Università degli Studi di Cagliari
Dipartimento di Matematica e Informatica

OBJECTIVE BAYESIAN VARIABLE SELECTION FOR CENSORED DATA

by

Silvia Perra

A thesis submitted for the degree of
PhilosophiæDoctor
PhD School in Mathematics and Computer Science

Supervised by
Stefano Cabras and Maria Eugenia Castellanos

SECS-S/01

2013

A mia Mamma,
donna straordinaria

Some folks hide, and some folks seek, and seeking, when it's mindless, neurotic, desperate, or pusillanimous can be a form of hiding. But there are folks who want to know and aren't afraid to look and won't turn tail should they find it - and if they never do, they'll have a good time anyway because nothing, neither the terrible truth nor the absence of it, is going to cheat them out of one honest breath of Earth's sweet gas.

Tom Robbins - Still Life with Woodpecker

Les questions les plus intéressantes restent des questions. Elles enveloppent un mystére. A chaque réponse, on doit joindre un "peut-être". Il n'y a que les questions sans intérêt qui ont une réponse définitive.

Eric-Emmanuel Schmitt - Oscar et la Dame Rose

Abstract

In this thesis we study the problem of selecting a set of regressors when the response variable follows a parametric model (such as Weibull or log-normal) and observations are right censored. Under a Bayesian approach, the most widely used tools are the Bayes Factors (BFs) which are, however, undefined when using improper priors. Some commonly used tools in literature, which solve the problem of indeterminacy in model selection, are the Intrinsic Bayes factor (IBF) and the Fractional Bayes factor (FBF). The two proposals are not actual Bayes factors but it can be shown that they asymptotically tend to actual BFs calculated over particular priors called intrinsic and fractional priors, respectively. Each of them depends on the size of a minimal training sample (MTS) and, in particular, the IBF also depends on the MTSs used. When working with censored data, it is not immediate to define a suitable MTS because the sample space of response variables must be fully explored when drawing MTSs, but only uncensored data are actually relevant to train the improper prior into a proper posterior. In fact, an unweighted MTS consisting only of uncensored data may produce a serious bias in model selection. In order to overcome this problem, a sequential MTS (SMTS) is used, leading to an increase in the number of possible MTSs as each one has random size. This prevents the use of the IBF for exploring large model spaces. In order to decrease the computational cost, while maintaining a behavior comparable to that of the IBF, we provide a suitable definition of the FBF that gives results similar to the ones of the IBF calculated over the SMTSs. We first define the conditional FBF on a fraction proportional to the MTS size and, then, we show that the marginal FBF (mFBF), obtained by averaging the conditional FBFs with respect to the probability distribution of the fraction, is consistent and provides also good results. Next, we recall the definition of intrinsic prior for the case

of the IBF and the definition of the fractional prior for the FBF and we calculate them in the case of the exponential model for right censored data. In general, when the censoring mechanism is unknown, it is not possible to obtain these priors.

Also another approach to the choice of the MTS, which consists in weighting the MTS by a suitable set of weights, is presented. In fact, we define the Kaplan-Meier minimal training sample (KMMTS) which depends on the Kaplan-Meier estimator of the survival function and which contains only suitable weighted uncensored observations. This new proposal could be useful when the censoring percentage is not very high, and it allows faster computations when the predictive distributions, calculated only over uncensored observations, can be obtained in closed-form.

The new methodologies are validated by means of simulation studies and applications to real data.

Acknowledgements

Nonostante non sia brava a scrivere ringraziamenti e sia naturalmente poco incline a slanci pubblici d'affetto, ci tengo a ringraziare tutte le persone che mi sono state vicine.

Prima di tutto desidero ringraziare i miei supervisori, Stefano e Maria Eugenia, per gli insegnamenti dati, per la pazienza, la costante presenza, l'aiuto e le discussioni che mi hanno aiutata a crescere in questi tre anni.

Un ringraziamento speciale va al Prof. Andrea Loi, per avermi fatto apprezzare il rigore della matematica, per essere stato un ottimo insegnante e modello, sempre pronto ad aiutarmi e chiarire ogni mio dubbio, ma soprattutto per essere un amico. Se ho deciso di intraprendere la strada della ricerca é solo merito (o colpa) tuo!

Vorrei, inoltre, ringraziare la Prof.ssa Monica Musio per avermi fatta innamorare di questa bellissima scienza, la Statistica.

Ringrazio i membri del Dipartimento di Matematica ed Informatica, in particolare il Prof. Stefano Montaldo che mi ha sempre trasmesso entusiasmo e passione, il Prof. Luigi Cerlienco che mi ha insegnato a lottare per difendere le proprie idee ed il Prof. Michele Pinna per avermi aiutata ed incoraggiata durante la fase piú delicata del dottorato.

Grazie a tutti i miei colleghi matematici per i momenti passati insieme e per l'aiuto reciproco, in particolare grazie a Betta, Vale M. e Marco G.

Ed ora arriva il momento di ringraziare voi, i miei adorati Abitanti della Batcaverna. Prima di tutti viene il boss, il mitico Ric! Con i tuoi "smettila!!!" mi hai coccolata durante i miei tre anni di dottorato pieni di perfidia e pettegolezzi. Sei un grandissimo Prof., cuoco ed amico. Grazie per avermi accolta in Caverna!

Cavernicoli, vi elencheró uno ad uno (sperando di non dimenticare nessuno):

Marianna, Daniela, Sara, Jenny, Vale C., Laura, Maura, Francesca, Sabrina, Elena, Samuel, Marconitti, Dadde, Guga, Simone, Ricky (aka Pippo el Cubano), Giammy, Marco (ciaff!), Dagi, Gianluca, Fabry, Giorgio, Franco, Stefano F., Broccia (e robro), Ale Soro ed Alessandro (aka Rossano). Un grazie enorme a tutti voi per avermi adottata, da abusiva, nella vostra splendida famiglia, grazie per i pranzi, le pause e gli immancabili momenti di "relax mentre si dovrebbe lavorare". Vi porteró sempre nel mio cuore!

Ma ci sono cinque Cavernicoli "speciali" che desidero ringraziare:

Ste: quante ne abbiamo passato insieme?! Eppure non ci stanchiamo di gioire, lamentarci, dire fesserie, ridere, abbracciarci (che siano abbracci virtuali o in carne ed ossa poco importa). Sei un amico speciale, buono, sincero e leale, mi stai vicino nei momenti peggiori, mi hai raccolta con un cucchiaino quando ero a terra e mi regali sempre un sorriso. Mi ritengo fortunata ad averti conosciuto e ti ringrazio di cuore per essere sempre te stesso!

Fabio (aka Flabio, Favio, Flappy, Flaps, Fabione, Fabissimo e potrei andare avanti all'infinito): credi che pianga sui giubbotti altrui per niente?? No! Sapevo giá quanto saresti diventato importante per me! La tua sensibilitá, la tua dolcezza, la tua irritabilitá (si, anche quella), la tua presenza ti rendono una perla rara ai miei occhi. Grazie di essere mio amico!

Kočo (vai a trovare il simbolo LaTeX): che dire? Grazie di avermi accolta da subito col tuo modo di fare da giocherellone (limiteró i termini) ma anche per avermi sempre aiutata, coccolata, confortata ed incoraggiata ogni volta che hai visto un'ombra sul mio viso.

Checcosai: grazie grazie e ancora grazie per le nostre chiacchierate, per i consigli e la tua sinceritá! E grazie per avermi fatto conoscere Galt MacDermot!!

Cino: Cicci! Il mio punkabbeshtia preferito che sopporta la mia logorrea. Quanto mi diverto ad ascoltare i tuoi racconti, progettare le grisate, vederti diventare un uomo di fede e disegnare i diti! Mi sei stato vicino pur essendo a piú di 9000 Km di distanza, ti meriti tanti cuori ma soprattutto la mia sincera amicizia e stima (ma si, mi lancio, tanto tra 5 minuti te ne sarai giá dimenticato).

Grazie di cuore ai miei amici: Anna, Matte, Francesco, Cami, Michele, Alberto, Paolo, Mony e Ila, grazie per essere tutti cosí diversi e bellissimi!
Grazie alla mia stella, Sere, la mia amica di sempre, la mia compagna di banco, il mio orgoglio. Ti voglio bene.
Ed ecco arrivato il tuo turno, Mary! Dunque... so giá in partenza che qualunque cosa io scriva non renderá mai giustizia alla tua persona, ma ci proveró comunque! Ci conosciamo da tanto, ma non ci siamo capite da subito. Da quando abbiamo ingranato, peró, é stato un percorso bellissimo. Con te non mi nascondo, so di potermi permettere di essere sempre me stessa, nel bene e nel male, e so che tu fai lo stesso! Le serate con te sono le migliori, cosí come i nostri tentativi in cucina, le nostre missioni e i nostri abbracci (specialmente se sedute su un tappeto e con le lacrime agli occhi). Hai la capacitá di rassicurarmi e tranquillizzarmi semplicemente con la tua voce. Grazie di avere sempre un pensiero ed una coperta per me. Ti voglio bene!
Un grazie specialissimo va alla mia bellissima amica Alessia! La tua dolcezza ed i tuoi "boccirí" sono un toccasana per il mio cuore! So che mi vuoi bene cosí come sono, la tua bambina speciale senza filtri (consapevole del pericolo di beccarmi un meritatissimo pugno in testa) e so che con la tua "povera Milvia" sarai sempre sincera! Per questo ti voglio bene! Ah!!! Grazie a Shugo per la sua presenza ed opera di controfigura!
Grazie alla mia bellissima e dolcissima Giann, per essere il mio sole con un sorriso che abbraccia. A te va la mia immensa gratitudine per avermi fatto smuovere il fondale dell'acquario!
Grazie alle persone che ultimamente mi stanno molto vicine, in particolare la mia tonna speciale, Vale, la cui grinta mi sprona tantissimo. É incredibile quanto ci capiamo a vicenda! E sono certa che presto troverai la felicitá che ti meriti.
Grazie al buon Andrea, artista cinico rock!
Grazie alle mie bellissime MUA, in particolare la divina Dona, Patry, Giorgia, Francesca e Ale! Con voi entro in un mondo fantastico, spero di non doverlo abbandonare mai!
Un grazie particolarmente speciale va a Maurizio, per aver reso quest' ultimo

periodo infinitamente piú bello, per starmi vicino, incoraggiarmi, regalarmi sorrisi, farmi brillare gli occhi e aggiungere colore alle giornate, specialmente quelle passate insieme!
Un grazie immenso alla mia splendida famiglia per essere sempre dalla mia parte e per coccolarmi come nessuno sa fare. Gli sguardi e i baci di Charlie, Chiara e Manu sono la mia forza.
Ma, soprattutto, grazie dal profondo del cuore a mia sorella, Fede! Crescere con te é stato uno dei piú bei regali che la vita mi potesse fare.
Un infinito grazie va alla mia mamma. Per me sei madre, padre, modello e amica, sei la donna piú forte che conosca, che mi ama incondizionatamente e a cui devo tutto e anche di piú! GRAZIE!
Ringrazio tutte le persone che mi hanno sostenuta durante il mio cammino e che qui non ho elencato.
Per concludere, desidero porgere un ringraziamento a tutte le persone che si sono avvicinate e poi si sono allontanate, a tutte loro dedico questo pensiero:

Socchiudo gli occhi e riconosco il mio mondo. Poi dedico una preghiera a tutte le persone che ad un certo punto si sono allontanate da me. Le persone con cui avrei potuto avere un rapporto diverso, e con le quali, invece, per qualche ragione non é andata bene. In questo mondo, a causa delle circostanze in cui li ho incontrati, tra me e loro le cose non hanno funzionato in nessun modo. Ma sento, ne sono certa, che da qualche parte, in un mondo profondo e lontano, su una bellissima riva, ci sorridiamo, ci offriamo gentilezza e trascorriamo insieme momenti felici.

(Banana Yoshimoto - Ricordi di un vicolo cieco)

Preface

The aim of survival analysis is to explain and predict the survival, usually defined along the time domain. In this work we study it by means of regression models (see Klein and Moeschberger (2003), Ibrahim et al. (2001), Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Therneau and Grambsch (2000) and Hosmer and Lemeshow (1999) for a complete discussion).

In statistical data analysis it is common to consider the regression set up in which a given response variable depends on some factors and/or covariates. The model selection problem mainly consists in choosing the covariates which better explain the dependent variable in a precise and hopefully fast manner. This process usually has several steps: the first one is to collect considerations from an expert about the set of covariates, then the statistician derives a prior on model parameters and constructs a tool to solve the model selection problem. We consider the model selection problem in survival analysis when the response variable is the time to event. Different terminal events can be considered, depending on the purposes of the analysis: deaths, failures in mechanical systems, divorces, discharges from hospital and so on. Survival studies include clinical trials, cohort studies (prospective and retrospective), etc.

The main problem in survival data is that terminal events are not fully observable, in this case we say that data are censored. Obviously, censored data are more difficult to handle than complete data and, hence, the statistician must pay attention to the choice of the most appropriate model selection tool tailored from these data.

Example 1. (Larynx dataset) *We present the* larynx *dataset introduced by Kardaun (1983) and described in Klein and Moeschberger (2003), which we study in detail in Chapter 5. The dataset contains the survival times of* $n = 90$ *patients suffering from larynx cancer of which* $n_{cens} = 40$ *are censored. The corresponding variables are:*

- `time`*: survival times (in months)*
- `delta`*: censoring indicator (0=alive, 1=dead)*

The dataset has 2 *predictors, namely:*

- `stage`*: the stage of the disease based on the T.N.M. (primary tumor (T), nodal involvement (N) and distant metastasis (M) grading) classification used by the American Joint Committee for Cancer Staging in 1972. The stages are ordered from least serious to most serious (1=stage 1, 2=stage 2, 3=stage 3, 4=stage 4)*
- `age`*: the age at diagnosis (in years)*

The goal is to choose the optimal set of predictors for survival times from all possible models, in this case $2^2 = 4$ *models, when considering linear models with only additive effects*

- $M_0 : Y_i = \mu + \sigma W_i$
- $M_1 : Y_i = \mu + \gamma_1 \cdot \texttt{stage} + \sigma W_i$
- $M_2 : Y_i = \mu + \gamma_2 \cdot \texttt{age} + \sigma W_i$
- $M_3 : Y_i = \mu + \gamma_1 \cdot \texttt{stage} + \gamma_2 \cdot \texttt{age} + \sigma W_i$

In real applications, it is typical to consider a response variable depending on a large number of covariates and, in many cases, the "true" model can be sparse, i.e. only a small number of covariates is related to the response (e.g. a small number of genes in the genome). In order to solve such a practical problem, we need a tool to select the most suitable model. We also pretend that such tool leads to a fast and accurate model selection procedure. Two are the main Bayesian approaches to variable selection: subjective and objective. Under a subjective point of view, the idea is to calculate a Bayes factor (BF) over a proper informative prior provided by an expert. However, in order to calculate BFs we need to specify a prior distribution $\pi_k(\boldsymbol{\theta}_k)$ separately for each model and this can be complicated, because one often initially entertains K models leading to the impossibility of careful subjective prior elicitation. For this purpose, Bayesian model selection is usually done by means of default methods. When an objective approach is adopted, minimal non-informative priors, and often improper priors (i.e. priors that do not integrate over the parameter space), are used and so one has to reconsider the

concept of BF in order to obtain a good tool for model selection (see Berger et al. (2001) for more details).

The main default Bayesian procedures considered in this thesis are the Intrinsic Bayes factor (IBF), the Fractional Bayes factor (FBF), the Bayesian information criterion (BIC) and a new version of the FBF, called *marginal Fractional Bayes factor* (mFBF). In this work it is illustrated how to adapt the four criteria (with their variations) when censored data are available, using theoretical arguments, simulations and applications to real datasets, as the larynx dataset illustrated above.

In Chapter 1 the different censoring mechanisms and the most common survival regression models are presented. In Chapter 2 the general variable selection problem without censoring is shown, along with the description of the IBF, the FBF and the BIC. Then in Chapter 3 and Chapter 4, which are the heart of the thesis, the objective Bayesian procedures for model selection under censoring are presented. In particular, in Chapter 3 it is introduced the variable selection under censoring using sequential minimal training samples. The calculus of the IBF, FBF, a new tool called mFBF and BIC are provided, jointly with some theoretical results and exemplifications. Also a simulation study is considered. In Chapter 4, the weighted Kaplan-Meier minimal training sample (KMMTS) is introduced, and its behavior is evaluated in a simulation study. In Chapter 5 applications to four real datasets are considered. Finally, in Chapter 6, some final remarks and observations, jointly with future work, are included.

Contents

List of Algorithms

List of Figures

List of Tables

1

Survival Regression Models

1.1 Introduction

In this chapter the main parametric regression models used to describe survival data are introduced. First we present the censoring mechanism (see Klein and Moeschberger (2003) for details) and, after that, a review of the main parametric models is done. The Weibull and log-normal models are presented more in detail in the following subsections. Finally, for completeness, a summary of semiparametric and nonparametric models is included in Subsection 1.3.2 and Subsection 1.3.3.

1.2 Censoring types

When working with survival data, we have to take into account that some data are not completely observable. This could be the case of an observational study in a limited time period. When not all the units or individuals under study fall or experience the terminal event, within the period of study, we say that data are *censored*.
There are several types of censoring, here we discuss the most common ones:

1. Type I censoring
2. Type II censoring
3. random censoring:
   (a) right censoring
   (b) left censoring

(c) interval censoring

4. truncation

We now give some details.

1. **Type I censoring**: this case occurs when an experiment has a certain number of subjects and it stops at a fixed pre-assigned censoring time t_c. Instead of observing the times to event, or lifetimes, $T_1, \ldots, T_n$, we observe $Z_1, \ldots, Z_n$, where

$$Z_i = \begin{cases} T_i & \text{if } T_i \leqslant t_c, \\ t_c & \text{otherwise.} \end{cases}$$

2. **Type II censoring**: when an experiment has a certain number n of subjects and it continues until the failure of a fixed number of subjects is observed.

3. **Random censoring**: it occurs when each individual has a censoring time which is statistically independent of the failure time. This is the most common case of censoring.

   (a) **Right censoring**: when an individual's lifetime is above a certain value but we don't know by how much. We denote by C_i the censoring time and by T_i the survival time. We observe the couples (Z_i, δ_i), where

$$Z_i = min(T_i, C_i)$$
$$\delta_i = \begin{cases} 1 & \text{if } T_i \leqslant C_i, \\ 0 & \text{otherwise.} \end{cases}$$

   If $T_i > C_i$ the individual is a survivor and the event time is censored at C_i. Here δ_i denotes whether the lifetime T_i corresponds to an event ($\delta = 1$) or is censored ($\delta = 0$).

   (b) **Left censoring**: as right censoring, except that

$$Z_i = max(T_i, C_i)$$
$$\delta_i = \begin{cases} 1 & \text{if } T_i \geqslant C_i, \\ 0 & \text{otherwise.} \end{cases}$$

(c) **Interval censoring**: this is the case when a lifetime is on an interval between two fixed values, $[L_i, U_i]$. This is the combination of right censoring and left censoring.

4. **Truncation**: it is due to the structure of the study. In this case only those individuals whose event time is smaller (right truncation) and/or greater (left truncation) than a particular truncation threshold are observed. So if the variable of interest falls outside the range, it is not recorded and no information on this subject is available.

In Figures 1.1, 1.2 and 1.3 three different types of censoring are shown: right censoring, interval censoring and left censoring, respectively, for $n = 4$ lifetimes.

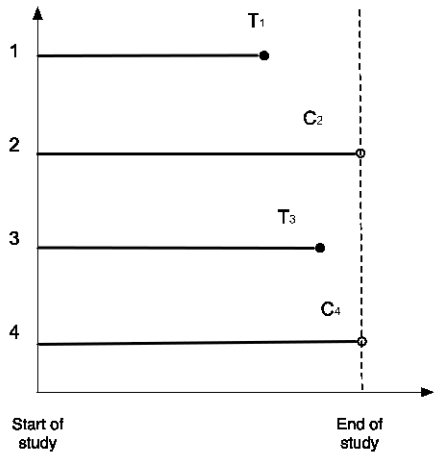

Figure 1.1: Example of right censored data.

In many real datasets, the censoring plan is a mixing of random and Type I censoring, because some patients are randomly censored when, for example, they die or they move from the center of the study, while others are Type I censored when the fixed study period ends. In this thesis we work with Type I censoring or with the combination of random and Type I censoring. Other types of censoring are possible, but they just complicate the exposition and calculus, while, for our purpose, a type of censoring and a corresponding random mechanism must be assumed.

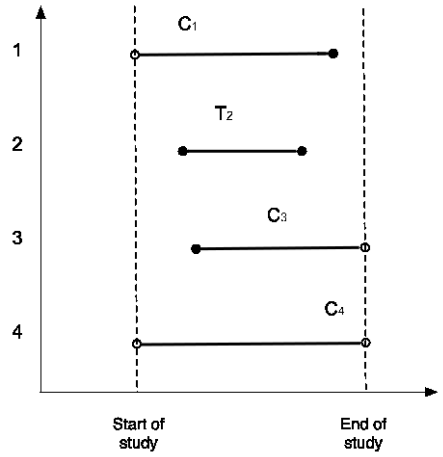

Figure 1.2: Example of interval censored data.

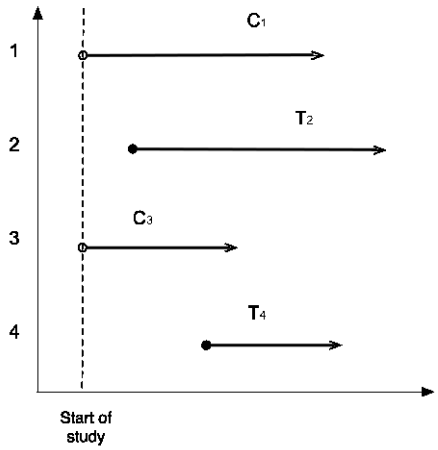

Figure 1.3: Example of left censored data.

1.3 Main survival models

In survival analysis different classes of models are used. It is common to divide them into: parametric, semiparametric and nonparametric models. In this section a review of the main models of each type is presented.

We recall that the distribution function of a random variable T with continuous density $f(t)$ is

$$F(t) = \Pr(T \leqslant t) = \int_0^t f(z)dz,$$

the survival function is defined as the complement of the distribution function

$$S(t) = \Pr(T > t) = 1 - F(t)$$

and the hazard function is given by

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t < T \leqslant t + dt | T > t)}{dt} = \frac{f(t)}{S(t)}.$$

1.3.1 Parametric Models

The most interesting feature of parametric models is that they easily describe the nature of some functions related to the survival distribution, in particular the hazard rate, using parametric functions. Some of the most important parametric models include the exponential, Weibull, gamma, log-normal, log-logistic, normal, Gompertz, inverse Gaussian, Pareto and the generalized gamma distribution.

1.3.1.1 Exponential model

The exponential model is a fundamental parametric model in survival analysis because of its historical significance, calculation simplicity and important properties. Its survival function is

$$S(t) = \exp(-\lambda t), \;\; \lambda > 0, \;\; t > 0.$$

The density function is

$$f(t) = \lambda \exp(-\lambda t)$$

and it is characterized by a constant hazard function

$$h(t) = \lambda.$$

One important characteristic of the exponential distribution is the lack of memory property

$$\Pr(T \geqslant t + z \mid T \geqslant t) = \Pr(T \geqslant z).$$

It follows that the mean residual life, that is the conditional expected life at time t, is constant

$$E(T - t | T > t) = E(T) = \frac{1}{\lambda}.$$

The fact that the exponential distribution has a constant hazard rate leads to a very restrictive assumption in many applications.

1.3.1.2 Weibull model

One of the most widely used parametric models is the Weibull one. Its density function $f(t)$, survival function $S(t)$ and hazard rate $h(t)$, for the time $T \geq 0$ to the terminal event, are

$$f(t) = \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha})$$

$$S(t) = \exp(-\lambda t^{\alpha})$$

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

with $\alpha, \lambda > 0, t \geqslant 0$. The parameters of the distribution, α and λ, are the shape and scale parameters, respectively. Note that the exponential distribution is a special case of the Weibull distribution with $\alpha = 1$.

The distribution is named after Ernst Hjalmar Waloddi Weibull (1887-1979) who published his first paper about this distribution (Weibull (1939)).

The Weibull distribution is commonly used in industrial and biomedical applications, for reliability engineering to describe time to failure in electronic and mechanical systems and for the analysis of time to failure data after the application of stress. It is also widely used for modelling survival data (see Klein and Moeschberger (2003), Hamada et al. (2008), Kalbfleisch and Prentice (2002) and Rausand and Hoyland (2004)).

This distribution is widely used because of its flexibility: it is possible to have increasing ($\alpha > 1$), decreasing ($\alpha < 1$) and constant hazard rates ($\alpha = 1$). In Figure 1.4 three hazard functions for different values of the parameters are shown.

Its flexible form and the model's simple survival, hazard and probability density function have made it a very popular parametric model.

Some authors, like Pike (1966) and Peto and Lee (1973) state that the Weibull model can be used to model the time to appearance of certain phenomena, like the time to appearence of a disease or the time until death. Other authors, like Lee and O'Neill (1971) and Doll (1971), claim that the Weibull model fits data describing time to appearance of tumors in animals and humans.

More details about the Weibull survival model are given in Section 1.4.

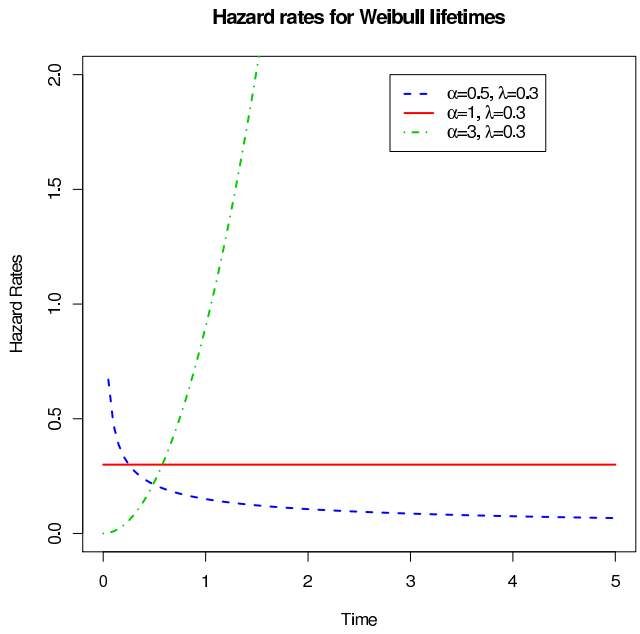

Figure 1.4: Hazard functions for the Weibull model.

1.3.1.3 Log-normal model

The log-normal model is another well known parametric model. The density function, the survival function and the hazard rate of a log-normal variable T are

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

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

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

where $\Phi(t)$ is the distribution function of a standard normal variable, $\mu \in \mathbb{R}$, $\sigma > 0$ and $t > 0$. The hazard rate of the log-normal at 0 is zero, it increases to a maximum and then decreases to 0 as t approaches infinity. In Figure 1.5 three different hazard functions for different values of the parameters are shown. Observe that 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.

Some authors, like Feinleib (1960) and Horner (1987), have used this distribution in the context of survival analysis.

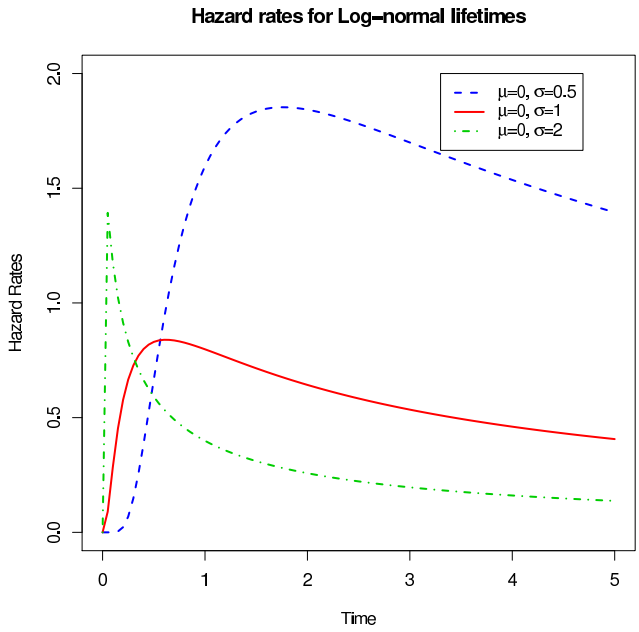

Figure 1.5: Hazard functions for the log-normal model.

More details about the log-normal model can be found in Section 1.5.

1.3.1.4 Log-logistic model

A variable T is said to follow the log-logistic distribution if its logarithm $Y = log(T)$ follows the logistic distribution with density

$$f(y) = \frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(\frac{y-\mu}{\sigma}\right)\right)^2}, \quad -\infty < y < \infty$$

where μ and σ^2 are the location and scale parameters of Y, respectively. The hazard rate for the log-logistic distribution is

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}$$

and the survival function is

$$S(t) = \frac{1}{1+\lambda t^{\alpha}},$$

where $\alpha = 1/\sigma > 0$ and $\lambda = \exp(-\mu/\sigma)$. The numerator of the hazard function is the same as the Weibull hazard but the entire hazard has the following characteristics: monotone decreasing for $\alpha \leqslant 1$, 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. This distribution has simple expressions for the hazard and survival functions, as well as the Weibull and exponential models. Note that its hazard rate is similar to the log-normal one, except in the extreme tail of the distribution (see Bennett (1983) and Gupta et al. (1999)). For this reason, it presents the same problems of the log-normal model in practical applications. In Figure 1.6 three different hazard functions for different values of the parameters are shown.

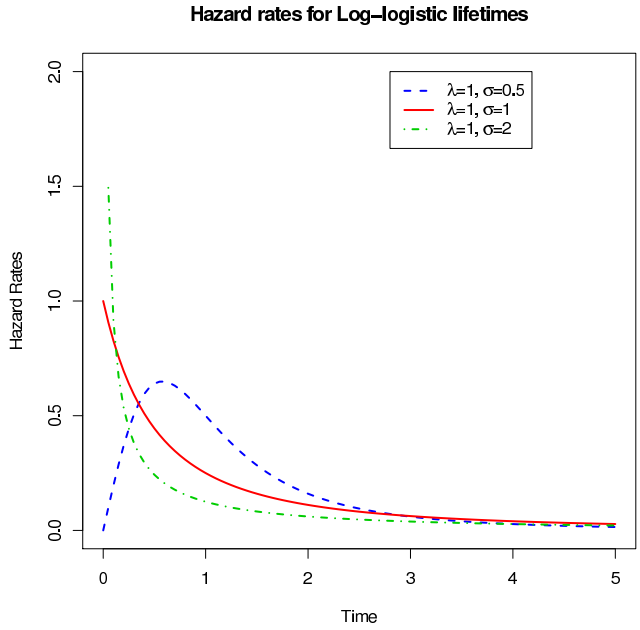

Figure 1.6: Hazard functions for the log-logistic model.

1.3.1.5 Gamma model

The gamma distribution has similar properties to the Weibull one except its mathematical tractability. Its density function is:

$$f(t) = \frac{\lambda^{\beta}}{\Gamma(\beta)} t^{\beta-1} \exp(-\lambda t)$$

where $\lambda > 0$ is the scale parameter, $\beta > 0$ is the shape parameter, $t > 0$ and $\Gamma(\cdot)$ is the gamma function. This distribution, like the Weibull one, includes the exponential as a

special case (for $\beta = 1$) and it approaches a normal distribution as β tends to infinity. The hazard function for the gamma distribution is monotone increasing for $\beta > 1$, with $h(0) = 0$ and $h(t) \to \lambda$ as $t \to \infty$, and monotone decreasing for $\beta < 1$, with $h(0) \to \infty$ and $h(t) \to \lambda$ as $t \to \infty$. When $\beta > 1$ the mode is at $t = (\beta - 1)/\lambda$. Its survival function is

$$S(t) = \frac{\int_z^\infty \lambda(\lambda z)^{\beta-1} \exp(-\lambda z) dz}{\Gamma(\beta)}$$

and its hazard function is

$$h(t) = \lambda(\lambda t)^{n-1} \left((n-1)! \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \right)^{-1}.$$

A more useful distribution is the *generalized gamma*, which has the following form:

$$f(t) = \frac{\alpha \lambda^\beta}{\Gamma(\beta)} t^{\alpha\beta-1} \exp(-\lambda t^\alpha).$$

Other distributions can be obtained as special cases: Weibull (if $\beta = 1$), exponential (if $\alpha = \beta = 1$) and log-normal (if $\beta \to \infty$). For this reason, the generalized gamma distribution is often used to choose the most adequate parametric model for survival data.

Figure 1.7 shows three different hazard functions for the gamma model and for different values of the parameters.

1.3.2 Semiparametric models

For completeness of exposition of the subject, in this subsection and in the following one we will recall some classes of semiparametric and nonparametric models that will not be used further in the thesis.

Semiparametric methods have been studied in the context of Bayesian survival analysis. The different approaches can be distiguished by the stochastic process used as prior distribution for the nonparametric part of the model. One of the most popular semiparametric models is the Cox proportional hazards model (Cox (1972)). Let $S(t \mid \boldsymbol{x})$ be the survival function of the time T given a vector of covariates $\boldsymbol{x}$, and let $h(t \mid \boldsymbol{x})$ be the corresponding hazard function

$$h(t \mid \boldsymbol{x}) = h_0(t) \exp(\boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x})$$

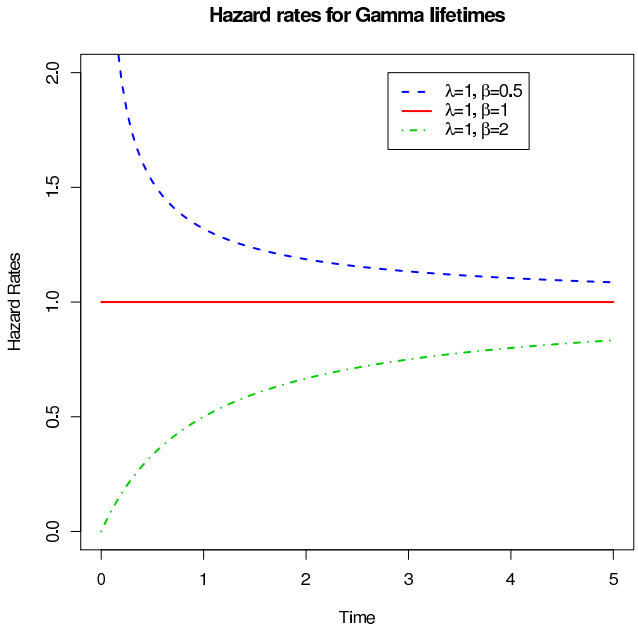

Figure 1.7: Hazard functions for the gamma model.

where $h_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\gamma}$ is the vector of regression coefficients. In this model, $h_0(\cdot)$ is the nonparametric part and the function containing the regression coefficients is the parametric part. Usually $\boldsymbol{\gamma}$ is supposed to be constant over time, but when $\boldsymbol{\gamma}$ is function of t there is a time-varying covariate effect and when $\boldsymbol{x}$ is a function of t there is a time-dependent covariate effect.
Suppose to have a partition of the time axis

$$0 < a_1 < a_2 < \ldots < a_J.$$

So there are J intervals $I_1 = (0, a_1]$, $I_2 = (a_1, a_2]$, ..., $I_J = (a_{J-1}, a_J]$. Two are the possible cases:

- if the survival function is absolutely continuous, then an ordinary Cox model is assumed

  $$S(t \mid \boldsymbol{x}) = \exp\left(-H_0(t)\exp(\boldsymbol{\gamma}^\top \boldsymbol{x})\right)$$

  where $H_0(t) = \int_0^t h(u)du$ is the cumulative baseline hazard function and $\boldsymbol{\gamma}$ and $\boldsymbol{x}$ are constant over time

- if the survival function is not absolutely continuous, then the discretized version of the Cox model is used

$$S(\alpha_j \mid \boldsymbol{\gamma}, \boldsymbol{x}) = \prod_{k=1}^{j} (1-\alpha_k)^{\exp(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x})} \quad j = 1, \ldots, J,$$

where $\alpha_k = \Pr(a_{k-1} \leqslant T < a_k \mid T \geqslant a_{k-1})$ is the discretized baseline hazard rate for the interval I_k.

For this model different nonparametric priors have been considered, leading to different models. In the following we summarize some of the most commonly used ones.

The piecewise constant hazard model. In the *j-th* interval suppose to have a constant baseline hazard $h_0(t) = \lambda_j$ for $t \in I_j = (a_{j-1}, a_j]$ and let $D = (n, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{\delta})$ be the observed data, where $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\mathsf{T}$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^\mathsf{T}$ with $\delta_i = 1$ if the *i-th* subject uncensored and 0 otherwise, and $\boldsymbol{X}$ is the $n \times r$ matrix of covariates with *i-th* row $\boldsymbol{x}_i{}^\mathsf{T}$. Letting $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)^\mathsf{T}$, then the likelihood function of $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ for the n subjects can be written as

$$L(\boldsymbol{\gamma}, \boldsymbol{\lambda} \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left(\lambda_j \exp(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{X})\right)^{\nu_{ij}\delta_i} \times \\ \times \exp\left[-\nu_{ij}\left(\lambda_j(t_i - a_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(a_g - a_{g-1})\right) \exp(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{X})\right]$$

where $\nu_{ij} = 1$ if the *i-th* subject uncensored or was censored in the *j-th* interval, and 0 otherwise.

This model is also known as *piecewise exponential model.* A common prior for the baseline hazard λ is the independent gamma prior $\lambda_j \sim Ga(\alpha_{0j}, \lambda_{0j})$, $j = 1, 2, \ldots, J$, where α_{0j} and λ_{0j} are prior parameters which regulate the prior mean and variance of λ_j.

Another nonparametric prior process used for the Cox model is the *gamma process.* Let $Ga(\alpha, \lambda)$ be the gamma distribution where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter, $\alpha(t)$ for $t \geqslant 0$, an increasing left-continuous function, with $\alpha(0) = 0$ and $Z(t)$ a stochastic process where

(i) $Z(0) = 0$;

(ii) $Z(t)$ has independent increments in disjoint intervals;

(iii) $Z(t) - Z(s) \sim Ga(c(\alpha(t) - \alpha(s)), c)$, for $t > s$.

Then $\{Z(t) : t \geqslant 0\}$ is called a *gamma process*

$$Z(t) \sim \mathcal{GP}(c\alpha(t), c)$$

where $\alpha(t)$ is the mean of the process and c is a weight or confidence parameter about the mean.

The gamma process can be used as a prior on the cumulative or baseline hazard.

Here we recall its use when modelling the cumulative hazard that is most common. For more details about the specification of a gamma process on the baseline hazard rate see Chapter 3 of Ibrahim et al. (2001).

The gamma process on cumulative hazard. The probability distribution of survival of n subjects given $\boldsymbol{X}$ under the Cox model is

$$\Pr(T > t \mid \boldsymbol{\gamma}, \boldsymbol{X}, H_0) = \exp\left(-\sum_{j=1}^{n} \exp(\boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_j) H_0(t_j)\right).$$

The gamma process is often used as a prior for the cumulative baseline hazard function $H_0(t)$:

$$H_0 \sim \mathcal{GP}(c_0 H^*, c_0).$$

$H^*(t)$ can be chosen to be Weibull distributed, for example, where $H^*(t)$ is an increasing nonparametric function with $H^*(0) = 0$ and $\boldsymbol{\beta}_0$ is the vector of hyperparameters. Then we have $H^*(t) = \eta_0 t^{k_0}$, where $\boldsymbol{\beta}_0 = (\eta_0, k_0)^\mathsf{T}$.

Then, the marginal survival function is

$$\Pr(T > t \mid \boldsymbol{\gamma}, \boldsymbol{X}, \boldsymbol{\beta}_0, c_0) = \prod_{j=1}^{n} [\Phi(iV_j)]^{c_0(H^*(t_{(j)}) - H^*(t_{(j-1)}))} \tag{1.1}$$

where $V_j = \sum_{l \in R_j} \exp(\boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_l)$ and R_j is the risk set at time $t_{(j)}$. The corresponding likelihood can be obtained by differentiating (1.1). For more details about the use of the gamma process and the Cox model, see Ibrahim et al. (2001), Kalbfleisch (1978) and Clayton (1991). Other priors used jointly with the semiparametric Cox model are the Beta process (Section 3.5 of Ibrahim et al. (2001)) and the Dirichlet process (Section 3.7 of Ibrahim et al. (2001)).

1.3.3 Nonparametric models

The Bayesian nonparametric approach to survival analysis consists in finding a specific functional form for the survival distribution conditional on the sample and in providing suitable priors for the corresponding space of random functions. The first works are mostly based on the Dirichlet process, introduced by Ferguson (1973), which is a class of random probability measures. Given a partition $B = \{B_1, \ldots, B_k\}$ of the sample space Ω, then a stochastic process P on (Ω, B) is said to be a Dirichlet process if the vector $(P(B_1), \ldots, P(B_k))$ follows a Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$, for all the partitions of Ω. In Susarla and Van Ryzin (1976), for example, the Dirichlet process is used to make point estimation of the survival curve.
Let T be a continuous random variable in $(0, \infty)$, then $F(t) = P((-\infty, t])$ and the process P are said *neutral to the right* if the normalized increments

$$F(t_1), [F(t_2) - F(t_1)]/[1 - F(t_1)], \ldots, [F(t_{k+1}) - F(t_k)]/[1 - F(t_k)]$$

are independent for all $t_1 < t_2 < \ldots < t_{k+1}$. In Doksum (1974) the independent increment processes (or Levy processes) are used to construct the neutral to the right processes and it is shown that the posterior distribution of a random probability neutral to the right is also neutral to the right. In Doksum (1974) it is also observed that, in this kind of models, the survival function is discrete with probability 1. In Ferguson and Phadia (1979) Bayesian nonparametric survival models are studied in the case of uncensored and censored data and then, the Dirichlet process, the simple homogeneous process and the gamma process are considered. Other works on nonparametric models are Hjort (1990), Doss (1994), based on Dirichlet process mixtures of Antoniak (1974), Muliere and Walker (1997), based on Polya tree priors of Ferguson (1974) and Lavine (1992), and Walker and Damien (1998), based on the beta-Stacy process priors of Walker and Muliere (1997). In Kim (1999) independent increment processes are taken as prior distributions for the cumulative intensity function of multiplicative counting processes. Dykstra and Laud (1981) describe a method to solve the problem of the discreteness of the survival function by modelling the hazard rate function by means of an independent increment process, obtaining continuous survival and cumulative hazard functions. A drawback of this technique is that the hazard rate function must be monotone. So in Arjas and Gasbarra (1994) it is suggested to use a Markov jump process with a

martingale structure. In order to overcome the problem of the monotony and the difficulty of this last structure, in Nieto-Barajas and Walker (2004) it has been proposed a piecewise continuous Markov process to model the hazard rate function and, then, the survival and cumulative hazard functions are modeled by means of a continuous process. In Kottas (2006) a computational method to calculate the posterior distribution of different functionals of a Weibull Dirichlet process mixture is presented. The idea is to model the survival function with a flexible Dirichlet process mixture having a Weibull kernel. This eliminates the problem of making full posterior inference in survival analysis for the different functionals of interest. For more details see Ibrahim et al. (2001), De Blasi (2006) and Kottas (2006).

1.4 Weibull model

We now provide more details about the Weibull model, introduced in Subsection 1.3.1.2, expecially in the context of regression as this will be a reference model for the rest of the thesis.

It is sometimes useful to work with the logarithm of lifetimes in order to convert positive values to observations on the entire real line.

Suppose $\boldsymbol{T} = (T_1, \ldots, T_n)$ denotes lifetimes or censored times. We consider

$$Y_i = \log(T_i), \quad i = 1, \ldots, n$$

where $T_i \sim Weibull(\alpha, \lambda_i)$. Then Y_i has the density function

$$f_Y(y_i \mid \alpha, \lambda_i) = \alpha \exp\left[\alpha\left(y_i - \left(-\frac{\log(\lambda_i)}{\alpha}\right)\right) - e^{\alpha\left(y_i - \left(-\frac{\log(\lambda_i)}{\alpha}\right)\right)}\right] \tag{1.2}$$

where $-\infty < y_i < +\infty$.

In survival analysis one of the most interesting problems is to ascertain the relationship between the failure time, T, and one or more covariates in order, for example, to determine the prognosis of a patient with various characteristics. Consider m covariates associated with a vector of times $\boldsymbol{T}$, which may include quantitative, categorical and/or time dependent variables. We choose an approach similar to the classical linear regression, assuming a linear model for $\boldsymbol{Y}$

$$Y_i = \mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i + \sigma W_i$$

where W_i follows a standard Gumbel distribution (which is obtained as the distribution of the logarithm of a Weibull variable), with the following density function

$$f_W(w) = \exp(w - \exp(w)) \tag{1.3}$$

for $-\infty < w < +\infty$. $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$ denotes the fixed design matrix with the observed covariates in the n subjects, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ are the values of the m covariates in the $i-th$ subject. We mainly operate under the following parametrization $\alpha = 1/\sigma$, $\lambda_i = \exp\left\{-(\mu + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}_i)/\sigma\right\}$ where $\boldsymbol{\gamma}^\mathsf{T} = (\gamma_1, \ldots, \gamma_m)$ is a vector of regression coefficients.

The Gumbel distribution[1] is used to model the distribution of the maximum (or minimum) of a number of samples of various distributions belonging to the exponential family. In fact it is useful in predicting the chance that an extreme event will occur. The potential applicability of this distribution to represent the distribution of maxima (or minima) relates to extreme values of the normal or exponential type (see Gumbel (1958)).

The Weibull model is also called the *accelerated failure-time model*. Let $S_0(t)$ denote the survival function when $\boldsymbol{x}$ is 0, that is, $S_0(t)$ is the survival of $\exp(\mu + \sigma W)$, then for another subject with regressor values $\boldsymbol{x}$ we have

$$\begin{aligned}
\Pr(T > t) &= \Pr(Y > \log(t)) = \Pr(\mu + \sigma W > \log(t) - \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}) \\
&= \Pr(\exp(\mu + \sigma W) > t\exp(-\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x})) \\
&= S_0(t\exp(-\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x})).
\end{aligned}$$

Observe that the effect of the covariates in the original time scale is to change the time scale by a factor $\exp(-\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x})$. Depending on the sign of $-\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}$ the time can be incremented or decremented by a constant factor.

This model is also a *multiplicative hazard rates model*. The hazard rate of an individual with a covariate vector $\boldsymbol{x}$ for this class of models is related to a baseline hazard rate h_0 and a non-negative function of the covariates by

$$\begin{aligned}
h(t \mid \boldsymbol{x}) &= \alpha\lambda t^{\alpha-1}\exp\left(-\frac{\boldsymbol{\gamma}^\mathsf{T}}{\sigma}\boldsymbol{x}\right) \\
&= h_0(t)\exp\left(-\frac{\boldsymbol{\gamma}^\mathsf{T}}{\sigma}\boldsymbol{x}\right),
\end{aligned}$$

[1] Emile Julius Gumbel, 1891-1966

where $h_0(t) = \alpha\lambda t^{\alpha-1}$.

When all the covariates are fixed at time zero, the hazard rates of two individuals with distinct values of $\boldsymbol{x}$ are proportional. To see this, consider two individuals with covariate values $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

$$\frac{h(t \mid \boldsymbol{x}_1)}{h(t \mid \boldsymbol{x}_2)} = \frac{h_0(t)\exp(-\frac{\boldsymbol{\gamma}^\mathsf{T}}{\sigma}\boldsymbol{x}_1)}{h_0(t)\exp(-\frac{\boldsymbol{\gamma}^\mathsf{T}}{\sigma}\boldsymbol{x}_2)}$$

which is constant over time.

Observe that the Weibull is the only parametric model which has the property of being both an accelerated failure-time model and a multiplicative hazards model.

In the context of survival analysis, a common feature of datasets is that they contain *censored* or *truncated* observations; this leads to a certain structure in the likelihood. In the following, we introduce the necessary inferential tools which allow us to work with incomplete data.

1.4.1 Inference

In this section we present the likelihood function, our choice for the prior distribution and the approximation of the corresponding posterior distribution for the parameters. Model selection is performed under such likelihood and prior.

1.4.1.1 Likelihood function

The likelihood for a vector of observations $\boldsymbol{y}$ has the following form

$$\begin{aligned} L(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X}) &= \prod_{i=1}^{n} f_Y\left(y_i\right)^{\delta_i} \left[S_Y\left(y_i\right)\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma} f_W\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}_i)}{\sigma}\right)\right]^{\delta_i} \left[S_W\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}_i)}{\sigma}\right)\right]^{(1-\delta_i)} \end{aligned} \tag{1.4}$$

where f_Y is given in (1.2), S_Y is the corresponding survival function, f_W is given in (1.3), S_W is the associated survival function and $\delta_i = 0$ if observation i is censored and 1 otherwise.

1.4.1.2 Prior distribution

In order to make inference about parameters in our model from a Bayesian perspective, it is necessary to specify a prior distribution for the parameters. In model selection problems, it is quite difficult to elicitate a prior on the parameters of each model, especially when the number of models is large. For the case of location-scale models, as the Weibull model, the usual default prior is the Jeffrey's one (see Yang and Berger (1998))

$$\pi(\mu, \boldsymbol{\gamma}, \sigma) \propto \frac{1}{\sigma} \text{ for } \mu \in \mathbb{R},\ \boldsymbol{\gamma} \in \mathbb{R}^{dim(\boldsymbol{\gamma})},\ \sigma \in \mathbb{R}^{+}.$$

This prior has been proposed in Evans and Nigm (1980) and also used in Albert (2009) and leads to a proper posterior distribution when calculated over a sample containing a number of uncensored observations equal to the number of parameters in the model (in this case $dim(\boldsymbol{\gamma}) + 2$).

1.4.1.3 Posterior distribution

The corresponding unnormalized kernel of the posterior distribution can be written as

$$\begin{aligned}
&\pi(\mu, \boldsymbol{\gamma}, \sigma \mid \boldsymbol{y}, \boldsymbol{X}) \propto \pi(\mu, \boldsymbol{\gamma}, \sigma) L(\mu, \boldsymbol{\gamma}, \sigma \mid \boldsymbol{y}, \boldsymbol{X}) \\
&= \frac{1}{\sigma} \prod_{i=1}^{n} \left(\frac{1}{\sigma} f_W \left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i)}{\sigma} \right) \right)^{\delta_i} \left(S_W \left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i)}{\sigma} \right) \right)^{(1-\delta_i)} \\
&= \frac{1}{\sigma} \prod_{i=1}^{n} \left[\frac{1}{\sigma} \exp\left(\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i)}{\sigma} \right) - \exp\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i)}{\sigma} \right) \right) \right]^{\delta_i} \times \\
&\quad \times \left[\exp\left(- \exp\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i)}{\sigma} \right) \right) \right]^{(1-\delta_i)}.
\end{aligned}$$

Approximation of the posterior distribution

The posterior distribution has not a closed-form and it has been approximated by using Markov Chain Monte Carlo simulation methods (*MCMC*), in particular a random walk Metropolis-Hastings (MH) for $\boldsymbol{\theta} = (\mu, \boldsymbol{\gamma}, \log(\sigma))$ with a multivariate normal distribution as proposal (see Chib and Jeliazkov (2001)). Algorithm 1 contains the pseudocode of the method proposed in Albert (2009). Firstly a Laplace approximation is run with the maximum likelihood estimator of the regression model (in particular, we have used the function `survreg` of the library `survival` in R). The Random Walk MH algorithm is

used with a proposal having the Laplace approximation's variance and a multiplicative factor of the Metropolis scale factor.

Algorithm 1 Random Walk Metropolis to approximate the posterior distribution.

Require: N, number of RW-MH MCMC steps;
Data $D(\boldsymbol{y}, \boldsymbol{X})$;
$\pi(\boldsymbol{\theta}|D)$ posterior kernel of the regression model M;
$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(1)}$ initial value of the parameters vector of length s;
$\hat{\Sigma}$ the variance-covariance matrix of the proposal distribution:
τ fixed scale factor;

1: Calculate the posterior distribution at $\boldsymbol{\theta}^*$, $\pi = \pi(\boldsymbol{\theta}^*|D)$;
2: **for** i=2 to N **do**
3: Generate $\boldsymbol{v} = (v_1, \ldots, v_s)^\mathsf{T}$, where $v_i \sim N(0,1)$, for $i = 1, \ldots, s$, and calculate the posterior probability at $\boldsymbol{\zeta} = \boldsymbol{\theta}^* + \tau\hat{\Sigma}\boldsymbol{v}$, $\pi^* = \pi(\boldsymbol{\zeta}^*|D)$;
4: Generate $u \sim \mathcal{U}(0,1)$;
5: **if** $u < \pi^*/\pi$ **then**
6: $\pi = \pi^*$;
7: $\boldsymbol{\theta}^* = \boldsymbol{\zeta}$
8: **end if**
9: $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$
10: **end for**
11: **return** $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$

1.5 Log-normal model

We now provide more details about the log-normal model, expecially in the context of regression, as this will be another reference model for the rest of the thesis. Suppose that the time to the event is log-normal distributed, then $Y_i = \log(T_i)$ follows a normal distribution. In the context of regression analysis it is possible to express Y_i as

$$Y_i = \log(T_i) = \mu + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}_i + \sigma W_i$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$ denotes the observed covariates in the n subjects, with $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im})$ the covariates for each subject i, and $W_i \sim N(0,1)$.

1.5.1 Inference

In this section the likelihood function of the censored log-normal model, a choice for the prior distribution and the corresponding posterior distribution are shown.

1.5.1.1 Likelihood function

For right censored data the likelihood has the form

$$\begin{aligned} L(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X}) &= \prod_{i=1}^{n} f_Y\left(y_i\right)^{\delta_i} \left[S_Y\left(y_i\right)\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{x}_i)}{\sigma}\right)^2\right)\right]^{\delta_i} \times \\ &\quad \times \left[1 - \Phi\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{x}_i)}{\sigma}\right)\right]^{(1-\delta_i)}. \end{aligned}$$

1.5.1.2 Prior distribution

In order to avoid elicitation of a proper prior for each possible model, default methods are considered. For this model the Reference, Jeffreys and location-scale priors agree and are given by

$$\pi(\mu, \boldsymbol{\gamma}, \sigma) \propto \frac{1}{\sigma}, \qquad \mu \in \mathbb{R}, \boldsymbol{\gamma} \in \mathbb{R}^{dim(\boldsymbol{\gamma})}, \sigma \in \mathbb{R}^{+}.$$

As in the case of the Weibull model, in order to obtain a proper posterior, it is necessary to calculate it over a sample containing a number of uncensored observations equal to the number of parameters in the model (in this case too the number is $dim(\boldsymbol{\gamma}) + 2$).

1.5.1.3 Posterior distribution

The corresponding posterior distribution takes the form

$$\pi(\mu, \boldsymbol{\gamma}, \sigma \mid \boldsymbol{y}, \boldsymbol{X}) \propto \pi(\mu, \boldsymbol{\gamma}, \sigma) L(\mu, \boldsymbol{\gamma}, \sigma \mid \boldsymbol{y}, \boldsymbol{X})$$

In Chapter 4 we present a technique to calculate Bayes factors which involves the expressions of the marginal distributions when we have uncensored samples and also when censored data are present in the samples. In order to calculate the marginal distributions, here we introduce the calculations of the posterior distribution in two cases:

(i) all the data are uncensored observations;

(ii) part of the data are censored and the remaining part is uncensored.

In case (i) we rewrite the model as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ are uncensored observations, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are normally distributed, $\epsilon_i \sim N(0, \sigma)$, $\boldsymbol{\beta} = (\mu, \boldsymbol{\gamma})$ and $\boldsymbol{Z} = (\boldsymbol{1}, \boldsymbol{X})$ is the covariate matrix with first column of ones and with rank r. Then we obtain

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left((\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right)\right], \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}$.
Observe that $\hat{\boldsymbol{\beta}}$ is a sufficient statistic for $\boldsymbol{\beta}$ when σ^2 is known. So

$$\hat{\boldsymbol{\beta}} \mid \sigma^2 \sim N_r(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}).$$

The posterior distribution is

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = \pi(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}, \sigma^2)\pi(\sigma^2 \mid (\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})),$$

where

$$\pi(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}, \sigma^2) \sim N_r(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1})$$

and

$$\pi(\sigma^2 \mid (\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})) \sim Inv\chi^2_{n-r}$$

with scale factor $(\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})/n - r$.
The marginal posterior distribution for $\boldsymbol{\beta}$ is:

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{\Gamma(\frac{n}{2}) \mid \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} \mid^{1/2} s_{var}^{-r}}{\Gamma(\frac{1}{2})^r \Gamma(\frac{n-r}{2})\sqrt{n-r}^r}\left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(n-r)s_{var}^2}\right]^{-\frac{n}{2}},$$

where $s_{var}^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})/(n-r)$ is the sample variance.

In case (ii) the kernel of the posterior distribution can be written as

$$\pi(\mu, \boldsymbol{\gamma}, \sigma \mid \boldsymbol{y}, \boldsymbol{X}) \propto \frac{1}{\sigma} \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \left(\frac{y_i - (\mu + \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_i)}{\sigma} \right)^2 \right) \right]^{\delta_i} \times$$
$$\times \left[1 - \Phi\left(\frac{y_i - (\mu + \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_i)}{\sigma} \right) \right]^{(1-\delta_i)}$$

which doesn't have a closed-form. Again, we need to use a MCMC simulation in order to approximate the distribution and we run the Algorithm 1.

1.5.1.4 Marginal predictive distribution

In case (i) the marginal distribution has the form

$$m(\boldsymbol{y}) = \frac{\Gamma(\frac{1}{2})^r \Gamma(\frac{n-r}{2})}{\Gamma(\frac{n}{2}) \mid \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} \mid^{1/2} \left((\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T} (\boldsymbol{y} - \hat{\boldsymbol{y}}) \right)^{(n-r)/2}}. \tag{1.5}$$

More details on the calculation of the marginal predictive distribution can be found in Chapter 8 of Ghosh et al. (2006) and in Section 2 of Berger and Pericchi (1997).

In case (ii) we need to approximate the marginal distribution with MCMC methods.

2

Variable Selection

2.1 Bayesian Formulation

In Statistics it is often required to summarise and represent a random phenomena by means of a statistical model. For this reason, one important issue is to choose the best model that may represent the behavior of the quantity of interest. Suppose we want to describe the survival of a group of patients suffering from a particular disease. It is, then, fundamental to discover which factors are correlated with the patient's survival and how survival could be predicted.

We represent the data, $\boldsymbol{y}$, by a statistical model, which usually depends on some unknown parameters, $\boldsymbol{\theta}$, and which specify a particular probability distribution for $\boldsymbol{y}$, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$.

The *model space*, i.e. the family of all possible models, is denoted by $\mathcal{M}$

$$\mathcal{M} = \{M_0, \ldots, M_K\}.$$

In our case each model has the form

$$M_k : y_i = \boldsymbol{\beta}_k{}^{\mathsf{T}} \boldsymbol{x}_{k,i} + \sigma_k \epsilon_i,$$

where $\boldsymbol{X}$ is a fixed design matrix with r columns, that is r covariates including the intercept ($r = m + 1$, where m is the number of independent quantitative covariates), $k \in \{0, 2, \ldots, K = 2^r - 1\}$ indicates the model index with the corresponding design matrix

$\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,n})^{\mathsf{T}} \in \mathbb{R}^{n \times r_k}$ and model vector parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k) \in \boldsymbol{\Theta}_k =$

$\mathbb{R}^{r_k} \times \mathbb{R}^+$ and ϵ_i is the error term.

The aim of *model selection* is to select a model, among those in the model space, $\mathcal{M}$, which better describes the phenomenon under study. A particular case of model selection which we study in this thesis is the *variable selection problem*.

Variable selection is a common problem in regression. Its goal is to explain a response variable, Y, using a set of covariates $\{X_1, \ldots, X_m\}$ related to Y. Our aim is to find out which variables, from the given set, are relevant to explain Y. In this model selection problem each entertained model M_i corresponds to a particular subset of covariates.

The Bayesian approach to model selection or hypothesis testing was developed by Jeffreys, whose solution was based on posterior odds probabilities or, equivalently, on BFs (see Kass and Raftery (1995)). This approach to model selection also arises formally in decision theory frameworks, given a certain loss function. In this work we show some model selection procedures based on Jeffreys' proposal, which are also based on the Neyman-Pearson-Wald Lemma (see DeGroot (1975) and Pereira et al. (2008)) and which involve posterior probabilities and BFs.

2.2 Bayes Factors and Posterior Model Probabilities

In this section we propose some model selection techniques based on hypothesis testing. Suppose we are comparing K models, in our case $K = 2^r - 1$

$$\boldsymbol{y} \mid \boldsymbol{\theta}_i \sim f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$$

considering only additive effects of r covariates $(X_1, \ldots, X_r)$, including the intercept, then our hypotheses are

$$H_i : \boldsymbol{\theta} = \boldsymbol{\theta}_i \in \Theta_i \ \text{ (the true model is } M_i)$$
$$H_j : \boldsymbol{\theta} = \boldsymbol{\theta}_j \in \Theta_j \ \text{ (the true model is } M_j) \ \ i \neq j = 0, \ldots, K.$$

Assuming $\pi_i(\boldsymbol{\theta}_i)$, $i = 1, \ldots, K$, prior distributions for the unknown parameters, the marginal or predictive density of $\boldsymbol{y}$ is

$$m_i(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}_i} f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i.$$

Definition 1. (Bayes factor) *Given two models, M_i and M_j, the Bayes factor (BF) in favor of M_i and against M_j is given by*

$$B_{ij} = \frac{m_i(\boldsymbol{y})}{m_j(\boldsymbol{y})} \tag{2.1}$$

where $m_i(\boldsymbol{y})$ and $m_j(\boldsymbol{y})$ are the marginal distributions of the models M_i and M_j, respectively.

The BF can also be defined as the quantity which updates the prior odds producing the posterior odds, that is

$$\frac{\Pr(M_j \mid \boldsymbol{y})}{\Pr(M_i \mid \boldsymbol{y})} = \frac{\Pr(M_j)}{\Pr(M_i)} B_{ji} \tag{2.2}$$

The posterior probability of model M_j in function of BFs is

$$\Pr(M_j \mid \boldsymbol{y}) = \frac{\Pr(M_j) m_j(\boldsymbol{y})}{\sum_{k=0}^{K} \Pr(M_k) m_k(\boldsymbol{y})} = \left\{ 1 + \sum_{k \neq j} B_{kj} \frac{\Pr(M_k)}{\Pr(M_j)} \right\}^{-1}. \tag{2.3}$$

In Table 2.1 we show the BF interpretation of Jeffreys (1961), who considers both the BF value and its logarithm, $\log_{10}$, called the *weight of evidence*.

$\log_{10}(B_{ij})$	B_{ij}	Evidence against M_j
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Table 2.1: Bayes factors interpretation.

Under the perspective of decision theory, the action space is $\mathcal{M}$ and we indicate an action with "a". Our model selection problem is a *finite action problem*, because we have to choose the best model among a finite number of proposed models, $K + 1$. Let $\{a_0, \ldots, a_K\}$ be all the available actions, with

$$a_i = \text{choice of the model } M_i$$

and $l(\boldsymbol{\theta}, a_i)$ the corresponding losses, $i = 0, \ldots, K$. The Bayes action is that one which minimizes the posterior expected loss

$$E^{\boldsymbol{\theta}|\boldsymbol{y}}[l(\boldsymbol{\theta}, a_i)].$$

In our case the actions of interest are a_i, $i = 0, \ldots, K$. If we choose the "$0 - k_i$" loss function

$$l(\boldsymbol{\theta}, a_i) = \begin{cases} 0, & \text{if } \boldsymbol{\theta} \in \Theta_i \\ k_i, & \text{if } \boldsymbol{\theta} \in \Theta_i^c \end{cases}$$

we obtain

$$E^{\boldsymbol{\theta}|\boldsymbol{y}}[l(\boldsymbol{\theta}, a_i)] = \int l(\boldsymbol{\theta}, a_i)\pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta} =$$
$$= \int_{\Theta_i^c} k_i \pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta} = k_i(1 - \Pr(\Theta_i \mid \boldsymbol{y})) = k_i(1 - \Pr(M_i \mid \boldsymbol{y})).$$

The Bayes decision is that one corresponding to the smallest posterior expected loss. We can also define the Bayes action in terms of posterior probabilities and BFs

$$\frac{E^{\boldsymbol{\theta}|\boldsymbol{y}}[l(\boldsymbol{\theta}, a_i)]}{E^{\boldsymbol{\theta}|\boldsymbol{y}}[l(\boldsymbol{\theta}, a_j)]} = \frac{k_i}{k_j} \frac{(1 - \Pr(M_i \mid \boldsymbol{y}))}{(1 - \Pr(M_j \mid \boldsymbol{y}))}.$$

Hence, if all the k_j are equal, the Bayes decision corresponds to choose the model with the largest posterior probability. There are several advantages in choosing a Bayesian approach to model selection. The first one is that BFs are easy to be comunicated due to their interpretation as odds. Another one is that Bayesian model selection is *consistent*: if one of the entertained models is actually the true model, or the most similar model to the true one, then Bayesian model selection will guarantee selection of such model if enough data is observed, while other selection tools such as p-values and AIC or likelihood ratios based methods may not guarantee consistency. Even when the true model is not among those being considered, results in Berk (1966) and Dmochowski (1994) show that, asymptotically and under mild conditions, Bayesian model selection will choose the model, between all the considered, that is closest to the true one in terms of Kulback-Leibler divergence.

Bayesian model selection procedures are automatic Ockham's razors (Berger et al. (2001)), favoring simpler models over more complex ones. We observe that this approach does not require nested models or regular asymptotics and can account for model uncertainty, while selecting a model on the basis of data, and then using the same data to estimate model parameters or make predictions based upon the model, often yields overoptimistic estimates of accuracy in the choice of the right model. In the classical approach it is recommended to use part of the data to select a model and the remaining part of the data for estimation and prediction but, when limited data is

available, this can be difficult. A flaw of this approach is that it ignores the fact that the selected model might be wrong, so that predictions based on assuming the model as true could be excessively optimistic. Moreover, under a strict prediction approach, all models could be left in the analysis with prediction being done using a weighted average of the predictive distributions from each model, and the weights determined from posterior probabilities or BFs. This is known as *Bayesian model averaging* and it is widely used today as the basic methodology of accounting for model uncertainty and particularly suited for prediction of Y. See Draper (1995), Raftery et al. (1997), and Clyde (1999) for details. In this thesis we mainly focus on Bayesian model selection, rather than prediction and, hence, Bayesian model averaging.

2.3 Objective Variable Selection

Before calculating the BFs, it is necessary to choose the prior distribution, $\pi_i(\boldsymbol{\theta}_i)$, $i = 0, \ldots, K$, for the model parameters. Under a Bayesian approach, there are two possible choices: the subjective, or informative, approach when an expert elicits a prior distribution $\pi_i(\boldsymbol{\theta}_i)$ based on some prior considerations and the objective, or non-informative, approach when expert prior informations are not available or are not convenient to be used. In this latter case, priors are derived from formal rules.

The subjective Bayesian variable selection has a long history, having been considered by Atkinson (1978), Smith and Spiegelhalter (1980), Pericchi (1984), Poirier (1985), Box and Meyer (1986), George and McCulloch (1993), George and McCulloch (1995), George and McCulloch (1997), Clyde et al. (1996), Geweke (1996), Smith and Kohn (1996), among others. In linear regression the proposed prior distributions on the regression coefficients and the error variance within each model are typically either conjugate priors or closely related distributions. For example, for the regression coefficients multivariate normal distributions (typically centered at 0) and inverse gammas for the error variances are usually considered as the posterior has a closed-form expression. The covariance matrices and the hyperparameters in the inverse gamma are often fixed with the help of some subjective criteria or by empirical Bayesian methods.

The first attempts at solving the problem in a form as "objective as possible" can be found in Mitchell and Beauchamp (1988) and Spiegelhalter and Smith (1982). The objective Bayesians argue that a subjective Bayesian analysis is frequently not a realistic

possibility, especially in model selection problems when the number of models is quite large, because it is difficult to elicit a prior for each subset of parameters under each model. The thesis is focused on the objective Bayesian methods for assessing prior distributions in variable selection problems. A more complete discussion about objective Bayesian techniques can be found in Berger et al. (2001) and Berger (2006).
As stated in Berger et al. (2001), the choice of a suitable prior distribution is a delicate issue, due to the following main problems:

- *Sensitivity of Bayes factors:* the influence of prior distributions on the BFs remains even asymptotically (see Kass (1993) and Kass and Raftery (1995)).

- *Computational difficulties:* BFs can be very difficult to obtain when the parameter spaces are high dimensional and the total number of models under consideration is large (see Carlin and Chib (1995), Kass and Raftery (1995), Verdinelli and Wasserman (1995) and Raftery et al. (1997)).

- *Indeterminacy of Bayes factors:* when we use improper non-informative priors and when models have different parameter spaces of different dimensions, the BFs are undefined. Let $\pi_i^N(\boldsymbol{\theta}_i)$ and $\pi_j^N(\boldsymbol{\theta}_j)$ be two improper priors for two competing models M_i and M_j, respectively. We can use $c_i\pi_i^N(\boldsymbol{\theta}_i)$ and $c_j\pi_j^N(\boldsymbol{\theta}_j)$ as non-informative priors, because the priors are improper and the BF becomes $(c_j/c_i)B_{ji}$. Notice that the choice of c_j/c_i is arbitrary, so the BF is undetermined. Choosing $c_i = c_j$ is accepted when $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ have similar parameter spaces, in the sense that the dimensions are equal (see Berger et al. (1998)).

- *Use of "vague proper priors" does not solve the difficulties arising with improper priors:* as shown in Berger et al. (2001), the resulting BF could depend on the choice of the prior and we can conclude that using a vague proper prior is never better than using an improper prior. This is also called the Bartlett's paradox, which states that in model selection a posterior distribution is acceptable even if the prior is broad and quite non-informative, but when the variance of the prior tends to large values, the BF tends to prefer the null model, regardless of the given information (see Bartlett (1957), Jeffreys (1961) and Liang et al. (2008) for more details).

When there are more than two models or the models are non-nested, some of the BFs we are going to introduce could have the undesirable feature of violating one of the coherence conditions of BFs

$$B_{jk} = B_{ji}B_{ik},$$

where i, j and k are indexes of three models.

In order to avoid this problem, we adopt the *encompassing approach*, where each submodel, M_i, is compared to the encompassing or reference model, M_R. In this way, it is possible to obtain the pairwise BFs, B_{Ri}, $i = 0, \ldots, K$. The BFs of M_i to M_j is then defined as

$$B_{ij} = \frac{B_{Rj}}{B_{Ri}}.$$

There are different possible choices for the reference model introduced by Moreno and Girón (2008) and Casella et al. (2009): the first proposal is to use the most complex model (the one containing all the covariates) and, then, to do pairwise comparison between that model and the others. This approach is called pairwise comparison *from above*. The second approach consists in using the simplest model M_0, the null one, and this is called pairwise comparison *from below*. In Moreno and Girón (2008) the two methods, in the case of linear regression and when using intrinsic priors, are compared. The two procedures can lead to different orderings in the space of the models and it is not clear which one is to be preferred. The variable selection from above is based on multiple pairwise comparison, which consists in comparing two models at a time. So model posterior probabilities are calculated and compared. This approach could not be coherent, due to the fact that the model posterior probabilities come from different probability spaces. The ordering of the models produced by the model comparison from below is equivalent to ordering the models according to the model posterior probabilities computed in the space of all models. Both methods work similarly, although the encompassing from below has more appealing theoretical properties. The most important property is that the from below procedure provides model posterior probabilities in the space of all models, $\mathcal{M}$, being the set of posterior probabilities coherent. Another important fact, highlighted by Moreno and Girón (2008), is that when the number of covariates is bigger than the sample size n, the posterior probability of any model having a number of covariates bigger than n is less than $1/2$ and this penalizes the complex models. Furthermore, in the from below encompassing procedure the number

of regressors does not need to be specified from the beginning and it produces coherent posterior probabilities in the set of all the models under study, while in the from above encompassing procedure the full model needs to be fixed in advance. In Moreno and Girón (2008) it is observed that it seems preferable the from above procedure when dealing with complex models, while the from below procedure seems to perform better for smaller models. Finally, there is no conclusive evidence for deciding which one of the two criteria is best. For more details and tests see Moreno and Girón (2008).
We choose to work with the null model M_0 as reference model doing pairwise comparison from below between models

$$M_0 : \boldsymbol{Y} = \mu_0 + \sigma_0 \boldsymbol{W} \tag{2.4}$$

$$M_k : \boldsymbol{Y} = \mu_k + {\boldsymbol{\gamma}_k}^{\mathsf{T}} \tilde{\boldsymbol{X}}_k + \sigma_k \boldsymbol{W}, \tag{2.5}$$

where $\tilde{\boldsymbol{X}}_k$ is the design matrix for model k.
As M_0 is nested in M_k, parameters μ_k and σ_k can be considered as common to both models, so the new parameters will be $\boldsymbol{\gamma}_k$, $k = 0, \ldots, K$. Without loss of generality we can write the prior as

$$\pi_k(\boldsymbol{\gamma}_k, \mu_0, \sigma_0) = \pi_k(\boldsymbol{\gamma}_k \mid \mu_0, \sigma_0)\pi_k(\mu_0, \sigma_0).$$

Other choices for the reference model are proposed and discussed in Perez (2000), Casella and Moreno (2006) and Liang et al. (2008).

2.3.1 Conventional priors

We review some of the main choices for the prior distributions. Such functions are viewed as objective priors with a wide consensus in the objective Bayesian community. Often in Bayesian analysis, one can use non-informative or default priors. Common choices are the uniform prior $\pi_k^U(\boldsymbol{\theta}_k) = 1$, the Jeffreys' prior $\pi_k^J(\boldsymbol{\theta}_k) = (det(I_k(\boldsymbol{\theta}_k)))^{1/2}$ (where $I_k(\boldsymbol{\theta}_k)$ is the expected Fisher information matrix corresponding to the model M_k) and the Reference prior $\pi_k^R(\boldsymbol{\theta}_k)$ whose definitions can be found in Bernardo (1979), Berger et al. (1992) and Berger et al. (2009). The use of conventional proper priors for model selection and hypothesis testing has been introduced in Jeffreys (1961). The idea is to assign a proper prior distribution for the new parameters conditional on the old parameters, $\pi_k(\boldsymbol{\gamma}_k \mid \mu_0, \sigma_0)$ and a non-informative, usually improper, prior for the

old parameters, $\pi_k(\mu_0, \sigma_0)$. As prior distributions for common parameters, Jeffreys (1961) and Zellner and Siow (1980) use the Reference or independent Jeffreys' prior $\pi_k(\mu_0, \sigma_0) = 1/\sigma_0$ under each model M_k, $k = 0, \ldots, K$. One of the most popular conventional priors is the *g-prior* introduced by Zellner (1986)

$$\pi(\boldsymbol{\gamma}_k|\mu_0, \sigma_0) = N(0, g\sigma_0(\tilde{\boldsymbol{X}}_k{}^{\mathsf{T}}\tilde{\boldsymbol{X}}_k)^{-1}),$$

that ensures closed-form expressions for the BFs when working with linear models. The main issue is the calibration of g. The original idea of Zellner (1986) was to place a prior over g and, then, to integrate over g. Several other proposals are present in literature: George and Foster (2000) propose to choose g by means of model selection criteria, as AIC and BIC, George and Foster (2000) and Clyde and George (2000) use empirical Bayes (EB) methods to make a global estimation of g, which is however a criticized approach because of its non formal Bayesian calculation. Hansen and Yu (2000) propose to make a local estimation of g. An interesting approach, presented in Liang et al. (2008), consists in considering a mixture of g-priors which simplifies the calculation of the corresponding marginal distributions. In particular, they define the *hyper-g* prior family, which is a family of priors for g based on the Gaussian hypergeometric function. In Zellner and Siow (1980) the following prior is proposed

$$\pi_k^{ZS}(\boldsymbol{\gamma}_k|\mu_0, \sigma_0) = Ca_{r_k}(\boldsymbol{\gamma}_k|0, n\sigma_0^2(\boldsymbol{V}_k{}^{\mathsf{T}}\boldsymbol{V}_k)^{-1}), \tag{2.6}$$

which is a multivariate Cauchy distribution, where $\tilde{\boldsymbol{X}}_k$ is the design matrix corresponding to the vector $\boldsymbol{\gamma}_k$ of length k, $r_k = rank(\tilde{\boldsymbol{X}}_k)$ and $\boldsymbol{V}_k = (I_n - P_0)\tilde{\boldsymbol{X}}_k$ is the design matrix corresponding to the orthogonal parametrization, where $P_0 = \boldsymbol{X}_0(\boldsymbol{X}_0{}^{\mathsf{T}}\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0{}^{\mathsf{T}}$ and $\boldsymbol{X}_0 = (1, \ldots, 1)^{\mathsf{T}}$ of length n. The Zellner-Siow prior can be viewed as a special case of mixtures of g-priors, where the prior over g is the $InvGa(1/2, n/2)$.

An adaptation of Berger's robust priors is proposed in Bayarri et al. (2012), which follows the spirit of conventional priors, and which also can be expressed as a scale mixture of normals.

In this work we choose as starting prior the non-informative prior, obtaining the so called *default Bayes factors*, such as the *Fractional Bayes Factor* (FBF) introduced by O'Hagan (1995) and the *Intrinsic Bayes Factor* (IBF) developed by Berger and Pericchi (1996). Notice that these are not actual Bayes factors, but the IBFs and FBFs can be shown to correspond asymptotically to BFs arising from proper priors called *intrinsic*

prior and *fractional prior*, which are the actual default priors used in the model selection procedure.

2.3.2 IBF

Suppose that non-informative (usually improper) priors $\pi_k^N(\boldsymbol{\theta}_k)$, $k = 0, \ldots, K$, are available for the $K+1$ models: $M_0, \ldots, M_K$. The corresponding marginal or predictive densities of $\boldsymbol{Y}$ are

$$m_k^N(\boldsymbol{y}) = \int f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)\pi_k^N(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k.$$

In order to define the IBF, we now introduce the notion of *proper minimal training sample* (MTS) which is a particular subset of the entire data $\boldsymbol{y}$. We consider a variety of training samples and we index them by l.

Definition 2. (Minimal Training Sample) *A training sample $\boldsymbol{y}(l)$ is a subset of the set $\boldsymbol{y}$ of all the observations. It is called* proper *if $0 < m_k^N(\boldsymbol{y}(l)) < +\infty$ for all M_k, $k = 0, \ldots, K$, and is called* minimal *if it is proper and no subset is proper.*

The minimal dimension of a training sample for a model with s parameters is s. When we compare two models, M_i and M_j, we choose the dimension of the MTS as the number of parameters of the most complex model.
The role of the training sample is to convert the improper prior $\pi_k^N(\boldsymbol{\theta}_k)$ into a proper posterior, that is

$$\pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{y}(l)) = \frac{f_k(\boldsymbol{y}(l) \mid \boldsymbol{\theta}_k)\pi_k^N(\boldsymbol{\theta}_k)}{m_k^N(\boldsymbol{y}(l))}$$

and then use the latter to define the BFs for the remaining data $\boldsymbol{y}(-l)$.

$$BF_{ij}(l) = \frac{m_i(\boldsymbol{y}(-l) \mid \boldsymbol{y}(l))}{m_j(\boldsymbol{y}(-l) \mid \boldsymbol{y}(l))}. \tag{2.7}$$

The following Proposition 1 can be found in Berger and Pericchi (1996).

Proposition 1. *Given two models M_i and M_j and assuming that the posterior distributions $\pi_i(\boldsymbol{\theta}_i|\boldsymbol{y}(l))$ and $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{y}(l))$ are proper, the expression of the BF of M_i to M_j is*

$$BF_{ij}(l) = B_{ij}^N(\boldsymbol{y})B_{ji}^N(\boldsymbol{y}(l)), \tag{2.8}$$

where

$$B_{ji}^N(\boldsymbol{y}(l)) = \frac{m_j^N(\boldsymbol{y}(l))}{m_i^N(\boldsymbol{y}(l))}.$$

Proof. Result follows from (2.7)

$$\begin{aligned} BF_{ij}(l) &= \frac{m_i(\boldsymbol{y}(-l) \mid \boldsymbol{y}(l))}{m_j(\boldsymbol{y}(-l) \mid \boldsymbol{y}(l))} \\ &= \frac{\int f_i(\boldsymbol{y}(-l) \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i \mid \boldsymbol{y}(l))d\boldsymbol{\theta}_i}{\int f_j(\boldsymbol{y}(-l) \mid \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j \mid \boldsymbol{y}(l))d\boldsymbol{\theta}_j} \\ &= \frac{\int f_i(\boldsymbol{y}(-l) \mid \boldsymbol{\theta}_i)f_i(\boldsymbol{y}(l) \mid \boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)/m_i^N(\boldsymbol{y}(l))d\boldsymbol{\theta}_i}{\int f_j(\boldsymbol{y}(-l) \mid \boldsymbol{\theta}_j)f_j(\boldsymbol{y}(l) \mid \boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)/m_j^N(\boldsymbol{y}(l))d\boldsymbol{\theta}_j} \\ &= B_{ij}^N(\boldsymbol{y})B_{ji}^N(\boldsymbol{y}(l)). \end{aligned}$$

□

Clearly in (2.8) the arbitrariness in the choice of constants that multiply π_i^N and π_j^N is removed. Observe that the BF conditional on $\boldsymbol{y}(l)$ depends on the specific training sample and this would lead to an indeterminacy. There are several techniques to avoid this dependence and to increase stability. One idea, firstly proposed by Berger and Pericchi (1996), is to consider $B_{ij}^N(\boldsymbol{y}(l))$ over all possible minimal training samples $\boldsymbol{y}(l)$, $l = 1, \ldots, L$, and we choose two approaches to express this:

- *arithmetic intrinsic Bayes factor* (AIBF). The IBF is calculated over the arithmetic mean of the $B_{ji}^N(\boldsymbol{y}(l))$

$$BF_{ij}^{AI} = B_{ij}^N(\boldsymbol{y})\frac{1}{L}\sum_{l=1}^{L} B_{ji}^N(\boldsymbol{y}(l)) \tag{2.9}$$

- *median intrinsic Bayes factor* (MIBF). The IBF is calculated over the median of the $B_{ji}^N(\boldsymbol{y}(l))$

$$BF_{ij}^{MI} = B_{ij}^N(\boldsymbol{y}) \operatorname*{Median}_{l=1,\ldots,L} B_{ji}^N(\boldsymbol{y}(l)) \tag{2.10}$$

Example 2. (Exponential vs. Weibull) *Suppose we want to compare the exponential model with the Weibull one*

$$\begin{aligned} M_0 &: f_0(y \mid \lambda) = \lambda \exp(-\lambda y) \\ M_1 &: f_1(y \mid \alpha, \beta) = \alpha\beta y^{\alpha-1} \exp(-\beta y^\alpha). \end{aligned}$$

M_1 is the most complex model, having two parameters, so a MTS is a set containing two observations $\{y_i, y_j\}$, $y_i \neq y_j \in \{y_1, \ldots, y_n\}$.
The Jeffreys prior for the first model M_0 is

$$\pi(\lambda) \propto \frac{1}{\lambda},$$

and the corresponding marginal distribution for the entire sample takes the form

$$m_0^N(\boldsymbol{y}) = \frac{\Gamma(n)}{(\sum_{i=1}^n y_i)^n},$$

then the marginal distribution calculated over the MTS is

$$m_0^N(\boldsymbol{y}(l)) = \frac{1}{(y_i + y_j)^2}.$$

For model M_1 the Jeffreys prior, according to Yang and Berger (1998), is

$$\pi_1(\alpha, \beta) \propto \frac{1}{\alpha\beta}$$

while the marginal distribution $m_1(\boldsymbol{y})$ cannot be obtained in a closed-form (see Berger and Pericchi (1996) for details), the marginal distribution over a MTS, as shown in Berger and Pericchi (1996), is

$$m_1^N(\boldsymbol{y}(l)) = \frac{1}{2y_iy_j\left|\log(\frac{y_i}{y_j})\right|}.$$

The AIBF has the following form

$$BF_{10}^{AI} = \frac{\int \frac{1}{\alpha\beta}\alpha^n\beta^n\prod_{i=1}^n y_i^{\alpha-1}\exp(-\beta\sum_{i=1}^n y_i^\alpha)d\alpha d\beta}{\Gamma(n)/(\sum_{i=1}^n y_i)^n}\frac{1}{L}\sum_{i<j}\frac{2y_iy_j\left|\log(\frac{y_i}{y_j})\right|}{(y_i+y_j)^2} \tag{2.11}$$

and the MIBF is

$$BF_{10}^{MI} = \frac{\int \frac{1}{\alpha\beta}\alpha^n\beta^n\prod_{i=1}^n y_i^{\alpha-1}\exp(-\beta\sum_{i=1}^n y_i^\alpha)d\alpha d\beta}{\Gamma(n)/(\sum_{i=1}^n y_i)^n}\underset{i<j}{Median}\frac{2y_iy_j\left|\log(\frac{y_i}{y_j})\right|}{(y_i+y_j)^2}, \tag{2.12}$$

where $L = \#\{i < j\} = n(n-1)/2$.

An important point noted in Berger and Pericchi (1998) is that for the AIBF it is typically necessary to place the more complex model in the numerator, i.e., to let M_j be the more complex model and then define $B_{ij}^{AI} = 1/B_{ji}^{AI}$, because in general the AIBF does not satisfy the reciprocity condition. In fact, O'Hagan (1997) observes: "*Not only would the arithmetic IBF then violate a natural coherence condition that ordinary Bayes factors satisfy automatically, but we would be in the embarrassing position of having two BFs for comparing M_i with M_j instead of just one*".
These IBFs along with alternate versions, like the *expected* IBF of Berger and Pericchi (1996) are useful in certain scenarios, such as when nested models are compared and when the sample size is small. In this last case, the two correction factors in (2.9) and (2.10) may have large variances and this would lead to unstable IBFs. In Berger et al. (2001) it is observed that the MIBF is often to be preferred and widely applicable due to its robustness with respect to outliers and it is considered the simplest default model selection tool altough it is not optimal.

2.3.3 FBF

O'Hagan (1995) introduces the FBF in order "*to avoid the arbitrariness of choosing a particular training sample, or having to consider all possible subsets of a given size*". The basic idea is very similar to the one behind the IBF but, instead of using a part of the data to turn non-informative priors into proper priors, it uses a fraction b of the likelihood $L_k(\boldsymbol{\theta}_k) = f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)$. The remaining $1-b$ part of the likelihood function is used for model discrimination.
Let b be a suitable constant, the FBF is defined as

$$\begin{aligned} BF_{ij}^{F,b} =& B_{ij}^N(\boldsymbol{y}) \frac{\int L_j^b(\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int L_i^b(\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} \\ =& B_{ij}^N(\boldsymbol{y}) \frac{\int f_j(\boldsymbol{y} \mid \boldsymbol{\theta}_j)^b\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)^b\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} \\ =& B_{ij}^N(\boldsymbol{y}) \frac{m_b(\boldsymbol{y} \mid M_j)}{m_b(\boldsymbol{y} \mid M_i)} \\ =& B_{ij}^N(\boldsymbol{y}) \frac{m_{j,b}(\boldsymbol{y})}{m_{i,b}(\boldsymbol{y})}. \end{aligned} \tag{2.13}$$

We denote by $B_{ij}^b(\boldsymbol{y})$ the correction factor of the FBF.
An important issue is how to choose the fraction b. In O'Hagan (1995) it is observed that the FBF is strictly bounded, because b is varied from s/n, where s is the MTS size, to 1. Furthermore, in the case of nested models, b should tend to 0 as $n \to \infty$, to achieve consistent model choice. In O'Hagan (1995) it is stated that the FBF is consistent for b of order $1/n$. This criterion is satisfied by the minimal value $b = s/n$. Among the different possibilities, O'Hagan (1995) proposes three ways to set b:

(i) $b = \frac{s}{n}$, when robustness with respect to the prior distribution or to the models is not a concern;

(ii) $b = \frac{1}{n} max\{s, \sqrt{n}\}$, when robustness is a serious concern;

(iii) $b = \frac{1}{n} max\{s, \log(n)\}$, as an intermediate option

and modifies the (ii) and (iii) choices by taking:

(iibis) $b = \frac{\sqrt{s}}{n}$;

(iiibis) $b = s\log(n)$

so that $b = 1$ when $s = n$.

Example 3. (Example 2 continued) *For the calculation of the FBF we choose* $b = s/n$, *where* s, *the MTS size, is equal to* 2.
So the FBF takes the form

$$B_{10}^{F,b} = \frac{\int \frac{1}{\alpha\beta}\alpha^n\beta^n \prod_{i=1}^n y_i^{\alpha-1} \exp(-\beta\sum_{i=1}^n y_i^\alpha)d\alpha d\beta}{\Gamma(n)/(\sum_{i=1}^n y_i)^n} \frac{\int \frac{1}{\lambda}(f_0(\boldsymbol{y} \mid \lambda))^{2/n} d\lambda}{\int \frac{1}{\alpha\beta}(f_1(\boldsymbol{y} \mid \alpha\beta))^{2/n} d\alpha d\beta}$$

$$= \frac{\int \frac{1}{\alpha\beta}\alpha^n\beta^n \prod_{i=1}^n y_i^{\alpha-1} \exp(-\beta\sum_{i=1}^n y_i^\alpha)d\alpha d\beta}{\Gamma(n)/(\sum_{i=1}^n y_i)^n} \frac{\frac{n^2}{4(\sum_{i=1}^n y_i)^2}}{\int \alpha\beta(\prod_{i=1}^n y_i^{\alpha-1})^{\frac{2}{n}} \exp(-\frac{2\beta}{n}\sum_{i=1}^n y_i^\alpha)d\alpha d\beta}$$

which doesn't have a closed-form.

2.3.4 BIC

We have already observed that the BF requires the specification of proper priors that may be seen as subjective or ad hoc. The Schwarz criterion Schwarz (1978), based on the first order asymptotic Laplace approximation of the marginal densities m_i and m_j, provides a very simple approximation to the BF when comparing model j and model i

$$BIC_{ji}^S = -2\left(l_j(\hat{\boldsymbol{\theta}}_j) - l_i(\hat{\boldsymbol{\theta}}_i)\right) + (k_j - k_i)\log(n) \tag{2.14}$$

where k_i and k_j are the dimensions of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, respectively, and $l_j(\hat{\boldsymbol{\theta}}_j)$ and $l_i(\hat{\boldsymbol{\theta}}_i)$ are the logarithms of the likelihood calculated over the MLEs, $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\theta}}_i$, under the two models M_j and M_i, respectively. Notice that n represents the effective sample size that must be determined carefully (see Volinsky and Raftery (2000)). BIC_{ji}^S is a function of the likelihood ratio test statistic and priors do not appear in its formula as the posterior is asymptotically likelihood dominated. Observe that the smaller is the BIC, the more we can state that the true model is M_j. This criterion is quite well established in the model selection literature (see Smith and Spiegelhalter (1980) for details) and it is asymptotically consistent (that is, it tends to an actual BF). Note that the Schwarz criterion is an approximately Bayesian testing procedure and it is easy to compute. So we can say that the Schwarz criterion is a useful automatic Bayesian testing procedure for nested models. However it requires standard regularity conditions for asymptotic expansions and there are some restrictions. In Kass and Wasserman (1995) and Berger and Pericchi (1997) it is observed that the BIC is inconsistent when applied to models with irregular asymptotics and in cases in which the likelihood can be concentrated at the boundary of the parameter space for one of the models.

Example 4. (Example 2 continued) *In this case the BIC takes the form*

$$BIC_{10}^S = -2\left(l_1(\hat{\alpha}, \hat{\beta}) - l_0(\hat{\lambda})\right) + \log(n),$$

where $(\hat{\alpha}, \hat{\beta})$ *are the maximum likelihood estimators for the Weibull likelihood function* $l_1(\alpha, \beta \mid \boldsymbol{y})$ *and* $\hat{\lambda} = n/\bar{y}$ *is the maximum likelihood estimator for the exponential likelihood function* $l_0(\lambda \mid \boldsymbol{y})$.

2.4 Intrinsic and fractional prior

Default BFs, as the IBF and the FBF, are not actual BFs, for this reason it is necessary to study their behavior. One way to do that is analyzing if they asymptotically correspond to BFs obtained from reasonable default prior distributions.

2.4.1 Intrinsic prior

In Berger and Pericchi (1996) the *intrinsic prior* is defined as a prior distribution that would produce the same default BF with a large amount of data. As Berger and Pericchi (1996) point out, the intrinsic prior exists when the correction factor of the IBF converges to a positive number as the sample size goes to infinity. The special case of intrinsic priors considered in this thesis is the one in which there are two models, M_i and M_j, with M_j nested in M_i. Under the following conditions and using the notation given in Bertolino et al. (2000):

- $f_j(\boldsymbol{y} \mid \boldsymbol{\theta}_j)$ is nested in $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$
- $\pi_i^N(\boldsymbol{\theta}_i)$ is an improper prior and $\pi_j(\boldsymbol{\theta}_j)$ is a proper prior
- the likelihood $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$, for a given sample size n, is integrable with respect to the prior $\pi_i^N(\boldsymbol{\theta}_i)$

the intrinsic priors corresponding to BF_{ij}^{AI} defined in (3.3) exist and are given by

$$\pi_j^I(\boldsymbol{\theta}_j) = \pi_j(\boldsymbol{\theta}_j), \quad \pi_i^I(\boldsymbol{\theta}_i) = \pi_i^N(\boldsymbol{\theta}_i) E_{\boldsymbol{\theta}_i}^{M_i}(B_{ji}^N(\boldsymbol{y}(l))) \tag{2.15}$$

where $E_{\boldsymbol{\theta}_i}^{M_i}(B_{ji}^N(\boldsymbol{y}(l)))$ is the expectation of the correction factor with respect to the density of $\boldsymbol{y}(l)$ under the model M_i.
More details on the calculation of the intrinsic prior can be found in the sequel of this thesis, in Berger and Pericchi (1996) and in Appendix 1 of Berger et al. (2001).

2.4.2 Fractional prior

In De Santis and Spezzaferri (1997) it is shown that, following the approach of Berger and Pericchi (1996), the FBF asymptotically corresponds to a real BF calculated over suitable *fractional priors*. Let $b = s/n$ be the generic fraction of the likelihood function for a fixed MTS, and denoting by $B^b_{ji}(\boldsymbol{y}) = \frac{\int f_j(\boldsymbol{y}|\boldsymbol{\theta}_j)^b \pi^N_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)^b \pi^N_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}$ the correction factor of (2.21), the FBF is

$$B^{F,b}_{ij} = B^N_{ij}(\boldsymbol{y}) B^b_{ji}(\boldsymbol{y}).$$

It can be observed that $B^{F,b}_{ij}$ corresponds asymptotically to an actual BF if

$$\frac{\pi_i(\hat{\boldsymbol{\theta}}_i)\pi^N_j(\hat{\theta}_j)}{\pi^N_i(\hat{\boldsymbol{\theta}}_i)\pi_j(\hat{\theta}_j)} + o_p(1) = \frac{1}{B^b_{ij}(\boldsymbol{y})},$$

where $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ are the maximum likelihood estimators for the two models M_i and M_j, respectively, and for some priors $\pi_i(\cdot)$ and $\pi_j(\cdot)$.
The equation which defines the fractional prior for the FBF is

$$\frac{\pi^{FI}_i(\boldsymbol{\theta}_i)\pi^N_j(\psi_j(\boldsymbol{\theta}_i))}{\pi^N_i(\boldsymbol{\theta}_i)\pi^{FI}_j(\psi_j(\boldsymbol{\theta}_i))} = B^*_i(\boldsymbol{\theta}_i), \tag{2.16}$$

where $B^*_i(\boldsymbol{\theta}_i)$ is the limit, as n goes to infinity, of $\frac{1}{B^b_{ij}(\boldsymbol{y})}$ and $\psi_j(\boldsymbol{\theta}_i)$ is the limit of $\hat{\boldsymbol{\theta}}_j$ under model M_i.
Then, the two fractional priors are

$$\begin{aligned} \pi^{FI}_j(\boldsymbol{\theta}_j) &= \pi^N_j(\boldsymbol{\theta}_j) u(\boldsymbol{\theta}_j) \\ \pi^{FI}_i(\boldsymbol{\theta}_i) &= \pi^N_i(\boldsymbol{\theta}_i) u(\psi_j(\boldsymbol{\theta}_i)) B^*_i(\boldsymbol{\theta}_i), \end{aligned} \tag{2.17}$$

where $u(\cdot)$ is a continuous non negative function.
As observed by De Santis and Spezzaferri (1997), under some general conditions (see Theorem 2.1 in De Santis and Spezzaferri (1997)) $B^*_i(\boldsymbol{\theta}_i)$ can be obtained from $B^N_{ij}(\boldsymbol{y})$ by replacing n with s and the maximum likelihood estimators $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ with their limits under the model M_i.

2.5 Approximation methods for predictive distributions

As marginal likelihoods are the key ingredients in all versions of the BFs, in this section we show a method proposed in Chib and Jeliazkov (2001) to calculate these quantities using a MCMC algorithm. We choose to use a Random Walk Metropolis algorithm

(described in Algorithm 1).

From Bayes theorem we have

$$\pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{y}) = \frac{\pi_k(\boldsymbol{\theta}_k) f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)}{m_k(\boldsymbol{y})}, \quad k = 0, \ldots, K.$$

It follows that $m_k(\boldsymbol{y})$ is the normalizing constant of the posterior distribution

$$m_k(\boldsymbol{y}) = \frac{f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)}{\pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{y})}.$$

This expression, called *basic marginal likelihood identity*, is evaluated on a given arbitrary point $\boldsymbol{\theta}^*$. In particular we calculate its logarithm

$$\log m_k(\boldsymbol{y}) = \log f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k^*) + \log \pi_k(\boldsymbol{\theta}_k^*) - \log \pi_k(\boldsymbol{\theta}_k^* \mid \boldsymbol{y}). \tag{2.18}$$

This expression says that it suffices to approximate the posterior distribution in a point $\boldsymbol{\theta}^*$. Then, using (2.1), the BF_{ij} is calculated as

$$BF_{ij} = \exp\left(\log(m_i(\boldsymbol{y})) - \log(m_j(\boldsymbol{y}))\right).$$

Finally, the IBFs for a given number L of training samples and the FBF are approximated using definitions in (2.9), (2.10) and (2.13) by

$$\begin{aligned} BF_{ij}^{AI} = &\exp\left(\log(m_i(\boldsymbol{y})) - \log(m_j(\boldsymbol{y}))\right) \times \\ &\times \frac{1}{L}\sum_{l=1}^{L} \exp\left(\log(m_j(\boldsymbol{y}(l))) - \log(m_i(\boldsymbol{y}(l)))\right) \end{aligned} \tag{2.19}$$

for the arithmetic mean and

$$\begin{aligned} BF_{ij}^{MI} = &\exp\left(\log(m_i(\boldsymbol{y})) - \log(m_j(\boldsymbol{y}))\right) \times \\ &\times Median_{l=1}^{L}\left(\exp\left(\log(m_j(\boldsymbol{y}(l))) - \log(m_i(\boldsymbol{y}(l)))\right)\right) \end{aligned} \tag{2.20}$$

for the median and

$$\begin{aligned} BF_{ij}^{F,b} = &\exp\left(\log(m_i(\boldsymbol{y})) - \log(m_j(\boldsymbol{y}))\right) \times \\ &\times \exp\left(\log(m_{j,b}(\boldsymbol{y})) - \log(m_{i,b}(\boldsymbol{y}))\right) \end{aligned} \tag{2.21}$$

for the FBF.

Approximation of expression given in equation (2.18)

As already mentioned in Subsection 1.4.1.3, the posterior distribution is approximated by simulation, using a random walk MH algorithm. The goal is to estimate the posterior distribution $\pi(\boldsymbol{\theta}^* \mid \boldsymbol{y})$ in $\boldsymbol{\theta}^*$, given the posterior sample $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$ (in each case from the involved full posterior, trained posterior or fractional one). Here we illustrate the algorithm proposed by Chib and Jeliazkov (2001).
Let $q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y})$ denote the proposal density for the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, and

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y}) = min\Big\{1, \frac{\pi(\boldsymbol{\theta}')f(\boldsymbol{y} \mid \boldsymbol{\theta}')q(\boldsymbol{\theta}', \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi(\boldsymbol{\theta})f(\boldsymbol{y} \mid \boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y})}\Big\}.$$

If we write

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y}) = \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y}),$$

from reversibility it is possible to obtain, for any point $\boldsymbol{\theta}^*$

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \boldsymbol{y})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \pi(\boldsymbol{\theta}^* \mid \boldsymbol{y})p(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \boldsymbol{y}).$$

Integrating both sides over $\boldsymbol{\theta}$, we have

$$\begin{aligned}\pi(\boldsymbol{\theta}^* \mid \boldsymbol{y}) &= \frac{\int p(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \boldsymbol{y})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}} \\ &= \frac{\int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \boldsymbol{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \boldsymbol{y})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \boldsymbol{y})q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \boldsymbol{y})d\boldsymbol{\theta}}\end{aligned}$$

and so

$$\hat{\pi}(\boldsymbol{\theta}^* \mid \boldsymbol{y}) = \frac{\frac{1}{M}\sum_{g=1}^{M} \alpha(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^* \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^* \mid \boldsymbol{y})}{\frac{1}{J}\sum_{j=1}^{J} \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)} \mid \boldsymbol{y})},$$

where $\{\boldsymbol{\theta}^{(g)}\}$ are draws from the posterior distribution and $\{\boldsymbol{\theta}^{(j)}\}$ are draws from the proposal $q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \boldsymbol{y})$.
Then substituting $\hat{\pi}(\boldsymbol{\theta}^* \mid \boldsymbol{y})$ in the logarithm of the *marginal likelihood identity* we obtain

$$\log(\hat{m}(\boldsymbol{y})) = \log(f(\boldsymbol{y} \mid \boldsymbol{\theta}^*)) + \log(\pi(\boldsymbol{\theta}^*)) - \log(\hat{\pi}(\boldsymbol{\theta}^* \mid \boldsymbol{y})).$$

Using this methodology we estimate different BFs given in equations (2.19), (2.20) and (2.21).
Observe that $\boldsymbol{\theta}^*$ must be chosen between values of high posterior density.

In Algorithm 2 the calculation of a generic Bayes factor is shown. In our simulation studies, approximations of BFs have been done using 10^4 MCMC samples and taking $\boldsymbol{\theta}^*$ equal to the posterior median.

Another method to approximate BFs can be the Laplace approximation, described in Lewis and Raftery (1997).

Algorithm 2 Approximation of the $B_{ij}(\boldsymbol{y})$

Require: N, number of RW-MH MCMC steps; τ, fixed scale factor for $k = i, j$; $D_k = (\boldsymbol{y}, \boldsymbol{X}_k)$ data; $\pi_k(\boldsymbol{\theta}_k|D_k)$ posterior kernel of M_k; $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|D_k)$ probability of moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ and its corresponding proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|D_k)$; $f_k(D_k|\boldsymbol{\theta}_k)$ likelihood function.

1: Do for $k = i, j$;

2: Calculate MLE, $\hat{\boldsymbol{\theta}}_k$, and the observed information matrix $\hat{\Sigma}_k^{-1}$;

3: Generate $\boldsymbol{\theta}_k^{(1,\ldots,N)} \sim \pi_k(\boldsymbol{\theta}_k|D_k)$ by a RW-MH using a normal proposal with covariance $\tau\hat{\Sigma}_k$;

4: Generate $\widetilde{\boldsymbol{\theta}}_k^{(1,\ldots,N)}$ from the normal proposal;

5: Approximate the posterior median of $\pi_k(\boldsymbol{\theta}_k|D_k)$ by $\theta_k^* = \text{Median}(\boldsymbol{\theta}_k^{(1,\ldots,N)})$;

6: Approximate the posterior density, by

$$\hat{\pi}_k(\theta_k^*|D_k) = \sum_{n=1}^{N} \alpha(\boldsymbol{\theta}_k^{(n)}, \theta_k^*|D_k) q(\boldsymbol{\theta}_k^{(n)}, \theta_k^*|D_k) / \sum_{n=1}^{N} \alpha(\theta_k^*, \widetilde{\boldsymbol{\theta}}_k^{(n)}|D_k);$$

7: Approximate the predictive density at θ_k^* by $\hat{m}_k(\theta_k^*) = f_k(D_k|\theta_k^*)\pi_k(\theta_k^*)/\hat{\pi}_k(\theta_k^*|D_k)$;

8: **return** Approximation of the $B_{ij}(\boldsymbol{y})$ as $\hat{m}_i(\theta_i^*)/\hat{m}_j(\theta_j^*)$.

2.6 Highest Posterior Probability Model and Median Probability Model

Once we have calculated the IBFs, the FBFs and the BICs, it is necessary to rank all the considered models. Here we present two different techniques proposed in the literature. In the space of all models the posterior probability of each one is computed and by doing this for each M_k, $k = 0, \dots, K$, we obtain an ordering in $\mathcal{M}$.
One common choice for the prior probability of the models is the discrete uniform distribution, so that each model has the same initial probability. In Spiegelhalter et al. (1993) and Lauritzen et al. (1994) the benefits of using informative prior distributions are analysed. Another approach, presented in Raftery et al. (1999), consists in choosing as prior distribution for model M_j

$$\Pr(M_j) = \prod_{k=1}^{p} \pi_k^{\delta_{jk}} (1 - \pi_k)^{1-\delta_{jk}}$$

where $\pi_k \in [0, 1]$ is the prior probability that the vector of regression coefficients is different from the null one in model M_j, and δ_{jk} is an indicator of whether or not variable k is included in model M_j. If we choose $\pi_k = 0.5$, for all k, then the prior is a uniform distribution. Choosing $\pi_k < 0.5$ gives a penalty for large models, while if we put $\pi_k = 1$, then it ensures that the variable k is included in each model (more details on this approach can be found in George and McCulloch (1993)).
Models analysed here are not sparse and hence we use the uniform prior

$$\Pr(M_k) = \frac{1}{K+1}, \quad k = 0, \dots, K.$$

If one deals with sparse models, we recommend to use other prior specification approaches as discussed in Scott and Berger (2010).

Highest Posterior Probability Model (HPPM)

Recalling the equation in (2.3), jointly with the uniform prior over the space of models, we obtain

$$\begin{aligned}
\Pr(M_j \mid \boldsymbol{y}) &= \left\{ \sum_{k=0}^{K} B_{kj} \frac{\Pr(M_k)}{\Pr(M_j)} \right\}^{-1} = \left\{ \sum_{k=0}^{K} B_{kj} \right\}^{-1} \\
&= \frac{1}{\sum_{k=0}^{K} B_{kj}} \\
&= \frac{B_{j0}}{\sum_{k=0}^{K} B_{k0}}
\end{aligned}$$

or, equivalently

$$\Pr(M_j \mid \boldsymbol{y}) = \frac{B_{j0}}{B_{00} + \sum_{k=1}^{K} B_{k0}} = \frac{B_{j0}}{1 + \sum_{k=1}^{K} B_{k0}}.$$

More specifically, for the null model we have

$$\Pr(M_0 \mid \boldsymbol{y}) = \frac{B_{00}}{1 + \sum_{k=1}^{K} B_{k0}} = \frac{1}{1 + \sum_{k=1}^{K} B_{k0}}.$$

Once we have obtained all the posterior probabilities, the models can be ordered according to these values and we choose the one having the highest posterior probability.

Median Posterior Probability Model (MPPM)

Another approach, introduced by Barbieri and Berger (2004) and called the *median probability model* method, consists in choosing the model containing those variables which have overall posterior probability at least 1/2 of being included along all the considered models.
Let $\boldsymbol{h}$ be the set of indexes of all the models containing a given variable. The following definition introduces the concept of inclusion probability of a variable.

Definition 3. (Posterior inclusion probability) *The posterior inclusion probability for a variable i is*

$$d_i = \sum_{h \in \boldsymbol{h}} \Pr(M_h \mid \boldsymbol{y})$$

that is, the overall posterior probability that the variable i is in the model.

If it exists, the *median probability model* M_{h^*} is the model consisting of those variables whose posterior inclusion probability is greater than or equal to 1/2.
If we define $\boldsymbol{h}^* = (h_1^*, \ldots, h_m^*)$, then

$$h_i^* = \begin{cases} 1, & \text{if } d_i \geqslant \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, m$. Sometimes it may happen that the median model does not exist: this is the case when the set of covariates defined by h^* does not correspond to a model under consideration. For more details see Barbieri and Berger (2004).

3

Variable Selection under Censoring using Sequential Minimal Training Samples

3.1 Introduction

In problems of reliability and survival analysis we often have to deal with censored data. In this case, as already seen in 1.4.1.1 and 1.5.1.1, the likelihood functions for survival models contain the censoring indicator $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$.
In the remainder of this thesis we will consider the right censoring case, already introduced in Chapter 1 (see Section 1.2 for more details).
In general, let $(t_i, \delta_i, \boldsymbol{x}_i)$ be the survival time, censoring indicator and covariates, respectively for individual $i = 1, \ldots, n$, where $\delta_i = 0$ if right censored and 1 if uncensored. Consider $y_i = \log(t_i)$ and the following regression model M_k with a set of covariates denoted by $\boldsymbol{x}_{k,i}$

$$M_k : y_i = {\boldsymbol{\beta}_k}^{\mathsf{T}} \boldsymbol{x}_{k,i} + \sigma_k \epsilon_i,$$

where $\boldsymbol{X}$ is a fixed design matrix with r columns, including the intercept, and ϵ_i is the error term with d.f. $f(\epsilon)$. Then, denoting by $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k)$, the corresponding likelihood function for right censored data has the following form

$$\begin{aligned} L(\boldsymbol{y} \mid \boldsymbol{\theta}_k, \boldsymbol{X}) &= \prod_{i=1}^{n} f_Y(y_i)^{\delta_i} \left[S_Y(y_i)\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma_k} f_\epsilon\left(\frac{y_i - {\boldsymbol{\beta}_k}^{\mathsf{T}} \boldsymbol{x}_{k,i}}{\sigma_k}\right)\right]^{\delta_i} \left[S_\epsilon\left(\frac{y_i - {\boldsymbol{\beta}_k}^{\mathsf{T}} \boldsymbol{x}_{k,i}}{\sigma_k}\right)\right]^{(1-\delta_i)}. \end{aligned}$$

3. VARIABLE SELECTION UNDER CENSORING USING SEQUENTIAL MINIMAL TRAINING SAMPLES

In this setting, it is necessary to redefine the concepts of IBF, FBF and BIC when data are censored and, hence, observations are not iid.
In Chapter 2 the concepts of training samples and minimal training samples have been introduced, they are used to convert improper priors into the proper distributions needed for model selection. However, when some observations in the set $\boldsymbol{y}$ are censored, it is important to reformulate the concept of MTS. We recall that in the uncensored case the minimal dimension of the MTS is equal to the number of parameters s in the model. In Berger and Pericchi (2004) the hypothetical sampling space of proper training samples, $\mathcal{X}^I$, obtained when it is assumed that an infinite amount of data is available, is introduced. It is required that, in drawing MTSs, the space of all possible MTSs should be fully explorable, that is, for each model the sampling mechanism of the MTS must cover the space of all the MTSs with probability 1 (this is stated in the following Assumption 0 of Berger and Pericchi (2004)):

Assumption 0: $\Pr^{M_i}_{\boldsymbol{\theta}_i}(\mathcal{X}^I) = 1$, for $i = 0, \ldots, K$.

In situations involving censoring, this assumption can be violated.
We now give an example of such situation already presented in Berger and Pericchi (2004).

Example 5. (Right censored exponential) *Suppose that data $y_1, y_2, \ldots, y_n$ are a random sample from the right censored exponential distribution, with censoring time ρ. Thus if $y_i < \rho$, then the density is $f(y_i \mid \theta) = \theta \exp(-\theta y_i)$, while if the data are censored, the density is $\Pr(Y_i = \rho \mid \theta) = \exp(-\rho\theta)$. We are interested in testing the two hypotheses*

$$M_0 : \theta = \theta_0 \quad vs \quad M_1 : \theta \neq \theta_0.$$

We choose the usual default prior for the exponential model, $\pi^N(\theta) = 1/\theta$. It can be seen that one single uncensored observation is sufficient to obtain a proper posterior, while no censored observation can achieve this. So the imaginary set of minimal training samples consists of single uncensored observations. Denoting the sampling space of training samples of the form $(0, \rho)$ with $\mathcal{X}^{MI}$, we can prove that the Assumption 0 is violated, in fact

$$\Pr^{M_i}_{\boldsymbol{\theta}_i}(\mathcal{X}^{MI}) = \Pr^{M_i}_{\boldsymbol{\theta}_i}(Y < \rho) = 1 - \exp(-\rho\theta_i) < 1, \quad i = 0, 1.$$

Observe that an enumeration of all possible MTSs that jointly satisfy Assumption 0 may not be feasible in general. For this reason Berger and Pericchi (2004) propose a sequential minimal training sample scheme which satisfies Assumption 0 in the context

of censored data (see Appendix A for a proof in the case of the right exponential distribution in Example 5).

3.2 Sequential Minimal Training Sample

The following definition plays a central role in the sequel:

Definition 4. (Sequential Minimal Training Sample (SMTS)) *Suppose we have s parameters in the model, then the SMTS is constructed drawing observations, without replacement, from $\boldsymbol{y}$ stopping when s uncensored observations are obtained. The SMTS induces a TS of the form*

$$\boldsymbol{y}(l) = \{\underbrace{\ldots}_{s-1 \text{ uncensored and } N_t - s \text{ censored observations}}, y_s(l)\},$$

with random size $N_t \geq s$ and $y_s(l)$ the $s-th$ uncensored observation.

Note that $\boldsymbol{y}(l)$ is not, in general, a MTS because it contains censored observations that can be removed but it is *minimal* in the sense that the last uncensored observation cannot be removed from the sample.
Observe that in this case the dimension of the SMTS is random. For our purposes, it is useful to obtain the probability distribution of the SMTS size that is derived in the following proposition.

Proposition 2. *Let $\boldsymbol{y}$ be the set of independent observations of size n, $\boldsymbol{y}(l)$ a SMTS, s the number of parameters in the model under study (i.e. for model M_k, $s = r_k + 1$), $n_{cens} = n - \sum_{i=1}^{n} \delta_i$ the number of censored observations and N_t be the SMTS size. Then the probability distribution of N_t is*

$$\Pr_{N_t}(N_t = n_t) = \frac{\binom{n_{cens}}{n_t - s}\binom{n - n_{cens} - 1}{s-1}(n_t - 1)!(n - n_{cens})}{D_{n,n_t}}, \tag{3.1}$$

where $N_t \in \{s, \ldots, n_{cens} + s\}$ and $D_{n,n_t} = \frac{n!}{(n-n_t)!}$.

Proof. As each observation, censored or uncensored, has the same probability to be extracted, it can be used the classical definition of probability of the event $N_t = n_t$

$$\Pr(N_t = n_t) = \frac{\text{favorable cases}}{\text{possible cases}}.$$

For the denominator, observe that this is the number of all possible ways in which N_t elements can be chosen out of a set of n. When repetitions are not allowed, this

number is given by the dispositions of n elements of class n_t

$$D_{n,n_t} = \frac{n!}{(n-n_t)!}.$$

For the numerator, we have to sample without replacement until we reach s uncensored observations. The sample takes the form

$$\text{SMTS of size } N_t = n_t : \begin{cases} (n_t-1) \text{ observations of which } \begin{cases} (s-1) \text{ uncensored} \\ (n_t-1)-(s-1) \text{ censored} \end{cases} \\ 1 \text{ uncensored observation} \end{cases}$$

The $(s-1)$ uncensored observations can be chosen in $\binom{n-n_{cens}-1}{s-1}$ ways and the $(n_t-1)-(s-1)$ censored observations in $\binom{n_{cens}}{n_t-s}$ ways. All these can be permuted in $(n_t-1)!$ ways. For the last observation, we have to take into account how many uncensored observations are contained in the sample, $(n-n_{cens})$. So the final probability is the one given in (3.1).

□

In the case of the Weibull and log-normal models, the smallest model to be considered along the thesis has two parameters, so $s \geqslant 2$ and $N_t \geqslant 2$.
Figure 3.1 illustrates the probability distribution of the SMTS size for $s = 3$, different sample sizes ($n = 10$, $n = 50$ and $n = 100$) and different censoring percentages (30% and 50%). As we can see, in all these settings the distribution of N_t is asymmetric, having a long right tail, as the percentage of censoring grows or n grows, there are more possible values for N_t resulting in a more diffuse distribution of N_t.

Example 6. (Probability distribution of N_t) *Suppose to have $n = 6$ observations of which 3 are uncensored and where the number of parameters in the model is $s = 2$. So*

$$N_t \in \{2, 3, 4, 5\}.$$

In this case we have to sample until we reach 2 uncensored observations. There are four possible ways:

1. $\boldsymbol{N_t = 2}$
   - ***Favorable cases:*** *the only case is when there are two uncensored observations from the three possible ones $D_{3,2} = 6$.*
   - ***Possible cases:*** $D_{6,2} = 30$.

So

$$\Pr(N_t = 2) = \frac{6}{30} = \frac{1}{5}.$$

2. $\boldsymbol{N_t = 3}$

   - ***Favorable cases:*** *we can choose three censored observations in groups of one, two uncensored observations in groups of one and then we take into account all the possible permutations,* $2!$*. For the last observation there are three possible choices among the three uncensored observations. This number is* $\binom{3}{1}\binom{2}{1}2!\,3$*.*
   - ***Possible cases:*** $D_{6,3} = 120$*.*

   So

$$\Pr(N_t = 3) = \frac{\binom{3}{1}\binom{2}{1}2!\,3}{D_{6,3}} = \frac{3}{10}.$$

3. $\boldsymbol{N_t = 4}$

   - ***Favorable cases:*** *here we have to choose three censored observations in groups of two and two uncensored observations in groups of one, then there are* $3!$ *possible permutations of them. For the last observation, again, there are three possible choices among the three uncensored observations. This number is* $\binom{3}{2}\binom{2}{1}3!\,3$*.*
   - ***Possible cases:*** $D_{6,4} = 360$*.*

   So

$$\Pr(N_t = 4) = \frac{\binom{3}{2}\binom{2}{1}3!\,3}{D_{6,4}} = \frac{3}{10}.$$

4. $\boldsymbol{N_t = 5}$

   - ***Favorable cases:*** *in this case we have to choose three censored observations in groups of three and then two uncensored observations in groups of one, so there are* $4!$ *possible permutations. For the last observation there are three possible choices among the three uncensored observations. This number is* $\binom{3}{3}\binom{2}{1}4!\,3$*.*
   - ***Possible cases:*** $D_{6,5} = 720$*.*

So

$$\Pr(N_t = 5) = \frac{\binom{3}{3}\binom{2}{1}4!\,3}{D_{6,5}} = \frac{1}{5}.$$

It is straightforward to check that

$$\Pr(N_t = 2) + \Pr(N_t = 3) + \Pr(N_t = 4) + \Pr(N_t = 5) = 1.$$

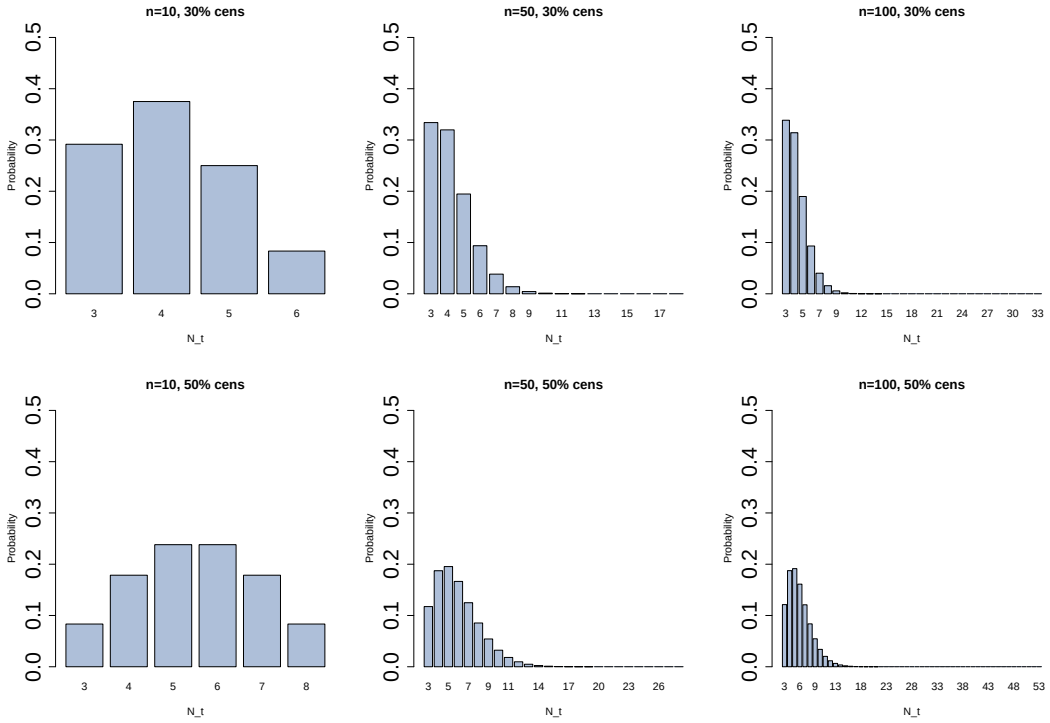

Figure 3.1: Probability distribution of N_t for samples of sizes 10, 50 and 100, for 30% and 50% of censoring with $s = 3$.

3.3 IBF under Censoring

The calculation of the IBF is computationally demanding because it is necessary to sum over $l = 1, \ldots, L$, where L is the number of all possible SMTSs, and this can be an important number. The natural solution would be to sum over all possible outcomes of the SMTS, but this may be unfeasible even in very simple situations with small samples and simple models. This is partially accomplished by the solution to sum over a subset

of SMTSs. In fact, as mentioned in Section 2 of Berger and Pericchi (2004) and in Varshavsky (1995), it is often sufficient to randomly choose $L = n \times n_t$, with replacement, where n is the sample size and n_t is the training sample size. However, in the case of SMTS, n_t is replaced by its random counterpart N_t and, hence, L becomes a random quantity.

It would be too costly to evaluate the IBF at each value of L unless the number of possible outcomes of the SMTS were small enough with respect to the available computational resources. We instead consider the two following definitions of L:

$$L_{mode} = n \times \text{mode}\{N_t\}$$
$$L_{median} = n \times [\text{median}\{N_t\}]$$

where N_t is the SMTS size that is random with distribution given by (3.1), and $[x]$ denotes the integer part of x.

Recalling Section 2.3, the idea is to compare each model M_i with the encompassing model M_0, the null one, through pairwise comparison from below. However, it is worth noting that using the encompassing from below approach, the analyst is forced to use a common number of uncensored observations, namely s for the full model, and if s is large ($\approx n - n_{cens}$), then the induced intrinsic and fractional priors can be quite informative. This problem is also common to a setup without censoring, then it may be viewed as a downfall of the encompassing from below procedure and not of the discussed versions of the BFs for censored data.

Distribution 3.1 can be further employed for implementing a stratified SMTS sampling so that the distribution of sample sizes follows 3.1. This would require to enumerate all the possible L SMTSs, which may be unfeasible for large sample sizes. Another way to introduce distribution 3.1 is to use it in reweighting the Monte Carlo samples of SMTSs according to their sizes. For purposes of comparisons with the actual version of the SMTS, this latter strategy would not be further pursued in this work. For each SMTS, $\boldsymbol{y}(l)$, we obtain

$$B_{0j}^N(\boldsymbol{y}(l))$$

and, after that, we calculate

$$BF_{j0}^{AI} = B_{j0}^N(\boldsymbol{y}) \frac{1}{L} \sum_{l=1}^{L} B_{0j}^N(\boldsymbol{y}(l))$$

and

$$B_{j0}^{MI} = B_{j0}^N(\boldsymbol{y}) \underset{l=1,\ldots,L}{\text{Median}} B_{0j}^N(\boldsymbol{y}(l)).$$

The values of L_{mode} and L_{median} are obtained considering the most complex model M_i, that is the one in which all the covariates are included. Using this rule, the number of random SMTSs used to approximate BF_{j0}^{AI} and BF_{j0}^{MI} is the same for each $j = 1, \ldots, K$.

For the practical calculation of the IBFs, L SMTSs are calculated in order to compute the mean and the median of the correction factor. As L is large, we resort to parallel computation because the evaluation of the different $BF_{ij}s$ is performed independently: we assign a certain number of partial BFs, $B_{ij}^N(\boldsymbol{y}(l))$, to different processors using the functions of the library *Rmpi* [1] in R. This allows us to speed up the computations, but even with this the IBF is still long to be computed.

Example 7. (Example 2 continued) *We calculate the IBF in the case of right censored data. The SMTS is constructed by randomly sampling from the entire set of observations and stopping when* 2 *uncensored observations are obtained. The SMTS size has probability given in* (3.1)*, where* $s = 2$.

The expressions of the AIBF and MIBF are analogous to the ones of the (2.11) *and* (2.12) *where the* $\boldsymbol{y}(l)$ *now has a different form.*
With an abuse of notation we are denoting with y_i *both censored and uncensored observations as these can be clearly recognized taking into account the index of the sums. Denoting by* n_u *the number of uncensored observations and by* $T = \sum_{i=1}^{n} y_i$ *the sum of all the observations, we obtain the expressions of the two marginal distributions*

$$m_0^N(\boldsymbol{y}) = \int_0^\infty \frac{1}{\lambda}\left(\lambda^{n_u}\exp\left(-\lambda\sum_i^{n_u} y_i\right)\right)\left(\exp\left(-\lambda\sum_i^{n_{cens}} y_i\right)\right)d\lambda = \frac{\Gamma(n_u)}{T^{n_u}}$$

and

$$\begin{aligned} m_1^N(\boldsymbol{y}) &= \int_0^\infty\int_0^\infty \frac{1}{\alpha\beta}\alpha^{n_u}\beta^{n_u}\prod_i^{n_u} y_i^{\alpha-1}\exp\left(-\beta\sum_i^{n_u} y_i^\alpha\right)\exp\left(-\beta\sum_i^{n_{cens}} y_i^\alpha\right)d\alpha d\beta \\ &= \Gamma(n_u)\int_0^\infty \frac{\alpha^{n_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{n_u}}\prod_i^{n_u} y_i^{\alpha-1}d\alpha. \end{aligned}$$

Each SMTS contains two uncensored observations, y_h *and* y_k*, and a random number of censored observations, say* j*, with* ρ *denoting the right censoring time, then marginal distributions for both models calculated in a SMTS* $\boldsymbol{y}(l)$ *are:*

$$m_0^N(\boldsymbol{y}(l)) = \frac{1}{(y_h + y_k + \rho j)^2}$$

and

$$m_1^N(\boldsymbol{y}(l)) = \int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha + y_k^\alpha + j\rho^\alpha)^2}d\alpha.$$

[1] Package "Rmpi"

The corresponding IBFs are

$$BF_{10}^{AI} = T^{n_u} \int_0^\infty \frac{\alpha^{n_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{n_u}} \prod_i^{n_u} y_i^{\alpha-1} d\alpha \frac{1}{L} \sum_D \frac{(y_h + y_k + \rho j)^{-2}}{\int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha + y_k^\alpha + j\rho^\alpha)^2} d\alpha} \tag{3.2}$$

and

$$BF_{10}^{MI} = T^{n_u} \int_0^\infty \frac{\alpha^{n_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{n_u}} \prod_i^{n_u} y_i^{\alpha-1} d\alpha \; \underset{D}{Median} \frac{(y_h + y_k + \rho j)^{-2}}{\int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha + y_k^\alpha + j\rho^\alpha)^2} d\alpha} \tag{3.3}$$

where $D = \{h < k\} \times \{j | j = 0, \ldots, n_{cens}\}$ *and*

$$L = \binom{n_u}{2} \sum_{j=0}^{n_{cens}} \binom{n_{cens}}{j} = \frac{n_u(n_u-1)}{2} \sum_{j=0}^{n_{cens}} \binom{n_{cens}}{j}.$$

These two Bayes Factors must be approximated numerically.

3.3.1 Intrinsic Prior for the IBF

An essential point in our discussion is the intrinsic prior, either intrinsic to the IBF or to the FBF. If we were able to derive such a prior, we could state that the versions of the considered BFs are actually real BFs in an asymptotic sense. In the case of censored data the calculation tends to be more difficult than in the case of all uncensored observations. Recalling the definition of intrinsic prior given in Subsection 2.4.1, we now show the following toy example from Berger and Pericchi (2004) which contains details about the calculation of the intrinsic prior in the case of censored data in the right exponential model.

Example 8. (Example 5 continued) *Suppose we want to test the following hypotheses for the right exponential model*

$$M_0 : \theta = \theta_0$$
$$M_1 : \theta \neq \theta_0$$

and let $\boldsymbol{y}$ *be the sample of observations. If we choose the usual default prior for this model* $\pi^N(\theta) = 1/\theta$*, we obtain*

$$B_{10}^N(\boldsymbol{y}) = \frac{m_1^N(\boldsymbol{y})}{m_0^N(\boldsymbol{y})} = \frac{\int \theta^{n_u-1} \exp\left(-\theta \sum_{i=1}^n y_i\right) d\theta}{\theta_0^{n_u} \exp\left(-\theta_0 \sum_{i=1}^n y_i\right)} = \Gamma(n_u) \left(\theta_0 \sum_{i=1}^n y_i\right)^{-n_u} \exp\left(\theta_0 \sum_{i=1}^n y_i\right).$$

If we denote by $\boldsymbol{y}(l)$ *the SMTS, let* $p(\theta) = \Pr(Y > \rho|\theta) = \exp(-\theta\rho)$*,* N_c *the number of censored observations in* $\boldsymbol{y}(l)$ *and* $y(l)$ *the single uncensored observation in the SMTS,*

then $\Pr(N_c = j|\theta) = (1-p(\theta))p(\theta)^j$ *and* $f(\boldsymbol{y}(l)|\theta) = p(\theta)^j\theta\exp(-\theta y(l))$.
So the correction factor of the IBF is:

$$B_{01}^N(\boldsymbol{y}(l)) = \frac{m_0^N(\boldsymbol{y}(l))}{m_1^N(\boldsymbol{y}(l))} = \theta_0\left(N_c\rho + y(l)\right)\exp\left(-(N_c\rho + y(l))\theta_0\right).$$

From (2.15) *the intrinsic prior is*

$$\begin{aligned}\pi^I(\theta) &= \pi^N(\theta)E_\theta^{M_1}\left[B_{01}^N(\boldsymbol{y}(l))\right]\\ &= \frac{1}{\theta}\sum_{j=0}^{\infty}\int_0^\rho \theta_0(j\rho+y)\exp\left(-(j\rho+y)\theta_0\right)p(\theta)^j\theta\exp(-\theta\rho)dy\\ &= \frac{\theta_0}{(\theta+\theta_0)^2}.\end{aligned}$$

As noted in Berger and Pericchi (2004) this prior is proper and has median equal to θ_0. *It agrees with the intrinsic prior for the exponential model without censoring when using an ordinary MTS (see Pericchi et al. (1993)).*

We consider an example in which we derive the intrinsic prior when comparing two Weibull models.

Example 9. (Weibull vs. Weibull) *Suppose we want to compare two Weibull models, the first one with known parameter* $\beta = \beta_0$ *and the second one with unknown parameter* β

$$\begin{aligned}M_0 &: f_0(y \mid \alpha, \beta_0) = \alpha\beta_0 y^{\alpha-1}\exp(-\beta_0 y^\alpha)\\ M_1 &: f_1(y \mid \alpha, \beta) = \alpha\beta y^{\alpha-1}\exp(-\beta y^\alpha).\end{aligned}$$

The intrinsic prior in the case of right censored data, with ρ *denoting the right censoring time, has the form of the* (2.15), *where* $\pi_j^N(\boldsymbol{\theta}_j)$ *is the Jeffreys' prior under the model* M_1, *that is* $\pi(\alpha,\beta) \propto 1/\alpha\beta$.
Let y_h *and* y_k *be the two uncensored observations in the SMTS and* j *the number of censored observations in the SMTS, then the two marginal distributions calculated over the SMTS are*

$$m_0^N(\boldsymbol{y}(l)) = \beta_0^2\int_0^\infty \alpha(y_h y_k)^{\alpha-1}\exp\left(-\beta_0(y_h^\alpha + y_k^\alpha + j\rho^\alpha)\right)d\alpha$$

and

$$m_1^N(\boldsymbol{y}(l)) = \int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha + y_k^\alpha + j\rho^\alpha)^2}d\alpha.$$

So we have

$$B_{01}^N(\boldsymbol{y}(l)) = \frac{m_0^N(\boldsymbol{y}(l))}{m_1^N(\boldsymbol{y}(l))} = \frac{\beta_0^2\int_0^\infty \alpha(y_h y_k)^{\alpha-1}\exp\left(-\beta_0(y_h^\alpha + y_k^\alpha + j\rho^\alpha)\right)d\alpha}{\int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha + y_k^\alpha + j\rho^\alpha)^2}d\alpha}.$$

Using the notation introduced in Example 8, we now calculate the intrinsic prior for right censored data, where $p(\alpha,\beta) = \Pr(Y > \rho|\alpha,\beta) = 1 - \exp(-\beta\rho^\alpha)$

$$\begin{aligned}\pi_1^I(\alpha,\beta) &= \frac{1}{\alpha\beta}\sum_{j=0}^{\infty}\int_0^\rho B_{01}^N(\boldsymbol{y}(l))f(\boldsymbol{y}(l)|\alpha,\beta)d\boldsymbol{y}(l)\\ &= \beta_0^2\sum_{j=0}^{\infty}\int_0^\rho\int_0^\rho \frac{\int_0^\infty \alpha(y_h y_k)^{\alpha-1}\exp\left(-\beta_0(y_h^\alpha + y_k^\alpha + j\rho^\alpha)\right)d\alpha}{\int_0^\infty \frac{\alpha(y_h y_k)^{\alpha-1}}{(y_h^\alpha+y_k^\alpha+j\rho^\alpha)^2}d\alpha}\times\\ &\quad\times p(\alpha,\beta)^j\alpha\beta(y_h y_k)^{\alpha-1}\exp\left(-\beta(y_h^\alpha+y_k^\alpha)\right)dy_h dy_k\end{aligned}$$

which does not have a closed-form.

As pointed out by Berger and Pericchi (2004), the intrinsic prior cannot be obtained when the censoring mechanism in unknown. The type of censoring may induce complications and difficulties in the calculation of the intrinsic prior. In particular, for Type II censoring there exists a deterministic stopping rule which further complicates the likelihood. Finally, random censoring implies that the stopping time $\rho \in \mathcal{T}$ becomes random and this induces a space of possible MTSs which is also random and Assumption 0 should be regarded as marginal to the probability distribution of $R = \rho$, namely $H_R(\rho)$. The analysis may proceed in two steps: first obtaining the intrinsic prior conditioning at $R = \rho$ and then marginalizing it with respect to $H_R(\rho)$. This implies that the random censoring mechanism, represented by model $H_R(\rho)$, should be fully known.
The use of IBF and intrinsic priors in problems with censored or truncated data modeled through the Weibull distribution can be found in Lingham and Sivaganesan (1999) and Kim and Sun (2000).

3.4 FBF under Censoring

The calculation of the FBF is less computational demanding than that of the IBF because we do not have to calculate it over MTSs or SMTSs: the partial information on the data is provided by the fraction b of the likelihood which only depends on the size of the training sample.
Remind that O'Hagan (1995) suggests to take $b = n_t/n$, where n_t is the MTS size and n is the full sample size.
Again, in the case of censored data n_t is random, so we propose three different strategies:

1. **Mode:** we take $n_t = mode\{N_t\}$. In other words, we choose the most probable

value of the SMTS size, and

$$b = \frac{mode\{N_t\}}{n}. \tag{3.4}$$

2. **Median:** we choose the median of the SMTS size distribution

$$b = \frac{median\{N_t\}}{n}. \tag{3.5}$$

3. **Marginalization:** considering that N_t is random, we define B as a random variable

$$B = \frac{N_t}{n}, \tag{3.6}$$

where N_t takes values in $\{s, s+1, \ldots, s+n_{cens}\}$ with probability given in (3.1). Finally, we define our proposal for the FBF.

Definition 5. (Marginal Fractional Bayes Factor) *Let N_t be the SMTS size with probability (3.1), then the marginal FBF, mFBF, is given by*

$$mFBF_{ij} = \sum_{b=s/n}^{(s+n_{cens})/n} B_{ij}^{F,b} \Pr_B(B=b) = \sum_{b=s/n}^{(s+n_{cens})/n} B_{ij}^{F,b} \Pr_{N_t}(N_t = bn). \tag{3.7}$$

The practical calculation of mFBF can be done using parallel computation as each $B_{ij}^{F,b}(\boldsymbol{y})$ is obtained using a different processor, in total $n_{cens}+1$, and finally the weighted mean is obtained. In cases of mode and median the FBF is a particular case of one of the $n_{cens}+1$ BFs previously calculated. The different BFs are compared in a simulation study that appears in Section 3.6, also the computational cost needed to calculate the IBFs and the FBFs are analyzed in Subsection 3.6.1.

We now give some results on the consistency of the FBF and mFBF, which means that the BF in favor of the true model tends to infinity as the sample size infinitely grows. We first consider the consistency of the mFBF which depends on the fraction $B = \frac{N_t}{n}$ and on its probability distribution $\Pr_B(B = b)$.

Lemma 1. *Let $n_u = n - n_{cens}$ the number of uncensored observations, assuming that the number of uncensored observations is proportional to the sample size, $n_u = [w \times n]$, where w is the proportion of uncensored observations, then for $n \to \infty$ we have that $N_t \xrightarrow{d} \widetilde{N}_t \sim NegBinomial(s, w)$ with $E(\widetilde{N}_t) = s/w$ and $Var(\widetilde{N}_t) = s(1-w)/w^2$, being s the number of parameters for the assumed model.*

Proof. The result descends from the definition of the Negative Binomial random variable. In fact, the size N_t of the MTS from a SMTS, is the total number of trials, from an infinite population, with probability of success w and we stop until we obtain s successes, namely s uncensored observations. □

We recall that O'Hagan (1995) stated that the FBF is consistent if the fraction $b \to 0$ for $n \to \infty$. The following Proposition 3 states that also all the proposed versions of the FBF, which depend on a particular fraction B, are consistent.

Proposition 3. *Let $B = N_t/n$, for $B \in \{s/n, \ldots, (n_{cens} + s)/n\}$, and assuming that w is a fixed proportion of uncensored observations, then as $n \to \infty$, $B \xrightarrow{d} 0$.*

Proof. From Lemma 1 we know that $N_t \xrightarrow{d} \widetilde{N}_t$, then $E(B) = E(\widetilde{N}_t/n) \to 0$ and $Var(B) = Var(\widetilde{N}_t/n) \to 0$ because are fixed constants with respect to n. □

Example 10. (Example 2 continued) *We now calculate the FBF for the Weibull vs Exponential model in presence of right censored data. This time we have to choose the fraction b between the different proposals in (3.4), (3.5) or (3.6).*
The corresponding marginal fractional distributions are

$$m_b^N(\boldsymbol{y} \mid M_0) = \frac{\Gamma(bn_u)}{(bT)^{bn_u}}$$

and

$$m_b^N(\boldsymbol{y} \mid M_1) = \Gamma(bn_u) \int_0^\infty \frac{\alpha^{bn_u - 1}}{(b\sum_{i=1}^n y_i^\alpha)^{bn_u}} \prod_i^{n_u} y_i^{b(\alpha-1)} d\alpha.$$

The corresponding FBF can be calculated but, again, it does not have a closed-form

$$B_{10}^{F,b} = T^{n_u} \int_0^\infty \frac{\alpha^{n_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{n_u}} \prod_i^{n_u} y_i^{\alpha-1} d\alpha \frac{1}{(T)^{bn_u} \int_0^\infty \frac{\alpha^{bn_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{bn_u}} \prod_i^{n_u} y_i^{b(\alpha-1)} d\alpha}$$

for (3.4) and (3.5), and

$$mFBF_{10} = T^{n_u} \int_0^\infty \frac{\alpha^{n_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{n_u}} \prod_i^{n_u} y_i^{\alpha-1} d\alpha \sum_{b=2/n}^{(2+n_{cens})/n} \Pr_B(B = b) \times$$
$$\times \frac{1}{(T)^{bn_u} \int_0^\infty \frac{\alpha^{bn_u-1}}{(\sum_{i=1}^n y_i^\alpha)^{bn_u}} \prod_i^{n_u} y_i^{b(\alpha-1)} d\alpha}.$$

3.4.1 Fractional Prior for the FBF

Recalling the definition of fractional prior given in 2.4.2, we now give the following proposition.

Proposition 4. *Suppose we want to test the following hypotheses for the right exponential model*

$$M_0 : \theta = \theta_0$$
$$M_1 : \theta \neq \theta_0$$

and let $\boldsymbol{y}$ *be the sample of observations. The corresponding fractional prior for a fixed* n_t *is*

$$\pi^{FI,n_t}(\theta) = \frac{(\theta_0 n_t w)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w-1} \exp\left(-\frac{\theta_0 n_t w}{\theta}\right) \sim \text{InvGamma}(\alpha = n_t w, \beta = \theta_0 n_t w) \tag{3.8}$$

and the marginal fractional prior is

$$\pi_1^{FI}(\theta) = \sum_{n_t=1}^{\infty} w(1-w)^{n_t-1} \frac{(n_t w \theta_0)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w-1} \exp\left(-\frac{n_t w \theta_0}{\theta}\right), \tag{3.9}$$

which is a mixture of Inverse Gamma distributions, with parameters $\alpha = n_t w$ *and* $\beta = n_t w \theta_0$.

Proof. A complete proof can be found in Appendix C. □

Note 1. *Observe that the prior for the mFBF results in a mixture of fractional priors obtained for each FBF in equation (C.1), with weights given by* $\Pr_{N_t}(N_t = n_t)$. *Note that this prior is not a unit information one, i.e. a prior that provides as much information as one observation, as shown in Appendix C.*
In Figure 3.2 and 3.3 we compare the intrinsic prior obtained by Berger and Pericchi (2004), the fractional prior for the FBF in Example 3.3.1, for same values of B, *that is some fixed values of* N_t, *and the fractional prior for the mFBF here obtained. We have also included in these figures the probability mass function for* N_t *in order to interpret the most probable values in each case. It can be observed that, for a fixed censoring percentage, the fractional prior corresponding to a fixed* n_t *tends to be less dispersed as* n_t *grows and, as the censoring percentage increases, all the prior distributions become more vague. It is important to notice that the fractional prior for the mFBF is close to the intrinsic prior of Berger and Pericchi (2004) in the tails and it agrees with the fractional prior when calculated over the mode of* N_t *for smaller censoring percentages. This is due to the fact that, in this case, the mass function of* N_t *is concentrated on*

its mode. While for greater censoring percentages, as the mass of N_t is more spread, the fractional prior for the mFBF is very close to the intrinsic prior and very different from the fractional under the mode of N_t. Summarizing, the election of N_t = mode could result in a poor prior depending on the percentage of censoring, while it seems that the fractional prior for the mFBF behaves better independently of the censoring percentage. Finally, the fact that the fractional prior exists also assures the existence of the corresponding BF.

Again, observe that in the case of an unknown censoring mechanism, the fractional prior for the Weibull model corresponding to the FBF cannot be obtained.

3.5 BIC under Censoring

In Volinsky and Raftery (2000) it is proposed a version of BIC in case of censored survival models in which the sample size is replaced by an estimation of the effective sample size. In presence of censored data the sample size n cannot be used in the penalty term such as in (2.14), but we have to take into account the presence of censoring. The proposal of Volinsky and Raftery (2000) is to replace n with the number, n_u, of uncensored observations because this is the rate at which the Hessian matrix of the log-likelihood function grows. We adopt this definition of BIC

$$BIC^S_{ji} = -2\left(l_j(\hat{\boldsymbol{\theta}}_j) - l_i(\hat{\boldsymbol{\theta}}_i)\right) + (k_j - k_i)\log(n_u) \tag{3.10}$$

As observed by Volinsky and Raftery (2000), this criterion still has the asymptotic properties derived in Kass and Wasserman (1995).

Example 11. (Example 2 continued) *In the case of the comparison between the exponential and Weibull models with right censored data, the BIC takes the form*

$$BIC^S_{10} = -2\left(\hat{\alpha}^{n_u}\hat{\beta}^{n_u}\prod_i^{n_u} y_i^{\hat{\alpha}-1}\exp(-\hat{\beta}\sum_{i=1}^{n} y_i^{\hat{\alpha}}) - \hat{\lambda}^{n_u}\exp(-\hat{\lambda}\sum_{i=1}^{n} y_i)\right) + \log(n_u),$$

where $(\hat{\alpha}, \hat{\beta})$ are the maximum likelihood estimators for the Weibull model and $\hat{\lambda}$ is the maximum likelihood estimator for the exponential model.

An example of the use of the BIC for right censored data when working in a real application with the Weibull regression model can be found in Armero et al. (2012).

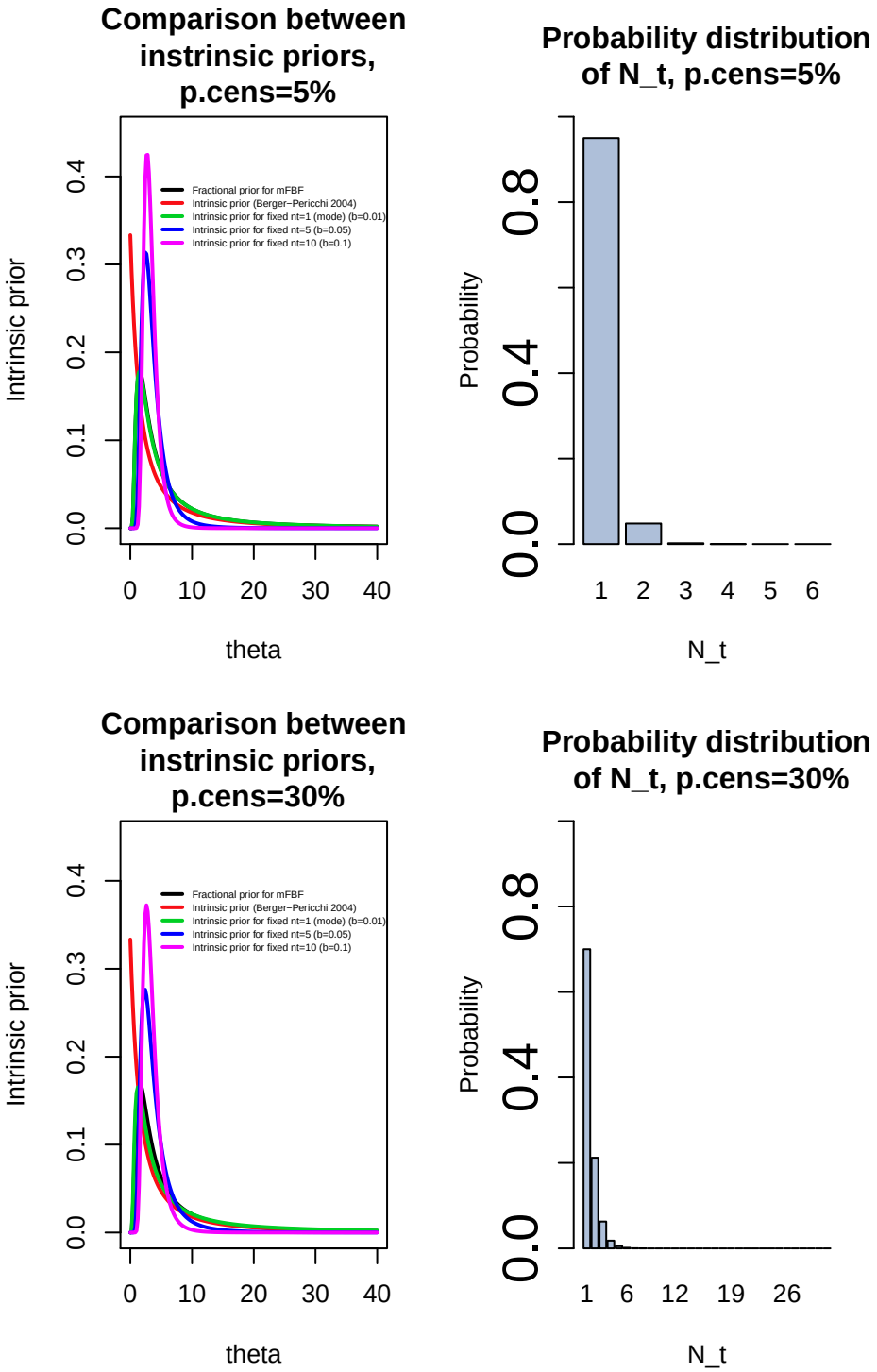

Figure 3.2: Comparison of different intrinsic and fractional priors for $\theta_0 = 3$, $n = 100$, $s = 1$, for different censoring percentages: 5% and 30% and corresponding mass function of N_t.

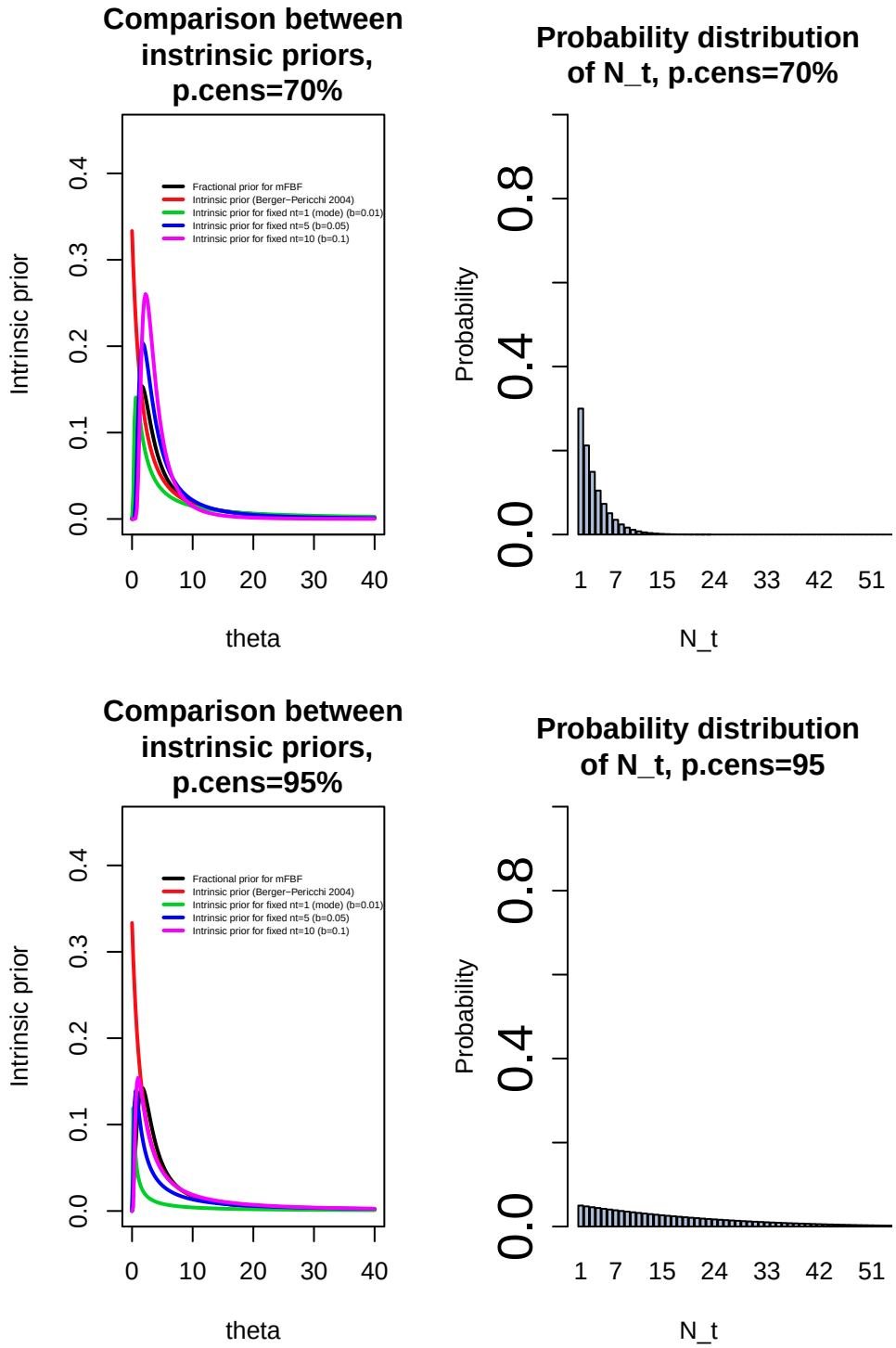

Figure 3.3: Comparison of different intrinsic and fractional priors for $\theta_0 = 3$, $n = 100$, $s = 1$, for different censoring percentages: 70% and 95% and corresponding mass function of N_t.

3.6 Simulation Study

In this section we present results of an ample simulation study in which we investigate and compare the performance of IBFs, FBFs and BIC defined in Sections 3.3, 3.4 and 3.5. All the simulations, calculations and graphics have been made using the statistical software R[1].
Our aim is to provide evidence against any significant difference between the FBF, in particular the mFBF, and the IBF which is more costly to compute.
We present results comparing the behavior of the IBFs, the FBFs and BIC over a set of simulated data from the Weibull and log-normal regression models.
First of all we simulate n observations from Y following a Weibull or log-normal distribution, with Weibull or log-normal censoring times, respectively, obtained as described in Appendix B and with two different censoring percentages: 10% and 30%.
In all cases the regression model from which data are simulated has the form

$$Y_i = \log(T_i) = \mu + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \sigma W_i \qquad i = 1, \ldots, n$$

where $W_i \sim f_W(w) = \exp(w - \exp(w))$ in the case of the Weibull model and $W_i \sim N(0,1)$ in the case of the log-normal model, with $w \in \mathbb{R}$.
The values of μ and σ are fixed to 0 and 1, respectively, for the four different models used to simulate data. In particular n observations have been drawn from the following four models:

M_0=**Null model:** $(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$.

M_1=**Model with 1 covariate:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 0, 0)$.

M_2=**Model with 2 covariates:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 1, 0)$.

M_3=**Model with 3 covariates:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 1, 1)$.

Finally, two different sample sizes have been used, $n = 50$ and $n = 100$, and the covariates X_1, X_2 and X_3 are independent, distributed according to a multivariate standard normal distribution.

For each model we have calculated all versions of IBF and FBF discussed above along with the version of BIC for censored data introduced in Subsection 3.5. Acceptance proportions are used to select the model along the 8 possible models in each

[1] The R Project for Statistical Computing

case. Results are based on 100 replications for each combination of simulation scenarios. Each replication of the dataset leads to an estimation of the distribution of the posterior probability for each possible model using 8 different tools for the Weibull case: $BF^{AI}_{L_{mo}}$, $BF^{MI}_{L_{mo}}$, $BF^{AI}_{L_{me}}$, $BF^{MI}_{L_{me}}$, FBF_{mo}, FBF_{me}, $mFBF$, BIC. We observe that the calculation of the IBF is computationally demanding, even using parallelism, and that, as we can see from the Weibull simulation study, the $B^{\cdot}_{L_{mo}}$ and the $B^{\cdot}_{L_{me}}$ lead to similar results (where the "$\cdot$" stays for AI or MI), so for the log-normal regression model we simply calculate the $B^{\cdot}_{L_{mo}}$ and the FBF_{mo}.
In order to analyse all these results we first applied an ANOVA analysis with the logit of the acceptance proportion of the true model as response variable, in order to estimate the main effects of: scenarios, BFs, selection criteria and type of models along with their possible interactions. From this analysis we observe that the most significant effects for the Weibull regression model are: the number n of observations, the type of true model and the type of BF. In particular, the full model, the $mFBF$, the $BF^{MI}_{L_{me}}$, the $BF^{MI}_{L_{mo}}$ have positive effects, which means that all the posterior probabilities grow, while the censoring percentage has a negative effect. For the log-normal model the most significant effects are: the number n of observations, the true model equal to the one with two covariates and the full one, the $BF^{MI}_{L_{mo}}$ (all of them with positive effects) and the $BF^{AI}_{L_{mo}}$ (with negative effect).

Figure 3.4 and Figure 3.5 provide an overview of the acceptance proportion of the true model, Weibull and log-normal respectively, marginally to all BFs and selection strategies. In fact, the considered BFs are consistent and the increasing proportion of censored observations complicates the model selection procedure.

Figures 3.6 and 3.7 show that there is no significant difference between the considered BFs marginally to all scenarios, although the $BF^{AI}_{L_{mo}}$ has more difficulties in selecting the true model in the case of the log-normal regression.

The dotcharts in Figures 3.8, 3.9, 3.10, 3.11 show the behavior of the considered BFs in two different scenarios: 30% and 10% of censoring. We have that $\widetilde{p}$, which is the acceptance proportion of the true model, decreases when the percentage of censored data increases, while it increases when the sample size increases. We denote by $se(\widetilde{p})$ the standard deviation of $\widetilde{p}$ and we also show $\pm se(\widetilde{p})$ in the dotcharts. A general idea about the behavior of all considered BFs can be obtained from Figures 3.8 and 3.9 at a specified scenario, namely 30% of censored observations. Considering $\widetilde{p} \pm se(\widetilde{p})$ we can see that:

i) $B^{MI}_{\cdot}$ provides the best results along all the true models;

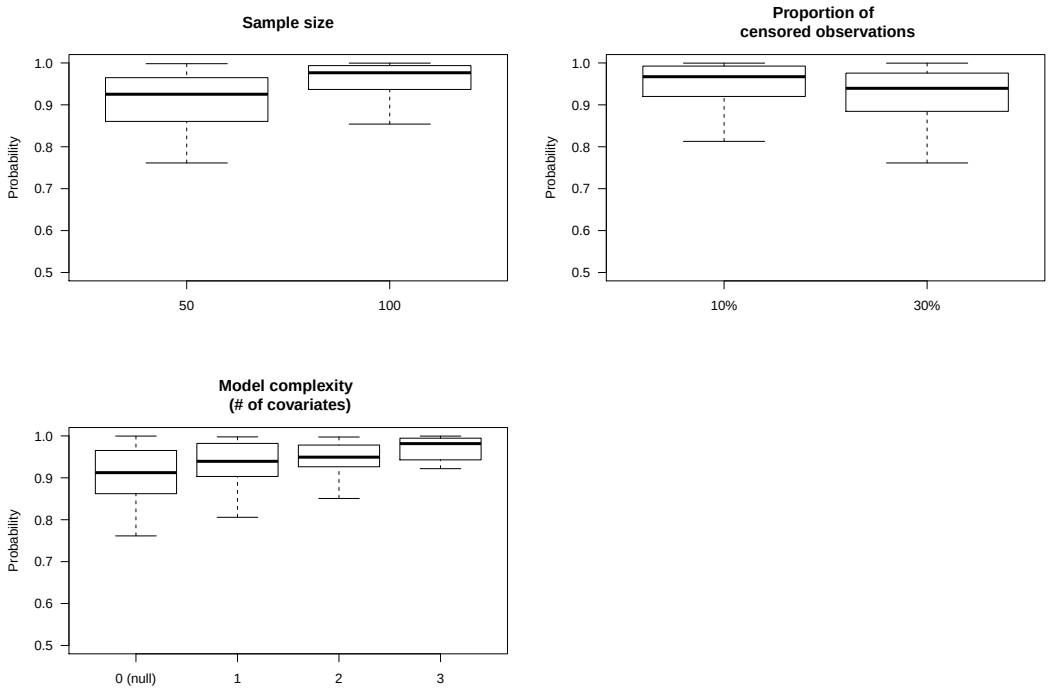

Figure 3.4: Conditional distributions of the acceptance proportion of the true Weibull model for the different simulated scenarios and marginally to the scenarios not mentioned in the corresponding Box-Plot. Values are based on all versions of BFs as well as all model selection strategies.

ii) BIC behaves similarly to the other BFs;

iii) $B_{\cdot}^{AI}$ has a worse behavior compared to $B_{\cdot}^{MI}$, $FBF_{\cdot}$ and $mFBF$ and, in the case of the log-normal regression, behaves worse than the BIC. This behavior is due to the instability of this measure, and it has also been noted in Berger et al. (2001) that the AIBF is less robust than the MIBF;

iv) $mFBF$ behaves similarly to $FBF_{\cdot}$. Both have similar results to $B_{\cdot}^{MI}$ for $n = 100$, while for $n = 50$, and the Weibull model, the acceptance proportion of the true model using $mFBF$ and FBF_{mo} decreases a little for the null and the full model. For the log-normal model this occurs in the null and 1 covariate scenario. Globally, $mFBF$ and $FBF_{\cdot}$ are the second best tools to select the correct model in the simulated scenarios, as shown in Figures 3.6 and 3.7.

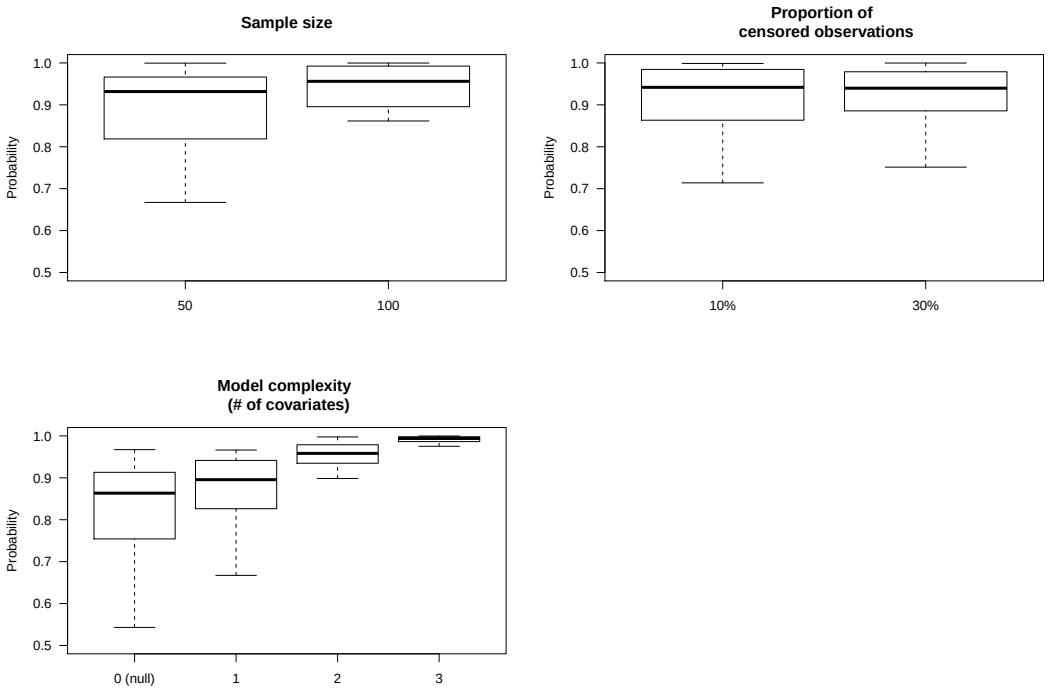

Figure 3.5: Conditional distributions of the acceptance proportion of the true log-normal model for the different simulated scenarios and marginally to the scenarios not mentioned in the corresponding Box-Plot. Values are based on all versions of BFs as well as all model selection strategies.

Finally, we have calculated the posterior expected model size for each BF and for each simulation scenario. In Figures 3.12, 3.13, 3.14 and 3.15 the boxplots of the posterior expected model size for the 100 replications for the Weibull models are represented, for $n = 50$ and $n = 100$ and for the two censoring percentages, 10% and 30 %. Analogous plots are presented in Figures 3.16, 3.17, 3.18 and 3.19 for the log-normal models. All these figures confirm the results of the previous dotcharts, in particular we can observe that the $B^{MI}_{L_{mo}}$ and the $B^{MI}_{L_{me}}$ are more precise in estimating the true model size, specially in the case of the null model. Also it can be seen that in the case of the Weibull model, BIC selects models with a model size, in mean, greater than the rest of the tools, while for the log-normal scenario it works similarly to the rest.

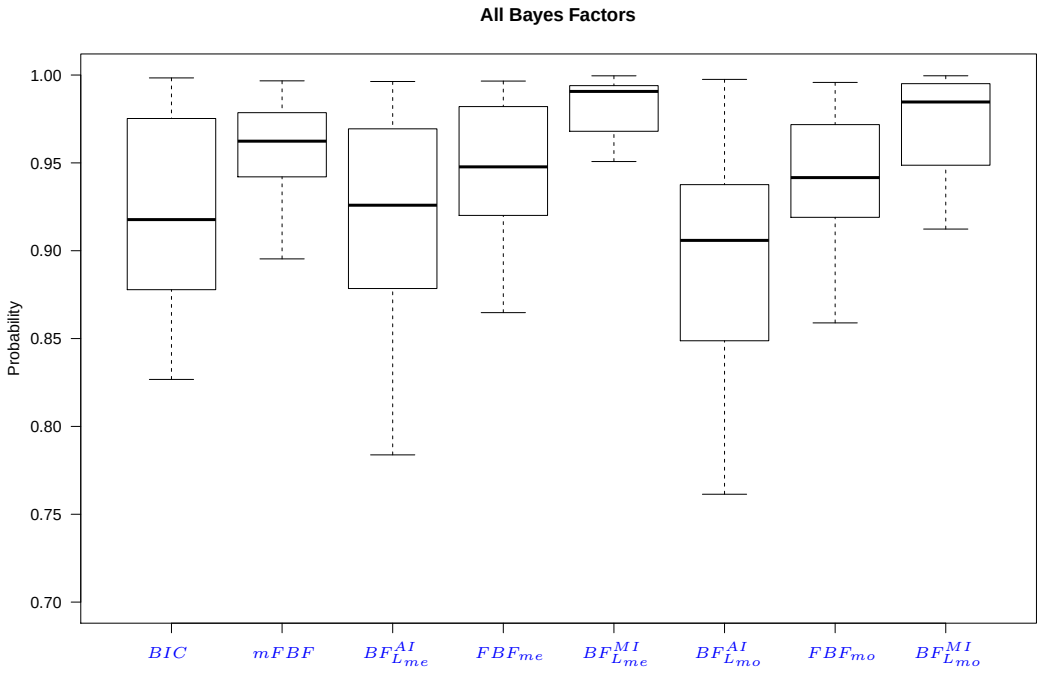

Figure 3.6: Conditional distributions of the acceptance proportion of the true Weibull model for the 8 different tools.

3.6.1 Computational cost

In this chapter we have considered the calculation of IBFs using SMTS, but these quantities are very computational demanding.

Observe that for the calculation of the IBF it is necessary to approximate $2(L_* + 1)$ integrals, where L_* stays for L_{mode} or L_{median}. While in order to obtain the mFBF it is necessary to approximate $2(n_{cens} + 2)$ integrals. The computational cost and the elaboration time can be compared in terms of the number of integrals needed to calculate the IBF and the mFBF when, for instance, $L_* = L_{mode}$. Figure 3.20 illustrates the difference in the number of integrals (we use the logarithmic scale for simplicity) to be approximated for different sample sizes, with $s = 5$ and 30% of censored observations. As it can be observed, IBF is much more expensive to compute than mFBF, and this cost grows quickly as n increases, being 1202 integrals for IBF and $n = 100$, while this number is 64 for mFBF and for the same sample size.

The calculation of the BIC is quite immediate but, as it can be seen, it has a worse behavior than the other BFs when the Laplace approximation does not perform very

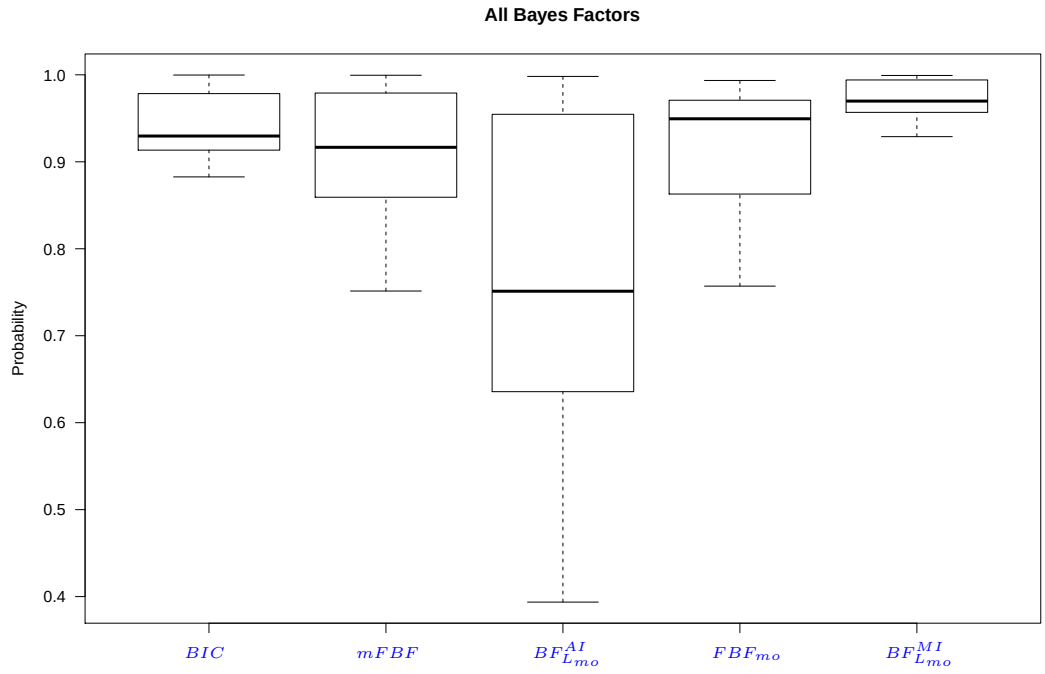

Figure 3.7: Conditional distributions of the acceptance proportion of the true log-normal model for the 5 different tools.

well, as in the Weibull case.

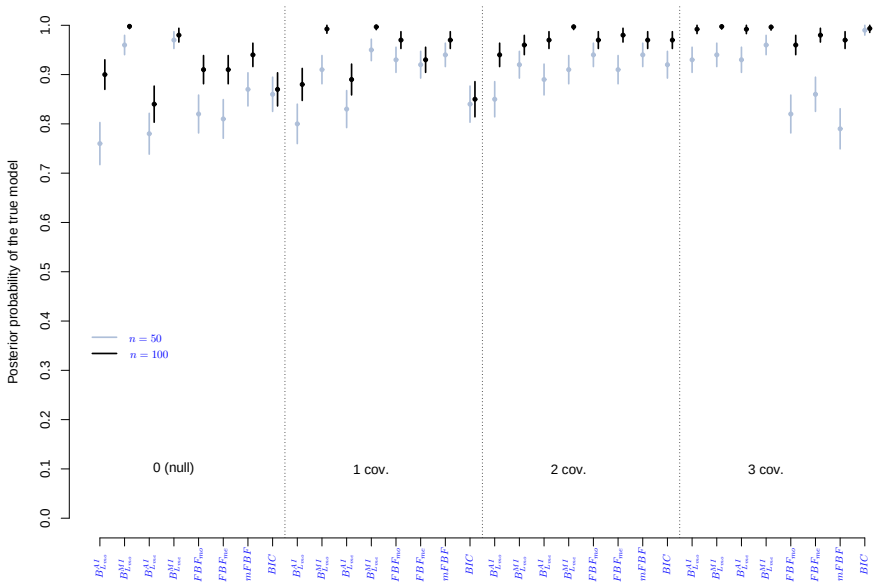

Figure 3.8: Values of $\widetilde{p} \pm se(\widetilde{p})$ for Weibull model, different BFs with: 30% of censored data and two sample sizes.

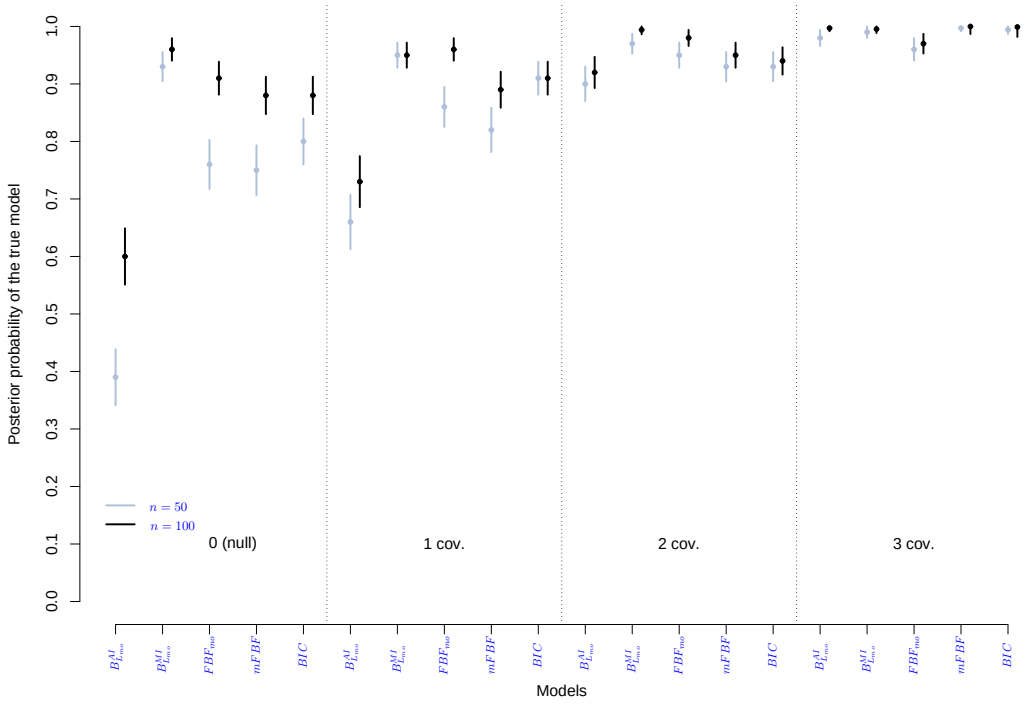

Figure 3.9: Values of $\widetilde{p} \pm se(\widetilde{p})$ for log-normal model, different BFs with: 30% of censored data and two sample sizes.

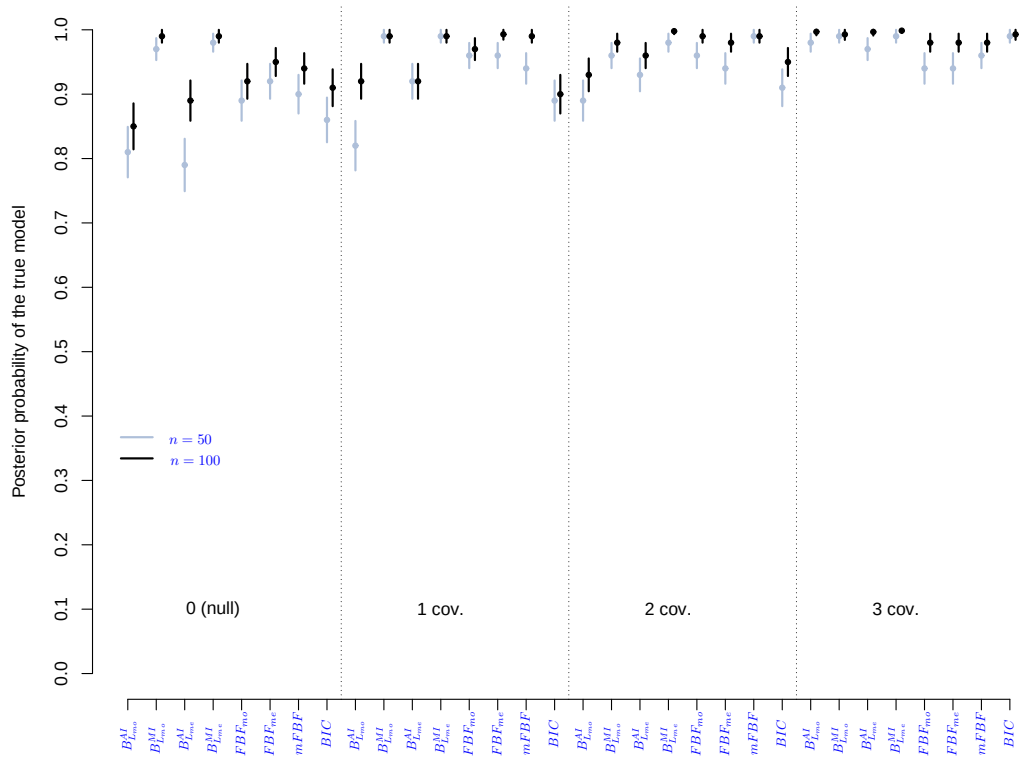

Figure 3.10: Values of $\tilde{p} \pm se(\tilde{p})$ for Weibull model, different BFs with: 10% of censored data and two sample sizes.

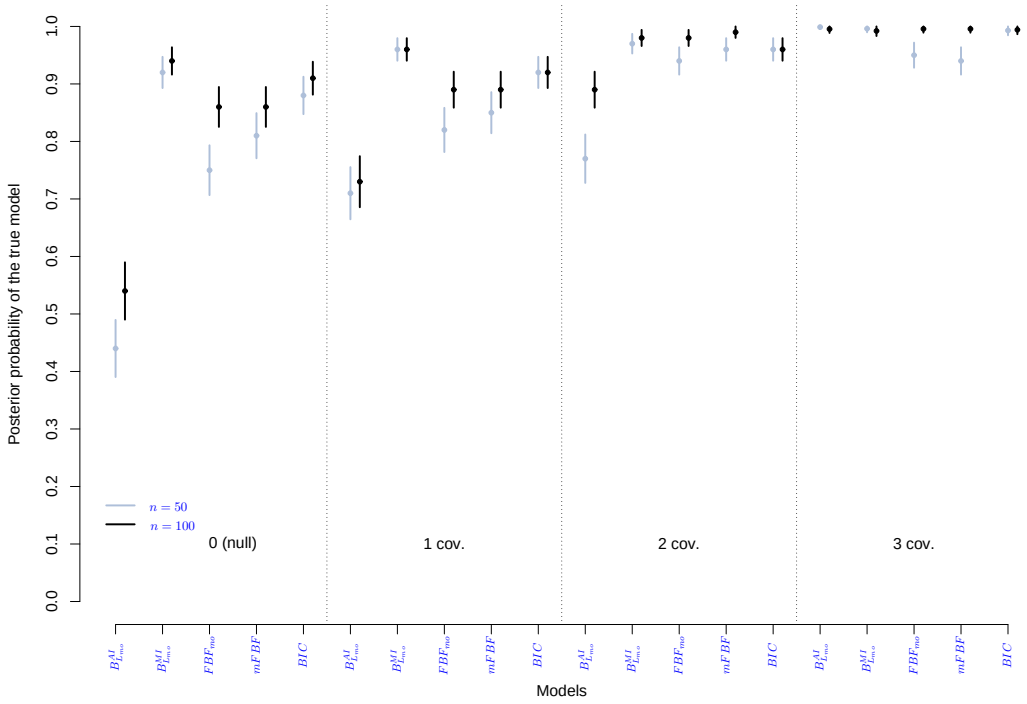

Figure 3.11: Values of $\tilde{p} \pm se(\tilde{p})$ for log-normal model, different BFs with: 10% of censored data and two sample sizes.

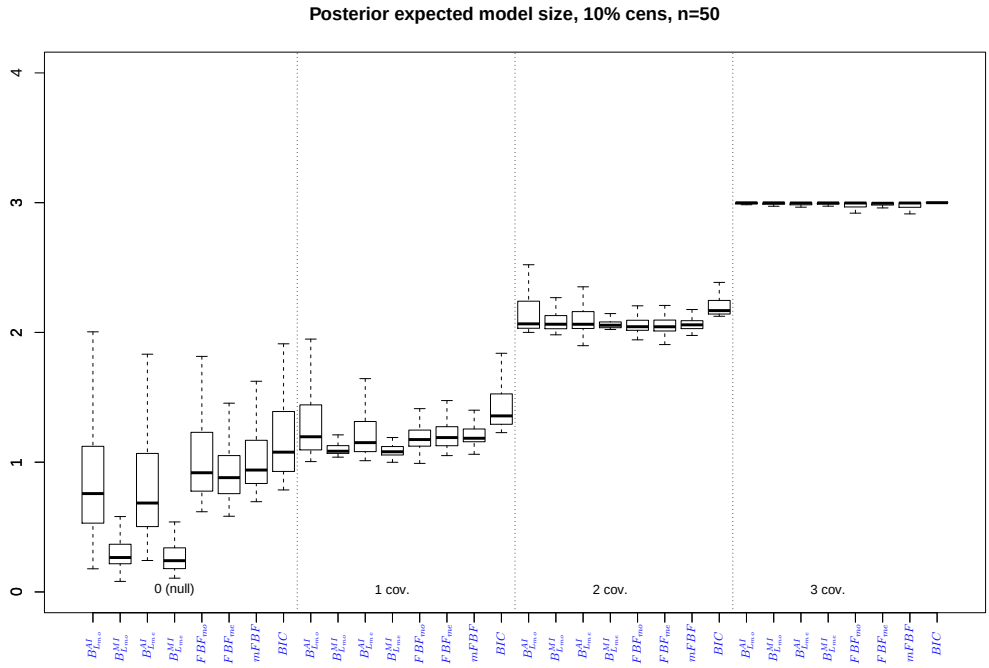

Figure 3.12: Distribution of the posterior expected model size for the Weibull model, different BFs with: 10% of censored data and $n = 50$.

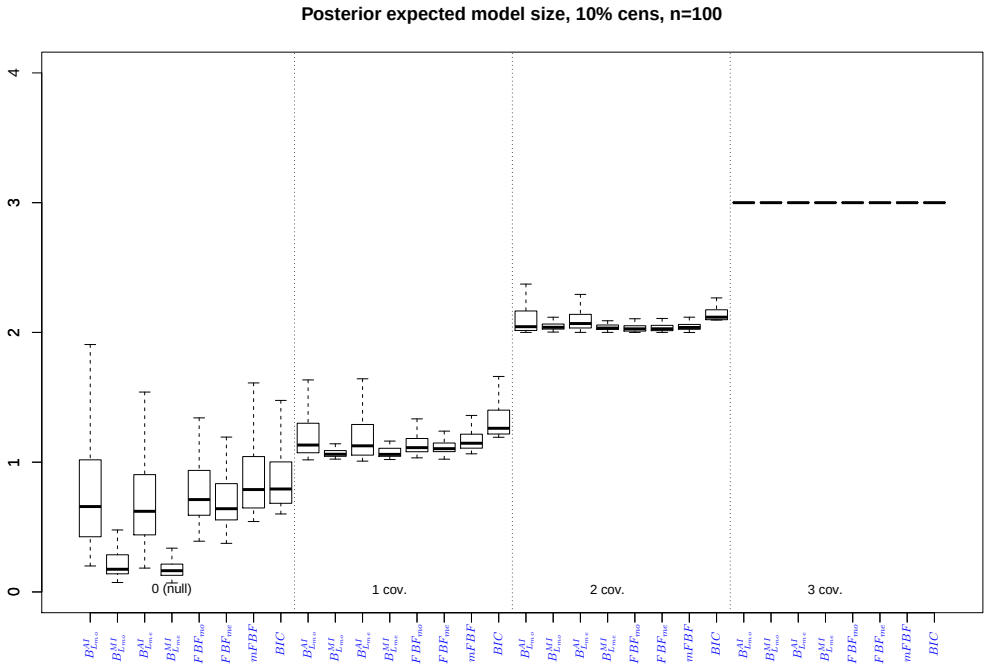

Figure 3.13: Distribution of the posterior expected model size for the Weibull model, different BFs with: 10% of censored data and $n = 100$.

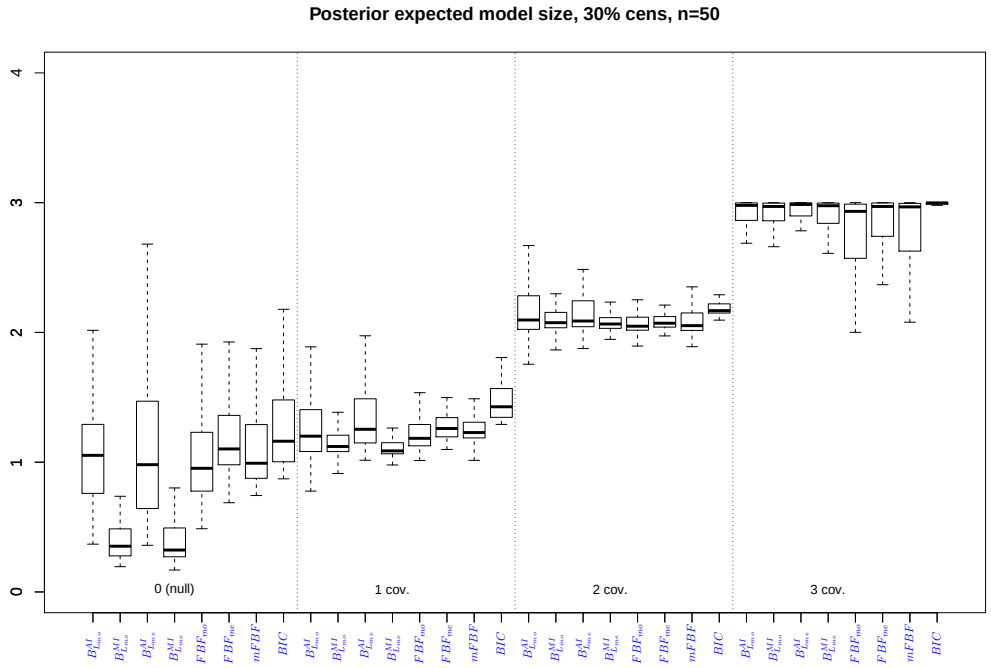

Figure 3.14: Distribution of the posterior expected model size for the Weibull model, different BFs with: 30% of censored data and $n = 50$.

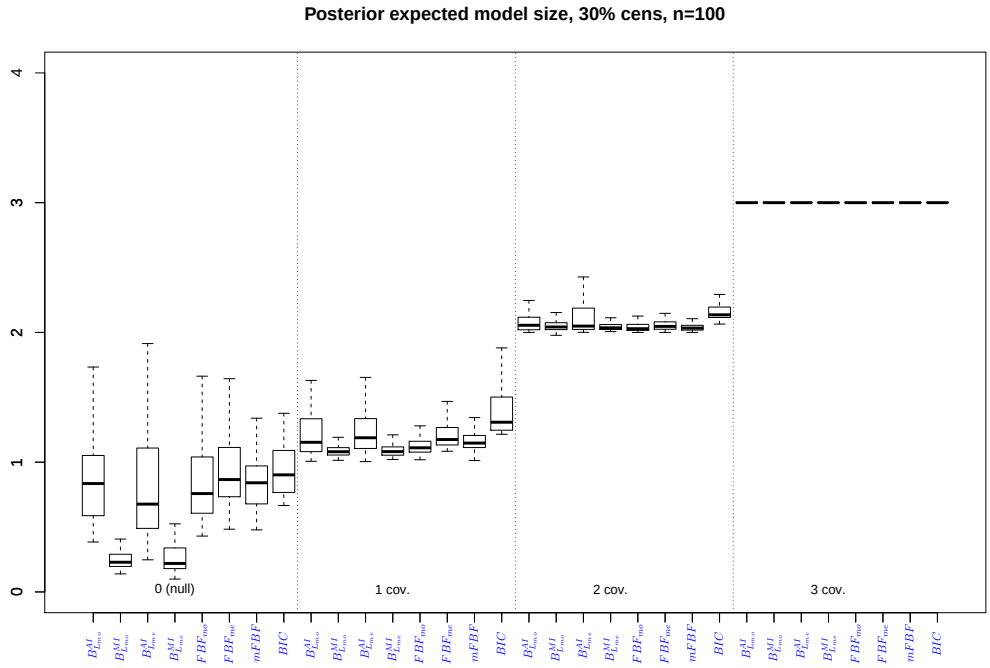

Figure 3.15: Distribution of the posterior expected model size for the Weibull model, different BFs with: 30% of censored data and $n = 100$.

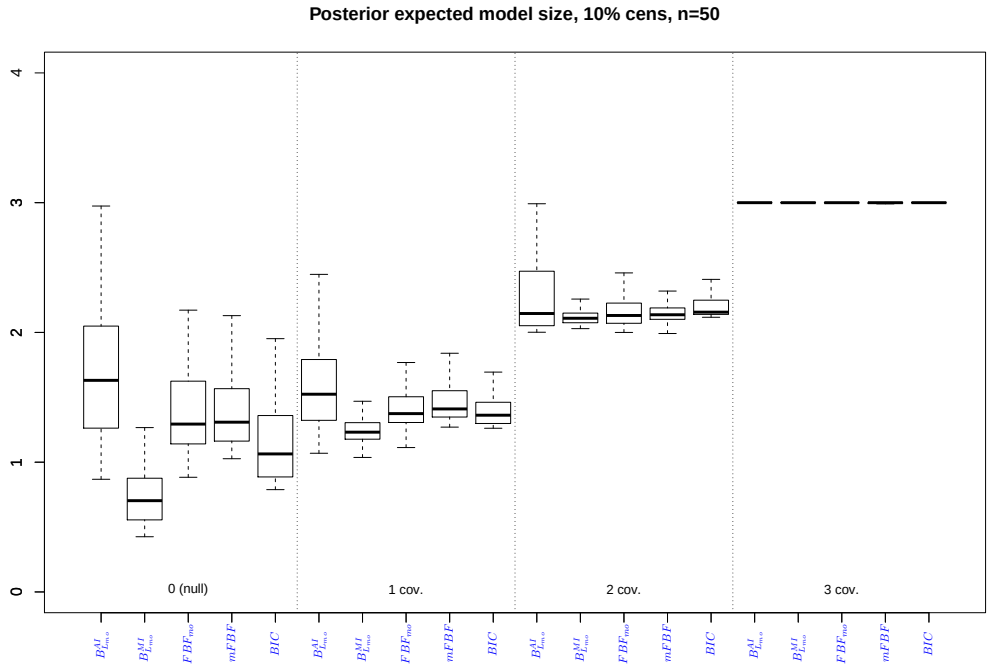

Figure 3.16: Distribution of the posterior expected model size for the log-normal model, different BFs with: 10% of censored data and $n = 50$.

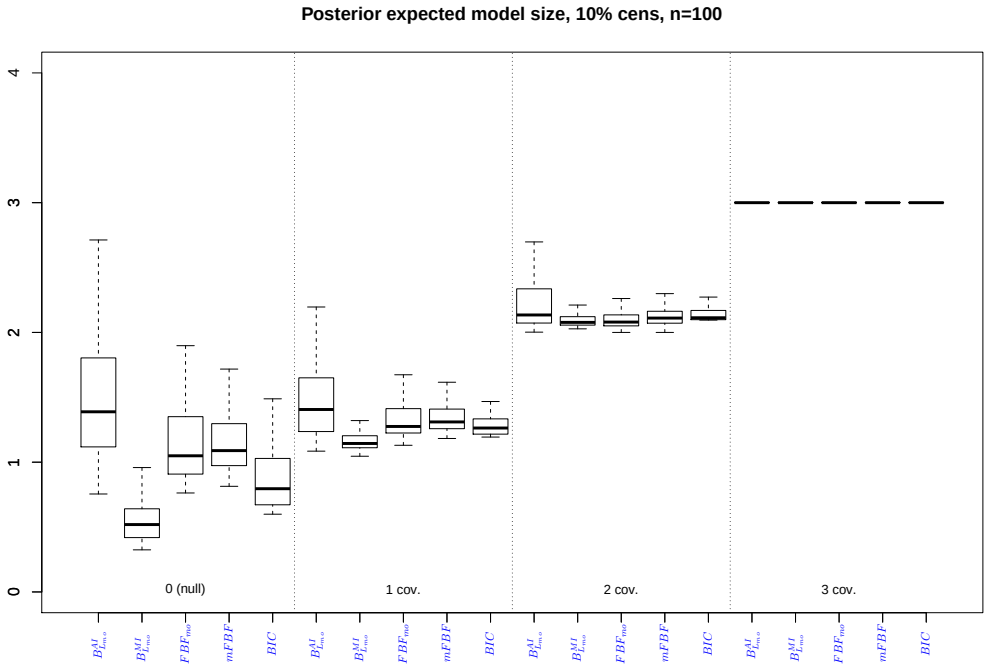

Figure 3.17: Distribution of the posterior expected model size for the log-normal model, different BFs with: 10% of censored data and $n = 100$.

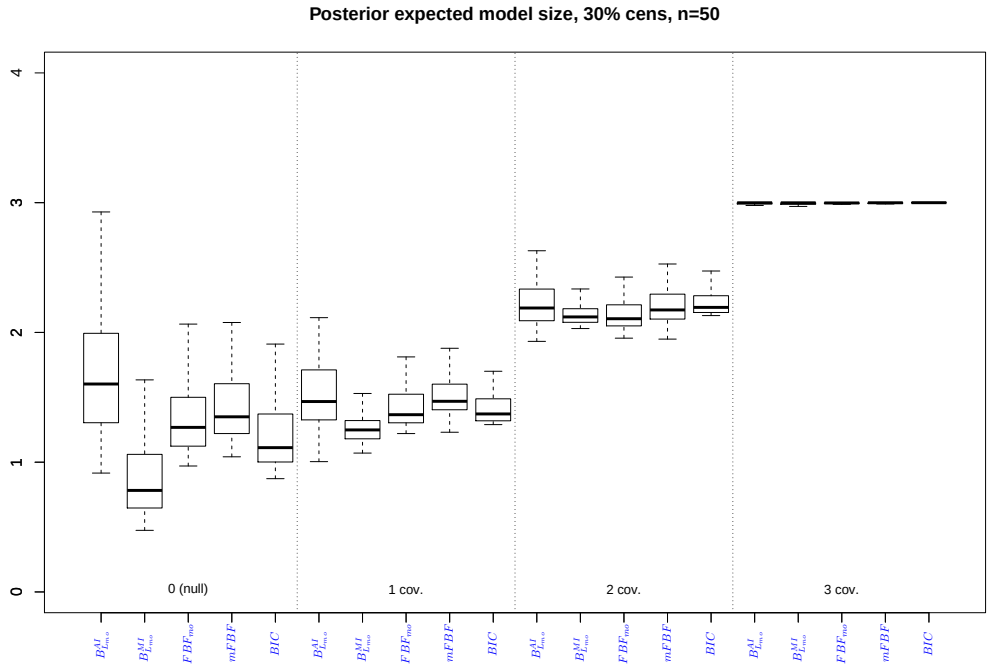

Figure 3.18: Distribution of the posterior expected model size for the log-normal model, different BFs with: 30% of censored data and $n = 50$.

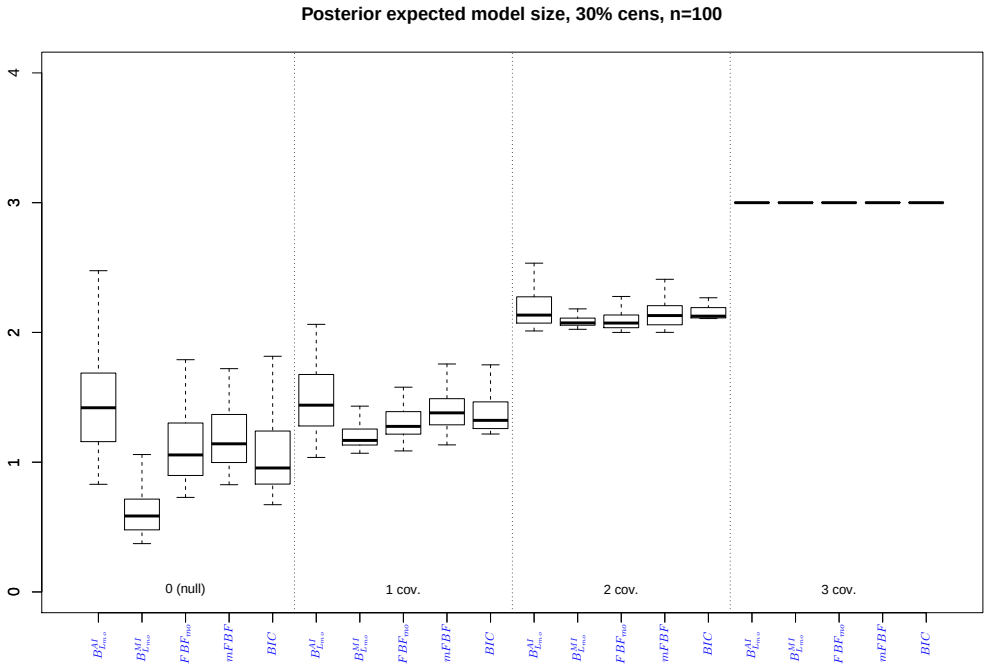

Figure 3.19: Distribution of the posterior expected model size for the log-normal model, different BFs with: 30% of censored data and $n = 100$.

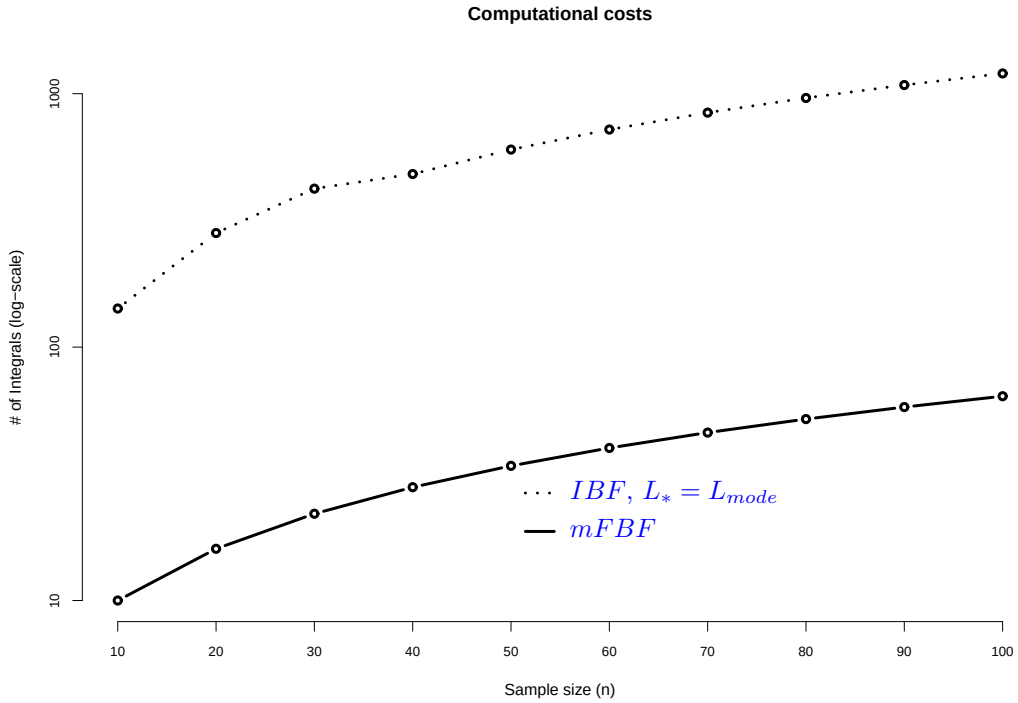

Figure 3.20: Number of integrals (vertical axis log-scale) to be approximated for the calculation of the mFBF and the IBF with $L_* = L_{mode}$ as a function of the sample size n (horizontal axis) for $s = 5$ and 30% of censored observations.

4

Construction of Minimal Training Samples under censoring using the Kaplan-Meier estimator

4.1 Introduction

In the previous Chapter we have introduced a method to obtain IBFs and FBFs based on the SMTS scheme in presence of censoring. As it has been analyzed in Subsection 3.6.1, the computational cost to approximate IBFs based on SMTSs is very high resulting in very long computation times, especially for large datasets or problems with a moderate or large number of involved variables. Along the present Chapter we will discuss a different approach to define MTSs in presence of censored data. This new strategy is very useful when it is possible to obtain closed-form expressions for the predictive distributions when samples do not contain censored data. In particular this is true for the log-normal model as it is shown in Subsection 1.5.1.4.
The new approach may be viewed as a reweighting of the usual MTS extraction mechanism, in order to verify the Assumption 0 introduced in Chapter 3, when working with censored data. In particular, we introduce a class of training samples, defined by Berger and Pericchi (2004), useful when the information in each observation of the TS is different and when Assumption 0 is violated.

Definition 6. (Randomized training sample) *A randomized training sample with sampling mechanism* $\boldsymbol{u} = (u_1, \ldots, u_{L_U})$*, where* $\boldsymbol{u}$ *is a probability vector, is obtained by drawing a training sample from* $\mathcal{X}^U$*, the space of all training samples, according to* $\boldsymbol{u}$*. In this context, the training samples can be considered to be weighted training samples*

with weights u_i.

The SMTS scheme introduced in Chapter 3 can be viewed as a particular method to construct randomized training samples, where the probabilities u_i are the probability of drawing the *i-th* SMTS from all the possible SMTSs when sampling without replacement from the data. Our proposal is to obtain weights in the randomized training samples through the reweighting of observations in the sample via a nonparametric estimator of the distribution function under the null model. In particular, we have used the Kaplan-Meier estimator.

4.1.1 MTS based on the Kaplan-Meier estimator

Let O_i be the number of individuals that are still alive at time t_i or experience the event of interest at t_i and d_i be the number of events that occur at that time. The quantity d_i/O_i is an empirical estimate of the conditional probability that an individual, who survives just prior to t_i, experiences the event at time t_i. This is the base from which the estimation of the survival function is constructed. Observe that the empirical version of the survival function $S(t)$ is

$$\hat{S}(t) = \frac{\text{number of individuals surviving longer than } t}{\text{total number of individuals under study}}.$$

Definition 7. (Kaplan-Meier estimator) *The Kaplan-Meier estimator, also known as the* product-limit estimator*, was introduced by Kaplan and Meier (1958). It is an estimator of the survival function having the following form*

$$\hat{S}_{KM}(t) = \prod_{t_i \leqslant t} \left(\frac{O_i - d_i}{O_i} \right) = \prod_{t_i \leqslant t} \left(1 - \frac{d_i}{O_i} \right)$$

where $(1 - \frac{d_i}{O_i})$ *is the conditional probability that an individual survives at the end of a time interval, under the condition that the individual was present at the start of the time interval.*

Note that $\hat{S}_{KM}(t)$ is not well defined for values of t beyond the largest observation, in fact if the largest study time corresponds to a death, then the estimated survival curve is zero after this time. If the largest study time is censored, the survival $\hat{S}_{KM}(t)$ beyond this point will be undetermined because we do not know what would have been the time to the death if the survivor had not been censored. In order to avoid this problem, we adopt the convention proposed by Efron (1967). It consists in fixing the value of the Kaplan-Meier estimator, $\hat{S}_{KM}(t)$, equal to 0 beyond the largest study time. This means that the survivor with the largest time on study has died immediately after

the survivor's censoring time.
Another proposal in Gill (1980) is to define $\hat{S}(t)$ as $\hat{S}(t_{max})$ for $t > t_{max}$, this corresponds to assume that this individual would die at infinity, and it leads to an estimator which is positively biased. Both techniques, the one of Efron (1967) and that of Gill (1980), correspond to the two most radical situations that can be found. Both estimators have the same large-sample properties and converge to the true survival function for large samples. Other works as Brown et al. (1974) or Moeschberger and Klein (1985) use parametric models as the exponential or the Weibull distributions to estimate the tail of $S(t)$.
The Kaplan-Meier estimator is based on an assumption of non-informative censoring, this means that knowledge about a censoring time for an individual does not provide further information about this person's likelihood of survival. This means, for example, that censoring times do not depend on covariates. When this assumption can be violated, $\hat{F}_{KM}$ estimates the wrong distribution function. When there are suspects that censoring could depend on some covariates in the study, Kaplan-Meier estimators conditional to these covariates can be considered instead of the proposal used here. The rest of calculations shown here are not affected for the estimator used to construct the MTS, if the resulting MTS is formed by uncensored observations.
Once it is obtained the estimator of the survival function $\hat{S}_{KM}(t) = 1 - \hat{F}_{KM}(t)$, the estimation of the cumulative distribution function $\hat{F}_{KM}(t)$ can be defined.

Definition 8. (KMMTS) *A Kaplan-Meier minimal training sample (KMMTS) is a training sample obtained by sampling without replacement s (the number of parameters in a given model) observations from the observed data according to the following probability mass function*

$$\begin{aligned}\hat{f}_{KM}(t) &= \hat{F}_{KM}(t_i) - \hat{F}_{KM}(t_{i-1}) \\ &= \begin{cases} \hat{F}_{KM}(t_1) & \text{if } t \leqslant t_1 \\ \hat{F}_{KM}(t_i) - \hat{F}_{KM}(t_{i-1}) & \text{if } t_{i-1} < t \leqslant t_i,\ i = 2, \ldots, n-1 \\ 1 - \hat{F}_{KM}(t_{n-1}) & \text{if } t_{n-1} < t \leqslant t_n. \end{cases}\end{aligned} \tag{4.1}$$

The Kaplan-Meier estimator of F results in a step function in which the mass function is defined only in values corresponding to uncensored observations, while the mass function estimated via the $\hat{F}_{KM}$ in a censored observation is 0. As a consequence of this definition, a KMMTS contains only uncensored observations.

Example 12. (Calculation of the Kaplan-Meier estimator) *We consider a simulated dataset based on the 6-mercaptopurine (6-MP) dataset introduced by Freireich et al. (1963). It consists in results from a clinical trial of the drug 6-MP versus a placebo in patients suffering from acute leukemia. Data about the survival times for the treatment group jointly with the censoring indicator are reported in Table 4.1.*

Lifetime (months)	6	6	6	6	7	9	10	10	11	13	16	17	19	20	22	23	25
Censoring	1	1	1	0	1	0	1	0	0	1	1	0	0	0	1	1	0

Table 4.1: Simulated data from the treatment group in the 6-MP dataset.

For the calculation of the Kaplan-Meier estimator we only consider the time to relapse and the corresponding censoring indicator. By applying Definition 8 we obtain the results shown in Table 4.2.

Time t_i	Number of events d_i	Number at risk O_i	Kaplan-Meier estimator $\hat{S}_{KM}(t) = \prod_{t_i \leqslant t} \left(1 - \frac{d_i}{O_i}\right)$
6	3	17	$\left(1 - \frac{3}{17}\right) = 0.824$
7	1	13	$0.824 \left(1 - \frac{1}{13}\right) = 0.760$
10	1	11	$0.760 \left(1 - \frac{1}{11}\right) = 0.691$
13	1	8	$0.691 \left(1 - \frac{1}{8}\right) = 0.605$
16	1	7	$0.605 \left(1 - \frac{1}{7}\right) = 0.518$
22	1	3	$0.518 \left(1 - \frac{1}{3}\right) = 0.346$
23	1	2	$0.346 \left(1 - \frac{1}{2}\right) = 0.173$

Table 4.2: Kaplan-Meier estimator for the simulated dataset.

In Figure 4.1 it is represented the Kaplan-Meier survival curve along with its 95% confidence interval.

Suppose we want to extract two different KMMTSs of length 3. We observe that each KMMTS has a different probability of being sampled. For example, the two KMMTSs consisting in lifetimes $\{16, 7, 22\}$ *and* $\{6, 16, 22\}$ *have probabilities:*

$$\Pr(\{16, 7, 22\}) = \frac{1}{7} \cdot 0.087 + \frac{1}{6} \cdot 0.064 + \frac{1}{5} \cdot 0.172 = 0.057$$
$$\Pr(\{6, 16, 22\}) = \frac{1}{7} \cdot 0.176 + \frac{1}{6} \cdot 0.087 + \frac{1}{5} \cdot 0.172 = 0.074.$$

In Algorithm 3 it is presented the pseudo-code to obtain a KMMTS.

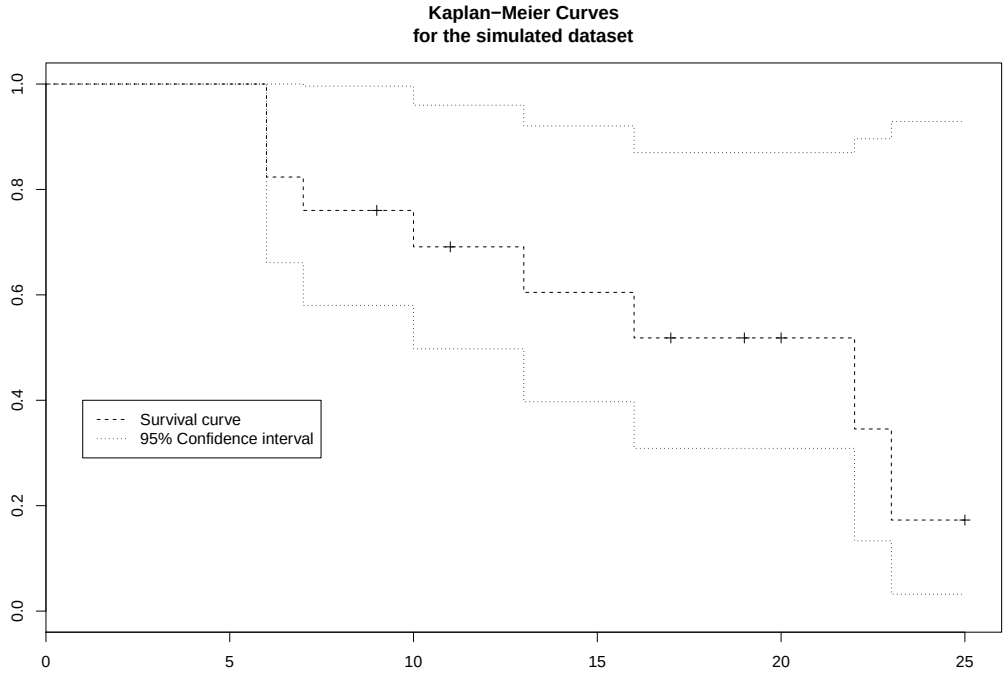

Figure 4.1: Kaplan-Meier survival curve and 95% confidence interval for the 6-MP dataset.

As seen in Section 1.5.1, the marginal distribution for the log-normal model in the case of uncensored data is expressed in closed-form. For this reason, the definition of the KMMTS results a suitable choice, in fact it allows a simpler expression of the BF and, indeed, faster computations.

Example 13. (Example 5 continued) *For the right censored exponential model, as the censoring time ρ is fixed, the Assumption 0 introduced in Section 3.1 is not verified using the KMMTS. So, in this example, one must use the strategy introduced in Section 3.1. But, if a random censoring time, ρ, is considered, results in Efron (1967) assure that $\hat{S}_{KM}$ converges to the true survival function, so $\hat{F}_{KM}$ converges to F. This means that with probability 1 all the possible samples of one uncensored observation distributed accordingly to F are recovered simulating from $\hat{F}_{KM}$.*

4.2 IBF based on the KMMTS

The AIBF and MIBF for randomized training samples are defined as:

Algorithm 3 Kaplan-Meier Minimal Training Sample.

Require: D, the data ordered by increasing lifetime;
s, the number of parameters of the most complicated model

1: Calculate the Kaplan-Meier estimator $\hat{S}_{KM}$ for each individual in D;
2: Create a vector of cumulated probabilities $\hat{F}_{KM} = 1 - \hat{S}_{KM}$;
3: Calculate the vector $\hat{f}_{KM}$ of point mass at each observation from $\hat{F}_{KM}$;
4: Sample s observations, without replacement, from D with the corresponding probabilities $\hat{f}_{KM}$;
5: **return** the Kaplan-Meier training sample.

$$BF_{ij}^{AI} = B_{ij}^{N}(\boldsymbol{y}) \sum_{l=1}^{L_U} u_l B_{ji}^{N}(\boldsymbol{y}(l))$$
$$BF_{ij}^{MI} = B_{ij}^{N}(\boldsymbol{y}) \underset{u_1,\ldots,u_{L_U}}{\text{Median}} B_{ji}^{N}(\boldsymbol{y}(l)),$$

where the last expression means that the median of $B_{ji}^{N}(\boldsymbol{y}(l))$ is calculated with respect to the probability distribution of the training samples $(u_1, \ldots, u_{L_U})$.
The calculation of all possible training samples jointly with their weights is in almost all the cases prohibitive, because of the large number of training samples. For this reason, these theoretical quantities are approximated using L draws of training samples and calculating:

$$\begin{aligned} BF_{ij}^{AI} &= B_{ij}^{N}(\boldsymbol{y}) \sum_{l=1}^{L} B_{ji}^{N}(\boldsymbol{y}(l)) \\ BF_{ij}^{MI} &= B_{ij}^{N}(\boldsymbol{y}) \underset{1,\ldots,L}{\text{Median}} B_{ji}^{N}(\boldsymbol{y}(l)). \end{aligned} \tag{4.2}$$

In particular, suppose we want to compare the following two log-normal models, where Y_i is the logarithm of the lifetime of the *i-th* subject

$$M_0 : Y_i = \mu + \sigma W_i$$
$$M_i : Y_i = \mu + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{x}_i + \sigma W_i$$

where W_i are standard normal independent error terms. The two models can be written in the following form, as seen in Subsection 1.5.1.3

$$M_0 : \boldsymbol{Y} = \boldsymbol{Z}_0 \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_0$$
$$M_i : \boldsymbol{Y} = \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i.$$

The corresponding Kaplan-Meier AIBF (KMAIBF) and Kaplan-Meier MIBF (KMMIBF) can be obtained from (4.2).
In this case we choose $L = n \times s$, where s is the number of parameters of the full model under study.
In particular, we have to calculate the $B_{0i}^N(\boldsymbol{y}(l))$ over the KMMTS, which does not contain censored data and it can be obtained in closed-form, while $B_{i0}^N(\boldsymbol{y})$ is calculated over the full data and must be approximated as there is not a closed-form expression for it. In order to approximate the last quantity, it is necessary to approximate the corresponding marginal distributions, as described in Subsection 1.5.1.3 and in Subsection 1.5.1.4.
Let r_0 and r_i be the ranks of the design matrices $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_i$, respectively. For simplicity, we denote by $\boldsymbol{Z}_0(l)$ and $\boldsymbol{Z}_i(l)$ the covariate matrices obtained by taking the rows corresponding to the KMMTSs' observations and the columns corresponding to models M_0 and M_i, respectively, from the full matrix $\boldsymbol{Z}$. Using the expression of the predictive distribution given in (1.5), we have

$$B_{0i}^N(\boldsymbol{y}(l)) = \frac{\Gamma(\frac{s-r_0}{2})\Gamma(\frac{1}{2})^{r_0} \mid \boldsymbol{Z}_i(l)^\mathsf{T}\boldsymbol{Z}_i(l) \mid^{1/2} \left[(\boldsymbol{y}(l) - \hat{\boldsymbol{y}}_i(l))^\mathsf{T}(\boldsymbol{y}(l) - \hat{\boldsymbol{y}}_i(l))\right]^{(s-r_i)/2}}{\Gamma(\frac{1}{2})^{r_i}\Gamma(\frac{s-r_i}{2}) \mid \boldsymbol{Z}_0(l)^\mathsf{T}\boldsymbol{Z}_0(l) \mid^{1/2} \left[(\boldsymbol{y}(l) - \hat{\boldsymbol{y}}_0(l))^\mathsf{T}(\boldsymbol{y}(l) - \hat{\boldsymbol{y}}_0(l))\right]^{(s-r_0)/2}},$$

with

$$\hat{\boldsymbol{y}}_i(l) = \boldsymbol{Z}_i(l)\hat{\boldsymbol{\beta}}_i, \text{ where } \hat{\boldsymbol{\beta}}_i = (\boldsymbol{Z}_i(l)^\mathsf{T}\boldsymbol{Z}_i(l))^{-1}\boldsymbol{Z}_i(l)^\mathsf{T}\boldsymbol{y}(l)$$
$$\hat{\boldsymbol{y}}_0(l) = \boldsymbol{Z}_0(l)\hat{\boldsymbol{\beta}}_0, \text{ where } \hat{\boldsymbol{\beta}}_0 = (\boldsymbol{Z}_0(l)^\mathsf{T}\boldsymbol{Z}_0(l))^{-1}\boldsymbol{Z}_0(l)^\mathsf{T}\boldsymbol{y}(l).$$

Example 14. (Example 1 continued) *We now present the results of the calculation of the B^{AI} and B^{MI} for the* larynx *cancer dataset presented in Example 1.*

We compute the KMAIBF and the KMMIBF and we compare them with the AIBF, MIBF, FBF (all of them calculated only over L_{mode}, because the mode and the median of N_t are equal), mFBF, all using the SMTS strategy, and the BIC for this dataset doing a pairwise comparison from below. The possible models, containing only additive effects, are listed in Example 1.

In particular, we adopt a log-normal regression model. Next we use the two strategies, HPPM and MPPM, to select among the models and in Table 4.3 we report the values for HPPM only, since they are not different from the MPPM ones.

As we can see B_{KM}^{MI} is close to B_{Lmo}^{MI} and agrees with the rest of BFs in choosing the model containing `stage` *as the most probable one. The behavior of B_{Lmo}^{AI} is not desirable because it concentrates all the probability in one model. As we will see in Chapter 5*

k	Model	B^{AI}_{KM}	B^{MI}_{KM}	$B^{AI}_{L_{mo}}$	$B^{MI}_{L_{mo}}$	FBF_{mo}	$mFBF$	BIC
0	`Null`	0.001	0.001	0.000	0.002	0.001	0.001	0.002
1	`stage`	0.599	0.756	1.000	0.813	0.692	0.663	0.949
2	`age`	0.001	0.001	0.000	0.001	0.001	0.001	0.001
3	`age+stage`	0.399	0.242	0.000	0.184	0.306	0.335	0.048

Table 4.3: Comparison of the posterior probabilities of the different BFs for the Larynx dataset.

this is very common and it is due to the instability of this measure. This behavior is mitigated when using KMMTS to calculate the AIBF.

4.3 Zellner and Siow prior for the log-normal model

In this section we consider the use of a conventional prior. As we are taking into account the log-normal distribution, which corresponds to normal data when working with the logarithm, there is a great consensus in the conventional prior to be used. For the simulation study we present below, we have used the one introduced in Subsection 2.3.1. Following the definition of effective sample size used in Section 3.5, we use n_u, that is the number of uncensored observations, instead of n in the definition of the prior

$$\pi^{ZS}(\boldsymbol{\gamma}_k|\mu_0,\sigma_0) = Ca_{r_k}(\boldsymbol{\gamma}_k|0, n_u\sigma_0^2({\boldsymbol{V}_k}^{\mathsf{T}}\boldsymbol{V}_k)^{-1}), \tag{4.3}$$

which is a multivariate Cauchy distribution, where $\tilde{\boldsymbol{X}}_k$ is the design matrix corresponding to the vector $\boldsymbol{\gamma}_k$, without including the intercept, $r_k = rank(\tilde{\boldsymbol{X}}_k)$ and $\boldsymbol{V}_k = (I_n - P_0)\tilde{\boldsymbol{X}}_k$ is the design matrix corresponding to the orthogonal parametrization, where $P_0 = \boldsymbol{X}_0({\boldsymbol{X}_0}^{\mathsf{T}}\boldsymbol{X}_0)^{-1}{\boldsymbol{X}_0}^{\mathsf{T}}$ and $\boldsymbol{X}_0 = (1,\ldots,1)^{\mathsf{T}}$ is a vector of length n.
The corresponding posterior distribution is

$$\begin{aligned}
\pi(\mu_0,\boldsymbol{\gamma}_k,\sigma_0 \mid \boldsymbol{y},\tilde{\boldsymbol{X}}_k) &\propto \pi(\mu_0,\boldsymbol{\gamma}_k,\sigma_0)L(\mu_0,\boldsymbol{\gamma}_k,\sigma_0 \mid \boldsymbol{y},\tilde{\boldsymbol{X}}_k)\\
&\propto Ca_{r_k}(\boldsymbol{\gamma}_k|0, n_u\sigma_0^2({\boldsymbol{V}_k}^{\mathsf{T}}\boldsymbol{V}_k)^{-1})\times\\
&\times \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2}\left(\frac{y_i-(\mu_0+{\boldsymbol{\gamma}_k}^{\mathsf{T}}\tilde{\boldsymbol{x}}_{k,i})}{\sigma_0}\right)^2\right)\right]^{\delta_i}\times\\
&\times\left[1-\Phi\left(\frac{y_i-(\mu_0+{\boldsymbol{\gamma}_k}^{\mathsf{T}}\tilde{\boldsymbol{x}}_{k,i})}{\sigma_0}\right)\right]^{(1-\delta_i)}
\end{aligned}$$

which does not have a closed-form, due to the presence of censoring. In order to calculate the BFs, the corresponding marginal distribution has been approximated using the algorithm of Chib and Jeliazkov (2001) as in Chapter 3.

4.4 Simulation Study

In this section we present results of a simulation study in order to compare the performances of the IBFs calculated over the KMMTS with the rest of the tools introduced in Chapter 3. In particular we have used: the IBFs calculated over the SMTS, the mFBF, the BIC, the BF calculated over the Zellner and Siow prior defined in (4.3) and we have calculated the FBF_{mo} and the FBF_{me} and here we only consider the FBF_{mo} because there are no significative differences between the two tools. The goal is to show that the IBFs calculated over the KMMTS work not worse than the mFBF and the IBFs calculated over the SMTS.
Data have been simulated from a log-normal distribution, as we work with the logarithm of the times this means to simulate data from a normal regression model. The censoring indicator has been simulated as described in Appendix B, considering two different censoring percentages, 10% and 30%.

The log-normal regression model from which data are simulated takes the form

$$Y_i = \log(T_i) = \mu + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \sigma W_i \qquad i = 1, \ldots, n$$

where $W_i \sim N(0, 1)$.
The parameters μ and σ are fixed to 0 and 1, respectively, for all the considered models. As in the simulation study presented in Section 3.6, we have drawn data from the following models:

M_0=**Null model:** $(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$.

M_1=**Model with 1 covariate:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 0, 0)$.

M_2=**Model with 2 covariates:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 1, 0)$.

M_3=**Model with 3 covariates:** $(\gamma_1, \gamma_2, \gamma_3) = (1, 1, 1)$.

Two sample sizes, $n = 50$ and $n = 100$, have been used, while the covariates have been simulated independently from standard normal distributions. As already done in Section 3.6, we use the Jeffreys' prior introduced in Subsection 1.5.1 to select the best model among the 8 possible models for each simulation scenario. All the results are

based on 100 replications for each combination of simulation scenarios. We denote by BF^{AI}_{KM} and BF^{MI}_{KM} the AIBF and the MIBF calculated over the KMMTS, respectively. As in Section 3.6, we use an ANOVA analysis where the logit of the acceptance proportion of the true model as response variable, with respect to the main effects of scenarios, BFs, selection criteria and models along with all possible interactions. From the ANOVA analysis it can be seen that the most significative covariates without interactions are: the number of observations n, the true model $(1, 1, 0)$, the $(1, 1, 1)$ (all of these with positive effects, which means that these factors make the posterior probability grow) and the $BF^{AI}_{L_{mo}}$ (with negative effect).

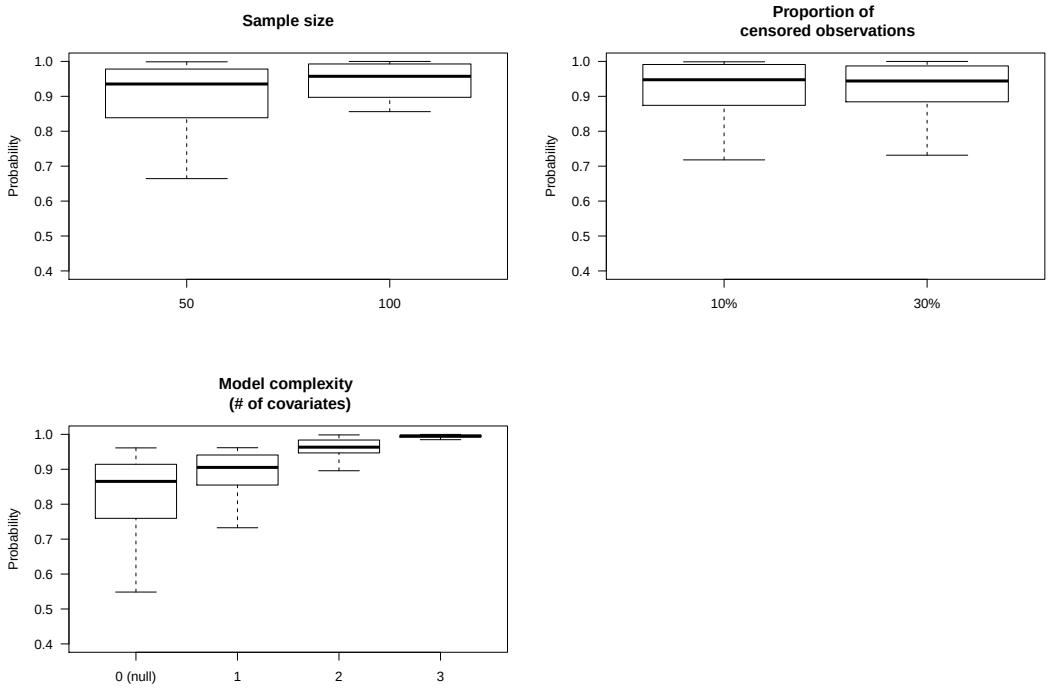

Figure 4.2: Conditional distributions of the acceptance proportion of the true log-normal model for the different simulated scenarios and marginally to the scenarios not mentioned in the corresponding Box-Plot. Values are based on all versions of BFs as well as all model selection strategies.

Figure 4.2 provides an overview of the acceptance proportion of the true model marginally to all BFs and selection strategies. Observe that BFs seem to be consistent, that the increasing censoring proportion slightly complicates the model selection procedure and that models with less covariates are in general more difficult to be detected.

Figure 4.3 shows that there is no significative difference between the considered BFs

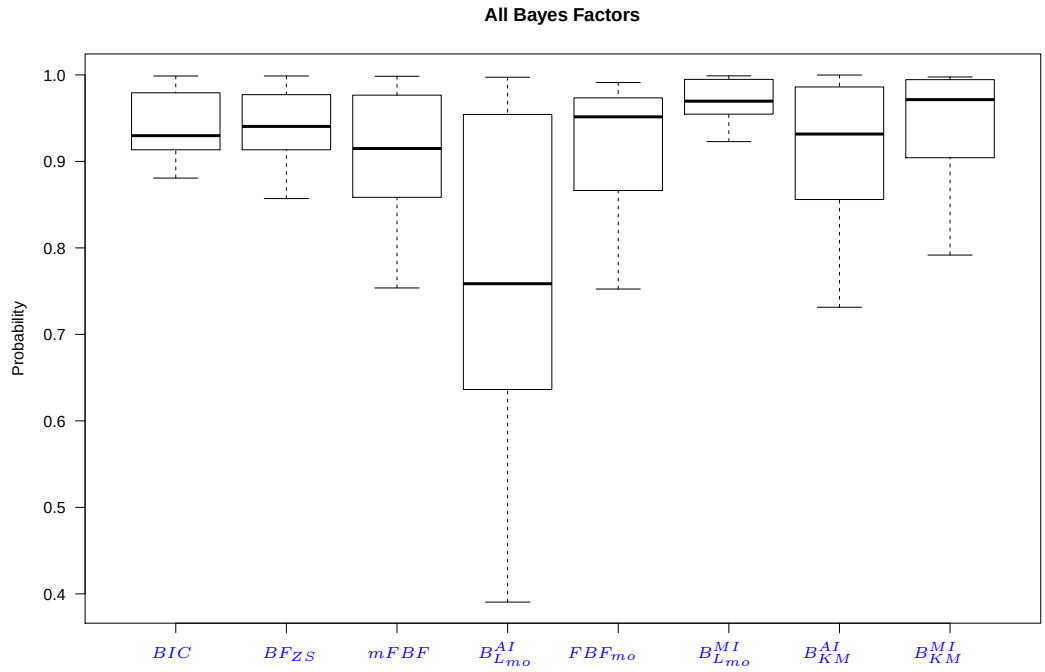

Figure 4.3: Conditional distributions of the acceptance proportion of the true log-normal model for the 9 different tools.

marginally to all scenarios, considering that their performance is consistent with the behavior of the real BF and it can be observed that the $BF^{AI}_{L_{mo}}$ has a great variability so it is less precise than the others.

In Figure 4.4 it appears the mean of the acceptance proportion of the true model calculated over 100 replications, $\widetilde{p}$, using each considered tool, sample sizes 50 and 100 with 30% of missing data, jointly with its standard deviation, $se(\widetilde{p})$. The same plot appears in Figure 4.5 where the percentage of missing data is 10%.

As it can be seen in these graphics, $\tilde{p}$ decreases when the percentage of censored data grows, while it increases with the sample size. A general idea about the behavior of all considered BFs can be obtained from Figure 4.4 at a specified scenario, namely 30% of censored observations. Considering $\widetilde{p} \pm se(\widetilde{p})$ we can see that:

i) $B^{MI}_{L_{mo}}$ provides best results along all the true models;

ii) B^{MI}_{KM} has a similar behavior to $B^{MI}_{L_{mo}}$, with some differences only under the null model;

iii) B^{AI}_{KM} and $B^{AI}_{L_{mo}}$ have a worse behavior compared to the other BFs, especially for simpler models;

iv) BF_{ZS} has a good behavior in general, in particular it produces very good results when the null model is the true one. This is due to the fact that the Zellner-Siow prior is centered at 0;

iv) $mFBF$ behaves similarly to FBF_{mo} and to BF_{ZS} and gives similar results to $B^{MI}_{L_{mo}}$, especially for complex models;

v) BIC works well in this case in all the scenarios, this could be due to the normal distribution of errors as BIC is based on Laplace approximations.

We can observe that the new tools introduced in this Section, B^{MI}_{KM} and the Zellner-Siow BF, BF_{ZS}, work well in all the considered scenarios. However, it is necessary to explore more deeply the behavior of B^{MI}_{KM} when considering data in which the censoring depends on covariates, because in the definition of the KMMTS the estimation of the mass function is done through $\hat{F}_{KM}$ estimated with the marginal distribution of $\boldsymbol{y}$ and, hence, without taking into account the effect of covariates. This could result in a poor behavior of B_{KM} when censoring depends on some covariates, as pointed out in Subsection 4.1.1. In these cases it would be necessary to use the estimator $\hat{F}_{KM}$ conditional to covariates influencing censoring. Again, averaging over all values of N_t produces good results.

Next, as done in Section 3.6, we have calculated the posterior expected model size for each BF and for each simulation scenario. In Figures 4.6, 4.7, 4.8 and 4.9 the boxplots of the posterior expected model size for the 100 replications for the log-normal models are represented, for $n = 50$ and $n = 100$ and for the two censoring percentages, 10% and 30 %. The figures confirm the results previously obtained by means of the acceptance proportion, in particular we can observe that the $B^{MI}_{L_{mo}}$ is more precise in estimating the true model size.

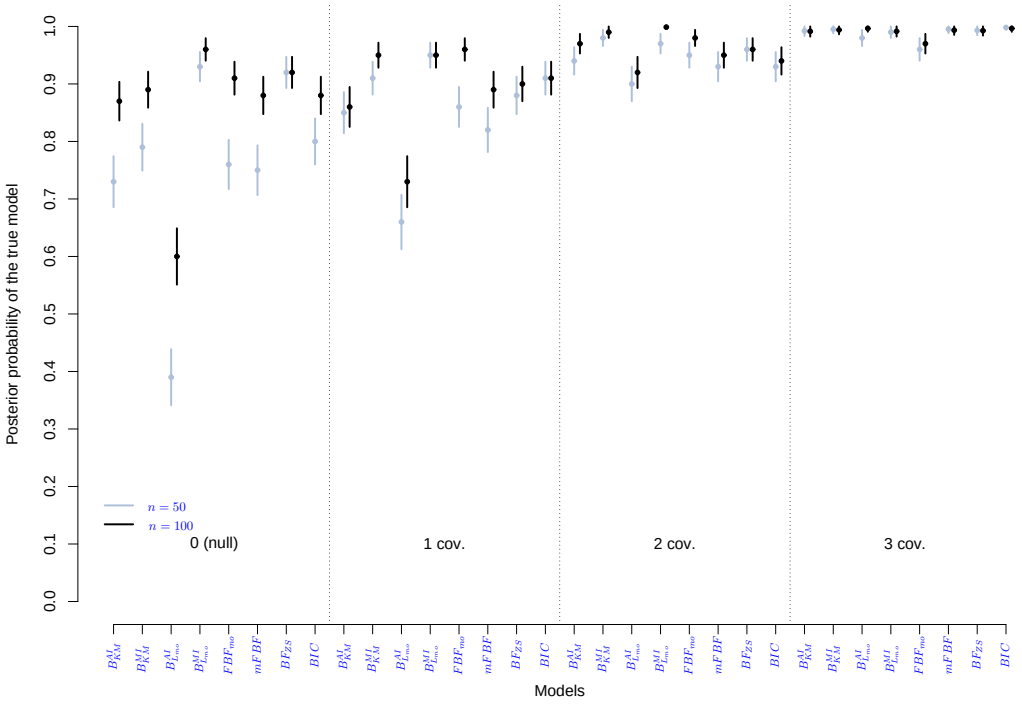

Figure 4.4: Values of $\tilde{p} \pm se(\tilde{p})$ for log-normal model, different BFs with: 30% of censored data and two sample sizes.

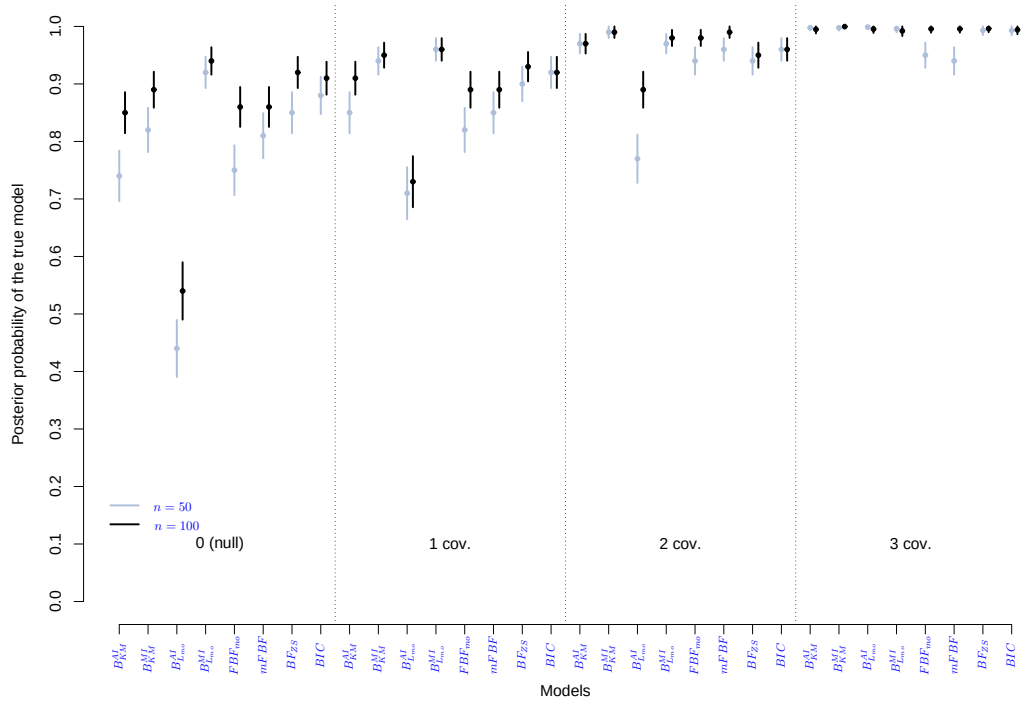

Figure 4.5: Values of $\tilde{p} \pm se(\tilde{p})$ for log-normal model, different BFs with: 10% of censored data and two sample sizes.

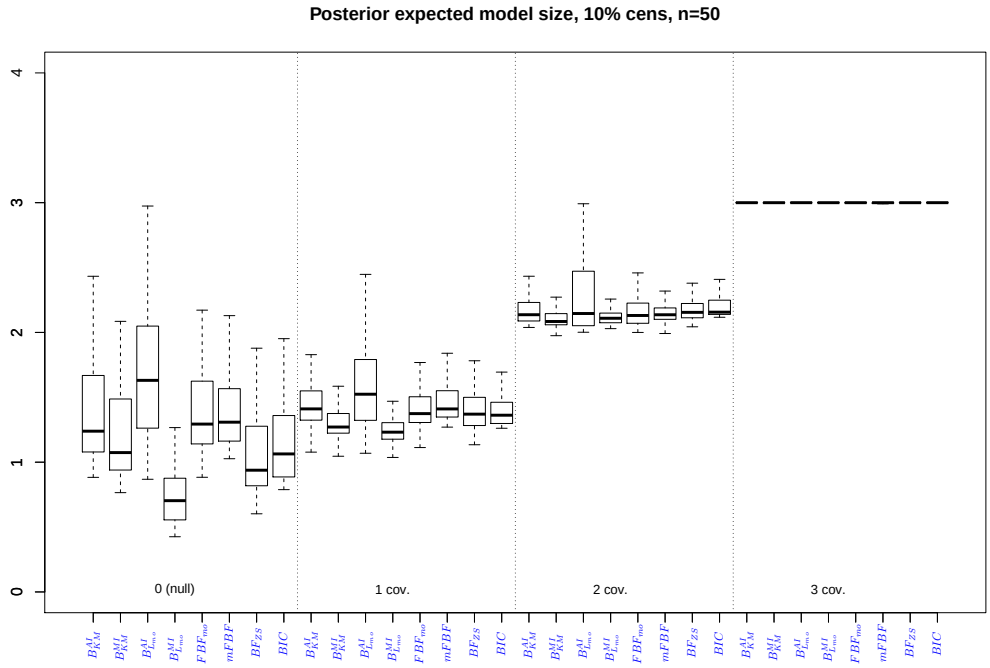

Figure 4.6: Distribution of the posterior expected model size for the log-normal model, different BFs with: 10% of censored data and $n = 50$.

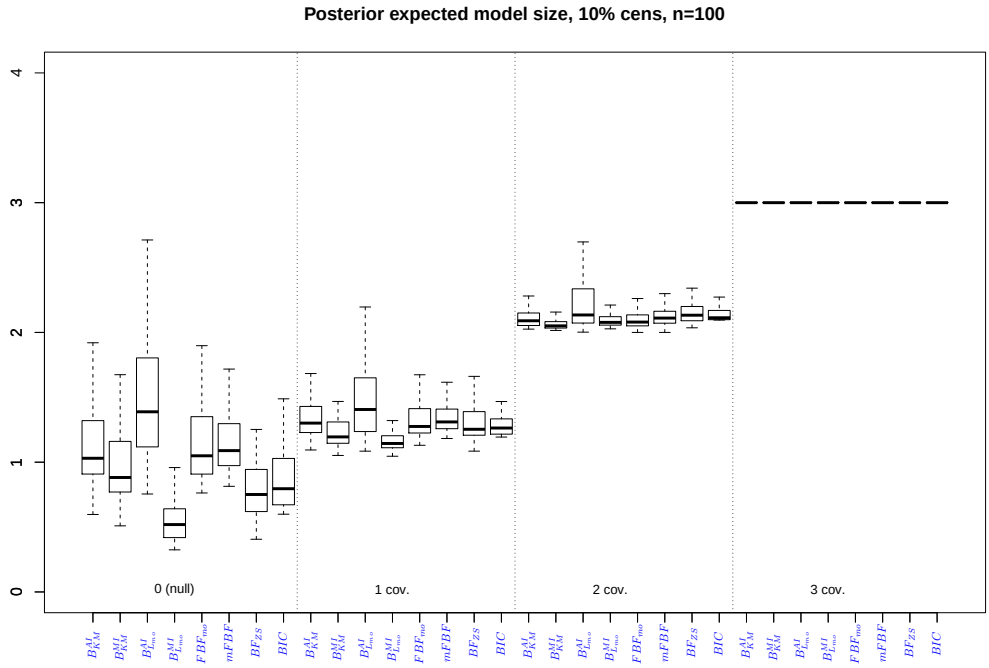

Figure 4.7: Distribution of the posterior expected model size for the log-normal model, different BFs with: 10% of censored data and $n = 100$.

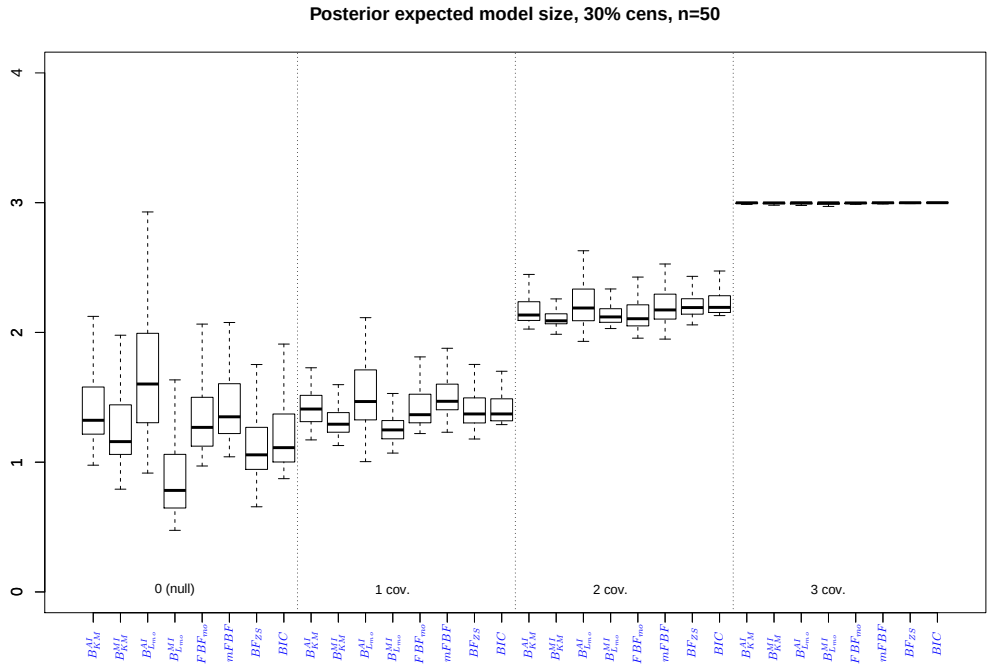

Figure 4.8: Distribution of the posterior expected model size for the log-normal model, different BFs with: 30% of censored data and $n = 50$.

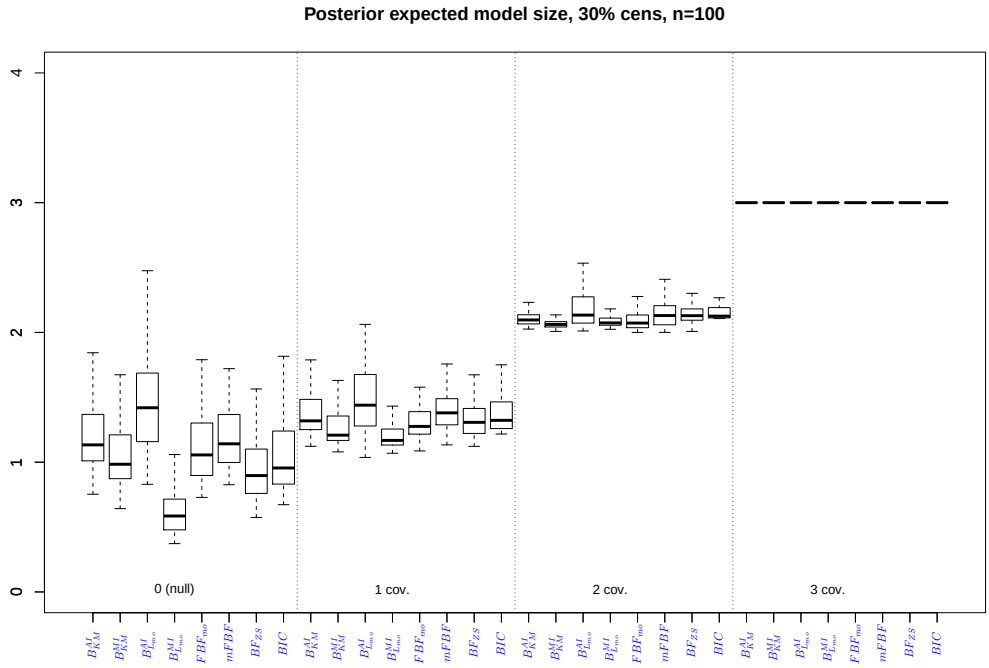

Figure 4.9: Distribution of the posterior expected model size for the log-normal model, different BFs with: 30% of censored data and $n = 100$.

5

Applications

5.1 NSCLC Dataset

The first application contains the NSCLC (non-small cells lung cancer) dataset, which we have analysed during the research project "Treatment optimization of the non-small cells lung cancer by means of a characterization of a Bayesian network and development of a decision making system: modelling, simulation and validation", at the Infanta Cristina Hospital, Parla, Madrid.
This dataset contains the survival times for 35 patients at the fourth stage of the NSCLC, of which 19 are censored. In the original dataset there are two different types of survival times: the overall survival, which is the time from the entrance in the study until death and the progression-free survival, which is the time until the cancer progresses. In this thesis we only consider as response variable the overall survival.
There are 14 predictive variables present in the study:

- `age`: patient's age expressed in years
- `sex`: patient's gender
- `smoking habit`: categorical (no smoker or ex smoker/smoker)
- `bmi`: body mass index, numerical
- `basal ecog`: categorical measure about patients' general well-being (0−1, 2 or NA)
- `localization`: categorical variable denoting the area of the body where the tumor is located (hilar mass, peripheral mass or multi-nodural)

	Median	Range
Body Mass Index (Bmi)	24.8	17.3 - 30.1
Albumin	3.5	2.1 - 4.6
Carcinoembryonic antigen (Cea)	2.9	0.5 - 8357.4
Lactate dehydrogenase (Ldh)	298.0	147 - 2744
Calcaemia	9.6	8.8 - 10.8

Table 5.1: Median and range of the continuous covariates for the 35 patients in the study.

- `number of organs`: number of affected organs (1, 2 or 3)
- `ldh`: value of lactate dehydrogenase (U/l)
- `calcaemia`: value of calcaemia (mg/dl)
- `anaemia`: value of anemia (g/dl)
- `cea`: value of carcinoembryonic antigen (ng/ml)
- `albumin`: value of albumin (g/dl)
- `histological type`: hystological type of cancer, with three values: adenocarcinoma, squamous or undetermined
- `complications`: number of complications (none, one or more).

As the sample size is relatively small and there is a 74% of censoring, we have considered only five possible predictive variables indicated by the oncologist. These variables are `albumin`, `bmi`, `cea`, `ldh` and `calcaemia` and are summarized in Table 5.1.

To avoid the effect of extreme observations `cea`, `ldh` and `bmi` have been discretized following medical indications. The discretized variables appear in Table 5.2.

We have considered the Weibull model because of the flexibility of its hazard ratio to represent the behavior of the survival in this type of study. The potential models are $2^5 = 32$, for each of them the BFs presented in Chapter 3 are approximated, taking into account that $L_{mode} = L_{median}$ in this case.

We use the two strategies, HPPM and MPPM, to select models.

Results corresponding to the posterior probability calculated with HPPM appear in Table 5.3, MPPM produces the same ordering across models.

From the table, we observe that all BFs, except $B^{MI}_{L_{mo}}$, agree in choosing the model `calcaemia` as the most probable one, in particular the BIC assigns a considerable

	N. of patients	%
`bmi`		
0: $18 \leq$ `bmi` ≤ 25	16	46
1: `bmi` < 18 or `bmi` > 25	19	54
`cea`		
0: `cea` ≤ 30	32	91
1: `cea` > 30	3	9
`ldh`		
0: `ldh` ≤ 250	18	51
1: $250 <$ `ldh` ≤ 400	10	29
2: `ldh` > 400	7	20

Table 5.2: Discretized variables for the NSCLC dataset.

probability to this model, while the other BFs, in particular the B^{MI}_{Lmo}, tend to assign non negligible probability to the null model. Based on simulation results reported in Section 3.6, the MIBF produces best results, which means that in this application it is expected a sparse model, that could be the null or the one with `calcaemia`. In order to make predictions it would be appropriate to use Bayesian model averaging to take into account all models with non-negligible probability.

Then, we have calculated the posterior expected model size for each BF, results appear in Table 5.4. Observe that the B^{MI}_{Lmo} has the lowest posterior expected model size and the BIC has the largest posterior expected model size, as supposed.

5.2 Larynx Dataset

In this Section we present results obtained working on the *larynx* dataset introduced in Example 1, but using a Weibull model. This dataset describes the survival times of $n = 90$ male patients suffering from larynx cancer of which $n_{cens} = 40$ are censored during the time period 1970−1978. As suggested by Klein and Moeschberger (2003), we adopt a Weibull regression model using the main effects of variables `age` and `stage` in Table 5.5.

All measures introduced in Chaper 3 have been computed, taking into account that $L_{mode} = L_{median}$. The possible models, containing only additive effects, are listed in Example 14.

In Table 5.6 posterior probabilities of models calculated via HPPM are presented,

k	Model	$B^{AI}_{L_{mo}}$	$B^{MI}_{L_{mo}}$	FBF_{mo}	$mFBF$	BIC
1	`calcaemia`	0.421	0.365	0.425	0.325	0.386
2	`Null`	0.191	0.367	0.271	0.181	0.043
3	`bmi-calcaemia`	0.128	0.074	0.002	0.057	0.099
4	`cea`	0.062	0.044	0.013	0.017	0.011
5	`albumin-calcaemia`	0.047	0.060	0.099	0.076	0.102
6	`cea-calcaemia`	0.035	0.000	0.080	0.074	0.140
7	`albumin`	0.024	0.041	0.043	0.037	0.022
8	`albumin-bmi-calcaemia`	0.018	0.011	0.000	0.021	0.027
9	`albumin-cea-calcaemia`	0.016	0.000	0.021	0.018	0.035
10	`bmi`	0.013	0.018	0.016	0.020	0.011
11	`cea-bmi-calcaemia`	0.013	0.001	0.015	0.017	0.037
12	`albumin-bmi`	0.008	0.007	0.004	0.008	0.007
13	`ldh-calcaemia`	0.005	0.008	0.000	0.006	0.025
14	`ldh`	0.003	0.000	0.004	0.003	0.003
15	`albumin-cea-bmi`	0.002	0.002	0.001	0.003	0.002

Table 5.3: 15 highest posterior probabilities, according to the $B^{AI}_{L_{mo}}$, of the models for the NSCLC dataset.

BF	Posterior expected model size
$B^{AI}_{L_{mo}}$	1.159
$B^{MI}_{L_{mo}}$	0.815
FBF_{mo}	0.995
$mFBF$	1.534
BIC	1.637

Table 5.4: Posterior expected model sizes for the NSCLC dataset.

MPPM produces similar results.

We can see from Table 5.6 that all BFs agree that survival is mostly related to the stage of the disease. It is worth noting that, in this case, the BIC agrees with the other BFs in that it assigns the largest probability to `stage`.

The posterior expected model sizes obtained for each BF appear in Table 5.7. Again, the $B^{MI}_{L_{mo}}$ has the lowest posterior expected model size, while the other BFs have similar values.

	N. of patients	%
`stage`		
1:	33	37
2:	17	19
3:	27	30
4:	13	14
	Median	Range
`age`	65.00	41.00 - 86.00

Table 5.5: Summary statistics for the covariates of the *larynx* dataset.

k	Model	$B^{AI}_{L_{mo}}$	$B^{MI}_{L_{mo}}$	FBF_{mo}	$mFBF$	BIC
0	`Null`	0.036	0.080	0.011	0.013	0.040
1	`stage`	0.642	0.720	0.744	0.744	0.679
2	`age`	0.021	0.012	0.005	0.007	0.023
3	`age+stage`	0.301	0.188	0.240	0.236	0.257

Table 5.6: Posterior probabilites of the 4 possible models for the Larynx dataset.

BF	Posterior expected model size
$B^{AI}_{L_{mo}}$	1.265
$B^{MI}_{L_{mo}}$	1.108
FBF_{mo}	1.230
$mFBF$	1.223
BIC	1.217

Table 5.7: Posterior expected model sizes for the Larynx dataset.

5.3 Veteran's Administration Lung Cancer Dataset (VA)

In this section we present the Veteran's Administration Lung Cancer dataset firstly presented by Prentice (1973) and analysed by Volinsky (1997). This dataset reports data from a randomized clinical trial to assess a test chemotherapy. It describes the survival times and conditions of 137 individuals suffering from advanced lung cancer, of which 9 are censored. The dataset contains 5 independent variables:

- `treat`: treatment (standard or test)

	N. of patients	%
`cell`		
type 1	35	25
type 2	48	35
type 3	27	20
type 4	27	20
`treat`		
standard	69	50
test	68	50
`prior`		
no	97	71
yes	40	29

Table 5.8: Categorical variables for the VA dataset.

- `age`: patient's age expressed in years
- `Karn`: Karnofsky score of patient's performance on a scale of 0 to 100
- `cell`: type of cells in the tumor, with four categories: squamous, small cell, large cell and adeno
- `prior`: prior therapy (yes/no)

whose descriptive statistics can be found in Tables 5.8 and 5.9.

	mean	median	sd	$q_{0.025}$	$q_{0.975}$
`age`	58.31	62.00	10.54	51.00	66.00
`Karn`	58.57	60.00	20.04	40.00	75.00

Table 5.9: Summary statistics for the continuous variables for the VA dataset.

As stated by Prentice (1973) and Kalbfleisch and Prentice (1980) and discussed in Volinsky (1997), the data fit an exponential model. So we adopt this model, which is a particular case of the Weibull one, and we use the techniques shown in Chapter 3 to calculate the FBF over the mode of N_t (because, also in this case, the median of N_t is equal to the mode), mFBF and BIC. We do not calculate the two versions of IBF because they are computationally expensive.

k	Model	FBF_{mo}
0	`Karn-cell`	0.707
1	`Karn`	0.093
2	`Karn-cell-prior`	0.056
3	`treat-Karn-cell`	0.052
4	`age-Karn-cell`	0.031
5	`Karn-prior`	0.019
6	`treat-Karn`	0.017
7	`age-Karn`	0.014
8	`age-Karn-prior`	0.003
9	`treat-Karn-prior`	0.03

Table 5.10: 10 highest posterior probabilities of models for the VA dataset according to FBF_{mo}.

In Tables 5.10, 5.11 and 5.12 we present results of the posterior probabilities of models, calculated via HPPM, between the 2^5 possible models. In particular, for each BF we have reported the 10 probability models with highest posterior probability. The FBF, mFBF and BIC select the model `Karn-cell` to be the most probable one and the model with only `Karn` the second most probable. FBF and mFBF give around 70% of probability to the model `Karn-cell` and around 9% to the model with `Karn` and BIC gives a comparable probability, being around 63%, for the `Karn-cell` model.

These results are in line with those obtained in Volinsky (1997), and as it is observed also in that work, we can state that the variable `treat` is not significantly effective.

Finally, we have calculated the posterior expected model size for each BF, results appear in Table 5.13. In this case, as in the case of the NSCLC dataset, the BIC has the largest posterior expected model size.

5.4 Primary Biliary Cirrhosis (PBC) Dataset

In this section we consider the PBC data collected by the Mayo Clinic of Rochester (Minnesota, US) from 1974 to 1984 to compare the effect of the drug DPCA with a placebo in the treatment of primary biliary cirrhosis of the liver (PBC).
The dataset was analysed by Dickson et al. (1985), Grambsch et al. (1989), Markus et al. (1989) and Fleming and Harrington (1991). Fleming and Harrington (1991), in particular, observed that the data fit a Cox regression model and they considered all

k	Model	$mFBF$
0	`Karn-cell`	0.740
1	`Karn`	0.091
2	`age-Karn-cell`	0.046
3	`treat-Karn-cell`	0.035
4	`Karn-cell-prior`	0.032
5	`treat-Karn`	0.016
6	`age-Karn`	0.015
7	`Karn-prior`	0.013
8	`treat-Karn-cell-prior`	0.005
9	`treat-Karn-prior`	0.002

Table 5.11: 10 highest posterior probabilities of models for the VA dataset according to $mFBF$.

k	Model	BIC
0	`Karn-cell`	0.625
1	`Karn`	0.099
2	`treat-Karn-cell`	0.095
3	`age-Karn-cell`	0.062
4	`Karn-cell-prior`	0.060
5	`treat-Karn`	0.011
6	`treat-age-Karn-cell`	0.011
7	`Karn-prior`	0.009
8	`treat-Karn-cell-prior`	0.009
9	`age-Karn`	0.009

Table 5.12: 10 highest posterior probabilities of models for the VA dataset according to BIC.

BF	Posterior expected model size
FBF_{mo}	2.059
$mFBF$	2.041
BIC	2.175

Table 5.13: Posterior expected model sizes for the VA dataset.

the 14 covariates and, then, Volinsky (1997) reduced the analysis to 8 covariates, due to the fact that 6 of them have no effect. The dataset contains 312 patients, 2 of them containing missing data, so we reduce the dataset to 310 patients, of which 186 are censored.
The considered covariates are:

- `age`: age expressed in years
- `albumin`: serum albumin (g/dl)
- `bili`: serum bilirubin (mg/dl)
- `copper`: urine copper (ug/day)
- `edema`: categorical (no edema, untreated or successfully treated, edema despite diuretic therapy)
- `stage`: categorical, histological stage of disease (needs biopsy) with 4 categories
- `ast`: aspartate aminotransferase, also called SGOT (U/ml)
- `protime`: standardised blood clotting time.

The descriptive statistics are presented in Tables 5.14 and 5.15.

As in Volinsky (1997), we consider the logarithm of `bili`, `albumin` and `protime` and analyse the data with the Weibull model. BFs introduced in Chapter 3 have been calculated, except for the IBFs which are very expensive to be obtained, as already observed in Section 5.3. Results using the posterior probabilities appear in Tables 5.16, 5.17, 5.18 and 5.19.

There are some differences in the models with highest posterior probabilities chosen using one or another tool, and there is substantial model uncertainty in the posterior probabilities of models. Using FBF (mode, median and marginalized) there is substantial uncertainty in the posterior probabilities of models, being the maximum

	N. of patients	%
`edema`		
no edema	262	85
untreated or successfully treated	28	9
edema despite diuretic therapy	20	6
`stage`		
1	16	5
2	66	21
3	120	39
4	108	35

Table 5.14: Categorical variables for the PBC dataset.

	mean	median	sd	$q_{0.025}$	$q_{0.975}$
`age`	49.95	49.71	10.57	42.05	56.68
`bili`	0.58	0.34	1.03	-0.22	1.25
`albumin`	1.25	1.27	0.13	1.20	1.34
`copper`	97.65	73.00	85.61	41.25	123.00
`ast`	122.40	114.10	56.83	80.60	151.90
`protime`	2.37	2.36	0.09	2.30	2.41

Table 5.15: Summary statistics for the continuous variables for the PBC dataset.

probability only around 20%. For all these tools the four most probable models are `age-edema-bili-albumin-copper-protime`, `age-bili-albumin-copper-protime`, `age-edema-bili-albumin-protime` and `age-bili-albumin-protime`, all of them having probabilities between 10%-20%. Also BIC selects these four models as the most probable ones and again, as in the VA dataset, it gives a comparable probability to the most complex model between these four, that is `age-edema-bili-albumin-copper-protime` with around a 20% of probability. These results are in line with findings in Volinsky (1997), in fact in that work it is proposed to use Bayesian model averaging to take into account model uncertainty.

These results are quite in agreement with the ones obtained in Volinsky (1997). Dickson et al. (1985) show that the test drug, DPCA, has not a significative effect in the treatment of the cirrhosis.

k	Model	FBF_{mo}
1	`age-edema-bili-albumin-copper-protime`	0.234
2	`age-bili-albumin-copper-protime`	0.213
3	`age-edema-bili-albumin-protime`	0.205
4	`age-bili-albumin-protime`	0.118
5	`age-bili-albumin-ast-protime`	0.059
6	`age-bili-albumin-copper-ast-protime`	0.040
7	`age-edema-bili-albumin-copper`	0.036
8	`age-edema-bili-albumin`	0.017
9	`bili-albumin-copper-protime`	0.015
10	`age-edema-bili-albumin-ast-protime`	0.014

Table 5.16: 10 highest posterior probabilities of models for the PBC dataset according to FBF_{mo}.

k	Model	FBF_{me}
1	`age-bili-albumin-copper-protime`	0.229
2	`age-edema-bili-albumin-copper-protime`	0.134
3	`age-edema-bili-albumin-protime`	0.117
4	`age-bili-albumin-protime`	0.106
5	`age-bili-albumin-copper-ast-protime`	0.095
6	`age-edema-bili-albumin-copper-ast-protime`	0.073
7	`age-bili-albumin-ast-protime`	0.068
8	`age-edema-bili-albumin-copper`	0.049
9	`age-edema-bili-albumin`	0.044
10	`age-edema-bili-albumin-ast-protime`	0.016

Table 5.17: 10 highest posterior probabilities of models for the PBC dataset according to FBF_{me}.

k	Model	$mFBF$
1	`age-bili-albumin-copper-protime`	0.237
2	`age-edema-bili-albumin-protime`	0.160
3	`age-bili-albumin-protime`	0.152
4	`age-edema-bili-albumin-copper-protime`	0.148
5	`age-edema-bili-albumin-copper`	0.058
6	`age-bili-albumin-ast-protime`	0.050
7	`age-edema-bili-albumin`	0.049
8	`age-bili-albumin-copper-ast-protime`	0.049
9	`age-edema-bili-albumin-ast-protime`	0.028
10	`bili-albumin-copper-protime`	0.023

Table 5.18: 10 highest posterior probabilities of models for the PBC dataset according to $mFBF$.

k	Model	BIC
1	`age-edema-bili-albumin-copper-protime`	0.239
2	`age-edema-bili-albumin-protime`	0.180
3	`age-bili-albumin-copper-protime`	0.174
4	`age-bili-albumin-protime`	0.106
5	`age-edema-bili-albumin-ast-protime`	0.052
6	`age-edema-bili-albumin-copper`	0.049
7	`age-edema-bili-albumin-copper-ast-protime`	0.049
8	`age-bili-albumin-copper-ast-protime`	0.034
9	`age-edema-bili-albumin`	0.033
10	`age-bili-albumin-ast-protime`	0.029

Table 5.19: 10 highest posterior probabilities of models for the PBC dataset according to BIC.

In Table 5.20 the posterior expected model sizes obtained for each BF appear. In this case, the $mFBF$ has the lowest posterior expected model size, while the other BFs have similar values. In this case the BIC has the largest posterior expected model size.

BF	Posterior expected model size
FBF_{mo}	5.153
FBF_{me}	5.236
$mFBF$	5.001
BIC	5.290

Table 5.20: Posterior expected model sizes for the PBC dataset.

6

Conclusions and Future Work

6.1 Summary and Conclusions

In this work we have studied the main approaches to the model selection problem for censored data under an objective Bayesian point of view, in which only non-informative priors are used. In particular, we have discussed the best known tools: the IBF, the FBF and the BIC. It is observed that, when working with censored data it is necessary to adapt the definitions of the usual BFs for improper priors. When using the definition of SMTS of Berger and Pericchi (2004), the probability distribution of the random size of the SMTS, which is crucial in the expression of the IBF and FBF, has been calculated. Then, we have noticed that the IBF, along with its variants, is very slow to be calculated and it requires uncommon tools (i.e a great number of processors), so we have introduced another variation of the FBF, the mFBF. The main advantage of this tool is that it requires less time to be computed, it takes into account all the possible values of the fraction of the likelihood along with its probability function and it produces a good approximation to the IBF in many of the analysed cases. Next, relying on the definitions of IBF and mFBF, we have recalled the definitions of intrinsic and fractional priors and we have obtained them for the exponential right censored model, showing that for this case, the fractional prior corresponding to the mFBF is a mixture of the fractional priors (i.e. inverse gamma distributions) corresponding to each FBF with the appropriate fraction b.

We have also introduced a new way of obtaining a MTS, called the KMMTS, which is based on the reweighting of the usual MTS extraction mechanism. In particular, the Kaplan-Meier estimator of the distribution function is used to draw observations from the sample, leading to a MTS that only contains uncensored observations. This

procedure does not require the sampling mechanism of the SMTS that would take a large amount of time, and it is suitable to be used when the BFs have closed-form expressions.
Finally, it is presented a comparison of the BFs by means of two simulation studies and the analysis of four real datasets. The main results are summarized below:

- the mFBF has the advantage that we do not need to specify a particular fraction of the likelihood function;
- the MIBF is computationally expensive, but it provides the best results;
- the AIBF is unstable, due to the presence of outliers in the data;
- BIC tends to select the most complex models especially when Laplace approximation results not adequate because of lack of symmetry and of normality, as in the case of the Weibull regression.

Based on these results, we can state that the proposed mFBF performs as the second best tool, compensating a small decrease in precision in model selection with a quicker answer.

6.2 Future work

There are many topics to be investigated to continue this work. Some of them are:

- *Reweighting of IBF using the weights defined by the distribution of N_t.* In relation to the IBF calculated on SMTSs in Section 3.3, one possibility could be to study the definition of the IBF considering the reweighting induced by the probability distribution of N_t. We have already observed that this is computationally expensive, but it would lead to a more precise and reliable IBF.
- *Investigating the behavior of IBF based on KMMTS when censoring depends on covariates.* In Subsection 4.1.1 we have noticed that when the assumption of non-informative censoring is violated, the Kaplan-Meier estimator of the $\hat{F}_{KM}$ estimates the wrong distribution function. Then, another choice could be to explore more in detail the behavior of IBFs calculated over the KMMTS, focusing in particular on how they behave when the censoring depends on some covariates.

- *Exploring more general definitions of the Kaplan-Meier estimator.* It would be useful to explore other more general definitions of the Kaplan-Meier estimator in the tail. Following the idea of Brown et al. (1974), which propose to complete the tail by an exponential curve chosen to give the same value of $S(t_{max})$, we would try to adopt again a Weibull model for the tail.

- *Analysing more deeply the use of conventional priors.* It would be interesting to study the behavior of the conventional priors for regression models and in particular for the Weibull model.

- *Studying the behavior of the BFs for different types of censoring.* Another possibility could be to explore the behavior of the BFs for different types of censoring. In particular, it would be useful to consider the Type II censoring characterized by a deterministic stopping rule which, as already observed in Subsection 3.3.1, complicates the likelihood.

Bibliography

J. Albert. *Bayesian Computation with R*. Springer, 2009.

C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. ISSN 00905364. URL http://www.jstor.org/stable/2958336.

E. Arjas and D. Gasbarra. Nonparametric bayesian inference from right censored survival data, using the gibbs sampler. *Statist. Sinica*, 4:505–524, 1994.

C. Armero, S. Cabras, M. E. Castellanos, S. Perra, A. Quirós, M. J. Oruezábal, and J. Sánchez-Rubio. Bayesian analysis of a disability model for lung cancer survival. *Statistical Methods in Medical Research*, 2012. doi: 10.1177/0962280212452803. URL http://smm.sagepub.com/content/early/2012/07/04/0962280212452803.abstract.

A. C. Atkinson. Posterior probabilities for choosing a regression model. *Biometrika*, 65 (1):39–48, 1978. ISSN 00063444. URL http://www.jstor.org/stable/2335274.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

M. S. Bartlett. A comment on d. v. lindley's statistical paradox. *Biometrika*, 44(3-4): 533–534, 1957.

M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, pages 1550–1577, 2012.

S. Bennett. Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2):165–171, 1983. ISSN 00359254. URL http://www.jstor.org/stable/2347295.

BIBLIOGRAPHY

J. O. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

J. O. Berger and L. Pericchi. On the justification of default and intrinsic bayes factors. *Modeling and Predictions, New York: Springer-Verlag*, pages 276–293, 1997.

J. O. Berger and L. R. Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291387.

J. O. Berger and L. R. Pericchi. Accurate and stable bayesian model selection: The median intrinsic bayes factor. *Sankhya: The Indian Journal of Statistics, Series B*, 60(1):1–18, 1998. ISSN 05815738. URL http://www.jstor.org/stable/25053019.

J. O. Berger and L. R. Pericchi. Training samples in objective bayesian model selection. *The Annals of Statistics*, 32(3):841–869, 2004. URL http://www.jstor.org/stable/3448577.

J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith. On the development of reference priors. *Bayesian Statistics*, 4:35–60, 1992.

J. O. Berger, L. R. Pericchi, and J. A. Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhya: The Indian Journal of Statistics, Series A*, 60(3):307–321, 1998. ISSN 0581572X. URL http://www.jstor.org/stable/25051210.

J. O. Berger, L. R. Pericchi, J. K. Ghosh, T. Samanta, and F. De Santis. Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, 38:135–207, 2001. ISSN 07492170. URL http://www.jstor.org/stable/4356165.

J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, 37:905–938, 2009.

R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, February 1966.

J. M. Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147, 1979.

F. Bertolino, W. Racugno, and E. Moreno. Bayesian model selection approach to analysis of variance under heteroscedasticity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 49(4):503–517, 2000. ISSN 00390526.

G. E. P. Box and R. D. Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18, 1986. ISSN 00401706. URL http://www.jstor.org/stable/1269599.

J. B. W. Brown, M. Hollander, and R. M. Korwar. Nonparametric tests of independence for censored data, with applications to heart transplant studies. *In Reliability and Biometry: Statistical Analysis of Lifelength, F. Proschan and R. J. Serfling, eds. Philadelphia*, pages 327–354, 1974.

B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346151.

G. Casella and E. Moreno. Objective bayesian variable selection. *Journal of the American Statistical Association*, 101(473), 2006.

G. Casella, F. J. Girón, M. L. Martínez, and E. Moreno. Consistency of bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207–1228, 2009.

S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001. URL http://pubs.amstat.org/doi/abs/10.1198/016214501750332848.

D. G. Clayton. A monte carlo method for bayesian inference in frailty models. *Biometrics*, 47:467–485, 1991.

M. Clyde. Bayesian model averaging and model search strategies. *Bayesian Statistics, Oxford University Press*, 6:157–185, 1999.

M. Clyde and E. I. George. Flexible empirical bayes estimation for wavelets. *J. Roy.Statist. Soc. Ser. B*, 62:681–698, 2000.

M. Clyde, H. Desimone, and G. Parmigiani. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291738.

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

BIBLIOGRAPHY

D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall, 1984.

P. De Blasi. *Semiparametric Models in Bayesian Event History Analysis Using Beta Processes*. Dissertation, L. Bocconi University, Milano, Italy, April 2006.

F. De Santis and F. Spezzaferri. Alternative bayes factors for model selection. *Canadian Journal of Statistics*, 25(4):503–515, 1997. ISSN 1708-945X. doi: 10.2307/3315344. URL http://dx.doi.org/10.2307/3315344.

M. H. DeGroot. *Probability and statistics*. Addison-Wesley series in statistics. Addison-Wesley Pub. Co., 1975. ISBN 9780201113662. URL http://books.google.co.uk/books?id=OPpQAAAAMAAJ.

E. R. Dickson, T. R. Fleming, R. H. Wiesner, W. P. Baldus, C. R. Fleming, J. Ludwig, and J. T. McCall. Trial of penicillamine in advanced primary biliary cirrhosis. *New England Journal of Medicine*, 312(16):1011–1015, 1985. doi: 10.1056/NEJM198504183121602. URL http://www.nejm.org/doi/full/10.1056/NEJM198504183121602.

J. Dmochowski. Intrinsic priors via kullback-leibler geometry. Technical report, Department of Statistics, Purdue University, June 1994.

K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, 2:183–201, 1974.

R. Doll. The age distribution of cancer: Implications for models of carcinogens. *Journal of the Royal Statistical Society, Series A.*, 134:133–166, 1971.

H. Doss. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, 22(4):1763–1786, 1994. ISSN 00905364. URL http://www.jstor.org/stable/2242483.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97, 1995.

R. L. Dykstra and P. W. Laud. A bayesian nonparametric approach to reliability. *Annals of Statistics*, 9:356–367, 1981.

B. Efron. The two sample problem with censored data. In *The Two Sample Problem with Censored Data*, volume 4, pages 831–853. New York: Prentice-Hall, 1967.

I. G. Evans and A. M. Nigm. Bayesian prediction for two-parameter weibull lifetime models. *Comm. Statistics- Theory and Methods*, 9(6):649–658, 1980.

M. Feinleib. A method of analyzing log normally distributed survival data with incomplete follow-up. *Journal of the American Statistical Association*, 55:534–545, 1960.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, 1974. ISSN 00905364. URL http://www.jstor.org/stable/2958401.

T. S. Ferguson and E. G. Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 7(1):163–186, 1979. ISSN 00905364. URL http://www.jstor.org/stable/2958840.

T. R. Fleming and D. P. Harrington. Counting processes and survival analysis. *Biometrical Journal*, 34(6):674–674, 1991. ISSN 1521-4036. doi: 10.1002/bimj.4710340605. URL http://dx.doi.org/10.1002/bimj.4710340605.

E. J. Freireich, Acute Leukemia Group B, E. Gehan, L. R. Schroeder, I. J. Wolman, R. Anbari, E. O. Burgert, S. D. Mills, D. Pinkel, O. S. Selawry, J. H. Moon, B. R. Gendel, C. L. Spurr, R. Storrs, F. Haurani, B. Hoogstraten, and S. Lee. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, 21(6):699–716, 1963. URL http://bloodjournal.hematologylibrary.org/content/21/6/699.abstract.

E. I. George and D. P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87:731–747, 2000.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2290777.

E. I. George and R. E. McCulloch. Stochastic search variable selection. *Practical Markov Chain Monte Carlo in Practice, eds. W.R. Gilks et al. Chapman and Hall: London*, pages 339–348, 1995.

E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

BIBLIOGRAPHY

J. Geweke. Variable selection and model comparison in regression. *Bayesian Statistics, eds. J.M.Bernardo et al., Oxford University Press: Oxford*, 5:169–194, 1996.

J. K. Ghosh, M. Delampady, and T. Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer texts in statistics. Springer, New York, NY, 2006.

R. D. Gill. The intrinsic bayes factor explained by examples. Technical report, Mathematical Centre Tracts. Amsterdam: Mathematisch Centrum, Amsterdam, 1980.

P. M. Grambsch, E. R. Dickson, M. Kaplan, G. Lesage, T. R. Fleming, and A. L. Langworthy. Extramural cross-validation of the mayo primary biliary cirrhosis survival model establishes its generalizability. *Hepatology*, 10 (5):846–850, 1989. ISSN 1527-3350. doi: 10.1002/hep.1840100516. URL http://dx.doi.org/10.1002/hep.1840100516.

E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

R. C. Gupta, O. Akman, and S. Lvin. A study of log-logistic model in survival analysis. *Biometrical Journal*, 41(4):431–443, 1999. ISSN 1521-4036.

M. S. Hamada, A. G. Wilson, C. S. Reese, and H. F. Martz. *Bayesian Reliability*. Springer, 2008.

M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *J. Roy.Statist. Soc. Ser. B*, 96:681–698, 2000.

N. L. Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18:1259–1294, 1990.

R. D. Horner. Age at onset of alzheimerÕs disease: Clue to the relative importance of etiologic factors. *Journal of the American Statistical Association*, 126:409–414, 1987.

D. W. Jr. Hosmer and S. Lemeshow. *Applied Survival Analysis*. Wiley Series in Probability and Statistics, 1999.

J. G. Ibrahim, M. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, 2001.

H. Jeffreys. *Theory o[Probability*. Oxford Univ. Press, Oxford, 1961.

J. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 1980.

J. D. Kalbfleisch. Non-parametric analysis of survival time data. *Journal of the Royal Statistical Society: Series B*, 40:214–221, 1978.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. New York: John Wiley and Sons., 2002.

E. R. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

O. Kardaun. Statistical analysis of male larynx-cancer patients - a case study. *Statistical Nederlandica*, 37:103–126, 1983.

R. E. Kass. Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42(5):551–560, 1993. ISSN 00390526. URL http://www.jstor.org/stable/2348679.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

R. E. Kass and L. Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90 (431):928, 1995. URL http://www.jstor.org/stable/2291327?origin=crossref.

S. Kim and D. Sun. Intrinsic priors for model selection using an encompassing model with applications to censored failure time data. *Lifetime Data Analysis*, 6:251–269, 2000. ISSN 1380-7870.

Y. Kim. Nonparametric bayesian estimators for counting processes. *The Annals of Statistics*, 27(2):562–588, 1999. ISSN 00905364. URL http://www.jstor.org/stable/120104.

J. P. Klein and M. L. Moeschberger. *Survival Analysis. Techniques for Censored and Truncated Data*. Springer, 2003.

A. Kottas. Nonparametric bayesian survival analysis using mixtures of weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578 – 596, 2006. ISSN 0378-3758. doi: 10.1016/j.jspi.2004.08.009. URL http://www.sciencedirect.com/science/article/pii/S0378375804003465.

S. L. Lauritzen, B. Thiesson, and D. J. Spiegelhalter. Diagnostic systems created by model selection methods: a case study. *Uncertainty in Artificial Intelligence*, 4:143–152, 1994.

BIBLIOGRAPHY

M. Lavine. Some aspects of polya tree distributions for statistical modelling. *Annals of Statistics*, 20:1222–1235, 1992.

P. N. Lee and J. A. O'Neill. The effect both of time and dose applied on tumor incidence rate in benzopyrene skin painting experiments. *British Journal of Cancer*, 25:759–770, 1971.

S. M. Lewis and A. E. Raftery. Estimating bayes factors via posterior simulation with the laplace-metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997. doi: 10.1080/01621459.1997.10474016. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1997.10474016.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. doi: 10.1198/016214507000001337. URL http://pubs.amstat.org/doi/abs/10.1198/016214507000001337.

R. T. Lingham and S. Sivaganesan. Intrinsic bayes factor approach to a test for the power law process. *Journal of Statistical Planning and Inference*, 77(2): 195 – 220, 1999. ISSN 0378-3758. doi: 10.1016/S0378-3758(98)00181-5. URL http://www.sciencedirect.com/science/article/pii/S0378375898001815.

B. H. Markus, E. R. Dickson, P. M. Grambsch, T. R. Fleming, V. Mazzaferro, G. B. G. Klintmalm, R. H. Wiesner, D. H. Van Thiel, and T. E. Starzl. Efficacy of liver transplantation in patients with primary biliary cirrhosis. *New England Journal of Medicine*, 320(26):1709–1713, 1989. doi: 10.1056/NEJM198906293202602. URL http://www.nejm.org/doi/full/10.1056/NEJM198906293202602.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. ISSN 01621459. URL http://www.jstor.org/stable/2290129.

M. L Moeschberger and J. P. Klein. A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics*, 41:253–259, 1985.

E. Moreno and F. Girón. Comparison of bayesian objective procedures for variable selection in linear regression. *Test*, 3:472–492, 2008.

P. Muliere and S. Walker. A bayesian non-parametric approach to survival analysis using polya trees. *Scandinavian Journal of Statistics*, 24(3):

331–340, 1997. ISSN 1467-9469. doi: 10.1111/1467-9469.00067. URL http://dx.doi.org/10.1111/1467-9469.00067.

L. E. Nieto-Barajas and S. G. Walker. Bayesian nonparametric survival analysis via lévy driven markov processes. *Statist. Sinica*, 14:1127–1146, 2004.

A. O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346088.

A. O'Hagan. Properties of intrinsic and fractional bayes factors. *TEST*, 6:101–118, 1997. ISSN 1133-0686. URL http://dx.doi.org/10.1007/BF02564428. 10.1007/BF02564428.

C. Pereira, J. Stern, and S. Wechslerz. Can a significance test be genuinely bayesian? *Bayesian Analysis*, 3:79–100, 2008.

J. M. Perez. *Development of expected posterior prior distributions for model comparisons*. Dissertation, Purdue University, United States, Indiana, January 2000.

L. R. Pericchi. An alternative to the standard bayesian procedure for discrimination between normal linear models. *Biometrika*, 71(3):575–586, 1984. ISSN 00063444. URL http://www.jstor.org/stable/2336567.

L. R. Pericchi, A. Fiteni, and E. Presa. The intrinsic bayes factor explained by examples. Technical report, Technical report, Dept. Estadística y Econometría, Universidad Carlos III, Madrid, 1993.

R. Peto and P. N. Lee. Weibull distributions for continuous-carcinogenesis experiments. *Biometrics*, 29:457–470, 1973.

M. C. Pike. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, 22:142–161, 1966.

D. J. Poirier. Bayesian hypothesis testing in linear models with continuously induced conjugate prior across hypotheses. *Bayesian Statistics, eds. J.M.Bernardo et al., Elsevier: New York*, 2:711–722, 1985.

R. L. Prentice. Exponential survivals with censoring and explanatory variables. *Biometrika*, 60(2):279–288, 1973. doi: 10.1093/biomet/60.2.279. URL http://biomet.oxfordjournals.org/content/60/2/279.abstract.

BIBLIOGRAPHY

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997. ISSN 01621459. URL http://www.jstor.org/stable/2291462.

A. E. Raftery, D. Madigan, J. A. Hoeting, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

M. Rausand and A. Hoyland. *System Reliability Theory; Models, Statistical Methods and Applications*. Wiley, Hoboken, New Jersey, 2004.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. ISSN 00905364. URL http://www.jstor.org/stable/2958889.

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

A. F. M. Smith and D. J. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2): 213–220, 1980. ISSN 00359246. URL http://www.jstor.org/stable/2984964.

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317 – 343, 1996. ISSN 0304-4076. doi: 10.1016/0304-4076(95)01763-1. URL http://www.sciencedirect.com/science/article/pii/0304407695017631.

D. J. Spiegelhalter and A. F. M. Smith. Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):377–387, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345495.

D. J. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell. Bayesian analysis in expert systems (with discussion). *Statistical Science*, 8:219–283, 1993.

V. Susarla and J. Van Ryzin. Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of American Statistical Association*, 71:897–902, 1976.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data*. Springer, 2000.

J. Varshavsky. *On the develpment of intrinsic Bayes factors*. PhD thesis, Purdue University, 1995.

I. Verdinelli and L. Wasserman. Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430): 614–618, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291073.

C. T. Volinsky. *Bayesian Model Averaging for Censored Survival Models*. Dissertation, Washington University, United States, Washington, 1997.

C. T. Volinsky and A. E. Raftery. Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262, 2000. ISSN 0006341X. URL http://www.jstor.org/stable/2677130.

S. Walker and P. Damien. A full bayesian non-parametric analysis involving a neutral to the right process. *Scandinavian Journal of Statistics*, 25 (4):669–680, 1998. ISSN 1467-9469. doi: 10.1111/1467-9469.00128. URL http://dx.doi.org/10.1111/1467-9469.00128.

S. Walker and P. Muliere. Beta-stacy processes and a generalization of the polya-urn scheme. *The Annals of Statistics*, 25(4):1762–1780, 1997. ISSN 00905364. URL http://www.jstor.org/stable/2959072.

S. Weerahandi and R. Johnson. Testing reliability in a stress-strength model when x and y are normally distributed. *Technometrics*, 34(1):83–91, 1992.

W. Weibull. A statistical theory of the strength of materials. *Ingeniörsvetenskapsakademiens Handlingar*, 151, 1939.

R. Yang and J. O. Berger. A catalog of noninformative priors. *Development*, 97(97-42):1–44, 1998. URL http://www.stats.org.uk/priors/noninformative/YangBerger1998.pdf.

A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, pages 389–399, 1986.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics 1 (Eds. J.M. Bernardo and J.O. Berger and A.P. Dawid and A.F.M. Smith)*, volume 31, pages 585–603. Springer Berlin / Heidelberg, 1980. URL http://dx.doi.org/10.1007/BF02888369. 10.1007/BF02888369.

Declaration

I declare that to the best of my knowledge the contents of this thesis are original and my work except where indicated otherwise.

Appendix A

In Section 3.1 the Assumption 0 of Berger and Pericchi (2004) is introduced and it is observed that, in some cases, it is not possible to satisfy that. In this Appendix we show that for the case of the right censored exponential model the training sample introduced in 3.1 satisfies this assumption.

A.1 Assumption 0 and SMTS

Consider the case of data $(y_1, \dots, y_n)$ following the right censored exponential distribution $Exp(\theta)$. Suppose to test

$$M_0 : \theta = \theta_0 \quad vs \quad M_1 : \theta \neq \theta_0.$$

As seen in Example 5, the probability of the imaginary space of MTSs is less than 1

$$\Pr_{\boldsymbol{\theta}_i}^{M_i}\left(\mathcal{X}^{MTS}\right) = \Pr_{\boldsymbol{\theta}_i}^{M_i}(X < \rho) = 1 - \exp(-\rho\theta_i) < 1, \quad i = 0, 1.$$

Now we study the case of the SMTS. We want to prove the following:

Proposition 5. *Given a SMTS (Definition 4), then*

$$\Pr_{\boldsymbol{\theta}_i}^{M_i}\left(\mathcal{X}^{SMTS}\right) = 1, \quad i = 0, 1,$$

where $\mathcal{X}^{SMTS}$ is the space of all possible sequential minimal training samples.

Proof. Let n_{cens} denote the number of censored observations, n_u the number of uncensored observations and t_i denote the first uncensored observation which we encounter when sampling from the entire dataset. The possible SMTSs are:

SMTS	**Probability of the SMTS in the actual space**
$\{t_i\}$	$1-p_i$
$\{t_1^*, t_i\}$	$(1-p_i)p_i$
$\{t_1^*, t_2^*, t_i\}$	$(1-p_i)p_i^2$
$\vdots$	$\vdots$
$\{t_1^*, t_2^*, \ldots, t_{n_{cens}}^*, t_i\}$	$(1-p_i)p_i^{n_{cens}}$

where $p_i = \exp(-\rho\theta_i)$.
So the probability under each model M_i of the SMTS space becomes

$$\begin{aligned}\Pr_{\boldsymbol{\theta}_i}^{M_i}\left(\mathcal{X}^{SMTS}\right) &= (1-p_i) + (1-p_i)p_i + \ldots + (1-p_i)p_i^{n_{cens}} \\ &= (1-p_i)\left[1 + p_i + p_i^2 + \ldots + p_i^{n_{cens}}\right].\end{aligned}$$

As $n \to \infty$, the expression in square brackets is a geometric series with common ratio p_i and the number of censored observations n_{cens} tends to $n\exp(-\rho\theta_i) = np_i$. Recalling that, for a geometric series of common ratio ϕ

$$\sum_{j=0}^{\infty}\phi^j = \frac{1}{1-\phi},$$

then

$$\Pr_{\boldsymbol{\theta}_i}^{M_i}\left(\mathcal{X}^{SMTS}\right) \to (1-p_i)\frac{1}{(1-p_i)} = 1,$$

where $\phi = p_i$.

□

Appendix B

When working with simulated data following a given distribution, it is interesting to see how censoring times, following the same distribution (with different parameters), can be obtained.
Here we show how this can be obtained in the case of the Weibull and log-normal models.

B.1 Weibull censoring times

Suppose to have survival times following a Weibull distribution, $Y_i \sim Weibull(\alpha, \lambda_i)$, for $i = 1, \ldots, n$, and to assign a Weibull distribution to the censoring times $C_i \sim Weibull(\alpha, \beta_i)$. The question is how to choose β_i depending on α and λ_i, given a censoring percentage p_{cens}.

$$\begin{aligned} p_{cens} &= \Pr(Y > C) = \int_0^\infty \Pr(Y > c) f(c) dc \\ &= \int_0^\infty \left(1 - (1 - \exp(-\lambda_i c^\alpha))\right) \beta_i \alpha c^{\alpha-1} \exp(-\beta_i c^\alpha) dc \\ &= \int_0^\infty \exp(-\lambda_i c^\alpha) \beta_i \alpha c^{\alpha-1} \exp(-\beta_i c^\alpha) dc \\ &= \frac{\beta_i}{\beta_i + \lambda_i} \int_0^\infty (\beta_i + \lambda_i) \alpha c^{\alpha-1} \exp(-(\beta_i + \lambda_i) c^\alpha) dc. \end{aligned}$$

Then, we obtain

$$p_{cens} = \frac{\beta_i}{\beta_i + \lambda_i}$$

so

$$\beta_i = \frac{\lambda_i p_{cens}}{1 - p_{cens}}$$

B.2 Normal censoring times

We now consider survival times following a normal distribution, $Y_i \sim N(\mu_1, \sigma_1^2)$, for $i = 1, \ldots, n$, and we assign a normal distribution to the censoring times $C_i \sim N(\mu_2, \sigma_2^2)$. For simplicity, we consider $\sigma_2 = \sigma_1$. In order to obtain the mean μ_2 of the censoring times C_i, we refer to the theory of *stress-strength models* (see Weerahandi and Johnson (1992) for details) in which a unit of strength Y is subject to a stress C. In our simulation study, we have $\mu_1 = \mu + \boldsymbol{\gamma}^\intercal \boldsymbol{x}_i$ (see Section 1.5.1) and $\sigma_1 = \sigma_2 = 1$.
Then we obtain

$$p_{cens} = \Pr(Y > C) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

$$p_{cens} = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{2}}\right)$$

and

$$\mu_2 = \mu_1 - \sqrt{2}\Phi^{-1}(p_{cens}).$$

So we construct the variable C which follows a normal distribution $N(\mu_2, \sigma_2)$. Y_i is simulated from a $N(\mu_1, \sigma_1)$ and it is labelled as censored if $Y_i > C_i$.

Appendix C

C.1 Proof of Proposition 4

We now give the proof of Proposition 4.

Proof. The marginal fractional distributions for the two models are

$$m_{0,b}^N(\boldsymbol{y}) = \left(\theta_0^{n_u} \exp\left(-\theta_0 \sum_{i=1}^n y_i\right)\right)^{n_t/n} = \theta_0^{\frac{n_t n_u}{n}} \exp\left(-\theta_0 \frac{n_t n_u}{n\hat{\theta}}\right)$$

and

$$m_{1,b}^N(\boldsymbol{y}) = \int_0^\infty \frac{1}{\theta}\left(\theta^{n_u} \exp\left(-\theta \sum_{i=1}^n y_i\right)\right)^{n_t/n} d\theta = \int_0^\infty \theta^{\frac{n_t n_u}{n}-1} \exp\left(-\theta \frac{n_t n_u}{n\hat{\theta}}\right) d\theta = \frac{\Gamma(\frac{n_t n_u}{n})}{\left(\frac{n_t n_u}{n\hat{\theta}}\right)^{\frac{n_t n_u}{n}}},$$

where $\hat{\theta}$ denotes the maximum likelihood estimator under the exponential model M_1. The two marginal distributions are

$$m_0^N(\boldsymbol{y}) = \theta_0^{n_u} \exp\left(-\theta_0 \frac{n_u}{\hat{\theta}}\right)$$

and

$$m_1^N(\boldsymbol{y}) = \int_0^\infty \frac{1}{\theta}\theta^{n_u} \exp\left(-\theta \frac{n_u}{\hat{\theta}}\right) d\theta = \frac{\Gamma(n_u)}{(\frac{n_u}{\hat{\theta}})^{n_u}}$$

The fractional part of the FBF is

$$B_{10}^b(\boldsymbol{y}) = \frac{\Gamma(\frac{n_t n_u}{n})}{\left(\frac{n_t n_u}{n\hat{\theta}}\right)^{\frac{n_t n_u}{n}}} \frac{1}{\theta_0^{\frac{n_t n_u}{n}} \exp\left(-\theta_0 \frac{n_t n_u}{n\hat{\theta}}\right)}$$

while the $B_{10}^N(\boldsymbol{y})$ calculated over the entire likelihood is

$$B_{10}^N(\boldsymbol{y}) = \frac{\Gamma(n_u)}{\left(\frac{n_u}{\hat{\theta}}\right)^{n_u}} \frac{1}{\theta_0^{n_u} \exp\left(-\theta_0 \frac{n_u}{\hat{\theta}}\right)},$$

where $n_u = n \times w$ with $w = 1 - p_{cens}$.
Then, by applying the (2.17) we obtain the fractional prior for a fixed n_t

$$\pi^{FI,n_t}(\theta) = \frac{(\theta_0 n_t w)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w - 1} \exp\left(-\frac{\theta_0 n_t w}{\theta}\right) \sim InvGamma(\alpha = n_t w, \beta = \theta_0 n_t w). \tag{C.1}$$

If we want to calculate the fractional prior for the mFBF, let

$$B_{10}^{F,b} = B_{10}^{N}(\boldsymbol{y}) CF_{01}(\boldsymbol{y}) = B_{10}^{N}(\boldsymbol{y}) \sum_{n_t=s}^{s+n_{cens}} \Pr_{N_t}(N_t = n_t) B_{01}^{b}(\boldsymbol{y}) \tag{C.2}$$

be the mFBF, where $CF_{01}(\boldsymbol{y})$ is the correction factor.
From Lemma 1, follows

$$\begin{aligned} \lim_{n\to\infty} \Pr_{N_t}(N_t = n_t) B_{01}^{b}(\boldsymbol{y}) &= \binom{n_t - 1}{s-1} w^s (1-w)^{n_t - s} \lim_{n\to\infty} B_{01}^{b}(\boldsymbol{y}) \\ &= \binom{n_t - 1}{s-1} w^s (1-w)^{n_t - s} \frac{(n_t w \theta_0)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w} \exp\left(-\frac{n_t w \theta_0}{\theta}\right). \end{aligned} \tag{C.3}$$

Then we obtain the fractional prior for the mFBF according to the (2.16), (2.17) and (C.3)

$$\begin{aligned} \pi_1^{FI}(\theta) &= \frac{1}{\theta} B_1^*(\theta) \\ &= \frac{1}{\theta} \lim_{n\to\infty} \frac{1}{CF_{10}(\boldsymbol{y})} \\ &= \frac{1}{\theta} \lim_{n\to\infty} CF_{01}(\boldsymbol{y}) \\ &= \frac{1}{\theta} \lim_{n\to\infty} \sum_{n_t=s}^{s+n_{cens}} \Pr_{N_t}(N_t = n_t) B_{01}^{b}(\boldsymbol{y}) \\ &= \frac{1}{\theta} \sum_{n_t=s}^{\infty} \binom{n_t - 1}{s-1} w^s (1-w)^{n_t - s} \frac{(n_t w \theta_0)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w} \exp\left(-\frac{n_t w \theta_0}{\theta}\right) \\ &= \sum_{n_t=s}^{\infty} \binom{n_t - 1}{s-1} w^s (1-w)^{n_t - s} \frac{(n_t w \theta_0)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w - 1} \exp\left(-\frac{n_t w \theta_0}{\theta}\right). \end{aligned} \tag{C.4}$$

In the case of the exponential model, the number of parameters s is equal to 1 and the marginal fractional prior is

$$\pi_1^{FI}(\theta) = \sum_{n_t=1}^{\infty} w(1-w)^{n_t - 1} \frac{(n_t w \theta_0)^{n_t w}}{\Gamma(n_t w)} \theta^{-n_t w - 1} \exp\left(-\frac{n_t w \theta_0}{\theta}\right), \tag{C.5}$$

which is a mixture of Inverse Gamma distributions, with parameters $\alpha = n_t w$ and $\beta = n_t w \theta_0$. □

C.2 Fractional prior and unit information

In this section we show that the fractional prior calculated in Proposition 4 is not a unit information prior.
In Kass and Wasserman (1995) it is defined the unit information prior as a prior having information about parameters $\boldsymbol{\theta}$ equal to the amount of information about these parameters in one observation.

Definition 9. (Unit information prior) *Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ be iid observations coming from a family parametrized by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$. Suppose we want to test two hypothesis*

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \Theta_0 \\ H_1 &: \boldsymbol{\theta} \in \Theta_0^c. \end{aligned} \tag{C.6}$$

The prior distribution on $\boldsymbol{\theta}$ under the alternative hypothesis H_1, $p(\boldsymbol{\theta})$ is called a unit information prior *if its variance satisfies the following*

$$|\Sigma_{\boldsymbol{\theta}}|^{-1} = \left|I(\boldsymbol{\theta})\right|\Big|_{\boldsymbol{\theta}\in\Theta_0},$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix.

We now prove that the fractional prior for the mFBF introduced in Proposition 4 is not a unit information prior.
Recall that we want to compare the two exponential models

$$\begin{aligned} M_0 &: \theta = \theta_0 \\ M_1 &: \theta \neq \theta_0. \end{aligned}$$

The Fisher information for a generic distribution $f(y|\theta)$ is

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2}\log L(y|\theta)\right],$$

where $L(y|\theta)$ is the likelihood function. The likelihood for the right censored exponential model is

$$\begin{aligned} L(y|\theta) &= f(y|\theta)^{\delta(y)} S(y|\theta)^{1-\delta(y)} \\ &= (\theta\exp(-\theta y))^{\delta(y)} \left(\exp(-\theta y)\right)^{1-\delta(y)} \end{aligned}$$

so the log-likelihood is

$$l(y|\theta) = \log L(y|\theta) = \delta(y)(\log\theta - \theta y) + (1-\delta(y))(-\theta y) = \delta(y)\log\theta - \theta y.$$

Then, the first derivative of the log-likelihood with respect to the parameter θ is

$$\frac{dl(\boldsymbol{y}|\theta)}{d\theta} = \frac{\delta(y)}{\theta} - y$$

and the second derivative is

$$\frac{d^2 l(\boldsymbol{y}|\theta)}{d\theta^2} = -\frac{\delta(y)}{\theta^2}.$$

We obtain

$$-E\left[\frac{d^2}{d\theta^2}\log L(\boldsymbol{y}|\theta)\right] = E\left(\frac{\delta(y)}{\theta^2}\right) = \frac{1}{\theta^2}E(\delta(y)) = \frac{w}{\theta^2},$$

where $w = 1 - p_{cens}$, and

$$\Sigma(\theta) = \frac{\theta^2}{w}\bigg|_{\theta=\theta_0} = \frac{\theta_0^2}{w}.$$

Next, we calculate the mean and the variance of the fractional prior (C.5). If $n_t w > 1$ for each n_t, then

$$E_{\pi_1^{FI}}(\theta) = \sum_{n_t=1}^{\infty} w(1-w)^{n_t-1} E_{InvGa}(\theta) = \sum_{n_t=1}^{\infty} w(1-w)^{n_t-1}\frac{n_t w\theta_0}{n_t w - 1}. \tag{C.7}$$

If $n_t w > 2$ for each n_t, then

$$\begin{aligned} Var_{\pi_1^{FI}}(\theta) &= E_{\pi_1^{FI}}(\theta^2) - \left(E_{\pi_1^{FI}}(\theta)\right)^2 \\ &= \sum_{n_t=1}^{\infty} w(1-w)^{n_t-1}\frac{n_t^2 w^2\theta_0^2}{(n_t w-1)(n_t w-2)} - \left(\sum_{n_t=1}^{\infty} w(1-w)^{n_t-1}\frac{n_t w\theta_0}{n_t w-1}\right)^2. \end{aligned} \tag{C.8}$$

Observe that when $n_t w$ is small, in particular if $n_t w < 2$, the variance of this prior, also called *marginal prior*, does not exist. Unfortunately, this is the most probable case, because the smallest values of n_t are usually the most probable (see Figure 3.1). So, we can conclude that the marginal prior is a vague prior and, clearly, it is not a unit information prior.