular event of interest, also called the survival time. To provide some examples, the time to death, but also the time to failure of a machine, or the duration of unemployment are all survival times. Since it is not possible to follow all individuals for an infinite period of time, an important characteristic of time-to-event data is that the occurrence of the event may not be observed for all observations. This particular feature, named censoring, gives to survival analysis its specificity.

In 'classical' survival analysis, one usually assumes that, despite censoring, all subjects under study are susceptible to the event of interest, that is, they will all eventually experience it. In some contexts, however, it may happen that a fraction of the observations never experience the event. This is the case, for example, when interest lies in the time until the recurrence or the progress of a certain disease. In such a case, some subjects who are cured from the disease will never experience a relapse or a progress. Likewise, when one studies the time until a component fails in an industrial process, the event of interest may never happen for some observations. Since the event will never occur, these individuals are considered as long-term survivors or as 'cured', and one usually says that the survival data contain a cure fraction.

In order to take the presence of a cure fraction into account, 'classical' survival analysis has been extended to cure models. Initially introduced by the works of Boag (1949) and Berkson & Gage (1952), the literature on cure models is mainly composed of two classes of models, namely, the mixture cure model introduced by Farewell (1977) and Farewell (1982), and the promotion time cure model proposed by Yakovlev et al. (1996). Both of these models have been extensively study in the literature and cure models constitute nowadays a branch of survival analysis.

Within this framework, the goal of this thesis is to contribute to the literature on cure models by investigating different aspects of these models, ranging from modelling to prediction assessment.

Before moving to the core of this thesis, the remainder of this preliminary chapter is dedicated to the introduction of the concepts underlying the research presented in this manuscript. Section 1.1 presents some key elements of survival analysis, while Section 1.2 defines the concept of cure survival data and provides a general definition of cure models. Prediction assessment is often performed through the Receiver Operating Characteristic (ROC) curve. Section 1.3 provides an overview of this method. In Section 1.4, we provide a general description of the Expectation-Maximisation algorithm which is often used to estimate some cure models. Finally, in Section 1.5, we end this chapter with the outline of this thesis, summarising our contribution to cure models in survival analysis.

1.1 Some Elements of Survival Analysis[1]

Formally speaking, the objective of survival analysis is to study a non-negative random variable, denoted by T, which represents the survival time. Throughout this manuscript, we will assume that T is continuous.

1.1.1 Basic Elements

Survival Quantities

Traditionally, the distribution of T is described by three equivalent quantities, namely the survival function

$$S(t) = P(T > t) = 1 - F(t),$$

where $F(t) = P(T \leq t)$ is the cumulative distribution function of T, corresponding to the probability of being still event free at time t, the hazard function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t},$$

which represents the instantaneous risk of experiencing the event right after time t given that the subject has survived until then, and the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u) du,$$

which corresponds to the accumulated instantaneous risk of experiencing the event over the time. These three quantities are related by the relationship $S(t) = \exp[-\Lambda(t)]$ or equivalently $\Lambda(t) = -\log S(t)$. It follows that $\lambda(t) = (d/dt)\Lambda(t) = f(t)/S(t)$, where $f(\cdot)$ denotes the density function.

[1] All the concepts presented in this section are only briefly described. For a deeper insight into survival analysis, we refer the reader to the textbook by Collett (2003) which provides an easy to read insight to survival analysis and the textbook by Klein & Moeschberger (2003) for a more advanced description.

As mentioned at the beginning of this chapter, a fundamental assumption of survival analysis is to suppose that all subjects under study eventually experience the event of interest. Hence, the survival function is proper, that is,

$$\lim_{t\to\infty} S(t) = 0.$$

Equivalently, from the relationship between $S(t)$ and $\Lambda(t)$, the cumulative hazard function is such that

$$\lim_{t\to\infty} \Lambda(t) = \infty,$$

meaning that when t becomes large, the accumulated instantaneous risk of experiencing the event is infinity, and therefore affects all subjects.

The Concept of Censoring

An important characteristic of survival data is the presence of censoring. By censoring, we mean that the exact event time is not observed for some individuals and that it is only known to occur within a certain interval. Thus, the information provided by survival data is incomplete.

Three types of censoring can be encountered: right censoring, left censoring, and interval censoring. Additionally, the survival time can also be subject to right and/or left truncation. In this manuscript, only right censoring is considered. We therefore refer the reader to the textbook by Klein & Moeschberger (2003) for more details about the other types of censoring and truncation. Right censoring corresponds to the case where only a lower bound of the survival time is observed. Typically, this type of censoring occurs because it is not possible to follow individuals for an infinite period of time. For example, this is the case when a patient is still alive by the end of a clinical study with death as endpoint. In such a case, instead of observing T, we rather observe the follow-up time

$$Y = \min(T, C),$$

where C is the censoring time. A common assumption is to assume that T is independent of C. Alongside the variable Y, we also observe the censoring indicator

$$\Delta = I(T \leq C),$$

where $I(\cdot)$ represents the indicator function. Throughout this thesis, we further assume that we are facing random right censoring, meaning that the censoring time is a random variable. This type of censoring occurs because of 'lost to follow-up', when, for example, in medical studies, a patient leaves the study before having experienced the event of interest (e.g., death or progression of the disease); but also because of administrative censoring which happens when the patient is still event-free at the time of performing the analysis, among others.

Censoring has some consequences, the most important ones being on the building of the likelihood function and on the estimation of survival quantities as described in the next sections.

1.1.2 Likelihood Function for Survival Data

Due to censoring, the information provided by each observation is different depending on the censoring status. When building the likelihood function, it is therefore necessary to be careful to what information is provided by each observation. By considering independent and non-informative censoring, meaning that the distribution of the censoring times does not depend on the parameters of interest related to the survival function, it can be shown that an uncensored observation contributes to the likelihood function by means of the density function, while a censored observation contributes by means of the survival function (see, for example, Klein & Moeschberger (2003) Chapter 3). Let (Y_i, Δ_i), $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) copies of (Y, Δ). The likelihood function for survival data is given by

$$\mathcal{L} = \prod_{i=1}^{n} f(Y_i)^{\Delta_i} S(Y_i)^{1-\Delta_i}. \tag{1.1}$$

Equivalently, given that $f(Y_i) = \lambda(Y_i)S(Y_i)$, this likelihood function is also often expressed as

$$\mathcal{L} = \prod_{i=1}^{n} \{\lambda(Y_i)\}^{\Delta_i} S(Y_i).$$

1.1.3 The Kaplan-Meier Estimator

Censoring also impact the estimation of the survival function. Indeed, considering a classical empirical distribution function would consist in disregarding censored observations, and hence losing the information provided by censored observations. To overcome the difficulty, a standard estimator for the survival function is the non-parametric estimator proposed by Kaplan & Meier (1958) which is given by

$$\hat{S}(t) = \prod_{j:Y^*_{(j)} \leq t} \frac{R(Y^*_{(j)}) - D_j}{R(Y^*_{(j)})},$$

where $Y^*_{(j)}$, $j = 1, \ldots, r$, represent the distinct uncensored event times ordered in increasing order, $R(Y^*_{(j)})$ is the number of observations at risk prior to the time $Y^*_{(j)}$, and D_j is the number of observations that experience the event at time $Y^*_{(j)}$. This estimator takes the form of a step function with jumps at each distinct uncensored event time. Censored observations with follow-up times greater than $Y^*_{(j)}$ are taken into account in $R(Y^*_{(j)})$. Those with follow-up times lower than $Y^*_{(j)}$ are disregarded in $R(Y^*_{(j)})$. Note that, it is well known that the Kaplan-Meier estimator is inconsistent in the right tail. As a consequence, when the last follow-up time is a censoring time, $\{R(Y^*_{(j)}) - D_j\}/R(Y^*_{(j)}) \neq 0$ and it follows that $S(Y^*_{(r)}) \neq 0$. This point, even if it represents a drawback in classical survival analysis, will be useful in the presence of a cure fraction as we will detail later in this chapter.

In the presence of a vector of covariates, this nonparametric estimator has been extended by Beran (1981). A brief description of its form will be given in Chapter 2.

1.1.4 Survival Models

Modelling the influence of a vector of covariates on the survival, hereafter denoted by $\mathbf{Z}$, is often performed via one of the two models described in this section.

The Cox Proportional Hazards Model

A popular semi-parametric model for the hazard function is the proportional hazards (PH) model introduced by Cox (1972). This model consists in modelling the hazard function as follows:

$$\lambda(t|\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^t\boldsymbol{z}), \tag{1.2}$$

where $\lambda_0(t)$ is a baseline hazard, that is the hazard function for $\boldsymbol{z} = 0$, which remains unspecified, and $\boldsymbol{\beta}$ is a vector of parameters associated with $\boldsymbol{z}$, that does not contain an intercept. A particular feature of this model is that the hazards are proportional as the ratio of the hazard functions for two subjects, with covariate vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, respectively, is constant over the values of t:

$$\frac{\lambda(t|\boldsymbol{z}_i)}{\lambda(t|\boldsymbol{z}_j)} = \frac{\exp(\boldsymbol{\beta}^t\boldsymbol{z}_i)}{\exp(\boldsymbol{\beta}^t\boldsymbol{z}_j)}.$$

The cumulative hazard function for the Cox PH model is given by

$$\Lambda(t|\boldsymbol{z}) = \Lambda_0(t)\exp(\boldsymbol{\beta}^t\boldsymbol{z}),$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the baseline cumulative hazard function, and the survival function $S(t|\boldsymbol{z}) = P(T > t|\mathbf{Z} = \boldsymbol{z})$ takes the form

$$S(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})},$$

where $S_0(t) = P(T > t|\mathbf{Z} = 0) = \exp\{-\Lambda_0(t)\}$, is the baseline survival function. Both $\Lambda_0(t)$ and $S_0(t)$ are left unspecified.

Inference for the vector of parameters $\boldsymbol{\beta}$ relies on a so-called 'partial likelihood' which, in the case where there are no ties among the uncensored observations, that is, when $D_j = 1$, for all j, takes the form

$$\mathcal{L}(\boldsymbol{\beta}) \propto \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^t\mathbf{Z}_j)}{\sum_{k\in R_j}\exp(\boldsymbol{\beta}^t\mathbf{Z}_k)}, \tag{1.3}$$

where R_j represents the set of observations at risk prior to the time $Y^*_{(j)}$. As it can be seen, this partial likelihood does not depend on the unknown baseline hazard function $\lambda_0(t)$. A *profile likelihood* (Murphy & Van der Vaart (2000)) method, considered by Breslow (1974) for the Cox PH model, is often used to obtain this formulation. The general principle consists in expressing a parameter of the model as a function of the other parameters in such a way that the former one does not appear anymore in the likelihood function. It is often used when there is a 'nuisance' parameter. This is a two-step approach which consists, in the context of the Cox PH model, in

1. estimating non-parametrically $\lambda_0(t)$ considered as a 'nuisance' parameter, given $\boldsymbol{\beta}$,

2. replacing $\lambda_0(t)$ by its estimator in the likelihood function.

Hence, the partial likelihood (1.3) is obtained. An alternative estimation method for $\boldsymbol{\beta}$ has also been proposed by Kalbfleisch & Prentice (1973) and relies on a marginal likelihood function. Its principle is briefly described in Chapter 2 Section 2.1.2.

Parametric versions of the Cox PH model also exist. In such a case, a parametric distribution is specified for the survival time T. The main parametric formulations that will be used in the forthcoming chapters are mentioned in Table 1.1 with a description of the corresponding baseline hazard functions.

Table 1.1: *Main parametric distributions considered for T with the associated hazard function.*

Distribution	*Hazard function*	*Parameter(s)*
exponential	$\lambda_0(t) = \lambda$	$\lambda > 0$
Weibull	$\lambda_0(t) = \lambda\rho t^{\rho-1}$	$\lambda > 0$, $\rho > 0$
Gompertz	$\lambda_0(t) = \lambda \exp(\alpha t)$	$\lambda > 0$, $\alpha > 0$
log-logistic	$\lambda_0(t) = \frac{\kappa t^{\kappa-1}\lambda}{1+\lambda t^{\kappa}}$	$\lambda > 0$, $\kappa > 0$

The Accelerated Failure Time Model

Despite its popularity, the Cox PH model is not always appropriate as the hazard functions are not always proportional. A popular alternative modelling is the so-called Accelerated Failure Time (AFT) model, first mentioned by Cox (1972) and further studied by Prentice (1978), which models the survival function as

$$S(t|\boldsymbol{z}) = S_0\left\{\frac{t}{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}\right\}, \tag{1.4}$$

where $S_0(\cdot)$ is a parametric baseline survival function. As it can be seen, covariates act multiplicatively on t, meaning that they 'accelerate' or 'slow down' the time scale or the 'speed' at which the event happens relatively to the baseline, that is for $\boldsymbol{z} = 0$. The hazard function of the model is given by

$$\lambda(t|\boldsymbol{z}) = \lambda_0\left\{\frac{t}{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}\right\}\{\exp(\boldsymbol{\beta}^t\boldsymbol{z})\}^{-1},$$

showing that, contrarily to the Cox PH model, the hazards are not proportional as their ratio is not constant over the values of t. Nevertheless, there exists a relationship between the two models. When a Weibull distribution is assumed for T and the baseline hazard in the Cox PH model is $\lambda_0(t) = \rho\lambda t^{\rho-1}$, both models are equivalent.

The AFT model is also sometimes expressed through a log-linear formulation given by

$$\log(T) = \beta_0 + \boldsymbol{\beta}^t \mathbf{Z} + \sigma\epsilon,$$

with β_0 an intercept term, σ a scale parameter and ϵ a random error term. Depending on the choice for the distribution of ϵ, different models will be obtained. When the distribution of ϵ is left unspecified, we obtain a semi-parametric AFT model.

1.2 Cure Models

When survival data contain a cure fraction, we distinguish two different types of observations:

- those who experience the event and that are therefore considered as susceptible to the event or 'uncured', and
- those who never experience event, and that are then non-susceptible or 'cured' regarding the event of interest.

In such a situation, survival analysis concepts and quantities are modified, justifying the need for cure models. In this section, we first explain the consequences of a cure fraction on the classical survival quantities presented before, and we briefly introduce the two broad classes of cure models. Note that this latter point will only consist in a general presentation of both of these classes of models. A detailed literature review containing further developments will be provided in Chapter 2.

1.2.1 Cure Fraction and Survival Quantities

When considering a cure fraction, a convention consists in assuming that a cured subject is such that $T = \infty$, in order to represent the fact that the event never happens, while $T < \infty$ for a non-cured subject. A first consequence is that, when t goes to infinity, a fraction of the observations is still event free. The survival function is then improper, that is,

$$\lim_{t\to\infty} S(t) > 0.$$

This limiting value, denoted by $1 - p$, corresponds to the proportion of cured observations, called the cure rate. Equivalently, the cumulative hazard function is bounded from above, that is,

$$\lim_{t\to\infty} \Lambda(t) < \infty.$$

In other words, when t becomes large, the accumulated instantaneous risk of experiencing the event does not become infinite, but reaches a plateau, meaning that some subjects can not experience the event. The limiting value of the cumulative hazard function is equal to $-\log(1-p)$.

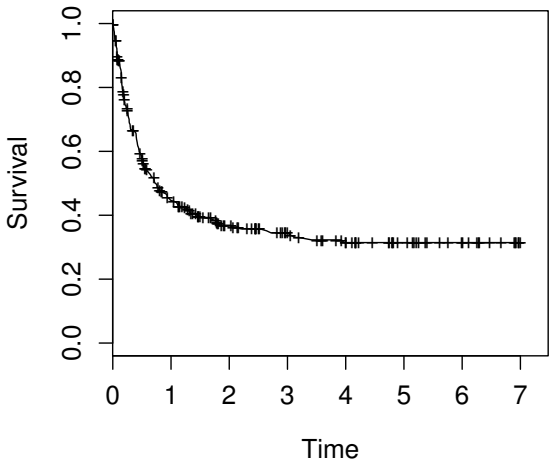

Figure 1.1: *Kaplan-Meier estimator for 300 data points, simulated from a model containing a cure fraction (+ : censored observations)*

Since the survival times are subject to censoring, the cure status is not directly observed. In fact, if we denote by B the uncure status, such as $B = I(T < \infty)$, it is obvious that an uncensored observation is such that $B = 1$, since $\Delta = 1$ and consequently $Y = T$. On the contrary, when an observation is censored, $\Delta = 0$, and therefore $Y = C$. However, given that censoring affects both cured and non-cured subjects – for the cured subjects because the event never happens, and for the uncured individuals because the follow-up cannot be infinite – it is not possible to determine the value of B in that case. Therefore, the cure status is only partially observed through the censoring indicator.

An implication of the latency of the cure status is when building the likelihood function for cure survival data. In fact, the information provided by the observations can only be of two types as in classical survival analysis, censored or uncensored. Therefore, uncensored observations contribute by means of the density function, and censored observations by means of the survival function to the likelihood function. No distinction is made between cured and uncured censored subjects. The likelihood function has then the same form as Equation (1.1). However, the survival and the density functions will be different as they take into account the presence of a cure fraction (further details will be given in Chapter 2).

In order to illustrate the existence of a cure fraction, we simulate 300 data points from a model in which 32% of the observations are cured and 40% are censored, accounting not only for the cured observations, but also for the censored uncured subjects, which represent 8% of the population. Figure 1.1 shows the Kaplan-Meier estimator of the survival function of these 300 observations.

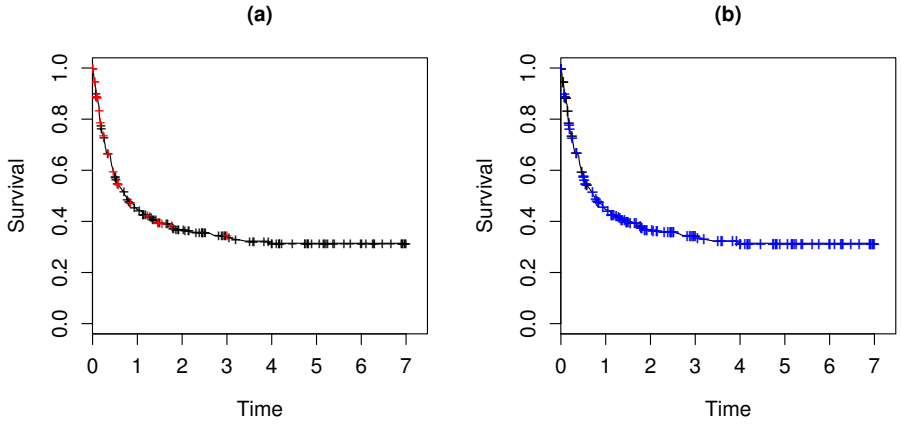

Figure 1.2: *Kaplan-Meier estimator for 300 data points, simulated from a model containing a cure fraction (a) + : censored uncured observations - (b) +: censored cured observations.*

As it can be seen, there is a clear plateau in the right tail, showing the fact that the survival function is improper, and of which the height is an estimator of the cure proportion $1-p$. As mentioned in Section 1.1.4, when there is no cure fraction, the Kaplan-Meier estimator is inconsistent in the right tail. Therefore certain conditions need to be fulfilled in order to be sure that the height of the plateau estimates well $1-p$ and corresponds to a cure fraction. In fact, it could happen that some of the observations in the plateau correspond to censored uncured observations and in that case the height of the plateau will be larger than $1-p$. Formal identifiability conditions will be given in Chapter 2, but informally speaking we can say that if we have a long plateau, corresponding to a sufficiently long follow-up, that contains a 'large' number of data points, we can be confident that (almost) all observations in the plateau correspond to cured observations as represented in Figure 1.2 which shows the Kaplan-Meier estimator for the simulated example but distinguishing cured from uncured censored subjects. Note that observing a long plateau with a substantial number of data points is essential on an empirical point of view to consider the existence of a cure fraction in the data, but there should also exist a contextual evidence for the existence of a cure fraction.

1.2.2 Cure Survival Models

As mentioned at the beginning of this chapter, there exist two classes of cure models, the mixture cure model and the promotion time cure model. The works by Boag (1949) and Berkson & Gage (1952), to whom we owe cure models, have introduced the idea of a mixture cure model which has been further developed by Farewell (1977) and Farewell (1982). The basis of this approach consists in considering that the population of interest is actually a mixture between a cured

and a non-cured sub-population. As the uncured status is latent, the survival function of the entire population $S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = P(T > t|\mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})$, where $\mathbf{X}$ denote a second vector of covariates which might be identical to $\mathbf{Z}$, or partially or completely different from $\mathbf{Z}$, can be modelled as a mixture model taking the form

$$S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = 1 - p(\boldsymbol{x}) + p(\boldsymbol{x})S_u(t|\boldsymbol{z}), \tag{1.5}$$

where $p(\boldsymbol{x}) = P(B = 1|\boldsymbol{x})$ is the probability of being susceptible or uncured (often called the incidence), and $S_u(t|\boldsymbol{z}) = P(T > t|\boldsymbol{z}, B = 1)$ is the (proper) conditional survival function of the susceptibles (often called the latency) such that $\lim_{t\to\infty} S_u(t|\boldsymbol{z}) = 0$. Therefore, it follows that $\lim_{t\to\infty} S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = 1 - p(\boldsymbol{x})$. Note that model (1.5) is saying that the probability of being uncured only depends on $\boldsymbol{x}$ (and not on $\boldsymbol{z}$) and that the conditional survival function of the susceptibles only depends on $\boldsymbol{z}$ (and not on $\boldsymbol{x}$).

The promotion time cure model, also called bounded cumulative hazard model or PH cure model, is a more recent proposal introduced by Yakovlev et al. (1996). This model is an adaptation of the Cox PH model to allow for a cure fraction. To understand how it is built, we first need to start from the cumulative hazard function. As mentioned previously, in the presence of a cure fraction, the cumulative hazard is bounded from above, that is, $\lim_{t\to\infty} \Lambda(t) < \infty$. By assuming that this limiting value is equal to θ, $\theta > 0$, the cumulative hazard function can be written as $\Lambda(t) = \theta F(t)$, where $F(t)$ is a proper distribution function. Given that $S(t) = \exp[-\Lambda(t)]$, it follows that the survival function $S_{pop}(t|\boldsymbol{x}) = P(T > t|\mathbf{X} = \boldsymbol{x})$ can be written as

$$S_{pop}(t|\boldsymbol{x}) = \exp\left\{-\theta(\boldsymbol{x})F(t)\right\}, \tag{1.6}$$

where $\boldsymbol{x}$ now represents the complete vector of covariates, and $\theta(\boldsymbol{x})$ captures the effect of the covariates $\boldsymbol{x}$ on the survival function $S_{pop}(t|\boldsymbol{x})$. For this model the cure rate is given by $\lim_{t\to\infty} S(t|\boldsymbol{x}) = \exp\{-\theta(\boldsymbol{x})\}$. One often chooses $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$. Note that this model includes an intercept, while the Cox PH model without a cure fraction does not, since it supposes that $\Lambda_0(t)$ tends to infinity when t tends to infinity, and an intercept would therefore not be identifiable. This formulation of the promotion time cure model is one that is encountered the most in the literature. However, covariates may also be introduced in $F(t)$. These formulations will be further described in Chapter 2.

1.3 The Receiver Operating Characteristic Curve

In Chapter 4, we propose a methodology to diagnose the classification performance of a continuous classifier to predict the cure status based on cure survival data. Often, when one wants to evaluate to which extend a continuous classifier correctly performs a binary classification, the Receiver Operating Characteristic or *ROC* curve is used. Initially developed during World War II to detect enemy objects in battlefields, this graphical tool has increasingly

spread, and is now used in a broad range of fields. Among others, we can cite medicine to evaluate the performance of a diagnostic test in predicting a disease for example, in psychology to investigate the performance of a scale related for example to depression, in meteorology when developing a tool to forecast an event such as a tornado or rain based on climate variables for example, but also in machine learning to evaluate the accuracy of a classification tree for example.

In order to motivate the concepts behind the ROC curve, this section starts with a description of the evaluation of the diagnostic ability of a binary classifier, and then extends this concept to the case where it is continuous, which leads to the ROC curve. The concepts and methods related to binary classification evaluation and ROC curves are only briefly described in this section. The textbook by Pepe (2003) provides a very good insight into that topic for all kinds of classifiers, while the textbook by Krzanowski & Hand (2009) focuses on ROC curves.

1.3.1 Binary Classifier

Let us consider a set of subjects, that we would like to classify into two classes, called cases and controls, based on a classifier M. The classes are represented by a binary variable, denoted by D, such that $D = 1$ for a case, and $D = 0$ for a control.

For a binary classifier, that is, when $M = 1$ when the subject is classified as a case, and $M = 0$ when it is classified as a control, there are four outcomes for the classification, which are given in Table 1.2. A correct classification is

Table 1.2: *Possible outcomes of a binary classification.*

	$D = 0$	$D = 1$
$M = 0$	true negative	false negative
$M = 1$	false positive	true positive

associated with true positive and true negative subjects, that is, subjects which are classified as a case (resp. a control) when they are effectively a case (resp. a control). On the contrary, false negative and false positive subjects, that is subjects classified as a control (resp. a case) when they are actually a case (resp. control), are representative of an incorrect classification. To evaluate the classification performance of M, a possibility is to consider the overall misclassification rate given by

$$P(M \neq D) = P(D = 1)\ P(M = 0|D = 1) + \ P(D = 0)\ P(M = 1|D = 0),$$

which gather both false positive and false negative error probabilities. However, it is often of interest to report both the false negative proportion, $P(M = 0|D = 1)$, and the false positive proportion, $P(M = 1|D = 0)$, instead of an overall value. One reason is that these two types of errors do not have the same

implication. In the context of a diagnostic test for a lethal but curable disease, for example, false negative error means that the subject is diagnosed as disease free, when it is not. In such a case, a potential consequence is death. A false positive error, conversely, means that the subject is diagnosed as having the disease while it is not. Even if there are some consequences in such a case, they are less serious in the long term than for a false negative error. Therefore, the classifier is usually evaluated, at least in medicine, based on the amount of false positive and false negative errors it produces. Furthermore it is important to look at both of them. Indeed, a test that would be always positive, that is, when $P(M = 1|D = 1) = P(M = 1|D = 0) = 1$, would have a false negative proportion $P(M = 0|D = 1)$ equal to 0, but also a true negative proportion $P(M = 0|D = 0) = 0$ which is not convenient.

A perfect classifier is such that $P(M = 0|D = 1) = P(M = 1|D = 0) = 0$, that is, a classifier which does not produce any false positive and false negative errors, or equivalently, a classifier which only produces true positive and true negative subjects.

1.3.2 Continuous Classifier

For a continuous classifier, it is not possible to directly classify a set of subjects into two classes based on a continuous quantity. It is therefore necessary to dichotomise M by considering a threshold, denoted hereafter by k. A convention is to classify a subject as a case when the classifier is such that $M > k$. In such a context, the true positive proportion, also referred to as the *sensitivity*, is given by

$$Se(k) = P(M > k|D = 1),$$

and the true negative proportion, also named the *specificity*, corresponds to

$$Sp(k) = P(M \leq k|D = 0).$$

Different thresholds correspond to different sensitivities and specificities. The ROC curve is a graphical representation of all possible combinations of the sensitivity and one minus the specificity that can be obtained from all possible dichotomised version of the classifier, based on the values of the threshold k. It plots the sensitivity against one minus the specificity, for all possible values of $k \in \mathbb{R}$, and its equation is given by

$$ROC(u) = Se\left\{(1 - Sp)^{-1}(u)\right\},\ 0 < u < 1,$$

where u is an index. The ROC curve is an increasing function lying in the quadrant $(0, 1) \times (0, 1)$. A graphical example is given in Figure 1.3. The position of the curve in the quadrant $(0, 1) \times (0, 1)$ indicates the diagnostic performance of the classifier. In fact, the sensitivity and the specificity both correspond to the distribution function of the classifier in the two groups, respectively. When the classifier is uninformative, that is, when it is not possible to distinguish cases from controls based on M, the distribution of the classifier is the same in the two groups, and therefore $P(M > k|D = 1) = P(M > k|D = 0)$, for all k. It means that the sensitivity equals to one minus the specificity. The

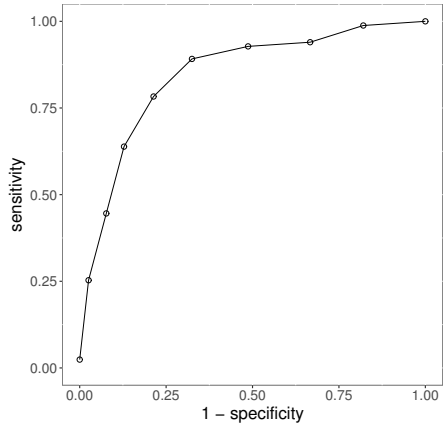

Figure 1.3: *Graphical representation of a ROC curve. Each point represents a combination of the sensitivity and one minus the specificity for different values of the threshold.*

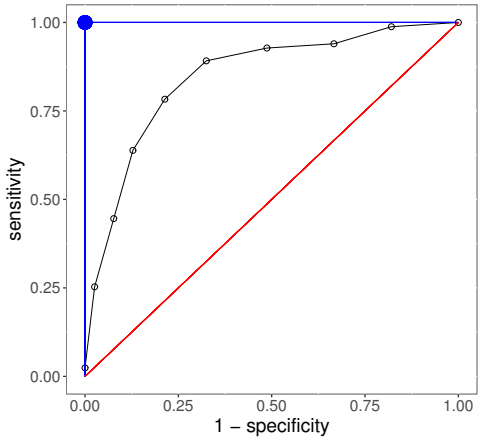

Figure 1.4: *Graphical representation of a ROC curve showing the two extreme performances of a classifier.*

classifier is then equivalent to a random guess and it corresponds, graphically, to the bisector, represented in red in Figure 1.4. On the contrary, a perfect classifier has zero false positive and zero false negative error. All subjects are therefore perfectly classified and the sensitivity and the specificity are equal to 1. Graphically, it corresponds to the point $(0, 1)$, the blue dot in Figure 1.4, where $P(M > k|D = 1) = 1$, and where $P(M > k|D = 0) = 0$. Only one ROC curve reaches this point, the one lying along the left and the upper border of quadrant as shown in blue in Figure 1.4. The ROC curve lies between these two extreme situations, only in the upper part of the quadrant. The closer the curve will be to the point $(0, 1)$, the better the classifier performance will be. Note that, if case subjects are such that $M \leq k$, the ROC curve lies in the lower part of the quadrant. In such a case, it suffices to consider $(-M)$ to obtain a ROC curve lying in the upper part of the quadrant.

Alongside, the ROC curve, one usually computes the area under the ROC curve (AUC) which provides a single value summary of the performance of the classifier. Its equation is given by

$$AUC = \int_0^1 ROC(u)\ du. \tag{1.7}$$

When the classifier is uninformative, the AUC is equal to 0.5, while for a perfect classifier, the AUC of the corresponding ROC curve equals 1.

Non-parametric, semi-parametric and parametric methods have been proposed to estimate the ROC curve. Let (D_i, M_i), $i = 1, \ldots, n$, be i.i.d copies of (D, M). The simplest approach consists in estimating the sensitivity and the specificity by their empirical distribution functions given, respectively, by

$$\check{Se}(k) = 1 - \frac{1}{\check{N}_1} \sum_{i=1}^{n} \check{W}_{i1}\ I(M_i \leq k),$$

$$\check{Sp}(k) = \frac{1}{\check{N}_0} \sum_{i=1}^{n} \check{W}_{i0}\ I(M_i \leq k),$$

where $\check{W}_{i1} = D_i$, $\check{W}_{i0} = 1 - \check{W}_{i1}$, $\check{N}_1 = \sum_{i=1}^{n} \check{W}_{i1}$, and $\check{N}_0 = n - \check{N}_1$. The corresponding ROC curve estimator takes the form of a step function with jumps at each M_i. When there are no ties in the data, the step function has vertical jumps of size $1/\check{N}_1$ associated with subjects in the case group for whom the classifier is equal to M_i, and horizontal jumps corresponding to subjects in the control group with classifier value equal to M_i, and of which the size is $1/\check{N}_0$. Ties between subjects in the same class produce larger jumps, while ties between cases and controls result in a 'diagonal' jump corresponding to vertical and horizontal jumps at the same time.

1.4 The EM Algorithm

After the description of these statistical quantities and methods, let us move now to the description of a tool that will be used in Chapter 2 and in Chapter 3.

When one is interested in estimating a model by maximum likelihood based on 'incomplete data', that is, when the model depends on unobserved variables, the *Expectation-Maximisation* (EM) algorithm, proposed by Dempster et al. (1977), is often considered, notably for mixture models. In this section, we describe the general principle of the EM algorithm which is used to estimate some modelling of the mixture cure model (1.5).

Let us consider that we have a vector of observed variables $\mathbf{X}$ coming from a model parametrised by a vector of unknown parameters $\boldsymbol{\theta}$ with likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{X})$. Let us further assume that there exists another vector of variables $\mathbf{Z}$ which is only partially observed (also referred to as the *latent* variables) through $\mathbf{X}$, and that the likelihood function for the complete-data, that is for $(\mathbf{X}, \mathbf{Z})$, is given by $\mathcal{L}_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$. Note that, since the vector $\mathbf{Z}$ is only partially observed, we talk about *incomplete data* and the complete-data likelihood $\mathcal{L}_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$ is not available.

Suppose now that one is interested in estimating $\boldsymbol{\theta}$ by maximum likelihood. In some situations, finding the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\theta; \mathbf{X})$ from $\log \mathcal{L}(\theta; \mathbf{X})$ is a complex problem which can be simplified by considering the complete-data likelihood. That is what the EM algorithm proposes. Since $\mathcal{L}_c(\theta; \mathbf{X}, \mathbf{Z})$ is not available, the principle of the EM algorithm consists in, first, 'estimating' the complete-data likelihood from $\mathbf{X}$ by computing the expectation of $\log \mathcal{L}_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$, given the observed data $\mathbf{X}$ and the value of the parameters $\boldsymbol{\theta}$, and, second, to use the 'estimated' log-complete-data likelihood function to estimate $\boldsymbol{\theta}$. However, since $\boldsymbol{\theta}$ is required to 'estimate' $\log \mathcal{L}_c(\theta; \mathbf{X}, \mathbf{Z})$, an iterative procedure is necessary. The EM algorithm then alternates iteratively between two steps, an Expectation or E-step and a Maximisation or M-step. At the m^{th} iteration of the algorithm, these two steps are as follows:

> *E-step* : computation of the expectation of the logarithm of the complete data likelihood, given the observed data and the current value of the parameters, with respect to the latent variable, that is,
>
> $$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}\right) = E_{\mathbf{Z}}\left\{\log \mathcal{L}_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})|\mathbf{X}, \boldsymbol{\theta}^{(m-1)}\right\},$$
>
> where $\boldsymbol{\theta}^{(m-1)}$ denotes the current parameter values obtained at the $(m-1)^{th}$ iteration.
>
> *M-step* : maximisation of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}\right)$ with respect to $\boldsymbol{\theta}$. An estimator for $\boldsymbol{\theta}$ is then given by
>
> $$\boldsymbol{\theta}^{(m)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}\right).$$

The algorithm alternates between the two steps until convergence.

1.5 Outline of the Thesis

The objective of this thesis is to investigate different aspects of cure models by focusing on the three following directions. Each of them constitutes a pa-

per that has been published, has been accepted for publication, or has been submitted for publication in an international journal.

Chapter 2. Cure Models in Survival Analysis: a Literature Review
First, both the mixture cure model and the promotion time cure model have been the topic of much research. For the mixture cure model, several model assumptions have been proposed leading to parametric, semiparametric, and nonparametric families of mixture cure models. Moreover, many different estimation methods have been developed, and other topics, such as the important issue of identifiability and the question of the verification of the model have also been investigated. On the side of the promotion time cure model, less has been done, but the literature proposes both Bayesian and frequentist estimation frameworks, contrarily to the mixture cure model which is mainly studied in the frequentist setting. Furthermore, measurement errors have only been studied for promotion time cure models. More recently, a literature on models that unify the mixture cure model and the promotion time cure model into one single over-arching model, avoiding hence the delicate task of choosing between these two models, has flourished. The first contribution of this thesis consists then in providing a state of the art on cure models, reviewing all the points mentioned above and providing a formal and a numerical comparison of the two models through an application on real data. This work has been published in the paper Amico & Van Keilegom (2018*b*).

Chapter 3. The Single-Index/Cox Mixture Cure Model
The second contribution of this thesis focuses on the mixture cure model and more precisely on the probability of being uncured. Often, this quantity is modelled parametrically, assuming a logistic regression model. However, there is no reason to strictly constrain and limit the uncure proportion to a logistic form. Our aim is then to propose a more flexible modelling approach for the uncure proportion, by assuming a single-index structure, that is, a generalised linear model in which the link function is left unspecified, and by considering a Cox proportional hazards model for the conditional survival function of uncured subjects. This chapter is based on Amico et al. (2018).

Chapter 4. Assessing Cure Status Prediction from Survival Data Using ROC Curves
Third, beside modelling and model selection, an important topic of statistical analysis is the question of model assessment, that is, the evaluation of the predictions that can be made from a given model. For cure models, predictions can be performed for two outcomes, the survival at a given time and the cure status, both of them being binary. Often, when one wants to assess the binary classification performance of a continuous classifier, the Receiver Operating Characteristic (ROC) curve is considered. However, while standard ROC curves suppose that the classes of the outcome are fully observed, building a ROC curve from cure survival data is a non trivial problem since survival data are subject to censoring and hence, the cure status is unobserved. This last

contribution concerns therefore the development of ROC curves to evaluate the performance of cure status prediction that can be made from survival data in the presence of a cure fraction. The content of this chapter is based on Amico & Van Keilegom (2018*a*).

Finally, to close this thesis, **Chapter 5** provides some general conclusions and the presentation of some further research.

Chapter 2

Cure Models in Survival Analysis: a Literature Review

As introduced in Chapter 1, there exists two broad classes of cure models, the mixture cure models and the promotion time cure models. Beside these two approaches, there also exist some other works which have developed broader classes of cure models, embedding both the mixture cure model and the promotion time. Many different contributions have been made to the literature on that topic and there was a need to make an inventory of the existing works. A first contribution of this thesis is then to make a detailed review of this literature. This second chapter first starts in Section 2.1 with a presentation of topics related to the mixture cure model, and continues in Section 2.2 with the aspect related to the promotion time cure model. Section 2.3 closes this chapter with a discussion about the difference and the relationship between the mixture cure model and the promotion time cure model, and describes unifying models.

Throughout this chapter, models and methods that have been proposed in the literature on cure models are illustrated on a dataset on breast cancer coming from Wang et al. (2005). The dataset consists in time to distant metastasis expressed in days, for 286 patients that experienced a lymph-node-negative breast cancer between 1980 and 1995. Four covariates are considered: the age of the patient (ranging from 26 to 83 with a median of 52 years old), the estrogen receptor (ER) status (0 = ER−: less than 10 fmol per mg protein - 77 patients, 1 = ER+: at least 10 fmol per mg protein - 209 patients), the size of the tumour (ranging from 1 to 4 with a median of 1), and the menopausal status (0 = premenopausal - 129 patients, 1 = postmenopausal - 157 patients). Figure 2.1 shows a graphical representation of the Kaplan-Meier estimator of the survival function. As can be seen, the curve levels-off at a value greater than 0, at around 60%, and there is a large plateau of approximately 2 770 days, a strong sign of the presence of a cure fraction. Moreover, among the

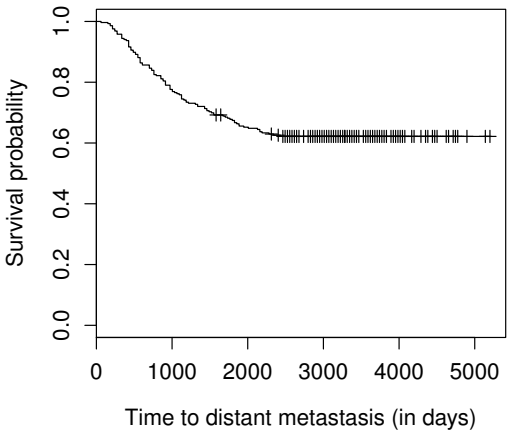

Figure 2.1: *Kaplan-Meier estimator for the data from Wang et al. (2005) (+ : censored observations)*

286 patients, 179 are censored, among which 88.3% are censored after the last observed event time. Hence, a lot of the censoring times are located in the plateau indicating that a cure model can be considered. Finally, there is a strong medical evidence for the presence of cured patients in breast cancer relapse at early stage. It turns out that this dataset is a perfect example of survival data with a cure fraction.

Before starting, we have to mention that many other topics have been investigated in the framework of cure models. Among others, we can mention the introduction of frailties, competing risks, quantile regression, or different types of censoring such as interval-censoring for example. However, in order to give the reader a better understanding of the basis of cure models, we will focus on a detailed description of the two main classes of cure models and unifying approaches in the classical random right censoring setting.

Let us finish this introduction by briefly mentioning some other works on cure models. The textbook by Maller & Zhou (1996), which is completely devoted to the topic of cure models, gives a nice introduction to many of the specific aspects of cure models. Recently, Peng & Taylor (2014) wrote a review paper on cure models, in which they give a detailed overview of the existing cure models.

This chapter is based on

> Amico, M. and Van Keilegom, I. (2018b). Cure models in survival analysis, *Annual Review of Statistics and its Application*, **5**, 311–342.

2.1 Mixture Cure Models

We start in Section 2.1.1 with the most fundamental issue related to the definition of a model: its identifiability. Although this is often neglected in the statistical literature, it should be the first task when studying a new model. In Section 2.1.2, we examine several models for the components of the mixture cure model, we see how they can be estimated, how the estimators can be computed in practice, and how they behave asymptotically. Section 2.1.3 describes different issues related to the verification of the model, such as goodness-of-fit tests, variable selection, and model diagnostics. Finally, in Section 2.1.4 we apply the most common mixture cure model to the breast cancer data introduced in the introduction of this chapter.

2.1.1 Identifiability

We have already mentioned the issue of identifiability of the model in a very informal way in the Chapter 1 (see Section 1.2). Indeed, we said that in order to identify (in a non formal way) the cure proportion, the 'plateau' in the plot of the survival function for the whole population should only consist of cured subjects. When the plateau stays constant for a long time without decreasing even incrementally, we can be relatively confident that all uncured subjects had their event before the start of the plateau, and hence the cure fraction corresponds to the height of the plateau. This informal analysis can be made more rigorous by saying that

$$\tau_{F_u} < \tau_G \tag{2.1}$$

(we omit covariates here for simplicity), where $F_u = 1 - S_u$, G is the censoring distribution, and $\tau_F = \inf\{t : F(t) = 1\}$ for any distribution F . This assumption will be crucial in most semi- and nonparametric papers on modelling of mixture cure models.

When we talk about the identifiability of a model, we should distinguish two common but different definitions of identifiability. The first definition (the weakest of the two) states that the mixture cure model (1.5) is identifiable within families $\mathcal{P}$ and $\mathcal{S}$ of functions corresponding to the incidence and the latency respectively, if the equality

$$1 - p_1(\boldsymbol{x}) + p_1(\boldsymbol{x})S_{u1}(t|\boldsymbol{z}) = 1 - p_2(\boldsymbol{x}) + p_2(\boldsymbol{x})S_{u2}(t|\boldsymbol{z}), \qquad \text{for all } t, \boldsymbol{x}, \boldsymbol{z}, \tag{2.2}$$

for some functions $p_1, p_2 \in \mathcal{P}$ and $S_{u1}, S_{u2} \in \mathcal{S}$, implies that $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ for all $\boldsymbol{x}$, and that $S_{u1}(t|\boldsymbol{z}) = S_{u2}(t|\boldsymbol{z})$ for all t and $\boldsymbol{z}$. This was studied in full detail and in a rigorous way by Hanin & Huang (2014), who consider several choices of the classes $\mathcal{P}$ and $\mathcal{S}$ under which the mixture cure model (1.5) is or is not identifiable. The paper by Hanin & Huang (2014) is an important improvement over earlier attempts to study the identifiability of the mixture cure model, in the sense that earlier papers contained mistakes in the proofs or did not study the problem in full generality. Note that this first definition of identifiability does not depend on the censoring mechanism, and hence condition (2.1) does not play any role here.

A second type of identifiability is related to the uniqueness of the parameters of the model in another sense – namely, in the sense that there is a unique set of parameters for which the expected log-likelihood is maximal. So, instead of equating two models, which are unrelated to the type of data at hand, we look at the likelihood, which is based on the density of the observed variables (that are subject to random right censoring in our case) under the given model. The conditions under which there exists a unique $p \in \mathcal{P}$ that maximises the expected log likelihood when $\mathcal{P}$ is a parametric class of probability functions coming, for example, from a logistic model, have been rigorously studied in Patilea & Van Keilegom (2018) (see Proposition 3.1), while Xu & Peng (2014) studied the case where $\mathcal{P}$ is nonparametric. In both papers, assumption (2.1) turns out to be a crucial assumption to ensure identifiability of the model (although the condition (7) of Xu & Peng (2014) could be relaxed to $\tau_{F_u}(\boldsymbol{z}) < \tau_G(\boldsymbol{z})$ for all $\boldsymbol{z}$).

We are now ready to study different models for the incidence $p(\cdot)$ and for the latency $S_u(\cdot|\cdot)$, and their corresponding estimation procedures.

2.1.2 Modelling Approaches and Inference

The literature on mixture cure models offers a wide variety of modelling approaches ranging from fully parametric to completely nonparametric models. In what follows, we assume that we have i.i.d. data $(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, having the same distribution as $(Y, \Delta, \mathbf{X}, \mathbf{Z})$, $\dim(\mathbf{X}) = p$, $\dim(\mathbf{Z}) = q$, and for given values of $\mathbf{X}$ and $\mathbf{Z}$, the event time T follows the mixture cure model (1.5). Let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be the order statistics of the observations $Y_1, \ldots, Y_n$, and let $Y^*_{(1)} < \ldots < Y^*_{(r)}$ be the distinct ordered uncensored observations, assuming there are $r \leq n$ in total.

Fully Parametric Models

The pioneer works on the mixture cure models are fully parametric approaches due to Boag (1949) and Berkson & Gage (1952). In both papers, the incidence is modelled as a constant and the survival function for uncured observations takes the form of a log-normal model and an exponential model, respectively, not depending on covariates.

Introduction of covariates in the incidence are due to Farewell (1977) which assume a logistic regression model for the probability of being uncured, that is,

$$p(\boldsymbol{x}) = \frac{\exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})}{1 + \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})},$$

where γ_0 is an intercept term, and models the latency according to an exponential distribution, that is,

$$S_u(t) = \exp(-\lambda t).$$

The introduction of covariates in the latency was proposed by Farewell (1982) who considered a Weibull model for the conditional survival function of

the form

$$S_u(t|\boldsymbol{z}) = \exp\left\{-\lambda\ \exp(\boldsymbol{\beta}^t\boldsymbol{z})\ t^\rho\right\},$$

where $\lambda > 0$ is a shape parameter and $\rho > 0$ is a scale parameter. Another proposal came from Ghitany et al. (1994) who modelled the effect of covariates on the latency with an exponential model.

For all of these models, a maximum likelihood estimation method is proposed based on the likelihood function

$$\prod_{i=1}^{n}\left\{p(\mathbf{X}_i)f_u(Y_i|\mathbf{Z}_i)\right\}^{\Delta_i}\left\{1-p(\mathbf{X}_i)+p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)\right\}^{1-\Delta_i}, \tag{2.3}$$

where $f_u(t|\boldsymbol{z}) = -(d/dt)S_u(t|\boldsymbol{z})$. As already mentioned in Chapter 1, the likelihood function is derived as for classical survival models and makes no distinction between cured and uncured censored observations since the cure status is unknown. Uncensored observations contribute through the density function, which is equal to $f(t|\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x})f_u(t|\boldsymbol{z})$, and censored observations contribute through the survival function given by the mixture cure model (1.5). To estimate the logistic/Weibull mixture cure model, Farewell (1982) proposed to maximise this likelihood function numerically using the Newton-Raphson technique.

Other parametric mixture cure models include an AFT model for the latency. Yamaguchi (1992) considered the extended family of generalized gamma models (Prentice (1974)) for $\log(T^*) = \beta_0 + \boldsymbol{\beta}^t\mathbf{Z} + \sigma\epsilon$, where T^* is the survival time for uncured observations, β_0 is an intercept term, $\sigma > 0$ is a scale parameter, and ϵ is an error term with density function

$$f_\epsilon(t) = \begin{cases} \frac{|\lambda_\epsilon|}{\Gamma(\lambda_\epsilon^{-2})}(\lambda_\epsilon^{-2})^{\lambda_\epsilon-2}\exp(\lambda_\epsilon t - e^{\lambda_\epsilon t}), & \text{if } \lambda_\epsilon \neq 0 \\ \frac{1}{(2\pi)^{1/2}}\exp(-t^2/2) & \text{if } \lambda_\epsilon = 0, \end{cases}$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function, and λ_ϵ is a shape parameter. Peng et al. (1998) proposed considering a generalized F distribution for T^*, given that T^* is said to have a generalised F distribution with location parameter μ, scale parameter σ and shape parameters s_1 and s_2, if $W = [\log(T^*)-\mu]/\sigma$ is the logarithm of a random variable having a F distribution with $2s_1$ and $2s_2$ degrees of freedom. The density function of W is given by

$$f_W(w) = \left(\frac{s_1e^w}{s_2}\right)^{s_1}\left(1+\frac{s_1e^w}{s_2}\right)^{-(s_1+s_2)}B(s_1,s_2)^{-1},$$

where $B(\cdot,\cdot)$ is the beta function. In both cases, the choice of these distributions are motivated by their flexibility and because they embed the exponential, the Weibull (when $s_1 \to 1$ and $s_2 \to \infty$, we obtain the model proposed by Farewell (1982)), the log-normal and the gamma distributions as special cases among others. For the incidence, both models assume a logistic regression model as Farewell (1982). Yamaguchi (1992) and Peng et al. (1998) developed a maximum likelihood approach based on the likelihood function (2.3) for

the two proposals in order to estimate these two models. First, the Newton-Raphson algorithm is used to maximise the likelihood function with respect to $(\gamma_0, \boldsymbol{\gamma}, \beta_0, \boldsymbol{\beta}, \sigma)^t$. In a second step, a search for the value of the shape parameter(s) that maximise the likelihood is made (λ_ϵ for the extended family of generalized gamma models, s_1 and s_2 for the generalized F model).

Logistic/Cox (LC) Mixture Cure Models

Semi-parametric mixture cure models are a second class of mixture cure models that have been extensively studied in the literature. The main motivation is that they avoid the restrictions imposed by parametric conditional survival functions. Most of them focus on the latency while they keep the logistic regression form for the incidence. A first group of models is composed of mixture cure models assuming a Cox PH model for the conditional survival function, that is,

$$S_u(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t \boldsymbol{z})},$$

where the unspecified baseline survival function is given by $S_0(t) = P(T > t|\mathbf{Z} = 0, B = 1)$. Note that, while the conditional hazard function $\lambda_u(t|\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^t \boldsymbol{z})$, where $\lambda_0(t) = f_0(t)/S_0(t)$ is the baseline hazard function with $f_0(t) = -(d/dt)S_0(t)$, does satisfy the proportional hazard property, this mixture cure model, which has been introduced by Kuk & Chen (1992), does not satisfy this property, contrarily to the Cox PH model. As a consequence, the profile likelihood approach described in Chapter 1 Section 1.1.4 which is usually considered to obtain the partial likelihood (1.3) for the Cox PH model cannot be applied for this model. In fact, it is not possible to isolate the baseline survival function in the likelihood. Likewise, because the latency is defined conditionally on the uncured status, if one considers the baseline conditional survival function as a nuisance parameter, information about the cure status will be lost. The literature contains several proposals to estimate the model taking into account this situation.

A first approach, due to Kuk & Chen (1992), adapts a marginal likelihood approach proposed by Kalbfleisch & Prentice (1973) for the classical Cox PH model. The marginal likelihood consists of integrating the likelihood function (2.3) over $Y^*_{(j)}, j = 1, ..., r$. Estimators are obtained by maximising this quantity with respect to the parameters. In practice, however, it is not possible to compute this marginal likelihood function, and it is therefore approximated by Monte Carlo methods.

Peng & Dear (2000) and Sy & Taylor (2000) proposed a second group of estimation approaches based on the EM algorithm (see Section 1.4 of Chapter 1 for a general introduction to this algorithm). The choice for this methodology is justified by the fact that the model depends on a latent variable, the cure status. Another interesting argument lies in the fact that the EM algorithm considers a complete-data likelihood, which, for the mixture cure model, takes

the form

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\gamma},\boldsymbol{\beta},S_0) =& \prod_{i=1}^{n}\{p(\mathbf{X}_i)\lambda_u(Y_i|\mathbf{Z}_i)S_u(Y_i|\mathbf{Z}_i)\}^{\Delta_i B_i}\\ &\times\prod_{i=1}^{n}\{p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)\}^{(1-\Delta_i)B_i}\\ &\times\prod_{i=1}^{n}\{1-p(\mathbf{X}_i)\}^{(1-\Delta_i)(1-B_i)},\end{aligned} \tag{2.4}$$

where $\lambda_u(t|\boldsymbol{z}) = f_u(t|\boldsymbol{z})/S_u(t|\boldsymbol{z})$ is the hazard function of the uncured observations. An interesting feature of (2.4) is that it can be rewritten as the product of two elements, namely

$$\mathcal{L}_1(\boldsymbol{\gamma}) = \prod_{i=1}^{n} p(\mathbf{X}_i)^{B_i}\{1-p(\mathbf{X}_i)\}^{1-B_i} \tag{2.5}$$

$$\mathcal{L}_2(\boldsymbol{\beta},S_0) = \prod_{i=1}^{n}\left[\{\lambda_u(Y_i|\mathbf{Z}_i)S_u(Y_i|\mathbf{Z}_i)\}^{\Delta_i B_i} S_u(Y_i|\mathbf{Z}_i)^{(1-\Delta_i)B_i}\right], \tag{2.6}$$

each of them only containing the parameters of one of the two parts of the model. It is then possible to estimate separately the incidence and the latency. In such a case, it becomes possible to extend methods developed for the classical Cox PH model. In the framework of the mixture cure model, the implementation of the EM algorithm is as follows. The E-step consists in computing at the m^{th} iteration of the algorithm the expectation of logarithm of the complete-data likelihood (2.4) given the current values of the parameters $\theta^{(m-1)} = (\gamma_0,\boldsymbol{\gamma},\boldsymbol{\beta},S_0)^{(m-1)}$ and the observed data $O_i = (Y_i,\Delta_i,\mathbf{X}_i,\mathbf{Z}_i)$, with respect to the latent variable B_i. As the log-complete-data likelihood is linear in B_i, it is the same as computing

$$\begin{aligned}&E\left(B_i|O_i,\theta^{(m-1)}\right)\\ &= \Delta_i\ \left\{1\times P(B_i=1|Y=Y_i,\Delta_i=1,\mathbf{X}=\boldsymbol{X}_i,\mathbf{Z}=\boldsymbol{Z}_i,\theta^{(m-1)})\right\}\\ &\quad+(1-\Delta_i)\ \left\{1\times P(B_i=1|Y=Y_i,\Delta_i=0,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)})\right\},\end{aligned}$$

where $\theta^{(m-1)}$ denotes the set of parameter values at the $(m-1)^{th}$ iteration. Given that $P(B_i=1|Y=Y_i,\Delta_i=1,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)})=1$, and that

$$\begin{aligned}&P(B_i=1|Y=Y_i,\Delta_i=0,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)})\\ &\qquad\qquad= \frac{P(B_i=1,T>Y_i|\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)})}{P(T>Y_i|\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)})},\end{aligned}$$

it follows that

$$\begin{aligned}&E\left(B_i|O_i,\theta^{(m-1)}\right)\\ &\quad=\Delta_i+(1-\Delta_i)\frac{p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|\mathbf{Z}_i)}{1-p^{(m-1)}(\mathbf{X}_i)+p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|\mathbf{Z}_i)} = W_i^{(m)}.\end{aligned}$$

The expected log-complete-data likelihood is therefore obtained by replacing B_i by its expectation $W_i^{(m)}$ in the logarithm of (2.4).

The M-step consists of maximising the expected complete-data likelihood with respect to the parameters of the model. For the incidence, (2.5) is the same likelihood function as for a classical logistic regression model. The Newton-Raphson technique is applied to estimate the parameters. For the latency, three methods can be distinguished, all of them being based on the likelihood (2.6):

- Sy & Taylor (2000) proposed a first approach based on the profile likelihood method proposed by Breslow (1974) for the Cox PH model (see Section 1.1.4 of Chapter 1). First, the baseline conditional cumulative hazard $\Lambda_0(t) = -\log[S_0(t)]$ is estimated non-parametrically by

$$\hat{\Lambda}_0(t) = \sum_{j:Y^*_{(j)} \leq t} \frac{D_{(j)}}{\sum_{k \in R_j} W_k^{(m)} \exp(\mathbf{Z}_k^t \boldsymbol{\beta})}.$$

  This estimator is then substituted in (2.6) and the following partial likelihood is obtained (assuming no ties):

$$\tilde{\mathcal{L}}_2(\boldsymbol{\beta}|\mathbf{W}^{(m)}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^t \boldsymbol{\beta})}{\sum_{k \in R_i} W_k^{(m)} \exp(\mathbf{Z}_k^t \boldsymbol{\beta})} \right\}^{\Delta_i}, \tag{2.7}$$

  where $\mathbf{W}^{(m)} = \{W_1^{(m)}, \ldots, W_n^{(m)}\}$. Note that when $W_i^{(m)} = 1$ for all $i = 1, ..., n$, (2.7) is equal to the partial likelihood for the classical Cox PH model given by (1.3). The latency part is then estimated by maximising (2.7) with respect to $\boldsymbol{\beta}$ using the Newton-Raphson method.

- A second proposal from Sy & Taylor (2000) is a product-limit type method in which the baseline conditional survival function is first estimated non-parametrically by a step function that takes a product-limit form:

$$S_0(t) = \prod_{j:Y^*_{(j)} \leq t} \alpha_j,$$

  where $\alpha_j = S_0(Y^*_{(j)})/S_0(Y^*_{(j-1)})$. In the absence of ties, the likelihood function (2.6) is reparametrised in terms of α_j and the EM algorithm is applied in order to estimate α given $\boldsymbol{\beta}$. In a second step, the estimator of α is substituted in the expected complete-data likelihood and a profile likelihood for $\boldsymbol{\beta}$ is obtained. We refer to Sy & Taylor (2000) for more details regarding the case with ties.

- A third approach has been proposed by Peng & Dear (2000), who considered a marginal likelihood. As for Kuk & Chen (1992), the marginal likelihood function is obtained by integrating (2.6) over $Y^*_{(j)}, j = 1, ..., r$. In the absence of ties, the following marginal likelihood is obtained:

$$\breve{\mathcal{L}}_2(\boldsymbol{\beta}|W^{(m)}) \approx \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^t \boldsymbol{\beta})}{\sum_{k \in R_i} W_k^{(m)} \exp(\mathbf{Z}_k^t \boldsymbol{\beta})} \right\}^{\Delta_i}.$$

Note that this marginal likelihood is approximately equivalent to the partial likelihood (2.7) obtained by Sy & Taylor (2000). For the case with ties, we refer the reader to the paper of Peng & Dear (2000).

Another type of estimation method has been proposed by Lu (2008) in order to estimate the logistic/Cox mixture cure model. Based on a nonparametric maximum likelihood function, the main idea is to consider a non-parametric estimator for $\Lambda_0(t)$, that is a step function with jumps at all the event times, and to replace the baseline conditional hazard in (2.3) by the size of the jump made by the cumulative baseline hazard at each event time. The likelihood function is then given by

$$\prod_{i=1}^{n} \left[p(\mathbf{X}_i)\{\Lambda_0(Y_i) - \Lambda_0(Y_i-)\} \exp(\boldsymbol{\beta}^t \mathbf{Z}_i) \exp\{-\Lambda_0(Y_i) \exp(\boldsymbol{\beta}^t \mathbf{Z}_i)\}\right]^{\Delta_i}$$
$$\times \prod_{i=1}^{n} \left[1 - p(\mathbf{X}_i) + p(\mathbf{X}_i) \exp\{-\Lambda_0(Y_i) \exp(\boldsymbol{\beta}^t \mathbf{Z}_i)\}\right]^{1-\Delta_i}.$$

The major contribution of Lu (2008) is to show that the estimators of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and Λ_0 converge weakly to a zero-mean Gaussian process. It also provides an estimator of the asymptotic covariance function.

Finally, Corbière et al. (2009) proposed a penalised likelihood approach that has the advantage of producing a smooth estimator of the conditional hazard function. The method consists in considering the penalised likelihood

$$\log[\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda_0)] - \kappa \int \lambda_0''(\nu)^2 d\nu, \tag{2.8}$$

where

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda_0) = \prod_{i=1}^{n} \left[p(\mathbf{X}_i)\lambda_0(Y_i) \exp(\boldsymbol{\beta}^t \mathbf{Z}_i) \exp\{-\Lambda_0(Y_i) \exp(\boldsymbol{\beta}^t \mathbf{Z}_i)\}\right]^{\Delta_i}$$
$$\times \prod_{i=1}^{n} \left[1 - p(\mathbf{X}_i) + p(\mathbf{X}_i) \exp\{-\Lambda_0(Y_i) \exp(\boldsymbol{\beta}^t \mathbf{Z}_i)\}\right]^{1-\Delta_i},$$

$\kappa \int \lambda_0''(\nu)^2 d\nu$ is the penalisation term, and $\kappa > 0$ is a positive smoothing parameter balancing between the fit of the data and the smoothness of the function. The model is estimated by maximising the likelihood (2.8) with respect to $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and λ_0. Since there is no explicit formula of the baseline conditional hazard that maximises the likelihood, it is approximated by a linear combination of cubic normalized B-splines. They also provide a method to compute the variance of the parameter estimates based on the inverse of the matrix of the second derivatives of the penalised likelihood.

Logistic/Semiparametric AFT Models

In addition to the popular logistic/Cox (LC) mixture cure model, other papers, beginning with Li & Taylor (2002), focused on a semiparametric AFT model

for the latency. They all consider that $\log(T^*) = \beta_0 + \boldsymbol{\beta}^t\mathbf{Z} + \epsilon$ and assume unspecified density and survival functions f and S, respectively, for the error term ϵ. As for the LC mixture cure model, a logistic regression model is assumed for the incidence. Three different estimation approaches have been proposed, all based on the EM algorithm. Starting from the complete-data likelihood (2.4), these methodologies are the same as for the LC mixture cure model until the M-step for the latency estimation. For this latter part, they extend methods that have been proposed for the classical semiparametric AFT models. Starting from the expected complete-data likelihood associated with the latency given by

$$\prod_{i=1}^{n} f_\epsilon \left\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\right\}^{\Delta_i W_i^{(m)}} \times \prod_{i=1}^{n} S_\epsilon \left\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\right\}^{(1-\Delta_i)W_i^{(m)}}, \tag{2.9}$$

- Li & Taylor (2002) proposed to extend the work of Ritov (1990) based on M-estimators. Starting from the score equation for $\boldsymbol{\beta}$ given by

$$\sum_{i=1}^{n} \mathbf{Z}_i \left[-W_i^{(m)}\Delta_i \frac{f'_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}}{f_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}} + W_i^{(m)}(1-\Delta_i)\frac{f_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}}{S_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}}\right] = 0,$$

  the principle consists in replacing in the score equation $-f'_\epsilon/f_\epsilon$ by an M-estimator and the unknown survival function S_ϵ by its Kaplan-Meier estimator given $\boldsymbol{\beta}$. Because the obtained score equation is not necessarily monotone and continuous, they propose estimating the parameters by using a grid search over the range of values of $\boldsymbol{\beta}$.

- Zhang & Peng (2007) proposed to rewrite (2.9) as the likelihood function for a classical semiparametric AFT model. Using the fact that $\Delta_i = 1$ and $W_i^{(m)} = 1$ if the i-th observation is uncensored, it turns out that $\Delta_i W_i^{(m)} \equiv \Delta_i$, and $\Delta_i \log W_i^{(m)} \equiv 0$. The likelihood function can be rewritten as

$$\prod_{i=1}^{n} \left[W_i^{(m)}\lambda_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}\right]^{\Delta_i} \left[S_\epsilon\{\log(Y_i) - \beta_0 - \boldsymbol{\beta}^t\mathbf{Z}_i\}\right]^{W_i^{(m)}},$$

  where $\lambda_\epsilon = f_\epsilon/S_\epsilon$, which corresponds to the likelihood function of an AFT model with $\log(T_i^*) = \beta_0 + \boldsymbol{\beta}^t\mathbf{Z}_i + \epsilon_i^*$, where the hazard function of ϵ_i^* is $W_i^{(m)}\lambda_\epsilon(\epsilon_i^*)$, and $W_i^{(m)}$ is a constant. A rank estimation method proposed by Wei (1992) for classical semiparametric AFT models is then used to estimate the latency. We refer to Zhang & Peng (2007) for more details.

- Lu (2010) proposed a profile likelihood approach. First, the hazard function in (2.9) is replaced by a piece-wise constant hazard:

$$\lambda(t) = \sum_{j=1}^{J_n} \lambda_j I\left\{t \in [x_{j-1}, x_j)\right\},\ 0 \leq t < M,$$

  where the support $[0, M]$ of $\exp[\log(Y_i) - \boldsymbol{\beta}^t \mathbf{Z}_i]$ is partitioned in J_n intervals of equal length. The likelihood function is first maximised with respect to λ_j, $j = 1, ..., J_n$, given $\boldsymbol{\beta}$. Then, the estimators of the λ_j's are substituted in (2.9). A profile likelihood is obtained. However, this profile likelihood is not smooth, and presents local maxima. As a solution, a kernel-smoothed approximation is proposed, and the latency is estimated from this latter function.

Flexible Semiparametric Models

All the preceding models consider a logistic regression for the incidence. However, as mentioned by Peng (2003a) and proposed by Lam et al. (2005), other types of link functions can be considered. If the logit link is the canonical link function for binary response variables in the generalized linear framework, one can also consider a probit or a complementary log-log link function, among others. These link functions only ask for a slight modification of the likelihood function. The EM algorithm can then be easily implemented to estimate these models. One possible limitation, however, is the lack of flexibility of parametric models. Even if parametric incidences offer some appealing characteristics, such as easy estimation and interpretation, one can question the quality of their fit. In order to widen the flexibility of the incidence of the mixture cure model, some semiparametric modelling approaches have been proposed. Wang et al. (2012) considered a smoothing splines analysis of variance (SS ANOVA) model for both the incidence and the latency. It consists of expressing the two parts of the model as

$$\log \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \zeta_0 + \sum_{j=1}^{p} \zeta_j(x_j) + \sum_{j,k=1}^{p} \zeta_{jk}(x_j, x_k) + ... + \zeta_{1...p}(x_1, ..., x_p),$$

and

$$\lambda(t|\boldsymbol{x}) = \exp\left\{\eta_0 + \sum_{l=1}^{q} \eta_l(z_l) + \sum_{l,m=1}^{q} \eta_{lm}(z_l, z_m) + ... + \eta_{1...q}(z_1, ..., z_q)\right\},$$

where all functions $\zeta(\cdot)$ and $\eta(\cdot)$ appearing in the formulas are unspecified. The SS ANOVA model is then estimated based on a penalised EM algorithm.

Another modelling approach has been proposed by Amico et al. (2018), who considered a single-index structure for the probability of being uncured, of the form

$$p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t \boldsymbol{x}),$$

where $g(\cdot)$ is a totally unspecified and not necessarily monotone link function, alongside a Cox PH model for the latency. They proposed an estimation

method based on the EM algorithm, close to the proposal of Sy & Taylor (2000), but with an additional substep in the M-step added to estimate the unknown link function in the single-index. They considered kernel estimator and proved the identifiability of the model. This model is a part of this thesis and will be fully described in Chapter 3.

In the above papers, the authors tried to relax the common LC mixture cure model by replacing the logistic model by a more flexible model. Another way to make the latter model more flexible is by replacing the Cox PH model by a nonparametric model. This was proposed by Taylor (1995), who considered a fully nonparametric model for the latency that does not depend on any of the covariates, and a logistic regression model for the incidence. He developed an estimation method based on the EM algorithm where the latency is estimated from (2.6) by a Kaplan-Meier-type estimator.

Another paper that considers a nonparametric model for the latency is Patilea & Van Keilegom (2018), but contrary to Taylor (1995), they allow the latency to depend on covariates, and they assume a parametric model for the incidence. So, no assumptions are made on the conditional survival function $S_u(\cdot|\boldsymbol{z})$ of the uncured subjects, except for smoothness and identifiability assumptions. They use a two-step procedure to estimate their model: in the first step, they fix the parameter vector coming from the incidence, and they estimate the survival function $S_u(\cdot|\boldsymbol{z})$ by means of a kernel approach. In the second step, they plug in this estimated function in the likelihood, which they maximise with respect to the parameters of the incidence. Patilea & Van Keilegom (2018) showed the weak consistency and the asymptotic normality of their model parameters, and they compared their estimated model with the logistic/Cox mixture cure model through finite sample simulations, which allowed them to study the sensitivity of the latter model with respect to the validity of the PH assumption.

Another approach is that of Lu & Ying (2004), who assume a semiparametric linear transformation model for the latency of the form

$$H(T^*) = -\boldsymbol{\beta}^t \mathbf{Z} + \epsilon,$$

where H is an unknown monotone increasing function. Depending on the distribution of ϵ, different models will be obtained, and two particular cases are mentioned. When an extreme value distribution is assumed, a Cox PH model is obtained. In contrast, if ϵ follows a standard logistic distribution, the latency follows a proportional odds model. Lu & Ying (2004) proposed an estimation method based on counting processes and martingale theory. They derived estimating equations in order to estimate H, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, and they presented an iterative approach to solve them. The main objective of the paper is to derive the asymptotic normality of the proposed estimator and to obtain consistent variance estimates.

Nonparametric Models

In all the preceding models, at least one part of each of them is modelled semi- or parametrically. A last possibility is to assume a fully nonparametric mixture

cure model. The main contribution on this topic comes from López-Cheda et al. (2017) who considered a nonparametric model for both the incidence and the latency, including covariates. They proposed estimating the two parts of the model based on the Beran (1981) estimator and proceeded as follows. For the incidence, they considered the cure rate estimator developed by Xu & Peng (2014):

$$1 - \hat{p}_h(x) = \prod_{j=1}^{n} \left\{ 1 - \frac{\Delta_{(j)} B_{h(j)}(x)}{\sum_{k=j}^{n} B_{h(k)}(x)} \right\}, \tag{2.10}$$

where $B_{h(j)}(x) = K\{(x - X_{(j)})/h\} / \sum_{l=1}^{n} K\{(x - X_{(l)})/h\}$ are Nadaraya-Watson weights, h is a bandwidth, K is a kernel function, and $X_{(j)}$ and $\Delta_{(j)}$ are the values of the covariate and of the censoring indicator corresponding to the j-th order statistic $Y^*_{(j)}$ (assuming no ties). The intuition behind this estimator is that the cure proportion corresponds to the value at which the survival function levels-off, or equivalently, to the value of the survival function for the last uncensored event time.

For the latency, the idea is to rewrite model (1.5) assuming that $X = Z$, which gives $S_u(t|x) = [S_{pop}(t|x) - \{1 - p(x)\}]/p(x)$, and to use the following estimator:

$$\hat{S}_{u,b}(t|x) = \frac{\hat{S}_{pop,b}(t|x) - (1 - \hat{p}_b(x))}{\hat{p}_b(x)},$$

where $\hat{S}_{pop,b}(t|x)$ is the Beran (1981) estimator of the survival function $S_{pop}(t|x)$ given by $\prod_{j:Y_{(j)} \leq t} [1 - \{\Delta_{(j)} B_{b(j)}(x) / \sum_{l=j}^{n} B_{b(l)}(x)\}]$, and where $B_{b(j)}(x)$ are Nadaraya-Watson weights, with b a bandwidth not necessarily equal to h. López-Cheda et al. (2017) developed the asymptotic theory, as well as a bandwidth selection method based on bootstrap.

The Zero-Tail Constraint and Baseline Conditional Survival Function Tail Estimation

Before ending this section on models and estimation methods for the mixture cure model, one issue still needs to be discussed. When the latency is modelled non- or semiparametrically, little information is available to distinguish cured from uncured individuals among censored subjects. Even if we suppose that $S_u(t) = 0$ when $t \to \infty$, the tail of the conditional survival function $\hat{S}_u(t|\boldsymbol{z})$ may be hard to estimate and we often observe that $\hat{S}_u(t|\boldsymbol{z}) > 0$ when $t > Y^*_{(r)}$, implying identifiability issues. To solve this problem, Taylor (1995) proposes to consider $Y^*_{(r)}$ as a 'cure threshold' τ, and to force $\hat{S}_u(t|\boldsymbol{z})$ to be equal to zero beyond τ by setting, in the M-step of the EM algorithm, $W_i^{(m)}$ equal to 0 when an observation i is such that $\Delta_i = 0$ and $Y_i > \tau$. This proposal is equivalent to considering that such an observation i is cured. To understand the intuition behind this proposal, we have to recall some elements introduced previously. When cured observations are present in survival data, $\lim_{t \to \infty} S_{pop}(t|\boldsymbol{x}, \boldsymbol{z}) > 0$. In practice, this assumption translates into a Kaplan-Meier estimator of the survival function with a large plateau, and levelling off to a value greater than 0 for $t > Y^*_{(r)}$. It corresponds to a situation where a certain number of censored

individuals have a follow-up time greater than the last uncensored event time $Y^*_{(r)}$. If the follow-up is sufficiently long and if the number of observations in the plateau is sufficiently large, it reasonably makes sense to consider these censored individuals as cured, which is what Taylor (1995) proposes. Also known as the *zero-tail constraint*, this constraint has also been applied by Sy & Taylor (2000), Peng & Dear (2000) and Li & Taylor (2002), among others. However, this constraint may overestimate the number of cured observations. This motivated Peng (2003b) to propose another approach which consists of parametrically estimating the tail of the conditional baseline survival function in a logistic/Cox mixture cure model, similarly to what Moeschberger & Klein (1985) did for the classical Kaplan-Meier estimator. He proposed to consider an exponential or a Weibull model for the baseline survival function $S_0(t)$ when $t > Y^*_{(r)}$. Using simulations, he showed that the proposed method reduces bias well compared with the zero-tail constraint.

2.1.3 Assessment of the Model

The literature on mixture cure models is quite extensive, not only regarding model definition and inference, but also about the verification of the model where many different aspects have been investigated. In this section, we detail these aspects, ranging from testing for crucial hypotheses such as the presence of a sufficient follow-up or the presence of a cure fraction, to variable selection.

Testing for Sufficient Follow-Up

As mentioned previously, a sufficient follow-up is an important element in order to consider a cure model. Heuristically, it consists of looking at the plateau of the Kaplan-Meier estimator and insuring that it is long enough. To evaluate this formally, Maller & Zhou (1996) developed a test which consists of testing the null hypothesis $H_0 : \tau_{F_u} \geq \tau_G$ against the alternative hypothesis $H_1 : \tau_{F_u} < \tau_G$. Note that H_1 is exactly the identification condition (2.1), of that we argued is a crucial assumption in most semi- and nonparametric cure models. Intuitively, the main idea of the test is as follows : first note that one can test H_0 by looking at the difference between the largest observation $Y_{(n)}$ and the largest uncensored observation $Y^*_{(r)}$. Indeed, if $Y_{(n)} - Y^*_{(r)}$ is large, the largest censoring time occurs well after the largest uncensored survival time, which is an indication that the follow-up is sufficiently long or that $\tau_{F_u} < \tau_G$. In contrast, if $\tau_{F_u} \geq \tau_G$, then $Y_{(n)} - Y^*_{(r)}$ will be close to zero. Based on a heuristic and informal reasoning, Maller & Zhou (1996) then used this information to propose the following test statistic:

$$q_n = \frac{N_n}{n},$$

where N_n represents the number of uncensored observations in the interval $[2Y^*_{(r)} - Y_{(n)}, Y^*_{(r)}]$. The decision rule is then as follows: H_0 is rejected if q_n exceeds a certain critical value, and the follow-up will be considered as sufficiently long in that case. Since the distribution of q_n is not known, Maller & Zhou

(1996) simulate the critical values when T follows an exponential distribution with mean 1 and C follows a uniform distribution on $[0, b]$, and they tabulated the critical values for several values of n, b and the cure rate $1 - p$. Although it is unlikely that H_0 can be tested in a general setting, the proposed solution to find the critical value is very restrictive, since it only holds for very specific parametric distributions.

Testing for the Presence of a Cure Fraction

Having a sufficient follow-up is a necessary condition to consider a cure model. However, it does not necessarily imply the presence of a cure fraction. To evaluate whether there exists a cure fraction or not, some authors have developed statistical tests. Zhao et al. (2009) proposed a score test in the setting of a logistic/Cox mixture cure model for the hypothesis $H_0 : p = 1$ or equivalently $H_0 : \varphi = 0$ against $H_1 : \varphi > 0$, where $\varphi = (1 - p)/p$ and $0 \leq \varphi < \infty$. They assumed that p does not depend on any of the covariates. The test statistic is given by

$$S_n(\hat{\boldsymbol{\beta}}) = U^t(\hat{\boldsymbol{\beta}})\hat{\Gamma}^{-1}U(\hat{\boldsymbol{\beta}})$$

where $U(\hat{\boldsymbol{\beta}})$ is the score vector of the logistic/Cox mixture cure model evaluated at $(\boldsymbol{\beta}, \varphi) = (\hat{\boldsymbol{\beta}}, 0)$, with $\hat{\boldsymbol{\beta}}$ the estimator of the regression coefficients, and $\hat{\Gamma}$ the Fisher information matrix evaluated at $\hat{\boldsymbol{\beta}}$ and $\varphi = 0$. Note that under H_0, the model reduces to a classical Cox survival model (1.2), and hence it is not necessary to estimate a mixture cure model to perform the test. Asymptotically, $S_n(\hat{\boldsymbol{\beta}})$ converges under H_0 to a mixture of a χ^2_0 and a χ^2_1 distribution with equal probability. It is important to mention that the test is based on a parametric model for the conditional baseline hazard function in the latency.

Hsu et al. (2016) also developed a test for the presence of a cure fraction, but they allow the cure fraction to depend on the covariates. Based on the model

$$p(\boldsymbol{x}) = \frac{\exp(\alpha)\exp(\boldsymbol{\gamma}^t\boldsymbol{x})}{1 + \exp(\alpha)\exp(\boldsymbol{\gamma}^t\boldsymbol{x})}$$

or equivalently $p(\boldsymbol{x}) = [1 + \exp(-\alpha)\exp(-\boldsymbol{\gamma}^t\boldsymbol{x})]^{-1}$, where α is an intercept and $\boldsymbol{\gamma}$ is a vector of slopes, the test is looking for infinite values of the intercept by testing whether $H_0 : \psi^* = 0$ for all $\boldsymbol{\gamma}$ vs. $H_1 : \psi^* > 0$ for some $\boldsymbol{\gamma}$, where $\psi = \exp(-\alpha)$, and ψ^* is the true value of ψ. Note that this is equivalent to testing $H_0 : p(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$ vs. $H_1 : p(\boldsymbol{x}) < 1$ for some $\boldsymbol{x}$. They derive a sup-score test statistic given by

$$T_n = \sup_{\boldsymbol{\gamma} \in \mathcal{B}} S_n(\boldsymbol{\gamma}),$$

where $\mathcal{B}$ is the ensemble of values that $\boldsymbol{\gamma}$ can take, and $S_n(\boldsymbol{\gamma})$ is a certain score test statistic obtained under H_0. Hsu et al. (2016) showed that under the null hypothesis and for fixed $\boldsymbol{\gamma}$, the score $S_n(\boldsymbol{\gamma})$ converges in distribution to the mixture $(\chi^2_0 + \chi^2_1)/2$. However, the asymptotic distribution of $\sup_{\boldsymbol{\gamma} \in \mathcal{B}} S_n(\boldsymbol{\gamma})$ is more complicated to obtain, and they propose a simple resampling technique to approximate this distribution under the null hypothesis. Note that if $\mathcal{B} = \{0\}$,

the test reduces to a test for a constant cure fraction. The test is derived assuming a logistic regression model for the incidence. However, the methodology can be implemented for any increasing, differentiable, and invertible link function. There are two main limitations: it assumes continuous covariates and it considers a parametric form for the baseline conditional hazard (a Weibull or a log-logistic model).

Model Diagnostics

Beside testing for sufficient follow-up and for the presence of a cure fraction, the literature on the mixture cure model also deals with model diagnostics. This topic has been first investigated by Wileyto et al. (2013) who proposed to derive Schoenfeld residuals for parametric mixture cure models. Schoenfeld residuals are used to evaluate the departure from the proportional hazards assumption in a classical survival analysis context, and they are used here to evaluate the fit of the model. Indeed, Wileyto et al. (2013) proposed replacing the weight $\exp(\boldsymbol{\beta}^t \boldsymbol{z})$ in the expected values of covariates by the hazard function of the entire population $\lambda_{pop}(t|\boldsymbol{x}, \boldsymbol{z})$. Since $\lambda_{pop}(t|\boldsymbol{x}, \boldsymbol{z})$ does not verify the proportional hazards assumption, these Schoenfeld residuals can not be used to check for this property.

In order to complete the toolbox, other diagnosis tools were considered by Peng & Taylor (2017) who developed a series of residual-based model diagnostic tools for the overall mixture cure model and for the latency, including three types of model checking:

- To check for the functional form of covariates and to diagnose the presence of outliers, they developed martingale residuals for the overall model and modified martingale residuals for the latency.
- To evaluate the fit of the model, they developed Cox-Snell and modified Cox-Snell residuals for the overall model and for the latency, respectively. The Cox-Snell residuals for the mixture cure model are sampled from a mixture type distribution whereas a unit exponential distribution is used in the classical case. As explained by Peng & Taylor (2017), a unit exponential distribution can still be used in practice, and this has no impact on the analysis. Regarding the modified Cox-Snell residuals, a Cramér-von Mises criterion is proposed to measure the distance between the estimated distribution and the unit exponential distribution.
- To evaluate the departure from the proportional hazards assumption for the latency, they proposed a score process from which they developed a Kolmogorov-type supremum test.

All these diagnostic tools can be used for mixture cure models with parametric and semiparametric latency, but some drawbacks have to be mentioned. First, the martingale residuals for the overall model are bounded from below, contrary to what happens in classical survival analysis, which may limit their application. They are also insensitive to the covariate effects in the incidence. Second, the Cox-Snell residuals have some difficulty to detect a misspecification

in the incidence modelling, whereas they perform well for the latency. Finally, for detection of outliers, Peng & Taylor (2017) also mentioned that the modified martingale residuals are preferred to the martingale residuals because they are not bounded from below. However they are not efficient at detecting outliers that are too large. As an alternative, they proposed to consider deviance residuals.

Testing for the Form of the Incidence

The Cox-Snell and the martingale residuals proposed by Peng & Taylor (2017) have some difficulties evaluating the fit of the incidence. However, another proposal has been made for this purpose. Müller & Van Keilegom (2018) developed a test for the parametric form of the incidence. Their test includes as special cases the case of a logistic model and the case where the cure rate does not depend on any covariates. The test statistic is a weighted L_2-distance between a nonparametric kernel estimator of the cure rate (obtained from Xu & Peng (2014) and given by equation 2.10) and a parametric estimator obtained under the null hypothesis. Although they proved the limiting distribution of their test statistic, they used a bootstrap procedure to calibrate the test, since the limiting distribution is only a reasonable approximation of the distribution of the test statistic for very large samples. The test can be used as a preliminary step before deciding to go for example for a single-index model (see Amico et al. (2018) and Chapter 3) or a completely nonparametric model (see Xu & Peng (2014)) for the incidence.

Variable Selection

Finally, some papers focused on the selection of relevant covariates. Liu et al. (2012) first developed a variable selection methodology for both parts of the logistic/Cox mixture cure model based on a penalised likelihood approach. Because of the interesting feature of the complete-data likelihood for such model, they proposed a penalised EM algorithm where two penalty terms are considered, one for $\boldsymbol{\gamma}$, and one for $\boldsymbol{\beta}$, which is equivalent to maximising a penalised logistic model and a penalized Cox model separately. They proposed to use Smoothly Clipped Absolute Deviate (SCAD) penalties developed by Fan & Li (2001). Note that this approach is only possible because a logistic/Cox mixture cure model is assumed and because the EM algorithm is considered to estimate the model. However, when a parametric form is assumed for the latency, the Liu et al. (2012) approach is not natural because the complete-data likelihood is not used. For mixture cure models with a parametric latency, Scolas et al. (2016) proposed a method based on a penalised likelihood in the context of interval-censored cure data. Adaptive LASSO penalties are assumed, one for each part of the model, and the penalised likelihood is derived from (2.3).

Dirick et al. (2015) developed an AIC to select the covariates in the incidence and in the latency. They proposed to construct the criterion for a logistic/Cox mixture cure model from the complete-data likelihood used in the EM algorithm. Two different approaches are considered to compute the expectation of the complete-data likelihood.

Finally, a third approach to do variable selection in a logistic/Cox mixture cure model has been proposed by Claeskens & Van Keilegom (2018). They considered a procedure based on the focused information criterion (FIC). This criterion selects the variables in the model in such a way that the resulting estimated model is the best possible model with respect to the estimation of a certain focus parameter. Here, 'best possible model' should be understood in the sense that the mean squared error of the estimated focus parameter is the smallest among all candidate models. The focus parameter can be any parameter depending on the latency and/or the incidence, for example, the cure rate, the regression parameters, the conditional or unconditional survival or hazard function, etc. Claeskens & Van Keilegom (2018) developed asymptotic theory for their proposed procedure, and they showed via simulations how the method works in practice.

2.1.4 Data Analysis

To illustrate the practical use of the mixture cure model, we estimate a LC mixture cure model based on the Wang et al. (2005) dataset, assuming a Breslow-type estimator for the conditional survival function as proposed by Sy & Taylor (2000), and including all covariates in both parts of the model. The model is fitted with the R-package `smcure` from Cai et al. (2012).

Table 2.1: *Parameter estimates from the mixture cure model together with their corresponding standard errors and p-values.*

Incidence	Estimate	Std.Err.	Z value	p-value
(Intercept)	1.1110	0.8850	1.2555	0.2093
Age	-0.0382	0.0173	-2.2112	0.0270
ER status [ER+ vs. ER-]	0.1824	0.2704	0.6746	0.4999
Tumour size	-0.0784	0.2054	-0.3814	0.7029
Menopausal [post vs. pre]	0.7721	0.4445	1.7371	0.0824
Latency	Estimate	Std.Error	Z value	p-value
Age	-0.0127	0.0179	-0.7059	0.4802
ER status [ER+ vs. ER-]	-1.0365	0.2317	-4.4739	<0.0001
Tumour size	0.5203	0.2184	2.3820	0.0172
Menopausal [post vs. pre]	0.0778	0.3970	0.1960	0.8446

As can be seen from Table 2.1, two different groups of estimates are obtained, one for each part of the model. The table also shows the p-values for the Wald test of the parameters computed by bootstrap based on 250 bootstrap resamples. If we first focus on the incidence, age and the menopausal status have a significant impact on the probability of being uncured at a .05 and .10 level of significance respectively. To interpret these effects, we proceed as for a classical logistic regression model: $\exp(-0.0382) = 0.9625$, and $1/0.9625 = 1.0390$,

meaning that an additional year of age (at diagnosis) increases the odds of being cured by 4%. Regarding the menopausal status, the odds of being uncured for a post-menopausal woman is $\exp(0.7721) = 2.1643$ times higher than for a pre-menopausal woman. For the latency, the estrogen receptor status and the tumour size have significant effects on the time to distant metastasis for uncured patients at a .05 level of significance. A Cox PH model is assumed for this part. The interpretation of the effect of the estrogen receptor status for example is as follows: among patients who experience metastasis, the hazard for ER+ patients is $1/\exp(-1.0365) = 2.82$ times smaller than the hazard for ER− patients. For the tumour size, the table shows that patients with a bigger tumour have a larger instantaneous risk than patients with a smaller tumour. As can be seen, covariates have different effects in the two parts of the model. This situation is representative of an interesting feature of mixture cure models. It is possible to distinguish long-term and short-term effects of covariates, a feature that is not present in classical survival analysis. Indeed, the incidence models the long-term effect of covariates on the cure status which is something permanent, whereas the latency focuses on the short-term effect that only concerns uncured observations. More details about this point can be found in Sy & Taylor (2000).

2.2 Promotion Time Cure Models

The promotion time cure model is the second main class of cure models. In this section, we will first give in Section 2.2.1 some details about the definition of the model and its interpretation. In Section 2.2.2, we will detail the different modelling approaches that have been proposed and present the corresponding estimation methods both in frequentist and Bayesian settings. Measurement errors is an important issue in medical studies and so also in the context of cure data. In Section 2.2.3, we will present some works that have been done on this topic. Finally, we will end this section with an application of the promotion time cure model to the breast cancer data used previously for the mixture cure model.

2.2.1 Model Justification and its Interpretation

The promotion time cure model defined in Chapter 1 offers a different approach to model survival data with a cure fraction. As a recall, the mathematical definition of this model comes from the assessment that, in the presence of a cure fraction, the cumulative hazard is bounded from above such that $\lim_{t\to\infty} \Lambda_{pop}(t) = \theta < \infty$, with $\theta > 0$. To take into account for such feature, the cumulative hazard function can be written such as $\Lambda_{pop}(t) = \theta F(t)$, where $F(t)$ is a (proper) distribution function such that $\lim_{t\to\infty} F(t) = 1$. It follows that the survival function is given by

$$S_{pop}(t) = \exp\{-\theta F(t)\}. \tag{2.11}$$

An interesting feature of this model is its biological interpretation. Indeed, in the particular context of cancer, the mathematical form described previously

can also be obtained by assuming that the survival time is the result of a latent process generating cancer tumour(s). Originally proposed by Yakovlev et al. (1993) for modelling tumour latency, it has been introduced for the promotion time cure model by Chen et al. (1999). The main idea is to assume that after a first treatment for cancer, a number $N \geq 0$ of carcinogenic cells can stay active in the organism of an individual, and that it will take a certain (latent) time $\tilde{T}_k$, for each cell $k = 1, ..., N$, to become an active tumour. For individuals for whom $N \geq 1$, that is, for uncured observations, the survival time T is defined as $\min\{\tilde{T}_k, k = 1, ..., N\}$. For cured individuals, no carcinogenic cells are still active, that is, $N = 0$, inducing that $T = \infty$. By assuming that N follows a Poisson distribution with parameter $\theta > 0$, that the $\tilde{T}_k$ are i.i.d. with distribution function $F(\cdot)$ and that they are independent of N, we can derive the survival function for T in the following way:

$$\begin{aligned}
P(T > t) &= P(N = 0) + P(\tilde{T}_1 > t, ..., \tilde{T}_N > t, N \geq 1) \\
&= P(N = 0) + \sum_{k=1}^{\infty} P(\tilde{T}_1 > t, ..., \tilde{T}_k > t) \times P(N = k) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} \{S(t)\}^k \ \exp(-\theta)\frac{\theta^k}{k!} \\
&= \exp(-\theta)\Big[\sum_{k=0}^{\infty} \{S(t)\}^k \ \frac{\theta^k}{k!}\Big] \\
&= \exp\{-\theta + \theta S(t)\} \\
&= \exp\{-\theta F(t)\}, \qquad (2.12)
\end{aligned}$$

where $S(t) = 1 - F(t)$. The survival function given in Equation (2.12) corresponds to the survival function (2.11) from the promotion time cure model.

The parameter θ represents the mean number of carcinogenic cells. In the presence of covariates, which are mainly introduced through θ, this parameter has a double interpretation. First, when θ is large, the mean number of carcinogenic cells is large and the probability of being cured is small. Second, a larger value for θ is also representative of a lower survival probability because a larger number of carcinogenic cells induces a smaller activation time. As it can be seen, θ contains two types of effects, on the cure probability and on the survival which can not be separate. Such an interpretation can also be drawn from a mathematical point of view. As proposed by Chen et al. (1999), quantities for each type of observations can be derived from the model in the manner of the mixture cure model. For cured observations, the associated quantity is given by $\lim_{t\to\infty} S_{pop}(t) = \exp(-\theta)$. For uncured observations, Chen et al. (1999) considered the biological development of the model, and they proposed to consider the survival function for uncured observations, which corresponds to observations with at least one carcinogenic cell, that is,

$$P(T > t | N \geq 1) = \frac{\exp\{-\theta F(t)\} - \exp\{-\theta\}}{1 - \exp\{-\theta\}}. \qquad (2.13)$$

Note that there also exists a vast literature on tumour latency modelling

from which the biological interpretation of the promotion time cure model is derived. Some references can be found in Tsodikov et al. (2003).

2.2.2 Modelling Approaches and Inference

Modelling Approaches

The literature on the promotion time cure model contains two main types of modelling approaches, depending on how the covariates are introduced.

The first group of models consists of introducing covariates only through θ as defined in Chapter 1. Proposed by Tsodikov (1998a), the survival function is given by (1.6) where $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$. Regarding $F(t)$, Tsodikov (1998a) and Tsodikov (2001) proposed to let the distribution function totally unspecified. Some other forms have been proposed, such as a Weibull or a gamma distribution (Chen et al. (1999)), or a semiparametric version $F(\cdot|\eta)$, for some parameter η introduced by Ibrahim et al. (2001).

An important characteristic of model (1.6) is its proportional hazards property. In fact, the hazard function of the model is given by $\lambda_{pop}(t|\boldsymbol{x}) = \theta(\boldsymbol{x})f(t)$. By considering two subjects i and j, with two different vector of covariates, $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, it follows that

$$\frac{\lambda_{pop}(t|\boldsymbol{x}_i)}{\lambda_{pop}(t|\boldsymbol{x}_j)} = \frac{\theta(\boldsymbol{x}_i)}{\theta(\boldsymbol{x}_j)}.$$

Hence, the ratio of the hazard functions is constant over the time.

As in the case of the mixture cure model, the identifiability of the promotion time cure model is an important issue that needs attention, before we can talk about the estimation of the model. Zeng et al. (2006) showed the strong identifiability of model (1.6) when $\theta(\boldsymbol{x}) = \eta(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$ for a strictly increasing function η like, for example, the exponential function, and when $F(\cdot)$ is left unspecified. With strong identifiability we mean that there is a unique vector $\boldsymbol{\gamma}$ and a unique function $F(\cdot)$ that maximise the expected log-likelihood under the model. Portier et al. (2017) improved the result of Zeng et al. (2006) by allowing the censoring time to be finite, and by allowing the covariates to have non-compact support.

The second group of models have been proposed by Tsodikov (2002), where covariates are introduced both in θ and $F(\cdot)$. The survival function is given by

$$S_{pop}(t|\boldsymbol{x}, \boldsymbol{z}) = \exp\left\{-\theta(\boldsymbol{x})F(t|\boldsymbol{z})\right\}, \tag{2.14}$$

and the form $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$ is usually assumed. Regarding $F(t|\boldsymbol{z})$, Tsodikov (2002) and Bremhorst & Lambert (2016) proposed the form $F(t|\boldsymbol{z}) = 1 - S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$, where $\mathbf{Z}$ does not contain an intercept. This model corresponds to a Cox PH model. Another form has been proposed by Tsodikov (2002), namely $F(t|\boldsymbol{z}) = \{\exp(\boldsymbol{\beta}^t\boldsymbol{z})F_0(t)\}/\{1 - \exp(\boldsymbol{\beta}^t\boldsymbol{z})F_0(t)\}$, which corresponds to a proportional odds model. Contrary to model (1.6), this model does not respect the proportional hazards assumption because of the presence of covariates in $F(\cdot)$. In such a case, the hazard function is given by $\lambda_{pop}(t|\boldsymbol{x}, \boldsymbol{z}) = \theta(\boldsymbol{x})f(t|\boldsymbol{z})$, where $f(t|\boldsymbol{z}) = (d/dt)F(t|\boldsymbol{z})$.

Bremhorst & Lambert (2016) have proved the weak identifiability when $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$ and $F(t|\boldsymbol{z}) = 1 - S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$. With weak identifiability we mean that if $\exp\big[-\exp(\gamma_{01} + \boldsymbol{\gamma}_1^t\boldsymbol{x})\{1 - S_{01}(t)^{\exp(\boldsymbol{\beta}_1^t\boldsymbol{z})}\}\big] = \exp\big[-\exp(\gamma_{02} + \boldsymbol{\gamma}_2^t\boldsymbol{x})\{1 - S_{02}(t)^{\exp(\boldsymbol{\beta}_2^t\boldsymbol{z})}\}\big]$ for all $t \in [0, \infty]$ and all $\boldsymbol{x}$ and $\boldsymbol{z}$ in the support of $\boldsymbol{x}$ and $\boldsymbol{z}$, then necessarily $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $S_{01} \equiv S_{02}$.

Frequentist Estimation Methods

Estimation methods for the promotion time cure model have first been developed in a frequentist setting. The likelihood function given by

$$\prod_{i=1}^{n} \Big[\theta(\mathbf{X}_i) f(Y_i) \exp\big\{-\theta(\mathbf{X}_i)F(Y_i)\big\}\Big]^{\Delta_i} \Big[\exp\big\{-\theta(\mathbf{X}_i)F(Y_i)\big\}\Big]^{1-\Delta_i}, \quad (2.15)$$

is considered for the two versions of the model described previously, i.e. $F(\cdot)$ may or may not depend on covariates, depending on the model we consider, and with $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$.

Tsodikov (1998a) proposed a profile likelihood approach in order to estimate the semiparametric version of model (1.6) where $F(\cdot)$ is totally unspecified. In a first step, the distribution function $F(\cdot)$ is estimated by a Non Parametric Maximum Likelihood Estimator (NPMLE). It consists in maximising the likelihood function (2.15) with respect to $F(\cdot)$, where $F(\cdot)$ is replaced by a step function taking values at the failure times. In order to identify the model, he proposed to perform this maximisation under the constraint that $F_r = 1$, where $F_r = \sum_{j=1}^{r} \Delta F_j$ and ΔF_j represents the jump size of $F(\cdot)$ at time $Y^*_{(j)}$. This constrained maximisation is similar to imposing the zero-tail constraint that we described in Section 2.1.2 in the mixture cure model. He derived score equations for the step sizes ΔF_j, $j = 1, ..., r$, and he developed his own maximisation algorithm in order to avoid instability problems encountered with the Newton-Raphson technique when the number of parameters is very large. A profile likelihood for $\boldsymbol{\gamma}$ is obtained by substituting the NPMLE for $F(\cdot)$ in the likelihood function. Alongside the estimation method, Tsodikov (1998b) has demonstrated the asymptotic relative efficiency of the semiparametric estimator of $\theta(\boldsymbol{x})$ in comparison with a parametric one. Rigorous asymptotic theory and efficiency results for the parametric component $\theta(\boldsymbol{x})$ and for the nonparametric component $F(\cdot)$ have been developed by Portier et al. (2017). They also developed a weighted bootstrap procedure that allows for a consistent approximation of the asymptotic law of the estimators.

A frequentist approach has also been proposed for model (2.14) for $F(t|\boldsymbol{z}) = 1 - S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$, where $S_0(t) = 1 - F_0(t)$ and for $F(t|\boldsymbol{z}) = \{\exp(\boldsymbol{\beta}^t\boldsymbol{z})F_0(t)\}/\{1 - \exp(\boldsymbol{\beta}^t\boldsymbol{z})F_0(t)\}$ by Tsodikov (2002). As for model (1.6) a profile likelihood method is considered where the likelihood function (2.15), with F_0 replaced by a step-function, is first maximised with respect to F_0 under the constraint that $F_{0r} = 1$. A NPMLE $\hat{F}_0$ for F_0 is obtained, and by substituting $\hat{F}_0$ in (2.15), a profile likelihood is obtained depending of the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Tsodikov (2002) proposed two methods to solve the score equations for F_0: an alternative approach to Newton-Raphson algorithm and a Quasi-EM algorithm approach.

Note that the Quasi-EM algorithm requires untied data. When tied data are present, the first proposal is preferred.

Bayesian Estimation Methods

Bayesian inference has been introduced by Chen et al. (1999). Let us denote by $F(\cdot|\eta)$ the distribution function F depending on some vector of parameters η. Chen et al. (1999) focused on the parametric version of model (1.6) where a Weibull distribution is considered for $F(\cdot|\eta)$, with $\eta = (\rho, \lambda)^t$. They proposed classes of both non-informative and genuine priors (based on historical data) for $(\boldsymbol{\gamma}, \eta)$, and they discussed some of their theoretical properties. The posterior distribution is given by

$$p(\boldsymbol{\gamma}, \eta|\mathbf{D}_{obs}) \propto \mathcal{L}(\boldsymbol{\gamma}, \eta)\pi(\boldsymbol{\gamma}, \eta),$$

where $\pi(\cdot)$ represents the joint prior distribution, $\mathbf{D}_{obs} = (\mathbf{Y}, \Delta, \mathbf{X})$ are the observed data, and $\mathcal{L}(\boldsymbol{\gamma}, \eta)$ is given by (2.15). When historical data (obtained from a previous study) are available, genuine priors are defined as the joint posterior distribution from historical data:

$$\pi(\boldsymbol{\gamma}, \eta, \alpha_0|\mathbf{D}_{0,obs}) \propto \{\mathcal{L}(\boldsymbol{\gamma}, \eta|\mathbf{D}_0)\}^{\alpha_0} \pi_0(\boldsymbol{\gamma}, \eta)\pi_0(\alpha_0),$$

where $\mathbf{D}_{0,obs} = (\mathbf{Y}_0, \Delta, \mathbf{X}_0)$ is the vector of the observed historical data, $\mathcal{L}(\boldsymbol{\gamma}, \eta|\mathbf{D}_{0,obs})$ is the likelihood function (2.15) from these historical data, $\pi_0(\boldsymbol{\gamma}, \eta)$ represents the joint prior distribution considered for $(\boldsymbol{\gamma}, \eta)$ from historical data, and α_0 is a parameter taking values between 0 and 1 which controls the influence of the historical data on the current data.

Ibrahim et al. (2001) considered the semiparametric version of (1.6) and proposed to consider a piecewise constant hazard model for $F(\cdot|\lambda)$, where $\lambda = (\lambda_1, ..., \lambda_J)^t$ represents the vector of constant hazards associated with the J partitions of the time axis. When a semiparametric model is considered for F, nothing guarantees that $\hat{F}(Y^*_{(r)}) = 1$ as already explained for the mixture cure model (see Section 2.1.2). To overcome this difficulty, Ibrahim et al. (2001) proposed to introduce a smoothing parameter in the prior distribution of λ_j, $j = 1, ..., J$, in order to control the degree of parametricity of the right tail of the survival function. In terms of inference, as in Chen et al. (1999), Ibrahim et al. (2001) discussed non-informative and informative priors as well as some of their properties. The posterior distributions are constructed assuming the same procedure as above. An extension of this method has been proposed by Kim et al. (2007). Beside the consideration of the degree of parametricity for the right tail, they also proposed to take into account the correlation between $\log(\lambda_{j-1})$ and $\log(\lambda_j)$. As they explained, this latter proposal improves the right tail estimation because information in that part of the survival function can be borrowed from neighbouring $\log(\lambda_j)$'s. They proposed to introduce a smoothing and a correlation parameter in the prior distribution of $\log(\lambda_j)$ by considering a martingale-type process prior. Additionally, they allow J to be random. Prior distributions are discussed as well as some of their properties. Because the dimensions of the posterior distribution can vary with random J,

a reversible jump Metropolis-Hasting algorithm is proposed to sample from the posterior distribution (we refer the reader to the article for more details about this latter point).

For model (2.14), Bremhorst & Lambert (2016) proposed a flexible semi-parametric approach when $F(t|\boldsymbol{z}) = 1 - S_0(t)^{\exp(\boldsymbol{\beta}^t \boldsymbol{z})}$ where the logarithm of the baseline hazard function $\lambda_0(t) = -(d/dt)\log\{1 - F_0(t)\}$ is written as a linear combination of cubic B-splines. Bayesian inference is considered where P-splines are assumed to estimate $\log\{\lambda_0(t)\}$. P-splines consist in taking a large number of B-splines and adding a penalty term as proposed by Eilers & Marx (1996). In a Bayesian setting, the penalty term is taken into account through the specification of the prior distributions as detailed by Lang & Brezger (2004). Following their proposal, Bremhorst & Lambert (2016) proposed prior distributions for the parameters and they gave an MCMC algorithm to sample from the joint posterior distribution. For an identifiability reason, they also proposed to set the last spline parameter to a large value in order to guarantee that the baseline survival function is proper.

Some Extensions

Under the biological perspective of the model, we assume that the latent times $\{\tilde{T}_1, ..., \tilde{T}_N\}$ are i.i.d random variables. However, this assumption could be unrealistic in certain situations since they concern the same individual. Zeng et al. (2006) introduced a subject-specific frailty term ξ_i, in order to relax this assumption and they obtained a more general class of cure models given by

$$S_{pop}(t|\mathbf{X}_i) = E_{\xi_i}\left[\exp\{-\theta(\mathbf{X}_i)F(t)\xi_i\}\right],\ i = 1, ..., n. \tag{2.16}$$

Depending on the distribution for ξ_i, different models are obtained. They proposed to consider a gamma distribution with mean 1 and obtained the model

$$S_{pop}(t|\boldsymbol{x}) = G_\eta\{\theta(\boldsymbol{x})F(t)\},$$

where $G_\eta(u) = (1 + \eta u)^{1/\eta}$ when the transformation parameter η is positive, and $G_\eta(u) = \exp(-u)$ when η equals zero. They also mentioned that other distributions and transformations can be assumed. In terms of modelling, Zeng et al. (2006) assumed that $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$, and they let $F(\cdot)$ totally unspecified. They proposed a profile likelihood approach to estimate the model. As in Tsodikov (1998a), they proposed a NPMLE for $F(\cdot)$ obtained by maximising the likelihood function under the constraint that $F_r = 1$. Score equations are obtained by making use of the Lagrange multiplier. The parameter $\boldsymbol{\gamma}$ is estimated from the profile likelihood obtained by substituting $\hat{F}(\cdot)$ in the likelihood function.

A second extension has been proposed by Portier et al. (2018), who considered a promotion time cure model of the form

$$S(t|\boldsymbol{x}) = \exp\{-g(\boldsymbol{\gamma}, \boldsymbol{x})\,\theta\,F(t)\}, \tag{2.17}$$

where g is a given function (not necessarily monotone) depending on a parameter vector $\boldsymbol{\gamma}$, $\theta > 0$ and $F(\cdot)$ is an unspecified proper distribution function.

When $g(\boldsymbol{\gamma}, \boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})$, then model (2.17) reduces to the promotion time cure model (1.6). For all other g functions, the two models are different. In their paper, Portier et al. (2018) showed the identifiability of their model, they developed the nonparametric maximum likelihood estimator of the model parameters, they showed the asymptotics for their estimators with closed-form formulas of the variance of the limiting Gaussian distributions, and they considered a likelihood-based methodology to select an appropriate g function among a family of proposals.

Other extensions based on the biological interpretation of the promotion time cure model have been proposed in the literature. However, because they also embed the mixture cure model as a special case, they will be detailed in Section 2.3.

2.2.3 Measurement Errors

In medical studies it often happens that some variables in the model are measured with error. For instance, the error can be caused by imprecise medical instruments, like for measuring blood pressure, weight, cholesterol level, etc. In economic studies variables like welfare or income can often not be measured in a precise way, in which case one has to work with approximate measures that might contain some error. Ignoring this measurement error can lead to wrong conclusions, since the presence of measurement error leads to biased estimators (see for example Carroll et al. (2006)).

In the context of the promotion time cure model, several authors have considered the problem of estimating the model when one or several covariates in the model are measured with error. The model that is considered is the classical additive measurement error model of the form

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{2.18}$$

where $\mathbf{W}$ is the vector of observed covariates and $\mathbf{U}$ is the vector of measurement errors. We further assume that $\mathbf{U} \sim N_p(0, V)$, where V is known, and $\mathbf{U}$ is independent of $\mathbf{X}$. If some covariates are not subject to measurement error, then the corresponding elements of V are set to 0. It is also assumed that (T, C) and $\mathbf{W}$ are independent given $\mathbf{X}$.

The first paper that estimated the promotion time cure model (1.6) in the presence of measurement error is Mizoi et al. (2007). The authors considered the case where only one covariate is measured with error (say X_1), and they assumed that the model is fully parametric, with $F(\cdot)$ equal to the distribution of a Weibull random variable. They used a corrected score approach to take the measurement error into account, which consists in replacing in the log-likelihood the unobserved covariate X_1 by a surrogate depending on W_1 and the (known) variance of U_1. The form of the surrogate depends on the assumed normality of U_1, and hence the method cannot be extended in an obvious way to the case where other error distributions are assumed.

Ma & Yin (2008) extended the above paper to the case where the distribution F is unknown and possibly more than one covariate is subject to

measurement error. They also use a corrected score approach and prove the asymptotic unbiasedness and the asymptotic normality of their estimators.

A third contribution comes from Bertrand, Legrand, Carroll, de Meester & Van Keilegom (2017), who assume the same model as Ma & Yin (2008), but they use a different approach to estimate the model parameters. Their method is based on the so-called SIMEX (simulation-extrapolation) approach that has been proposed by Cook & Stefanski (1994). The SIMEX method consists of two steps. In the first step increasing levels of measurement error variance are considered, and at each level a large number of datasets is generated. The idea is then to estimate at each level the vector of regression coefficients ignoring the measurement error. In the second step these estimators corresponding to the different levels of error are extrapolated to the situation where the covariates are observed without error. An important advantage of this approach is that the distribution of the error term $\mathbf{U}$ can by anything, as long as the covariates observed with error are continuous. So it does not restrict attention to the normal case. Bertrand, Legrand, Carroll, de Meester & Van Keilegom (2017) established the asymptotic unbiasedness and the asymptotic normality of their estimators, under the assumption that the extrapolation function is correct, and they compared the finite sample performance of their method with that of the corrected score approach of Ma & Yin (2008) through a small simulation study.

Finally, Bertrand, Legrand, Léonard & Van Keilegom (2017) justified the need to take the measurement error into account via a theoretical study of the bias of the naive estimator that ignores the measurement error. They also performed an extensive simulation study investigating the robustness of both the corrected score and the SIMEX approach with respect to the model assumptions.

2.2.4 Data Analysis

To illustrate the implementation of the promotion time cure model, we fit the model on the Wang et al. (2005) data introduced previously. We consider the semiparametric model proposed by Tsodikov (1998a) where covariates are introduced only through θ assuming the form $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})$. We fit the model using the estimation method proposed by Tsodikov (1998a). Parameter estimates are given in Table 2.2, together with their standard errors and their p-values.

Contrary to the mixture cure model, we only have one set of parameters which influences both the cure probability and the survival as explained in Section 2.2.1. The interpretation of the parameter estimates is as follows. The quantity $\exp(\hat{\gamma_0} + \hat{\boldsymbol{\gamma}}^t \boldsymbol{x})$ represents the estimated mean number of carcinogenic cells for a patient with covariates $\boldsymbol{x}$. It then encompasses two levels of interpretation :

- First, a larger value for $\exp(\hat{\gamma_0} + \hat{\boldsymbol{\gamma}}^t \boldsymbol{x})$ is representative of a higher probability of not being cured. Indeed, the more carcinogenic cells are present, the more the observation has a chance to experience the event.

Table 2.2: *Parameter estimates from the promotion time cure model together with their corresponding standard errors and p-values.*

	Estimate	Std.Err.	Z value	p-value	exp(Est.)
(Intercept)	0.6218	0.6926	0.8977	0.3693	1.8622
Age	-0.0326	0.0143	-2.2813	0.0225	0.9679
ER status [ER+ vs. ER−]	0.0285	0.2260	0.1260	0.8997	1.0289
tumour size	0.0056	0.1659	0.0337	0.9731	1.0056
Menopausal [post vs. pre]	0.6382	0.3503	1.8220	0.0684	1.8931

- Second, a larger value for $\exp(\hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}^t \boldsymbol{x})$ is also associated with an earlier event because more carcinogenic cells means more chance of having a small activation time.

As can be seen, age has a significant impact on the survival probability at a 0.05 level of significance, and menopausal status at a 0.10 level of significance. This means that postmenopausal women have a higher risk of being uncured and they experience the event earlier in comparison with premenopausal women. For age, the effect is reverse. Older women have a higher probability of being cured and they have a lower instantaneous risk of experiencing a relapse than younger women. Those results are different from what we obtained with the mixture cure model where age and menopausal status have a significant impact only on the incidence, and tumour size and estrogen receptor status influence significantly the latency. In the case of the promotion time cure model, it is not possible to distinguish and quantify the covariate effects related to the cure probability and those related to the survival of the uncured observations. We only have one global effect.

2.3 Unifying Models

The mixture cure model and the promotion time cure model represent two different modelling approaches for survival data with a cure fraction. Several differences have been underlined in the literature, but the two models are also related. Indeed, there exists a mathematical link between them and they are equivalent in some situations. Moreover, shortly after the introduction of the promotion time cure model, the literature on cure models has been broadened by unifying approaches embedding both the mixture cure model and the promotion time cure model. In this section, we will first detail in Subsection 2.3.1 the differences between the mixture cure model and the promotion time cure model but also present their mathematical relationship. In Subsection 2.3.2, we will review the unifying models that have been proposed in the literature and their estimation methods. The section will be closed by a Subsection 2.3.3 discussing model selection.

2.3.1 Dissimilarities and Relationship Between the Mixture Cure Model and the Promotion Time Cure Model

Chen et al. (1999) distinguished three main differences between the mixture cure model and the promotion time cure model. First, in the presence of covariates, the mixture cure model does not respect the proportional hazards property. On the contrary, when we assume that only θ is a function of covariates, the promotion time cure model does respect this property as already mentioned in Section 2.2.2. Second, Chen et al. (1999) argued that the promotion time cure model can be interpreted biologically in the context of cancer studies which is not the case for the mixture cure model. This point of view is however disputed by Peng & Xu (2012), who explained that by assuming a Bernoulli distribution for N with parameter p, a mixture cure model is obtained if the biological development described in Section 2.2 is followed. Note however that this biological interpretation is an oversimplification of tumour kinetics. Third, the promotion time cure model has been developed both in a frequentist and and in a Bayesian setting contrarily to the mixture cure model.[1] As explained by Chen et al. (1999), the lack of such methods for the latter model is due to the necessity of taking proper priors for $\boldsymbol{\gamma}$, both for the informative and for the non-informative case, in order to obtain proper posterior distributions, contrarily to the promotion time cure model.

All these elements favour the promotion time cure model. But another important difference, not mentioned by Chen et al. (1999) and that is in favour of the mixture cure model, is in the interpretation of covariate effects. Indeed, the mixture cure model distinguishes the effect of covariates on the probability of being uncured from the effect of covariates on the survival function of the uncured observations. It is then possible to consider different covariates in the two parts of the model and to evaluate the effect of the same covariate(s) on the two parts. For the promotion time cure model, the question is more delicate because, as detailed in Section 2.2.2, θ influences both the cure probability and the conditional survival function (2.13). Therefore, the same parameters represent both the long-term (on the cure probability) and the short-term (on the conditional survival function) effects of the covariates.

Beside these differences, the two models are mathematically related as explained by Chen et al. (1999). Indeed, given that the cure proportion $\exp(-\theta)$ in the promotion time cure model is equivalent to $1-p$ in the mixture cure model, and considering the conditional survival function given by Equation (2.13), the promotion time cure model can be rewritten as (we omit covariates for simplicity)

$$S_{pop}(t) = \exp(-\theta) + \{1-\exp(-\theta)\}\frac{\exp\{-\theta F(t)\} - \exp(-\theta)}{1-\exp(-\theta)},$$

that is, as a mixture cure model. Furthermore, as discussed by Peng & Xu (2012), the two models are equivalent in some situations. First, when no co-

[1] Note that since the article by Chen et al. (1999), some mixture cure models have been proposed in a Bayesian setting, for example, by Yu & Tiwari (2012).

variates are considered at all, and $S_u(\cdot)$ and $F(\cdot)$ are unspecified, they are obviously equivalent. Then, when p and θ are not a function of covariates, and when $F(\cdot|\boldsymbol{z})$ and $S_u(|\boldsymbol{z})$ are unspecified, they are also equivalent. They represent the same model in a different form. On the contrary, when p and θ are a function of covariates the two models are different. In a first case, when $F(\cdot)$ does not depend on covariates, they will represent different models for different data structures (a model with the proportional hazards property for the promotion time cure model vs. a mixture cure model that does not verify the property). In a second case, when $F(\cdot)$ is a function of covariates, they will be both flexible models for survival data with a cure fraction.

2.3.2 Unifying Models: Specification and Estimation

Two different streams have driven the development of unifying models. On one side, we have models that have a pure mathematical motivation. On the other side, some articles tried to extend the biological motivation of the promotion time cure model and new classes of models appeared.

Mathematical Perspective

Unifying models developed in a mathematical perspective are all based on Box-Cox transformations. Yin & Ibrahim (2005) proposed to apply this transformation to the survival function $S_{pop}(t|\boldsymbol{x},\boldsymbol{z})$, such that

$$\begin{cases} \frac{S_{pop}(t|\boldsymbol{x},\boldsymbol{z})^{\alpha}-1}{\alpha} & = -\theta(\alpha,\boldsymbol{x})F(t|\boldsymbol{z}), \text{ if } 0<\alpha\leq 1 \\ \log\{S_{pop}(t|\boldsymbol{x},\boldsymbol{z})\} & = -\theta(0,\boldsymbol{x})F(t|\boldsymbol{z}), \text{ if } \alpha=0, \end{cases} \tag{2.19}$$

where α is a transformation parameter. The survival function is then given by

$$S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \{1-\alpha\theta(\alpha,\boldsymbol{x})F(t|\boldsymbol{z})\}^{1/\alpha}, & \text{if } 0<\alpha\leq 1 \\ \exp\{-\theta(0,\boldsymbol{x})F(t|\boldsymbol{z})\}, & \text{if } \alpha=0, \end{cases}$$

and the associated cure probability is given by $\lim_{t\to\infty} S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = \{1-\alpha\theta(\alpha,\boldsymbol{x})\}^{1/\alpha}$ for $0<\alpha\leq 1$, and $\exp\{-\theta(0,\boldsymbol{x})\}$ for $\alpha=0$. They proposed to model $\theta(\alpha,\boldsymbol{x})$ as $\theta(\alpha,\boldsymbol{x}) = \{\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}/\{1+\alpha\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}$, and they assumed that $F(t|\boldsymbol{z}) = 1-S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$. The mixture cure model and the promotion time cure model are two special cases of this model when $\alpha=1$ and when $\alpha=0$, respectively. When $0<\alpha<1$, an intermediate model is obtained. Both Bayesian and frequentist estimation methods have been proposed. Yin & Ibrahim (2005) considered a Bayesian approach and assumed a piecewise exponential distribution for $S_0(\cdot)$. They considered α as random, and they proposed a uniform discrete prior for this parameter in order to guarantee stability of the model. Note that they mentioned that there is no advantage of a random α in comparison with a fixed one except that it facilitates Bayesian inference. Parameters are considered independent a priori and they proposed noninformative prior distributions. Peng & Xu (2012) proposed a frequentist estimation method for model (2.19) based on a maximum likelihood approach. They considered a piecewise constant hazard model for $S_0(\cdot)$ defined

as $S_0(t|\lambda) = \exp[-\int_0^t \lambda_0(s|\lambda)ds]$, where $\lambda_0(t|\lambda) = \exp(\lambda_j)$, and $\lambda = (\lambda_1, ..., \lambda_J)$ is the vector of constant hazards corresponding to each of the J time intervals. Parameter estimates are obtained from the following likelihood function:

$$\mathcal{L}(\alpha, \boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda) = \prod_{i=1}^{n} \{\theta(\alpha, \mathbf{X}_i) f(Y_i|\mathbf{Z}_i)\}^{\Delta_i} \{1 - \alpha\theta(\alpha, \mathbf{X}_i) F(Y_i|\mathbf{Z}_i)\}^{1/\alpha - \Delta_i}.$$

Alternatively, Taylor & Liu (2007) proposed to apply a Box-Cox transformation to both side of the equation in order to obtain the unifying model

$$\begin{cases} \frac{S_{pop}(t|\boldsymbol{x},\boldsymbol{z})^\alpha - 1}{\alpha} & = \frac{q(\boldsymbol{x})^\alpha - 1}{\alpha} F(t|\boldsymbol{z}), \text{ if } 0 < \alpha \leq 1 \\ \log\{S_{pop}(t|\boldsymbol{x}, \boldsymbol{z})\} & = \log\{q(\boldsymbol{x})\} F(t|\boldsymbol{z}), \text{ if } \alpha = 0, \end{cases}$$

where α is a transformation parameter, and $q(\boldsymbol{x}) = 1 - p(\boldsymbol{x})$. In that case, the survival function is given by

$$S_{pop}(t|\boldsymbol{x}, \boldsymbol{z}) = \begin{cases} [1 + \{q^\alpha(\boldsymbol{x}) - 1\} F(t|\boldsymbol{z})]^{1/\alpha}, & \text{if } 0 < \alpha \leq 1 \\ \exp\left[\log\{q(\boldsymbol{x})\} F(t|\boldsymbol{z})\right], & \text{if } \alpha = 0, \end{cases}$$

with a cure probability equal to $q(\boldsymbol{x})$. As for the proposal of Yin & Ibrahim (2005), when $\alpha = 0$ the model reduces to the promotion time cure model. When $\alpha = 1$, it becomes the mixture cure model. In term of modelling, they proposed as an example a complementary log-log model for the cure probability $q(\boldsymbol{x})$ depending on a parameter vector $\boldsymbol{\gamma}$, and they considered a Weibull model for $1 - F(t|\boldsymbol{z}) = \exp[-\lambda t^\rho \exp(\boldsymbol{\beta}^t \boldsymbol{z})]$. No formal estimation is proposed. However, Taylor & Liu (2007) defined the likelihood function for the model:

$$\mathcal{L}(\alpha, \boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda, \rho) = \prod_{i=1}^{n} \left[\{q(\mathbf{X}_i)^\alpha - 1\} f(Y_i|\mathbf{Z}_i)\right]^{\Delta_i} \alpha^{-\Delta_i} S_{pop}(Y_i|\mathbf{X}_i, \mathbf{Z}_i)^{1-\alpha\Delta_i}.$$

Note that by assuming that $q(\boldsymbol{x}) = \{1 + \alpha \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})\}^{-1/\alpha}$, the two unifying models are equivalent as explained by Peng & Xu (2012).

Biological Perspective

The biological development of the promotion time cure model assumes that the survival time T is generated by the latent survival times $\tilde{T}_k$ such that $T = \min\{\tilde{T}_k, k = 1, ..., N\}$. Cooner et al. (2007) proposed to widen this relationship and considered that a number r out of the N carcinogenic cells need to be activated in order to produce a failure time. The survival time is then defined as $T = \tilde{T}_{(r)}$, for $1 \leq r \leq N$, where $\tilde{T}_{(r)}$, $r = 1, ..., N$ represent the ordered latent activation times. As for the promotion time cure model, they assumed that $\tilde{T}_k$ are i.i.d. random variables with distribution function $F(\cdot)$. The associated survival function for the whole population is given by

$$S_{pop}(t) = E_{(N,r)}\{S(t|N, r)\} \tag{2.20}$$

where $E_{N,r}$ is the expectation with respect to the joint distribution of (N, r), and $S(t|N, r) = P(T > t|N, r)$ is given by

$$P(T > t|N, r) = I(N = 0) + \sum_{j=0}^{r-1} \binom{N}{j} F(t)^j S(t)^{N-j} I(N \geq r \geq 1).$$

The cure probability for this model is given by $\lim_{t\to\infty} S_{pop}(t) = P(N = 0)$. The variable r is considered as a threshold variable determining the survival time T. It can be considered as a constant, as a function of N or as a random variable. Moreover, depending on its value, different activation schemes are possible.

When r is considered as random, a conditional distribution is specified for $r|N$. In order to model the survival time, Cooner et al. (2007) proposed to decompose the joint distribution of r and N in (2.20) as the product of the conditional distribution of r given N and the marginal distribution of N. In such a case, a so-called hierarchical-activation scheme is obtained. They considered two types of conditional distributions for $r|N$: a mixture distribution where a positive mass on $\{1, N\}$ is attributed to $r|N$ with probability π and $1 - \pi$ respectively, and a binomial distribution for $r - 1|N$ with parameter $N - 1$ and π. For the marginal distribution of N, four main distributions are considered: a Poisson, a Bernoulli, a binomial and a geometric one.

In the particular case where $r = 1$, i.e. when only one of the N tumour cells needs to be activated in order to produce a tumour, the survival time will be equal to the first latent activation time associated with the first cell that gives a tumour, that is, $T = \min\{\tilde{T}_k, k = 1, ..., N\}$. In such a case the survival function reduces to

$$\begin{aligned} S_{pop}(t) &= E_N\{S(t|N, 1)\} \\ &= E_N\{I(N = 0) + S(t)^N I(N \geq 1)\} \\ &= E_N\left\{S(t)^N\right\}. \end{aligned} \tag{2.21}$$

A so-called first-activation scheme is obtained. When it is assumed that $N \sim Poisson(\theta)$, $\theta > 0$, the model becomes the promotion time cure model. When a Bernoulli distribution with parameter $0 \leq \theta \leq 1$ is assumed for N, the model reduces to the mixture cure model. Other distributions, such as a binomial or a geometric one, are also considered by Cooner et al. (2007), these distributions giving other types of cure models. Another interesting distribution for N that has been proposed in the literature is the negative binomial. Tournoud & Ecochard (2008), Rodrigues et al. (2009), and de Castro et al. (2009) considered this distribution. The survival function for the whole population in such a case is given by

$$S_{pop}(t) = \{1 + \rho\theta F(t)\}^{-1/\rho},\ \rho \geq -1,\ \theta > 0,$$

where $\theta = E(N)$, $\rho = -1/N$, and $V(N) = \theta + \rho\theta^2$. Interestingly, this model embeds the mixture (when $\rho = -1$) and the promotion time cure models (when $\rho \to 0$). Moreover, it is equivalent to model (2.19) proposed by Yin & Ibrahim (2005) when $\rho = -\alpha$ for ρ in $[-1, 0]$.

At the other extreme, Cooner et al. (2007) considered the case where $r = N$, that is, all tumour cells N need to be activated in order to produce a tumour. In such a case, the survival time will be defined as the largest latent activation time associated with the last activated cell, that is, $T = \max\{\tilde{T}_k, k = 1, ..., N\}$. This class of models is referred to as the last-activation scheme. The survival function is given by

$$S_{pop}(t) = P(N = 0) + 1 - E_N\left\{F(t)^N\right\}.$$

As for the first-activation scheme, they considered different distributions for N: Bernoulli, binomial, Poisson, and geometric. Different types of cure models are obtained that are different from the mixture and the promotion time cure models. In summary, the proposal of Cooner et al. (2007) represents a general class of cure models that allows more flexibility to model survival data with a cure fraction than the mixture or the promotion time cure model.

In terms of modelling, Cooner et al. (2007) proposed a Weibull distribution for $S(\cdot)$ that depends on covariates. For the parameter θ, they considered the case where it is a function of covariates assuming the form $\theta(\boldsymbol{x}) = \exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})$ and also the case where it does not depend on covariates. In an attempt to make the model more flexible, Cooner et al. (2009) proposed later a piecewise exponential distribution for the latent time $\tilde{T}_k$. For the first-activation scheme, Tournoud & Ecochard (2008) assumed that $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})$ when N follows a negative binomial distribution, whereas de Castro et al. (2009) considered that $\theta(\boldsymbol{x}) = \{q(\boldsymbol{x})^{-\alpha} - 1\}/\alpha$, where $q(\boldsymbol{x}) = \{\exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})\}/\{1 + \exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}$ when $\alpha \neq 0$ and that $\theta(\boldsymbol{x}) = -\log\{q(\boldsymbol{x})\}$ when $\alpha = 0$. Furthermore, de Castro et al. (2009) assumed that $F(\cdot)$ has a parametric distribution (they proposed a Weibull or a piecewise exponential distribution) and did not consider covariates.

Regarding estimation approaches, Cooner et al. (2007) and de Castro et al. (2009) proposed to perform Bayesian inference. Cooner et al. (2007) considered a marginalized likelihood function because the survival function (2.20) depends on N and r which are latent, and discussed the choice of the prior distributions with an emphasis on the parameter θ in order to guarantee the identifiability of the model. They proposed to use a MCMC algorithm to sample from the posterior distribution. de Castro et al. (2009) considered an observed-data likelihood and as Yin & Ibrahim (2005), they assumed that the parameters are independent a priori and chose a discrete uniform prior for α. A MCMC algorithm is also used to sample from the posterior distribution.

Beside the Cooner et al. (2007) proposal, an even more general model has been proposed by Kim et al. (2011). Based on the same biological approach, they considered that an individual experiences a failure if a certain number of the N carcinogenic cells greater or equal to a threshold variable R is activated, where R may or may not be independent of N. In such a way, they relax the dependence assumption on N formulated for r by Cooner et al. (2007). The survival time is then defined as $T = \tilde{T}_{(R)}$, and the survival function is given by

$$S_{pop}(t) = P(N < R) + E_{(N,R)}\left\{\sum_{j=0}^{r-1}\left(\begin{array}{c} N \\ j \end{array}\right) F(t)^j S(t)^{N-j} I(N \geq R)\right\},$$

where the cured proportion is $P(N < R)$. R, which can be fixed or random, may be considered as the antibody level of the immune system for example. If R is assumed dependent of N, the joint distribution of (N, R) can be specified as the product of the conditional distribution of R given N, and of the marginal distribution of N. In such a case, the model becomes the hierarchical-activation scheme from Cooner et al. (2007). If $R = 1$, the model reduces to the first-activation scheme because only one cell needs to be activated in order

to produce a tumour. If additionally N follows a Poisson distribution with parameter θ, the model is the promotion time cure model. Equivalently, if R is considered as random with a geometric distribution, and if N follows a Poisson distribution, the model reduces to the promotion time cure model. Kim et al. (2011) only considered a Poisson distribution for N. However if we assume that $R = 1$ and that $N \sim Bernoulli(p)$, the model becomes the mixture cure model. As previously, Kim et al. (2011) assumed that $\theta(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{x})$, and considered a piecewise exponential distribution for the latent time $\tilde{T}_k$. A Bayesian approach is proposed to estimate the model where they assumed that the distribution of R is known and that R and N are independent.

2.3.3 Model Selection

Unifying models offer a flexible way to model survival data with a cure fraction. As explained in the previous section, there exist several possible models for each proposal. A question of interest is then how to choose the most adequate model? Two directions have been investigated. From models proposed by Yin & Ibrahim (2005) and Taylor & Liu (2007), the transformation parameter α is usually considered as random, and it is estimated alongside the other parameters. The best model is directly determined by the estimation process. However, this approach supposes that one can estimate the parameter α with precision which seems to be difficult because there is usually not enough information about the parameter in the data as explained by Yin & Ibrahim (2005). This point is corroborated by Diao & Yin (2012), who proposed a model similar to Yin & Ibrahim (2005) with a frailty, and who mentioned the fact that when the sample size is small, the likelihood function is quite flat for α. Taylor & Liu (2007) performed a simulation study where they evaluated a fixed versus a random α, and they draw the same conclusion: when the sample size is small, it is complicated to obtain a precise estimate for α. An alternative is to consider a grid of values for α and to perform model comparison as proposed by Diao & Yin (2012).

For biologically specified models, because there is no transformation parameter, model selection is also performed based on model comparison. The main idea is to define different distributions for random quantities, such as for N or R, and to select the best fit according to some criteria. Cooner et al. (2007) proposed to fit several models to the data with different activation schemes and different distributions for N and to compare them based on the posterior predictive L-measure proposed by Laud & Ibrahim (1995) and Gelfand & Ghosh (1998), a measure that rewards goodness-of-fit, assessed via posteriori predictive comparison, and at the same time penalizes for complexity. de Castro et al. (2009) proposed a similar approach and used the deviance information criterion (DIC) and the conditional predictive ordinate statistic to compare the models. Kim et al. (2011) fit different models with different distributions for R to the data, and compare them based on the DIC and the logarithm of the pseudo-marginal likelihood. Others, like Tournoud & Ecochard (2008) proposed simply to choose the most appropriate model based on the scientific knowledge of the disease, the hazard structure (proportional or not, depending

on the distribution assumed for N), and on the variance structure from the distribution of N.

Even if these proposals are targeted to compare flexible cure rate models, one can implement them to compare the mixture cure model and the promotion time cure model. Peng & Xu (2012) went in this direction, but proposed to perform likelihood ratio and score tests from model (2.19) in order to evaluate the adequacy of the mixture cure model ($H_0 : \alpha = 1$ vs. $H_1 : \alpha \neq 1$) and of the promotion time cure model ($H_0 : \alpha = 0$ vs. $H_1 : \alpha > 0$). The proposed tests perform well with large sample sizes but they are sensitive to a misspecification of the baseline distribution function F.

Chapter 3

The Single-Index/Cox Mixture Cure Model

A second contribution of this thesis to cure models focuses on the mixture cure model and more precisely on the modelling of the cure proportion. In Chapter 2, we have reviewed in detail the existing literature on that model. As we have explained, various models have been proposed to model the latency. Parametric models are, for example, given in Boag (1949) and Farewell (1982). A semiparametric approach based on a Cox PH model is provided by, for example, Kuk & Chen (1992), Sy & Taylor (2000), and Lu (2008), whereas a completely nonparametric estimation approach for $S_u(t|\mathbf{z})$ is given in Patilea & Van Keilegom (2018). On the other hand, much less attention has been paid to the incidence and the modelling and estimation of the cure rate $1-p(\boldsymbol{x})$. In fact, it is common practice to assume a logistic model given by $p(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})/\{1+\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}$ for some parameter vector $\boldsymbol{\gamma}$ and an intercept γ_0. The logistic model has certainly a number of important qualities. For instance, it is easy to interpret, it is easy to estimate, and it is available in various statistical software packages. However, it also has an important drawback. Indeed, the logistic function $\exp(a)/\{1+\exp(a)\}$ is of a fixed known form, while there is no reason to constraint the cure rate $1-p(\boldsymbol{x})$ to have a S-shape. Furthermore, even if $p(\boldsymbol{x})$ is often monotone in practical applications, we might think of situations where $p(\boldsymbol{x})$ would for example be increasing in $\boldsymbol{\gamma}^t\boldsymbol{x}$ up to some threshold, after which it would become decreasing. In order to accommodate these concerns, we propose in this chapter a single-index model for $p(\boldsymbol{x})$, i.e. we assume that there exists an unknown link function $g(\cdot)$ (monotone or non-monotone) such that

$$p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t\boldsymbol{x}). \tag{3.1}$$

The link function g can be any (smooth) function with values between 0 and 1, and will be estimated nonparametrically using kernel methods. Considering the popular Cox PH model for the latency, and a single-index model for the incidence, we will refer to our model as the Single-Index/Cox (SIC) cure model.

Single-index models have been used in various contexts and have several

advantages. First of all, as explained above, they are much more flexible than purely parametric models, like logistic or probit models, which assume that the link function g is a fixed known function. Second, contrary to the completely nonparametric models for $p(\boldsymbol{x})$, they do not suffer from curse-of-dimensionality problems, since they summarize the covariate vector $\mathbf{X}$ into one single score $\boldsymbol{\gamma}^t\mathbf{X}$, often referred to as the *index*. Third, despite their nonparametric nature, they remain quite easy to interpret. Indeed, to compare the relative importance of one covariate with respect to another, it suffices to standardise the covariates and to compare the absolute value of the corresponding $\boldsymbol{\gamma}$-coefficients. We refer to Ichimura (1993) and Klein & Spady (1993) for some early references, to the book by Horowitz (2009) for a nice overview of existing results, and to Lu & Burke (2005), Wang et al. (2007), Lopez (2009), and Lopez et al. (2013) for papers which have used the single-index model in survival analysis.

This chapter is organized as follows. In Section 3.1 we detail our proposed model, i.e. the SIC cure model, we provide results about the identifiability of the model and we define a maximum likelihood based estimation procedure. The proof of the identifiability of the model is given in the Appendix at the end of this chapter. The finite sample performance of the proposed estimators is investigated through a numerical study in Section 3.2. It also contains a discussion on the issue of bandwidth selection for the kernel based non-parametric estimator of g. The breast cancer data introduced in Chapter 2 is analysed with the SIC cure model in Section 3.3, and we end the chapter with some concluding remarks in Section 3.4.

This chapter is based on

> Amico, M., Van Keilegom, I., and Legrand, C. (2018). The single-index/Cox mixture cure model. *Biometrics* (to appear).

3.1 The Model and its Estimation

Let us consider the same setting as in the previous chapters, that is, the survival time T is subject to random right censoring, T and C are independent given the covariates $(\mathbf{X}^t, \mathbf{Z}^t)^t$ and the data consists in $(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, which are i.i.d. realisations of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$.

3.1.1 The Single-Index/Cox (SIC) Cure Model

Consider the mixture cure model (1.5) and assume a single-index structure (3.1) for $p(\boldsymbol{x})$, where $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_d)$ and d is the dimension of $\mathbf{X}$. For identifiability reasons we do not include an intercept in the model. For the latency, a Cox PH model is considered with survival function

$$S_u(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}, \tag{3.2}$$

where $S_0(t) = P(T > t|B = 1)$ is the unspecified baseline conditional survival function and $\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_q)$ is a vector of parameters associated with $\mathbf{Z}$ that does not include an intercept (with $q = \dim(\mathbf{Z})$). The conditional hazard

function is given by $\lambda_u(t|\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^t\boldsymbol{z})$, where $\lambda_0(t) = f_0(t)/S_0(t)$ is the baseline conditional hazard function, with $f_0(t) = -(d/dt)S_0(t)$. Note that $S_u(t|\boldsymbol{z})$ is a proper survival function, i.e. $S_u(t|\boldsymbol{z}) \to 0$ when $t \to \infty$, implying that the survival function is such that $S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) \to 1 - p(\boldsymbol{x})$ when $t \to \infty$. Note also that, as already mentioned in Chapter 2, even if the conditional hazard $\lambda_u(t|\boldsymbol{z})$ satisfies the proportional hazard property, this is not the case for the hazard function of the entire population.

3.1.2 Identifiability of the Model

Under the data generating process described before, and assuming that the censoring is non-informative, the likelihood of an observation $(y, \delta, \boldsymbol{x}, \boldsymbol{z})$ is given by

$$\mathcal{L}(y,\delta,\boldsymbol{x},\boldsymbol{z}) = \{g(\boldsymbol{\gamma}^t\boldsymbol{x})f_u(y|\boldsymbol{z})\}^{\delta}\{1 - g(\boldsymbol{\gamma}^t\boldsymbol{x}) + g(\boldsymbol{\gamma}^t\boldsymbol{x})S_u(y|\boldsymbol{z})\}^{1-\delta},$$

where $f_u(t|\boldsymbol{z}) = -(d/dt)S_u(t|\boldsymbol{z})$.

In practice, the censoring time C is bounded. This prevents us from observing cured subjects in the data. A way around this problem is to assume the existence of a so-called 'cure threshold' $\tau < \infty$, that we already described in Chapter 2, such that $T > \tau$ implies that $T = +\infty$ (Taylor (1995)). This assumption is commonly accepted and often used in cure models literature. As a consequence, whenever Y is observed to be greater than τ, the individual is assumed to be cured.

We first derive our identifiability result for the case where $\mathbf{X}$ is only composed of continuous random variables. Additional assumptions are necessary when $\mathbf{X}$ is composed of both continuous and discrete random variables. We give them at the end of this subsection.

For a continuous random vector $\mathbf{X}$, our identifiability result is derived under the following set of assumptions:

(A1) (i) The function g is differentiable and not constant on the support of $\boldsymbol{\gamma}^t\mathbf{X}$.

(ii) The components of $\mathbf{X}$ are continuous random variables that have a joint probability density function.

(iii) The support of $\mathbf{X}$ is not contained in any proper linear subspace of $\mathbb{R}^p$.

(iv) $\boldsymbol{\gamma}^t\mathbf{X}$ does not contain an intercept.

(v) Either $\gamma_1 = 1$, or $\|\boldsymbol{\gamma}\| = 1$ and the sign of γ_1 is fixed, with $\|\cdot\|$ the Euclidean norm.

(A2) (i) $\boldsymbol{\beta}^t\mathbf{Z}$ does not contain an intercept.

(ii) The matrix $\mathrm{Var}(\mathbf{Z})$ has full rank.

(A3) (i) There exists a $\tau < \infty$ (cure threshold) such that

(a) $T > \tau \iff T = \infty$,

(b) $P(C > \tau|\mathbf{X},\mathbf{Z}) > 0$ for almost all $\mathbf{X}$ and $\mathbf{Z}$.

(ii) For all $\boldsymbol{x}$, $0 < p(\boldsymbol{x}) < 1$.

Assumption (A1) is required to identify the single-index model (see Horowitz (2009), Theorem 2.1, p. 14), whereas (A2) is needed to make sure that the Cox model is identifiable, and (A3) is a typical assumption needed to identify the mixture cure model (see e.g. Taylor (1995)).

Proposition 3.1.1. *Under (A1)-(A3), the model given by (1.5), (3.1) and (3.2) is identifiable.*

The proof of this Proposition is given in the Appendix provided at the end of this chapter.

The identifiability result stated above only considers that $\mathbf{X}$ is a vector of continuous covariates. When $\mathbf{X}$ is a mixture of continuous and discrete variables, two additional conditions, as described in Horowitz (2009), are necessary:

(A4) The support of $\boldsymbol{\gamma}^t\mathbf{X}$ must not be divided in disjoint subsets when the values of the discrete components vary.

(A5) The function g is not periodic.

3.1.3 Maximum Likelihood Estimation

In the context of mixture cure models, the likelihood function takes the form

$$\mathcal{L} = \prod_{i=1}^{n} \{p(\mathbf{X}_i) f_u(Y_i|\mathbf{Z}_i)\}^{\Delta_i} \left[\{1 - p(\mathbf{X}_i)\} + p(\mathbf{X}_i) S_u(Y_i|\mathbf{Z}_i)\right]^{1-\Delta_i}. \tag{3.3}$$

When the latency is modelled as a Cox PH model, a particular feature of this model is that the baseline hazard is left unspecified. In Chapter 2, however, we explained that the profile likelihood approach usually considers to fit the Cox PH model and to obtain the partial likelihood in classical survival analysis, can not be considered for the mixture cure model. Indeed, by eliminating $\lambda_0(t)$ from the likelihood function (3.3), one would lose part of the information about the incidence because the baseline hazard function is conditional on the uncure status B. Furthermore, cured and uncured censored observations have the same contribution to the likelihood function and no distinction is made between the two. As we have seen in Chapter 2, one solution consists in using the EM algorithm, as proposed by Sy & Taylor (2000), in order to handle both the fact that the cure status is partially unobserved and the fact that $\lambda_0(t)$ is unspecified. For the mixture cure model, the complete-data likelihood is given by

$$\begin{aligned} \mathcal{L}_c &= \prod_{i=1}^{n} \{p(\mathbf{X}_i) \lambda_u(Y_i|\mathbf{Z}_i) S_u(Y_i|\mathbf{Z}_i)\}^{B_i \Delta_i} \\ &\quad \times \prod_{i=1}^{n} \left[\{p(\mathbf{X}_i) S_u(Y_i|\mathbf{Z}_i)\}^{B_i} \times \{1 - p(\mathbf{X}_i)\}^{1-B_i}\right]^{1-\Delta_i}. \end{aligned} \tag{3.4}$$

As it can be seen from equation (3.4), cured and uncured censored subjects have a different contribution to this likelihood, and one can therefore use a profile likelihood-type approach to estimate the survival function for uncured observations without losing information about the incidence.

As explained in Chapter 1, the EM algorithm consists in maximising the complete-data likelihood by alternating between an E and a M step until convergence. For the mixture cure model, at the m^{th} iteration of the algorithm, as it has been detailed in Chapter 2, the E-step consists first in computing

$$E\left(B_i|O_i,\theta^{(m-1)}\right) = \Delta_i + (1-\Delta_i)\frac{p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|\mathbf{Z}_i)}{1-p^{(m-1)}(\mathbf{X}_i)+p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|\mathbf{Z}_i)} = W_i^{(m)},$$

and then substituting $W_i^{(m)}$ in the logarithm of (3.4) to obtained the expected log-complete-data likelihood or equivalently the expected complete-data likelihood

$$\begin{aligned}\tilde{\mathcal{L}}_c &= \prod_{i=1}^{n}\left\{p(\mathbf{X}_i)\lambda_u(Y_i|\mathbf{Z}_i)S_u(Y_i|\mathbf{Z}_i)\right\}^{W_i^{(m)}\Delta_i} \\ &\quad\times\prod_{i=1}^{n}\left[\left\{p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)\right\}^{W_i^{(m)}}\times\left\{1-p(\mathbf{X}_i)\right\}^{1-W_i^{(m)}}\right]^{1-\Delta_i}. \quad (3.5)\end{aligned}$$

In the M-step, (3.5) is maximised with respect to the parameters of the model. Since (3.5) can be re-written as the product of two factors, each of them containing the parameters of one part of the model, it may be maximised separately for the two parts of the model :

$$\begin{aligned}\tilde{\mathcal{L}}_c &= \prod_{i=1}^{n}\left[p(\mathbf{X}_i)^{W_i^{(m)}}\left\{1-p(\mathbf{X}_i)\right\}^{1-W_i^{(m)}}\right] \\ &\quad\times\prod_{i=1}^{n}\left\{\lambda_u(Y_i|\mathbf{Z}_i)^{\Delta_i}S_u(Y_i|\mathbf{Z}_i)\right\}^{W_i^{(m)}} \\ &= \tilde{\mathcal{L}}_1\times\tilde{\mathcal{L}}_2. \quad (3.6)\end{aligned}$$

For the incidence, we follow the maximum likelihood approach of Klein & Spady (1993) to estimate a single-index model with a binary outcome. The likelihood function given in (3.6) is the same as for a logistic regression model, except that the link function $g(\cdot)$ is totally unknown :

$$\tilde{\mathcal{L}}_1(g,\boldsymbol{\gamma}) = \prod_{i=1}^{n} g(\boldsymbol{\gamma}^t\mathbf{X}_i)^{W_i^{(m)}}\left\{1-g(\boldsymbol{\gamma}^t\mathbf{X}_i)\right\}^{1-W_i^{(m)}}.$$

To estimate the single-index model, there is first a need to estimate the link function $g(\cdot)$. The following (infeasible) leave-one-out kernel estimator of $g(\boldsymbol{\gamma}^t\mathbf{X}_i)$

based on Nadaraya-Watson weights :

$$\sum_{j\neq i}^{n} \frac{K\left(\frac{\boldsymbol{\gamma}^t\mathbf{X}_i-\boldsymbol{\gamma}^t\mathbf{X}_j}{h}\right)}{\sum_{l\neq i}^{n} K\left(\frac{\boldsymbol{\gamma}^t\mathbf{X}_i-\boldsymbol{\gamma}^t\mathbf{X}_l}{h}\right)} B_j,$$

where h is a one-dimensional bandwidth, can not be used in practice as B is latent. Alternatively, B can be replaced by its expectation $W^{(m)}$ obtained in the E-step of the algorithm. The Nadaraya-Watson estimator becomes

$$\tilde{g}_{-i}^{(m)}(\boldsymbol{\gamma}^t\mathbf{X}_i) = \sum_{j\neq i}^{n} \frac{K\left(\frac{\boldsymbol{\gamma}^t\mathbf{X}_i-\boldsymbol{\gamma}^t\mathbf{X}_j}{h}\right)}{\sum_{l\neq i}^{n} K\left(\frac{\boldsymbol{\gamma}^t\mathbf{X}_i-\boldsymbol{\gamma}^t\mathbf{X}_l}{h}\right)} W_j^{(m)}. \tag{3.7}$$

The kernel estimator (3.7) is substituted in $\tilde{\mathcal{L}}_1$, and $\boldsymbol{\gamma}$ is estimated by maximising the likelihood function with numerical techniques such as the Newton-Raphson algorithm. The resulting estimator of $\boldsymbol{\gamma}$ is denoted by $\hat{\boldsymbol{\gamma}}^{(m)}$. Once $\hat{\boldsymbol{\gamma}}^{(m)}$ is computed, $g(\boldsymbol{\gamma}^t\mathbf{X}_i)$ is estimated by the estimator given in (3.7), but with $\boldsymbol{\gamma}$ replaced by $\hat{\boldsymbol{\gamma}}^{(m)}$. The estimator is denoted by $\hat{g}_{-i}^{(m)}\{(\hat{\boldsymbol{\gamma}}^{(m)})^t\mathbf{X}_i\}$. Note that when $d = 1$, as $\gamma_1 = 1$, the estimator (3.7) reduces to a non-parametric estimator.

For the latency, the likelihood function given in (3.6) is similar to the likelihood function for the classical Cox PH model except that the baseline cumulative hazard $\Lambda_0(t)$and $\lambda_0(t)$ are conditional on $B = 1$:

$$\tilde{\mathcal{L}}_2(\Lambda_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} \Big[\big\{ \lambda_0(Y_i) \exp(\boldsymbol{\beta}^t\mathbf{Z}_i) \big\}^{\Delta_i} \exp\big\{ -\Lambda_0(Y_i) \exp(\boldsymbol{\beta}^t\mathbf{Z}_i) \big\} \Big]^{W_i^{(m)}}.$$

The profile likelihood approach based on the work of Breslow (1974) and proposed by Sy & Taylor (2000) that have been presented in Chapter 2 is considered to estimate $\boldsymbol{\beta}$. In a first step, the value of $\boldsymbol{\beta}$ is fixed and $\Lambda_0(t)$ is estimated non-parametrically by

$$\sum_{j:Y^*_{(j)}\leq t} \frac{D_j}{\sum_{k\in R_j} W_k^{(m)} \exp(\boldsymbol{\beta}^t\mathbf{Z}_k)}, \tag{3.8}$$

In a second step, (3.8) is plugged in $\tilde{\mathcal{L}}_2$, obtaining the partial likelihood

$$\breve{\mathcal{L}}_2(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^t\mathbf{Z}_i)}{\sum_{k\in R_i} W_k^{(m)} \exp(\boldsymbol{\beta}^t\mathbf{Z}_k)} \right\}^{\Delta_i}. \tag{3.9}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by maximising (3.9), and is denoted by $\hat{\boldsymbol{\beta}}^{(m)}$. $\hat{\Lambda}_0^{(m)}(t)$ denotes the estimator of $\Lambda_0(t)$ given in (3.8) but with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}^{(m)}$. The EM algorithm iterates until the difference between two consecutive values of the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $S_0(\cdot)$ is smaller than a certain value a priori defined. The final estimators are denoted by $\hat{\boldsymbol{\gamma}}$, $\hat{g}(\cdot)$, $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(\cdot)$.

In Chapter 2, we have explained that when the latency is modelled non- or semiparametrically, one issue arises from the conditional survival function estimation. In fact, the lack of information in the right tail of the survival function (with possibly a large number of censored observations after the last uncensored observation) may lead to a situation where $\hat{S}_u(Y^*_{(r)}) \neq 0$, resulting in identifiability problems. To overcome this situation, Taylor (1995) proposed the so-called 'cure threshold' also referred to as the *zero-tail constraint* which consists, in practice, to impose in the E-step of the EM algorithm that the weight W_i equals 0 if $Y_i > Y^*_{(r)}$. It is equivalent to consider that observations censored after $Y^*_{(r)}$ are cured. Under the assumed Cox PH model for the latency, this constraint is also applied to the SIC cure model.

3.2 Numerical Study

The objective of this numerical study is to compare the fit of the SIC cure model with the fit of the classical logistic/Cox (LC) cure model in various settings. The two models differ in their incidence (a single-index versus a logistic structure), while they both assume a Cox PH model in the latency. The impact of a single-index structure on both the estimates of $p(\boldsymbol{x})$ and $S_u(t|\boldsymbol{z})$ are investigated as well as the impact of different censoring rates.

3.2.1 Some Preliminaries

Data for the incidence is generated according to three scenarios, each of them corresponding to a different link function. The first scenario assumes a logistic link function of the form $g(a) = \exp(\gamma_0 + a)/\{1 + \exp(\gamma_0 + a)\}$, where γ_0 is an intercept term. The second scenario is an adaptation of a model coming from Müller & Schmitt (1988), namely $g(a) = \exp[0.75\ \Phi\{(\gamma_0 + a) + 0.5\} + 0.25\ \Phi\{0.5(\gamma_0 + a)^3\}]/\big(1 + \exp[0.75\ \Phi\{(\gamma_0 + a) + 0.5\} + 0.25\ \Phi\{0.5(\gamma_0 + a)^3\}]\big)$, with $\Phi(\cdot)$ the standard normal cumulative distribution, leading to a non-logistic but monotone link function. The third scenario assumes a non-monotone link function of the form $g(a) = \exp[0.4\{(\gamma_0+a)^3 - (\gamma_0+a)^2 - 0.9(\gamma_0+a)+1\}]/\big(1+\exp[0.4\{(\gamma_0+a)^3 - (\gamma_0+a)^2 - 0.9(\gamma_0+a)+1\}]\big)$. Figure 3.1 shows the three link functions. Note that, for the second and the third scenarios, the link functions are represented alongside the logistic link function (the dotted curve) in order to allow the comparison. As it can be seen, these last two link functions were chosen sufficiently different from a logistic link function in order to assess the performance of a single-index and a logistic model in such situations. Note also that we included a logistic transformation in the two last link functions in order to restrict the probability to the interval $[0, 1]$.

For all scenarios, we consider four independent covariates: X_1 and X_2 have a standard normal distribution, and X_3 and X_4 have a Bernoulli distribution with parameters 0.3 and 0.6 respectively. The parameters $\gamma_0, \ldots, \gamma_4$ are chosen so that a sufficient cure proportion is obtained and so that the identification condition (A4) is fulfilled. Table 3.1 summarizes the parameter values and the cure proportions. Note that the intercept γ_0 is only estimated under the logistic

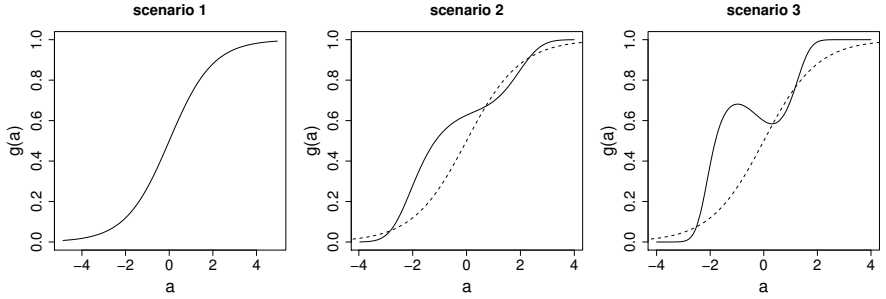

Figure 3.1: *Link functions considered for the incidence in the data generation process when* $\gamma_0 = 0$ *(dotted curve: logistic link function).*

model. Under the single-index model it is incorporated into the link function g.

For the latency, we consider one binary covariate Z that is independent of $\mathbf{X} = (X_1, X_2, X_3, X_4)^t$ and that has a Bernoulli distribution with parameter 0.6. The survival times for the uncured observations are generated according to a Weibull PH model, i.e. $S_u(t|z) = \exp(-\lambda t^\rho)^{\exp(\beta z)}$ for given choices of λ, ρ and β. For the three scenarios, the scale parameter λ equals 1.5, the shape parameter ρ equals 1.2, and $\beta = 1.5$. The censoring time C is independent of the vector $(\mathbf{X}, Z, T)$, and is generated from an exponential distribution with density function $f(t) = \lambda_c \exp(-\lambda_c t)$. Three different values for the rate λ_c, denoted by level 1, level 2 and level 3, are considered so that the difference between the censoring rate and the cure rate increases. An increment of 5% between each level has been considered. The values for each scenario are given in Table 3.1. For both the survival and the censoring times, the parameter values were chosen in such a way that a certain number of the censored observations have a follow-up time larger than $Y^*_{(r)}$, and such that these observations are cured. In such a way we mimic the type of real survival data on which cure models are typically used. Additionally by increasing λ_c our goal is to assess the impact of larger censoring rates on the model estimates. Note that when the censoring rate increases, the proportion of observations with a follow-up time larger than $Y^*_{(r)}$ tends to decrease. As a consequence, very large censoring rates are not considered in order to still have a non-negligible number of observations in the plateau. Table 3.1 gives the censoring rates and the percentage of observations in the plateau for the different values of λ_c.

For each setting, we consider samples of size $n = 250$ and 500. This leads to a total of 18 settings (3 model scenarios, 3 censoring rates and 2 sample sizes). For each setting we generate 500 datasets.

For each dataset, we fit a LC cure model and a SIC cure model. For the single-index model, we use an Epanechnikov kernel, given by $K(u) = 3(1 - u^2/5)/(4\sqrt{5})I(u^2 \leq 5)$. The bandwidth is selected according to a likelihood cross-validation procedure which has been previously considered for

Table 3.1: *The parameter values for the incidence, the cure proportions, the censoring rates, and the proportion of observations in the plateau for each scenario.*

	incidence parameters						*censoring scheme*			
scen.	γ_0	γ_1	γ_2	γ_3	γ_4	cure rate	censoring level	λ_c	censoring rate	% obs. plateau
1	1.4	-1.5	0.5	2.3	-1.3	18.6%	level 1	0.05	20.1%	16.3%
						18.6%	level 2	0.25	25.7%	10.3%
						18.6%	level 3	0.5	31.4%	6.7%
2	-0.3	1.5	-0.9	2	-1	34.4%	level 1	0.05	35.6%	30.3%
						34.4%	level 2	0.25	40.1%	19.5%
						34.4%	level 3	0.55	45.6%	12.0%
3	1	1.3	-2	-2	1	37.9%	level 1	0.1	40.2%	29.7%
						37.9%	level 2	0.35	45.2%	18.2%
						37.9%	level 3	0.7	50.7%	11.2%

single-index models by Strzalkowska-Kominiak & Cao (2013). It is performed at the beginning of the M-step, prior to the incidence estimation and it is computed at each iteration of the algorithm. The cross-validation criterion is given by minus the logarithm of $\tilde{\mathcal{L}}_1$ evaluated at $\hat{\boldsymbol{\gamma}}^{(m-1)}$, that is,

$$CV^{(m)}(h) = -\sum_{i=1}^{n} W_i^{(m)} \log \hat{g}_{h,-i}^{(m-1)}\Big\{\big(\hat{\boldsymbol{\gamma}}^{(m-1)}\big)^t \mathbf{X}_i\Big\} - \sum_{i=1}^{n} \big(1 - W_i^{(m)}\big) \log\Big[1 - \hat{g}_{h,-i}^{(m-1)}\Big\{\big(\hat{\boldsymbol{\gamma}}^{(m-1)}\big)^t \mathbf{X}_i\Big\}\Big], \quad (3.10)$$

where $\hat{g}_{h,-i}^{(m-1)}(\cdot)$ is the leave-one-out estimator (3.7) with $\boldsymbol{\gamma}$ replaced by $\hat{\boldsymbol{\gamma}}^{(m-1)}$ and based on a bandwidth h. The selected bandwidth is the minimizer on the interval [0.4,1] of the cross-validation criterion. It is given by

$$h_{CV}^{(m)} = \arg\min_h CV^{(m)}(h).$$

The interval [0.4,1] is chosen based on visual inspection of what are reasonable bandwidths in this setting. The bandwidth obtained at the last iteration of the algorithm is the final bandwidth. In order to make both models identifiable, the conditional survival function is forced to be equal to 0 beyond $Y_{(r)}^*$ as proposed by Taylor (1995). For the SIC cure model, we apply the identification conditions mentioned previously. For condition (A1)-(v), we set $\hat{\gamma}_1 = 1$ or $\hat{\gamma}_1 = -1$, depending on the sign of $\hat{\gamma}_1$ for the LC cure model. These identification conditions will be assumed for the remaining of the article. For the two models, the initial values of the EM algorithm are the parameters coming from a logistic regression model taking the censoring indicator as response variable for the incidence and coming from a Cox PH model fitted on the uncensored observations only for the latency. We consider that the algorithm converges when the difference between two consecutive values of the parameters is smaller than 10^{-5}.

In order to evaluate the performance of the SIC cure model for the incidence, we compute the Average Squared Error (ASE) of $\hat{p}(\boldsymbol{x})$ for each model,

$$ASE(\hat{p}) = V^{-1} \sum_{j=1}^{V} \Big\{\hat{g}(\hat{\boldsymbol{\gamma}}^t \boldsymbol{x}_j) - g(\boldsymbol{\gamma}^t \boldsymbol{x}_j)\Big\}^2.$$

The ASE is computed on a grid $(\mathbf{x}_j)_j = (\{x_{j1}, x_{j2}, x_{j3}, x_{j4}\})_j$, $j = 1, ..., V$, of values of the vector (X_1, X_2, X_3, X_4). For X_1 and X_2 we take grid points on $[-1.5, 1.5]$ with step size 0.01, while X_3 and X_4 take values in $\{0, 1\}$. For the latency, we compute the bias, the variance and the mean squared error (MSE) of $\hat{\beta}$ for each model.

3.2.2 Results

Table 3.2 shows the bias, variance, and MSE of $\hat{\beta}$ for the two models. As can be seen from the table, the bias of $\hat{\beta}$ is small for the two models, and

Table 3.2: *Bias, variance and mean squared error (MSE) of $\hat{\beta}$ for the SIC cure model and for the LC cure model.*

			censoring schemes								
			level 1			*level 2*			*level 3*		
n	scen.		bias	var.	MSE	bias	var.	MSE	bias	var.	MSE
250	1	SIC	0.0154	0.0343	0.0345	0.0248	0.0372	0.0378	0.0271	0.0394	0.0401
		LC	0.0149	0.0343	0.0346	0.0230	0.0370	0.0375	0.0233	0.0390	0.0396
	2	SIC	0.0206	0.0450	0.0454	0.0278	0.0480	0.0487	0.0285	0.0528	0.0536
		LC	0.0208	0.0450	0.0454	0.0289	0.0480	0.0489	0.0297	0.0524	0.0533
	3	SIC	0.0155	0.0473	0.0475	0.0165	0.0547	0.0549	0.0170	0.0665	0.0668
		LC	0.0160	0.0468	0.0471	0.0185	0.0551	0.0555	0.0189	0.0655	0.0659
500	1	SIC	0.0227	0.0169	0.0174	0.0192	0.0180	0.0184	0.0198	0.0199	0.0203
		LC	0.0224	0.0169	0.0174	0.0176	0.0180	0.0183	0.0174	0.0199	0.0202
	2	SIC	0.0117	0.0210	0.0211	0.0092	0.0224	0.0225	0.0047	0.0262	0.0262
		LC	0.0120	0.0210	0.0212	0.0102	0.0225	0.0227	0.0062	0.0262	0.0262
	3	SIC	0.0168	0.0220	0.0223	0.0123	0.0242	0.0244	0.0151	0.0292	0.0294
		LC	0.0178	0.0220	0.0223	0.0150	0.0244	0.0247	0.0176	0.0296	0.0299

we observe very small differences between them. These small differences however tend to increase when λ_c increases. This situation occurs because when λ_c increases, the censoring rate increases, and the number of observations in the plateau becomes substantially smaller than the number of cured observations. In such a situation, the number of censored observations such that $Y \leq Y^*_{(r)}$ gets larger. The weight $W_i^{(m)}$ for these observations will be equal to $\{p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|Z_i)\}/\{1-p^{(m-1)}(\mathbf{X}_i)+p^{(m-1)}(\mathbf{X}_i)S_u^{(m-1)}(Y_i|Z_i)\}$, that is, it will rely more on $\hat{p}(\boldsymbol{x})$ than when an observation is uncensored ($W_i^{(m)} = 1$) or when an observation is censored after $Y^*_{(r)}$ ($W_i^{(m)} = 0$). Depending on the estimate for $p(\boldsymbol{x})$, these observations censored before $Y^*_{(r)}$ will have a different contribution to the likelihood function for the latency. Larger differences in the estimation of β are observed between the two models. Despite this effect, the bias is still small for the two models and it seems that choosing between a logistic regression and a single-index model does not have a large impact on the latency parameter estimate when the latency modeling is the same. Furthermore, the variance and MSE of $\hat{\beta}$ are very similar for the two models. Nevertheless, when λ_c increases, $\hat{\beta}$ becomes more variable, and the MSE increases for the same reason as above.

For the incidence, Figure 3.2 gives the boxplots of the ASE of $\hat{p}(\boldsymbol{x})$ for the two models. As can be expected, the LC cure model performs better than the SIC cure model when the true model is a logistic regression, regardless of the sample size (see the boxplots associated with scenario 1 in Figure 3.2). For scenarios 2 and 3 on the contrary, when the true link function is different from a logistic one, the SIC cure model performs better for all censoring rates and sample sizes considered. As is the case for the latency, the censoring rate has an impact on the quality and the precision of the estimates of $p(\mathbf{x})$, with the ASE taking on higher values and being more variable when λ_c increases. The explanation is the same as above. When the censoring rate gets much larger than the cure proportion, more observations have a weight $W_i^{(m)}$ taking values between 0 and 1. Because $W_i^{(m)}$ is involved in the likelihood function for estimating $p(\boldsymbol{x})$, the uncertainty in the estimation of $p(\boldsymbol{x})$ increases, and larger and more variable values of ASE are observed. On the contrary, as can be expected, when the sample size increases, a decrease in the value of the ASE is observed, meaning a better fit to the data.

Estimating a SIC cure model is computationally more demanding. However, although a difference by a factor of around 85 has been observed in our simulations, the required time remains reasonable, going from 0.10 seconds for the LC cure model to 8.4 seconds for the SIC cure model when $n = 250$, and from 0.18 seconds to 15 seconds when $n = 500$ (2.7 GH processor and 16 Go of memory computer). With a more complex model, for e.g. with a higher dimension of $\mathbf{X}$ or a more complex link function, the computation time increases for both models, mainly because the EM algorithm gets slower. In term of EM algorithm convergence, between 99.2 % and 100% of the datasets converged when estimating a SIC cure model.

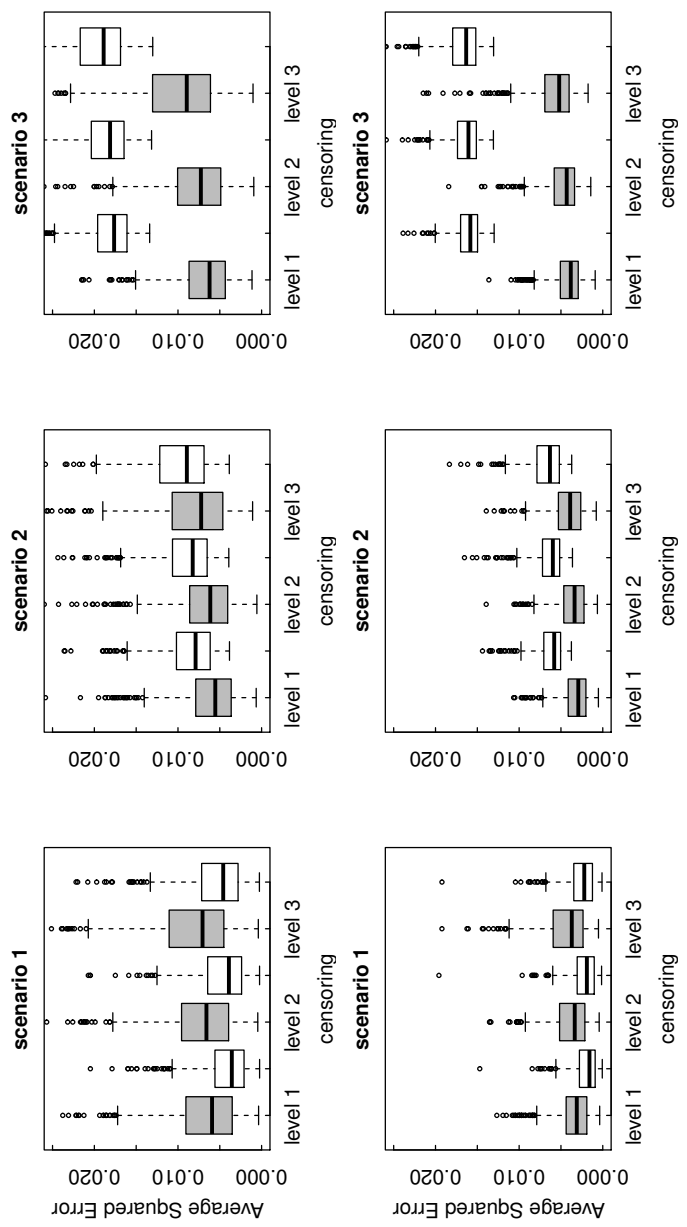

Figure 3.2: *Boxplots of the Average Squared Error (ASE) for the single-index model (grey boxplots) and the logistic model (white boxplots).*

3.3 Real Data Application

Our application concerns a breast cancer study which consists of 286 patients that experienced a lymph-node-negative breast cancer between 1980 to 1995 (Wang et al., 2005) and that has been already analysed in Chapter 2. The event of interest is distant metastasis, and the associated survival time is the time to distant metastasis (DM). Among the 286 patients, 107 experienced a distant recurrence from breast cancer. As can be seen from Figure 2.1, the Kaplan-Meier estimator of the survival function shows a large plateau at about 60%, and 88% of the censored observations are in the plateau. A cure model seems therefore appropriate for these data. Four covariates are considered: age of the patient (ranges from 26 to 83 with a median of 52 years), estrogen receptor (ER) status (0 = ER−: less than 10 fmol per mg protein - 77 patients, 1 = ER+: 10 fmol per mg protein or more - 209 patients), size of the tumour (ranges from 1 to 4 with a median of 1), and menopausal status (0 = pre-menopausal - 129 patients, 1 = post-menopausal - 157 patients) which has been obtained by dichotomising age (pre-menopausal: age $\leq$ 50 - post-menopausal: age $>$ 50). The original data are split in a training and a test set following the 2/3 - 1/3 recommendations of Hastie et al. (2009). For our application, it corresponds to 190 and 96 observations respectively. The training set is used to estimate and to interpret the model. The test set is used to compute the cross-entropy error (CEE) of the predictions of the logistic and the single-index models, a measure of loss often considered for binary classification (see, for example, Hastie et al. (2009)). It is given by

$$CEE = -\sum_{j=1}^{n_{test}} \log\left[\hat{p}(\boldsymbol{x}_j^{test})^{\hat{w}_j}\left\{1-\hat{p}(\boldsymbol{x}_j^{test})\right\}^{1-\hat{w}_j}\right],$$

where n_{test} is the size of the test set, $\hat{p}(\boldsymbol{x}_j^{test})$ and $\hat{w}_j$ are the predicted uncure probability and the predicted weight for the j^{th} observation in the test set, respectively, computed based on the parameter estimates (and the link function for the single-index) obtained from the training set. On the training set, a LC and a SIC cure model including the four covariates in both parts of the two models are fitted. For interpretation convenience, a standardized version of all covariates, continuous and discrete, is considered in the incidence. For the single-index model, we proceed as in Section 3.2 for model fitting, bandwidth selection, and model identification, and the parameters are further normalized so that $\|\hat{\boldsymbol{\gamma}}^{SI}\| = 1$, with $\hat{\boldsymbol{\gamma}}^{SI}$ the vector of parameter estimates under the single-index model, in order to ease the interpretation.

The standard errors of the parameter estimates for both models have been computed using a naive bootstrap approach, which consists in drawing with replacement $M = 250$ resamples $\{(Y_i^*, \Delta_i^*, \mathbf{X}_i^*, \mathbf{Z}_i^*), i = 1, \ldots, n\}$ from the training set, each resample having the same size as the training set. On each resample, a LC and a SIC cure model are estimated and the standard error of each parameter for each model are then computed over the M resamples.

Table 3.3 gives the parameters estimates jointly with p-values for the Wald test of the parameters. For both models, the effects for the four covariates on

Table 3.3: *Parameter estimates and standard errors for the SIC cure model and for the LC cure model.*

	SIC cure model			*LC cure model*		
Incidence	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
(intercept)	-	-	-	-0.5707	0.1514	0.0002
age	-0.8111	0.2394	0.0007	-0.3668	0.2593	0.1573
ER status [ER+ vs. ER-]	0.1758	0.2923	0.5475	0.1422	0.1805	0.4308
size of the tumour	-0.0003	0.3496	0.9992	-0.0058	0.1563	0.9705
menopausal [post vs. pre]	0.5577	0.3293	0.0894	0.3413	0.2678	0.2026
bandwidth	*0.3245*	-	-	-	-	-
Latency	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
age	-0.0113	0.0212	0.5953	-0.0107	0.0211	0.6122
ER status [ER+ vs. ER-]	-1.3639	0.3855	0.0004	-1.3640	0.3886	0.0004
size of the tumour	0.4527	0.2977	0.1284	0.4524	0.3027	0.1351
menopausal [post vs. pre]	-0.2519	0.6178	0.6835	-0.2644	0.6166	0.6682

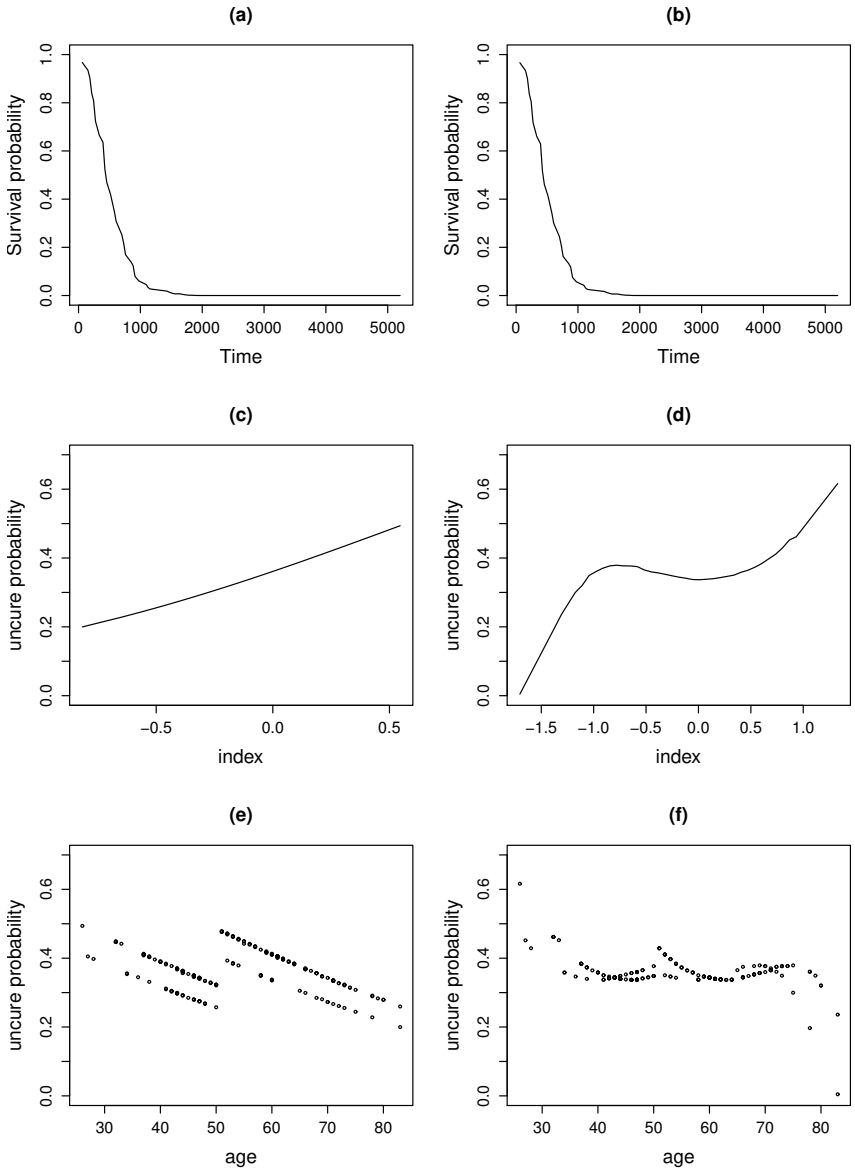

Figure 3.3: *Estimated baseline survival function for (a) the LC mixture cure model and (b) the SIC mixture cure model; estimated link function for (c) the logistic model and (d) the single-index model (index = $\hat{\boldsymbol{\gamma}}^t\boldsymbol{x}$); and plot of the effect of age on the uncure probability for (e) the logistic model and (f) the single-index model.*

the latency have the same direction, and the estimates are very close. Moreover, the ER status affects significantly the time to DM of uncured patients showing that a positive ER status implies a longer time to DM. Regarding the estimated baseline survival function, as it can be seen in Figure 3.3 (a)–(b), it is the same for both models.

For the incidence, we first compute the cross entropy error for both models to compare their fits. The cross entropy error for the SIC cure model equals to 67.71, while it is equal to 71.57 for the LC cure model, meaning that the SIC cure model performs better than the LC cure model in predicting the uncure status. To understand why the single-index model performs better, we compare the estimates of the two models. Beside different parameter estimates, the link functions, given in Figure 3.3 (c)–(d), are also different, the single-index link function showing a non-monotone trend which is quite different from the logistic one. Nevertheless, as already pointed out by Li & Duan (1989) in a broader context, even if two models are estimated with different link functions, the relative importance of the parameter estimates is similar, permitting to evaluate which variable impacts the most the outcome. Here, age and menopausal status are the most influential covariates on the uncure probability because of their larger estimates in comparison with the other covariates. However, to interpret covariate effects on $p(\boldsymbol{x})$, the link function and the parameters have to be considered jointly as the sign of the effect depends on the shape of the link function. If the link function is monotone increasing (resp. decreasing), the sign of the effect will be the same as (resp. the opposite) the sign of the estimate. For this dataset, the parameter estimates show, at a 10% level, that age and menopausal status have a significant effect on the uncure probability for the single-index model, whereas none of them are significant for the logistic model. As the link function is not monotone, we can not interpret the parameter estimates directly. But we can instead plot the uncure probability against covariates to better understand their effects. Figure 3.3 (e)–(f) show the effect of age on $p(\boldsymbol{x})$ for both models. Two different groups appear, corresponding to pre-menopausal and post-menopausal women. Contrary to the logistic model where the relationship seems to be almost equivalent and linear in the two groups, it seems that the trend for pre-menopausal in the single-index model is flat between roughly 35 and 50, and much higher in the 25-35 range, while for post-menopausal it is roughly flat from 50 to 70, then much lower for ages greater than 70.

To further extend the data analysis, we sequentially remove the non significant covariates based on the p-values. We obtain a final model containing age and the menopausal status in the incidence, and age and the estrogen receptor status in the latency.

3.4 Conclusion

We proposed a new semi-parametric modelling approach for the mixture cure model by considering a single-index structure for the uncure probability. We proved the identifiability of the proposed model, parameter estimators are ob-

tained via maximum likelihood using the EM algorithm, and for the unknown link function of the single-index a Nadaraya-Watson estimator is proposed with a likelihood cross-validation method for the bandwidth parameter. Based on a numerical study, we showed the good performance of the method for the estimation of the uncure probability, and its ability to outperform the logistic model when the true link function is not the logistic one. The application of the method to a breast cancer dataset illustrates its practical use and the usefulness of a single-index model. The analysis uncovered a non-monotone link function and both the parameter estimates and the graphical representation indicate a more complex effect of age not observed with the logistic model. However, even if the link function is left unspecified, providing yet a greater flexibility, the single-index model assumes a linear combination of covariates, which might be questionable in this particular context. Indeed, since age seems to have a non-linear effect on the uncure probability, one could envisage to take this feature into account in a different way, not with a flexible link function, but by considering a flexible modelling of the covariates. For example, we could consider a generalised additive model introduced by Hastie & Tibshirani (1986) instead of a single-index for the incidence, where the logistic link function is kept but where instead of having a linear combination of covariates, the index is given by $\sum_{j=1}^{d} s_j(X_j)$, where $s(\cdot)$ are unspecified smooth functions. Such mixture cure models have already been proposed by Li & Taylor (2002) who consider cubic B-splines to estimate $s_j(\cdot)$. Alternatively, we could keep the single-index structure, but with an additional feature for the specific effect of age. For example, a quadratic term could be added, or we could also envisage something more flexible by introducing the effect of age as a smooth unknown function, that is, having an index taking the form $\boldsymbol{\gamma}^t\mathbf{X}' + s(X_{age})$, with $\mathbf{X}'$ the vector $\mathbf{X}$ without age and where $s(\cdot)$ has to be estimated.

Due to its greater flexibility over the logistic regression, the SIC cure model may be considered as a diagnostic tool to investigate misspecification of the incidence of the mixture cure model. A possible extension of our work will be to develop a test for the parametric form of $p(\cdot)$. Müller & Van Keilegom (2018) proposed such a test based on the non-parametric estimator of the cure rate proposed by Xu & Peng (2014) for the case where $\boldsymbol{X}$ is one-dimensional (see Chapter 2 for a brief description). By avoiding curse of dimensionality problems, our estimator $\hat{g}(\hat{\boldsymbol{\gamma}}^t\mathbf{X})$ may constitute a natural extension of their work when $\boldsymbol{X}$ is multi-dimensional. Furthermore, its flexibility has another advantage as it permits to model a wide range of link functions, monotone or not. However in certain contexts it is natural to assume that the link function is monotone. A possible adaptation of this work would be to restrict the link function to the monotone case.

3.5 Appendix: Proof of Proposition 3.1.1

To show that the model is identifiable, suppose that there exist $(g, \boldsymbol{\gamma}, S_0, \beta)$ and $(\tilde{g}, \tilde{\boldsymbol{\gamma}}, \tilde{S}_0, \tilde{\beta})$ such that

$$\begin{aligned} & \left\{g(\boldsymbol{\gamma}^t\mathbf{x}) f_u(y|\mathbf{z})\right\}^{\delta} \ \left\{1 - g(\boldsymbol{\gamma}^t\mathbf{x}) + g(\boldsymbol{\gamma}^t\mathbf{x}) S_u(y|\mathbf{z})\right\}^{1-\delta} \qquad (3.11) \\ = & \left\{\tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{x}) \tilde{f}_u(y|\mathbf{z})\right\}^{\delta} \ \left\{1 - \tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{x}) + \tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{x}) \tilde{S}_u(y|\mathbf{z})\right\}^{1-\delta} \end{aligned}$$

for all realizations $(y, \delta, \mathbf{x}, \mathbf{z})$ of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$, where $S_u(y|\mathbf{z}) = S_0(y)^{\exp(\boldsymbol{\beta}^t\mathbf{z})}$, $\tilde{S}_u(y|\mathbf{z}) = \tilde{S}_0(y)^{\exp(\tilde{\boldsymbol{\beta}}^t\mathbf{z})}$, and $f_u(y|\mathbf{z})$ and $\tilde{f}_u(y|\mathbf{z})$ are the corresponding probability density functions. Then, we need to show that $g \equiv \tilde{g}, \boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}, S_0 \equiv \tilde{S}_0$ and $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$.

First, consider $y > \tau$ and $\delta = 0$. Note that

$$\begin{aligned} P(Y > \tau, \Delta = 0|\mathbf{X}, \mathbf{Z}) &= P(C > \tau, C \leq T|\mathbf{X}, \mathbf{Z}) \\ &= P(C > \tau, T = \infty|\mathbf{X}, \mathbf{Z}) \\ &= P(C > \tau|\mathbf{X}, \mathbf{Z}) \ (1 - p(\mathbf{X})) > 0. \end{aligned}$$

Indeed, since C is greater than τ and smaller or equal than T, il follows that $T > \tau$. Then, thanks to (A3), it follows that $T = \infty$ and that $(C > \tau, C \leq T)$ is equivalent to $(C > \tau, T = \infty)$. Hence, $(y > \tau, \delta = 0, \mathbf{x}, \mathbf{z})$ is a possible realisation of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. Equation (3.11) reduces in this case to $1 - g(\boldsymbol{\gamma}^t\mathbf{x}) = 1 - \tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{x})$. Theorem 2.1 in Horowitz (2009) together with assumption (A1) ensures that $g \equiv \tilde{g}$ and $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$.

Next, take $y \leq \tau$ and $\delta = 1$, and observe that

$$\begin{aligned} & P(Y \leq \tau, \Delta = 1|\mathbf{X}, \mathbf{Z}) \\ & \quad = P(T \leq \tau, T \leq C|\mathbf{X}, \mathbf{Z}) \\ & \quad \geq P(T \leq \tau, C > \tau|\mathbf{X}, \mathbf{Z}) = P(T \leq \tau|\mathbf{X}, \mathbf{Z}) P(C > \tau|\mathbf{X}, \mathbf{Z}) \\ & \quad = p(\mathbf{X}) P(C > \tau|\mathbf{X}, \mathbf{Z}) > 0, \end{aligned}$$

given that T and C are independent given $(\mathbf{X}, \mathbf{Z})$ and again by (A3), showing that $(y \leq \tau, \delta = 1, \mathbf{x}, \mathbf{z})$ is possible. The likelihood contribution in (3.11) is such that $g(\boldsymbol{\gamma}^t\mathbf{x}) f_u(y|\mathbf{z}) = \tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{x}) \tilde{f}_u(y|\mathbf{z})$. Since $g(\boldsymbol{\gamma}^t\mathbf{X}) = \tilde{g}(\tilde{\boldsymbol{\gamma}}^t\mathbf{X}) = p(\mathbf{X}) > 0$, it follows that $f_u(y|\mathbf{z}) = \tilde{f}_u(y|\mathbf{z})$. Since the Cox model is identified under condition (A2), it follows that $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $S_0 \equiv \tilde{S}_0$. □

Chapter 4

Assessing Cure Status Prediction from Survival Data Using ROC Curves

As we have seen in the previous chapters, two outcomes are of interest when modelling cure survival data: the survival at a given time t, as in classical survival analysis, and the cure status. In Chapter 2, we have seen that the literature on cure models mainly focuses on modelling the effect of covariates on these two quantities, while very little has been done on evaluating the quality of predictions. However, good predictions are essential for practitioners. In fact, when there exists a possible cure fraction, we can think of situations where one would be interested in predicting who is cured and who is not based on marker(s) in order to determine if a treatment is necessary to prevent a cancer relapse. Likewise, being able to correctly predict the survival probability of an uncured patient after a certain time by taking into account the presence of cured subjects in the data is also important. A first contribution to that topic is due to Yu et al. (2008) who propose to validate individual prediction for patients with prostate cancer performed based on a joint longitudinal survival-cure model. Recently, Beyene et al. (2018) investigate the accuracy of time-dependent event prediction, extending to cure survival data the results that have been previously obtained for classical survival analysis (see for example Heagerty et al. (2000), Heagerty & Zheng (2005), Chambless & Diao (2006), Blanche et al. (2013), Li et al. (2018) among others). Zhang & Shao (2018) propose a concordance measure, in the spirit of the c-index proposed by Harrell et al. (1982) and Harrell et al. (1984), to assess the prediction accuracy of the overall survival for uncured patients by taking into account the presence of a cure fraction. They extend the work of Göner & Heller (2005) for the Cox PH model. For the cure status, on the contrary, nothing has been done to the best of our knowledge, while it is an important issue.

In Chapter 1, we have introduced the ROC curve, a graphical tool which aims to evaluate if a continuous classifier correctly performed a binary classifi-

cation. As a recall, it plots the sensitivity given by

$$Se(k) = P(M > k|D = 1), \tag{4.1}$$

against one minus the specificity given by

$$1 - Sp(k) = P(M > k|D = 0), \tag{4.2}$$

for all possible values of $k \in \mathbb{R}$. Its equation is given by

$$ROC(u) = Se\left\{(1 - Sp)^{-1}(u)\right\}, \quad 0 < u < 1,$$

where u is an index.

In this chapter, we propose to develop a ROC curve approach in order to evaluate the accuracy of a (combination of) covariate(s) to predict the cure status based on cure survival data. Since the cure status is missing for censored observations, 'classical' ROC curve approaches, which rely on the knowledge of the classes of the observations, can not be directly implemented in this context. Therefore, an important issue to address is how to handle the latency of the cure status. Our proposal is presented in Section 4.1 alongside some important points related to the estimation of the sensitivity and the specificity. In Section 4.2, some asymptotic properties are presented, followed in Section 4.3 by the investigation of the finite sample performance of the proposed method through an extensive simulation study. Section 4.4 illustrates the practical use of our proposal on a melanoma dataset, while Section 4.5 concludes with some final remarks and discussion. Finally, Appendix 1 given at the end of this chapter provides additional results on the finite sample performance of our estimators, while Appendix 2 gives the proofs of the asymptotic properties derived in Section 4.2.

This chapter is based on

> Amico, M. and Van Keilegom, I. (2018b). Assessing cure status prediction from survival data using ROC curves, *Submitted*.

4.1 Methodology

Once again, let us consider the same setting as in the previous chapters, where T is subject to random right censoring, T and C are independent given $\mathbf{X}$ and $\mathbf{Z}$, and let us consider that we have a random sample of n i.i.d. observations $(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, having the same distribution as $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. We further assume that the data comes from the mixture cure model (1.5).

The objective is to derive a ROC curve estimator in order to evaluate the prediction accuracy of M for the cure status $D = I(T = \infty)$. Hereafter, we assume that $M = \gamma_0 + \boldsymbol{\gamma}^t\mathbf{X}$, where $\boldsymbol{\gamma}$ is a vector of parameters associated with $\mathbf{X}$ and γ_0 is an intercept term. We further assume that $\mathbf{X}$ can be unidimensional or multidimensional, and for this latter case, that the vector of parameters $(\gamma_0, \boldsymbol{\gamma}^t)^t$ can be known, in such a case M is a known score such as a genetic score, for example, or unknown that needs therefore to be estimated.

4.1.1 Infeasible Estimators

In Chapter 1, we have mentioned that a simple and common nonparametric method to estimate a ROC curve consists in estimating the sensitivity and the specificity by their empirical distribution functions given by

$$\check{Se}(k) = 1 - \frac{1}{\check{N}_1} \sum_{i=1}^{n} \check{W}_{i1} I(M_i \leq k), \tag{4.3}$$

$$\check{Sp}(k) = \frac{1}{\check{N}_0} \sum_{i=1}^{n} \check{W}_{i0} I(M_i \leq k), \tag{4.4}$$

where $\check{W}_{i1} = D_i$, $\check{W}_{i0} = 1 - D_i$, $\check{N}_1 = \sum_{i=1}^{n} \check{W}_{i1}$ and $\check{N}_0 = n - \check{N}_1$. The ROC curve estimator takes therefore the form of a step function with jumps at each M_i. When working with cure survival data, however, these estimators cannot be used as the cure status is unobserved.

When dealing with cure survival data, we have explained in Chapter 2 that Taylor (1995) proposes to consider as cured an observation with a follow-up time greater than the last uncensored follow-up time $Y^*_{(r)}$. Based on this *cure threshold* denoted by τ, it is therefore possible to distinguish three types of observations from cure survival data. In fact, since an uncensored subject experiences the event, it belongs to the non-cured population with certainty, that is, $D = 0$. Based on the cure threshold, censored observations can be separated into two groups, those with a follow-up time $Y > \tau$, for which $D = 1$, and those with a follow-up time $Y \leq \tau$. For this latter case, a probability, given by $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$, replaces the unobserved cure status. It follows that estimators for the sensitivity and the specificity are given by the following weighted empirical distribution functions:

$$\tilde{Se}(k) = 1 - \frac{1}{\tilde{N}_1} \sum_{i=1}^{n} \tilde{W}_{i1} \; I(M_i \leq k), \tag{4.5}$$

$$\tilde{Sp}(k) = \frac{1}{\tilde{N}_0} \sum_{i=1}^{n} \tilde{W}_{i0} \; I(M_i \leq k), \tag{4.6}$$

where $\tilde{W}_{i1} = (1 - \Delta_i) P(D = 1|\mathbf{X} = \mathbf{X}_i, \mathbf{Z} = \mathbf{Z}_i, C = C_i, T > C_i)$, $\tilde{W}_{i0} = 1 - \tilde{W}_{i1}$, $\tilde{N}_1 = \sum_{i=1}^{n} \tilde{W}_{i1}$, and $\tilde{N}_0 = n - \tilde{N}_1$. Furthermore, when the cure threshold is assumed, $\tilde{W}_{i1}$ can further be written as $\tilde{W}_{i1} = I(Y_i > \tau) + \; (1 - \Delta_i) \; I(Y_i \leq \tau) \; P(D = 1|\mathbf{X} = \mathbf{X}_i, \mathbf{Z} = \mathbf{Z}_i, C = C_i, T > C_i)$. An infeasible estimator for the ROC curve is then given by

$$\widetilde{ROC}(u) = \tilde{Se}\{(1 - \tilde{Sp})^{-1}(u)\}, \; 0 < u < 1. \tag{4.7}$$

This estimator is a monotone increasing function of u and is invariant to strictly increasing transformations of M, which are both required properties of ROC curves as described by Pepe (2003). Note that these estimators consider a random design. However, they can also be applied when the design is fixed. In such a case, notations will be different.

The development of this method relies on the following theoretical elements. Based on the definition of conditional probability, the sensitivity (4.1) can be written as $Se(k) = P(M > k, T = \infty)/P(T = \infty)$. Let us consider the numerator:

$$\begin{aligned} P(M > k,&T = \infty) \\ &= E\left\{I(M > k)\ I(T = \infty)\right\} \\ &= E\left\{I(M > k)\ I(T > C)\ I(T = \infty)\right\} \\ &= E\left[I(M > k)\ I(T > C)\ E\{I(T = \infty)|\mathbf{X}, \mathbf{Z}, C, T > C\}\right] \\ &= E\left\{I(M > k)\ I(T > C)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\right\} \\ &= E\left\{I(M > k)\ (1 - \Delta)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\right\}. \end{aligned}$$

Hence,

$$Se(k) = \frac{E\left[I(M > k)\ (1 - \Delta)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\right]}{E\left[(1 - \Delta)P(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\right]}. \tag{4.8}$$

By assuming the cure threshold, and by replacing the expectation by a sum, it follows that a natural estimator for $Se(k)$ is given by (4.5).

Based on these derivations, an estimator for the AUC can also be obtained. The AUC defined in Chapter 1 can be written as $AUC = \int_0^1 Se\big[(1 - Sp)^{-1}(u)\big]\ du$. Define $f_0(k) = (d/dk)\ Sp(k)$. By proceeding to a change of variable (assuming that $(1 - Sp)^{-1}(u) = k$), we have for arbitrary $1 \leq i \neq j \leq n$ that

$$\begin{aligned} AUC &= \int_{-\infty}^{+\infty} Se(k)\ f_0(k)\ dk \\ &= E\Big\{Se(M_i)|T_i < \infty\Big\} \\ &= E\Big\{P(M_j > M_i|T_j = \infty, M_i)|T_i < \infty\Big\} \\ &= E\Big[E\big\{I(M_j > M_i)|T_j = \infty, M_i\big\}|T_i < \infty\Big], \\ &\qquad\qquad \text{with } (M_i, T_i) \perp\!\!\!\perp (M_j, T_j) \\ &= E\Big[E\big\{I(M_j > M_i)|T_j = \infty, M_i, T_i < \infty\big\}|T_i < \infty, T_j = \infty\Big] \\ &= E\big\{I(M_j > M_i)|T_j = \infty, T_i < \infty\big\} \\ &= P(M_j > M_i|T_j = \infty, T_i < \infty). \end{aligned}$$

Based on the definition of conditional probability, the AUC can be rewritten as

$$AUC = \frac{P(M_j > M_i, T_j = \infty, T_i < \infty)}{P(T = \infty)\ P(T < \infty)}.$$

Let us consider the numerator:

$$\begin{aligned} &P(M_j > M_i, T_j = \infty, T_i < \infty) \\ &= E\big\{I(M_j > M_i)\ I(T_j = \infty)\ I(T_i < \infty)\big\} \\ &= E\Big[I(M_j > M_i)\ E\big\{I(T_j = \infty)I(T_i < \infty)|\mathbf{X}_i, \mathbf{X}_j, \mathbf{Z}_i, \mathbf{Z}_j\big\}\Big]. \end{aligned}$$

Given that $(T_i, \mathbf{X}_i, \mathbf{Z}_i) \perp\!\!\!\perp (T_j, \mathbf{X}_j, \mathbf{Z}_j)$ and that $T_i \perp\!\!\!\perp T_j | (\mathbf{X}_i, \mathbf{X}_j, \mathbf{Z}_i, \mathbf{Z}_j)$ since

$$\begin{aligned} & f_{T_j|\mathbf{X}_j\mathbf{Z}_j}(t_1|\boldsymbol{x}_1, \boldsymbol{z}_1)\ f_{T_i|\mathbf{X}_i\mathbf{Z}_i}(t_2|\boldsymbol{x}_2, \boldsymbol{z}_2) \\ &= \frac{f_{T_j\mathbf{X}_j\mathbf{Z}_j}(t_1, \boldsymbol{x}_1, \boldsymbol{z}_1)}{f_{\mathbf{X}_j,\mathbf{Z}_j}(\boldsymbol{x}_1, \boldsymbol{z}_1)} \frac{f_{T_i\mathbf{X}_i\mathbf{Z}_i}(t_2, \boldsymbol{x}_2, \boldsymbol{z}_2)}{f_{\mathbf{X}_i\mathbf{Z}_i}(\boldsymbol{x}_2, \boldsymbol{z}_2)} \\ &= \frac{f_{T_jT_i\mathbf{X}_j\mathbf{Z}_j\mathbf{X}_i\mathbf{Z}_i}(t_1, t_2, \boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{x}_2, \boldsymbol{z}_2)}{f_{\mathbf{X}_j\mathbf{Z}_j\mathbf{X}_i\mathbf{Z}_i}(\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{x}_2, \boldsymbol{z}_2)} \\ &= f_{T_jT_i|\mathbf{X}_j\mathbf{Z}_j\mathbf{X}_i\mathbf{Z}_i}(t_1, t_2|\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{x}_2, \boldsymbol{z}_2), \end{aligned}$$

it follows that

$$\begin{aligned} & E\Big[I(M_j > M_i)\ E\{I(T_j = \infty)I(T_i < \infty)|\mathbf{X}_i, \mathbf{X}_j, \mathbf{Z}_i, \mathbf{Z}_j\}\Big] \\ &= E\Big[I(M_j > M_i)\ E\{I(T_j = \infty)|\cancel{\mathbf{X}_i}, \mathbf{X}_j, \cancel{\mathbf{Z}_i}, \mathbf{Z}_j\} \\ &\qquad \times E\{I(T_i < \infty)|\mathbf{X}_i, \cancel{\mathbf{X}_j}, \mathbf{Z}_i, \cancel{\mathbf{Z}_j}\}\Big]. \end{aligned}$$

Furthermore,

$$\begin{aligned} I(T_j = \infty) &= I(T_j = \infty)\{I(T_j \leq C_j) + I(T_j > C_j)\} \\ &= I(T_j = \infty)I(T_j > C_j), \end{aligned}$$

since $\{T_j = \infty\} \cap \{T_j \leq C_j\} = \varnothing$, and

$$\begin{aligned} I(T_i < \infty) &= I(T_i < \infty)\{I(T_i \leq C_j) + I(T_i > C_i)\} \\ &= I(T_i < \infty)I(T_i > C_j) + I(T_i \leq C_i), \end{aligned}$$

since $\{T_i < \infty\} \cap \{T_i \leq C_i\} = \Omega$.

Then,

$$\begin{aligned} & E\Big[I(M_j > M_i)\ E\{I(T_j = \infty)|\mathbf{X}_j, \mathbf{Z}_j\}\ E\{I(T_i < \infty)|\mathbf{X}_i, \mathbf{Z}_i\}\Big] \\ &= E\Big(I(M_j > M_j) \times E\{I(T_j = \infty)|\mathbf{X}_j, \mathbf{Z}_j, C_j, T_j > C_j\}I(T_j > C_j) \\ &\qquad \times \big[E\{I(T_i < \infty)|\mathbf{X}_i, \mathbf{Z}_i, C_i, T_i > C_i\}I(T_i > C_i) + I(T_i \leq C_i)\big]\Big). \end{aligned}$$

Hence,

$$\begin{aligned} & P(M_j > M_i, T_j = \ \infty, T_i < \infty) \\ &= E\Big[I(M_j > M_i) \times P(T_j = \infty|\mathbf{X}_j, \mathbf{Z}_j, C_j, T_j > C_j)(1 - \Delta_j) \\ &\qquad \times \big\{P(T_i < \infty|\mathbf{X}_i, \mathbf{Z}_i, C_i, T_i > C_i)(1 - \Delta_i) + \Delta_i\big\}\Big]. \end{aligned}$$

By replacing the expectation by a sum and by assuming the cure threshold, it follows that an infeasible estimator of the AUC is given by

$$\widetilde{AUC} = \frac{1}{\tilde{N}_0\tilde{N}_1} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{W}_{j1}\tilde{W}_{i0}\ I(M_j > M_i). \tag{4.9}$$

4.1.2 Feasible Estimators

The probability $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$ is involved in the infeasible estimators (4.5) and (4.6) of the sensitivity and the specificity, as well as in the infeasible AUC estimator (4.9). It is therefore necessary to estimate this quantity in order to obtain estimators that can be used in practice. Based on the definition of conditional probability, this probability can be written as

$$P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C) = \frac{P(T = \infty|\mathbf{X}, \mathbf{Z}, C)}{P(T > C|\mathbf{X}, \mathbf{Z}, C)} = \frac{P(T = \infty|\mathbf{X}, \mathbf{Z})}{P(T > C|\mathbf{X}, \mathbf{Z}, C)}$$

since T and C are independent given $\mathbf{X}$ and $\mathbf{Z}$. Since we suppose that the data come from the mixture cure model (1.5), it can be further written as

$$\frac{P(T = \infty|\mathbf{X}, \mathbf{Z})}{P(T > C|\mathbf{X}, \mathbf{Z}, C)} = \frac{1 - p(\mathbf{X})}{\{1 - p(\mathbf{X})\} + p(\mathbf{X})S_u(C|\mathbf{Z})}. \tag{4.10}$$

The literature on cure models offers various modelling approaches for the mixture cure model (1.5). The most common one is the LC mixture cure model proposed by Kuk & Chen (1992), and further studied by Sy & Taylor (2000) and Peng & Dear (2000) that has been introduced in Chapter 2. This proposal assumes a logistic model for p, that is $p(\boldsymbol{x}) = \exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})/\{1+\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}$ and considers a Cox PH model for S_u, where $S_u(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$, with $S_0(t) = P(T > t|T < \infty, \mathbf{Z} = 0)$, a baseline conditional survival function which remains totally unspecified, and $\boldsymbol{\beta}$ a vector of parameters associated with $\mathbf{Z}$. A drawback of this model, however, is that the estimator for $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$ relies on a parametric assumption for p which may not be fulfilled by the data. An alternative model is the SIC mixture cure model presented in Chapter 3, which assumes a single-index structure for p, that is $p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t\boldsymbol{x})$, where g is a smooth unknown function, and a Cox PH model for S_u. This SIC cure model assumes a less restrictive model for p and it may therefore be more appropriate. Both approaches are considered and their respective finite sample performances are compared in Section 4.3. The estimators for $\tilde{W}_{i0}$ and $\tilde{W}_{i1}$ are given by

$$\begin{aligned}
\hat{W}_{i1} &= I(Y_i > \tau) + (1 - \Delta_i)\, I(Y_i \leq \tau)\, \frac{1 - \hat{p}(\mathbf{X}_i)}{\{1 - \hat{p}(\mathbf{X}_i)\} + \hat{p}(\mathbf{X}_i)\hat{S}_u(Y_i|\mathbf{Z}_i)} \\
\hat{W}_{i0} &= 1 - \hat{W}_{i1},
\end{aligned}$$

and are obtained by either a LC cure model or a SIC cure model. The feasible estimators of Se, Sp, ROC and AUC are now given by

$$\hat{Se}(k) = 1 - \frac{1}{\hat{N}_1} \sum_{i=1}^{n} \hat{W}_{i1} \; I(M_i \leq k), \tag{4.11}$$

$$\hat{Sp}(k) = \frac{1}{\hat{N}_0} \sum_{i=1}^{n} \hat{W}_{i0} \; I(M_i \leq k), \tag{4.12}$$

$$\widehat{ROC}(u) = \hat{Se}\{(1 - \hat{Sp})^{-1}(u)\}, \qquad 0 < u < 1, \tag{4.13}$$

$$\widehat{AUC} = \frac{1}{\hat{N}_0 \hat{N}_1} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{W}_{j1} \hat{W}_{i0} \; I(M_j > M_i), \tag{4.14}$$

where $\hat{N}_1 = \sum_{i=1}^{n} \hat{W}_{i1}$ and $\hat{N}_0 = n - \hat{N}_1$.

Both $\mathbf{X}$ and $\mathbf{Z}$ enter in the computation of $\hat{W}_0$ and $\hat{W}_1$, while M only relies on $\mathbf{X}$. For the choice of the covariates to include in $\mathbf{X}$, we consider those included in M. A more delicate question concerns the choice of the covariates to consider for $\mathbf{Z}$. When M only contains one covariate, and when there is only one covariate available in the data, it is easy to assume that $X = Z$. If there are several covariates in the data, or when M is a combination of covariates, on the contrary, the choice of $\mathbf{Z}$ will depend on the knowledge of the topic of the analysis, and on which covariates are thought to influence the survival of uncured subjects. In such contexts, $\mathbf{Z}$ can be partially or fully identical to $\mathbf{X}$, or completely different from $\mathbf{X}$. However, we are not free of misspecification. The influence of a misspecification of this vector on the estimation of the ROC curve is therefore investigated through simulations in Section 4.3.

4.2 Asymptotic Theory

In this section we will develop the limiting distribution of the proposed estimators of the sensitivity, the specificity, the ROC curve and the AUC given in equations (4.11), (4.12), (4.13) and (4.14). In the previous section these estimators were constructed either based on a logistic/Cox mixture cure model or on a single-index/Cox mixture cure model. However, asymptotic theory for the estimation of these models has only been developed so far under the logistic/Cox model (see Lu (2008)), and so we restrict attention in this section to the latter model.

The proofs of the results of this section can be found in Appendix 2 at the end of this chapter.

The asymptotic theory for the estimation of the logistic/Cox mixture cure model derived by Lu (2008) relies on a the following set of assumptions.

There exists a constant κ such that:

(C1) The function $\Lambda_0(t)$ is strictly increasing and continuously differentiable, and $\Lambda_0(\kappa) < \infty$.

(C2) $\theta = (\gamma_0, \boldsymbol{\gamma}^t, \boldsymbol{\beta}^t)^t$ lies in the interior of a compact set $\mathcal{C}$ and the covariate vectors $\mathbf{X}$ and $\mathbf{Z}$ are bounded in the sense that $P(|\mathbf{X}| < m$ and $|\mathbf{Z}| < m) = 1$ for some constant $m > 0$.

(C3) With probability one, there exists a positive constant ϵ such that $P(C \geq T^* \geq \kappa | \mathbf{X}, \mathbf{Z}) > \epsilon$.

(C4) $P(Y > s | \mathbf{X}, \mathbf{Z})$ is continuous in s.

Theorem 4.2.1. *Assume that conditions (C1)–(C4) are satisfied and that the logistic/Cox mixture cure model is valid. Then,*

$$\begin{aligned} \hat{Se}(k) - Se(k) &= n^{-1} \sum_{i=1}^{n} \eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + R_{n,Se}(k) \\ \hat{Sp}(k) - Sp(k) &= n^{-1} \sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + R_{n,Sp}(k), \end{aligned}$$

where $\sup_k |R_{n,Se}(k)| = o_P(n^{-1/2})$, $\sup_k |R_{n,Sp}(k)| = o_P(n^{-1/2})$, *and* $\eta_{Se}(\mathbf{x}, \mathbf{z}, y, \delta, k)$ *and* $\eta_{Sp}(\mathbf{x}, \mathbf{z}, y, \delta, k)$ *are defined after (4.18) and (4.19) in Appendix 2.*

Moreover, the process $n^{1/2}(\hat{Se}(k) - Se(k))$ $(k \in \mathbb{R})$ *converges weakly to a Gaussian process* $Z_{Se}(k)$ *with zero mean and covariance function given by*

$$Cov(Z_{Se}(k_1), Z_{Se}(k_2)) = E\big[\eta_{Se}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_1)\ \eta_{Se}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_2)\big],$$

and the process $n^{1/2}(\hat{Sp}(k) - Sp(k))$ $(k \in \mathbb{R})$ *converges weakly to a Gaussian process* $Z_{Sp}(k)$ *with zero mean and covariance function given by*

$$Cov(Z_{Sp}(k_1), Z_{Sp}(k_2)) = E\big[\eta_{Sp}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_1)\ \eta_{Sp}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_2)\big].$$

As a corollary to the above result we now state the limiting distribution of the estimator $\widehat{ROC}(u)$ defined in (4.13) and of the estimator $\widehat{AUC}_\delta$, given by

$$\widehat{AUC}_\delta = \int_\delta^{1-\delta} \widehat{ROC}(u)\, du.$$

For technical reasons we need to restrict the integration to the interval $[\delta, 1-\delta]$ (for some small $\delta > 0$), which can however be made arbitrarily close to the interval $[0, 1]$. The corresponding theoretical AUC is denoted by $AUC_\delta = \int_\delta^{1-\delta} ROC(u)\, du$.

Corollary 4.2.1. *Assume that conditions 1–4 in Lu (2008) are satisfied and that the logistic/Cox mixture cure model is valid. Assume in addition that* $\inf_{k_1 \leq k \leq k_2} Sp'(k) > 0$, *where* $Sp'(\cdot)$ *is the first derivative of Sp with respect to its argument,* $k_1 = (1 - Sp)^{-1}(\delta)$ *and* $k_2 = (1 - Sp)^{-1}(1 - \delta)$ *for some* $\delta > 0$, *and that the functions Se and Sp are twice continuously differentiable on* $[k_1, k_2]$. *Then,*

$$\widehat{ROC}(u) - ROC(u) = n^{-1} \sum_{i=1}^{n} \eta_{ROC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, u) + R_{n,ROC}(u),$$

where $\sup_{\delta<u<1-\delta} |R_{n,ROC}(u)| = o_P(n^{-1/2})$, *and*

$$\eta_{ROC}(\boldsymbol{x}, \boldsymbol{z}, y, \delta, u) = \eta_{Se}\big(\boldsymbol{x}, \boldsymbol{z}, y, \delta, (1-Sp)^{-1}(u)\big) + \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}}\eta_{Sp}\big(\boldsymbol{x}, \boldsymbol{z}, y, \delta, (1-Sp)^{-1}(u)\big),$$

where $Se'(\cdot)$ *and* $(1-Sp)'(\cdot)$ *are the first derivatives of* Se *and* $(1-Sp)$*, respectively, with respect to their argument. Moreover, the process* $n^{1/2}(\widehat{ROC}(u) - ROC(u))$ $(u \in [\delta, 1-\delta])$ *converges weakly to a Gaussian process* $Z_{ROC}(u)$ *with zero mean and covariance function given by*

$$Cov(Z_{ROC}(u_1), Z_{ROC}(u_2)) = E\big[\eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_1)\ \eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_2)\big],$$

and

$$n^{1/2}(\widehat{AUC}_\delta - AUC_\delta) \xrightarrow{d} N(0, \sigma^2_{AUC}),$$

where

$$\sigma^2_{AUC} = \int_\delta^{1-\delta}\int_\delta^{1-\delta} E(Z_{ROC}(u_1)Z_{ROC}(u_2))\, du_1\, du_2.$$

4.3 Finite Sample Performance

4.3.1 Some Preliminaries

In this section, an extensive simulation study is performed in order to evaluate the finite sample performance of the ROC curve estimator (4.13). Two versions of this estimator are considered:

> *LC* : assuming a LC cure model for W_0 and W_1. The LC cure model is estimated assuming the method proposed by Sy & Taylor (2000) based on the EM algorithm (see also Section 2.1.2 in Chapter 2),
>
> *SIC* : assuming a SIC cure model for W_0 and W_1, where the model is estimated according to the maximum likelihood approach described in Chapter 3 (see Section 3.1.3).

Both the case of known and unknown M are investigated, and for both of them, the following points are analysed. First, we are interested in the general performance of the proposed estimators of the sensitivity and the specificity. Particular interest lies in the effect of censoring and of an incorrect specification of the vector $\mathbf{Z}$. Then, other points include a misspecification of the model for S_u and a non-logistic model for the cure proportion p.

To assess the performance of our proposed method, we consider two infeasible competitors:

> *CSK* (Cure Status Known): corresponding to the ROC curve estimator that would be obtained if the cure status would be fully observed. Equations (4.3) and (4.4) give the estimators for the sensitivity and the specificity in that case. The objective here is to evaluate the effect of the imputation of the cure status described in Section 4.1.1,

TW (True Weights): corresponding to the estimator (4.7) combined with (4.10), based on the true values of p and S_u. This benchmark estimator allows to investigate the effect of estimating p and S_u by means of the LC or SIC cure model.

Two criteria are considered to compare the four estimators, namely, the L1 distance between the true and the estimated ROC curves and the AUC. The L1 distance is given by

$$L1 = V^{-1} \sum_{i=1}^{V} |\widehat{ROC}(u_i) - ROC(u_i)|,$$

where $\widehat{ROC}$ is one of the ROC curve estimates and ROC is the true ROC curve. It is computed over a grid of points $u_i = \frac{i}{100}$ for $i = 1, \ldots, V = 99$. For LC and SIC estimators, the AUC is given by

$$\widehat{AUC} = \frac{1}{\hat{N}_0 \hat{N}_1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big[\big\{ I(M_j > M_i) + 0.5 \times I(M_j = M_i) \big\} \hat{W}_{j1} \hat{W}_{i0} \Big].$$

For CSK and TW estimators, the formula is almost the same, but with different W_{i0}, W_{i1}, N_0 and N_1. For CSK, they are replaced by $\check{W}_{i0}$, $\check{W}_{i1}$, $\check{N}_0$ and $\check{N}_1$, while for TW, they are given by $\tilde{W}_{i0}$, $\tilde{W}_{i1}$, $\tilde{N}_0$ and $\tilde{N}_1$. Note that these formulas take into account possible ties in M with the added term $0.5 \times I(M_j = M_i)$.

4.3.2 Data Generating Process

Within this section, we assume that the data are generated from the mixture cure model (1.5). The data generating process is as follows.

1. First, the incidence is considered :

   (i) The uncure probability p is generated according to the model $p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t \boldsymbol{x})$, where $g(\cdot)$ is a link function. Primary interest lies in the logistic link function, that is, $g(a) = \exp(\gamma_0 + a)/\{1 + \exp(\gamma_0 + a)\}$, which gives the logistic regression model, but other link functions can also be assumed;

   (ii) The second step consists in generating, for given $\boldsymbol{x}$, the uncure status $(1 - D)$ from a Bernoulli distribution with parameter equal to $p(\boldsymbol{x})$.

2. Next, the latency is generated :

   (i) We consider two models for the survival function of the uncured observations. The first model is a Gompertz model with survival function $S(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t \boldsymbol{z})}$, where $S_0(t) = \exp[-\theta\alpha^{-1}\{\exp(\alpha t) - 1\}]$, $\theta = 0.5$ and $\alpha = 0.03$.
   The second model is an AFT model assuming a log-logistic distribution for T, with survival function $S_u(t|\boldsymbol{z}) = [1 + \lambda \{t/\exp(\beta^t \boldsymbol{z})\}^{\kappa}]^{-1}$ where $\lambda = 0.05$ and $\kappa = 2.5$. Note that the AFT model does not

respect the proportional hazards property contrarily to the Gompertz model. Since $\hat{W}_0$ and $\hat{W}_1$ are obtained from a mixture cure model assuming a Cox PH model for the latency, this allows us to verify whether a model misspecification of S_u affects the ROC curve estimate;

(ii) Next, we generate the censoring time from a uniform distribution on $[U_{min}, U_{max}]$ that is independent of T, $\mathbf{X}$ and $\mathbf{Z}$. We further truncate the survival times of the susceptible observations at $U_{max} - 1$ so that the support of C is larger than the support of T;

(iii) We finally generate the follow-up time $Y = \min(T, C)$ and the censoring indicator $\Delta = I(T \leq C)$.

4.3.3 Known Classifier

First, we consider the case where the classifier takes the form of a single variable or of a known one-dimensional score denoted by X. Note that when $\dim(\mathbf{X}) = 1$, the single-index model reduces to a non-parametric model. We assume three different scenarios for the incidence. The first two scenarios assume a logistic regression model for $p(x)$ corresponding to different discriminations between the cured and the uncured sub-populations:

Scenario 1: $X \sim N(2, 2.5)$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $AUC = 0.9016$. This scenario corresponds to a good discrimination with a cure proportion equal to 25.6%.

Scenario 2: $X \sim N(1.2, 1)$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $AUC = 0.7374$. This scenario is associated with a moderate separation between the two sub-populations, and the proportion of cured subjects is equal to 26.9%.

The third scenario assumes a non-logistic model for $p(x)$ with a non-monotone shape in order to evaluate the performance of the LC and SIC estimators in such a case. The link function is given by $g(a) = [\sin\{(3/2)\ \pi\ a\} + 1]/2$. Its characteristics are as follows:

Scenario 3: $X \sim \text{Unif}(0, 1)$, $\gamma_1 = 1$, and $AUC = 0.8124$, corresponding to a good separation between cured and uncured sub-populations. The cure proportion equals 39.4%.

The graphical representation of the respective ROC curves is given in Figure 4.1.

For the survival times, the Gompertz model and the AFT model have the following characteristics.

Gompertz model: We consider two covariates, Z_1 and Z_2, that are independent, following a Bernoulli distribution with parameter equal to 0.6 and 0.2, respectively. The associated vector of parameters is $\boldsymbol{\beta} = (1.5, -0.5)^t$. For the uniform distribution considered for the censoring time C, we assume that $U_{min} = 0$ and three different values are considered for U_{max}: 65, 25 and 10, corresponding to three different levels of censoring denoted by level 1, level 2 and level 3.

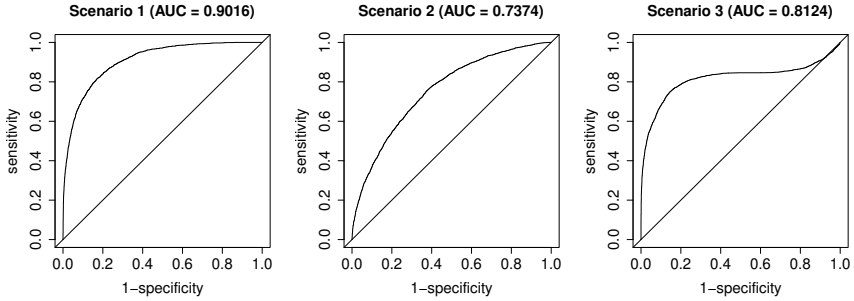

Figure 4.1: *True ROC curves for Scenarios 1, 2 and 3.*

AFT model: Two independent covariates, Z_1 and Z_2, are considered, following a Bernoulli distribution with parameter equal to 0.6 and 0.3, respectively. The associated vector of parameters is $\boldsymbol{\beta} = (0.7, -0.3)^t$. As for the Gompertz model, the censoring time is generated from a uniform distribution with $U_{min} = 0$ and with three different values for U_{max}. These values are chosen such that the proportion of censored observations with a follow-up time lower than or equal to τ is the same as for the Gompertz model, in order to allow comparison between the two models.

We now consider the following five settings, corresponding to the three scenarios for the incidence and the two models for the latency considered above (note that the non-logistic scenario for the incidence is only considered in combination with the Gompertz model, since it serves to assess the performance of the LC estimator when the logistic model is not satisfied). Each setting has a particular objective:

- Scenario 1/Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is good;
- Scenario 1/AFT model – to verify whether a model misspecification of S_u affects the ROC curve estimate when the discrimination is good;
- Scenario 2/Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is moderate;
- Scenario 2/AFT model – to verify whether a model misspecification of S_u affects the ROC curve estimate when the discrimination is moderate;
- Scenario 3/Gompertz model – to evaluate our proposal when the link function is not logistic. Note that for this latter setting, a fourth censoring rate is considered in order to further evaluate its impact in such a case.

To assess the effect of a misspecification of $\mathbf{Z}$, the estimators are further estimated assuming that $Z = X$, on Scenario 1/Gompertz and Scenario 2/Gompertz. Table 4.1 summarises the setting characteristics, comprising parameter

Table 4.1: *Setting characteristics for Scenarios 1 to 3: parameters of the censoring distribution, cure rate, censoring rate and percentage of censored observations for which* $Y \leq \tau$.

	latency type							
	Gompertz model				*AFT model*			
incid. type	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$
Scen. 1	65	25.6%	27.0%	5.3%	345	25.6%	27.0%	5.3%
	25	25.6%	29.1%	12.6%	121	25.6%	29.5%	12.6%
	10	25.6%	34.0%	25.4%	44	25.6%	35.9%	25.5%
Scen. 2	65	26.9%	28.2%	5.3%	360	26.9%	28.2%	5.3%
	25	26.9%	30.3%	13.0%	120	26.9%	30.8%	12.9%
	10	26.9%	35.1%	25.9%	45	26.9%	37.1%	26.1%
Scen. 3	65	39.4%	40.5%	6.8%				
	25	39.4%	42.2%	16.2%				
	10	39.4%	46.3%	31.4%				
	5	39.4%	51.8%	43.1%				

values for the censoring distributions, cure rates, censoring rates, and the percentage of observations for which $Y \leq \tau$.

For each setting, we consider 500 datasets, and for each dataset, we assume two sample sizes, $n = 250$ and $n = 500$.

Figure 4.2 shows the boxplots of the L1 distance when $n = 250$ for the settings with Scenarios 1 and 2 for the incidence. As it can be seen, when the censoring rate is close to the cure rate, and when everything is specified correctly, our proposals perform almost as well as the two infeasible competitors whatever the model assumed for W_1. In such a case, very few censored observations are below τ, which are those with weight equal to $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$. A larger censoring rate is conversely associated with higher L1 distance and larger variance, particularly for SIC under the third censoring level. When the censoring rate gets larger, fewer censored observations are located in the plateau, meaning that less censored observations are considered as cured, that is, with $W_{i1} = 1$. Furthermore, as shown in Chapter 3, the SIC cure model performs worse than the LC cure model when the true model for the incidence is a logistic model and when the censoring rate increases as it is the case for the third censoring level. Interestingly, LC is close to CSK even for the third censoring level. Another interesting point is that the L1 distance for TW decreases slightly when the censoring rate increases. It seems that having more censored observations below τ produces better results when considering the true W_0 and W_1. In such a case, the size of the jumps are smaller and it seems that the ROC curve becomes 'smoother' and closer to the true curve. Nevertheless, this feature is not observed for LC and SIC. By comparing Scenario 1 and Scenario 2, we observe that the L1 distance is larger for Scenario 2. Indeed, it is more difficult to correctly separate cured from uncured sub-populations based on this scenario since the discrimination is moderate. The discrimination between cured and uncured sub-populations seems therefore to have an influence on the performance of the ROC curve estimators. However, the general conclusions are the same for both scenarios.

For the settings where $\mathbf{Z}$ is misspecified, we observe that when few censored observations are below τ, the L1 distance for our proposals is only slightly higher in comparison with the two infeasible competitors, while when the number of censored observations below τ is larger as for the third censoring level, both the LC and SIC estimators have higher L1 distance than when $\mathbf{Z}$ is correctly specified. Interestingly, for the second censoring level, the L1 distance for LC seems to be comparable to the L1 distance of CSK for both scenarios while SIC seems to already present some difficulties. Note that we consider the case where $\mathbf{Z}$ is completely misspecified, whereas it seems more likely to have only a partial misspecification of this vector of covariates in practical applications. We are therefore in an extreme case.

When the survival times are generated according to an AFT model, our proposals show a higher increase in the L1 distance in comparison with when there is no misspecification, especially when the censoring rate increases. SIC is still the least favourable estimator. However, a misspecification in $\mathbf{Z}$ affects the performance of our proposals more than a misspecification in the latency.

Figure 4.3 provides the boxplots of the AUC for Scenarios 1 and 2 when

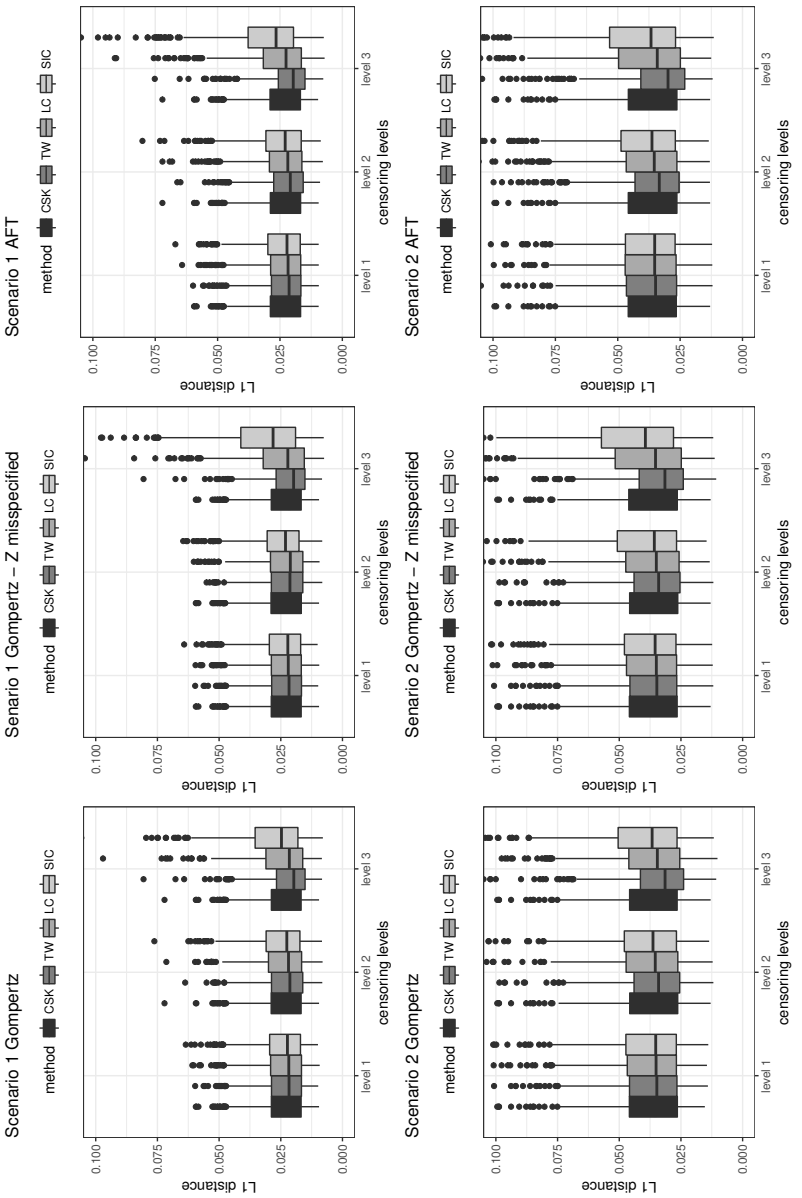

Figure 4.2: *Boxplots of the L1 distances for Scenarios 1 and 2 for* $n = 250$.

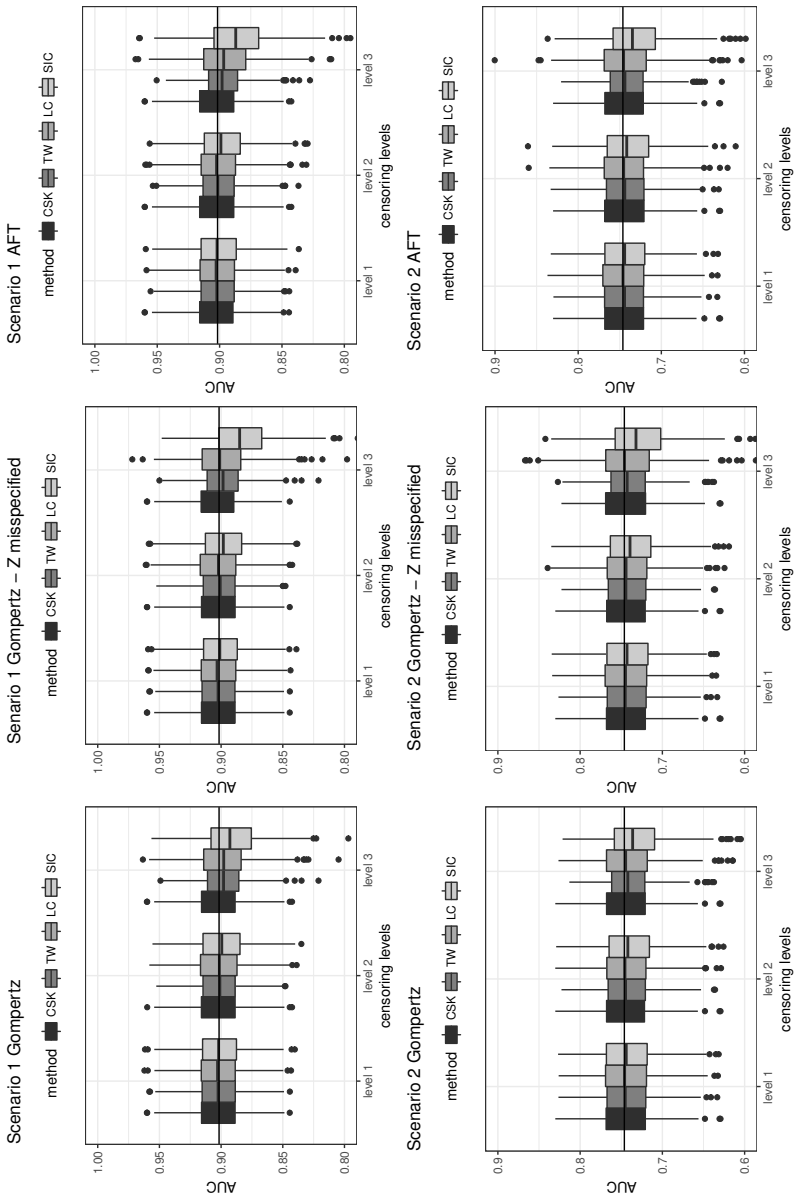

Figure 4.3: *Boxplots of the AUC for Scenarios 1 and 2 for* $n = 250$.

$n = 250$. The same conclusions as for the L1 distances can be drawn. Note that SIC performs less good than LC, especially for the third censoring rate. LC on the contrary is close to CSK for all censoring rates considered, but we observe more variability. For all these settings, the true incidence is a logistic regression model.

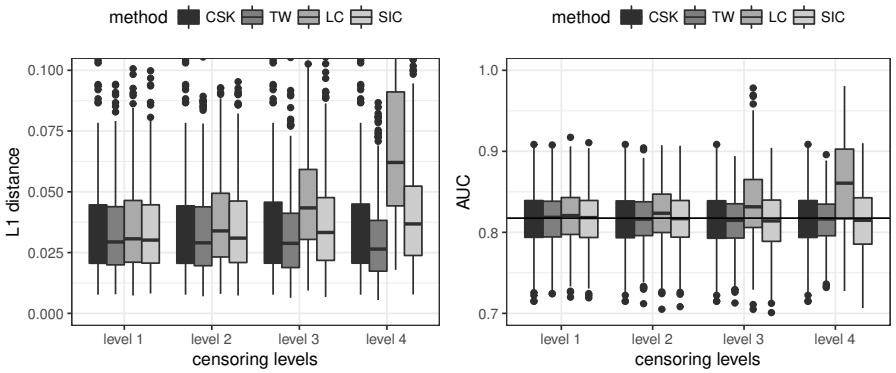

Figure 4.4: *Boxplots of the L1 distance and the AUC for Scenario 3 Gompertz for* $n = 250$.

When the true incidence is not a logistic regression model (Scenario 3), LC has always higher L1 distances than SIC as it can be seen in Figure 4.4. When the censoring rate gets larger, the difference between LC and SIC increases, and we observe that SIC outperforms LC, especially for the third and the fourth censoring levels. It seems therefore that, when $\hat{p}(\boldsymbol{x})$ is inconsistent and when the proportion of censored observations below τ is large, LC performs badly. Based on the analysis of the AUC, the same conclusions can be drawn.

An interesting feature that was already observable from the previous settings and which is confirmed with the fourth censoring level here is that the more the censoring rate increases, the more the L1 distance of LC and SIC increases since many more observations have $\hat{W}_0$ and $\hat{W}_1$ relying on $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$ in that case. Therefore, the censoring level is crucial in the performance of our proposals.

For both the L1 distance and the AUC, the same conclusions also apply when $n = 500$ (see Figures 4.12, 4.13 and 4.14 in Appendix 1 at the end of this chapter). Furthermore, as expected, the L1 distances are smaller and less variable than for $n = 250$. For the AUC, the variability is also lower.

4.3.4 Unknown Classifier

We next consider the case where the classifier is an unknown combination of variables. Two configurations are investigated: when the variables are independent and when there exists some correlation between some of the variables.

Independent Variables

Two scenarios are considered for the incidence, both of them assuming a logistic regression model and corresponding to two different levels of discrimination between the two sub-populations. For both of them, we assume three independent variables. Each setting has the following characteristics:

Scenario 4: $X_1 \sim N(0,1)$, $X_2 \sim Bernoulli(0.6)$ and $X_3 \sim Bernoulli(0.3)$. We take $\gamma_0 = 1$ and $\boldsymbol{\gamma} = (2, 3, -2)^t$. The cure rate is equal to 24.5% and $AUC = 0.9080$, corresponding to a good discrimination between cured ans uncured subjects.

Scenario 5: $X_1 \sim N(0.6,1)$, $X_2 \sim Bernoulli(0.5)$ and $X_3 \sim Bernoulli(0.4)$. We take $\gamma_0 = -1.7$ and $\boldsymbol{\gamma} = (0.7, 1.1, 0.3)^t$. The cure rate is equal to 62.2% and $AUC = 0.7219$. With this scenario the discrimination between the two sub-populations is moderate. These characteristics have been chosen such that the scenario mimics the characteristics of the melanoma data on which our methodology is illustrated (see Section 4.4).

The graphical representation of the true ROC curves for these two scenarios is given in Figure 4.5.

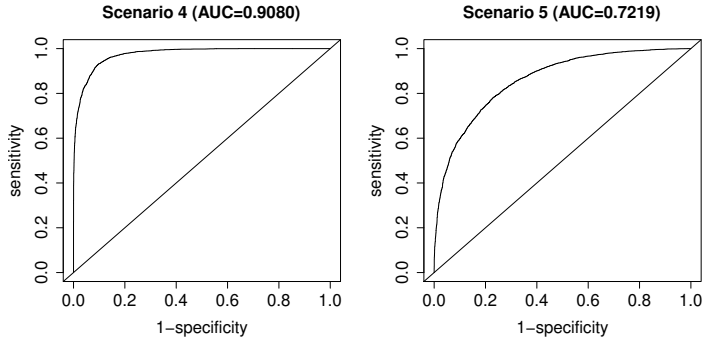

Figure 4.5: *True ROC curves for Scenarios 4 and 5.*

For the latency, we consider the same model as when the classifier is known, that is, a Gompertz model or an AFT model, with the same characteristics. We consider the four following settings corresponding to all combinations of the two scenarios for the incidence and the two models for the latency. As for the known classifier case, each setting has the following objective:

- Scenario 4/Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is good;
- Scenario 4/AFT model – to verify whether a model misspecification in S_u affects the ROC curve estimate when the discrimination is good;
- Scenario 5/Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is moderate;

- Scenario 5/AFT model – to verify whether a model misspecification in S_u affects the ROC curve estimate when the discrimination is moderate;

To assess the effect of a misspecification of $\mathbf{Z}$, the estimators are further estimated assuming that $\mathbf{Z} = \mathbf{X}$ in Scenario 4/Gompertz and Scenario 5/Gompertz. Note also that a fourth censoring rate is considered when Scenario 5 is assumed for the incidence in order to evaluate the situation when many observations are censored before τ as it is the case for the melanoma data set. Table 4.2 summarises the setting characteristics, comprising parameter values for the censoring distributions, cure rates, censoring rates, and the percentage of observations below τ.

Since the classifier is unknown, it first needs to be estimated before the ROC curve can be computed. We consider a LC cure model, M being estimated by the score $\hat{\gamma}_0+\hat{\boldsymbol{\gamma}}^t\mathbf{X}$ from the logistic model considered in the incidence. However, a difficulty remains. Indeed, if M and the ROC curve are estimated on the same dataset, the ROC curve can overestimate the classification performance of M as explained by Copas & Corbett (2002), and hence this would lead to misleading conclusion(s). It is therefore necessary to obtain generalizable conclusions about the performance of M. To do so, an ideal approach would consist in splitting the dataset into two groups, a training set on which the model is fitted and a test set on which predictions are made and then used to build the ROC curve. However, to split a dataset, a large sample size is required. When this is not the case, the use of other approaches exist as explained by Hastie et al. (2009), among which *cross-validation*. To mimic real data settings, the finite sample performance of our proposal is therefore evaluated as if the sample size was not large enough to be split into two groups. We instead perform it using cross-validation. We proceed as follows. First, the initial dataset is split into K folds. Each fold is then considered as the test set successively, the other folds being considered as the training set. For example, if $K = 5$, and denote by $F1$, $F2$, $F3$, $F4$, and $F5$ the five folds, the cross-validation procedure will produce five different runs with the following composition for the test and training sets :

run	test set	training set
1	$F1$	$F2 \cup F3 \cup F4 \cup F5$
2	$F2$	$F1 \cup F3 \cup F4 \cup F5$
3	$F3$	$F1 \cup F2 \cup F4 \cup F5$
4	$F4$	$F1 \cup F2 \cup F3 \cup F5$
5	$F5$	$F1 \cup F2 \cup F3 \cup F4$

For each run, a LC cure model is fitted on the training set in order to estimate γ_0 and $\boldsymbol{\gamma}$. Then, predictions for M are performed on the test set and the ROC curves are estimated. Furthermore, the L1 distance and the AUC are computed for each ROC curve at each run. At the end of the K runs, the L1 distances and AUCs are averaged over the K folds. In what follows, we take $K = 5$. As for the known classifier case, 500 datasets are generated for each setting, and two sample sizes are considered, $n = 250$ and $n = 500$.

Table 4.2: *Setting characteristics for Scenarios 4 and 5: parameters of the censoring distribution, cure rate, censoring rate and percentage of censored observations for which* $Y \leq \tau$.

	latency type							
	Gompertz model				*AFT model*			
incid. type	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$
Scen. 4	65	24.5%	25.9%	4.9%	370	24.5%	25.9%	4.9%
	25	24.5%	28.0%	12.1%	126	24.5%	28.5%	12.1%
	10	24.5%	33.2%	24.5%	47	24.5%	34.9%	24.4%
Scen. 5	65	62.2%	63.3%	8.7%	337	62.2%	63.4%	8.7%
	25	62.2%	64.4%	21.3%	111	62.2%	64.8%	21.3%
	10	62.2%	66.9%	40.4%	40	62.2%	68.6%	40.5%
	5	62.2%	70.3%	54.4%	23	62.2%	72.6%	54.4%

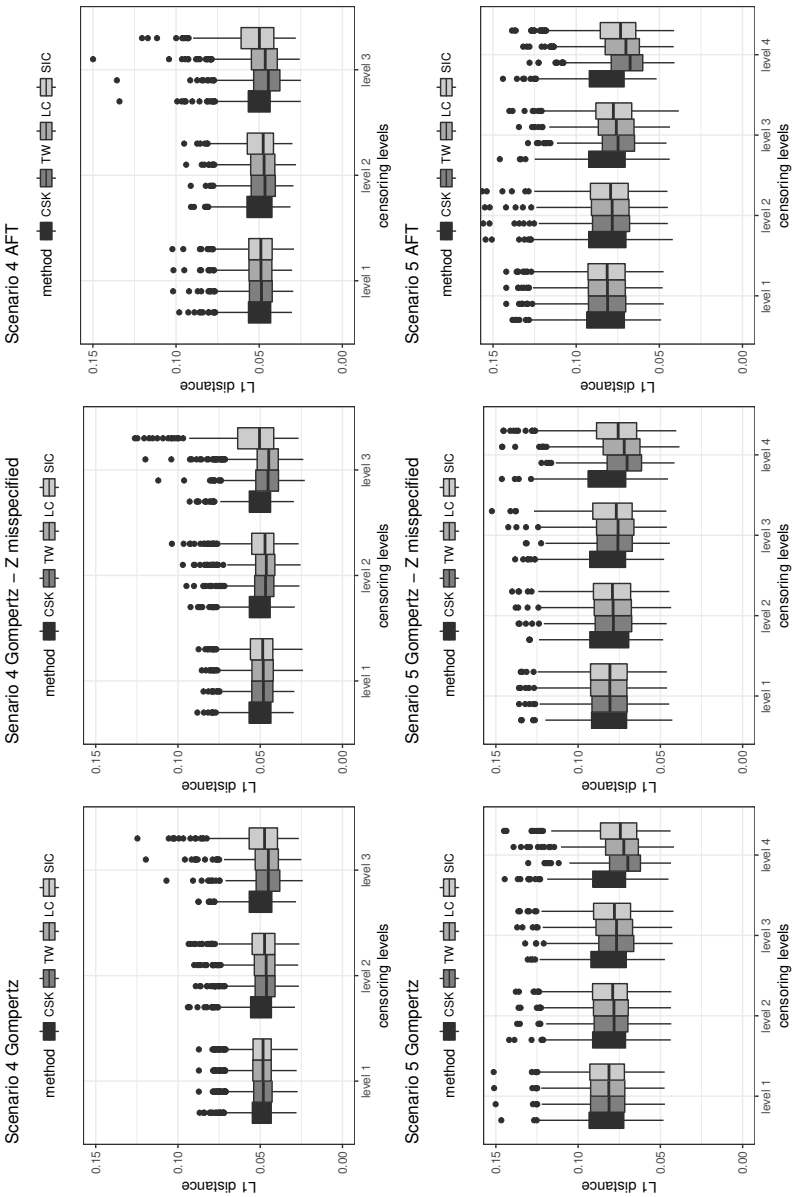

Figure 4.6: *Boxplots of the L1 distances for Scenarios 4 and 5 for* $n = 250$.

The boxplots of the L1 distance for all settings and for $n = 250$ are given in Figure 4.6. As it can be seen, when there is no misspecification on $\mathbf{Z}$ or on the latency, the conclusions are almost the same as when the classifier is known. We observe only slight differences between the four estimators when few observations are censored before τ, while when the censoring rate gets larger, LC and SIC present slightly larger L1 distance than TW, particularly for the third and the fourth (only for Scenario 5) censoring level. We also notice that, as for the known classifier case, SIC presents higher L1 distance than LC in such a case. However, we also observe that compared to CSK, our proposals have ROC curves that are closer to the true one when the number of censored observations for which $Y \leq \tau$ increases. Since the ROC curves are computed based on cross-validation, only 50 observations are considered to build the ROC curves on each run of the cross-validation. Furthermore, as already mentioned for the known classifier case, when the censoring rate increases, more observations have a weight equal to $P(D = 1|\mathbf{X}, \mathbf{Z}, C, T > C)$ and it follows that our proposed method gives a smoother curve which is closer to the true one. Since the sample size is small the ROC curves have fewer jumps than previously and it seems that, for the unknown classifier case, not only TW presents better results in such a case, but also LC and SIC. As before, the L1 distances are large for Scenario 5 compared to Scenario 4. This is due to the lower discrimination in Scenario 5.

When $\mathbf{Z}$ is misspecified, LC performs almost as well as TW and the results are comparable to the case where there is no misspecification and so for all censoring levels. Conversely, when the number of censored observations before τ increases, SIC presents slightly higher L1 distance compared to the situation where there is no misspecification. When the latency is misspecified, the same conclusions can be drawn. We only observe slight differences with the case where there is no misspecification. When $n = 500$, the conclusions stay the same as it can be seen in Figure 4.15 given in Appendix 1 even if we notice lower L1 distances, as expected.

Figure 4.7 (and Figure 4.16 in Appendix 1 at the end of this chapter) contains the boxplots of the AUC for $n = 250$ (for $n = 500$). The same conclusions as for the L1 distance can be drawn. As for the known classifier, our proposals perform well in comparison with the two infeasible competitors, even when the censoring rate increases or when the weights are not correctly specified. As for the L1 distance, both LC and SIC present some difficulties when many censored observations are such that $Y \leq \tau$. SIC is still the one performing the worst in such a case.

Dependent Variables

To further investigate the finite sample performance of our proposal in the case where there is a correlation between variables, we consider a *Scenario 6* for the incidence where two of the three covariates are generated from a bivariate normal distribution. The characteristics are as follows:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1.5 \end{pmatrix} \right),$$

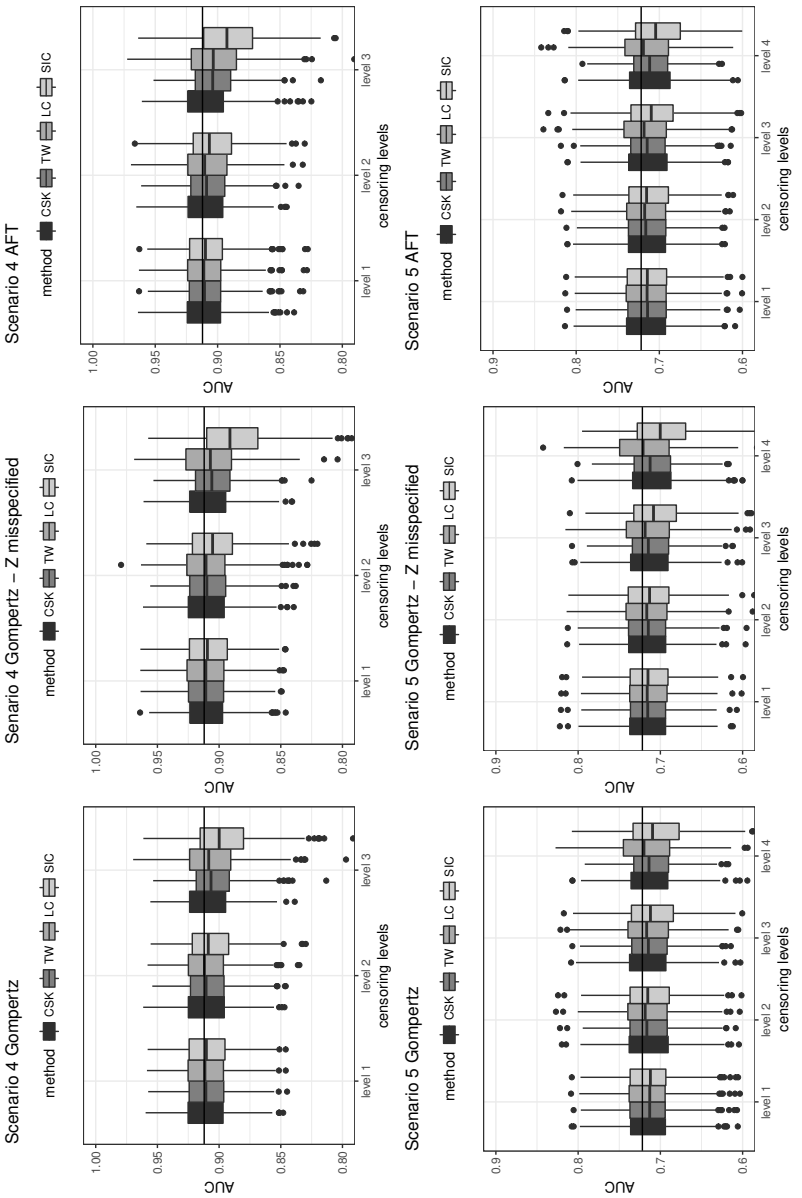

Figure 4.7: *Boxplots of the AUC for Scenarios 4 and 5 for* $n = 250$.

where σ_{12} represents the covariance between X_1 and X_2, and we suppose that $X_3 \sim Bernoulli(0.3)$. Different values for σ_{12} are assumed in order to assess the impact of different degrees of dependency between the covariates on the ROC curve estimates. We assume that σ_{12} can be equal to 0, which serves as benchmark, $0.5, 1$, and 1.2. For the parameters in M, we take $\gamma_0 = 1$ and $\boldsymbol{\gamma} = (-1.3, 0.5, -2)^t$. Each value of σ_{12} corresponds to a different sub-scenario denoted by Scenario 6(a), 6(b), 6(c) and 6(d), respectively. For each of them, the cure rate is equal to 35.0%, 34.2%, 33.1% and 32.6%, respectively, and the AUC is equal to 0.8333, 0.8140, 0.7899 and 0.7794, respectively. Figure 4.8 represents the true ROC curve for each of the four sub-scenarios as well as the correlation between X_1 and X_2. By combining all these informations, we observe that when the strength of the dependency increases, the cure rate decreases as well as the discrimination between cured and uncured subjects (the ROC curves get closer to the bisector) resulting in smaller AUCs.

Table 4.3: *Setting characteristics for Scenario 6: parameters of the censoring distribution, cure rate, censoring rate and percentage of censored observations for which* $Y \leq \tau$.

incid. *type*	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$
Scen. 6(a)	65	35.0%	36.1%	6.2%
	25	35.0%	37.9%	15.3%
	10	35.0%	42.4%	30.0%
Scen. 6(b)	65	34.2%	35.3%	6.1%
	25	34.2%	37.1%	15.0%
	10	34.2%	41.7%	29.6%
Scen. 6(c)	65	33.1%	34.2%	6.0%
	25	33.1%	30.1%	14.7%
	10	33.1%	40.7%	29.1%
Scen. 6(d)	65	32.6%	33.8%	5.9%
	25	32.6%	35.7%	14.5%
	10	32.6%	40.3%	29.0%

For the latency, we only consider the Gompertz model presented above, with the same characteristics. Associated with the four sub-scenarios considered for the incidence, it results in a total of four settings. The parameter values for the censoring distributions, the censoring rates, as well as the percentage of observations with a follow-up time below τ for each of them are given in Table 4.3. Furthermore, we only consider the case where $\mathbf{Z}$ is well specified.

The cross-validation procedure presented in the previous section is considered to compute the ROC curves. 500 datasets are generated for each setting

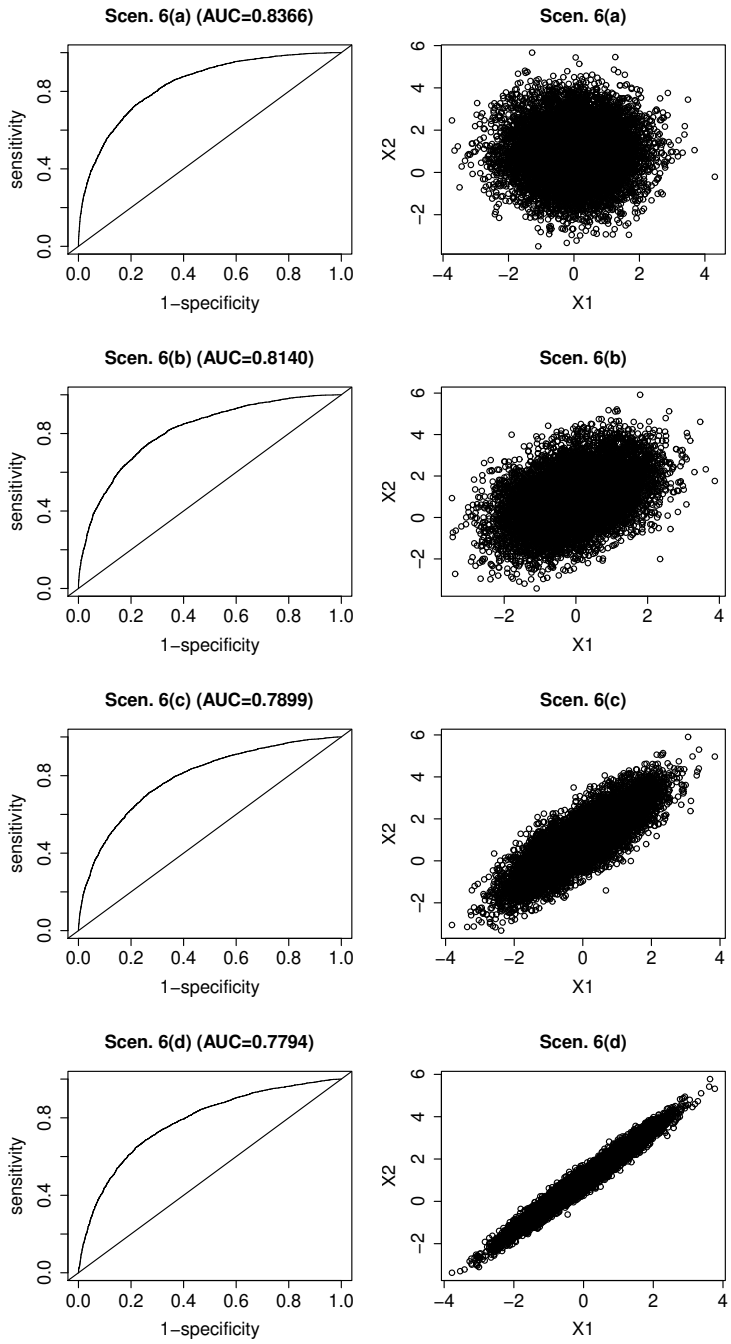

Figure 4.8: *True ROC curves and correlation between X_1 and X_2 for Scenarios 6(a) to 6(d) for $n = 250$.*

and we only perform the simulations for $n = 250$.

Figure 4.9 represents the boxplots of the L1 distance and the AUC for the four sub-scenarios. In terms of general comparison of the methods, the same conclusions as above can be drawn whatever the degree of correlation. We observe that our proposals perform as well as TW and CSK when few observations are censored before τ. When the censoring rate increases, we observe larger differences between our proposals and TW, but we also notice that CSK performs poorly comparatively with the three other methods, as it has already been observed when the three covariates are generated independently. By specifically considering the impact of the degree of correlation, we note some difference between the settings. Indeed, when the correlation increases, the L1 distance increases as well. This phenomenon was expected since the discrimination between cured and uncured populations tends to decrease when the strength of the association between X_1 and X_2 is getting larger. However, we also observe that SIC seems to perform almost as well as LC for the third level of censoring for Scenarios 6 (c) and (d), situation which was not observed for the independent case. Nevertheless, it seems that having correlated covariates does not have an important impact on ROC curve estimates, whatever the method considered. A possible explanation comes from the fact that it is the order of the values of M which is considered to build the ROC curve and not the values themselves. Therefore, if there are some problems of collinearity but if the order of the observations stays the same, the ROC curve still performs the same.

4.4 Application

In order to illustrate our methodology on real data, we use a melanoma data set coming from the textbook by Andersen et al. (1993). This data set consists of 205 patients diagnosed with malignant melanoma (skin cancer) during the period 1962–1977, and who experienced a radical operation (complete removal of the tumour together with the skin within a distance of about 2.5cm around it) performed at the Plastic Surgery department of the University Hospital of Odense in Denmark. All patients have been followed since surgery and until the end of the year 1977. The endpoint of interest is the death from melanoma.

Among the 205 patients, 57 died from the melanoma. Figure 4.10 represents the Kaplan-Meier estimator of the survival function. As it can be seen, it levels off at around 65% and there is a long plateau of 2227 days (approximately 6 years), which contains 23% of the censored observations. These two elements are indicative of the presence of a cure fraction alongside the contextual evidence. Indeed, melanoma belongs to the group of cancers for which it is known that some patients get cured. It seems therefore that there exists a cure fraction for patients suffering from melanoma and that this data set contains such subjects.

Alongside the survival time, three covariates are available, the thickness of the tumour (in millimetres (mm), ranges from 0.10 to 17.47 mm with a median value of 1.94mm), a binary variable indicating whether the tumour was

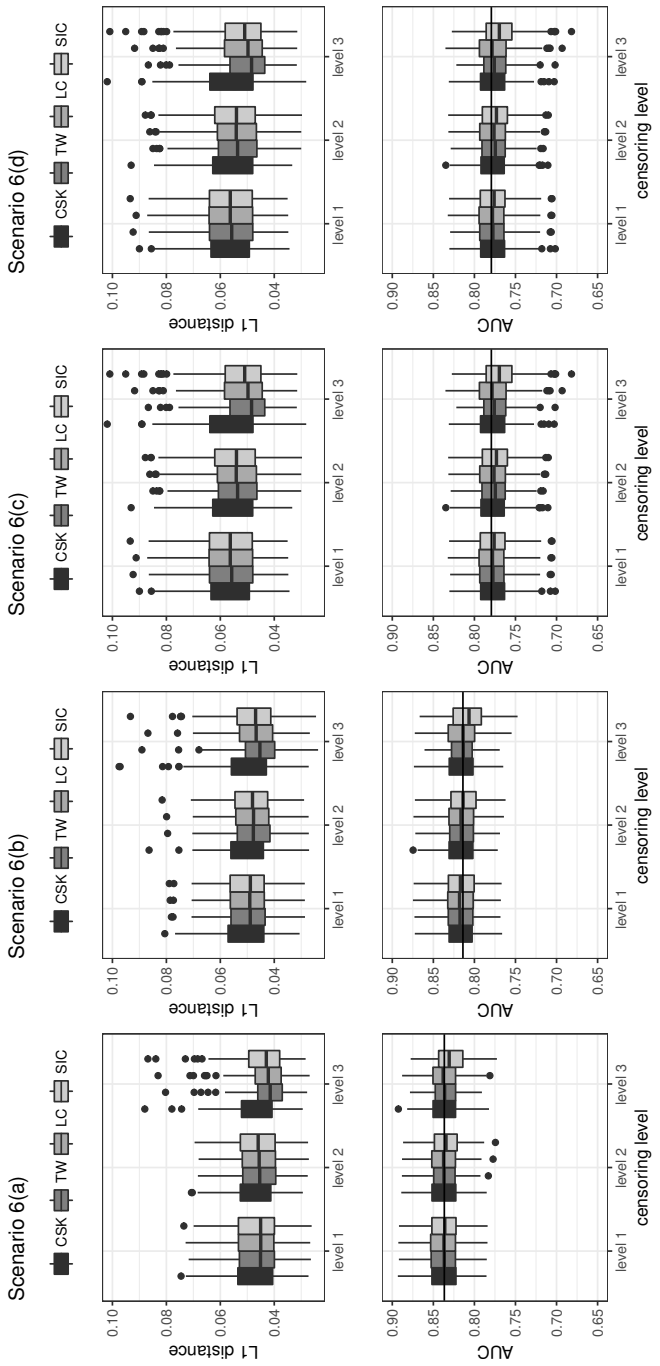

Figure 4.9: *Boxplots of the L1 distances and the AUC for Scenarios 6(a) to 6(d) for* $n = 250$.

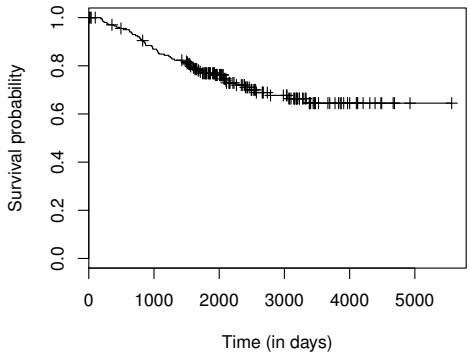

Figure 4.10: *Kaplan-Meier estimator of the survival function for the melanoma dataset.*

ulcerated or not (0 = absence - 115 patients, 1 = presence - 90 patients) and the gender of the patients (0 = female - 126 patients, 1 = male - 79 patients).

Our objective is to assess if these three variables are good predictors of the cure status. Before computing the ROC curve, we first estimate a LC mixture cure model with the three covariates included in both parts of the model. The results are given in Table 4.4. Note that we fit the model considering the logarithm of the tumour thickness. As it can be seen, by considering a 5% level, both the ulceration and the thickness of the tumour significantly affect the cure probability. An absence of ulceration and a thinner tumour increase the cure probability.

Since gender is not significant, we then compute the ROC curve to assess the predictive performance of the tumor thickness and the ulceration only according to the cross-validation procedure described in Section 4.3. We consider a LC cure model to estimate the classifier and we compute the ROC based on the estimator (4.13), assuming both a LC cure model and a SIC cure model for W_0 and W_1. Figure 4.11 (a) provides the graphical representation of the two curves corresponding to the mean ROC curves over the five runs of the cross-validation. Their respective AUC are 0.7412 and 0.6912, showing that the ROC curve based on the SIC cure model is lower than the ROC curve obtained from the LC cure model. In fact, the proportion of observations censored before τ represents 55.6% of the total number of observations. Furthermore, as it can be seen from Figure 4.11 (b)-(c), the estimated link function for the SIC cure model (Figure 4.11(c)) is monotone and close to the logistic one (Figure 4.11(b)), meaning that a logistic model for the incidence is appropriate. As it has been demonstrated in Section 4.3, when there is a substantial number of observations censored before τ and when the true link function is a logistic one, the perfomance of the ROC curve based on a SIC cure model is less good than that of the ROC curve based on a LC cure model. Therefore, it seems that the

Table 4.4: *Parameter estimates for the melanoma data set based on the LC cure model.*

	incidence			*latency*		
Parameters	estimate	std error	p-value	estimate	std error	p-value
intercept	-1.7759	0.3079	< 0.0001	-	-	-
ulceration[pres. vs. abs.]	1.0493	0.4261	0.0138	0.2658	0.4132	0.5201
log(tum. thick.)	0.5329	0.2341	0.0282	0.6703	0.2751	0.0148
gender[male vs. female]	0.2971	0.3845	0.4397	0.6215	0.4035	0.1235

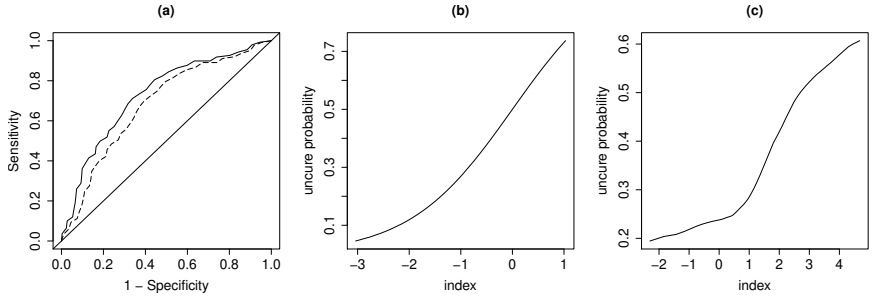

Figure 4.11: *(a) ROC curve estimates - solid curve: LC cure model, dashed curve: SIC cure model - (b) Estimated link function for the LC cure model (index = $\hat{\gamma}^t \boldsymbol{x}$) - (c) Estimated link function for the SIC cure model (index = $\hat{\gamma}^t \boldsymbol{x}$).*

former one has some more difficulties to correctly evaluate the discrimination ability of the classifier considered here. Nevertheless, the two ROC curves are quite close and they indicate that the tumour thickness and the tumour ulceration only moderately discriminate cured from uncured patients.

4.5 Concluding Remarks

In this chapter we proposed a method to assess cure status prediction from survival data using ROC curves. Based on the definition of conditional probability, we derived estimators for the sensitivity and the specificity taking the form of weighted empirical distribution functions. We proposed to estimate the weights based on the so-called mixture cure model, assuming both a LC and a SIC cure model. We further developed an estimator of the area under the curve, and we derived the asymptotic properties of the proposed estimators. Through an extensive simulation study we showed that our proposal performs well when the censoring rate is reasonably high and when not too many censored observations are below τ, both when the classifier is known and unknown. When many censored observations have a follow-up time lower than τ, however, our proposal shows some difficulties when a SIC cure model is considered to compute the weights and when the true model for the incidence is a logistic regression model. In such a case, assuming a LC cure model provides more accurate results compared to the infeasible competitors. We further investigated the effect of a misspecification of the weights, both at the covariate and at the modelling levels. We have seen that the performance of our proposal is only slightly affected and that when the proportion of censored observations below τ is not too large, the LC particularly still performs very well. In summary, censoring is the element affecting the most the performance of our proposal and we therefore recommend to be cautious when the censoring rate is high and when many observations are censored before τ. We further recommend to check the model in the incidence since, as we have seen, when the true link

function is not logistic, the LC cure model can provide bad results when there are many censored observations below τ.

Throughout this chapter, we supposed that M is a linear combination of variables. However, it is possible to extend our proposal to the case where the classifier would be obtained from a different model. Further investigation would be necessary to assess the impact of such a situation on the computation of the weights, but our proposal is not restricted to the linear case. Furthermore, we have considered mixture cure models to compute $\hat{W}_0$ and $\hat{W}_1$. However, a promotion time cure model, such as the model proposed by Tsodikov (1998a), could also be considered to estimate a ROC curve for the cure status prediction from survival data.

4.6 Appendix 1: Boxplots of the L1 Distance and the AUC for all Settings when $n = 500$

This appendix contains the boxplots of the L1 distance and the AUC for all settings when $n = 500$.

Known Classifier (See next pages)

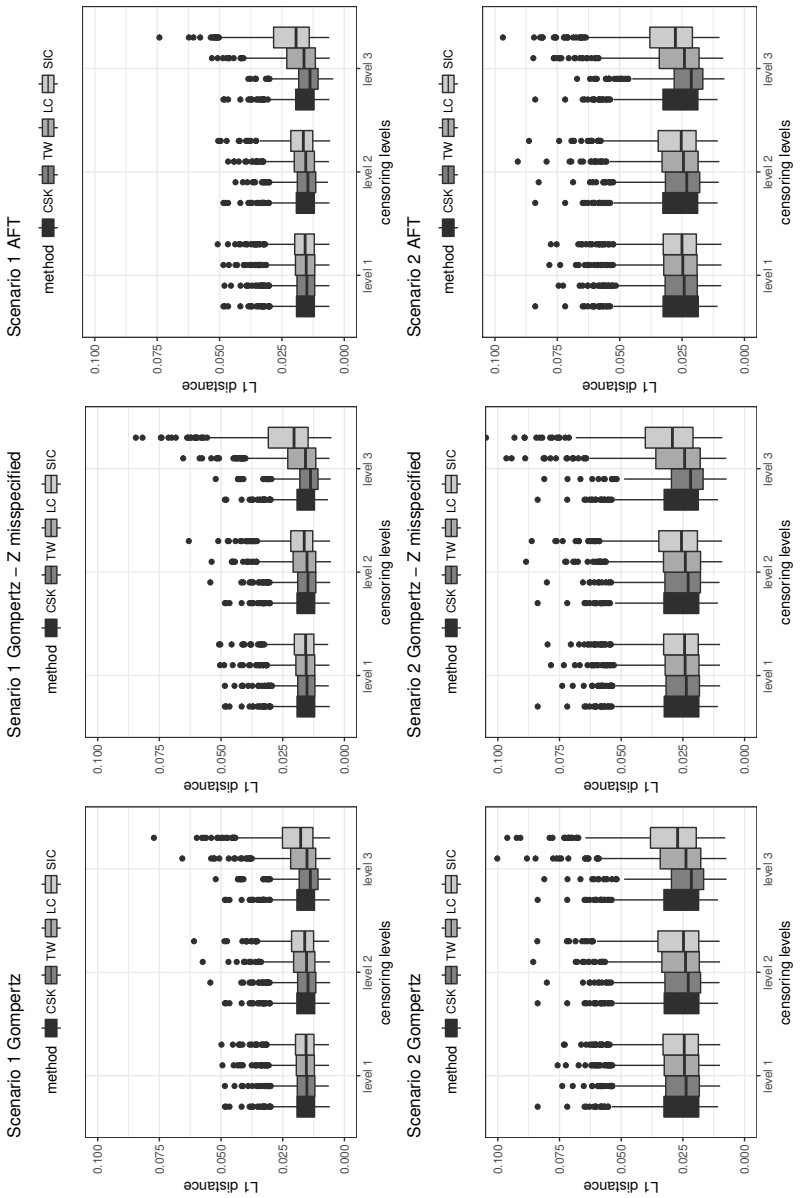

Figure 4.12: *Boxplots of the L1 distances for Scenarios 1 and 2 for* $n = 500$.

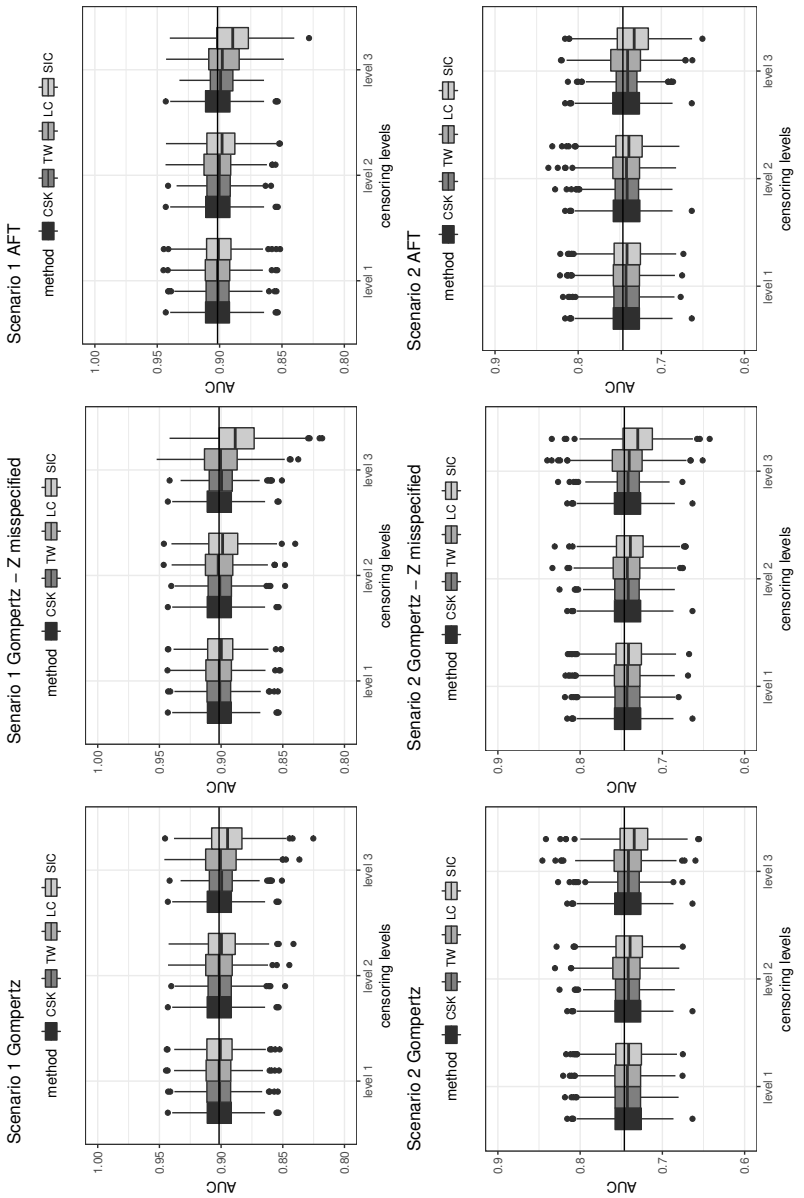

Figure 4.13: *Boxplots of the AUC for Scenarios 1 and 2 for* $n = 500$.

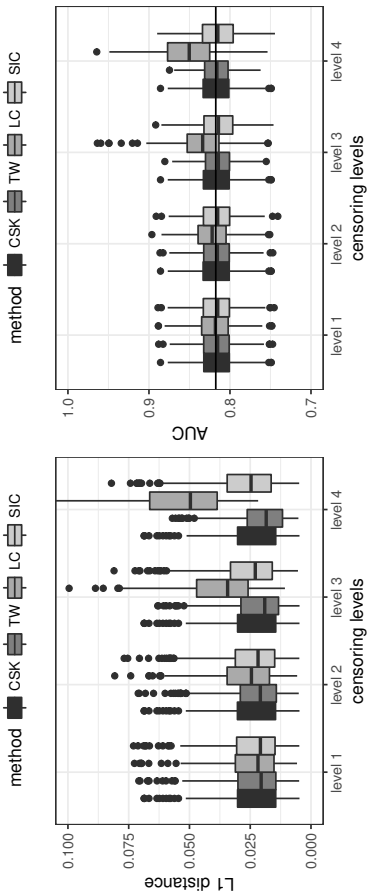

Figure 4.14: *Boxplots of the L1 distance and the AUC for Scenario 3 Gompertz for* $n = 500$.

Unknown Classifier

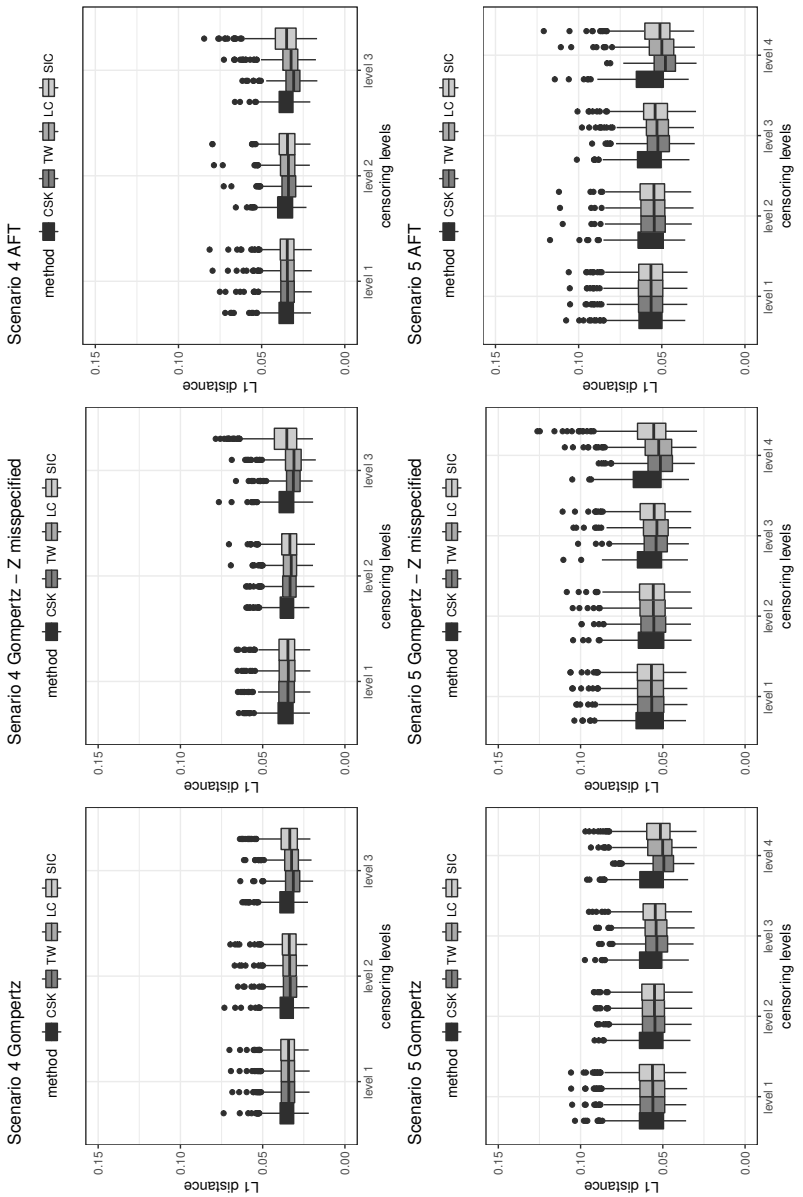

Figure 4.15: *Boxplots of the L1 distances for Scenarios 4 and 5 for* $n = 500$.

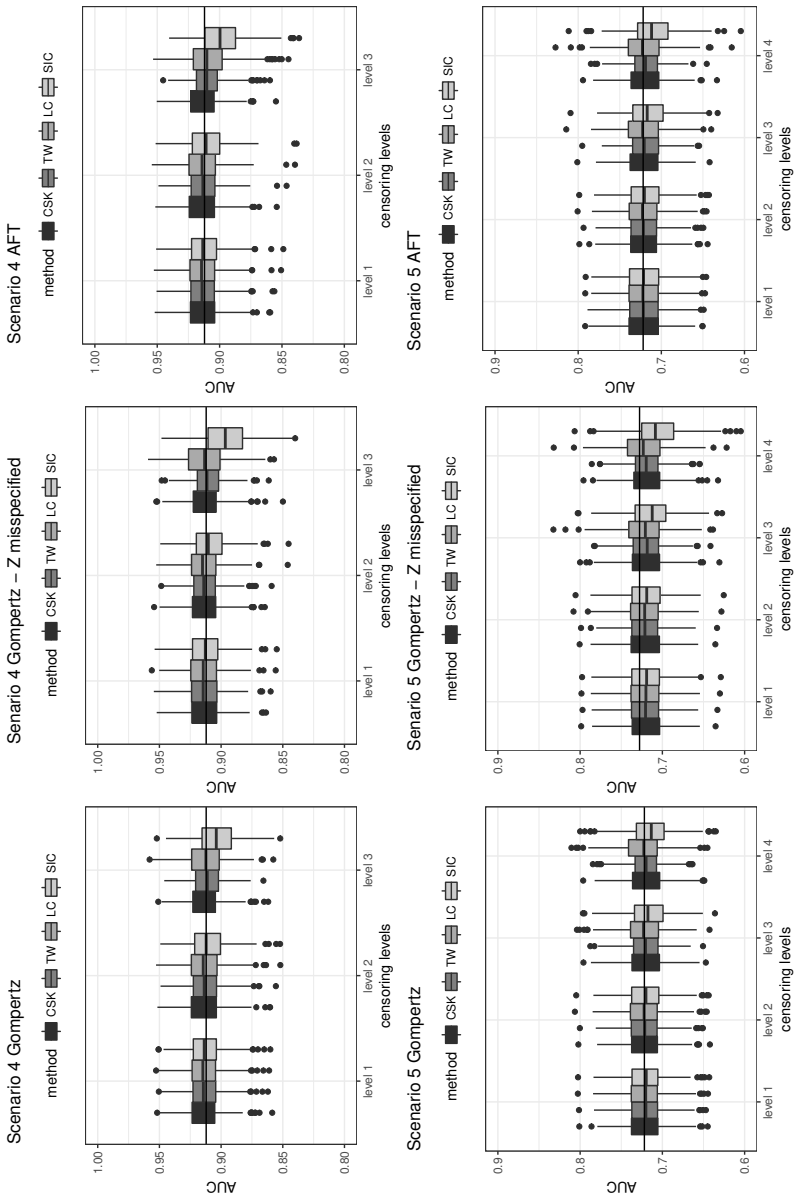

Figure 4.16: *Boxplots of the AUC for Scenarios 4 and 5 for* $n = 500$.

4.7 Appendix 2: Proofs of Theorem 4.2.1 and Corollary 4.2.1

This appendix contains the proofs of Theorem 4.2.1 and Corollary 4.2.1.

Proof of Theorem 4.2.1. Write

$$\begin{aligned}\hat{Se}(k) - Se(k) &= [\hat{Se}(k) - \tilde{Se}(k)] + [\tilde{Se}(k) - Se(k)] \\ &= S_1(k) + S_2(k) \quad \text{(say)}.\end{aligned}$$

Note that it follows from (4.8) that under the LC mixture cure model,

$$Se(k) = \frac{E[W_1 I(M > k)]}{E(W_1)} := \frac{N(k)}{D}, \quad \text{with} \quad W_1 = \frac{\{1 - p(\mathbf{X})\}(1-\Delta)}{1 - p(\mathbf{X}) + p(\mathbf{X})S_u(Y|\mathbf{Z})}.$$

Similarly, write

$$\hat{Se}(k) = \frac{\hat{N}(k)}{\hat{D}} = \frac{n^{-1}\sum_{i=1}^n \hat{N}_i(k)}{n^{-1}\sum_{i=1}^n \hat{D}_i} \quad \text{and} \quad \tilde{Se}(k) = \frac{\tilde{N}(k)}{\tilde{D}} = \frac{n^{-1}\sum_{i=1}^n \tilde{N}_i(k)}{n^{-1}\sum_{i=1}^n \tilde{D}_i}.$$

Then,

$$\begin{aligned}&S_2(k) \\ &= \frac{\tilde{N}(k) - N(k)}{\tilde{D}} + \Big(\frac{1}{\tilde{D}} - \frac{1}{D}\Big)N(k) \\ &= \Big\{\frac{\tilde{N}(k) - N(k)}{D} - \frac{N(k)}{D^2}(\tilde{D} - D)\Big\}\big\{1 + o_P(1)\big\} \\ &= \Big\{\frac{1}{D}n^{-1}\sum_{i=1}^n(\tilde{N}_i(k) - E\tilde{N}(k)) - \frac{N(k)}{D^2}n^{-1}\sum_{i=1}^n(\tilde{D}_i - E\tilde{D})\Big\} \\ &\qquad \times \big\{1 + o_P(1)\big\}, \qquad (4.15)\end{aligned}$$

since $D = E\tilde{D}$ and $N(k) = E\tilde{N}(k)$, which is a sum of zero-mean i.i.d. terms indexed by k.

Next, consider $S_1(k)$. Using a similar derivation as for $S_2(k)$, we have :

$$\begin{aligned}&S_1(k) \\ &= \Big\{\frac{1}{D}n^{-1}\sum_{i=1}^n(\hat{N}_i(k) - \tilde{N}_i(k)) - \frac{N(k)}{D^2}n^{-1}\sum_{i=1}^n(\hat{D}_i - \tilde{D}_i)\Big\}\big\{1 + o_P(1)\big\} \\ &= \Big\{S_{11}(k) + S_{12}(k)\Big\}\big\{1 + o_P(1)\big\}.\end{aligned}$$

Let us consider first $S_{11}(k)$. The term $S_{11}(k)$ depends on $\hat{W}_{i1} - \tilde{W}_{i1}$, which

equals

$$
\begin{aligned}
&\hat{W}_{i1} - \tilde{W}_{i1} \\
&= (1-\Delta_i)\Big\{ \frac{1-\hat{p}(\mathbf{X}_i)}{1-\hat{p}(\mathbf{X}_i)+\hat{p}(\mathbf{X}_i)\hat{S}_u(Y_i|\mathbf{Z}_i)} \\
&\qquad\qquad - \frac{1-p(\mathbf{X}_i)}{1-p(\mathbf{X}_i)+p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)} \Big\} \\
&:= (1-\Delta_i)\Big\{ \frac{\hat{A}_i}{\hat{B}_i} - \frac{A_i}{B_i} \Big\} \\
&= (1-\Delta_i)\Big\{ \frac{\hat{A}_i - A_i}{B_i} - \frac{A_i}{B_i^2}(\hat{B}_i - B) \Big\} \big\{ 1 + o_P(1) \big\} \\
&= \frac{1-\Delta_i}{B_i^2}\Big\{ -(\hat{p}(\mathbf{X}_i) - p(\mathbf{X}_i))(1-p(\mathbf{X}_i)+p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)) \\
&\qquad\qquad + (1-S_u(Y_i|\mathbf{Z}_i))(\hat{p}(\mathbf{X}_i) - p(\mathbf{X}_i))(1-p(\mathbf{X}_i)) \\
&\qquad\qquad - p(\mathbf{X}_i)(1-p(\mathbf{X}_i))(\hat{S}_u(Y_i|\mathbf{Z}_i) - S_u(Y_i|\mathbf{Z}_i)) \Big\} \big\{ 1 + o_P(1) \big\} \\
&= \frac{1-\Delta_i}{B_i^2}\Big\{ -(\hat{p}(\mathbf{X}_i) - p(\mathbf{X}_i))S_u(Y_i|\mathbf{Z}_i) \\
&\qquad\qquad - (\hat{S}_u(Y_i|\mathbf{Z}_i) - S_u(Y_i|\mathbf{Z}_i))p(\mathbf{X}_i)(1-p(\mathbf{X}_i)) \Big\} \\
&\qquad\qquad \times \big\{ 1 + o_P(1) \big\}.
\end{aligned}
$$

It follows from the proof of Theorem 3 in Lu (2008) that

$$
\hat{p}(\boldsymbol{x}) - p(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} \xi(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, \boldsymbol{x}) + o_p(n^{-1/2})
$$

uniformly in $\boldsymbol{x}$, for a certain function ξ satisfying $E(\xi(\mathbf{X}, \mathbf{Z}, Y, \Delta, \boldsymbol{x})) = 0$ for all $\boldsymbol{x}$, and

$$
\hat{S}_u(t|\boldsymbol{z}) - S_u(t|\boldsymbol{z}) = \frac{1}{n}\sum_{j=1}^{n} \zeta(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, t|\boldsymbol{z}) + o_p(n^{-1/2})
$$

uniformly in t and $\boldsymbol{z}$, for a certain function ζ satisfying $E(\zeta(\mathbf{X}, \mathbf{Z}, Y, \Delta, t|\boldsymbol{z})) = 0$ for

all t and $\boldsymbol{z}$. Hence,

$$
\begin{aligned}
&S_{11}(k)\\
&= \frac{1}{D}\frac{1}{n}\sum_{i=1}^{n}(\hat{W}_{i1}-\tilde{W}_{i1})I(M_i>k)\\
&= \frac{1}{D}\frac{1}{n}\sum_{i=1}^{n}\frac{(1-\Delta_i)I(M_i>k)}{B_i^2}\Big\{-S_u(Y_i|\mathbf{Z}_i)(\hat{p}(\mathbf{X}_i)-p(\mathbf{X}_i))\\
&\quad -p(\mathbf{X}_i)(1-p(\mathbf{X}_i))(\hat{S}_u(Y_i|\mathbf{Z}_i)-S_u(Y_i|\mathbf{Z}_i))\Big\}\big\{1+o_P(1)\big\}\\
&= \frac{1}{D}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(1-\Delta_i)I(M_i>k)}{B_i^2}\Big\{-S_u(Y_i|\mathbf{Z}_i)\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X}_i)\\
&\qquad -p(\mathbf{X}_i)(1-p(\mathbf{X}_i))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y_i|\mathbf{Z}_i)\Big\}+o_P(n^{-1/2})\\
&:= \frac{1}{D}\frac{2}{n(n-1)}\sum_{i<j}\Big\{\frac{1}{2}\Big(h(V_i,V_j,k)+h(V_j,V_i,k)\Big)\Big\}+o_P(n^{-1/2})\\
&:= \frac{1}{D}\frac{2}{n(n-1)}\sum_{i<j}\tilde{h}(V_i,V_j,k)+o_P(n^{-1/2}),
\end{aligned}
$$

where $h(V_i,V_j,k)=\frac{(1-\Delta_i)I(M_i>k)}{B_i^2}\Big\{-S_u(Y_i|\mathbf{Z}_i)\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X}_i)-p(\mathbf{X}_i)(1-p(\mathbf{X}_i))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y_i|\mathbf{Z}_i)\Big\}$.

We have a U-process of order 2 with symmetric kernel $\tilde{h}$, where $V_i=(\mathbf{X}_i,\mathbf{Z}_i,Y_i,\Delta_i)$. It follows from Corollary 4 in Sherman (1994) that this U-process can be decomposed in its Hajek projection and a remainder term that is uniformly of smaller order :

$$
\begin{aligned}
&S_{11}(k)\\
&= \frac{2}{D}\frac{1}{n}\sum_{j=1}^{n}E\big[\tilde{h}(V,V_j,k)|V_j\big]+o_P(n^{-1/2})\\
&= \frac{1}{D}\frac{1}{n}\sum_{j=1}^{n}E\Big[\frac{(1-\Delta)I(M>k)}{B^2}\Big\{-S_u(Y|\mathbf{Z})\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X})\\
&\qquad -p(\mathbf{X})(1-p(\mathbf{X}))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y|\mathbf{Z})\Big\}\Big]\\
&\qquad +o_P(n^{-1/2}).
\end{aligned}
\tag{4.16}
$$

In a similar way, we can show that

$$
\begin{aligned}
S_{12}(k) = -\frac{N(k)}{D^2}\frac{1}{n}\sum_{j=1}^{n}E\Big[\frac{1-\Delta}{B^2}\Big\{&-S_u(Y|\mathbf{Z})\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X})\\
&-p(\mathbf{X})(1-p(\mathbf{X}))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y|\mathbf{Z})\Big\}\Big]\\
&+o_P(n^{-1/2}).
\end{aligned}
\tag{4.17}
$$

We can now combine the expressions for $S_{11}(k)$, $S_{12}(k)$ and $S_2(k)$, which leads to

$$\begin{aligned}\hat{Se}(k) - Se(k) &= S_{11}(k) + S_{12}(k) + S_2(k) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{j=1}^{n} \eta_{Se}(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, k) + o_P(n^{-1/2}), \end{aligned} \tag{4.18}$$

uniformly in k, where $\eta_{Se}(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, k)$ is obtained by combining the expressions given in (4.16), (4.17) and (4.15).

Next, it can be shown that the class

$$\{(\boldsymbol{x}, \boldsymbol{z}, y, \delta) \to \eta_{Se}(\boldsymbol{x}, \boldsymbol{z}, y, \delta, k) : k \in \mathbb{R}\}$$

is Donsker, by decomposing the function η_{Se} in products and sums of subfunctions that are bounded and Donsker (see Examples 2.10.7 and 2.10.8 in Van der Vaart & Wellner (1996), VW hereafter). For this, it suffices to show that the bracketing number of the classes corresponding to each of these subfunctions is small enough, in the sense of Theorem 2.5.6 in VW. For calculating these bracketing numbers, one can use the well known results about the bracketing number of classes of bounded and monotone functions (see Theorem 2.7.5 in VW), classes of sufficiently smooth functions (see Corollary 2.7.2 in VW), or related results.

Finally, in a similar way as for the specificity, it can be shown that the estimator $\hat{Sp}(k)$ of the specificity can be decomposed in a sum of iid terms, plus a remainder term that is uniformly of smaller order :

$$\hat{Sp}(k) - Sp(k) = \frac{1}{n}\sum_{j=1}^{n} \eta_{Sp}(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, k) + o_P(n^{-1/2}), \tag{4.19}$$

where η_{Sp} is obtained by replacing in the expression of η_{Se} all indicators $I(M > k)$ by $I(M \leq k)$ and by noting that $W_0 = 1 - W_1$. □

Proof of Corollary 4.2.1. Write

$$\begin{aligned}\hat{ROC}(u) - ROC(u) &= \big[\hat{Se}\{(1-\hat{Sp})^{-1}(u)\} - Se\{(1-\hat{Sp})^{-1}(u)\}\big] \\ &\quad + \big[Se\{(1-\hat{Sp})^{-1}(u)\} - Se\{(1-Sp)^{-1}(u)\}\big] \\ &:= T_1(u) + T_2(u).\end{aligned}$$

We start with $T_2(u)$:

$$\begin{aligned}T_2(u) &= Se'\big\{(1-Sp)^{-1}(u)\big\}\Big\{(1-\hat{Sp})^{-1}(u) - (1-Sp)^{-1}(u)\Big\} \\ &\quad + o_P(n^{-1/2}) \\ &= Se'\big\{(1-Sp)^{-1}(u)\big\}\Big[-\frac{(1-\hat{Sp})\{(1-Sp)^{-1}(u)\} - u}{(1-Sp)'\{(1-Sp)^{-1}(u)\}}\Big] \\ &\quad + o_P(n^{-1/2}).\end{aligned}$$

This development is similar as in, for example, Cheng (1984), among others.

We know that

$$\hat{Sp}(k) - Sp(k) = n^{-1} \sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + o_P(n^{-1/2}),$$

uniformly in k. It follows that

$$\begin{aligned} T_2(u) = & \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}} \\ & \times \; n^{-1} \sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) \; + o_P(n^{-1/2}), \end{aligned}$$

uniformly in $\delta \leq u \leq 1-\delta$. Next,

$$T_1(u) = \hat{Se}\{(1-Sp)^{-1}(u)\} - Se\{(1-Sp)^{-1}(u)\} + o_P(n^{-1/2}),$$

since it follows from the weak convergence of $\hat{Se} - Se$ that

$$\sup_{|k_2-k_1| \leq Cn^{-1/2}} |\hat{Se}(k_2) - Se(k_2) - \hat{Se}(k_1) + Se(k_1)| = o_p(n^{-1/2})$$

for $0 < C < \infty$, and

$$\begin{aligned} & \sup_{\delta \leq u \leq 1-\delta} |(1-\hat{Sp})^{-1}(u) - (1-Sp)^{-1}(u)| \\ & \qquad = \; O_P\Big(\sup_k |\hat{Sp}(k) - Sp(k)| \Big) = O_P(n^{-1/2}). \end{aligned}$$

It follows that

$$T_1(u) = n^{-1} \sum_{i=1}^{n} \eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) + o_P(n^{-1/2}).$$

Hence,

$$\begin{aligned} & \hat{ROC}(u) - ROC(u) \\ & = n^{-1} \sum_{i=1}^{n} \Big[\eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) \\ & \quad + \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) \Big] \\ & \quad + o_P(n^{-1/2}), \end{aligned}$$

uniformly in $\delta \leq u \leq 1-\delta$. We know that the classes

$$\big\{(\boldsymbol{x}, \boldsymbol{z}, y, \delta) \to \eta_{Se}(\boldsymbol{x}, \boldsymbol{z}, y, \delta, k) : k \in \mathbb{R}\big\}$$

and

$$\big\{(\boldsymbol{x}, \boldsymbol{z}, y, \delta) \to \eta_{Sp}(\boldsymbol{x}, \boldsymbol{z}, y, \delta, k) : k \in \mathbb{R}\big\}$$

are Donsker, and that the functions $u \to (1 - Sp)^{-1}(u)$ and $k \to Se(k)'/(1 - Sp)^{-1}(k)$ are continuously differentiable and bounded. Hence, the weak convergence of the process $n^{1/2}\{\widehat{ROC}(u) - ROC(u)\}$ $(\delta \leq u \leq 1 - \delta)$ follows.

It remains to show the limiting distribution of $n^{1/2}(\hat{AUC}_\delta - AUC_\delta)$. Note that

$$\begin{aligned}\widehat{AUC}_\delta - AUC_\delta &= \int_\delta^{1-\delta} \left\{\widehat{ROC}(u) - ROC(u)\right\} du \\ &= n^{-1} \sum_{i=1}^{n} \int_\delta^{1-\delta} \eta_{ROC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, u)\, du + o_P(n^{-1/2}) \\ &:= n^{-1} \sum_{i=1}^{n} \eta_{AUC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i) + o_P(n^{-1/2})\end{aligned}$$

and that

$$\begin{aligned}&\mathrm{Var}(\eta_{AUC}(\mathbf{X}, \mathbf{Z}, Y, \Delta)) \\ &= \int_\delta^{1-\delta} \int_\delta^{1-\delta} \mathrm{Cov}\{\eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_1), \eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_2)\}\, du_1\, du_2,\end{aligned}$$

since η_{Se} and η_{Sp} are unbiased. This finishes the proof. □

Chapter 5

Conclusions and Further Research

Throughout this thesis we have investigated several issues related to cure models in survival analysis. In this last chapter, we will conclude the presentation of our contribution to that topic by first summarising the three projects on which we have worked on and by then presenting some further research which is in the continuity of the present work and which seems to be worth investigating.

5.1 General Conclusions

In this manuscript, we have first extensively presented the growing literature that exists on cure models which is mainly composed of two classes of models, the mixture cure model which defines the survival function of the entire population as a mixture of the survival of cured and uncured sub-populations, and the promotion time cure model which adapts the Cox PH model to take long-term survivors into account. In Chapter 2, we have seen that, for the mixture cure model, many different modelling approaches have been proposed, ranging from completely parametric models for the incidence and the latency, to completely nonparametric proposals. If a logistic regression model is mainly assumed to model the probability of being uncured, many different approaches have been proposed to model the survival function for uncured subjects, all of them coming from the classical survival analysis literature. Among others, we can mention a Weibull model, a Cox PH model, parametric or nonparametric approaches. Among them, the most widely studied is the semiparametric logistic/Cox mixture cure model for which several original estimation approaches have been proposed. Alongside modelling, other elements such that testing for the presence of a cure fraction, testing for sufficient follow-up, model diagnostics and variable selection have also be investigated. For the promotion time cure model, less modelling approaches have been proposed, but it is important to note that estimation methods based on the Bayesian paradigm have been investigated as well as measurement errors. More recently, unifying approaches,

embedding both the mixture cure model and the promotion time cure model, have been proposed. As we have explained in Section 2.3, two different streams have driven these developments, one based on a mathematical approach mainly relying on the use of the Box-Cox transformation; the other one being based on the biological development underlying the promotion time cure model. Since all these unifying approaches contain the mixture and the promotion time cure model as special cases, one can use them to determine which model to use. However, as described in Section 2.3.3, when there is a transformation parameter leading to one or another model, very few elements are available in the data to estimate it and hence to determine the most appropriate model. Even if some other proposals have been made, it seems that choosing between the mixture and the promotion time cure model is an issue which relies more on a personal choice than on a purely mathematical perspective. Therefore, the knowledge of the field of the analysis and the scientific question, among others, are central points which guide the choice for one or the other model.

A second contribution of this thesis focuses on the mixture cure model and more precisely on the modelling of the uncure probability $p(\boldsymbol{x})$. Indeed, among all the proposals for the mixture cure model, many of them consider a logistic regression model for $p(\boldsymbol{x})$. Nevertheless, there is no reason to limit the incidence to such modelling. Even if Peng (2003a) mentions other parametric models – a probit and a complementary log-log model – one could envisage a more flexible approach. In Chapter 3 we have proposed such a model by considering a single-index structure for the incidence and a Cox PH model for the latency. Among the advantages of the single-index model, we can mention its flexibility compared with a parametric model, but also the fact that it avoids curse-of-dimensionality problems encountered in nonparametric regression. Based on the work by Sy & Taylor (2000), we proposed a maximum likelihood estimation method based on the EM algorithm which permits to estimate separately the incidence and the latency of the model. For the incidence, we have adapted the method proposed by Klein & Spady (1993) for single-index models with binary outcome which considers a kernel estimator with Nadaraya-Watson weights for the unknown link function. Since the cure status is unobserved, this estimator had to be adapted, the expected cure status being rather considered instead of the true one. As all kernel estimators, a crucial point was the choice of the bandwidth. We proposed to consider likelihood cross-validation, embedded within the EM algorithm, the bandwidth being computed at each iteration of the algorithm prior to the estimation of $p(\boldsymbol{x})$. For the latency, based again on Sy & Taylor (2000), we proposed to estimate the vector of parameters $\boldsymbol{\beta}$ from a Breslow-type profile likelihood approach. Alongside the estimation of the model, we also investigated the important issue of model identifiability and we proved the identifiability of our proposal. The investigation of the finite sample performance of the SIC cure model has demonstrated its good performance when the true model is not a logistic model, both when the true link function is monotone and when it is not. Furthermore, an interesting point on which our analysis shed light is that the latency is only slightly affected by the model considered for the incidence. Nevertheless, the censoring rate affects both parts of the model and one should be careful when using a cure model

when only few observations are located in the plateau. To further assess the practicability of our proposal, we estimated a SIC cure model on a breast cancer data set and we compared the results with those obtained from a LC cure model. Based on prediction errors we have seen that the SIC cure model performed better than the LC cure model. An important issue to address was the interpretation of the single-index part of the model. We considered a graphical analysis by representing the link function and plotting the estimate of the uncured probability against covariates. We observed a non-monotone link function, the SIC cure model seeming therefore more suitable to model these data. We further investigated this first result by studying the effect of age on $p(\boldsymbol{x})$ through a graphical representation. We observed a non-linear effect of age on the uncured probability, indicating that a flexible link function might not be the best option to model the data. Nevertheless, the SIC cure model provides an interesting diagnostic tool to assess for model misspecification.

The third contribution of this thesis (Chapter 4) concerns the issue of prediction assessment for cure survival data. Indeed, the evaluation of the predictive performance of a covariate or a model is an important issue in many fields and there was a gap in the evaluation of the cure status prediction based on survival data. A third contribution of this thesis was therefore the development of a ROC curve for this purpose. In the presence of a continuous classifier, the ROC curve, alongside the AUC, is often considered to evaluate binary classification performance. However, estimating a ROC curve supposes that the membership to the two possible classes is fully observed. In the presence of a cure fraction, the cure status is only partially observed through the censoring indicator. It was therefore not possible to compute the ROC based on 'classical' methods. Alternatively, we proposed to decompose the sensitivity and the specificity based on the definition of conditional probability which lead to estimators taking the form of weighted empirical distribution functions. We have also explained that intuitively these estimators corresponded to a situation where three groups of observations could be distinguished based on the so-called *cure threshold* proposed by Taylor (1995): those who are uncensored and therefore uncured, those who are censored after τ and considered as cured, and those censored with a follow-up time smaller than τ, to which a probability of being cured was attributed. We proposed to compute the weights according to the mixture cure model, investigating both the LC cure model and the SIC cure model. Alongside the sensitivity, the specificity and the ROC curve estimators, we also provided an AUC estimator and we established the asymptotic properties of the proposed estimators. Through an extensive simulation study, we have demonstrated the good performance of our proposal compared to the two infeasible competitors we considered, both when the classifier was known and unknown and when not too many observations where censored before the cure threshold τ. We also investigated misspecification in the weights and we have seen that censoring was the element affecting the most the performance of our proposal. We finally illustrated our proposal on a real data set representing the time to death for patients suffering from melanoma.

5.2 Discussion and Further Research

When thinking about potential extensions of the present work, different ideas emerge. First, two points seem to be worth discussing since they are directly related to the work we have presented on the SIC cure model in Chapter 3. Starting from the application of the SIC cure model on the breast cancer dataset (see Section 3.3), a first point which could enhance real data analysis would consist in computing confidence intervals for the uncure probability curve $p(\boldsymbol{x})$ when it is estimated by a SIC cure model. This could be an interesting additional element to evaluate the uncertainty in the estimate of the curve. To do so, an approach would consist in using a bootstrap method to estimate the distribution of the estimator $p(\boldsymbol{x})$ for each fixed $\boldsymbol{x}$, from which a confidence interval can be obtained. However, the bootstrap procedure that we used for obtaining the variance of the estimator of the coefficients of the single-index model can not be used for the uncure probability. The reason is that the naive bootstrap that we used for the parameters of the single-index model is not able to capture the bias of the estimator. In fact, the bootstrap estimator obtained from the naive bootstrap is an unbiased estimator (in the bootstrap world) of the original estimator. A possible way out is to use a local bootstrap, which consists of drawing bootstrap samples among the data points that lie close to the covariate vector $\boldsymbol{x}$ in which we are interested. In this way, we should be able to capture the correct variance and bias, and so we should obtain correct pointwise confidence intervals for $p(\boldsymbol{x})$.

Another point which seems to be interesting to discuss is the development of the asymptotic theory of the proposed estimator, which would constitute a further step forward in the SIC cure model methodology. However, this is a challenging problem. Indeed, since the response variable in the kernel estimator for the single-index model is latent and therefore replaces by its expectation obtained from the EM algorithm, the estimator is defined via an EM algorithm and there does not seem another equivalent way to define the estimator. There are other situations where the EM algorithm is merely an algorithm used to calculate an estimator, but in our case it is more than that since the methodology itself is defined via this algorithm.

Beside these two first direct extensions of the SIC cure model, we can envisage some more derivatives of our proposal. One of them consists in considering another model for the latency alongside a single-index structure for the incidence. As we have explained in Chapter 1, the AFT model constitutes the other popular model considered for survival data alongside the Cox PH model. An appealing characteristic of the AFT model is that the effect of covariates can be interpret as decelerating or accelerating the timing of the event offering therefore an alternative to the Cox model and its proportional hazards property. As we have seen in Chapter 2, the literature on mixture cure models contains some proposals assuming an AFT model for the latency, both parametric and semiparametric. We could therefore think of a semiparametric mixture cure model where, alongside a single-index structure for the incidence, the latency could be modelled as a semiparametric AFT model. In such a case, we could assume that, in the latency, $\log(T^*) = \beta_0 + \boldsymbol{\beta}^t\mathbf{Z} + \epsilon$ where the distribution of ϵ

would be left unspecified and that the incidence would take the form given by Equation (3.1). Based on what has already been done for estimating a mixture cure model assuming a semiparametric AFT model for the latency, the EM algorithm could be considered to estimate the model, in the same manner as for the SIC cure model. The E-step would remain the same as well as the M-step for the incidence. For the latency, one of the methods proposed by Li & Taylor (2002), Zhang & Peng (2007) or Lu (2010) could be considered for maximising the likelihood for the latency. We could also envisage to introduce more flexibility by considering a nonparametric modelling for the latency alongside a single-index model for the incidence. A possibility would consist in extending the proposal made by Patilea & Van Keilegom (2018) who consider a mixture cure model with a parametric incidence and a nonparametric latency.

By considering a flexible link function in the incidence modelling, this also calls for the development of testing procedures for the form of the link function. Such methods, with, for example, the development of a goodness-of-fit test for the link function based on the single-index model, would therefore constitute another possible extension of our work. Likewise, we could also envisage to evaluate the fit of the latency by proposing a goodness-of-fit test for this part of the model under a single-index model for the incidence.

Another potential extension for the SIC cure model concerns high dimensional data which are of growing interest in the statistical literature. Two types of analysis are mainly performed when high dimensional data are at hand, namely, the identification of features that are individually associated with the outcome of interest and modelling. Concerning the first point, it consists in testing the effect of each feature on the outcome. Within the context of cure survival data, it has already been performed by López-Cheda (2018) who applies the non-parametric test for covariate significance for the incidence of mixture cure models she developed with co-authors (López-Cheda et al. (2018)) on gene-expression data. In our case, we could also envisage to do it based on a SIC cure model to determine which variables affect individually the cure status or the survival of uncured subjects by fitting a SIC cure model on each feature and then perform a significance test. However, as it consists of multiple testing, a correction is necessary. We could for example consider the False Discovery Rate approach proposed by Benjamini & Hochberg (1995) which is one of the most recommended methods in the literature on high dimensional data.

For modelling, 'classical' statistical approaches can not be applied directly on high dimensional data since the number of features is much larger than the number of observations. An approach to model data in such context is dimension reduction with variable selection approaches. Penalised regressions are one type of methods considered for this purpose. They consist in penalising the magnitude of the parameters by considering an additional penalty term in the estimation procedure, the parameters being therefore shrunk toward zero. These methods have the advantage of both reducing the dimension and selecting features. Different types of penalties have been proposed: the LASSO by Tibshirani (1996), the SCAD penalty proposed by Fan & Li (2001) and the adaptive LASSO by Zou (2006). Based on this approach, we could think of modelling high dimensional data based on a penalised SIC cure model. Since

the two parts of the model are estimated separately based on the EM algorithm, this appealing characteristic eases the estimation of the model. We can think of applying a penalty to each part of the likelihood as it has been proposed by Liu et al. (2012) for the LC cure model and borrow methods from the single-index and the Cox PH model literature on penalised regression to estimate the two parts of the model.

On a very different perspective, we can also envisage an extension of the ROC curve approach we proposed in Chapter 4 to assess the cure status prediction from cure survival data. In the literature on ROC curves, alongside the sensitivity and the specificity which evaluate the frequency of misclassification, there exist two other quantities that can also be of interest when evaluating the predictive performance of a classifier: the *positive predictive value* given by

$$PPV(k) = P(D = 1|M > k),$$

and the *negative predictive value*:

$$NPV(k) = P(D = 0|M \leq k),$$

which both evaluate how well the classifier predicts the true classes. A possible extension of Chapter 4 would be to propose estimators for these two quantities which can be obtained straightforwardly. Indeed, it is easy to see the relationship between these two quantities and the sensitivity and the specificity. Based on the definition of conditional probability, we have that

$$PPV(k) = \frac{P(M > k|D = 1)\ P(D = 1)}{P(M > k)} = \frac{Se(k)\ P(D = 1)}{P(M > k)},$$

and that

$$NPV(k) = \frac{P(M \leq k|D = 0)\ P(D = 0)}{P(M \leq k)} = \frac{Sp(k)\ P(D = 0)}{P(M \leq k)}.$$

At it can be seen both quantities rely on the prevalence given by $P(D = 1)$. As already mentioned several times, since cure survival data are subject to censoring, the prevalence is latent. It is therefore not possible to directly obtain the positive predictive value and the negative predictive value of the cure status from cure survival data. However, we can apply the same reasoning as in Chapter 4 to estimate these two quantities. In fact,

$$\begin{aligned} P(D = 1) &= E\big\{I(T = \infty)\big\} \\ &= E\{I(T = \infty)I(T > C)\} \\ &= E\Big[I(T > C)\ E\big\{I(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\big\}\Big]. \end{aligned}$$

It follows that

$$P(D = 1) = E\big\{(1 - \Delta)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, C, T > C)\big\}.$$

By replacing the expectation by an average and assuming the cure threshold, $P(D = 1)$ can be estimated by the infeasible estimator

$$\tilde{P}(D = 1) = \frac{1}{\tilde{N}_1} \sum_{i=1}^{n} \tilde{W}_{i1}.$$

Hence, an infeasible estimator for the predictive positive value is given by

$$\widetilde{PPV}(k) = \frac{\tilde{Se}(k)\ \tilde{P}(D = 1)}{P(M > k)}.$$

Likewise, we can obtain the following infeasible estimator for the negative predictive value:

$$\widetilde{NPV}(k) = \frac{\tilde{Sp}(k)\ \tilde{P}(D = 0)}{P(M \leq k)},$$

where

$$\tilde{P}(D = 0) = \frac{1}{\tilde{N}_0} \sum_{i=1}^{n} \tilde{W}_{i0}.$$

We can further envisage to estimate $\tilde{W}_0$ and $\tilde{W}_1$ assuming a LC cure model or a SIC cure model as in Chapter 4 in order to obtain feasible estimators for the positive and the negative predictive values.

Finally, to go far beyond the scope of what has been done in this thesis, an original idea would be to develop random forests for cure survival data. Introduced by Breiman (2001), a random forest is a combination of decision trees which are more and more used for predictions. Decision trees and random forests exist for binary and continuous outcomes (see for example the textbook by Hastie et al. (2009) for a detailed explanation of the method) as well as for survival data (see for example the article by Bou-Hamad et al. (2011) for an interesting literature review on survival trees). An interesting but also challenging topic would consist in extending such methods to cure survival data. A first step would consist in developing a decision tree. Since two types of information can be obtained from cure survival data, we could consider a classification tree for the cure status prediction, but also a survival tree which could handle the presence of a cure fraction. A second step would consist in extending them to random forests. Many different points would have to be investigated, but it could be, for example, a useful approach when handling high dimensional data.

Bibliography

Amico, M. & Van Keilegom, I. (2018*a*), 'Assessing cure status prediction from survival data using ROC curves'. *Submitted*.

Amico, M. & Van Keilegom, I. (2018*b*), 'Cure models in survival analysis', *Annual Review of Statistics and Its Application* **5**, 311–342.

Amico, M., Van Keilegom, I. & Legrand, C. (2018), The single-index/Cox mixture cure model. *Biometrics* (to appear).

Andersen, P. K., Borgan, R., Gill, R. D. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer, New York.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society - Series B* **57**, 289–300.

Beran, R. (1981), Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

Berkson, J. & Gage, R. (1952), 'Survival curve for cancer patients following treatment', *Journal of the American Statistical Association* **47**, 501–515.

Bertrand, A., Legrand, C., Carroll, R., de Meester, C. & Van Keilegom, I. (2017), 'Inference in a survival cure model with mismeasured covariates using a SIMEX approach', *Biometrika* **104**, 31–50.

Bertrand, A., Legrand, C., Léonard, D. & Van Keilegom, I. (2017), 'Robustness of estimation methods in a survival cure model with mismeasured covariates', *Computational Statistics and Data Analysis* **113**, 3–18.

Beyene, K., El Ghouch, A. & Oulhaj, A. (2018), On the validity of time-dependent AUC estimation in the presence of a cure fraction. *Submitted*.

Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. (2013), 'Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring', *Biometrical journal* **55**, 687–704.

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

Bou-Hamad, I., Larocque, D. & Ben-Hameur, H. (2011), 'A review of survival trees', *Statistics Surveys* **5**, 44–71.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Bremhorst, V. & Lambert, P. (2016), 'Flexible estimation in cure survival models using bayesian P-splines', *Computional Statististics and Data Analysis* **93**, 270–284.

Breslow, N. (1974), 'Covariates analysis of censored survival data', *Biometrics* **30**, 89–99.

Cai, C., Zou, Y., Peng, P. & J., Z. (2012), 'smcure: Fit semiparametric mixture cure models. r package version 2.0'.
URL: *https://CRAN.R-project.org/package=smcure*

Carroll, R., Ruppert, D., Stefanski, L. & Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective. 2nd edition*, Chapman and Hall/CRC, Boca Raton, London, New York, Washington, D.C.

Chambless, L. & Diao, G. (2006), 'Estimation of time-dependent area under the ROC curve for long-term risk prediction', *Statistics in medicine* **25**, 3474–3486.

Chen, M., Ibrahim, J. & Sinha, D. (1999), 'A new Bayesian model for survival data with a surviving fraction', *Journal of the American Statistical Association* **94**, 909–919.

Cheng, K.-F. (1984), 'On almost sure representation for quantiles of the product-limit estimator with applications', *Sankhya, Series A* **46**, 426–443.

Claeskens, G. & Van Keilegom, I. (2018), The focused information criterion for a mixture cure model. Working paper.

Collett, D. (2003), *Modelling Survival Data in Medical Research. 2nd edition*, Chapman and Hall/CRC, Boca Raton, London, New York, Washington, D.C.

Cook, J. & Stefanski, L. (1994), 'Simulation-extrapolation in parametric measurement error models', *Journal of the American Statistical Association* **89**, 1314–1328.

Cooner, F., Banerjee, S., Carlin, B. & Sinha, D. (2007), 'Flexible cure rate modeling under latent activation schemes', *Journal of the American Statistical Association* **102**, 560–572.

Cooner, F., Yu, X., Banerjee, S., Grambsch, P. & McBean, A. (2009), 'Hierarchical dynamic time-to-event models for post-treatment preventive care data on breast cancer survivors', *Statistical Modelling* **9**, 560–572.

Copas, J. & Corbett, P. (2002), 'Overestimation of the Receiver Operating Characteristic curve for logistic regression', *Biometrika* **89**, 315–331.

Corbière, F., Commenges, D., Taylor, J. & Joly, P. (2009), 'A penalized likelihood approach for mixture cure models', *Statistics in Medicine* **28**, 510–524.

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

de Castro, M., Cancho, V. & Rodrigues, J. (2009), 'A bayesian long-term survival model parametrized in the cured fraction', *Biometrical Journal* **3**, 443–355.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via EM algorithm', *Journal of the Royal Statistical Society - Series B* **39**, 1–38.

Diao, G. & Yin, G. (2012), 'A general transformation class of semiparametric cure rate frailty models', *Annals of the Institute of Statistical Mathematics* **64**, 959–989.

Dirick, L., Claeskens, G. & Baesens, B. (2015), 'An Akaike information criterion for multiple event mixture cure models', *European Journal of Operational Research* **241**, 449–457.

Eilers, P. & Marx, B. (1996), 'Flexible smoothing B-splines and penalties (with discussion)', *Statistical Science* **1**, 89–121.

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

Farewell, V. (1977), 'A model for binary variable with time-censored observations', *Biometrika* **64**, 43–46.

Farewell, V. (1982), 'The use of a mixture model for the analysis of survival data with long-term survivors', *Biometrics* **38**, 1041–1046.

Gelfand, A. & Ghosh, S. (1998), 'Model choice: a minimum posteriori predictive loss approach', *Biometrika* **85**, 1–11.

Ghitany, M., Maller, R. & Zhou, S. (1994), 'Exponential mixture models with long-term survivors and covariates', *Journal of Multivariate Analysis* **49**, 218–241.

Göner, M. & Heller, G. (2005), 'Concordance probability and discriminatory power in proportional hazards regression', *Biometrika* **92**, 965–970.

Hanin, L. & Huang, L.-S. (2014), 'Identifiability of cure models revisited', *Journal of Multivariate Analysis* **130**, 261–274.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. (1982), 'Evaluating the yield of medical tests', *Journal of the American Medical Association* **247**, 2543–2546.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. (1984), 'Regression modeling strategies for improved prognostic prediction', *Statistics in Medicine* **3**, 143–152.

Hastie, T. & Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **3**, 297–318.

Hastie, T., Tibshirani, R. & Frideman, R. (2009), *The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 2nd edition*, Springer, New York.

Heagerty, P., Lumley, T. & Pepe, M. (2000), 'Time-dependent ROC curves for censored survival data and a diagnostic marker', *Biometrics* **56**, 337–344.

Heagerty, P. & Zheng, Y. (2005), 'Survival model predictive accuracy and ROC curves', *Biometrics* **61**, 92–105.

Horowitz, J. (2009), *Semiparametric and Nonparametric Methods in Econometrics*, Springer, New York.

Hsu, W., Todem, D. & Kim, K. (2016), 'A sup-score test for the cure fraction in mixture models for long-term survivors', *Biometrics* **76**, 1348–1357.

Ibrahim, J., Chen, M. & Sinha, D. (2001), 'Bayesian semiparametric models for survival data with a cure fraction', *Biometrics* **57**, 383–388.

Ichimura, H. (1993), 'Semiparametric least squares (SLS) and weighted SLS estimation of single-index models', *Journal of Econometrics* **58**, 71–120.

Kalbfleisch, J. & Prentice, R. (1973), 'Marginal likelihoods based on Cox's regression and life model', *Biometrika* **60**, 267–278.

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

Kim, S., Chen, M., Dey, D. & Gamerman, D. (2007), 'Bayesian dynamic models for survival data with a cure fraction', *Lifetime Data Analysis* **13**, 17–35.

Kim, S., Chen, M. & Key, D. (2011), 'A new threshold regression model for survival data with a cure fraction', *Lifetime Data Analysis* **17**, 101–122.

Klein, J. & Moeschberger, M. (2003), *Survival analysis : Techniques for Censored and Truncated Data. 2nd edition*, Springer, New York.

Klein, R. & Spady, R. (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica* **61**, 387–421.

Krzanowski, W. & Hand, D. (2009), *ROC Curves for Continuous Data*, Chapman & Hall/CRC, Boca Raton.

Kuk, A. & Chen, C. (1992), 'A mixture model combining logistic regression with proportional hazards regression', *Biometrika* **79**, 531–541.

Lam, K., Fong, D. & Tang, O. (2005), 'Estimating the proportion of cured patients in censored sample', *Statistics in Medicine* **24**, 1865–1879.

Lang, S. & Brezger, A. (2004), 'Bayesian P-splines', *Journal of Computational and Graphical Statistics* **13**, 183–212.

Laud, P. & Ibrahim, J. (1995), 'Predictive model selection', *Journal of the Royal Statistical Society - Series B* **57**, 247–262.

Li, K.-C. & Duan, H. (1989), 'Regression analysis under link violation', *The Annals of Statistics* **17**, 1009–1052.

Li, L., Greene, T. & Hu, B. (2018), 'A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data', *Statistical Methods in Medical Research* **27**, 2264–2278.

Li, S. & Taylor, J. (2002), 'A semi-parametric accelerated failure time cure model', *Statistics in Medicine* **21**, 3235–3247.

Liu, X., Peng, Y., Tu, D. & Liang, H. (2012), 'Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials', *Statistics in Medicine* **31**, 2882–2891.

López-Cheda, A. (2018), Nonparametric Inference in Mixture Cure Models, PhD thesis, Universidade da Coruña, Spain.

López-Cheda, A., Cao, R., Jácome, M. & Van Keilegom, I. (2017), 'Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models', *Computational Statistics and Data Analysis* **105**, 144–165.

López-Cheda, A., Jácome-Pumar, A., Van Keilegom, I. & Cao, R. (2018), Nonparametric covariate significance tests for the incidence in cure models. *Submitted*.

Lopez, O. (2009), 'Single-index regression models with right-censored responses', *Journal of Statistical Planning and Inference* **139**, 1082–1097.

Lopez, O., Patilea, V. & Van Keilegom, I. (2013), 'Single index regression models in the presence of censoring depending on the covariates', *Bernoulli* **19**, 721–747.

Lu, W. (2008), 'Maximum likelihood estimation in the proportional hazards cure model', *Annals of the Institute of Statistical Mathematics* **60**, 545–574.

Lu, W. (2010), 'Efficient estimation for an accelerated failure time model with a cure fraction', *Statistica Sinica* **20**, 661–674.

Lu, W. & Ying, Z. (2004), 'On semiparametric transformation cure models', *Biometrika* **91**, 331–343.

Lu, X. & Burke, M. (2005), 'Censored multiple regression by the method of average derivatives', *Journal of Multivariate Analysis* **95**, 182–205.

Ma, Y. & Yin, G. (2008), 'Cure rate model with mismeasured covariates under transformation', *Journal of the American Statistical Association* **103**, 743–756.

Maller, R. & Zhou, S. (1996), *Survival Analysis with Long Term Survivors*, Wiley, New York.

Mizoi, M., Bolfarine, H. & Pedroso-De-Lima, A. (2007), 'Cure rate model with measurement error', *Communication in Statistics - Simulation and Computation* **36**, 185–196.

Moeschberger, M. & Klein, J. (1985), 'A comparison of several methods of estimating the survival function when there is extreme right censoring', *Biometrics* **41**, 253–259.

Müller, H.-G. & Schmitt, T. (1988), 'Kernel and probit estimates in quantal bioassay', *Journal of the American Statistical Association* **83**, 750–759.

Müller, U. & Van Keilegom, I. (2018), Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika* (to appear).

Murphy, S. & Van der Vaart, A. (2000), 'On profile likelihood', *Journal of the American Statistical Association* **95**, 449–465.

Patilea, V. & Van Keilegom, I. (2018), A general approach for cure models in survival analysis. Under revision for *The Annals of Statistics*.

Peng, Y. (2003a), 'Fitting semiparametric cure models', *Computational Statistics and Data Analysis* **41**, 481–490.

Peng, Y. (2003b), 'Estimating baseline distribution in proportional hazards cure models', *Computational Statistics and Data Analysis* **42**, 187–201.

Peng, Y. & Dear, K. (2000), 'A nonparametric mixture model for cure rate estimation', *Biometrics* **56**, 237–243.

Peng, Y., Dear, K. & Denham, J. (1998), 'A generalized F mixture model for cure rate estimation', *Statistics in Medicine* **17**, 813–830.

Peng, Y. & Taylor, J. (2014), Cure models, *in* J. Klein, H. van Houwelingen, J. Ibrahim & T. Scheike, eds, 'Handbook of Survival Analysis', Handbooks of Modern Statistical Methods series, Chapman & Hall, Boca Raton, FL, chapter 6, pp. 113–134.

Peng, Y. & Taylor, J. (2017), 'Residual-based model diagnosis methods for mixture cure models', *Biometrics* **73**, 495–505.

Peng, Y. & Xu, J. (2012), 'An extended cure model and model selection', *Lifetime Data Analysis* **18**, 215–233.

Pepe, M. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxfod University Press, Oxford.

Portier, F., El Ghouch, A. & Van Keilegom, I. (2017), 'Efficiency and bootstrap in the promotion time cure model', *Bernoulli* **23**, 3437–3468.

Portier, F., El Ghouch, A. & Van Keilegom, I. (2018), 'On proportional hazards cure models'. *Submitted*.

Prentice, R. (1974), 'A log-gamma model and its maximum likelihood estimation', *Biometrika* **61**, 538–544.

Prentice, R. (1978), 'Linear rank tests with right censored data', *Biometrika* **65**, 167–179.

Ritov, Y. (1990), 'Estimation in a linear regression model with censored data', *The Annals of Statistics* **18**, 303–328.

Rodrigues, J., Canch, V., de Castro, M. & Louzada-Neto, F. (2009), 'On the unification of long-term survival models', *Statistics and Probability Letters* **79**, 753–759.

Scolas, S., El Ghouch, A., Legrand, C. & Oulhaj, A. (2016), 'Variable selection in a flexible parametric mixture cure model with interval-censored data', *Statistics in Medicine* **35**, 1210–1225.

Sherman, R. (1994), 'Maximal inequalities for degenerate U-processes with applications to optimization estimators', *The Annals of Statistics* **22**, 439–459.

Strzalkowska-Kominiak, E. & Cao, R. (2013), 'Maximum likelihood estimation for conditional distribution single-index models under censoring', *Journal of Multivariate Analysis* **114**, 74–98.

Sy, J. & Taylor, J. (2000), 'Estimation of a Cox proportional hazards cure model', *Biometrics* **56**, 227–236.

Taylor, J. (1995), 'Semi-parametric estimation in failure time mixture models', *Biometrics* **51**, 899–907.

Taylor, J. & Liu, N. (2007), Statistical issues involved with extending standards models, *in* V. Nair, ed., 'Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A Doksum', series in biostatistics, World Scientific, Singapore, chapter 15, pp. 299–311.

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

Tournoud, M. & Ecochard, R. (2008), 'Promotion time models with time-changing exposure and heterogeneity : application to infectious diseases', *Biometrical Journal* **3**, 395–407.

Tsodikov, A. (1998a), 'A proportional hazards model taking account off long-term survivors', *Biometrics* **54**, 1508–1516.

Tsodikov, A. (1998b), 'Asymptotic efficiency of proportional hazards model with cure', *Statistics and Probability Letters* **39**, 237–244.

Tsodikov, A. (2001), 'Estimation of survival based on proportional hazards when cure is a possibility', *Mathematical and Computer Modelling* **33**, 1227–1236.

Tsodikov, A. (2002), 'Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage.', *Statistics in Medicine* **21**, 895–920.

Tsodikov, A., Ibrahim, J. & Yakovlev, A. (2003), 'Estimating cure rates from survival data', *Journal of the American Statistical Association* **98**, 1063–1078.

Van der Vaart, A. & Wellner, J. (1996), *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New-York.

Wang, L., Pang, D. & Liang, H. (2012), 'Two-component mixture cure rate model with spline estimated nonparametric components', *Biometrics* **68**, 726–735.

Wang, Y., He, S., Zhu, L. & Yuen, K. (2007), 'Asymptotics for a censored generalized linear model with unknown link function', *Probability Theory and Related Fields* **138**, 235–267.

Wang, Y., Klijn, J., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijet-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D. & Foekens, J. (2005), 'Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer', *The Lancet* **365**, 671–679.

Wei, L. (1992), 'The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis', *Statistics in Medicine* **11**, 1871–1879.

Wileyto, E., Li, Y. & Chen, J., H. D. (2013), 'Assessing the fit of parametric cure models', *Biostatistics* **14**, 340–350.

Xu, J, . & Peng, Y. (2014), 'Nonparametric cure rate estimation with covariates', *Canadian Journal of Statistics* **42**, 1–17.

Yakovlev, A., Asselain, B., Bardou, V., Fouquet, A. & Hoang, T. e. a. (1993), A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, *in* 'Biometrie et Analyse de Aonnées Spatio-Temporelles', chapter 11, pp. 66–83.

Yakovlev, A., Tsodikov, A. & Asselain, B. (1996), Stochastic models of tumor latency and their biostatistical applications, *in* 'Mathematical Biology and Medicine', Vol. 1, World Scientific, Singapore.

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

Yin, G. & Ibrahim, J. (2005), 'Cure rate model : a unified approach', *Canadian Journal of Statistics* **33**, 559–570.

Yu, B. & Tiwari, R. (2012), 'A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative to survival data', *Canadian Journal of Statistics* **40**, 40–54.

Yu, M., Taylor, J. & Sandler, H. (2008), 'Individual prediction in prostate cancer studies using a joint longitudinal survival-cure model', *Journal of the American Statistical Association* **103**, 178–187.

Zeng, D., Yin, G. & Ibrahim, J. (2006), 'Semiparametric transformation for survival data with a cure fraction', *Journal of the American Statistical Association* **101**, 670–684.

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

Zhang, Y. & Shao, Y. (2018), 'Concordance measure and discriminatory accuracy in transformation cure models', *Biostatistics* **19**, 14–26.

Zhao, Y., Lee, A., Yau, K., Burke, V. & McLachlan, G. (2009), 'A score test for assessing the cured proportion in long-term survivor mixture model', *Statistics in Medicine* **28**, 3454–3466.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**, 1418–1429.

List of Figures

List of Tables

Doctoral Dissertations of the Faculty of Economics and Business

A list of doctoral dissertations from the Faculty of Economics and Business can be found at the following website:

http://www.kuleuven.be/doctoraatsverdediging/archief.htm