UNIVERSITAT DE VALÈNCIA

Programa de Doctorat en Estadística i Optimització

Flexible Bayesian survival models with application in biometric studies

by

Elena Lázaro Hervás

Thesis submitted for the degree of Doctor of Philosophy in Statistics and Optimisation

Supervised by

Carmen Armero i Cervera,
Virgilio Gómez Rubio
and Luis Rubio Miguélez

in the

Faculty of Mathematics
Department of Statistics and Operations Research

May 2018

This thesis was supported by grant FPU2013-02042 from the Spanish Ministry of Education, Culture and Sport, and by grant MTM2016-77501-P from the Spanish Ministry of Economy and Competitiveness.

Acknowledgements

Una tesis es al final una experiencia vital que hace que las personas que te rodean jueguen un papel fundamental. Por eso, quiero aprovechar esta sección para mostrar mi agradecimiento a todas las personas que forman parte de mi vida y que de un modo u otro me han acompañado en esta aventura.

Y entre todas estas personas quisiera destacar el papel fundamental de Carmen Armero, Virgilio Gómez y Luis Rubio, mi directora y directores de tesis, por aceptar acompañarme en este proceso. Gracias Carmen, por tu infinita ayuda, por escucharme y tener en cuenta siempre mis opiniones, por compartir de forma tan generosa todos tus conocimientos, por creer en mi y sobretodo por cuidarme, mil gracias de corazón. Gracias Virgilio, por llenar de optimismo el final de este proceso, por tu cálida acogida durante mi estancia en Albacete y por creer en mi trabajo. Y Gracias Luis, porque pese a que las circunstancias no han dejado que fuera lo que iba a ser, por tu confianza y comprensión.

A toda mi familia y de forma muy especial a mis padres, Juli y Aurelio, y a mi hermana Belén, por cuidar de mi, sobretodo en mis momentos de máxima fragilidad, por quererme sin condiciones, por ayudarme a levantarme las muchísimas veces que me he caído y por enseñarme a luchar por mis sueños. A mi abuela Julia, por su amor infinito. A mis dos tesoros, mi hijo Julio y mi hija Helena, porque son los que más han sufrido la gran cantidad de horas invertidas en este trabajo y a pesar de eso, siempre me han recibido con una sonrisa y un montón de abrazos y besos, os quiero. A mi compañero de vida, Julio, por permanecer a mi lado y por apoyar y respetar mi trabajo. A toda mi familia política, y en especial a Dori y a Julio, por su ayuda inestimable. Gracias Dori, por comprender desde el minuto cero cuan importante era para mi este proceso.

También quiero dar las gracias a todo el personal del Departamento de Estadística e Investigación Operativa por su acogida tan calurosa durante estos cuatro años y como no podría ser de otra manera a las chicas de secretaría, por su eficaz gestión de todo el papeleo que acompaña este proceso. No puedo olvidar a todas mis compañeras y todos mis compañeros predoctorales, por llenar de sentido del humor y de momentos de evasión mis días, infinitas gracias por llenar mi camino doctoral de tantas sonrisas. Y en especial quiero agradecer de forma muy particular a Danilo Alvares da Silva y Blanca Sarzo Carles, su apoyo, su ayuda y sus palabras de aliento en los momentos más difíciles.

Para todas vosotras y todos vosotros. Muchísimas gracias.

"Hazte quien eres: hay que hacerse quien se es, y todos somos distintos. Pero lo que quiera que seas desarróllalo al máximo. Cada cual debe aspirar a ser lo máximo que pueda ser con sus condiciones. Y de esa manera devolverá a la vida de todos la vida que ha recibido él."

José Luis Sampedro

UNIVERSITAT DE VALÈNCIA

Resumen

Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa

Programa de Doctorado en Estadística y Optimización

Introducción

El análisis de supervivencia es una metodología estadística diseñada para analizar datos procedentes de estudios científicos relativos a tiempos de ocurrencia de uno o varios eventos de interés. La duración de estos tiempos suele conocerse como tiempos de supervivencia debido a los particulares orígenes de esta metodología en contextos exclusivamente médicos y demográficos. Durante las últimas décadas, la literatura científica en este campo ha sido muy prolífica y su aplicación se ha extendido a múltiples áreas de conocimiento.

Los procedimientos estadísticos propios de esta metodología empezaron a abordarse desde el marco inferencial frecuentista. Sin embargo, en los últimos años la utilización de la metodología bayesiana, tanto en desarrollos teóricos como en estudios reales, ha experimentado un enorme interés. Uno de los elementos más importantes que pueden ayudar a entender el aumento de su presencia es, sin duda, el desarrollo de entornos y herramientas computacionales rápidos y eficientes.

El atractivo principal de la metodología bayesiana es estrictamente conceptual. Proporciona un marco teórico que permite cuantificar de forma probabilística cualquier tipo de incertidumbre asociada al

problema objeto de estudio y permite, también, la incorporación de información experta al proceso inferencial, que es de especial relevancia en escenarios de tipo biológico y médico. Además, en el contexto del análisis de la supervivencia, la estadística bayesiana incorpora de forma natural y sencilla el tratamiento de mecanismos de censura y sobretodo, de truncamiento.

Objetivos

Este proyecto de tesis tiene como objetivo principal desarrollar e implementar nuevas propuestas metodológicas en el contexto del análisis de supervivencia y en el marco del paradigma bayesiano, al que consideramos una metodología adecuada y robusta para abordar el tratamiento de modelos de supervivencia complejos. Nuestra visión de la estadística no se circunscribe únicamente al mundo de la metodología y la teoría. También concebimos la estadística como una herramienta poderosa y necesaria para el estudio de problemas reales basados en datos. Por ello, ilustramos el comportamiento de estas propuestas metodológicas combinando el uso de datos simulados y de datos procedentes de estudios de áreas de conocimiento de distinta naturaleza, como son el área de la mejora genética de plantas, de la microbiología de alimentos y de las ciencias de la salud.

Uno de los objetivos específicos de esta memoria es proponer y evaluar modelizaciones de tipo paramétrico bajo diferentes esquemas de censura, concretamente en contextos de censura por la derecha y censura por intervalos.

El segundo de los objetivos específicos de esta memoria es proponer y analizar modelizaciones flexibles en el contexto del modelo de riesgos proporcionales de Cox (Cox, 1972), así como en extensiones de dicho modelo en el marco de los modelos conjuntos para datos longitudinales y de supervivencia. Nuestra propuesta se fundamenta en el estudio de

diferentes especificaciones, paramétricas y no paramétricas, de la función de riesgo basal. Esta componente tiene un papel clave en la modelización estadística debido a su influencia directa en la estimación de la función de riesgo y, en consecuencia, en la función de supervivencia, por lo que su inespecificación o su incorrecta especificación puede condicionar negativamente el proceso inferencial y, por tanto, conducir a conclusiones erróneas o poco precisas.

El tercer gran objetivo específico de esta memoria se orienta al tratamiento de modelos de supervivencia complejos. Estudiamos algunos modelos de supervivencia inicialmente intratables a través del entorno integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009) como son los modelos de curación de tipo mixtura. Nuestra propuesta se basa en la adaptación del algoritmo propuesto por Gómez-Rubio (2017) para ajustar modelos de mixtura con INLA.

Para finalizar, querríamos comentar que en esta memoria también trabajamos, aunque de forma transversal, temas relativos a los procedimientos bayesianos de regularización a través de estructuras de correlación en las distribuciones *a priori*, la computación de distribuciones *a posteriori* de cantidades de interés relevantes en los problemas objeto de estudio, la evaluación de modelos a través de algunos de los criterios de selección más relevantes, así como también la comparación entre los dos procedimientos más comunes para llevar a cabo inferencia bayesiana: los métodos de simulación basados en métodos de cadenas de Markov Monte Carlo (MCMC) y la metodología INLA.

Estructura de la memoria

Después de introducir brevemente el marco teórico en que se fundamenta esta memoria y los objetivos, en esta sección presentamos de forma detallada sus contenidos:

- **Capítulo 1. Introducción.** Este capítulo introduce el contexto de la presenta memoria y hace un resumen de los contenidos que se abordan en la misma.

- **Capítulo 2. Análisis bayesiano de supervivencia.** Este capítulo proporciona una introducción muy general al análisis de supervivencia y una visión general de los conceptos característicos de este tipo de análisis. Concretamente, definimos de forma detallada la función de supervivencia y la función de riesgo, así como también los fenómenos de censura y truncamiento y su influencia en la construcción de la función de verosimilitud. También abordamos con detalle la descripción de las distribuciones de probabilidad más habituales en este contexto y los modelos de regresión de supervivencia que usaremos a lo largo de este trabajo. Finalmente, presentamos una visión general de la metodología bayesiana que incluye una breve descripción de los métodos MCMC y la metodología INLA.

- **Capítulo 3. Análisis bayesiano de supervivencia en mejora genética de plantas y en microbiología de alimentos.** Este capítulo exporta el análisis bayesiano de supervivencia a los contextos de la mejora genética de plantas y la microbiología de alimentos para el tratamiento de diferentes esquemas de censura, concretamente censura por intervalos y censura por la derecha. Estas dos áreas de conocimiento han sido fundamentales en el desarrollo de la estadística, sin embargo, en algunas ocasiones infrautilizan mucha de la metodología existente. En el contexto de la mejora genética de plantas proponemos el uso de los modelos de tiempo de fallo acelerado (AFT) con distribución de base de valores extremos para evaluar una nueva variedad de planta en términos de su resistencia y tolerancia frente a un virus específico. Añadimos al estudio una comparación con los métodos inferenciales clásicos para evaluar su robustez con respecto al tratamiento de observaciones censuradas. En el contexto de la microbiología de alimentos proponemos un modelo de riesgos proporcionales

de Cox (CPH) para evaluar los cambios de virulencia de un patógeno humano de transmisión alimentaria como consecuencia de diferentes frecuencias de aplicación de un nuevo tratamiento de preservación. Aprovechamos este ejemplo ilustrativo para realizar una comparativa entre los métodos MCMC y la metodología INLA.

- **Capítulo 4. Funciones de riesgo basal en el modelo bayesiano de riesgos proporcionales de Cox.** Este capítulo presenta una doble finalidad. La primera se centra en evaluar la influencia de la especificación de la función de riesgo basal en el marco del modelo CPH. Abordamos la definición de la mencionada función a través de una elección paramétrica basada en la distribución de Weibull y dos no paramétricas, definidas a través de una mixtura de funciones constantes y a través combinaciones lineales de bases cúbicas de B-splines, respectivamente. La segunda, se centra en la evaluación del efecto de la regularización bayesiana a través de la definición de estructuras de correlación en las distribuciones *a priori* que describen los parámetros implicados en las propuestas no paramétricas. Los procesos inferenciales sujetos a especificaciones no paramétricas de la función de riesgo basal pueden presentar problemas de sobreajuste e inestabilidad y la regularización bayesiana a través de la especificación de escenarios *a priori* que contengan estructuras de correlación se presenta como una posible solución. Estas propuestas se ilustran haciendo uso del conjunto de datos usados en el Capítulo 2, que recogen información sobre un ensayo de virulencia en el contexto de la microbiología de alimentos. Además, hacemos usos de la simulación para generar diferentes escenarios de interés en base a la metodología presentada. Acometemos la evaluación de los dos objetivos descritos con anterioridad realizando una comparativa entre los diferentes escenarios de modelización propuestos en base a las distribuciones *a posteriori* de los parámetros de interés así como también de algunas cantidades derivadas, como las distribuciones *a posteriori* de las funciones de riesgo y supervivencia. También

valoramos las distintas modelizaciones en términos de bondad de ajuste y de capacidad predictiva a través de dos criterios de selección de modelos como son el "deviance information criterion" (DIC) (Spiegelhalter *et al.*, 2002) y el "log pseudo-marginal likelihood "(LPML) (Geisser and Eddy, 1979).

- **Capítulo 5. Modelos bayesianos de curación de tipo mixtura usando R-INLA.** En este Capítulo proponemos la implementación de una extesión del software bayesiano R-INLA para estimar modelos de curación de tipo mixtura. INLA es una metodología alternativa a los métodos de MCMC para realizar inferencia bayesiana. Sin embargo, en el caso de los modelos de curación basados en mixturas no se puede aplicar de forma directa. Ilustramos el comportamiento de esta propuesta a través de dos estudios paradigmáticos en el área de la medicina, el primero en temas oncológicos y el segundo relativo a transplantes de médula ósea. Valoramos la bondad de nuestra propuesta a través de una comparación completa con la técnicas MCMC.

- **Capítulo 6. Funciones de riesgo basal en modelos bayesianos conjuntos.** Este Capítulo comparte la base metodológica explorada en el Capítulo 4 pero extiende el material propuesto al contexto de los modelos bayesianos conjuntos para datos longitudinales y datos de supervivencia. En particular, nos centramos en una formulación de la modelización conjunta estándar, en la que definimos el submodelo de supervivencia a través de un modelo de riesgos proporcionales de Cox (CPH) y el submodelo longitudinal a través de un modelo lineal mixto, estableciendo la correlación entre los dos procesos a través de los efectos aleatorios. El capítulo también contempla el desarrollo de cuestiones metodológicas referidas a la modelización específica de riesgos competitivos, ya que nuestras propuestas se ilustran con datos pertenecientes a un estudio que persigue evaluar la relación entre los eventos "morir" y "ser dado de alta" y un marcador longitudinal que valora el ínidice de gravedad de pacientes con

ventilación mecánica ingresados en unidades de cuidados intensivos (UCI). Evaluamos las diferencias en todos los procesos inferenciales acometidos comparando las estimaciones *a posteriori* de los parámetros más relevantes en la modelización y las distribuciones *a posteriori* de cantidades interés propias del contexto de la aplicación. En este capítulo también hacemos una comparativa entre los diferentes escenarios de modelización en términos de la bondad del ajuste y de la capacidad predictiva de los mismos a través de los siguientes criterios de selección de modelos: "deviance information criterion" (DIC) (Spiegelhalter *et al.*, 2002) y "log pseudo-marginal likelihood " (LPML) (Geisser and Eddy, 1979).

- **Capítulo 7. Conclusiones y trabajo futuro.** En el último capítulo de esta memoria se subrayan las principales conclusiones obtenidas y las líneas de trabajo futuro.

- **Apéndice A. Método de la transformación inversa.** En este apéndice mostramos la adaptación del método de la transformación inversa (Crowther and Lambert, 2013) en el contexto del análisis de supervivencia para acometer la simulación de datos en el marco del modelo de riesgos proporcionales de Cox bajo especificaciones paramétricas y no paramétricas de la función de riesgo basal.

Conclusiones

En este trabajo, hemos propuesto y desarrollado diferentes propuestas metodológicas en el contexto del análisis de supervivencia bajo el paradigma bayesiano. Las principales conclusiones obtenidas a través de los estudios plasmados en esta memoria se resumen a continuación.

- Los resultados obtenidos en el Capítulo 3 respaldan tres conclusiones relevantes. En primer lugar, se pone en evidencia la potencia de la metodología bayesiana en el contexto del análisis de

supervivencia, así como la existencia de software bayesiano robusto y accesible para implementar procesos inferenciales complejos. En segundo lugar, se manifiesta la utilidad del análisis bayesiano de supervivencia en ciertas áreas de investigación en que su aplicación es escasa. En tercer lugar, se constata la gran robustez de esta metodología con respecto al enfoque clásico para proporcionar inferencias sólidas en contextos donde se presentan esquemas de censura complejos.

- Los resultados obtenidos en el Capítulo 4 subrayan la utilidad de los métodos bayesianos para incorporar flexibilidad a través de especificaciones no paramétricas de la función de riesgo basal en el contexto del modelo de riesgos proporcionales de Cox (CPH). A este respecto, observamos que las especificaciones no paramétricas de la función de riesgo basal son capaces de incrementar la adaptabilidad de la modelización en lo que se refiere a la captura de patrones de la función de riesgo con tendencias que están fuera de la monotonicidad. También se pone de manifiesto la eficacia del proceso de regularización bayesiano para minimizar los problemas de sobreajuste e inestabilidad propios de las modelizaciones que contemplan una especificación no paramétrica de la función de riesgo basal. Además nuestras propuestas metodológicas parecen superar las limitaciones del enfoque clásico para abordar el proceso inferencial basado en el método de la "verosimilitud parcial", en el que el proceso de estimación se aborda omitiendo la especificación de la función de riesgo basal. La aplicación de esta metodología en el análisis de datos procedentes de un estudio real y bajo diferentes escenarios de simulación subraya también la importancia de abordar correctamente la estimación de la función de riesgo basal y de capturar su tendencia con objeto de completar todo el proceso inferencial y de proporcionar resultados precisos en lo que se refiere, sobretodo, a la estimación de cantidades *a posteriori* de interés, como por ejemplo las probabilidades de supervivencia *a posteriori*.

- Los principales resultados del Capítulo 5 refuerzan la capacidad del software R-INLA como alternativa a los métodos MCMC para realizar análisis de supervivencia bayesiana así como también las posibilidades de su extensión a modelos más complejos. Nuestra propuesta extiende el uso de INLA para la estimación de modelos de curación de tipo mixtura a través de una descomposición de las distribuciones marginales *a posteriori* en términos de las distribuciones condicionales *a posteriori* dada toda la información latente del modelo y el uso de un algoritmo adaptado basado en la propuesta de Gómez-Rubio (2017). Los resultados inferenciales obtenidos para los dos ejemplos ilustrativos son buenos y precisos en comparación con los que proporcionan las métodos MCMC.

- Los resultados obtenidos en el Capítulo 6 apoyan nuevamente las conclusiones que se derivan del Capítulo 4, pero en el contexto de los modelos conjuntos para datos longitudinales y de supervivencia con objetivos de supervivencia. Los resultados obtenidos enfatizan de nuevo las fortalezas del enfoque bayesiano para introducir flexibilidad en el submodelo de supervivencia por medios de escenarios similares a los discutidos en el Capítulo 4. Además, se refuerza la solidez de esta metodología en el ajuste de modelos conjuntos permitiendo completar todos los procesos inferenciales (proceso longitudinal, proceso de supervivencia, y la asociación entre los dos procesos), cuantificando la incertidumbre, y estimando y lidiando con el fenómeno de censura de manera eficiente.

Contents

List of Figures

List of Tables

A la memoria de mi abuelo, Manuel Hervás LLora, porque tu presencia sigue viva en mi.

Chapter 1

Introduction

1.1 Main objectives

Survival analysis groups a great variety of statistical methods for analysing data whose main response variable is the time until the ocurrence of an event of interest. Its relevance in the field of statistics is very substantial due to its extensive application in many fields of science. Literature for survival analysis shows an use of both frequentist and bayesian statistical approaches. However, in recent years bayesian methods for new analysis have proliferated considerably due to several reasons which are summarised in the following paragraph.

Possibly, the most important elements are related to the improvement of computational methods, the increase of the processing capacity, and the development of statistical software. On the other hand, the bayesian paradigm allows to deal with complex censoring and truncation schemes easily and, furthermore the availability of software eases their implementation. In addition, bayesian methodology enables the assessment of uncertainty estimates through explicit probabilistic tools. Point and interval

estimates are naturally derived form the subsequent posterior distribution such as, for instance, posterior variances and posterior probabilities and features of the survival curves with regard to relevant covariate patterns. It is also remarkable that bayesian methods make also possible the incorporation of prior information to the inferential process, thus improving and enhancing estimation and prediction of any outcome of interest (Guo and Carlin, 2004). See Ibrahim *et al.* (2001) for further explanation about the advantages of bayesian survival analysis.

This PhD dissertation relies on the fact that the bayesian approach is a suitable and robust methodology to perform survival analyses beyond the standard survival models. This conception is based on the bayesian hierarchical model formulation which allows the introduction and implementation of complex structures in survival modeling in an easy and intuitive manner. Specifically, the aim of this PhD is to provide an appropiate methodology that will allow us to describe and illustrate the use and application of flexible survival models in many biometrical contexts.

The specific objectives of this Thesis are:

- To place on value the potentialities of bayesian survival analysis in contexts in which that methodology has not been widely used. In that regard, we play special attention to some of the advantages that this approach offers compared to frequentist inference.
- To propose and implement a general survival modeling framework in the context of Cox proportional hazards (CPH) models (Cox, 1972). There are many studies that need to go beyond the standard approach of CPH model (Cox, 1972) in which the baseline hazard is usually unespecified

or parametrically defined. Baseline hazard functions are a key component in the CPH model definition and its misspecification can imply a lost of valuable model information that can make impossible to fully report estimated outcomes of interest, such as posterior probabilities and survival curves for all relevant groups patterns.

In that regard, different model scenarios are adressed and discussed based on:

- Parametric and non-parametric specifications of the baseline hazard function. Weibull distribution is the default choice to illustrate the parametric specification while non-parametric specifications are defined by means of mixtures of piecewise constant functions (Sahu *et al.*, 1997) and cubic B-spline functions (Hastie *et al.*, 2009).
- Different prior scenarios that introduce regularization procedures to avoid overfitting and unstability (Breiman, 1996) in the estimation process of the models defined via non-parametric baseline hazard proposals.

- To propose and implement a feasible extension to estimate standard mixture cure models by means of the integrated nested Laplace approximation (INLA, Rue *et al.*, 2009).

At this point it is worth mentioning that our intention in this PhD project has a transversal objective based on comparing two of the most usual methods for accounting bayesian inference in the context of survival analysis: Markov chain Monte Carlo (MCMC) simulation methods and the INLA methodology.

1.2 Layout

After this introductory part which briefly describes the aims of this PhD dissertation, the contents are outlined as follows:

Chapter 2. **Bayesian survival analysis.** This Chapter provides a very general introduction to survival analysis and an overview of the main concepts, such as the survival and the hazard functions. We also emphasise the concept of censoring and underline its influence in the construction of the likelihood function. We describe the most usual survival probabilistic distributions. Then, we introduce the survival regression models that we will use throughout this dissertation. Finally, we present an overview of the bayesian methodology which includes a brief description of the most widely used computing tools to account for the inference process.

Chapter 3. **Bayesian survival analysis in plant breeding and food microbiology.** This Chapter highlights the potentialities of bayesian survival analysis in plant breeding and food microbiology, two research areas in which the bayesian survival analysis is not very common. In plant breeding area, we use accelerated failure time (AFT) models to evaluate a new plant variety for resistance and tolerance to a specific virus. We add to the study a comparison with its relative frequentist counterpart to underline the strengths of the bayesian methodology with regard to the treatment of censored observations. On the other hand, in the context of food mircrobiology we propose a Cox proportional hazards (CPH) model to assess virulence changes in a foodborne pathogen as a consequence of different frequencies of application of a new preservation treatment.

The inferential process is based on MCMC methods and the INLA methodology with the aim of carrying out a comparison between the results from both methodologies.

Chapter 4. **Baseline hazard functions in the bayesian Cox proportional hazards model.** This Chapter presents a twofold objective. The first one is focused on assessing the influence of the specification of the baseline hazard function in the context of the CPH model. We consider a parametric election based on the Weibull distribution and two paradigmatic non-parametric ones, defined by means of mixture of piecewise constant functions and cubic B-spline functions. The second objective, is centered on evaluating the effect of regularization with different prior proposals for the coefficients associated to non-parametric baseline hazard models. Note that inferential processes in which a non-parametric proposal is used to define the baseline hazard function can suffer overfitting and unstability problems. We illustrate these issues by means of a real dataset which collects information about a virulence assay in the context of food microbiology as well as a simulation study. Differences in all statistical processes were evaluated through relevant posterior estimates as well as other derived quantities resulted from posterior hazard and survival functions. We also discussed two model selection scores to measure the goodness of fit and the predictive ability of the different models considered.

Chapter 5. **Bayesian mixture cure models using R-INLA** In this Chapter, we propose a feasible INLA extension for estimating mixture cure models. INLA is currently an alterntative to MCMC methods to perform bayesian inference, however in the case of mixture cure models it is not directly applicable. We illustrate our proposal by

means of two benchmark paradigmatic datasets and confirm the accuracy of our proposal through a comparison with MCMC methods.

Chapter 6. **Baseline hazard functions in bayesian joint models.** This Chapter shares the main objectives and methodology with Chapter 4 but extends the proposals to the context of bayesian joint models for longitudinal and survival data. In particular, we focus on a simple joint model, with the survival part defined in terms of a CPH model which accounts for longitudinal information described in terms of a mixed lineal model. We discuss several important issues in a benchmark survival study devoted to assess the relationship between the risk of *death* or be *discharged alive* and a longitudinal disease severity index marker in patients hospitalized at intensive care units. It is worth noting that the survival model is defined by means of a competing risks survival for the two events of interest (*death discharged alive*). Differences in all inferential processes were evaluated comparing relevant posterior estimates as well as other derived quantities. Goodness of fit and predictive ability was assessed in terms of different models selection scores.

Chapter 7. **Conclusions and future researh.** This is the last chapter of this dissertation. It presents some conclusions and suggests different issues for future research.

The final part of the project includes the usual section with all the bibliographic references mentioned in the document as well as one Appendix, **Appendix** A, devoted to develop the inversion method to simulate survival times.

Chapter 2

Bayesian survival analysis

2.1 Introduction

This Chapter introduces time-to-event models as well as the general objectives of its statistical analysis. Some fundamental concepts and procedures are introduced and commonly used methods of estimation are described. The framework in this Chapter is the basis for the methodological proposals and applications presented in subsequent chapters.

Time-to-event analysis refers to the statistical methodology developed to study outcome variables that describe the time from a starting time until an event of interest or end point occurs. It is also named as survival analysis (Collet, 2015; Ibrahim *et al.*, 2001) given its extended use in the fields of medicine and biology. This methodology is known as event history analysis when it is applied in the area of sociology, failure time analysis or reliability analysis in engineering, and duration analysis or transition analysis in economics.

The main objectives of survival analysis include the analysis of event time patterns, comparison of survival times in different groups of individuals, and the assessment of covariates associated to the risk of the occurrence of the event of interest (Kartsonaki, 2016). Its statistical treatment requires taking into account the following two special features: i) the response variable time is generally positively skewed, and ii) not all individuals experience the event of interest within the follow-up period (i.e., they are censored observations) (Crowther, 2014).

2.2 Survival and hazard functions

Let T be a non-negative random variable, which represents the time up to a given event. Consequently, its probabilistic behaviour can be equivalently described by the survival function, the density function, and the hazard function. Mathematically, they can all be written in terms of one another. It is worth noting that in all this document the survival time is always considered as a continuous random variable. Next, we describe those functions with a special emphasis on their relationships, according to definitions in Lee and Wang (2013).

The survival function $S(t)$ is defined as the probability of surviving longer than time t:

$$S(t) = P(T > t),\ t > 0. \tag{2.1}$$

$S(t)$ is a non-increasing function with $S(0) = 1$ and $S(t) \to 0$ as $t \to \infty$. It is related to the cumulative distribution function (cdf), $F(t)$, as

$$F(t) = 1 - S(t), \tag{2.2}$$

which represents the probability that an individual experiences the event of interest before time t.

Survival time T has a density function, $f(t)$, defined as the limit of the probability that an individual experiences the event of interest in the interval $(t, t + \Delta t)$, i.e., the probability of failure within a interval:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t}. \tag{2.3}$$

The density function, also known as the unconditional failure rate, satisfies:

1. $f(t) > 0, \quad t > 0$ and $f(t) = 0, \quad t < 0$
2. $\int_0^\infty f(t)\, dt = 1.$

Hazard Function

The hazard function, $h(t)$, also known as the conditional failure rate, is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t}. \tag{2.4}$$

This function is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate (Klein and Moeschberger, 2005). The hazard function must be positive, $h(t) \geq 0$ and its integral over $[0, \infty]$ must be infinite. Moreover, it can be increasing, decreasing, constant or

a combination of both. It can be expressed in terms of the density function as follows:

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{2.5}$$

The cumulative hazard function, $H(t)$, is a relative expression of the hazard function:

$$H(t) = \int_0^t h(u)\, du. \tag{2.6}$$

We show all possible transformations of the three essential functions described above. The hazard function, $h(t)$, combining (2.2) and (2.5) can be rewritten as

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.7}$$

The density function, $f(t)$, can be expressed as

$$f(t) = \frac{d}{dt}\left[1 - S(t)\right] = -S'(t), \tag{2.8}$$

given that it is defined as the derivative of the cumulative hazard function, $H(t)$.

Combining (2.8) in (2.7), the hazard function can also be defined as

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log(S(t)). \tag{2.9}$$

Integrating (2.9) and combining it with (2.6), the following identity is obtained:

$$-\int_0^t h(u)\,du \;=\; \log(S(t)). \tag{2.10}$$

Alternatively, it can be expressed as:

$$S(t) \;=\; \exp[-H(t)] \;=\; \exp\Big[-\int_0^t h(u)du\Big]. \tag{2.11}$$

Finally, using (2.7) and (2.11), the density function can also be rewritten as:

$$f(t) \;=\; h(t)\exp[-H(t)]\,. \tag{2.12}$$

2.3 Censoring and truncation

One of the reasons why survival analysis requires "special" techniques is because of the possibility that the event of interest could not be fully observed for some individuals. These incomplete observations are usually referred to censored or truncated and cannot be removed from the analysis. Furthermore, they need to be correctly identified and handled appropiately in the statistical model. Based on the excellent book by Klein and Moeschberger (2005), censoring patterns can be classified as:

1. Right censoring.
   (a) Type I censoring.
   (b) Type II censoring.
   (c) Random right censoring.
2. Left censoring.

3. Interval censoring.
4. Truncation.
   (a) Right truncation.
   (b) Left Truncation.

Next, we explain riefly the meaning of each of these patters.

Right censoring

In the case of right censored observations, times to event are known to be above a certain time C_R. Hence, if T denotes the observed relative lifetimes instead of their lifetimes then $T = \min(T^*, C_R)$, where T^* is the time-to-event randon variable. An indicator variable is used for describing whether an survival time is censored, that is

$$\delta = \begin{cases} 1, & \text{if } T^* \leq C_R \\ 0, & \text{otherwise.} \end{cases}$$

Observations will be expressed in terms of pairs (T, δ).

Right censoring C_R can be fixed or random depending on the characteristics of the study. This situation generates the following right censoring types:

(a) **Type I censoring**: the end of the period of the study C_R is known and pre-fixed before it begins.

(b) **Type II censoring**: it is a special case of Type I censoring, in which the pre-fixed time C_R is determined by the failure of a pre-specified number of individuals.

(c) **Random right censoring**: this censoring situation arises when some individuals in the study experience some competing event which causes them to be removed from the study. In this situation, event and censoring times may not be independent. Depending on whether the condition of independence is fullfilled, inference must be tackled in different ways. Typical examples of independent random censoring times of the main event time of interest are accidental deaths and migration of individuals.

Left censoring

For left censored observations, time-to-event is known to be below a certain value. With C_L denoting censoring time, observed and true survival times (T and T^*, respectively) are related as $T = \max(T^*, C_L)$. Observations are pairs (T, δ) where now δ is a non-censoring indicator with value $\delta = 1$ when the event is observed and $\delta = 0$ when it is not.

Interval censoring

Time to event is somewhere in an interval $[C_L, C_R]$ which could be understod as a generalization of left and right censoring.

Truncation

Truncation occurs when only those individuals whose event time lies within a certain observational window (T_L, T_R) are observed. An individual whose event time is not in this interval is not observed

and no information on this subject is available to the investigator. This situation contrasts to censoring where there is at least partial information on the censored individuals. Because individual event times belong to the observational window, the inference for truncated data is restricted to conditional estimation (Klein and Moeschberger, 2005). This is a problem for doing frequentist inference but has a natural and simple approach within the bayesian reasoning (Armero and Bayarri, 1994).

(a) **Left truncation**: it occurs when T_R is infinite. Here we only observe those individuals whose observed event time T exceeds the truncation time Y_L, that is $T > T_L$.

(b) **Right truncation**: it occurs when $T_L = 0$, hence survival times T are only observed when $T \leq T_R$.

2.3.1 Likelihood function

The likelihood function is a key element in the inferential process. In the context of survival data analysis its construction requires special attention because it depends on the type of censoring and truncation observations. Assuming independency between lifetimes and censoring, the likelihood of the parameters of the model can be written by incorporating the corresponding elements such as:

(a) The density of of the survival time at the observed time t, $f(t)$, when the exact lifetime is known.

(b) The survival function at the censoring time, $S(C_R)$, in the case of a right-censored observation.

(c) The cumulative distribution function at the censoring time, $F(C_L) = 1 - S(C_L)$, in the case of a left-censored observation.

(d) The difference between the survival function at times $S(C_L)$ and $S(C_R)$, in the case of an interval-censored observation between C_L and C_R.

(e) The density of the survival time at observed time t conditional on the survival time is greater than T_L, $f(t)/S(T_L)$, in the case of a left truncated observation in which $T > T_L$.

(f) The density of the survival time at observed time t conditional on the survival time is less than T_L, $f(t)/(1 - S(T_R))$, in the case of a right truncated observation in which it is assumed that $T \leq T_R$.

2.4 Survival distributions

In this Section we will present the most ususal probability distributions in the survival analysis framework. Exponential, Weibull, log-normal and log-logistic distributions are introduced by means of their density, survival, hazard and cumulative hazard function. All the information included here comes from Klein and Moeschberger (2005) and Christensen *et al.* (2011).

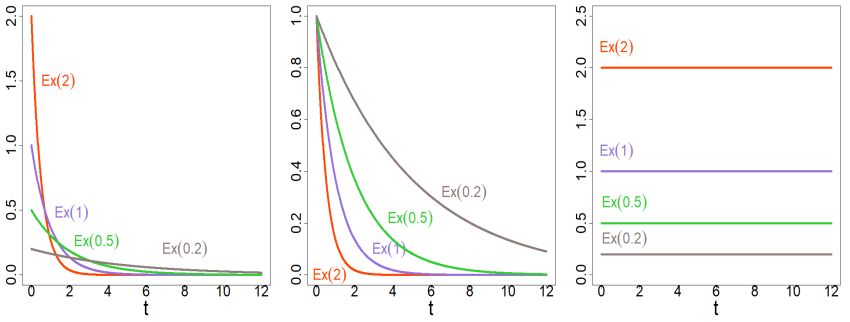
2.4.1 Exponential distribution

The exponential distribution, $(T \mid \lambda) \sim \text{Exp}(\lambda)$, with $\lambda > 0$ as the rate of failure, is a fundamental distribution in survival analysis because of its historical significance, simplicity and important properties. Its hazard, survival and density function are expressed, respectivwly, as:

- $f(t \mid \lambda) = \lambda\, e^{-\lambda t}$,

- $S(t \mid \lambda) = e^{-\lambda t}$,
- $h(t \mid \lambda) = \lambda$.

Therefore, if we assume that the hazard rate is constant then the survival times will follow an exponential distribution. Figure 2.1 shows density, survival and hazard functions for different values of the parameter λ.



(a) $f(t)$ (b) $S(t)$ (c) $h(t)$

FIGURE 2.1: Density, survival and hazard function for the exponential distribution Exp(λ) for different values of λ.

2.4.2 Weibull distribution

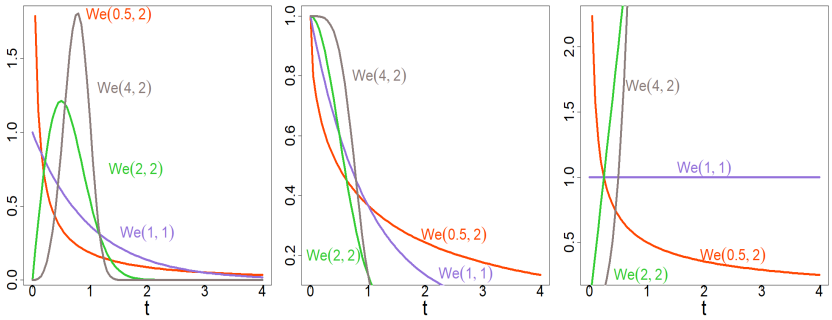
A more flexible choice than the exponential distribution is the Weibull distribution, $(T \mid \alpha, \lambda) \sim \text{We}(\alpha, \lambda)$, with $\alpha > 0$ and $\lambda > 0$, as the shape and scale parameters, respectively. Its density, survival and hazard function are:

- $f(t \mid \alpha, \lambda) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^{\alpha}}$,
- $S(t \mid \alpha, \lambda) = e^{-\lambda t^{\alpha}}$,

- $h(t \mid \alpha, \lambda) = \lambda\alpha t^{\alpha-1}$.

The Weibull distribution introduces more flexibility for the hazard function, which can now be monotonically increasing if $\alpha > 1$ or decreasing if $\alpha < 1$. Note that if $\alpha = 1$ the Weibull distribution reduces to the exponential distribution.

We illustrate some of the shapes of the density, survival and hazard functions for different α and λ values in Figure 2.2.



(a) $f(t)$ (b) $S(t)$ (c) $h(t)$

FIGURE 2.2: Density, survival and hazard function for the Weibull distribution $\text{We}(\alpha, \lambda)$ for different values of α and λ.

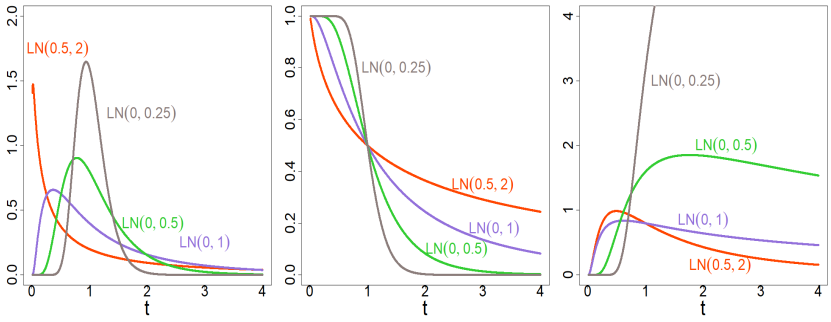
2.4.3 Log-normal distribution

The log-normal distribution, $(T \mid \mu, \sigma) \sim \text{LN}(\mu, \sigma)$, with $\mu \in \mathcal{R}$ and $\sigma > 0$ as the location and scale parameters, respectively, is another reference distribution in survival analysis. Its density, survival and hazard function are:

- $f(t \mid \mu, \sigma) = \frac{1}{t\,\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma^2}(\log(t) - \mu)^2 \Big\}$,

- $S(t \mid \mu, \sigma) = 1 - \Phi\left[\frac{(\log(t)-\mu)}{\sigma}\right]$, where $\Phi(.)$ is the cdf for the $\mathrm{N}(0,1)$,
- $h(t \mid \mu, \sigma) = f(t \mid \mu, \sigma)/S(t \mid \mu, \sigma)$.

The hazard rate of the log-normal at time 0 is zero, it increases to a maximum and then decreases to 0 as t approaches infinity. Figure 2.3 shows the density, survival and hazard function for different values of the parameters. Observe how the log-normal model is not ideal to describe the lifetime distribution, because the hazard, as t increases, is a decreasing function. This fact does not seem reasonable, except in special cases in which larger values of t are not considered (Perra, 2013).



(a) $f(t)$ (b) $S(t)$ (c) $h(t)$

FIGURE 2.3: Density, survival and hazard function for the log-normal distribution $\mathrm{LN}(\mu, \sigma)$ for different values of μ and σ.

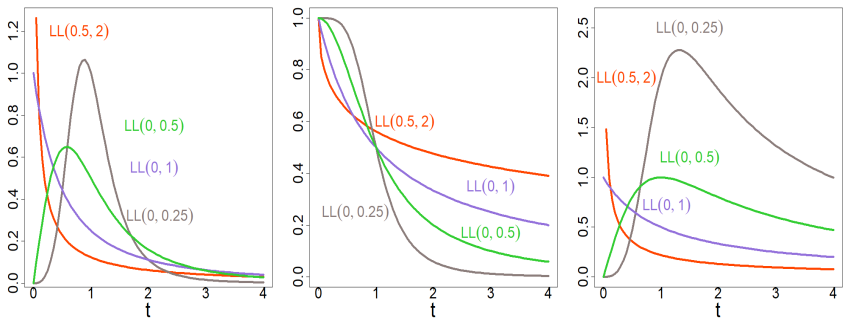
2.4.4 Log-logistic distribution

A random variable T is said to follow a log-logistic distribution T, $(T \mid \alpha, \lambda) \sim \mathrm{LL}(\alpha, \lambda)$, with $\alpha > 0$ and $\lambda > 0$, as the shape and scale

parameters, respectively. Its density, survival and hazard function are:

- $f(t \mid \alpha, \lambda) = \frac{\alpha\lambda\, t^{\alpha-1}}{(1+\alpha\, t^{\alpha})^2}$,
- $S(t \mid \alpha, \lambda) = \frac{1}{1+\alpha\, t^{\alpha}}$,
- $h(t \mid \alpha, \lambda) = \frac{\alpha\lambda\, t^{\alpha-1}}{1+\alpha\, t^{\alpha}}$.

The numerator of the hazard function is the same as the Weibull hazard function but the entire hazard has the following characteristics: monotone decreasing for $\alpha \leq 1$, while for $\alpha > 1$ the hazard rate increases initially to a maximum at time $[(\alpha - 1)/\lambda]^{1/\alpha}$ and then decreases to zero as time approaches infinity. For this reason, it presents the same problems that the log-normal model in practical applications. Figure 2.4 shows the density, survival and hazard function for different values of the parameters.



(a) $f(t)$ (b) $S(t)$ (c) $h(t)$

FIGURE 2.4: Density, survival and hazard function for the log-logistic distribution $\text{LL}(\alpha, \lambda)$ for different values of α and λ.

2.5 Survival regression models

In this Section we give a description of the survival regression models that will be used in the following chapters. Firstly, we present the most standard regression models: accelerated failure time (AFT) and Cox proportional hazard (CPH) models. Then, we introduce a general framework for to the joint models of longitudinal and survival data, and the mixture cure rate models.

2.5.1 Accelerated failure time models

The AFT model mimics the general structure of linear models. It is a log-linear-regression model for survival times T defined as,

$$\log(T) = \mu + \boldsymbol{x}'\boldsymbol{\beta} + \sigma\epsilon, \;\; \epsilon \sim F_\epsilon(\cdot) \tag{2.13}$$

where μ is an intercept parameter, $\boldsymbol{x}$ is a vector with r covariates, $\boldsymbol{\beta}$ is a vector of r regression coefficients, σ is a scale parameter, and ϵ is a random error term with known baseline cdf $F_\epsilon(\cdot)$, density $f_\epsilon(\cdot)$, survival function $S_\epsilon(\cdot)$ and hazard function $h_\epsilon(\cdot) = f_\epsilon(\cdot)/S_\epsilon(\cdot)$.

For each distribution of the error term (ϵ), there is a corresponding distribution for T. Common choices for the error distribution include the standard normal distribution which yields a log-normal AFT model, the logistic distribution, which yields a log-logistic AFT model or the extreme value distribution, which yields a Weibull AFT model. Table 2.1 summarizes common baseline distributions for ϵ and their corresponding distributions of T. Textbooks of Cox and Oakes (1984); Klein and Moeschberger (2005); Lawless (2011); Collet (2015) contain further details of AFT models.

Distribution for ϵ	$f_\epsilon(u)$	$F_\epsilon(u)$	Distribution for T
Standard Normal	$(2\pi)^{-1/2}\mathrm{e}^{-u^2/2}$	$\Phi(u)$	log-normal
Logistic(0,1)	$\mathrm{e}^u/(1+\mathrm{e}^u)^2$	$\mathrm{e}^u/(1+\mathrm{e}^u)$	log-logistic
Standard Gumbel	$\mathrm{e}^{-u}\mathrm{e}^{-\mathrm{e}^u}$	$1-\mathrm{e}^{-\mathrm{e}^u}$	Weibull

TABLE 2.1: Relationship between the distribution of the random error ϵ and the survival time T in an AFT model.

Survival, density and hazard function of exponential, Weibull, log-logistic and log-normal AFT models are described without covariate information in Section 2.4. Generally, covariate information is usually included in a linear predictor which is additive on the logarithmic transformation of the "scale" of the reference distribution. Table 2.2 shows this information and the relationships with the survival distributions introduced in Section 2.4. Note that σ parameter refered to equation (2.13) has the following relationship with the parameter α of the Weibull distribution $\sigma = 1/\alpha$.

Model	Naive distribution	AFT distribution
Exponential	$\mathrm{Ex}(\lambda)$	$\mathrm{Ex}(\exp\{-(\mu+\boldsymbol{x}'\boldsymbol{\beta})\})$
Weibull	$\mathrm{We}(\alpha,\lambda)$	$\mathrm{We}(\alpha,\exp\{-(\mu+\boldsymbol{x}'\boldsymbol{\beta})\alpha\})$
Log-normal	$\mathrm{LN}(\mu,\sigma)$	$\mathrm{LN}((\mu+\boldsymbol{x}'\boldsymbol{\beta}),\sigma)$
Log-logistic	$\mathrm{LL}(\alpha,\lambda)$	$\mathrm{LL}(\alpha,(\mu+\boldsymbol{x}'\boldsymbol{\beta}))$

TABLE 2.2: Relationships between standard parametric survival distributions and their corresponding AFT.

A key feature of the AFT models is that the effect of the covariates on survival times T is expressed in the exponential scale as $\exp\{-\boldsymbol{x}'\boldsymbol{\beta}\}$. Hence, depending on the sign of the regression

coefficients, the time is either accelerated or decelerated. This is the main reason of the name, accelerated failure time, of these models.

A relevant quantity for AFT models is the Relative Median (RM) between two individuals with covariate vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively, which is defined as:

$$\text{RM} \;=\; \exp\{(\boldsymbol{x}_1' - \boldsymbol{x}_2')\,\boldsymbol{\beta}\}.$$

The survival, density and hazard function of T can be expressed in relation to the distribution of the random error ϵ. The survival function for the time to event T is:

$$\begin{aligned}
S(t\,|\boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) &= P(T > t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) \\
&= P(\log T > \log t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) \\
&= P((\log T - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau} > (\log t - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau} \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) \\
&= P(\epsilon > (\log t - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau} \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) \\
&= S_\epsilon((\log t - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau} \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon). \qquad (2.14)
\end{aligned}$$

The density function of T is:

$$\begin{aligned}
f(t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) &= \mathrm{d}(F(t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon))/\mathrm{d}t \\
&= (\sqrt{\tau}/t)\, f_\epsilon((\log(t) - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau}). \qquad (2.15)
\end{aligned}$$

And the hazard function of T is:

$$\begin{aligned}
h(t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon) &= \frac{f(t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon)}{S(t \mid \boldsymbol{x}, \mu, \boldsymbol{\beta}, \tau, F_\epsilon)} \\
&= \frac{(\sqrt{\tau}/t)\, f_\epsilon((\log(t) - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau})}{S_\epsilon((\log(t) - \boldsymbol{x}'\boldsymbol{\beta})\sqrt{\tau})} \\
&= \sqrt{(\tau}/t)\, h_\epsilon((\log(t) - (\mu + \boldsymbol{x}'\boldsymbol{\beta}))\sqrt{\tau}). \qquad (2.16)
\end{aligned}$$

Although AFT models provide a direct extension of the classical linear model for survival data, its use is restricted by the specific distribution of the random error assumed.

2.5.2 Cox proportional hazards models

The main approach to model the effects of covariates in survival models is through the hazard rate function. Two general classes of models have been used to account for covariate effects in survival analysis, which are the family of multiplicative hazard models and the family of additive hazard rate models (Klein and Moeschberger, 2005). The multiplicative hazard model is the most popular approach and it is usually known as the Cox proportional hazards model (CPH). Focused on the hazard function, Cox (1972) introduced the proportional hazards model defined as:

$$h(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) = h_0(t) \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}, \tag{2.17}$$

in which the hazard function is expressed as the product of a baseline hazard function, $h_0(\cdot)$, and an exponential term that contains covariate information, $\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$. Note that $h_0(\cdot)$ is a completely arbitrary hazard function that determines a baseline distribution with density $f_0(\cdot)$, cdf $F_0(\cdot)$, and survival function $S_0(\cdot)$, and $\boldsymbol{\beta}$ is a vector of unknown parameters associated to covariates $\boldsymbol{x}$.

The cumulative hazard function for the time to event T is,

$$\begin{aligned} H(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) &= \textstyle\int_0^t h(s \mid h_0, \boldsymbol{x}, \boldsymbol{\beta})\, \mathrm{d}s = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\} \int_0^t h_0(s)\, \mathrm{d}s \\ &= \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\, H_0(t), \end{aligned} \tag{2.18}$$

whith $H_0(\cdot)$ as the baseline cumulative hazard function.

The survival function of T is:

$$\begin{aligned} S(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) &= \exp\left[-H(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta})\right] = \exp\left[-\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\, H_0(t)\right] \\ &= \exp\left[-H_0(t)\right]^{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}} = \left[S_0(t)\right]^{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}. \end{aligned} \tag{2.19}$$

Consequently, if $\boldsymbol{x}_1'\boldsymbol{\beta} < \boldsymbol{x}_2'\boldsymbol{\beta}$ then $S(t \mid h_0, \boldsymbol{x}_1, \boldsymbol{\beta}) < S(t \mid h_0, \boldsymbol{x}_2, \boldsymbol{\beta})$.

The density function of T is then:

$$\begin{aligned} f(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) &= h(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta})\, S(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) \\ &= h_0(t) \exp\{\boldsymbol{x}'\boldsymbol{\beta}\} [S_0(t)]^{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}. \end{aligned} \tag{2.20}$$

A relevant characteristic of the CPH model is that the hazard ratio (HR) (relative risk) of an individual with risk factor $\boldsymbol{x}_1$ having the event as compared to an individual with risk factor $\boldsymbol{x}_2$, i.e.,

$$\mathrm{HR}(\boldsymbol{\beta}, \boldsymbol{x}_1', \boldsymbol{x}_2') = \frac{h(t \mid h_0, \boldsymbol{x}_1, \boldsymbol{\beta})}{h(t \mid h_0, \boldsymbol{x}_2, \boldsymbol{\beta})} = \exp\{(\boldsymbol{x}_1' - \boldsymbol{x}_2')\boldsymbol{\beta}\}, \tag{2.21}$$

is time independent.

The key assumption of equation (2.17) comes from the above expression (2.21) and is based on the statement of the proportional hazard condition:

$$h(t \mid h_0, \boldsymbol{x}_1, \boldsymbol{\beta}) = \mathrm{HR} \cdot h(t \mid h_0, \boldsymbol{x}_2, \boldsymbol{\beta}),$$

which assumes that survival curves for individuals with distinct covariates never cross.

The plausibility of the proportional hazards assumption should always be ckecked. There are different proposals in the literature to assess it (Grambsch and Therneau, 1994), but generally it can be checked graphically by examining different types of residuals. A

typical graphical check used specially with categorical covariates, is to plot $\log(-\log(S(t, h_0, \boldsymbol{\beta} \mid \boldsymbol{x})))$ against t for the different values of $\boldsymbol{x}$. Note that $-\log(S(t, h_0, \boldsymbol{\beta} \mid \boldsymbol{x}))$ corresponds to *Cox-Snell residuals* definition (Cox and Snell, 1968). Under the proportional hazards assumption the curves should be separated by a constant vertical deviation, equal to the effect β of the explanatory variable. Thus if separation varies with time, or curves cross, the assumption is not appropriate. For more than one explanatory variables the plot could be done on combinations of possible values of the variables (Kartsonaki, 2016). Bayesian computation of this methodology is described in Wang *et al.* (2018).

A possible solution to a model for which the proportional hazards assumption seems not to be plausible is to change the set of covariates included in the model or alternatively to stratify by a categorical variable. Stratification in this context means to group individuals into strata and to allow a different baseline hazard $h_{0k}(\cdot)$ in each stratum k but to still assume that the effect of the covariates on the outcome is the same for the entire dataset. It might also be used if it is thought that there are differences between the groups defined by the strata which cannot be fully accounted for by the covariates (Kartsonaki, 2016).

To deal with interactions, another alternative can be the introduction of time dependent covariates, which is an extension of the standard Cox model (Therneau and Grambsch, 2013). Remarkably, the automatic inclusion of time-dependent covariates should be avoided because the Cox model only works properly with exogenous covariates (Rizopoulos, 2012). A time-varying covariate is considered as exogenous if its value at any time point t is not affected by an event occurring at an earlier time point $s < t$. Environmental factors as humidity, pollution levels or temperature are some standard examples. Reversely, covariates

measured for individuals in survival studies are endogenous. For a more formal definition of exogenous and endogenous time-varying covariates (see Kalbfleisch and Prentice (2002) and Rizopoulos (2012)). In contrast, joint modeling of survival and longitudinal data is a robust alternative modeling to introduce endogenous time-dependent covariates (see Section 2.5.3 for further details) in survival studies.

Regarding the AFT models described in Section 2.5.1, CPH models are more flexible in the sense that baseline hazard function, $h_0(\cdot)$, can be specified both parametrically and non-parametrically (see Chapter 4). It is worth mentioning that the case of a CPH model with a Weibull baseline hazard function is equivalent to the AFT Weibull model.

2.5.3 Joint models of longitudinal and survival data

In many medical and biological studies, longitudinal and survival data are frequently collected in the same period of time and related to the main scientific questions of the study. Given the association between both types of data, a separate analysis may lead to inefficient or biased conclussions. Hence, joint models of longitudinal and survival data are an alternative modeling option that allows in a natural way the connection of both types of information thus providing valid and efficient inferences.

Joint data analysis can aim for different objectives, longitudinal, survival or both. In particular, when the analysis has a longitudinal aim joint models allow for the introduction of informative dropouts in the longitudinal scenarios by means of survival outcomes (Wu and Carroll, 1988). By contrast, when the analysis is focused on

survival interest, the joint analysis allows for the introduction of endogenous temporal covariates defined in terms of longitudinal information (Tsiatis *et al.*, 1995). In addition, in other contexts, the main objetive recalls in the association between the longitudinal process and survival process.

Joint models for longitudinal and survival data are exppressed as a full joint probability distribution for the longitudinal ($\boldsymbol{y}$) and the survival ($\boldsymbol{s}$) process as well as for individual random effects ($\boldsymbol{b}$) and relevant parameters ($\boldsymbol{\theta}$). Particularly, that generic probability distribution is usually factorized as follows:

$$f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{\theta},)\, f(\boldsymbol{b} \mid \boldsymbol{\theta}), \tag{2.22}$$

where $\boldsymbol{x}$ are baseline covariates, $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$ is the conditional joint distribution of $\boldsymbol{y}$ and $\boldsymbol{s}$ given the random effects, parameters, and covariates and $f(\boldsymbol{b} \mid \boldsymbol{\theta})$ is the conditional distribution of the random effects given the parameters of the model. The set of covariates could also affect the particular specification of $f(\boldsymbol{b} \mid \boldsymbol{\theta})$ but it has been omitted in (2.22) for simplicity.

There are several approaches to properly model the correlation between both processes. The most popular are the *so-called* conditional models (Little, 2009), which include the random pattern-mixture and the random selection models (Sousa, 2011), the shared parameter formulation (Albert and Follmann, 2009), the random effects models (Wu and Carroll, 1988) and the joint latent class models (Proust-Lima *et al.*, 2015). In this Section, we only describe the shared parameter formulation because it is possibly the one more prevalent in the literature and, moreover, it is the approach we used in Chapter 6.

Shared-parameter models (Albert and Follmann, 2009) use random effects as the common elements that connect the survival and

the longitudinal processes, and provide conditional independence between them in the form:

$$f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})\, f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x}). \tag{2.23}$$

This generic type of model has been intensively used in a great number of studies about the human immunodeficiency virus infection carried out in the 1990s (DeGruttola and Tu, 1994). A very appealing feature of it is that the separate interpretation of the parameters in the longitudinal and the survival models is the same that the one in the joint model (Verbeke and Davidian, 2009). Furthermore, this model also allows to establish strong correlations between the longitudinal and the survival processes.

Standard joint model formulation

In essence, a joint model is made of two submodels: a model for the trajectory of the longitudinal measurements, a model for the event occurrence, and some probabilistic element that connects them. A basic version of a joint shared random effects joint model generally expresses the conditional distribution of the longitudinal process, $f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$, by means of q linear mixed-effects model (LMM) and the conditional distribution of the survival outcomes, $f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$, throughout a CPH model. Next we discuss with a more detail those standard longitudinal and survival models, that obviously have to be understood only as a basic specification with the only aim of introducing them.

Longitudinal submodel

The longitudinal submodel for the longitudinal information corresponding to the ith individual, $i = 1, \ldots, n$, in the sample

is given by

$$\begin{array}{rcl} \big(y_i(t) \mid \mu_i(t), \sigma\big) & \sim & \mathrm{N}\big(\mu_i(t), \sigma^2\big), \\ \big(\mu_i(t) \mid \boldsymbol{b}_i, \boldsymbol{\beta}\big) & = & \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, t, \\ \big(\boldsymbol{b}_i \mid \sigma_0, \sigma_1\big) & \sim & \mathrm{N}\left((0,0)^\top, \mathrm{diag}\,(\sigma_0^2, \sigma_1^2)\right), \end{array} \tag{2.24}$$

where $y_i(t)$ expresses the value of the longitudinal covariate for the ith individual at time t, which is normally distributed with mean $\mu_i(t)$ and variance σ^2. The parameters β_0 and β_1 are regression coefficients for the intercept and the slope of $\mu_i(t)$, respectively, and b_{0i} and b_{1i} are their subsequent random effects. Random effects b_{0i} and b_{1i} are considered as independent and normally distributed with mean 0 and variances σ_0^2 and σ_1^2, respectively.

Survival submodel

The time-to-event modeling is expressed through the CPH Model (see Section 2.5.2),

$$h_i\big(t \mid h_0, \boldsymbol{b}_i, \boldsymbol{\gamma}, \alpha_0, \alpha_1\big) = h_0(t) \exp\left[\boldsymbol{x}_i'\boldsymbol{\gamma} + \alpha_0 b_{0i} + \alpha_1 b_{1i} t\right], \quad t \geq 0, \tag{2.25}$$

where parameters α_0 and α_1 quantify the association between the individual characteristics of the longitudinal outcome and the risk for the survival event, $\boldsymbol{x}_i$ represents the set of baseline covariates of the ith individual and $\boldsymbol{\gamma}$ its corresponding coefficient vector.

2.5.4 Mixture cure rate models

In survival analysis, it is usually assumed that every individual in the study is susceptible to experience the event of interest. However, this assumption can be unrealistic in some specific situations in which there is a subpopulation of individuals immune to the occurrence of such event. The standard survival methodology is inappropriate

to address them and it is necessary the incorporation of a cure fraction in the survival model in order to assess the ability of a certain "treatment" to "cure". The existing statistical methodology to handle such type of data is broad and is generally referred to as cure rate models (Lambert, 2007).

In a cure model, the target population is considered as a mixture of susceptible and non-susceptible (cured) individuals. Hence, the main objective of this model is to provide a simultaneous estimation of the proportion of "inmune" individuals and of the distribution of the survival times for the "susceptible" ones. The standard mixture model (Boag, 1949; Berkson and Gage, 1952) is the most common cure survival model.

Let T a continuous and non-negative random variable that describes the time-to-event of an individual in some target population. Let Z a latent variable defined as $Z = 0$ if that individual is susceptible of experiencing the event of interest and $Z = 1$ if she/he is cured or immune for that event. If we define $1 - \eta$ and η as the probabilities for $Z = 0$ and $Z = 1$, respectively, the survival function for individuals in the cured and uncured population, $S_c(t)$ and $S_u(t)$ are

$$\begin{aligned} S_u(t) &= P(T > t \mid Z = 0) \\ S_c(t) &= P(T > t \mid Z = 1) = 1. \end{aligned}$$

The general survival function for T can be expressed in terms of a mixture of both cured and uncured populations in the form:

$$S(t \mid \eta, S_u) = \ P(T > t) = \eta + (1 - \eta)\, S_u(t). \tag{2.26}$$

It is important to point out that $S_u(t)$ is a proper survival function but $S(t)$ is not. It goes to η and not to zero when t goes to infinity.

The cure fraction η is also known as the "incidence" model and the time-to-event part ($S_u(t)$) as the "latency" model .

This modeling representation makes use of the latent variable Z, which classifies each observation to one of the two groups, uncured and cured. Hence, mixture cure models are a combination of two independent models, "incidence" and "latency". Covariate information can be included assuming two specific covariate vectors, $\boldsymbol{x}_c$ for the cure fraction and $\boldsymbol{x}_u$ for the uncured survival function. The general equation in (2.26) can be rewritten following a more bayesian notation as:

$$S(t \mid \boldsymbol{x}, \boldsymbol{\theta}) = \eta(z \mid \boldsymbol{x}_c, \boldsymbol{\theta}_c) + (1 - \eta(z \mid \boldsymbol{x}_c, \boldsymbol{\theta}_c))\, S_u(t \mid \boldsymbol{x}_u, \boldsymbol{\theta}_u) \quad (2.27)$$

with $\boldsymbol{x} = (\boldsymbol{x}_c, \boldsymbol{x}_u)$ and general parametric vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_c, \boldsymbol{\theta}_u)$.

Generally, the effect of the covariate vector $\boldsymbol{x_c}$ on the cure proportion is typically modeled using a logistic link, although other link functions such as the probit link or the complementary log-log link can be used. Covariates in the uncured survival function can be specified by means of the two main types of regression survival models, which are the AFT models and the CPH model (see Chapter 5 for further details).

2.6 Bayesian inference

2.6.1 Bayes' theorem

In bayesian inference, all types of uncertainty are always expressed in terms of probability distributions (Schoot *et al.*, 2014). There

are three essential ingredients underlying the bayesian statistics methodology (Bayes and Price, 1763; Stigler, 1986). The first ingredient refers to all knowledge available before the data is observed, which is expressed in the so-called prior distribution $\pi(\boldsymbol{\theta})$, a probability distribution that contains all the available prior information about the parametric vector $\boldsymbol{\theta}$. The second ingredient is the information from the data $\mathcal{D}$, whose relationship with the unknown parametric vector is expressed in terms of the likelihood function of $\boldsymbol{\theta}$ ($\mathcal{L}(\boldsymbol{\theta})$). The third ingredient is obtained by combining the first two elements. Prior distribution and the likelihood function of $\boldsymbol{\theta}$ are combined via Bayes' theorem and summarized by the so-called posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$, which is a compromise between the prior knowledge and the experimental evidence. The posterior distribution reflects the updated knowledge, balancing prior knowledge with observed data.

All these ingredients are part of the Bayes' theorem, which states, that our updated understanding of the parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence of the data, i.e.,

$$\pi(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathcal{D})} \propto \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.28}$$

Here, $m(\mathcal{D}) = \int \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the normalising constant, also called model evidence or marginal likelihood of the data $\mathcal{D}$ (Robert, 2007). This constant makes the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ integrate to one. However, as it is referred in the second expression of equation (2.28), $m(\mathcal{D})$ can often be ignored and work in terms of proportionality rather than equality.

2.6.2 Sampling from the posterior distribution: MCMC and INLA

Although the basis of the bayesian methodology is simple and intuitive, its application to complex real problems in non-standard probabilistic scenarios and high-dimensional problems was initially very difficult (Robert, 2014). Particularly, in a great number of models and applications, $m(\mathcal{D})$ does not have an analytic closed expression, and therefore the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ does not have a closed form (Ibrahim *et al.*, 2001).

The intensive development of Markov chain Monte Carlo (MCMC) sampling methods during last decades has made bayesian inference a feasible methodology to solve properly many statistical problems. Furthermore, other bayesian procedures such as the integrated nested Laplace approximations (INLA) (Rue *et al.*, 2009) have gained importance. In the following paragraphs we describe briefly, the most two habitual MCMC algorithms: Metropolis-Hastings sampling (Metropolis *et al.*, 1953; Hastings, 1970) and Gibbs sampler (Gelfand and Smith, 1990), as well as the INLA approximation and highlight the main differences between both methodologies.

Markov chain Monte Carlo methods

MCMC simulation methods are a class of stochastic algorithms for sampling from posterior distributions. These methods allow to draw samples from some probability distribution without knowing their exact density. Therefore, with MCMC we do not get a closed form of the posterior but a sample of values from it. These samples can then be directly used to obtain inferences upon key derived quantities of interest (Jackson, 2015).

Metropolis-Hastings

The Metropolis-Hastings (M-H) algorithm is possibly the most general and simplest MCMC procedure. It basically constructs a Markov chain whose limit distribution is the target density, that is, the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$. The M-H algorithm begins with an initial value $\boldsymbol{\theta}^{(0)}$ and specifies a rule for simulating values from the target distribution based on a proposal density (a Markovian kernel) $q(\cdot \mid \cdot)$. The algorithm can be described as

Step 0. Select and arbitrary starting point $\boldsymbol{\theta}^{(0)}$ and consider $m = 0$.

Step 1. Simulate a candidate value $\boldsymbol{\theta}^{(*)}$ from the proposal density $q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(m)})$ and an observation u from the uniform distribution $\mathrm{U}(0, 1)$.

Step 2. Compute the acceptance probability

$$a(\boldsymbol{\theta}^{(*)}, \boldsymbol{\theta}^{(m)}) = \min\Big(\frac{\pi(\boldsymbol{\theta}^{(*)} \mid \mathcal{D})\, q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(*)})}{\pi(\boldsymbol{\theta}^{(m)} \mid \mathcal{D})\, q(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(m)})}, 1\Big),$$

and set $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(*)}$ if $u \leq a(\boldsymbol{\theta}^{(*)}, \boldsymbol{\theta}^{(m)})$ and $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)}$ otherwise. Note that both posterior probabilities $\pi(\boldsymbol{\theta}^{(*)} \mid \mathcal{D})$ and $\pi(\boldsymbol{\theta}^{(m)} \mid \mathcal{D})$ can be approximated using Bayes' rule, hence they would be proportional to the product of their corresponding likelihood function and prior distribution.

Step 3. Consider $m = m + 1$, and return to Step 1.

Gibbs sampler

Let $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ be the posterior distribution of the parametric vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ given data $\mathcal{D}$. The Gibbs sampler is an algorithm which, at each iteration, draws a sample from the distribution of each component of $\boldsymbol{\theta}$ conditional on the rest

of components, i.e., the full conditional. The algorithm can be described as:

STEP 0. We start with an arbitrary vector $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ and consider $m = 0$.

STEP 1. Simulate $\boldsymbol{\theta}^{(m+1)} = (\theta_1^{(m+1)}, \ldots, \theta_p^{(m+1)})$ as follows:

- Simulate $\theta_1^{(m+1)}$ from $\pi(\theta_1|\theta_2^{(m)}, \ldots, \theta_p^{(m)}, \mathcal{D})$
- Simulate $\theta_2^{(m+1)}$ from $\pi(\theta_2|\theta_1^{(m+1)}, \theta_3^{(m)} \ldots, \theta_p^{(m)}, \mathcal{D})$
- Simulate $\theta_3^{(m+1)}$ from $\pi(\theta_3|\theta_1^{(m+1)}, \theta_2^{(m+1)}, \theta_4^{(m)} \ldots, \theta_p^{(m)}, \mathcal{D})$
  ...
- Simulate $\theta_p^{(m+1)}$ from $\pi(\theta_p|\theta_1^{(m+1)}, \theta_2^{(m+1)}, \ldots, \theta_{p-1}^{(m+1)}, \mathcal{D})$

STEP 2. Consider $m = m + 1$, and return to Step 1.

Integrated nested Laplace approximation

The INLA approximation, proposed by Rue *et al.* (2009) and implemented in the `R-INLA` package, is a numerical approximation for bayesian inference. INLA uses Laplace approximations to approximate the marginal posterior distribution of the relevant components in $\boldsymbol{\theta}$ (Laplace, 1986; Tierney and Kadan, 1986).

INLA is applicable to a very popular subset of structured additive regression models named latent Gaussian models LGM) (Rue and Held, 2005). Specifically, it can be applied only if these models can be expressed as latent Gaussian Markov random field (GMRF) because of their important computational properties (for details, see Rue *et al.* (2009)). Under these assumptions, they are a special class of bayesian additive models that cover a wide range of applications

(Rue *et al.*, 2017), including survival models (see Martino *et al.* (2011)).

The structure and main elements of the INLA approach are summarised below. Let us assume a set of n variables $\boldsymbol{T} = (T_1, \ldots, T_n)$ mutually conditionally independent given a latent GMRF, $\boldsymbol{\theta}$, and a set of hyperparameters $\boldsymbol{\phi}_1$. The latent GMRF, $\boldsymbol{\theta}$, depends on hyperparameters $\boldsymbol{\phi}_2$ and can include effects of different types (regression coefficientes, random effects, seasonal effects, etc). The joint posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$, after data $\mathcal{D}$ have been observed can be written as

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathcal{D}) &\propto \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_1)\, \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ &\propto \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_1)\, \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}_2)\, \pi(\boldsymbol{\phi}). \end{aligned} \tag{2.29}$$

The posterior marginal distributions of interest are $\pi(\theta_m \mid \mathcal{D})$ and $\pi(\phi_j \mid \mathcal{D})$. They can be obtained as

$$\pi(\theta_m \mid \mathcal{D}) = \int \pi(\theta_m \mid \boldsymbol{\phi}, \mathcal{D})\, \pi(\boldsymbol{\phi} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\phi}, \tag{2.30}$$

$$\pi(\phi_j \mid \mathcal{D}) = \int \pi(\boldsymbol{\phi} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\phi}_{-j}, \tag{2.31}$$

where $\boldsymbol{\phi}_{-j}$ are all elements in $\boldsymbol{\phi}$ except ϕ_j.

INLA makes use of the Laplace approximation (Rue *et al.*, 2009) to obtain approximations $\tilde{\pi}(\boldsymbol{\phi} \mid \mathcal{D})$ and $\tilde{\pi}(\theta_m \mid \phi, \mathcal{D})$ of $\pi(\boldsymbol{\phi} \mid \mathcal{D})$ and $\pi(\theta_m \mid \phi, \mathcal{D})$, respectively. Posterior distributions $\pi(\theta_m \mid \mathcal{D})$ are approximated by numerical integration as:

$$\tilde{\pi}(\theta_m \mid \mathcal{D}) \approx \sum_k \tilde{\pi}(\theta_m \mid \boldsymbol{\phi}_k, \mathcal{D})\, \tilde{\pi}(\boldsymbol{\phi}_k \mid \mathcal{D})\, \Delta_k, \tag{2.32}$$

where $\boldsymbol{\phi}_k$ are points in the parametric space Φ and Δ_k integration weights. Posterior marginal distribution $\tilde{\pi}(\phi_k \mid \mathcal{D})$ can also be derived by numerical integration according to the expression in equation (2.31).

Chapter 3

Bayesian survival analysis in plant breeding and food microbiology

3.1 Introduction

Bayesian survival analysis has increased its popularity in many fields of research. Its direct and intuitive quantification of the uncertainty through explicit probabilistic inference, the flexibility in the modeling, and the existence of specific software such as WinBUGS (Lunn *et al.*, 2000), JAGS (Plummer, 2003) or INLA (Rue *et al.*, 2009) have made possible its feasibility for both practitioners and researchers.

We dedicate this Chapter to highlight the strong potential of the bayesian methodology for dealing with survival studies in the framework of two different scientific areas such as plant breeding and food microbiology. Firstly, we illustrate the use of accelerated failure time modeling (AFT) to evaluate a new plant variety for resistance

and tolerance to a specific virus. MCMC simulation methods are used to estimate the posterior distribution of the parameters of interest and the frequentist approach is also considered to compare the results. Secondly, we implement a Cox prorportional hazard (CPH) model to assess virulence changes in a foodborne pathogen as a consequence of different frequencies of application of a new preservation treatment. Posterior inference is made with the INLA approximation and MCMC to check how both methodologies behave. We show a detailed comparison in which we highlight strengths and weaknesses of both techniques.

3.2 Evaluating a new plant variety against a virus disease

Virus diseases are one of the most important threats to large-scale production of crops causing important economical losses and undermining sustainability (Gallitelli, 2000). According to Lecoq *et al.* (2004), introgression of genes conferring resistance and/or tolerance by plant breeding is the most efficient and simplest strategy for disease control. Most breeding programs are aimed at finding and implementing resistance based on the absence of systemic infection. However, new proposals suggest that considering degrees of resistance (reduction of virus infectivity and/or multiplication), and/or tolerance (reduction of symptom severity) may be useful to rescue valuable phenotypes (Soler *et al.*, 2015).

The main scientific question addressed in this study was to evaluate a new plant variety, characterised by its genotype, for resistance and tolerance to a specific virus through a comparison with other

well-known varieties. Resistance was defined as the time, in days, from virus inoculation to virus infection and, tolerance, as the time, in days, from virus inoculation to the appearance of severe symptoms.

3.2.1 Resistance and tolerance data

Three genotype characterizations ($G1$ for susceptible plants, $G2$ for resistant, and $G3$ for plants to evaluate) and two different virus biotypes ($V1$ with capacity to only infect plants $G1$ plants, and $V2$ with a resistance-breaking capacity to infect $G2$ plants) were considered. A total of 180 plants with genotypes $G1$, $G2$ and $G3$ were inoculated with virus biotypes $V1$ and $V2$ according to a balanced two-factor factorial design which generated six groups with 30 plants each.

All plants were evaluated in terms of resistance and tolerance at monitoring times 7, 14, 21, and 28 days post inoculation (*dpi*). Hence, both resistance and tolerance times were considered interval-censored when the event of interest occurred between two consecutive monitoring times or right-censored when it was not observed at the end of the study (28 *dpi*). In both survival processes time zero was synchronised with the time at which the virus was inoculated.

Tables 3.1 and 3.2 show the observed resistance and tolerance frequency, respectively, for the plants of each of the six groups considered. Groups $G2V1$ and $G3V1$ contain a great number of individuals right censored for both events. This is not the case of the observations in the $G1V1$ group where all plants experienced both events before the end of the study. Remarkably, the number of right censored plants for virus $V2$ was at most 7 in nearly all groups.

However, in the G3V2 group neither of the plants developed severe symptoms.

Genotype	Virus	$(0, 7]$	$(7, 14]$	$(14, 21]$	$(21, 28]$	$28 <$
$G1$	$V1$	8	14	7	1	0
	$V2$	21	9	0	0	0
$G2$	$V1$	0	0	0	0	30
	$V2$	2	12	3	6	7
$G3$	$V1$	1	2	1	3	23
	$V2$	2	12	3	6	7

TABLE 3.1: Frequency of resistance survival times regarding to plant genotype and virus biotype.

Genotype	Virus	$(0, 7]$	$(7, 14]$	$(14, 21]$	$(21, 28]$	$28 <$
$G1$	$V1$	0	2	23	5	0
	$V2$	1	3	26	0	0
$G2$	$V1$	0	0	0	0	30
	$V2$	0	4	11	15	0
$G3$	$V1$	0	0	0	0	30
	$V2$	0	0	0	0	30

TABLE 3.2: Frequency of tolerance survival times regarding to plant genotype and virus biotype.

3.2.2 Modeling

Resistance and tolerance times are analysed independently through an accelerated failure time (AFT) model (see Section 2.5.1 for

further details of this type of modeling):

$$\begin{aligned}\log T_i &= \mu + \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma\epsilon_i, \;\; i.i.d.\; \epsilon_i \sim S_\epsilon(\cdot) \\ &= \beta_{G1V1} + \beta_{G2V1}\, I_{G2V1}(i) + \beta_{G3V1}\, I_{G3V1}(i) + \beta_{G1V2}\, I_{G1V2}(i) + \\ &\quad \beta_{G2V2}\, I_{G2V2}(i) + \beta_{G3V2}\, I_{G3V2}(i) + \sigma\, \epsilon_i,\end{aligned} \tag{3.1}$$

where T_i can be for resistance or tolerance time for individual i, $i = 1, \ldots, n$. Note that, in both modelings, the baseline covariates are indicator variables for identifying the relevant plant genotype and virus biotype combination in the study. $G1$ plants inoculated with biotype $V1$ ($G1V1$) was considered the reference category, and hence, it was introduced as the intercept term $\mu = \beta_{G1V1}$. The distribution $S_\epsilon(\cdot)$ was specified as a standard Gumbel distribution implying a conditional (on the vector $\boldsymbol{\beta}$ of all regression coefficients and σ) Weibull survival model for T_i with shape $\alpha = 1/\sigma$ and scale $\lambda(\mu, \boldsymbol{\beta}) = \mathrm{e}^{-(\mu + \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma}$ parameters (Christensen *et al.*, 2011).

Both bayesian models were completed with the specification of a prior distribution for their corresponding parameters. A prior independent default scenario was considered. The marginal prior distribution for each regression coefficient $\beta_{G_jV_k}$, $j = 1, 2, 3$, $k = 1, 2$, was elicited as a normal distribution centered at zero and a wide variance, $\pi(\beta_{G_jV_k}) = \mathrm{N}(0, 1000)$. A uniform distribution $\mathrm{Un}(0, 100)$ was selected as the marginal prior distribution for σ. Note that all marginal prior distributions are scarcely informative. This fact is even more evident due to the logarithmic scale for the survival times that compacts the information.

The likelihood function of $(\mu, \boldsymbol{\beta}, \sigma)$ for the observed data $\mathcal{D}$ is the product of the likelihood function for each individual. Individual time-to-event data are right or interval censored. A right censored data corresponds to individuals that have not experienced the

event of interest at the end of the period of the study, 28 *dpi*. Its contribution to the likelihood is $P(T_i > 28 \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma)$, its survival function at 28 *dpi*. Interval censored data for individual i arises when the event of interest occurred between the current monitoring time (t_{iu}) and the previous one (t_{il}) and its contribution to the likelihood is $S_i(t_{il} \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma) - S_i(t_{iu} \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma)$ with $S_i(t \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma)$ as the survival function for individual i, $S_i(t \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma) = \exp\{-t^{(1/\sigma)}\, \mathrm{e}^{-(\mu+\boldsymbol{x}_i'\boldsymbol{\beta})/\sigma}\}$. Consequently

$$\begin{aligned} \mathcal{L}(\mu, \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} \mathcal{L}_i(\mu, \boldsymbol{\beta}, \sigma) \qquad (3.2) \\ &= \prod_{i}^{\mathcal{R}} S_i(28 \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma) \prod_{i}^{\mathcal{I}} [S_i(t_{il} \mid \boldsymbol{x}_i, \mu, \boldsymbol{\beta}, \sigma) - S_i(t_{iu} \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma)], \end{aligned}$$

where $\mathcal{R}$ is the set of right (interval) censored data and $\mathcal{I}$ is the set of interval censored data.

3.2.3 Posterior inferences

For each model, the posterior distribution of the parameters was estimated by means of MCMC methods with the WinBUGS software (Lunn *et al.*, 2000). Specifically, simulations were run considering three Markov chains with 100,000 iterations and a burn-in period with 10,000. In addition, the chains were thinned by keeping every 10th iteration in order to reduce autocorrelation in the saved sample and avoid space computer problems. Trace plots of the simulated values of the chains seem to appear overlap one another, indicating stabilization. Convergence of the chains to the posterior distribution was assessed using the potential scale reduction factor, $\hat{R}$, and the effective number of independent simulation draws, neff. In all cases, the $\hat{R}$ values were equal or close to 1 and neff > 100,

thus indicating that the distribution of the simulated values between and within the three chains was practically identical, and that enough MCMC samples had been obtained, respectively (Gelman and Rubin, 1992).

Both models were also estimated using the frequentist approach in order to compare bayesian and frequentist results. Frequentist estimation was performed through the function `survreg` in the `survival R` package (Therneau, 2015; Therneau and Grambsch, 2013).

Results are arranged in two parts for tolerance and resistance separately. However, as both survival times were studied with the same type of model, outcomes are both presented following the same scheme to detect similarities and differences between them. We focused on the effect of covariates on the estimated probabilities of remaining free of infection and free of the appearance of severe symptoms. A Section for comparing bayesian and frequentist results is also included.

Resistance

Posterior summaries of the estimated posterior distribution for the regression coefficients and the error scale parameter are shown in Table 3.3. Genotype plants $G1$ shows the shortest resistance times among the plants inoculated with $V1$. Posterior probabilities $P(\beta_{G2V1} > 0 \mid \mathcal{D}) = 1$ and $P(\beta_{G3V1} > 0 \mid \mathcal{D}) = 1$ provide strong evidence that $G2$ and $G3$ plants show a better resistance behaviour compared to $G1$ under $V1$ infection. In addition, genotype $G2$ is the most resistant variety with $P(\beta_{G2V1} > \beta_{G3V1} \mid \mathcal{D}) = 1$ despite of the wide variability of its estimated coefficient. Under biotype infection $V2$, resistance is worse for all genotypes although $G3$ genotype

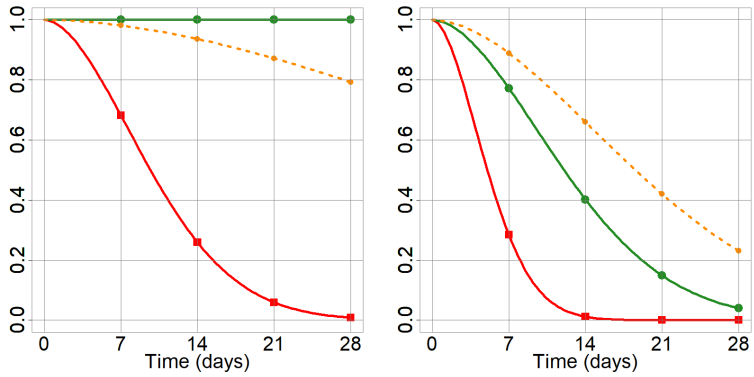
improves resistance in relation to $G2$ ($P(\beta_{G3V2} > \beta_{G2V2} \mid \mathcal{D}) = 0.99$).

Parameter	Mean	Sd	$CI_{95\%}$	$P(\cdot > 0)$
β_{G1V1}	2.27	0.12	[2.02, 2.51]	1.00
β_{G2V1}	4.97	1.24	[2.62, 7.00]	1.00
β_{G3V1}	1.64	0.26	[1.15, 2.24]	1.00
β_{G1V2}	-0.66	0.15	[-0.96, -0.36]	0.00
β_{G2V2}	0.22	0.16	[-0.08, 0.55]	0.93
β_{G3V2}	0.65	0.16	[0.34, 0.98]	1.00
σ	0.55	0.06	[0.46, 0.67]	

TABLE 3.3: Summary of the MCMC estimated posterior distribution for the resistance model: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive. Group $G1V1$ is the reference category.

Figure 3.1 shows the posterior mean of the probability of remaining free of infection over time (from 0 to 28 dpi) for each genotype plant under virus infection $V1$ and $V2$. For both virus biotypes, $G1$ plants show the lowest probability values in all the monitoring times (7, 14, 21 and 28 *dpi*). Plants $G2$ exhibit the highest probability values under $V1$ infection and $G3$ under $V2$ infection. Remarkably, the pattern of the differences between genotypes $G2$ and $G3$ under virus $V1$ and $V2$ is very different. Under $V2$ infection, differences among posterior probabilities (in favour of no infection for $G3$) are stable enough from 14 dpi and for any time they exceed the value of 0.27. In the case of $V1$, there is an increasing difference over time in favour of no infection for $G2$ with a maximum distance of 0.21 at 28 *dpi*. Posterior mean of the probability of remaining free of infection decreases with time for all genotypes under infection $V2$ highlighting $V2$ resistance-breaking capacity. At 14 *dpi* (the midpoint of the monitoring times), the estimated mean of that probability is 0.26,

1, and 0.93 for groups $G1V1$, $G2V1$ and $G3V1$, and 0.02, 0.40, and 0.65 for $G1V2$, $G2V2$ and $G3V2$, respectively.



(a) $V1$ infection (b) $V2$ infection

FIGURE 3.1: Posterior mean of the probability of remaining free of infection over time (from 0 to 28 *dpi*) for $G1$ (in solid red line), $G2$ (in solid green line) and $G3$ (in dotted orange line) genotypes under infection $V1$ and $V2$. Monitoring times 7, 14, 21 and 28 *dpi* are highlighted with dots.

Tolerance

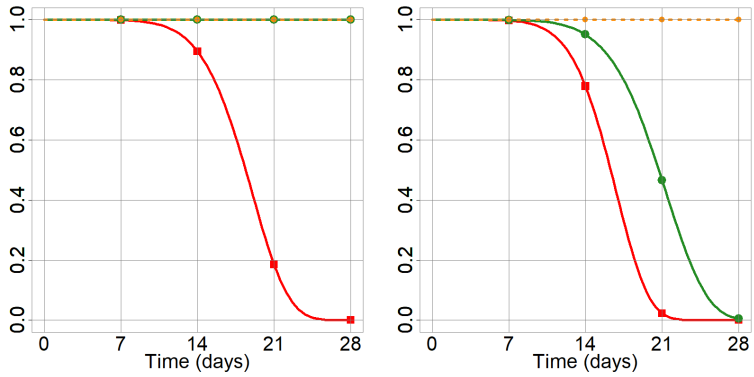
Table 3.4 shows a summary of the posterior distribution for the regression coefficients and the error scale parameter in the AFT model for tolerance times. Estimation in terms of the sign of the posterior outcomes are quite similar to the subsequent results of the resistance model, but we can also appreciate some noticeable differences. It is worth mentioning the similar effect of biotype $V1$ on $G2$ and $G3$ plants and the overwhelming estimated effect related to $G3$ genotype under $V2$ infection. Plants $G3$ show a similar tolerance pattern for both virus biotypes.

The posterior mean of the probability of remaining free of the appearance of severe symptoms during the period of the study (from

Parameter	Mean	Sd	$CI_{95\%}$	$P(\cdot > 0)$
β_{G1V1}	2.91	0.04	[2.84, 2.98]	1.00
β_{G2V1}	3.90	1.77	[1.08, 6.95]	1.00
β_{G3V1}	4.09	1.74	[1.16, 6.93]	1.00
β_{G1V2}	-0.12	0.05	[-0.23, -0.03]	0.00
β_{G2V2}	0.12	0.05	[0.02, 0.21]	1.00
β_{G3V2}	4.00	1.81	[1.07, 6.89]	1.00
σ	0.15	0.02	[0.12, 0.19]	

TABLE 3.4: Summary of the $MCMC$ approximate posterior distribution for the tolerance model: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive. Group $G1V1$ is the reference category.

0 to 28 *dpi*) for biotype and virus groups is displayed in Figure 3.2. Under $V1$ infection, plants $G2$ and $G3$ exhibit similar probability values, very close to one. They are higher than the subsequent for $G1$ values, which show a decreasing trend with a strong slope between 14 and 21 *dpi*'s. Plants $G1$ and $G3$ behave analogously under $V1$ and $V2$ infection. However, probabilities for $G2$ are very different for both virus: $G2$ is similar to $G3$ for infection $V1$ but its behaviour changes under $V2$ infections. In particular, $G2$ shows a decreasing probability of remaining free of infection from 14 *dpi* on, which at the end of the monitoring time is equal to the value of variety $G1$. At 14 *dpi* (the midpoint of the monitoring times), the posterior mean of the probability of remaining free of the appearance of severe symptoms is 0.89, 1, and 1 for $G1V1$, $G2V1$ and $G3V1$ crosses, and 0.77, 0.95, and 1 for $G1V2$, $G2V2$ and $G3V2$, respectively.



(a) $V1$ infection (b) $V2$ infection

FIGURE 3.2: Posterior mean of the probability of remaining free of the appearance of severe symptoms over time (from 0 to 28 *dpi*) for $G1$ (in solid red line), $G2$ (in solid green line) and $G3$ (in dotted orange line) genotypes under infection $V1$ and $V2$. Monitoring times 7, 14, 21 and 28 *dpi* are highlighted with dots.

Resistance and tolerance: frequestist and bayesian modelings

Results in this subsection are focused on the frequentist approach to the resistance (Table 3.5) and the tolerance (Table 3.6) AFT model. Both tables try to reproduce the structure of Table 3.3 (bayesian resistance model) and Table 3.4 (bayesian tolerance model) with regard to the frequentist concepts (estimate, standard error, 95% confidence interval, and p-value) which could be considered (in a not rigorous and very broad sense) as somehow *parallel* to bayesian posterior mean, standard deviation, 95% credible interval, and posterior probability for a positive regression coefficient.

At first glance, most of the numerical (but not conceptual) results provided by the two inferential approaches seem not to be very different. But a more leisurely observation of them shows relevant differences in the punctual and interval estimation of the regression

Parameter	Estimate	Sd. error	$CI_{95\%}$	p-value
β_{G1V1}	2.47	0.10	[2.27, 2.67]	< 0.05
β_{G2V1}	11.54	2523.17	[-4933.79, 4956.87]	1.00
β_{G3V1}	1.55	0.24	[1.09, 2.02]	< 0.05
β_{G1V2}	-0.65	0.15	[-0.94, -0.35]	< 0.05
β_{G2V2}	0.22	0.14	[-0.06, 0.49]	0.13
β_{G3V2}	0.63	0.15	[0.34, 0.93]	< 0.05
$\log(\sigma)$	-0.65	0.10		< 0.05

TABLE 3.5: Summary of the regression parameter estimation for the resistance model under the frequentist approach: estimate, standard error, 95% confidence interval and p-value. Group $G1V1$ is the reference category.

Parameter	Estimate	Sd. error	$CI_{95\%}$	p-value
β_{G1V1}	2.97	0.03	[2.90,3.03]	<0.05
β_{G2V1}	3.60	1710	[-3340.72,3347.92]	1.00
β_{G3V1}	3.60	1710	[-3340.72,3347.92]	1.00
β_{G1V2}	-0.12	0.05	[-0.22,-0.02]	<0.05
β_{G2V2}	0.12	0.05	[0.02,0.21]	<0.05
β_{G3V2}	3.60	1710	[-3340.72,3347.92]	1.00
$\log(\sigma)$	-1.92	0.11		<0.05

TABLE 3.6: Summary of the regression parameter estimation for the tolerance model under the frequentist approach: estimate, standard error, 95% confidence interval and p-value. Group $G1V1$ is the reference category.

coefficients, particularly in those groups in which all the observations were right censored. This is the case of the $G2V1$ group for the resistance model and groups $G2V1$, $G3V1$ and $G3V2$ for tolerance. In the case of the resistance model for group $G2V1$, the punctual frequentist and bayesian estimates are very different 11.54 and 4.97 *dpi*, respectively. But the more relevant differences are in variability, with enormous confidence intervals and p-values close to 1. This is

also the case of the frequentist results for tolerance in groups $G2V1$, $G3V1$ and $G3V2$, all having the same enormous 95% confidence interval.

The inferences achieved indicate that the AFT frequentist model has difficulties in the estimation corresponding to groups with data with very little signal. This is not the case of the bayesian analyses that accommodate very well for the particular data of the study. This situation agrees with the general comment in Ibrahim *et al.* (2001) about the advantages of the bayesian methodology over the frequestist in survival analysis with regard to estimation problems in the presence of complex censoring data. Moreover, the bayesian results are more compatible with the agronomic expectations based on preliminary studies.

3.2.4 Discussion

Agronomical conclusions indicated that genotype G3 did not suppose an improvement in terms of resistance with respect to G2. However, they showed a very high tolerance to the specific virus considered. This process is not easy because it is necessary to identify the sources of tolerance and subsequently select the appropriate procedures to be included in the study.

Bayesian survival regression models provide a useful tool for quantifying differences among the different genotype $\times$ virus biotype groups as well as to assess the degree of resistance and tolerance. They also make possible the incorporation of censoring and truncation mechanisms that are frequent in this type of studies with good inference results. Frailty models (Christensen *et al.*, 2011) are a future line of work in order to approach a more suitable model that can better capture all the uncertainties of the real problem.

3.3 Assessing virulence changes in a foodborne pathogen

Increased consumption of fresh fruits and vegetables has been associated with a rise in foodborne disease outbreaks (Olaimat and Holley, 2012). *Salmonella spp.*, specifically the serotype *Salmonella enterica* serovar Typhimurium (*S.* Typhimurium), is one of the most habitual serotypes related to salnonellosis outbreaks.

Different alternative preservation treatments have been developed to reduce or eliminate *S.* Typhimurium load and also preserve food properties. The addition of bioactive substances from nature or agroindustrial by-products with antimicrobial effect (Viuda-Martos *et al.*, 2008) as well as the application of non-thermal treatments (Mosqueda-Melgar *et al.*, 2012) are some of the innovative techniques that are currently being tested against *S.* Typhimurium. However, these treatments have important drawbacks because their repeated use can generate serious antimicrobial resistance problems (Kisluk *et al.*, 2013; Vanlint, 2013), such as changes in virulence patterns.

Host organisms are frequently used to study the multi-factorial nature of the microbial pathogenicity and, in consequence, to assess virulence. The natural feeding (bacteria) of the host organisms is replaced by the pathogen organism which is going to be assessed. Hence, virulence assays are based on the study of the ability of the foodborne pathogen to kill the host organism and, in particular, virulence is assessed by means of the analysis of host organism life span after infection.

In the case of *S.* Typhimurium, *Caenorhabditis elegans* (*C. elegans*), a nematode that inhabits soils around the world, is considered a

good model (Aballay *et al.*, 2000; Labrousse *et al.*, 2000) to explore the pathogenesis of *S.* Typhimurium by making the worms feed on the pathogen and not *Escherichia coli* (strain OP50), its usual laboratory food.

A common goal in virulence studies is the comparison of survival profiles among the different treatment and control groups. Most of the studies in the area only use Kaplan-Meier estimation to construct graphs of the observed survival curves and the log-rank test (Chai-Hoon *et al.*, 2010; Sem and Rhen, 2012) to compare survival curves from two different groups. Survival regression models are rarely used and Cox proportional hazards (CPH) models are in general the frequent option.

The standard of the CPH models in virulence studies is based on the so called *partial likelihood* that does not take into account the specification of the baseline hazard function (Cox, 1972) (See Yang *et al.* (2011); Han *et al.* (2016); Ziehm and Thornton (2013) for online applications OASIS, OASIS2, and `SurvCurv`, respectively). This approach makes impossible the estimation of all the outcomes of interest, such as hazard and survival curves for relevant covariate patterns (Royston, 2011). Additionally, in the context of microbial virulence the baseline hazard function, $h_0(t)$, can be considered as a meaningful measure of the natural course of the infection.

3.3.1 Virulence data

Virulence data came from an experiment designed to assess the effect of the use of a cauliflower by-product infusion treatment in *S.* Typhimurium virulence behavior. One and three expositions to the treatment were evaluated as well as a pathogen population non exposed to the treatment that was considered as the control group.

Each group (*S.* Typhimurium non treated, *S.* Typhimurium treated once, and *S.* Typhimurium treated three times) was the source of nutrition of 250 synchronized young adult nematodes. All worms were kept in identical environmental conditions (20^oC) during all their lifespan (three weeks as maximum approximately). Virulence was defined in terms of their subsequent survival time which was individually evaluated at intervals of between 48 and 56 hours.

Worms were placed in plates to facilitate its monitoring. Survivor worms at the times of each observation period were always transferred to a new plate to avoid confusions and interferences with the eggs laid by them. The experiment finished when all the worms eventually died. It is worth to mentioning that a small amount of worms accidentally died during the transfer process (see Sanz-Puig *et al.* (2017) for more details about the validation and special conditions of the experiment).

Due to the experimental collection strategy, survival data accounted for interval and right-censored patterns. The vast majority of the data were interval-censored, which means that the relevant information about the subsequent survival times is that the events can occur between two consecutive monitoring times and in this study are very closed to each other. The number of right censored data, which belonged to survival times of worms killed accidentally during the transfer between plates, was scarce: 0, 1, and 5 individuals for *S.* Typhimurium non treated, *S.* Typhimurium treated once and *S.* Typhimurium treated three times, respectively.

Figure 3.3 shows the individual life span, in days, of the worms of the sample (ranked in increasing order) according to each *S.* Typhimurium population considered. Lifetimes pattern for individuals feed on *S.* Typhimurium were generally lower than the ones of the experimental groups, with median survival time 5 days

compared to medians 8 and 8 days for *S.* Typhimurium exposed once or three times, respectively. Differences between the survival patterns of the worms feed on *S.* Typhimurium treated one and three times are practically imperceptible and seem to indicate a high degree of similarity between them.

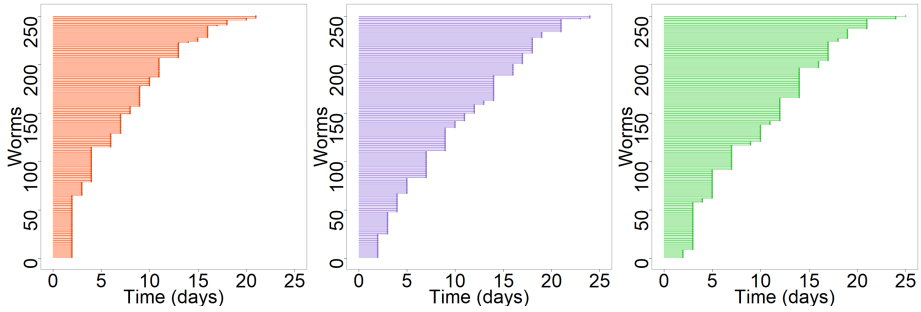


FIGURE 3.3: Ranked survival times, in days, for individuals feed on a) untreated *S.* Typhimurium, b) *S.* Typhimurium exposed one time, and c) *S.* Typhimurium exposed three times to the antimicrobial treatment.

3.3.2 Modeling

Virulence times (worms lifespan) are modelled by means of a CPH model (see Section 6 on Chapter 2 for further details of these models),

$$\begin{aligned} h_i(t \mid h_0, \boldsymbol{x}_i, \boldsymbol{\beta}) &= h_0(t) \exp\{\boldsymbol{x}_i' \boldsymbol{\beta}\} \\ &= h_0(t) \exp\{\beta_1 I_1(i) + \beta_3 I_3(i)\}, \end{aligned} \tag{3.3}$$

where the baseline hazard function is specified by means of a Weibull distribution with hazrad function $h_0(t \mid \alpha, \lambda) = \lambda \alpha t^{\alpha-1}$, $\boldsymbol{\beta} = (\beta_1, \beta_3)$, and covariate vector $\boldsymbol{x}_i$ which includes both treatment

groups in terms of dummy variables, $I_1(i)$ for *S.* Typhimurium treated one time and $I_3(i)$ for *S.* Typhimurium treated three times. It is important to highlight that $h_i(t \mid \cdot) = h_0(t)$ in the case of untreated *S.* Typhimurium, which acts as the control group, $h_i(t \mid \cdot) = h_0(t) \exp\{\beta_1 I_1(i)\}$ when *S.* Typhimurium is exposed one time to the antimicrobial treatment, and $h_i(t \mid \cdot) = h_0(t) \exp\{\beta_3 I_3(i)\}$ when it is exposed three times. This fact indicates that the specification of a baseline hazard function, $h_0(t)$, in our study is a relevant issue of the statistical modeling. We will return to these data in Chapter 4 to discuss different parametric and non-parametric proposals for $h_0(t)$ that have been widely used within the bayesian literature.

It is known that the Bayes theorem combines the prior distribution of the unknown elements in the model and the likelihood function of them for the observed data to compute the posterior distribution

$$\pi(h_0, \boldsymbol{\beta} \mid \mathcal{D}) \propto \mathcal{L}(h_0, \boldsymbol{\beta})\, \pi(h_0, \boldsymbol{\beta}),$$

where $\mathcal{L}(h_0, \boldsymbol{\beta})$ is the likelihood function of all unknown elements in $h_0(t)$ and $\boldsymbol{\beta}$ $(h_0, \boldsymbol{\beta})$, for data $\mathcal{D}$, and $\pi(h_0, \boldsymbol{\beta})$ the prior distribution.

The prior distribution was elicited considering a prior independent default scenario among the parameters associated to the baseline hazard function and the regression coefficients. Prior independence was also reckoned between the regression coefficients within a non informative scenario, with normal distributions centered at zero and a wide known variance:

$$\begin{aligned}\pi(h_0, \boldsymbol{\beta}) &= \pi(\alpha)\, \pi(\log(\lambda))\, \pi(\beta_1)\, \pi(\beta_3) \\ &= \mathrm{Ga}(\alpha \mid 0.01, 0.01)\, \mathrm{N}(\log(\lambda) \mid 0, 1000)\, \mathrm{N}(\beta_1 \mid 0, 1000)\, \mathrm{N}(\beta_3 \mid 0, 1000).\end{aligned} \tag{3.4}$$

The likelihood function $\mathcal{L}(h_0, \beta)$ for the observed data, $\mathcal{D}$, can be expressed as the product of the likelihood for each individual. For right censored observations, its contribution to the likelihood is its corresponding survival function, $P(T_i > t_i \mid \boldsymbol{x}_i, h_0\,\boldsymbol{\beta})$. The contribution to the likelihood function of an interval censored observation is the difference between its corresponding survival functions evaluated in the lower (t_{il}) and upper (t_{iu}) monitoring times, $S_i(t_{il} \mid \boldsymbol{x}_i, h_0\,\boldsymbol{\beta}) - S_i(t_{iu} \mid \boldsymbol{x}_i, h_0\,\boldsymbol{\beta})$. Consequently,

$$\begin{aligned} \mathcal{L}(h_0, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \mathcal{L}_i(h_0, \boldsymbol{\beta}) \qquad (3.5) \\ &= \prod_{i}^{\mathcal{R}} S_i(t_i \mid \boldsymbol{x}_i, h_0, \boldsymbol{\beta}) \prod_{i}^{\mathcal{I}} [S_i(t_{il} \mid \boldsymbol{x}_i, h_0, \boldsymbol{\beta}) - S_i(t_{iu} \mid \boldsymbol{x}_i, h_0, \boldsymbol{\beta})], \end{aligned}$$

where $\mathcal{R}$ ($\mathcal{I}$) is the set of right (interval) censored data, and the survival function for individual i

$$S_i(t \mid \boldsymbol{x}_i, h_0, \boldsymbol{\beta}) = \exp\{-H_0(t_i)\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}\},\ t > 0, \qquad (3.6)$$

with $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$ as the cumulative baseline hazard function, that in the case of the Weibull baseline hazard function is $H_0(t) = \lambda t^{\alpha}, t > 0$.

3.3.3 Posterior inferences

Bayesian inference has also been performed using INLA approximation by means of the `R-INLA` package, in which Weibull CPH models are properly implemented (see Martino *et al.*, 2011, for further details of INLA implementation). Results are discussed in terms of posterior inferences for the regression parameters as well as

posterior hazard and survival distributions with regard to covariate patterns of interest. Note that the plausibility of the proportional hazards assumption has been checked.

Furthermore, the model has been also estimated by means of MCMC simulation using WinBUGS software (Lunn *et al.*, 2000). Specifically, simulations were run considering three Markov chains of 100,000 iterations with a burn-in period of 10,000, thinning each 10th iteration. All posterior samples showed good convergence properties with values of the potential scale reduction factor $\hat{R}$ values equal or close to 1 and effective number of independent simulation draws greater than 100 (neff > 100). Hence, a subsection for comparing INLA and MCMC is also included where we highlight similarities and discrepancies between INLA and MCMC approaches as well as their subsequent strengths and weaknesses in bayesian inference.

Regression coefficients

Table 3.7 summarizes the estimated posterior marginal distribution of the regression coefficients, $\pi(\beta_1 \mid \mathcal{D})$ and $\pi(\beta_1 \mid \mathcal{D})$.

Parameter	Mean	Sd	$CI_{95\%}$	$P(\cdot > 0)$
β_1	-0.452	0.091	[-0.631,-0.273]	0
β_3	-0.422	0.091	[-0.601,-0.243]	0

TABLE 3.7: Summary of the marginal posterior distribution for the regression parameters: mean, standard deviation, 95% credible interval, and posterior probability that the parameter is positive.

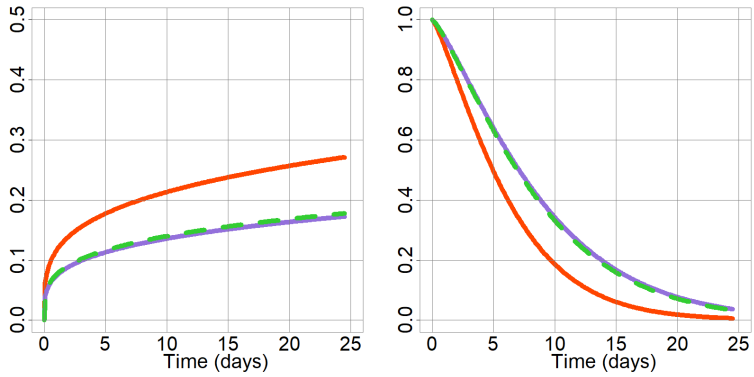
The last column of Table 3.7 shows that the regression parameters associated to changes in virulence, have posterior probabilities associated to negative values close to one. The estimated model

clearly indicates a relevant and negative effect of the alternative antimicrobial treatment applied once ([-0.631, -0.273] is a 95% CI). The marginal posterior distribution associated to β_1 is concentrated on real negative values and therefore exhibits a shrinkage of the hazard function with regard to the one corresponding to the untreated *S.* Typhimurium. Hence, when the foodborne pathogen is exposed to the antimicrobial treatment the infection hazard decreases, so *S.* Typhimurium seems to become less virulent than in untreated conditions. Posterior values associated to β_1 and β_3 are similar thus practically reporting no changes in the virulence behaviour.

Hazard and survival function

Figure 3.4a shows the mean of the posterior distribution of the *C. elegans* hazard function, $\pi(h(t) \mid \mathcal{D})$, for each of the two cauliflower by-product infusion treatments as well as the control group. The posterior mean of the hazard function associated to the control group is a monotonic increasing curve. Obviously, this monotonic trend is a direct consequency of the previous specification of the Weibull hazard baseline model.

Figure 3.4b presents the posterior mean of the survival function distribution, $\pi(S(t \mid h_0, \beta) \mid \mathcal{D})$, for *C. elegans* fed with each microbial populations. It is hard to distinguish between survival prospects related to both cauliflower treatments. In fact, the posterior mean of the survival probability at day 2, 12 and 22 is 0.869, 0.260, 0.055 for *S.* Typhimurium treated one time, and 0.865, 0.249, 0.051 for *S.* Typhimurium treated three times, respectively. With regard to the control group, the survival probabilities are lower than the ones of both cauliflower treatments, with posterior median survival probabilities 0.802, 0.121, and 0.011 at day 2, 12 and 22,



(a) (b)

FIGURE 3.4: (a): Mean of the posterior distribution for the hazard function of *C. elegans* fed with untreated *S.* Typhimurium (in red), *S.* Typhimurium treated one time (in purple), and *S.* Typhimurium treated three times (in green). (b): Mean of the posterior distribution for the survival function of *C. elegans* fed with untreated *S.* Typhimurium (in red), *S.* Typhimurium treated one time (in purple) and *S.* Typhimurium treated three times (in green).

respectively. Again, the results corroborate that the repetitively application of the antimicrobial treatment does not seem to have consequences on the virulence of *S.* Typhimurium.

INLA and MCMC comparison

MCMC sampling procedures were also used for bayesian inference. Results in this subsection are focused on comparing INLA and MCMC outcomes. Figure 3.5 shows the posterior marginal distribution of the regression coefficients approximated by INLA (black solid line) and MCMC-based density estimates (red dashed line). INLA and MCMC marginal posterior distributions for β_1 and β_2 are in almost perfect agreement, but in terms of speed, the respective model could be fitted in roughly 0.50 seconds on

an IntelCore i7-7700 3.60 GHz processor, while MCMC sampling required approximately 14 minutes.

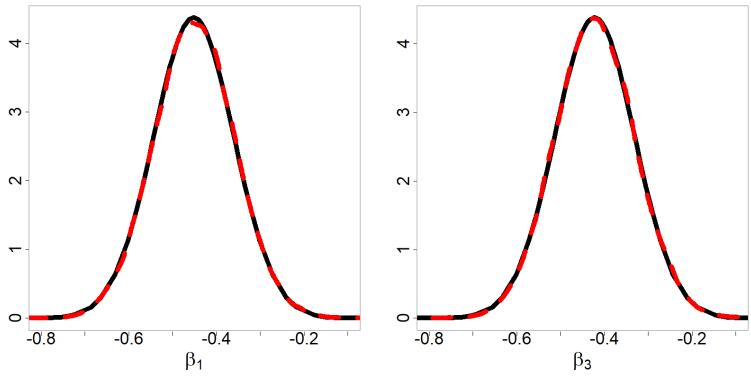


FIGURE 3.5: Posterior marginal distributions approximated by INLA (black solid line) and MCMC (red dashed line) for regression parameters associated to β_1 and β_3.

Regarding posterior estimates of the hazard function and survival probabilities, INLA and MCMC obviously, displays similar results as it can be observed in Tables 3.8 and 3.9. However, it is worth mentioning that in case of the INLA approach, computation of the posterior marginals of certain derived quantities such as hazard functions and survival probabilities is not directly available in INLA default outcomes. It can be done by means of the function `inla.posterior.samples()`, with which it is possible to generate n samples, from the approximated joint posterior distribution of the fitted model. These samples can then be further processed to derive quantities of interest. Remarkably, we have noted that accuracy and uncertainty of the samples are influenced by the number of samples generated (n), being necessary to account this.

Model	Group	$t = 2$	$t = 12$	$t = 22$
INLA	$ST0$	0.139 [0.120, 0.160]	0.223 [0.191,0.260]	0.263 [0.217, 0.316]
	$ST1$	0.089 [0.076, 0.103]	0.142 [0.124,0.163]	0.168 [0.141, 0.198]
	$ST3$	0.091 [0.078, 0.106]	0.147 [0.127,0.169]	0.173 [0.145, 0.204]
MCMC	$ST0$	0.139 [0.120, 0.159]	0.223 [0.190,0.259]	0.263 [0.217, 0.315]
	$ST1$	0.089 [0.076, 0.103]	0.142 [0.123,0.163]	0.167 [0.140, 0.197]
	$ST3$	0.091 [0.078, 0.106]	0.146 [0.127,0.167]	0.172 [0.145, 0.202]

TABLE 3.8: INLA and MCMC mean and 95% credible interval of the posterior distribution for the hazard function at days 2, 12 and 22 days of treatments untreated *S.* Typhimurium ($ST0$), *S.* Typhimurium treated one ($ST1$) and *S.* Typhimurium treated three times ($ST3$).

Model	Group	$t = 2$	$t = 12$	$t = 22$
INLA	$ST0$	0.802 [0.769, 0.833]	0.120 [0.090,0.154]	0.011 [0.005, 0.019]
	$ST1$	0.869 [0.843, 0.892]	0.260 [0.217,0.305]	0.055 [0.036, 0.079]
	$ST3$	0.865 [0.839, 0.888]	0.249 [0.207,0.293]	0.051 [0.032, 0.073]
MCMC	$ST0$	0.802 [0.769, 0.833]	0.121 [0.090,0.157]	0.011 [0.005, 0.020]
	$ST1$	0.869 [0.843, 0.891]	0.260 [0.218,0.304]	0.056 [0.036, 0.079]
	$ST3$	0.866 [0.839, 0.889]	0.250 [0.209,0.294]	0.051 [0.033, 0.073]

TABLE 3.9: INLA and MCMC mean and 95% credible interval of the posterior distribution for the survival function at days 2, 12 and 22 days of treatments untreated *S.* Typhimurium ($ST0$), *S.* Typhimurium treated one ($ST1$) and *S.* Typhimurium treated three times ($ST3$).

3.3.4 Discussion

The bayesian CPH model seems to be an appropiate methodology to assess virulence changes in the field of Pathogenicity and Microbial Virulence. Results indicate that the virulence of *S.* Typhimurium decreases when it is treated with cauliflower by-product infusion but also that the repetitively application of this antimicrobial treatment does not seem to have additional consequences in its virulence.

The INLA approach seems to be a fast alternative to MCMC, although we can not forget that MCMC is an asimptoticaly exact method whereas INLA is an approximation. Regarding the computations of derived quantities, in MCMC-based analysis it is easy to obtain them as of the parameter samples being possible its direct specification through the model syntax. While in INLA, it is necessary to make use of the simulation to obtain samples of the joint posterior distribution to compute quantities of interest. Remember that the posterior marginal distribution of non-linear combinations between different latent components are not direct available in INLA outcomes.

The bayesian approach allows the easy implementation of the baseline hazard function in the model definition, which allows the estimation and prediction of hazard and survival curves for given covariate patterns. However, other parametric as well as non-parametric options can be easily specificated.

In the modeling presented, we have chosen the Weibull distribution as the default option. However, since in that context, the baseline hazard function is considered as a meaningful measure of the natural course of the infection, in Chapter 4 we address the influence of baseline hazard specification in the whole inferential process comparing different baseline hazard definitions (parametric and non-parametric specifications) using this dataset as an illustrative example.

Chapter 4

Baseline hazard functions in the bayesian Cox proportional hazards model

4.1 Introduction

The Cox proportional hazards (CPH) model (Cox, 1972; Cox and Oakes, 1984) is the most popular regression model in survival analysis. It expresses the hazard function $h(t)$ of the survival time of each individual of the target population as the product of a common baseline hazard function $h_0(t)$, which determines the shape of $h(t)$, and an exponential term which includes the relevant covariates, and possibly, other effects.

The estimation of the regression coefficients in the CPH model under the frequentist approach can be obtained without specifying a model for the baseline hazard function by using partial likelihood methodology (Cox, 1972). However, depending on the context of the study, the baseline hazard misspecification can imply a lost of

valuable model information that makes impossible to fully report the estimation of the outcomes of interest, such as probabilities and survival curves for relevant covariate patterns (Royston, 2011). This is specially important in survival studies where $h_0(t)$ represents the natural course of a disease or an infection, or even the control group when comparing several treatments. Bayesian Inference provides a natural framework to jointly analyse all elements and uncertainties involved in the statistical modeling. In particular it allows in a natural manner, the analysis of both parametric and non-parametric baseline hazard functions.

Parametric approaches imply restricted shapes which do not enable the presence of irregular patterns (Dellaportas and Smith, 1993; Kim and Ibrahim, 2000). Non-parametric choices result in more flexible baseline hazard shapes (Sahu *et al.*, 1997; Ibrahim *et al.*, 2001) but which may suffer from overfitting and unstability (Breiman, 1996). Regularization methods try to modify the estimation procedures to give reasonable answers to these type of situations. Bayesian reasoning usually accounts for regularization through the prior distribution.

In this Chapter, we have a twofold objective: assessing the role of the baseline hazard function specification as well as the effect of regularization for non-parametric proposals in the CPH inferential process. We illustrate our objectives by means of two different studies: the first one is based on a real data set which collects information about a virulence assay in the context of food microbiology and the second one is a simulation study. We consider two different flexible specifications for $h_0(t)$ that allow for multimodal patterns: a mixture of piecewise constant functions (Sahu *et al.*, 1997) and a cubic B-spline function (Hastie *et al.*, 2009). We set regularization considering different prior scenarios which vary the great flexibility provided by prior

independence to some particular correlated structures. A baseline hazard Weibull distribution, the usual parametric proposal for $h_0(t)$ given its ability to represent different types of monotonic risks, is also included for comparison purposes.

4.2 Baseline hazard functions

Chapter 2 introduces extensively the CPH model. Recall that it expresses the hazard function of the survival time

$$h(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) = h_0(t) \exp\{\boldsymbol{x}' \boldsymbol{\beta}\}, \tag{4.1}$$

as the product of a baseline hazard function, $h_0(\cdot)$, and an exponential term where $\boldsymbol{x}$ is a vector of covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients. Here, we deepen into the baseline hazard function definition. We describe three paradigmatic proposals, one parametric based on the Weibull distribution and two non-parametric ones, a mixture of piecewise constant functions and a cubic B-spline function.

Weibull baseline hazard functions

The baseline hazard function corresponding to a Weibull distribution, $\mathrm{We}(\alpha, \lambda)$, with shape and scale parameter $\alpha > 0$ and $\lambda > 0$, respectively, is:

$$h_0(t \mid \alpha, \lambda) = \lambda \, \alpha \, t^{\alpha - 1}, \ t > 0. \tag{4.2}$$

This is a traditional model for survival data in biometric applications. It is very appealing due to its computational

simplicity, especially in small-sample settings, and flexibility in representing different types of risks, but always within the monotonicity (Lee *et al.*, 2016).

Mixture of piecewise constant functions

This proposal is based on piecewise functions defined by polynomial functions. They generate a flexible framework for modeling univariate survival data and have a long tradition (Henschel *et al.*, 2009; Ibrahim *et al.*, 2001) in the bayesian survival literature as alternative models to the Weibull $h_0(t)$. The overall shape of the baseline hazard function does not have to be imposed in advance as with the parametric models.

We assume a finite partition of the time axis with knots $c_0 \leq c_1 \leq \ldots \leq c_K$, where $c_0 = 0$, and c_K is usually taken as the last observed survival or censoring time. The hazard function is a flexible mixture of piecewise constant functions defined as

$$h_0(t \mid \boldsymbol{\varphi}) = \sum_{k=1}^{K} \varphi_k \, I_{(c_{k-1}, c_k]}(t), \;\; t > 0, \tag{4.3}$$

where $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)$, $I_{(c_{k-1}, c_k]}(t)$ is the indicator function defined as 1 when $t \in (c_{k-1}, c_k]$ and 0 otherwise. This baseline hazard function is usually known as *piecewise constant* function (PC from now on) because it is assumed to be constant within the K predetermined intervals $(c_{k-1}, c_k]$ for $k = 1, 2, \ldots, K$.

Cubic B-spline functions

We assume the same finite partition of the time axis specified for the PC baseline hazard function. The spline function for the baseline hazard function is usually defined in logarithmic scale (Murray *et al.*,

2016) to accommodate for normality the subsequent selection of prior distributions. It is defined as

$$\log h_0(t \mid \boldsymbol{\gamma}) = \sum_{k=1}^{K+3} \gamma_k \, B_{(k,4)}(t), \;\; t > 0, \tag{4.4}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+3})$, and $\{B_{(k,4)}(t), k = 1, ..., K+3\}$ is a cubic basis of B-splines with boundary knots c_0 and c_K and internal knots c_k, $k = 1, .., K-1$ (Hastie *et al.*, 2009). It is worth noting that the definition of this B-spline function needs the augmentation of the original knot sequence $\boldsymbol{c} = (c_0, c_1, \ldots, c_K)$ to the new one $\boldsymbol{\tau}$ defined as

$$\begin{aligned} &\tau_1 \leq \ldots \leq \tau_4 \leq c_0, \\ &\tau_{j+4} = c_j, \; j = 1, 2, \ldots, K-1, \\ &c_K \leq \tau_{K+4} \leq \ldots \leq \tau_{K+7}. \end{aligned} \tag{4.5}$$

This modeling strategy in known as a *piecewise cubic B-spline* function (*PS* from now on). Note that functions in the hazard function shown in equation (4.3) are also B-spline functions, in particular B-splines of order 1.

4.2.1 Regularization

We considered a prior independent default scenario between the parameters associated to the baseline hazard function and the regression coefficients associated to covariates. We also reckoned prior independence between the regression coefficients within a non

informative scenario, with normal distributions centered at zero and a wide known variance:

$$\pi(h_0, \boldsymbol{\beta}) = \pi(h_0)\,\pi(\beta_j) = \pi(h_0)\prod_{j=1}^{J} \mathrm{N}(\beta_j \mid 0, \sigma_j^2). \tag{4.6}$$

It is worth noting that $\pi(h_0)$ represents the prior distribution of all relevant parameters and hyperparameters in $h_0(t)$.

PC and *PS* baseline hazard functions can accommodate different shapes depending on the characteristics of the partition of the time axis. This is a relevant issue with a great research activity: Breslow (1974) considers various failure times as end points of intervals; Kalbfleisch and Prentice (1973) support that the selection of the grid should be made independently of the data; Henschel *et al.* (2009) fix the intervals assuming the condition that all the intervals contain comparable information, i.e. similar number of events; and Lee *et al.* (2016) avoid reliance on fixed partitions of the time scale by introducing the number of splits as a parameter to be estimated. When K is large, all kind of shapes of $h_0(t)$ tend to be similar. However, too small choices of K will lead to poor model fitting. In addition, it is important to point out that in the cases where the number of parameters is greater than the number of data we would need some shrinkage or regularization procedure, which in the bayesian setting is usually carried out by means of *informative* prior distributions that restrict the freedom of the parameters.

The elicitation of prior distributions for *PC* and *PS* baseline hazard functions includes different prior distributions proposals for coefficients $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$, respectively. They range from a default situation of prior independence among all the coefficients to correlated prior distributions that account for shape restrictions in order to avoid overfitting and strong irregularities in the estimation process.

We consider four prior scenarios for baseline hazard functions defined in terms of a mixture of piecewise constant functions. They are based on different correlation patterns among the coefficients associated to the piecewise functions.

Scenario PC1. Independent gamma prior distributions

$$\pi(\varphi_k) = \text{Ga}(\eta_k, \psi_k),\ k = 1, 2, ...K. \tag{4.7}$$

This is the most flexible and general prior scenario. A common selection is $\eta_k = \psi_k = 0.01$ (Sahu *et al.*, 1997).

Scenario PC2. Independent gamma prior distributions defined by means of a discrete-time Gamma process prior (Ibrahim *et al.*, 2001) for the cumulative hazard baseline function.

$$\pi(\varphi_k) = \text{Ga}(w_0\,\eta_0\,(c_k - c_{k-1}), w_0\,(c_k - c_{k-1})),\ k = 1, \ldots, K. \tag{4.8}$$

All these marginal prior distributions share the same prior expectation, η_0, but the prior variance of each φ_k is inversely proportional to the corresponding interval length, $c_k - c_{k-1}$. The selection $w_0 = 0.01$ is a usual value which provides a high level of uncertainty to the prior. We will assume the *ad hoc* proposal by Christensen *et al.* (2011) for the elicitation of η_0 that considers $\eta_0 = 0.69315/\tilde{t}$, where $\tilde{t}$ is the median survival time of the reference group.

Scenario PC3. Correlated conditional gamma prior distributions

$$\pi(\varphi_k \mid \varphi_1, \ldots, \varphi_{k-1}) = \text{Ga}(\eta_k,\, \eta_k/\varphi_{k-1}),\ k = 2, \ldots, K. \tag{4.9}$$

This prior is based on a discrete-time martingale process (Sahu *et al.*, 1997) which correlates coefficients of adjacent intervals so that $\mathrm{E}(\varphi_k \mid \varphi_1, \ldots, \varphi_{k-1}) = \varphi_{k-1}$ and $\mathrm{Var}(\varphi_k \mid \varphi_1, \ldots, \varphi_{k-1}) = \varphi_{k-1}^2/\eta_k$. The parameter η_k is very important because it controls the level of smoothness, which decreases when η_k goes to zero. A common elicitation is $\eta_k = 0.01,\ k = 2, \ldots, K$ and $\pi(\varphi_1) = \mathrm{Ga}(0.01, 0.01)$.

Scenario PC4. Correlated conditional normal prior distributions for the coefficients in logarithmic scale

$$\pi(\log(\varphi_k) \mid \varphi_1, \ldots, \varphi_{k-1}) = \mathrm{N}(\log(\varphi_{k-1}), \sigma_\varphi^2),\ k = 2, \ldots, K, \tag{4.10}$$

with $\pi(\log(\varphi_1)) = \mathrm{N}(0, \sigma_\varphi^2)$. This is also a proposal based on a discrete-time martingale process. It comes from the areas of spatial statistics (Banerjee *et al.*, 2014) and bayesian B-splines (Lang and Brezger, 2004) where it is more known as a first-order random walk. Correlation between the $\log(\varphi_k)$'s corresponding to neighboring intervals is expressed assuming conditional normal prior distributions.

Non-informative prior distributions for the variance σ_φ^2 have been generally taken as inverse gamma distributions, $\mathrm{IG}(\nu_0, \nu_0)$, with small values for ν_0. However, there are some research that questions the role of these distributions for describing lack of prior information. Gelman (2006) proposes the use of proper uniforms and half-t distributions for the standard deviations as sensible choices for *non-informative* priors, which understand as reference models to be used as a standard of comparison or a starting point of the inferential process (Bernardo, 1979).

We also considered different prior specifications for the coefficients associated to the baseline hazard function of the *PS* modeling

following the idea of smoothing its level of flexibility and prevent overfitting. These scenarios are not a mere repetition of those considered for *PC* baseline hazard functions. They have been chosen because they are usual proposals in the statistical literature regarding cubic B-splines especifications.

Scenario PS1. Independent normal prior distributions

$$\pi(\gamma_k) = \mathrm{N}(0, \sigma_k^2),\ k = 1, \ldots, K+3. \tag{4.11}$$

This is the most simple scenario, similar to *PC1*, in which γ_k are considered independent and normally distributed with a known and wide variance.

Scenario PS2. Hierarchical normal prior distributions

$$\pi(\gamma_k \mid \sigma_\gamma^2) = \mathrm{N}(0, \sigma_\gamma^2),\ k = 1, \ldots, K+3, \tag{4.12}$$

where σ_γ^2 is the common and unknown variance population. As mentioned before, a usual election for the hyperprior distribution for σ_γ^2 is an inverse gamma distribution or also a proper uniform distribution (Gelman, 2006).

Scenario PS3. Correlated conditional normal prior distributions defined as

$$\pi(\gamma_k \mid \gamma_1, \ldots, \gamma_{k-1}) = \mathrm{N}(\gamma_{k-1}, \sigma_\gamma^2),\ k = 2, \ldots, K+3, \tag{4.13}$$

and based on a first order Gaussian random walk which involves an intrinsic Gaussian Markov random field as the conditional joint prior distribution for the spline coefficients given σ_γ^2. This proposal comes from the so called bayesian P-splines (Lang and Brezger, 2004;

Fahrmeir and Kneib, 2011) and has been widely used in bayesian spatial statistics (Banerjee *et al.*, 2014), where it is usually expressed in terms of conditional distributions in the form

$$\pi(\gamma_k \mid \boldsymbol{\gamma}_{-k}) = \mathrm{N}(\frac{1}{2}(\gamma_{k-1} + \gamma_{k+1}), 2\sigma_\gamma^2), \; k = 2, \ldots, K+3, \quad (4.14)$$

where $\boldsymbol{\gamma}_{-k}$ denotes all splines coefficients except γ_k. Popular marginal prior distributions choices for σ_γ that try to be as neutral as possible are $\mathrm{Ga}(1, 0.0005)$ (Lang and Brezger, 2004) and Ga(0.001, 0.001) as a default option in the software `BayesX` (Belitz *et al.*, 2015). This scenario is analogue to *Scenario PC4*. Consequently, all the discussion regarding the elicitation of the prior distribution for the variance σ_γ^2 (precision or standard deviation τ_γ and σ_γ, respectively) also applies here.

4.2.2 Likelihood function

The model needs to be formulated on data $\mathcal{D} = \{(t_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$, where t_i is the observed survival time for the ith individual, δ_i the indicator taking 1 if the event has occurred and 0 otherwise, and $\boldsymbol{x}_i$ the subsequent covariates or risk factors.

The likelihood function of $(h_0, \boldsymbol{\beta})$ for $\mathcal{D}$ which, in the absence of tied observations, is defined by Ibrahim *et al.* (2001) as

$$\mathcal{L}(h_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} h_0(t_i)^{\delta_i} \exp\{-H_0(t_i)\}[\exp\{\boldsymbol{x}_i' \boldsymbol{\beta}\}]^{\delta_i} \exp\{\exp\{\boldsymbol{x}_i' \boldsymbol{\beta}\}\}, \quad (4.15)$$

with $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$ as the cumulative baseline hazard function.

In the case of the Weibull hazard baseline function, the cumulative baseline hazard function is:

$$H_0(t) = \lambda t^{\alpha}, t > 0. \tag{4.16}$$

When the baseline function is defined via mixture of piecewise constant functions,

$$\begin{aligned} H_0(t) &= \textstyle\sum_{m=1}^{k-1} \varphi_m(c_m - c_{m-1}) + \varphi_k(t - c_{k-1}), \\ & c_{k-1} \leq t < c_k,\ k = 1, \ldots, K. \end{aligned} \tag{4.17}$$

The expression of the cumulative baseline hazard for the $h_0(t)$ defined in terms of cubic B-spline functions needs to take into account some properties of B-splines (Boor, 1978). In particular,

$$H_0(t) \;=\; \int_0^t \sum_{k=1}^{K+3} \gamma_k\, B_{(k,4)}(u)\mathrm{d}u = \sum_{k=1}^{K+4} \phi_k\, B_{(k,5)}(t), \tag{4.18}$$

with $\phi_1 = 0$, and

$$\phi_{m+1} = \frac{\tau_{m+1} - \tau_5}{4} \sum_{j=1}^{m} \gamma_j,\ m = 1, 2, \ldots, K+3$$

Note that $H_0(t)$ in (4.18) is defined in terms of B-splines of order 5 which need to add two additional nodes to the augmented knot sequence $\boldsymbol{\tau}$ in (4.5).

4.3 Virulence in foodborne pathogens study

4.3.1 Database

A dataset involving a virulence assay is taken into account to explore the different baseline hazard specifications presented above for the CPH model. The data came from an experiment designed to assess the effect of the use of a cauliflower by-product infusion treatment in *S.* Typhimurium virulence behaviour. *Salmonella enterica serovar* Typhimurium (*S.* Typhimurium) is currently one the most usual serotypes related to salmonellosis outbreaks and cauliflower by-product infusion treatment is an alternative preservation treatment against it. The experiment pays special attention to the effect of a reiterative use of the preservation treatment in the virulence behaviour.

One and three expositions to the treatment were evaluated as well as a pathogen *S.* Typhimurium population non-exposed to the treatment that was considered as the control group. A nematode, *Caenorhabditis elegans (C. elegans)* was the host model used for quantifying virulence of *S.* Typhimurium. Each group, *S.* Typhimurium non treated (*ST0*), *S.* Typhimurium treated once (*ST1*), and *S.* Typhimurium treated three times (*ST3*)) was the source of nutrition of 250 synchronized young adult nematodes kept in identical environmental conditions (20^{o}C) during all their lifespan (three weeks as maximum approximately). Virulence for each worm was defined in terms of their subsequent survival time (see Sanz-Puig *et al.* (2017) for more details about the validation and special conditions of the experiment).

4.3.2 Modeling

We modeled virulence for each worm i by means of the following CPH model

$$h_i(t \mid h_0, \boldsymbol{x}_i, \boldsymbol{\beta}) = h_0(t) \exp\{\beta_1 I_1(i) + \beta_3 I_3(i)\}, \tag{4.19}$$

where $I_1(i)$ and $I_3(i)$ are dummy variables for groups *ST1* and *ST3*, respectively. It is important to highlight that $h_i(t \mid \cdot) = h_0(t)$ in the case of the group *ST0*, which acts as the control group, $h_i(t \mid \cdot) = h_0(t) \exp\{\beta_1 I_1(i)\}$ when *ST1*, and $h_i(t \mid \cdot) = h_0(t) \exp\{\beta_3 I_3(i)\}$ when *S.* Typhimurium is exposed three times, $ST3$.

We assumed for PC and PS baseline hazard functions a common finite partition of the time axis with $K = 10$, 9 internal knots and $c_{10} = 24.50$ days, which was the longest survival time observed. As recommended by Murray *et al.* (2016), we selected the intervals of the partition with the same length, 2.45 days.

4.3.3 Posterior inferences

We carried out eight survival inferential processes which where the result of the combination of the three generic specifications of the baseline hazard function presented above with the different prior scenarios. The posterior distribution for each model was estimated through the `JAGS` software (Plummer, 2003). For each estimated model, we run three parallel Markov chains with 50,000 iterations plus 5,000 dedicated to the burn-in period. Moreover, the chains were additionally thinned by storing every 10th iteration in order to reduce autocorrelation in the sample. Convergence was guaranteed monitoring that the potential scale reduction factor $\hat{R}$ were close to

1 and the effective number of independent simulation draws higher than 100 ($n_{eff} > 100$).

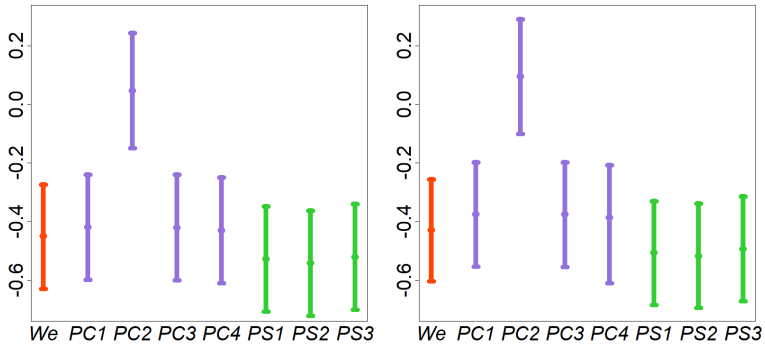
Regression coefficients

We first focus on the posterior stability of the posterior distribution for the regression coefficients as well as the behaviour of the subsequent marginal posterior distribution for the baseline hazard function and for the survival function.

Discrepancies between the posterior marginal distributions for the regression coefficients associated to groups $ST1$ and $ST3$. $\pi(\beta_1 \mid \mathcal{D})$ and $\pi(\beta_3 \mid \mathcal{D})$, are only the result of the different specifications for $h_0(t)$ and its prior distribution. Figure 4.1 shows the posterior mean and a 95% credible interval for the coefficients associated to experimental groups *ST1* and *ST3*. The first thing that attracts our attention is that both graphics are almost equal, thus indicating no substantial changes in the virulence when the antimicrobial treatment is applied one or three times. Secondly, it is clearly appreciated that *PC2* model shows marginal posterior results very different (near zero) than those for the rest of models, which provide quite similar inferences. *PS* models give very stable results with slightly lower values than *PC1*, *PC3* and *PC4* models. Weibull and *PC* inferences (except *PC2* model) are closer than *PS* models, with a very broad degree of overlap in both posterior estimations.

Results from scenario *PC2* need a bit of attention. The marginal prior distribution for each φ_k is $\text{Ga}(0.00304, 0.0245)$, with prior expectation and variance 0.1242 and 5.0646, respectively. It is derived from a discrete-time Gamma process prior, (see expression (4.8), constructed from the sample median $\tilde{t}_0 = 5.58$ days and the election of $w_0 = 0.01$ and $\eta_0 = 0.69315/5.58 = 0.1242$. This is an

informative prior greatly skewed in favour of values near zero and low variability which acts as a dominant element in the inferential process. Both posteriors, concentrated around zero, would indicate no differences in the lifetime of the worms in the experimental groups with regard to the control group. This would be a strong conclusion.



(a) β_1 (b) β_3

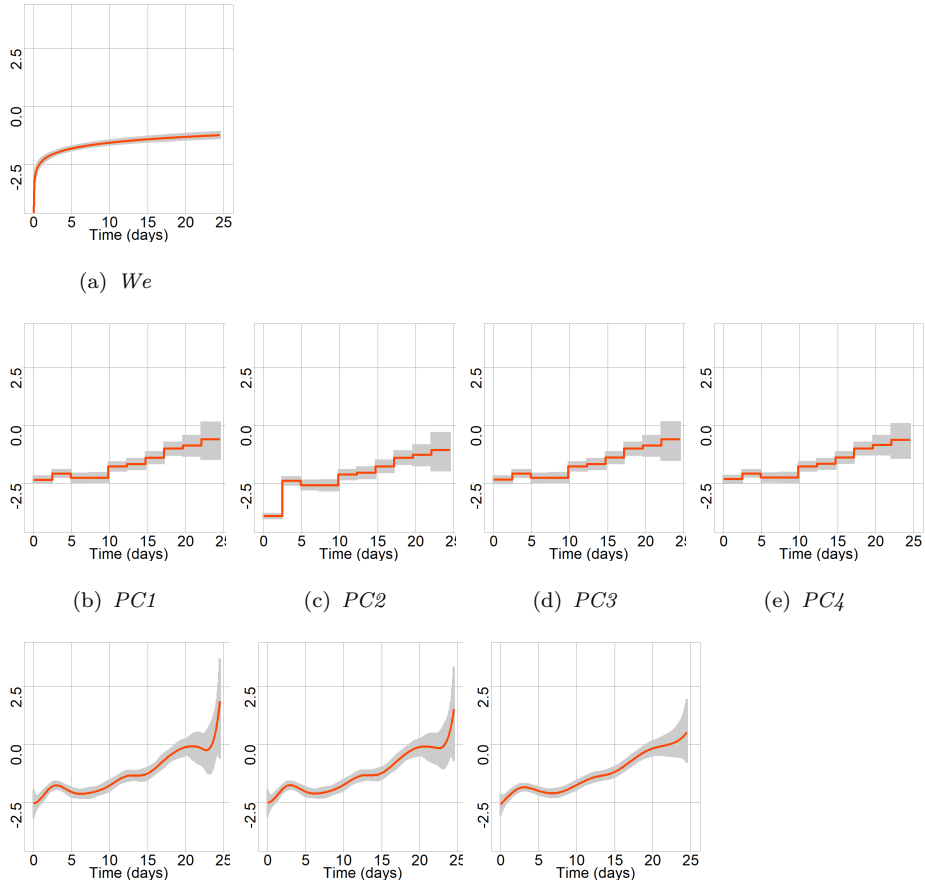
FIGURE 4.1: Posterior mean and 95% credible interval for the regression coefficients β_1 (a) and β_3 (b) associated to groups $ST1$ and $ST3$, respectively, for all survival models under evaluation.

Baseline hazard and baseline survival functions

Below we discuss the posterior distribution for the baseline hazard and the survival function of the different models of the study. Figure 4.2 is a matrix of figures. Row one is for Weibull baseline hazard, row two for piecewise constant, and row three for cubic B-spline specification. Each figure shows the mean of the logarithm of the baseline hazard function, which is also the hazard function associated to control group.

Parametric and non-parametric specifications of the baseline hazard report different shapes of the log baseline hazard function. The

Weibull model displays an increasing monotone log hazard behavior with a mild concavity profile which seems to be levelling off from the eighth day approximately. All *PC* and *PS* models report an increasing convex-exponential pattern with different intensities. *PC* models, show, in general similar results for all prior scenarions, even model *PC2*, which had shown remarkable differences for the coefficient regression estimates. *PS* models show more irregularities that *PC* models as a consequence of its definition in terms of cubic splines, possibly more flexible than piecewise constant functions. It is interesting to note that the hierarchical modelling in *PS2* introduce scarce differences with regard to the independent *PS1* scenario. In the case of *PS3* we can appreciate a patternmuch more smoothed than the ones in $PS1$ and $PS2$ scenarios.

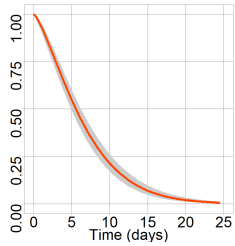


(a) *We*

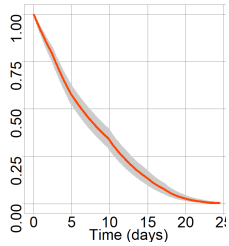
(b) *PC1* (c) *PC2* (d) *PC3* (e) *PC4*

(f) *PS1* (g) *PS2* (h) *PS3*

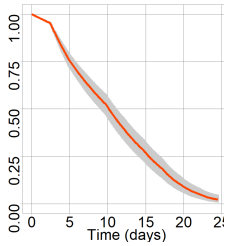
FIGURE 4.2: Posterior mean and 95% credible interval for the log baseline hazard function, $\log(h_0(t))$, under the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).



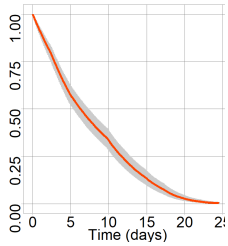
(a) *We*



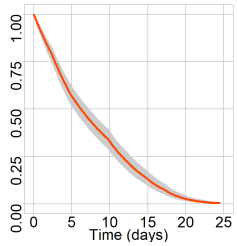
(b) *PC1*



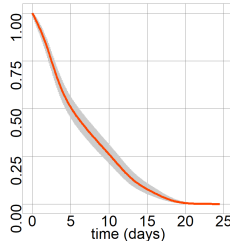
(c) *PC2*



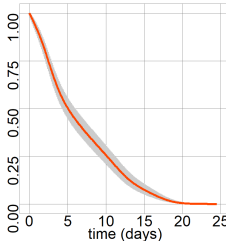
(d) *PC3*



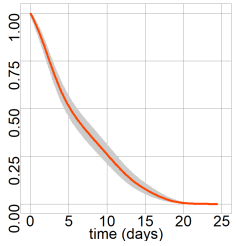
(e) *PC4*



(f) *PS1*



(g) *PS2*



(h) *PS3*

FIGURE 4.3: Posterior mean and 95% credible interval for the baseline survival, $S_0(t)$, function under the diferent modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

Hazard information (even logarithmic transformation) is usually difficult to interpret and that is why the survival function is in general a more widely output of interest. Figure 4.3 describes graphically the posterior mean and 95% credible intervals (CI) of the posterior distribution of the baseline survival function and Table 4.1 shows the posterior mean and 95% CI at days 2, 12, and 22. These

outputs confirm the fact that the particular choice of h_0 is relevant in the posterior distribution $\pi(S_0(t) \mid \mathcal{D})$. *PC* models present similar inferences, except for *PC2*, as we expected. *PS* models also show close outcomes.

Model	$t = 2$	$t = 12$	$t = 22$
We	0.841 [0.813, 0.866]	0.139 [0.106,0.174]	0.011 [0.006, 0.020]
PC1	0.824 [0.792, 0.852]	0.222 [0.203,0.241]	0.008 [0.004, 0.012]
PC2	0.961 [0.956, 0.965]	0.401 [0.343,0.458]	0.051 [0.026, 0.084]
PC3	0.824 [0.793, 0.853]	0.237 [0.193,0.284]	0.012 [0.005, 0.022]
PC4	0.819 [0.787, 0.848]	0.233 [0.190,0.281]	0.011 [0.004, 0.021]
PS1	0.811 [0.776, 0.844]	0.171 [0.134,0.214]	0.001 [0.000, 0.002]
PS2	0.801 [0.772, 0.841]	0.167 [0.130,0.208]	0.000 [0.000, 0.002]
PS3	0.806 [0.772, 0.839]	0.173 [0.134,0.215]	0.000 [0.000, 0.003]

TABLE 4.1: Mean and 95% credible interval of the posterior baseline survival probabilities at days 2, 12 and 22 for the eight estimated models.

Model selection criteria

We considered the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) and the log pseudo-marginal likelihood (LPML) (Geisser and Eddy, 1979) for model selection. DIC measures the information of a model by means of its deviance penalized with regard to its complexity. LPML is based on predictive criteria. It combines, in a logarithmic scale, the conditional predictive ordinate value (CPO) associated to observations of each individual (Gelfand, 1996). Smaller values for DIC are preferred, while LPML larger values indicate better predictive performance.

Table 4.2 shows the DIC and LPML values corresponding to all estimated models. Results indicate that *PS* baseline hazard

functions exhibit better behaviour (lower DIC and larger LPML values) than *Weibull* or *PC* specifications. In addition, the *PS3* model, in which the prior distribution is defined through a first order random walk for the coefficients of $h_0(t)$, has the lowest DIC and the largest LPML. Differences among *PS* models are not very important and not only because of DIC differences are less than 2 but also because all *PS* models report similar inferences. *Weibull* hazard modeling is the second best choice (supported by DIC and LPML criteria) while *PC* models are the worst, in particular the *PC2* model, with DIC and LPML values that differ substantially from the rest of the *PC* models. According to model selection, it is important to point out the necessity to also consider the nature of the problem to tackle. In that example, the baseline hazard reflects the natural course of the infection, and it seems reasonable that *PS* models reflect very well that process.

Model	DIC	LPML
We	4553.309	-2276.334
PC1	4751.743	-2373.499
PC2	4939.347	-2467.535
PC3	4751.913	-2373.565
PC4	4750.871	-2373.198
PS1	4461.602	-2231.980
PS2	4461.333	-2231.770
PS3	4459.937	-2229.443

TABLE 4.2: DIC and LPML values for the survival models defined by means of different specifications of the baseline hazard function.

4.4 Simulation study

In this Section, we explore the impact of the baseline hazard specification in the whole inferential process, specifically in the posterior estimates of the regression coefficients as well as in the posterior distributions for the hazard and survival function. For this purpose, we conduct three simulation studies (based in three different baseline hazard definitions) to assess the performance of the generic modelings (*We*, *PC* and *PS*) previously developed in this Chapter.

4.4.1 Simulation scenarios

Under three simulation scenarios survival times have been obtained from a generic CPH model:

$$h(t \mid h_0, \boldsymbol{x}, \boldsymbol{\beta}) = h_0(t) \exp\{\boldsymbol{x}_1' \boldsymbol{\beta}_1\}. \tag{4.20}$$

The baseline hazard funcion for each simulated scenario is

(a) *Scenario 1*, a standard Weibull distribution with increasing hazard funtion ($\alpha = 1.5$ and $\lambda = 0.5$).

(b) *Scenario 2*, a mixture of piecewise functions.

$$h_0(t \mid \boldsymbol{\varphi}) = \sum_{k=1}^{3} \varphi_k \, I_{(c_{k-1}, c_k]}(t), \; t > 0,$$

with three pieces, $\varphi_1 = 0.5$ in $0 < t \leq 0.4$, $\varphi_2 = 2.5$ in $0.4 < t \leq 1$ and $\varphi_3 = 1.5$ in $t > 1$.

(c) *Scenario 3*, a mixture of two Weibull distributions with shape and scale parameters, α_1, α_2, λ_1 and λ_2, respectively, and a mixing probability parameter p,

$$h_0(t \mid \alpha_1, \alpha_2, \lambda_1, \lambda_2) = \frac{\lambda_1\,\alpha_1\,t^{\alpha_1-1}\,p\exp\{-\lambda_1\,t^{\alpha_1}\} + \lambda_2\,\alpha_2\,t^{\alpha_2-1}\,(1-p)\exp\{-\lambda_2\,t^{\alpha_2}\}}{p\exp\{-\lambda_1\,t^{\alpha_1}\} + (1-p)\exp\{-\lambda_2\,t^{\alpha_2}\}}$$

with $\alpha_1 = 3$, $\lambda_1 = 0.1$, $\alpha_2 = 1.6$, $\lambda_2 = 0.1$ and $p = 0.8$.

All the scenarios included a binary treatment covariate drawn from a Bernoulli distribution with probability 0.5, Be(0.5), with associated log-hazard ratio $\beta_1 = 1$. We apply administrative right censoring at time (C_R) which was previously fixed for each scenario from the following restriction $S_0(C_R) = 0.1$, where $S_0(\cdot)$ denoting the baseline survival function. For each scenario we generated 100 replicates, each one with sample size of $N = 100$.

We analyse each simulated dataset via the eight generic modeling proposed in this Chapter (*We*, *PC1*, *PC2*, *PC3*, *PC4*, *PS1*, *PS2* and *PS3*). It is worth mentioning that for the estimation of models *PC* and *PS* for the data in *Scenarios 1* and *3*, we assume a finite partition of the time axis $c_0 \leq c_1 \leq \ldots c_K$ with $K = 10$ knots defined according to the theoretical baseline hazard function from which the data have been simulated. Concretely, each c_{k-1}, defined for $k = 1, \ldots, 10$, was assigned considering $S_0(c_{k-1}) = 1 - 0.09\,(k-1)$.

4.4.2 Generating survival times

We follow the *inversion method* (Bender *et al.*, 2005; Austin, 2012; Crowther and Lambert, 2013) to simulate survival data for *Scenario*

1 and *Scenario 2*. The method is based on using the relationship $F(T) = U$ for $t \geq 0$, where $F(t)$ represents the cumulative distribution function (CDF) of a survival random variable T, and U is a standard uniform random variable. Hence, solving $T = F^{-1}(U)$ we can obtain a random draw from the distribution of T. This procedure can be directly applied when the cumulative hazard function has a closed form expression and can be directly inverted. It is easily implemented in any standard software with a random number generator, and in the case of the `R` software (R Core Team, 2013) we can use the `simsurv` (Brilleman, 2018) and `SimSCRPiecewise` (Chapple, 2016) packages.

Specifically, for the *Scenario 3* in which a more complex baseline hazard function is considered, the *inversion method* is not directly suitable. The cumulative hazard function has a closed form expression, but it can not be directly inverted. We must use iterative root-finding techniques (Crowther and Lambert, 2013) to solve it. This procedure is implemented for the `R` software (R Core Team, 2013) in the `simsurv` (Brilleman, 2018) package. Further details of the inversion method and its corresponding extension to simulate complex baseline hazard functions are described in Appendix A.

4.4.3 Posterior inferences

Each simulation dataset was used to estimate different survival Cox models based on the three generic specifications of the baseline hazard function and the different prior scenarios discussed in the first part of this Chapter. We have obtained the posterior distribution by using `JAGS` software (Plummer, 2003). For the estimation of the uncertainties in each model, we have run three parallel chains with 20,000 iterations plus 2,000 (10%) dedicated to the burn-in period. Moreover, the chains were additionally thinned

by storing every 10th iteration in order to reduce the autocorrelation of the sample. In all inferential process, the convergence of the chains to the posterior distribution was guaranteed by monitoring that the potential scale reduction factor $\hat{R}$ were close to 1 and the effective number of independent simulation draws, $n_{eff} > 100$.

4.4.4 Regression coefficients

Remember that we ran 100 replicates of each inferential process, and consequently, we had 100 approximate random samples of the subsequent posterior distributions. Next, we present in a very simple and general notation the 100 replicates of the approximate marginal posterior sample for a generic regression coefficient β. In particular, a replica is represented by $\{\beta_1^{(r)}, \ldots, \beta_r^{(N)}\}$ with $r = 1, \ldots, 100$ and N the size of each posterior sample.

We considered four different measures for assessing the stability of the posterior distribution for the regression coefficients:

- Bias. It is the difference between the average of the posterior means of the replicas and the true regression coefficient, $(\sum_{r=1}^{R} \bar{\beta}_r/R) - \beta$, where R is the number of replicas, $R = 100$, and $\bar{\beta}_r$ the sample mean of the posterior sample corresponding to the replica r.

- Standard error (SE). It is the square root of the average of the posterior variances of the replicas, $\sqrt{\sum_{r=1}^{R} s_r^2/100}$, where s_r^2 is the sample variance of the posterior sample for the replica r.

- Standard deviation (SD). It is defined as the standard deviation of the set that includes the posterior mean of the regression coefficient of all replicas.

- Coverage probability (CP). It is the proportion of the $R = 100$ replicate 95% credible intervals which contain the true value of the regression coefficient.

Tables 4.3, 4.4 and 4.5 display the values of the Bias, SE, SD and CP referred to the three simulation scenarios, respectively. It is important to note that the only regression coefficient in the model (4.20) is β_1, which corresponds to the binary covariate x_1.

Model	Bias	SE	SD	CP
We	-0.011	0.218	0.214	0.96
PC1	-0.137	0.223	0.202	0.93
PC2	0.488	0.250	0.352	0.54
PC3	-0.136	0.224	0.203	0.94
PC4	-0.282	0.224	0.194	0.77
PS1	-0.014	0.224	0.225	0.94
PS2	-0.176	0.221	0.216	0.84
PS3	-0.024	0.225	0.222	0.94

TABLE 4.3: Bias, SE, SD and CP corresponding to all inferential and replicate processes for the regression coefficient β_1 of the simulated model (4.20) under simulation *Scenario 1*.

Model	Bias	SE	SD	CP
We	-0.024	0.221	0.219	0.95
PC1	-0.988	0.208	0.123	0.00
PC2	0.045	0.282	0.365	0.89
PC3	-0.988	0.208	0.123	0.00
PC4	-0.998	0.200	0.114	0.00
PS1	0.034	0.227	0.242	0.95
PS2	0.021	0.224	0.229	0.95
PS3	0.032	0.228	0.240	0.95

TABLE 4.4: Bias, SE, SD and CP corresponding to all inferential and replicate processes for the regression coefficient β_1 of the simulated model (4.20) under simulation *Scenario 2*.

Model	Bias	SE	SD	CP
We	0.147	0.227	0.200	0.95
PC1	-0.089	0.224	0.180	0.98
PC2	0.366	0.234	0.285	0.65
PC3	-0.089	0.224	0.180	0.98
PC4	-0.165	0.222	0.171	0.94
PS1	0.019	0.224	0.200	0.99
PS2	-0.038	0.225	0.193	0.99
PS3	0.048	0.228	0.200	0.98

TABLE 4.5: Bias, SE, SD and CP corresponding to all inferential and replicate processes for the regression coefficient β_1 of the simulated model (4.20) under simulation *Scenario 3*.

The Weibull modeling of the baseline hazard function in *Scenario 1* produces the least unbiased estimates and the best coverage probabilities. *PS1* and *PS3* models also reports good estimates with low bias values and high coverage probabilities. SE and SD presents similar values in all models. On the contrary, *PS2* model shows

the worst values, highlighting again the strong influence of its prior distribution. The performance of the models in *Scenario 2* evidences similar patterns to the ones in Scenario 1, *We* and *PS* models show the lowest bias values and the highest coverage probabilities. SE and SD estimates are close under all models. However, *PC2* model shows better estimates than its counterparts *PC1*, *PC3* and *PC4*. In the case of *Scenario 3* results obtained are in accordance with the obtained in *Scenario 1*, but with a clear improvement of *PS* and *PC* models.

4.4.5 Hazard function

We evaluate the performance of all models under the three simulation scenarios, Weibull, a mixture of piecewise functions, and a mixture of Weibull models, in terms of the baseline hazard estimates. We work with the logarithmic transformation of the baseline hazard $h_0(t)$ to scale values and facilitate the visual comparison among estimates. For the posterior sample of each replica, we can construct an approximate posterior sample of the baseline hazard function at t, whose average can be used as a punctual estimates of the true baseline hazard at that time, $h_0(t)$. And we can merge the information of all replicas to obtain a better estimation by averaging among all replicas. We represent this estimation as $\log(\widehat{h}_0(t)$. The above procedure is also useful for extracting information about its posterior variability and constructing, for example, a 95% credible interval for the posterior of the baseline hazard at t.

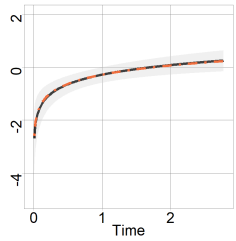
We mesure the accurary of our proposal by computing the square of the difference between the posterior estimation of $h_0(t)$ and the true hazard function at t. A general measure that accounts for

this difference over the time interval of the study is the root-mean squared deviation (RMSD) computed as

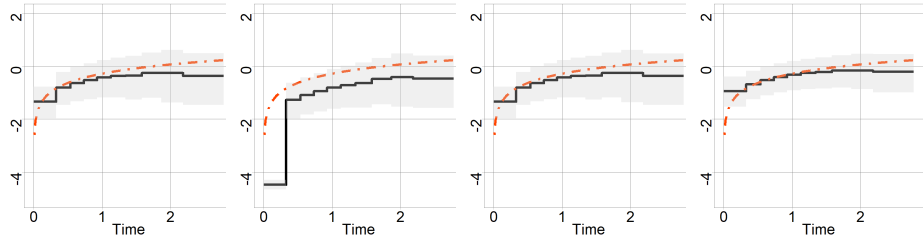
$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{m}[\log(\widehat{h}_0(t_m)) - \log(h_0(t_n))]^2}{m}} \tag{4.21}$$

a discrete approximation based on the idea of the Riemann sums to aproximate the continuous sum, an integral. At this point, we would like to note that we have considered a wide partition of the time axis, with m knots spaced 0.01 time points.

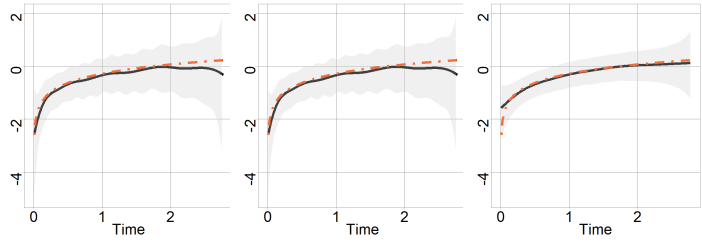
Figure 4.4 displays the posterior mean of the baseline hazard function and a 95% credible bound for the models from simulated *Scenario 1*. The Figure also shows the true hazard baseline function and the estimated RMSD. Figures 4.5 and 4.6 contain the same information than Figure 4.4 but for simulated *Scenarios 2* and *3*, respectively.



(a) *We*. RMSD=0.021



(b) *PC1*. RMSD=0.342 (c) *PC2*. RMSD=1.205 (d) *PC3*. RMSD=0.342 (e) *PC4*. RMSD=0.281

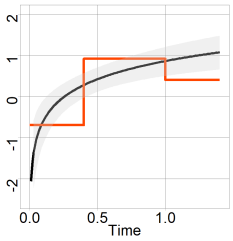


(f) *PS1*. RMSD=0.169 (g) *PS2*. RMSD=0.140 (h) *PS3*. RMSD=0.102

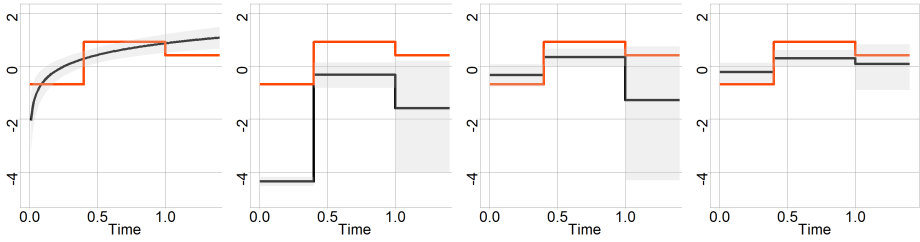
FIGURE 4.4: Average pointwise of the posterior approximate means of the log-baseline hazard estimate (black solid line) of the replicas, average of the posterior 95% credible intervals (grey area) of the replicas, true log-baseline hazard function (red dashdotted line) and reported RMSD for the estimated survival models in the simulated *Scenario 1* (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

We know that the *Scenario 1* was simulated from a Weibull model and it is clear from Figure 4.4 that the estimated Weibull model (*We*) provides the closest fit to the true function, as expected. *PS* models also seem to capture the underlying shape quite

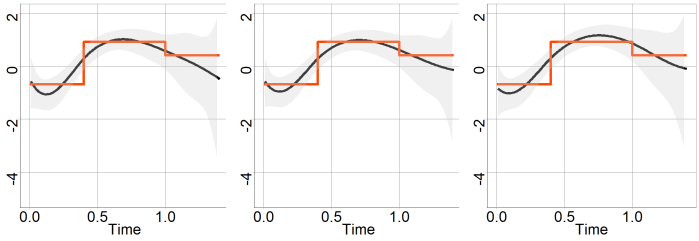
accurately concretely and, in particular, *PS3* performs very well evidencing the effect of the prior distribution in the estimation of the baseline shape. *PC* models show the worst performance, possibly because the baseline hazard estimates for them are discontinuos piecewise estimates, and this situation complicates the capture of the curvature of the true baseline hazard function. Remarkably, it is also noticeable, in *PC* models, the effect of the regularization by means of correlated prior distributions.



(a) *We.* RMSD=0.520



(b) *PC1.* RMSD=0.787 (c) *PC2.* RMSD=2.374 (d) *PC3.* RMSD=1.000 (e) *PC4.* RMSD=0.509

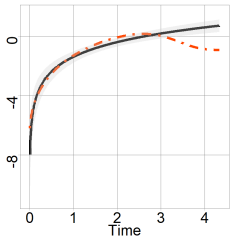


(f) *PS1.* RMSD=0.356 (g) *PS2.* RMSD=0.307 (h) *PS3.* RMSD=0.320

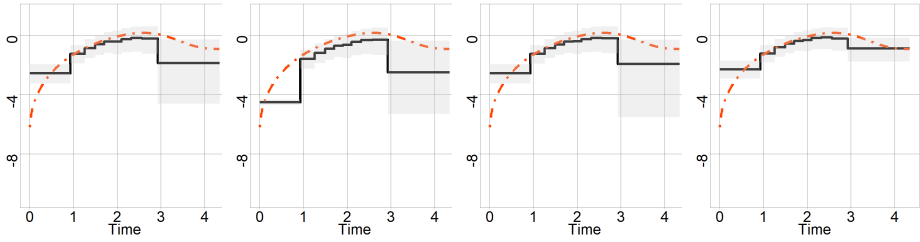
FIGURE 4.5: Average pointwise of the posterior approximate means of the log-baseline hazard estimate (black solid line) of the replicas, average of the posterior 95% credible intervals (grey area) of the replicas, true log-baseline hazard function (red solid line) and reported RMSD for the estimated survival models in the simulated *Scenario 2* (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

Outcomes related to *Scenario 2* in which the baseline hazard function was simulated from a mixture of piecewise functions highlight that the estimated *PC4* model provides the most similar fit to the true function (in terms of visual comparison). Remarkably,

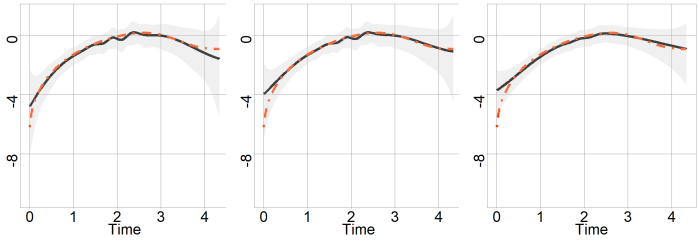
the rest of *PC* models show poor fitting as well as certain inability to capture the true shape of the baseline hazard (visual outcomes and RMSD values confirm these statements). This fact underlines the inferential sensitivity to prior scenarios in bayesian procedures and the necessity of accounting for regularization when non-parametric specifications are used in baseline hazard definitions. *PS* models also seem to capture the behaviour (increases and downs) of the true function quite accurately. Furthermore they also present the lowest values of RMSD. As it would be expected, *We* model exhibits estimates with different trend (monotonic increasing function), however it shows lower RMSD values than *PC1*, *PC2* and *PC3* models.



(a) *We.* RMSD=0.663



(b) *PC1.* RMSD=0.974 (c) *PC2.* RMSD=1.492 (d) *PC3.* RMSD=1.002 (e) *PC4.* RMSD=0.677



(f) *PS1.* RMSD=0.213 (g) *PS2.* RMSD=0.253 (h) *PS3.* RMSD=0.299

FIGURE 4.6: Average pointwise of the posterior approximate means of the log-baseline hazard estimate (black solid line) of the replicas, average of the posterior 95% credible intervals (grey area) of the replicas, true log-baseline hazard function (red solid line) and reported RMSD for the estimated survival models in the simulated *Scenario 3* (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

Scenario 3 is defined by the baseline hazard function simulated from a mixture of two Weibull distributions. In that context, *PS* models perform very well, with estimates very close to the true function, in terms of visual comparison as well as in RMSD

values. In particular, *PS3* model show the best estimates which also exhibits lower variability than its counterparts. Despite the implicit discontinuity of $PC4$ model, it also shows good performance in capturing the trend of the true function. Once again, the effect of the prior distribution also plays a strong effect in PC estimates, not only in the estimation of the shape but also in its involved uncertainty. The worst performace is accounted for the *We* model, its estimate differs substainally from the true one, however it shows the smallest variability.

4.5 Discussion

In this Chapter we have presented a few options to perform a fully bayesian analysis of time-to-event data in the context of the CPH model considering both parametric and non-parametric definitions of the baseline hazard function. Bayesian analysis allows the implementation of baseline hazard functions easily, even non-parametric proposals which are necessary in contexts in which it is expected certain complexity in the shape of the underlying function. We have considered some of the most popular proposals in the literature of the subject: the Weibull distribution as the most common parametric model and piecewise constant and piecewise cubic B-spline baseline hazards as non parametric definitions. Flexibility and overfitting was discussed within both non-parametric options with regard to different regularization schemes expressed in terms of prior distributions. These developments provide a unified framework to conduct a fully bayesian analysis of complex survival data that we hope will encourage more comprehensive analyses, which currently often rely on some version of the CPH model without further exploration. The modifiability of our approach

eases investigations into prior sensitivity and assumptions about the relationship between covariates and the hazard function.

The *S.* Typhimurium data analysis in Section 3 illustrates the proposed methods, and all models proposed (except *PC2*) verify the conclusion of Sanz-Puig *et al.* (2017) that cauliflower by-product infusion can be considered an alternative preservation treatment. Outcomes highlight that piecewise constant and B-splines specification capture flexibility in the baseline hazard function. However, piecewise constant options are less flexible given that their own definition implies discontinuous piecewise linear estimates. Furthermore, the induction of smoothing restrictions by means of a correlated prior process in non-parametric scenarios seems to overcome the problem of overfitting and unstability in estimates. The Weibull proposal behaves very well but it is not appropiate if we expect irregularities in the hazard, and data seem to provide substantial evidence of that fact. *PS* models show the better fit based on DIC and LPML criteria, compared to Weibull and piecewise constant models. It is important to note that a mechanical application of some of the proposal discussed, just as we had done with the gamma process prior in model *PS2*, can produce inadequate analysis and results of questionable validity.

We have also exemplified our model proposals through a variety of simulation studies. In particular, by simulating from Weibull, piecewise constant and a mixture of two Weibull distributions baseline hazard function distributions, respectively. In general, outcomes have shown that moderate bias can be observed in estimates of the regression coefficient for a treatment effect when fitting a CPH model in which baseline hazard function specification does not match with the specification from which data has been generated. Remarkably, *PC* models (except *PC2*) display a suspect behaviour in relation to treatment effect estimates in piecewise

simulation scenario. For baseline hazard estimates, in general, small differences between the true baseline hazard and the estimated (visual assessment) and lower RMSD values are in close relationship to the data-generating model. The Weibull model provides the most accurate baseline hazard estimates in Weibull simulated data; the *PC4* model in the case of piecewise constant simulated data, although *PS* models show the lower RMSD values; and *PS* provides the best estimates for the Weibull mixture data. Remarkably, the *PC2* model presents in all scenarios a questionable performance that may be the subject of further studies.

Although in this Chapter we have extolled the potential of bayesian inference in dealing with non-parametric specifications in the context of the CPH model, it must be stated that in many settings a simpler distribution may be adequate. However, using a more complex distribution can provide much more realistic data inference in certain situations. Some interesting issues that are beyond the contents developed here are to consider different partitions of the time axis, introduce uncertainty in its size, include new regularization proposals such as penalized complexity priors or even, to carry out a sensitivity analysis within each scenario.

Chapter 5

Bayesian mixture cure models using R-INLA

5.1 Introduction

The integrated nested Laplace approximation (INLA, Rue *et al.*, 2009) is currently an alternative to MCMC methodology within the bayesian framework. In the field of survival analysis, INLA has been adapted to analyze most of the standard models (Martino *et al.*, 2011). However, in the case of mixture cure models INLA is not directly applicable. Currently, it is possible to extend the number of models that R-INLA can fit with little extra effort. Bivand *et al.* (2015) describe a way to increase the number of models that R-INLA can manage in the framework of spatial analysis. Gómez-Rubio and Rue (2017) propose a novel methodology that combines INLA and MCMC to be applied in the context of complex hierarchical models, and Gómez-Rubio (2017) extends this approach to the field of mixture models.

In this Chapter, we propose a feasible INLA extension for estimating mixture cure models based on the above mentioned method for finite mixture models developed by Gómez-Rubio (2017). Two paradigmatic datasets, the Eastern Cooperative Oncology Group (ECOG) phase III clinical trial e1684 dataset (Kirkwood *et al.*, 1996) and the bonemarrow transplant study dataset (Kersey *et al.*, 1987) are used to illustrate our novel approach. Subsequently, the accuracy of our proposal has been evaluated by means of a thorough comparison with MCMC inference methods.

5.2 Mixture cure models

Chapter 2 includes a brief introduction to cure survival models. In mixture cure rate models, the target population consists of two subpopulations: cured and uncured individuals. However, the observed data do not include complete information about the subpopulation to which each observation belongs. For this reason, mixture cure models are often represented using a latent auxiliary variable that indicates the population to which observations belong. Random variable Z is a cure indicator variable (latent variable), with $Z = 0$ if the individual is susceptible to experience the event of interest and $Z = 1$ if it is cured for that event; η and $1 - \eta$ are the probabilities for $Z = 1$ and for $Z = 0$, respectively. The overall survival function for an individual of the target population is expressed as the mixture model

$$S(t \mid \eta, S_u) = \ P(T > t) = \eta + (1 - \eta)\, S_u(t), \tag{5.1}$$

where $S_u(t)$ denotes the survival function for individuals in the uncured subpopulation and η the cure fraction.

The mixture cure model can be considered as a combination of two models, the *incidence* model, which accounts for the probability of curation η, and the *latency* model for the event time in the uncured population. For these reasons, the most basic strategy involves separately modeling the cure proportion and the survival function of the uncured patients.

The *incidence* model in the presence of a covariate vector $\boldsymbol{x}_c$ is typically modeled using a logistic link function, $\text{logit}[\eta(\boldsymbol{\gamma})] = \boldsymbol{x}_c'\boldsymbol{\gamma}$, also expressed as

$$\eta(\boldsymbol{\gamma}) = \frac{\exp\{\boldsymbol{x}_c'\boldsymbol{\gamma}\}}{1 + \exp\{\boldsymbol{x}_c'\boldsymbol{\gamma}\}}, \tag{5.2}$$

where $\boldsymbol{\gamma}$ is the vector of regression coefficients associated to covariates $\boldsymbol{x}_c$. Note that other link functions such as the probit link or the complementary log-log link (see Robinson, 2014, for more details) can be used to connect the cure fraction with the vector of covariates $\boldsymbol{x}_c$.

The *latency* model in the presence of a covariate vector $\boldsymbol{x}_u$ can be modeled in a very different number of ways, but the most common proposals are based on the Accelerated Failure Time (AFT) model and on the Cox Proportional Hazard (CPH) model. Note that both types of models are extensively explained in Chapter 2, hence here we only introduce them briefly and adapted to the cure rate modeling context.

5.2.1 Accelerated failure time mixture cure models

The Accelerated Failure Time Mixture Cure Model (AFTMC) formulation is based on modeling the survival time T_u of individuals from the uncured subpopulation as

$$\log(T_u) = \mu + \boldsymbol{x}_u'\boldsymbol{\beta} + \sigma\epsilon. \tag{5.3}$$

The logarithmic transformation of T_u is expressed as the sum of a general mean μ plus a linear combination of the covariates $\boldsymbol{x}_u$, and an error term ϵ amplified or reduced by a scale factor σ. Common distributions for ϵ are normal, logistic and extreme value that respectively imply log-normal, log-logistic and Weibull distributions for T_u. Covariate information $\boldsymbol{x}_u$ is additively included in a linear predictor with unknown coefficients $\boldsymbol{\beta}$.

The specific case of the Weibull AFT model assumes a Weibull distribution with shape α and scale parameter $\lambda(\mu, \boldsymbol{\beta}) = -(\mu + \boldsymbol{x}_u'\boldsymbol{\beta})\alpha$ and consequently hazard and survival function

$$h_u(t \mid \alpha, \mu, \boldsymbol{\beta}) = \alpha\, t^{\alpha-1} \exp\{-(\mu + \boldsymbol{x}_u'\boldsymbol{\beta})\alpha\} \tag{5.4}$$

and

$$S_u(t \mid \alpha, \mu, \boldsymbol{\beta}) = \exp\{-t^{\alpha} \mathrm{e}^{-(\mu+\boldsymbol{x}_u'\boldsymbol{\beta})\alpha}\}. \tag{5.5}$$

5.2.2 Cox proportional hazards mixture cure models

Under the Cox Proportional Hazards Mixture Cure (CPHMC) model the hazard function for event time T_u is expressed as

$$h_u(t \mid h_{u0}, \boldsymbol{\beta}) = h_{u0}(t) \exp\{\boldsymbol{x}'_u \boldsymbol{\beta}\}, \tag{5.6}$$

where $h_{u0}(t)$ is the baseline hazard function that determines the shape of the hazard function. Model (5.6) can also be presented in terms of the survival function as

$$S_u(t \mid S_{u0}, \boldsymbol{\beta}) = [S_{u0}(t)]^{\exp\{\boldsymbol{x}'_u \boldsymbol{\beta}\}}, \tag{5.7}$$

where $S_{u0}(t) = \exp\{-\int_0^t h_{u0}(s)\mathrm{d}s\}$ is the survival baseline function.

As it was mentioned in Chapter 4, fully bayesian methods specify a model for $h_{u0}(t)$ which may be of parametric or non-parametric nature. Exponential, Weibull and Gompertz hazard functions are common proposals in the empirical literature. Mixture of piecewise constant functions or B-splines basis functions are the usual counterpart in non-parametric selections. They provide a great flexibility to the modeling but some caution is needed when eliciting prior distributions for the subsequent coefficients in order to avoid overfitting.

In the case of a Weibull $We(\alpha, \lambda)$ baseline hazard function, the hazard and survival function of T_u expressions (5.6) and (5.7) turn out to be

$$h_u(t \mid \alpha, \lambda, \boldsymbol{\beta}) = \lambda\, \alpha\, t^{\alpha-1} \exp\{\boldsymbol{x}'_u \boldsymbol{\beta}\} \tag{5.8}$$

and

$$S_u(t \mid \alpha, \lambda, \boldsymbol{\beta}) = \exp\{-\lambda t^{\alpha}\, \mathrm{e}^{\boldsymbol{x}'_u \boldsymbol{\beta}}\}. \tag{5.9}$$

Note that the Weibull is the only continuous distribution that yields both an Accelerated Failure Time and a Cox Proportional hazards model (Klein and Moeschberger, 2005). Hence, equation (5.4) is equivalent to (5.8) and equation (5.5) with (5.9) from what it follows that $\mu = -\log(\lambda)$ and $\boldsymbol{\beta}^* = -\boldsymbol{\beta}/\alpha$ ($\boldsymbol{\beta}^*$ denotes coefficients belonged to AFT specification).

5.2.3 Likelihood function

We will continue by expressing the full likelihood function for the mixture cure model. Likelihood is a key element in bayesian Inference but also in the frequentist approach, in which the common procedure of estimating parametes involves maximizing it.

Let us consider non-informative and independent right censoring. Consequently, the survival time for individual i, $i = 1, \ldots, n$, is defined as the pair (T_i, δ_i), where $T_i = \min(T_i^*, C_{Ri})$, C_{Ri} being the censoring time, and δ_i an indicator function defined as $\delta_i = 0$ when the observation is censored ($T_i^* > C_{Ri}$), and $\delta_i = 1$ when it is not.

We represent by $\mathcal{D}_{obs,i} = (t_i, \delta_i, \boldsymbol{x}_i)$ the observed data for individual i where $\boldsymbol{x}_i = (\boldsymbol{x}_{ci}, \boldsymbol{x}_{ui})$ are the possible covariates in the *incidence* and *latency* terms of the model, respectively, and $\mathcal{D}_{obs} = \cup_{i=1}^{n} \mathcal{D}_{obs,i}$. The complete data for individual i, $\mathcal{D}_i = (t_i, \delta_i, \boldsymbol{x}_i, z_i)$, also includes the value z_i of the subsequent latent variable that classifies this individual as cured or not, and $\mathcal{D} = \cup_{i=1}^{n} \mathcal{D}_i$.

The likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \mu, \boldsymbol{\beta}, S_{u0})$ for the observed data $\mathcal{D}$ is the product of the likelihood function for each individual

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \tag{5.10}$$
$$= \prod_{i=1}^{n} \eta_i(\boldsymbol{\theta})^{z_i} \, (1 - \eta_i(\boldsymbol{\theta}))^{1-z_i} \, h_{iu}(t_i \mid \boldsymbol{\theta})^{\delta_i \, (1-z_i)} \, S_{iu}(t_i \mid \boldsymbol{\theta})^{(1-z_i)}.$$

where η is the probability of being cured and $h_u(\cdot)$ is the hazard function associated to "uncured" individuals corresponding to $S_u(\cdot)$.

5.3 INLA to estimate mixture cure models

Our proposal for estimating mixture cure models by means of INLA is based on Gómez-Rubio (2017). It basically uses INLA for estimating the relevant uncertainties of the model when the latent vector which determines the subpopulation to which each individual belongs to is known. Our proposal is based on the combination of two different posterior distributions. The first one is the marginal posterior distribution for the latent cure indicator vector, $\pi(\boldsymbol{z} \mid \mathcal{D}_{obs})$, where $\boldsymbol{z} = (z_1, \ldots, z_n)$, which is approximated by means of a variation of the "modal" Gibbs sampling algorithm proposed by Gómez-Rubio (2017) for analyzing mixture models via INLA. The second is the posterior distribution for each single parameter in $\boldsymbol{\theta}$ which is obtained by means of INLA, that from now on we will represent by $\theta_{\bullet}$. Both posterior distributions are combined to approximate the posterior marginal distribution of $\theta_{\bullet}$ as

$$\pi(\theta_{\bullet} \mid \mathcal{D}_{obs}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \pi(\theta_{\bullet} \mid \mathcal{D}) \, \pi(\boldsymbol{z} \mid \mathcal{D}_{obs}), \tag{5.11}$$

where $\mathcal{Z}$ represents the parameter space of the cure indicator variables, which is the n-dimensional Cartesian product of the binary set $\{0, 1\}$.

Expression (5.11) needs some additional discussion. Note that survival observations can be censored or uncensored. In the case of censored ones, they can or can not experience the event of interest, hence their belonging to the uncured or the cured subpopulation is unknown and consequently, there is uncertainty about the value of the corresponding cure indicator variable, z_{cen}. Conversely, in the case of an uncensored observation we know that he subsequent individual has surely experienced the event of interest and consequently she/he always belongs to the uncured subpopulation. This situation implies total certainty about the latent indicator z_{unc}, zero with probability one. For this reason,

$$\begin{aligned} \pi(\boldsymbol{z} \mid \mathcal{D}_{obs}) &= \pi(\boldsymbol{z}_{cen}, \boldsymbol{z}_{unc} \mid \mathcal{D}_{obs}) \\ &= \pi(\boldsymbol{z}_{cen} \mid \mathcal{D}_{obs}), \end{aligned} \tag{5.12}$$

and consequently, expression (5.11) can be rewritten as

$$\pi(\theta_{\bullet} \mid \mathcal{D}_{obs}) = \sum_{\boldsymbol{z}_{cen} \in \mathcal{Z}_{cen}} \pi(\theta_{\bullet} \mid \mathcal{D})\, \pi(\boldsymbol{z}_{cen} \mid \mathcal{D}_{obs}), \tag{5.13}$$

where now $\mathcal{Z}_{cen}$ is the parameter space of the cure indicator variables for the censored observations, with lower dimensionality that $\mathcal{Z}$.

The above procedure can be described via the following algorithm:

Step 0. Assign initial values to the latent cure indicator of the n_{cen} censored observations, $\boldsymbol{z}_{cen}^{(0)}$, and consider $\boldsymbol{z}_{unc} = \mathbf{0}$ for the uncensored observations. Define $\boldsymbol{z}^{(0)} = \{\boldsymbol{z}_{cen}^{(0)}, \boldsymbol{z}_{unc}\}$.

For $m = 1, 2, \ldots,$ repeat:

STEP 1. Use INLA to approximate $\pi(\theta_\bullet \mid \mathcal{D}_{obs}, \boldsymbol{z}^{(m-1)}),\ \theta_\bullet \in \boldsymbol{\theta}$.

STEP 2. Obtain posterior (conditional) modes $\hat{\boldsymbol{\theta}}^{(m-1)}$ of $\boldsymbol{\theta}$ from $\pi(\theta_\bullet \mid \mathcal{D}_{obs}, \boldsymbol{z}^{(m-1)})$.

STEP 3. Sample $\boldsymbol{z}_{cen}^{(m)} = (z_{cen,1}^{(m)}, \ldots, z_{cen,n_{cen}}^{(m)})$ from the full conditional distribution for the cure latent variable (Marin *et al.*, 2005),

$$\pi(z_{cen,i}^{(m)} = 0 \mid \mathcal{D}_{obs}, \hat{\boldsymbol{\theta}}^{(m-1)}) = \frac{(1 - \eta_i(\hat{\boldsymbol{\theta}}^{(m-1)}))\, S_i u(t_i \mid \hat{\boldsymbol{\theta}}^{(m-1)})}{\eta_i(\hat{\boldsymbol{\theta}}^{(m-1)}) + (1 - \eta_i(\hat{\boldsymbol{\theta}}^{(m-1)}))\, S_i u(t_i \mid \hat{\boldsymbol{\theta}}^{(m-1)})}. \tag{5.14}$$

Note that the starting point of the algorithm begins with a random assignment of the vector $\boldsymbol{z}_{cen}$. Remember that the randomness of this assignament only concerns the censored observations of the sample because the uncensored always will belong to the uncured group. Once we have a possible configuration of vector $\boldsymbol{z}$, we estimate the *incidence* and the *latency* submodels (conditional on $\boldsymbol{z}$) using INLA and approximate $\pi(\theta_\bullet \mid \mathcal{D}_{obs}, \boldsymbol{z})$, $\theta_\bullet \in \boldsymbol{\theta}$. After that process, we use the (conditional) modes of the vector of parameters to sample from the "marginalised" full conditional posterior distribution of $\boldsymbol{z}_{cen}$. All the resulting conditionals are combined using bayesian model averaging to obtain $\pi(\theta_\bullet \mid \mathcal{D}_{obs})$ (Hoeting *et al.*, 1999; Bivand *et al.*, 2014), as in equation (5.11).

It is worth to mentioning that since we use INLA to estimate those conditional models, the modeling specification includes a wide range of distributions for both the *incidence* and the *latency* part. In particular, in the case of the *incidence* implementation, INLA supports not only the logistic link function but also the probit link

and the complementary log-log, among others. On the other hand, for the *latency* computing, INLA currently implements four popular parametric survival regression models, including the exponential, Weibull, log-normal, and log-logistic model as well as the CPH model with piecewise constant baseline hazard function (see, for example `http://www.r-inla.org/models/latent-models` for all the available latent models).

5.4 Illustrative examples

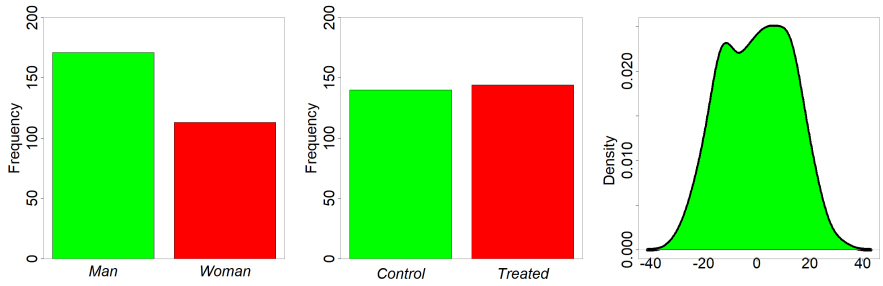
This Section illustrates our proposal for estimating mixture cure models via INLA. In particular, we consider two popular datasets: the so-called Eastern Cooperative Oncology Group (ECOG) phase III clinical trial e1684 dataset (Kirkwood *et al.*, 1996) and the bonemarrow transplant study dataset (Kersey *et al.*, 1987). In both studies, we have compared our results with the ones obtained via MCMC methods.

All analyses in this Chapter were performed on a Windows laptop with an Intel(R) Core(TM) i7-7700 3.60GHz processor. All implementations were made in the `R` environment (version 3.4.3). We used the `R-INLA` package for implementing our proposal, and the `JAGS` software (version 4.3.0) (Plummer, 2003) through the `rjags` package for MCMC.

5.4.1 ECOG study

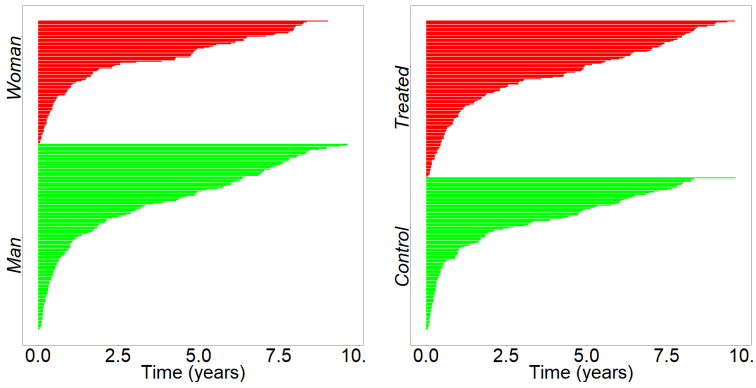
The aim of the ECOG phase III clinical trial was to evaluate the high dose interferon alpha-2b (IFN) regimen against the placebo as the postoperative adjuvant therapy (Kirkwood *et al.*, 1996). We

estimate a generic CPHMC model with baseline hazard function (h_{u0}) specified as a Weibull distribution. We have included in the analysis information of a total number of 284 observations, 88 of which are right-censored. The response variable was taken to be the relapse-free survival, in years. Covariate information included gender (0 = man, 1 = woman), group (0 = control, 1 = treatment), and age (continuous variable measured in years and centered on the sample mean). Figure 5.1 displays the frequency of the two categories of the gender and treatment covariates as well as an estimated kernel density of the age. Figure 5.2 presents a description of the response variable (in years) against gender (a) and group (b), respectively. It is worth mentioning that all covariate information was incorporated both in the *incidence* model and also in the *latency* model.



(a) sex (b) group (c) age

FIGURE 5.1: Graphical description of the ECOG study covariates: gender, group and age.



(a) sex (b) group

FIGURE 5.2: Survival times (in years) with regard to gender and group.

Modeling

The cure proportion in the *incidence* model was expressed in terms of a binary regression logistic model defined as

$$\text{logit}[\eta_i(\boldsymbol{\gamma})] = \gamma_0 + \gamma_{Woman}\, I_{Woman}(i) + \gamma_{Trt}\, I_{Trt}(i) + \gamma_{Age}\, Age_i, \tag{5.15}$$

where γ_0 represents the reference category, to be a man in the control group, and $I_A(i)$ is an indicator variable with value 1 if individual i has the characteristic A and 0 otherwise.

Survival times for the uncured subpopulation in the *latency* model, was modeled by a CPH model here expressed in terms of the survival function,

$$S_{ui}(t \mid S_{u0}(\cdot), \boldsymbol{\beta}) = [S_{u0}(t)]^{\exp\{\boldsymbol{\beta}'\boldsymbol{x}_{ui}\}}, \tag{5.16}$$

with $\boldsymbol{\beta}'\boldsymbol{x}_{ui} = \beta_{Woman}\, I_{Woman}(i) + \beta_{Trt}\, I_{Trt}(i) + \beta_{Age}\, Age_i$ and baseline survival function $S_{u0}(t) = \exp\{-\lambda t^{\alpha}\}$ specified by means of a Weibull distribution $We(\alpha, \lambda)$. The model is completed with the elicitation of a prior distribution for all uncertainties in the model.

We assume prior independence and select vague normal distributions centered at zero and variance 1,000 for all the regression coefficients in (5.15) and (5.16) as well as for $\log(\lambda)$. The elicited prior distribution for α is Ga(0.01, 0.01), a very common election in these models.

Posterior inferences

After some preliminary testing, our algorithm configuration included 50 burn-iterations followed by other 450 iterations for inference. In addition, the simulations were thinned by storing one in five iteration in order to reduce autocorrelation in the saved sample. The convergence was evaluated by examinig whether the marginal log-likelihood (conditional on $\boldsymbol{z}$) estimates achieved stability during the iteration steps of our algorithm. The posterior distribution of the remainder parameters in the mixture cure model has been obtained by using bayesian model averaging with conditional posterior marginals (on the latent cure indicator variable).

Note that the marginal likelihood is a fundamental quantity in the bayesian statistics, which is extensively adopted for bayesian model selection and averaging in various settings (Hubin and Storvik, 2016). It is approximated by INLA when the model is completely fitted with it (Gómez-Rubio, 2017). Hence, under our model approach, its computation comes down to combine by addition the marginal log-likelihood of the *incidence* and the *latency* models, given that are directly approximated by INLA.

We will compare the results obtained with our proposal to those obtained via MCMC methods with the JAGS software (Plummer, 2003). MCMC algorithm was run considering three Markov chains

		Parameter	Mean	Sd	$CI_{95\%}$	$P(\cdot > 0)$
Incidence	INLA	γ_0	-1.200	0.235	[-1.676,-0.753]	0
		γ_{Woman}	0.061	0.275	[-0.483,0.597]	0.587
		γ_{Trt}	0.573	0.271	[0.045,1.107]	0.983
		γ_{Age}	-0.015	0.010	[-0.035,0.005]	0.076
	MCMC	γ_0	-1.220	0.239	[-1.701,-0.777]	0
		γ_{Woman}	0.058	0.283	[-0.518,0.595]	0.585
		γ_{Trt}	0.572	0.277	[0.044,1.107]	0.983
		γ_{Age}	-0.016	0.011	[-0.037,0.006]	0.073
Latency	INLA	β_{Woman}	0.131	0.161	[-0.187,0.442]	0.794
		β_{Trt}	-0.106	0.154	[-0.410,0.195]	0.244
		β_{Age}	-0.007	0.005	[-0.018,0.004]	0.098
		α	0.918	0.052	[0.818,1.022]	
		λ	0.938	0.113	[0.729,1.173]	
	MCMC	β_{Woman}	0.133	0.168	[-0.201,0.437]	0.779
		β_{Trt}	-0.108	0.165	[-0.441,0.209]	0.269
		β_{Age}	-0.007	0.006	[-0.018,0.003]	0.102
		α	0.909	0.055	[0.802,1.016]	
		λ	0.921	0.114	[0.715,1.152]	

TABLE 5.1: Summary of the INLA and MCMC approximate posterior distributions: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive.

with 100,000 iterations and a burn-in period with 20,000. In addition, the chains were thinned by storing one in two hundred iteration in order to reduce autocorrelation in the saved sample and avoid space computer problems. Convergence was assessed based on the potential scale reduction factor, $\hat{R}$, and the effective number of independent simulation draws, neff (Gelman and Rubin, 1992).

Note that the number of iterations that we needed for accomplishing convergence under our proposal is much smaller than in MCMC configuration. This fact is a consequence that our algorithm only explore the parameter space of the cure indicator variable for censored observations Z_{cen} and not the full parameter space of Z because uncensored observations will always belong to the uncured subpopulation. Regarding computational times, with INLA

approach we get reliable estimates in 17 minutes and the MCMC sampler needed around 13 minutes.

Table 5.1 shows a summary of the mixture cure model parameters estimated with INLA and with MCMC-based inference. Figures 5.3 and 5.4 show the posterior marginals of the *incidence* and *latency* parameters derived from INLA (by bayesian model averaging on the conditional posterior marginals) and from MCMC. In all cases, the agreement is quite high and confirms that our approach provides similar estimates to MCMC.

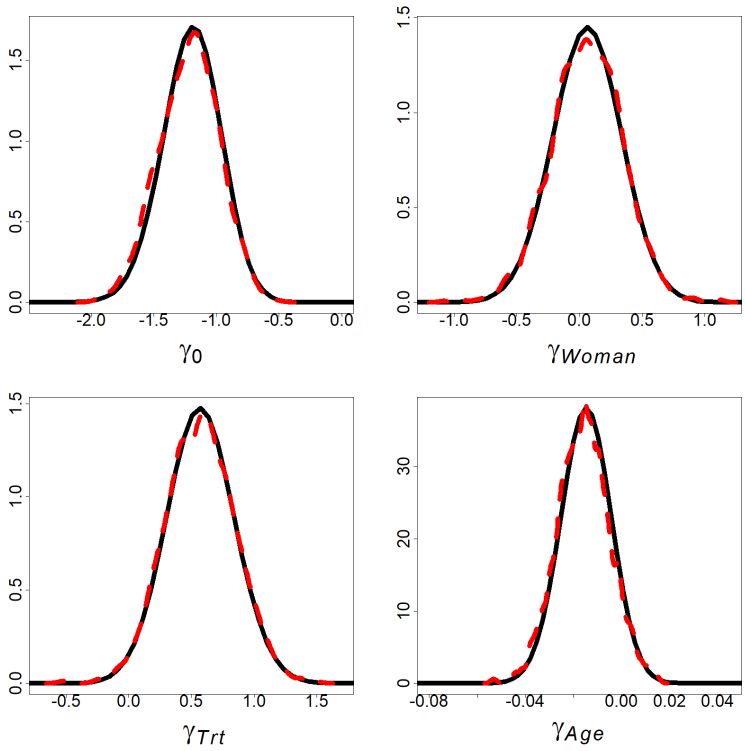


FIGURE 5.3: Posterior marginal distribution estimates for the *incidence* regression parameters approximated by INLA (black solid line) and by MCMC (red dashed line).

From the point of view of the study, it is interesting the estimation of the cure proportion as well as the survival profiles for groups of

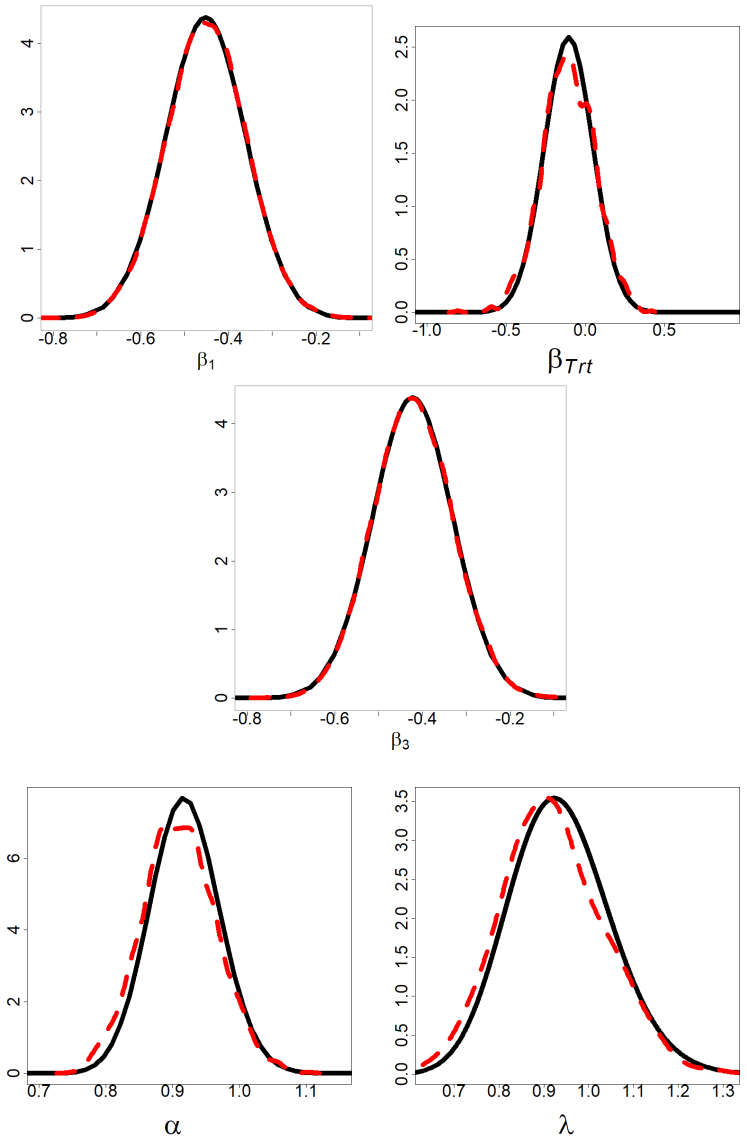


FIGURE 5.4: Posterior marginal distribution for the parameters of the *latency* model approximated by INLA (black solid line) and by MCMC (red dashed line).

individuals with certain covariate values. In this regard, it is worth mentioning that our approach does not provide a directly way to compute them. Remember that in INLA non-linear combinations or

multivariate posterior marginals are not direct available as default outcomes. On the contrary, it is very easy to estimate them via `JAGS` from the joint posterior MCMC samples and subsequently, selecting the approximate subsample from the posterior distribution of interest.

However, our algorithm allows in a simple way from the computation of the marginal log-likelihood. This function can be used to select the most likely configuration of the latent vector $\boldsymbol{z}$ that has been generated during the sampled process to approximate the posterior distribution. The function `inla.posterior.samples()` may be used to generate n samples from the approximated joint posterior distribution of the estimated model (we select the most likely model), hence these samples can then be further processed to derive posterior distributions for the quantities of interest.

Figure 5.5 and Table 5.2 present graphically and numerically respectively the posterior distribution of the cure proportion for mean aged individuals in the groups of interest: *Man-Non treated* (*M-N*), *Man-Treated* (*M-T*), *Woman-Non Treated* (*W-N*) and *Woman-Treated* (*W-T*) obtained with INLA (selecting the poster $\boldsymbol{z}$ configuration) and with MCMC. Outcomes obatined are in close agreement for both estimation methods and highlight that the group *W-T* presents the highest cure proportion estimates and the group *M-N* the lowest.

Figure 5.6 displays the mean of the posterior distribution of the "uncured" survival function for mean aged individuals in the groups of interest: *M-N*, *M-T*, *W-N* and *W-T* estimated from INLA and from MCMC. Estimation of both approaches differs slightly revealing in both cases the best survival profiles for the *M-T* group and the worst for the *W-N* one.

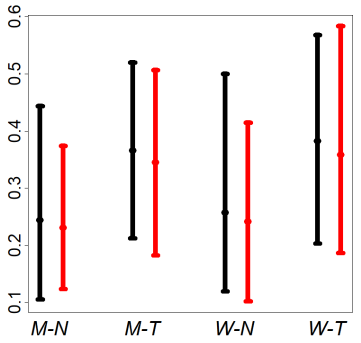


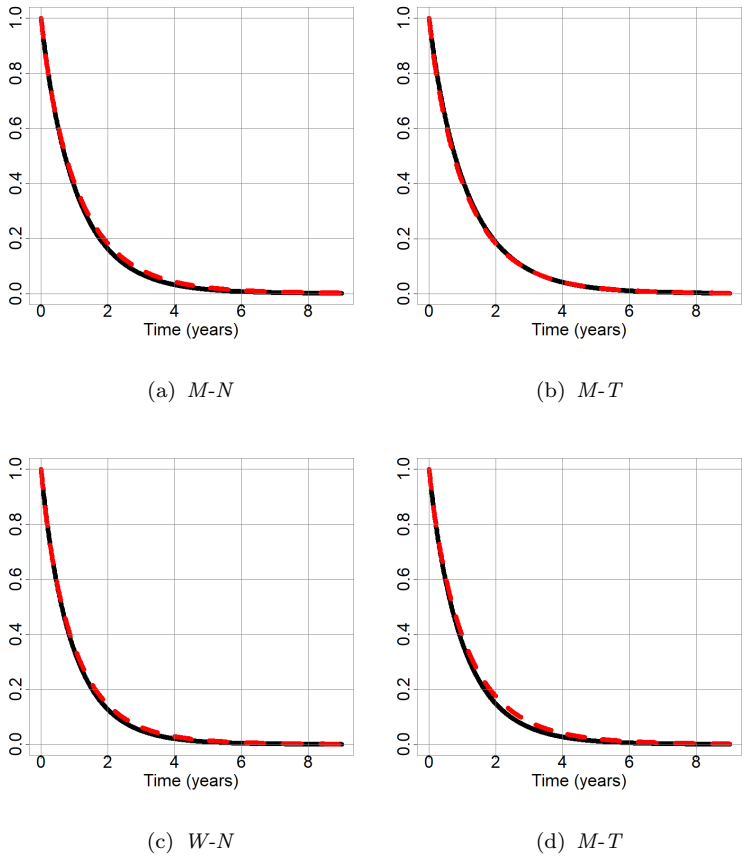
FIGURE 5.5: Posterior distribution of the cure proportion of mean aged individuals in the groups: *Man-Non Treated* (*M-N*), *Man-Treated* (*M-T*), *Woman-Non Treated* (*W-N*) and *Woman-Treated* (*W-T*) approximated by INLA (black) and by MCMC (red).

	Group	Mean	Sd	$CI_{95\%}$
INLA	*M-N*	0.242	0.042	[0.166,0.333]
	M-T	0.363	0.046	[0.280,0.453]
	W-N	0.258	0.048	[0.172,0.357]
	W-T	0.382	0.056	[0.278,0.495]
MCMC	*M-N*	0.231	0.042	[0.230,0.315]
	M-T	0.345	0.048	[0.252,0.443]
	W-N	0.242	0.049	[0.151,0.346]
	W-T	0.358	0.057	[0.248,0.475]

TABLE 5.2: Summary of posterior distribution of the probability of curation for mean aged individuals in the groups: *Man-Non Treated* (*M-N*), *Man-Treated* (*M-T*), *Woman-Non Treated* (*W-N*) and *Woman-Treated* (*W-T*) computed with INLA and MCMC.

5.4.2 Bone marrow transplant study

We consider the bone marrow transplant study dataset in Kersey *et al.* (1987) to illustrate the Weibull AFTMC model. This study was undertaken to compare autologous and allogeneic marrow



(a) *M-N* (b) *M-T*

(c) *W-N* (d) *M-T*

FIGURE 5.6: Posterior mean of the "uncured" survival function for mean aged individuals in the groups: *Man-Non Treated* (*M-N*), *Man-Treated* (*M-T*), *Woman-Non Treated* (*W-N*) and *Woman-Treated* (*W-T*) computed with INLA (black solid line) and MCMC (red dashed line).

transplantation with regard to survival times of patients affected with lymphoblastic leukemia and poor prognosis. A total of 91 patients were treated with high-doses of chemoradiotherapy and followed-up during a period between 1.4 to 5.0 years. Forty-six patients with a HLA-matched donor received allogeneic marrow (allogeneic transplanted), and forty-five patients without a matched donor received their own marrow taken during remission and purged

of leukemic cells with the use of monoclonal antibodies (autologous transplanted). The survival variable was, time to death, in days, which ranges from 11 to 1845 days. Data contain 22 right-censored observations and 69 uncensored, and in general, times to death are longer for autollogous transplanted patients than for allogeneic transplanted ones (see Figure 5.7). It is worth mentioning that all covariate information was incorporated both in the *incidence* and the *latency* terms respectively.

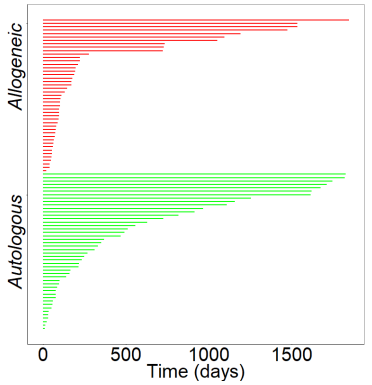


FIGURE 5.7: Survival times (in days) with regard to the type of transplant.

Modeling

Cure proportion in the *incidence* model was expressed in terms of a regression logistic model defined as

$$\text{logit}[\eta_i(\boldsymbol{\gamma})] \;=\; \gamma_0 + \gamma_{Auto}\, I_{Auto}(i), \tag{5.17}$$

where γ_0 represents the effect of the reference category, to be an individual with an allogeneic transplant, and $I_{Auto}(i)$ is an indicator variable with value 1 if individual i has been autologous transplanted.

In the *latency* model, survival times for the uncured subpopulation was modeled by a Weibull AFT model here expressed in terms of the survival function,

$$S_{ui}(t \mid \alpha, \mu, \boldsymbol{\beta}) = \exp\{-t^{\alpha} \exp\{(\mu + \boldsymbol{\beta}'\boldsymbol{x}_{ui})\}\}. \tag{5.18}$$

where μ represents the effect of the reference category, to be an individual treated with the allogeneic treatment and $\boldsymbol{\beta}'\boldsymbol{x}_{ui} = \beta_{Auto} I_{Auto}(i)$ with $I_{Auto}(i)$ as an indicator variable with value 1 if individual has received an autologous transplant and 0 otherwise. Note that formulation presented in equation (5.18) differs from the standard Weibull AFT specification described in Chapter 2 and in Section 5.2.1 for the particular case of Weibull AFTMC. This is because INLA has implemented Weibull likelihood consistently to the Cox model and we have adapted our modeling to this feature. So, in our outputs, a positive value of the risk coefficient must be associated with poor survival profiles.

The model is completed with the elicitation of a prior distribution for all parameters in the model. We assume prior independence and select vague normal distributions centered at zero and variance 1,000 for all the regression coefficients in (5.17) and (5.18) as well as for $\log(\lambda)$. The elicited prior distribution for α is Ga(0.01, 0.01).

Poesterior inferences

After some preliminary testing, our algorithm configuration for this specific model included 20 burn-iterations and other 180 iterations for inference. In addition, the simulations were thinned by storing every 2nd iteration in order to reduce autocorrelation in the saved sample. The convergence was evaluated by examinig whether the conditional (on $\boldsymbol{z}$) marginal log-likelihood estimates

achieved stability during the iteration steps of our algorithm. The posterior distribution of the remainder parameters in the mixture cure model has been obtained by using bayesian model averaging with conditional posterior marginals (on the latent cure indicator variable). As we have mentioned previously, marginal log-likelihood computation was obtained by adding the marginal log-likelihood of the *incidence* and the *latency* models conditional on $\boldsymbol{z}$, respectively, both quantities are provided by INLA.

		Parameter	Mean	Sd	$CI_{95\%}$	$P(\cdot > 0)$
Incidence	INLA	γ_0	-0.988	0.341	[-1.691,-0.351]	0
		γ_{Auto}	-0.404	0.505	[-1.407,0.575]	0.211
	MCMC	γ_0	-1.025	0.355	[-1.763,-0.367]	0
		γ_{Auto}	-0.413	0.524	[-1.437,0.665]	0.203
Latency	INLA	β_0	-6.372	0.652	[-7.709,-5.131]	0
		β_{Auto}	0.759	0.262	[0.247, 1.277]	0.998
		α	1.138	0.103	[0.941,1.343]	
	MCMC	β_0	-6.305	0.631	[-7.572,-5.118]	0
		β_{Auto}	0.754	0.267	[0.238, 1.287]	1
		α	1.124	0.101	[0.934,1.325]	

TABLE 5.3: Summary of the approximate posterior distribution for the *incidence* and *latency* parameters of the cure model obtained from INLA (by bayesian model averaging) and MCMC: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive.

We will compare the results obtained with our current approach to those obtained via MCMC with the JAGS software (Plummer, 2003). MCMC simulation was run considering three Markov chains with 200,000 iterations and a burn-in period with 40,000 iterations. In addition, the chains were thinned by storing every 400th iteration in order to reduce autocorrelation in the saved sample and avoid space computer problems. Convergence was assessed based on the potential scale reduction factor, $\hat{R}$, and the effective number of independent simulation draws, neff (Gelman and Rubin, 1992).

As we have remarked in the ECOG study, in this case our proposed method has needed less iterations than MCMC configuration to reach convergence and accuracy results. This is because we introduce information about the parametric space of the cure indicator variable Z by means of the information provided by uncensored obervations which always belong to the uncured subpopulation. So, we only have to conveniently explore Z_{cen}, the parametric space of the cure indicator variable of censored observations.

Table 5.3 shows a summary of the mixture cure model parameters obtained under our proposal and with MCMC-based inference. Figures 5.8 and 5.9 show the posterior marginal distribution of the model parameters obtained with INLA (by bayesian model averaging on the conditional posterior marginals) and with MCMC. In all cases the agreement is quite high, which confirms that our approach and MCMC provide similar outputs.

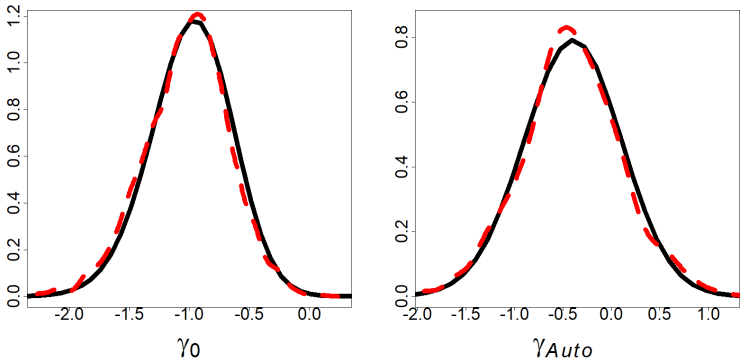


FIGURE 5.8: Posterior marginal distribution estimates for the *incidence* regression parameters approximated by INLA (black solid line) and by MCMC (red dashed line).

In the case of the estimation of derived quantites of interest, we proceed in a similar way as in the ECOG study. We estimate the cure proportion for individuals in the group of allogeneic and

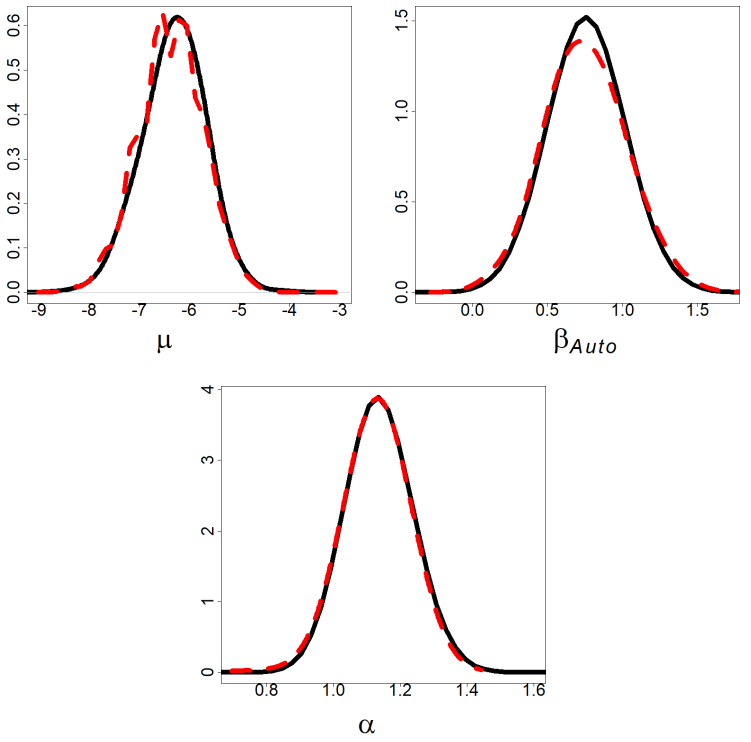


FIGURE 5.9: Posterior marginal distribution for the parameters of the *latency* model approximated by INLA (black solid line) and by MCMC (red dashed line).

autollogous transplants, as well as their "uncured" survival function. Note that the computation of these quantities with our proposal with INLA has been performed analogously as it has been done in MCMC. That is to say, taking the approximate posterior samples for all involved parameters and subsequently, computing from them the posterior distribution of interest. Note also that, to obtain posterior samples in the case of INLA outcomes, we have selected the most likely model (by means of conditional log-likelihood criteria) among all sampled models. And subsequently, through the `inla.posterior.samples()` function we have generated a sufficient number of samples from the approximated joint posterior

distribution.

Figure 5.10 and Table 5.4 present graphically and numerically the posterior distribution of the cure proportion for allogeneic and autollogous transplanted patients obtained with INLA (selecting the model with the most likely $\boldsymbol{z}$ configuration) and with MCMC. Outcomes obtained present slight differences and underline that autologous transplanted patients present a higher cure proportion posterior mean estimates than allogeneic ones, although they display a very broad degree of overlap.

Figure 5.11 displays the posterior mean of the "uncured" survival function for autologous and alogeneic transplanted patients computed from INLA and MCMC. The estimates of both quantities seems to be very close, thus indicating that our procedure has good accurary. We also can observe that autologous transplanted patients have better survival profiles.

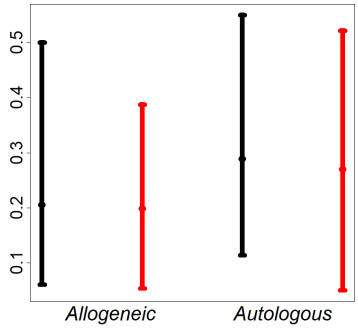
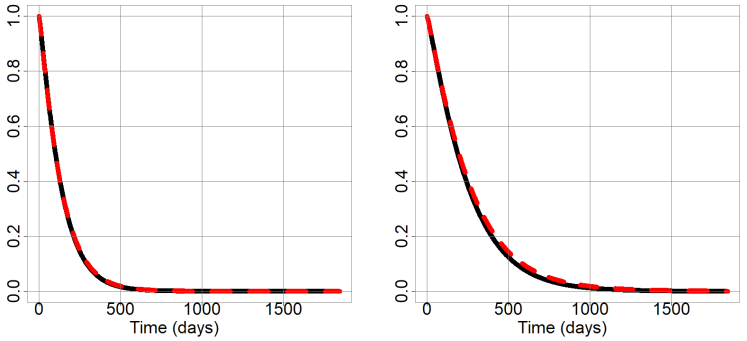


FIGURE 5.10: Posterior distribution for the cure proportion for *Autologous* and *Allogeneic* transplanted patients approximated by INLA (black) and by MCMC (red).

	Group	Mean	Sd	$CI_{95\%}$
INLA	*Allogeneic*	0.198	0.057	[0.094,0.319]
	Autologous	0.270	0.067	[0.146,0.410]
MCMC	*Allogeneic*	0.206	0.059	[0.105,0.334]
	Autologous	0.288	0.065	[0.172,0.425]

TABLE 5.4: Summary of posterior distribution of the cure proportion computed from INLA and MCMC: mean, standard deviation, 95% credible interval.



(a) *Allogeneic* (b) *Autologous*

FIGURE 5.11: Posterior mean of the uncured survival function for *Allogeneic* and *Autologous* transplanted patients computed from INLA (black solid line) and MCMC (red dashed line).

5.5 Discussion

In this Chapter we have proposed a feasible INLA extension for mixture cure models based on a general proposal for finite mixture models by Gómez-Rubio (2017). Our method combines the computation of the posterior distribution of the latent indicator variable which identifies the "cured" and "uncured" subpopulations and the *incidence* and *latency* model fitting using INLA. The Bayesian learning process is completed by approximating the posterior marginals of the parameters involved in both processes

by means of bayesian model averaging of the conditional posterior marginals computed with INLA.

Our methodological proposal has been illustrated with two specific data sets which come from well known medical studies and outcomes obtained seem to show that it is not only a sensible method but also performs quite well in practice. In fact, inference outcomes obtained under our proposal match considerably with MCMC. Remarkably, it presents several other advantages, such as, lower number of iterations to reach convergence and to explore conveniently the parametric space of the latent variable $\boldsymbol{z}$. Furthermore, the use of INLA to fit conditional models does not force the use of conjugate prior and allows the direct computation of the marginal log-likelihood, a very useful measure to tackle model selection (see Gómez-Rubio, 2017).

On the other hand, MCMC approach provides slightly faster computational times. However, our proposal can be improved by minimizing computational efforts and storage requirements. Note that INLA adjusts two complete new models (*incidece* and *latency* models) in each iteration. This leads to a computational burden due to in each iteration two complete process are generated and consequently new temporary files and other secondary process . So, if we limit the default outcomes provided by INLA and we define prior distribution based on the inference of the previous interation, computational savings can be achieved.

Chapter 6

Baseline hazard functions in bayesian joint models

6.1 Introduction

In the joint modeling framework, the Cox proportional hazards (CPH) model (Cox, 1972; Cox and Oakes, 1984) is also the most recurrent option to define the survival submodel. In that context, it is also possible accounting for the inference process without specificating the baseline hazard function (see for example Wulfsohn and Tsiatis, 1997; Henderson *et al.*, 2000). However, leaving this model component unspecified precludes the estimation of relevant outcomes such as absolute measures of the risk as well as survival individual predictions.

The bayesian treatment of the CPH model has become a natural framework to account for non-parametric specification of the baseline hazard function easily as it has been illustrated in Chapter 4. Furthermore, bayesian methodology in the joint modeling framework allows the incorporation of prior information

to the study, improving and enhancing estimation and prediction of any outcome of interest (Guo and Carlin, 2004). More specifically, it makes possible to estimate and predict characteristics of the longitudinal variable as well as the survival function estimates and the prediction of survival times for individuals in the current sample or even for new individuals that could enter to the study (Alvares, 2017).

Our main objective in this Chapter is addressing the analysis of the impact of different parametric and non-parametric proposals for the baseline hazard function in the framework of bayesian joint models. This is an important issue that naturally would need more work and dedication than the devoted in this dissertation. But we think that is interesting to take here a first look to the problem.

We know that some parametric approaches provide strictly monotone baseline hazard estimations and non-parametric choices allow for more flexible patterns. We considered the same choices for the baseline hazard function as in Chapter 4, the Weibull distribution as a parametric choice, and piecewise constant and B-splines basis functions as non-parametric proposals. We also account for regularization of the non-parametric proposals by means of the same prior scenarios. These proposals have been illustrated in a benchmark survival study devoted to assess the relationship between the risk of death or be discharged alive and a longitudinal disease severity index marker in patients hospitalized at intensive care units.

6.2 Bayesian joint models for longitudinal and survival data

Bayesian joint models for longitudinal and survival data assume a full joint distribution for the longitudinal ($\boldsymbol{y}$) and the survival process ($\boldsymbol{s}$) as well as for the individual random effects ($\boldsymbol{b}$) and relevant parameters and hyperparameters ($\boldsymbol{\theta}$). This probability distribution is usually factorized as follows

$$f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})\, f(\boldsymbol{b} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{6.1}$$

where $\boldsymbol{x}$ are baseline covariates; $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$ is the conditional joint distribution of $\boldsymbol{y}$ and $\boldsymbol{s}$ given the random effects, parameters and hyperparameters, and covariates; $f(\boldsymbol{b} \mid \boldsymbol{\theta})$ is the conditional distribution of the random effects given the hyperparameters of the model, and $\pi(\boldsymbol{\theta})$ is a prior distribution of $\boldsymbol{\theta}$. The set of covariates could also affect the particular specification of $f(\boldsymbol{b} \mid \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ but it has been omitted in equation (6.1) for simplicity.

As we have mentioned in Chapter 2, there are different approaches to properly model the correlation between both processes, which imply different factorization patterns of the joint conditional distribution $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$. We centered here in the shared-parameter approach (Albert and Follmann, 2009) in which all random-effects are common elements that connect the survival and the longitudinal processes providing conditional independence between them in the form in which the distribution in equation (6.1) turns out

$$f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})\, f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})\, f(\boldsymbol{b} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}). \tag{6.2}$$

Turning back to the bayesian paradigm and the role of the Bayes' theorem to compute the relevant posterior distribution, the posterior

distribution (Armero *et al.*, 2016),

$$\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}) \propto \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{b})\, f(\boldsymbol{b} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{6.3}$$

where $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{b})$ is the likelihood function of $(\boldsymbol{\theta}, \boldsymbol{b})$ for the observed data $\mathcal{D}$.

In next Sections, we will adapt to the specific context of our illustrative example the particular specifications of the conditional distributions of the longitudinal and survival process, $f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$ and $f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})$, and the conditional distribution of the random effects, $f(\boldsymbol{b} \mid \boldsymbol{\theta})$. Furthermore, we will also select $\pi(\boldsymbol{\theta})$ and discuss the likelihood funtion $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{b})$ in more details.

6.3 Data description

The dataset comes from a benchmark study studied in Rué *et al.* (2017) that focused on patients admitted in intensive care units (ICU) who recieved mechanical ventilation (MV). These patients were followed from the first day in MV until ICU discharge or day 30 after MV initiation, whichever ocurred the first. Two main survival events were of interest in the study: *death* in the ICU or to be *discharged alive* from the ICU. A total of 139 patients were recorded, among which 28 died, 97 were discharged alive and 14 were administrative censored (they did not experience any of both events before day 30 in MV). Figure 6.1 shows a summary of the survival data.

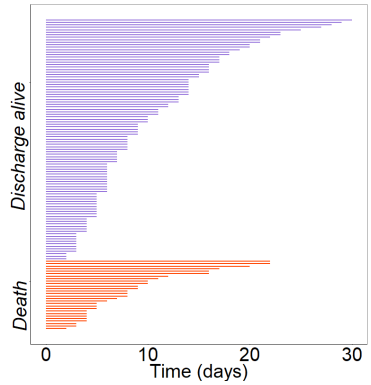
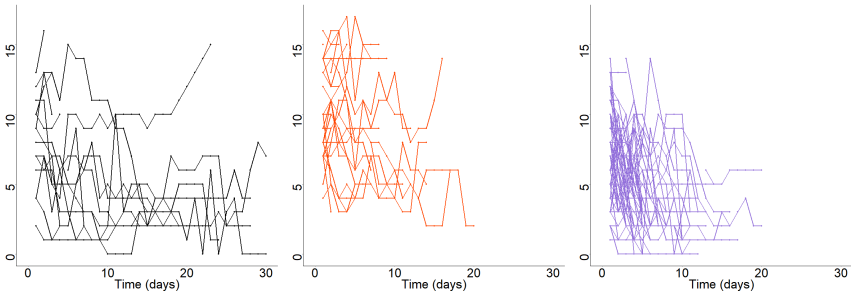


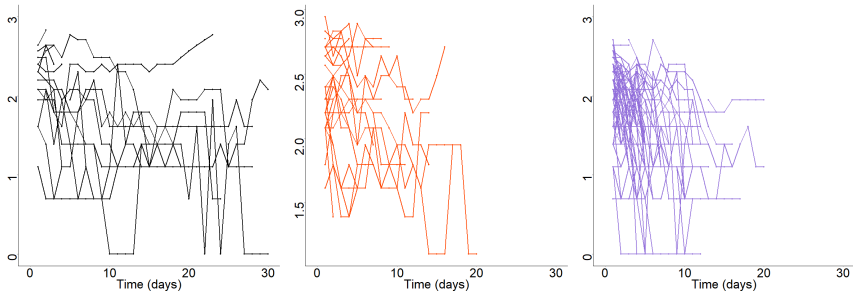
FIGURE 6.1: Survival times (days) with regard to the survival event of interest.

A severity marker, the sequential organ failure assessment (SOFA) score was daily evaluated for each individual. The SOFA score measures the degree of organ dysfunction in six human body systems: respiratory, cardiovascular, renal, coagulation, hepatic, and neurological. Each system is assessed with scores from 0 (normal) to 4 (most abnormal) and the final SOFA value is obtained by the aggregation of the six resulting punctuations.

To illustrate our proposal we consider the $SOFA$ score index in the model as $\log(SOFA + 1)$. With this transformation, we can accomodate normality for the longitudinal modeling and increase the signal of the longitudinal biomarker. See Figures 6.2a and 6.2b and observe from Figure 6.2b that trajectories for patients who died are generally higher than those for patients who were discharged alive. It is worth mentioning that in many cases, the last $SOFA$ measurement was recorded several days before the patient experienced one of the events of interest or was administratively censored.



(a) $SOFA$



(b) $\log(SOFA + 1)$

FIGURE 6.2: a) $SOFA$ and b) $\log(SOFA + 1)$ longitudinal measurements for patients who were administratively censored (black), died (red) and were discharged alive (purple).

6.4 Modeling

We propose a simple bayesian model specification in which the longitudinal process for the SOFA biomarker trajectory is specificied by means of a linear mixed-effects (LMM) model. The survival process for variables time to *death* and time to be *discharged alive* was set by a competing risks (Pintilie, 2006) model because of both possible causes of failure (*death* and to be *discharged alive*) were mutually exclusive. The longitudinal and the survival process have been connected by means of a shared random-effects approach which will be described below.

6.4.1 Longitudinal submodel

The longitudinal submodel for the ith patient, with $i = 1, \ldots, 139$, is defined as:

$$\begin{aligned} \big(y_i(t) \mid \mu_i(t), \sigma\big) &\sim \mathrm{N}\big(\mu_i(t), \sigma^2\big), \\ \big(\mu_i(t) \mid \boldsymbol{b}_i, \boldsymbol{\beta}^{(y)}\big) &= \beta_0^{(y)} + b_{0i} + \Big(\beta_1^{(y)} + b_{1i}\Big)\, t + \beta_2^{(y)} Age_i, \\ \big(\boldsymbol{b}_i \mid \sigma_0, \sigma_1\big) &\sim \mathrm{N}\left(\big(0,0\big)^\top, \mathrm{diag}\,(\sigma_0^2, \sigma_1^2)\right), \end{aligned} \tag{6.4}$$

with $y_i(t)$ denoting the ith $\log(SOFA + 1)$ patient observation at time t, which was assumed normally distributed with mean $\mu_i(t)$ and variance σ^2. Parameters $\beta_0^{(y)}$ and $\beta_1^{(y)}$ are the regression coefficients associated to the intercept and the slope of $\mu_i(t)$, respectively. Elements b_{0i} and b_{1i} are intercept and slope random effects, considered as independent and normally distributed with mean 0 and variance σ_0^2 and σ_1^2, respectively. Parameter $\beta_2^{(y)}$ is the regression coefficient associated to covariate Age_i which is the age of the ith patient, in years.

6.4.2 Survival submodel

The survival submodel was specified by means of a cause-specific hazards model (Prentice *et al.*, 1978; Gaynor *et al.*, 1993; Chen *et al.*, 2013), which is one of the most usual modeling strategies for survival analysis in the context of competing risks.

We define T_{iv}^* as the time from MV initiation to the occurrence of the event v for the ith patient, with $v = 1$ associated to *death* and $v = 2$ to be *discharged alive*. The cause-specific hazard function from a given cause v at time t for the ith patient, for $i = 1, \ldots, 139$

is defined as

$$h_{iv}(t) = \lim_{\Delta t \to 0} \frac{P\big(t \leq T_{iv}^{*} < t + \Delta t,\ \delta_i = v \mid T_{iv}^{*} \geq t\big)}{\Delta t},\ \ t \geq 0, \quad (6.5)$$

with $\delta_i = v$ indicating that event v has been occurred for the ith patient and Δt is an incremental time. It is worth mentioning that the cause-specific hazard function measures the instantaneous risk of failing at a given time from a specific cause v, among all individuals at risk.

We assume a Cox proportional hazards (CPH) model structure for each cause. Hence the cause-specific hazard function (6.5) is defined as

$$h_{iv}\big(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}\big) = h_{0v}(t) \exp\left[\beta_v^{(s)} Age_i + \rho_{0v} b_{0i} + \rho_{1v} b_{1i} t\right],\ \ t \geq 0, \quad (6.6)$$

where $h_{0v}(t)$ represents a generic baseline cause-specific hazard function at time t for $v = 1, 2$. Terms ρ_{0v} and ρ_{1v} are parameters which quantify the association between the individual characteristics of the biomarker and the risk for event v, for $k = 1, 2$. Again, Age_i represents the age of the ith patient and $\beta_v^{(s)}$ is its corresponding fixed effects coefficient, for $v = 1, 2$.

We define C_{Ri} as the administrative right censoring time (day 30 after MV initiation). We expand the definition of the event indicator and introduce the value $\delta_i = 0$ when the subsequent patient did not experience any of the events of interest and was, consequently, censored. Hence, we express the observed event time for the ith patient as $T_i = \min\left(T_{i1}^{*}, T_{i2}^{*}, C_{Ri}\right)$.

6.4.3 Cause-specific baseline hazard functions

We also deepen into the cause-specific baseline hazard function specification in order to assess its impact in the whole inferential process. We consider the same three paradigmatic scenarios discussed in Chapter 4, one parametric, based on the Weibull distribution, and two non-parametric ones, a mixture of piecewise constant functions and a cubic B-spline function, but now adapted to the competing risk environment.

Weibull

The Weibull cause-specific baseline hazard function defined for event $v = 1$ (*death*) and $v = 2$ (to be *discharged alive*) is

$$h_{0v}(t \mid \alpha_v, \lambda_v) = \lambda_v \, \alpha_v \, t^{\alpha_v - 1}, \; t > 0, \tag{6.7}$$

where $\alpha_v > 0$ and $\lambda_v > 0$ are the shape and the scale parameters of the subsequent Weibull distributions.

Mixture of piecewise constant functions

This specification allows to acommodate possible multimodalities in the shape of the cause-specific baseline hazard funcion. It is assumed to be constant within K predetermined intervals $(c_{k-1}, c_k]$ for $k = 1, 2, \ldots, K$. We consider for causes $v = 1$ (*death*) and $v = 2$ (to be *discharged alive*) a common time axis partition $= \{0, 2, 4, \ldots, 14, 18, \ldots, 30\}$, thus $K = 11$ with $c_0 = 0$ and $c_{11} = 30$ (administrative censoring time). The cause-specific baseline hazard function for cause v is defined as a flexible mixture of piecewise constant functions,

$$h_{0v}(t \mid \boldsymbol{\varphi_v}) = \sum_{k=1}^{K} \varphi_{kv} \, I_{(c_{k-1}, c_k]}(t), \; t > 0, \tag{6.8}$$

where $\boldsymbol{\varphi}_v = (\varphi_{1v}, \ldots, \varphi_{Kv})$, $I_{(c_{k-1},c_k]}(t)$ is the indicator function defined as 1 when $t \in (c_{k-1}, c_k]$ and 0 otherwise. Consistently with Chapter 4, we also refered to this proposal as *PC*.

Cubic B-spline functions

The same finite partition of the time axis specified for the *PC* cause-specific baseline hazard function is also here assumed. We define a spline basis function for the cause-specific baseline hazard function for cause v in the logarithmic scale (Murray *et al.*, 2016) to accommodate the subsequent selection of prior distributions for normality. It is defined as

$$\log\left(h_{0v}(t \mid \boldsymbol{\gamma}_v)\right) = \sum_{k=1}^{K+3} \gamma_{kv}\, B_{(k,4)}(t),\ t > 0, \tag{6.9}$$

where $\boldsymbol{\gamma}_v = (\gamma_{1v}, \ldots, \gamma_{K+3,v})$, $\{B_{(k,4)}(t), k = 1, ..., K+3\}$ is a cubic basis of B-splines with boundary knots c_0 and c_K and internal knots c_k, $k = 1, .., K-1$ (Hastie *et al.*, 2009). This specification of the cause-specific baseline hazard function is called *PS*. Note that functions in hazard equation (6.8) are also B-spline functions, in particular B-splines of order 1.

6.4.4 Prior scenarios

We considered a prior independent default scenario with non-informative marginal prior distributions. Specifically, we elicited normal distributions centered at zero with a wide known variance for the regression coefficients associated to the longitudinal and survival submodels and for the association coeffients between the longitudinal biomarker and the risk of event v, for $v = 1, 2$:

$$\begin{aligned}
\pi(\beta_0^{(y)}) = \pi(\beta_1^{(y)}) = \pi(\beta_2^{(y)}) &= \mathrm{N}(0, 1000), \\
\pi(\beta_1^{(s)}) = \pi(\beta_2^{(s)}) &= \mathrm{N}(0, 1000), \\
\pi(\rho_{01}) = \pi(\rho_{02}) &= \mathrm{N}(0, 1000), \\
\pi(\rho_{11}) = \pi(\rho_{12}) &= \mathrm{N}(0, 1000).
\end{aligned}$$

On the other hand, for the standard deviation of the error term associated to the longitudinal variable, as well as for the standard deviations associated to the intercept and slope random effects, we assume the following uniform distributions:

$$\begin{aligned}
\pi(\sigma) &= \mathrm{U}(0, 20), \\
\pi(\sigma_0) = \pi(\sigma_1) &= \mathrm{U}(0, 10).
\end{aligned}$$

In the case of the cause-specific baseline hazard parameters, we considered the same default prior scenarios used in Chapter 4. For the Weibull cause-specific baseline hazard function:

$$\begin{aligned}
\pi(\alpha_1) = \pi(\alpha_2) &= \mathrm{Ga}(0.1, 0.1), \\
\pi(\log(\lambda_1)) = \pi(\log(\lambda_2)) &= \mathrm{N}(0, 1000).
\end{aligned}$$

For the generic *PC* specification, note that we discussed four prior scenarios for the regularization process (see Chapter 4 for further details). The assumed partition of the axis time has generated 11 intervals, thus, all *PC* priors secenarios assume $k = 1, 2, ...11$.

Scenario PC1. Independent gamma prior distributions,

$$\begin{aligned}
\pi(\varphi_{k1}) &= \mathrm{Ga}(0.01, 0.01), \\
\pi(\varphi_{k2}) &= \mathrm{Ga}(0.01, 0.01).
\end{aligned}$$

Scenario PC2. Independent gamma prior distributions:

$$\begin{aligned} \pi(\varphi_{k1}) &= \text{Ga}(0.016, 0.16), \\ \pi(\varphi_{k2}) &= \text{Ga}(0.016, 0.16), \end{aligned}$$

defined by means of a discrete-time Gamma process prior (Ibrahim *et al.*, 2001) for the cumulative hazard baseline function. This prior specification corresponds to the generic expression

$$\pi(\varphi_{kv}) = \text{Ga}(w_0\,\eta_0\,(c_k - c_{k-1}), w_0\,(c_k - c_{k-1})),$$

in which all the marginal prior distributions share the same prior expectation, η_0, but the prior variance of each φ_k is inversely proportional to the corresponding interval length, $c_k - c_{k-1}$. We fix $w_0 = 0.01$ because it is a usual value which provides a high level of uncertainty to the prior, and $\eta_0 = 0.08$ after some preliminary tests.

Scenario PC3. Correlated conditional gamma prior distributions. This proposal correlates the φ_k's of adjacent intervals based on a discrete-time martingale process (Sahu *et al.*, 1997). We elicit $\pi(\varphi_{11}) = \text{Ga}(0.01, 0.01)$ and $\pi(\varphi_{12}) = \text{Ga}(0.01, 0.01)$, so that for $k = 2, \ldots, K$ prior distributions are defined in a recurrent way as

$$\begin{aligned} \pi(\varphi_{k1} \mid \varphi_{11}, \ldots, \varphi_{(k-1)1}) &= \text{Ga}(0.01,\ 0.01/\varphi_{(k-1)1}), \\ \pi(\varphi_{k2} \mid \varphi_{12}, \ldots, \varphi_{(k-1)2}) &= \text{Ga}(0.01,\ 0.01/\varphi_{(k-1)2}). \end{aligned}$$

Note that these conditional distributions imply that $\text{E}(\varphi_{kv} \mid \varphi_{1v}, \ldots, \varphi_{(k-1)v}) = \varphi_{(k-1)v}$ and $\text{Var}(\varphi_{kv} \mid \varphi_{1v}, \ldots, \varphi_{(k-1)v}) = \varphi^2_{(k-1)v}/0.01$, for $v = 1, 2$.

Scenario PC4. This proposal is analogous to the one in *Scenario PC3* but with the particularity that now prior marginal

distributions are set on $\log(\varphi_k)$'s to accomodate for normality. We set $\pi(\log(\varphi_{11})) = \mathrm{N}(0, \sigma^2_{\varphi 1})$ and $\pi(\log(\varphi_{12})) = \mathrm{N}(0, \sigma^2_{\varphi 2})$, with $\pi(\sigma^2_{\varphi 1}) \sim \mathrm{IG}(0.01, 0.01)$ and $\pi(\sigma^2_{\varphi 1}) \sim \mathrm{IG}(0.01, 0.01)$, thus the correlation between $\log(\varphi_{kv})$'s (for $v = 1, 2$) is expressed assuming the following conditional normal prior distributions, for $k = 2, \ldots, K$:

$$\pi(\log(\varphi_{k1}) \mid \varphi_{11}, \ldots, \varphi_{(k-1)1}) = \mathrm{N}(\log(\varphi_{(k-1)1}), \sigma^2_{\varphi_1}),$$
$$\pi(\log(\varphi_{k2}) \mid \varphi_{12}, \ldots, \varphi_{(k-1)2}) = \mathrm{N}(\log(\varphi_{(k-1)2}), \sigma^2_{\varphi_2}).$$

We discuss now *PS* scenarios. We also considered different prior specifications for the coefficients associated to the cause-specific baseline hazard function associated to that *PS* proposal with the aim of imposing certain flexibility restrictions and preventing the problem of overfitting (see Chapter 4 for further details). It is worth mentioning that in the *PS* proposal, $k = 1, \ldots, K + 3$, with $K = 11$. This is because the number of basis functions needed to define properly the B-spline is determined by the addition of the grade, 3, and the number of internal knots, 11. For this specific study we will have $3 + 11 = 14$ basis.

Scenario PS1. Independent normal prior distributions:

$$\pi(\gamma_{k1}) = \mathrm{N}(0, 1000),$$
$$\pi(\gamma_{k2}) = \mathrm{N}(0, 1000).$$

Scenario PS2. Hierarchical normal prior distributions:

$$\pi(\gamma_{k1} \mid \sigma^2_{\gamma_1}) = \mathrm{N}(0, \sigma^2_{\gamma_1}),$$
$$\pi(\gamma_{k2} \mid \sigma^2_{\gamma_2}) = \mathrm{N}(0, \sigma^2_{\gamma_2}),$$

where $\sigma^2_{\gamma_v}$, for $v = 1, 2$, are the common and unknown variances population which are defined as $\pi(\sigma_{\gamma_v}) \sim \mathrm{U}(0, 40)$.

Scenario PS3. Correlated conditional normal prior distributions defined as

$$\pi(\gamma_{k1} \mid \gamma_{11}, \ldots, \gamma_{(k-1)1}) = \mathrm{N}(\gamma_{(k-1)1}, \sigma^2_{\gamma_1}),$$
$$\pi(\gamma_{k2} \mid \gamma_{12}, \ldots, \gamma_{(k-1)2}) = \mathrm{N}(\gamma_{(k-1)2}, \sigma^2_{\gamma_2}).$$

We set $\pi(\gamma_{1v}) = \mathrm{N}(0, 0.1)$ and $\pi(\sigma_{\gamma_v}) \sim \mathrm{U}(0, 40)$ for $v = 1, 2$.

6.4.5 Likelihood

We will continue by expressing the full likelihood function, $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{b})$, for the observed data $\mathcal{D}$ in which $\boldsymbol{\theta}$ represents both parameters and hyperparameters and $\boldsymbol{b}$ the latent elements. Assuming that the observed data are $\mathcal{D} = \cup_{i=1}^{n} \mathcal{D}_i$, where $\mathcal{D}_i = \{ (\boldsymbol{y}_{i,1:n_i}, (t_i, \delta_i), Age_i)\}$ with $\boldsymbol{y}_{i,1:n_i} = (y_{i1}, \ldots, y_{1n_i})$ is the vector of follow-up measurements for the ith patient and y_{ij} is the observed $\log(SOFA + 1)$ score at time t_{ij}; (t_i, δ_i) is the pair with the observed survival time and the value of the event indicator for the ith patient (remember that $\delta_i = 0, 1, 2$ when the i-th pacient was censored, died or was discharged alive); and Age_i the age of the patient in years, which is the only baseline covariate considered in both longitudinal and survival submodels. The vector of parameters and hyperparameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \sigma_0, \sigma_1, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2, h_{01}, h_{02})$ with $\boldsymbol{\beta} = \left(\beta_0^{(y)}, \beta_1^{(y)}, \beta_2^{(y)}, \beta_1^{(s)}, \beta_2^{(s)}\right)$, $\boldsymbol{\rho}_1 = (\rho_{01}, \rho_{11})$, $\boldsymbol{\rho}_2 = (\rho_{02}, \rho_{12})$ and $h_{01}(\cdot)$ and $h_{02}(\cdot)$ denoting baseline hazard parameters which will depend on the cause-specific baseline hazard specification.

The likelihood function of $(\boldsymbol{\theta}, \boldsymbol{b})$ for the information $\mathcal{D}_i$ gathered for individual i can be expressed as,

$$\begin{aligned} \mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{b}) &= f\big(\boldsymbol{y}_{i,1:n_i}, (t_i, \delta_i) \mid \boldsymbol{\theta}, \boldsymbol{b}_i\big), \\ &= f\big(\boldsymbol{y}_{i,1:n_i} \mid \boldsymbol{\theta}_y, \boldsymbol{b}_i\big) f\big((t_i, \delta_i) \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big). \end{aligned} \tag{6.10}$$

with $\boldsymbol{\theta}_y$ specifying parameters and hyperparameters involved in the longitudinal process and $\boldsymbol{\theta}_s$ the subsequent ones in the survival process. It is very important to note that the factorization of the likelihood as the product of the likelihood contribution between the longitudinal and the survival information is a result of the shared-parameter approach to joint modelling, see expression (6.2).

The longitudinal contribution to the likelihood in (6.10) $f\big(\boldsymbol{y}_{i,1:n_i} \mid \boldsymbol{\theta}_y, \boldsymbol{b}_i\big)$, is the product of normal densities evaluated at observations $y_{1:n_i}$, expressed by,

$$\begin{aligned} f\big(\boldsymbol{y}_{i,1:n_i} \mid \boldsymbol{\theta}_y, \boldsymbol{b}_i\big) &= \prod_{j=1}^{n_i} \mathrm{N}(\boldsymbol{y}_{i,1:n_i} \mid \boldsymbol{\mu}_i(t_{1:n_i}), \sigma) \\ &= \prod_{j=1}^{n_i} \left(\tfrac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp\left[-\tfrac{1}{2\sigma^2} \left\|\boldsymbol{y}_{i,1:n_i} - \boldsymbol{\mu}_i(t_{i,1:n_i})\right\|^2\right]. \end{aligned} \tag{6.11}$$

where $\|\boldsymbol{v}_{1:n_i}\|^2 = \sum_{j=1}^{n_i} v_j^2$ represents the Euclidean distance.

The survival contribution to the likelihood in equation (6.10), $f\big((t_i, \delta_i) \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big)$ is the following:

$$\begin{aligned} f\big(t_i, \delta_i \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big) &= \prod_{v=1}^{2} h_{iv}\big(t_i \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big)^{I(\delta_i = v)} S_{iv}\big(t_i \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big) \\ &= \prod_{v=1}^{2} h_{iv}\big(t_i \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i\big)^{I(\delta_i = v)} \exp\big(-\textstyle\int_0^t h_{iv}(s \mid \boldsymbol{\theta}_s, \boldsymbol{b}_i)\, \mathrm{d}s\big). \end{aligned} \tag{6.12}$$

in which not censored observations contribute with the product of the cause-specific hazard function and the cause-specific survival function both evaluated at the observed survival times. Conversely, censored observations contributed with the cause-specific survival function at the observed censored time. Note that likelihood contribution for each specific cause implies the treament of the other cause observations as censored.

Remarkably, the right term of equation (6.12), $\exp\big(-\int_0^t h_{iv}(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\,\mathrm{d}s\big)$, is an analytically intractable integral. To address its approximation we made use of the Q-point Gauss-Legendre quadrature rule (Stoer and Bulirsch, 2013) with 15 quadrature points.

6.4.6 Posterior inferences

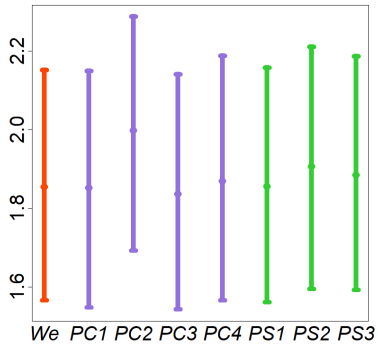
We carried out eight survival inferential processes as the result of the combination of the three generic specifications of the cause-specific baseline hazard function presented above with the different prior scenarios. The posterior distribution for each model was approximated through the `JAGS` software (Plummer, 2003). For the estimation of each joint model, we run three parallel chains with 200,000 iterations plus 40,000 (20%) dedicated to the burn-in period. Moreover, the chains were additionally thinned by storing every 400th iteration in order to reduce autocorrelation in the saved sample. In all inferential processes convergence was assessed by monitoring that the potential scale reduction factor $\hat{R}$ were close to 1 and the effective number of independent simulation draws higher than 100, $n_{eff} > 100$.

Regression coefficients

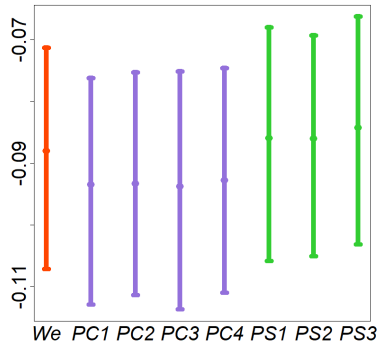
We first focused on the stability of the posterior distribution of the regression coefficients, $(\beta_0^{(y)}, \beta_1^{(y)}, \beta_2^{(y)})$ and $(\beta_1^{(s)}, \beta_2^{(s)})$, associated to the longitudinal and survival submodels, respectively, as well as the posterior distribution for the association coefficients, $(\boldsymbol{\rho}_{0v}, \boldsymbol{\rho}_{1v})$, between the longitudinal biomarker and the risk of event v, for $v = 1, 2$.

Discrepancies between the posterior marginal distributions of these parameters should only be the result of the different specifications for $h_0 v(t)$ and its prior distribution. Figure 6.3 shows the posterior mean and 95% credible interval of the posterior distribution of regression coefficients associated to the longitudinal submodel $(\beta_0^{(y)}, \beta_1^{(y)}, \beta_2^{(y)})$. The marginal posterior distributions associated to $\beta_0^{(y)}$ among all models are quite similar, with no strong differences in relation to posterior means and 95% credible intervals estimates, which show a very broad degree of overlap. *PC2* model shows marginal posterior results slightly different with higher posterior means and wide 95% credible interval. Regarding the marginal posterior distribution associated to $\beta_1^{(y)}$, the first fact that attracts our attention is that practically all posterior estimates are equal, with posterior means and 95% credible intervals very close, even in the case of the *PC2* model. Notoriously, all the values in the credible intervals are negative, indicating a decreasing of the longitudinal marker over the time. Lastly, the marginal posterior distribution associated to $\beta_2^{(y)}$ also presents a similar behaviour in all the models. Posterior means and 95% credible interval estimates are comparable, except for the *PC2* model which displays a clear lower posterior mean. It is also noticeable that all interval values are concentrated on real positive values but very close to zero, evidencing that older patients present little higher values of the longitudinal marker.

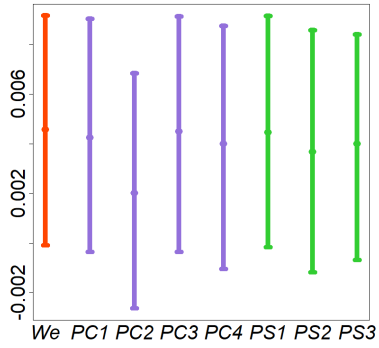
Finally, it is interesting to comment that is the *PC2* model which shows the most divergent longitudinal results as it also was observed in the survival analysis from Chapter 3



(a) $\beta_0^{(y)}$



(b) $\beta_1^{(y)}$

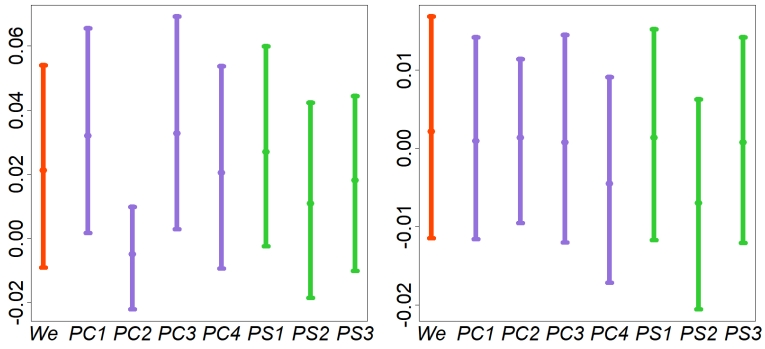


(c) $\beta_2^{(y)}$

FIGURE 6.3: Posterior mean and 95% credible interval for the longitudinal regression coefficients $\beta_0^{(y)}$ (a), $\beta_1^{(y)}$ (b) and $\beta_2^{(y)}$ (c) for all inferential processes.

Figure 6.4 shows the posterior mean and 95% credible interval of the regression coefficients associated to the survival submodel $(\beta_1^{(s)}, \beta_2^{(s)})$ corresponding to covariate *Age*. The marginal posterior distribution associated to $\beta_1^{(s)}$ presents certain differences among all inferential

process. However with the exception of *PC2*, all of them have in common that they are defined mainly on real positive values reflecting that older patients present an increased risk of death. In relation to $\beta_2^{(s)}$, its marginal posterior distribution display small discrepancies among all models, with *PC4* and *PS2* estimatations as the most diverging results. Furthermore, all posterior marginals contains the 0 value next to the middle point of its 95% credible intervals, thus highlighting that patients age does not give relevant information about the risk of the cause to be *discharged alive*.

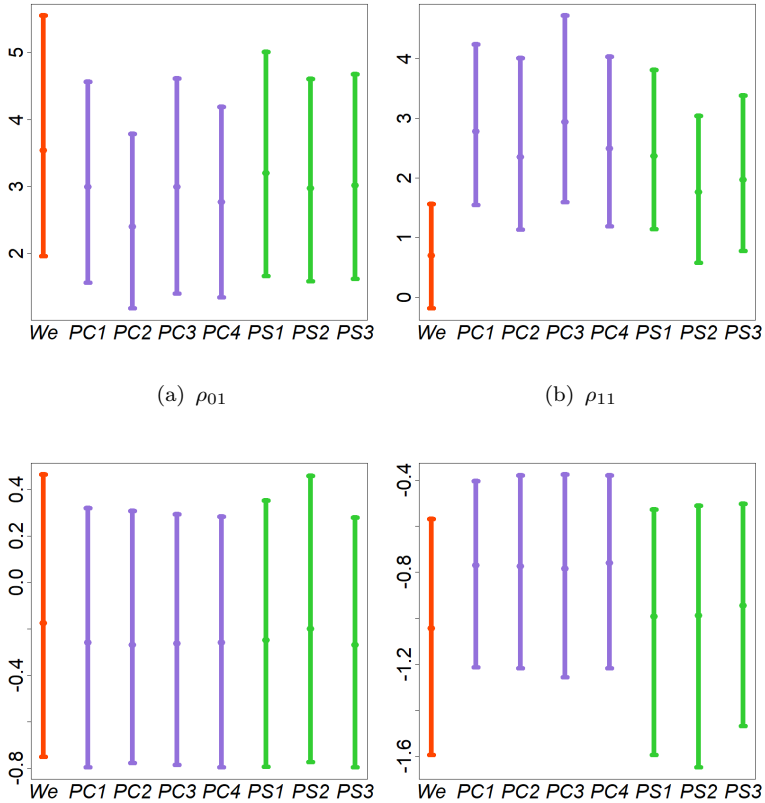


(a) $\beta_1^{(s)}$ (b) $\beta_2^{(s)}$

FIGURE 6.4: Posterior mean and 95% credible interval for the longitudinal regression coefficients $\beta_1^{(s)}$ (a) and $\beta_2^{(s)}$ (b) for all inferential processes.

Figure 6.5 shows the posterior mean and 95% credible interval of the association parameters between the longitudinal process and the risk of *death* (ρ_{01}, ρ_{11}) and of being *discharged alive* (ρ_{02}, ρ_{12}). The marginal posterior distributions of ρ_{01} and ρ_{11} present certain differences among all models. However the more evident outcomes are related to the posterior distribution of ρ_{11} in which the Weibull model, *We*, resuts in small and most concentrate estimations. For patients who finally died, these models give more relevance to

the patient condition at the starting time of the study than to the follow-up. Despite these differences it is worth mentioning that all posterior marginals have in common that they are defined on real positive values evidencing an increasing risk of *death* as the longitudinal marker value raises. For the marginal posterior distribution of ρ_{02} and ρ_{12}, it is also observable certain variability among all inferential process, although they are more clear in the posterior estimation of ρ_{11} in which PC models outline higher posterior means and narrower 95% credible intervals. However, for all models both coefficients present a dominant negative support denoting that the increase of the longitudinal marker provoques a decreasing of the risk of being *discharged alive*. For this event, it is interesting that practically all the models give the same poor relevance to the initial condition of the patient and the similar results from the We and PS models for the specific characteristics of the patient follow-up.



(a) ρ_{01} (b) ρ_{11}

(c) ρ_{02} (d) ρ_{12}

FIGURE 6.5: Posterior mean and 95% credible interval for the association parameters ρ_{01} (a), ρ_{11} (b), ρ_{02} (c) and ρ_{12} (d) for all inferential processes.

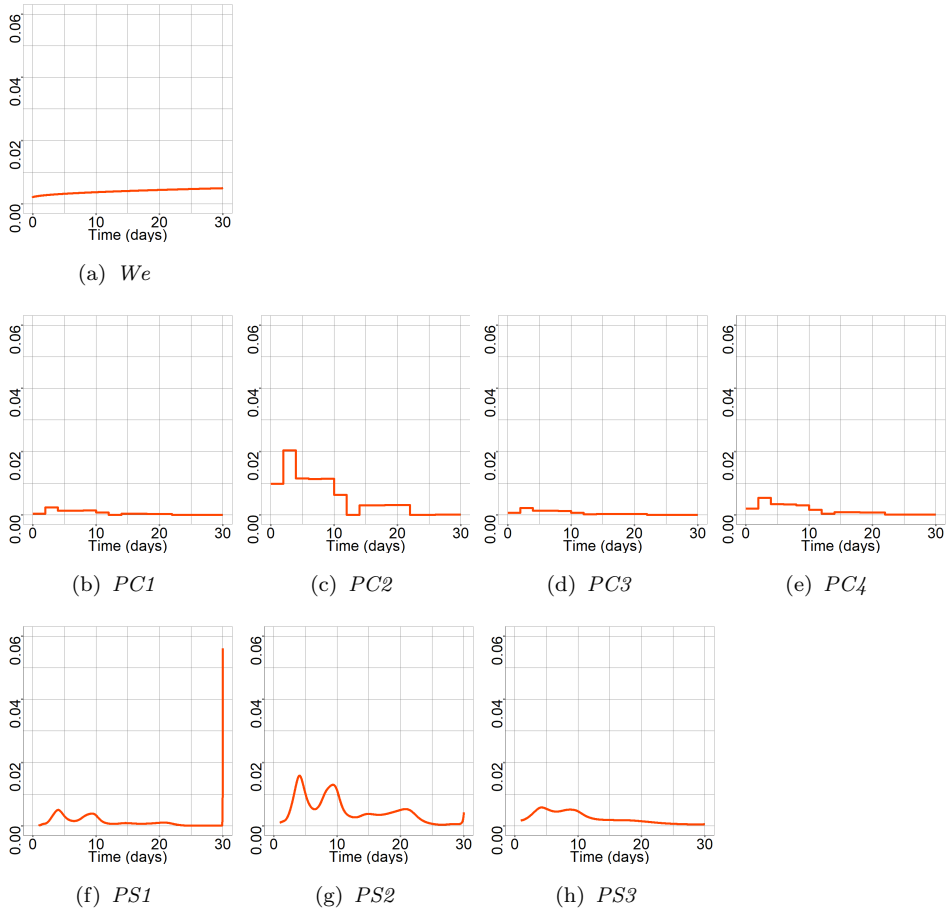
Cause-specific baseline hazard and cause-specific cumulative incidence functions

In this Section, we focus on comparing our modeling proposals with regard to the posterior distribution for the cause-specific baseline hazard estimates. Figure 6.6 is a matrix of subfigures which shows the mean of the posterior distribution, $\pi(h_{01}(t \mid \boldsymbol{\theta}, \boldsymbol{b}) \mid \mathcal{D})$

corresponding to event *death* for each one of our modeling proposals. Figure 6.7 is an analogous graphic but it includes the subsequent 95% credible intervals. Row one corresponds to Weibull modeling, row two is for piecewise constant, and row three for cubic B-spline specifications.

As it is observed, parametric and non-parametric specifications report different shapes of the posterior means of the baseline hazard function. Model *We* shows an increasing monotone hazard trend. Conversely, *PC* and *PS* models report more flexible shapes, acommodating increases and decreases of different intensities during the study period. It is worth mentioning that, despite the different nature of *PC* and *PS* specifications, their related inferences seem to outline similar trends with the exception of *PC2*. Regarding the influence of the prior setting in *PC* models, the different scenarios seem not have a strong influence, even though in the uncertainty of estimation as we can be observed in the Figure 6.7. On the other hand, *PS* models seem to be more influenced by the regularization process, thus *PS3* model display the smoothest posterior mean estimates.

Figure 6.8 is a matrix of subfigures which outlines the mean of the posterior distribution, $\pi(h_{02}(t \mid \boldsymbol{\theta}, \boldsymbol{b}) \mid \mathcal{D})$, for the event to be *discharged alive*. Figure 6.7 is an analogous graphic but it includes the subsequent 95% credible intervals. Subfigures layout follows the same pattern that the previous one. It is visually clear that model *We* shows an increasing monotone hazard and once again, *PC* and *PS* models acommodate for more flexibility with increases and decreases of different intensity during the study time. In general terms, *PC* and *PS* provide similar trends, although *PS* models evidence in a stronger way the influence of the regularization processes involved. In fact, *PS3* model presents more smoothed and accurate estimates (narrower 95% credible intervals). Regarding *PC*



(a) *We*

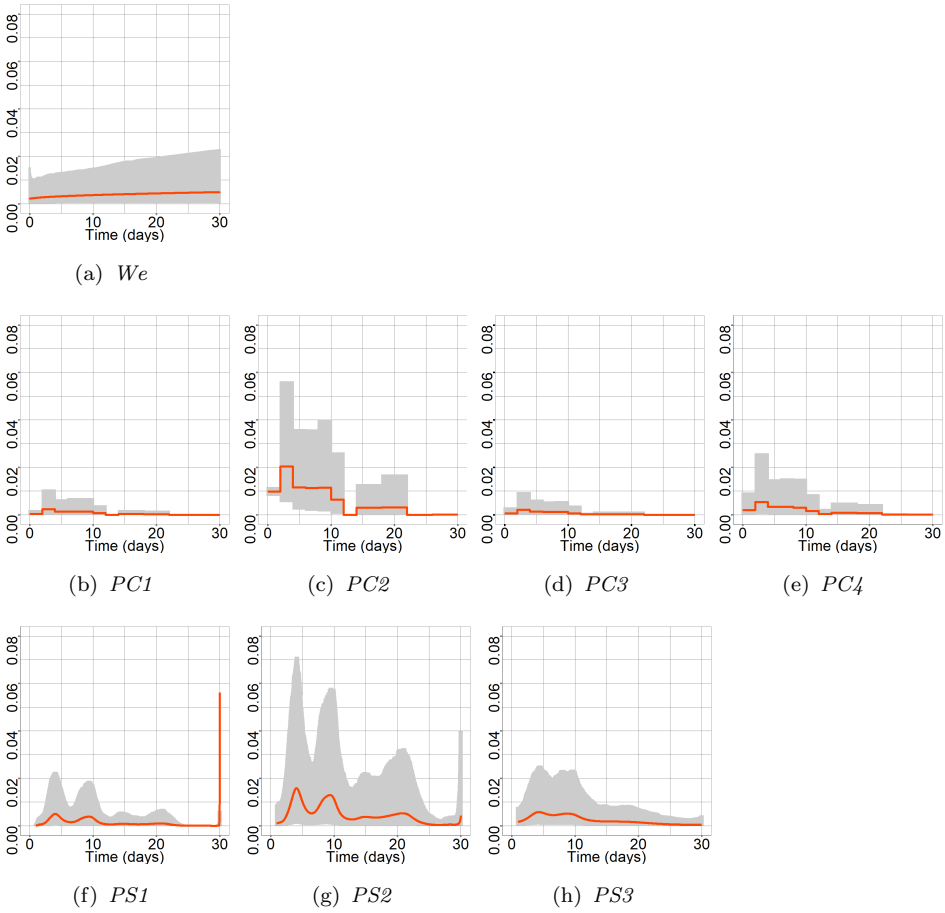
(b) *PC1* (c) *PC2* (d) *PC3* (e) *PC4*

(f) *PS1* (g) *PS2* (h) *PS3*

FIGURE 6.6: Posterior mean for the cause-specific baseline hazard function, corresponding to event *death* fot the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

models, outcomes are very similar in all scenarios, even outcomes related to *PC2* model.

The cause-specific cumulative incidence function is a very interesting concept in competing risk models. It quantifies the probability that a cause v occurs at time t or before and it is defined as



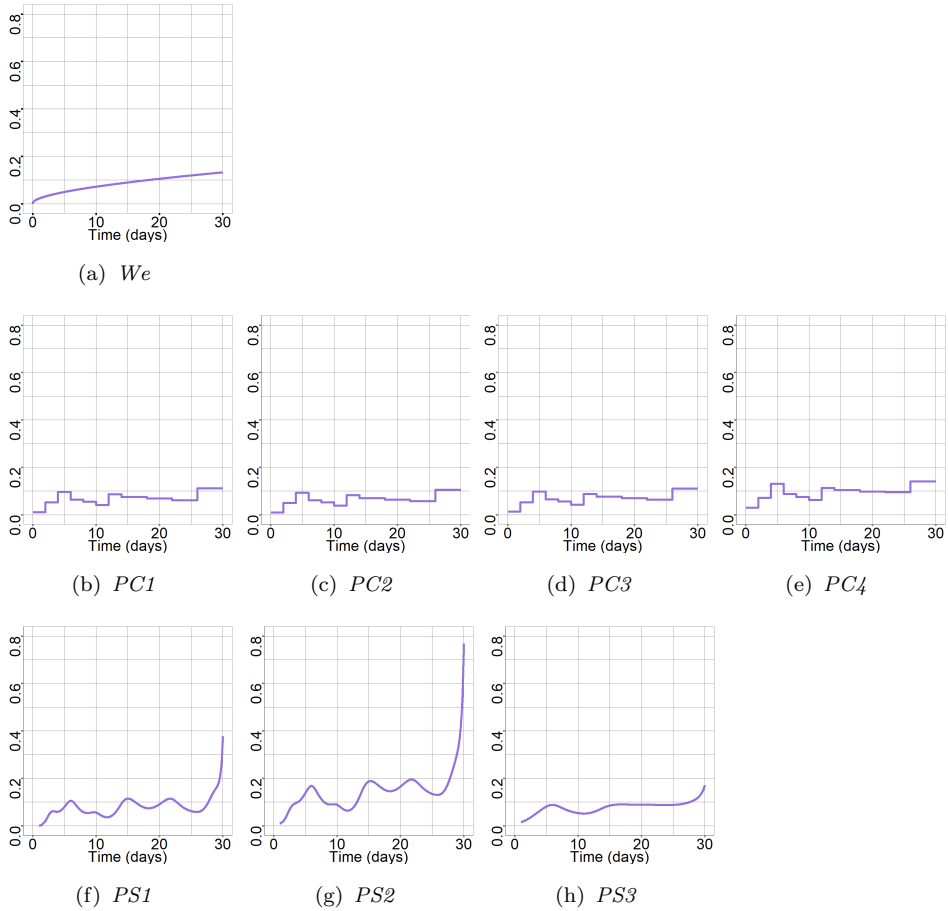
(a) *We*

(b) *PC1* (c) *PC2* (d) *PC3* (e) *PC4*

(f) *PS1* (g) *PS2* (h) *PS3*

FIGURE 6.7: Posterior mean and 95% credible interval for the cause-specific baseline hazard function, corresponding to event *death* fot the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

$$
\begin{aligned}
F_v(t \mid \boldsymbol{\theta}, \boldsymbol{b}) &= P(T \leq t, \delta = v \mid \boldsymbol{\theta}, \boldsymbol{b}) \\
&= \int_0^t h_v(u \mid \boldsymbol{\theta}, \boldsymbol{b})\, S(u \mid \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{du}, \quad t \geq 0 \text{ and } v = 1, 2,
\end{aligned}
\tag{6.13}
$$



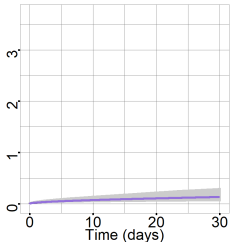
(a) *We*

(b) *PC1* (c) *PC2* (d) *PC3* (e) *PC4*

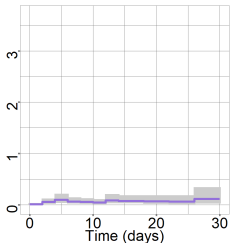
(f) *PS1* (g) *PS2* (h) *PS3*

FIGURE 6.8: Posterior mean for the cause-specific baseline hazard function, corresponding to event to be *discharged alive* fot the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

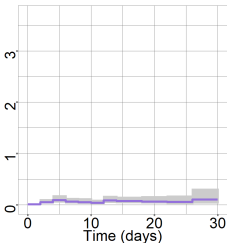
in which $h_v(u \mid \boldsymbol{\theta}, \boldsymbol{b})$ expresses th cause-specific hazard function and $S(u \mid \boldsymbol{\theta}, \boldsymbol{b})$ the overall survival function which is defined as:



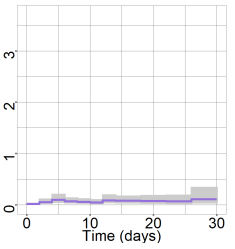
(a) *We*



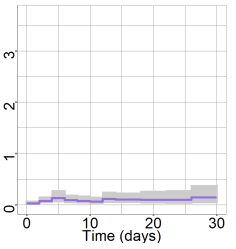
(b) *PC1*



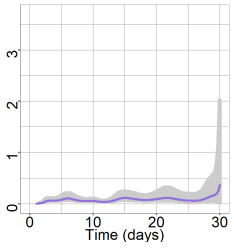
(c) *PC2*



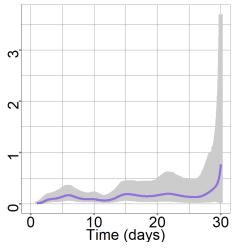
(d) *PC3*



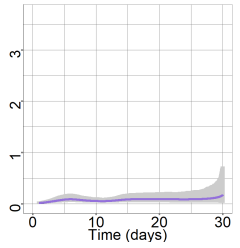
(e) *PC4*



(f) *PS1*



(g) *PS2*



(h) *PS3*

FIGURE 6.9: Posterior mean and 95% credible interval for the cause-specific baseline hazard function, corresponding to event to be *discharged alive* fot the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models)

$$S(t \mid \boldsymbol{\theta}, \boldsymbol{b}) = P(T > t \mid \boldsymbol{\theta}, \boldsymbol{b}) \\ = \exp\{-[\int_0^t h_1(u \mid \boldsymbol{\theta}, \boldsymbol{b})\,\mathrm{du} + \int_0^t h_2(u \mid \boldsymbol{\theta}, \boldsymbol{b})\,\mathrm{du}]\} \tag{6.14}$$

We computed the posterior distribution, $\pi(F_v(t \mid \boldsymbol{\theta}.\boldsymbol{b}) \mid \mathcal{D})$, of the cumulative incidence function for the events *death* and to be *discharged alive* for a generic individual aged 63 years (sample median). It is important to note that the raw distribution from which we computed this posterior distribution was the marginal conditional distribution

$$F_v(t \mid \boldsymbol{\theta}) = \int F_v(t \mid \boldsymbol{\theta}, \boldsymbol{b})\, f(\boldsymbol{b} \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{b},$$

computed by integrating the random effects in $F_v(t \mid \boldsymbol{\theta}, \boldsymbol{b})$.
Figure 6.10 is a matrix of subfigures with the same pattern that the ones in this Section which outlines the estimated posterior mean of the cumulative incidence for both events. Tables 6.1 and 6.2 show the posterior mean and 95% CI of the cumulative incidence of *death* and *alive dicharge*, respectively, at days 10, 20, and 30. Results show that the estimation of the cumulative incidence function for both outcomes presents a certain sentivity with regard to the baseline hazard specification. *We* and *PS* models report very similar estimations with posterior cumulative incidences for *death* lower and a more gentle growth slope than the cumulative incidences for being *discharged alive*. Remarkably, *We* estimation presents narrower 95% credible intervals compared to *PS* models, in which the effect of the prior correlated process is clear. On the other hand, *PC* specifications present more divergent estimatations in relation to the

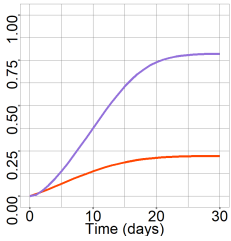
Model	$t = 10$	$t = 20$	$t = 30$
We	0.137 [0.075, 0.230]	0.211 [0.116,0.361]	0.222 [0.117, 0.372]
PC1	0.143 [0.081, 0.226]	0.196 [0.116,0.296]	0.205 [0.123, 0.302]
PC2	0.203 [0.107, 0.386]	0.291 [0.160,0.471]	0.300 [0.172, 0.472]
PC3	0.285 [0.119, 0.581]	0.393 [0.197,0.689]	0.400 [0.213, 0.690]
PC4	0.227 [0.110, 0.462]	0.323 [0.168,0.585]	0.339 [0.181, 0.585]
PS1	0.144 [0.078, 0.240]	0.279 [0.137,0.512]	0.294 [0.142, 0.512]
PS2	0.138 [0.080, 0.214]	0.232 [0.123,0.409]	0.249 [0.127, 0.429]
PS3	0.145 [0.081, 0.233]	0.254 [0.136,0.478]	0.277 [0.142, 0.491]

TABLE 6.1: Mean and 95% credible interval of the posterior cumulative incidence function for *death* at days 10, 20 and 30 for all the baseline hazard-based models.

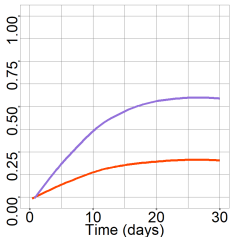
Model	$t = 10$	$t = 20$	$t = 30$
We	0.374 [0.288, 0.470]	0.739 [0.583,0.871]	0.786 [0.633, 0.887]
PC1	0.365 [0.286, 0.448]	0.527 [0.384,0.667]	0.551 [0.384, 0.713]
PC2	0.334 [0.251, 0.431]	0.453 [0.304,0.618]	0.470 [0.304, 0.670]
PC3	0.332 [0.245, 0.426]	0.410 [0.263,0.578]	0.413 [0.263, 0.615]
PC4	0.339 [0.258, 0.428]	0.443 [0.290,0.601]	0.455 [0.290, 0.655]
PS1	0.404 [0.322, 0.500]	0.674 [0.481,0.842]	0.703 [0.489, 0.854]
PS2	0.405 [0.318, 0.509]	0.705 [0.541,0.862]	0.752 [0.574, 0.874]
PS3	0.399 [0.319, 0.490]	0.680 [0.512,0.839]	0.730 [0.521, 0.865]

TABLE 6.2: Mean and 95% credible interval of the posterior cumulative incidence function for *alive dicharge* at days 10, 20 and 30 for all the baseline hazard-based models.

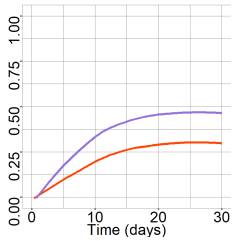
We and *PS* models and also among its counterparts, highlighting an obvious influence of the prior scenarios. In particular, *PC3* and *PC4* models provide posterior estimations of the cumulative incidence function for *death* higher than the rest. In the case of the posterior cumulative incidence for being *discharged alive* cause, it is noticeable that all *PC* models displays lower posterior estimations than the ones via *We* and *PS* models.



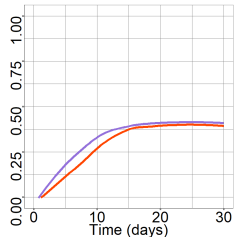
(a) *We*



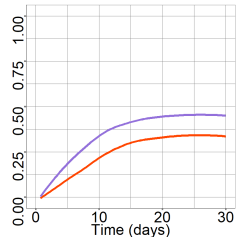
(b) *PC1*



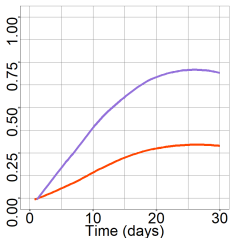
(c) *PC2*



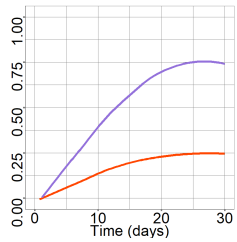
(d) *PC3*



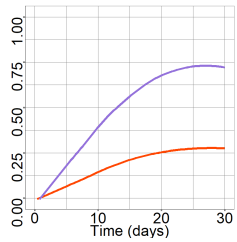
(e) *PC4*



(f) *PS1*



(g) *PS2*



(h) *PS3*

FIGURE 6.10: Posterior mean for the cumulative incidence function for for events *death* (in red) and to be *discharged alive* (in purple) under the different modeling scenarios (row one is for the *We* model, row two for *PC* models, and row three for *PS* models).

Model selection criteria

We consider the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) and the log pseudo-marginal likelihood (LPML) (Geisser and Eddy, 1979) criteria for comparing all models. As we have mentioned in Chapter 4, DIC measures fit and model complexity and fit and LPML model predictive ability. Smaller values for DIC are preferred, while LPML larger values indicate better predictive performance. Table 6.3 shows the value of the DIC and the LPML for all models considered. *PS* cause-specific baseline hazard models exhibit the best behaviours (lower DIC and larger LPML) values. In particular, the *PS1* model has the lowest DIC and the *PS3* model the largest. On the other hand, the *PC2* model shows the largest DIC value and *PC4* the smallest LPML model. According to model selection, it is important to point out that the necessity to also consider the nature of the problem to estimate. In that example, the goal of the study was to analyse the contribution of the longitudinal marker in the assessment of the patients prognosis at UCI considering that each patient in this unit can only die or to be discharged alive and send to hospital. In that respect, intensive care specialists think that *PS* model, and specifically the $PS3$ could appropriately reflect the course of the diseases in ICU patients with mechanical ventilation.

6.5 Discussion

This Chapter has presented several modeling proposals with regard to survival submodel formulation within a bayesian joint model framework with survival objectives. Within a CPH survival formulation, we have discussed model flexibility by adressing

Model	DIC	LPML
We	1630	-936.34
PC1	1625	-935.18
PC2	1636	-927.97
PC3	1625	-931.33
PC4	1631	-944.14
PS1	1595	-705.11
PS2	1604	-711.32
PS3	1606	-696.06

TABLE 6.3: DIC and LPML values for all joint models defined by mean of different specifications of the cause-specific baseline hazard function for causes *death* and to be *dischargedalive*.

non-parametric definitions of the baseline hazard function. These approaches can overcome limitations of standard parametric choices such as the exponential and Weibull distributions, which often lack enough flexibility to capture complex survival behaviours in real datasets. Furthermore, bayesian inference addresses them with simplicity and offers plausible solutions to account for the problems of overfitting and unstability associated with them by means of different prior scenarios. These extensions enable the introduction of more flexibility in bayesian joint model formulations and encourage the study of prior sensitivity as well as their influence in the whole inferential process.

We have used the UCI data described in Section 6.3 to illustrate our proposals. The inferential process was defined troughout a joint model with competing risk events for a study whose main objective was to connect the information of a longitudinal biomarker (SOFA) with two competing events, *death* and to be *discharged alive*, in mechanically ventilated patients hospitalised at intensive care units (ICU).
Outcomes analysed in Section 6.4.6 raised three important issues.

Firstly, we observed consistent estimations of the regression coefficients associated to the longitudinal process among all modeling scenarios. Secondly, posterior estimations related to regression coefficients associated to the survival competing risk model as well as for the association coefficients between the longitudinal biomarker and the risk of each event have different patterns with regard to the longitudinal process. However, they show in general a wide range of overlapping and, in consequence, go point in the same direction. Lastly, posterior inferences for the baseline hazard functions and cause-specific cumulative incidence functions emphasise that piecewise constant and B-splines specification capture a high degree of flexibility in cause-baseline hazard functions. However, *PC* models report posterior cause-specific cumulative incidence function estimates strongly different from the previous ones and exhibit a clear sensitivity to prior scenarios. Modeling based on the *We* model reports similar results that *PS* models, but DIC and LPML criteria provide more substantial evidence in favour of *PS* options.

In conclussion, this crearly indicates the potentiality of the bayesian methodology to account for flexibility in the context of joint models with survival objectives by means of non-parametric specifications of the baseline hazard function. However, we also would highlight that possibly not all studies require the setting of these proposals and maybe a simpler distribution is sufficient to describe the whole process. On the other hand, the use of more flexible modeling approaches in certain situations can provide much more realistic outputs, above all in the case in which prediction was one of the aims of the study. Some interesting issues that are beyond the contents addressed here are to consider the effect of the different modeling proposals not only in estimation but also in dynamic estimation and prediction. Note also that a very important issue

for joint models is dealing with prediction of relevant survival and longitudinal observations. This is a very relevant issue in many areas of research, particularly in medical areas focused on personalised medicine statistical procedures.

Acknowledgements

We thank Montserrat Rué, Lluís Blanch and the investigators of the Asynchronies in the ICU Group (ASYNICU) who contributed to generating the data and gave permission to use it: Candelaria de Haro, Gemma Gomà, Josefina López-Aguilar, Encarna Chacón, Marc Turon, Sol Fernández-Gonzalo, Anna Estruga, Maria Cinta Millán (*Parc Taulí Hospital Universitari. Institut d'Investigació i Innovació Parc Taulí (I3PT). Universitat Autònoma de Barcelona. Sabadell , Spain*); Carles Subirà, Rafael Fernández (*Hospital Sant Joan de Deu-Fundació Althaia. Universitat Internacional de Catalunya. Manresa, Spain*); Umberto Lucangelo (*Cattinara Hospital, Trieste University. Trieste, Italy*); Gastón Murias (Clínica Bazterrica y Clínica Santa Isabel. Buenos Aires, Argentina); Jaume Montanyà, Rudys Magrans (*Ciberes*); Robert M. Kacmarek (*Massachusetts General Hospital and Department of Anesthesiology, Harvard Medical School. Boston, MA, USA*); Guillermo M. Albaiceta (*Hospital Central de Asturias*); and Enrique Fernández-Mondejar (*Complejo Hospitalario Universitario de Granada, Granada, Spain*).

Chapter 7

Conclusions and future research

7.1 Conclusions

In this work, we have explored and developed different methodological proposals in the context of bayesian survival analysis, including the framework of joint models for longitudinal and survival data. Procedures have been addressed by means of simulated data and studies from different biometrical areas of research with the aim of illustrating its broad applicability and power. The main conclusions of this dissertation, inspired and firmly rooted in these ideas, are summarized below.

- Results in Chapter 3 support three relevant conclusions. Firstly, the potentiality of the bayesian methodology in the context of survival analysis as well as the existence of different available software for implementing simple and complex inferential processes. Secondly, the extremely utility of bayesian survival analysis in certain areas of research in

which its application is currently limited. Thirdly, the great capability of this methodology to provide robust inferences and deal with simplicity the presence of different censoring and truncation schemes in a simple and natural way.

- The results in Chapter 4 underline the usefulness of bayesian methods to incorpore flexibility in the Cox Proportional Hazard model. In this respect, we observed that non-parametric specifications of the baseline hazard function are able to provide an appropiate proposal to increase modeling adaptability. In adition, the possiblity of introducing some restrictions throughout the definition of correlated prior distributions is a precious tool to minimise the common problems of overfitting and unstability associated with them. Remarkably, these proposals overcome certain limitations related to the *so called* partial likelihood approach inherent to frequentist approach that deals with the estimation process leaving the baseline hazard function unspecified. These methods were applied to a benchmark dataset as well as in different simulated scenarios which highlighted the importance of estimating and capturing the true shape of the baseline hazard function to complete the whole inference process and provide accurate results to other quantities of interest, such as posterior survival probabilities.

- The main results in Chapter 5 reinforce the already proven capability of the INLA as an alternative to MCMC methodology to perform bayesian survival analysis. Our proposal is focused on extending the use of INLA to mixture cure models by means of a decomposition of the relevant posterior marginal distributions in terms of conditional posterior distributions given all latent information and the use

of "modal" Gibbs sampling. It is discussed in two benchmark studies with good and accurate inferential results.

- Results obtained in Chapter 6 support again the potential of the bayesian methodology but in the context of joint models for longitudinal and survival data with survival objectives. The bayesian approach to joint models allows to complete any inferential process (longitudinal, time-to-event, and association between both issues), quantify uncertainty estimation and deal with censoring phenomenon efficiently. This Chapter stresses the strengths of the bayesian approach to to introduce model flexibility in the survival modeling by means of similar scenarios to those discussed in Chapter 4 in a benchmark data set.

7.2 Future research

In overall, this PhD dissertation has discussed a range of survival structures with the aim of addressing certain important issues in the field of survival analysis. But certainly, the scope of time-to-event analysis is immense. Consequently, there is a great quantity of interesting research topics that have not been addressed so far. We only focus here on some possible extensions related to the research discussed in this work.

- Extend the flexibility scenario to the specification of mixtures cure models and multistate models based on CPH modeling.
- Implement flexible approaches based on piecewise cubic B-splines baseline hazard function to INLA software.

- Investigate new model structures that accommodate flexibility in the context of CPH modeling.
- Improve the INLA algorithm proposed in Chapter 5 for mixture cure models in terms of computational efforts and storage requirements.
- Explore the capability of the INLA software to account for survival analysis with the aim of implementing modeling extensions with potential applied interest.
- Assess the influence of the specification of baseline hazard functions on prediction in bayesian joint models with survival processes defined by means of CPH models.
- Start implementing and validating non-standard bayesian joint models with INLA.
- Discuss model selection and comparison in bayesian joint models via Bayes factors.

Appendix A

Inversion method

The inversion method for generating observations from probability distributions. In the most simple case where the variable of interest T^* is continuous and has an increasing distribution function $F(t)$, $F^{-1}(U)$, where U is a standard uniform random variable, is the main tool to generate random samples from the distribution of T^*.

The survival function of a random variable T^* modeled by means of a generic CPH model is (see equation (2.19)):

$$S(t) = \exp\{-H_0(t)\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\}, \tag{A.1}$$

where the conditional notation in the survival function has been omitted for simplicity. Consequently, the cumulative distribution function for T^* will be

$$F(t) = 1 - \exp\{-H_0(t)\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\}. \tag{A.2}$$

The inversion method in this case establishes the following expression

$$U = \exp\{-H_0(T^*)\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\} \sim \mathrm{U}(0,1), \tag{A.3}$$

or also

$$T^* = H_0^{-1}\{-\log(U)\exp\{-\boldsymbol{x}'\boldsymbol{\beta}\}\}, \tag{A.4}$$

given that the baseline hazard function is a positive function in all its domain, $h_0(t) > 0$ for all t, and $H_0(\cdot)$ can be inverted. Thus, simulation of survival times depends only on the calculation of the inverse of the cumulative hazard function.

In the case of the Weibull baseline hazard function, We(α, λ), cumulative hazard function has a closed form,

$$H_0(t) = \lambda\, t^{\alpha}, \tag{A.5}$$

and its inverse can be obtained directly as,

$$H_0^{-1}(t) = (\lambda^{-1}\, t)^{\frac{1}{\alpha}}. \tag{A.6}$$

Consequently, survival times are generated as,

$$T^* = \left(-\frac{\log(U)}{\lambda \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}} \right)^{\frac{1}{\alpha}}. \tag{A.7}$$

Next, we consider the scenario defined by a mixture piecewise constant baseline hazard functions defined through a finite partition of the time axis with knots $0 = c_0 \leq c_1 \leq \ldots \leq c_K$ and a

baseline hazard vector $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \ldots, \varphi_K)$. For a given interval $c_{m-1} \leq t < c_m$ with $m = 1, \ldots, K$, the cumulative hazard function is defined as

$$H_0(t) = \sum_{l=1}^{m-1} \varphi_l (c_l - c_{l-1}) + \varphi_m (t - c_{m-1}), \tag{A.8}$$

hence its inverse cumulative hazard function can be expressed as

$$H_0^{-1}(t) = c_{m-1} + \frac{t}{\varphi_m} - \frac{\sum_{l=1}^{m-1} \varphi_l (c_l - c_{l-1})}{\varphi_m}. \tag{A.9}$$

Survival times are thus generated from

$$T^* = c_{m-1} - \frac{\log(U)}{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\,\varphi_m} - \frac{\sum_{l=1}^{m-1} \varphi_l (c_l - c_{l-1})}{\varphi_m}. \tag{A.10}$$

In this case, it is important to note that since times are generated related to the condition $c_{m-1} \leq t < c_m$, the simulation process requires the imposition of the following constraint,

$$-\exp\{\boldsymbol{x}'\,\beta\} \sum_{l=1}^{m-1} \varphi_l (c_l - c_{l-1}) < \log(U) \leq \exp\{\boldsymbol{x}'\,\boldsymbol{\beta}\} \sum_{l=1}^{m-1} \varphi_l (c_l - c_{l-1}). \tag{A.11}$$

The cumulative baseline hazard function for CPH times in which the baseline hazard function is a mixture of two Weibull distributions, We(α_1, λ_1) and We(α_2, λ_2), with p as the mixing probability parameter has a closed form expression:

$$H_0(t) = -\log(p \exp\{-\lambda_1\, t^{\nu_1}\} + (1-p)\exp\{-\lambda_2\, t^{\nu_2}\}). \tag{A.12}$$

When $H_0(t)$ in equation (A.12) is substituted into equation (A.3), it produces an expression which cannot be analytically solved,

being necessary the use of root finding techniques to overcome this situation. Crowther and Lambert (2013) propose the use of the Brent's univariate root-finding method or the Newton-Raphson root finder.

Bibliography

A. Aballay, P. Yorgey, and F. M. Ausubel. *Salmonella* Typhimurium proliferates and establishes a persistent infection in the intestine of *Caenorhabditis elegans*. *Current Biology*, 10(23):1539–1542, 2000.

P. S. Albert and D. Follmann. *Longitudinal data analysis*, chapter Shared-parameter models, pages 433–452. New York, Chapman & Hall/CRC Press, 2009.

D. Alvares. *Sequential Monte Carlo methods in Bayesian joint models for longitudinal and time-to-event data*. PhD thesis, Universitat de València, 2017.

C. Armero and M. J. Bayarri. Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15(1-4):401–417, 1994.

C. Armero, A. Forte, H. Perpiñán, M. J. Sanahuja, and S. Agustí. Bayesian joint modeling for assessing the progression of chronic kidney disease in children. *Statistical Methods in Medical Research*, 27(1): 298–311, 2016.

P. C. Austin. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958, 2012.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Boca Raton, Chapman & Hall Crc Press, 2014.

T. Bayes and R. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763.

C. Belitz, A. Brezger, T. Kneib, S. Lang, and N. Umlauf. BayesX software for Bayesian inference in structured additive regression models version 2.0.1. *URL http://www.bayesx.org*, 2015.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24 (11):1713–1723, 2005.

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

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

R. S. Bivand, V. Gómez-Rubio, and H. Rue. Approximate Bayesian inference for spatial econometrics models. *Spatial Statistics*, 9:146–165, 2014.

R. S. Bivand, V. Gómez-Rubio, and H. Rue. Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software*, 63(20), 2015.

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

C. De Boor. *A practical guide to splines*. New York, Springer-Verlag, 1978.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

S. Brilleman. *simsurv: Simulate Survival Data*, 2018. R package version 3.3.0.

K. Chai-Hoon, S. Jiun-Horng, M. S. Shiran, R. Son, S. Sabrina, A. S. Noor Zaleha L., Learn-Han, and C. Yoke-Kqueen. *Caenorhabditis elegans*-based analysis of *Salmonella enterica*. *International Food Research Journal*, 17(4):845–852, 2010.

A. G. Chapple. *SimSCRPiecewise: Simulates Univariate and Semi-Competing Risks Data Given Covariates and Piecewise Exponential Baseline Hazards*, 2016. R package version 0.1.1.

M.-H. Chen, M. Castro, M. Ge, and Y. Zhang. *Handbook of Survival Analysis*, chapter Bayesian Regression Models for Competing Risks, pages 180–196. Boca Raton, Chapman and Hall/CRC Press, 2013.

R. Christensen, J. Wesley, A. Branscum, and T. E. Hanson. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, Chapman & Hall/CRC Press, 2011.

D. Collet. *Modelling survival data in medical research*. Boca Raton, Chapman and Hall/CRC Press, 2015.

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

D. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275, 1968.

D. R. Cox and D. Oakes. *Analysis of Survival Data*. Boca Raton, Chapman & Hall/CRC Press, 1984.

M. J. Crowther. *Development and application of methodology for the parametric analysis of complex survival and joint longitudinal-survival data in biomedical research*. PhD thesis, Department of Health Sciences, 2014.

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

V. DeGruttola and X. M. Tu. Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. *Biometrics*, 50(4): 1003–1014, 1994.

P. Dellaportas and A. F. M. Smith. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, pages 443–459, 1993.

L. Fahrmeir and T. Kneib. *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford, Oxford University Press, 2011.

D. Gallitelli. The ecology of Cucumber Mosaic Virus and sustainable agriculture. *Virus Research*, 71(1-2):9–21, 2000.

J. J. Gaynor, E. J. Feuer, C. C. Tan, D. H. Wu, C. R. Little, D. J. Straus, B. D. Clarkson, and M. F. Brennan. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association*, 88 (422):400–409, 1993.

S. Geisser and W. F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

A. E. Gelfand. *Markov Chain Monte Carlo in Practice*, chapter Model determination using sampling-based methods, pages 146–151. London, Chapman & Hall/CRC Press, 1996.

A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457 – 472, 1992.

V. Gómez-Rubio. Mixture model fitting using conditional models and modal Gibbs sampling. *arXiv:1712.09566*, 2017.

V. Gómez-Rubio and H. Rue. Markov chain Monte Carlo with the Integrated Nested Laplace Approximation. *Statistics and Computing*, pages 1–19, 2017.

P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

X. Guo and B. P. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):1 – 9, 2004.

S. K. Han, D. Lee, H. Lee, D. Kim, H. G. Son, J.-S. Yang, S.-J. Lee, and S. Kim. Oasis 2: online application for survival analysis 2 with features for the analysis of maximal lifespan and healthspan in aging research. *Oncotarget*, 7(35):56147–56152, 2016.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. New York, Springer-Verlag, 2009.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

V. Henschel, J. Engel, D. Hölzel, and U. Mansmann. A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects. *BMC Medical Research Methodology*, 9(1):1–15, 2009.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417, 1999.

A. Hubin and G. Storvik. Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA). *arXiv:1611.01450*, 2016.

J. G. Ibrahim, M. H. Chen, and D. Sinha. *Bayesian Survival Analysis*. New York, Springer-Verlag, 2001.

R. Jackson. *Some statistical methods for the analysis of survival data in cancer clinical trials*. PhD thesis, University of Liverpool, 2015.

J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. New Jersey, John Wiley & Sons, 2002.

C. Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7): 263–270, 2016.

J. H. Kersey, D. Weisdorf, M. E. Nesbit, T. W. LeBien, W. G. Woods, P. B. McGlave, T. Kim, D. A. Vallera, A. I. Goldman, B. Bostrom, et al. Comparison of Autologous and Allogeneic Bone Marrow Transplantation for Treatment of High-Risk Refractory Acute Lymphoblastic Leukemia. *New England Journal of Medicine*, 317(8): 461–467, 1987.

S. W. Kim and J. G. Ibrahim. On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis*, 6(4):331–341, 2000.

J. M. Kirkwood, M. H. Strawderman, M. S. Ernstoff, T. J. Smith, E. C. Borden, and R. H. Blum. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. *Journal of Clinical Oncology*, 14(1): 7–17, 1996.

G. Kisluk, E. Kalily, and S. Yaron. Resistance to essential oils affects survival of *Salmonella enterica* serovars in growing and harvested basil. *Environmental Microbiology*, 15(10):2787–2798, 2013.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, Springer-Verlag, 2005.

A. Labrousse, S. Chauvet, C. Couillault, C. L. Kurz, and J. J. Ewbank. *Caenorhabditis elegans* is a model host for *Salmonella* Typhimurium. *Current Biology*, 10(23):1543–1545, 2000.

P. C. Lambert. Modeling of the cure fraction in survival studies. *Stata Journal*, 7(3):351, 2007.

S. Lang and A. Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

P. S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1986.

J. F. Lawless. *Statistical Models and Methods for Lifetime Data*, volume 362. New Jersey, John Wiley & Sons, 2011.

H. Lecoq, B. Moury, C. Desbiez, A. Palloix, and M. Pitrat. Durable virus resistance in plants through conventional approaches: a challenge. *Virus Research*, 100(1):31–39, 2004.

E. T. Lee and J. W. Wang. *Statistical Methods for Survival Data Analysis*. New Jersey, John Wiley & Sons, 2013.

K. H. Lee, F. Dominici, D. Schrag, and S. Haneuse. Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. *Journal of the American Statistical Association*, 111(515):1075–1095, 2016.

R. Little. *Longitudinal data analysis*, chapter Selection and pattern-mixture models, pages 409–431. New York, Chapman and Hall/CRC Press, 2009.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325 – 337, 2000.

J. M. Marin, K. Mengersen, and C. P. Robert. *Handbook of Statistics*, chapter Bayesian modelling and inference on mixtures of distributions, pages 459–507. Elsevier, 2005.

S. Martino, R. Akerkar, and H. Rue. Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3):514–528, 2011.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

J. Mosqueda-Melgar, R. M. Raybaudi-Massilia, and O. Martín-Belloso. Microbiological shelf life and sensory evaluation of fruit juices treated by high-intensity pulsed electric fields and antimicrobials. *Food and Bioproducts Processing*, 90(2):205–214, 2012.

T. A. Murray, B. P. Hobbs, D. J. Sargent, and B. P. Carlin. Flexible Bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian Analysis*, 11(2):381, 2016.

A. N. Olaimat and R. A. Holley. Factors influencing the microbial safety of fresh produce: a review. *Food Microbiology*, 32(1):1–19, 2012.

S. Perra. *Objective bayesian variable selection for censored data*. PhD thesis, Universita'degli Studi di Cagliari, 2013.

M. Pintilie. *Competing risks: a practical perspective*. John Wiley & Sons, 2006.

M. Plummer. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson, N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978.

C. Proust-Lima, V. Philipps, A. Diakite, and B. Liquet. `lcmm`: Extended Mixed Models Using Latent Classes and Latent Processes. *R package version*, 1(2), 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton, Chapman & Hall/CRC Press, 2012.

C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, Springer-Verlag, 2007.

C. Robert. Bayesian computational tools. *Annual Review of Statistics and Its Application*, 1(1):153 – 177, 2014.

M. Robinson. *Mixture cure models: Simulation comparisons of methods in R and SAS*. PhD thesis, University of South Carolina, 2014.

P. Royston. Estimating a smooth baseline hazard function for the Cox model. *Department of Statistical Science, University College London*, Research report, 314, 2011.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, Chapman & Hall/CRC press, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 71(2):319 – 392, 2009.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

M. Rué, E. R. Andrinopoulou, D. Alvares, C. Armero, A. Forte, and L. Blanch. Bayesian joint modeling of bivariate longitudinal and competing risks data: An application to study patient-ventilator asynchronies in critical care patients. *Biometrical Journal*, 59(6): 1184–1203, 2017.

S. K. Sahu, D. K. Dey, H. Aslanidou, and D. Sinha. A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, 3(2):123–137, 1997.

M. Sanz-Puig, E. Lázaro, C. Armero, D. Alvares, A. Martínez, and D. Rodrigo. *S.* Typhimurium virulence changes caused by exposure to different non-thermal preservation treatments using *C. elegans*. *International Journal of Food Microbiology*, 262:49–54, 2017.

R. Schoot, David D. Kaplan, D. Jaap, J. B. Asendorpf, F. J. Neyer, and M. A. Aken. A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, 85(3):842–860, 2014.

X. Sem and M. Rhen. Pathogenicity of *Salmonella enterica* in *Caenorhabditis elegans* relies on disseminated oxidative stress in the infected host. *PLoS One*, 7(9):e45417, 2012.

S. Soler, D.E. Debreczeni, E. Vidal, J. Aramburu, C. López, L. Galipienso, and L. Rubio. A new *Capsicum baccatum* accession shows tolerance to wild-type and resistance-breaking isolates of Tomato Spotted Wilt Virus. *Annals of Applied Biology*, 167(3): 343–353, 2015.

I. Sousa. A review on joint modelling of longitudinal measurements and time-to-event. *Revstat*, 9(1):57–81, 2011.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

S. M. Stigler. Laplace's 1774 Memoir on Inverse Probability. *Statistical Science*, pages 359–363, 1986.

J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. New York, New York, Springer-Verlag, 2013.

T. Therneau. A package for survival analysis in S. R package version 2.38. *Retrived from http://CRAN. R-project. org/package= survival*, 2015.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York, Springer-Verlag, 2013.

L. Tierney and J. B. Kadan. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

A. A. Tsiatis, V. Degruttola, and M. S. Wulfsohn. Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

D. Vanlint. *The evolution of bacterial resistance against high hydrostatic pressure*. PhD thesis, University of Reading, 2013.

G. Verbeke and M. Davidian. Joint models for longitudinal data: Introduction and overview. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, editors, *Longitudinal data analysis*, pages 319 – 326. New York, Chapman & Hall/CRC Press, 2009.

M. Viuda-Martos, Y. Ruiz-Navajas, J. Fernández-López, and J. Pérez-Álvarez. Antibacterial activity of lemon (*Citrus lemon* l.), mandarin (*Citrus reticulata* l.), grapefruit (*Citrus paradisi* l.) and orange (*Citrus sinensis* l.) essential oils. *Journal of Food Safety*, 28 (4):567–576, 2008.

X. Wang, Y. Y. Ryan, and J. J. Faraway. *Bayesian Regression Modeling with INLA*. Boca Raton, Chapman & Hall/CRC Press, 2018.

M. C. Wu and R. J. Carroll. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44(1):175–188, 1988.

M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.

J.-S. Yang, H.-J. Nam, M. Seo, S. K. Han, Y. Choi, H. G. Nam, S.-J. Lee, and S. Kim. Oasis: online application for the survival analysis of lifespan assays performed in aging research. *PLoS One*, 6(8):e23525, 2011.

M. Ziehm and J. M. Thornton. Unlocking the potential of survival data for model organisms through a new database and online analysis platform: Survcurv. *Aging Cell*, 12(5):910–916, 2013.