Relaxing Assumptions on the Censoring Mechanism in Survival Link-Based Additive Models

DOCTORAL THESIS

Author:

Robinson Dettoni

Supervisors:

Prof. Giampiero Marra

Prof. Rosalba Radice

A thesis submitted for the Degree of Doctor of Philosophy in the

Faculty of Mathematical and Physical Sciences

Department of Statistical Science

University College London

September 2020

Declaration

I, Robinson Dettoni, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Dedicated to

My Parents, Wife & Daughter

¿Robinsón por qué volviste de tu isla?
De la isla de tus obras y tus sueños privados
La isla de ti mismo rica de tus actos
Sin leyes ni abdicación ni compromisos
Sin control de ojo intruso
Ni mano extraña que rompa los encantos
¿Robinson cómo es posible que volvieras de tu isla?
Extracto de Altazor, Canto I,
Vicente Huidobro

Robinson why did you come back from your island?
From the island of your works and your private dreams
The island of your self rich with your labors
With no laws no betrayals no compromise
Uncontrolled by intruding eyes
Or the foreign hand that breaks the spell
Robinson how could you come back from your island?
Excerpt from Altazor, Canto I,
Vicente Huidobro
Translated by Eliot Weinberger

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors Prof. Giampiero Marra and Prof. Rosalba Radice for their unwavering support from the beginning of this PhD programme. They are brilliant academics, but most importantly, they are great people. It has been an honour to work with you both. Once again, many thanks to Giampiero for his guidance, care and help for the successful completion of this thesis.

I am very proud to have studied in the Department of Statistical Science at UCL. It was a challenging but a beautiful experience. In particular, I am deeply grateful to Panagiota Filippou for her help during the first year of my PhD study and Prof. Gianluca Baio for his useful comments in my upgrade exam.

I want to thank to Prof. Aidan O'Keeffe and Prof. Angela Noufaily for their insightful comments and suggestions during my PhD viva, I really enjoyed discussing my thesis with you both.

Studying a PhD in Statistical Science involves a strong mathematical background, in this sense I am very grateful to Prof. Eugenio Saavedra, Prof. Enrique Reyes, Prof. Humberto Prado, Prof. Pedro Ubilla and Boris Brayovic for their help and to encourage me to finish my MSc in Mathematics at Universidad de Santiago de Chile (USACH). I would like to offer my special thanks to Prof. Costante Bellettini from UCL for his mathematical technical support on my study.

I would like to extend my sincere thanks to Jorge Friedman and Carlos Yévenes to encourage me during my career to accomplish a PhD study. I am deeply grateful to Cristian Cespedes for his constant support at every stage of my PhD and for sharing good times with me around the world. I would like to thank to Javier Espinosa for the great coffee breaks in London, during which we had discussions about almost everything.

Outside of work, I want to thank to my lovely wife, Pía Loreto, for her love,

patience and continuous support, specially during the hardest moments of my PhD studies. Without her light and joy, I could not have finished this piece of work. I also want to thank to my children Martina and Matteo, they have been the source of my inspiration in London. I am very grateful to Carmen Gloria Parra for her help and continuous support to me and my family during this project.

I am eternally grateful to my parents, brother and sisters for their love and support throughout my life. Many thanks to my mother and my father to believe in me, and constantly pushing me to be a better person and to follow my dreams.

Finally, I am also grateful to USACH and to the ANID (Agencia Nacional de Investigación y Desarrollo) in Chile for the financial support they provided me with during my PhD studies.

Abstract

Survival models are frequently encountered in applications. In these models, the response of interest, the time until a particular event occurs, is often right censored. Most estimation methods assume that the event time and the censoring time are stochastically independent and non-informative conditional on covariates. However, these assumptions may be questioned. The aim of this thesis is to relax these assumptions in a class of flexible parametric survival models, called survival link-based additive models.

The assumption of non-informative censoring is relaxed by assuming that the marginal survival functions of the event and censoring times have parameters in common. In particular, we provide evidence on the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart. The independence assumption is relaxed by modelling both the event time and the censoring time simultaneously using copula functions. We provide some preliminary arguments towards model identification although this topic is very challenging and requires more future work.

In these survival link-based additive models, the baseline functions are estimated non-parametrically by monotonic P-splines, whereas covariate effects are flexibly determined using additive predictors that allow for a vast variety of effects. Parameter estimation is reliably carried out within a penalised maximum likelihood framework with integrated automatic multiple smoothing parameter selection. We derive the $\sqrt{n}$-consistency and asymptotic normality of the estimators proposed in this thesis. Their finite sample performance are investigated via Monte Carlo simulation studies, and the approaches illustrated using two cases study based on infants hospitalised for pneumonia as well as prostate cancer data. The `R` package `GJRM` has been extended to incorporate the developments discussed in this thesis to facilitate transparent and reproducible research.

Key Words: additive predictor, informative censoring, dependent censoring, copula, identification, link function, penalised maximum likelihood estimation; survival data.

Impact statement

Survival models can be applied to many problems in different fields. For example, Biostatistics, Demography, Operation Research, Health Insurance and Social Sciences are examples of applied areas in which survival models appear frequently. However, most estimation methods assume that the event and censoring times are stochastically independent and non-informative conditional on covariates. Therefore, given that the assumptions of non-informative censoring and independence are often made for convenience and considering that models and methods introduced in this thesis have been implemented in the `R` package `GJRM` (Marra & Radice, 2020b), the proposed methodology is likely to appeal the wider audience, both inside and outside academia, wishing to estimate possibly more realistic survival models or at least assess whether allowing for informative and dependent censoring can produce more plausible results.

On the other hand, people working on industry or academia can use the methods developed here as a starting point to build new methodologies in survival analysis, or to apply them in their own field of expertise to solve real problems. Finally, the developments contained in this thesis have been collected in the following papers:

- Dettoni R, Marra G, Radice R (2020), Copula Link-Based Additive Models for Dependent Right-Censored Event Time Data. (*Working paper*).

- Dettoni R, Marra G, Radice R (2020), Generalized Additive Survival Models with Informative Censoring. *Journal of Computational and Graphical Statistics*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims of the thesis

Survival models are frequently encountered in applications. In these models, the response of interest, the time until a particular event occurs, is often right censored. This means that only the lower bound between the event time, T_1, and the censoring time, T_2, is recorded. Because of this, observations alone can not provide direct information on the event of interest unless some assumptions are made. In the latent survival time approach (e.g., Crowder, 1991), standard modelling techniques assume that the observed and unobserved parts of the data are related via means of the random variables $y = \min(T_1, T_2) \in \mathbb{R}^+$ and $\delta = I(T_1 < T_2)) \in \{0, 1\}$, where I is the usual indicator function.

Within this framework, most estimation methods assume that T_1 and T_2 are stochastically independent and non-informative conditional on covariates (e.g., Wu & Witten, 2019; Ma et al., 2014; Scheike & Zhang, 2003; Younes & Lachin, 1997; Cox, 1972). However, these assumptions may be questioned.(e.g., Dettoni et al., 2020; Deresa & Van Keilegom, 2019; Xu et al., 2018; Lu & Zhang, 2012; Li & Peng, 2015; Chen, 2010; Huang & Zhang, 2008; Zheng & Klein, 1995; Emoto

& Matthews, 1990; Koziol & Green, 1976). If the event and censoring times are assumed to be dependent, then survival models accounting for this feature of the data face a problem of identification. In general, without additional assumptions, it is not possible to identify the survival distribution from the censored data alone or testing whether the censoring and survival mechanisms are independent (Tsiatis, 1975; Cox, 1959).

Censoring is informative when the censoring times contain information on the parameters of the distribution of the event variable (Lagakos, 1979). In particular, let us write the conditional probability density functions for the event and censored times as $f_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1)$ and $f_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_2)$, where $\mathbf{z}_1$ and $\mathbf{z}_2$ are vectors of covariates of dimensions p_1 and p_2 respectively. If the vector of parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ have components in common then censoring is informative. In this case, the observable data $(y, \delta, \mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^+ \times \{0,1\} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ provide sufficient information to identify the marginal survival functions of T_1 and T_2 (Berman, 1963).

In this thesis, we will extend a family of Survival Link-Based Additive Models (e.g., Liu et al., 2018; Royston & Parmar, 2002; Shen, 1998; Younes & Lachin, 1997) by relaxing the assumptions of independence and non-informative censoring. Our approach is based on the flexible and tractable proposal of Marra & Radice (2020a), which compared with other parametric and non-parametric models (e.g., Deresa & Van Keilegom, 2019; Lu & Zhang, 2012; Chen, 2010; Zheng & Klein, 1995; Emoto & Matthews, 1990; Koziol & Green, 1976) has several advantages. In particular, it can flexibly determine in a data driven manner the functional shapes of the baseline and covariate effects, avoiding the need for numerical integration, and easily allows for time-dependent effects via smooth interaction terms.

To deal with informative right censoring we assume that T_1 and T_2 are stochastically independent, then we propose a model where the parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ in the survival functions $S_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1)$ and $S_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_2)$ have components in common.

On the other hand, to account for dependent censoring, we propose to model the strength of the association between the event and censoring times via the copula structure $C[S_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1), S_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_2);\theta]$, whose dependence parameter, θ, is estimated from the data, and $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ do not have parameters in common.

As in Marra & Radice (2020a), in both models, baseline functions are non-parametrically estimated using monotonic P-splines and covariate effects are flexibly determined using additive predictors. Model fitting is based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters. The performance of the proposals are demonstrated through Monte Carlo simulation studies and relevant empirical applications. All the models and methods introduced in this thesis have been implemented in the `R` package `GJRM` (Marra & Radice, 2020b) to allow for transparent and reproducible research.

1.2 Outline

This thesis is organized as follows. In Chapter 2, a review of the essential concepts of survival analysis is presented. In particular, we will discuss the main quantities used in survival modelling such as the survival, hazard and cumulative hazards functions. Then, the crucial problem of censoring and their causes are analysed, along with a summary of univariate survival models. The last part of this chapter focuses on the independent and non-informative censoring assumptions.

In Chapter 3, we present a summary of models that permit not only for different assumptions about the nature of the covariate effects on the survival time, but also where the baseline hazard and survival functions can be modelled in a flexible way. Then, in the last part of the chapter, the survival link-based additive models are discussed in detail

In Chapter 4, we introduce the informative right censoring model, where the penalized log-likelihood estimation approach, and the $\sqrt{n}$-consistency and asymp-

totic normality of the non-informative and informative estimators are provided. Then the effectiveness of the proposed methodology is explored by means of a simulation study, and illustrated on data about infants hospitalised for pneumonia.

In Chapter 5, we introduce a flexible copula regression survival model that accounts for administrative and dependent right censoring, and provide some preliminary identification arguments. Parameter estimation as well as the $\sqrt{n}$-consistency and asymptotic normality of the dependent estimator are also discussed. In the last sections, the finite sample properties of the estimator are investigated via a Monte Carlo simulation study and the proposal illustrated using prostate cancer data.

Finally, in Chapter 6, we give a summary of the main results, where some related open topics for further work are also discussed.

Chapter 2

Preliminary Concepts

In this chapter, a review of the essential concepts in survival analysis is provided. In general, the response variable, the time until a specific event occurs, can be represented by many functions. However, three of them are of considerable importance in applications: the survival function, the hazard function and the cumulative hazard function. In what follows, these functions and the interrelations among them are presented when the response of interest, T, is distributed continuously over $\mathbb{R}^+$. Since the methods proposed in this thesis were built assuming that T is a continuous variable, the survival, hazard and cumulative hazard functions for the discrete case are presented in Appendix A.1. The product integral representation, which incorporates discrete and continuous data at the same time, is also discussed briefly in Appendix A.2. Then, the crucial problem of censoring and their causes will be discussed. Next, a general summary of univariate survival models will be given, where we focus on relevant splines-based methods. The last part of this chapter focuses on the independent and non-informative censoring assumptions.

2.1 Time to event functions

Let T be a non-negative random variable defined to represent the time until an specific event occurs for an individual. In particular, time can be measured in years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs. This event can be, for example, death because of a disease, relapse from remission, recovery or any designated experience of interest that may occur to an individual. In this section, it is assumed that T, the outcome of interest, is distributed continuously over $\mathbb{R}^+$.

The survival function of T, denoted by $S_T(t)$, represents the probability of surviving past time t. This can be defined as

$$S_T(t) = P(T > t). \tag{2.1}$$

In particular, $S_T(0) = 1$, $\lim_{t\to\infty} S_T(t) = 0$ and $S_T(t) = 1 - F_T(t)$, where $F_T(t)$ is the cumulative distribution function. If $F_T(t)$ is differentiable, then $f_T(t) = -\dfrac{dS_T(t)}{dt}$ and $S_T(t) = \displaystyle\int_t^\infty f_T(u)du$, where $f_T(t)$ is the usual probability density function for T, with $f_T(t) > 0$ and $\displaystyle\int_0^\infty f_T(t)dt = 1$. Furthermore, if $\tau_1 \leq \tau_2$, then $S_T(\tau_1) \geq S_T(\tau_2)$. This implies that $S_T(t)$ declines monotonically.

On the other hand, for all $\varepsilon > 0$, the hazard function of T, denoted by $\hbar_T(t)$, can be defined as

$$\hbar_T(t) = \lim_{\varepsilon\to 0} \frac{P(t < T \leq t + \varepsilon | T > t)}{\varepsilon}. \tag{2.2}$$

For each t, $\hbar_T(t)$ represents the instantaneous rate at which subjects experience the event of interest, given that they have survived up to time t. Moreover, $\hbar_T(t) \geq 0$ for all $t \geq 0$ and $\displaystyle\int_0^\infty \hbar_T(t)dt = \infty$. The hazard function can also be written as

$$\hbar_T(t) = \frac{f_T(t)}{S_T(t)}. \tag{2.3}$$

This can be proved by noting that the numerator of equation (2.2) can be expressed as $P(t < T \leq t + \varepsilon | T > t) = \dfrac{P(t < T \leq t + \varepsilon)}{P(T > t)} = \dfrac{F_T(t+\varepsilon) - F_T(t)}{1 - F_T(t)}$. Then, by taking the limit as ε approaches to zero from above, and dividing $\dfrac{F_T(t+\varepsilon) - F_T(t)}{1 - F_T(t)}$ by ε, we obtain $\hbar_T(t) = \lim_{\varepsilon \to 0} \dfrac{F_T(t+\varepsilon) - F_T(t)}{\varepsilon} \dfrac{1}{1 - F_T(t)} = \dfrac{f_T(t)}{1 - F_T(t)} = \dfrac{f_T(t)}{S_T(t)}$, as required.

Furthermore, since $f_T(t) = -\dfrac{dS_T(t)}{dt}$, another useful expression for the hazard function is

$$\hbar_T(t) = -\frac{1}{S_T(t)} \frac{dS_T(t)}{dt} = -\frac{d \log S_T(t)}{dt}. \tag{2.4}$$

On the other hand, the cumulative hazard function can be defined as

$$\mathcal{H}_T(t) = \int_0^t \hbar_T(u) du. \tag{2.5}$$

This function measures the total amount of risk that has been accumulated up to time t. Moreover, integrating (2.4) yields $\log S_T(t) = -\int_0^t \hbar_T(u) du$. Then, solving for $S_T(t)$, we obtain

$$S_T(t) = \exp\left[-\int_0^t \hbar_T(u) du \right]. \tag{2.6}$$

Finally, using (2.5) and (2.6), we have

$$S_T(t) = \exp\left[-\mathcal{H}_T(t) \right]. \tag{2.7}$$

Therefore, as shown in (2.7), $S_T(t)$ can also be written as a function of $\mathcal{H}_T(t)$. The survival, hazard and cumulative hazard functions when T is distributed continuously over $\mathbb{R}^+$ are summarized in Table 2.1.

For example, suppose that T follows an exponential distribution, then $\hbar_T(t) = \alpha > 0$, $f_T(t) = \alpha \exp^{-\alpha t}$, $S_T(t) = \exp^{-\alpha t}$ and $\mathcal{H}_T(t) = \alpha t$. However, if T has a Weibull distribution, then $\hbar_T(t) = \alpha\gamma(\alpha t)^{\gamma - 1}$ for all t. This is monotone decreasing if $0 < \gamma < 1$, increasing if $\gamma > 1$, and reduces to the constant exponential hazard

Function	Definition	Relationships
$S_T(t)$	$S_T(t) = P(T > t)$	$S_T(t) = \exp\left[-\mathcal{H}_T(t)\right]$
$\hbar_T(t)$	$\hbar_T(t) = \lim_{\varepsilon \to 0} \dfrac{P(t < T \leq t + \varepsilon \mid T > t)}{\varepsilon}$	$\hbar_T(t) = \dfrac{f_T(t)}{S_T(t)}$
$\mathcal{H}_T(t)$	$\displaystyle\int_0^t \hbar_T(u)du$	$\mathcal{H}_T(t) = -\ln S_T(t)$

Table 2.1: Summary of the survival, hazard and cumulative hazard functions when T is distributed continuously over $\mathbb{R}^+$. The last column shows the relationships between them.

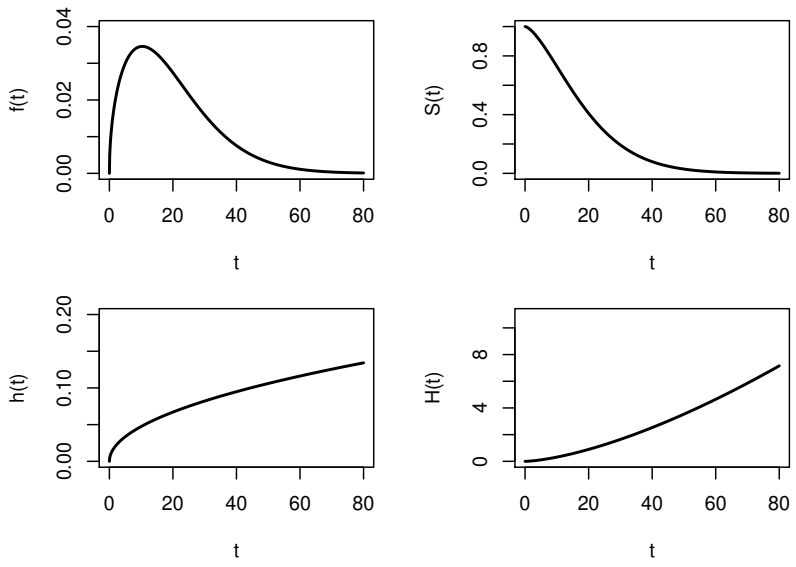

Figure 2.1: $f_T(t)$, $S_T(t)$, $\hbar_T(t)$ and $\mathcal{H}_T(t)$ when T follows a Weibull distribution with parameters $\alpha = 0.01$ and $\gamma = 1.5$

if $\gamma = 1$. Moreover, $f_T(t) = \gamma(\alpha t)^{\gamma-1}\exp[-(\alpha t)^\gamma]$, $S_T(t) = \exp[-(\alpha t)^\gamma]$ and $\mathcal{H}_T(t) = (\alpha t)^\gamma$. The graphs for $f_T(t)$, $S_T(t)$, $\hbar_T(t)$ and $\mathcal{H}_T(t)$ when $\alpha = 0.01$ and $\gamma = 1.5$ are shown in Figure 2.1.

2.2 Censoring

Censoring is an unavoidable problem in survival analysis. This occurs when the response of interest, T, can not be totally observed. In general, the occurrence of censoring can be explained by the following reasons. First, individuals can be

censored because the study ends before they have experienced the event of interest. This situation is typically called administrative censoring. Second, individuals may be censored because they are lost to follow up or withdraw from the study. Finally, censoring can also be generated by competing risks, that is, the occurrence of another event which precludes the main event of interest from occurring (Kalbfleisch & Prentice, 2002).

Survival data may be right-censored, left-censored, or interval-censored. Right censoring arises when the true survival time becomes incomplete at the right side of the follow-up period. For this data, the complete survival time, has been cut off (censored) at the right side of the observed survival time. Left-censoring occurs when the true survival time of an individual is less than or equal to its observed survival time. For example, suppose that a group of individuals are followed until they become HIV positive, and an event is recorded when a individual first tests positive for the virus. However, the exact time of first exposure to the virus may not be known, and thus it is not known exactly when the event took place. Therefore, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the individual's test is positive. Finally, interval censoring occurs if the true (but unobserved) survival time of an individual is within a certain known specified time interval. For example, suppose that an individual may have had two HIV tests, and he or she was HIV negative at the time t_1 of the first test and HIV positive at the time t_2 of the second test. In this case the true survival time occurred after time t_1 and before time t_2. Therefore, the subject is interval-censored in the time interval (t_1, t_2) (Kleinbaum & Klein, 2010). In this thesis we will only focus on right censoring.

2.3 Univariate survival models

One of the most widely used approaches to estimate the survival function when no covariates are present is the nonparametric estimator proposed by Kaplan & Meier (1958). This estimator is very useful for descriptive purposes and it is also the basis to develop more advanced models. Although, it is often insightful to know the shape of the hazard or survival functions, in most applications it is necessary to incorporate covariates when modelling survival time.

In survival analysis, regression models are crucial and the existing literature is vast (e.g., Kalbfleisch & Prentice, 2002; Andersen & Keiding, 2006). The proportional hazards model of Cox (1972) is by far the most used regression technique to model survival data. In this model, the hazard function is $\hbar_T(t|\mathbf{z}) = \hbar_0(t)\exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})$, where $\hbar_0(t)$ is an arbitrary unspecified baseline hazard function, $\mathbf{z}$ is a covariate vector and $\boldsymbol{\gamma}$ is a vector of parameters. The exponential link function makes the covariate effects multiplicative and assures non-negative rates. Due to no structure being imposed on $\hbar_0(t)$, the proportional hazards model is remarkably flexible. The proportional hazards assumption corresponds to assuming that the hazard ratio of two subjects with different time-constant covariate vectors is constant over time. Estimation of $\boldsymbol{\gamma}$ can be undertaken using the partial log-likelihood estimator. In particular, Cox (1972) showed that his proposed estimator is consistent and asymptotically normally distributed. The proportional hazards model can be extended to include time-dependent covariates (Kalbfleisch & Prentice, 2002).

Another approach is to consider failure time models (Cox, 1972), where $\log(T) = \mathbf{z}^\mathsf{T}\boldsymbol{\gamma}+\varepsilon$. In this model, the effects of the covariates are defined to act multiplicatively on T, or additively on $\log(T)$. The hazard function can be written as $\hbar_T(t|\mathbf{z}) = \hbar_0[t\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})]\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})$, where it is easily seen that the effects of the vector of covariates $(\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma}))$ is multiplicative on t rather than on the baseline hazard

function. More specifically, there is an acceleration of $\hbar_0(t)$ if $\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma}) > 1$ and a deceleration if $\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma}) < 1$. Different continuous distributions on $(-\infty, \infty)$ for ε lead to different accelerated failure time models.

The proportional odds model (McCullagh, 1980; Bennett, 1983) is structurally similar to the proportional hazard model, and can be used in similar situations. In the proportional hazards model the hazard rates for different individuals have a constant ratio to each other, while in the proportional odds model they converge with time. This can be more useful than the notion of constant hazard ratio when initial effects disappear with time (Bennett, 1983). The model can be written as $\left[\frac{1 - S_T(t|\mathbf{z})}{S_T(t|\mathbf{z})}\right] = \left[\frac{1 - S_T(t)}{S_T(t)}\right] \exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})$, and estimation of model's parameters can be obtained by maximising the full likelihood (e.g., Rossini & Tsiatis, 1996; Shen, 1998; Yang & Prentice, 1999).

The additive models are very easy to work with and the survival function can be easily estimated (Aalen, 1989; McKeague & Sasieni, 1994; Cortese & Scheike, 2008; Cortese et al., 2010). In these models, the hazard function can be written as $\hbar_T(t|\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^\mathsf{T}\boldsymbol{\gamma}_1(t) + \mathbf{z}_2^\mathsf{T}\boldsymbol{\gamma}_2$, where $\boldsymbol{\gamma}_1(t)$ is a vector of functions that depend on time. These models have the drawback that they could lead to negative hazards in some time periods (Scheike & Zhang, 2003). The combined Cox–Aalen model (e.g., Scheike & Zhang, 2002; Zahl, 2003; Martinussen & Scheike, 2002; Shang & Wang, 2017), with hazard function $\hbar_T(t|\mathbf{z}_1, \mathbf{z}_2) = [\mathbf{z}_1^\mathsf{T}\boldsymbol{\gamma}_1(t)] \exp(\mathbf{z}_2^\mathsf{T}\boldsymbol{\gamma}_2)$, provides a flexible extension of the proportional hazards model. In this approach, the covariates are partitioned into those that lead to relative risk, $\mathbf{z}_2$, and those where additional flexibility is needed, $\mathbf{z}_1$. Estimation of model's parameters can be carried out using an approximate maximum likelihood method or the generalized method of moments.

Frailty models provide a suitable way to introduce random effects to account for unobserved heterogeneity. In its simplest form, a frailty is an unobserved random

factor that modifies multiplicatively the hazard function (e.g., Hougaard, 1984, 1986). For example, the proportional hazard frailty model with a multiplicative $\varepsilon > 0$ (frailty) is given by $\hbar_T(t|\mathbf{z}, \epsilon) = \hbar_0(t) \exp(\mathbf{z}^\top \boldsymbol{\gamma})\varepsilon$. The frailty ε is a random variable varying over the population which decreases ($\varepsilon < 1$) or increases ($\varepsilon > 1$) the individual risk. Although, the multiplicative heterogeneity assumption is particularly restrictive, it is mathematically convenient and more attractive than an additive error, which can not assure non-negativity of T. A standard approach involves assuming a distribution for ε, and then deriving the marginal distribution of T (Hougaard, 1995).

Splines-based models are general enough to include many data structures and they are easy to interpret due to their parametric nature. In these models, the hazard and survival functions are approximated using splines functions which yield smooth estimates and are intermediate in structure between parametric and non-parametric models. These can be defined as piecewise polynomial of degree q, the pieces join in the so called knots and fulfil continuity conditions for the function itself and the first $q - 1$ derivatives (De Boor et al., 1978). The use of splines functions to approximate the baseline hazard function was first introduced by Anderson & Senthilselvan (1980) and Whittemore & Keller (1986) in the context of a proportional hazard model and fixed knots. For example, Anderson & Senthilselvan (1980) proposed a quadratic splines function with discontinuities in the knots which is estimated using the penalized likelihood approach, while Whittemore & Keller (1986) use a non-parametric likelihood estimator to estimate the survival function for right censored data.

Gray (1992) uses cubic B-splines functions and a proportional additive hazard model to study the effects of covariates on the hazard for time to recurrence of breast cancer patients. For estimation, he uses a penalised partial likelihood method and tests statistics that are similar to those used in standard likelihood

analysis (Therneau et al., 1990). Hess (1994), Rosenberg (1995) and Herndon & Harrell (1995) also proposed a cubic spline-based approach to approximate the hazard function, while Kooperberg et al. (1995) employ linear splines functions and their tensor products to estimate the log-hazard function conditional to covariates. In all of these works, estimation of model's parameters is carried out by using the maximum likelihood method.

2.4 Censoring assumptions

In all works discussed in the previous section, it was assumed that the censoring mechanism is independent and non-informative. In this thesis, for independent censoring we mean that T_1 and T_2 are stochastically independent (e.g., Kalbfleisch & Prentice, 2002). However, this definition differs from the one generally used in the multiplicative intensity model discussed, for example, in Aalen et al. (2008) and Andersen et al. (2012), where censoring is said to be independent if the hazard rate of the event of interest for the censored observations is equal to the hazard rate for the uncensored ones.

On the other hand, censoring is informative when the censoring times, T_2, contain information on the parameters of the distribution of the event variable, T_1 (e.g., Lagakos, 1979; Koziol & Green, 1976). For example, let us write the survival functions for the event and censored times as $S_{T_1}(t|\mathbf{z}_1;\boldsymbol{\gamma}_1)$ and $S_{T_2}(t|\mathbf{z}_2;\boldsymbol{\gamma}_2)$. If the vector of parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ have components in common then censoring is informative.

In addition to these assumptions, and for estimation purposes, most of the standard modelling techniques (e.g., Dettoni et al., 2020; Marra & Radice, 2020a; Deresa & Van Keilegom, 2019) use the latent survival time approach (e.g., Crowder, 1991). Under this approach, it is assumed that the observed and unobserved parts of the data are related via means of the random variables $y = \min(T_1, T_2) \in \mathbb{R}^+$

and $\delta = I(T_1 < T_2)) \in \{0, 1\}$, where I is the usual indicator function.

In the next chapter, we will focus on a class of flexible parametric survival models, known as Survival Link-Based Additive Models.

Chapter 3

Survival Link-Based Additive Models

This chapter presents a summary of models that allow for different assumptions about the nature of the covariate effects on the survival time, and where the baseline hazard and survival functions can be modelled in a flexible form. The general ideas of these models are employed to build survival link-based additive models, which are analysed in the last part of the chapter.

3.1 Introduction

Generalized additive models have the form $g[\eta(\mathbf{z})] = \sum_{j=1}^{J} f(z_j)$, where $g(\cdot)$ is a link function and $f(z_1), \ldots, f(z_J)$ are smooth functions (e.g., Hastie & Tibshirani, 1986, 1990; Wood, 2017). In these models, for example, $g[\eta(\mathbf{z})]$ might represent the logistic transformation of the probability $P(y = 1|\mathbf{z})$ in a logistic regression or the regression function in a multiple regression. In fact, the generalized additive models extend the whole family of generalized linear models $g[\eta(\mathbf{z})] = \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ (Nelder & Wedderburn, 1972), where $g[\eta(\mathbf{z})]$ is some transformation of the regression function. Analogously, the conditional survival function can be modelled using

smooth functions for time, covariates and time–covariate interactions via the model $g[S_T(t|\mathbf{z})] = \sum_{j=1}^{J} f(x_j)$, where $f(x_j)$ being $f(t)$, $f(z)$ or $f(t,z)$ (e.g., Liu et al., 2018; Marra & Radice, 2020a). Models like $g[S_T(t|\mathbf{z})] = \sum_{j=1}^{J} f(x_j)$ belong to the class of survival link-based additive models, and are the subject of this chapter. In particular, we will focus on families of models that allow for different assumptions about the nature of the covariate effects on the survival time, and also give flexibility in modelling the baseline hazard and survival functions. Then, we will present a summary of a particular class of these models: the survival link-based additive models.

3.2 Summary of Survival Link-Based Additive Models

In practical statistical modelling, we need flexibility to model not only the baseline hazard and survival functions, but also the effects of covariates on the survival time. Some models make specific assumptions about the nature of this effect. For example, the proportional hazards and the proportional odds models specifically assume that the covariates act multiplicatively on the baseline hazard, or the baseline odds of survival, respectively. Since any specific assumption will not always hold, families of models, such as the proposed by Etezadi-Amoli & Ciampi (1987), Doksum & Gasko (1990) and Cheng et al. (1995) are particularly appealing.

Specifically, Etezadi-Amoli & Ciampi (1987) proposed a flexible model where the baseline hazard function is approximated with a quadratic splines function and model's parameters are estimated using the maximum likelihood method. Their conditional hazard function is modelled as $\hbar_T(t|\mathbf{z}) = \exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma_1})\hbar_0[\exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma_2})t]$, where $\hbar_T(t|\mathbf{z})$ reduces to the proportional hazards model if $\boldsymbol{\gamma_2} = \mathbf{0}$ and to the accelerate failure time model when $\boldsymbol{\gamma_1} = \boldsymbol{\gamma_2}$. In the general case, $\mathbf{z}$ affects the

survival by changing both the time scale by the factor $\exp(\mathbf{z}^\top \boldsymbol{\gamma_2})$, and the scale in which the hazard is measured by the factor $\left[\frac{\exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma_1})}{\exp(\mathbf{z}^\mathsf{T}\boldsymbol{\gamma_2})}\right]$.

On the other hand, the correspondence between models in binary data analysis and continuous time survival analysis, and the correspondence between continuous time survival models and linear transformation models have been spelled out by Doksum & Gasko (1990). This can be seen by noting first that, in binary regression analysis, the parameter of interest is $\varphi(\mathbf{z}) = P[I(\mathcal{A}) = 1|\mathbf{z}]$, where $\varphi(\mathbf{z})$ is the probability that an individual has a certain characteristic, $\mathcal{A}$, conditional to a covariate vector $\mathbf{z}$, and $I(\cdot)$ is the indicator function. Thus, if $t > 0$ is fixed and $\mathcal{A} = \mathcal{A}_t = T \leq t$, then $F_T(t|\mathbf{z}) = P(T \leq t|\mathbf{z})$ is the same as $\varphi(\mathbf{z})$, since $F_T(t|\mathbf{z}) = P[I(\mathcal{A}_t) = 1|\mathbf{z}] = \varphi(\mathbf{z})$. In addition, let us define the linear transformation as

$$h(T) = \mathbf{z}^\mathsf{T}\boldsymbol{\gamma} + \varepsilon, \tag{3.1}$$

where $h(\cdot)$ is an increasing continuous function on some domain $\mathcal{D}$, and ε is a random error with continuous distribution function Υ with support $(-\infty, \infty)$ (Doksum & Gasko, 1990). Using this notation, the logistic regression model (Berkson, 1944) in binary data analysis can be written as $\log\left[\frac{\varphi(\mathbf{z})}{1-\varphi(\mathbf{z})}\right] = \mathbf{z}^\mathsf{T}\boldsymbol{\gamma}$ or $\varphi(\mathbf{z}) = \Upsilon_{\text{L}}(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})$, where $\Upsilon_{\text{L}}(\mathtt{w}) = [1 + \exp(-\mathtt{w})]^{-1}$ is the logistic distribution. The corresponding proportional odds model in continuous time survival data (Bennett, 1983) is

$$\log\left[\frac{F_T(t|\mathbf{z})}{1 - F_T(t|\mathbf{z})}\right] = -\mathbf{z}^\mathsf{T}\boldsymbol{\gamma} + \log \mathcal{O}(t), \tag{3.2}$$

where $\mathcal{O}(t)$ is an increasing function on $(0, \infty]$ with $\mathcal{O}(0) = 0$ and $\mathcal{O}(\infty) = \infty$. In binary analysis, where t is fixed, $\mathcal{O}(t)$ is constant and $\log \mathcal{O}(t)$ is absorbed into the intercept term of $\mathbf{z}^\mathsf{T}\boldsymbol{\gamma}$. Let us define the odds function as $\mathcal{O}(t|\mathbf{z}) = F_T(t|\mathbf{z})[1 - F_T(t|\mathbf{z})]^{-1}$. Then, using (3.2), $\mathcal{O}(t|\mathbf{z}) = \mathcal{O}(t)\exp(-\mathbf{z}^\mathsf{T}\boldsymbol{\gamma})$, where $\mathcal{O}(t)$ is the baseline odds function. Therefore, model (3.2) can be written in logistic

form (Bennett, 1983) as

$$F_T(t|\mathbf{z}) = \Upsilon_{\text{L}}(\log \mathcal{O}(t) - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}), \quad t \geq 0. \tag{3.3}$$

It is easy to show that the proportional odds model for continuous survival data given by (3.2) is a linear transformation model of the form (3.1). In particular, note that

$$\begin{aligned} P[h(T) \leq t] &= P[\log \mathcal{O}(t) \leq t] \\ &= P[T \leq \mathcal{O}^{-1}(\exp(t))] = F_T[\mathcal{O}^{-1}(\exp(t))|\mathbf{z}], \end{aligned} \tag{3.4}$$

with $\mathcal{D} = [0, \infty)$, $h(t) = \log \mathcal{O}(t)$ and $\Upsilon = \Upsilon_{\text{L}}$. Using (3.4) and (3.3) we arrive at $P[h(T) \leq t] = \Upsilon_{\text{L}}(t - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$ which is another way of writing (3.1) when $\Upsilon = \Upsilon_{\text{L}}$ (Prentice, 1978; Dabrowska & Doksum, 1988a,b; Doksum & Gasko, 1990).

These ideas can be used for the proportional hazards model (Cox, 1972). More precisely, since $\hbar_T(t|\mathbf{z}) = f_T(t|\mathbf{z})[1 - F_T(t|\mathbf{z})]^{-1}$, then $\mathcal{H}_T(t|\mathbf{z}) = -\log[1 - F_T(t|\mathbf{z})]$. Therefore, $\hbar_T(t|\mathbf{z}) = \hbar_0(t)\exp(\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$ is equivalent to

$$\mathcal{H}_T(t|\mathbf{z}) = \mathcal{H}_T(t)\exp(-\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}), \tag{3.5}$$

where $\mathcal{H}_T(t) = \int_0^t \hbar_T(u)du$. Model (3.5) is equivalent to $\log[-\log[1 - F_T(t|\mathbf{z})]] = \log \mathcal{H}_T(t) - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$, which can also be written as

$$F_T(t|\mathbf{z}) = \Upsilon_{\text{E}}[\log \mathcal{H}_T(t) - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}], \tag{3.6}$$

where Υ_{E} is the extreme value distribution $\Upsilon_{\text{E}}(\mathtt{w}) = 1 - \exp[-\exp(\mathtt{w})]$. Moreover, if t is fixed, the constant $\log \mathcal{H}_T(t)$ can be absorbed in the intercept of $\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$, and the binary model can be expressed as $\log[-\log[1 - \varphi(\mathbf{z})]] = \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ or $\varphi(\mathbf{z}) = \Upsilon_{\text{E}}(-\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$. The connection between $\log[-\log[1 - \varphi(\mathtt{w})]]$ and the proportional hazards model

can be found in McCullagh (1980). Due to (3.6), the proportional hazards model is a linear transformation model of the form (3.1) with $\mathcal{D} = [0, \infty)$, $h(t) = \log \mathcal{H}_T(t)$ and $\Upsilon = \Upsilon_{\text{E}}$. The connection between Υ_{E} and the proportional hazards model was observed by Kalbfleisch (1978) and Prentice (1978).

The binary probit model (Bliss, 1935) is $\Phi^{-1}[\varphi(\mathbf{z})] = \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ or $\varphi(\mathbf{z}) = \Phi(-\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$, then for $(t \geq 0)$, the corresponding survival analysis model is $\Phi^{-1}[F_T(t|\mathbf{z})] = \Phi^{-1}[F_T(t)] - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ or $F_T(t|\mathbf{z}) = \Phi[\Phi^{-1}[F_T(t)] - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}]$. The corresponding linear transformation model has $\mathcal{D} = [0, \infty)$, $h(t) = \Phi^{-1}[F_T(t)]$ and $\Upsilon = \Phi$ (Peto & Peto, 1972; Prentice, 1976; Pettitt, 1982, 1983).

In general, a binary regression model can be written as $\Upsilon^{-1}[\varphi(\mathbf{z})] = -\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ or $\varphi(\mathbf{z}) = \Upsilon(-\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$ for a given continuous distribution function with support $(-\infty, \infty)$. Given $(t \geq 0)$, the corresponding survival model is $\Upsilon^{-1}[F_T(t|\mathbf{z})] = h(t) - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ or $F_T(t|\mathbf{z}) = \Upsilon_{\text{E}}[h(t) - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}]$, where $h(t)$ is an increasing continuous function on $[0, \infty)$ with $h(0) = -\infty$ and $h(\infty) = \infty$. Since $P[h(T) \leq t] = P[T \leq h^{-1}(t)] = F_T[h^{-1}(t)|\mathbf{z}] = \Upsilon(t - \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma})$, a natural choice for $h(t)$ is $\Upsilon^{-1}[F_T(t)]$. The equivalent linear transformation model is $h(T) = \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma} + \varepsilon$, with $\varepsilon \sim \Upsilon$. Finally, if $\mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ is replaced by some other function $f(\mathbf{z}, \boldsymbol{\gamma})$, the correspondence between the binary, survival, and transformation models would still be valid. These one-to-one correspondences were also discussed for other models, such as the gamma-logit model, the power family model and the log-linear model (Dabrowska & Doksum, 1988a,b; Doksum & Gasko, 1990).

In this line of analysis, Cheng et al. (1995) also considered a class of semi-parametric transformation models, which include the proportional hazards and proportional odds models as especial cases. Their model can be written as $g[S_T(t|\mathbf{z})] = h(t) + \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$, where $g(\cdot)$ is a known decreasing function, and $h(\cdot)$ is a unspecified strictly increasing function, which maps $[0, \infty)$ onto $(-\infty, \infty)$. If $g(S) = \log[-\log(S)]$, then $g[S_T(t|\mathbf{z})] = h(t) + \mathbf{z}^{\mathsf{T}}\boldsymbol{\gamma}$ reduces to the proportional

hazards model and to the proportional odds model when $g(S) = -\log\left[\frac{S}{1-S}\right]$. In their approach, $g[S_T(t|\mathbf{z})] = h(t) + \mathbf{z}^\mathsf{T}\boldsymbol{\gamma}$ is equivalent to the linear transformation model $h(T) = -\mathbf{z}^\mathsf{T}\boldsymbol{\gamma} + \varepsilon$, with distribution function $F_S(s) = 1 - g(s)^{-1}$. As in Doksum & Gasko (1990), if $F_S(s) = 1 - \exp[-\exp(s)]$ the linear transformation model is the proportional hazards model, while if $F_S(s)$ is the standard logistic distribution, then $h(T) = -\mathbf{z}^\mathsf{T}\boldsymbol{\gamma} + \varepsilon$ becomes in the proportional odds model. The parametric version of $h(T) = -\mathbf{z}^\mathsf{T}\boldsymbol{\gamma}+\varepsilon$, with $h(\cdot)$ specified up to a finite-dimensional parameter vector, has been discussed by Box & Cox (1964).

Within the parametric approach, Younes & Lachin (1997) proposed a flexible link-based model for survival analysis that makes use of an arbitrarily defined link function, $g(\cdot)$, to express the way by which the covariates act on the survival times. The link function can be controlled by an additional parameter, yielding families of models, such as the proportional hazards and the proportional odds models when $g(\cdot)$ is the parametrized link function $g_\vartheta(S) = \log\left[\frac{S^{-\vartheta}-1}{\vartheta}\right]$ of Aranda-Ordaz (1981). In particular, $\vartheta \to 0$ corresponds to the proportional hazard model and $\vartheta = 1$ to the proportional odds model. Their model is $g[S_T(t|\mathbf{z})] = g[S_0(t)] + \mathbf{z}^\mathsf{T}\boldsymbol{\gamma}$, where $S_0(t)$ is the baseline survival function, which is approximated by B-splines functions and determined by integration. Under the assumptions of independent and non-informative censoring, the parameters of the model are estimated by full maximizing likelihood. Other links, such as the probit link $g(S) = -\Phi^{-1}(S)$ (where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution), or other link families, such as the asymmetric family $g(S) = \dfrac{\vartheta - S}{S(1-S)}$ can also be used in this approach. A similar and computationally complex proposal was proposed by Shen (1998), where sieve maximum likelihood and monotone splines functions are used to estimate a general version of the proportional odds model. Within the model proposed by Younes & Lachin (1997), Royston & Parmar (2002) developed flexible parametric models based on the assumption of proportional hazards and

proportional odds scaling of covariate effects via the parametric link of Aranda-Ordaz (1981). In their proposal, $g[S_0(t)]$ is approximated using natural cubic splines functions and full likelihood is performed for estimation.

These models have several practical advantages. In particular, they are estimated by parametric maximum likelihood estimation, which allows for straightforward estimation and model comparisons; model formulation includes simple and computationally efficient expressions for the survival and hazard functions and several quantities of interest can be easily calculated from these models (for example, odds ratios, time-dependent differences and standardized survival) that are often difficult to estimate with non-parametric models. However, they also have some drawbacks, including the need for model selection for smoothing across time and covariates and restricted functional forms for the time effects (Liu et al., 2018; Marra & Radice, 2020a).

Liu et al. (2018) propose a family of survival link-based additive models that extend the parametric models proposed by Younes & Lachin (1997) and Royston & Parmar (2002). In particular, their model associates the conditional survival function $S_T(t|\mathbf{z};\boldsymbol{\gamma})$ with a linear predictor $\xi(t,\mathbf{z};\boldsymbol{\gamma})$ through a specific link function, $g(\cdot)$. In this approach, the design matrix $\mathcal{B}(t,\mathbf{z})$ is used to model the linear predictor, that is: $g[S_T(t|\mathbf{z};\boldsymbol{\gamma})] = \mathcal{B}(t,\mathbf{z})\boldsymbol{\gamma}$. In general, $\mathcal{B}(t,\mathbf{z})$ includes a baseline function for time or a stratified set of baseline functions for time, and covariates and interactions between time and covariates. Estimation is carried out using penalized log-likelihood, which prevents over-fitting and allow for straightforward estimation and model comparisons.

Finally, Marra & Radice (2020a) proposed a methodology to estimate joint survival models where two survival link-based additive models are modelled by a copula function, where all the model's parameters can be specified as functions of various types of covariate effects, and monotonic P-splines of transformations

of the baseline survival functions are utilized to provide coherent marginal survival fits. Under the assumptions of independent and non-informative censoring, their estimation approach consists of a carefully constructed optimization scheme that allows for the simultaneous penalized maximum likelihood estimation of the model's parameters as well as for stable and efficient automatic multiple smoothing parameter selection.

In the next chapter, we will use the approach of Marra & Radice (2020a) to extend the class of univariate survival link-based additive models by relaxing the assumption of non-informative censoring, but assuming that the censoring and the event times are stochastically independent.

Chapter 4

Survival Link-Based Additive Models with Informative Censoring

In this chapter we introduce a class of flexible survival models which account for the information provided by the censoring times. Survival functions are modelled using generalised survival or link-based functions models and baseline functions are estimated non-parametrically by monotonic P-splines. Covariate effects (such as linear, nonlinear, random and spatial) are flexibly determined using additive predictors. The performance of the proposed methodology is evaluated through a Monte Carlo simulation study and an empirical application on data about infants hospitalised for pneumonia. The relevant numerical computation can be easily undertaken using the function `gamlss()` in the `R` package `GJRM` (Marra & Radice, 2020b) (see Appendix B.5 for some software details). Both, the simulation study and the empirical application highlight the merits of the proposal.

4.1 Introduction

Most of the related estimation techniques assume that the censoring scheme is independent and non-informative conditional on covariates (e.g., Xue et al., 2018; Ma et al., 2014; Scheike & Zhang, 2003; Younes & Lachin, 1997; Cox, 1972). In many applications, however, these assumptions can at least be questioned (e.g., Xu et al., 2018, 2017; Li & Peng, 2015; Wang et al., 2015; Lu & Zhang, 2012; Huang & Zhang, 2008; Zeng et al., 2004; Zheng & Klein, 1995; Slud & Rubinstein, 1983). If the event and censoring times are assumed to be dependent, then survival models accounting for this feature of the data face a problem of identification. In general, without additional assumptions, it is not possible to identify the survival distribution from the censored data alone or testing whether the censoring and survival mechanisms are independent (Tsiatis, 1975; Cox, 1959). However, if censoring is informative, the observable data $(y, \delta) = \{\min(T_1, T_2), I(T_1 < T_2)\}$, where I is the usual indicator function, provide sufficient information to identify the marginal survival functions of T_1 and T_2 (Kalbfleisch & Prentice, 2002).

Although dependent censoring is a well studied problem in the survival analysis and competing risk literature (e.g., Emura & Chen, 2018; Crowder, 2012), the specific literature analysing the problem of informative censoring is scarce, even though ignoring it may have detrimental consequences on inferential conclusions (e.g., Siannis et al., 2005; Lu & Zhang, 2012). In a seminal work, Koziol & Green (1976) proposed an informative survival model where the hazard functions of T_1 and T_2 satisfy $\hbar_{T_2}(t) = p\hbar_{T_1}(t)$, for some constant $0 < p < 1$. Since this model did not incorporate covariates, it was further extended. For instance, Yuan (2005) introduced a semiparametric Cox model estimated via profile likelihood in which, for a given vector of covariates $\mathbf{z}$, $\hbar_{T_2}(t|\mathbf{z}) = \varrho(t, \mathbf{z}; \gamma)\hbar_{T_1}(t|\mathbf{z})$, where ϱ is a function known up to a finite-dimensional parameter, γ. The purpose of ϱ was to capture the possible information contained in the censoring times. Lu & Zhang (2012)

proposed a semi-parametric informative survival model where the baseline hazards are estimated non-parametrically and the covariate effects parametrically. In their approach, the hazard functions of T_1 and T_2 conditional on $\mathbf{z}$ are modelled using $\hbar_{T_\upsilon}(t|\mathbf{z}) = \hbar_{0,T_\upsilon}(t)\exp(\mathbf{z}^\mathsf{T}\boldsymbol{\varphi}_\upsilon)$, where $\mathbf{z}^\mathsf{T}\boldsymbol{\varphi}_\upsilon = \mathbf{z}_1^\mathsf{T}\boldsymbol{\gamma}_0 + \mathbf{z}_2^\mathsf{T}\boldsymbol{\gamma}_\upsilon$, for $\upsilon = 1, 2$.

The remainder of this chapter is organized as follow. In Section 4.2, we will develop a flexible, general and tractable survival modelling framework where the baseline functions are modelled non-parametrically via means of monotonic P-splines, covariate effects are flexibly determined using additive predictors, and informative censoring is accounted for. In Section 4.3, we propose a model fitting based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection. In this section, the $\sqrt{n}$-consistency and asymptotic normality of the non-informative and informative estimators are also provided, where we also show that the newly introduced informative estimator is more efficient than its non-informative counterpart. Confidence intervals and p-values are also provided in this section. Finally, in Sections 4.4 and 4.5, a Monte Carlo simulation study highlights the merits of the proposal, and the modelling framework is illustrated on data about infants hospitalised for pneumonia. The models and methods introduced in the article have been implemented in the `R` package `GJRM` (Marra & Radice, 2020b) to allow for transparent and reproducible research.

4.2 Methodology

In this thesis, only the case of right censored data is considered; the true event time is not always observed, in which case censoring (lower) times are observed. For individual i, where $i = 1, \ldots, n$ and n represents the sample size, let T_{1i} and T_{2i} denote the true event and censoring times. Let also $\mathbf{z}_{\nu i}^\top = (z_{\nu 1i}, \ldots, z_{\nu K_\nu i})$ be a vector of baseline covariates of dimension K_ν, where $\mathbf{z}^\top$ stands for the transpose of a

vector $\mathbf{z}$, $\nu = 1, 2$ and $\mathbf{z}_i^\top = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$. It is assumed that the $(T_{1i}, \mathbf{z}_i)$, for $i = 1, ..., n$, are independently and identically distributed (*i.i.d.*). The censoring times, T_{2i}, are also assumed to be *i.i.d.* The distribution of T_2 depends on $\mathbf{z}$. In addition, we assume that T_{1i} and T_{i2} are conditionally independent given $\mathbf{z}_i$, and that T_{i1} is informatively right censored by T_{i2} through some covariates (Andersen & Keiding, 2006). We observe $(y_i, \mathbf{z}_i, \delta_{1i})$, where $y_i = \min\{T_{1i}, T_{2i}\}$ and $\delta_{1i} = I(T_{1i} \leq T_{2i})$. We also define $\delta_{2i} = [1 - \delta_{1i}]$. Finally, $\boldsymbol{\varphi}$ is a generic vector of parameters, that can be either $\boldsymbol{\alpha}$ or $\boldsymbol{\gamma}$ as appropriate.

4.2.1 Survival functions

The survival function of $T_{\nu i}$ taking values in $(0, 1)$, conditional on $\mathbf{z}_{\nu i}$ and $\boldsymbol{\varphi_\nu}$, can be expressed as

$$P(T_{\nu i} > t_{\nu i}|\mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu) = S_{T_\nu}(t_{\nu i}|\mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu) = \mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu)], \tag{4.1}$$

where, for $\nu = 1, 2$, $\boldsymbol{\varphi}_\nu$ and $\mathbf{z}_{\nu i}$ represent generic vectors of coefficients and covariates, respectively. The survival functions are specified using link-based models (see Marra & Radice, 2020a, and references therein). That is, $S_{T_\nu}(t_{\nu i}|\mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu)$ is defined as $\mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu)]$, where $\mathcal{G}_\nu$ is an inverse link function. The set up of the two ξ predictors is discussed in the detail in the next section. As conveyed by the notation, ξ_{1i} and ξ_{2i} must include baseline functions of time. Different choices for function $\mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\varphi}_\nu)]$ can be specified; some common examples are shown in Table 4.1.

The cumulative hazard function, $\mathcal{H}_{T_\nu}$, and the hazard function, $\hbar_{T_\nu}$, are given

`Model`	`Link` $g(S)$	`Inverse link` $g^{-1}(\xi) = G(\xi)$	$G'(\xi)$
`Prop.hazards ("PH")`	$\log\{-\log(S)\}$	$\exp\{-\exp(\xi)\}$	$-G(\xi)\exp(\xi)$
`Prop.odds ("PO")`	$-\log\left(\frac{S}{1-S}\right)$	$\frac{\exp(-\xi)}{1+\exp(-\xi)}$	$-G^2(\xi)\exp(-\xi)$
`Probit ("probit")`	$-\Phi^{-1}(S)$	$\Phi(-\xi)$	$-\phi(-\xi)$

Table 4.1: Link functions implemented in `GJRM`. Φ and ϕ are the cumulative distribution and density functions of a univariate standard normal distribution. Alternative links can be implemented. The first two functions can be called log-log and -logit links, respectively.

by

$$\begin{aligned} \mathcal{H}_{T_\nu}(t_{\nu i}|\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu) &= -\log \mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)], \\ h_{T_\nu}(t_{\nu i}|\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu) &= -\frac{\mathcal{G}'_\nu[\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)]}{\mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)]}\frac{\partial \xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)}{\partial t_{\nu i}}, \end{aligned} \tag{4.2}$$

where $\mathcal{G}'_\nu[\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)] = \partial \mathcal{G}_\nu[\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)] \,/\, \partial \xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\varphi}_\nu)$.

4.2.2 Additive predictors

This section provides details on the set up of the additive predictors used to model the event and censoring times. To make the presentation simpler but without loss of generality, the same design matrix is set up for the predictors. Note also that $t_{\nu i}$ can be treated like a covariate. The main advantages of using additive predictors are several types of covariate effects can be dealt with and that such effects can be flexibly determined from the data without making strong parametric a priori assumptions on their forms (Ruppert et al., 2003; Wood, 2017). Let us consider a generic predictor $\xi_{\nu i} \in \mathbb{R}$ (where the dependence on the covariates and parameters is momentarily dropped), and the overall covariate vector $\mathbf{x}_{\nu i}$, which contains $\mathbf{z}_{\nu i}$ and $t_{\nu i}$. The additive predictors for the censoring and event times can be defined generically as

$$\xi_{\nu i} = \gamma_{\nu 0} + \sum_{k_\nu=0}^{K_\nu} s_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}),\ i = 1,\ldots,n, \tag{4.3}$$

where $\gamma_{\nu 0} \in \mathbb{R}$ is an overall intercept, $\mathbf{x}_{\nu k_\nu i}$ denotes the k_ν^{th} sub-vector of the complete vector $\mathbf{x}_{\nu i}$ and the K_ν functions $s_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})$ represent generic effects which are chosen according to the type of covariate(s) considered. Note that, in (4.3), k_ν starts from 0 since the summation also includes a smooth function of time. Each $s_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})$ can be represented as a linear combination of $J_{\nu k_\nu}$ basis functions $\mathcal{Q}_{\nu k_\nu j_{\nu k_\nu}}(\mathbf{x}_{\nu k_\nu i})$ and regression coefficients $\gamma_{\nu k_\nu j_{\nu k_\nu}} \in \mathbb{R}$, that is (e.g., Wood, 2017)

$$s_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}) = \sum_{j_{\nu k_\nu}=1}^{J_{\nu k_\nu}} \gamma_{\nu k_\nu j_{\nu k_\nu}} \mathcal{Q}_{\nu k_\nu j_{\nu k_\nu}}(\mathbf{x}_{\nu k_\nu i}). \tag{4.4}$$

Therefore, equation (4.3) can be written as

$$\xi_{\nu i} = \gamma_{\nu 0} + \sum_{k_\nu=0}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu},$$

where $\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}) = \{\mathcal{Q}_{\nu k_\nu 1}(\mathbf{x}_{\nu k_\nu i}), \ldots, \mathcal{Q}_{\nu k_\nu J_{\nu k_\nu}}(\mathbf{x}_{\nu k_\nu i})\}^\top$ and $\boldsymbol{\gamma}_{\nu k_\nu} = (\gamma_{\nu k_\nu 1}, \ldots, \gamma_{\nu k_\nu J_{\nu k_\nu}})^\top$. Furthermore, if $\boldsymbol{\mathcal{Q}}_{\nu i}^\top \boldsymbol{\gamma}_\nu = \sum_{k_\nu=0}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu}$, $\boldsymbol{\gamma}_\nu = (\gamma_{\nu 0}, \boldsymbol{\gamma}_{\nu 0}, \ldots, \boldsymbol{\gamma}_{\nu K_\nu})^\top$ and $\boldsymbol{\mathcal{Q}}_{\nu i} = \{1, \boldsymbol{\mathcal{Q}}_{\nu 0}(\mathbf{x}_{\nu 0 i})^\top, \ldots, \boldsymbol{\mathcal{Q}}_{\nu K_\nu}(\mathbf{x}_{\nu K_\nu})^\top\}^\top$, we obtain

$$\xi_{\nu i} = \boldsymbol{\mathcal{Q}}_{\nu i}^\top \boldsymbol{\gamma}_\nu. \tag{4.5}$$

Finally, if we define $\boldsymbol{\mathcal{Q}}_\nu = \{\boldsymbol{\mathcal{Q}}_{\nu 1}, \ldots, \boldsymbol{\mathcal{Q}}_{\nu n}\}^\top$, the complete system can be written as

$$\boldsymbol{\xi}_\nu = \boldsymbol{\mathcal{Q}}_\nu \boldsymbol{\gamma}_\nu.$$

If censoring is informative, some covariates in $\mathbf{x}_{1i}$ must also appear in $\mathbf{x}_{2i}$. In particular, let us define the vectors of informative and non-informative covariates of dimensions Q and Q_ν as $\mathbf{x}_i^{0\top} = (x_{1i}^0, \ldots, x_{Qi}^0)$ and $\mathbf{x}_{\nu i}^{1\top} = (x_{\nu 1 i}^1, \ldots, x_{\nu Q_\nu i}^1)$, where $K_\nu = Q + Q_\nu$. Informative censoring implies that some components of $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ must appear in $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$. Without loss of generality, we assume that the first Q components in $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ appear in $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$.

That is,

$$\begin{aligned}\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1i}) &= \sum_{q=1}^{Q} s_q(\mathbf{x}^0_{qi}) + \sum_{q_1=1}^{Q_1} s_{1q_1}(\mathbf{x}^1_{1q_1i}) \\ \sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2i}) &= \sum_{q=1}^{Q} s_q(\mathbf{x}^0_{qi}) + \sum_{q_2=1}^{Q_2} s_{2q_2}(\mathbf{x}^1_{2q_2i})\end{aligned}. \tag{4.6}$$

Therefore, using (4.6), equation (4.3) becomes

$$\xi_{\nu i} = \alpha_{\nu 0} + \sum_{q=1}^{Q} s_q(\mathbf{x}^0_{qi}) + \sum_{q_\nu=0}^{Q_\nu} s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}), \tag{4.7}$$

where $\mathbf{x}^0_{qi}$ and $\mathbf{x}^1_{\nu q_\nu i}$ denote the informative and non-informative sub-vectors of the complete vectors $\mathbf{x}^0_i$ and $\mathbf{x}^1_{\nu i}$ respectively, and $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}) = s_{\nu 0}(t_{\nu i})$ when $q_\nu = 0$.

As before, in (4.7), each $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})$ can be approximated as a linear combination of $J_{\nu q_\nu}$ non-informative basis functions $\mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})$ and regression coefficients $\alpha_{\nu q_\nu j_{\nu q_\nu}} \in \mathbb{R}$. In a similar manner, each $s_q(\mathbf{x}^0_{qi})$ can be approximated as a linear combination of J_q informative basis functions $\mathcal{Q}_{qj_q}(\mathbf{x}^0_{qi})$ and regression coefficients $\alpha_{0qj_q} \in \mathbb{R}$. More specifically, $s_q(\mathbf{x}^0_{qi})$ and $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})$ are given by $s_q(\mathbf{x}^0_{qi}) = \sum_{j_q=1}^{J_q} \alpha_{0qj_q} \mathcal{Q}_{qj_q}(\mathbf{x}^0_{qi})$ and $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}) = \sum_{j_{\nu q_\nu}=1}^{J_{\nu q_\nu}} \alpha_{\nu q_\nu j_{\nu q_\nu}} \mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})$, and therefore (4.7) can be written as

$$\xi_{\nu i} = \alpha_{\nu 0} + \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=0}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu}, \tag{4.8}$$

where $\boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q} = \sum_{j_q=1}^{J_q} \alpha_{0qj_q} \mathcal{Q}_{qj_q}(\mathbf{x}^0_{qi})$, $\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu} = \sum_{j_{\nu q_\nu}=1}^{J_{\nu q_\nu}} \alpha_{\nu q_\nu j_{\nu q_\nu}}$ $\mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})$, $\boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi}) = \{\mathcal{Q}_{q1}(\mathbf{x}^0_{qi}), \ldots, \mathcal{Q}_{qJ_q}(\mathbf{x}^0_{qi})\}^\top$, $\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}) = \{\mathcal{Q}_{\nu q_\nu 1}(\mathbf{x}^1_{\nu q_\nu i}),$ $\ldots, \mathcal{Q}_{\nu q_\nu J_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})\}^\top$, $\boldsymbol{\alpha}_{0q} = (\alpha_{0q1}, \ldots, \alpha_{0qJ_q})^\top$ and $\boldsymbol{\alpha}_{\nu q_\nu} = (\alpha_{\nu q_\nu 1}, \ldots, \alpha_{\nu q_\nu J_{\nu q_\nu}})^\top$. To write equation (4.8) in a more compact way, we define $\boldsymbol{\mathcal{Q}}_i^{0\top} \boldsymbol{\alpha}_0 = \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q}$ and $\boldsymbol{\mathcal{Q}}_{\nu i}^{1\top} \boldsymbol{\alpha}_\nu = \sum_{q_\nu=0}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu}$, where $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}, \ldots, \boldsymbol{\alpha}_{0Q})^\top$, $\boldsymbol{\alpha}_\nu = (\alpha_{\nu 0}, \boldsymbol{\alpha}_{\nu 0}, \ldots, \boldsymbol{\alpha}_{\nu Q_\nu})^\top$, $\boldsymbol{\mathcal{Q}}_i^0 = \{\boldsymbol{\mathcal{Q}}_1(\mathbf{x}^0_{1i})^\top, \ldots, \boldsymbol{\mathcal{Q}}_Q(\mathbf{x}^0_{Qi})^\top\}^\top$ and $\boldsymbol{\mathcal{Q}}_{\nu i}^1 = \{1, \boldsymbol{\mathcal{Q}}_{\nu 0}(\mathbf{x}^1_{\nu 0i})^\top,$

$\ldots, \boldsymbol{\mathcal{Q}}_{\nu Q_\nu}(\mathbf{x}^1_{\nu Q_\nu})^\top\}^\top$. Therefore,

$$\xi_{\nu i} = \boldsymbol{\mathcal{Q}}_i^{0\top}\boldsymbol{\alpha}_0 + \boldsymbol{\mathcal{Q}}_{\nu i}^{1\top}\boldsymbol{\alpha}_\nu. \tag{4.9}$$

If $Q > 0$ then censoring is informative and $\sum_{q=1}^{Q} s_q(\mathbf{x}^0_{qi})$ can be estimated using the information from both the censoring and event times. If $Q = 0$ (i.e., the components in $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ and $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$ are assumed all distinct) then (4.8) reduces to the model with non-informative censoring defined in equation (4.5).

Note that, for the case in which $Q = 0$, we have introduced the new parameter vector $\boldsymbol{\gamma}_\nu$ to stress the difference between the parameters of the informative and non-informative models. Some methods for determining the value of Q are discussed in Appendix B.1.

Each $\boldsymbol{\varphi}_{\nu k_\nu}$ has an associated quadratic penalty $\lambda_{\nu k_\nu}\boldsymbol{\varphi}_{\nu k_\nu}^\mathsf{T}\boldsymbol{\mathcal{Q}}_{\nu k_\nu}\boldsymbol{\varphi}_{\nu k_\nu}$ that allows one to enforce specific properties on the k_ν^{th} function, such as smoothness. Note that each matrix $\boldsymbol{\mathcal{Q}}_{\nu k_\nu}$ only depends on the choice of the basis functions. Smoothing parameter $\lambda_{\nu k_\nu} \in [0, \infty)$ controls the trade-off between fit and smoothness, and as such it determines the shape of the related estimated smooth function. The overall penalty can be defined as $\boldsymbol{\varphi}_\nu^\mathsf{T}\boldsymbol{\mathcal{D}}_\nu\boldsymbol{\varphi}_\nu$, where $\boldsymbol{\mathcal{D}}_\nu = \text{diag}(0, \lambda_{\nu 1}\boldsymbol{\mathcal{D}}_{\nu 1}, \ldots, \lambda_{\nu K_\nu}\boldsymbol{\mathcal{D}}_{\nu K\nu})$. Smooth functions are typically subject to centering (identifiability) constraints (see Wood (2017) for more details). Depending on the types of covariate effects one wishes to model, several definitions of basis functions and penalty terms are possible. Examples include thin plate, cubic and P- regression splines, Markov random fields, random effects and Gaussian process smooths (Wood, 2017).

To give a concrete example, consider the informative additive model

$$g_\nu\{S_\nu(t_{\nu i}|\mathbf{z}_i^0,\mathbf{z}_{\nu i}^1)\} = g_\nu\{S_{\nu 0}(t_{\nu i})\} + \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{z}_{qi}^0)^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=1}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{z}_{\nu q_\nu i}^1)^\top \boldsymbol{\alpha}_{\nu q_\nu}, \tag{4.10}$$

where $g_\nu : (0,1) \to (-\infty,\infty)$ is a differentiable and invertible link function (see Table 4.1), $S_{\nu 0}(t_{\nu i})$ is a baseline survival function, and $g_\nu\{S_{\nu 0}(t_{\nu i})\}$ is represented using a smooth function of time, $s_{\nu 0}(t_{\nu i})$. When the log-log link is chosen, equation (4.10) yields the proportional hazards model

$$\log\{\mathcal{H}_\nu(t_{\nu i}|\mathbf{z}_i^0,\mathbf{z}_{\nu i}^1)\} = \log\{\mathcal{H}_{\nu 0}(t_{\nu i})\} + \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{z}_{qi}^0)^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=1}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{z}_{\nu q_\nu i}^1)^\top \boldsymbol{\alpha}_{\nu q_\nu},$$

where $\mathcal{H}_\nu(t_{\nu i}|\mathbf{z}_i^0,\mathbf{z}_{\nu i}^1) = -\log\{S_\nu(t_{\nu i} \mid \mathbf{z}_i^0,\mathbf{z}_{\nu i}^1)\}$ and $\log\{\mathcal{H}_{\nu 0}(t_{\nu i})\} = -\log\{S_{\nu 0}(t_{\nu i})\}$ is the cumulative baseline hazard function. Analogously, equation (4.10) yields the proportional odds model when the -logit link is chosen.

The hazard function defined in (4.2) requires the term $\partial\xi_\nu(t_{\nu i},\mathbf{z}_{\nu i},\boldsymbol{\gamma}_\nu)/\partial t_{\nu i}$ ($\nu = 1,2$) which, using (4.5), can be calculated as follows

$$\frac{\partial\xi_\nu(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)}{\partial t_{\nu i}} = \lim_{\varepsilon\to 0}\left\{\frac{\boldsymbol{\mathcal{Q}}_\nu(t_{\nu i}+\varepsilon,\mathbf{z}_{\nu i}) - \boldsymbol{\mathcal{Q}}_\nu(t_{\nu i}-\varepsilon,\mathbf{z}_{\nu i})}{2\varepsilon}\right\}^{\mathsf{T}} \boldsymbol{\gamma}_\nu = \boldsymbol{\mathcal{Q}}_\nu(t_{\nu i},\mathbf{z}_{\nu i})^{\mathsf{T}}\boldsymbol{\gamma}_\nu, \tag{4.11}$$

where $\boldsymbol{\mathcal{Q}}_\nu(t_{\nu i},\mathbf{z}_\nu)^{\mathsf{T}}$ can be obtained either by a finite-difference method or analytically (if feasible). Note also that (4.11) must be positive to ensure that the hazard functions are positive. This is achieved by modelling the time effects using B-splines with coefficients constrained such that the resulting smooth functions of time are monotonically increasing. In particular, let $s_\nu(t_{\nu i}) = \sum_{j_\nu=1}^{J_\nu} \beta_{\nu j_\nu}\mathcal{Q}_{\nu j_\nu}(t_{\nu i})$, where the $\mathcal{Q}_{\nu j_\nu}$ are B-spline basis functions of at least second order built over the interval $[a,b]$, based on equally spaced knots, and $\beta_{\nu j_\nu}$ are spline coefficients. A sufficient condition for $s'_\nu(t_{\nu i}) \geq 0$ over $[a,b]$ is that $\beta_{\nu j_\nu} \geq \beta_{\nu j_\nu - 1}, \forall j$ (e.g., Leitenstorfer

& Tutz, 2006). Such a condition can be imposed by re-parametrizing the spline coefficient vector so that $\boldsymbol{\beta}_v = \boldsymbol{\Gamma}_v \tilde{\boldsymbol{\varphi}}_\nu$, where $\tilde{\boldsymbol{\varphi}}_v^\mathsf{T} = \{\varphi_{\nu 1}, \exp(\varphi_{\nu 2}), \ldots, \exp(\varphi_{\nu J_\nu})\}$ and $\boldsymbol{\Gamma}_\nu[\kappa_{\nu 1}, \kappa_{\nu 2}] = 0$ if $\kappa_{\nu 1} < \kappa_{\nu 2}$ and $\boldsymbol{\Gamma}_\nu[\kappa_{\nu 1}, \kappa_{\nu 2}] = 1$ if $\kappa_{\nu 1} \geq \kappa_{\nu 2}$, with $\kappa_{\nu 1}$ and $\kappa_{\nu 2}$ denoting the row and column entries of the respective matrix. Note that the parameter vector to estimate is $\boldsymbol{\varphi}_\nu^\mathsf{T} = (\varphi_{\nu 1}, \gamma_{\nu 2}, \ldots, \varphi_{\nu J_\nu})$. The penalty term is set up to penalise the squared differences between adjacent $\varphi_{\nu j_\nu}$, starting from $\varphi_{\nu 2}$, using $\boldsymbol{\mathcal{D}}_\nu = \boldsymbol{\mathcal{D}}_\nu^{\odot\top} \boldsymbol{\mathcal{D}}_\nu^{\odot}$ where $\boldsymbol{\mathcal{D}}_\nu^{\odot}$ is a $(J_\nu - 2) \times J_\nu$ matrix made up of zeros except that $\boldsymbol{\mathcal{D}}_\nu^{\odot}[\kappa_\nu, \kappa_\nu + 1] = -\boldsymbol{\mathcal{D}}_\nu^{\odot}[\kappa_\nu, \kappa_\nu + 2] = 1$ for $\kappa_\nu = 1, ..., J_\nu - 2$ (Pya & Wood, 2015). Matrix $\boldsymbol{\mathcal{Q}}_\nu$, in equation (4.2.2), can absorb $\boldsymbol{\Gamma}_\nu$. So, the non-informative and informative additive predictors can be written as

$$\begin{aligned}
\xi_{\nu i} &= \gamma_{\nu 0} + \boldsymbol{\mathcal{Q}}_{\nu 0}(y_i)^\top \boldsymbol{\Gamma}_{\nu 0} \tilde{\boldsymbol{\gamma}}_{\nu 0} + \sum_{k_\nu=1}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu}, \\
\xi_{\nu i} &= \alpha_{\nu 0} + \boldsymbol{\mathcal{Q}}_{\nu 0}(y_i)^\top \boldsymbol{\Gamma}_{\nu 0} \tilde{\boldsymbol{\alpha}}_{\nu 0} + \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{x}_{qi}^0)^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=1}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)^\top \boldsymbol{\alpha}_{\nu q_\nu}.
\end{aligned} \tag{4.12}$$

The model set up described above has several advantages. For instance, it can flexibly determine and in a data driven manner the functional shapes of the baseline and covariate effects, avoids the need for numerical integration, easily allows for time-dependent effects via smooth interaction terms, and can deal with time-varying covariates in the usual manner. It is worth noting that the more extensive use of parametric survival models in applications has been encouraged by Cox; see the discussion in Reid (1994). Moreover, as pointed out for instance by Hjort (1992), parametric approaches simplify somewhat model estimation and comparison, easily allow for the visualization of the estimated baseline hazard and survival functions, and allow us to calculate several quantities of interest and their variances which would otherwise be difficult to obtain with a non-parametric approach.

In this section, we have introduced Generalized Additive survival models with

informative censoring. In the next section, we will propose a penalized likelihood method to estimate the developed methodology.

4.3 Estimation approach

This section proposes a penalized likelihood method to estimate the parameters of the developed model. Specifically, model fitting is based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection. The $\sqrt{n}$ consistency and asymptotic normality of the non-informative and informative estimators are derived, where the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart is highlighted. The construction of confidence intervals and p-values are discussed in the last part of this section.

4.3.1 Penalized maximum log-likelihood estimation

The data consist of $\{y_i, \delta_{1i}, \mathbf{z}_i\}$, where $y_i = \min\{T_{1i}, T_{2i}\}$ and $\delta_{1i} = I(T_{1i} \leq T_{2i})$, for $i = 1, \ldots, n$. Let $f(t_1, t_2|\mathbf{z})$ be the conditional joint distribution of (T_1, T_2) given $\mathbf{z}$. We can write $P(T_1 = y_i, T_2 > y_i|\mathbf{z}_i) = \int_{y_i}^{\infty} f(y_i, t_2|\mathbf{z}_i)dt_2$ and $P(T_1 > y_i, T_2 = y_i|\mathbf{z}_i) = \int_{y_i}^{\infty} f(t_1, y_i|\mathbf{z}_i)dt_1$. Therefore, the conditional likelihood function of (y_i, δ_{1i}) given $\mathbf{z}_i$, for all $i = 1, ..., n$, is

$$\mathcal{L} = \prod_{i=1}^{n} \left[\int_{y_i}^{\infty} f(y_i, t_2|\mathbf{z}_i)dt_2\right]^{\delta_{1i}} \left[\int_{y_i}^{\infty} f(t_1, y_i|\mathbf{z}_i)dt_1\right]^{\delta_{2i}}.$$

Below we provide the relevant details for the cases of informative and non-informative censoring, which highlight the differences between the two estimators and that are also required for the theoretical derivations in Section 4.3.3.

If it is assumed that T_{1i} and T_{i2} are conditionally independent given $\mathbf{z}_i$, then

$\int_{y_i}^{\infty} f(y_i, t_2|\mathbf{z}_i)dt_2 = f_1(y_i|\mathbf{z}_{1i};\boldsymbol{\gamma}_1)S_2(y_i|\mathbf{z}_{2i};\boldsymbol{\gamma}_2)$ and $\int_{y_i}^{\infty} f(t_1, y_i|\mathbf{z}_i)dt_1 = f_2(y_i|\mathbf{z}_{2i};\boldsymbol{\gamma}_2)$ $S_1(y_i|\mathbf{z}_{1i};\boldsymbol{\gamma}_1)$ when censoring is non-informative. However, if censoring is informative $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ would have some components in common. Since it was assumed that the first Q components of $\boldsymbol{\gamma}_1$ are the same as the first Q components of $\boldsymbol{\gamma}_2$, we have $\boldsymbol{\mathcal{Q}}_{\nu i}^{\top}\boldsymbol{\gamma}_\nu = \boldsymbol{\mathcal{Q}}_i^{0\top}\boldsymbol{\alpha}_0 + \boldsymbol{\mathcal{Q}}_{\nu i}^{1\top}\boldsymbol{\alpha}_\nu$. Using (4.1), (4.2), and $\xi_{\nu i}(\boldsymbol{\gamma}_\nu)$ and $\xi_{\nu i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu)$ as the shorthand notation for $\xi_{\nu i}(y_i, \mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)$ and $\xi_{\nu i}(y_i, \mathbf{z}_{\nu i};\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu)$ respectively, the non-informative and informative log-likelihood functions can be written, respectively, as

$$
\begin{aligned}
\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\left\{\log \mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right] + \delta_{1i}\log\left[-\frac{\mathcal{G}_1'\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]}{\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]}\frac{\partial\xi_{1i}(\boldsymbol{\gamma}_1)}{\partial y_i}\right]\right\} \\
&\quad + \sum_{i=1}^{n}\left\{\log \mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right] + \delta_{2i}\log\left[-\frac{\mathcal{G}_2'\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]}{\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]}\frac{\partial\xi_{2i}(\boldsymbol{\gamma}_2)}{\partial y_i}\right]\right\}, \\
\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left\{\log \mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right] + \delta_{1i}\log\left[-\frac{\mathcal{G}_1'\left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right]}{\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right]}\frac{\partial\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)}{\partial y_i}\right]\right\} \\
&\quad + \sum_{i=1}^{n}\left\{\log \mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right] + \delta_{2i}\log\left[-\frac{\mathcal{G}_2'\left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right]}{\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right]}\frac{\partial\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)}{\partial y_i}\right]\right\}.
\end{aligned}
\tag{4.13}
$$

Our model specification allows for a high degree of flexibility in modelling survival data. If an unpenalised estimation approach is employed to estimate $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\top}, \boldsymbol{\gamma}_2^{\top})^{\top}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^{\top}, \boldsymbol{\alpha}_1^{\top}, \boldsymbol{\alpha}_2^{\top})^{\top}$, then the resulting smooth function estimates are likely to be unduly wiggly (e.g., Wood, 2017). Therefore, to prevent over-fitting, the following functions are maximized

$$\ell_p(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}^{\top}\boldsymbol{\Lambda}\boldsymbol{\gamma}, \tag{4.14}$$

$$\ell_p(\boldsymbol{\alpha}) = \ell(\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^{\top}\boldsymbol{\Lambda}\boldsymbol{\alpha}, \tag{4.15}$$

where $\ell_p(\boldsymbol{\gamma})$ and $\ell_p(\boldsymbol{\alpha})$ are the non-informative and informative penalized log-likelihoods. Moreover, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\mathcal{D}}_1, \boldsymbol{\mathcal{D}}_2)$, and $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$ are overall penalties

which contain $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$. The smoothing parameter vectors can be collected in the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$. Estimation of the models' parameters and smoothing coefficients is achieved by using a stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection (this will be discussed in Section 4.3.2). This required working with first and second order analytical derivatives which have been tediously derived as well as verified using numerical derivatives. Their structures are shown below. Note that these results were also required for the theoretical proofs presented in Section 4.3.3.

When censoring is non-informative, the gradient of (4.14) can be obtained as

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell_p(\boldsymbol{\gamma}) = \boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) - \boldsymbol{\gamma}\boldsymbol{\Lambda},$$

where $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) = \left(\boldsymbol{\nabla}_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})^\top,\ \boldsymbol{\nabla}_{\boldsymbol{\gamma}_2} \ell(\boldsymbol{\gamma})^\top\right)^\top$. The components of $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_\nu} \ell(\boldsymbol{\gamma})$ can generically be calculated using the following expression

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu k_\nu}} \ell(\boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^n \left[\Delta_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle}(y_i) + \Omega_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle'}(y_i) \right] & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0}, \\ \sum_{i=1}^n \left[\Delta_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}) \right] & \text{otherwise.} \end{cases} \tag{4.16}$$

In (4.16), $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle'}(y_i)$ are design vectors. Furthermore, $\Omega_{\nu i} = \delta_{\nu i} \left(\dfrac{\partial \xi_{\nu i}}{\partial y_i} \right)^{-1}$ and $\Delta_{\nu i} = \left[\dfrac{\mathcal{G}_\nu'}{\mathcal{G}_\nu} + \delta_{\nu i} \left(\dfrac{\mathcal{G}_\nu''}{\mathcal{G}_\nu'} - \dfrac{\mathcal{G}_\nu'}{\mathcal{G}_\nu} \right) \right]$, for all $\nu = 1, 2$. The non-informative penalized Hessian can be calculated as

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell_p(\boldsymbol{\gamma}) = \boldsymbol{\nabla}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) - \boldsymbol{\Lambda},$$

where

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{\nabla}_{\boldsymbol{\gamma}_1\boldsymbol{\gamma}_1}\ell(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\nabla}_{\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2}\ell(\boldsymbol{\gamma}) \end{bmatrix}.$$

Further, the elements of $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_\nu\boldsymbol{\gamma}_\nu}\ell(\boldsymbol{\gamma})$ are calculated using

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu k_\nu}\boldsymbol{\gamma}_{\nu 0}}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\Phi_{\nu i}\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i)^\top\Big], \\
\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu 0}\boldsymbol{\gamma}_{\nu s_\nu}}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu s_\nu}(\mathbf{x}_{\nu s_\nu i})^\top\Big], \\
\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu k_\nu}\boldsymbol{\gamma}_{\nu s_\nu}}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu s_\nu}(\mathbf{x}_{\nu s_\nu i})^\top\Big], \\
\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu 0}\boldsymbol{\gamma}_{\nu 0}}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i)^\top + \Delta_{\nu i}\boldsymbol{\mathcal{Q}}^{\triangle\triangle}_{\nu 0}(y_i) - \boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(y_i)\Psi_{\nu i}\boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(y_i)^\top \\
&\quad + \Omega_{\nu i}\boldsymbol{\mathcal{Q}}^{\triangle\triangle'}_{\nu 0}(y_i)\Big].
\end{aligned} \tag{4.17}$$

In these sub-matrices $\Phi_{\nu i} = \delta_{\nu i}\left(\dfrac{\mathcal{G}'''_\nu}{\mathcal{G}_\nu} - \dfrac{\mathcal{G}''^2_\nu}{\mathcal{G}'^2_\nu} - \dfrac{\mathcal{G}''_\nu}{\mathcal{G}_\nu} + \dfrac{\mathcal{G}'^2_\nu}{\mathcal{G}^2_\nu}\right)$ and $\Psi_{\nu i} = \left[\delta_{\nu i}\left(\dfrac{\partial\xi_{\nu i}}{\partial y_i}\right)^{-2}\right]$. In addition, $\boldsymbol{\mathcal{Q}}^{\triangle\triangle}_{\nu 0}(y_i)$ and $\boldsymbol{\mathcal{Q}}^{\triangle\triangle'}_{\nu 0}(y_i)$ are design diagonal matrices.

If the censoring is informative, the gradient of (4.15) can be calculated as

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell_p(\boldsymbol{\alpha}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\Lambda},$$

where $\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) = \left(\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})^\top, \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha})^\top, \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha})^\top\right)^\top$. To obtain $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_\nu}\ell(\boldsymbol{\alpha})$, we use

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}^0_i\left(\Delta_{1i} + \Delta_{2i}\right)\right], \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu k_\nu}}\ell(\boldsymbol{\alpha}) &= \begin{cases} \sum_{i=1}^{n}\Big[\Delta_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle}_{\nu 0}(y_i) + \Omega_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle'}_{\nu 0}(y_i)\Big] & \text{if } \boldsymbol{\alpha}_{\nu k_\nu} = \boldsymbol{\alpha}_{\nu 0}, \\ \sum_{i=1}^{n}\Big[\Delta_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\Big] & \text{otherwise,} \end{cases}
\end{aligned} \tag{4.18}$$

where $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\prime}(y_i)$ are design vectors. The informative penalized Hessian can be obtained as follow

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_p(\boldsymbol{\alpha}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) - \boldsymbol{\Lambda},$$

where

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \mathbf{0} \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \mathbf{0} & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \end{bmatrix}.$$

Furthermore, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$ and the components of $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu}\ell(\boldsymbol{\alpha})$ are obtained using

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}_i^0(\Phi_{1i}+\Phi_{2i})\boldsymbol{\mathcal{Q}}_i^{0\top}\right], \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_{\nu q_\nu}}\ell(\boldsymbol{\alpha}) &= \begin{cases} \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}_i^0\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i)^\top\right] & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0}, \\ \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}_i^0\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)^\top\right] & \text{otherwise,} \end{cases} \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \begin{cases} \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_i^{0\top}\right] & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0}, \\ \sum_{i=1}^{n}\left[\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_i^{0\top}\right] & \text{otherwise.} \end{cases}
\end{aligned} \tag{4.19}$$

Finally, the elements of $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_\nu}\ell(\boldsymbol{\alpha})$ are calculated using

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_{\nu 0}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})\Phi_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle}_{\nu 0}(y_i)^\top\Big],\\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu q_\nu}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}^{\iota\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top\Big],\\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_{\nu r_\nu}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})\Phi_{\nu i}\boldsymbol{\mathcal{Q}}_{\nu r_\nu}(\mathbf{x}^1_{\nu r_\nu i})^\top\Big],\\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu 0}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\Big[\boldsymbol{\mathcal{Q}}^{\iota\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle}_{\nu 0}(y_i)^\top + \Delta_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle\triangle}_{\nu 0}(y_i) - \boldsymbol{\mathcal{Q}}^{\iota\triangle'}_{\nu 0}(y_i)\Psi_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle'}_{\nu 0}(y_i)^\top\\
&\quad + \ \Omega_{\nu i}\boldsymbol{\mathcal{Q}}^{\iota\triangle\triangle'}_{\nu 0}(y_i)\Big].
\end{aligned} \tag{4.20}$$

As before, $\boldsymbol{\mathcal{Q}}^{\iota\triangle\triangle}_{\nu 0}(y_i)$ and $\boldsymbol{\mathcal{Q}}^{\iota\triangle\triangle'}_{\nu 0}(y_i)$ represent design diagonal matrices.

The derivations of the results reported here are given in Appendixes B.2 and B.3.

Remark 1. The scores and Hessian components described in this section have been implemented in a modular way, hence no substantial programming work will be required to incorporate link functions not considered in this article. Furthermore, quantities such as those defined in (4.16), (4.17), (4.18), (4.19) and (4.20), are needed for the theoretical proofs provided in Section 4.3.3.

4.3.2 Algorithmic details

The optimization method used for the informative and non-informative estimator is the trust region algorithm. In this method, model fitting is based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection. As already mentioned, $\boldsymbol{\varphi}$ is a generic vector of parameters, that can be either $\boldsymbol{\alpha}$ or $\boldsymbol{\gamma}$ as appropriate.

At iteration a, for a given vector $\boldsymbol{\varphi}$ and maintaining $\boldsymbol{\lambda}$ fixed at a vector of values,

equations (4.14) or (4.15) (or generally, any of the models' likelihoods considered in the thesis) are maximized using

$$\boldsymbol{\varphi}^{[a+1]} = \underset{\boldsymbol{\varepsilon}:\|\boldsymbol{\epsilon}\|\leq\boldsymbol{\Xi}^{[a]}}{\arg\min}\ \bar{\ell}_p(\boldsymbol{\varphi}^{[a]}),$$

where $\bar{\ell}_p(\boldsymbol{\varphi}^{[a]}) = -\Big\{\ell_p(\boldsymbol{\varphi}^{[a]}) + \boldsymbol{\varepsilon}^\top \mathbf{g}_p(\boldsymbol{\varphi}^{[a]}) + \frac{1}{2}\boldsymbol{\varepsilon}^\top \boldsymbol{\mathcal{H}}_p(\boldsymbol{\varphi}^{[a]})\boldsymbol{\varepsilon}\Big\}$, $\mathbf{g}_p(\boldsymbol{\varphi}^{[a]}) = \mathbf{g}(\boldsymbol{\varphi}^{[a]}) - \boldsymbol{\Lambda}\boldsymbol{\varphi}^{[a]}$, $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\varphi}^{[a]}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\varphi}^{[a]}) - \boldsymbol{\Lambda}$. Vector $\mathbf{g}(\boldsymbol{\varphi}^{[a]})$ consists of $\mathbf{g}_0(\boldsymbol{\varphi}^{[a]}) = \boldsymbol{\nabla}_{\boldsymbol{\varphi}_0}\ell(\boldsymbol{\varphi})\big|_{\boldsymbol{\varphi}_0=\boldsymbol{\varphi}_0^{[a]}}$ and $\mathbf{g}_\nu(\boldsymbol{\varphi}^{[a]}) = \boldsymbol{\nabla}_{\boldsymbol{\varphi}_\nu}\ell(\boldsymbol{\varphi})\big|_{\boldsymbol{\varphi}_\nu=\boldsymbol{\varphi}_\nu^{[a]}}$, and $\boldsymbol{\mathcal{H}}(\boldsymbol{\varphi}^{[a]})_{l,j} = \boldsymbol{\nabla}_{\boldsymbol{\varphi}_l\boldsymbol{\varphi}_j}\ell(\boldsymbol{\varphi})\big|_{\boldsymbol{\varphi}_l=\boldsymbol{\varphi}_l^{[a]},\boldsymbol{\varphi}_j=\boldsymbol{\varphi}_j^{[a]}}$, where $l, j = 0, 1, 2$ and $\nu = 1, 2$. The euclidean norm is denoted by $\|\cdot\|$, and the radius of the trust region is represented by $\boldsymbol{\Xi}^{[a]}$ which is adjusted through the iterations. Close to the solution, the trust region algorithm behaves as a classic Newton-Raphson unconstrained method (Nocedal & Wright, 2006).

Estimation of $\boldsymbol{\lambda}$ is achieved by adapting the general and automatic multiple smoothing parameter estimation method of Marra et al. (2017) to the context of the proposed survival models. The smoothing criterion is based on the knowledge of $\mathbf{g}(\boldsymbol{\varphi})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\varphi})$. The main ideas and some useful results are given here.

To simplify the notation, $\mathbf{g}_p(\boldsymbol{\varphi}^{[a]})$, $\mathbf{g}(\boldsymbol{\varphi}^{[a]})$, $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\varphi}^{[a]})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\varphi}^{[a]})$ are denoted as $\mathbf{g}_p^{[a]}$, $\mathbf{g}^{[a]}$, $\boldsymbol{\mathcal{H}}_p^{[a]}$ and $\boldsymbol{\mathcal{H}}^{[a]}$. First, it is necessary to express the parameter estimator in terms of $\mathbf{g}_p^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]}$. To achieve this, a first order Taylor expansion of $\mathbf{g}_p^{[a+1]}$ about $\boldsymbol{\varphi}^{[a]}$ is used, which yields the following expression: $\mathbf{0} = \mathbf{g}_p^{[a+1]} \approx \mathbf{g}_p^{[a]}(\boldsymbol{\varphi}^{[a+1]} - \boldsymbol{\varphi}^{[a]})\boldsymbol{\mathcal{H}}_p^{[a]}$. After some manipulations, $\boldsymbol{\varphi}^{[a+1]} = (-\boldsymbol{\mathcal{H}}^{[a]} + \boldsymbol{\Lambda})^{-1}\sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\Big[\sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\boldsymbol{\varphi}^{[a]} + \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}^{-1}\mathbf{g}^{[a]}\Big]$ is obtained, which then becomes $\boldsymbol{\varphi}^{[a+1]} = (-\boldsymbol{\mathcal{H}}^{[a]} + \boldsymbol{\Lambda})^{-1}\sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\boldsymbol{\mathcal{Z}}^{[a]}$, where $\boldsymbol{\mathcal{Z}}^{[a]} = \boldsymbol{\upsilon}_{\boldsymbol{\mathcal{Z}}}^{[a]} + \boldsymbol{\xi}_{\boldsymbol{\mathcal{Z}}}^{[a]}$, $\boldsymbol{\upsilon}_{\boldsymbol{\mathcal{Z}}}^{[a]} = \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\boldsymbol{\varphi}^{[a]}$ and $\boldsymbol{\xi}_{\boldsymbol{\mathcal{Z}}}^{[a]} = \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}^{-1}\mathbf{g}^{[a]}$. Eigenvalue decomposition is used to obtain the square root of $-\boldsymbol{\mathcal{H}}^{[a]}$ and its inverse. Furthermore, from likelihood theory, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\mathcal{Z}} \sim \mathcal{N}(\boldsymbol{\upsilon}_{\mathcal{N}}, \mathbf{I})$, where $\boldsymbol{\upsilon}_{\boldsymbol{\mathcal{Z}}} = \sqrt{-\boldsymbol{\mathcal{H}}}\boldsymbol{\varphi}^0$, $\boldsymbol{\varphi}^0$ is the true parameter vector and $\mathbf{I}$ is the identity matrix. $\hat{\boldsymbol{\upsilon}}_{\boldsymbol{\mathcal{Z}}} = \sqrt{-\boldsymbol{\mathcal{H}}}\hat{\boldsymbol{\varphi}} = \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{Z}}$ is the predicted value vector for $\boldsymbol{\mathcal{Z}}$,

where $\boldsymbol{\mathcal{B}} = \sqrt{-\boldsymbol{\mathcal{H}}}(-\boldsymbol{\mathcal{H}} + \boldsymbol{\Lambda})^{-1}\sqrt{-\boldsymbol{\mathcal{H}}}$. Since our objective is to estimate $\boldsymbol{\lambda}$ so that the smooth terms' complexity which is not supported by the data is removed, the following criterion is used

$$\mathbb{E}\Big(\|\boldsymbol{\upsilon}_{\boldsymbol{\mathcal{Z}}} - \hat{\boldsymbol{\upsilon}}_{\boldsymbol{\mathcal{Z}}}\|^2\Big) = \mathbb{E}\Big(\|\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{Z}}\|^2\Big) - \bar{\mathrm{n}} + 2\mathrm{tr}(\boldsymbol{\mathcal{B}}), \tag{4.21}$$

where $\bar{\mathrm{n}} = 2\,\mathrm{n}$ and $\mathrm{tr}(\boldsymbol{\mathcal{B}})$ represent the number of effective degrees of freedom of the penalized model. In applications, $\boldsymbol{\lambda}$ is estimated by minimizing an estimate of equation (4.21), in other words

$$\widehat{\|\boldsymbol{\upsilon}_{\boldsymbol{\mathcal{Z}}} - \hat{\boldsymbol{\upsilon}}_{\boldsymbol{\mathcal{Z}}}\|^2} = \|\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{Z}}\|^2 - \bar{\mathrm{n}} + 2\mathrm{tr}(\boldsymbol{\mathcal{B}}). \tag{4.22}$$

The RHS of equation (4.22) depends on $\boldsymbol{\lambda}$ through $\boldsymbol{\mathcal{B}}$ while $\boldsymbol{\mathcal{Z}}$ is associated with the un-penalized part of the model. Equation (4.21) is approximately equivalent to the AIC (Akaike, 1973). This implies that $\boldsymbol{\lambda}$ is estimated by minimizing what is effectively the AIC with number of parameters given by $\mathrm{tr}(\boldsymbol{\mathcal{B}})$. Holding the model's parameter vector value fixed at $\boldsymbol{\varphi}^{[a+1]}$, the following problem

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \|\boldsymbol{\mathcal{Z}}^{[a+1]} - \boldsymbol{\mathcal{B}}^{[a+1]}\boldsymbol{\mathcal{Z}}^{[a+1]}\|^2 - \bar{\mathrm{n}} + 2\mathrm{tr}(\boldsymbol{\mathcal{B}}^{[a+1]}) \tag{4.23}$$

is solved using the automatic efficient and stable computational method proposed by Wood (2004). This approach uses the performance iteration idea of Gu (1992), which is based on Newton's method and can evaluate in an efficient and stable way the components in (4.23) along with their first and second derivatives with respect to $\log(\boldsymbol{\lambda})$, because the smoothing parameters can only take positive values.

The methods for estimating $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$ are iterated until the algorithm satisfies the criterion $\left|\ell(\boldsymbol{\varphi}^{[a+1]}) - \ell(\boldsymbol{\varphi}^{[a]})\right| \,/\, \left(0.1 + \left|\ell(\boldsymbol{\varphi}^{[a+1]})\right|\right) \leq 10^{-7}$. Starting values are obtained by fitting two non-informative models for the survival and censoring

times.

In this section, we have presented a penalized likelihood approach to estimate the parameters of the informative model, which is based on an optimization scheme that allows for the reliable simultaneous estimation of all model parameters. In the next section, we will analyse the theoretical properties of the non-informative and informative estimators.

4.3.3 Asymptotic properties of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$

In this section, we derive the $\sqrt{n}$ consistency and asymptotic normality of the non-informative and informative estimators, and shed light on the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart. If the estimators are $\sqrt{n}$ consistent, $\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^0)$ is bounded in probability ($\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^0 = O_p(n^{-\frac{1}{2}})$). Intuitively, this means that the probability that $\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^0)$ takes on extreme values is small, implying that both estimators will converge in probability to their true values at a rate of $\frac{1}{\sqrt{n}}$.

To study the asymptotic properties of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ two approaches can be used. In the first approach, to obtain consistency, it must be assumed that the number of knots increases with sample size n. In the other approach, we have to fix the number and location of knots. In this case, as in parametric modelling, the number of parameters is fixed, and the parameters can be estimated at the usual $\sqrt{n}$ rates. However, as in non-parametric modelling, the model is flexible enough to adapt to regression functions of unknown form (Xingwei et al., 2010). In this work, we use the fixed-knot asymptotic framework since it is closer to practical statistical modelling (e.g., Vatter & Chavez-Demoulin, 2015, and references therein). In what follows, we define $\hat{S}_{\nu 0}(\hat{\boldsymbol{\varphi}}_{\nu 0}) = \mathcal{G}_{\nu 0}[s(\hat{\boldsymbol{\varphi}}_{\nu 0})]$ as the short notation for $\hat{S}_{\nu 0}(y_i, \hat{\boldsymbol{\varphi}}_{\nu 0}) = \mathcal{G}_{\nu 0}[s(y_i, \hat{\boldsymbol{\varphi}}_{\nu 0})]$ and $\boldsymbol{\varphi}^0$ as the true vector of parameters.

The informative penalized maximum log-likelihood estimator (IPMLE) can be

defined as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\mathcal{S}_{\boldsymbol{\alpha}}}{\operatorname{argmax}}\ \ell_p(\boldsymbol{\alpha}),$$

and the non-informative counterpart (NPMLE) as

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathcal{S}_{\boldsymbol{\gamma}}}{\operatorname{argmax}}\ \ell_p(\boldsymbol{\gamma}).$$

Theorem 1 (Asymptotic properties of the IPMLE estimator)**.** Under assumptions (A1)-(A8) in Appendix B.4.1,

(i) the informative penalized maximum log-likelihood estimator $\hat{\boldsymbol{\alpha}}$ exists, is $\sqrt{n}$-consistent and

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\alpha}^0)]^{-1}\right\},$$

where $\mathcal{I}(\boldsymbol{\alpha}^0) = \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\mathbf{w};\boldsymbol{\alpha}^0)]$ with $\mathbf{w}$ containing the response and covariate vectors.

(ii) $\hat{S}_{10}(\hat{\boldsymbol{\alpha}}_{10})$ is asymptotically independent of $\hat{S}_{20}(\hat{\boldsymbol{\alpha}}_{20})$ and

$$\sqrt{n}[\hat{S}_{\nu 0}(\hat{\boldsymbol{\alpha}}_{\nu 0}) - S_{\nu 0}(\boldsymbol{\alpha}^0_{\nu 0})] \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}^0_{\nu 0}}\right\},\ \nu = 1, 2,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}^0_{\nu 0}} = \mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^0_{\nu 0})]\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^0_{\nu 0})[\mathcal{I}(\boldsymbol{\alpha}^0_{\nu 0})]^{-1}\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^0_{\nu 0})^\top \mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^0_{\nu 0})]$ and $\mathcal{I}(\boldsymbol{\alpha}^0_{\nu 0}) = \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu 0}}\ell(\mathbf{w};\boldsymbol{\alpha}^0_{\nu 0})]$.

Theorem 2 (Asymptotic properties of the NPMLE estimator)**.** Under assumptions (A1)-(A8) in Appendix B.4.1,

(i) the non-informative penalized maximum log-likelihood estimator $\hat{\boldsymbol{\gamma}}$ exists, is

$\sqrt{n}$-consistent and

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\gamma}^0)]^{-1}\right\},$$

where $\mathcal{I}(\boldsymbol{\gamma}^0) = \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\mathbf{w}; \boldsymbol{\gamma}^0)]$ with $\mathbf{w}$ containing the response and covariate vectors.

(ii) $\hat{S}_{10}(\hat{\boldsymbol{\gamma}}_{10})$ is asymptotically independent of $\hat{S}_{20}(\hat{\boldsymbol{\gamma}}_{20})$ and

$$\sqrt{n}[\hat{S}_{\nu 0}(\hat{\boldsymbol{\gamma}}_{\nu 0}) - S_{\nu 0}(\boldsymbol{\gamma}^0_{\nu 0})] \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0_{\nu 0}}\right\}, \ \nu = 1, 2,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0_{\nu 0}} = \mathcal{G}'_{\nu 0}[s(\boldsymbol{\gamma}^0_{\nu 0})]\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu 0}} s(\boldsymbol{\gamma}^0_{\nu 0})[\mathcal{I}(\boldsymbol{\gamma}^0_{\nu 0})]^{-1}\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu 0}} s(\boldsymbol{\gamma}^0_{\nu 0})^\top \mathcal{G}'_{\nu 0}[s(\boldsymbol{\gamma}^0_{\nu 0})]$ and $\mathcal{I}(\boldsymbol{\gamma}^0_{\nu 0}) = \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\gamma}_{\nu 0}\boldsymbol{\gamma}_{\nu 0}}\ell(\mathbf{w}; \boldsymbol{\gamma}^0_{\nu 0})]$.

Theorem 3 (Efficiency of the IPMLE estimator)**.** For $\nu = 1, 2$, let $\boldsymbol{\gamma}_\nu = (\boldsymbol{\gamma}^\iota_\nu, \boldsymbol{\gamma}^{n\iota}_\nu)^\top$ be the informative and non-informative parameters of the non-informative model, respectively. Under assumptions (A1)-(A8) in Appendix B.4.1, and if we further assume that $\boldsymbol{\gamma}^{n\iota}_{\nu 0} = \boldsymbol{\alpha}_{\nu 0}$, then

$$\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\alpha}}_0) < \mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^\iota_\nu),$$
$$\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\alpha}}_\nu) < \mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{n\iota}_\nu),$$

where $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\alpha}}_0) = \Sigma_{\boldsymbol{\alpha}^0_0}$, $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\alpha}}_\nu) = \Sigma_{\boldsymbol{\alpha}^0_\nu}$, $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^\iota_\nu) = \Sigma_{\boldsymbol{\gamma}^{0\iota}_\nu}$, and $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{n\iota}_\nu) = \Sigma_{\boldsymbol{\gamma}^{0n\iota}_\nu}$ represent the asymptotic covariance matrices of $\hat{\boldsymbol{\alpha}}_0$, $\hat{\boldsymbol{\alpha}}_\nu$, $\hat{\boldsymbol{\gamma}}^\iota_\nu$ and $\hat{\boldsymbol{\gamma}}^{n\iota}_\nu$ respectively.

The proofs of Theorems 1, 2 and 3 are given in Appendix B.4.2.

Remark 2. The fact that the informative and non-informative survival functions are orthogonal (part (ii) of Theorems 1 and 2) suggests that the estimation algorithm will yield more accurate parameter vector updates throughout the iterations (e.g.,

Nocedal & Wright, 2006). Moreover, Theorem 3 shows that under informative censoring it is possible to estimate the model's coefficients more efficiently since more information is exploited by the informative model.

The construction of confidence intervals and p-values are discussed in the next section of this chapter.

4.3.4 Confidence intervals and p-values

As far as the construction of confidence intervals and p-values are concerned, for practical purposes it is convenient to adapt to the current context the results discussed in Marra et al. (2017).

In particular, at convergence, point-wise intervals for linear and non-linear functions for both the non-informative and informative models parameters can be obtained using the following Bayesian large sample approximation

$$\boldsymbol{\varphi} \sim \mathcal{N}(\hat{\boldsymbol{\varphi}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\varphi}}}), \tag{4.24}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\varphi}}} = [\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\varphi}})]^{-1}$. For generalised additive models, intervals derived using equation (4.24) have good frequentist properties, since they account for both smoothing bias and sampling variability (Marra & Wood, 2012). For the non-informative and informative models, equation (4.24) can be verified using the distribution of $\boldsymbol{\mathcal{Z}}$ (described in Section 4.3.2), making the large sample assumption that $\boldsymbol{\mathcal{H}}(\boldsymbol{\varphi})$ can be treated as fixed, and making the usual prior Bayesian assumption for smooth models $\boldsymbol{\varphi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}^{-1}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{\Lambda}$ (Silverman, 1985; Wood, 2017). In equation (4.24), smoothing parameter uncertainty is neglected. Nevertheless, according to Marra & Wood (2012) this is not problematic if heavy over-smoothing is avoided so that the smoothing bias is not a large proportion of the sampling variability. See also Marra et al. (2017) for

an application of this approach to a more general smoothing spline context.

Following Pya & Wood (2015), confidence interval estimates for the monotonic smooth terms in the models can be obtained using the distribution of $\tilde{\boldsymbol{\varphi}}_{\nu 0}$ (defined in Section 4.2.2) since all smooth components would then depend linearly on $\tilde{\boldsymbol{\varphi}}_{\nu 0}$. Such a distribution is

$$\tilde{\boldsymbol{\varphi}}_{\nu 0} \sim \mathcal{N}(\hat{\tilde{\boldsymbol{\varphi}}}_{\nu 0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varphi}}_{\nu 0}}),$$

where $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varphi}}_{\nu 0}} = \text{diag}(\boldsymbol{\Gamma}_{\nu 0})\ [\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\varphi}}_{\nu 0})]^{-1}\,\text{diag}(\boldsymbol{\Gamma}_{\nu 0})$. The derivation of this result can be found in Pya & Wood (2015).

P-values for the smooth components in the non-informative and informative models are obtained by adapting the results discussed in Wood (2013) to the present context, where $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varphi}}_{\nu 0}}$ is used for the calculations. The reader is referred to the above citation for the definition of reference degrees of freedom.

In this section, the asymptotic properties of the non-informative and informative estimators were derived. We also shed light on the efficiency gains produced by the informative estimator when compared to its non-informative counterpart. The construction of confidence intervals and p-values were discussed in the last section. In the next sections, we will evaluate the performance of the proposed methodology through a Monte Carlo simulation study and an empirical application.

4.4 Simulation study

This section provides evidence on the empirical effectiveness of the proposed methodology in recovering true linear effects, non linear effects and baseline functions under informative censoring for three Data Generating Processes (DGPs). The performance of the informative penalized maximum log-likelihood estimator against that of its non-informative counterpart was also examined.

(i) DGP1 (z_{1i} non-informative, z_{2i} informative and censoring rate of about 78%). Event times, T_{1i}, were generated from a proportional hazard model, while censored times, T_{2i}, were generated from a proportional odd model. These, defined on the survival function scale, are given by

$$\begin{aligned} &\log\left[-\log\left\{S_{10}(t_{1i})\right\}\right] + \alpha_{01} + \alpha_{11} z_{1i} + s_{11}(z_{2i}), \\ &\log\left[\frac{\left\{1 - S_{20}(t_{2i})\right\}}{S_{20}(t_{2i})}\right] + \alpha_{02} + \alpha_{12} z_{1i} + s_{12}(z_{2i}), \end{aligned} \tag{4.25}$$

where $S_{10}(t_{1i}) = 0.72\exp\left(-0.4t_{1i}^{2.4}\right) + 0.28\exp\left(-0.1t_{1i}^{1.0}\right)$ and $S_{20}(t_{2i}) = 0.99\exp\left(-0.1t_{2i}^{2.2}\right) + 0.01\exp\left(-0.4t_{2i}^{1.1}\right)$ (Crowther & Lambert, 2013). Covariate z_{1i} was generated using a binomial distribution and z_{2i} using a uniform distribution. As for the smooth functions, we used $s_{11}(z_{2i}) = s_{12}(z_{2i}) = -0.2\exp(3.2z_i)$, whereas the parametric coefficients were: $\alpha_{01} = 0.25$, $\alpha_{02} = 0.85$, $\alpha_{11} = -2.0$ and $\alpha_{12} = 1.8$.

Sample sizes were set to 500, 1000 and 4000, and the number of replicates to 1000. Replicates in which the models did not converge were discarded and replaced with additional ones. The models were fitted using `gamlss()` in `GJRM` by employing the proportional hazard link (`"PH"`) for the event times and the proportional odd link (`"PO"`) for the censoring times (see Appendix B.5). The smooth components of z_2 were represented using penalized low rank thin plate splines with second order penalty and 10 bases (the default in `GJRM`), and the smooths of times using monotonic penalized B-splines with penalty defined in Section 4.2.2 and 10 bases. Note that smooth terms of explanatory variables can also be represented using different spline definitions (see Appendix B.5). In the case of one-dimensional smooth functions, all definitions lead to virtually the same result as long as the amount of smoothing is selected in a data-driven manner (e.g., Wood, 2017). For each replicate,

curve estimates were constructed using 200 equally spaced fixed values in the $(0, 8)$ range for the baseline functions and $(0, 1)$ otherwise.

Results: Regarding the estimates for α_{11} (the parameter of the non-informative covariate), Figure 4.1 and Table 4.2 show that overall the mean estimates for the IPMLE and NPMLE are very close to the respective true values and improve as the sample size increases, and that the variability of the estimates decreases as the sample size grows large.

As for the smooth effect of the informative covariate, Figures 4.2 and 4.3, and Table 4.2 show that overall the true functions are recovered well by the proposed estimation methods and that the results improve in terms of bias and efficiency as the sample size increases. However, the IPMLE is more efficient than the NPMLE for all sample sizes examined in the simulation study; for example, for n $= 500, 1000$ the RMSE for the NPMLE is more than twice as large as the IPMLE. Some gains in efficiency are also observed for the baseline functions.

(ii) DGP2 (z_{1i} informative, z_{2i} informative and censoring rate of about 74%). As for DGP1, T_{1i} and T_{2i} were generated using the model defined in (4.25). However, in this case, the baseline survival functions were defined as $S_{10}(t_{1i}) = 0.75\exp\left(-0.4t_{1i}^{2.4}\right) + 0.25\exp\left(-0.1t_{1i}^{1.0}\right)$ and $S_{20}(t_{2i}) = 0.99\exp\left(-0.1t_{2i}^{2.2}\right) + 0.01\exp\left(-0.4t_{2i}^{1.1}\right)$. The informative covariates, z_{1i} and z_{2i}, were generated using binomial and uniform distributions, respectively. Finally, $s_{11}(z_{2i}) = s_{12}(z_{2i}) = -0.2\exp(3.2z_i)$, $\alpha_{01} = 0.25$, $\alpha_{02} = 0.85$ and $\alpha_{11} = \alpha_{12} = -1.5$.

Results: Similarly to DGP1, Figures B.1, B.8 and B.9, and Table B.1 (in Appendix B.6) show that overall the mean estimates for the two estimators are very close to the respective true values and improve as the sample size increases. The variability of the estimates also decreases as the sample size

(a) Informative Penalized Maximum Log-likelihood Estimator (IPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.047	-0.013	-0.001	0.369	0.239	0.118
s_{11}	0.036	0.028	0.013	0.161	0.114	0.061
h_{10}	0.095	0.069	0.034	0.336	0.245	0.104
S_{10}	0.027	0.024	0.018	0.071	0.054	0.033
(b) Non-informative Penalized Maximum Log-likelihood Estimator (NPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.079	-0.015	-0.005	0.360	0.245	0.116
s_{11}	0.085	0.069	0.046	0.383	0.206	0.118
h_{10}	0.120	0.070	0.034	0.427	0.292	0.121
S_{10}	0.034	0.025	0.017	0.086	0.068	0.039

Table 4.2: Bias and root mean squared error (RMSE) for the IPMLE and NPMLE obtained by applying the `gamlss()` to informative survival data simulated according to DGP1 characterised by a censoring rate of about 78%. Bias and RMSE for the smooth terms are calculated, respectively, as $n_s^{-1}\sum_{i=1}^{n_s}|\bar{\hat{s}}_i - s_i|$ and $n_s^{-1}\sum_{i=1}^{n_s}\sqrt{n_{rep}^{-1}\sum_{rep=1}^{n_{rep}}(\hat{s}_{rep,i}-s_i)^2}$, where $\bar{\hat{s}}_i = n_{rep}^{-1}\sum_{rep=1}^{n_{rep}}\hat{s}_{rep,i}$, n_s is the number of equally spaced fixed values in the $(0,8)$ or $(0,1)$ range, and n_{rep} is the number of simulation replicates. In this case, $n_s = 200$ and $n_{rep} = 1000$. The bias for the smooth terms is based on absolute differences in order to avoid compensating effects when taking the sum.

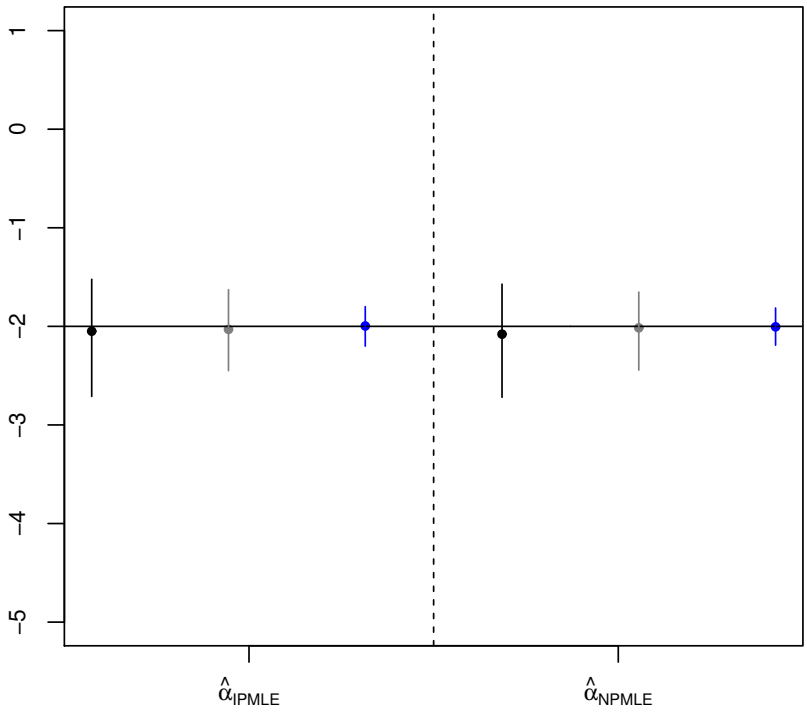

Figure 4.1: Linear coefficient estimates obtained by applying `gamlss()` to informative survival data simulated according to DGP1 which is characterised by a censoring rate of about 78%. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 1000$ and $n = 4000$ are given in dark gray and blue, respectively.

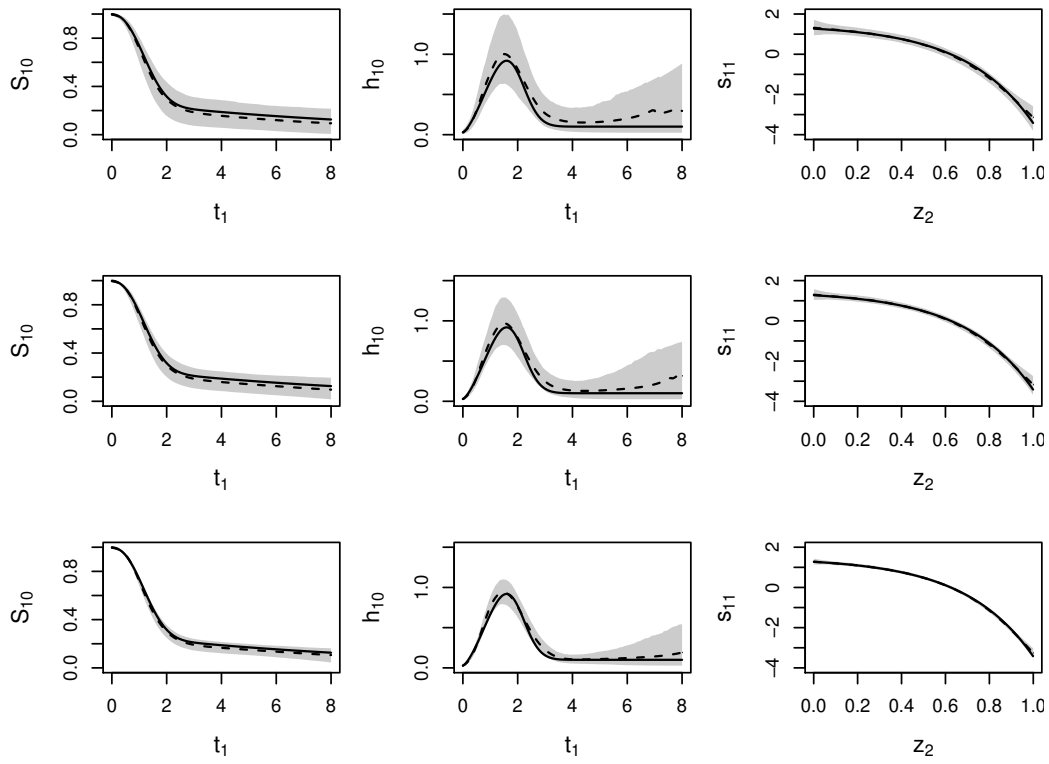

Figure 4.2: Smooth function estimates for the IPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP1 characterised by a censoring rate of about 78%. True functions are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting from 5% and 95% quantiles by shaded areas. The results in the first row refer to $n = 500$, whereas those in the second and third rows to $n = 1000$ and $n = 4000$.

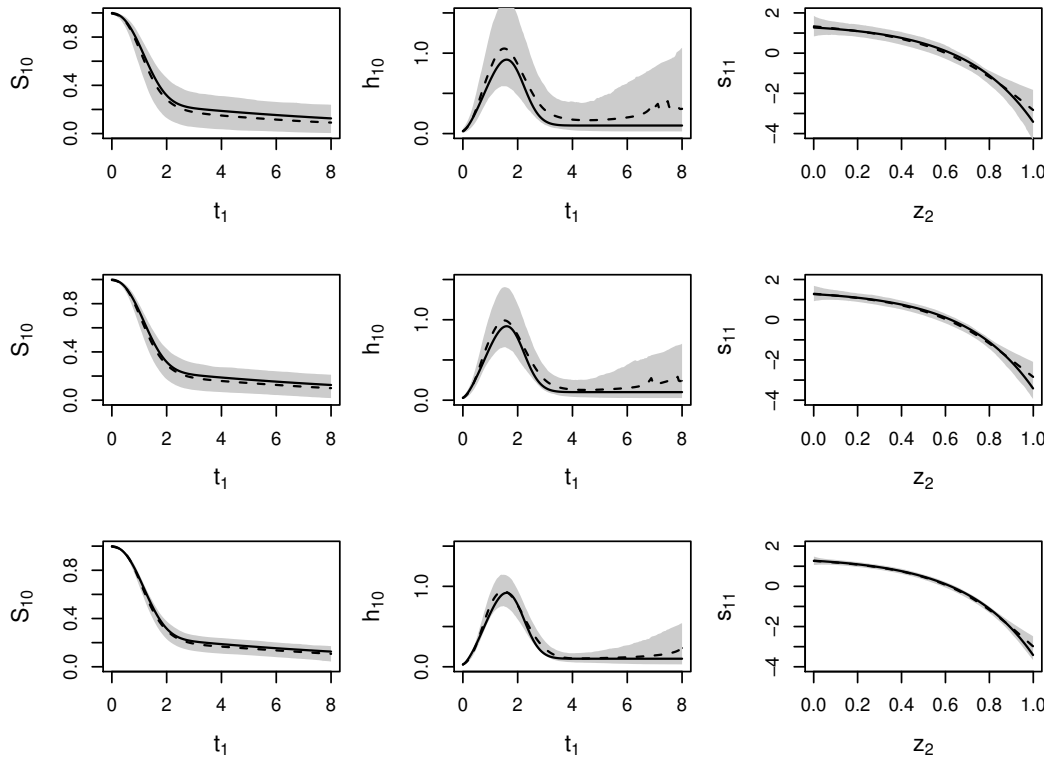

Figure 4.3: Smooth function estimates for the NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP1 characterised by a censoring rate of about 78%. Further details are given in the caption of Figure 4.2.

grows large. However, the IPMLE is significantly more efficient than the NPMLE for all cases considered.

Computing times for the proposed approach were on average 8 seconds for $n = 4000$ and around 5 seconds for $n = 1000, 500$. In addition, two DGPs (DGP3 and DGP4) with a different smooth function for z_{2i} and with censoring rates of about 47% and 29% respectively were explored (in Appendix B.6). These DGPs suggested that the gain in efficiency of the IPMLE tends not to be too significant when lower censoring rates are considered. Finally, for the above DGPs, we explored the ability of information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), defined in Supplementary Material B.1, to select the correct model. When doing this, we also considered the informative estimator with incorrectly chosen set of informative covariates (e.g., for DGP1, in estimation, z_1 was assumed to be informative instead of z_2). For all sample sizes and cases considered, based on AIC/BIC, the correct model is always chosen.

4.5 Empirical illustration

The modelling framework is illustrated using the data employed by Lu & Zhang (2012), where the aim was to assess how several factors affect the contraction of pneumonia in infants in the presence of informative censoring. According to the World Health Organization (WHO), pneumonia accounted for 16% of all deaths of children under five years old in 2015. The data set consists of 3470 annual personal interviews conducted for the National Longitudinal Survey of Youth from 1979 through 1986 (NLSY, 1995). The response variable, y_i, is the age, in months, at which the infant was hospitalised for pneumonia, and 97.9% of this variable is right censored.

The covariates considered in the modeling were age of the mother in years (`mthage`), `urban` environment (urban = 1, rural = 0), `region` (1 = north-east, 2 = north central, 3 = south, 4 = west), `poverty` (1 = yes, 0 = no), whether the infant had a normal birth weight as defined by weighting at least 5.5 pounds (`wmonth` = 1 if yes and 0 otherwise), `race` (1 = white, 2 = black, 3 = other), `education` (years of school of mother), month the child started to be on solid food (`sfmonth`), average number of cigarettes smoked per week during pregnancy (`smoke` = 0, 1 or 2) and `alcohol` used by mother during pregnancy (0, 1, 2), where the higher the number the higher the frequency of alcohol consumption. To capture the effect of housing crowding (since pneumonia is a communicable disease), number of siblings of the child (`nsibs`) was considered and grouped in three categories (0 for infants without siblings, 1 for infants with one to three siblings, and 2 for more than three siblings.

To assess whether the censoring mechanism was informative, we employed the AIC, BIC, and K-Fold Cross validation (Υ^{KCV}) with $K = 20$ (decreasing or increasing this value did not alter the conclusions); see Appendix B.1 for their definitions. Since several combinations of covariates and link functions had to

be considered, a number of models were tried out and the final models selected using the above mentioned criteria. Table 4.3 shows the results for the chosen models and supports the presence of informative censoring through the `alcohol` and `region` variables (Model 3). Table 4.4 and Figure 4.4 present the results for Model 3 and Model 1 (the latter neglects informative censoring).

Main findings: From a quick overall look at Table 4.4, the results exhibit a smaller estimation uncertainty for the informative model. Analysing the table in more detail, the coefficients of `wmonth`, `nsibs1`, `nsibs2` are statistically significant for both models. For instance, the expected hazard for infants with one to three siblings is 2 times that for infants without siblings. Similarly, the expected hazard is 7.3 times higher in infants with more than 3 siblings as compared to infants with no siblings. The parameter of category `alcohol1` of the `alcohol` variable is statistically significant at the 10% level for the informative model and is not significant for the non-informative model. The implication of this result is that using the non-informative model the variable `alcohol` would most likely be removed from the model, hence missing out on some potentially important behavioral patterns. The survival and hazard curves are very similar across the two models with the main difference that the informative approach yields considerably less variable estimates (Figure 4.4). Our results are consistent with those of Lu & Zhang (2012) who found that the censoring mechanism is informative in this dataset, and that the informative model provides a better fit as compared to its non informative counterpart.

Model	Non-Inf. Cov.	Inf. Cov.	Link T_{1i}	Link T_{2i}	AIC	Υ^{KCV}	BIC
1	s(wmonth) s(mthage) region alcohol nsibs	...	PH	PH	13601.11	-6828.74	13878.34
2	s(wmonth) s(mthage) region alcohol nsibs	...	PO	PH	13602.28	-8710.51	13879.61
3	s(wmonth) s(mthage) nsibs	alcohol region	PH	PH	13597.75	-6826.89	13844.08
4	s(wmonth) s(mthage) nsibs	alcohol region	PO	PH	13598.87	-8716.34	13845.33

Table 4.3: Values of three model selection criteria (AIC, BIC and Υ^{KCV}) for the best informative and non-informative models fitted to the real data application of this paper. The models were fitted using `gamlss()` in `GJRM` by employing different combinations of covariates and link functions.

(a) Model 1 (NPMLE)				
Linear Covariates	Estimate	Standard Error	Z-value	P-value
intercept	-77.15	36.13	-2.135	0.033 *
alcohol1	0.324	0.309	1.048	0.294
alcohol2	-0.185	0.336	-0.551	0.582
nsibs1	0.697	0.261	2.670	0.008 **
nsibs2	1.959	0.760	2.578	0.009 **
region2	0.138	0.343	0.401	0.689
region3	-0.384	0.342	-1.121	0.262
region4	-0.490	0.437	-1.120	0.263
wmonth	-0.809	0.294	-2.757	0.006 **
Smooth Variables	EDF	Ref. DF	Chi-square	P-value
s(u)	7.747	8.619	101.71	<2e-16 ***
s(mthage)	2.141	2.720	4.045	0.276

(b) Model 3 (IPMLE)				
Linear Covariates	Estimate	Standard Error	Z-value	P-value
intercept	-77.37	36.14	-2.141	0.032 *
alcohol1	0.077	0.046	1.665	0.096 .
alcohol2	-0.048	0.046	-0.046	0.295
nsibs1	0.685	0.259	2.641	0.008 **
nsibs2	1.986	0.754	2.635	0.008 *
region2	0.035	0.056	0.626	0.531
region3	0.070	0.052	1.335	0.182
region4	0.041	0.059	0.688	0.492
wmonth	-0.791	0.289	-2.739	0.006 **
Smooth Variables	EDF	Ref. DF	Chi-square	P-value
s(u)	7.747	8.620	101.65	<2e-16 ***
s(mthage)	2.119	2.692	3.741	0.31

Table 4.4: Estimation results of the non-informative and informative models (Models 1 and 3, respectively, in Table 4.3) applied to pneumonia data. The models were fitted using `gamlss()` in `GJRM` by employing the "`PH-PH`" link functions combination. Furthermore, EDF and Ref. DF refer to the effective degrees of freedom and reference degrees of freedom of the smooths. More details can be found in Sections 4.3.2 and 4.3.4.

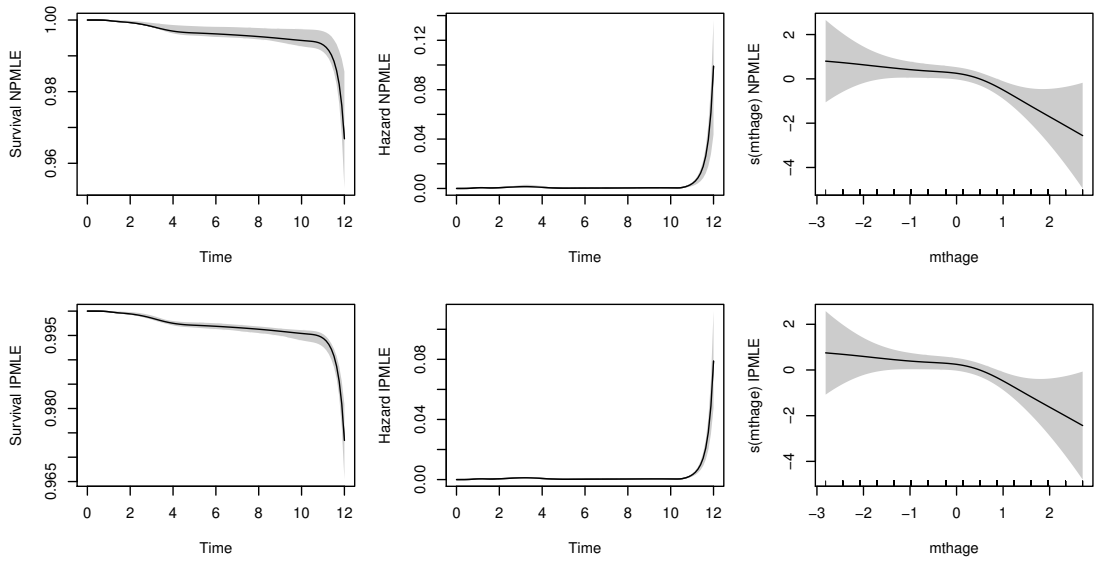

Figure 4.4: Smooth function estimates and their corresponding 95% intervals for Model 1 (non-informative model) and Model 3 (informative model) obtained by applying `gamlss()` in `GJRM` to pneumonia data. The intervals have been obtained using the approach described in Section 4.3.4.

4.6 Concluding remarks

In this chapter, we have introduced generalized link-based additive survival models with informative censoring and their potential was illustrated using simulated and real data. The proofs of the $\sqrt{n}$-consistency and asymptotic normality of the non-informative and informative estimators have been provided. Further, we showed that the newly introduced informative estimator is more efficient than its non-informative counterpart.

Important features of the modelling framework are that: the survival models can account for informative censoring; the baseline functions are estimated non-parametrically via means of monotonic P-splines, which allows one to obtain coherent estimated survival functions; covariate effects are flexibly determined using additive predictors; the optimization scheme allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection; the models can be easily utilized

using the freely available `GJRM R` package which allows for several modelling choice as well as for transparent and reproducible research. Given that the assumption of absence of informative censoring is often made for mathematical convenience as well as lack of software, the proposed methodology is likely to appeal to researchers in various fields wishing to estimate possibly more realistic survival models.

The next chapter will look into the feasibility of modelling jointly the event and the censoring times, through a flexible bivariate copula model, where the aim is to correct for dependent censoring.

Chapter 5

Survival Link-Based Additive Models with Dependent Censoring

In this chapter, we propose a flexible regression survival model that accounts for administrative and dependent right censoring, and provide preliminary arguments towards model identification although this topic is very challenging and requires more future work. The strength of the association between the event and censoring times is modelled via a copula structure whose dependence parameter is estimated from the data, and the margins are determined using link-based functions models. Baseline functions are non-parametrically estimated using monotonic P-splines and covariate effects are flexibly determined using additive predictors. Parameter estimation is efficiently achieved within a penalised maximum likelihood framework, and the consistency and asymptotic normality of the proposed estimator are also derived under the assumption that the model is identified. The finite sample properties of the dependent estimator are investigated via a Monte Carlo simulation study, and the proposal illustrated using prostate cancer data. The results highlight the

effectiveness of the methodology proposed, and the relevant numerical computation can be easily carried out using the function `gjrm()` in the `R` package `GJRM` (Marra & Radice, 2020b). Although establishing formal identification will require more future work, the practical performance of the proposed estimator suggests that the approach can effectively and flexibly deal with dependent censoring.

5.1 Introduction

When time to event data are analysed, it is often assumed that the censoring mechanism is independent, a strong assumption in many empirical situations. Standard modelling techniques assume that the observed and unobserved parts of the data are related via means of the random variables $\{y = \min(T_1, T_2), \delta = I(T_1 < T_2)\}$, where I is the usual indicator function.

Most estimation methods assume that T_1 and T_2 are stochastically independent (e.g., Cox, 1972; Ma et al., 2014; Scheike & Zhang, 2003; Wu & Witten, 2019; Younes & Lachin, 1997). However, this assumption may be questioned. If individuals are right censored because the study ends before they have experienced the event of interest (a situation typically called administrative censoring) then it is reasonable to make the assumption of independence. However, if individuals are lost to follow up or withdraw from the study then the cause of this may be related to the event time. For example, individuals may withdraw from a study because they are showing side effects that need alternative treatments or because their condition is worsening or may withdraw from a study because they are feeling healthy and hence do not wish to continue with the treatment (e.g., Deresa & Van Keilegom, 2019; Moradian et al., 2019; Xu et al., 2018; Willems et al., 2018; Staplin et al., 2015). Note that right censoring can also be generated by competing risks, that is, the occurrence of another event which precludes the main event of interest from occurring. Often, these mutually exclusive events are dependent. For instance,

diabetes and hypertension are closely related as they share similar risk factors such as vascular inflammation, arterial remodelling and obesity, among others (Petrie et al., 2018). Ignoring the possible latent causes of censoring may lead to misleading inference (e.g., Crowder, 1991; Siannis et al., 2005).

Let $S_{T_1,T_2}(t_1, t_2)$ be the joint survival function of (T_1, T_2) and $f_{y,\delta}(y, \cdot)$ the sub-density function of (y, δ). According to Tsiatis (1975), if T_1 and T_2 are dependent then $S_{T_1,T_2}(t_1, t_2)$ is not identifiable from the sub-density function $f_{y,\delta}(y, \cdot)$. That is, given any joint survival function $S_{T_1,T_2}(t_1, t_2)$ with arbitrary dependence between T_1 and T_2, there exists a different joint survival function $S^{*}_{T_1,T_2}(t_1, t_2)$, in which T_1 and T_2 are independent, that reproduces $f_{y,\delta}(y, \cdot)$ precisely. Therefore, in order to identify the joint distribution of T_1 and T_2, we need additional information about their dependence. Several approaches have been proposed in the survival analysis and competing risk literature to deal with dependent censoring (e.g., Crowder, 2012; Emura & Chen, 2018). In a seminal work, Zheng & Klein (1995) proposed a copula model to account for the dependence between T_1 and T_2, where the marginal distribution of T_1 is estimated non-parametrically. Their proposed estimator is consistent and reduces to the Kaplan-Meier estimator when T_1 and T_2 are independent. This approach was further investigated by Rivest & Wells (2001) in the context of an Archimedean copula. Since this model did not incorporate covariates, it was further extended (e.g., Braekers & Veraverbeke, 2005; Huang & Zhang, 2008; Chen, 2010; Sujica & Van Keilegom, 2018). Dependent censoring can also be adjusted for via multiple imputation (Jackson et al., 2014), auxiliary information (Scharfstein & Robins, 2002; Hsu et al., 2015) and the inverse probability censoring weighted model (Robins & Finkelstein, 2000). Some scholars have exploited parametric restrictions on the joint bivariate distribution of T_1 and T_2 (Basu & Ghosh, 1978; Basu, 1988; Emoto & Matthews, 1990; Deresa & Van Keilegom, 2019). For example, Basu & Ghosh (1978), using a bivariate normal distribution

for T_1, T_2, showed that the distribution of $\{y = \min(T_1, T_2), \delta = I(T_1 < T_2)\}$ identifies the bivariate joint distribution of the pair. This result was extended by Deresa & Van Keilegom (2019) to include covariates. In particular, they do so by employing a class of monotonic parametric transformations of the logarithm of the survival and censoring times which are assumed to follow a multivariate normal distribution. However, even when the functional form of the joint distribution of T_1 and T_2 is known, the model's parameters may or may not be identified by the joint distribution of (Y, δ) (e.g., Basu, 1988; Rao, 1992).

Generally, there are not known global conditions for unique solutions of system of non-linear equations, hence it is difficult to verify whether a nonlinear model is globally identifiable (e.g., Koop et al., 2013; McCullagh, 2002; Koopmans & Reiersol, 1950). However, sufficient conditions for parametric and nonparametric local identification are point-wise differentiability at $\boldsymbol{\vartheta}^0$ (the true vector of parameters), and a rank condition (e.g., Chen et al., 2014; Stanghellini et al., 2013). Stanghellini et al. (2013) used rank conditions to show the local identifiability of discrete graphical models with one latent variable. Chen et al. (2014) extended the rank conditions to nonparametric and nonlinear structural models, and then used them to identify, for example, non-separable quantile instrumental variable models and semiparametric consumption-based asset pricing models. Furthermore, local identification allows for the consistent estimation of the model's parameters, and is therefore sufficient for the parameter estimator to have the usual asymptotic properties (Rao, 1992).

The remainder of this chapter is organized as follows. In Section 5.2, we will propose a flexible regression survival model that accounts for administrative and dependent right censoring, where the strength of the association between the event and censoring times is modelled via a copula structure whose dependence parameter is estimated from the data. Baseline functions are non-parametric and covariate

effects are flexibly estimated using the approach already discussed in Section 4.2. Then, in Section 5.3, we provide some ideas to prove identification for the proposed family of models. Parameter estimation as well as the consistency and asymptotic normality of the proposed estimator are discussed in Sections 5.4 and 5.5. Finally, in Sections 5.6 and 5.7, the finite sample properties of the dependent estimator are investigated via a Monte Carlo simulation study, and the proposal illustrated using prostate cancer data.

5.2 Model formulation

We consider the case of right censored data where the true event times are not always recorded, in which case the censoring times are observed. For individual i, with $i = 1, 2, \ldots, n$ where n is the sample size, let (T_{1i}, T_{2i}, T_{3i}) denote a vector of event, dependent censoring and administrative censoring times, respectively, and $\mathbf{z}_i$ a generic vector of individual characteristics. We observe $y_i = \min\{T_{1i}, T_{2i}, T_{3i}\} \in \mathbb{R}^+$ and the corresponding censoring indicator functions $\delta_{1i} = I\{y_i = T_{1i}\}$, $\delta_{2i} = I\{y_i = T_{2i}\}$ and $\delta_{3i} = (1 - \delta_{1i} - \delta_{2i})$.

Let T_{1i} and T_{2i} have conditional marginal survival functions generically expressed as $S_\nu(t_{\nu i}|\mathbf{z}_{\nu i}; \boldsymbol{\gamma}_\nu) = P(T_{\nu i} > t_{\nu i}|\mathbf{z}_{\nu i}; \boldsymbol{\gamma}_\nu) \in (0, 1)$, for $\nu = 1, 2$, and conditional joint survival function defined as $S(t_{1i}, t_{2i}|\mathbf{z}_i; \boldsymbol{\vartheta}) = P(T_{1i} > t_{1i}, T_{2i} > t_{2i}|\mathbf{z}_i; \boldsymbol{\vartheta})$. To link T_{1i} and T_{2i} the following copula model is assumed

$$S(t_{1i}, t_{2i}|\mathbf{z}_i; \boldsymbol{\vartheta}) = C\left(S_1(t_{1i}|\mathbf{z}_{1i}; \boldsymbol{\gamma}_1), S_2(t_{2i}|\mathbf{z}_{2i}; \boldsymbol{\gamma}_2); \theta\right), \tag{5.1}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \Theta$, $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are vectors of covariates (which can be equal to $\mathbf{z}_i$ but have not to) with associated coefficient vectors $\boldsymbol{\gamma}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^{p_2}$ of dimensions p_1 and p_2 and $C : [0, 1]^2 \to [0, 1]$ is a uniquely defined 2-dimensional one-parameter copula function with coefficient $\theta \in \Theta \subseteq \mathbb{R}$, capturing

the conditional dependence of (T_{1i}, T_{2i}) (e.g., Nelsen, 2006; Marra & Radice, 2020a; Sklar, 1973). Note that Θ depends on the choice of the copula structure; for example, for the Gaussian copula we have that $\Theta = [-1, 1]$. Table 5.1 displays the copulae functions available in `GJRM` for practical modelling.

`Copula`	$C(p_1, p_2; \theta)$	`Range of` θ	`Kendall's` τ
`AMH ("AMH")`	$\frac{p_1 p_2}{1-\theta(1-p_1)(1-p_2)}$	$\theta \in [-1, 1]$	$-\frac{2}{3\theta^2}\left\{\theta + (1-\theta)^2 \log(1-\theta)\right\} + 1$
`Clayton ("C0")`	$\left(p_1^{-\theta} + p_2^{-\theta} - 1\right)^{-1/\theta}$	$\theta \in (0, \infty)$	$\frac{\theta}{\theta+2}$
`FGM ("FGM")`	$p_1 p_2 \{1 + \theta(1-p_1)(1-p_2)\}$	$\theta \in [-1, 1]$	$\frac{2}{9}\theta$
`Frank ("F")`	$-\theta^{-1} \log\{1 + (\exp\{-\theta p_1\} - 1)$ $(\exp\{-\theta p_2\} - 1)/(\exp\{-\theta\} - 1)\}$	$\theta \in \mathbb{R} \setminus \{0\}$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$
`Plackett ("PL")`	$\left(Q - \sqrt{R}\right) / \{2(\theta - 1)\}$	$\theta \in (0, \infty)$	–
`Gaussian ("N")`	$\Phi_2\left(\Phi^{-1}(p_1), \Phi^{-1}(p_2); \theta\right)$	$\theta \in [-1, 1]$	$\frac{2}{\pi}\arcsin(\theta)$
`Gumbel ("G0")`	$\exp\left[-\left\{(-\log p_1)^\theta + (-\log p_2)^\theta\right\}^{1/\theta}\right]$	$\theta \in [1, \infty)$	$1 - \frac{1}{\theta}$
`Joe ("J0")`	$1 - \left\{(1-p_1)^\theta + (1-p_2)^\theta - (1-p_1)^\theta(1-p_2)^\theta\right\}^{1/\theta}$	$\theta \in (1, \infty)$	$1 + \frac{4}{\theta^2} D_2(\theta)$
`Student-t ("T")`	$t_{2,\zeta}\left(t_\zeta^{-1}(p_1), t_\zeta^{-1}(p_2); \zeta, \theta\right)$	$\theta \in [-1, 1]$	$\frac{2}{\pi}\arcsin(\theta)$

Table 5.1: Definition of the copulae implemented in `GJRM`, with corresponding parameter range of association parameter θ and relation between Kendall's τ (which takes values in the customary range $[-1, 1]$) and θ. $\Phi_2(\cdot, \cdot; \theta)$ denotes the cumulative distribution function (cdf) of a standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. $t_{2,\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the cdf of a standard bivariate Student-t distribution with correlation θ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_\zeta(\cdot)$ denotes the cdf of a univariate Student-t distribution with ζ degrees of freedom. $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1} dt$ is the Debye function and $D_2(\theta) = \int_0^1 t \log(t)(1-t)^{\frac{2(1-\theta)}{\theta}} dt$. Quantities Q and R are given by $1 + (\theta - 1)(p_1 + p_2)$ and $Q^2 - 4\theta(\theta - 1)p_1 p_2$, respectively. The Kendall's τ for `"PL"` is computed numerically as no analytical expression is available. Counter-clockwise rotated versions of copulae such as Clayton and Gumbel can be obtained using the following expressions: $C_{90} = p_2 - C(1 - p_1, p_2)$, $C_{180} = p_1 + p_2 - 1 + C(1 - p_1, 1 - p_2)$, $C_{270} = p_1 - C(p_1, 1 - p_2)$, where the subscript indicates the degree of rotation and θ has been suppressed for simplicity (e.g., Brechmann & Schepsmeier, 2013). Argument `BivD` of `gjrm()` in `GJRM` allows the user to employ the desired copula function and can be set to any of the values within brackets next to the copula names in the first column; for example, `BivD = "J0"`. For Clayton, Gumbel and Joe, the number after the capital letter indicates the degree of rotation required: the possible values are `0`, `90`, `180` and `270`.

The margins are specified using the link-based functions approach introduced

in Section 4.1. That is,

$$S_\nu(t_{\nu i}|\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu) = \mathcal{G}_\nu\left(\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)\right), \tag{5.2}$$

where $\xi_{\nu i}(t_{\nu i},\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)$ in $\mathbb{R}$, represents the non-informative additive predictors for T_{1i} and T_{2i}, whose set up was discussed in Section 4.2.2. To complete the model, we recall the final expression for the additive predictor

$$\xi_{\nu i} = \gamma_{\nu 0} + \boldsymbol{\mathcal{Q}}_{\nu 0}(y_i)^\top \boldsymbol{\Gamma}_{\nu 0}\tilde{\boldsymbol{\gamma}}_{\nu 0} + \sum_{k_\nu=1}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu}. \tag{5.3}$$

In the next section, we will provide some identification arguments for the family of models proposed in this section.

5.3 Some identification arguments

The proposed model is defined by equations (5.1), (5.2) and (5.3), which are collected below

$$\begin{aligned} S(t_1, t_2|z_1, z_2; \boldsymbol{\vartheta}) &= C\left[S_1(t_1|z_1; \boldsymbol{\gamma}_1), S_2(t_2|z_2; \boldsymbol{\gamma}_2); \theta\right], \\ S_\nu(t_\nu|z_\nu; \boldsymbol{\gamma}_\nu) &= \mathcal{G}_\nu\left[\xi_\nu(t_\nu, z_\nu; \boldsymbol{\gamma}_\nu)\right], \\ \xi_\nu(t_\nu, z_\nu; \boldsymbol{\gamma}_\nu) &= \gamma_{\nu 0} + \boldsymbol{\mathcal{Q}}_\nu(t_\nu)^\top \boldsymbol{\Gamma}_\nu \tilde{\boldsymbol{\gamma}}_\nu + \sum_{k_\nu=1}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu})^\top \boldsymbol{\gamma}_{\nu k_\nu}. \end{aligned} \tag{5.4}$$

Assume that the pair (T_1, T_2) is generated by model (5.4) with parameter vector $\boldsymbol{\vartheta}^0 = (\boldsymbol{\gamma}_1^0, \boldsymbol{\gamma}_2^0, \theta^0) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \Theta$ and that we only observe $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$. On the basis of the joint distribution of the latter vector, because in the current setting it is not possible to observe simultaneously T_1 and T_2, identification of $\boldsymbol{\vartheta}^0$ has to be proved: two different sets of parameters imply different joint distributions of $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$ in an open neighbourhood of $\boldsymbol{\vartheta}^0$.

Let us denote a particular realization of $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$ as (t, l_1, l_2, z_1, z_2), and also assume that the joint density function of vector $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$ exists and that, for $l_1, l_2 = 0, 1$, this is defined as

$$f_{y,\delta_1,\delta_2,\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2, z_1, z_2; \boldsymbol{\vartheta}) = f_{y,\delta_1,\delta_2,|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta}) f_{\mathbf{z}_1,\mathbf{z}_2}(z_1, z_2),$$

where $f_{y,\delta_1,\delta_2,|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$ is the sub-density function of (y, δ_1, δ_2) conditional on $\mathbf{z}_1$ and $\mathbf{z}_2$ for a given $\boldsymbol{\vartheta}$, which contains all the available sample information about $\boldsymbol{\vartheta}$. In addition, let us define $f_{1,0}(\boldsymbol{\vartheta})$, $f_{0,1}(\boldsymbol{\vartheta})$ and $f_{0,0}(\boldsymbol{\vartheta})$ as the shorthand notations for $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 1, 0|z_1, z_2; \boldsymbol{\vartheta})$, $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 0, 1|z_1, z_2; \boldsymbol{\vartheta})$ and $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 0, 0|z_1, z_2; \boldsymbol{\vartheta})$. If $y = T_1$ then we have that

$$f_{1,0}(\boldsymbol{\vartheta}) = \lim_{\varrho \to 0} \varrho^{-1} P(t < T_1 \leq t + \varrho, T_2 > T_1, T_3 > T_1 | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}).$$

If $y = T_2$ then

$$f_{0,1}(\boldsymbol{\vartheta}) = \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_2 \leq t + \varrho, T_1 > T_2, T_3 > T_2 | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}).$$

Finally, if $y = T_3$ then

$$f_{0,0}(\boldsymbol{\vartheta}) = \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_3 \leq t + \varrho, T_1 > T_3, T_2 > T_3 | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}).$$

In addition, we assume that

(B1) (T_1, T_2) and T_3 are independent given $\mathbf{z}_1$ and $\mathbf{z}_2$.

(B2) The censoring by T_3 is not informative for (T_1, T_2) given $\mathbf{z}_1$ and $\mathbf{z}_2$. This implies that the distribution of T_3 does not depend on $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta)$ (e.g., Dettoni et al., 2020).

Tsiatis (1975) showed that the sub-densities can be obtained directly from the joint survival function of the latent survival times. In the following, we provide details on the calculation of $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,1,0|z_1,z_2;\boldsymbol{\vartheta})$, $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,1|z_1,z_2;\boldsymbol{\vartheta})$ and $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,0|z_1,z_2;\boldsymbol{\vartheta})$.

Theorem 4. Assume that $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$ is observed, where $y = \min\{T_1, T_2, T_3\}$, $\delta_1 = I\{y = T_1\}$ and $\delta_2 = I\{y = T_2\}$. If (B1) holds then the sub-densities can be expressed as

$$\begin{aligned}
f_{1,0}(\boldsymbol{\vartheta}) =& \left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)]} \mathcal{G}_1'[\xi_1(t,z_1;\boldsymbol{\gamma}_1)] \frac{\partial \xi_1(t,z_1;\boldsymbol{\gamma}_1)}{\partial t} \right] \\
& \times S_{T_3}(t), \\
f_{0,1}(\boldsymbol{\vartheta}) =& \left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]} \mathcal{G}_2'[\xi_2(t,z_2;\boldsymbol{\gamma}_2)] \frac{\partial \xi_2(t,z_2;\boldsymbol{\gamma}_2)}{\partial t} \right] \\
& \times S_{T_3}(t), \\
f_{0,0}(\boldsymbol{\vartheta}) =& C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]; \theta\right\} \times f_{T_3}(t)
\end{aligned}$$

Identification of the proposed model can be proved using different approaches and assumptions. In particular, if the dependence parameter θ is fixed, then the margins $\mathcal{G}_\nu\left[\xi_\nu(t_\nu, z_\nu; \boldsymbol{\gamma}_\nu)\right]$, and therefore $\boldsymbol{\gamma}_\nu$ in model (5.3) could in principle be shown to be identified in the complete parameter space following, for instance, Zheng & Klein (1995), Chen (2010) and Xu et al. (2018), although we have not pursued this idea and hence it would have to be verified. However, identification for the case where θ is estimated from the data is considerably more involved. For example, Deresa & Van Keilegom (2019), following the results of Nádas (1971) and Basu & Ghosh (1978), proposed a model in which the association parameter is identified. In their approach, a class of monotonic parametric transformations of the logarithm of the survival and censoring times are employed and the margins assumed to follow a multivariate normal distribution. Since it is in general difficult or not feasible to verify whether a nonlinear model is globally identifiable, one can focus on a neighbourhood of $\boldsymbol{\vartheta}$ and hence focus on local identification whose definition is given below.

Definition. Let the complete parameter space be $\mathcal{S}_{\boldsymbol{\vartheta}} = \{(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta) : \boldsymbol{\gamma}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\gamma}_2 \in \mathbb{R}^{p_2}, \theta \in \Theta\}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta) \in \mathcal{S}_{\boldsymbol{\vartheta}}$ be a $p \times 1$ vector of parameters, with $p = p_1 + p_2 + 1$. Suppose that inference about $\boldsymbol{\vartheta}$ is made on the basis of n observations of $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$ with sub-density function $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, \cdot, \cdot|z_1, z_2; \boldsymbol{\vartheta})$. Let $\mathcal{O}_{\boldsymbol{\vartheta}^0}$ denotes an open neighbourhood of $\boldsymbol{\vartheta}_0$. A point $\boldsymbol{\vartheta}^0 \in \mathcal{S}_{\boldsymbol{\vartheta}}$ is said to be locally identified if for $l_1, l_2 = 0, 1$ and for almost every (t, z_1, z_2), the equality $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta}^0) = f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$ implies $\boldsymbol{\vartheta}^0 = \boldsymbol{\vartheta}$, for any $\boldsymbol{\vartheta}$ in $\mathcal{O}_{\boldsymbol{\vartheta}^0}$.

Theorem 5. Assume that, for $l_1, l_2 = 0, 1$, and for almost every (t, z_1, z_2), $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}($

$\cdot, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$ is differentiable at $\boldsymbol{\vartheta}^0$ and the rank of $\dfrac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}$ at $\boldsymbol{\vartheta}^0$ is equal to p, then $\boldsymbol{\vartheta}^0$ is locally identified.

The proof of Theorems 4 and 5 are given in Appendix C.1.

Turning now to our model, since for $l_1, l_2 = 0, 1$ and for almost every (t, z_1, z_2), $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$ is differentiable at $\boldsymbol{\vartheta}^0$, its derivative with respect to $\boldsymbol{\vartheta}$ evaluated at $\boldsymbol{\vartheta}^0$, can be written as

$$\left.\frac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}_0} = \left[\begin{array}{c} \dfrac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \theta} \end{array}\right]_{\boldsymbol{\vartheta}^0}.$$

Let C, $\mathcal{G}_\nu$ and ξ_ν be the shorthand notations of $C\left(\mathcal{G}_1(\xi_1(t, z_1; \boldsymbol{\gamma}_1)), \mathcal{G}_2(\xi_2(t, z_2; \boldsymbol{\gamma}_2)); \theta\right)$, $\mathcal{G}_\nu\left[\xi_\nu(t, z_\nu; \boldsymbol{\gamma}_\nu)\right]$ and $\xi_\nu(t, z_\nu; \boldsymbol{\gamma}_\nu)$ respectively. Then, as shown in Appendix C.2, $\left.\dfrac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0}$, $\left.\dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0}$ and $\left.\dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0}$ can be written as

$$\begin{aligned}
\left.\frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} &= -\Big[\Delta_1, \Delta_1 \boldsymbol{\mathcal{Q}}_{10}^{\triangle}(t)^\mathsf{T} + \Omega_1 \boldsymbol{\mathcal{Q}}_{10}^{\triangle'}(t)^\mathsf{T}, \Delta_1 \boldsymbol{\mathcal{Q}}_{11}(z_{11})^\mathsf{T}, \ldots, \Delta_1 \boldsymbol{\mathcal{Q}}_{1K_1}(z_{1K_1})^\mathsf{T}, \\
&\qquad \Upsilon_1, \Upsilon_1 \boldsymbol{\mathcal{Q}}_{20}^{\triangle}(t)^\mathsf{T}, \Upsilon_1 \boldsymbol{\mathcal{Q}}_{21}(z_{21})^\mathsf{T}, \ldots, \Upsilon_1 \boldsymbol{\mathcal{Q}}_{2K_2}(z_{2K_2})^\mathsf{T}, \Psi_1 \frac{\partial^2 C}{\partial \mathcal{G}_1 \partial \theta}\Big]^\mathsf{T}_{\boldsymbol{\vartheta}^0}, \\
\left.\frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} &= -\Big[\Upsilon_2, \Upsilon_2 \boldsymbol{\mathcal{Q}}_{10}^{\triangle}(t)^\mathsf{T}, \Upsilon_2 \boldsymbol{\mathcal{Q}}_{11}(z_{11})^\mathsf{T}, \ldots, \Upsilon_2 \boldsymbol{\mathcal{Q}}_{1K_1}(z_{1K_1})^\mathsf{T}, \Delta_2, \Delta_2 \boldsymbol{\mathcal{Q}}_{20}^{\triangle}(t)^\mathsf{T} + \\
&\qquad \Omega_2 \boldsymbol{\mathcal{Q}}_{20}^{\triangle'}(t)^\mathsf{T}, \Delta_2 \boldsymbol{\mathcal{Q}}_{21}(z_{21})^\mathsf{T}, \ldots, \Delta_2 \boldsymbol{\mathcal{Q}}_{2K_2}(z_{2K_2})^\mathsf{T}, \Psi_2 \frac{\partial^2 C}{\partial \mathcal{G}_2 \partial \theta}\Big]^\mathsf{T}_{\boldsymbol{\vartheta}^0}, \\
\left.\frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} &= \Big[\Omega_1, \Omega_1 \boldsymbol{\mathcal{Q}}_{10}^{\triangle}(t)^\mathsf{T}, \Omega_1 \boldsymbol{\mathcal{Q}}_{11}(z_{11})^\mathsf{T}, \ldots, \Omega_1 \boldsymbol{\mathcal{Q}}_{1K_1}(z_{1K_1})^\mathsf{T}, \Omega_2, \Omega_2 \boldsymbol{\mathcal{Q}}_{20}^{\triangle}(t)^\mathsf{T}, \\
&\qquad \Omega_2 \boldsymbol{\mathcal{Q}}_{21}(z_{21})^\mathsf{T}, \ldots, \Omega_2 \boldsymbol{\mathcal{Q}}_{2K_2}(z_{2K_2})^\mathsf{T}, \frac{\partial C}{\partial \theta}\Big]^\mathsf{T}_{\boldsymbol{\vartheta}^0},
\end{aligned}$$

where $\Psi_\nu = \dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\dfrac{\partial \xi_\nu}{\partial t}$, $\Delta_\nu = \left[\Psi_\nu \dfrac{\partial^2 C}{\partial \mathcal{G}_\nu^2}\dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} + \dfrac{\partial C}{\partial \mathcal{G}_\nu}\dfrac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2}\dfrac{\partial \xi_\nu}{\partial t}\right]$, $\Omega_\nu = \left[\dfrac{\partial C}{\partial \mathcal{G}_\nu}\dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right]$ and $\Upsilon_\nu = \left[\Psi_\nu \dfrac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega}\dfrac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\right]$, and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle}(t)$, $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle'}(t)$ and $\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu})$ can be defined as

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle}(t) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 03}\right) \\ \vdots \\ \mathcal{Q}_{\nu 0 J_{\nu 0}}(t) \exp\left(\gamma_{\nu 0 J_{\nu 0}}\right) \end{bmatrix}, \quad \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu}) = \begin{bmatrix} \mathcal{Q}_{\nu k_\nu 1}(z_{\nu k_\nu}) \\ \mathcal{Q}_{\nu k_\nu 2}(z_{\nu k_\nu}) \\ \mathcal{Q}_{\nu k_\nu 3}(z_{\nu k_\nu}) \\ \vdots \\ \mathcal{Q}_{\nu k_\nu J_{\nu k_\nu}}(z_{\nu k_\nu}) \end{bmatrix},$$

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle'}(t) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 03}\right) \\ \vdots \\ \mathcal{Q}'_{\nu 0 J_{\nu 0}}(t) \exp\left(\gamma_{\nu 0 J_{\nu 0}}\right) \end{bmatrix}.$$

for $\nu = 1, 2$, $\omega = 1, 2$ and $\nu \neq \omega$.

According to Theorem (5), in order to locally identify model (5.3), we should prove that $\left.\dfrac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} \neq \mathbf{0}$, $\left.\dfrac{f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} \neq \mathbf{0}$ and $\left.\dfrac{f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0} \neq \mathbf{0}$ for almost every (t, z_1, z_2), and that $\operatorname{rank}\left[\left.\dfrac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}^0}\right] = p$, for $l_1, l_2 = 0, 1$. This can be achieved by proving that the following statements are true.

Case 1: $(l_1, l_2) = (1, 0)$ or $(l_1, l_2) = (0, 1)$. For all $\varphi_j \in \mathbb{R}$ and $j = 0, 1, 2, \ldots, \kappa$,

$$\text{if } \Bigg\{ \Delta_\nu \varphi_0 + \left[\Delta_\nu \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y) + \Omega_\nu \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y) \right] \varphi_1$$
$$+ \left[\Delta_\nu \sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y) + \Omega_\nu \sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y) \right]$$
$$\times \exp(\gamma_{\nu 02})\varphi_2 + \cdots + \left[\Delta_\nu \mathcal{Q}_{\nu 0 J_{\nu 0}}(y) + \Omega_\nu \mathcal{Q}'_{\nu 0 J_{\nu 0}}(y) \right] \exp(\gamma_{\nu 0 J_{\nu 0}})\varphi_{\iota_1} + \left[\Delta_\nu \mathcal{Q}_{\nu 11}(z_{\nu 1}) \right] \varphi_{\iota_1 + 1}$$
$$+ \cdots + \left[\Delta_\nu \mathcal{Q}_{\nu K_\nu J_{\nu K_\nu}}(z_{\nu K_\nu}) \right] \varphi_{\iota_2} + \Upsilon_\nu \varphi_{\iota_2 + 1} + \left[\Upsilon_\nu \sum_{j_{\omega 0}=1}^{J_{\omega 0}} \mathcal{Q}_{\omega 0 j_{\omega 0}}(y) \right] \varphi_{\iota_2 + 2} + \left[\Upsilon_\nu \sum_{j_{\omega 0}=2}^{J_{\omega 0}} \mathcal{Q}_{\omega 0 j_{\omega 0}}(y) \right]$$
$$\times \exp(\gamma_{\omega 02})\varphi_{\iota_2 + 3} + \cdots + \left[\Upsilon_\nu \mathcal{Q}_{\omega 0 J_{\omega 0}}(y) \right] \exp(\gamma_{\omega 0 J_{\omega 0}})\varphi_{\iota_3} + \left[\Upsilon_\nu \mathcal{Q}_{\omega 11}(z_{\omega 1}) \right] \varphi_{\iota_3 + 1} + \cdots +$$
$$\left[\Upsilon_\nu \mathcal{Q}_{\omega K_\omega J_{\omega K_\omega}}(z_{\omega K_\omega}) \right] \varphi_{\kappa - 1} + \left[\Psi_\nu \frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \theta} \right] \varphi_\kappa = 0 \Bigg\}, \quad \text{then} \quad \varphi_0 = 0,\ \varphi_1 = 0,\ \varphi_2 = 0, \ldots, \varphi_\kappa = 0.$$

Case 2: $(l_1, l_2) = (0, 0)$.

$$\text{if } \Bigg\{ \Omega_1 \varphi_0 + \left[\Omega_1 \sum_{j_{10}=1}^{J_{10}} \mathcal{Q}_{10 j_{10}}(t) \right] \varphi_1 + \left[\Omega_1 \sum_{j_{10}=2}^{J_{10}} \mathcal{Q}_{10 j_{10}}(t) \right] \times \exp(\gamma_{102})\varphi_2 + \cdots + \left[\Omega_1 \mathcal{Q}_{10 J_{10}}(t) \right]$$
$$\times \exp(\gamma_{10 J_{10}})\varphi_{\iota_1} + \left[\Omega_1 \mathcal{Q}_{111}(z_{11}) \right] \varphi_{\iota_1 + 1} + \cdots + \left[\Omega_1 \mathcal{Q}_{1 K_1 J_{1 K_1}}(z_{1 K_1}) \right] \varphi_{\iota_2} + \Omega_2 \varphi_{\iota_2 + 1} +$$
$$\left[\Omega_2 \sum_{j_{20}=1}^{J_{20}} \mathcal{Q}_{20 j_{20}}(t) \right] \times \varphi_{\iota_2 + 2} + \left[\Omega_2 \sum_{j_{20}=2}^{J_{20}} \mathcal{Q}_{20 j_{20}}(t) \right] \exp(\gamma_{202})\varphi_{\iota_2 + 3} + \cdots + \left[\Omega_2 \mathcal{Q}_{20 J_{20}}(t) \right]$$
$$\times \exp(\gamma_{20 J_{20}})\varphi_{\iota_3} + \left[\Omega_2 \mathcal{Q}_{211}(z_{21}) \right] \varphi_{\iota_3 + 1} + \cdots + \left[\Omega_2 \mathcal{Q}_{2 K_2 J_{2 K_2}}(z_{2 K_2}) \right] \varphi_{\kappa - 1} + \left[\frac{\partial C}{\partial \theta} \right] \varphi_\kappa = 0 \Bigg\},$$

then $\varphi_0 = 0,\ \varphi_1 = 0,\ \varphi_2 = 0, \ldots, \varphi_\kappa = 0.$

This is a preliminary result and future research will focus on establishing conditions that are not too restrictive for the links and copula functions implemented in this work in order to prove these statements. If this does not prove successful then a possible approach would be to work with a simpler version of the proposed class of models that would allow to define realistic conditions more easily.

Remark 3. Local identification at one point in the parameter space does not ensure that the model is locally identified everywhere in $\mathcal{S}_{\boldsymbol{\vartheta}}$. Also, local identifiability everywhere in $\mathcal{S}_{\boldsymbol{\vartheta}}$ is a necessary but not a sufficient condition for global identification. It is worth noting that local identification still allows for consistent estimation of

$\boldsymbol{\vartheta}$ and it is sufficient to derive the asymptotic properties of the estimator $\hat{\boldsymbol{\vartheta}}$; if the sample size is sufficiently large then it is possible to limit the parameter space to a neighbourhood of $\boldsymbol{\vartheta}^0$ and rely on local identification (e.g., Hsiao, 1989).

5.4 Penalized estimation approach for the dependent censoring model

Assuming that the model is identified, model fitting is undertaken via maximum likelihood estimation. As pointed out in Section 5.3, since $f_{\mathbf{z}_1,\mathbf{z}_2}(z_1, z_2)$ does not involve the model's parameters, the likelihood function can be formulated using the sub-density function $f_{y,\delta_1,\delta_2,|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$. Let us assume that the observed data consist of n i.i.d. replications $\{(y_i, \delta_{1i}, \delta_{2i}, \mathbf{z}_{1i}, \mathbf{z}_{2i})\}_{i=1}^n$ of $(y, \delta_1, \delta_2, \mathbf{z}_1, \mathbf{z}_2)$. This allows us to write the likelihood function for $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta)$ as

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(y_i, \delta_{1i}, \delta_{2i}|\mathbf{z}_{1i}, \mathbf{z}_{2i}; \boldsymbol{\vartheta}). \tag{5.5}$$

Using the results in Theorem 4 and since (B2) implies that $f_{T_3}(t)$ and $S_{T_3}(t)$ can be discarded from the likelihood, we can write

$$\begin{aligned}
\ell(\boldsymbol{\vartheta}) = &\sum_{i=1}^{n} \delta_{1i} \log \left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_1[\xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)]} \right. \\
&\left. \times \mathcal{G}_1'[\xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)] \frac{\partial \xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)}{\partial y_i} \right] \\
&+ \sum_{i=1}^{n} \delta_{2i} \log \left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_2[\xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)]} \right. \\
&\left. \times \mathcal{G}_2'[\xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)] \frac{\partial \xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)}{\partial y_i} \right] \\
&+ \sum_{i=1}^{n} \delta_{3i} \log \left[C\left\{\mathcal{G}_1[\xi_1(y_i, \mathbf{z}_{1i}; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(y_i, \mathbf{z}_{2i}; \boldsymbol{\gamma}_2)]; \theta\right\}\right].
\end{aligned} \tag{5.6}$$

The proposed model allows for a high degree of flexibility in modelling data. Therefore, in order to prevent over-fitting, we maximize

$$\ell_p(\boldsymbol{\vartheta}) = \ell(\boldsymbol{\vartheta}) - \frac{1}{2}\boldsymbol{\vartheta}^\mathsf{T}\Lambda\boldsymbol{\vartheta}, \tag{5.7}$$

where ℓ_p is the penalized log-likelihood, $\Lambda = \text{diag}(\boldsymbol{\mathcal{D}}_1, \boldsymbol{\mathcal{D}}_2, 1)$, and $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$ are overall penalties which contain $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ defined as $\boldsymbol{\lambda}_\nu = (\lambda_{\nu 1}, \ldots, \lambda_{\nu K_\nu})^\mathsf{T}$ for $\nu = 1, 2$. The smoothing parameter vectors can be collected in the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\mathsf{T}, \boldsymbol{\lambda}_2^\mathsf{T})^\mathsf{T}$.

Given the flexibility and complexities of the proposed model, estimation of all model's parameters is carried out using the algorithm introduced in section 4.3.2, by simply replacing $\boldsymbol{\varphi}$ by $\boldsymbol{\vartheta}$. Since this optimization scheme is based on first and second order analytical derivatives, these have been tediously derived and reported in Appendix C.2.

5.5 Theoretical properties of $\hat{\boldsymbol{\vartheta}}$

In this section, we derive the $\sqrt{n}$ consistency and asymptotic normality of the dependent censoring estimator, assuming that the model is identified. As in Section 4.3.3, we use the fixed-knot asymptotic framework since it is closer to practical statistical modelling. The construction of confidence intervals and p-values is undertaken using the approach introduced in section 4.3.4.

Theorem 6. Under assumptions (C1)-(C8) in Appendix C.3, the parameter estimator $\hat{\boldsymbol{\vartheta}} = \underset{\boldsymbol{\vartheta} \in \mathcal{S}_{\boldsymbol{\vartheta}}}{\text{argmax}}\ \ell_p(\boldsymbol{\vartheta})$ exists, is $\sqrt{n}$-consistent and

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\vartheta}^0)]^{-1}\right\},$$

where $\mathcal{I}(\boldsymbol{\vartheta}^0) = \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell(\mathbf{w}; \boldsymbol{\vartheta}^0)]$ with $\mathbf{w}$ containing the response and covariate

vectors.

The proof of this result is given in Appendix C.3.

To construct the confidence intervals and p-values, we have used the approach already introduced in Section 4.3.4, whose results straightforwardly extend to the current context, by substituting $\boldsymbol{\varphi}$ by $\boldsymbol{\vartheta}$.

In the next chapter, we will investigate the finite sample properties of the dependent estimator through a Monte Carlo simulation study. The proposed estimator is also illustrated using prostate cancer data.

5.6 Simulation study

This section provides evidence on the empirical effectiveness of the proposed methodology in recovering true linear effects, non linear effects, association parameters and baseline functions under dependent censoring for the Data Generating Processes (DGPs) detailed in Table 5.2. The performance of the dependent censoring penalized maximum log-likelihood estimator (DCPMLE) is compared to that of its independent counterpart (ICPMLE; see Appendix C.4 for details on the independent estimator).

For all the DGPs considered in the study, event times, T_{1i}, were generated from a proportional hazard model, while censored times, T_{2i}, were generated from a proportional odds model. These, defined on the survival function scale, are given by

$$\begin{aligned} &\log\left[-\log\left\{S_{10}(t_{1i})\right\}\right] + \gamma_{01} + \gamma_{11} z_{1i} + s_{11}(z_{2i}), \\ &\log\left[\frac{\left\{1 - S_{20}(t_{2i})\right\}}{S_{20}(t_{2i})}\right] + \gamma_{02} + \gamma_{12} z_{1i} + s_{12}(z_{2i}), \end{aligned} \tag{5.8}$$

where $S_{10}(t_{1i}) = 0.80 \exp\left(-0.4 t_{1i}^{2.5}\right) + 0.20 \exp\left(-0.1 t_{1i}^{1.0}\right)$ and $S_{20}(t_{2i}) =$

$0.99 \exp(-0.05 t_{2i}^{2.3}) + 0.01 \exp(-0.4 t_{2i}^{1.1})$ (e.g., Crowther & Lambert, 2013). Covariate z_{1i} was generated using a binomial distribution and z_{2i} using a uniform distribution. The administrative censoring variable, T_3, was generated from a uniform distribution on $[3, 8]$, independent of (T_1, T_2). The models were fitted considering two levels of association. Details on the DGPs (in terms of copula structure, parametric coefficients, smooth functions and association parameters) and the proportion of observations for T_1, T_2 and T_3 are provided in Table 5.2.

Sample sizes were set to 500 and 2000, and the number of replicates to 1000. The models were fitted using `gjrm()` in `GJRM` by employing the proportional hazard link (`"PH"`) for the event times and the proportional odd link (`"PO"`) for the censoring times (see Appendix C.5 for some software details). The smooth components of z_1 and z_2 were represented using penalized low rank thin plate splines with second order penalty and 10 bases. Here, it is possible to employ different spline definitions and related penalties (e.g., cubic regression splines and P-splines). As explained, e.g., in Wood (2017), for uni-dimensional smooths of continuous covariates, the specific choice of spline definition will not have an impact on the estimated curves as long as smoothing parameter estimation is reliably achieved. As for the number of basis functions, the chosen value of 10 is arbitrary and based on the fact that it generally offers enough modelling flexibility in applications. However, a sensitivity analysis using more bases was attempted and there was no tangible change in the results apart from the computing time which increased. The smooths of times were represented using monotonic penalized B-splines with penalty defined in Section 4.2.2 and 10 bases. For each replicate, curve estimates were constructed using 200 equally spaced fixed values in the $(0, 7)$ range for the baseline functions and $(0, 1)$ otherwise.

(i) **Parametric effects**: Regarding the estimates for the parametric effects, Figure 5.1, Figure C.14 (Appendix C.6), and Table 5.3 show that overall the

mean estimates for the DCPMLE are very close to the respective true values and improve as the sample size increases, and that the variability of the estimates decreases as the sample size grows large. On the contrary, when the ICPMLE is considered, there is a not negligible bias and this does not disappear as the sample size grows large. This bias is even higher when τ is equal to 0.7 (`DGPs 1`, `2` and `3`). Furthermore, the RMSE of the ICPMLE is considerably higher than the RMSE of the ICPMLE for all DGPs and sample sizes examined in the simulation study as shown in Table 5.3.

(ii) **Smooth effects**: As for the smooth effect of the non-linear covariates, Figure 5.2 (third column), Figures C.15 to C.19 (third column) in Appendix C.6, and Table 5.3 show that overall the true functions are recovered well by the proposed estimation methods and that the results improve in terms of bias and efficiency as the sample size increases. Furthermore, the RMSE of the ICPMLE is higher than the RMSE of the ICPMLE for all DGPs and sample sizes examined in the simulation study as shown in Table 5.3.

(iii) **Survival and hazard functions**: Figure 5.2 (first and third rows), Figures C.15 to C.19 (first and third rows) and Table C.1 in Appendix C.6 show that overall the true survival (first column) and hazard functions (second column) for the DCPMLE are recovered well, and the results improve in terms of bias and efficiency as the sample size increases. However, when the ICPMLE is considered, Figure 5.2 (second and fourth rows), Figures C.15 to C.19 (second and fourth rows) and Table C.1 in Appendix C.6, there is a not negligible bias for the survival and hazard functions and the situation does not improve as the sample size grows large. In addition, the DCPMLE is more efficient than the ICPMLE for almost all the DGPs and samples sizes examined in the simulation study.

(iv) **Kendall's** τ: Regarding the estimates for the Kendall's τ, Figure 5.3, and Figure C.20 and Table C.2 in Appendix C.6 show that overall the mean estimates for τ in the DCPMLE are very close to the respective true values and improve as the sample size increases, and that the variability of the estimates decreases as the sample size grows large. The results are even better in terms of bias and efficiency when τ is equal to 0.7 since it is easier to detect the dependence between the margins.

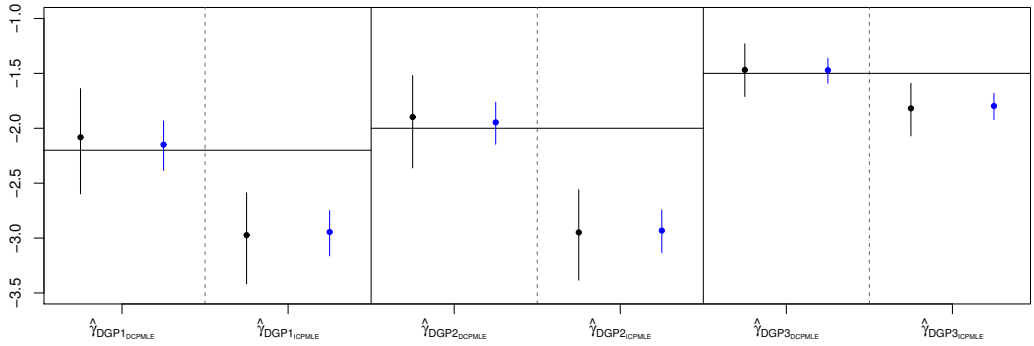

Figure 5.1: Parametric effects (γ_{11}) when DCPMLE ($\tau = 0.7$) and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP1` (Clayton copula), `DGP2` (Frank copula) and `DGP3` (Gaussian copula) defined in Table 5.2. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 2000$ are given in blue

Computing times for the proposed approach were on average 10 seconds for n = 2000 and around 5 seconds for n = 500. Finally, we would like to point out that in this thesis we have reported the simulation results for a sub-group of copulae; in our preliminary experiments we considered all the copulae listed in Table 5.1 and the results were in line with those discussed in this section. These results can be found in Appendix C.6 (DGPs 7 to 14). In particular, DGP13 and DGP14 consider lower censoring rates of approximately 30% and 29% respectively.

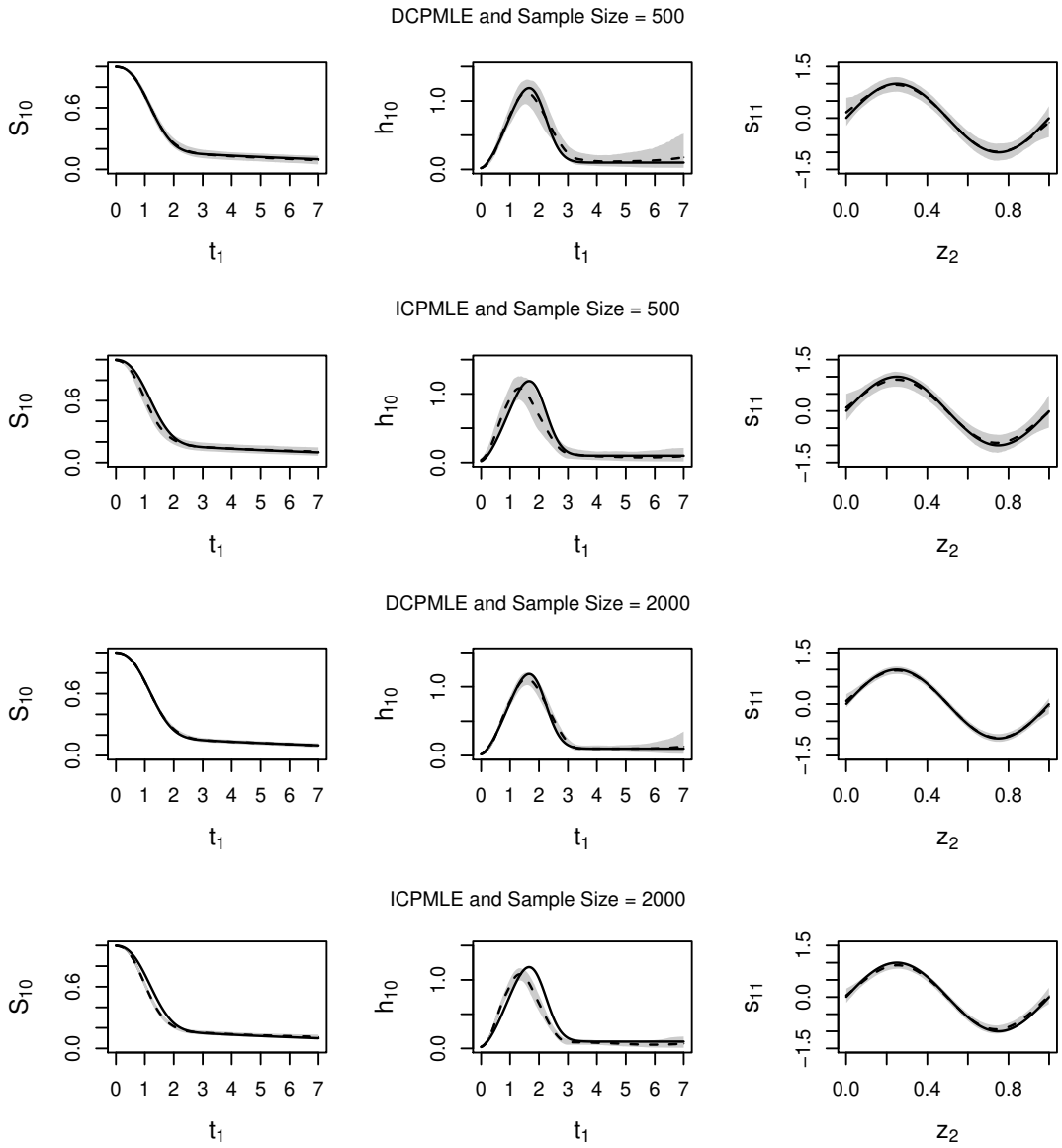

Figure 5.2: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP1` (Table 5.2). The results in the first and second rows refer to $n = 500$, whereas that in the third and fourth rows to $n = 2000$. True functions are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting from 5% and 95% quantiles by shaded areas.

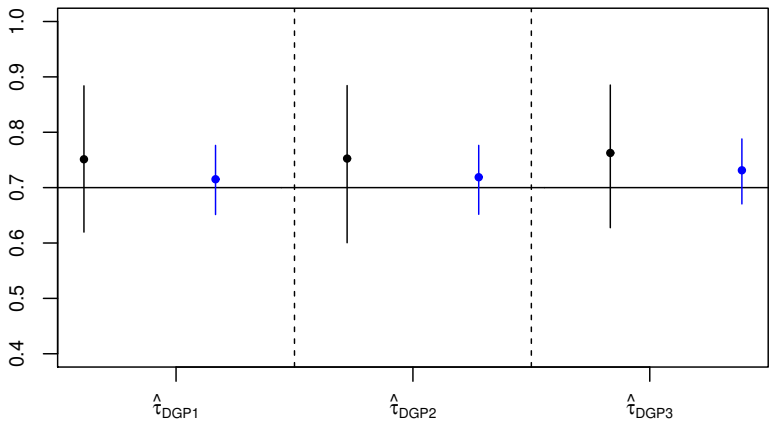

Figure 5.3: Kendall Tau coefficient ($\tau = 0.7$) estimates obtained when DCPMLE is fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP1` (Clayton copula), `DGP2` (Frank copula) and `DGP3` (Gaussian copula) defined in Table 5.2. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 2000$ are given in blue.

Data Generating Processes (DGPs)											
DGP	Copula	$s_{11}(z_{2i})$	$s_{12}(z_{2i})$	τ	γ_{01}	γ_{02}	γ_{11}	γ_{12}	T_1	T_2	T_3
1	Clayton	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-2.20	1.80	42%	43%	15%
2	Frank	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-2.00	1.80	43%	42%	15%
3	Gaussian	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-1.50	1.20	55%	29%	16%
4	Clayton	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.40	-0.01	0.02	-2.20	1.80	50%	37%	13%
5	Frank	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.40	-0.01	0.02	-2.00	1.80	48%	40%	12%
6	Gaussian	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.40	-0.01	0.02	-1.50	1.20	55%	31%	14%
7	FGM	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-1.40	1.20	56%	31%	13%
8	AMH	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-1.40	1.20	54%	33%	13%
9	Gumbel	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-2.20	1.80	43%	44%	13%
10	Joe	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-1.50	1.20	53%	30%	17%
11	Plackett	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-2.20	1.80	42%	45%	13%
12	Student-t	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-1.50	1.20	54%	30%	16%
13	Gaussian	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.70	-0.01	0.02	-0.80	0.60	70%	17%	13%
14	Gaussian	$\sin(2\pi z_i)$	$-0.2\exp(3.2z_i)$	0.40	-0.01	0.02	-0.80	0.60	71%	17%	12%

Table 5.2: Summary of the Data Generating Processes (DGPs) used to simulate dependent censoring data. The proportion of observations, in average, for T_1, T_2 and T_3 is also shown. The models were fitted using the gjrm() function in GJRM by employing the proportional hazard link ("PH") for the event times and the proportional odd link ("PO") for the censoring times.

	DCPML: Parametric Effects (γ_{11})				ICPMLE: Parametric Effects (γ_{11})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
1	0.118	0.056	0.195	0.174	-0.774	-0.732	0.815	0.751
2	0.103	0.054	0.280	0.131	-0.949	-0.932	0.982	0.940
3	0.031	0.029	0.147	0.075	-0.319	-0.297	0.350	0.306
4	0.066	0.033	0.252	0.124	-0.310	-0.294	0.367	0.312
5	0.078	0.047	0.269	0.133	-0.418	-0.403	0.463	0.415
6	0.047	0.045	0.168	0.093	-0.189	-0.173	0.236	0.187
	DCPML: Smooth Effects (s_{11})				ICPMLE: Smooth Effects (s_{11})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
1	0.029	0.017	0.149	0.076	0.052	0.043	0.158	0.090
2	0.032	0.017	0.147	0.075	0.053	0.051	0.159	0.094
3	0.031	0.018	0.132	0.069	0.026	0.032	0.139	0.080
4	0.034	0.017	0.147	0.075	0.048	0.035	0.154	0.084
5	0.033	0.020	0.147	0.077	0.043	0.036	0.151	0.085
6	0.034	0.020	0.137	0.071	0.026	0.021	0.140	0.074

Table 5.3: Bias and root mean squared error (RMSE) for parametric and smooth effects when DCPMLE and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to the `DGPs 1` to `6` defined in Table 5.2. Bias and RMSE for the smooth terms are calculated, respectively, as $n_s^{-1} \sum_{i=1}^{n_s} |\bar{\hat{s}}_i - s_i|$ and $n_s^{-1} \sum_{i=1}^{n_s} \sqrt{n_{rep}^{-1} \sum_{rep=1}^{n_{rep}} (\hat{s}_{rep,i} - s_i)^2}$, where $\bar{\hat{s}}_i = n_{rep}^{-1} \sum_{rep=1}^{n_{rep}} \hat{s}_{rep,i}$, n_s is the number of equally spaced fixed values in the $(0, 8)$ or $(0, 1)$ range, and n_{rep} is the number of simulation replicates. In this case, $n_s = 200$ and $n_{rep} = 1000$. The bias for the smooth terms is based on absolute differences in order to avoid compensating effects when taking the sum.

5.7 Empirical illustration

The modelling framework is illustrated using data obtained from a randomized clinical trial conducted to compare different levels of an active treatment for prostate cancer (Byar & Green, 1980). These data have been analysed extensively (e.g., Escarela & Carriere, 2003; Kleinbaum & Klein, 2010; Deresa & Van Keilegom, 2019). In total, 506 individuals with prostate cancer were randomized to receive either a placebo or one of three dose levels of diethylstilbestrol (DES). The primary event of interest, y_i, is the time at which the patient died of prostate cancer. Because of potentially fatal side effects of DES (e.g., cardiovascular-related or other types of diseases), the analysis of the treatment must consider not only the death time from prostate cancer but also that from other diseases.

Following Kleinbaum & Klein (2010), the clinically meaningful covariates were: binary treatment (`rx` = 0 if the subject received a placebo or 0.2 mg of DES, and 1 if received 1.0 or 5.0 mg of DES); performance status (`pf` = 0 if normal, 1 if there was limitation of activity); history of cardiovascular disease (`hx` = 0 for no and 1 for yes); standardized weight (`wt`); hemoglobin in μg/100 ml (`hg`); age of the patient at diagnosis (`age`); size of primary lesion estimated in cm^2 from rectal examination (`sz`); combined index of tumour stage and histological grade (`sg`). After dropping all the patients with missing information, the dataset consists of 483 observations. In this sample, 125 patients died of prostate cancer during the study period, 219 died of non-prostate cancer diseases (cardiovascular-related or other diseases), while 139 were alive at the end of the study. Therefore, some subjects were censored due to a competing risk (non-cancer death) and others because they would be alive at the end of the study (administrative censoring). Prostate cancer death and non-cancer death are assumed to be dependent, and administrative censoring as being independent of everything else. Recall that the aim is to assess the effect of DES on prostate cancer death while accounting for

individual characteristics and dependence censoring.

Deresa & Van Keilegom (2019) analysed the same data using a fully parametric regression approach based on the bivariate Gaussian distribution and parametric monotone increasing transformations of the logarithm of the times variables. To compare the magnitudes of the correlation parameters obtained employing the technique of these authors and the method proposed in this paper, we fitted a Gaussian copula model where the covariate effects were modelled parametrically and the baselines using the monotonic spline approach detailed in this paper with the same settings employed for the simulation study. Smoothing for the baselines was implemented on the log-time scale which usually yields very smooth fitted functions and hence it helps, for example, to reduce the chance of potential artifacts in the estimated hazard functions (e.g., Royston & Parmar, 2002). Since several combinations of link functions had to be considered, a number of models were tried out and the final model selected using the AIC and BIC. Table C.6 in Appendix C.7 shows the results for the fitted models and supports the presence of dependent censoring. The chosen model (N°1) exhibits an estimated correlation of `0.47` which is virtually identical to the correlation of `0.46` obtained by Deresa & Van Keilegom (2019), and they are both statistically significant. Moreover, the parameters for `rx`, `hg`, `sz` and `sg` are statistically significant in the two models.

Next, we model the covariate effects flexibly. We followed a similar process as before (see Table C.7 in Appendix C.7). Table 5.4 and Figure 5.4 show the results for the selected dependent censoring model (Model 7). For comparison purposes we also fitted the model under the assumption of independence. Table 5.5 and Figure 5.5 show these results (Model 9).

Main findings: From Tables 5.4 and 5.5, the results show a considerably smaller estimation uncertainty for the dependent model (for example, the standard error of `rx` for the independent censoring model is approximately 2.3 times higher

Model 7 (DCPMLE)				
Parametric Effects	Estimate	Standard Error	Z-value	P-value
intercept	-8.352	0.614	-13.60	0.000 ***
rx	-0.224	0.085	-2.653	0.008 **
hx	0.394	0.118	3.329	0.001 ***
pf	0.399	0.161	2.484	0.013 *
sz	0.278	0.065	4.248	0.000 ***
sg	0.217	0.066	3.298	0.001 ***
Smooth Effects	EDF	Ref.DF	Chi-square	P-value
s(log(u))	1.000	1.000	217.1	0.000 ***
s(hg)	7.790	8.447	36.24	0.000 ***
s(age)	5.173	6.231	19.27	0.005 **
Kendall tau	Estimate	Confidence Interval		
τ	0.841	(0.7,0.923)		

Table 5.4: Estimation results of the dependent censoring model (Model 7 in Table C.7, Appendix C.7) applied to prostate cancer data. The models were fitted using the functions `gamlss()` and `gjrm()` in `GJRM` by employing the "`PH-PO`" link functions combination. Furthermore, EDF and Ref.DF refer to the effective degrees of freedom and reference degrees of freedom of the smooths. More details can be found in Sections 4.3.2 and 4.3.4.

Model 9 (ICPMLE)				
Parametric Effects	Estimate	Standard Error	Z-value	P-value
intercept	-9.133	0.618	-14.79	0.000 ***
rx	-0.686	0.193	-3.557	0.000 ***
hx	0.063	0.204	0.311	0.756
pf	0.422	0.267	1.581	0.114
sz	0.509	0.082	6.188	0.000 ***
sg	0.617	0.099	6.227	0.000 ***
Smooth Effects	EDF	Ref.DF	Chi-square	P-value
s(log(u))	1.000	1.000	177.3	0.000 ***
s(hg)	3.506	4.411	17.95	0.002 **
s(age)	4.579	5.628	13.09	0.030 *

Table 5.5: Estimation results of the independent censoring model (Model 9 in Table C.7, Appendix C.7) applied to prostate cancer data. The models were fitted using the functions `gamlss()` and `gjrm()` in `GJRM` by employing the "`PH-PO`" link functions combination. Furthermore, EDF and Ref.DF refer to the effective degrees of freedom and reference degrees of freedom of the smooths. More details can be found in Sections 4.3.2 and 4.3.4.

than that of its dependent counterpart). Analysing the tables in more detail, the coefficient of `rx` is statistically significant for both models. For instance, the expected hazard for the treatment group is approximately 0.8 times the hazard for the placebo group. Furthermore, the parameters `rx`, `hx`, `pf`, `sz` and `sg` are statistically significant for the dependent censoring model. However, only the parameters `rx`, `sz` and `sg` are statistically significant for the independent censoring model. Tables 5.4 and 5.5 also show that `s(u)`, `s(hg)` and `s(age)` are statistically significant for both models, whereas Figures 5.4 and 5.5 display their estimated functional forms along with the survival and hazard curves. The plots show, for instance, that, after a certain point, the hazard of dying from prostate cancer for the dependent censoring model decreases when the levels of hemoglobin are higher.

The estimate for the association parameter is $\hat{\tau} = 0.841$ and is statistically significant. This is a strong association that will induce bias in the parameter estimates if ignored. In fact, Figures 5.4 and 5.5 show that the survival and hazard curves of the dependent censoring model have a substantially better fit than those of its independent censoring counterpart.

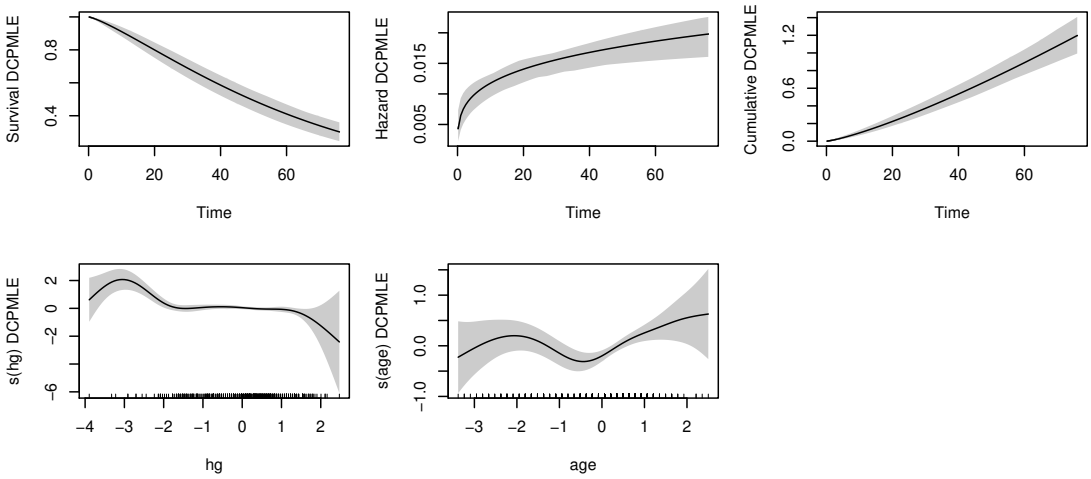

Figure 5.4: Smooth function estimates and their corresponding 95% intervals for the dependent censoring model (Model 7 in in Table C.7, Appendix C.7) obtained by applying `gjrm()` in `GJRM` to prostate cancer data. The intervals have been obtained using the approach described in Section 4.3.4

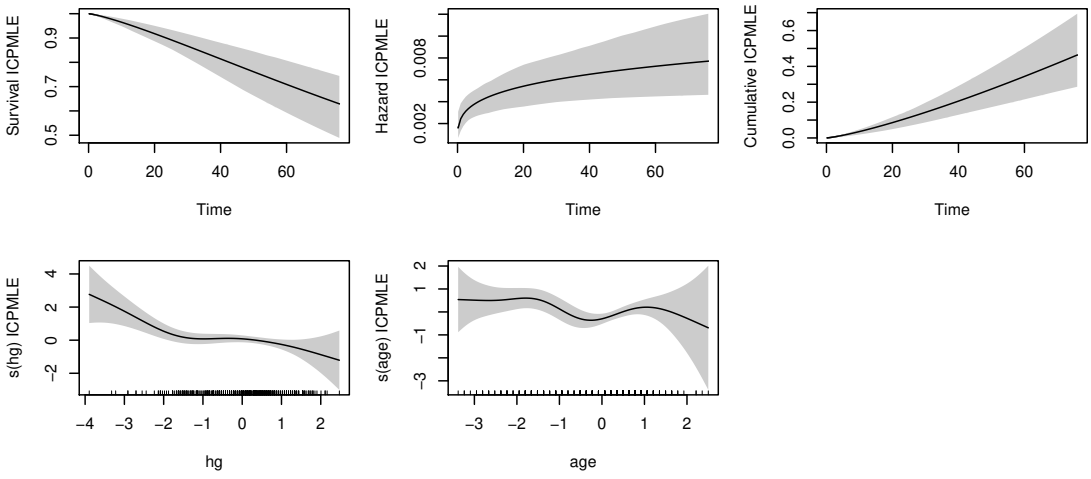

Figure 5.5: Smooth function estimates and their corresponding 95% intervals for the independent censoring model (Model 9 in Table C.7, Appendix C.7) obtained by applying `gjrm()` in `GJRM` to prostate cancer data. The intervals have been obtained using the approach described in Section 4.3.4

5.8 Concluding remarks

In this chapter, we have introduced copula link-based additive models for dependent censoring and their potential, illustrated using simulated and real data. Our extended simulation study suggests that the model is identified. The performance of the dependent censoring penalized maximum log-likelihood estimator (DCPMLE) was also compared to that of its independent counterpart which leads to substantial bias in the estimates since it neglects the dependent censoring mechanism. Further, we have also discussed the $\sqrt{n}$-consistency and asymptotic normality of the dependent censoring estimator, under the assumption that the model is identified.

Important features of the modelling framework are that: the strength of the association between the event and censoring times is modelled using a copula structure and estimated in a way that is flexible and tractable at the same time; the baseline functions are estimated non-parametrically via means of monotonic P-splines, which allows one to obtain coherent estimated survival functions; covariate effects are flexibly determined using additive predictors; the optimization scheme allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection; the models can be easily utilised using the freely available `GJRM R` package which allows for several modelling choices.

Chapter 6

Final Remarks

This thesis has been mainly motivated by the idea of relaxing assumptions on the censoring mechanism in survival analysis. In particular, we aimed to broaden the current applications by introducing two modelling frameworks that extend the class of Survival Generalized Additive Models by allowing for informative and dependent censoring. This work had two objectives: (i) to develop the theory needed for estimating flexible and tractable informative and dependent censoring models; and (ii) to implement the developed modelling frameworks in the `R` package `GJRM`, hence allowing for transparent and reproducible research.

In Chapter 2, we made a review of the essential concepts in survival analysis, where the most important representations of the response variable: the survival function, the hazard function and the cumulative hazard function, were analysed. Then, the crucial problem of censoring and their causes was discussed, along with a general summary of univariate survival models, where the focus was on splines-based methods. In the last part of this chapter, the independent and non-informative censoring assumptions were discussed.

In Chapter 3, we presented a summary of models that allow for different assumptions about the nature of the covariate effects on the survival time, and where the baseline hazard and survival functions can be modelled in a flexible

form. The general ideas of these models are employed to build survival link-based additive models, which were discussed in the last part of the chapter.

Chapter 4, introduced a flexible survival modelling approach to account for the information provided by the censoring times where the survival functions for the censoring and event times were determined using link-based functions models. Baseline functions were modelled non-parametrically by monotonic P-splines, and covariate effects were flexibly determined using additive predictors. In this chapter, a penalized likelihood method to estimate the informative model was proposed, where model fitting was based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection. Then, the $\sqrt{n}$ consistency and asymptotic normality of the non-informative and informative estimators were derived, and shed light on the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart. The construction of confidence intervals and p-values were also discussed, and the performance of the proposed methodology was evaluated using a Monte Carlo simulation study and an empirical application on data about infants hospitalised for pneumonia. Both, the simulation study and the empirical application highlighted the merits of the proposal. In addition, we explained how to fit the model proposed by using the function `gamlss()` in the `R` package `GJRM`.

In Chapter 5, we proposed a flexible regression survival model that accounts for administrative and dependent right censoring, and provided some identification arguments. The strength of the association between the event and censoring times is modelled via a copula structure whose dependence parameter is estimated from the data. As before (Section 4.2.2), baseline functions are non-parametrically estimated using monotonic P-splines and covariate effects are flexibly determined using additive predictors. Parameter estimation as well as the consistency and

asymptotic normality of the estimator were also presented. Finally, the finite sample properties of the dependent estimator were investigated via a Monte Carlo simulation study, and the proposal illustrated using prostate cancer data. These results highlighted the effectiveness of the methodology proposed, and the relevant numerical computation were easily carried out using the function `gjrm()` in the `R` package `GJRM` (Marra & Radice, 2020b).

Although in this thesis we have only modelled right censored responses, it is plausible to consider outcome types other than right censored. Therefore, an interesting extension will be the incorporation of left and interval censored responses in the models introduced in this thesis. These extensions will considerably increase the scope and applicability of the modelling approach proposed. Future research will also focus on extending the proposed informative model to include time varying covariates, and on the construction of efficient schemes for selecting automatically the set of informative covariates. All of these extensions will require the calculation of the log-likelihood functions of the informative and dependent censoring models and their respective score and Hessian components.

Finally, in order to locally identify the dependent censoring model introduced in Chapter 5, we will work on establishing not too restrictive identifying conditions for the links and copula functions implemented in this thesis.

Appendix A

Supplements to Chapter 2

A.1 Discrete T

Survival data can also be measured as an interval. This could indicate, for example, that a transition occurred in a particular period of time, but the exact time within the period is not given (Kalbfleisch & Prentice, 2002). In particular, let T be a discrete random variable taking values $t_1 < t_2 < \dots$. Then, the probability mass function can be defined, for all $i = 1, 2, ...$, as $f_T(t_j) = P(T = t_j)$. This allows to express the survival function as $S_T(t) = \sum_{j|t_j>t} f_T(t_j)$. In addition, for all $j = 1, 2, ...$, the hazard function can be written as

$$h_T(t_j) = P(T = t_j | T \geq t_j) = \frac{f_T(t_j)}{S_T(t_j^-)}, \tag{A.1}$$

where $S_T(a^-) = \lim_{t \to a^-} S_T(t_j)$, since formally $S_T(t)$ equals $P(T > t)$ rather than $P(T \geq t)$. Equation (A.1) can be interpreted as the probability of T at t_j, given survival to time t_j. Furthermore, the cumulative hazard function can be defined as $H_T(t) = \sum_{j|t_j \leq t} h_T(t_j)$.

On the other hand, the survival function can also be obtained from the hazard

function through $S_T(t) = P(T > t) = \prod_{j|t_j \leq t} [1 - \hbar_T(t_j)]$. This allows to represent $f_T(t_j)$ as $f_T(t_j) = \hbar_T(t_j) S_T(t_j^-)$, where $S_T(t_j^-) = \prod_{j=1}^{i-1} [1 - \hbar_T(t_j)]$.

A.2 Discrete and continuous T

So far, we have introduced the crucial functions to represent the response variable, T, assuming that this can be either continuous or discrete. When the distribution of T have both discrete and continuous components, the survival function can be represented using the product-integral function (Gill & Johansen, 1990). This function, which is useful, for example, to deal with mixed distributions, is discussed briefly in this section.

In particular, the general idea is that the hazard function can be built to include the continuous and discrete hazard functions $\hbar_T(t)$ and $\hbar_T(t_j)$, respectively. This allows to write the cumulative hazard function as

$$\mathcal{H}_T^{DC}(t) = \int_0^t \hbar_T(u)du + \sum_{j|t_j \leq t} \hbar_T(t_j). \tag{A.2}$$

$\mathcal{H}_T^{DC}(t)$ is a right-continuous non-decreasing function, but it is not necessarily differentiable, since by definition, the response variable is either continuous or discrete. This also implies that the cumulative distribution function, $F_T(t)$, is not necessarily a differentiable function, and therefore $f_T(t) = \dfrac{dF_T(t)}{dt}$ can not be valid.

Finally, the survival function that considers continuous and discrete outcomes can be written as

$$S_T^{DC}(t) = \exp\left[-\int_0^t \hbar_T(u)du\right] \prod_{j|t_j \leq t} [1 - \hbar_T(t_j)]. \tag{A.3}$$

$S_T^{DC}(t)$ is reduced to $\exp\left[-\int_0^t \hbar_T(u)du\right]$ in the continuous case and $\prod_{j|t_j \leq t} [1 - \hbar_T(t_j)]$

in the discrete case. For a proof, and a more theoretical explanation, see Gill & Johansen (1990).

Appendix B

Supplements to Chapter 4

B.1 Model selection

In practical situations, it is important to detect if $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1i})$ and $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2i})$ have components in common. This is basically a model selection problem and, to this end, we propose using the AIC, BIC and K-Fold Cross validation criterion $\left(\Upsilon^{\text{KCV}}\right)$. The AIC and BIC can be defined as

$$\begin{aligned}\text{AIC} &= -2\ell(\hat{\boldsymbol{\theta}}) + 2\,\text{EDF},\\ \text{BIC} &= -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)\,\text{EDF},\end{aligned}$$

where the log-likelihood is evaluated at the penalized parameter estimates and $\text{EDF} = \text{tr}(\hat{\boldsymbol{\mathcal{B}}})$ with $\hat{\boldsymbol{\mathcal{B}}}$ defined in Section 4.3.2.

As for Υ^{KCV} (Stone, 1974), we first randomly divide the set of observations in K groups (folds) of approximately equal size. Each fold is then in turn treated as a validation set, and the IPMLE for a given model is used to estimate the vector of parameters $\boldsymbol{\alpha}$ using the remaining $\text{K}-1$ folds. The so obtained estimates are

denoted as $\hat{\boldsymbol{\alpha}}_0^{\setminus k}$ and $\hat{\boldsymbol{\alpha}}_\nu^{\setminus k}$, and the log-likelihood function is calculated as

$$\ell_k(\hat{\boldsymbol{\alpha}}^{\setminus k}) = \left\{ \log \mathcal{G}_1 \left[\xi_{1i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_1^{\setminus k}) \right] + \delta_{1i} \log \left[-\frac{\mathcal{G}_1' \left[\xi_{1i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_1^{\setminus k}) \right]}{\mathcal{G}_1 \left[\xi_{1i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_1^{\setminus k}) \right]} \frac{\partial \xi_{1i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_1^{\setminus k})}{\partial y_i} \right] \right\}$$
$$+ \left\{ \log \mathcal{G}_2 \left[\xi_{2i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_2^{\setminus k}) \right] + \delta_{2i} \log \left[-\frac{\mathcal{G}_2' \left[\xi_{2i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_2^{\setminus k}) \right]}{\mathcal{G}_2 \left[\xi_{2i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_2^{\setminus k}) \right]} \frac{\partial \xi_{2i}(\hat{\boldsymbol{\alpha}}_0^{\setminus k}, \hat{\boldsymbol{\alpha}}_2^{\setminus k})}{\partial y_i} \right] \right\},$$

and Υ^{KCV} given by

$$\Upsilon^{\mathrm{KCV}} = \sum_{k=1}^{K} \ell_k(\hat{\boldsymbol{\alpha}}^{\setminus k}). \tag{B.1}$$

We choose the model which maximizes (B.1). The same procedure is used when Υ^{KCV} is calculated for the non-informative model. In such a case we have

$$\ell_k(\hat{\boldsymbol{\gamma}}^{\setminus k}) = \left\{ \log \mathcal{G}_1 \left[\xi_{1i}(\hat{\boldsymbol{\gamma}}_1^{\setminus k}) \right] + \delta_{1i} \log \left[-\frac{\mathcal{G}_1' \left[\xi_{1i}(\hat{\boldsymbol{\gamma}}_1^{\setminus k}) \right]}{\mathcal{G}_1 \left[\xi_{1i}(\hat{\boldsymbol{\gamma}}_1^{\setminus k}) \right]} \frac{\partial \xi_{1i}(\hat{\boldsymbol{\gamma}}_1^{\setminus k})}{\partial y_i} \right] \right\}$$
$$+ \left\{ \log \mathcal{G}_2 \left[\xi_{2i}(\hat{\boldsymbol{\gamma}}_2^{\setminus k}) \right] + \delta_{2i} \log \left[-\frac{\mathcal{G}_2' \left[\xi_{2i}(\hat{\boldsymbol{\gamma}}_2^{\setminus k}) \right]}{\mathcal{G}_2 \left[\xi_{2i}(\hat{\boldsymbol{\gamma}}_2^{\setminus k}) \right]} \frac{\partial \xi_{2i}(\hat{\boldsymbol{\gamma}}_2^{\setminus k})}{\partial y_i} \right] \right\},$$

and therefore $\Upsilon^{\mathrm{KCV}} = \sum_{k=1}^{K} \ell_k(\hat{\boldsymbol{\gamma}}^{\setminus k})$.

B.2 Informative and non-informative Scores

If censoring is informative then $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ would have some components in common. Because the first Q components of $\boldsymbol{\gamma}_1$ are the same as the first Q components of $\boldsymbol{\gamma}_2$, we have

$$\boldsymbol{\mathcal{Q}}_{\nu i}^{\top} \boldsymbol{\gamma}_\nu = \boldsymbol{\mathcal{Q}}_i^{0\top} \boldsymbol{\alpha}_0 + \boldsymbol{\mathcal{Q}}_{\nu i}^{1\top} \boldsymbol{\alpha}_\nu.$$

Therefore, defining $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^\top, \boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top)^\top$, the informative penalized log-likelihood function can be written as

$$\ell_p(\boldsymbol{\alpha}) = \ell(\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\Lambda} \boldsymbol{\alpha}, \tag{B.2}$$

where $\ell(\boldsymbol{\alpha})$ is defined as

$$\begin{aligned}\ell(\boldsymbol{\alpha}) = &\sum_{i=1}^{n} \left\{ \log \mathcal{G}_1 \left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right] + \delta_{1i} \log \left[-\frac{\mathcal{G}_1' \left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right]}{\mathcal{G}_1 \left[\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\right]} \frac{\partial \xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)}{\partial y_i} \right] \right\} \\ &+ \sum_{i=1}^{n} \left\{ \log \mathcal{G}_2 \left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right] + \delta_{2i} \log \left[-\frac{\mathcal{G}_2' \left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right]}{\mathcal{G}_2 \left[\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)\right]} \frac{\partial \xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)}{\partial y_i} \right] \right\}.\end{aligned}$$

The gradient of equation (B.2) can be calculated as

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell_p(\boldsymbol{\alpha}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\Lambda},$$

where $\boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}) = \left(\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0} \ell(\boldsymbol{\alpha})^\top, \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1} \ell(\boldsymbol{\alpha})^\top, \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2} \ell(\boldsymbol{\alpha})^\top \right)^\top$. The expressions $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0} \ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1} \ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_2} \ell(\boldsymbol{\alpha})$ can be obtained as $\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_0} = \left[\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{011}} \cdots \dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{0QJ_Q}} \right]^\top$, $\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_1} = \left[\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{111}} \cdots \dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{1Q_1J_{1Q_1}}} \right]^\top$ and $\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_2} = \left[\dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{21}} \cdots \dfrac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{2Q_2J_{2Q_2}}} \right]^\top$. In particular, the scalar derivatives of $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0} \ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1} \ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_2} \ell(\boldsymbol{\alpha})$ can be calculated as

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{0j}} &= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} \right\} + \sum_{i=1}^{n} \delta_{1i} \left\{ \left[-\frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial y_i} \right]^{-1} \left[-\frac{\mathcal{G}_1''}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} \frac{\partial \xi_{1i}}{\partial y_i} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} \frac{\partial \xi_{1i}}{\partial y_i} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{0j}} \right] \right\} \\ &\quad + \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} \right\} + \sum_{i=1}^{n} \delta_{2i} \left\{ \left[-\frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial y_i} \right]^{-1} \left[-\frac{\mathcal{G}_2''}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} \frac{\partial \xi_{2i}}{\partial y_i} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} \frac{\partial \xi_{2i}}{\partial y_i} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{0j}} \right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} + \delta_{1i} \left[\frac{\mathcal{G}_1''}{\mathcal{G}_1'} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{0j}} \left(\frac{\partial \xi_{1i}}{\partial y_i} \right)^{-1} \right] \right\} \\ &\quad + \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} + \delta_{2i} \left[\frac{\mathcal{G}_2''}{\mathcal{G}_2'} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{0j}} \left(\frac{\partial \xi_{2i}}{\partial y_i} \right)^{-1} \right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} \left[\frac{\mathcal{G}_1'}{\mathcal{G}_1} + \delta_{1i} \left(\frac{\mathcal{G}_1''}{\mathcal{G}_1'} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \right) \right] + \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} \left[\frac{\mathcal{G}_2'}{\mathcal{G}_2} + \delta_{2i} \left(\frac{\mathcal{G}_2''}{\mathcal{G}_2'} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \right) \right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{1i}}{\partial \alpha_{0j}} \Delta_1 + \frac{\partial \xi_{2i}}{\partial \alpha_{0j}} \Delta_2 \right\},\end{aligned} \tag{B.3}$$

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{1j}} &= \sum_{i=1}^{n}\left\{\frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\right\} + \sum_{i=1}^{n}\delta_{1i}\left\{\left[-\frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial y_i}\right]^{-1}\left[-\frac{\mathcal{G}_1''}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial y_i} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial y_i} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}} + \delta_{1i}\left[\frac{\mathcal{G}_1''}{\mathcal{G}_1'}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}} + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-1}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\left[\frac{\mathcal{G}_1'}{\mathcal{G}_1} + \delta_{1i}\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1'} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\right)\right] + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\delta_{1i}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-1}\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\Delta_1 + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\Omega_1\right\},
\end{aligned} \tag{B.4}$$

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_{2j}} &= \sum_{i=1}^{n}\left\{\frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\right\} + \sum_{i=1}^{n}\delta_{2i}\left\{\left[-\frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial y_i}\right]^{-1}\left[-\frac{\mathcal{G}_2''}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial y_i} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial y_i} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}} + \delta_{2i}\left[\frac{\mathcal{G}_2''}{\mathcal{G}_2'}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}} + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-1}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\left[\frac{\mathcal{G}_2'}{\mathcal{G}_2} + \delta_{2i}\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2'} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\right)\right] + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\delta_{2i}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-1}\right\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\Delta_2 + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\Omega_2\right\},
\end{aligned} \tag{B.5}$$

where $\xi_{\nu i} = \xi_{\nu i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu)$, $\Delta_\nu = \left[\frac{\mathcal{G}_\nu'}{\mathcal{G}_\nu} + \delta_{\nu i}\left(\frac{\mathcal{G}_\nu''}{\mathcal{G}_\nu'} - \frac{\mathcal{G}_\nu'}{\mathcal{G}_\nu}\right)\right]$ and $\Omega_\nu = \delta_{\nu i}\left(\frac{\partial \xi_{\nu i}}{\partial y_i}\right)^{-1}$. The last terms of equations (B.3), (B.4) and (B.5) allow to express $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha})$ as follow

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\Delta_1\frac{\partial \xi_{1i}}{\partial \boldsymbol{\alpha}_0} \;+\; \Delta_2\frac{\partial \xi_{2i}}{\partial \boldsymbol{\alpha}_0}\right], \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\Delta_1\frac{\partial \xi_{1i}}{\partial \boldsymbol{\alpha}_1} \;+\; \Omega_1\frac{\partial^2 \xi_{1i}}{\partial y_i \partial \boldsymbol{\alpha}_1}\right], \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\Delta_2\frac{\partial \xi_{2i}}{\partial \boldsymbol{\alpha}_2} \;+\; \Omega_2\frac{\partial^2 \xi_{2i}}{\partial y_i \partial \boldsymbol{\alpha}_2}\right],
\end{aligned}$$

where, for all $i = 1, ..., n$ and $\nu = 1, 2$, $\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_0} = \left[\frac{\partial \xi_{\nu i}}{\partial \alpha_{011}} \cdots \frac{\partial \xi_{\nu i}}{\partial \alpha_{0QJ_Q}}\right]^\top$, $\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu} = \left[\frac{\partial \xi_{\nu i}}{\partial \alpha_{\nu 11}} \cdots \frac{\partial \xi_{\nu i}}{\partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}}\right]^\top$ and $\frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu} = \left[\frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \alpha_{\nu 11}} \cdots \frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}}\right]^\top$. These ex-

pressions can be calculated using the design vectors defined in Section 2.2 as

$$\begin{aligned}
\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_0} &= \left(\boldsymbol{\mathcal{Q}}_1(\mathbf{x}_{1i}^0)^\top, \ldots, \boldsymbol{\mathcal{Q}}_Q(\mathbf{x}_{Qi}^0)^\top\right)^\top = \boldsymbol{\mathcal{Q}}_i^0, \\
\frac{\partial \xi_{\nu i}}{\partial y_i} &= \lim_{\varepsilon \to 0}\left\{\frac{\boldsymbol{\mathcal{Q}}_{\nu 0}(y_i+\varepsilon) - \boldsymbol{\mathcal{Q}}_{\nu 0}(y_i-\varepsilon)}{2\varepsilon}\right\}^\top \boldsymbol{\Gamma}_{\nu 0}\tilde{\boldsymbol{\alpha}}_{\nu 0} = \boldsymbol{\mathcal{Q}'}_{\nu 0}(y_i)^\top \boldsymbol{\Gamma}_{\nu 0}\tilde{\boldsymbol{\alpha}}_{\nu 0}, \\
\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_{\nu q_\nu}} &= \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i) & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0} \\ \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1) & \text{otherwise,} \end{cases} \\
\frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_{\nu q_\nu}} &= \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\prime}(y_i) & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}
\end{aligned}$$

where $\boldsymbol{\mathcal{Q}'}_{\nu 0}(y_i)$ can be conveniently obtained using a finite-difference method. Moreover, we define the design vectors $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\prime}(y_i)$ as

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle}(y_i) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 03}\right)] \\ \vdots \\ \mathcal{Q}_{\nu 0 J_{\nu 0}}(y_i) \exp\left(\alpha_{\nu 0 J_{\nu 0}}\right) \end{bmatrix} \quad \boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\prime}(y_i) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 03}\right)] \\ \vdots \\ \mathcal{Q}'_{\nu 0 J_{\nu 0}}(y_i) \exp\left(\alpha_{\nu 0 J_{\nu 0}}\right) \end{bmatrix}.$$

On the other hand, when censoring is non-informative the penalized log-likelihood function is

$$\ell_p(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\Lambda}\boldsymbol{\gamma}, \tag{B.6}$$

where $\ell(\boldsymbol{\gamma})$ can be written as

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left\{ \log \mathcal{G}_1 \left[\xi_{1i}(\boldsymbol{\gamma}_1)\right] + \delta_{1i} \log \left\{ -\frac{\mathcal{G}_1' \left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]}{\mathcal{G}_1 \left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]} \frac{\partial \xi_{1i}(\boldsymbol{\gamma}_1)}{\partial y_i} \right\} \right\} + \sum_{i=1}^{n} \left\{ \log \mathcal{G}_2 \left[\xi_{2i}(\boldsymbol{\gamma}_2)\right] + \delta_{2i} \log \left\{ -\frac{\mathcal{G}_2' \left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]}{\mathcal{G}_2 \left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]} \frac{\partial \xi_{2i}(\boldsymbol{\gamma}_2)}{\partial y_i} \right\} \right\}.$$

The gradient of (B.6) can be calculated as

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell_p(\boldsymbol{\gamma}) = \boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) - \boldsymbol{\gamma} \boldsymbol{\Lambda},$$

where $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) = \left(\boldsymbol{\nabla}_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})^\top,\ \boldsymbol{\nabla}_{\boldsymbol{\gamma}_2} \ell(\boldsymbol{\gamma})^\top \right)^\top$. In addition, $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})$ and $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_2} \ell(\boldsymbol{\gamma})$ can be calculated as $\dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_1} = \left[\dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{111}} \cdots \dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{1K_1 J_{1K_1}}} \right]^\top$ and $\dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_2} = \left[\dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{211}} \cdots \dfrac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{2K_2 J_{2K_2}}} \right]^\top$. Furthermore, the scalar derivatives of $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})$ and $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_2} \ell(\boldsymbol{\gamma})$ can be obtained as

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{1j}} &= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} \right\} + \sum_{i=1}^{n} \delta_{1i} \left\{ \left[-\frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial y_i} \right]^{-1} \left[-\frac{\mathcal{G}_1''}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} \frac{\partial \xi_{1i}}{\partial y_i} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} \frac{\partial \xi_{1i}}{\partial y_i} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \gamma_{1j}} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} + \delta_{1i} \left[\frac{\mathcal{G}_1''}{\mathcal{G}_1'} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \gamma_{1j}} \left(\frac{\partial \xi_{1i}}{\partial y_i} \right)^{-1} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} \left[\frac{\mathcal{G}_1'}{\mathcal{G}_1} + \delta_{1i} \left(\frac{\mathcal{G}_1''}{\mathcal{G}_1'} - \frac{\mathcal{G}_1'}{\mathcal{G}_1} \right) \right] + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \gamma_{1j}} \delta_{1i} \left(\frac{\partial \xi_{1i}}{\partial y_i} \right)^{-1} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{1i}}{\partial \gamma_{1j}} \Delta_1 + \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \gamma_{1j}} \Omega_1 \right\},
\end{aligned} \tag{B.7}$$

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{2j}} &= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} \right\} + \sum_{i=1}^{n} \delta_{2i} \left\{ \left[-\frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial y_i} \right]^{-1} \left[-\frac{\mathcal{G}_2''}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} \frac{\partial \xi_{2i}}{\partial y_i} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} \frac{\partial \xi_{2i}}{\partial y_i} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \gamma_{2j}} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} + \delta_{2i} \left[\frac{\mathcal{G}_2''}{\mathcal{G}_2'} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \gamma_{2j}} \left(\frac{\partial \xi_{2i}}{\partial y_i} \right)^{-1} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} \left[\frac{\mathcal{G}_2'}{\mathcal{G}_2} + \delta_{2i} \left(\frac{\mathcal{G}_2''}{\mathcal{G}_2'} - \frac{\mathcal{G}_2'}{\mathcal{G}_2} \right) \right] + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \gamma_{2j}} \delta_{2i} \left(\frac{\partial \xi_{2i}}{\partial y_i} \right)^{-1} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial \xi_{2i}}{\partial \gamma_{2j}} \Delta_2 + \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \gamma_{2j}} \Omega_2 \right\},
\end{aligned} \tag{B.8}$$

where $\xi_{\nu i} = \xi_{\nu i}(\boldsymbol{\gamma}_\nu)$. The last terms of equations (B.7) and (B.8) allow $\boldsymbol{\nabla}_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})$

and $\boldsymbol{\nabla}_{\gamma_2}\ell(\boldsymbol{\gamma})$ to be expressed as

$$\boldsymbol{\nabla}_{\gamma_1}\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left[\Delta_1\frac{\partial\xi_{1i}}{\partial\boldsymbol{\gamma}_1} + \Omega_1\frac{\partial^2\xi_{1i}}{\partial y_i\partial\boldsymbol{\gamma}_1}\right]$$
$$\boldsymbol{\nabla}_{\gamma_2}\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left[\Delta_2\frac{\partial\xi_{2i}}{\partial\boldsymbol{\gamma}_2} + \Omega_2\frac{\partial^2\xi_{2i}}{\partial y_i\partial\boldsymbol{\gamma}_2}\right],$$

where $\dfrac{\partial\xi_{\nu i}}{\partial\boldsymbol{\gamma}_\upsilon} = \left[\dfrac{\partial\xi_{\nu i}}{\partial\gamma_{\nu 11}}\cdots\dfrac{\partial\xi_{\nu i}}{\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}}\right]^\top$ and $\dfrac{\partial^2\xi_{\nu i}}{\partial y_i\partial\boldsymbol{\gamma}_\nu} = \left[\dfrac{\partial^2\xi_{\nu i}}{\partial y_i\partial\gamma_{\nu 11}}\cdots\dfrac{\partial^2\xi_{\nu i}}{\partial y_i\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}}\right]^\top$ for all $i = 1, ..., n$ and $\nu = 1, 2$. Furthermore, $\dfrac{\partial\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i}$, can be generically calculated using

$$\frac{\partial\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i} = \lim_{\varepsilon\to 0}\left\{\frac{\boldsymbol{\mathcal{Q}}_{\nu 0}(y_i+\varepsilon) - \boldsymbol{\mathcal{Q}}_{\nu 0}(y_i-\varepsilon)}{2\varepsilon}\right\}^\top\boldsymbol{\Gamma}_{\nu 0}\hat{\boldsymbol{\gamma}}_{\nu 0} = \boldsymbol{\mathcal{Q}}'_{\nu 0}(y_i)^\top\boldsymbol{\Gamma}_{\nu 0}\hat{\boldsymbol{\gamma}}_{\nu 0},$$

where $\boldsymbol{\mathcal{Q}}'_{\nu 0}(y_i)$ can also be calculated using a finite-difference method. The design vectors for $\dfrac{\partial\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\boldsymbol{\gamma}_\nu}$ and $\dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\boldsymbol{\gamma}_\nu}$ can be obtained using

$$\frac{\partial\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\boldsymbol{\gamma}_{\nu k_\nu}} = \begin{cases}\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0}\\ \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}) & \text{otherwise,}\end{cases}$$

$$\frac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\boldsymbol{\gamma}_{\nu k_\nu}} = \begin{cases}\boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(y_i) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0}\\ \mathbf{0} & \text{otherwise.}\end{cases}$$

Finally, we have that

$$\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(y_i) = \begin{bmatrix}\sum_{j_{\nu 0}=1}^{J_{\nu 0}}\mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}}\mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right]\exp(\gamma_{\nu 02})\\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}}\mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right]\exp(\gamma_{\nu 03})]\\ \vdots\\ \mathcal{Q}_{\nu 0 J_{\nu 0}}(y_i)\exp(\gamma_{\nu 0 J_{\nu 0}})\end{bmatrix}\quad \boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(y_i) = \begin{bmatrix}\sum_{j_{\nu 0}=1}^{J_{\nu 0}}\mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}}\mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right]\exp(\gamma_{\nu 02})\\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}}\mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right]\exp(\gamma_{\nu 03})]\\ \vdots\\ \mathcal{Q}'_{\nu 0 J_{\nu 0}}(y_i)\exp(\gamma_{\nu 0 J_{\nu 0}})\end{bmatrix}.$$

B.3 Informative and non-informative Hessians

The informative penalized Hessian can be obtained as

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_p(\boldsymbol{\alpha}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) - \boldsymbol{\Lambda},$$

where $\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha})$ is

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \end{bmatrix}. \tag{B.9}$$

In addition, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_\upsilon\boldsymbol{\alpha}_\kappa}\ell(\boldsymbol{\alpha}) = \dfrac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}_\upsilon\partial\boldsymbol{\alpha}_\kappa^\top}$, for all $\upsilon = 0, 1, 2$ and $\kappa = 0, 1, 2$. This expression is calculated using

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}_\upsilon\boldsymbol{\alpha}_\kappa}\ell(\boldsymbol{\alpha}) = \begin{bmatrix} \dfrac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{\upsilon 11}\partial\alpha_{\kappa 11}} & \cdots & \dfrac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{\upsilon 11}\partial\alpha_{\kappa Q_\kappa J_{\kappa Q_\kappa}}} \\ \cdots & \ddots & \cdots \\ \dfrac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{\upsilon Q_\upsilon J_{\upsilon Q_\upsilon}}\partial\alpha_{\kappa 11}} & \cdots & \dfrac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{\upsilon Q_\upsilon J_{\upsilon Q_\upsilon}}\partial\alpha_{\kappa Q_\kappa J_{\kappa Q_\kappa}}} \end{bmatrix}.$$

Since $\boldsymbol{\alpha}_1$ appears only in $\xi_{1i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ and $\boldsymbol{\alpha}_2$ only in $\xi_{2i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2)$, then $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) = \mathbf{0}$. Hence, (B.9) can be written as

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \mathbf{0} \\ \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \mathbf{0} & \boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \end{bmatrix}. \tag{B.10}$$

In equation (B.10), the scalar derivatives of $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha})$, $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha})$ and $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha})$, can be calculated as

$$
\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{0j}\partial\alpha_{0k}} &= \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_1''}{\mathcal{G}_1}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_1'''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} \\
&- \frac{\mathcal{G}_1''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{0k}} + \frac{\mathcal{G}_1''}{\mathcal{G}_1'}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{0k}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{0k}}\Bigg\} \\
&+ \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_2''}{\mathcal{G}_2}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_2'''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} \\
&- \frac{\mathcal{G}_2''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}} + \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{0k}} + \frac{\mathcal{G}_2''}{\mathcal{G}_2'}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{0k}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{0k}}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}}\left[\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right) + \delta_{1i}\left(\frac{\mathcal{G}_1'''}{\mathcal{G}_1} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2} - \frac{\mathcal{G}_1''}{\mathcal{G}_1} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right)\right] \\
&+ \frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}}\left[\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right) + \delta_{2i}\left(\frac{\mathcal{G}_2'''}{\mathcal{G}_2} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2} - \frac{\mathcal{G}_2''}{\mathcal{G}_2} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right)\right]\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{0k}}\Phi_1 + \frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{0k}}\Phi_2\Bigg\},
\end{aligned}
\tag{B.11}
$$

$$
\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{0j}\partial\alpha_{1k}} &= \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_1''}{\mathcal{G}_1}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} + \frac{\mathcal{G}_1'''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} \\
&- \frac{\mathcal{G}_1''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}} + \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{1k}} + \frac{\mathcal{G}_1''}{\mathcal{G}_1'}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{1k}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\alpha_{0j}\partial\alpha_{1k}}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}}\left[\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right) + \delta_{1i}\left(\frac{\mathcal{G}_1'''}{\mathcal{G}_1} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2} - \frac{\mathcal{G}_1''}{\mathcal{G}_1} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right)\right]\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{1i}}{\partial\alpha_{0j}}\frac{\partial\xi_{1i}}{\partial\alpha_{1k}}\Phi_1\Bigg\},
\end{aligned}
\tag{B.12}
$$

$$
\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\alpha})}{\partial\alpha_{0j}\partial\alpha_{2k}} &= \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_2''}{G_2}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} - \frac{\mathcal{G}_2'^2}{G_2^2}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} + \frac{\mathcal{G}_2'''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} \\
&- \frac{\mathcal{G}_2''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}} + \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{2k}} + \frac{\mathcal{G}_2''}{\mathcal{G}_2'}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{2k}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\alpha_{0j}\partial\alpha_{2k}}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}}\left[\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right) + \delta_{2i}\left(\frac{\mathcal{G}_2'''}{\mathcal{G}_2} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2} - \frac{\mathcal{G}_2''}{\mathcal{G}_2} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right)\right]\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{2i}}{\partial\alpha_{0j}}\frac{\partial\xi_{2i}}{\partial\alpha_{2k}}\Phi_2\Bigg\},
\end{aligned}
\tag{B.13}
$$

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \alpha_{1j}\partial \alpha_{1k}} &= \sum_{i=1}^{n}\Bigg\{ \frac{\mathcal{G}_1''}{\mathcal{G}_1}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} + \frac{\mathcal{G}_1'''}{\mathcal{G}_1}\delta_{1i}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2}\delta_{1i}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} \\
&- \frac{\mathcal{G}_1''}{\mathcal{G}_1}\delta_{1i}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\delta_{1i}\frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}} + \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial^2 \xi_{1i}}{\partial \alpha_{1j}\partial \alpha_{1k}} + \frac{\mathcal{G}_1''}{\mathcal{G}_1'}\delta_{1i}\frac{\partial^2 \xi_{1i}}{\partial \alpha_{1j}\partial \alpha_{1k}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\delta_{1i}\frac{\partial^2 \xi_{1i}}{\partial \alpha_{1j}\partial \alpha_{1k}} \\
&+ \frac{\partial^3 \xi_{1i}}{\partial y_i \partial \alpha_{1j}\partial \alpha_{1k}}\delta_{1i}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-1} - \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1k}}\frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\delta_{1i}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-2}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{ \frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}}\left[\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right) + \delta_{1i}\left(\frac{\mathcal{G}_1'''}{\mathcal{G}_1} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2} - \frac{\mathcal{G}_1''}{\mathcal{G}_1} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right)\right] \\
&+ \frac{\partial^2 \xi_{1i}}{\partial \alpha_{1j}\partial \alpha_{1k}}\left[\frac{\mathcal{G}_1'}{\mathcal{G}_1} + \delta_{1i}\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1'} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\right)\right] - \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1k}}\frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\left[\delta_{1i}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-2}\right] \\
&+ \frac{\partial^3 \xi_{1i}}{\partial y_i \partial \alpha_{1j}\partial \alpha_{1k}}\left[\delta_{1i}\left(\frac{\partial \xi_{1i}}{\partial y_i}\right)^{-1}\right]\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{ \frac{\partial \xi_{1i}}{\partial \alpha_{1j}}\frac{\partial \xi_{1i}}{\partial \alpha_{1k}}\Phi_1 + \frac{\partial^2 \xi_{1i}}{\partial \alpha_{1j}\partial \alpha_{1k}}\Delta_1 - \frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1k}}\frac{\partial^2 \xi_{1i}}{\partial y_i \partial \alpha_{1j}}\Psi_1 + \frac{\partial^3 \xi_{1i}}{\partial y_i \partial \alpha_{1j}\partial \alpha_{1k}}\Omega_1\Bigg\},
\end{aligned}
\tag{B.14}
$$

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \alpha_{2j}\partial \alpha_{2k}} &= \sum_{i=1}^{n}\Bigg\{ \frac{\mathcal{G}_2''}{\mathcal{G}_2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} + \frac{\mathcal{G}_2'''}{\mathcal{G}_2}\delta_{2i}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2}\delta_{2i}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} \\
&- \frac{\mathcal{G}_2''}{\mathcal{G}_2}\delta_{2i}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\delta_{2i}\frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}} + \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial^2 \xi_{2i}}{\partial \alpha_{2j}\partial \alpha_{2k}} + \frac{\mathcal{G}_2''}{\mathcal{G}_2'}\delta_{2i}\frac{\partial^2 \xi_{2i}}{\partial \alpha_{2j}\partial \alpha_{2k}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\delta_{2i}\frac{\partial^2 \xi_{2i}}{\partial \alpha_{2j}\partial \alpha_{2k}} \\
&+ \frac{\partial^3 \xi_{2i}}{\partial y_i \partial \alpha_{2j}\partial \alpha_{2k}}\delta_{2i}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-1} - \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2k}}\frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\delta_{2i}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-2}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{ \frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}}\left[\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right) + \delta_{2i}\left(\frac{\mathcal{G}_2'''}{\mathcal{G}_2} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2} - \frac{\mathcal{G}_2''}{\mathcal{G}_2} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right)\right] \\
&+ \frac{\partial^2 \xi_{2i}}{\partial \alpha_{2j}\partial \alpha_{2k}}\left[\frac{\mathcal{G}_2'}{\mathcal{G}_2} + \delta_{2i}\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2'} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\right)\right] - \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2k}}\frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\left[\delta_{2i}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-2}\right] \\
&+ \frac{\partial^3 \xi_{2i}}{\partial y_i \partial \alpha_{2j}\partial \alpha_{2k}}\left[\delta_{2i}\left(\frac{\partial \xi_{2i}}{\partial y_i}\right)^{-1}\right]\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{ \frac{\partial \xi_{2i}}{\partial \alpha_{2j}}\frac{\partial \xi_{2i}}{\partial \alpha_{2k}}\Phi_2 + \frac{\partial^2 \xi_{2i}}{\partial \alpha_{2j}\partial \alpha_{2k}}\Delta_2 - \frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2k}}\frac{\partial^2 \xi_{2i}}{\partial y_i \partial \alpha_{2j}}\Psi_2 + \frac{\partial^3 \xi_{2i}}{\partial y_i \partial \alpha_{2j}\partial \alpha_{2k}}\Omega_2\Bigg\},
\end{aligned}
\tag{B.15}
$$

where $\Phi_\nu = \delta_{\nu i}\left(\frac{\mathcal{G}_\nu'''}{\mathcal{G}_\nu} - \frac{\mathcal{G}_\nu''^2}{\mathcal{G}_\nu'^2} - \frac{\mathcal{G}_\nu''}{\mathcal{G}_\nu} + \frac{\mathcal{G}_\nu'^2}{\mathcal{G}_\nu^2}\right)$ and $\Psi_\nu = \left[\delta_{\nu i}\left(\frac{\partial \xi_{\nu i}}{\partial y_i}\right)^{-2}\right]$. Collecting the last terms of (B.11), (B.12), (B.13), (B.14) and (B.15), we obtain

$$\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_0 \partial \boldsymbol{\alpha}_0^\top} = \sum_{i=1}^{n}\left\{\Phi_1 \frac{\partial \xi_{1i}}{\partial \boldsymbol{\alpha}_0}\left[\frac{\partial \xi_{1i}}{\partial \boldsymbol{\alpha}_0}\right]^\top + \Phi_2 \frac{\partial \xi_{2i}}{\partial \boldsymbol{\alpha}_0}\left[\frac{\partial \xi_{2i}}{\partial \boldsymbol{\alpha}_0}\right]^\top\right\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_0 \partial \boldsymbol{\alpha}_\nu^\top} = \sum_{i=1}^{n}\left\{\Phi_\nu \frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_0}\left[\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu}\right]^\top\right\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top} = \sum_{i=1}^{n}\left\{\Phi_\nu \frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu}\left[\frac{\partial \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu}\right]^\top + \Delta_\nu \frac{\partial^2 \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top} - \Psi_\nu \frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu}\left[\frac{\partial^2 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu}\right]^\top + \Omega_\nu \frac{\partial^3 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top}\right\},$$

where

$$\frac{\partial^2 \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top} = \begin{bmatrix} \frac{\partial^2 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial \alpha_{\nu 11} \partial \alpha_{\nu 11}} & \cdots & \frac{\partial^2 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial \alpha_{\nu 11} \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}} \\ \cdots & \ddots & \cdots \\ \frac{\partial^2 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}} \partial \alpha_{\nu 11}} & \cdots & \frac{\partial^2 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}} \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}} \end{bmatrix},$$

$$\frac{\partial^3 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top} = \begin{bmatrix} \frac{\partial^3 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial y_i \partial \alpha_{\nu 11} \partial \alpha_{\nu 11}} & \cdots & \frac{\partial^3 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial y_i \partial \alpha_{\nu 11} \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}} \\ \cdots & \ddots & \cdots \\ \frac{\partial^3 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial y_i \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}} \partial \alpha_{\nu 11}} & \cdots & \frac{\partial^3 \xi_{\nu i}(\boldsymbol{\alpha}_\nu)}{\partial y_i \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}} \partial \alpha_{\nu Q_\nu J_{\nu Q_\nu}}} \end{bmatrix}.$$

In particular, the design sub-matrices of $\frac{\partial^2 \xi_{\nu i}}{\partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top}$ and $\frac{\partial^3 \xi_{\nu i}}{\partial y_i \partial \boldsymbol{\alpha}_\nu \partial \boldsymbol{\alpha}_\nu^\top}$ are calculated using

$$\frac{\partial^2 \xi_{\nu i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu)}{\partial \boldsymbol{\alpha}_{\nu q_\nu} \partial \boldsymbol{\alpha}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota \triangle \triangle}(y_i) & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu s_\nu} = \boldsymbol{\alpha}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

$$\frac{\partial^3 \xi_{\nu i}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu)}{\partial y_i \partial \boldsymbol{\alpha}_{\nu q_\nu} \boldsymbol{\alpha}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota \triangle \triangle'}(y_i) & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu s_\nu} = \boldsymbol{\alpha}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\triangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\triangle'}(y_i)$ are defined as

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\triangle}(y_i) = \begin{cases} \dfrac{\partial^2 \xi_{\nu i}}{\partial \alpha_{\nu 0 j_{\nu 0}} \partial \alpha_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^2 \xi_{\nu i}}{\partial \alpha_{\nu 0 j_{\nu 0}} \partial \alpha_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\iota\triangle\triangle'}(y_i) = \begin{cases} \dfrac{\partial^3 \xi_{\nu i}}{\partial y_i \alpha_{\nu 0 j_{\nu 0}} \partial \alpha_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\alpha_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^3 \xi_{\nu i}}{\partial y_i \alpha_{\nu 0 j_{\nu 0}} \partial \alpha_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise.} \end{cases}$$

On the other hand, the non-informative penalized Hessian is

$$\boldsymbol{\nabla}_{\gamma\gamma} \ell_p(\boldsymbol{\gamma}) = \boldsymbol{\nabla}_{\gamma\gamma} \ell(\boldsymbol{\gamma}) - \boldsymbol{\Lambda}.$$

Since $\xi_{1i}(\boldsymbol{\gamma}_1)$ and $\xi_{2i}(\boldsymbol{\gamma}_2)$) do not have parameters in common, $\boldsymbol{\nabla}_{\gamma\gamma} \ell(\boldsymbol{\gamma})$ can be written as

$$\boldsymbol{\nabla}_{\gamma\gamma} \ell(\boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{\nabla}_{\gamma_1\gamma_1} \ell(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\nabla}_{\gamma_2\gamma_2} \ell(\boldsymbol{\gamma}) \end{bmatrix},$$

where $\boldsymbol{\nabla}_{\gamma_\nu\gamma_\nu} \ell(\boldsymbol{\gamma}) = \dfrac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\top}$. This expression can be obtained using

$$\boldsymbol{\nabla}_{\gamma_\nu\gamma_\nu} \ell(\boldsymbol{\gamma}) = \begin{bmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \gamma_{\nu 11} \partial \gamma_{\nu 11}} & \cdots & \dfrac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \gamma_{\nu 11} \partial \gamma_{\nu K_\nu J_{\nu K_\nu}}} \\ \cdots & \ddots & \cdots \\ \dfrac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \gamma_{\nu K_\nu J_{\nu K_\nu}} \partial \gamma_{\nu 11}} & \cdots & \dfrac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \gamma_{\nu K_\nu J_{\nu K_\nu}} \partial \gamma_{\nu K_\nu J_{\nu K_\nu}}} \end{bmatrix}.$$

Furthermore, the scalar derivatives of $\boldsymbol{\nabla}_{\gamma_1\gamma_1}\ell(\boldsymbol{\gamma})$ and $\boldsymbol{\nabla}_{\gamma_2\gamma_2}\ell(\boldsymbol{\gamma})$ are

$$\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\gamma})}{\partial\gamma_{1j}\partial\gamma_{1k}} &= \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_1''}{\mathcal{G}_1}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} + \frac{\mathcal{G}_1'''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} \\
&\quad - \frac{\mathcal{G}_1''}{\mathcal{G}_1}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\delta_{1i}\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}} + \frac{\mathcal{G}_1'}{\mathcal{G}_1}\frac{\partial^2\xi_{1i}}{\partial\gamma_{1j}\partial\gamma_{1k}} + \frac{\mathcal{G}_1''}{\mathcal{G}_1'}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\gamma_{1j}\partial\gamma_{1k}} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\delta_{1i}\frac{\partial^2\xi_{1i}}{\partial\gamma_{1j}\partial\gamma_{1k}} \\
&\quad + \frac{\partial^3\xi_{1i}}{\partial y_i\partial\gamma_{1j}\partial\gamma_{1k}}\delta_{1i}\left(\frac{\partial\xi_{1i}}{\partial y_i}\right)^{-1} - \frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1k}}\frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1j}}\delta_{1i}\left(\frac{\partial\xi_{1i}}{\partial y_i}\right)^{-2}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}}\left[\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1} - \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right) + \delta_{1i}\left(\frac{\mathcal{G}_1'''}{\mathcal{G}_1} - \frac{\mathcal{G}_1''^2}{\mathcal{G}_1'^2} - \frac{\mathcal{G}_1''}{\mathcal{G}_1} + \frac{\mathcal{G}_1'^2}{\mathcal{G}_1^2}\right)\right] \\
&\quad + \frac{\partial^2\xi_{1i}}{\partial\gamma_{1j}\partial\gamma_{1k}}\left[\frac{\mathcal{G}_1'}{\mathcal{G}_1} + \delta_{1i}\left(\frac{\mathcal{G}_1''}{\mathcal{G}_1'} - \frac{\mathcal{G}_1'}{\mathcal{G}_1}\right)\right] - \frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1k}}\frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1j}}\left[\delta_{1i}\left(\frac{\partial\xi_{1i}}{\partial y_i}\right)^{-2}\right] \\
&\quad + \frac{\partial^3\xi_{1i}}{\partial y_i\partial\gamma_{1j}\partial\gamma_{1k}}\left[\delta_{1i}\left(\frac{\partial\xi_{1i}}{\partial y_i}\right)^{-1}\right]\Bigg\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial\xi_{1i}}{\partial\gamma_{1j}}\frac{\partial\xi_{1i}}{\partial\gamma_{1k}}\Phi_1 + \frac{\partial^2\xi_{1i}}{\partial\gamma_{1j}\partial\gamma_{1k}}\Delta_1 - \frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1k}}\frac{\partial^2\xi_{1i}}{\partial y_i\partial\gamma_{1j}}\Psi_1 + \frac{\partial^3\xi_{1i}}{\partial y_i\partial\gamma_{1j}\partial\gamma_{1k}}\Omega_1\right\},
\end{aligned} \tag{B.16}$$

$$\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\gamma})}{\partial\gamma_{2j}\partial\gamma_{2k}} &= \sum_{i=1}^{n}\Bigg\{\frac{\mathcal{G}_2''}{\mathcal{G}_2}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} + \frac{\mathcal{G}_2'''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} \\
&\quad - \frac{\mathcal{G}_2''}{\mathcal{G}_2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\delta_{2i}\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}} + \frac{\mathcal{G}_2'}{\mathcal{G}_2}\frac{\partial^2\xi_{2i}}{\partial\gamma_{2j}\partial\gamma_{2k}} + \frac{\mathcal{G}_2''}{\mathcal{G}_2'}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\gamma_{2j}\partial\gamma_{2k}} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\delta_{2i}\frac{\partial^2\xi_{2i}}{\partial\gamma_{2j}\partial\gamma_{2k}} \\
&\quad + \frac{\partial^3\xi_{2i}}{\partial y_i\partial\gamma_{2j}\partial\gamma_{2k}}\delta_{2i}\left(\frac{\partial\xi_{2i}}{\partial y_i}\right)^{-1} - \frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2k}}\frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2j}}\delta_{2i}\left(\frac{\partial\xi_{2i}}{\partial y_i}\right)^{-2}\Bigg\} \\
&= \sum_{i=1}^{n}\Bigg\{\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}}\left[\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2} - \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right) + \delta_{2i}\left(\frac{\mathcal{G}_2'''}{\mathcal{G}_2} - \frac{\mathcal{G}_2''^2}{\mathcal{G}_2'^2} - \frac{\mathcal{G}_2''}{\mathcal{G}_2} + \frac{\mathcal{G}_2'^2}{\mathcal{G}_2^2}\right)\right] \\
&\quad + \frac{\partial^2\xi_{2i}}{\partial\gamma_{2j}\partial\gamma_{2k}}\left[\frac{\mathcal{G}_2'}{\mathcal{G}_2} + \delta_{2i}\left(\frac{\mathcal{G}_2''}{\mathcal{G}_2'} - \frac{\mathcal{G}_2'}{\mathcal{G}_2}\right)\right] - \frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2k}}\frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2j}}\left[\delta_{2i}\left(\frac{\partial\xi_{2i}}{\partial y_i}\right)^{-2}\right] \\
&\quad + \frac{\partial^3\xi_{2i}}{\partial y_i\partial\gamma_{2j}\partial\gamma_{2k}}\left[\delta_{2i}\left(\frac{\partial\xi_{2i}}{\partial y_i}\right)^{-1}\right]\Bigg\} \\
&= \sum_{i=1}^{n}\left\{\frac{\partial\xi_{2i}}{\partial\gamma_{2j}}\frac{\partial\xi_{2i}}{\partial\gamma_{2k}}\Phi_2 + \frac{\partial^2\xi_{2i}}{\partial\gamma_{2j}\partial\gamma_{2k}}\Delta_2 - \frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2k}}\frac{\partial^2\xi_{2i}}{\partial y_i\partial\gamma_{2j}}\Psi_2 + \frac{\partial^3\xi_{2i}}{\partial y_i\partial\gamma_{2j}\partial\gamma_{2k}}\Omega_2\right\}.
\end{aligned} \tag{B.17}$$

The last terms of equations (B.16) and (B.17) allow to express $\boldsymbol{\nabla}_{\gamma_1\gamma_1}\ell(\boldsymbol{\gamma})$ and

$\boldsymbol{\nabla}_{\gamma_2\gamma_2}\ell(\boldsymbol{\gamma})$ as

$$\boldsymbol{\nabla}_{\gamma_\nu\gamma_\nu}\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\Bigg\{\Phi_{\nu i}\frac{\partial\xi_{\nu i}}{\partial\boldsymbol{\gamma}_\nu}\left[\frac{\partial\xi_{\nu i}}{\partial\boldsymbol{\gamma}_\nu}\right]^\top + \Delta_{\nu i}\frac{\partial^2\xi_{\nu i}}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top} - \Psi_{\nu i}\frac{\partial^2\xi_{\nu i}}{\partial y_i\partial\boldsymbol{\gamma}_\nu}\left[\frac{\partial^2\xi_{\nu i}}{\partial y_i\partial\boldsymbol{\gamma}_\nu}\right]^\top + \Omega_{\nu i}\frac{\partial^3\xi_{\nu i}}{\partial y_i\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top}\Bigg\},$$

where

$$\frac{\partial^2\xi_{\nu i}}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top} = \begin{bmatrix} \dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\gamma_{\nu 11}\partial\gamma_{\nu 11}} & \cdots & \dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\gamma_{\nu 11}\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}} \\ \cdots & \ddots & \cdots \\ \dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}\partial\gamma_{\nu 11}} & \cdots & \dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}} \end{bmatrix},$$

$$\frac{\partial^3\xi_{\nu i}}{\partial y_i\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top} = \begin{bmatrix} \dfrac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\gamma_{\nu 11}\partial\gamma_{\nu 11}} & \cdots & \dfrac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\gamma_{\nu 11}\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}} \\ \cdots & \ddots & \cdots \\ \dfrac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}\partial\gamma_{\nu 11}} & \cdots & \dfrac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}\partial\gamma_{\nu K_\nu J_{\nu K_\nu}}} \end{bmatrix}.$$

In addition, the design sub-matrices of $\dfrac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top}$ and $\dfrac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\top}$ can be obtained using the following equations

$$\frac{\partial^2\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial\boldsymbol{\gamma}_{\nu k_\nu}\partial\boldsymbol{\gamma}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(y_i) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu s_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

$$\frac{\partial^3\xi_{\nu i}(\boldsymbol{\gamma}_\nu)}{\partial y_i\partial\boldsymbol{\gamma}_{\nu k_\nu}\boldsymbol{\gamma}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(y_i) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu s_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(y_i)$ can be calculated as

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(y_i) = \begin{cases} \dfrac{\partial^2 \xi_{\nu i}}{\partial \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\gamma_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^2 \xi_{\nu i}}{\partial \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(y_i) = \begin{cases} \dfrac{\partial^3 \xi_{\nu i}}{\partial y_i \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(y_i)\right] \exp\left(\gamma_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^3 \xi_{\nu i}}{\partial y_i \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise.} \end{cases}$$

B.4 Proofs of Theorems 1, 2 and 3

B.4.1 Assumptions

This section provides the proofs of Theorems 1, 2 and 3 stated in Section 2.4. First, we establish the main set of assumptions (regularity conditions and vanishing penalties), then the main results are presented.

Since the same set of assumptions are used to proof Theorems 1 and 2, we use $\boldsymbol{\varphi}$ to represents the generic vector of parameters, where $\boldsymbol{\varphi} = \boldsymbol{\alpha}$ in Theorem 1 and $\boldsymbol{\varphi} = \boldsymbol{\gamma}$ in Theorem 2. The generic log-likelihood function can be written as

$$\ell(\boldsymbol{\varphi}) = \sum_{i=1}^{n} \log\left[\left[f_{T_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\varphi}_1) S_{T_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\varphi}_2)\right]^{\delta_i} \left[f_{T_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\varphi}_2) S_{T_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\varphi}_1)\right]^{(1-\delta_i)}\right]. \tag{B.18}$$

Let us define $\ell(\boldsymbol{\varphi}) = \sum_{i=1}^{n} \log \boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi})$, where $\boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi}) = \left[\left[f_{T_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\varphi}_1) S_{T_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\varphi}_2)\right]^{\delta_i} \left[f_{T_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\varphi}_2) S_{T_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\varphi}_1)\right]^{(1-\delta_i)}\right]$, $\mathbf{w}_i = (y_i, \mathbf{z}_{1i}^\top, \mathbf{z}_{2i}^\top)^\top \in \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, and $\mathbb{R}_+ = (0, \infty)$. Moreover, $\mathbf{z}_i = (\mathbf{z}_{1i}^\top, \mathbf{z}_{2i}^\top)^\top \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, $\ell(\mathbf{w}_i;\boldsymbol{\varphi}) = \log \boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi})$, $\ell_n(\boldsymbol{\varphi}) = n^{-1} \sum_{i=1}^{n} \ell(\mathbf{w}_i;\boldsymbol{\varphi})$, $\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}_i;\boldsymbol{\varphi}) = \dfrac{\partial \ell(\mathbf{w}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}}$, $\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell_n(\boldsymbol{\varphi}) = \dfrac{\partial \ell_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}}$, $\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \ell(\mathbf{w}_i;\boldsymbol{\varphi}) = \dfrac{\partial^2 \ell(\mathbf{w}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top}$ and $\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \ell_n(\boldsymbol{\varphi}) = \dfrac{\partial^2 \ell_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top}$. The penalised log-likelihood is $\ell_p(\boldsymbol{\varphi}) = \ell_n(\boldsymbol{\varphi}) - \frac{1}{2} \boldsymbol{\varphi}^\top \boldsymbol{\Lambda} \boldsymbol{\varphi}$.

Set of Assumptions 1 [Regularity conditions and vanishing penalty]

(A1) The true parameters vector $\boldsymbol{\varphi}^0$ is in the interior of $\mathcal{S}_{\boldsymbol{\varphi}} \subseteq \mathbb{R}^p$, which is a compact set, and $\mathcal{O}_{\boldsymbol{\varphi}^0}$ is an open neighbourhood around $\boldsymbol{\varphi}^0$.

(A2) For all $\mathbf{w}_i$, $\boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi})$ is continuous in $\boldsymbol{\varphi}$. Also, $\boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi})$ is measurable in $\mathbf{w}_i$ for all $\boldsymbol{\varphi} \in \mathcal{S}_{\boldsymbol{\varphi}}$.

(A3) The model is identified. That is, for any $\boldsymbol{\varphi}$ in $\mathcal{S}_{\boldsymbol{\varphi}}$ and for almost every $\mathbf{w}$, $\boldsymbol{\omega}(\mathbf{w};\boldsymbol{\varphi}) = \boldsymbol{\omega}(\mathbf{w};\boldsymbol{\varphi}^0)$ implies $\boldsymbol{\varphi}^0 = \boldsymbol{\varphi}$. In addition, $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\varphi}}}|\ell(\mathbf{w}_i;\boldsymbol{\varphi})|\} < \infty$.

(A4) For all $\mathbf{w}_i$, $\boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi})$ is three times continuously differentiable in $\boldsymbol{\varphi}$ in an open neighbourhood around $\boldsymbol{\varphi}^0$. That is $\boldsymbol{\omega}(\mathbf{w}_i;\boldsymbol{\varphi}) \in \mathcal{C}^3(\mathcal{O}_{\boldsymbol{\varphi}^0})$

(A5) $\int \sup_{\boldsymbol{\varphi}\in\mathcal{O}_{\boldsymbol{\varphi}^0}} \|\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi})\|\, d\mathbf{w}_i < \infty$ and $\int \sup_{\boldsymbol{\varphi}\in\mathcal{O}_{\boldsymbol{\varphi}^0}} \|\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi})\|\, d\mathbf{w}_i < \infty$.

(A6) For $\boldsymbol{\varphi} \in \mathcal{O}_{\boldsymbol{\varphi}^0}$, $\mathcal{I}(\boldsymbol{\varphi}^0) = \mathrm{Cov}\{\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi})\} = \mathbb{E}\{\{\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi}^0) - \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi}^0)]\}\{\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi}^0) - \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w}_i;\boldsymbol{\varphi}^0)]\}^\top\}$ exists and is positive-definite.

(A7) For all $1 \leq e, f, h \leq p+1$, there exist a function $\boldsymbol{\phi} : \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \longrightarrow \mathbb{R}$ such that, for $\boldsymbol{\varphi} \in \mathcal{O}_{\boldsymbol{\varphi}^0}$ and $\mathbf{w}_i \in \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, $\left|\dfrac{\partial^3\ell(\mathbf{w}_i;\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}_e\partial\boldsymbol{\varphi}_f\partial\boldsymbol{\varphi}_h}\right| \leq \boldsymbol{\phi}(\mathbf{w}_i)$, with $\mathbb{E}[\boldsymbol{\phi}(\mathbf{w}_i)] < \infty$.

(A8) The penalties vanish as the sample size n goes to infinite. That is $\boldsymbol{\lambda} = o(n^{-1/2})\mathbf{1}$.

In addition, the following lemmas are required to prove Theorems 1, 2 and 3.

Lemma 1. Let $s(\mathbf{w},\boldsymbol{\varphi})$ be a continuously differentiable function, a.s. $d\mathbf{w}$, on $\boldsymbol{\varphi} \in \mathcal{O}_{\boldsymbol{\varphi}^0}$.
If $\int \sup_{\boldsymbol{\varphi}\in\mathcal{O}_{\boldsymbol{\varphi}^0}} \left\|\dfrac{\partial s(\mathbf{w},\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}}\right\| d\mathbf{w} < \infty$, then for $\boldsymbol{\varphi} \in \mathcal{O}_{\boldsymbol{\varphi}^0}$,

(i) $\int s(\mathbf{w}, \boldsymbol{\varphi}) d\mathbf{w}$ is continuously differentiable.

(ii) $\int [\partial s(\mathbf{w}, \boldsymbol{\varphi})/\partial \boldsymbol{\varphi}] d\mathbf{w} = \partial [\int s(\mathbf{w}, \boldsymbol{\varphi}) d\mathbf{w}]/\partial \boldsymbol{\varphi}$.

Proof. Newey & McFadden (1994, Lemma 3.6). □

Lemma 2. If Assumptions (A1)-(A7) hold, then

(i) $\mathbb{E}[\boldsymbol{\nabla}_{\varphi} \ell(\mathbf{w}; \boldsymbol{\varphi}^0)] = \mathbf{0}$

(ii) $\mathbb{E}[-\boldsymbol{\nabla}_{\varphi\varphi} \ell(\mathbf{w}; \boldsymbol{\varphi}^0)] = \mathcal{I}(\boldsymbol{\varphi}^0)$

Proof.

(i) Since $\boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi})$ is a hypothetical density, its integral is unity:

$$\int \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy = 1.$$

This is an identity, valid for any $\boldsymbol{\varphi} \in \mathcal{S}_{\varphi}$. Differentiating both sides of this identity with respect to $\boldsymbol{\varphi}$, we obtain

$$\frac{\partial}{\partial \boldsymbol{\varphi}} \int \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy = \mathbf{0}.$$

Then, by (A4), (A5) and Lemma 1, the following expression is obtained

$$\frac{\partial}{\partial \boldsymbol{\varphi}} \int \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy = \int \frac{\partial}{\partial \boldsymbol{\varphi}} \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy. \tag{B.19}$$

By the definition of the score, we have

$$\boldsymbol{\nabla}_{\varphi} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) = \frac{\partial}{\partial \boldsymbol{\varphi}} \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}).$$

The last equation can be substituted into (B.19) to obtain

$$\int \boldsymbol{\nabla}_{\varphi} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy = \mathbf{0}. \tag{B.20}$$

This holds for any $\boldsymbol{\varphi} \in \mathcal{O}_{\boldsymbol{\varphi}^0}$, in particular, for $\boldsymbol{\varphi}^0$. Setting $\boldsymbol{\varphi} = \boldsymbol{\varphi}^0$, we have

$$\int \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}^0) dy = \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0) | \mathbf{z}] = \mathbf{0}.$$

Then, by applying the law of total expectations, we obtain

$$\mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0)] = \mathbb{E}\{\mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0) | \mathbf{z}]\} = \mathbf{0},$$

as required.

(ii) Differentiating both sides of (B.20) and by (A4), (A5) and Lemma 1, we have

$$\int \frac{\partial}{\partial \boldsymbol{\varphi}^\top} [\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi})] dy = \mathbf{0}. \tag{B.21}$$

The integrand of (B.21) can be written as

$$\frac{\partial}{\partial \boldsymbol{\varphi}^\top} [\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi})] = \boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) + \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi})^\top \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}).$$

We can substitute the last expression into (B.21) to obtain

$$-\int \boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy = \int \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}) \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi})^\top \boldsymbol{\omega}(\mathbf{w}; \boldsymbol{\varphi}) dy \tag{B.22}$$

By setting $\boldsymbol{\varphi} = \boldsymbol{\varphi}^0$, (B.22) can be written as

$$\mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0) | \mathbf{z}] = \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0) \boldsymbol{\nabla}_{\boldsymbol{\varphi}} \ell(\mathbf{w}; \boldsymbol{\varphi}^0)^\top | \mathbf{z}].$$

Then, by applying the law of total expectations, we obtain

$$\begin{aligned}\mathbb{E}\{\mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)|\mathbf{z}]\} &= \mathbb{E}\{\mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)\nabla_{\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)^\top|\mathbf{z}]\}.\\ \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)] &= \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)^\top].\\ \mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell(\mathbf{w};\boldsymbol{\varphi}^0)] &= \mathcal{I}(\boldsymbol{\varphi}^0)\end{aligned}$$

as required.

□

Lemma 3. Let $r \in \mathbb{R}_+$, and $\mathcal{O}_r$ be the surface of a sphere with radius $rn^{-1/2}$ and center $\boldsymbol{\varphi}^0$, that is $\mathcal{O}_r = \{\boldsymbol{\varphi} \in \mathcal{S}_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} = \boldsymbol{\varphi}^0 + n^{-1/2}\boldsymbol{r}, \|\boldsymbol{r}\| = r\}$. For any $\epsilon > 0$, there exist r such that $\mathbb{P}\left(\sup_{\boldsymbol{\varphi}\in\mathcal{O}_r} \ell_p(\boldsymbol{\varphi}) < \ell_p(\boldsymbol{\varphi}^0)\right) \geq 1-\epsilon$, when n is large enough.

Proof. We define $n\ell_p(\boldsymbol{\varphi}) - n\ell_p(\boldsymbol{\varphi}^0) = n\ell_n(\boldsymbol{\varphi}) - n\ell_n(\boldsymbol{\varphi}^0) - \frac{n}{2}[\boldsymbol{\varphi}^\top\boldsymbol{\Lambda}\boldsymbol{\varphi} - \boldsymbol{\varphi}^{0\top}\boldsymbol{\Lambda}\boldsymbol{\varphi}^0]$. A Third Order Taylor expansion around $\boldsymbol{\varphi}^0$ yields

$$\begin{aligned}n\ell_p(\boldsymbol{\varphi}) - n\ell_p(\boldsymbol{\varphi}^0) &= n\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)^\top(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0) + \frac{n}{2}(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)^\top\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)\\ &\quad - n\boldsymbol{\varphi}^{0\top}\boldsymbol{\Lambda}(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)\\ &\quad + \frac{n}{6}\sum_e\sum_f\sum_h(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)_e(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)_f(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0)_h\frac{\partial^3\ell_n(\bar{\boldsymbol{\varphi}})}{\partial\boldsymbol{\varphi}_e\partial\boldsymbol{\varphi}_f\partial_h\boldsymbol{\varphi}} - \frac{n}{2}(\boldsymbol{\varphi}\\ &\quad - \boldsymbol{\varphi}^0)^\top\boldsymbol{\Lambda}(\boldsymbol{\varphi}-\boldsymbol{\varphi}^0).\end{aligned} \tag{B.23}$$

Let $\boldsymbol{\varphi} = \boldsymbol{\varphi}^0 + n^{-1/2}\boldsymbol{r} \in \mathcal{O}_r$. Then (B.23) becomes in

$$\begin{aligned}n\ell_p(\boldsymbol{\varphi}) - n\ell_p(\boldsymbol{\varphi}^0) &= n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)^\top\boldsymbol{r} + \frac{1}{2}\boldsymbol{r}^\top\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)\boldsymbol{r}\\ &\quad + \frac{n^{-1/2}}{6}\sum_e\sum_f\sum_h\boldsymbol{r}_e\boldsymbol{r}_f\boldsymbol{r}_h\frac{\partial^3\ell_n(\bar{\boldsymbol{\varphi}})}{\partial\boldsymbol{\varphi}_e\partial\boldsymbol{\varphi}_f\partial_h\boldsymbol{\varphi}}\\ &\quad - n^{1/2}\boldsymbol{\varphi}^{0\top}\boldsymbol{\Lambda}\boldsymbol{r} - \frac{1}{2}\boldsymbol{r}^\top\boldsymbol{\Lambda}\boldsymbol{r},\end{aligned}$$

where $\bar{\boldsymbol{\varphi}}$ lies between $\boldsymbol{\varphi}^0$ and $\boldsymbol{\varphi}^0 + n^{-1/2}\boldsymbol{r}$. By (A5), Lemma 2(i) and the CLT, $n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathcal{I}(\boldsymbol{\varphi}^0)]$. Therefore, $|n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)^\top\boldsymbol{r}| = O_p(1)\|\boldsymbol{r}\|$. By

Lemma 2(ii) and the LLN, $n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0) \xrightarrow{p} -\mathcal{I}(\boldsymbol{\varphi}^0)$, which, by the continuous mapping theorem, yields $\frac{1}{2}\boldsymbol{r}^\top\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)\boldsymbol{r} \xrightarrow{p} -\frac{1}{2}\boldsymbol{r}^\top\mathcal{I}(\boldsymbol{\varphi}^0)\boldsymbol{r}$. Thus, by (A6), $\frac{1}{2}\boldsymbol{r}^\top\boldsymbol{\nabla}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\ell_n(\boldsymbol{\varphi}^0)\boldsymbol{r} \leq -\frac{1}{2}\zeta_{\min}\|\boldsymbol{r}\|^2$, where $\zeta_{\min} > 0$ is the smallest eigenvalue of $\mathcal{I}(\boldsymbol{\varphi}^0)$. By (A7) and the LLN, $\left|\dfrac{\partial^3\ell_n(\bar{\boldsymbol{\varphi}})}{\partial\boldsymbol{\varphi}_e\partial\boldsymbol{\varphi}_f\partial\boldsymbol{\varphi}_h}\right| \leq \frac{1}{n}\sum_1^n \boldsymbol{\phi}(\mathbf{w}_i) \xrightarrow{p} \mathbb{E}[\boldsymbol{\phi}(\mathbf{w}_i)] < \infty$. This fact and the Cauchy-Schwarz inequality imply that $\left|\dfrac{n^{-1/2}}{6}\sum_e\sum_f\sum_h \boldsymbol{r}_e\boldsymbol{r}_f\boldsymbol{r}_h\dfrac{\partial^3\ell_n(\bar{\boldsymbol{\varphi}})}{\partial\boldsymbol{\varphi}_e\partial\boldsymbol{\varphi}_f\partial_h\boldsymbol{\varphi}}\right| \xrightarrow{p} 0$. Finally, by (A8) we have that $\left|n^{1/2}\boldsymbol{\varphi}^{0\top}\boldsymbol{\Lambda}\boldsymbol{r}\right| \xrightarrow{p} 0$ and $\left|\frac{1}{2}\boldsymbol{r}^\top\boldsymbol{\Lambda}\boldsymbol{r}\right| \xrightarrow{p} 0$. Therefore, combining all of these results, we obtain

$$n\ell_p(\boldsymbol{\varphi}) - n\ell_p(\boldsymbol{\varphi}^0) \leq O_p(1)\|\boldsymbol{r}\| - \frac{1}{2}\zeta_{\min}\|\boldsymbol{r}\|^2 \tag{B.24}$$

for large enough n. Since the choice of $\boldsymbol{\varphi}$ was arbitrary, (B.24) becomes in

$$\sup_{\boldsymbol{\varphi}\in\mathcal{O}_r} n\ell_p(\boldsymbol{\varphi}) - n\ell_p(\boldsymbol{\varphi}^0) \leq \boldsymbol{\mathcal{C}},$$

where $\boldsymbol{\mathcal{C}} = O_p(1)\|\boldsymbol{r}\| - \frac{1}{2}\zeta_{\min}\|\boldsymbol{r}\|^2$. This implies that $\mathbb{P}\left(\sup_{\boldsymbol{\varphi}\in\mathcal{O}_r}\ell_p(\boldsymbol{\varphi}) < \ell_p(\boldsymbol{\varphi}^0)\right) \geq \mathbb{P}(\boldsymbol{\mathcal{C}} < 0)$. Finally, for all $\epsilon > 0$, there exists a $\|\boldsymbol{r}\| \in \mathbb{R}_+$ such that $\mathbb{P}[\boldsymbol{\mathcal{C}} < 0] \geq 1-\epsilon$, then $\mathbb{P}\left(\sup_{\boldsymbol{\varphi}\in\mathcal{O}_r}\ell_p(\boldsymbol{\varphi}) < \ell_p(\boldsymbol{\varphi}^0)\right) \geq 1-\epsilon$, as required. □

Lemma 4. (Delta Method). Suppose that $\boldsymbol{\varphi}_n$ is a sequence of k-dimensional random vectors and $\boldsymbol{\varphi}^0$ be a constant k-vector such that $\sqrt{n}(\boldsymbol{\varphi}_n - \boldsymbol{\varphi}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ for some $k \times k$ matrix $\boldsymbol{\Omega}$. Let $g : \mathbb{R}^k \to \mathbb{R}^l$ be continuously differentiable at $\boldsymbol{\varphi}^0$. Then

$$\sqrt{n}(g(\boldsymbol{\varphi}_n) - g(\boldsymbol{\varphi}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, G\boldsymbol{\Omega}G^\top)$$

where $G = \left.\dfrac{\partial g(\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}^\top}\right|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}^0}$ is the $l \times k$ Jacobian matrix.

Proof. Hayashi (2000, Lemma 2.5). □

B.4.2 Theorems 1, 2 and 3

Theorem 1 (Asymptotic properties of the IPMLE estimator).

Proof.

(i) By (A1), (A2) and Gourieroux & Monfort (1995, Property 24.1), there exists a well defined measurable function $\hat{\boldsymbol{\alpha}}$ that solves the optimization problem in equation (31). Due to Lemma 3, the informative censoring penalized log-likelihood function has a local maximum $\hat{\boldsymbol{\alpha}}$ in the interior of a sphere centered on $\boldsymbol{\alpha}^0$. Then, $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\| = O_p(n^{-1/2})$, implying that $\hat{\boldsymbol{\alpha}}$ is a $\sqrt{n}$-consistent estimator.

To prove the asymptotic normality of the informative censoring penalized likelihood estimator, we take the derivative of the log-likelihood function in equation (31) to obtain

$$\mathbf{0} = \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell_n(\hat{\boldsymbol{\alpha}}) - \boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}. \tag{B.25}$$

Applying a second order Taylor expansion to equation (B.25) yields

$$\mathbf{0} = \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell_n(\boldsymbol{\alpha}^0) - \boldsymbol{\Lambda}\boldsymbol{\alpha}^0 + \boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \ell_n(\boldsymbol{\alpha}^0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) - \boldsymbol{\Lambda}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) + \boldsymbol{\Delta}, \tag{B.26}$$

where the last term is defined as

$$\boldsymbol{\Delta} = \begin{bmatrix} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)^\top [\boldsymbol{\nabla}^2 \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell_n(\bar{\boldsymbol{\alpha}})]_1 (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \\ \vdots \\ (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)^\top [\boldsymbol{\nabla}^2 \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \ell_n(\bar{\boldsymbol{\alpha}})]_p (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \end{bmatrix}, \tag{B.27}$$

and $\bar{\boldsymbol{\alpha}}$ lies between $\boldsymbol{\alpha}^0$ and $\hat{\boldsymbol{\alpha}}$, therefore $\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\| \leq \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\|$. We can

rewrite equation (B.26) to obtain

$$\mathbf{0} = \boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0) - \boldsymbol{\Lambda}\boldsymbol{\alpha}^0 + \boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) - \boldsymbol{\Lambda}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) + \boldsymbol{\Delta}_p(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0), \tag{B.28}$$

where $\boldsymbol{\Delta}_p$ is defined as

$$\boldsymbol{\Delta}_p = \begin{bmatrix} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)^\top [\boldsymbol{\nabla}\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_n(\bar{\boldsymbol{\alpha}})]_1 \\ \vdots \\ (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)^\top [\boldsymbol{\nabla}\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_n(\bar{\boldsymbol{\alpha}})]_p \end{bmatrix}.$$

By multiplying the right hand side of (B.28) by $\sqrt{n}$, we obtain

$$[\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0) - \boldsymbol{\Lambda} + \boldsymbol{\Delta}_p]\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) = \sqrt{n}[\boldsymbol{\Lambda}\boldsymbol{\alpha}^0 - \boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0)]. \tag{B.29}$$

Assumption (A8) yields $\boldsymbol{\Lambda} \xrightarrow{p} 0$ and $\boldsymbol{\Lambda}\boldsymbol{\alpha}^0 \xrightarrow{p} 0$. By assumption (A7), we obtain $\boldsymbol{\Delta}_p \xrightarrow{p} 0$. By Lemma 2(i), (A5) and the CLT, $n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathcal{I}(\boldsymbol{\alpha}^0)]$. Furthermore, by Lemma 2(ii) and the LLN, $n^{1/2}\boldsymbol{\nabla}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_n(\boldsymbol{\alpha}^0) \xrightarrow{p} -\mathcal{I}(\boldsymbol{\alpha}^0)$. Finally, by Slutsky's theorem, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\alpha}^0)]^{-1}\right\},$$

as required.

(ii) Under Theorem 1, $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\alpha}^0)]^{-1}\}$. In particular, for $\hat{\boldsymbol{\alpha}_{\nu 0}} \in \hat{\boldsymbol{\alpha}}$ we have $\sqrt{n}(\hat{\boldsymbol{\alpha}_{\nu 0}} - \boldsymbol{\alpha}^0_{\nu 0}) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\alpha}^0_{\nu 0})]^{-1}\}$. In addition, $S : \mathbb{R}^k \to \mathbb{R}$ is continuously differentiable at $\boldsymbol{\alpha}^0_{\nu 0}$, with gradient defined as $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} S(\boldsymbol{\alpha}^0_{\nu 0}) =$

$\mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^0_{\nu 0})]\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^0_{\nu 0})$. Then, we can apply Lemma 4 to obtain

$$\sqrt{n}[\hat{S}_{\nu 0}(\hat{\boldsymbol{\alpha}}_{\nu 0}) - S_{\nu 0}(\boldsymbol{\alpha}^0_{\nu 0})] \xrightarrow{d} \mathcal{N}\Big\{\mathbf{0}, \mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^0_{\nu 0})]\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^0_{\nu 0})[\mathcal{I}(\boldsymbol{\alpha}^0_{\nu 0})]^{-1} \times \boldsymbol{\nabla}_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^0_{\nu 0})^\top \mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^0_{\nu 0})]\Big\}.$$

Furthermore, we know that $\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) = \mathbf{0}$, therefore $\mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}_0)] = \mathbf{0}$. This also implies that $\mathbb{E}[-\boldsymbol{\nabla}_{\boldsymbol{\alpha}_{10}\boldsymbol{\alpha}_{20}}\ell(\boldsymbol{\alpha}_0)] = \mathbf{0}$, which means that $\boldsymbol{\alpha}_{10}$ and $\boldsymbol{\alpha}_{20}$ are independent. Then, $S(\boldsymbol{\alpha}_{10})$ and $S(\boldsymbol{\alpha}_{20})$ are also independent, as required.

□

Theorem 2 (Asymptotic properties of the NPMLE estimator).

Proof. This proof follows similar arguments of Theorem 1. □

Theorem 3 (Efficiency of the IPMLE estimator).

Proof. For $\nu = 1, 2$, we define $\boldsymbol{\gamma}_\nu = (\boldsymbol{\gamma}^\iota_\nu, \boldsymbol{\gamma}^{n\iota}_\nu)^\top$ so that $\boldsymbol{\mathcal{Q}}_i^\top \boldsymbol{\gamma}_\nu = \boldsymbol{\mathcal{Q}}_i^{0\top}\boldsymbol{\gamma}^\iota_\nu + \boldsymbol{\mathcal{Q}}_{\nu i}^{1\top}\boldsymbol{\gamma}^{n\iota}_\nu$. Where $\boldsymbol{\gamma}^\iota_\nu = (\boldsymbol{\gamma}^{\iota\top}_{\nu 1}, ..., \boldsymbol{\gamma}^{\iota\top}_{\nu Q})^\top$ and $\boldsymbol{\gamma}^{n\iota}_\nu = (\boldsymbol{\gamma}^{n\iota\top}_{\nu(Q+1)}, ..., \boldsymbol{\gamma}^{n\iota\top}_{\nu Q_\nu})^\top$ are the informative and non-informative parameters of the non-informative model respectively. Thus, by Assumption (A6) and Lemma 2(ii), $\mathcal{I}(\boldsymbol{\gamma}^0)$ can be written as

$$\mathcal{I}(\boldsymbol{\gamma}^0) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\gamma}^\iota_1} & \mathcal{I}_{\boldsymbol{\gamma}^\iota_1\boldsymbol{\gamma}^{n\iota}_1} & \mathbf{0} & \mathbf{0} \\ \mathcal{I}_{\boldsymbol{\gamma}^{n\iota}_1\boldsymbol{\gamma}^\iota_1} & \mathcal{I}_{\boldsymbol{\gamma}^{n\iota}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{I}_{\boldsymbol{\gamma}^\iota_2} & \mathcal{I}_{\boldsymbol{\gamma}^\iota_2\boldsymbol{\gamma}^{n\iota}_2} \\ \mathbf{0} & \mathbf{0} & \mathcal{I}_{\boldsymbol{\gamma}^{n\iota}_2\boldsymbol{\gamma}^\iota_2} & \mathcal{I}_{\boldsymbol{\gamma}^{n\iota}_2} \end{bmatrix}, \tag{B.30}$$

where $\mathcal{I}_{\boldsymbol{\gamma}^\iota_\nu} = \mathcal{I}(\boldsymbol{\gamma}^{0\iota}_\nu)$, $\mathcal{I}_{\boldsymbol{\gamma}^{n\iota}_\nu} = \mathcal{I}(\boldsymbol{\gamma}^{0n\iota}_\nu)$ and $\mathcal{I}_{\boldsymbol{\gamma}^\iota_\nu\boldsymbol{\gamma}^{n\iota}_\nu} = \mathcal{I}(\boldsymbol{\gamma}^{0n\iota}_\nu, \boldsymbol{\gamma}^{0\iota}_\nu)$. Taking the inverse

of (B.30), we obtain

$$[\mathcal{I}(\boldsymbol{\gamma}^0)]^{-1} = \begin{bmatrix} \Sigma_{\boldsymbol{\gamma}_1^{0\iota}} & \Sigma_{\boldsymbol{\gamma}_1^{0\iota}\boldsymbol{\gamma}_1^{0n\iota}} & \mathbf{0} & \mathbf{0} \\ \Sigma_{\boldsymbol{\gamma}_1^{0n\iota}\boldsymbol{\gamma}_1^{0\iota}} & \Sigma_{\boldsymbol{\gamma}_1^{0n\iota}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{\boldsymbol{\gamma}_2^{0\iota}} & \Sigma_{\boldsymbol{\gamma}_2^{0\iota}\boldsymbol{\gamma}_2^{0n\iota}} \\ \mathbf{0} & \mathbf{0} & \Sigma_{\boldsymbol{\gamma}_2^{0\iota}\boldsymbol{\gamma}_2^{0n\iota}} & \Sigma_{\boldsymbol{\gamma}_2^{0n\iota}} \end{bmatrix}, \tag{B.31}$$

where $\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}} = [\mathcal{I}_{\boldsymbol{\gamma}_\nu^\iota} - \mathcal{I}_{\boldsymbol{\gamma}_\nu^\iota\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}_{\boldsymbol{\gamma}_\nu^{n\iota}\boldsymbol{\gamma}_\nu^\iota}]^{-1}$, $\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}\boldsymbol{\gamma}_\nu^{0n\iota}} = -\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}}\mathcal{I}_{\boldsymbol{\gamma}_\nu^\iota\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}}$, $\Sigma_{\boldsymbol{\gamma}_\nu^{0n\iota}\boldsymbol{\gamma}_\nu^{0\iota}} = -\mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}_{\boldsymbol{\gamma}_\nu^{n\iota}\boldsymbol{\gamma}_\nu^\iota}\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}}$ and $\Sigma_{\boldsymbol{\gamma}_\nu^{0n\iota}} = \mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}} + \mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}_{\boldsymbol{\gamma}_\nu^{n\iota}\boldsymbol{\gamma}_\nu^\iota}\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}}\mathcal{I}_{\boldsymbol{\gamma}_\nu^\iota\boldsymbol{\gamma}_\nu^{n\iota}}\mathcal{I}^{-1}_{\boldsymbol{\gamma}_\nu^{n\iota}}$.

On the other hand, by Assumption (A6) and Lemma 2(ii), $\mathcal{I}(\boldsymbol{\alpha}^0)$ can be written as

$$\mathcal{I}(\boldsymbol{\alpha}^0) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\alpha}_0} & \mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1} & \mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2} \\ \mathcal{I}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0} & \mathcal{I}_{\boldsymbol{\alpha}_1} & \mathbf{0} \\ \mathcal{I}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0} & \mathbf{0} & \mathcal{I}_{\boldsymbol{\alpha}_2} \end{bmatrix}, \tag{B.32}$$

where $\mathcal{I}_{\boldsymbol{\alpha}_0} = \mathcal{I}(\boldsymbol{\alpha}_0^0)$, $\mathcal{I}_{\boldsymbol{\alpha}_\nu} = \mathcal{I}(\boldsymbol{\alpha}_\nu^0)$, $\mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu} = \mathcal{I}(\boldsymbol{\alpha}_0^0, \boldsymbol{\alpha}_\nu^0)$ and $\mathcal{I}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0} = \mathcal{I}(\boldsymbol{\alpha}_\nu^0, \boldsymbol{\alpha}_0^0)$. Taking the inverse of (B.32), yields

$$[\mathcal{I}(\boldsymbol{\alpha}^0)]^{-1} = \begin{bmatrix} \Sigma_{\boldsymbol{\alpha}_0^0} & \Sigma_{\boldsymbol{\alpha}_0^0\boldsymbol{\alpha}_1^0} & \Sigma_{\boldsymbol{\alpha}_0^0\boldsymbol{\alpha}_2^0} \\ \Sigma_{\boldsymbol{\alpha}_1^0\boldsymbol{\alpha}_0^0} & \Sigma_{\boldsymbol{\alpha}_1^0} & \mathbf{0} \\ \Sigma_{\boldsymbol{\alpha}_2^0\boldsymbol{\alpha}_0^0} & \mathbf{0} & \Sigma_{\boldsymbol{\alpha}_2^0} \end{bmatrix}, \tag{B.33}$$

where $\Sigma_{\boldsymbol{\alpha}_0^0} = [\mathcal{I}_{\boldsymbol{\alpha}_0} - \mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1}\mathcal{I}^{-1}_{\boldsymbol{\alpha}_1}\mathcal{I}_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0} - \mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\mathcal{I}^{-1}_{\boldsymbol{\alpha}_2}\mathcal{I}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0}]^{-1}$, $\Sigma_{\boldsymbol{\alpha}_0^0\boldsymbol{\alpha}_\nu^0} = -\Sigma_{\boldsymbol{\alpha}_0^0}\mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu}\mathcal{I}^{-1}_{\boldsymbol{\alpha}_\nu}$, $\Sigma_{\boldsymbol{\alpha}_\nu^0\boldsymbol{\alpha}_0^0} = -\mathcal{I}^{-1}_{\boldsymbol{\alpha}_\nu}\mathcal{I}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0}\Sigma_{\boldsymbol{\alpha}_0^0}$ and $\Sigma_{\boldsymbol{\alpha}_\nu^0} = \mathcal{I}^{-1}_{\boldsymbol{\alpha}_\nu} + \mathcal{I}^{-1}_{\boldsymbol{\alpha}_\nu}\mathcal{I}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0}\Sigma_{\boldsymbol{\alpha}_0^0}\mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu}\mathcal{I}^{-1}_{\boldsymbol{\alpha}_\nu}$.

Thus, by (4.16), (4.17), (4.18), (4.19), (4.20) and using that $\boldsymbol{\gamma}_{\nu 0}^{n\iota} = \boldsymbol{\alpha}_{\nu 0}$, we obtain $\mathcal{I}_{\boldsymbol{\alpha}_0} = \mathcal{I}_{\boldsymbol{\gamma}_1^\iota} + \mathcal{I}_{\boldsymbol{\gamma}_2^\iota}$, $\mathcal{I}_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu} = \mathcal{I}_{\boldsymbol{\gamma}_\nu^\iota\boldsymbol{\gamma}_\nu^{n\iota}}$, $\mathcal{I}_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0} = \mathcal{I}_{\boldsymbol{\gamma}_\nu^{n\iota}\boldsymbol{\gamma}_\nu^\iota}$ and $\mathcal{I}_{\boldsymbol{\alpha}_\nu} = \mathcal{I}_{\boldsymbol{\gamma}_\nu^{n\iota}}$. This and the fact that $\Sigma^{-1}_{\boldsymbol{\alpha}_0^0}$ and $\Sigma^{-1}_{\boldsymbol{\gamma}_\nu^{0\iota}}$ are positive definite matrices, imply that $[\Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}} - \Sigma_{\boldsymbol{\alpha}_0^0}]$ is positive definite. Therefore, $\Sigma_{\boldsymbol{\alpha}_0^0} < \Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}}$. Using this reasoning, we conclude that $\Sigma_{\boldsymbol{\alpha}_0^0\boldsymbol{\alpha}_\nu^0} < \Sigma_{\boldsymbol{\gamma}_\nu^{0\iota}\boldsymbol{\gamma}_\nu^{0n\iota}}$, $\Sigma_{\boldsymbol{\alpha}_\nu^0\boldsymbol{\alpha}_0^0} < \Sigma_{\boldsymbol{\gamma}_\nu^{0n\iota}\boldsymbol{\gamma}_\nu^{0\iota}}$ and $\Sigma_{\boldsymbol{\alpha}_\nu^0} < \Sigma_{\boldsymbol{\gamma}_\nu^{0n\iota}}$, as required. □

The proof of Lemma 3 in the context of informative and non-informative censoring models was adapted from Xingwei et al. (2010) and Vatter & Chavez-Demoulin (2015). The proofs of the asymptotic normality (part (i) of Theorems 1 and 2) are based on Vatter & Chavez-Demoulin (2015).

B.5 Software details: `gamlss()` function

The models proposed in this article can be employed via the `gamlss()` function in the `R` package `GJRM` (Marra & Radice, 2020b). As an example, consider the following call

```
eq1 <- u ~ s(u, bs = "mpi") + z1 + s(z2),
eq2 <- u ~ s(u, bs = "mpi") + z1 + s(z2),
out <- gamlss(list(eq1, eq2), data = data, surv = TRUE,
margin = "PH", margin2 = "PH",cens = delta, informative = "yes",
inform.cov = c("z1")),
```

where `eq1` and `eq2` are the two additive predictors of the dependent censoring model. In these equations, `s(u, bs = "mpi")` represents the monotonic P-spline function which models a transformation of the baseline survival function. As for `s(z2)`, the default is `bs = "tp"` (penalized low rank thin plate spline) with `k = 10` (number of basis functions) and `m = 2` (order of derivatives). However, argument `bs` can also be set to, for example, `cr` (penalized cubic regression spline), `ps` (P-spline) and `mrf` (Markov random field), to name but a few. In the `gamlss` function, `surv = TRUE` indicates that a survival model is fitted. The arguments `margin ="PH"` and `margin2 ="PH"` specify the link functions for the survival and censoring times, respectively. Table 4.1 shows the possible choices for the links that have been implemented for this article. In this example, we specify

the proportional hazard link (`"PH"`) for the two equations. Argument `cens = delta` is a binary censoring indicator; this variable has to be equal to 1 if the event occurred and 0 otherwise. Finally, `informative = "yes"` indicates that we are fitting a survival model with informative censoring, and `inform.cov = c("z1")` specifies the set of informative covariates.

B.6 Additional simulation results for DGP2, DGP3 and DGP4

In DGP3, z_{1i} is informative, z_{2i} is informative and a mild censoring rate (about 47%) is considered. T_{1i} and T_{2i} were generated using the model defined in equation (4.25). The baseline survival functions were defined as $S_{10}(t_{1i}) = 0.8\exp(-0.4t_{1i}^{2.5}) + 0.2\exp(-0.1t_{1i}^{1.0})$ and $S_{20}(t_{2i}) = 0.99\exp(-0.05t_{2i}^{2.3}) + 0.01\exp(-0.4t_{2i}^{1.1})$. The informative covariates, z_{1i} and z_{2i}, were generated using a binomial and a uniform distribution respectively. Also, $s_{11}(z_{2i}) = s_{12}(z_{2i}) = \sin(2\pi z_i)$, $\alpha_{01} = -0.10$, $\alpha_{02} = -0.25$ and $\alpha_{11} = \alpha_{12} = -1.5$.

The main findings are:

- Figure B.1 and Table B.2 show that overall the mean estimates for the two estimators are very close to the respective true values and improve as the sample size increases. However, even though the variability of the estimates (IPMLE and NPMLE) decreases as the sample size grows large, the IPMLE is slightly more efficient than the NPMLE in recovering the true linear effects for all sample sizes examined here. In particular, the RMSE of the IPMLE is slightly smaller than the RMSE of the NPMLE for all sample sizes considered.

- Figures B.2 and B.3, and Table B.2 show that overall the true functions are recovered well by the IPMLE and NPMLE and that the results improve in

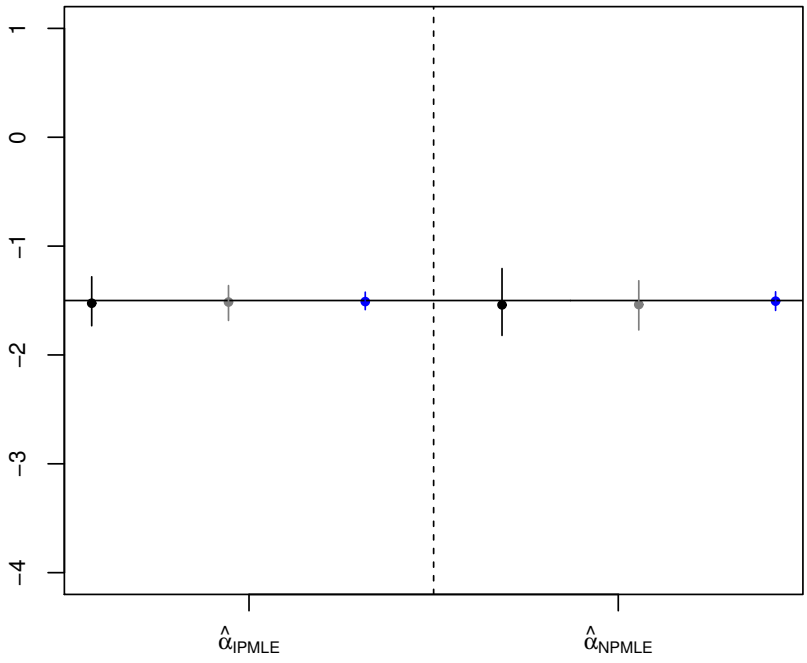

Figure B.1: Linear coefficient estimates obtained by applying `gamlss()` to informative survival data simulated according to DGP3 characterised by a censoring rate of about 47%. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 1000$ and $n = 4000$ are given in dark gray and blue, respectively.

terms of bias and efficiency as the sample size increases. Furthermore, the IPMLE is slightly more efficient than the NPMLE in recovering the non-linear covariate effects for all sample sizes examined in this section (Table B.2). However, this gain in efficiency by the IPMLE is not too significant when a mild censoring rate (47%) is examined.

In DGP4, z_{1i} is informative, z_{2i} is informative and it is considered a low censoring rate (about 29%). T_{1i} and T_{2i} were generated using also the model defined in equation (4.25). The baseline survival functions were defined as $S_{10}(t_{1i}) = 0.8\exp\left(-0.4t_{1i}^{2.4}\right) + 0.2\exp\left(-0.1t_{1i}^{1.0}\right)$ and $S_{20}(t_{2i}) = 0.99\exp\left(-0.065t_{2i}^{2.3}\right) +$

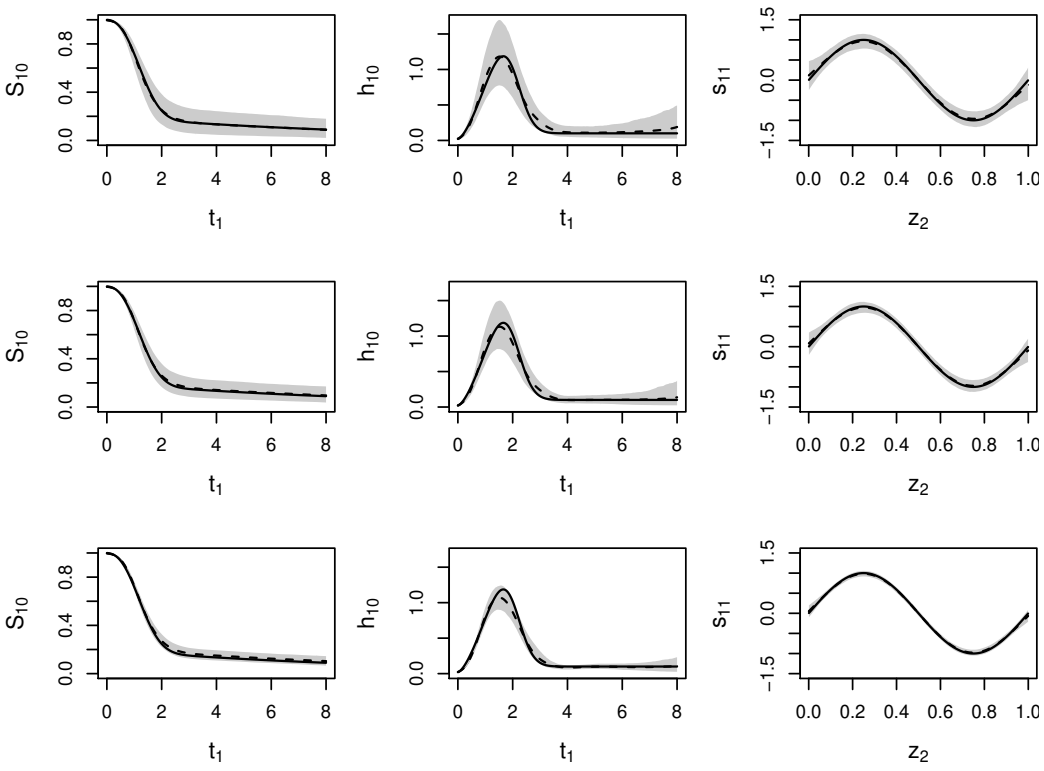

Figure B.2: Smooth function estimates for the IPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP3 characterised by a censoring rate of about 47%. True functions are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting from 5% and 95% quantiles by shaded areas. The results in the first row refer to $n = 500$, whereas those in the second and third rows to $n = 1000$ and $n = 4000$.

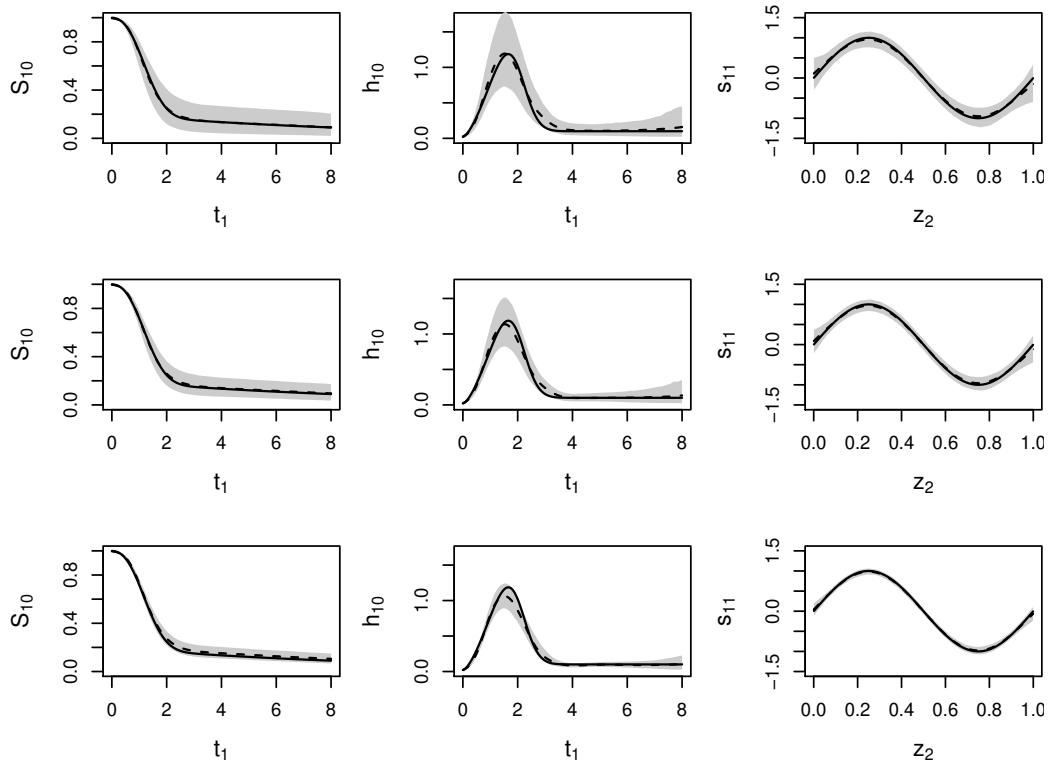

Figure B.3: Smooth function estimates for the NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP3 characterised by a censoring rate of about 47%. Further details are given in the caption of Figure B.2.

$0.01 \exp(-0.4 t_{2i}^{1.1})$. The informative covariates, z_{1i} and z_{2i}, were generated using a binomial and a uniform distribution respectively. Also, $s_{11}(z_{2i}) = s_{12}(z_{2i}) = \sin(2\pi z_i)$, $\alpha_{01} = -0.10$, $\alpha_{02} = -0.25$ and $\alpha_{11} = \alpha_{12} = -0.15$.

The main findings are:

- Figure B.4 and Table B.3 show that overall the mean estimates for the two estimators are very close to the respective true values and improve as the sample size increases. However, as for DGP3, the IPMLE is slightly more efficient than the NPMLE in recovering the true linear effects for all sample sizes examined here.

- Figures B.5 and B.6, and Table B.3 show that in general the true functions are recovered well by the IPMLE and NPMLE and that the results improve in terms of bias and efficiency as the sample size increases. Furthermore, the IPMLE is slightly more efficient than the NPMLE in recovering the non-linear covariate effects for all sample sizes examined in this section (Table B.3). However, this gain in efficiency by the IPMLE is not too significant when a low censoring rate (29%) is examined.

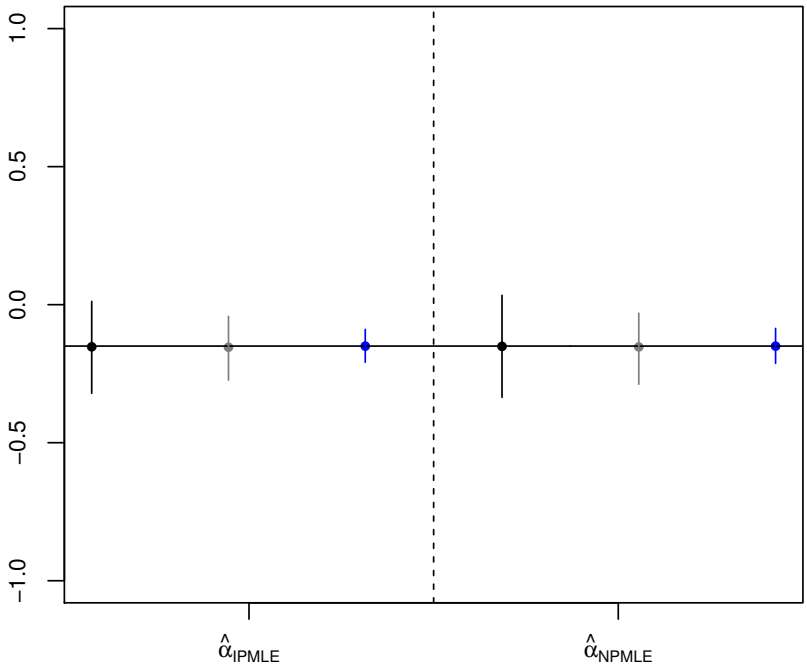

Figure B.4: Linear coefficient estimates obtained by applying `gamlss()` to informative survival data simulated according to DGP4 characterised by a censoring rate of about 29%. Further details are given in the caption of Figure B.1.

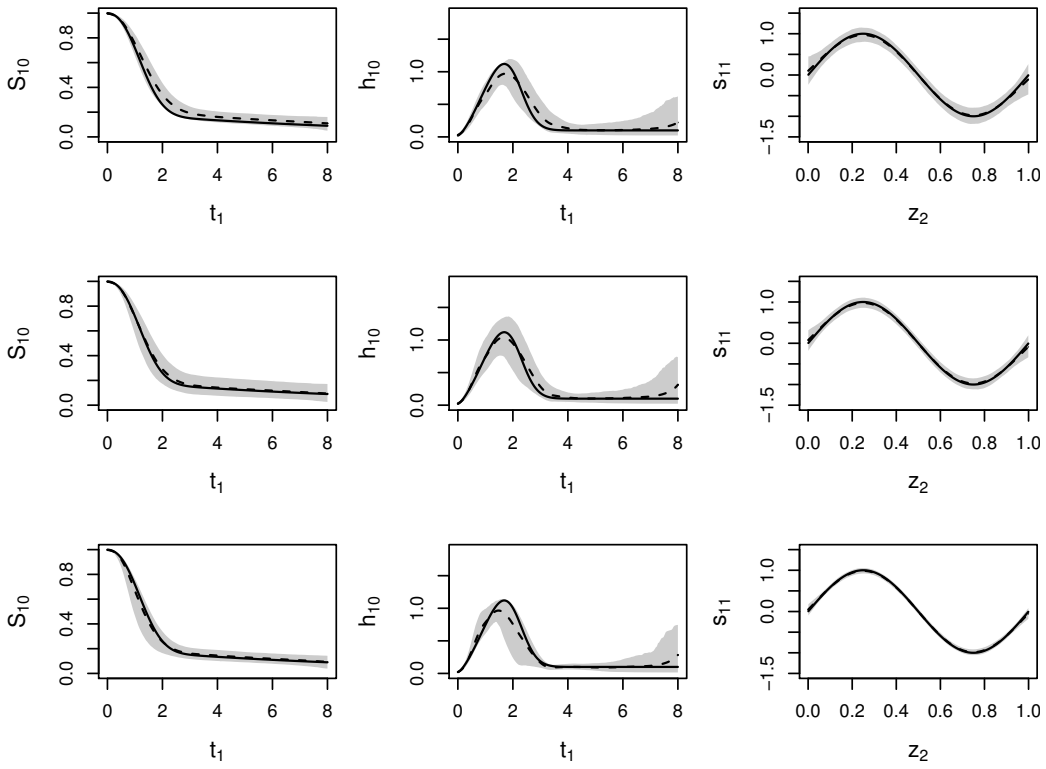

Figure B.5: Smooth function estimates for the IPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP4 characterised by a censoring rate of about 29%. Further details are given in the caption of Figure B.2.

(a) Informative Penalized Maximum Log-likelihood Estimator (IPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.024	-0.014	-0.006	0.138	0.100	0.049
s_1	0.039	0.025	0.012	0.154	0.114	0.059
h_{10}	0.084	0.048	0.035	0.262	0.144	0.083
S_{10}	0.028	0.020	0.017	0.063	0.050	0.031
(b) Non-informative Penalized Maximum Log-likelihood Estimator (NPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.045	-0.017	-0.007	0.208	0.144	0.071
s_1	0.085	0.068	0.044	0.191	0.206	0.111
h_{10}	0.085	0.057	0.033	0.195	0.292	0.083
S_{10}	0.027	0.021	0.015	0.058	0.068	0.033

Table B.1: Bias and root mean squared error (RMSE) for the IPMLE and NPMLE obtained by applying the `gamlss()` to informative survival data simulated according to DGP2 characterised by a censoring rate of about 74%. Further details are given in the caption of Table 4.2.

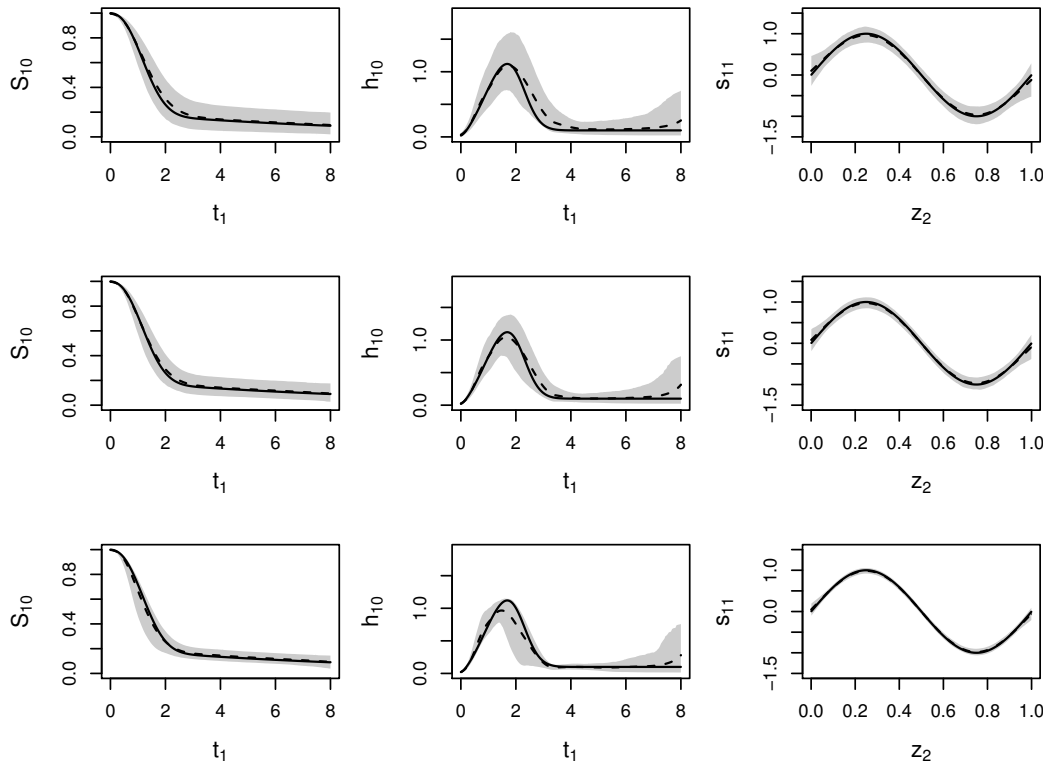

Figure B.6: Smooth function estimates for the NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP4 characterised by a censoring rate of about 29%. Further details are given in the caption of Figure B.2.

(a) Informative Penalized Maximum Log-likelihood Estimator (IPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.012	-0.006	0.003	0.121	0.058	0.045
s_1	0.031	0.021	0.015	0.124	0.091	0.051
h_{10}	0.040	0.027	0.026	0.135	0.088	0.058
S_{10}	0.003	0.008	0.015	0.057	0.047	0.030
(b) Non-informative Penalized Maximum Log-likelihood Estimator (NPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.022	0.001	0.007	0.140	0.100	0.050
s_1	0.036	0.027	0.014	0.142	0.104	0.055
h_{10}	0.037	0.027	0.027	0.131	0.089	0.056
S_{10}	0.004	0.008	0.017	0.065	0.047	0.032

Table B.2: Bias and root mean squared error (RMSE) for the IPMLE and NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP3 characterised by a censoring rate of about 47%. Further details are given in the caption of Table 4.2.

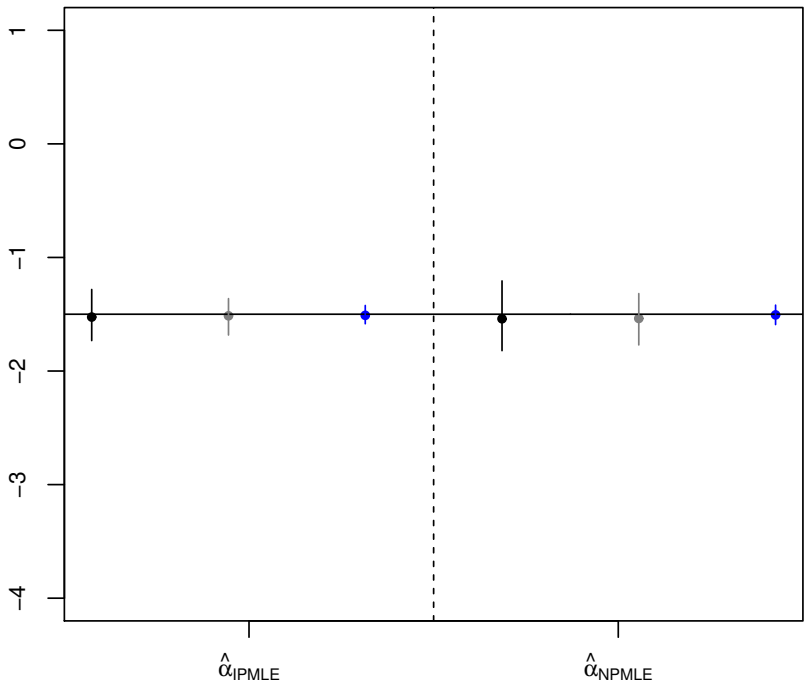

Figure B.7: Linear coefficient estimates obtained by applying `gamlss()` to informative survival data simulated according to DGP2 which is characterised by a censoring rate of about 74%. Further details are given in the caption of Figure B.1.

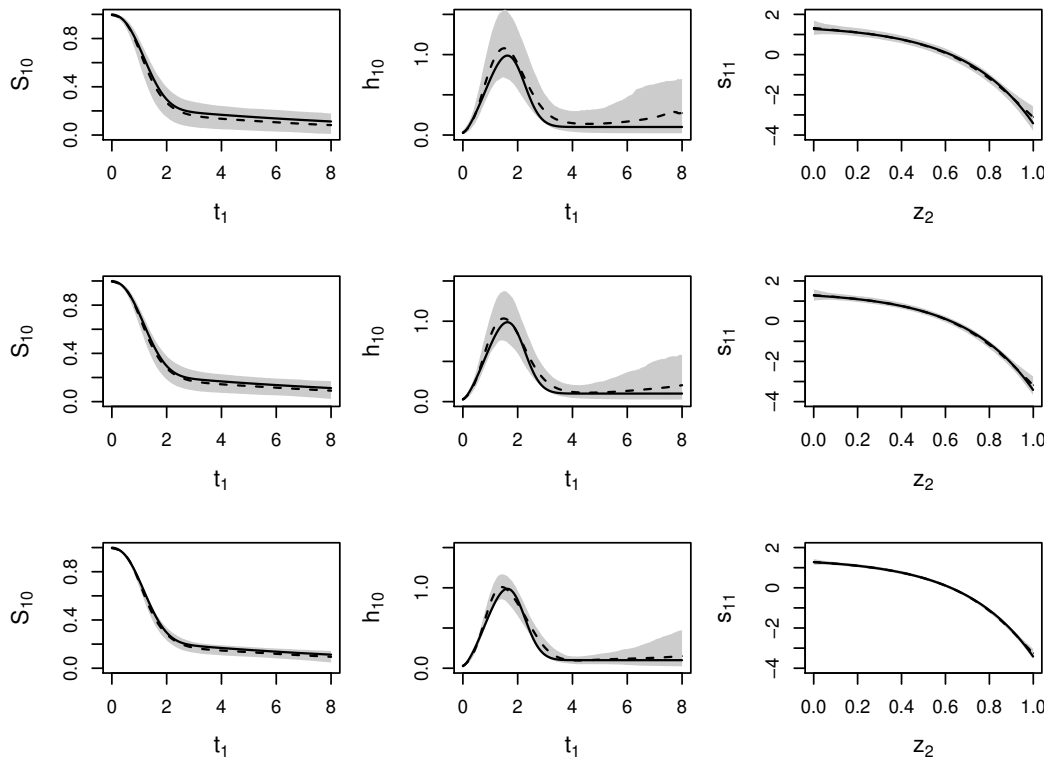

Figure B.8: Smooth function estimates for the IPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP2 characterised by a censoring rate of about 74%. Further details are given in the caption of Figure B.2.

(a) Informative Penalized Maximum Log-likelihood Estimator (IPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.003	-0.004	0.000	0.100	0.071	0.035
s_1	0.023	0.019	0.011	0.117	0.086	0.045
h_{10}	0.055	0.041	0.046	0.134	0.150	0.129
S_{10}	0.033	0.010	0.013	0.049	0.051	0.038
(b) Non-informative Penalized Maximum Log-likelihood Estimator (NPMLE)						
	Bias			RMSE		
	n = 500	n = 1000	n = 4000	n = 500	n = 1000	n = 4000
α_{11}	-0.001	-0.003	0.000	0.108	0.078	0.038
s_1	0.029	0.023	0.013	0.127	0.093	0.049
h_{10}	0.059	0.040	0.046	0.186	0.152	0.129
S_{10}	0.014	0.010	0.013	0.066	0.053	0.039

Table B.3: Bias and root mean squared error (RMSE) for the IPMLE and NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP4 characterised by a censoring rate of about 29%. Further details are given in the caption of Table 4.2.

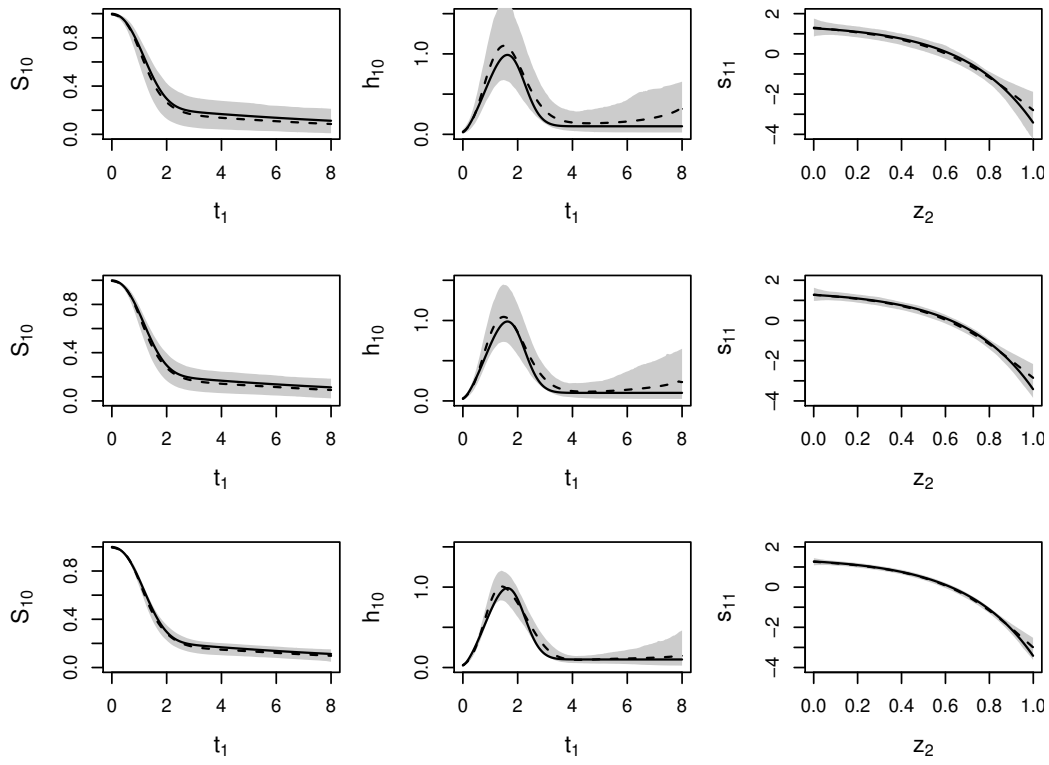

Figure B.9: Smooth function estimates for the NPMLE obtained by applying `gamlss()` to informative survival data simulated according to DGP2 characterised by a censoring rate of about 74%. Further details are given in the caption of Figure B.2.

Appendix C

Supplements to Chapter 5

C.1 Proofs of Theorems 4 and 5

Theorem 4 (Sub-densities)

Proof. Let us write the generic sub-survival function as $S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2) = P(T_j > t, \bigcap_{k\neq j} T_k > T_j|\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)$, where $j, k = 1, 2, 3$. Similarly, for $\varrho > 0$, $S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t + \varrho, l_1, l_2|z_1, z_2) = P(T_j > t + \varrho, \bigcap_{k\neq j} T_k > T_j|\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)$. Therefore, $[S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2) - S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t + \varrho, l_1, l_2|z_1, z_2)]$ can be defined as

$$\begin{aligned} &P(t < T_j \leq t + \varrho, \bigcap_{k\neq j} T_k > T_j|\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2) = \\ &[S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2) - S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t + \varrho, l_1, l_2|z_1, z_2)]. \end{aligned} \tag{C.1}$$

Let ψ be an arbitrary positive number such that $0 < \varrho < \psi$. The definition of $S_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2)$ implies that (C.1) has a lower and an upper bound. The lower bound can be written as

$$\begin{aligned} &P(t < T_j \leq t + \varrho, \bigcap_{k\neq j} T_k > t + \psi|\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2) = \\ &[S_{T_1,T_2,T_3|\mathbf{z}_1,\mathbf{z}_2}(t, t + \psi, t + \psi|z_1, z_2) - S_{T_1,T_2,T_3|\mathbf{z}_1,\mathbf{z}_2}(t + \varrho, t + \psi, t + \psi|z_1, z_2)]. \end{aligned} \tag{C.2}$$

Similarly, the upper bound can be defined as

$$\begin{aligned} &P(t < T_j \leq t + \varrho, \underset{k \neq j}{\cap} T_k > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2) = \\ &[S_{T_1,T_2,T_3|\mathbf{z}_1,\mathbf{z}_2}(t,t,t|z_1,z_2) - S_{T_1,T_2,T_3|\mathbf{z}_1,\mathbf{z}_2}(t+\varrho,t,t|z_1,z_2)]. \end{aligned} \tag{C.3}$$

Dividing (C.1), (C.2) and (C.3) by ϱ and also taking the limit as $\varrho \to 0$, we obtain, for all $\psi > 0$,

$$\begin{aligned} f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,l_1,l_2|z_1,z_2) &\geq \frac{\partial}{\partial t_j} P(T_j > t_j, \underset{k \neq j}{\cap} T_k > t + \psi | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)\Big|_{t_j = t}, \\ f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,l_1,l_2|z_1,z_2) &\leq \frac{\partial}{\partial t_j} P(T_j > t_j, \underset{k \neq j}{\cap} T_k > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)\Big|_{t_j = t}, \end{aligned} \tag{C.4}$$

where $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,l_1,l_2|z_1,z_2) = \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_j \leq t + \varrho, \underset{k \neq j}{\cap} T_k > T_j | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)$. In addition, by taking the limit as $\psi \to 0$, we obtain

$$f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,l_1,l_2|z_1,z_2) = -\frac{\partial}{\partial t_j} P(T_j > t_j, \underset{k \neq j}{\cap} T_k > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)\Big|_{t_j = t}. \tag{C.5}$$

On the other hand, because of A1, $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,1,0|z_1,z_2;\boldsymbol{\vartheta})$, $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,1|z_1,z_2;\boldsymbol{\vartheta})$ and $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,0|z_1,z_2;\boldsymbol{\vartheta})$ can be written as

$$\begin{aligned} f_{1,0}(\boldsymbol{\vartheta}) &= \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_1 \leq t + \varrho, T_2 > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}) P(T_3 > t), \\ f_{0,1}(\boldsymbol{\vartheta}) &= \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_2 \leq t + \varrho, T_1 > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}) P(T_3 > t), \\ f_{0,0}(\boldsymbol{\vartheta}) &= \lim_{\varrho \to 0} \ \varrho^{-1} P(t < T_3 \leq t + \varrho) P(T_1 > t, T_2 > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}). \end{aligned} \tag{C.6}$$

Furthermore, if we use (C.5), equation (C.6) can be expressed as

$$
\begin{aligned}
f_{1,0}(\boldsymbol{\vartheta}) &= -\frac{\partial}{\partial t_1} P(T_1 > t_1, T_2 > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)\Big|_{t_1=t} S_{T_3}(t), \\
f_{0,1}(\boldsymbol{\vartheta}) &= -\frac{\partial}{\partial t_2} P(T_1 > t, T_2 > t_1 | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)\Big|_{t_2=t} S_{T_3}(t), \\
f_{0,0}(\boldsymbol{\vartheta}) &= P(T_1 > t, T_2 > t | \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2; \boldsymbol{\vartheta}) f_T(t),
\end{aligned}
$$

where $S_{T_3}(t) = P(T > t)$ and $f_T(t) = \varrho^{-1} P(t < T_3 \leq t + \varrho)$. Finally, since $S(t_1, t_2 | \mathbf{z}; \boldsymbol{\vartheta}) = C\left[S_1(t_1 | \mathbf{z}_1; \boldsymbol{\gamma}_1), S_2(t_2 | \mathbf{z}_2; \boldsymbol{\gamma}_2); \theta\right]$ and $S_\nu(t_\nu | \mathbf{z}_\nu; \boldsymbol{\gamma}_\nu) = \mathcal{G}_\nu\left[\xi_\nu(t_\nu, \mathbf{z}_\nu; \boldsymbol{\gamma}_\nu)\right]$, we obtain

$$
\begin{aligned}
f_{1,0}(\boldsymbol{\vartheta}) = &\left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t, z_1; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t, z_2; \boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_1[\xi_1(t, z_1; \boldsymbol{\gamma}_1)]} \mathcal{G}_1'[\xi_1(t, z_1; \boldsymbol{\gamma}_1)] \frac{\partial \xi_1(t, z_1; \boldsymbol{\gamma}_1)}{\partial t}\right] \\
&\times S_{T_3}(t), \\
f_{0,1}(\boldsymbol{\vartheta}) = &\left[-\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t, z_1; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t, z_2; \boldsymbol{\gamma}_2)]; \theta\right\}}{\partial \mathcal{G}_2[\xi_2(t, z_2; \boldsymbol{\gamma}_2)]} \mathcal{G}_2'[\xi_2(t, z_2; \boldsymbol{\gamma}_2)] \frac{\partial \xi_2(t, z_2; \boldsymbol{\gamma}_2)}{\partial t}\right] \\
&\times S_{T_3}(t), \\
f_{0,0}(\boldsymbol{\vartheta}) = &C\left\{\mathcal{G}_1[\xi_1(t, z_1; \boldsymbol{\gamma}_1)], \mathcal{G}_2[\xi_2(t, z_2; \boldsymbol{\gamma}_2)]; \theta\right\} \times f_{T_3}(t),
\end{aligned}
$$

as required. □

Remark 4. The proof that $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2 | z_1, z_2)$ can be calculated directly from the joint survival function of the latent survival times (equation (C.5)) is due to Tsiatis (1975). His result was formulated in the context of competing risks when no covariates are included.

Theorem 5 (Local Identification Condition)

Proof. Let us define $f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})$ and $f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})$ as the shorthand notations of $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l_1, l_2 | z_1, z_2; \boldsymbol{\vartheta})$ and $\left.\frac{\partial f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l_1, l_2 | z_1, z_2; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^0}$, respectively. Since, by hypothesis, we know that the rank of $f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})$ is equal to p, the matrix

$f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})^\mathsf{T}$ is symmetric and the nonnegative square root ϕ of its smallest eigenvalue ϕ^2 is positive. Therefore, for $\mathbf{h} \in \mathbb{R}^p$, we have that $\left\|f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})\mathbf{h}\right\| \geq \|\phi\mathbf{h}\|$, where $\|\cdot\|$ is the euclidean norm. On the other hand, because $f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})$ is differentiable at $\boldsymbol{\vartheta}^0$, we have

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\left\|f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0+\mathbf{h}) - f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0) - f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})\mathbf{h}\right\|}{\|\mathbf{h}\|} = \mathbf{0}.$$

This implies that there exists a $\epsilon > 0$ such that, for all $\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^0\| < \varepsilon$, with $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}^0$, we have

$$\frac{\left\|f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}) - f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0) - f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|}{\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0\|} < \phi.$$

Let us write

$$\frac{\left\|f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}) - f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0) - f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|}{\left\|f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|} =$$

$$\frac{\left\|f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}) - f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0) - f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|}{\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0\|} \frac{\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0\|}{\left\|f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|}.$$

Since $\dfrac{\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0\|}{\left\|f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|} \leq \dfrac{1}{\phi}$, then $\dfrac{\left\|f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}) - f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0) - f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|}{\left\|f'_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0)\right\|} <$ 1. This implies that $f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}) \neq f_{y,\delta_1,\delta_2}(\boldsymbol{\vartheta}^0)$. Therefore $\boldsymbol{\vartheta}^0$ is locally identified on the neighbourhood $\{\boldsymbol{\vartheta} \in \mathcal{S} : \|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^0\| < \varepsilon\}$, as required. □

Remark 5. This result follows from the implicit function theorem, and can be found in (e.g., Chen et al., 2014; Stanghellini et al., 2013; Bekker, 1989). In particular, we have adapted the proof proposed by Chen et al. (2014) to our case.

Remark 6. As pointed out by Bekker & Wansbeek (2001), local identification is related to the existence of a consistent estimator. That is, if $\boldsymbol{\vartheta}^0$ is locally not identified, there exist vectors $\boldsymbol{\vartheta}$ arbitrarily close to $\boldsymbol{\vartheta}^0$ with $\boldsymbol{\vartheta}^0 \neq \boldsymbol{\vartheta}$ and

$f_{Y,\delta|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l|z_1, z_2; \boldsymbol{\vartheta}^0) = f_{Y,\delta|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l|z_1, z_2; \boldsymbol{\vartheta})$ for $l = 0, 1$ and for almost every (y, z_1, z_2). Therefore, in general, exact knowledge of $f_{Y,\delta|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l|z_1, z_2; \boldsymbol{\vartheta}^0)$ is not sufficient to distinguish between $\boldsymbol{\vartheta}^0$ and $\boldsymbol{\vartheta}$. Suppose that $\hat{\boldsymbol{\vartheta}}$ is an estimator of $\boldsymbol{\vartheta}$. Its limit distribution is a function of $f_{Y,\delta|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l|z_1, z_2; \boldsymbol{\vartheta}^0)$, therefore exact knowledge of this distribution function is not enough to distinguish between $\boldsymbol{\vartheta}^0$ and $\boldsymbol{\vartheta}$. This means that $\boldsymbol{\vartheta}^0$ can not be expressed as a function of the large sample distribution of $\hat{\boldsymbol{\vartheta}}$. In particular, $\boldsymbol{\vartheta}^0$ can not be expressed as the probability limit of $\hat{\boldsymbol{\vartheta}}$. On the contrary, if $\boldsymbol{\vartheta}^0$ is locally identified, and if the parameter space is restricted to a sufficiently small open neighbourhood of $\boldsymbol{\vartheta}^0$, $f_{Y,\delta|\mathbf{z}_1,\mathbf{z}_2}(\cdot, l|z_1, z_2; \boldsymbol{\vartheta}^0)$ corresponds uniquely to a single value $\boldsymbol{\vartheta}^0 = \boldsymbol{\vartheta}$.

C.2 Dependent censoring Score and Hessian

In this section, the detailed derivations of the Score and the Hessian for the dependent censoring model are presented.

C.2.1 Dependent censoring Score

First, let us define $f_{1,0}(\boldsymbol{\vartheta})$, $f_{0,1}(\boldsymbol{\vartheta})$ and $f_{0,0}(\boldsymbol{\vartheta})$ as the shorthand notations for $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 1, 0|z_1, z_2; \boldsymbol{\vartheta})$, $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 0, 1|z_1, z_2; \boldsymbol{\vartheta})$ and $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, 0, 0|z_1, z_2; \boldsymbol{\vartheta})$. As discussed in section 4, the penalised log-likelihood function of $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta)$ is

$$\ell_p(\boldsymbol{\vartheta}) = \ell(\boldsymbol{\vartheta}) - \frac{1}{2}\boldsymbol{\vartheta}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{\vartheta}, \tag{C.7}$$

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \left[\delta_{1i} \log f_{1,0}(\boldsymbol{\vartheta}) + \delta_{2i} \log f_{0,1}(\boldsymbol{\vartheta}) + \delta_{3i} f_{0,0}(\boldsymbol{\vartheta})\right], \tag{C.8}$$

$$
\begin{aligned}
f_{1,0}(\boldsymbol{\vartheta}) &= -\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)],\mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)];\theta\right\}}{\partial \mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)]}\mathcal{G}_1'[\xi_1(t,z_1;\boldsymbol{\gamma}_1)]\frac{\partial \xi_1(t,z_1;\boldsymbol{\gamma}_1)}{\partial t},\\
f_{0,1}(\boldsymbol{\vartheta}) &= -\frac{\partial C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)],\mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)];\theta\right\}}{\partial \mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]}\mathcal{G}_2'[\xi_2(t,z_2;\boldsymbol{\gamma}_2)]\frac{\partial \xi_2(t,z_2;\boldsymbol{\gamma}_2)}{\partial t},\\
f_{0,0}(\boldsymbol{\vartheta}) &= C\left\{\mathcal{G}_1[\xi_1(t,z_1;\boldsymbol{\gamma}_1)],\mathcal{G}_2[\xi_2(t,z_2;\boldsymbol{\gamma}_2)];\theta\right\}.
\end{aligned}
\tag{C.9}
$$

Then, the penalised score, $\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell_p(\boldsymbol{\vartheta})$, of (C.7) can be calculated as

$$
\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell_p(\boldsymbol{\vartheta}) = \boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}) - \boldsymbol{\vartheta}\Lambda, \tag{C.10}
$$

$$
\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\left[\frac{\delta_{1i}}{f_{1,0}(\boldsymbol{\vartheta})}\frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} + \frac{\delta_{2i}}{f_{0,1}(\boldsymbol{\vartheta})}\frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} + \frac{\delta_{3i}}{f_{0,0}(\boldsymbol{\vartheta})}\frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]. \tag{C.11}
$$

For $l_1, l_2 = 0, 1$, we can write

$$
\frac{\partial f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \begin{bmatrix} \dfrac{\partial f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix}. \tag{C.12}
$$

Then, (C.20) becomes in

$$
\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\frac{\delta_{1i}}{f_{1,0}(\boldsymbol{\vartheta})}\begin{bmatrix} \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} + \sum_{i=1}^{n}\frac{\delta_{2i}}{f_{0,1}(\boldsymbol{\vartheta})}\begin{bmatrix} \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} + \sum_{i=1}^{n}\frac{\delta_{3i}}{f_{0,0}(\boldsymbol{\vartheta})}\begin{bmatrix} \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix}. \tag{C.13}
$$

In particular, for $f_{1,0}(\boldsymbol{\vartheta})$, we have

$$\begin{bmatrix} \dfrac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial^2 C}{\partial \mathcal{G}_1^2} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \xi_1}{\partial t} \dfrac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} + \dfrac{\partial C}{\partial \mathcal{G}_1} \dfrac{\partial \mathcal{G}_1^2}{\partial \xi_1^2} \dfrac{\partial \xi_1}{\partial t} \dfrac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} + \dfrac{\partial C}{\partial \mathcal{G}_1} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \xi_1^2}{\partial t \partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial^2 C}{\partial \mathcal{G}_1 \partial \mathcal{G}_2} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_1}{\partial t} \dfrac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial^2 C}{\partial \mathcal{G}_1 \partial \theta} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \xi_1}{\partial t} \end{bmatrix}. \tag{C.14}$$

Similarly, $f_{0,1}(\boldsymbol{\vartheta})$, we obtain

$$\begin{bmatrix} \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial^2 C}{\partial \mathcal{G}_2 \partial \mathcal{G}_1} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_2}{\partial t} \dfrac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial^2 C}{\partial \mathcal{G}_2^2} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_2}{\partial t} \dfrac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} + \dfrac{\partial C}{\partial \mathcal{G}_2} \dfrac{\partial \mathcal{G}_2^2}{\partial \xi_2^2} \dfrac{\partial \xi_2}{\partial t} \dfrac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} + \dfrac{\partial C}{\partial \mathcal{G}_2} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_2^2}{\partial t \partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial^2 C}{\partial \mathcal{G}_2 \partial \theta} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_2}{\partial t} \end{bmatrix}. \tag{C.15}$$

Finally, if $f_{0,0}(\boldsymbol{\vartheta})$ is observed, then

$$\begin{bmatrix} \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial \mathcal{G}_1} \dfrac{\partial \mathcal{G}_1}{\partial \xi_1} \dfrac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} \\ \dfrac{\partial C}{\partial \mathcal{G}_2} \dfrac{\partial \mathcal{G}_2}{\partial \xi_2} \dfrac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} \\ \dfrac{\partial C}{\partial \theta} \end{bmatrix}. \tag{C.16}$$

If we define $\Psi_\nu = \dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} \dfrac{\partial \xi_\nu}{\partial t}$, $\Delta_\nu = \left[\Psi_\nu \dfrac{\partial^2 C}{\partial \mathcal{G}_\nu^2} \dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} + \dfrac{\partial C}{\partial \mathcal{G}_\nu} \dfrac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2} \dfrac{\partial \xi_\nu}{\partial t} \right]$, $\Omega_\nu = \left[\dfrac{\partial C}{\partial \mathcal{G}_\nu} \dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} \right]$ and $\Upsilon_\nu = \left[\Psi_\nu \dfrac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega} \dfrac{\partial \mathcal{G}_\omega}{\partial \xi_\omega} \right]$,

with $\nu = 1, 2$, $\omega = 1, 2$ and $\nu \neq \omega$, equations (C.14), (C.15) and (C.16) can be

written as

$$\begin{bmatrix} \frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = - \begin{bmatrix} \Delta_1 \frac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} + \Omega_1 \frac{\partial \xi_1^2}{\partial t \partial \boldsymbol{\gamma}_1} \\ \Upsilon_1 \frac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} \\ \Psi_1 \frac{\partial^2 C}{\partial \mathcal{G}_1 \partial \theta} \end{bmatrix}, \quad \begin{bmatrix} \frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = - \begin{bmatrix} \Upsilon_2 \frac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} \\ \Delta_2 \frac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} + \Omega_2 \frac{\partial \xi_2^2}{\partial t \partial \boldsymbol{\gamma}_2} \\ \Psi_2 \frac{\partial^2 C}{\partial \mathcal{G}_2 \partial \theta} \end{bmatrix},$$

$$\begin{bmatrix} \frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} \\ \frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} \\ \frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \Omega_1 \frac{\partial \xi_1}{\partial \boldsymbol{\gamma}_1} \\ \Omega_2 \frac{\partial \xi_2}{\partial \boldsymbol{\gamma}_2} \\ \frac{\partial C}{\partial \theta} \end{bmatrix}. \tag{C.17}$$

In addition, $\frac{\partial \xi_\nu(\boldsymbol{\gamma}_\nu)}{\partial \boldsymbol{\gamma}_\nu}$ and $\frac{\partial^2 \xi_\nu(\boldsymbol{\gamma}_\nu)}{\partial t \partial \boldsymbol{\gamma}_\nu}$ can be calculated using the following expressions

$$\frac{\partial \xi_\nu(\boldsymbol{\gamma}_\nu)}{\partial \boldsymbol{\gamma}_{\nu k_\nu}} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle}(t) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu}) & \text{otherwise,} \end{cases}$$

$$\frac{\partial^2 \xi_\nu(\boldsymbol{\gamma}_\nu)}{\partial t \partial \boldsymbol{\gamma}_{\nu k_\nu}} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle'}(t) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where, $\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(t)$, $\boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(t)$ and $\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu})$ can be defined as

$$\boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(t) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 03}\right) \\ \vdots \\ \mathcal{Q}_{\nu 0 J_{\nu 0}}(t) \exp\left(\gamma_{\nu 0 J_{\nu 0}}\right) \end{bmatrix}, \quad \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(z_{\nu k_\nu}) = \begin{bmatrix} \mathcal{Q}_{\nu k_\nu 1}(z_{\nu k_\nu}) \\ \mathcal{Q}_{\nu k_\nu 2}(z_{\nu k_\nu}) \\ \mathcal{Q}_{\nu k_\nu 3}(z_{\nu k_\nu}) \\ \vdots \\ \mathcal{Q}_{\nu k_\nu J_{\nu k_\nu}}(z_{\nu k_\nu}) \end{bmatrix},$$

$$\boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(t) = \begin{bmatrix} \sum_{j_{\nu 0}=1}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t) \\ \left[\sum_{j_{\nu 0}=2}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 02}\right) \\ \left[\sum_{j_{\nu 0}=3}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 03}\right) \\ \vdots \\ \mathcal{Q}'_{\nu 0 J_{\nu 0}}(t) \exp\left(\gamma_{\nu 0 J_{\nu 0}}\right) \end{bmatrix}.$$

Therefore, for $\nu = 1, 2$, $\omega = 1, 2$ and $\nu \neq \omega$, $\Delta_\nu \dfrac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu} + \Omega_\nu \dfrac{\partial \xi_\nu^2}{\partial t \partial \boldsymbol{\gamma}_\nu}$, $\Upsilon_\nu \dfrac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega}$ and $\Omega_\nu \dfrac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu}$ can be written as

$$\Delta_\nu \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu} + \Omega_\nu \frac{\partial \xi_\nu^2}{\partial t \partial \boldsymbol{\gamma}_\nu} = -\begin{bmatrix} \Delta_\nu \\ \Delta_\nu \boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(t) + \Omega_\nu \boldsymbol{\mathcal{Q}}^{\triangle'}_{\nu 0}(t) \\ \Delta_\nu \boldsymbol{\mathcal{Q}}_{\nu 1}(z_{\nu 1}) \\ \vdots \\ \Delta_\nu \boldsymbol{\mathcal{Q}}_{\nu K_\nu}(z_{\nu K_\nu}) \end{bmatrix}, \quad \Upsilon_\nu \frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega} = -\begin{bmatrix} \Upsilon_\nu \\ \Upsilon_\nu \boldsymbol{\mathcal{Q}}^{\triangle}_{\omega 0}(t) \\ \Upsilon_\nu \boldsymbol{\mathcal{Q}}_{\omega 1}(z_{\omega 1}) \\ \vdots \\ \Upsilon_\nu \boldsymbol{\mathcal{Q}}_{\omega K_\omega}(z_{\omega K_\omega}) \end{bmatrix},$$

$$\Omega_\nu \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu} = \begin{bmatrix} \Omega_\nu \\ \Omega_\nu \boldsymbol{\mathcal{Q}}^{\triangle}_{\nu 0}(t) \\ \Omega_\nu \boldsymbol{\mathcal{Q}}_{\nu 1}(z_{\nu 1}) \\ \vdots \\ \Omega_\nu \boldsymbol{\mathcal{Q}}_{\nu K_\nu}(z_{\nu K_\nu}) \end{bmatrix}. \tag{C.18}$$

C.2.2 Dependent censoring Hessian

The Hessian for the dependent censoring model can be written as

$$\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell_p(\boldsymbol{\vartheta}) = \boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}) - \boldsymbol{\Lambda}, \tag{C.19}$$

$$\begin{aligned}\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}) =& \sum_{i=1}^{n}\left[\frac{\delta_{1i}}{f_{1,0}(\boldsymbol{\vartheta})}\frac{\partial^2 f_{1,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\mathsf{T}} - \frac{\delta_{1i}}{f_{1,0}(\boldsymbol{\vartheta})^2}\frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}\frac{\partial f_{1,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}^\mathsf{T}\right] \\ &+ \sum_{i=1}^{n}\left[\frac{\delta_{2i}}{f_{0,1}(\boldsymbol{\vartheta})}\frac{\partial^2 f_{0,1}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\mathsf{T}} - \frac{\delta_{2i}}{f_{0,1}(\boldsymbol{\vartheta})^2}\frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}\frac{\partial f_{0,1}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}^\mathsf{T}\right] \\ &+ \sum_{i=1}^{n}\left[\frac{\delta_{3i}}{f_{0,0}(\boldsymbol{\vartheta})}\frac{\partial^2 f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\mathsf{T}} - \frac{\delta_{3i}}{f_{0,0}(\boldsymbol{\vartheta})^2}\frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}\frac{\partial f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}^\mathsf{T}\right].\end{aligned} \tag{C.20}$$

For $l_1, l_2 = 0, 1$, we can write

$$\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\mathsf{T}} = \begin{bmatrix} \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_1^\mathsf{T}} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_2^\mathsf{T}} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_1\partial\theta} \\ \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_2\partial\boldsymbol{\gamma}_1^\mathsf{T}} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_2\partial\boldsymbol{\gamma}_2^\mathsf{T}} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_2\partial\theta} \\ \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\theta\partial\boldsymbol{\gamma}_1} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\theta\partial\boldsymbol{\gamma}_2} & \frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})}{\partial\theta^2} \end{bmatrix}. \tag{C.21}$$

For $f_{0,0}(\boldsymbol{\vartheta})$, we have

$$\begin{aligned}\frac{\partial^2 f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\mathsf{T}} &= \left[\frac{\partial^2 C}{\partial\mathcal{G}_\nu^2}\left(\frac{\partial\mathcal{G}_\nu}{\partial\xi_\nu}\right)^2 + \frac{\partial C}{\partial\mathcal{G}_\nu}\frac{\partial\mathcal{G}_\nu^2}{\partial\xi_\nu^2}\right]\frac{\partial\xi_\nu}{\partial\boldsymbol{\gamma}_\nu}\left[\frac{\partial\xi_\nu}{\partial\boldsymbol{\gamma}_\nu}\right]^\mathsf{T} + \frac{\partial C}{\partial\mathcal{G}_\nu}\frac{\partial\mathcal{G}_\nu}{\partial\xi_\nu}\frac{\partial^2\xi_\nu}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\nu^\mathsf{T}} \\ \frac{\partial^2 f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_\nu\partial\boldsymbol{\gamma}_\omega^\mathsf{T}} &= \frac{\partial^2 C}{\partial\mathcal{G}_\nu\partial\mathcal{G}_\omega}\frac{\partial\mathcal{G}_\nu}{\partial\xi_\nu}\frac{\partial\mathcal{G}_\omega}{\partial\xi_\omega}\frac{\partial\xi_\nu}{\partial\boldsymbol{\gamma}_\nu}\left[\frac{\partial\xi_\omega}{\partial\boldsymbol{\gamma}_\omega}\right]^\mathsf{T} \\ \frac{\partial^2 f_{0,0}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\gamma}_\nu\partial\theta} &= \frac{\partial^2 C}{\partial\mathcal{G}_\nu\partial\theta}\frac{\partial\mathcal{G}_\nu}{\partial\xi_\nu}\frac{\partial\xi_\nu}{\partial\boldsymbol{\gamma}_\nu} \\ \frac{\partial^2 f_{0,0}(\boldsymbol{\vartheta})}{\partial\theta^2} &= \frac{\partial^2 C}{\partial\theta^2}\end{aligned} \tag{C.22}$$

For $f_{1,0}(\boldsymbol{\vartheta})$ and $f_{0,1}(\boldsymbol{\vartheta})$, let us write $f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2} = -\dfrac{\partial C}{\partial \mathcal{G}_\nu}\dfrac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\dfrac{\partial \xi_\nu}{\partial t}$. Then, for $\nu = 1, 2$, $\omega = 1, 2$ and $\nu \neq \omega$, we have

$$
\begin{aligned}
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\mathsf{T}} =& -\left\{ \left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu^3}\left(\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right)^3 \frac{\partial \xi_\nu}{\partial t} + 2\frac{\partial^2 C}{\partial \mathcal{G}_\nu^2}\frac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} + \frac{\partial^2 C}{\partial \mathcal{G}_\nu^2}\frac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} + \frac{\partial C}{\partial \mathcal{G}_\nu}\frac{\partial \mathcal{G}_\nu^3}{\partial \xi_\nu^3} \right] \right. \\
& \times \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu}\left[\frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu}\right]^\mathsf{T} \\
& + \left[\frac{\partial^2 C}{\partial \mathcal{G}_\nu^2}\left(\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right)^2 \frac{\partial \xi_\nu}{\partial t} + \frac{\partial C}{\partial \mathcal{G}_\nu}\frac{\partial^2 \mathcal{G}_\nu}{\partial \xi_\nu^2} \right] \frac{\partial^2 \xi_\nu}{\partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\mathsf{T}} + \left[\frac{\partial C}{\partial \mathcal{G}_\nu}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} \right] \frac{\partial^3 \xi_\nu}{\partial t\, \partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\mathsf{T}} \\
& \left. + \left[2\frac{\partial^2 C}{\partial \mathcal{G}_\nu^2}\left(\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right)^2 + \frac{\partial C}{\partial \mathcal{G}_\nu}\frac{\partial^2 \mathcal{G}_\nu}{\partial \xi_\nu^2} \right] \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu}\left[\frac{\partial^2 \xi_\nu}{\partial t \partial \boldsymbol{\gamma}_\nu}\right]^\mathsf{T} \right\} \\
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \boldsymbol{\gamma}_\omega \partial \boldsymbol{\gamma}_\omega^\mathsf{T}} =& -\left\{ \left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega^2}\left(\frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\right)^2 \frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} + \frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega}\frac{\partial \mathcal{G}_\omega^2}{\partial \xi_\omega^2}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} \right] \frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega}\left[\frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega}\right]^\mathsf{T} \right. \\
& \left. + \left[\frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega}\frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} \right] \frac{\partial^2 \xi_\omega}{\partial \boldsymbol{\gamma}_\omega \partial \boldsymbol{\gamma}_\omega^\mathsf{T}} \right\} \\
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\omega^\mathsf{T}} =& -\left\{ \left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega^2}\left(\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right)^2 \frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\frac{\partial \xi_\nu}{\partial t} + \frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega}\frac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2}\frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\frac{\partial \xi_\nu}{\partial t} \right] \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu}\left[\frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega}\right]^\mathsf{T} \right. \\
& \left. + \left[\frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega} \right] \frac{\partial \xi_\nu}{\partial t\ \partial \boldsymbol{\gamma}_\nu}\left[\frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega}\right]^\mathsf{T} \right\} \\
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \boldsymbol{\gamma}_\nu \partial \theta} =& -\left\{ \left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu^2 \partial \theta}\left(\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\right)^2 \frac{\partial \xi_\nu}{\partial t} + \frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \theta}\frac{\partial \mathcal{G}_\nu^2}{\partial \xi_\nu^2}\frac{\partial \xi_\nu}{\partial t} \right] \frac{\partial \xi_\nu}{\partial \boldsymbol{\gamma}_\nu} + \left[\frac{\partial^2 C}{\partial \mathcal{G}_\nu \partial \theta}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu} \right] \frac{\partial^2 \xi_\nu}{\partial\ t \partial \boldsymbol{\gamma}_\nu} \right\} \\
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \boldsymbol{\gamma}_\omega \partial \theta} =& -\left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu \partial \mathcal{G}_\omega \partial \theta}\frac{\partial \mathcal{G}_\omega}{\partial \xi_\omega}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} \right] \frac{\partial \xi_\omega}{\partial \boldsymbol{\gamma}_\omega} \\
\frac{\partial^2 f_{l_1,l_2}(\boldsymbol{\vartheta})_{l_1\neq l_2}}{\partial \theta^2} =& -\left[\frac{\partial^3 C}{\partial \mathcal{G}_\nu \partial \theta^2}\frac{\partial \mathcal{G}_\nu}{\partial \xi_\nu}\frac{\partial \xi_\nu}{\partial t} \right]
\end{aligned}
\tag{C.23}
$$

In addition, $\dfrac{\partial^2 \xi_\nu}{\partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\top}$ and $\dfrac{\partial^3 \xi_\nu}{\partial t\ \partial \boldsymbol{\gamma}_\nu \partial \boldsymbol{\gamma}_\nu^\top}$ can be obtained using the following equations

$$\frac{\partial^2 \xi_\nu}{\partial \boldsymbol{\gamma}_{\nu k_\nu} \partial \boldsymbol{\gamma}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(t) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu s_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

$$\frac{\partial^3 \xi_\nu}{\partial t \partial \boldsymbol{\gamma}_{\nu k_\nu} \boldsymbol{\gamma}_{\nu s_\nu}^\top} = \begin{cases} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(t) & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu s_\nu} = \boldsymbol{\gamma}_{\nu 0} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(t)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(t)$ can be calculated as

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle}(t) = \begin{cases} \dfrac{\partial^2 \xi_\nu}{\partial \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^2 \xi_\nu}{\partial \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{\mathcal{Q}}_{\nu 0}^{\triangle\triangle'}(t) = \begin{cases} \dfrac{\partial^3 \xi_\nu}{\partial t \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = \left[\sum_{j_{\nu 0}}^{J_{\nu 0}} \mathcal{Q}'_{\nu 0 j_{\nu 0}}(t)\right] \exp\left(\gamma_{\nu 0 j_{\nu 0}}\right) & \text{if } j = k \neq 1 \\ \dfrac{\partial^3 \xi_\nu}{\partial t \gamma_{\nu 0 j_{\nu 0}} \partial \gamma_{\nu 0 k_{\nu 0}}} = 0 & \text{otherwise.} \end{cases}$$

C.3 Proof of the asymptotic properties of $\hat{\vartheta}$

This section provides the proof of the asymptotic properties of $\hat{\boldsymbol{\vartheta}}$. This follows the same arguments to prove the asymptotic properties of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ (Theorems 1 and 2 in Appendix C). Let us write the log-likelihood function as

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log\left[f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_1}(y_i, \delta_{1i}, \delta_{1i}|\mathbf{z}_{1i}, \mathbf{z}_{2i}; \boldsymbol{\vartheta})\right]. \tag{C.24}$$

Let $\ell_n(\boldsymbol{\vartheta}) = n^{-1}\sum_{i=1}^{n} \log f(\mathbf{w}_i, \boldsymbol{\vartheta})$, where $f(\mathbf{w}_i, \boldsymbol{\vartheta}) = f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_1}(y_i, \delta_{1i}, \delta_{1i}|\mathbf{z}_{1i}, \mathbf{z}_{2i}; \boldsymbol{\vartheta})$ with $\mathbf{w}_i = (y_i, \mathbf{z}_{1i}^\top, \mathbf{z}_{2i}^\top)^\top \in \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, and $\mathbb{R}_+ = (0, \infty)$. Moreover, $\ell(\mathbf{w}_i; \boldsymbol{\vartheta}) = \log f(\mathbf{w}_i, \boldsymbol{\vartheta})$, $\boldsymbol{\nabla}_{\boldsymbol{\vartheta}} \ell(\mathbf{w}_i; \boldsymbol{\vartheta}) = \dfrac{\partial \ell(\mathbf{w}_i; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}$, $\boldsymbol{\nabla}_{\boldsymbol{\vartheta}} \ell_n(\boldsymbol{\vartheta}) = \dfrac{\partial \ell_n(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}$, $\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}} \ell(\mathbf{w}_i; \boldsymbol{\vartheta}) = \dfrac{\partial^2 \ell(\mathbf{w}_i; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top}$

and $\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell_n(\boldsymbol{\vartheta}) = \dfrac{\partial^2\ell_n(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}$. The penalised log-likelihood is $\ell_p(\boldsymbol{\vartheta}) = \ell_n(\boldsymbol{\vartheta}) - \frac{1}{2}\boldsymbol{\vartheta}^\top\Lambda\boldsymbol{\vartheta}$. Finally, $\boldsymbol{\vartheta}^0$ denotes the true vector of parameters.

Set of Assumptions 2 [Regularity conditions and vanishing penalty]

(C1) The true parameters vector $\boldsymbol{\vartheta}^0$ is in the interior of $\mathcal{S}_{\boldsymbol{\vartheta}} \subseteq \mathbb{R}^p$, which is a compact set, and $\mathcal{O}_{\boldsymbol{\vartheta}^0}$ is an open neighbourhood around $\boldsymbol{\vartheta}_0$.

(C2) For all $\mathbf{w}_i$, $f(\mathbf{w}_i;\boldsymbol{\vartheta})$ is continuous in $\boldsymbol{\vartheta}$. Also, $f(\mathbf{w}_i;\boldsymbol{\theta})$ is measurable in $\mathbf{w}_i$ for all $\boldsymbol{\vartheta} \in \mathcal{S}_{\boldsymbol{\vartheta}}$.

(C3) The model is identified. That is, for $l_1, l_2 = 0, 1$ and for almost every (t, z_1, z_2), $f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta}^0) = f_{y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t, l_1, l_2|z_1, z_2; \boldsymbol{\vartheta})$ implies $\boldsymbol{\vartheta}^0 = \boldsymbol{\vartheta}$, for any $\boldsymbol{\vartheta}$ in $\mathcal{S}_{\boldsymbol{\vartheta}}$. In addition, $\mathbb{E}\{\sup_{\boldsymbol{\vartheta}\in\mathcal{S}_{\boldsymbol{\vartheta}}}|\ell(\mathbf{w}_i;\boldsymbol{\vartheta})|\} < \infty$.

(C4) For all $\mathbf{w}_i$, $f(\mathbf{w}_i;\boldsymbol{\vartheta})$ is three times continuously differentiable in $\boldsymbol{\vartheta}$ in an open neighbourhood around $\boldsymbol{\vartheta}^0$. That is $f(\mathbf{w}_i;\boldsymbol{\vartheta}) \in \mathcal{C}^3(\mathcal{O}_{\boldsymbol{\vartheta}^0})$.

(C5) $\int \sup_{\boldsymbol{\vartheta}\in\mathcal{O}_{\boldsymbol{\vartheta}^0}} \|\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta})\|\, d\mathbf{w}_i < \infty$ and $\int \sup_{\boldsymbol{\vartheta}\in\mathcal{O}_{\boldsymbol{\vartheta}^0}} \|\boldsymbol{\nabla}_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta})\|\, d\mathbf{w}_i < \infty$.

(C6) For $\boldsymbol{\vartheta} \in \mathcal{O}_{\boldsymbol{\vartheta}^0}$, $\mathcal{I}(\boldsymbol{\vartheta}^0) = \mathrm{Cov}\{\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta})\} = \mathbb{E}\{\{\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta}^0) - \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta}^0)]\}$ $\{\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta}^0) - \mathbb{E}[\boldsymbol{\nabla}_{\boldsymbol{\vartheta}}\ell(\mathbf{w}_i;\boldsymbol{\vartheta}^0)]\}^\top\}$ exists and is positive-definite.

(C7) For all $1 \leq e, f, h \leq p$, there exists a function $\boldsymbol{\phi} : \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \longrightarrow \mathbb{R}$ such that, for $\boldsymbol{\vartheta} \in \mathcal{O}_{\boldsymbol{\vartheta}^0}$ and $\mathbf{w}_i \in \mathbb{R}_+ \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, $\left|\dfrac{\partial^3\ell(\mathbf{w}_i;\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}_e\partial\boldsymbol{\vartheta}_f\partial\boldsymbol{\vartheta}_h}\right| \leq \boldsymbol{\phi}(\mathbf{w}_i)$, with $\mathbb{E}[\boldsymbol{\phi}(\mathbf{w}_i)] < \infty$.

(C8) The penalties vanish as the sample size n goes to infinite. That is $\boldsymbol{\lambda} = o(n^{-1/2})\mathbf{1}$.

Theorem 6 (Asymptotic properties of the DCPMLE estimator).

Proof. See the proof of Theorem 1 in Appendix C. □

C.4 Independent censoring log-likelihood function

In this section we show how the sub-densities functions are built when it is assumed that the censoring mechanism is stochastically independent. Then we will use these sub-densities to construct the log-likelihood function.

Suppose that $Y = \min\{T_1, T_2, T_3\} \in \mathbb{R}^+$, and the censoring indicators $\delta_1 = I\{Y = T_1\}$ and $\delta_2 = I\{Y = T_2\}$ are observed. If we assume that the censoring is independent, the sub-density function $f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(\cdot,\cdot,\cdot|z_1,z_2;\boldsymbol{\gamma})$ of (Y,δ_1,δ_2) given $(\mathbf{z}_1,\mathbf{z}_2) = (z_1,z_2)$ and $\boldsymbol{\gamma}$, when $Y = T_1$, can be written as

$$\begin{aligned} f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,1,0|z_1,z_2;\boldsymbol{\gamma}) &= P(Y=t,\delta_1=1,\delta_2=0|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}),\\ &= P(T_1=t,T_2>t,T_3>t|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}),\\ &= f_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1)P(T_2>t|z_2;\boldsymbol{\gamma}_2)P(T_3>t),\\ &= f_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1)S_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_2)S_{T_3}(t). \end{aligned} \tag{C.25}$$

Similarly, when $Y = T_2$, we obtain

$$\begin{aligned} f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,1|z_1,z_2;\boldsymbol{\gamma}) &= P(Y=t,\delta_1=0,\delta_2=1|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}),\\ &= P(T_1>t,T_2=t,T_3>t|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}),\\ &= f_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_1)P(T_1>t|z_1;\boldsymbol{\gamma}_2)P(T_3>t),\\ &= f_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_1)S_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_2)S_{T_3}(t). \end{aligned} \tag{C.26}$$

Finally, if $Y = T_3$, then

$$\begin{aligned} f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(t,0,0|z_1,z_2;\boldsymbol{\gamma}) &= P(Y=t,\delta_1=0,\delta_2=0|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}), \\ &= P(T_1>y,T_2>t,T_3=t|\mathbf{z}_1=z_1,\mathbf{z}_2=z_2;\boldsymbol{\gamma}), \\ &= P(T_1>t|z_1;\boldsymbol{\gamma}_1)P(T_2>t|z_2;\boldsymbol{\gamma}_2)f_{T_3}(t), \\ &= S_{T_1|\mathbf{z}_1}(t|z_1;\boldsymbol{\gamma}_1)S_{T_2|\mathbf{z}_2}(t|z_2;\boldsymbol{\gamma}_2)f_{T_3}(t). \end{aligned} \tag{C.27}$$

On the other hand, the joint density of $(Y,\delta_1,\delta_2,\mathbf{z}_1,\mathbf{z}_2)$ can be written as $f_{Y,\delta_1,\delta_2\mathbf{z}_1,\mathbf{z}_2} = f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}f_{\mathbf{z}_1,\mathbf{z}_2}$. Since $f_{\mathbf{z}_1,\mathbf{z}_2}$ does not involve the model parameters, the likelihood function can be formulated using the conditional density $f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}$. Let us assume that the data consist of n random *i.i.d.* replications $\{(y_i,\delta_{1i},\delta_{2i},\mathbf{z}_{1i},\mathbf{z}_{2i})\}_{i=1}^n$ of $(Y,\delta_1,\delta_2,\mathbf{z}_1,\mathbf{z}_2)$. This allows us to write the likelihood function for $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^\mathsf{T},\boldsymbol{\gamma}_2^\mathsf{T}\right)^\mathsf{T}$ as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) = \prod_{i=1}^n \Big\{ &\Big[f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(y_i,1,0|\mathbf{z}_{1i},\mathbf{z}_{2i};\boldsymbol{\gamma})\Big]^{\delta_{1i}} \Big[f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(y_i,0,1|\mathbf{z}_{1i},\mathbf{z}_{2i};\boldsymbol{\gamma})\Big]^{\delta_{2i}} \\ &\times \Big[f_{Y,\delta_1,\delta_2|\mathbf{z}_1,\mathbf{z}_2}(y_i,0,0|\mathbf{z}_{1i},\mathbf{z}_{2i};\boldsymbol{\gamma})\Big]^{1-\delta_{1i}-\delta_{2i}}\Big\}. \end{aligned}$$

Using (C.25), (C.26) and (C.27), we obtain

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) = \prod_{i=1}^n \Big\{ &\Big[f_{T_1|\mathbf{z}_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\gamma}_1)S_{T_2|\mathbf{z}_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\gamma}_2)S_{T_3}(y_i)\Big]^{\delta_{1i}} \\ &\times \Big[f_{T_2|\mathbf{z}_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\gamma}_2)S_{T_1|\mathbf{z}_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\gamma}_1)S_{T_3}(y_i)\Big]^{\delta_{2i}} \\ &\times \Big[S_{T_1|\mathbf{z}_1}(y_i|\mathbf{z}_{1i};\boldsymbol{\gamma}_1)S_{T_2|\mathbf{z}_2}(y_i|\mathbf{z}_{2i};\boldsymbol{\gamma}_2)f_{T_3}(y_i)\Big]^{1-\delta_{1i}-\delta_{2i}}\Big\}. \end{aligned} \tag{C.28}$$

Since $\hbar_{T_\nu|\mathbf{z}_\nu}(y_i|\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu) = \dfrac{f_{T_\nu|\mathbf{z}_\nu}(y_i|\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)}{S_{T_\nu|\mathbf{z}_\nu}(y_i|\mathbf{z}_{\nu i};\boldsymbol{\gamma}_\nu)}$ and $S_{T_3}(y_i)$ does not involve $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$,

the logarithm of equation (C.28) can be written as

$$\begin{aligned}\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\Bigg\{ & (\delta_{1i}+\delta_{2i})\log\mathcal{G}_1\left[\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)\right] + \delta_{1i}\log\left[-\frac{\mathcal{G}_1'[\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)]}{\mathcal{G}_1[\xi_1(y_i,\mathbf{z}_{\nu i};\boldsymbol{\gamma}_1)]}\frac{\partial\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)}{\partial y_i}\right] \\ & + (1-\delta_{1i}-\delta_{2i})\log\mathcal{G}_1\left[\xi_1(y_i,\mathbf{z}_1;\boldsymbol{\gamma}_1)\right] \\ & + (\delta_{1i}+\delta_{2i})\log\mathcal{G}_2\left[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)\right] + \delta_{2i}\log\left[-\frac{\mathcal{G}_2'[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)]}{\mathcal{G}_2[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)]}\frac{\partial\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)}{\partial y_i}\right] \\ & + (1-\delta_{1i}-\delta_{2i})\log\mathcal{G}_2\left[\xi_2(y_i,\mathbf{z}_2;\boldsymbol{\gamma}_2)\right]\Bigg\}.\end{aligned} \tag{C.29}$$

Finally, if (C.29) is rearranged, then

$$\begin{aligned}\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\Bigg\{ & \log\mathcal{G}_1\left[\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)\right] + \delta_{1i}\log\left[-\frac{\mathcal{G}_1'[\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)]}{\mathcal{G}_1[\xi_1(y_i,\mathbf{z}_{\nu i};\boldsymbol{\gamma}_1)]}\frac{\partial\xi_1(y_i,\mathbf{z}_{1i};\boldsymbol{\gamma}_1)}{\partial y_i}\right] \\ & \log\mathcal{G}_2\left[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)\right] + \delta_{2i}\log\left[-\frac{\mathcal{G}_2'[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)]}{\mathcal{G}_2[\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)]}\frac{\partial\xi_2(y_i,\mathbf{z}_{2i};\boldsymbol{\gamma}_2)}{\partial y_i}\right]\Bigg\}.\end{aligned}$$

C.5 Software details: `gjrm()` function

The models proposed in Section 5.6 can be employed via the `gjrm()` function in the `R` package `GJRM` (Marra & Radice, 2020b). As an example, consider the following call

```
eq1 <- u ~ s(u, bs = "mpi") + z1 + s(z2),

eq2 <- u ~ s(u, bs = "mpi") + z1 + s(z2),

out <- gjrm(list(eq1, eq2), data = data, surv = TRUE,

margins = c("PH", "PH"),cens1 = delta1, cens2 = delta2,

cens3 = 1-delta1-delta2, Model = "B",

BivD = "N", dep.cens = TRUE, gamlssfit = TRUE),
```

where `eq1` and `eq2` are the two additive predictors of the dependent censoring model. In these equations, `s(u, bs = "mpi")` represents the monotonic P-spline function which models a transformation of the baseline survival function. As for `s(z2)`, the default is `bs = "tp"` (penalized low rank thin plate spline) with `k = 10` (number of basis functions) and `m = 2` (order of derivatives). However, argument `bs` can also be set to, for example, `cr` (penalized cubic regression spline), `ps` (P-spline) and `mrf` (Markov random field), to name but a few. In the `gjrm` function, `surv = TRUE` indicates that a survival model is fitted. The arguments `margin ="PH"` and `margin2 ="PH"` specify the link functions for the survival and censoring times, respectively. Table 4.1 shows the possible choices for the links that have been implemented for this article. In this example, we specify the proportional hazard link (`"PH"`) for the two equations. The arguments `cens1 = delta1`, `cens2 = delta2` and `cens3 =1-delta1-delta2` are binary censoring indicators. In particular, `cens1 = delta1` has to be equal to 1 if the event occurred and 0 otherwise. Similarly, `cens2 = delta` has to be equal to 1 if the dependent censoring occurred and 0 otherwise. When both `cens1 = delta1` and `cens1 = delta2` are equal to zero, the argument `cens3 =1-delta1-delta2` captures the administrative censoring observations. The option `Model = "B"` indicates that the model is bivariate. Argument `BivD = "N"` represents the type of bivariate survival copula employed. The choices considered in this work are Normal (`"N"`), Frank (`"F"`), Clayton (`"C0"`), Joe (`"J0"`), Student (`"T"`), Farlie-Gumbel-Morgenstern, (`"FGM"`), Ali-Mikhail-Haq (`"AMH"`), Plackett (`"PL"`), Gumbel (`"G0"`) and Student-t (`"T"`). Moreover, if `dep.cens = "TRUE"` then the dependence censored model is employed. Finally, if `gamlssfit = TRUE` then `gamlss` univariate models are also fitted. This is useful for obtaining starting values, for instance.

C.6 Additional simulation results for DGP1 to DGP14

	DCPML: Survival Functions (S_{10})				ICPMLE: Survival Functions (S_{10})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
1	0.006	0.003	0.024	0.013	0.020	0.010	0.037	0.016
2	0.005	0.002	0.025	0.012	0.018	0.020	0.037	0.027
3	0.005	0.003	0.024	0.012	0.021	0.021	0.037	0.027
4	0.005	0.002	0.025	0.013	0.020	0.022	0.037	0.028
5	0.005	0.002	0.024	0.013	0.017	0.019	0.035	0.026
6	0.004	0.002	0.024	0.013	0.018	0.019	0.035	0.026
	DCPML: Hazard Functions (h_{10})				ICPMLE: Hazard Functions (h_{10})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
1	0.039	0.022	0.100	0.052	0.071	0.085	0.115	0.098
2	0.038	0.022	0.119	0.053	0.063	0.078	0.115	0.098
3	0.036	0.022	0.097	0.051	0.069	0.082	0.107	0.093
4	0.037	0.021	0.101	0.054	0.067	0.082	0.112	0.095
5	0.033	0.021	0.096	0.052	0.059	0.071	0.118	0.098
6	0.038	0.023	0.126	0.053	0.062	0.075	0.107	0.091

Table C.1: Bias and root mean squared error (RMSE) for the hazard and survival functions when DCPMLE and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGPs 1` to `6` defined in Table 5.2. Bias and RMSE for the smooth terms are calculated, respectively, as $n_s^{-1} \sum_{i=1}^{n_s} |\bar{\hat{s}}_i - s_i|$ and $n_s^{-1} \sum_{i=1}^{n_s} \sqrt{n_{rep}^{-1} \sum_{rep=1}^{n_{rep}} (\hat{s}_{rep,i} - s_i)^2}$, where $\bar{\hat{s}}_i = n_{rep}^{-1} \sum_{rep=1}^{n_{rep}} \hat{s}_{rep,i}$, n_s is the number of equally spaced fixed values in the $(0, 8)$ or $(0, 1)$ range, and n_{rep} is the number of simulation replicates. In this case, $n_s = 200$ and $n_{rep} = 1000$. The bias for the smooth terms is based on absolute differences in order to avoid compensating effects when taking the sum.

DCPMLE: Kendall Tau (τ)				
	Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000
1	0.051	0.015	0.095	0.040
2	0.052	0.019	0.108	0.043
3	0.063	0.031	0.102	0.048
4	0.053	0.023	0.137	0.059
5	0.066	0.035	0.169	0.075
6	0.086	0.065	0.199	0.105

Table C.2: Bias and root mean squared error (RMSE) for the Kendall Tau obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGPs 1` to `6` defined in Table 5.2.

	DCPML: Parametric Effects (γ_{11})				ICPMLE: Parametric Effects (γ_{11})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
7	-0.019	0.004	0.144	0.073	-0.082	-0.069	0.161	0.098
8	-0.060	0.005	0.150	0.065	-0.127	-0.087	0.200	0.108
9	0.122	0.067	0.325	0.153	-1.588	-1.546	1.634	1.556
10	0.029	0.024	0.149	0.069	-0.293	-0.278	0.330	0.287
11	0.095	0.051	0.347	0.160	-1.210	-1.163	1.248	1.173
12	0.013	0.027	0.137	0.074	-0.296	-0.282	0.328	0.291
13	0.001	0.006	0.111	0.054	-0.035	-0.032	0.116	0.063
14	0.010	0.015	0.116	0.058	-0.050	-0.044	0.124	0.072
	DCPML: Smooth Effects (s_{11})				ICPMLE: Smooth Effects (s_{11})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
7	0.030	0.017	0.139	0.071	0.033	0.023	0.140	0.073
8	0.028	0.017	0.119	0.071	0.032	0.026	0.120	0.075
9	0.028	0.018	0.151	0.077	0.042	0.039	0.164	0.091
10	0.029	0.019	0.129	0.067	0.041	0.051	0.147	0.090
11	0.032	0.019	0.153	0.077	0.048	0.045	0.161	0.092
12	0.028	0.019	0.130	0.036	0.028	0.036	0.140	0.081
13	0.024	0.016	0.126	0.064	0.018	0.010	0.125	0.065
14	0.031	0.019	0.126	0.068	0.020	0.014	0.125	0.067

Table C.3: Bias and root mean squared error (RMSE) for parametric and smooth effects when DCPMLE and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGPs 7` to `14` defined in Table 5.2. Further details are given in the caption of Table C.1.

	DCPML: Survival Functions (S_{10})				ICPMLE: Survival Functions (S_{10})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
7	0.003	0.003	0.024	0.012	0.016	0.017	0.035	0.025
8	0.004	0.004	0.022	0.012	0.022	0.020	0.037	0.027
9	0.005	0.002	0.025	0.013	0.023	0.021	0.037	0.028
10	0.005	0.002	0.024	0.012	0.021	0.020	0.037	0.027
11	0.005	0.002	0.025	0.013	0.020	0.019	0.037	0.027
12	0.004	0.002	0.023	0.012	0.021	0.021	0.037	0.028
13	0.003	0.003	0.023	0.012	0.021	0.021	0.036	0.027
14	0.004	0.003	0.024	0.012	0.018	0.019	0.035	0.026
	DCPML: Hazard Functions (h_{10})				ICPMLE: Hazard Functions (h_{10})			
	Bias		RMSE		Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000	n=500	n=2000	n=500	n=2000
7	0.030	0.019	0.138	0.049	0.056	0.065	0.128	0.092
8	0.031	0.022	0.069	0.052	0.080	0.072	0.103	0.097
9	0.039	0.022	0.148	0.053	0.054	0.069	0.124	0.089
10	0.032	0.021	0.093	0.050	0.065	0.074	0.111	0.095
11	0.039	0.021	0.108	0.052	0.060	0.053	0.114	0.091
12	0.033	0.020	0.920	0.050	0.073	0.082	0.105	0.096
13	0.030	0.022	0.890	0.050	0.072	0.080	0.100	0.091
14	0.036	0.024	0.102	0.050	0.062	0.073	0.101	0.089

Table C.4: Bias and root mean squared error (RMSE) for the hazard and survival functions when DCPMLE and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGPs 7` to `14` defined in Table 5.2. Further details are given in the caption of Table C.1.

DCPMLE: Kendall Tau (τ)				
	Bias		RMSE	
DGP	n=500	n=2000	n=500	n=2000
7	-0.026	-0.004	0.110	0.057
8	-0.031	0.008	0.142	0.048
9	0.068	0.027	0.115	0.049
10	0.064	0.033	0.109	0.050
11	0.052	0.018	0.123	0.049
12	0.032	0.028	0.100	0.046
13	0.012	0.025	0.169	0.061
14	0.074	0.076	0.226	0.117

Table C.5: Bias and root mean squared error (RMSE) for the Kendall Tau obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGPs 7` to `14` defined in Table 5.2.

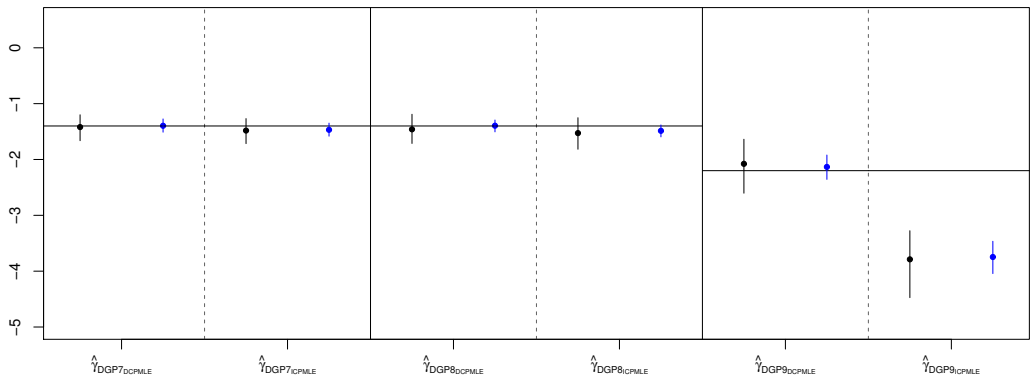

Figure C.1: Parametric effects (γ_{11}) when DCPMLE ($\tau = 0.7$) and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP7` (FGM copula), `DGP8` (AMH copula) and `DGP9` (Gumbel copula) defined in Table 5.2. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 2000$ are given in blue.

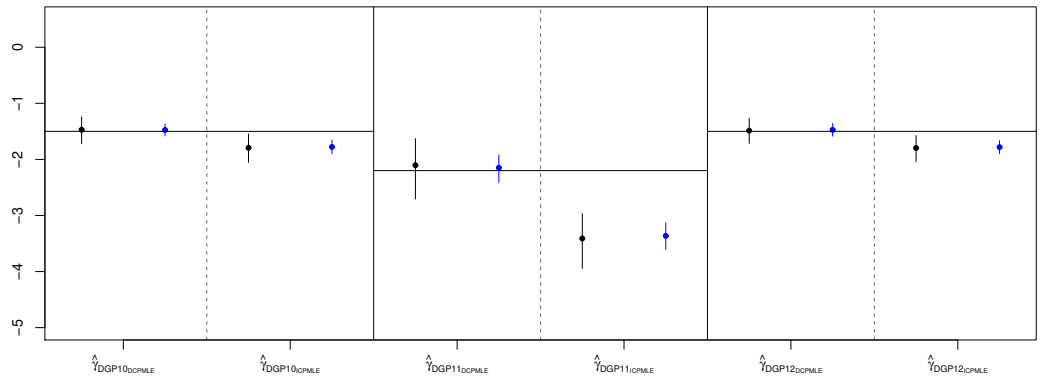

Figure C.2: Parametric effects (γ_{11}) when DCPMLE ($\tau = 0.7$) and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP10` (Joe copula), `DGP11` (Plakett copula) and `DGP12` (Student copula) defined in Table 5.2. Further details are given in the caption of Figure C.1.

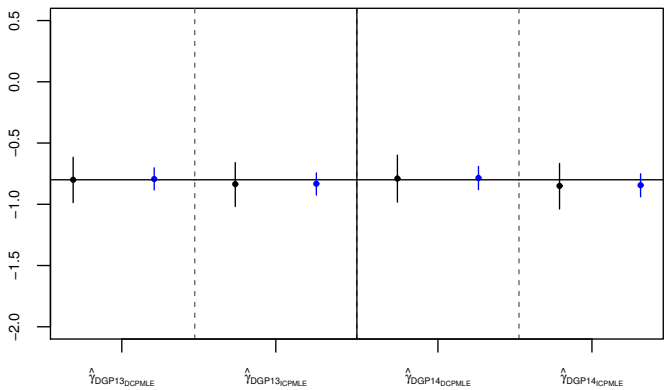

Figure C.3: Parametric effects (γ_{11}) when DCPMLE and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP13` (Gaussian copula and $\tau = 0.7$) and `DGP14` (Gaussian copula and $\tau = 0.4$) defined in Table 5.2. Further details are given in the caption of Figure C.1.

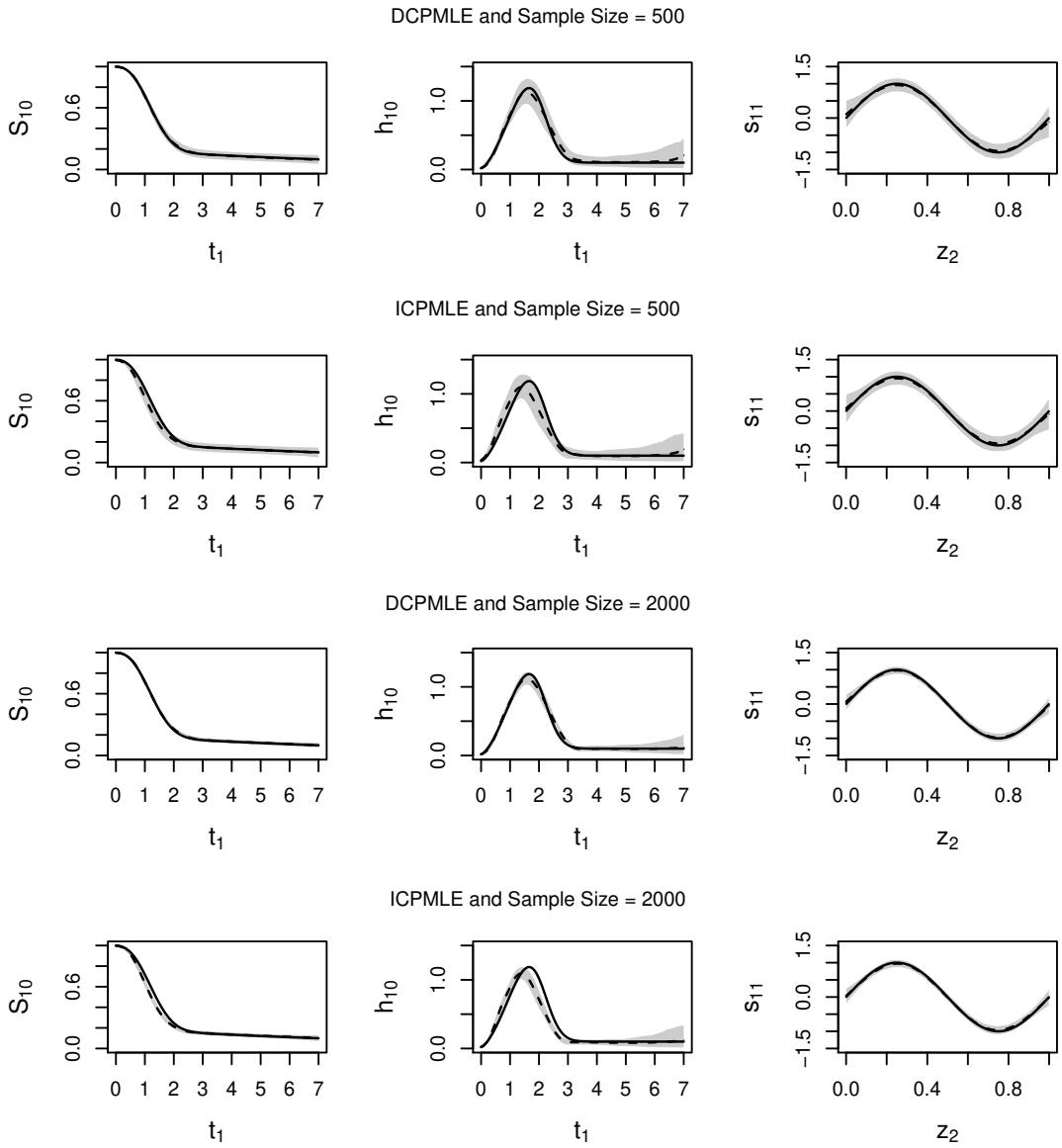

Figure C.4: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP7` (Table 5.2). The results in the first and second rows refer to $n = 500$, whereas that in the third and fourth rows to $n = 2000$. True functions are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting from 5% and 95% quantiles by shaded areas.

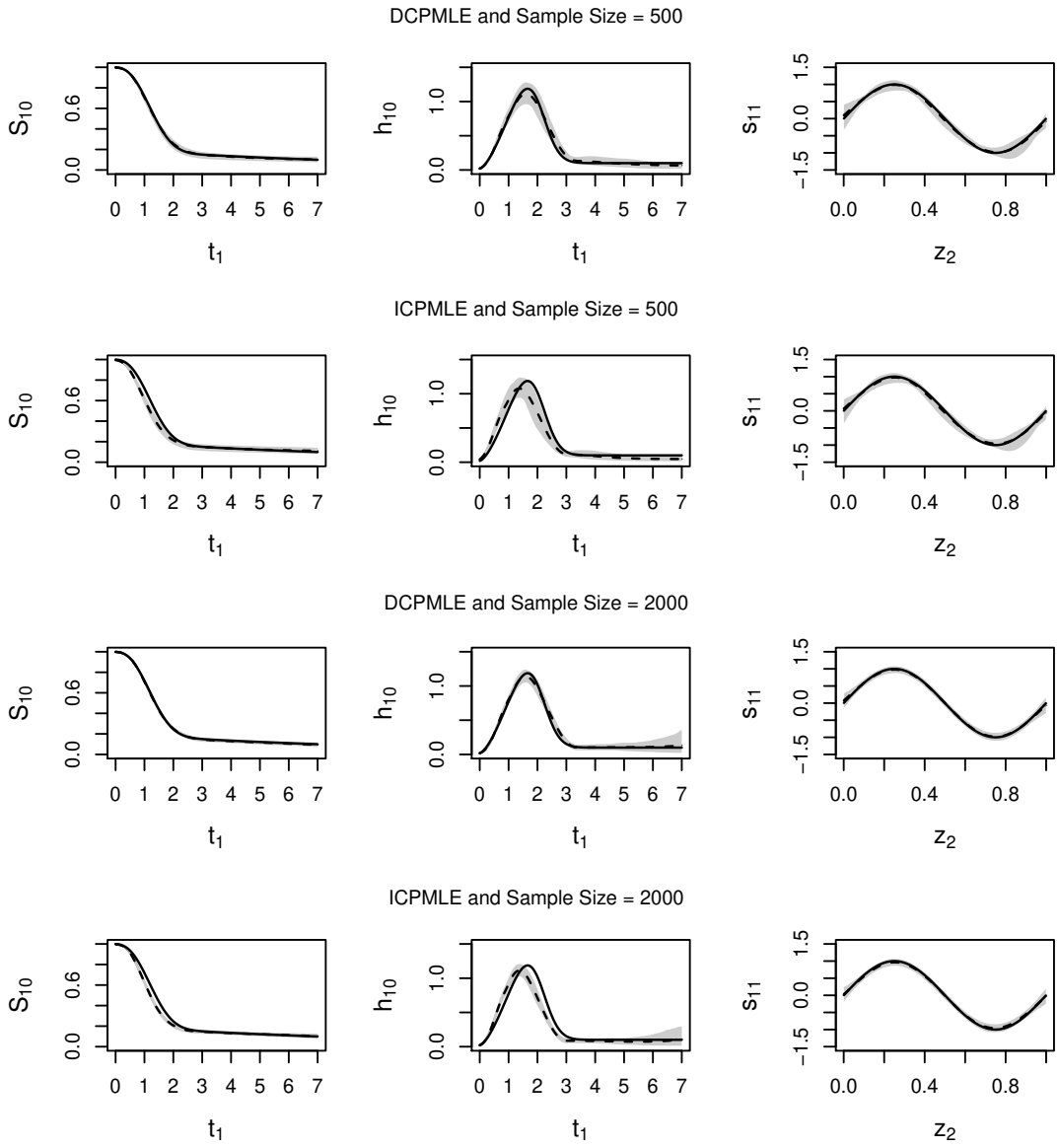

Figure C.5: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP8` (Table 5.2). Further details are given in the caption of Figure C.4.

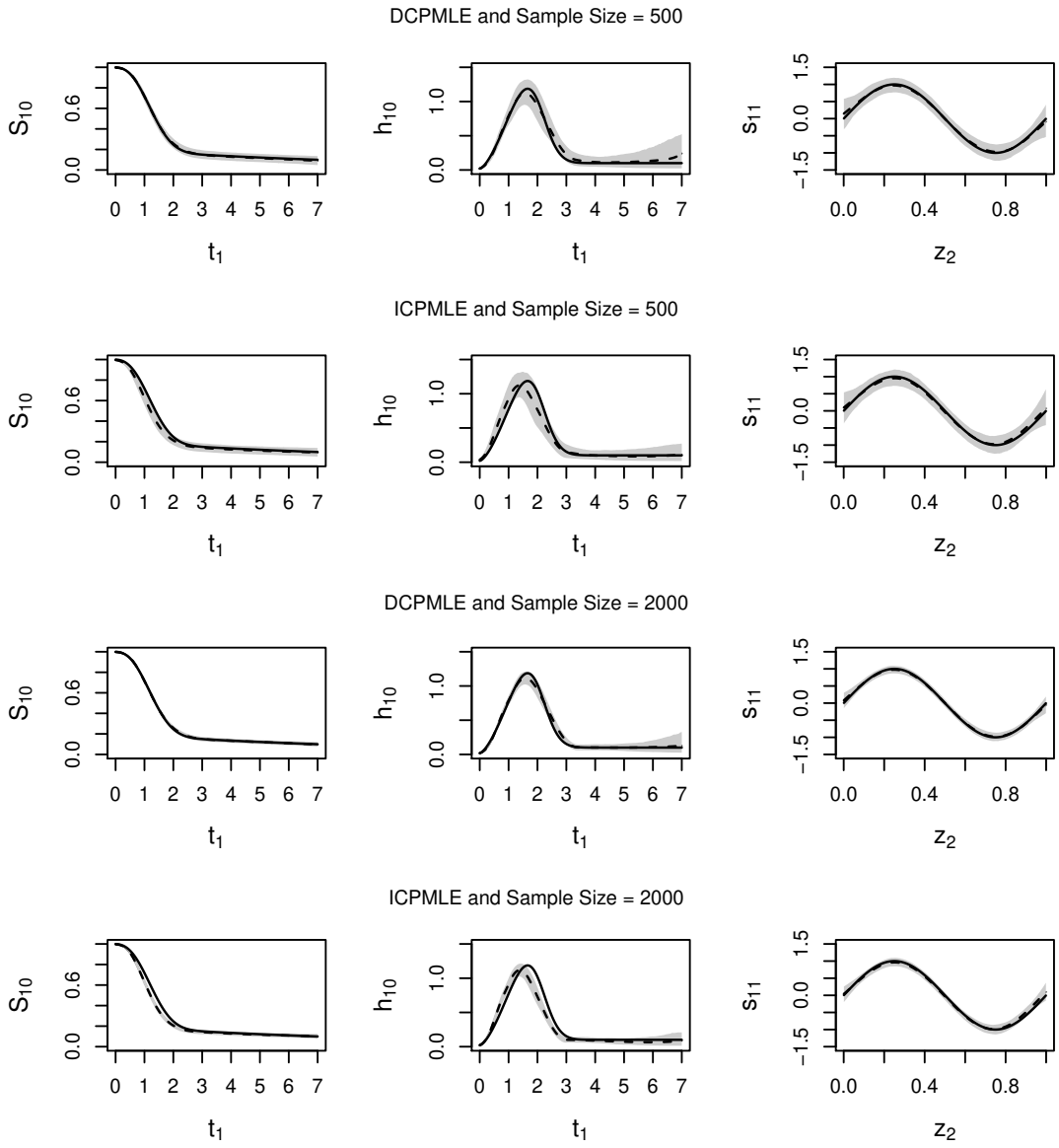

Figure C.6: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP9` (Table 5.2). Further details are given in the caption of Figure C.4.

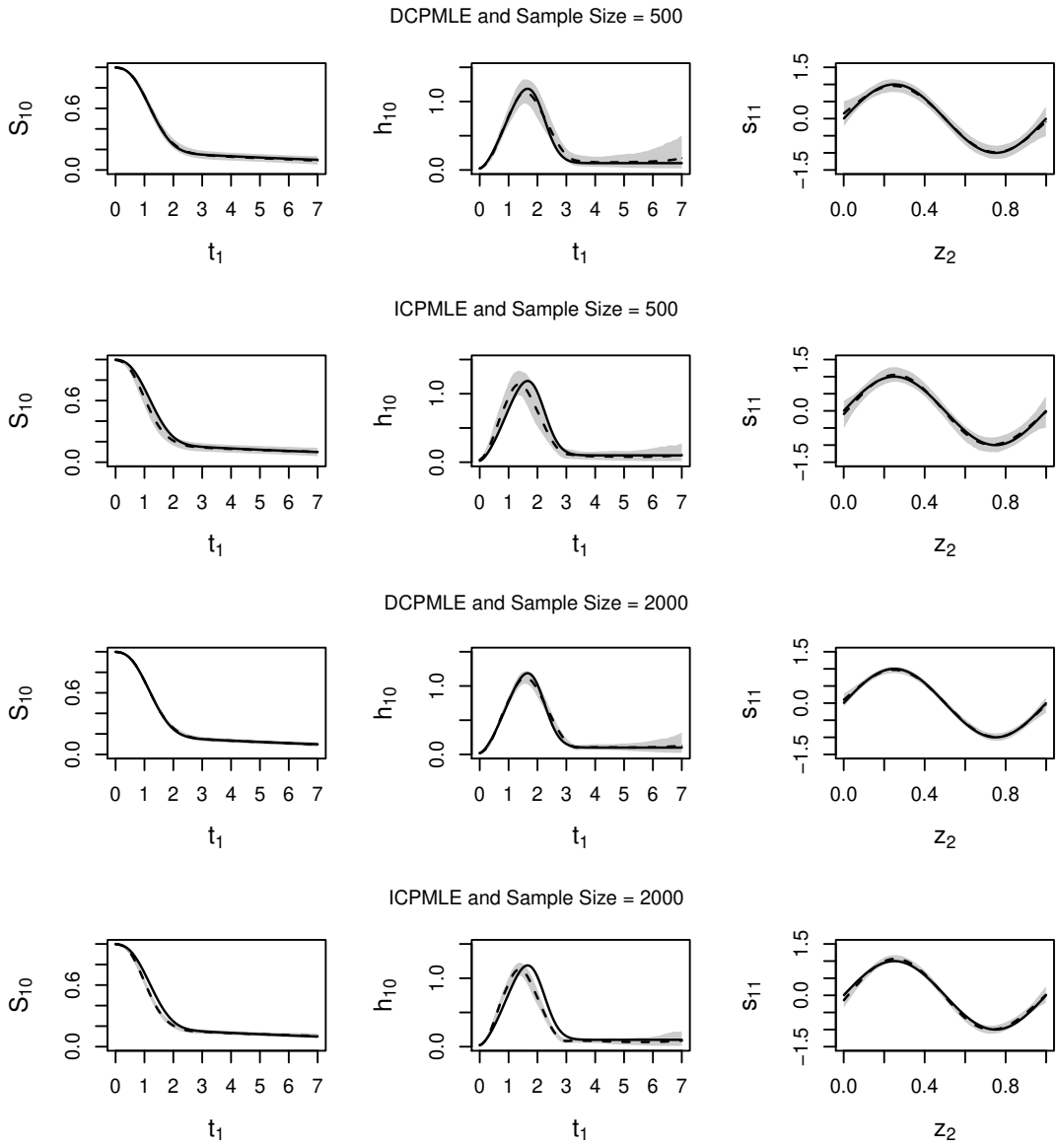

Figure C.7: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP10` (Table 5.2). Further details are given in the caption of Figure C.4.

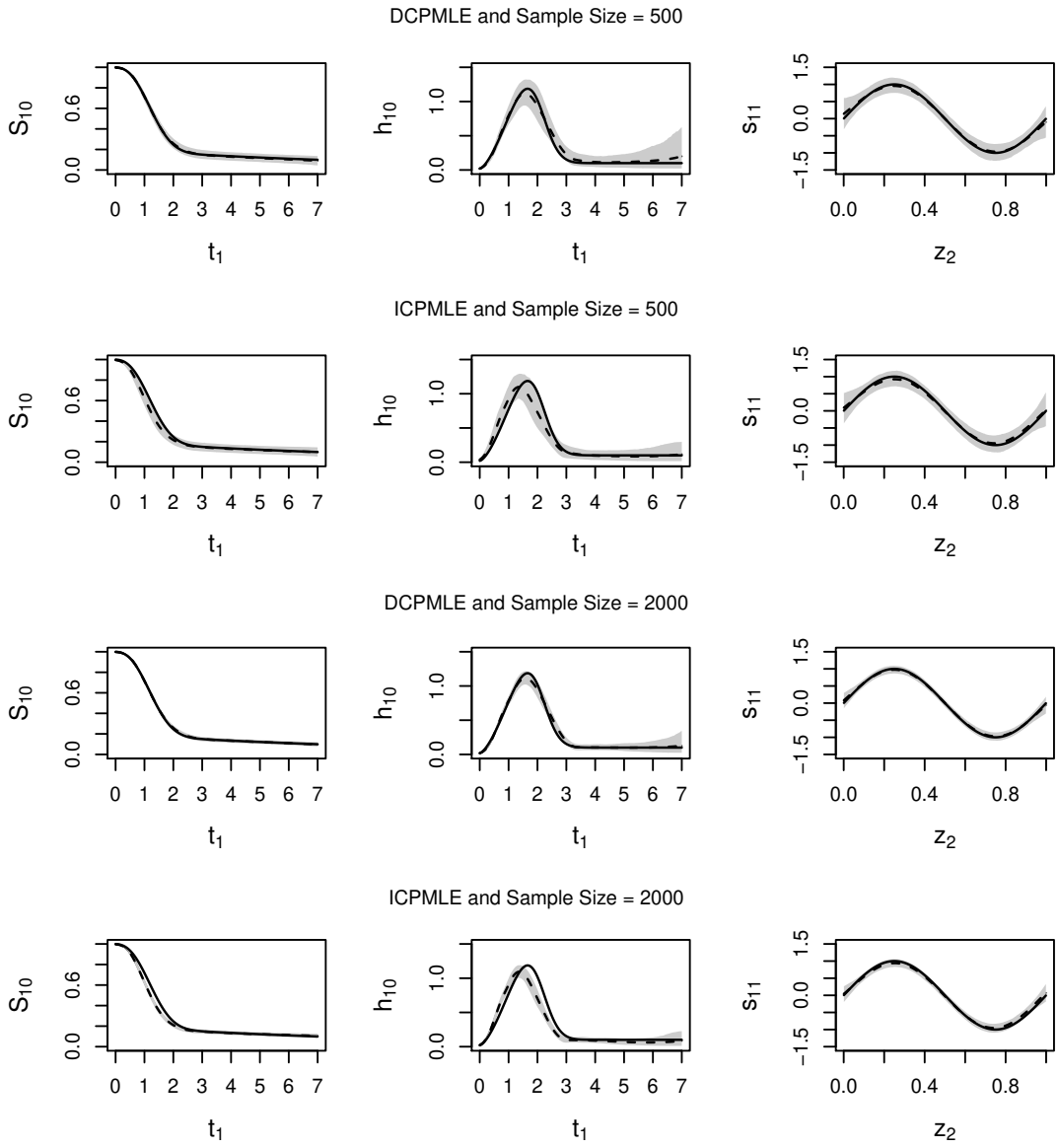

Figure C.8: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP11` (Table 5.2). Further details are given in the caption of Figure C.4.

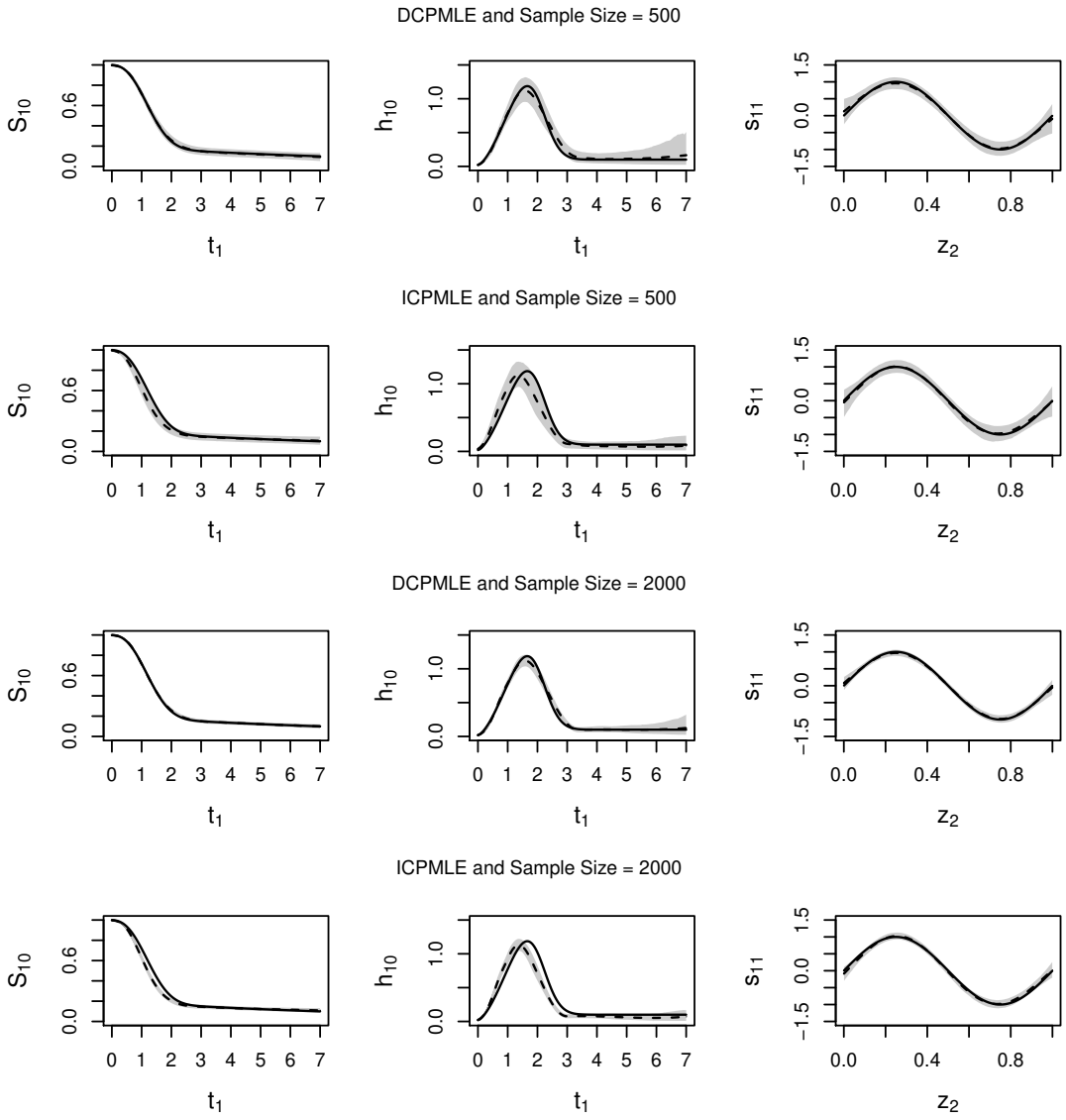

Figure C.9: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP12` (Table 5.2). Further details are given in the caption of Figure C.4.

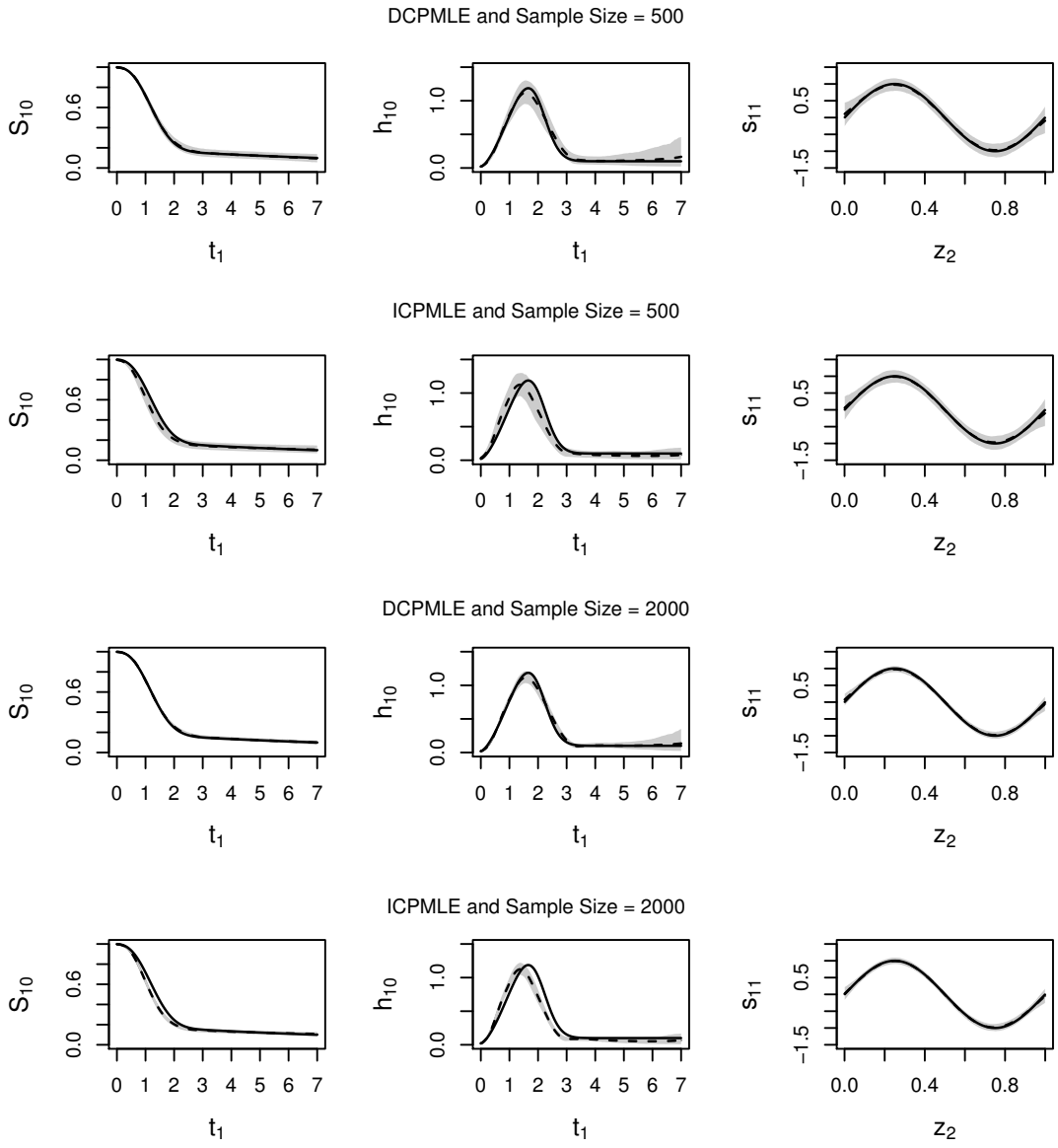

Figure C.10: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP13` (Table 5.2). Further details are given in the caption of Figure C.4.

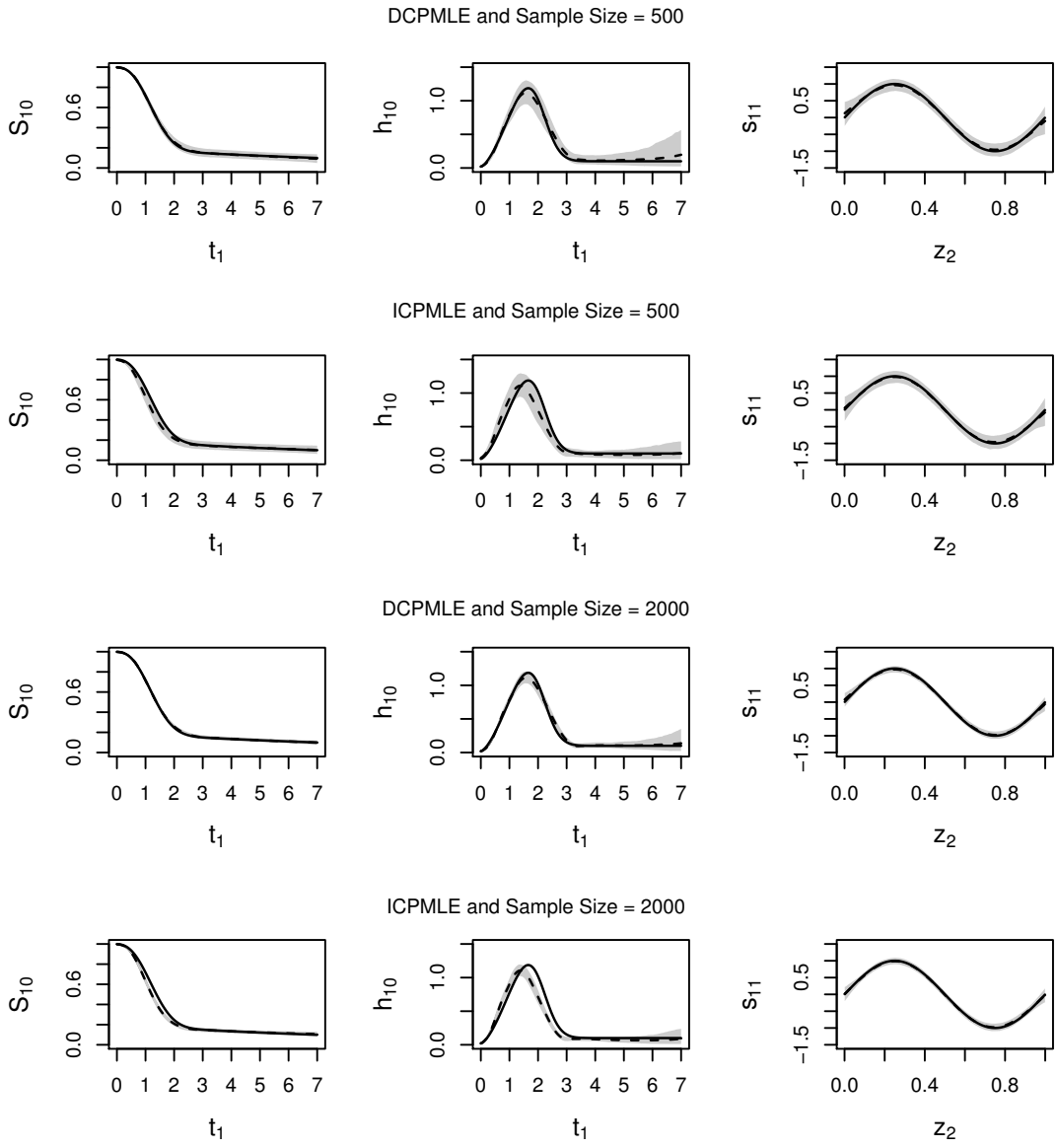

Figure C.11: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP14` (Table 5.2). Further details are given in the caption of Figure C.4.

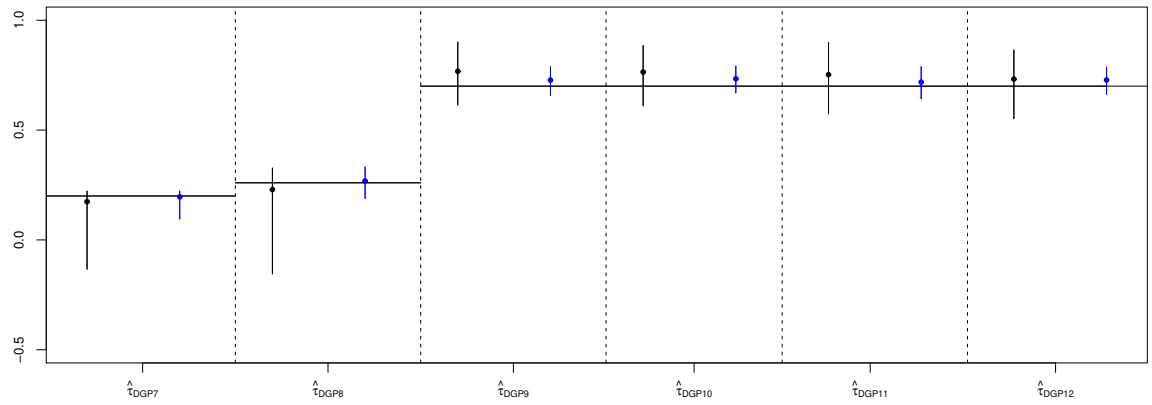

Figure C.12: Kendall Tau coefficient ($\tau = 0.7$) estimates obtained when DCPMLE is fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP7` (FGM copula), `DGP8` (AMH copula), `DGP9` (Gumbel copula), `DGP10` (Joe copula), `DGP11` (Plakett copula) and `DGP12` (Student copula) defined in Table 5.2. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 500$, whereas those for $n = 2000$ are given in blue.

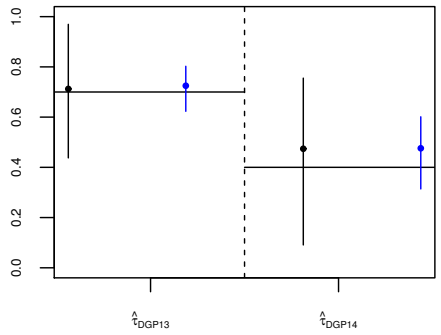

Figure C.13: Kendall Tau coefficient estimates obtained when DCPMLE is fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP13` (Gaussian copula and $\tau = 0.7$) and `DGP14` (Gaussian copula and $\tau = 0.4$) defined in Table 5.2. Further details are given in the caption of Figure C.12.

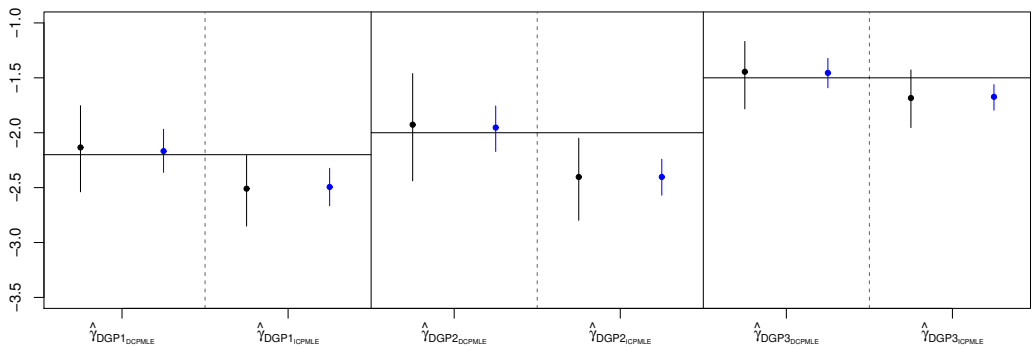

Figure C.14: Parametric effects (γ_{11}) when DCPMLE ($\tau = 0.4$) and ICPMLE are fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP4` (Clayton copula), `DGP5` (Frank copula) and `DGP6` (Gaussian copula) defined in Table 5.2. Further details are given in the caption of Figure C.1.

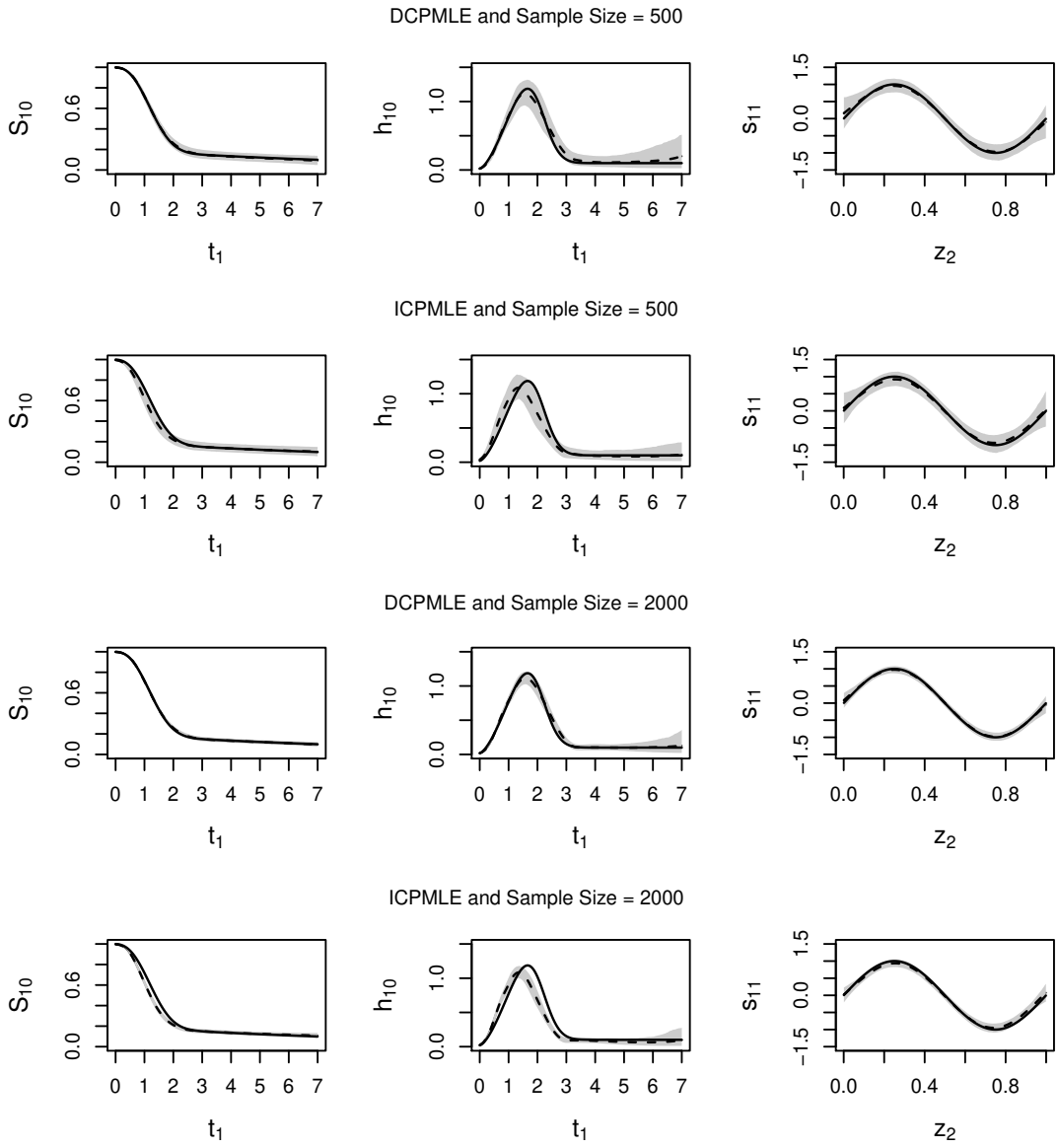

Figure C.15: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP2` (Table 5.2). Further details are given in the caption of Figure C.4.

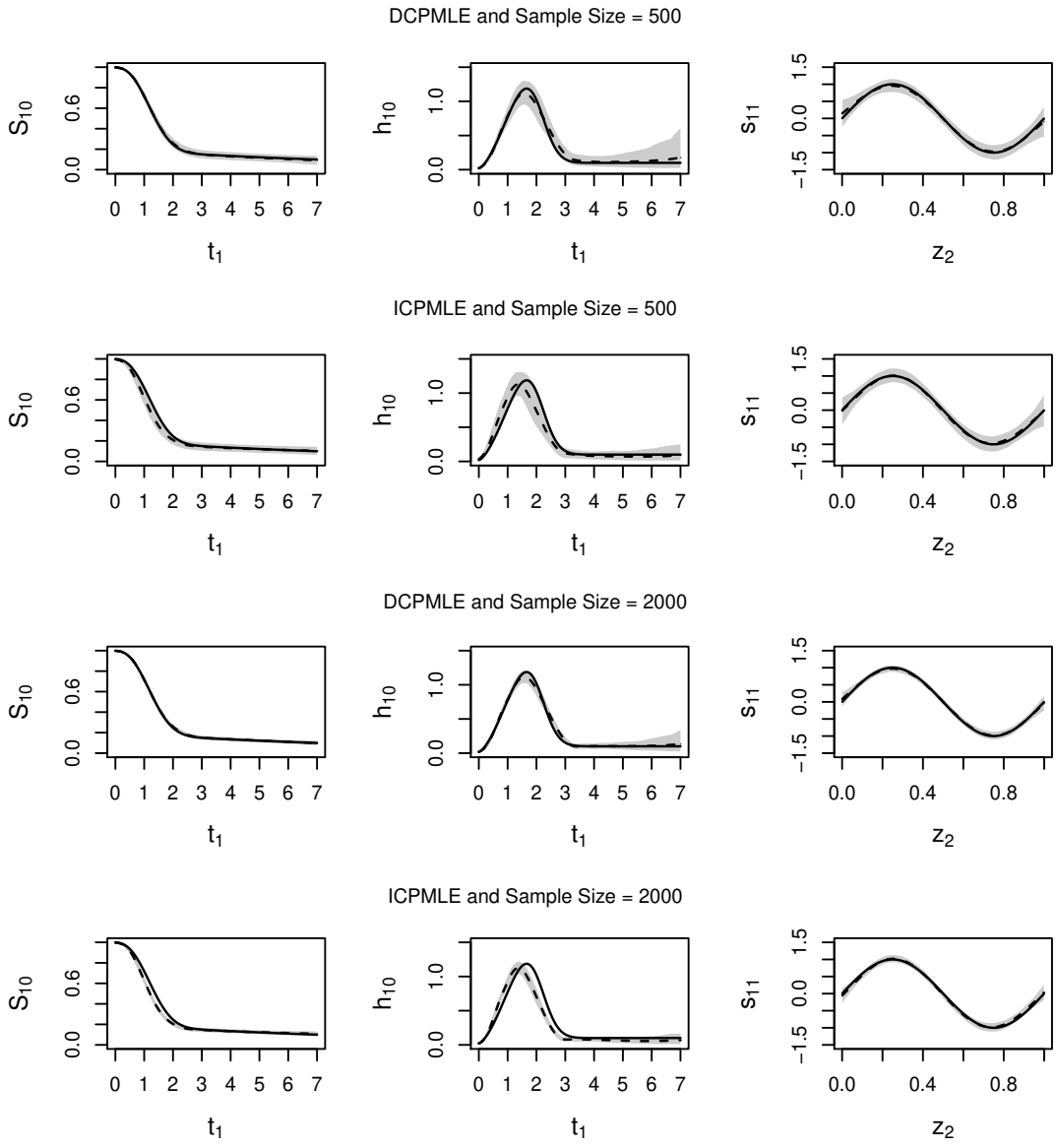

Figure C.16: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP3` (Table 5.2). Further details are given in the caption of Figure C.4.

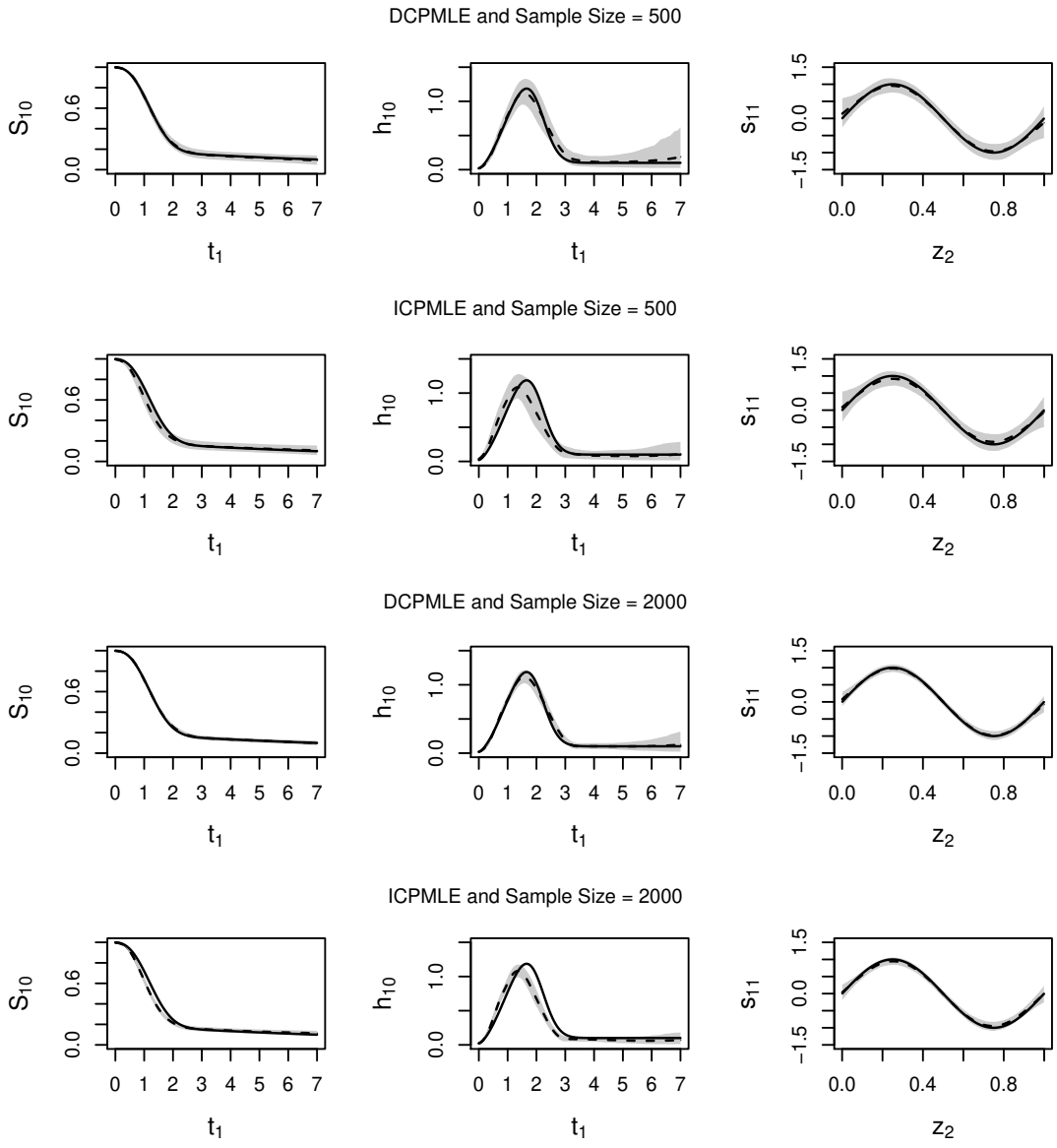

Figure C.17: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP4` (Table 5.2). Further details are given in the caption of Figure C.4.

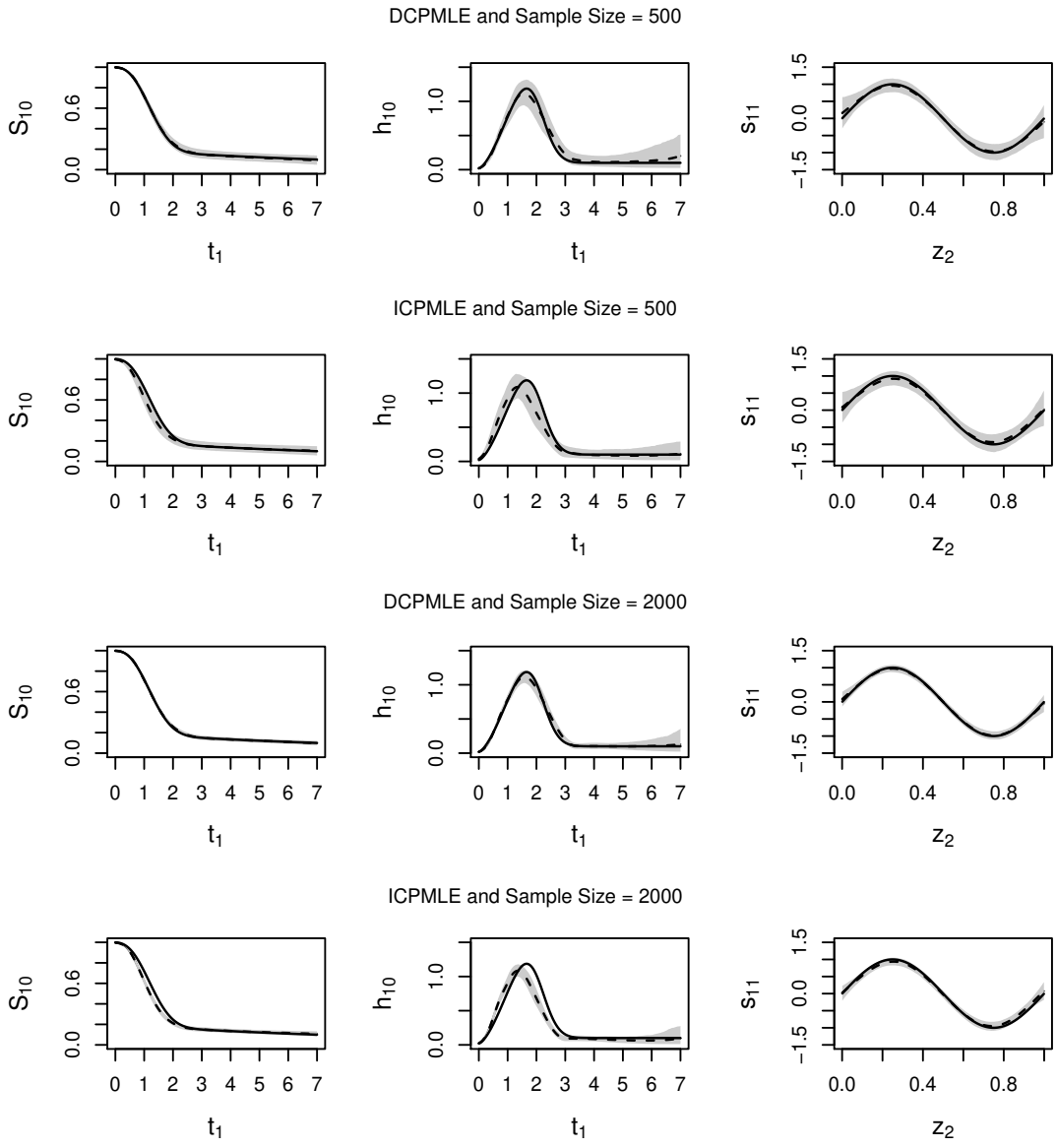

Figure C.18: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP5` (Table 5.2). Further details are given in the caption of Figure C.4.

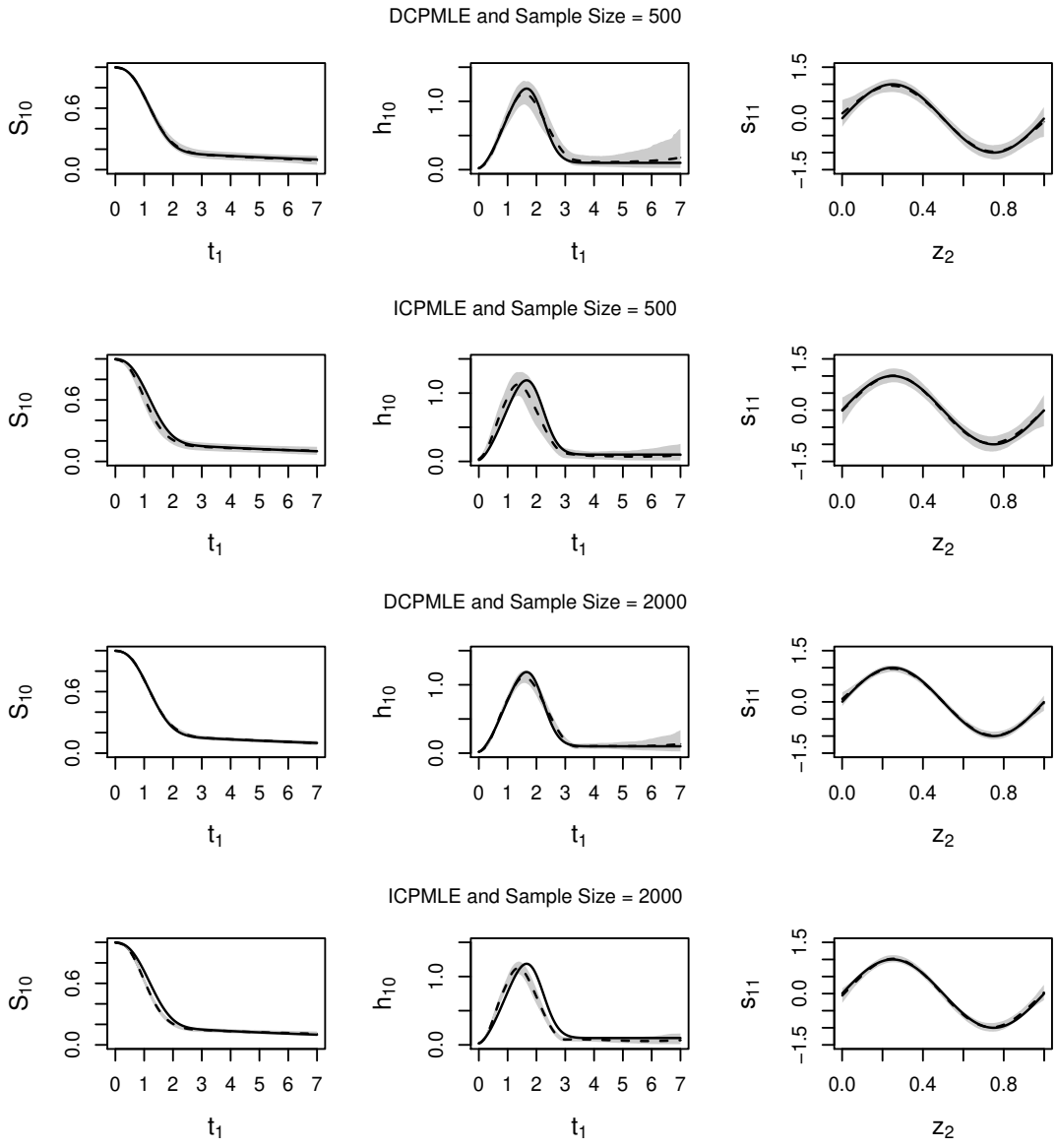

Figure C.19: Estimates of the survival functions (first column, S_{10}), hazard functions (second column, h_{10}) and smooth effects (third column, s_{11}) for the DCPMLE (rows 1 and 3) and ICPMLE (rows 2 and 4) obtained by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP6` (Table 5.2). Further details are given in the caption of Figure C.4.

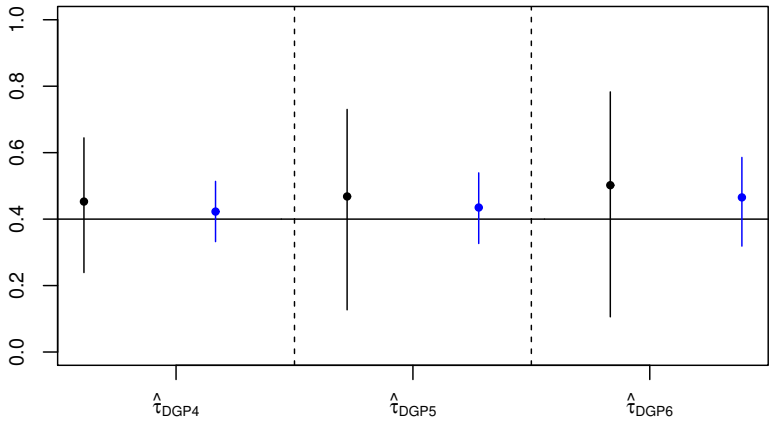

Figure C.20: Kendall Tau coefficient ($\tau = 0.4$) estimates obtained when DCPMLE is fitted by applying the `gjrm()` function in `GJRM` to dependent censoring survival data simulated according to `DGP4` (Clayton copula), `DGP5` (Frank copula) and `DGP6` (Gaussian copula) defined in Table 5.2. Further details are given in the caption of Figure C.12.

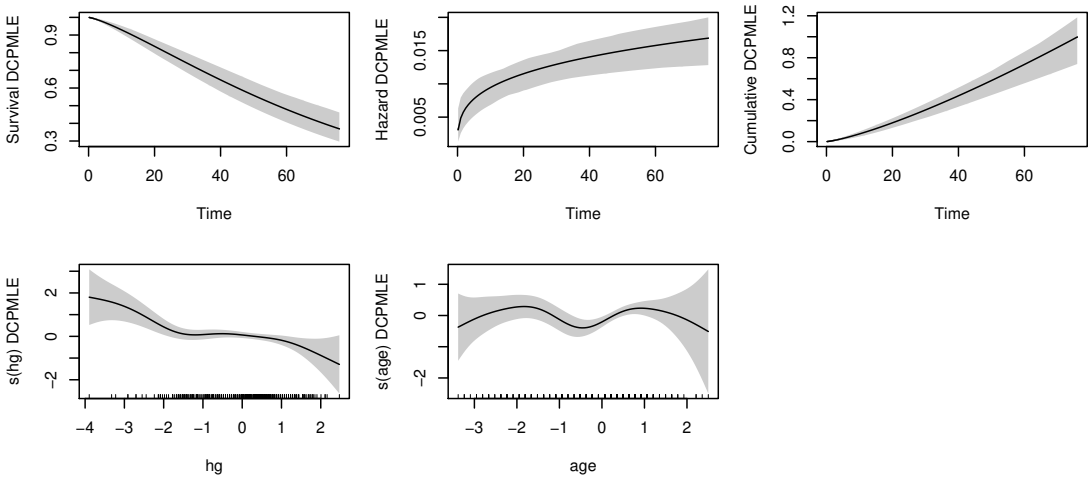

Figure C.21: Smooth function estimates and their corresponding 95% intervals for Model 8 in Table C.7 obtained by applying `gjrm()` in `GJRM` to prostate cancer data. The intervals have been obtained using the approach described in Section 4.3.4.

C.7 Model Selection and Additional Results for Section 5.7

Model	Link T_1	Link T_2	τ	AIC	BIC
1	PH	PH	0.472	3615.50	3727.02
2	PH	probit	0.422	3616.59	3728.05
3	PH	PO	0.455	3617.85	3729.39
4	PH	PH	-	3618.96	3730.15
5	PH	probit	-	3620.28	3731.60
6	PH	PO	-	3621.28	3732.57

Table C.6: Values of the model selection criteria (AIC and BIC) for the best dependent (`Models 1, 2` and `3`) and independent (`Models 4, 5` and `6`) censoring models fitted to the real data application in Section 5.7. The dependent censoring models were fitted using a Gaussian copula and all the covariates were included parametrically. The models were fitted using the functions `gamlss()` and `gjrm()` in `GJRM`.

Model	Copula	Covariates	Link T_1	Link T_2	AIC	BIC
7	Clayton ("C0")	rx hx pf sz sg s(hg) s(age)	PH	PO	3599.60	3752.86
8	Frank ("F")	rx hx pf sz sg s(hg) s(age)	PH	PO	3601.98	3730.14
9	Independent Censoring	rx hx pf sz sg s(hg) s(age)	PH	PO	3611.14	3731.16

Table C.7: Values of the model selection criteria (AIC and BIC) for the best dependent and independent models fitted to prostate cancer data, by allowing the covariates to be modelled nonparametricaly. The models were fitted using the functions `gamlss()` and `gjrm()` in `GJRM`.

Bibliography

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8), 907–925.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information Theory. Academiai Kiado, Budapest*.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. & Keiding, N. (2006). *Survival and event history analysis*. Wiley Chichester.

Anderson, J. & Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 322–327).

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, 68(2), 357–363.

Basu, A. & Ghosh, J. (1978). Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8(3), 413–429.

Basu, A. P. (1988). *Handbook of Statistics*, volume 7. Elsevier.

Bekker, P. & Wansbeek, T. (2001). Identification in parametric models. *A companion to theoretical econometrics*, (pp. 144–161).

Bekker, P. A. (1989). Identification in restricted factor models and the evaluation of rank conditions. *Journal of Econometrics*, 41(1), 5–16.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in medicine*, 2(2), 273–277.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227), 357–365.

Berman, S. M. (1963). Note on extreme values, competing risks and semi-markov processes. *The Annals of Mathematical Statistics*, 34(3), 1104–1106.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1), 134–167.

Box, G. E. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243.

Braekers, R. & Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33(3), 429–447.

Brechmann, E. C. & Schepsmeier, U. (2013). Modeling dependence with c- and d-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3), 1–27.

Byar, D. & Green, S. (1980). The choice of treatment for cancer patients based on covariate information. *Bulletin du cancer*, 67(4), 477–490.

Chen, X., Chernozhukov, V., Lee, S., & Newey, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2), 785–809.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2), 235–251.

Cheng, S., Wei, L., & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4), 835–845.

Cortese, G. & Scheike, T. H. (2008). Dynamic regression hazards models for relative survival. *Statistics in medicine*, 27(18), 3563–3584.

Cortese, G., Scheike, T. H., & Martinussen, T. (2010). Flexible survival regression modelling. *Statistical Methods in Medical Research*, 19(1), 5–28.

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

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2), 411–421.

Crowder, M. (1991). On the identifiability crisis in competing risks analysis. *Scandinavian Journal of Statistics*, 18(3), 223–233.

Crowder, M. J. (2012). *Multivariate survival analysis and competing risks*. Chapman and Hall/CRC.

Crowther, M. J. & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in medicine*, 32(23), 4118–4134.

Dabrowska, D. M. & Doksum, K. A. (1988a). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, 83(403), 744–749.

Dabrowska, D. M. & Doksum, K. A. (1988b). Partial likelihood in transformation models with censored data. *Scandinavian journal of statistics*, (pp. 1–23).

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., & De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

Deresa, N. W. & Van Keilegom, I. (2019). A multivariate normal regression model for survival data subject to different types of dependent censoring. *Computational Statistics & Data Analysis*, 144, 106879.

Dettoni, R., Marra, G., & Radice, R. (2020). Generalized link-based additive survival models with informative censoring. *Journal of Computational and Graphical Statistics*, (pp. 1–10).

Doksum, K. A. & Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review/Revue Internationale de Statistique*, (pp. 243–252).

Emoto, S. E. & Matthews, P. C. (1990). A weibull model for dependent censoring. *The Annals of Statistics*, (pp. 1556–1577).

Emura, T. & Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer.

Escarela, G. & Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4), 333–349.

Etezadi-Amoli, J. & Ciampi, A. (1987). Extended hazard regression for censored

survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics*, (pp. 181–192).

Gill, R. D. & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, (pp. 1501–1555).

Gourieroux, C. & Monfort, A. (1995). *Statistics and econometric models*, volume 1. Cambridge University Press.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420), 942–951.

Gu, C. (1992). Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics*, 1(2), 169–179.

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3), 297–310.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hayashi, F. (2000). *Econometrics*. Princeton University Press Princeton, NJ.

Herndon, J. E. & Harrell, F. E. (1995). The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in medicine*, 14(19), 2119–2129.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in medicine*, 13(10), 1045–1062.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, 60(3), 355–387.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 71(1), 75–83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2), 387–396.

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

Hsiao, C. (1989). Consistent estimation for some nonlinear errors-in-variables models. *Journal of econometrics*, 41(1), 159–185.

Hsu, C.-H., Taylor, J. M., & Hu, C. (2015). Analysis of accelerated failure time data with dependent censoring using auxiliary variables via nonparametric multiple imputation. *Statistics in medicine*, 34(19), 2768–2780.

Huang, X. & Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4), 1090–1099.

Jackson, D., White, I. R., Seaman, S., Evans, H., Baisley, K., & Carpenter, J. (2014). Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in medicine*, 33(27), 4681–4694.

Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, 73(361), 167–170.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Ed. Hoboken, Wiley.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282), 457–481.

Kleinbaum, D. G. & Klein, M. (2010). *Survival analysis*, volume 3. Springer.

Koop, G., Pesaran, M. H., & Smith, R. P. (2013). On identification of bayesian dsge models. *Journal of Business & Economic Statistics*, 31(3), 300–314.

Kooperberg, C., Stone, C. J., & Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, 90(429), 78–94.

Koopmans, T. C. & Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2), 165–181.

Koziol, J. A. & Green, S. B. (1976). A cramér-von mises statistic for randomly censored data. *Biometrika*, 63(3), 465–474.

Lagakos, S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1), 139–156.

Leitenstorfer, F. & Tutz, G. (2006). Generalized monotonic regression based on b-splines with an application to air pollution data. *Biostatistics*, 8(3), 654–673.

Li, R. & Peng, L. (2015). Quantile regression adjusting for dependent censoring from semicompeting risks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1), 107–130.

Liu, X.-R., Pawitan, Y., & Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical methods in medical research*, 27(5), 1531–1546.

Lu, Z. & Zhang, W. (2012). Semiparametric likelihood estimation in survival models with informative censoring. *Journal of Multivariate Analysis*, 106, 187–211.

Ma, J., Heritier, S., & Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics & Data Analysis*, 74, 142–156.

Marra, G. & Radice, R. (2020a). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115(530), 886–895.

Marra, G. & Radice, R. (2020b). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-2.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating hiv prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484–496.

Marra, G. & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Martinussen, T. & Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika*, (pp. 283–298).

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc., B*, 42(2), 109–142.

McCullagh, P. (2002). What is a statistical model? *Annals of statistics*, (pp. 1225–1267).

McKeague, I. W. & Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81(3), 501–514.

Moradian, H., Larocque, D., & Bellavance, F. (2019). Survival forests for data with dependent censoring. *Statistical methods in medical research*, 28(2), 445–461.

Nádas, A. (1971). The distribution of the identified minimum of a normal pair determines' the distribution of the pair. *Technometrics*, 13(1), 201–202.

Nelder, J. A. & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384.

Nelsen, R. (2006). *An Introduction to Copulas*. Second Edition, Springer, New York.

Newey, W. K. & McFadden, D. (1994). *Handbook of econometrics*, volume 4. Elsevier.

NLSY (1995). National Longitudinal Survey of Youth Handbook. The Ohio State University.

Nocedal, J. & Wright, S. (2006). *Numerical optimization, series in operations research and financial engineering*. Springer, New York, USA, 2006.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2), 185–198.

Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Canadian Journal of Cardiology*, 34(5), 575–584.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 234–243.

Pettitt, A. N. (1983). Approximate methods using ranks for regression with censored data. *Biometrika*, 70(1), 121–132.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, (pp. 761–768).

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 65(1), 167–179.

Pya, N. & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, 25(3), 543–559.

Rao, B. P. (1992). *Identifiability in stochastic models: characterization of probability distributions*. Academic Press.

Reid, N. (1994). A conversation with sir david cox. *Statistical Science*, 9(3), 439–455.

Rivest, L.-P. & Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1), 138–155.

Robins, J. M. & Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3), 779–788.

Rosenberg, P. S. (1995). Hazard function estimation using b-splines. *Biometrics*, (pp. 874–887).

Rossini, A. & Tsiatis, A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(434), 713–721.

Royston, P. & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21(15), 2175–2197.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Scharfstein, D. O. & Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3), 617–634.

Scheike, T. H. & Zhang, M.-J. (2002). An additive–multiplicative cox–aalen regression model. *Scandinavian Journal of Statistics*, 29(1), 75–88.

Scheike, T. H. & Zhang, M.-J. (2003). Extensions and applications of the cox-aalen survival model. *Biometrics*, 59(4), 1036–1045.

Shang, W. & Wang, X. (2017). The generalized moment estimation of the additive–multiplicative hazard model with auxiliary survival information. *Computational Statistics & Data Analysis*, 112, 154–169.

Shen, X. (1998). Propotional odds regression and sieve maximum likelihood estimation. *Biometrika*, 85(1), 165–177.

Siannis, F., Copas, J., & Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1), 77–91.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1), 1–21.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9(6), 449–460.

Slud, E. V. & Rubinstein, L. V. (1983). Dependent competing risks and summary survival curves. *Biometrika*, 70(3), 643–649.

Stanghellini, E., Vantaggi, B., et al. (2013). Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A), 1920–1937.

Staplin, N., Kimber, A., Collett, D., & Roderick, P. (2015). Dependent censoring in piecewise exponential survival models. *Statistical methods in medical research*, 24(3), 325–341.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, 36(2), 111–147.

Sujica, A. & Van Keilegom, I. (2018). The copula-graphic estimator in censored nonparametric location-scale regression models. *Econometrics and statistics*, 7, 89–114.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1), 147–160.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1), 20–22.

Vatter, T. & Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141, 147–167.

Wang, A., Chandra, K., Xu, R., & Sun, J. (2015). The identifiability of dependent competing risks models induced by bivariate frailty models. *Scandinavian Journal of Statistics*, 42(2), 427–437.

Whittemore, A. S. & Keller, J. B. (1986). Survival estimation using splines. *Biometrics*, (pp. 495–506).

Willems, S., Schat, A., van Noorden, M., & Fiocco, M. (2018). Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research*, 27(2), 323–335.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R*. Second Edition, Chapman & Hall/CRC, London.

Wu, J. & Witten, D. (2019). Flexible and interpretable models for survival data. *Journal of Computational and Graphical Statistics*, 28(4), 954–966.

Xingwei, T., Tao, H., & Hengjian, C. (2010). Hazard regression with penalized spline: The smoothing parameter choice and asymptotics. *Acta Mathematica Scientia*, 30(5), 1759–1768.

Xu, J., Ma, J., Connors, M. H., & Brodaty, H. (2018). Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood. *Statistics in medicine*, 37(14), 2238–2251.

Xu, J., Ma, J., & Prvan, T. (2017). Non parametric hazard estimation with dependent censoring using penalized likelihood and an assumed copula. *Communications in Statistics-Theory and Methods*, 46(22), 11383–11403.

Xue, X., Xie, X., & Strickler, H. D. (2018). A censored quantile regression approach for the analysis of time to event data. *Statistical methods in medical research*, 27(3), 955–965.

Yang, S. & Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, 94(445), 125–136.

Younes, N. & Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, 53(4), 1199–1211.

Yuan, M. (2005). Semiparametric censorship model with covariates. *Test*, 14(2), 489–514.

Zahl, P.-H. (2003). Regression analysis with multiplicative and time-varying additive regression coefficients with examples from breast and colon cancer. *Statistics in medicine*, 22(7), 1113–1127.

Zeng, D. et al. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics*, 32(4), 1533–1555.

Zheng, M. & Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1), 127–138.