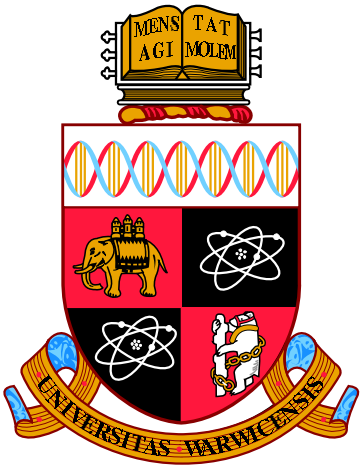


Incorporating unobserved heterogeneity and multiple event types in survival models: A Bayesian approach

by

Catalina A. Vallejos

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

March 2014

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Words might fail expressing my gratitude to all those who supported me throughout this trip. In the first place, I would like to thank Prof. Mark F.J. Steel, my supervisor, for these three and a half years of advice, encouragement and understanding. I can truly say how much I enjoyed our work and that, thanks to him, I am leaving Warwick as a better researcher. I would also like to thank Prof. Jane Hutton and Dr. Elke Thönnes for invaluable discussions throughout my PhD studies and to Dr. Marco A.R. Ferreira for his constructive comments about the contents of Appendix B. Many thanks to Prof. Jon Forster and Prof. Gareth Roberts, my revision panel, for an enjoyable discussion and invaluable feedback.

My gratitude to P.O.D. Pharoah and Prof. Jane Hutton for the access to the cerebral palsy dataset. I cannot fail to be grateful for the support of Lorena Correa and Prof. Guillermo Marshall, especially by motivating the survival analysis of university outcomes presented in Chapter 5. I also thank Valeria Leiva for her assistance while accessing the dataset.

Many thanks to the University of Warwick who funded this research via the Warwick Postgraduate Research Scholarship. I am also grateful for the complementary funding support provided by the Departments of Statistics at the Pontificia Universidad Católica de Chile and the University of Warwick.

Finally, but not least important, I also want to thank Petros and my dearest friends. Thank you all for being always there when I needed you the most. I would like to dedicate this thesis to my parents, Alicia and Carlos, who always trust in me and give me their unconditional support.

Declarations

This thesis is a result of my own work, which was performed between October of 2010 and March of 2014 under the supervision of Prof. Mark F.J. Steel. The materials in this thesis are original, except in those cases where indicated by references. No contents of this thesis have been submitted for examination to any other degree at the University of Warwick or other institutions.

Most materials in Chapters 1, 2, 3 and 4 and Appendices A, C and D are a compilation of

- Vallejos, C.A. and Steel, M.F.J. (2014). *Objective Bayesian Survival Analysis Using Shape Mixtures of Log-Normal Distributions.* To appear at *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2014.923316.

- Vallejos, C.A. and Steel, M.F.J. (2014). *Incorporating unobserved heterogeneity in Weibull survival models: A Bayesian approach.* To be submitted for publication shortly after the submission of this dissertation.

In addition, Appendix B is based on

- Vallejos, C.A. and Steel, M.F.J. (2013). *On posterior propriety for the Student-t linear regression model under Jeffreys priors.* http://arxiv.org/abs/1311.1454.

Finally, materials in Chapter 5 are compiled in

- Vallejos, C.A. and Steel, M.F.J. (2014). *Bayesian Survival Modelling of University Outcomes.* Under review.

Abstract

This thesis covers theoretical and practical aspects of Bayesian inference and survival analysis, which is a powerful tool for the analysis of the time until a certain event of interest occurs. This dissertation focuses on non-standard models inspired by features of real datasets that are not accommodated by conventional models.

Materials are divided in two parts. The first and more extended part relates to the development of flexible parametric lifetime distributions motivated by the presence of anomalous observations and other forms of unobserved heterogeneity. Chapter 2 presents the use of mixture families of lifetime distributions for this purpose. This idea can be interpreted as the introduction of an observation-specific random effect on the survival distribution. Two families generated via this mechanism are studied in Chapter 3. Covariates are introduced through an accelerated failure times representation, for which the interpretation of the regression coefficients is invariant to the distribution of the random effect. The Bayesian model is completed using reasonable (improper) priors that require a minimum input from practitioners. Under mild conditions, these priors induce a well-defined posterior distribution. In addition, the mixture structure is exploited in order to propose a novel method for outlier detection where anomalous observations are identified via the posterior distribution of the individual-specific random effects. The analysis is illustrated in Chapter 4 using three real medical applications.

Chapter 5 comprises the second part of this thesis, which is motivated in the context of university outcomes. The aim of the study is to identify determinants of the length of stay at university and its associated academic outcome for undergraduate students of the Pontificia Universidad Católica de Chile. In this setting, survival times are defined as the time until the end of the enrollment period, which can relate to different reasons - graduation or dropout - that are driven by different processes. Hence, a competing risks model is employed for the analysis. Model uncertainty is handled through Bayesian model averaging, which leads to a better predictive performance than choosing a unique model. The output of this analysis does not account for all features of this complex dataset yet it provides a better understanding of the problem and a starting point for future research.

Finally, Chapter 6 summarizes the main findings of this work and suggests future extensions.

Abbreviations

AA	Autologous and Allogeneic
AFT	Accelerated failure times
BF	Bayes factors
BMA	Bayes factors
CDF	Cumulative distribution function
CPO	Conditional predictive ordinate
DIC	Deviance information criteria
ESS	Effective sample size
FIM	Fisher information matrix
HPD	Highest probability density
MCMC	Markov chain Monte Carlo
PDF	Probability density function
PH	Proportional hazards
PO	Proportional odds
PsBF	Pseudo Bayes factors
PsML	Pseudo marginal likelihood
PUC	Pontificia Universidad Católica de Chile
PVF	Power Variance Function
VA	Veteran's Administration

Notation

$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Positive real numbers
$\mathbb{R}^p$	p-dimensional space of real numbers
$\Phi(\cdot)$	CDF of a standard normal random variable
$\stackrel{d}{=}$	Equality in distribution
$I(A)$	Indicator function of the statement A being true

Chapter 1

Introduction

> "...the past and the future formed part of a single unit, and the reality of the present was a kaleidoscope of jumbled mirrors where everything and anything could happen".
> Isabel Allende
> *The house of the spirits*

The use of survival methods had an important growth during the last few decades. A possible explanation is the availability of ready-to-use software, which makes sophisticated techniques accessible to applied users. However, this popularity is also justified by a wider range of applications. Whereas *survival analysis* was originally motivated in a medical setting, nowadays, other disciplines are making use of its strengths. Depending on the context, developments have been made under different names. For instance, engineers refer to it as *reliability analysis*. In economics, it is renamed as *duration analysis*. *Event history analysis* is often the choice in other social sciences. Regardless of the label, the objective is the same: to model or predict the time until a certain event of interest occurs. Perhaps, the canonical example is in clinical trials, where the event is usually defined as the relapse, recovery or death of a patient. Other examples include the time to failure of a system and the amount of time that a graduate spends searching for a first job.

One proof of this increasing popularity is the large number of books dedicated to the topic [*e.g.* Cox and Oakes, 1984; Klein and Moeschberger, 1997; Ibrahim et al., 2001; Kalbfleisch and Prentice, 2002; Collett, 2003]. Standard models vary from simple parametric models (*e.g.* exponential survival times) to more complex parametric structures and non-parametric or semi-parametric extensions [*e.g.* the well-known Cox proportional hazards model presented in Cox, 1972]. In this dissertation, the main focus is on parametric models and on conducting Bayesian inference

Table 1.1: Functions that characterize a lifetime distribution.

Density	$f_T(t)$
Distribution	$F_T(t) = P(T < t) = \int_0^t f_T(s)\, ds$
Survival	$S_T(t) = P(T \geq t) = \int_t^\infty f_T(s)\, ds = 1 - F_T(t)$
Hazard	$h_T(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t+\Delta \mid T \geq t)}{\Delta} = \frac{f_T(t)}{S_T(t)} = -\frac{d}{dt} \log(S_T(t))$

with them.

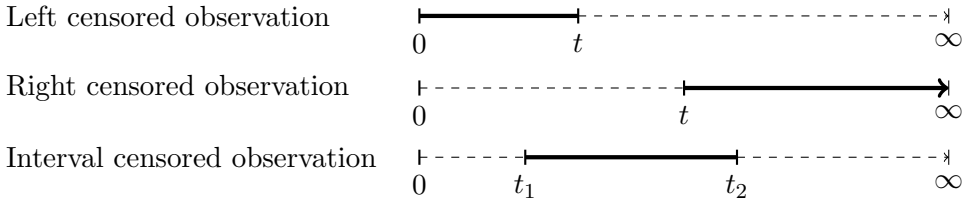
This introductory Chapter provides a framework for the methodology presented throughout this thesis. Firstly, Section 1.1 briefly introduces some of the main concepts in survival analysis. Section 1.2 relates to Bayesian inference for survival models, including implementation issues and standard Bayesian model comparison criteria. In particular, Subsection 1.2.2 highlights that the use of point observations under continuous sampling can affect the existence of the posterior distribution and considers a solution through set observations for this problem. Regression models for survival data are introduced in Section 1.3. Section 1.4 summarizes the main contributions that this thesis adds to existent literature. An outline of subsequent chapters concludes Chapter 1. All the proofs are contained in Appendix A without mention in the text.

1.1 Standard setting for survival analysis

Let T be a positive-valued random variable representing the *time-to-event* for an individual (or unit). It is usually called survival or failure time. On a first stage, T is assumed to have a continuous nature but discrete times are considered in Chapter 5. A model for T can be specified via any of the functions defined in Table 1.1. For easy of notation, these definitions ignore possible parameters associated to the lifetime distribution. In particular, $S_T(t)$ represents the probability of observing no event for the individual before time t. The hazard function $h_T(t)$ is defined as the instantaneous rate of failure at time t, given that no event has been observed before.

A distinct feature of survival analysis is the ability to deal with censored observations. Censoring appears when, because of limited time or resources, it is not possible to observe the exact survival time and only some bounds for the actual lifetime are available. Censoring must be taken into account when conducting inference. It is possible to distinguish between three types of censoring. These are illustrated in Figure 1.1. *Right censoring* is frequently encountered in survival datasets. It occurs when only a lower bound for T is known. For instance, when the event of interest has not yet happen by the end of a fixed observation period.

Figure 1.1: Graphical representation of censored observations.



On the other hand, *left censoring* is less often seen. In such a case, the record consists of an upper limit for T. This might happen, for example, when the event of interest already occurred before the first screening time. Finally, *interval censoring* is a combination of the previous schemes where a lower and upper bound for T are reported (*e.g.* the event took place between two consecutive inspection times and it is not possible to identify the exact moment). Censoring is assumed to be non-informative throughout this thesis, mainly focusing on right censoring. Define

$$c_i \quad = \quad \begin{cases} 0, & \text{if the observation } i \text{ is non-censored,} \\ 1, & \text{if the observation } i \text{ is right censored,} \\ 2, & \text{if the observation } i \text{ is left censored,} \\ 3, & \text{if the observation } i \text{ is interval censored.} \end{cases} \tag{1.1}$$

A survival dataset contains both, the recorded lifetimes (possibly censored) and the corresponding censoring indicators (when $c_i = 3$, $t_i = (t_{i1}, t_{i2})$ is recorded). In an abuse of notation, define $T = (T_1, \ldots, T_n)'$ as a vector containing the survival times of n independent individuals. In the presence of censoring, if $T = t$ is recorded, a general expression for the associated likelihood function is given by

$$L_T(t; c) = \prod_{i=1}^{n} [f_{T_i}(t_i)]^{I(\{c_i=0\})} [S_{T_i}(t_i)]^{I(\{c_i=1\})} [F_{T_i}(t_i)]^{I(\{c_i=2\})} [F_{T_i}(t_{i2}) - F_{T_i}(t_{i1})]^{I(\{c_i=3\})}. \tag{1.2}$$

Choosing a parametric model for the survival times is not a trivial task. The survival literature includes a large number of lifetime distributions. For instance, Marshall and Olkin [2007] compiles a comprehensive list of parametric models and their properties. One aspect to be considered when selecting a parametric model is the behaviour of its hazard function. If the risk of the event is expected to be constant over time, a simple exponential model is a reasonable option. Non-monotonic hazard trajectories are accommodated by more flexible models such as the Weibull

or the log-normal ones. Physical or theoretic reasons might also motivate the choice of a model. For example, in a reliability context, Owen and Padgett [1999] pointed out that cumulative damages can be represented in an additive or a multiplicative fashion, in which a failure will be observed when the cumulative damage exceeds certain threshold. Their argument can be extended to other backgrounds (*e.g.* continued losses of a firm can lead to bankruptcy; high levels of arsenic ingested in drinking water can produce liver damage). The log-normal distribution arises as the limiting distribution of an additive damage scheme [Crow and Shimizu, 1988]. The Birnbaum-Saunders [Birnbaum and Saunders, 1969] is its counterpart for a multiplicative damage model.

1.2 Bayesian inference

Assume the distribution of the survival times depends on a parameter Ψ with support $\mathfrak{F}$ (often Ψ contains a vector of regression parameters plus scale and/or shape parameters). Classical inference assumes Ψ is a fixed but unknown quantity. In contrast, the Bayesian approach considers it as a random magnitude. Uncertainty about Ψ is represented in terms of (probability) measures, which are defined as subjective degrees of belief. Prior to the observation of the data, previous knowledge about Ψ is summarized into a so-called *prior distribution* $\pi_\Psi(\psi)$. Once data has been observed, prior beliefs are updated via Bayes theorem [Bayes, 1763]. The so-called *posterior distribution* of Ψ given the observed data corresponds to

$$\pi_\Psi(\psi|T=t;c) = \frac{L_T(t|\Psi=\psi;c)\pi_\Psi(\psi)}{L_T(t;c)}, \tag{1.3}$$

where $L_T(t|\Psi=\psi;c)$ is the likelihood function given a fixed value ψ of Ψ (as in (1.2)) and $L_T(t;c)$ corresponds to the marginal likelihood (after integrating out ψ), defined as

$$L_T(t;c) = \int_{\mathfrak{F}} L_T(t|\Psi=\psi;c)\pi_\Psi(\psi)\,d\psi. \tag{1.4}$$

All inferences about Ψ are based on its posterior distribution. Hereafter, $\pi(\psi)$, $\pi(\psi|t;c)$, $L(t;c)$ and $L(t|\psi;c)$ will be used instead of $\pi_\Psi(\psi)$, $\pi_\Psi(\psi|T=t;c)$, $L_T(t;c)$ and $L_T(t|\Psi=\psi;c)$, respectively. In addition, no differentiation between Ψ and ψ will be made throughout the text.

1.2.1 Jeffreys priors

The choice of a prior distribution is a challenging task. If reliable prior information is available, such beliefs can be used in order to construct a prior distribution. The Bayesian literature refers to this process as *prior elicitation*. Details and guidance about this procedure are provided in O'Hagan et al. [2006]. Nonetheless, the (frequently encountered) setting of prior ignorance precludes the elicitation of a prior distribution on the basis of prior information. Alternative *non-informative* or *objective priors* (based on formal mathematical rules rather than in prior knowledge) can be used in such a situation. These priors attempt to minimize the influence of the prior over posterior inference and to provide baseline comparison when actual prior knowledge exists [Bernardo and Smith, 2000]. In this context, one of the most popular choices is the *Jeffreys prior* [Jeffreys, 1946, 1961], defined as the square root of the determinant of the Fisher information matrix (FIM). Jeffreys [1961] also proposed the *independence Jeffreys prior*, a variation that deals separately with blocks of the FIM. These priors do not require the elicitation of hyper-parameters, providing an attractive tool to applied users. However, as Jeffreys-style priors do not always correspond to proper probability density functions, their use requires careful consideration.

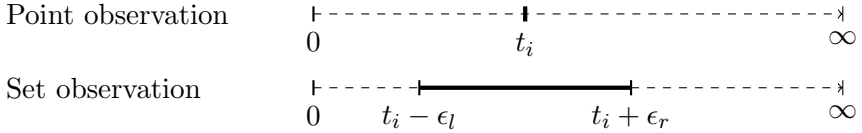
1.2.2 Posterior propriety and the use of point observations under continuous sampling

Posterior inference is well-defined as long as the marginal likelihood $L(t;c)$ is finite (see equations (1.3) and (1.4)). This condition is not a major drawback when proper prior distributions are in use (where $L(t;c)$ is always finite with probability one). However, improper priors may lead to an infinite marginal likelihood, preventing a meaningful Bayesian analysis. Hence, posterior propriety has to be verified in order to validate posterior inferences.

Censoring must be taken into account when conducting Bayesian inference for survival datasets. Nonetheless, the following proposition states that adding censored observations cannot destroy the propriety of the posterior distribution. Ignoring censored observations leads to sufficient conditions for posterior existence.

Proposition 1. *Let $t = (t_1, \ldots, t_n)'$ be the recorded survival times of n independent individuals, realizations of random variables with survival function $S_{T_i}(t_i|\psi)$ $(i = 1, \ldots, n)$. Without loss of generality, assume only the first n_o observations are uncensored ($n_o \leq n$) and denote by t_o the vector containing all uncensored observations. A sufficient condition for the existence of $\pi(\psi|t;c)$ is the propriety of*

Figure 1.2: Graphical representation of point and set observations.



$\pi(\psi|t_o)$.

Posterior propriety verification is usually conducted without taking into account events that have zero probability of being observed. In fact, standard checks only assess if $L(t;c)$ is finite with probability one. This situation can cause problems when conducting Bayesian inference under continuous sampling. Continuous models assign zero probability to particular (point) values. In spite of this, conventional statistical analysis is based on point observations. Hence, the propriety of the posterior distribution can be destroyed when a specific sample of point observations t_0 is observed. As argued in Fernández and Steel [1998], this issue introduces the risk of having senseless inference. Theorem 6 (Subsection 3.2.3) reveals an example for which using point observations is a liability.

In the context of scale mixture of normals, Fernández and Steel [1998]; Fernï¿½ndez and Steel [1999] proposed the use of set observations as a solution to this problem. This idea is based on the fact that, in practice, it is impossible to record realizations of continuous random variables with total precision. Each observation can only be considered as a label of a set with positive Lebesgue measure. In fact, a point observation t_i only indicates that the actual survival time is between $t_i - \epsilon_l$ and $t_i + \epsilon_r$, where ϵ_l and ϵ_r are determined by the accuracy with which the data was recorded (*e.g.* if the data is recorded in integers, $\epsilon_l = \epsilon_r = 0.5$). The latter has an easy interpretation in survival data. As illustrated in Figure 1.2, such a set observation is an interval censored record on $(t_i - \epsilon_l, t_i + \epsilon_r)$. In the same spirit, right censored observations are themselves set observations (without the need of ϵ_l and ϵ_r). As shown by Theorem 1, set observations can ensure a proper posterior distribution in situations where a particular sample of point observations might not.

Theorem 1. *Adopt the same assumptions as in Proposition 1. Denote by t_c the $n - n_o$ censored observations. Replace the uncensored observations by set observations $t_\epsilon = \{(t_1 - \epsilon_l, t_1 + \epsilon_r), \ldots, (t_{n_o} - \epsilon_l, t_{n_o} + \epsilon_r)\}$ $(0 < \epsilon_l, \epsilon_r < \infty)$. Define $E = (t_1 - \epsilon_l, t_1 + \epsilon_r) \times (t_2 - \epsilon_l, t_2 + \epsilon_r) \times \cdots \times (t_{n_o} - \epsilon_l, t_{n_o} + \epsilon_r)$. The posterior distribution of ψ given (t_ϵ, t_c) is proper if and only if the marginal likelihood under point observations*

(t_o, t_c) is finite for any $t_o \in E$, excluding a set of zero Lebesgue measure.

1.2.3 Implementation of posterior inference

Conjugate priors (for which prior and posterior belong to the same parametric family) produce well-known and tractable posterior distributions. In contrast, under more general priors, the posterior distribution is usually known only up to a normalization constant $L(t; c)$. This is often the case when Jeffreys-style priors are in use. Exact posterior inference is not possible in those cases, yet Markov Chain Monte Carlo (MCMC) methods make Bayesian inference feasible. The general strategy is to generate a Markov chain whose stationary distribution is $\pi(\psi|t; c)$ [Bernardo and Smith, 2000]. Once the sampler converges to the equilibrium distribution, draws generated via this mechanism can be used in order to estimate features of $\pi(\psi|t; c)$. An initial *burn-in* period (before convergence) of iterations is normally discarded for this purpose. The Markov structure induces correlation between the MCMC draws. Strong autocorrelations show evidence of a poor mixing as the chain will explore the parameter space slowly. Let $\psi = (\psi_1, \ldots, \psi_J)'$. For each element of ψ, the Effective Sample Size (ESS) is defined as

$$\mathrm{ESS}(\psi_j) = \frac{M}{1 + 2\sum_{m=1}^{\infty} \rho_m(\psi_j)}, \tag{1.5}$$

where M is the total number of iterations in use (after burn-in) and $\rho_m(\psi_j)$ represents the autocorrelation function of lag m between the draws of ψ_j. The $\mathrm{ESS}(\psi_j)$ quantifies the number of independent samples to which the chain of ψ_j is equivalent. It can be larger than M if negative autocorrelations are observed. In the presence of strong positive autocorrelations, storage space can be saved by introducing a *thinning* period (*i.e.* only storing draws every certain number of iterations).

The Bayesian literature includes several approaches for assessing the convergence of a chain. A first, intuitive, idea is to run various independent chains using different (disperse) starting values. Under stationary, these chains should exhibit very similar behaviour. This can be informally assessed using the trace plots of the chains. For the numerical examples in this document, two formal convergence diagnostics are applied to MCMC chains (after burn-in and thinning). Both of them are available in standard statistical software. Firstly, the test proposed in Geweke [1992] compares the means of the first 10% and the last 50% of the chain. If both means differ substantially, the chain has not yet reached stationarity. The second diagnostic, introduced in Heidelberger and Welch [1983], uses the Cramer-von-Mises statistic in order to assess lack of convergence. If extra burn-in is required, the test

reports the number of iterations that should be discarded.

The complexity of the implementation might be affected by the presence of censoring (*e.g.* if the survival function has no closed analytical form, as in the log-normal case). However, censored observations can be accommodated through the idea of data augmentation [Tanner and Wong, 1987]. This introduces an additional step in the sampler in which, given the current value of the parameters, point values of the survival times in line with the censoring are simulated. Given these values, the rest of the sampler acts as if there were not censoring. The latter also applies when replacing non censored observations by set observations (see Subsection 1.2.2). Nevertheless, even in the absence of censoring, direct sampling from $\pi(\psi|t;c)$ can be a cumbersome task. The following algorithm often provides a simple solution.

Gibbs sampling [Geman and Geman, 1984]

Define

$$\pi(\psi_j|\psi_{-j},t;c), \quad \psi_{-j}=(\psi_1,\ldots,\psi_{j-1},\psi_{j+1},\ldots,\psi_J)', \quad j=1,\ldots,J, \tag{1.6}$$

as the set of *full conditionals* for $\{\psi_1,\ldots,\psi_J\}$. A Markov chain $\left\{\psi^{(0)},\psi^{(1)},\ldots\right\}$ is generated via the following mechanism. Given an initial guess $\psi^{(0)}=(\psi_1^{(0)},\ldots,\psi_J^{(0)})'$, at the iteration m of the chain

$$\begin{array}{lll}
\text{sample} & \psi_1^{(m+1)} & \text{from } \pi(\psi_1|\psi_2^{(m)},\ldots,\psi_J^{(m)},t;c),\\
\text{sample} & \psi_2^{(m+1)} & \text{from } \pi(\psi_2|\psi_1^{(m+1)},\psi_3^{(m)},\ldots,\psi_J^{(m)},t;c),\\
 & \vdots & \\
\text{sample} & \psi_J^{(m+1)} & \text{from } \pi(\psi_J|\psi_1^{(m+1)},\ldots,\psi_{J-1}^{(m+1)},t;c).
\end{array}$$

For large m, the distribution of $\psi^{(m)}$ converges to $\pi(\psi|t;c)$. If all the full conditionals have a known form, the implementation of a Gibbs sampler is straightforward. Otherwise, if sampling from $\pi(\psi_j|\psi_{-j},t;c)$ is troublesome, stochastic simulation techniques can be used within a Gibbs sampler. Some common examples of this are described below.

Rejection sampling [Devroye, 1986]

Let $g(\cdot)$ be a probability density function such that $\pi(\psi_j|\psi_{-j},t;c)\leq Ag(\psi_j)$ for all possible values of ψ_j and a constant value of A ($A>1$). This method relies on the ability of generating random samples from $g(\cdot)$. Drawings from $\pi(\psi_j|\psi_{-j},t;c)$ are generated via the following mechanism.

1. Sample $\upsilon \sim \text{Unif}(0,1)$ and a candidate ψ_j^* from $g(\cdot)$.

2. If $\upsilon A g(\psi_j^*) \leq \pi(\psi_j^*|\psi_{-j}, t; c)$, return ψ_j^*.

3. Otherwise, reject ψ_j^* and repeat from step 1 until a candidate is accepted.

Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970]

Given a current status $\psi_j^{(m)}$ define $q(\psi_j^{(m)}, \cdot)$ as a transition PDF. A sample whose equilibrium distribution is $\pi(\psi_j|\psi_{-j}, t; c)$ is obtained as follows.

1. Sample $\upsilon \sim \text{Unif}(0,1)$ and a candidate ψ_j^* from $q(\psi_j^{(m)}, \cdot)$.

2. Define

$$a(\psi_j^{(m)}, \psi_j^*|\psi_{-j}, t; c) = \min\left\{1, \frac{\pi(\psi_j^*|\psi_{-j}, t; c)}{\pi(\psi_j^{(m)}|\psi_{-j}, t; c)} \frac{q(\psi_j^*, \psi_j^{(m)})}{q(\psi_j^{(m)}, \psi_j^*)}\right\}. \tag{1.7}$$

3. If $\upsilon \leq a(\psi_j^{(m)}, \psi_j^*|\psi_{-j}, t; c)$, return ψ_j^*. Otherwise, return $\psi_j^{(m)}$.

A common choice for $q(\psi_j^{(m)}, \cdot)$ is a Normal$(\psi_j^{(m)}, \omega^2)$ distribution. The literature often refers to this as a Gaussian Random Walk Metropolis-Hastings algorithm. The value of ω^2 should be tuned in order to achieve an optimal acceptance rate [Roberts and Rosenthal, 2001]. This can be tedious and time consuming (it requires to re-run the algorithm several times using different values of ω^2). Alternatively, the adaptive Metropolis-Hastings algorithm detailed in Section 3 of [Roberts and Rosenthal, 2009] can be used. The latter provides an automated tuning process for the variance of the proposal distribution. The combination of a Gibbs sampling scheme and (adaptive) Metropolis-Hastings updates is often called *(adaptive) Metropolis-within-Gibbs*.

1.2.4 Bayesian model comparison

The following Bayesian model comparison criteria are used throughout this thesis.

Bayes Factors (BF) [Jeffreys, 1935; Kass and Raftery, 1995]

In a way that is totally coherent with the Bayesian paradigm, BF compares two models M_0 and M_1 in terms of their prior and posterior odds. The BF against of M_0 and in favour of M_1 is defined as

$$\text{BF}_{10} = \frac{\pi(M_1|t; c)}{\pi(M_0|t; c)} \Big/ \frac{\pi(M_1)}{\pi(M_0)} = \frac{L_1(t; c)}{L_0(t; c)}, \tag{1.8}$$

Table 1.2: Kass and Raftery [1995] rule for the interpretation of Bayes factors.

$2\log_e(B_{10})$	B_{10}	Evidence against M_0
$0-2$	$1-3$	Not worth more than a bare mention
$2-6$	$3-20$	Positive
$6-10$	$20-150$	Strong
>10	>150	Very strong

where $\pi(M_0)$, $\pi(M_1)$, $\pi(M_0|t;c)$ and $\pi(M_1|t;c)$ are corresponding prior and posterior probabilities associated to each model. The marginal likelihoods $L_0(t;c)$ and $L_1(t;c)$ are defined as in (1.4). This criterion cannot be used in combination with improper priors, unless the improper part of the prior is related to parameters that are shared by both models. Jeffreys [1961] proposed an initial rule for the interpretation of B_{10}. However, Kass and Raftery [1995] introduced some modifications in order to have more accurate results. Their interpretation rule is summarized in Table 1.2. Computing marginal likelihoods is a very challenging endeavour. A survey of several methods is provided in Section 7.3 of Robert [2007]. In particular, two approaches are employed throughout this thesis: the Bridge sampling proposed in Meng and Wong [1996] and the MCMC estimator of [Chib, 1995] and Chib and Jeliazkov [2001], which is based on a Metropolis-within-Gibbs algorithm.

Let $g_0(\cdot)$ and $g_1(\cdot)$ be two densities sharing the same support and that are known only up to proportionality constants c_0 and c_1, respectively. Inspired by the physics literature, Meng and Wong [1996] shown that for any arbitrary *bridge* function $\alpha(\cdot)$ (such that the required expectations exist), it follows that

$$r = \frac{c_1}{c_0} = \frac{\mathrm{E}_{g_0}\left(\tilde{g}_1(\psi)\alpha(\psi)\right)}{\mathrm{E}_{g_1}\left(\tilde{g}_0(\psi)\alpha(\psi)\right)}, \tag{1.9}$$

where $\tilde{g}_0(\cdot)$ and $\tilde{g}_1(\cdot)$ are the known un-normalized versions of $g_0(\cdot)$ and $g_1(\cdot)$, respectively. The expectations in (1.9) are with respect to $g_0(\cdot)$ and $g_1(\cdot)$, respectively. Using (1.9), the *bridge sampling* estimator of c_1/c_0 is defined as

$$\hat{r}_\alpha = \frac{1/n_0 \sum_{i=1}^{n_0} \tilde{g}_1(\psi_{0i})\alpha(\psi_{0i})}{1/n_1 \sum_{i=1}^{n_1} \tilde{g}_0(\psi_{1i})\alpha(\psi_{1i})}, \tag{1.10}$$

where $\psi_{01}, \ldots, \psi_{0n_0}$ and $\psi_{11}, \ldots, \psi_{1n_1}$ are random samples from $g_0(\cdot)$ and $g_1(\cdot)$, respectively. If draws within each of these samples are independent, Meng and Wong [1996] deduced that the variance of $\log(\hat{r_\alpha})$ is minimized when

$$\alpha^*(\psi) \propto \frac{1}{s_1\tilde{g}_1(\psi) + r s_0 \tilde{g}_0(\psi)}, \tag{1.11}$$

where $s_j = n_j/(n_0 + n_1)$, $j = 0, 1$. As discussed in Meng and Schilling [2002], dependencies between the draws are not too critical for this optimization as long as they are weak. Since r is unknown, $\alpha^*(\psi)$ cannot be directly used. Nevertheless, given an initial guess $\hat{r}_\alpha^{(0)}$, an optimal bridge estimator can be defined iteratively as

$$\hat{r}_\alpha^{(m+1)} = \frac{1/n_0 \sum_{i=1}^{n_0} l_{0i}/(s_1 l_{0i} + s_0 \hat{r}_\alpha^{(m)})}{1/n_1 \sum_{i=1}^{n_1} 1/(s_1 l_{1i} + s_0 \hat{r}_\alpha^{(m)})}, \quad m = 1, 2, \ldots \tag{1.12}$$

where $l_{ji} = \tilde{g}_1(\psi_{ji})/\tilde{g}_0(\psi_{ji})$, $i = 1, \ldots, n_j$, $j = 0, 1$. The latter defines a consistent estimator of r. Nonetheless, the method in Meng and Wong [1996] is restrictive in the sense that it requires the same support for $g_0(\cdot)$ and $g_1(\cdot)$. In particular, this condition does not hold when the aim is to estimate the BF between two models M_0 and M_1 which have different number of parameters (*e.g.* in variable selection). As a solution, Chen and Shao [1997] proposed to augment the smaller support, introducing a correction factor in (1.10). Alternatively, Meng and Schilling [2002] suggested a different solution that computes c_0 and c_1 independently. They pointed out that (1.10) defines an estimator of c_1 when $\tilde{g}_0(\psi)$ is replaced by an auxiliary normalized density $g(\psi)$ which has the same support as $g_1(\psi)$. Of course, c_0 can be estimated in an analogous manner.

Another estimator for the marginal likelihood of a given model is proposed in Chib [1995] and Chib and Jeliazkov [2001]. This is defined as

$$\log(\hat{L}(t; c)) = \log(L(t|\hat{\psi}; c)) + \log(\pi(\hat{\psi})) - \log(\hat{\pi}(\hat{\psi}|t; c)), \tag{1.13}$$

where $\hat{\psi}$ denotes a value of ψ with high posterior density and $\hat{\pi}(\psi|t; c)$ is an estimator for the posterior density of ψ. Often, computing $L(t|\hat{\psi}; c)$ and $\pi(\hat{\psi})$ is straightforward. In order to estimate $\pi(\hat{\psi}|t; c)$, Chib [1995] exploits the following decomposition

$$\pi(\hat{\psi}|t; c) = \prod_{j=1}^{J} \pi(\hat{\psi}_j|\hat{\psi}_{(j-1)}, t; c), \tag{1.14}$$

where $\hat{\psi}_{(j-1)} = (\hat{\psi}_1, \ldots, \hat{\psi}_{j-1})'$. For each $j = 1, \ldots, J$, define

$$\hat{\pi}(\hat{\psi}_j|\hat{\psi}_{(j-1)}, t; c) = M^{-1} \sum_{m=1}^{M} \pi(\hat{\psi}_j|\hat{\psi}_{(j-1)}, \psi_{(-j)}^{(m)}, t; c), \tag{1.15}$$

where $\left\{\psi_{(-j)}^{(m)} = (\psi_{j+1}^{(m)}, \ldots, \psi_J^{(m)})', m = 1, \ldots, M\right\}$ are draws from a reduced Gibbs sampler, with fixed $\hat{\psi}_{(j-1)}$. The estimator in (1.15) involves the full conditional of ψ_j (see (1.6)). Chib and Jeliazkov [2001] extends this methodology for when the

draws of ψ_j are generated using a (non-adaptive) Metropolis-Hastings algorithm. In such a case,

$$\hat{\pi}(\hat{\psi}_j|\hat{\psi}_{(j-1)},t;c)=\frac{M^{-1}\sum_{m=1}^{M}a(\psi_j^{(m)},\hat{\psi}_j|\hat{\psi}_{(j-1)},\psi_{(-j)}^{(m)},t;c)q(\psi_j^{(m)},\hat{\psi}_j)}{L^{-1}\sum_{l=1}^{L}a(\hat{\psi}_j,\psi_j^{(l)}|\hat{\psi}_{(j-1)},\psi_{(-j)}^{(l)},t;c)}, \quad (1.16)$$

with $a(\cdot,\cdot|\psi_{-j},t;c)$ defined as in (1.7). In addition, $\left\{(\psi_j^{(m)},\psi_{(-j)}^{(m)})',m=1,\ldots,M\right\}$ and $\left\{\psi_{(-j)}^{(l)},l=1,\ldots,L\right\}$ are draws from reduced Gibbs samplers with fixed $\hat{\psi}_{(j-1)}$ and $\hat{\psi}_{(j)}$, respectively. For each $l=1,\ldots,L$, $\psi_j^{(l)}$ is a draw from the Metropolis-Hastings proposal $q(\hat{\psi}_j,\cdot)$. As shown in Meng and Schilling [2002] and Mira and Nicholls [2004], the estimator in Chib [1995] and Chib and Jeliazkov [2001] is a particular case of bridge sampling.

The former estimator is based on a non-adaptive Metropolis-within-Gibbs algorithm. For the adaptive version, using the stabilized proposal variances, the $L(t;c)$ can be estimated from shorter non-adaptive chains for which the starting values are defined as the converged parameter values of the original chains.

Deviance Information Criteria (DIC)

Introduced by Spiegelhalter et al. [2002], the DIC is defined

$$\text{DIC}\equiv \text{E}(D(\psi,t)|t;c)+p_D=\text{E}(D(\psi,t)|t;c)+[\text{E}(D(\psi,t)|t;c)-D(\hat{\psi},t)], \quad (1.17)$$

where $D(\psi,t)=-2\log(L(t|\psi;c))$ is known as the deviance function, p_D is interpreted as the effective number of parameters of the model and $\hat{\psi}$ is an estimated value of ψ (*e.g.* the posterior mean or median). The expectation on (1.17) is with respect to $\pi(\psi|t;c)$ and it can be easily estimated using an MCMC sample of the model parameters. Lower DIC values suggest better models.

Conditional Predictive Ordinate (CPO)

Model performance can be also measured in terms of predictive ability. For each observation i, the CPO_i [Geisser and Eddy, 1979] is defined as

$$\text{CPO}_i=L(t_i|t_{-i};c)=\left[\text{E}\left(\frac{1}{L(t_i|\psi;c)}\right)\right]^{-1}, \qquad t_{-i}=(t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n), \quad (1.18)$$

where the expectation is with respect to $\pi(\psi|t;c)$ and $L(t_i|t_{-i};c)$ is the predictive likelihood for t_i given t_{-i}. For uncensored observations, $L(t_i|t_{-i};c)$ is equal to the

predictive density function $f(t_i|t_{-i})$. In case of right censored observations, $f(t_i|t_{-i})$ is replaced by the predictive survival function $S(t_i|t_{-i})$ [as in Banerjee et al., 2007; Hanson, 2006]. A larger value of CPO_i indicates better predictive accuracy for the observation i. A Monte Carlo estimation of CPO_i is easily obtained on the basis of an MCMC sample of ψ.

Pseudo Bayes Factors (PsBF)

Geisser and Eddy [1979] also proposed $\text{PsML} = \prod_{i=1}^{n} \text{CPO}_i$ as an estimator of the marginal likelihood (often called Pseudo Marginal Likelihood). Higher values of PsML indicate a better overall predictive performance of the model. PsBF can be easily computed as ratios of PsML's.

1.2.5 Detection of influential observations

A robust model will have no (or just few) influential observations. Influential observations can be detected using $K_i = \text{KL}(\pi(\psi|t;c), \pi(\psi|t_{-i};c_{-i}))$, where $\text{KL}(\cdot,\cdot)$ denotes the Kullback-Leibler divergence function [Peng and Dey, 1995; Cho et al., 2009]. It quantifies the perturbation produced in the posterior distribution of ψ when the observation i is removed from the sample. As suggested in McCulloch [1989], K_i is transformed in terms of its calibration index $p_i = 0.5\left[1 + \sqrt{1 - \exp\{-2K_i\}}\right]$, $p_i \in [0.5, 1]$. In relation to the Kullback-Leibler divergence, the effect of removing observation i is equivalent to assigning probability p_i to an event which has true probability 0.5. A large value of p_i (usually larger than 0.9) suggests that observation i is influential. This method is closely related to CPO's. In fact,

$$K_i = \text{E}_\psi\left(\log(L(t_i|\psi;c))\right) - \log(\text{CPO}_i), \tag{1.19}$$

where the expectation is with respect to the posterior distribution of ψ. This can be easily estimated using the draws of an MCMC algorithm.

1.3 Survival regression

An important aspect of statistical modelling is the inclusion of covariates. These covariates can include clustering variables such as group of treatment or manufacturer and other characteristics that are specific of each subject (*e.g.* age, sex, health scores). Here, covariates are assumed to be non-stochastic and constant over time. Let $x_i \in \mathbb{R}^k$ be a vector containing the value of k covariates associated with individual i. Survival regression models arise as representations of the dependence between

T_i and x_i.

1.3.1 Proportional hazards model

The semi-parametric Cox Proportional Hazards (PH) model [Cox, 1972] is routinely used in applied survival analysis. It defines the effect of the covariates over the survival times in terms of the hazard function as

$$h_{T_i}(t_i|\beta^*; x_i) = h_0(t_i)\, e^{x_i'\beta^*}, \qquad i = 1, \ldots, n, \tag{1.20}$$

where $\beta^* = (\beta_1^*, \ldots, \beta_k^*)' \in \mathbb{R}^k$ is a vector of parameters. The factor $h_0(\cdot)$, denominated baseline hazard rate, represents the hazard rate of a baseline variable T_0 (which does not depend on x_i nor i). In this context, $e^{\beta_j^*}$ is interpreted as the proportional marginal change of the hazard rate after a unit change in covariate j. The original proposal in Cox [1972] does not specify $h_0(\cdot)$. Instead, the inference focuses on β^*, considering $h_0(\cdot)$ as a nuisance element. Alternatively, a parametric model can be assigned to T_0. Some standard choices are the exponential and Weibull distributions, for which the distribution of T_i remains in the same parametric family.

Despite its popularity, the model in (1.20) is not always appropriate. In terms of the survival function, (1.20) is equivalent to

$$S_{T_i}(t_i|\beta^*; x_i) = [S_0(t_i)]^{e^{x_i'\beta^*}}, \qquad i = 1, \ldots, n. \tag{1.21}$$

Hence, if the PH assumption holds, the survival functions associated to different sub-populations (defined by configurations of the covariates values) must not cross. The latter motivates an informal graphical test for the PH property. Typically, the non-parametric Kaplan-Meier estimator [Kaplan and Meier, 1958] of $S_{T_i}(t_i|x_i)$ is used for this purpose. Other (formal) checks are based on residual analysis [see Chapter 4 in Collett, 2003]. The validity of the PH premise relies on the inclusion of all relevant covariates. If (1.20) is truly satisfied for a set of covariates, the omission of one or more of these predictors destroys the PH property. In such cases, using a PH model produces biased estimations of the regression parameters [Hutton and Monaghan, 2002].

1.3.2 Accelerated failure times model

Alternatively, in an Accelerated Failure Times (AFT) model, the effect of the covariates is directly introduced through the time-scale as

$$T_i|\beta; x_i \stackrel{d}{=} e^{x_i'\beta} T_{i0}|\beta; x_i, \qquad i = 1, \ldots, n, \tag{1.22}$$

where $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbb{R}^k$ and T_{i0} is assigned a baseline distribution (which does not depend on x_i nor i). In the log-scale, (1.22) coincides with a linear regression for $\log(T_i)$ with error terms distributed as $\log(T_{i0})$. Nonetheless, standard procedures for linear regression cannot be used because they do not account for censored observations. The model in (1.22) is more intuitive than the PH specification as it directly relates to the survival times [Wei, 1992; Cox, 1997]. The impact of changes in the covariate j is to accelerate or decelerate the speed at which the event occurs. Such effects can be interpreted in terms of the percentiles of the lifetime distribution (*e.g.* its median).

In terms of the hazard and survival functions, (1.22) is equivalent to

$$S_{T_i}(t_i|\beta; x_i) = S_0\left(e^{-x_i'\beta} t_i\right), \qquad i = 1, \ldots, n \tag{1.23}$$

and

$$h_{T_i}(t_i|\beta; x_i) = e^{-x_i'\beta} h_0\left(e^{-x_i'\beta} t_i\right), \qquad i = 1, \ldots, n, \tag{1.24}$$

respectively. Unlike the PH setting, a fully parametric model is usually assumed for T_{i0}. However, a non-parametric approach can be found in Wei [1992]. In a parametric setting, the relationship in (1.22) makes attractive the use of distributions that are invariant under power-scale transformations (*i.e.* the distribution of the resultant random variable remains in the same parametric family). Typical examples for this are the log-normal, log-logistic and Weibull distributions.

In contrast to PH models, the omission of relevant covariates does not destroy the validity of an AFT regression. Hence, AFT models constitute a more robust alternative to the PH hazards regression [Hutton and Monaghan, 2002].

1.4 Contributions in this thesis

Conventional survival models often do not accommodate all features of real datasets, inducing the need of more flexible models. Since some standard lifetime distributions can be too restrictive (in terms of shape and tails), a first group of developments aims to build new parametric models (or to extend the old ones). In this line, Chapters 2 and 3 provide the following contributions.

- Flexible families of life distributions are intuitively generated on the basis of well-known models by introducing an individual-specific random effect.

This induces a hierarchical structure, accounting for unobserved heterogeneity which possibly relates to outlying observations. Whereas this idea has been previously explored in the survival literature, the approach presented here is more general as it does not rely on specific parametric models.

- Unlike most of the previous related literature, an AFT scheme is adopted for the inclusion of covariates. In this context, the interpretation of the regression coefficients is invariant to the distribution of the random effect (and, in particular, whether or not a random effect was introduced). This is a major advantage over the usual PH specification, where the interpretation of the regression coefficients is conditional on the random effect.

- Reasonable (improper) priors that require a minimum input from applied users are proposed and weak conditions for posterior propriety are derived. Censoring, which is a critical feature of survival data, is incorporated in the analysis. Nevertheless, it is shown that adding censored observations cannot destroy the existence of a well-defined posterior distribution.

- A novel outlier detection procedure (based on Bayes factors) is proposed. This is an intuitive use of the model hierarchical structure where anomalous observations are identified via the posterior distribution of the individual-specific random effects.

A second path of extensions relates to situations in which the standard setting does not correctly represent the nature of the event under analysis. In the context of university outcomes (graduation or dropout), Chapter 5 studies a discrete-time competing risks model which allows more than one type of event. The main aim of the analysis is to determine potential risks factors that might contribute to higher rates of dropout and delayed graduations. The output of this analysis does not account for all features of this complex dataset yet it provides a better understanding of the problem and a starting point for future research.

- The empirical approach presented here jointly deals with graduations and dropouts. Since it incorporates a temporal component, the detection of critical periods where students have a higher risk of dropout is allowed. In contrast, previous studies often treat university outcomes in a dichotomous manner, focusing on whether or not a student withdraws before graduation.

- The complex structure of the analyzed dataset cannot be directly handled using a maximum likelihood approach and the proportional odds model that

has been suggested by previous authors in similar contexts. Nonetheless, a Bayesian setting and the choice of appropriate priors aids the analysis, allowing the extraction of sensible information from the data.

- Finally, different criteria for covariate selection are employed in order to identify, within the set of available covariates, the main determinants of length of stay at university and its associated outcomes. This provides valuable information to university authorities, which might have an important impact on future policies.

1.5 Outline

This thesis is organized as follows. In Chapter 2, mixture families of survival distributions are introduced as a natural approach for accommodating unobserved heterogeneity between individuals. For these models, the effect of outlying observations is diminished, producing more robust inference. General features of these mixture models are discussed, with emphasis on the implementation of Bayesian inference. In specific, it is discussed how the mixing representation of these models is particularly useful for the detection of potential outlying observation and an intuitive Bayesian method for outlier detection is proposed. Chapter 2 finalizes making a link to a wide range of previous related literature. Two examples within these mixture families are studied in Chapter 3. These families contain some well-known survival models, some of which are routinely used in applied research. Nonetheless, the hierarchical structure facilitates prior elicitation and the implementation of Bayesian inference. These methods are illustrated in Chapter 4 using three real datasets, one concerning a lung cancer trials, a second one related to bone marrow transplants and another on cerebral palsy. A substantive real-life problem motivates Chapter 5. As a potential result of a less restrictive access to university education, issues such as dropout and late graduations appear as a major complication. Using a dataset provided by the Pontificia Universidad Católica de Chile (PUC), the length of stay at university (until graduation or dropout) is analyzed. In this context, a competing risks model is employed. Model uncertainty is handled through Bayesian model averaging, which leads to a better predictive performance than choosing a unique model. As a comparison, other criteria for model selection are also implemented. Finally, Chapter 6 concludes this thesis describing further developments to be explored in further research. Appendix F lists the main probability density functions used throughout this thesis.

Chapter 2

Mixtures of life distributions

"The essential is invisible to the eye".
Antoine de Saint-Exupèry
The Little Prince

2.1 Introduction

Frequently, standard survival models do not accommodate all features of real applications. In particular, datasets often exhibit more "rare" or "tail" observations than predicted by usual models. Hence, models such as Weibull or log-normal lead to inference that is not robust to the presence of outliers [Barros et al., 2008]. A second, related, issue is the existence of specific individual factors that result in unobserved heterogeneity of the survival times which cannot be captured by covariates [Marshall and Olkin, 2007]. Therefore, the typical assumption that the survival times correspond to realizations of random variables $T_1, \ldots, T_n$ which have the same "thin tailed" distribution (possibly depending on a set of known covariates) can be inappropriate. An example of such a case is the Veterans' Administration (VA) lung cancer data presented in Kalbfleisch and Prentice [2002], for which the previous literature found strong evidence of influential observations and unobserved heterogeneity related to outliers [*e.g.* Barros et al., 2008; Heritier et al., 2009]. These data are analyzed in Chapter 4.

Here, the use of mixtures of life distributions is considered in order to account for unobserved heterogeneity and add robustness to the presence of outliers. These families are sometimes called *compound distributions* and their use has been argued by several authors. See for example Padgett and Tsokos [1978], McDonald and Butler [1987], Singpurwalla [2006] and Marshall and Olkin [2007]. In particular,

the last authors dedicated a whole chapter of their book to the study these mixture families. Nevertheless, in spite of the theoretical development of the recent years, their use has not yet reached high levels of popularity in applied work, a task that remains as a big challenge.

Section 2.2 introduces mixture families of lifetime distributions as a natural extension of well-known distributions, where unobserved heterogeneity is represented in a hierarchical manner. As illustrated in Section 2.3, this hierarchical structure can be easily incorporated when conducting Bayesian inference. Covariates are included in Section 2.5 via two alternative representations. In addition, Section 2.4 presents a novel method for outlier detection that exploits the mixing structure. For completeness, Section 2.6 provides an overview of previous related literature. Finally, Section 2.7 concludes with a discussion of the main advantages of the proposed framework.

2.2 Mixtures of life distributions

Definition 1. *Let T_i be a positive-valued random variable. The distribution of T_i is defined as a mixture of lifetime distributions if and only if its density function can be represented as*

$$f(t_i|\psi,\theta) = \int_{\mathcal{L}} f(t_i|\psi,\Lambda_i=\lambda_i)\, dP_{\Lambda_i}(\lambda_i|\theta), \tag{2.1}$$

where $f(\cdot|\psi,\Lambda_i=\lambda_i)$ represents the density function of a lifetime distribution parameterized in terms of ψ and λ_i (denoted by the underlying distribution) and λ_i is a realized value of a random variable Λ_i which has distribution function $P_{\Lambda_i}(\cdot|\theta)$ defined on $\mathcal{L}$ (denoted by the mixing distribution). Alternatively, a hierarchical representation of (2.1) is given by

$$T_i|\psi,\Lambda_i=\lambda_i \sim F\left(\cdot|\psi,\Lambda_i=\lambda_i\right), \qquad \Lambda_i|\theta \sim P_{\Lambda_i}(\cdot|\theta), \tag{2.2}$$

where $F\left(\cdot|\psi,\Lambda_i=\lambda_i\right)$ is the underlying distribution function.

This approach intuitively leads to flexible distributions on the basis of a known distribution by mixing over a parameter. A wide variety of shapes and tails are generated by (2.1), accommodating unobserved heterogeneity (possibly related to outlying observations). Mixture models mitigate the effect of extreme observations on the posterior distribution of the model parameters. This is reflected in a reduction on the number of influential observations (see Subsection 1.2.5).

The extent of unobserved heterogeneity is controlled by the spread of the mixing distribution. If $\mathcal{L}$ is a finite set of values, the distribution of T_i is a finite mixture of life distributions. In particular, if $\mathcal{L}$ contains a single value, the mixture recovers the original underlying distribution (no unobserved heterogeneity). Discrete mixtures of lifetime distributions are explored in Nickell [1979], Vaupel and Yashin [1985], Marín et al. [2005] and Soliman [2006], among others. Here, however, the focus is the case in which Λ_i is a continuous random variable (throughout $\mathcal{L} = \mathbb{R}_+$, unless specified), where $f(\cdot|\psi, \theta)$ is interpreted as an infinite mixture of densities [as in, *e.g.*, Vaupel et al., 1979; Hougaard, 1995; Duchateau and Janssen, 2008].

The mixing distribution can, in principle, correspond to any proper probability distribution [several alternatives are listed in Chapter 5 of Hanagal, 2011]. Nevertheless, some restrictions are often required for identifiability reasons. These identification constraints are specific to each family of mixtures (typically, unknown separate scale parameters are not allowed). Heckman and Singer [1984a] remark that inference might be sensitive to the mixing distribution and therefore use a non-parametric model for the random effect. Non-parametric mixtures of parametric survival models are also explored in Elbers and Ridder [1982], Horowitz [1999] and Kottas [2006], among others. However, a non-parametric mixing distribution might not be appropriate for moderate sample sizes. A fully parametric approach is adopted here and the adequacy of a particular mixing distribution is evaluated using Bayesian model comparison tools (see Subsection 1.2.4). This parametric choice is a compromise between the standard model in which $\Lambda_i = \lambda_0$ (with probability one) and the use of a fully flexible non-parametric mixing distribution.

Varying the underlying model, generates a wide class of lifetime distributions. To illustrate Chapter 3 explores two mixture families generated by log-normal and Weibull distributions, respectively. In addition, Balakrishnan et al. [2009] and Patriota [2012] explored mixtures of Birnbaum-Saunders distributions [Birnbaum and Saunders, 1969] that are based on scale mixtures of normals [Fernï¿½ndez and Steel, 2000]. In an engineering context, Patriota [2012] suggests the latter family for failures produced by progressive material cracks. In such a case, the mixing distribution accounts for dependencies between the cracks.

If an (underlying) distribution is underpinned by theoretical or practical reasons, the same reasons hold for the mixture model in the presence of unobserved heterogeneity. Conditional on the mixing parameters, survival times are distributed as in the underlying model but with a different value λ_i for each individual (see (2.2)). For example, if theory suggests that individuals have a constant hazard rate, an exponential model is appropriate. Using mixtures of exponential distributions

leads to a decreasing hazard rate [Marshall and Olkin, 2007, p.92,Corollary D.4.a.], yet does not contradict this theory. In such a case, the decreasing behaviour of the population hazard is linked to unobserved heterogeneity. In fact, individual hazards remain constant on time but high-risk individuals die earlier, leaving only low-risk individuals to be observed at longer times. Hence, if neglected, unobserved heterogeneity yields to an incorrect estimation of the individual hazard rate [Omori and Johnson, 1993].

An extended study of the distributions distributions generated by (2.1) and its properties is presented in Marshall and Olkin [2007]. In particular, the survival function retains the same structure as in (2.1), being defined as

$$S(t_i|\psi,\theta) = \int_{\mathcal{L}} S(t_i|\psi,\Lambda_i=\lambda_i)\, dP_{\Lambda_i}(\lambda_i|\theta), \tag{2.3}$$

where $S(\cdot|\psi,\Lambda_i=\lambda_i)$ is the survival function associated to the underlying model. The latter also applies to the distribution function but it is not valid for the hazard function. A similar representation for $h(t_i|\psi,\theta)$ exists, however it involves a different mixing distribution which depends on t_i [Marshall and Olkin, 2007, p.84]. In addition, if the underlying distribution has a decreasing hazard rate, the marginal hazard (after integrating out λ_i) also decreases. This is regardless of the mixing distribution. The counterpart, when the underlying hazard is non-monotone or increasing, is not true. In those cases, the hazard rate induced by mixing is more flexible. For instance, mixtures of distributions with increasing hazard rate might produce monotone decreasing hazards [Marshall and Olkin, 2007, p.92].

2.3 Posterior inference for mixtures of life distributions

If an analytical solution is available for the integral in (2.1), the marginal model can be used for inference purposes. An example of such a situation is the log-logistic distribution, which can be represented as an infinite mixture of log-normal distributions (see Section 3.2). When an analytic representation of the marginal model does not exist, the hierarchical structure in (2.2) can be exploited. In a frequentist setting, a maximum likelihood analysis can be implemented by means of a Expectation-Maximization algorithm [Dempster et al., 1977]. Instead, a Bayesian approach can deal with the mixing parameters using a Gibbs Sampler (Subsection 1.2.3) and the data augmentation idea proposed in Tanner and Wong [1987]. Let $\pi(\psi,\theta)$ represent a prior distribution for (ψ,θ). The following full conditionals are defined.

$$\pi(\psi|\lambda,\theta,t;c) \propto \pi(\psi,\theta)\prod_{i=1}^{n} L(t_i|\psi,\Lambda_i=\lambda_i;c_i), \tag{2.4}$$

$$\pi(\lambda_i|\psi,\theta,t;c) \propto L(t_i|\psi,\Lambda_i=\lambda_i;c_i)\, dP_{\Lambda_i}(\lambda_i|\theta)\pi(\psi,\theta), \quad i=1,\ldots,n, \tag{2.5}$$

$$\pi(\theta|\psi,\lambda,t;c) \propto \pi(\psi,\theta)\prod_{i=1}^{n} dP_{\Lambda_i}(\lambda_i|\theta), \tag{2.6}$$

where $\lambda=(\lambda_1,\ldots,\lambda_n)'$ and $L(t_i|\beta,\Lambda_i=\lambda_i;c_i)$ denotes the likelihood contribution of the i-th observation.

The latter sampler requires the update of n mixing parameters at each step of the chain. This may be computationally inefficient (especially when sampling from the mixing variables is cumbersome). In order to avoid this problem, the λ_i's can be sampled only every Q iterations of the chain. As a consequence, the ESS (see (1.5)) of the chain is diminished. Hence, an appropriate value for Q must be chosen by considering a trade-off between ESS and the time required for running the algorithm.

In terms of setting up a sampler, it might be easier to simply start directly from the marginal model (if known), rather than its interpretation as a mixture. Nonetheless, when using the mixing structure, various mixtures of the same underlying model can be implemented by only modifying (2.5) and (2.6) in the sampler. This is particularly useful when $\pi(\psi,\theta)=\pi(\psi)\pi(\theta)$ and (2.4) has a known closed form (*e.g.* if ψ has a conjugate prior with respect to the underlying model). Moreover, the mixture representation often facilitates dealing with censored (or set) observations (as point values are sampled using a common underlying structure). In addition, prior elicitation for ψ and θ can also be benefited by the mixing structure. For instance, Jeffreys-style priors (Subsection 1.2.1) can share a similar patterns for all distributions in the same mixture family. If using informative priors, these can be "matched" through a common feature in order to represent the same prior information for any mixing distribution (within the same underlying model). Furthermore, if the mixing representation is ignored, posterior inference on the mixing variables would be lost. As described in Section 2.4, such information is particularly important in identifying outlying observations [West, 1984; Lange et al., 1989; Fernï¿½ndez and Steel, 1999]. These ideas are further discussed in Chapter 3.

2.4 A method for outlier detection

Mixture models account for unobserved heterogeneity between subjects that cannot be measured with covariates. Occasionally, this heterogeneity is linked to particularly anomalous observations. Here, the posterior distribution of the mixing variables is exploited in order to propose an intuitive method for outlier detection. Extreme values (with respect to a reference value, λ_{ref}) of the mixing variables are associated with outliers [see also West, 1984]. Formally, evidence of t_i being an outlying observation can be assessed by contrasting the models $M_0 : \Lambda_i = \lambda_{ref}$ versus $M_1 : \Lambda_i \neq \lambda_{ref}$ (with all other $\Lambda_j, j \neq i$ free). Evidence in favour of each of these models is measured using Bayes factors (Subsection 1.2.4), which can be computed as the generalized Savage-Dickey density ratio proposed in Verdinelli and Wasserman [1995]. The evidence in favour of M_0 versus M_1 (*i.e.* against observation i being an outlier) is

$$\mathrm{BF}_{01} = \pi(\lambda_i|t;c)\mathrm{E}\left(\frac{1}{dP_{\Lambda_i}(\lambda_i|\theta)}\right)\Bigg|_{\lambda_i=\lambda_{ref}}, \tag{2.7}$$

where the expectation is with respect to $\pi(\theta|\Lambda_i = \lambda_{ref}, t; c)$. In such a case, estimating $\left\{\mathrm{BF}_{01}^{(i)}\right\}_{i=1,\ldots,n}$ is computationally intensive. In fact, the estimation of each $\mathrm{BF}_{01}^{(i)}$ requires a reduced run of the algorithm introduced in Section 2.3, where λ_i is fixed (equal to λ_{ref}). Long running times are needed for this (specially when n is large and sampling from λ_i is not straightforward). Nevertheless, as these n runs are independent, the process can be easily speed-up with the help of parallel computing. In contrast, when the parameter θ does not appear in the model, (2.7) simplifies to the original Savage-Dickey density ratio

$$\mathrm{BF}_{01} = \frac{\pi(\lambda_i|t;c)}{dP_{\Lambda_i}(\lambda_i)}\Bigg|_{\lambda_i=\lambda_{ref}} = \mathrm{E}\left(\frac{L(t_i|\psi,\Lambda_i=\lambda_i;c_i)}{L(t_i|\psi;c_i)}\right)\Bigg|_{\lambda_i=\lambda_{ref}}, \tag{2.8}$$

where $L(t_i|\psi,\Lambda_i = \lambda_i; c_i)$ and $L(t_i|\psi; c_i)$ represent the likelihood contribution of the i-th observation under the underlying (conditional on the mixing parameter) and marginal models, respectively. The expectations in (2.8) are with respect to $\pi(\psi|t; c)$. The original MCMC chain generated by the algorithm in Section 2.3 can be used for a fast estimation of (2.8).

This outlier detection method relies on the choice of a reasonable value for λ_{ref}, which is specific of each mixture. In the absence of unobserved heterogeneity, the posterior density of the random effects should behave as a Dirac function with a spike on λ_{ref}. Following this intuition, $E(\Lambda_i|\theta)$ (if it exists) is proposed as λ_{ref}. If

θ is unknown, it can be replaced by a Bayesian estimate (*e.g.* its posterior median). Examples for which $E(\Lambda_i|\theta)$ is not finite require a more detailed analysis. Subsection 3.2.5 provides extra guidance in this respect. In such cases, a value for λ_{ref} can be determined in an empirical fashion (using simulated and real datasets).

2.5 Incorporating unobserved heterogeneity to survival regressions

There is no unique method for incorporating unobserved heterogeneity to survival regressions. Conditional on the mixing parameters, a regression model can be specified for the underlying structure (as in the standard setting). This model summarizes the effect of the measured covariates at an individual level. After integrating out the random effects, the marginal model condenses the covariates effect at a population level. These effects do not always coincide. Below, two schemes that are predominantly used in previous research are introduced.

2.5.1 The mixed PH model

In the presence of unobserved heterogeneity, the survival literature mostly focuses on a PH specification. In the so-called mixed PH model, mixing parameters affect the hazard rate in a multiplicative manner. It is defined as

$$h_{T_i}(t_i|\beta^*, \Lambda_i = \lambda_i; x_i) = g^*(\lambda_i) h_0(t_i)\, e^{x_i'\beta^*}, \quad \Lambda_i \sim P_{\Lambda_i}(\cdot|\theta), \quad i = 1, \ldots, n, \quad (2.9)$$

where β^* and $h_0(\cdot)$ are defined as in (1.20). In addition, $g^*(\cdot)$ is an arbitrary positive-valued function. When $g^*(\cdot)$ is the identity function, Omori and Johnson [1993] shown that the unconditional hazard function (after integrating out the mixing parameters) is given by

$$h_{T_i}(t_i|\beta^*; x_i) = \frac{\mathrm{E}(\Lambda_i \exp\{-\Lambda_i H_0(t_i)\, e^{x_i'\beta^*}\})}{\mathrm{E}(\exp\{-\Lambda_i H_0(t_i)\, e^{x_i'\beta^*}\})} h_0(t_i)\, e^{x_i'\beta^*}, \quad (2.10)$$

where $H_0(t_i) = \int_0^{t_i} h_0(s)\, ds$ and both expectations are with respect to the mixing distribution. Hence, even though the mixed PH model is a mixture of PH models, the proportional hazards property is generally not preserved for the marginal model. The deviation from the PH assumption caused by unobserved heterogeneity adds plausibility to the mixed PH model in applications where this assumption has been refuted. Accounting for unobserved heterogeneity is critical under a PH scheme. In fact, in this context, $e^{\beta_j^*}$ is interpreted as the proportional marginal change of

the hazard rate after a unit change in covariate j at an individual level. This interpretation is conditional on the random effect and it cannot be extended to the population level.

The mixed PH model is widely used in econometrics [*e.g.* Heckman and Singer, 1984b; Honoré, 1990; Omori and Johnson, 1993; Mosler, 2003; Abbring and Van Den Berg, 2007]. However, the baseline hazard is often assumed to have a non-parametric structure [as in Cox, 1972]. For the mixed PH model, the marginal survival function corresponds to the Laplace transform of the mixing density evaluated at $H_0(t_i)\, e^{x_i'\beta^*}$ [Wienke, 2010]. Therefore, mixing densities with known Laplace transform are an attractive choice. An example of this is the Power Variance Function (PVF) family, for which the variance is a power of the mean. This option is explored in Wasinrat et al. [2013] (under a maximum likelihood approach). In particular, the positive stable distribution is a limiting case of the PVF family [Wienke, 2010]. Some other examples in the PVF family are the Gamma and the inverse Gaussian distributions (with one of their parameters fixed). The Gamma distribution is perhaps the most popular choice for the distribution of the random effect. Although one of the main reasons for this is the simplification of analytical expressions, Abbring and Van Den Berg [2007] also gives an asymptotic argument for a Gamma mixing.

2.5.2 The mixed AFT model

Unobserved heterogeneity can be also incorporated through a mixture of AFT regressions [*e.g.* Anderson and Louis, 1995]. The mixed AFT model is defined as

$$T_i|\beta, \Lambda_i = \lambda_i; x_i \stackrel{d}{=} e^{x_i'\beta} g(T_{i0}, \lambda_i)|\beta, \Lambda_i = \lambda_i; x_i \qquad i = 1, \ldots, n, \tag{2.11}$$

with β and T_{i0} are defined as in (1.22). Furthermore, $g(\cdot)$ is an arbitrary positive-valued function. Although this option is less explored in the existing literature, some authors recommend its use [*e.g.* Keiding et al., 1997]. Unlike the mixed PH model, the marginal model generated by (2.11) is itself an AFT model, for which the baseline variable is defined as $\tilde{T}_{i0} = g(T_{i0}, \Lambda_i)$. Hence, the interpretation of the regression coefficients is invariant to the mixing distribution (and, in particular, whether or not a random effect was introduced). This important feature is an advantage over the mixed PH model, in which the interpretation of the regression parameters is conditional on the random effect. As in the standard AFT model, e^{β_j} can be interpreted as the proportional marginal change of the median survival time (or any other percentile) after a unit change in covariate j. This interpretation does

not differentiate between individual and population levels.

2.6 Related literature

Mixture modeling can be interpreted as the introduction of a random effect on the survival distribution. The survival literature often refers to $\lambda_1, \ldots, \lambda_n$ as *frailties*, a term that was originally introduced by Vaupel et al. [1979]. In this context, the model in (2.1) is usually called *univariate frailty model* and its use dates back to Beard [1959]. During the last decades, the literature about frailty models experienced a large expansion. Among others, some examples of this are Honoré [1990], Omori and Johnson [1993], Mosler [2003], Abbring and Van Den Berg [2007], Wienke [2010] and Hanagal [2011]. Mixtures as in (2.1) constitute an small part of the research related to frailty models. Beyond representing unobserved heterogeneity between specific individuals, frailty models can also accommodate more complex data structures. Some examples are listed below.

2.6.1 Shared frailty models

Aiming to account for correlation between clustered observations, shared frailty models are one of the most popular extensions of the univariate frailty model [Clayton, 1978; Hougaard, 1995]. An extensive survey about this subject can be found in Duchateau and Janssen [2008]. These models are used for grouped datasets where, conditional on the observed covariates, survival times are assumed to have the same distribution within each cluster (*e.g.* siblings, patients treated in the same hospital, systems built by the same manufacturer). In such a case, the frailty terms take a common value for all individuals belonging to the same group. The latter introduces intra-cluster dependencies (independence is conditional on the mixing parameters). As discussed in Chapter 6 of Duchateau and Janssen [2008], this approach can be also extended to hierarchical frailties, where more than one level of clustering occurs.

2.6.2 Correlated frailty models

Assigning the same frailty value to all observations within a cluster is not always appropriate. There is often intra-cluster variation that cannot be controlled by the observable covariates (as in the un-clustered case). A more flexible approach for grouped observations is provided by correlated frailty models [see Chapter 12 in Hanagal, 2011]. These models assign a joint distribution to the mixing parameters associated to each group. For example, in Banerjee et al. [2003], correlations between

the frailties account for spatial dependencies when modelling infant mortality. If groups are formed by only two observations (*e.g.* identical twins), these models are often renamed as *bivariate frailty models* [Wienke et al., 2005]. In particular, in the absence of within-group dependencies, they reduce to the univariate frailty case.

2.6.3 Cure rate models

In some contexts, there is a proportion of individuals who will never experience the event of interest. Following a medical nomenclature, these subjects are commonly labeled as *cured* units. For instance, patients that evidence a full recovery must be removed from the "at-risk" group. Frailty models accommodate these type of datasets by using a mixing distribution that assigns a positive probability to not observing the event (*i.e.* the hazard function is equal to zero). One example of such mixing is the compound Poisson distribution proposed by Aalen [1992], which is also used in Price and Manatunga [2001].

2.7 Concluding remarks

The use of mixtures of life distributions is recommended as a convenient framework for survival analysis, particularly when standard models such as the Weibull or log-normal are not able to capture some features of the data. These mixture families can accommodate unobserved heterogeneity (possibly related to outlying observations), which is crucial in survival analysis. This approach intuitively leads to flexible distributions on the basis of a known distribution by mixing over a parameter. Mixture modelling can also be interpreted through random effects or frailty terms which has a strong link with the previous survival literature. The setting presented here is general, without assuming a particular distribution for underlying nor the mixing model (whether the induced survival time distribution has a closed-form density function or not). This is a major advantage over previous works, where mixing parameters are typically assigned a Gamma distribution in order to obtain analytical expressions for the marginal model. The proposed MCMC inference scheme does not rely on a closed form expression for the survival density with the mixing variables integrated out and this sampler can be easily extended for some of the models described in Section 2.6.

Mixture models diminish the effect that anomalous observations have over posterior inference. Nonetheless, it might be of interest for practitioners to determine whether a small group of outlying observations drives the unobserved heterogeneity. An outlier detection method is designed for this objective. It exploits the

mixing structure, comparing individual frailties with respect to a reference level. This comparison is formalized by means of Bayes factors. If the mixing distribution has a finite expectation, a general recommendation for the (critical) choice of a reference value is provided. Cases where the expected value of the mixing parameters fails to exist are explored in detail in Section 3.2.5.

Previous literature provides no consensus about how unobserved heterogeneity must be incorporated in survival regression models. A mixed PH specification is frequently used for this purpose. However, for the mixed PH model, the interpretation of the regression parameters is subject to conditioning on the random effects and posterior inference is very sensitive to variations of the mixing distribution. Instead, the mixed AFT model is an attractive alternative. It provides a clearer interpretation of the covariates effects (which is not affected by the mixture) and the posterior distribution of the regression coefficients is more robust to the choice of mixing distribution.

The methodology introduced in this Chapter is illustrated in Chapter 3 using mixture families generated from log-normal and Weibull distributions, respectively.

Chapter 3

Two flexible families for survival modelling

> "In this quest to seek and find God in all things there is still an area of uncertainty. There must be".
> Pope Francis

3.1 Introduction

The log-normal and Weibull distributions are routinely applied in survival analysis. Respectively, Crow and Shimizu [1988] and Rinne [2008] provide detailed surveys about these models, their origins and properties. In an engineering context, the log-normal model can be conceived as the limiting distribution of an additive cumulative damage scheme, where repeated exposures to a risk factor trigger the event. It generates a non-monotone hazard function, which has an initial increasing hazard phase. Instead, the Weibull distribution accommodates flexible shapes for the hazard rate (including monotone ones). The Weibull model is a particular case of the generalized extreme value distribution [Fisher and Tippett, 1928]. Despite their popularity, the presence of unobserved heterogeneity can invalidate the use of these models. In particular, they produce inferences that are not robust to the presence of outlying observations [Barros et al., 2008]. As in Chapter 2, unobserved heterogeneity is incorporated to these models by means of an infinite mixture of lifetime distributions. This idea generates flexible classes of distributions for survival modelling that provide natural ways to deal with both the presence of outlying observations and unobserved heterogeneity.

Section 3.2 explores the Shape Mixtures of Log-Normal (SMLN) distributions for which the shape parameter is assigned a mixing distribution. This new class covers a wide range of shapes, in particular cases with fatter tail behaviour than the log-normal. It includes the already studied log-Student t, log-Laplace, log-exponential power and log-logistic distributions among others. Covariates are included in Subsection 3.2.1 via an AFT specification (for which the interpretation of the regression parameters is not affected by the mixture). A prior distribution, inspired by the Jeffreys-rule, is presented in Subsection 3.2.2 and conditions for posterior propriety are provided in Subsection 3.2.3. Subsection 3.2.4 describes implementation aspects and outlier detection for AFT-SMLN models is studied in Subsection 3.2.5. Following the same structure, Section 3.3 introduces the family of Rate Mixtures of Weibull (RMW) distributions, for which a random effect is introduced through the rate parameter. This family contains i.a. the well-known Lomax distribution and can accommodate flexible hazard functions. Finally, Section 3.4 concludes. All proofs are contained in the Appendix A without mention in the text.

3.2 The family of Shape Mixtures of Log-Normals

Definition 2. *A random variable T_i has a distribution in the family of Shape Mixtures of Log-Normals (SMLN) if and only if its density can be represented as*

$$f(t_i|\mu,\sigma^2,\theta) = \int_{\mathcal{L}} \frac{\sqrt{\lambda_i}}{\sqrt{2\pi\sigma^2}} \frac{1}{t_i} \exp\left\{-\frac{\lambda_i(\log(t_i)-\mu)^2}{2\sigma^2}\right\} dP_{\Lambda_i}(\lambda_i|\theta), \quad t_i > 0, \quad (3.1)$$

where $\mu \in \mathbb{R}$, $\sigma^2 > 0$, $\theta \in \Theta$ and λ_i is a realized value of a random variable Λ_i which has distribution function $P_{\Lambda_i}(\cdot|\theta)$ defined on $\mathcal{L} \subseteq \mathbb{R}_+$ (possibly discrete). Denote $T_i \sim SMLN_P(\mu,\sigma^2,\theta)$. A hierarchical representation of (3.1) is given by

$$T_i|\mu,\sigma^2,\Lambda_i = \lambda_i \sim Log-Normal\left(\mu, \frac{\sigma^2}{\lambda_i}\right), \qquad \Lambda_i|\theta \sim P_{\Lambda_i}(\cdot|\theta). \qquad (3.2)$$

The SMLN family can be interpreted as a mixture of log-normal distributions with random shape parameter or as the exponential transformation of a random variable distributed as a scale mixture of normals. This family includes a number of distributions that have been proposed in the context of survival analysis. For example, finite mixtures of log-normal distributions are explored in Fowlkes [1979] and Tian et al. [2010]. Here, instead, the focus is on infinite mixtures generated by continuous mixing distributions. Table 3.1 lists some of them. In particular, the log-Student t distribution was introduced by Hogg and Klugman [1983] and used in

Table 3.1: Some SMLN models. $f_{PS}(\cdot|\delta)$ denotes a positive stable PDF with parameter δ.

Distribution	Marginal PDF	Mixing PDF
Log-Student t	$\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)\sqrt{\pi\sigma^2\nu}}\frac{1}{t_i}\left[1+\frac{(\log(t_i)-\mu)^2}{\sigma^2\nu}\right]^{-(\frac{\nu}{2}+\frac{1}{2})}$	Gamma($\nu/2$, $\nu/2$), $\nu>0$
Log-Laplace	$\frac{1}{2\sigma}\frac{1}{t_i}\exp\left\{-\frac{\lvert\log(t_i)-\mu\rvert}{\sigma}\right\}$	Inv-Gamma(1,1/2)
Log-exponential power	$\frac{\alpha}{2\sigma\Gamma(\frac{1}{\alpha})}\frac{1}{t_i}\exp\left\{-\left(\frac{\lvert\log(t_i)-\mu\rvert}{\sigma}\right)^{\alpha}\right\}$	$\frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\lambda_i^{-\frac{1}{2}}f_{PS}(\lambda_i\lvert\frac{\alpha}{2}),\alpha\in(1,2)$
Log-logistic	$\frac{1}{\sigma\, e^{\mu}}\frac{(t_i/\, e^{\mu})^{1/\sigma-1}}{[1+(t_i/\, e^{\mu})^{1/\sigma}]^2}$	$\lambda_i^{-2}\sum_{k=0}^{\infty}\binom{-2}{k}(1+k)\, e^{-\frac{(1+k)^2}{2\lambda_i}}$

e.g. McDonald and Butler [1987] and Cassidy et al. [2009]. In the case of $\nu = 1$, Lindsey et al. [2000] applied it in the context of pharmacokinetic data. The log-Laplace appeared in Uppuluri [1981] and Lindsey [2004]. The log-exponential power was proposed by Vianelli [1983] and used in Martï¿½n and Pï¿½rez [2009]. Finally, the log-logistic distribution was introduced by Shah and Dave [1963] and is used regularly in survival analysis, hydrology and economics. This list can be increased by varying the mixing distribution. For example, all the mixing distributions used for scale mixtures of normals listed in Fernï¿½ndez and Steel [2000] can be used in this context. However, for identifiability reasons, the mixing distribution must not have separate unknown scale parameters (unknown scale parameters are allowed as long as they are linked to other features of the mixing distribution, *e.g.* $\Lambda_i \sim$ Gamma(θ,θ)).

Whereas all positive moments exist for the log-normal, this is not necessarily the case for the shape mixtures. In general, the existence of moments relates to a well-defined moment generation function for Λ_i^{-1} (given θ).

Theorem 2. *Let T_i be a random variable distributed according to (3.1). The r-th moment of T_i ($r \geq 0$) is finite if and only if $E_{\Lambda_i}\left(\exp\left\{\frac{\sigma^2 r^2}{2}\frac{1}{\Lambda_i}\right\}\middle|\theta\right) < \infty$. If it exists, it corresponds to $e^{r\mu}E_{\Lambda_i}\left(\exp\left\{\frac{\sigma^2 r^2}{2}\frac{1}{\Lambda_i}\right\}\middle|\theta\right)$.*

As a consequence of Theorem 2, no positive moments exist for the log-Student t (for any finite value of ν) and the log-Laplace only allows for moments up to $1/\sigma$. Theorem 2 is less helpful for the log-logistic and log-exponential power models (because they relate to more complex mixing distributions). However, log-logistic moments with order less than $1/\sigma$ are well defined [Tadikamalla and Johnson, 1982] and the log-exponential power distribution with $\alpha > 1$ does possess all moments

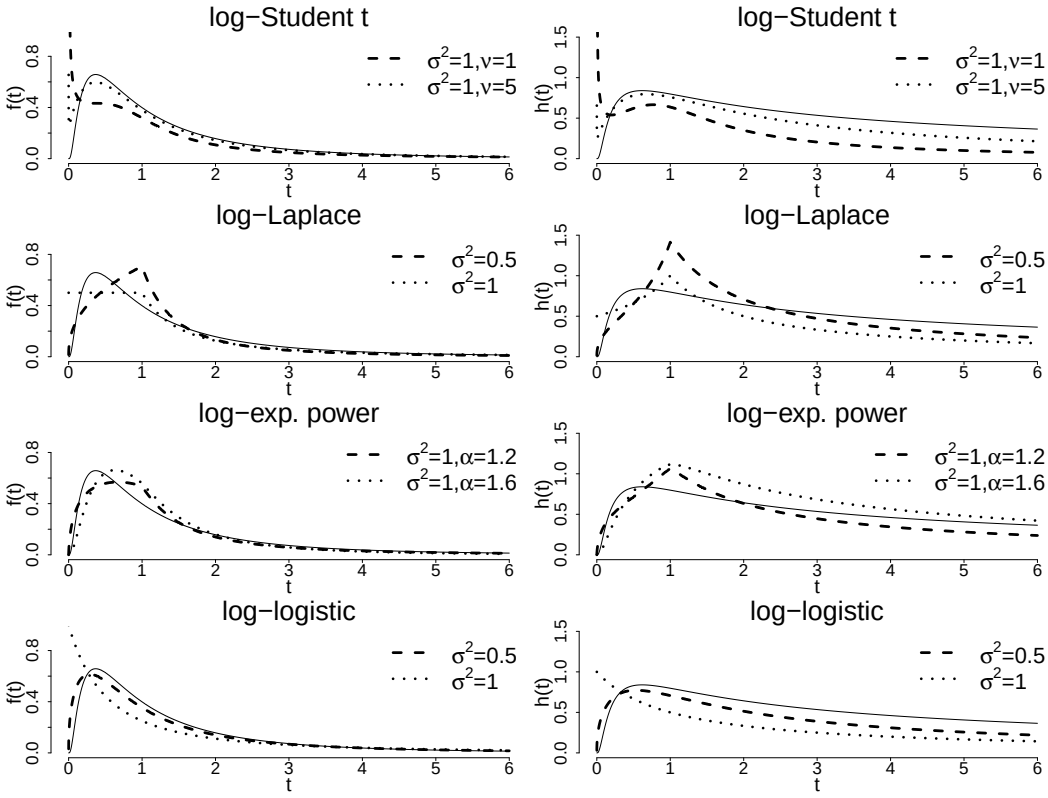


Figure 3.1: Density and hazard function (left and right panels, respectively) of some SMLN models ($\mu = 0$). Solid line is the log-normal$(0, 1)$ density (or hazard).

[Angeletti et al., 2012; Singh et al., 2012]. In addition, as a corollary of Theorem 2, the coefficient of variation (*i.e.* the ratio between the standard deviation and the expected value) of random variables in the SMLN family does not depend on μ. Hence, σ^2 and θ are the only parameters controlling the spread of these distributions. As illustrated in Figure 3.1, the SMLN family allows for a wide variety of shapes for the density and the hazard function. For example, it is clear that the tails of all these examples of SMLN with continuous mixing distributions are fatter than those of the log-normal distribution. In particular, the left tail behaviour of the density function can be quite different. Moreover, while the hazard rate of the log-normal distribution has an increasing initial phase, the log-Laplace and log-logistic distributions produce a monotone decreasing hazard rate for some values of σ^2.

3.2.1 The SMLN-AFT model

A mixed AFT formulation is adopted. Apart from the arguments in Subsection 2.5.2, this choice is based on the closure under scale-power transformations of the SMLN family (if $T \sim SMLN_P(\mu, \sigma^2, \theta)$, then $aT^b \sim SMLN_P(\log(a) + \mu b, \sigma^2 b^2, \theta)$ for $a > 0, b \neq 0$) and the lack of an analytic expression for the log-normal hazard. The SMLN-AFT model expresses the dependence between the covariates and the survival time by replacing the parameter μ with $x_i'\beta$, so that

$$T_i|\beta, \sigma^2, \theta; x_i \overset{ind}{\sim} \text{SMLN}_P(x_i'\beta, \sigma^2, \theta), \quad i = 1, 2, \ldots, n, \tag{3.3}$$

where x_i is a vector containing the value of k covariates associated with individual i and $\beta \in \mathbb{R}^k$ is a vector of parameters. This can also be interpreted as a linear regression model for the logarithm of the survival times with error term distributed as a scale mixture of normals [as in Fernï¿½ndez and Steel, 2000]. As the median of T_i in (3.3) is given by $e^{x_i'\beta}$, e^{β_j} is interpreted as the (proportional) marginal change of the median survival time as a consequence of a unitary change in covariate j. As discussed in Subsection 2.5.2, this interpretation is not affected by the mixing distribution (covering both, individuals and population levels).

3.2.2 Jeffreys-style priors for the SMLN-AFT model

Bayesian inference is conducted using *objective* priors generated by the Jeffreys rule [Jeffreys, 1961]. This is one of the most common choices in the absence of prior information and has interesting invariance and information-theoretic properties. The next theorem presents the FIM for the SMLN-AFT model which is the basis for the Jeffreys-style priors.

Theorem 3. *Let $T_1, \ldots, T_n$ be independent random variables with T_i distributed according to (3.3), then the FIM corresponds to*

$$FIM(\beta, \sigma^2, \theta) = \begin{pmatrix} \frac{1}{\sigma^2} k_1(\theta) \sum_{i=1}^n x_i x_i' & 0 & 0 \\ 0 & \frac{1}{\sigma^4} k_2(\theta) & \frac{1}{\sigma^2} k_3(\theta) \\ 0 & \frac{1}{\sigma^2} k_3(\theta) & k_4(\theta) \end{pmatrix}, \tag{3.4}$$

where $k_1(\theta), k_2(\theta), k_3(\theta)$ and $k_4(\theta)$ are functions depending only on θ.

The expressions involved in $k_1(\theta), k_2(\theta), k_3(\theta)$ and $k_4(\theta)$ are complicated (see the proof) and thus FIM$(\beta, \sigma^2, \theta)$ is not easily obtained from this theorem for any arbitrary mixing distribution. Indeed, it is usually more efficient to compute

FIM$(\beta, \sigma^2, \theta)$ directly from $f(\cdot|\beta, \sigma^2, \theta)$. However, this structure facilitates a general representation of the Jeffreys-style priors:

Corollary 1. *Under the same assumptions as in Theorem 3, the Jeffreys, independence Jeffreys (which deals separately with the blocks for β and (σ^2, θ)) and independence I Jeffreys (which deals separately with β, σ^2 and θ) priors are respectively given by*

$$\pi^J(\beta, \sigma^2, \theta) \propto \frac{1}{(\sigma^2)^{1+\frac{k}{2}}} \sqrt{[k_1(\theta)]^k [k_2(\theta)k_4(\theta) - k_3^2(\theta)]}, \tag{3.5}$$

$$\pi^I(\beta, \sigma^2, \theta) \propto \frac{1}{\sigma^2} \sqrt{k_2(\theta)k_4(\theta) - k_3^2(\theta)}, \tag{3.6}$$

$$\pi^{II}(\beta, \sigma^2, \theta) \propto \frac{1}{\sigma^2} \sqrt{k_4(\theta)}. \tag{3.7}$$

The three non-subjective priors presented here can be written as

$$\pi(\beta, \sigma^2, \theta) \propto \frac{1}{(\sigma^2)^p} \pi(\theta), \tag{3.8}$$

where $\pi(\theta)$ is the factor of the prior that depends on θ. For the Jeffreys prior $p = 1 + (k/2)$ and $p = 1$ for the other two priors. If θ does not appear (*e.g.* log-normal, log-Laplace and log-logistic models) this prior simplifies to $\pi(\beta, \sigma^2) \propto (\sigma^2)^{-p}$.

Corollary 1 also specifies the prior for θ. The implied priors for the special cases of the log-Student t and the log-exponential power (derived directly from the specific likelihood functions) are explicitly presented in the proof of Theorem 5. In order to obtain meaningful Bayes factors between models, priors with a improper component $\pi(\theta)$ for θ are discarded. For the examples explored throughout this Section, this argument discards the independence I Jeffreys prior for the log-Student t model.

3.2.3 Posterior propriety for the SMLN-AFT model

The three priors presented in Corollary 1 do not correspond to proper probability distributions and therefore the propriety of the posterior distribution must be verified. As shown by Proposition 1, ignoring censored observations leads to sufficient conditions for posterior propriety. The following results verify posterior propriety for the SMLN-AFT model under the priors in Corollary 1 on the basis of the non-censored observations (using n instead of n_o for ease of notation).

Theorem 4. *Let $t_1, \ldots, t_n > 0$ be the (non-censored) survival times of n independent individuals, realizations of random variables distributed as in (3.3). Assume*

the prior given in (3.8), with $\int_\Theta \pi(\theta)\, d\theta = 1$. *Define* $X = (x_1, \ldots, x_n)'$ *and suppose that the rank of* X *is* k.

(i) For $p = 1$, *a sufficient condition for posterior existence is* $n > k$,

(ii) For $p = 1 + k/2$, *a sufficient condition for the posterior propriety is* $n > k$ *and*

$$\int_\Theta E(\Lambda_1^{-\frac{k}{2}}|\theta)\pi(\theta)\, d\theta < \infty. \tag{3.9}$$

Theorem 5. *Under the assumptions in Theorem 4 and provided that* $n > k$, *it follows that*

(i) For the log-Student t AFT model, the posterior is proper under the independence Jeffreys prior. However, the posterior does not exist for the Jeffreys prior.

(ii) For the log-Laplace AFT model, log-exponential power AFT model and log-logistic AFT model, the propriety of the posterior can be verified with any of the three proposed priors.

Theorem 5 implies that the log-Student t AFT model does not lead to valid Bayesian inference in combination with the Jeffreys prior (see also Appendix B). In the log-Student t case, the independence I Jeffreys prior is not covered by Theorem 5 but it was already discarded in Subsection 3.2.2). The other models can be combined with all priors considered here; of course, the absence of θ in the log-Laplace and log-logistic models implies that the independence Jeffreys and independence I Jeffreys priors coincide in those cases.

As explained in Subsection 1.2.2, the use of point observations for continuous sampling models introduces the risk of having senseless inference. The following theorem illustrates the danger induced by the use of point observations in the context of the log-Student t AFT model.

Theorem 6. *Adopt the same assumptions as in Theorem 4 and assume that* $n_o > k$. *If the mixing distribution is Gamma*($\nu/2$,$\nu/2$) *and* s ($k \leq s < n$) *is defined as the largest number of uncensored observations that can be represented as an exact linear combination of their covariates (i.e.* $\log(t_i) = x_i'\beta$ *for some fixed* β), *a necessary condition for the propriety of the posterior distribution of* (β, σ^2, ν) *is*

$$\int_0^m \pi(\nu)\, d\nu = 0, \quad \textit{where} \quad m = \frac{n - k + (2p - 2)}{n - s} - 1. \tag{3.10}$$

This result indicates that it is possible to have samples of point observations for which no Bayesian inference can be conducted, unless $\pi(\nu)$ induces a positive lower bound for ν. For the log-Student t model, only the independence Jeffreys prior is used, so that $p = 1$ and (3.10) is violated whenever $s > k$. When no covariates are taken into account ($k = 1$), s coincides with the largest number of (uncensored) tied observations. Theorem 6 highlights the need for considering sets of zero Lebesgue measure when checking the propriety of the posterior distribution based on point observations.

Throughout, the set observations solution proposed in Fernández and Steel [1998]; Fernï¿½ndez and Steel [1999] is implemented when conducting Bayesian inference for the SMLN-AFT models (regardless of the mixing distribution).

3.2.4 Implementation

Bayesian inference is implemented by means of the sampler presented in Section 2.3. Throughout, the prior presented in (3.8) is adopted. As the log-normal survival function does not have a closed analytical form, data augmentation [Tanner and Wong, 1987] is employed in order to accommodate censored and set observations. Conditional on the mixing parameters, survival times $t^* = (t_1^*, \ldots, t_n^*)'$ are simulated in line with the censoring. Based on the mixing representation, it follows that

$$\log(T_i)|\beta, \sigma^2, \Lambda_i = \lambda_i \sim \text{ Normal}\left(x_i'\beta, \frac{\sigma^2}{\lambda_i}\right). \tag{3.11}$$

Therefore, for right-censored observations, $\log(t_i^*)$ is drawn from a truncated version of (3.11) to $(\log(t_i), \infty)$, where t_i denotes the recorded censored time. Analogously, for set observations, the range of (3.11) is truncated to $(\log(t_i - \epsilon_l), \log(t_i + \epsilon_r))$. Obviously, if $t_i < \epsilon_l$, the lower bound corresponds to $-\infty$. Conditional on these simulated values, the other steps can be treated as if there were no censoring nor set observations. Regardless of the mixing distribution and provided that $n > 2 - 2p$, the full conditionals for β and σ^2 are respectively given by

$$\beta|\sigma^2, \theta, \lambda, t^* \sim \text{Normal}_k\left((X'X)^{-1}X'Dy^*, \sigma^2(X'X)^{-1}\right), \tag{3.12}$$

$$\sigma^2|\beta, \theta, \lambda, t^* \sim \text{Inv-Gamma}\left(\frac{n+2p-2}{2}, \frac{1}{2}(y^* - X\beta)'D(y^* - X\beta)\right), \tag{3.13}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)'$, $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $y^* = (\log(t_1^*), \ldots, \log(t_n^*))'$. Metropolis-Hastings updates are not required for these parameters. In contrast, the full conditionals for $\Lambda_1, \ldots, \Lambda_n$ and θ are generally not of a known form for

arbitrary mixing distributions. The full conditional for θ is given by

$$\pi(\theta|\beta,\sigma^2,\lambda,t^*) \propto \pi(\theta)\prod_{i=1}^{n} dP(\lambda_i|\theta). \tag{3.14}$$

Metropolis updates for θ are implemented under the adaptive scheme in Roberts and Rosenthal [2009]. For the Λ_i's, $\pi(\lambda_1,\ldots,\lambda_n|\beta,\sigma^2,\theta,t^*) = \prod_{i=1}^{n}\pi(\lambda_i|\beta,\sigma^2,\theta,t^*)$ with

$$\pi(\lambda_i|\beta,\sigma^2,\theta,t^*) \propto \lambda_i^{\frac{1}{2}} \exp\left\{-\frac{\lambda_i}{2\sigma^2}(\log(t_i^*)-x_i'\beta)^2\right\} dP(\lambda_i|\theta),\ i=1,\ldots,n. \tag{3.15}$$

The difficulty of this algorithm is strongly related to the complexity of the mixing distribution. For example, for the log-Student t and log-Laplace models, (3.15) simplifies to

$$\Lambda_i|\beta,\sigma^2,\theta,t^* \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{1}{2}\left[\frac{(\log(t_i^*)-x_i'\beta)^2}{\sigma^2}+\nu\right]\right), \tag{3.16}$$

$$\Lambda_i|\beta,\sigma^2,\theta,t^* \sim \text{Inv-Gaussian}\left(\frac{\sigma}{|\log(t_i^*)-x_i'\beta|}, 1\right), \tag{3.17}$$

respectively. If (3.15) does not have a known closed form, Metropolis-Hastings updates can be implemented. However, if evaluating the mixing density is cumbersome (as in *e.g.* the log-exponential power and log-logistic distributions), Metropolis updates are very challenging (and inefficient, as they might require long running times). Instead, for the log-logistic model, the rejection sampling algorithm proposed in Holmes and Held (2006, p. 163) is implemented. It uses the fact that $(2\sqrt{\Lambda_i})^{-1}$ has the asymptotic Kolmogorov-Smirnov distribution [Devroye, 1986, p. 151]. In the case of the log-exponential power model, one possible approach to this is to use a hierarchical representation for the positive stable distributions [as in Ibragimov and Chernin, 1959]. Nonetheless, the latter requires the use of n extra augmenting variables and is not appropriate when the value of α is unknown [Tsionas, 1999]. Instead, the mixture of uniforms representation used in Martï¿½n and Pï¿½rez [2009] is adopted. This replaces the use of Λ_i by U_i $(i=1,\ldots,n)$. In this case,

$$\log(T_i)|\beta,\sigma^2,\alpha,U_i=u_i \sim \text{Unif}\left(x_i'\beta-\sigma u_i^{1/\alpha}, x_i'\beta+\sigma u_i^{1/\alpha}\right), \tag{3.18}$$

with $U_i \overset{iid}{\sim} \text{Gamma}(1+1/\alpha,1)$. Consistently with the SMLN representation, the range of α is restricted to $(1,2)$. The cases where $\alpha=1$ and $\alpha=2$ are excluded but they covered by the log-Laplace and log-normal models, respectively. The data

augmentation strategy for censored and set observations must be adapted to this setting. As in the standard SMLN case, $\log(t_1^*), \ldots, \log(t_n^*)$ are simulated from a truncated versions of (3.18) to $(\log(t_i), \infty)$ and $(\log(t_i - \epsilon_l), \log(t_i + \epsilon_r))$ for right-censored and set records, respectively. For the log-exponential power model, the posterior distribution $\pi(\beta, \sigma^2, \alpha, u_1, \ldots, u_n|t)$ can be decomposed as $\pi(\beta, \sigma^2, \alpha|t^*) \times \prod_{i=1}^n \pi(u_i|\beta, \sigma^2, \alpha, t^*)$. As a consequence, the following full conditionals are defined

$$\pi(\beta|\sigma^2, \alpha, t^*) \propto \exp\left\{-\frac{1}{\sigma^\alpha}\sum_{i=1}^n |\log(t_i^*) - x_i'\beta|^\alpha\right\}, \tag{3.19}$$

$$\pi(\sigma^2|\beta, \alpha, t^*) \propto (\sigma^2)^{-(n/2+p)} \exp\left\{-\frac{1}{\sigma^\alpha}\sum_{i=1}^n |\log(t_i^*) - x_i'\beta|^\alpha\right\}, \tag{3.20}$$

$$\pi(\alpha|\beta, \sigma^2, t^*) \propto \frac{\alpha^n}{\Gamma^n(1/\alpha)} \exp\left\{-\frac{1}{\sigma^\alpha}\sum_{i=1}^n |\log(t_i^*) - x_i'\beta|^\alpha\right\}\pi(\alpha), \tag{3.21}$$

$$\tag{3.22}$$

As none of the above has a known form, adaptive Metropolis-Hastings steps are implemented for each parameter. In addition, the full conditionals for the mixing parameters are given by the following truncated exponential distributions

$$\pi(u_i|\beta, \sigma^2, \alpha, t^*) \propto e^{-u_i},\ u_i > \left(\frac{|\log(t_i^*) - x_i'\beta|}{\sigma}\right)^\alpha,\ i = 1, \ldots, n. \tag{3.23}$$

A simulation study shown that standard Bayesian model comparison criteria can fairly easily identify the need of incorporating unobserved heterogeneity to the model, even with rather small sample sizes and a considerable amount of censoring (see Appendix C). Ignoring unobserved heterogeneity can lead to biased or less precise inference for the regression parameters, whereas inference with SMLN models works well even in the absence of unobserved heterogeneity. The best results in terms of identifying the correct model are obtained for the independence Jeffreys prior and the model selection criteria DIC and PsBF.

3.2.5 Outlier detection for SMLN-AFT models

Outlying observations (in relation to a log-normal model with no mixture) can be detected using the methodology introduced in Section 2.4. It compares the posterior behaviour of the mixing parameters with respect to a reference value λ_{ref}. Section 2.4 suggests $\lambda_{ref} = \mathrm{E}(\Lambda_i|\theta)$ (if such expectation exists). Using this rule, $\lambda_{ref} = 1$ for the log-Student t model. This choice was supported by the empirical examples explored in Chapter 4. However, the expectation of the mixing distributions that

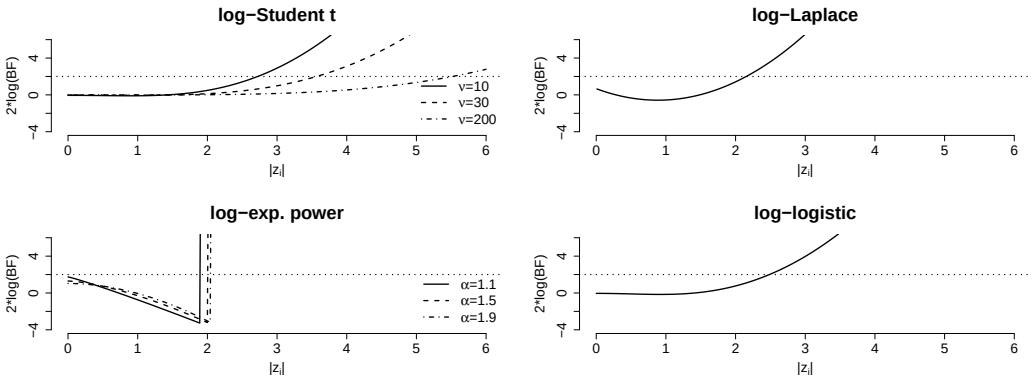


Figure 3.2: Bayes factor for outlier detection as a function of $|z_i|$. The log Bayes factor has been re-scaled by 2 in order to apply the interpretation rule proposed in Kass and Raftery [1995]. The dotted horizontal line is the threshold above which observations will be considered outliers.

generate the log-Laplace and log-logistic distributions do not exist. In these cases, λ_{ref} can be determined in an empirical manner. For example, simulated datasets indicate a large heterogeneity between the posterior distributions of the Λ_i's and the existence of a unique reference value is not clear (even in the absence of outlying observations). Nonetheless, on average, the posterior medians of the mixing parameters are generally close to unity (in this calculation, the lowest 25% of λ_i values is discarded in order to remove the influence of any possible outliers). Hence, $\lambda_{ref} = 1$ is proposed for the log-Laplace model. In the log-logistic case, the posterior distributions of the Λ_i's behave as in the log-Student t case, where the reference value is clearer. Using the same argument as in the log-Laplace case, $\lambda_{ref} = 0.4$ is defined for the log-logistic model. Figure 3.2 illustrates the performance of these reference values by plotting the Bayes factor in (2.7) for the log-Student t and in (2.8) for the log-Laplace and log-logistic models against a standardized log survival time z_i (given β, σ^2 and θ). This is defined as $\log(t_i)$ minus its mean, divided by its standard deviation (*i.e.* $\sigma\sqrt{\mathrm{E}_{\Lambda_i}(\Lambda_i^{-1}|\theta)}$). For the log-Student t, log-Laplace and log-logistic models, $z_i = \frac{\log(t_i)-x_i'\beta}{\sigma}\sqrt{\frac{\nu-2}{\nu}}$ (for $\nu > 2$), $z_i = \frac{\log(t_i)-x_i'\beta}{\sigma}\frac{1}{\sqrt{2}}$ and $z_i = \frac{\log(t_i)-x_i'\beta}{\sigma}\frac{\sqrt{3}}{\pi}$, respectively. As expected, large values of $|z_i|$ lead to evidence in favour of an outlier. The log-Student t model with very large number of degrees of freedom requires exceptionally large $|z_i|$ values to distinguish it from the log-normal case.

The log-exponential power model is a special case. As explained in Subsection 3.2.4, Bayesian inference for this model is implemented through a mixture of uniforms representation (with mixing parameters denoted by U_i). The models for

outlier detection in terms of U_i are $M_0 : U_i = u_{ref}$ versus $M_1 : U_i \neq u_{ref}$ (with all other $U_j, j \neq i$ free). The expectation of U_i (given α) is $1 + 1/\alpha$ and, according to the intuition presented in Section 2.4, it might be used as u_{ref}. With this rule, u_{ref} is a function of α which lies in $(1.5, 2)$. In practice, this choice detected large amounts of outliers (even for datasets generated from the log-normal model). In this case, $\pi(u_{ref}|t)$ is estimated by averaging $\pi(u_{ref}|\beta, \sigma^2, \alpha, t^*)$ in (3.23) using an MCMC sample from the posterior distribution of $(\beta, \sigma^2, \alpha)$ and the augmented survival times t^*. Hence, if the value of $(t_i^*, \beta, \sigma^2, \alpha)$ is such that $u_{ref} \leq \left(\frac{|\log(t_i^*) - x_i'\beta|}{\sigma}\right)^{\alpha}$, $\pi(u_{ref}|\beta, \sigma^2, \alpha, t^*)$ is equal to zero and the BF in favour of the observation i being an outlier is computed as infinity. Simulated datasets indicated that the means and medians of $\left(\frac{|\log(t_i^*) - x_i'\beta|}{\sigma}\right)^{\alpha}, i = 1, \ldots, n$, are around 0.6, regardless of the model from which the data was generated. Hence, the reference value is adjusted to $u_{ref} = 1 + 1/\alpha + 0.6$ (which lies in $(2.1, 2.6)$). This choice performed much better with simulated datasets (*e.g.* using log-normal data no outliers were detected). The resulting BF as a function of $z = \frac{\log(t) - x'\beta}{\sigma}\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$ (see Figure 3.2) are not much affected by the value of α.

For all models, moderate changes to the reference values do not have a large impact on the outlier detection curves in Figure 3.2.

3.3 The family of Rate Mixtures of Weibulls

Definition 3. *Let T_i be a positive-valued random variable distributed as a Rate Mixture of Weibull distributions (RMW). Its density function is defined as*

$$f(t_i|\alpha, \gamma, \theta) = \int_0^{\infty} \gamma\alpha\lambda_i t_i^{\gamma-1} e^{-\alpha\lambda_i t_i^{\gamma}} \, dP_{\Lambda_i}(\lambda_i|\theta), \quad t_i > 0, \alpha, \gamma > 0, \theta \in \Theta, \quad (3.24)$$

where λ_i is a realization of a random variable Λ_i which has distribution function $P_{\Lambda_i}(\cdot|\theta)$ defined on $\mathcal{L} \subseteq \mathbb{R}_+$ (possibly discrete). Denote this by $T_i \sim RMW_P(\alpha, \gamma, \theta)$. A hierarchical representation of (3.24) is given by

$$T_i|\alpha, \gamma, \Lambda_i = \lambda_i \sim Weibull(\alpha\lambda_i, \gamma), \qquad \Lambda_i|\theta \sim P_{\Lambda_i}(\cdot|\theta). \quad (3.25)$$

In line with Jewell [1982] and Kottas [2006], infinite mixtures of Weibull distributions are studied here. Nonetheless, discrete mixtures are explored in *e.g.* Tsionas [2002] and Marín et al. [2005]. In the previous literature, γ is often fixed at 1 and the mixing parameters are assigned a Gamma distribution [*e.g.* Abbring and Van Den Berg, 2007]. Throughout, the case with $\gamma = 1$ is referred as the Rate Mixtures of Exponentials (RME) family and it is denoted by $T_i \sim \text{RME}_P(\alpha, \theta)$.

The RME case can be extended to the RMW family via a power transformation. In fact, if $T_i \sim \text{RME}_P(\alpha, \theta)$ then $T_i^{1/\gamma} \sim \text{RMW}_P(\alpha, \gamma, \theta)$. When $\gamma = 2$, (3.24) reduces to an infinite mixture of Rayleigh distributions [as in Hansen and Meno, 1977; Gómez-Déniz and Gómez-Déniz, 2013]. As shown in Jewell [1982], RMW models are characterized as the distributions for which the survival function of $T_i^{1/\gamma}$ is completely monotone in $(0, \infty)$ (*i.e.* $(-1)^m \frac{d^m}{dt_i^m} S(t_i^{1/\gamma}) \geq 0$ for all $m = 0, 1, \ldots$). If $\gamma \leq 1$, the hazard rate of the resultant distribution decreases regardless of the mixing distribution [Marshall and Olkin, 2007]. In contrast, non-monotonic behaviours can be observed for $\gamma > 1$.

The following theorem provides some identifiability conditions for (α, γ, θ). In particular, it precludes the use of (separate) unknown scale parameters for the mixing distribution. This can be achieved by either fixing scale parameters of the mixing distribution or by imposing the restriction $\text{E}(\Lambda_i|\theta) = 1$. The latter is adopted hereafter for a Gamma mixing, since it leads to better mixing when implementing posterior inference by means of the MCMC sampler described in Subsection 3.3.4. For the other mixtures explored here, the sampler performs better if the scale parameters of the mixing distribution are fixed.

Theorem 7. *Let T_i be a random variable distributed according to (3.24). (α, γ, θ) is identified by the distribution of T_i if and only if:* ***(i)*** *$E(\Lambda_i|\theta)$ is finite and* ***(ii)*** *(α, θ) is identified by the distribution of $\alpha\Lambda_i$.*

Random variables in the RMW family do not necessarily have finite moments of any order and the existence of finite moments is linked to the moments of $\Lambda_i^{-1/\gamma}$.

Theorem 8. *Let T_i be a random variable distributed according to (3.24). The r-th moment of T_i ($r \geq 0$) is finite if and only if $E_{\Lambda_i}(\Lambda_i^{-r/\gamma}|\theta) < \infty$. If it exists, it corresponds to $\Gamma(1 + r/\gamma)\, \alpha^{-r/\gamma} E_{\Lambda_i}(\Lambda_i^{-r/\gamma}|\theta)$.*

Corollary 2. *If all the following expressions are well defined, the coefficient of variation c_v (i.e. the ratio between the standard deviation and the expected value) of the survival distributions in (3.24) is*

$$c_v(\gamma, \theta) = \sqrt{\frac{\Gamma(1 + 2/\gamma)}{\Gamma^2(1 + 1/\gamma)} \underbrace{\frac{var_{\Lambda_i}(\Lambda_i^{-1/\gamma}|\theta)}{E^2_{\Lambda_i}(\Lambda_i^{-1/\gamma}|\theta)}}_{(c_v^*(\gamma,\theta))^2} + \underbrace{\frac{\left[\Gamma(1 + 2/\gamma) - \Gamma^2(1 + 1/\gamma)\right]}{\Gamma^2(1 + 1/\gamma)}}_{(c_v^W(\gamma))^2}}. \quad (3.26)$$

The expression in (3.26) simplifies to $\sqrt{2\frac{var_{\Lambda_i}(\Lambda_i^{-1}|\theta)}{E^2_{\Lambda_i}(\Lambda_i^{-1}|\theta)} + 1}$ when $\gamma = 1$.

Corollary 2 indicates that

Table 3.2: Examples in the RME family. $K_p(\cdot)$ stands for the modified Bessel function. $\Theta = (0, \infty)$, unless specified.

Mixing density	$\mathrm{E}(\Lambda_i\|\theta)$	$f(t_i\|\alpha,\theta)$	$h(t_i\|\alpha,\theta)$
Exponential(1)	1	$\alpha(\alpha t_i+1)^{-2}$	$\alpha(\alpha t_i+1)^{-1}$
Gamma(θ,θ)	1	$\alpha([\alpha/\theta]\,t_i+1)^{-(\theta+1)}$, $\theta>2$	$\alpha([\alpha/\theta]\,t_i+1)^{-1}$
Inv-Gamma$(\theta,1)$	$\frac{1}{\theta-1}$	$\frac{2\alpha}{\Gamma(\theta)}K_{-(\theta-1)}(2\sqrt{\alpha t_i})(\alpha t_i)^{(\theta-1)/2}, \theta>1$	$\sqrt{\frac{\alpha}{t_i}}\frac{K_{-(\theta-1)}(2\sqrt{\alpha t_i})}{K_{-\theta}(2\sqrt{\alpha t_i})}$
Inv-Gauss$(\theta,1)$	θ	$\alpha\, e^{1/\theta}\left[\frac{1}{\theta^2}+2\alpha t_i\right]^{-1/2} e^{-\left[\frac{1}{\theta^2}+2\alpha t_i\right]^{1/2}}$	$\alpha\left[\frac{1}{\theta^2}+2\alpha t_i\right]^{-1/2}$
Log-Normal$(0,\theta)$	$e^{\theta/2}$	$\frac{\alpha}{\sqrt{2\pi\theta}}\int_0^\infty e^{-\alpha\lambda_i t_i}\, e^{-\frac{(\log(\lambda_i))^2}{2\theta}}\, d\lambda_i$	No closed form

(i) $c_v(\gamma,\theta)$ is an increasing function of $c_v^*(\gamma,\theta)$, which is the coefficient of variation of $\Lambda_i^{-1/\gamma}$ given θ,

(ii) for the same value of γ, the coefficient of variation of the Weibull distribution $c_v^W(\gamma)$ is a lower bound for $c_v(\gamma,\theta)$ and they are equal if and only if $\Lambda_i = \lambda_0$ with probability 1. Hence, evidence of unobserved heterogeneity can be quantified in terms of the ratio

$$R_{c_v}(\gamma,\theta) = \frac{c_v(\gamma,\theta)}{c_v^W(\gamma)}, \tag{3.27}$$

defined as the inflation that the mixture induces in the coefficient of variation (with respect to a Weibull model with the same γ). If θ is such that $c_v^*(\gamma,\theta)$ goes to zero, then $R_{c_v}(\gamma,\theta)$ tends to one and the mixture reduces to the underlying Weibull model itself. If $\gamma \to 0$, $c_v^W(\gamma)$ and, consequently, $c_v(\gamma,\theta)$ become unbounded. In that case, $R_{c_v}(\gamma,\theta)$ behaves as $\sqrt{[c_v^*(\gamma,\theta)]^2+1}$. If $\gamma = 1$, then $R_{c_v}(\gamma,\theta) = c_v(1,\theta)$.

Throughout, the range of (γ,θ) is restricted such that c_v is finite (this restriction is not required when θ does not appear). This decision facilitates the implementation of Bayesian inference (see Subsection 3.3.2).

The survival function generated by (3.24) corresponds to the Laplace transform of the mixing density evaluated in αt_i^γ [Wienke, 2010]. Therefore, mixing densities with known Laplace transform are an attractive choice. An example of this is the PVF family (see also Subsection 2.5.1). In particular, a positive stable mixing distribution yields a marginal model which is the Weibull distribution itself. Table 3.2 summarizes some examples in the RME family. This list can be enlarged by simply varying the mixing distribution. All these examples can be extended to

the RMW case using the power transformation that was introduced shortly after (3.25). A Gamma mixing generates the Lomax distribution [Lomax, 1954] which is widely used in the literature as a heavy tailed distribution in finance and other contexts. In contrast, some other mixing distributions (such as the log-normal) do not lead to analytical expressions for the resulting density.

Figure 3.3 shows the RME densities produced by the examples in Table 3.2 and different values of θ. The density is decreasing, like in the exponential case. Nevertheless, the behaviour exhibited by the tails is very flexible. Figure 3.3 also presents the hazard rate for these mixtures. As shown in Marshall and Olkin [2007], they are decreasing functions of the survival time but the gradient varies among the different mixing distributions. Figure 3.4 illustrates the effect of a Gamma(θ, θ) mixing distribution for distributions in the RMW family (with free γ). Whereas the shape of the density function was not greatly affected in this example, the effect of the mixture on the hazard rate is more pronounced. For instance, while the hazard rate of the Weibull model with $\gamma = 2$ is an increasing function of t_i, the hazard of the corresponding mixture exhibits a non-monotonic behaviour.

3.3.1 The RMW-AFT model

A Weibull survival regression can be equivalently written in terms of PH and AFT specifications. Let x_i be a vector containing the value of k covariates associated with the survival time i and $\beta \in \mathbb{R}^k$ be a vector of parameters. In the RMW-AFT model, the covariates affect the time scale through the parameter α. This model is defined as

$$T_i \sim RMW_P(\alpha_i, \gamma, \theta), \quad \alpha_i = e^{-\gamma x_i' \beta}, \quad i = 1, \ldots, n. \tag{3.28}$$

Alternatively, the RMW-AFT model can be expressed as

$$\log(T_i) = x_i'\beta + \log(\Lambda_i^{-1/\gamma} T_0), \tag{3.29}$$

where $\Lambda_i \sim dP_{\Lambda_i}(\theta)$ and $T_0 \sim$ Weibull$(1, \gamma)$. As explained in Subsection 2.5.2, the RMW-AFT is itself an AFT model with baseline survival function defined by the distribution of $T_0' = \Lambda_i^{-1/\gamma} T_0$ and $T_0' \sim \text{RMW}_P(1, \gamma, \theta)$. Under this model, e^{β_j} can be interpreted as the proportional marginal change of the median (or any other percentile) survival time after a unit change in covariate j. For $\beta^* = -\gamma\beta$, (3.28) is equivalent to the RMW-PH model which is defined as

$$h(t_i|\beta^*, \gamma, \Lambda_i = \lambda_i; x_i) = \lambda_i \gamma t_i^{\gamma-1} e^{x_i'\beta^*}, \quad \Lambda_i \sim dP(\Lambda_i|\theta), \quad i = 1, \ldots, n. \tag{3.30}$$

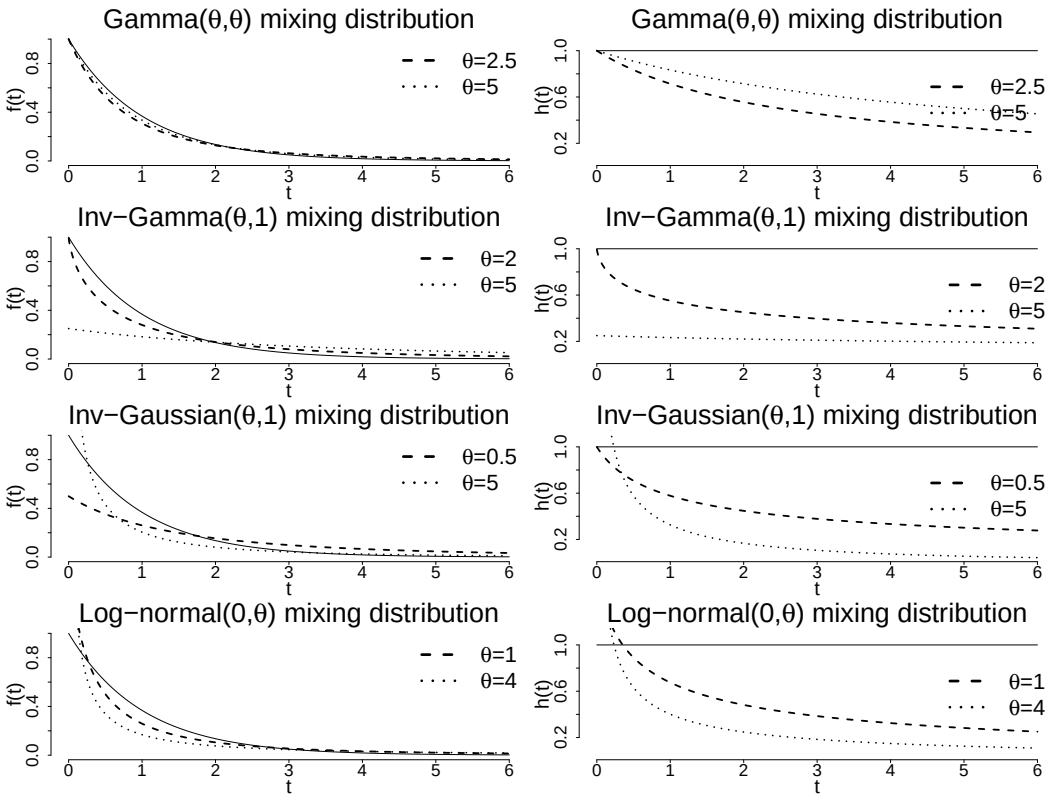


Figure 3.3: Density and hazard function (left and right panels, respectively) of some RME models ($\alpha = 1$). The solid line is the Exponential(1) density (or hazard).

However, the marginal model generated by (3.30) does not generally retain the PH property (see also Subsection 2.5.1). The only mixing distribution that retains this property is the positive stable distribution [Wienke, 2010], where the marginal model is the Weibull itself. In this setting, the interpretation of the regression coefficients is conditional on the random effect (individual level). Unlike for the RMW-AFT model, this interpretation cannot be extended to the population level. Most of the earlier literature for unobserved heterogeneity is in terms of the PH model. Nevertheless, here results are presented in terms of the RMW-AFT presentation since the interpretation of the regression coefficients is clearer and the mixture model is still an AFT model.

3.3.2 A weakly informative prior for the RMW-AFT model

First, a prior is defined for the RME-AFT model (*i.e.* fixing $\gamma = 1$). In the absence of prior information, a popular choice is to use priors based on the Jeffreys rule [Jef-

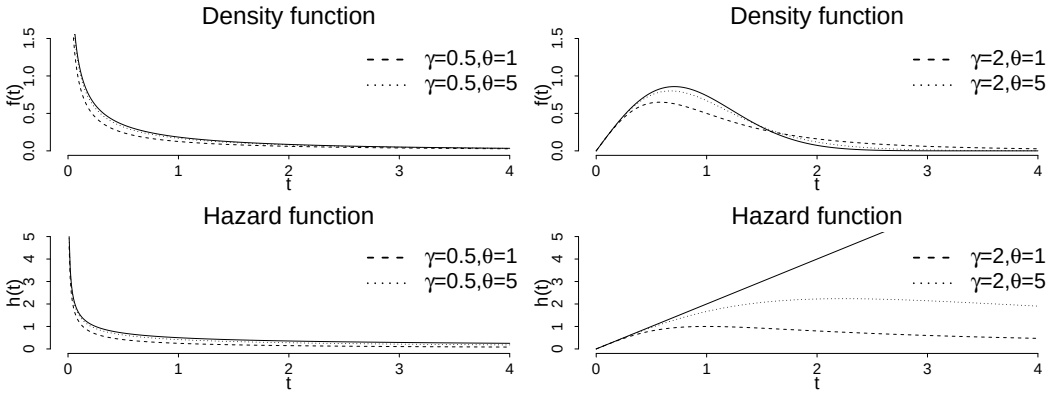


Figure 3.4: Some RMW models ($\alpha = 1$). The mixing distribution is Gamma(θ, θ) (Exponential(1) for $\theta = 1$). The solid line is the Weibull$(1, \gamma)$ density (or hazard).

freys, 1961]. Jeffreys-style priors require the FIM which is provided by the following Theorem.

Theorem 9. *Let $T_1, \ldots, T_n$ be independent random variables with T_i distributed according to (3.28) with $\gamma = 1$, then their FIM corresponds to*

$$FIM(\beta, \theta) = \begin{pmatrix} k_1(\theta) X'X & k_2(\theta) X' \mathbf{1}_n \\ k_2(\theta) \mathbf{1}_n' X & n k_3(\theta) \end{pmatrix}, \tag{3.31}$$

where $k_1(\theta), k_2(\theta)$ and $k_3(\theta)$ are functions depending only on θ, $X = (x_1 \cdots x_n)'$ and $\mathbf{1}_n$ is a column vector of n ones.

Corollary 3. *Under the same assumptions as in Theorem 9, assume also that X has rank k $(n > k)$ and θ is a scalar parameter. The Jeffreys prior and the independence Jeffreys prior (which deals separately with the blocks for β and θ) for the RME-AFT model are correspondingly given by*

$$\pi^J(\beta, \theta) \propto k_1^{k/2}(\theta) k_3^{1/2}(\theta) \left[1 - \frac{k_2^2(\theta)}{n k_1(\theta) k_3(\theta)} \mathbf{1}_n' X (X'X)^{-1} X' \mathbf{1}_n \right]^{1/2}, \tag{3.32}$$

$$\pi^I(\beta, \theta) \propto k_3^{1/2}(\theta). \tag{3.33}$$

These two Jeffreys-style priors can be expressed as

$$\pi(\beta, \theta) \propto \pi(\theta), \tag{3.34}$$

where $\pi(\theta)$ is the component of the prior that depends on θ. Although Corollary 3 provides certain structure for the priors based on the Jeffreys rule, the actual

Table 3.3: Relationship between c_v and θ for some distributions in the RME family.

Mixing density	Range of c_v	$c_v(\theta)$	$\left\lvert\frac{dc_v(\theta)}{d\theta}\right\rvert$
Gamma(θ,θ)	$(1,\infty)$	$\sqrt{\frac{\theta}{\theta-2}}$	$\theta^{-1/2}(\theta-2)^{-3/2}$
Inv-Gamma$(\theta,1)$	$(1,\sqrt{3})$	$\sqrt{\frac{\theta+2}{\theta}}$	$\theta^{-3/2}(\theta+2)^{-1/2}$
Inv-Gaussian$(\theta,1)$	$(1,\sqrt{5})$	$\sqrt{\frac{5\theta^2+4\theta+1}{\theta^2+2\theta+1}}$	$\frac{3\theta+1}{(5\theta^2+4\theta+1)^{1/2}(\theta+1)^2}$
Log-Normal$(0,\theta)$	$(1,\infty)$	$\sqrt{2\,e^{\theta}-1}$	$e^{\theta}(2\,e^{\theta}-1)^{-1/2}$

expressions are not easily derived. Even when assuming a particular mixing distribution, it is not trivial to obtain $k_1(\theta), k_2(\theta)$ and $k_3(\theta)$. One alternative is to compute the FIM directly from the resultant density (as in the proof of Theorem 5). For example, in the simple case of a Gamma$(\theta,1)$ mixing distribution the Jeffreys and independence Jeffreys prior are given respectively by

$$\pi^J(\beta,\theta) \equiv \pi^J(\theta) \propto \left[\frac{\theta}{\theta+2}\right]^{k/2} \frac{1}{\theta}\left[1-\frac{\theta(\theta+2)}{n(\theta+1)^2}\mathbf{1}_n' X(X'X)^{-1}X'\mathbf{1}_n\right]^{1/2} \tag{3.35}$$

and

$$\pi^I(\beta,\theta) \equiv \pi^I(\theta) \propto \frac{1}{\theta}. \tag{3.36}$$

Even though this is one the simplest cases in the RME family, $\pi^J(\theta)$ is very involved. It depends on the number of covariates, the sample size and the design matrix. These priors become more complicated and have no easy derivation for other mixtures. In particular, if the resultant distribution does not have a closed analytical form (*e.g.* with a log-normal mixing distribution), computing the FIM is very challenging. In addition, there is no guarantee of having a proper prior for θ when using an arbitrary mixing distribution. For instance, in the Lomax case, $\pi^J(\theta)$ and $\pi^I(\theta)$ are not proper density functions (both behave as $1/\theta$ for large values of θ). As the role of θ is specific to each mixture, improper priors for θ will not allow the comparison between RME models using Bayes factors.

To overcome these issues, a simplification of the Jeffreys-style priors is proposed. It keeps the structure in (3.34) but assigns a proper $\pi(\theta)$. The comparison between models is meaningful if, regardless of the mixing distribution, $\pi(\theta)$ contains the same prior information (*i.e.* the priors are "matched"). This is achieved by exploiting the relationship between θ and c_v, the coefficient of variation of the survival times. A proper prior, which is common for all models, is assigned to c_v. Denote it by $\pi^*(c_v)$. As c_v does not involve β (expression (3.26) does not involve α), $\pi^*(c_v)$ only provides information about θ. Once $\pi^*(c_v)$ has been defined, $\pi(\theta)$

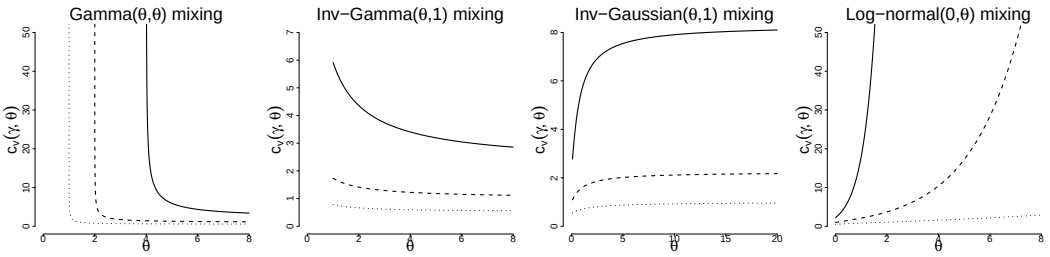


Figure 3.5: Relationship between (γ, θ) and c_v for some RMW models. Solid, dashed and dotted lines are for $\gamma = 0.5, 1$ and 2, respectively. Dashed lines indicate the relationship between θ and c_v for distributions in the RME family.

can be easily derived by means of a change of variables. Using (3.26), the functional relationship between c_v and θ for some distributions in the RME family is derived (see Table 3.3). The inverse function of $c_v(\theta)$ must exist ($c_v(\theta)$ must be injective), yet an explicit expression is not required. Injectivity holds for all the examples in Table 3.3 (Figure 3.5 illustrates this). The induced prior for θ is defined by

$$\pi(\theta) = \pi^*(c_v(\theta)) \left| \frac{dc_v(\theta)}{d\theta} \right| \tag{3.37}$$

When comparing with models without θ, meaningful results derive from the fact that the prior on θ is reasonable. Two natural choices for $\pi^*(c_v)$ are the truncated exponential and Pareto type I distributions (both on $(1, \infty)$) with hyper-parameters a and b, respectively. These priors cover a wide set of tails for c_v. Smaller values of a and b assign larger probabilities to small values of c_v (b is restricted to be larger than 1 in order to have a finite expectation for c_v). These hyper-parameters can be elicited from experts' opinion, for example, using the expected value of c_v. The expected values under these priors are $1 + 1/a$ and $b/(b-1)$ respectively, and with $b = a + 1$ the expected values are equated. When the range of c_v differs from $(1, \infty)$ (*e.g.* with the inverse Gamma and inverse Gaussian mixing distributions), these priors can be adjusted by truncating $\pi^*(c_v)$. If the values of a and b are such that the prior expectation of c_v falls outside the range allowed by a specific model, the prior is not consistent with that model. For example, the model generated by the inverse Gaussian mixing distribution should be discarded a priori if $a < (\sqrt{5}-1)^{-1}$ and $b < 1 + (\sqrt{5}-1)^{-1}$.

For a general RMW-AFT model with unknown γ, the structure of the FIM is more involved than the one presented in Theorem 9. As a consequence, priors based on the Jeffreys rule are not easy to obtain (there is also no guarantee of having

Table 3.4: $c_v^*(\gamma,\theta)$ and its partial derivative with respect to θ for some mixing distributions. $K_p(\cdot)$ and $\psi(\cdot)$ stand for the modified Bessel and the digamma functions, respectively.

Mixing	$[c_v^*(\gamma,\theta)]^2$	$\left\lvert\frac{d[c_v^*(\gamma,\theta)]^2}{d\theta}\right\rvert$
Gamma(θ,θ)	$\frac{\Gamma(\theta)\Gamma(\theta-2/\gamma)}{\Gamma^2(\theta-1/\gamma)}-1, \quad \theta>2/\gamma$	$\frac{\Gamma(\theta)\Gamma(\theta-2/\gamma)}{\Gamma^2(\theta-1/\gamma)}\left[\psi(\theta)+\psi(\theta-2/\gamma)-2\psi(\theta-1/\gamma)\right]$
Inv-Gamma$(\theta,1)$	$\frac{\Gamma(\theta)\Gamma(\theta+2/\gamma)}{\Gamma^2(\theta+1/\gamma)}-1$	$\frac{\Gamma(\theta)\Gamma(\theta+2/\gamma)}{\Gamma^2(\theta+1/\gamma)}\left[\psi(\theta)+\psi(\theta+2/\gamma)-2\psi(\theta+1/\gamma)\right]$
Inv-Gauss$(\theta,1)$	$\sqrt{\frac{\theta\pi}{2}}\,e^{-\frac{1}{\theta}}\frac{K_{-(\frac{2}{\gamma}+\frac{1}{2})}(1/\theta)}{K^2_{-(\frac{1}{\gamma}+\frac{1}{2})}(1/\theta)}-1$	$\sqrt{\frac{\pi}{2}}\frac{\theta^{-3/2}\,e^{-\frac{1}{\theta}}}{K^3_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta)}\Big[K_{-\left(\frac{2}{\gamma}+\frac{1}{2}\right)}(1/\theta)K_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta)$ $+K_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta)K_{-\left(\frac{2}{\gamma}-\frac{1}{2}\right)}(1/\theta)$ $-2K_{-\left(\frac{2}{\gamma}+\frac{1}{2}\right)}(1/\theta)K_{-\left(\frac{1}{\gamma}-\frac{1}{2}\right)}(1/\theta)\Big]$
Log-normal$(0,\theta)$	$e^{\theta/\gamma^2}-1$	$\frac{1}{\gamma^2}\,e^{\theta/\gamma^2}$

a proper component for θ). As an alternative, a prior for (β,γ,θ) is defined by extending the structure in (3.34) to

$$\pi(\beta,\gamma,\theta)\propto\pi(\gamma,\theta)\equiv\pi(\theta|\gamma)\pi(\gamma), \tag{3.38}$$

where $\pi(\theta|\gamma)$ is a proper density function of θ (given γ) and $\pi(\gamma)$ is a proper prior for γ. This implies a flat prior on β. The prior product structure between β and (γ,θ) in (3.38) is reasonable in the RMW-AFT model where the interpretation of β does not depend on γ nor θ. Conditional on the value of γ, $\pi(\theta|\gamma)$ is defined as in the RME-AFT case (*i.e.* via a prior for c_v, $\pi^*(c_v)$). Define $c_v(\gamma,\theta)$ and $c_v^*(\gamma,\theta)$ as in (3.26). Hence,

$$\pi(\theta|\gamma)=\pi^*(c_v(\gamma,\theta))\left|\frac{dc_v(\gamma,\theta)}{d\theta}\right|, \tag{3.39}$$

where

$$\frac{dc_v(\gamma,\theta)}{d\theta}=\frac{\Gamma(1+2/\gamma)}{\Gamma^2(1+1/\gamma)}\frac{1}{2c_v(\gamma,\theta)}\frac{d[c_v^*(\gamma,\theta)]^2}{d\theta}. \tag{3.40}$$

Table 3.4 contains $[c_v^*(\gamma,\theta)]^2$ and its partial derivative with respect to θ for the same mixing distributions used in Table 3.3. Although some of the expressions in Table 3.4 are complicated, they can be easily evaluated numerically. Figure 3.5 shows the relationship between (γ,θ) and c_v for some distributions in the RMW family. As in the RME case, a truncated exponential and Pareto type I prior distributions for c_v are suggested (given γ). These priors must be truncated to $(c_v^W(\gamma),\infty)$ (see (3.26)). However, as in the exponential mixtures, some mixing distributions impose a finite upper bound for c_v. This upper bound is equal to $\left[\frac{\Gamma^2(1+2/\gamma)}{\Gamma^4(1+1/\gamma)}-1\right]^{1/2}$ and $\left[\sqrt{\pi}\frac{\Gamma(1+2/\gamma)}{\Gamma^2(1+1/\gamma)}\frac{\Gamma(2/\gamma+1/2)}{\Gamma^2(1/\gamma+1/2)}-1\right]^{1/2}$ when using the inverse Gamma and inverse Gaus-

sian mixing distributions, respectively.

A proposal for $\pi(\gamma)$ is not trivial. In particular, a conjugate prior for γ in $(0, \infty)$ does not exist [Soland, 1969]. A discrete prior for γ is conjugate but restrictive and inappropriate in most real situations (especially where no prior information about γ is available). Berger and Sun [1993] and Kundu [2008] suggested the use of continuous log-concave priors for γ. Here, a Gamma prior is defined for γ (with a range of hyperparameters values that also allows for a not log-concave Gamma density). Hyper-parameters for this prior can be elicited using expert's opinion. For example, if the hazard function is expected to be monotonically decreasing, the prior must mostly support values in $(0, 1)$. Beliefs about non-monotone behaviours of the hazard rate are translated in priors for γ that assign more probability to the range $(1, \infty)$.

3.3.3 Posterior propriety for the RMW-AFT model

The following theorem covers posterior propriety for the RMW-AFT model under the weakly informative (improper) prior in (3.38). Following Proposition 1, it only considers the non-censored observations (using n instead of n_o for ease of notation). As a consequence, only sufficient conditions for posterior existence are derived.

Theorem 10. *Let $T_1, \ldots, T_n$ be the survival times of n independent individuals distributed as in (3.28). Assume that survival times $t_1, \ldots, t_n$ are observed and define $X = (x_1 \cdots x_n)'$. Assume that $n \geq k$, X has rank k (full rank) and that the prior for (β, γ, θ) is proportional to $\pi(\gamma, \theta)$, which is a proper density function for (γ, θ). If $t_i \neq 0$ for all $i = 1, \ldots, n$, the posterior distribution of (β, γ, θ) is proper.*

As mentioned in Section 3.3.2, the suggested prior for (γ, θ) is proper so that Theorem 10 assures a proper posterior distribution if X has full column rank and there are no zero observations of the survival time.

Posterior propriety can be precluded when conditioning on a particular sample of point observations which has zero Lebesgue measure (see Subsection 1.2.2). Nevertheless, point observations are not a major issue regarding to posterior propriety for the RMW-AFT model. In this case, the posterior distribution is well-defined as long as there are no individuals for which $t_i = 0$. Whereas the latter is a sensible assumption in most real applications, survival times can be recorded as zero due to rounding. In such a case, the point observation can be replaced by a set observation $(0, \epsilon)$, where ϵ stands for the minimum value that the recording mechanism detects (equivalent to a left censored observation on $(0, \epsilon)$).

3.3.4 Implementation

Here, only right-censoring is assumed, which is the most frequent situation in survival data. Bayesian inference for the RMW-AFT model, under the prior in (3.38), is implemented using the sampler described in Section 2.3. Mixing parameters are handled through data augmentation [Tanner and Wong, 1987]. As the Weibull survival function has a known simple form, data augmentation is not required for dealing with censored (and set) observations [Ibrahim et al., 2001; Kottas, 2006]. The full conditionals for the Gibbs sampler are

$$\pi(\beta_j|\beta_{-j},\gamma,\theta,\lambda,t;c) \propto e^{-\gamma\beta_j\sum_{i=1}^n(1-c_i)x_{ij}-\sum_{i=1}^n\lambda_i(t_i\,e^{-x_i'\beta})^\gamma}, j=1,\ldots,k, \quad (3.41)$$

$$\pi(\gamma|\beta,\theta,\lambda,t;c) \propto \gamma^{n-\sum_{i=1}^n c_i}\left[\prod_{i=1}^n t_i^{1-c_i}\right]^{\gamma-1} e^{-\gamma\sum_{i=1}^n(1-c_i)x_i'\beta} \times e^{-\sum_{i=1}^n\lambda_i(t_i\,e^{-x_i'\beta})^\gamma}\pi(\theta|\gamma)\pi(\gamma), \quad (3.42)$$

$$\pi(\theta|\beta,\gamma,\lambda,t;c) \propto \prod_{i=1}^n dP(\lambda_i|\theta)\pi(\theta|\gamma), \quad (3.43)$$

$$\pi(\lambda_i|\beta,\gamma,\theta,\lambda_{-i},t;c) \propto \lambda_i^{1-c_i}\,e^{-\lambda_i(t_i\,e^{-x_i'\beta})^\gamma}\,dP(\lambda_i|\theta), i=1,\ldots,n, \quad (3.44)$$

where $\beta_{-j}=(\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\beta_k)$, $\lambda_{-i}=(\lambda_1,\ldots,\lambda_{i-1},\lambda_{i+1},\lambda_n)$ and the c_i's, $i=1,\ldots,n$ are censoring indicators equal to 1 if the survival time for individual i is right censored and 0 otherwise (as in (1.1), in the absence of left and interval censoring).

For a general mixing distribution, Metropolis updates are required in all full conditionals. These are drawn using an adaptive scheme [Roberts and Rosenthal, 2009]. Nevertheless, Gibbs steps can be used for the λ_i's in case of particular mixing distributions. For instance, the first four mixing distributions in Table 3.2, respectively, lead to

$$\Lambda_i|\beta,\gamma,\theta,t;c \sim \text{Gamma}\left(2-c_i, 1+(t_i\,e^{-x_i'\beta})^\gamma\right), \quad (3.45)$$

$$\Lambda_i|\beta,\gamma,\theta,t;c \sim \text{Gamma}\left(\theta+1-c_i, \theta+(t_i\,e^{-x_i'\beta})^\gamma\right), \quad (3.46)$$

$$\Lambda_i|\beta,\gamma,\theta,t;c \sim \text{GIG}\left(-\theta+1-c_i, 2, 2(t_i\,e^{-x_i'\beta})^\gamma\right), \quad (3.47)$$

$$\Lambda_i|\beta,\gamma,\theta,t;c \sim \text{GIG}\left(1/2-c_i, 1, \theta^{-2}+2(t_i\,e^{-x_i'\beta})^\gamma\right). \quad (3.48)$$

In practice, the suggested prior led to very poor mixing of the chain when using a log-normal$(0,\theta)$ mixing. This relates to a strong a priori correlation between γ and θ, which persists when not much can be learned about θ (as θ controls the tails

of the distribution, this is especially problematic for small n and/or high proportion of censoring). A re-parametrization of the model from (θ, γ) to (θ^*, γ), where $\theta^* = \theta/\gamma^2$, is adopted. As in the original parametrization, a prior for θ^* can be induced via a prior for c_v ($[c^*(\gamma, \theta^*)]^2$ equals $e^{\theta^*} - 1$ in this case). This new parametrization is more orthogonal and substantially improves the mixing of the chain.

3.3.5 Outlier detection for RMW-AFT models

As in Section 2.4, outliers (with respect to the underlying Weibull model) are detected using the posterior distribution of the mixing variables. The suggested reference value is a valid option for RMW models because $\mathrm{E}(\Lambda_i|\theta)$ is always finite (given the identifiability constrains provided by Theorem 7). Therefore, $\lambda_{ref} = \mathrm{E}(\Lambda_i|\theta)$ is adopted. Table 3.2 displays $\mathrm{E}(\Lambda_i|\theta)$ for the mixing distributions presented as examples here. As advised in Section 2.4, when unknown, θ is replaced by its posterior median (based on a MCMC sample). Unlike in the SMLN case, empirical evidence does not support the latter choice for the censored observations. Only a lower bound of the survival time is known for right censored observations. Therefore, this is highly informative for the mixing parameters (as the λ_i's affect the scale of the underlying distribution). For this reason, the posterior distributions of the λ_i's linked to right censored observations are driven towards lower values (in line with the possibility of very large survival times). The proposal here is to keep $\lambda^o_{ref} = E(\Lambda_i|\theta)$ as the reference value for non-censored observations and adjust it for right-censored observations as follows:

$$\lambda^c_{ref} = C_i(\beta, \gamma, \theta)\lambda^o_{ref}, \text{ with } C_i(\beta, \gamma, \theta) = \frac{E(\Lambda_i|t_i, c_i = 1, \beta, \gamma, \theta)}{E(\Lambda_i|t_i, c_i = 0, \beta, \gamma, \theta)}. \tag{3.49}$$

For exponential mixing $C_i(\beta, \gamma, \theta) = 1/2$ and $C_i(\beta, \gamma, \theta) = \theta/(\theta + 1)$ for the Gamma mixing distribution (see the conditionals in Subsection 3.3.4). In these cases, the correction factor does not depend on i, β or γ. If $\Lambda_i \sim$ inv-Gamma$(\theta, 1)$ or $\Lambda_i \sim$ inv-Gaussian$(\theta, 1)$, $C_i(\beta, \gamma, \theta)$ is equal to

$$\frac{K^2_{-\theta+1}\left(2\sqrt{\left(t_i\, e^{-x_i'\beta}\right)^\gamma}\right)}{K_{-\theta+2}\left(2\sqrt{\left(t_i\, e^{-x_i'\beta}\right)^\gamma}\right) K_{-\theta}\left(2\sqrt{\left(t_i\, e^{-x_i'\beta}\right)^\gamma}\right)}, \text{ or} \tag{3.50}$$

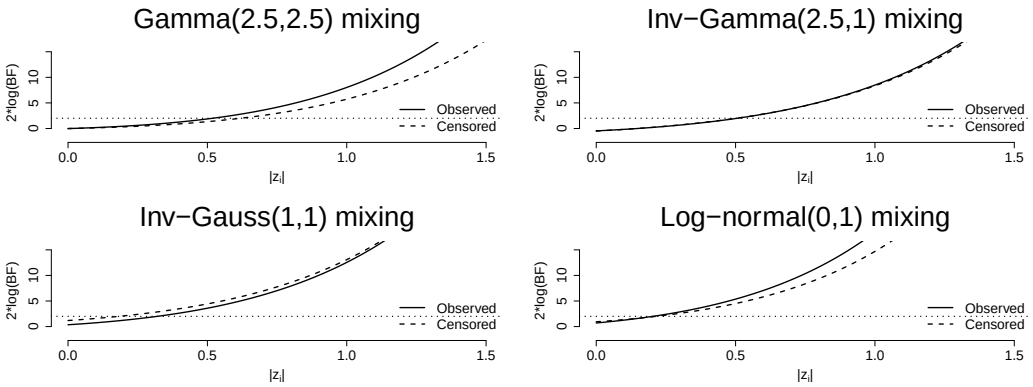


Figure 3.6: $2 \times$ log-Bayes factor for outlier detection as a function of $|z_i|$ in AFT-RMW models. The dotted horizontal line is the threshold above which observations will be considered outliers [according to the rule in Kass and Raftery, 1995].

$$\frac{K_{1/2}^2\left(\sqrt{2\left(t_i\, e^{-x_i'\beta}\right)^{\gamma}+\theta^{-2}}\right)}{K_{-1/2}\left(\sqrt{2\left(t_i\, e^{-x_i'\beta}\right)^{\gamma}+\theta^{-2}}\right) K_{3/2}\left(\sqrt{2\left(t_i\, e^{-x_i'\beta}\right)^{\gamma}+\theta^{-2}}\right)}, \tag{3.51}$$

respectively, where $K_p(\cdot)$ is the modified Bessel function. For the log-normal mixing distribution $C_i(\beta, \gamma, \theta)$ has no closed form but can be estimated via numerical integration. The performance of these reference values has been validated using simulated datasets.

To illustrate this outlier detection method, Figure 3.6 displays $BF_{01}^{(i)}$ as a function of a standardized observation z_i. Following the AFT structure in (3.29), this is defined in terms of $\log(t_i)$ minus its mean (if finite), divided by its standard deviation (given β, γ and θ). Using that $\log(T_0) \sim \text{Gumbel}(0, \gamma^{-1})$, it follows that

$$z_i = \gamma\left[\frac{\log(t_i) - x_i'\beta + \gamma^{-1}\left(\text{E}_{\Lambda_i}(\log(\Lambda_i)|\theta) + \psi(1)\right)}{\sqrt{\text{var}_{\Lambda_i}(\log(\Lambda_i)|\theta) + \pi^2/6}}\right], \tag{3.52}$$

where $\psi(\cdot)$ denotes the digamma function. In terms of z_i, $BF_{01}^{(i)}$ does not depend on β nor γ (the full conditional of Λ_i depends on t_i only through $\left[t_i\, e^{-x_i\beta}\right]^{\gamma}$). Naturally, outliers relate to large values of $|z_i|$. The threshold on $|z_i|$ at which an observation is detected as outlier depends on θ. For example, for a Gamma(θ, θ) mixing, this threshold is an increasing function of θ. The RMW model with gamma mixing tends to the Weibull model as $\theta \to \infty$ and thus, the model with large θ requires large $|z_i|$ values to distinguish it from the Weibull. As shown in Figure 3.6, the correction

factor $C_i(\beta, \gamma, \theta)$ induces a similar outlier detection threshold (in terms of $|z_i|$) for censored and non-censored observations.

3.4 Concluding remarks

As in Chapter 2, mixtures of life distribution are used in order to account for unobserved heterogeneity in survival models. In particular, mixtures generated from log-normal and Weibull underlying models were explored in detail. These families produce distributions with a variety of tail behavior, making their use applicable in a wide range of situations. In the SMLN case, the mixing is applied to the shape parameter of the log-normal distribution and the resulting density has a quite flexible shape which can be adjusted by choosing the mixing distribution. Instead, in RMW models, the mixture is introduced through a rate (scale) parameter and the shape of RMW densities is not much affected by the mixture with respect to the underlying Weibull shape (although the mass is redistributed). In both cases, the mixing has an important effect over the hazard function.

The prior distributions adopted in this Chapter are inspired by the Jeffreys rule, which is particularly useful in the absence of reliable prior information or as a benchmark analysis. A general representation of the FIM, which is the basis of Jeffreys-style priors, is provided for the SMLN-AFT and RME-AFT models. In both cases, regardless of the mixing distribution, the induced priors can be factorized as an (improper) flat prior for the regression coefficients times a (possibly improper) prior for the other model parameters. In view of the clear interpretation of β (which does not depend on the mixing), this product structure of the prior seems a reasonable assumption. An explicit expression for the Jeffreys prior (and two of its variations) for SMLN-AFT models can be found in Corollary 1 and the proof of Theorem 5. For the examples analyzed here, these priors produced a proper component for θ. In contrast, the Jeffreys prior of the RMW-AFT is of no simple form and the induced prior for θ is not guaranteed to be proper. The latter precludes meaningful model comparison between RMW models in terms of Bayes factors. Instead, a weakly informative prior, with a proper component for θ is presented. The latter preserves the structure of the Jeffreys prior (being flat on β) and the prior for θ is elicited via the coefficient of variation of the survival times. Priors for different mixing distributions are matched by a common prior on the coefficient of variation, so that models can be meaningfully compared through Bayes factors.

Subsections 3.2.3 and 3.3.3 provide conditions for posterior propriety based on an arbitrary mixing distribution. In particular, the problem associated with

the use of point observations is addressed. Whereas using point observations is not critical for RMW-AFT models (unless a zero time is recorded), Theorem 6 illustrates that point observations might invalidate posterior inferences for SMLN-AFT models. The use of set observations for the SMLN family is proposed as a solution. This can easily be implemented in an MCMC sampling scheme. Set observations can also be helpful in other contexts. For example, the issues of the Cox PH model with ties in the data are well known. Heritier et al. [2009] ignored ties when analyzing a real dataset, but that strategy might lead to serious loss of information if applied routinely. Other methods have been proposed for dealing with ties in the Cox regression model [see Kalbfleisch and Prentice, 2002, p. 104], but they might lead to biased estimations [Scheike and Sun, 2007]. In contrast, set observations are a natural solution that takes into account the imprecision with which the data was recorded and, as illustrated in Chapter 4, posterior inferences do not substantially change whether set or point observations are in use (of course, this comparison is only valid when the posterior distribution based on point observations is well-defined).

Outlier detection (with respect to the underlying model) is implemented as in Section 2.4. However, the general suggestion of reference value $\lambda_{ref} = \mathrm{E}(\Lambda_i|\theta)$ is not always applicable. In particular, Subsection 3.2.5 deals with situations where $\mathrm{E}(\Lambda_i|\theta)$ is not finite. In those cases, a reference value is defined in an empirical manner, considering the posterior distribution of the mixing parameters for simulated and real datasets. A different scenario is observed for RMW-AFT models, where $\mathrm{E}(\Lambda_i|\theta)$ is always finite, but $\lambda_{ref} = \mathrm{E}(\Lambda_i|\theta)$ is only applicable to non-censored observations. As discussed in Subsection 3.3.5, censoring is highly informative for the mixing parameters of these models. Hence, a new re-scaled reference value is proposed in such a case.

Chapter 4

Some applications

> "To mix or not to mix, that is the survival dilemma".
> Catalina

4.1 Introduction

In this Chapter, the two mixture families studied throughout Chapter 3 are applied to three real datasets. Bayesian inference is conducted under the priors introduced in Subsections 3.2.2 and 3.3.2. MCMC chains are generated using an adaptive Metropolis-within-Gibbs algorithm as described in Subsections 3.2.4 and 3.3.4. For these chains, the total number of iterations, thinning and burning periods are displayed in Tables D.1, D.13 and D.25 of Appendix D. These tables also show the update period for the mixing parameters Q (defined in Section 2.3). The use of different starting points (including random values) and the convergence diagnostics described in Subsection 1.2.3 strongly suggest convergence of the chains (see Appendix D). Posterior distributions are summarized in terms of their posterior medians and Highest Probability density intervals (HPD). Models are compared through the criteria described in Section 1.2.4.

4.2 Veteran's Administration Lung Cancer

This Veteran's Administration (VA) Lung Cancer dataset [presented in Kalbfleisch and Prentice, 2002] relates to a trial in which a therapy (standard or test chemotherapy) was randomly applied to 137 patients who were diagnosed with inoperable lung cancer. The survival times of the patients were measured in days since the application of the treatment and the following covariates were recorded: the treatment

that is applied to the patient (0: standard, 1: test); the histological type of the tumor (squamous, small cell, adeno, large cell); a continuous index representing the status of the patient at the moment of the treatment (the higher the index, the better the patient's condition[1]); the time between the diagnosis and the treatment (in months); age (in years); and a binary indicator of prior therapy (0: no, 1: yes). The data contain 9 right censored observations. During the trial, 69 patients received the standard treatment (only 5 of them recorded as censored observations). For these patients, the median time to follow-up (death or censoring) is equal to 97 days (first and third quartiles are 25 and 153 days, respectively). The remaining 68 patients were assigned a test treatment (4 of them with censored survival times). In this group, the median of their follow-up times is 52.5 days (first and third quartiles are 24.75 and 117.20 days, respectively). All patients are aged between 34 and 81 years old. Although the proportion of patients with small cell tumors is doubled (halved in case of adeno type tumors) for those under the standard treatment, both treatment groups presented a similar distribution of the patients with respect to the other recorded covariates.

This dataset has been previously analyzed from a frequentist point of view using traditional models such as the Cox, Weibull, log-normal and log-logistic regressions [see Lee and Wang, 2003; Barros et al., 2008; Heritier et al., 2009]. These models all suggest that the status of the patient at the moment of treatment and the histological type of the tumor are the most relevant explanatory variables for the survival time. Nevertheless, evidence of influential observations has been found. Barros et al. [2008] illustrated that the inference produced by a log-Birnbaum-Saunders model is greatly modified when dropping observations 77, 85 and 100. They proposed a log-Birnbaum-Saunders Student t distribution as a more robust alternative for this dataset because it allows for fatter tails and accommodates heterogeneity in the data (this distribution can also be represented through a mixture family as in Chapter 2, so the methodology presented here could be also extended to include this distribution). In an independent analysis, Heritier et al. [2009] detected observations 17 and 44 as influential when fitting a Cox proportional hazard model and proposed the use of an adaptive robust estimator as a solution.

Firstly, the data is analyzed using the log-normal and Weibull AFT models (with no mixture). Regression coefficients are defined as: β_0 (intercept), β_1 (treatment: test), β_2 (tumor type: squamous), β_3 (tumor type: small cell), β_4 (tumor type: adeno), β_5 (status), β_6 (time from diagnosis), β_7 (age) and β_8 (prior therapy:

[1]According to the value of this index, the patient can be classified in three different categories (10-30: completely hospitalized, 40-60: partial confinement, 70-90: able to care for self).

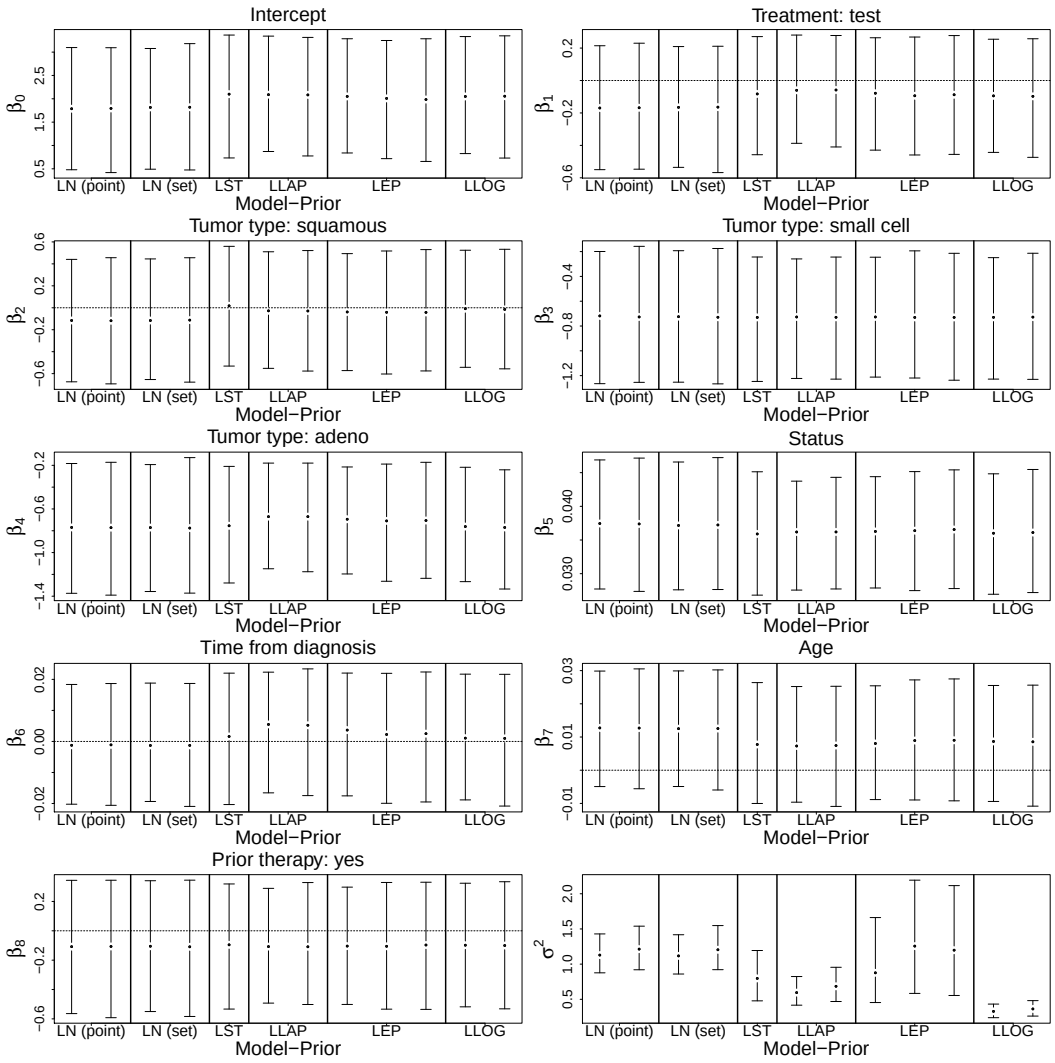


Figure 4.1: VA lung cancer dataset using SMLN-AFT models: vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, Jeffreys and ind. Jeffreys priors (plus ind. I Jeffreys prior for log-exp. power model). Only ind. Jeffreys prior is used for log-Student t. Horizontal lines at 0 were drawn for reference.

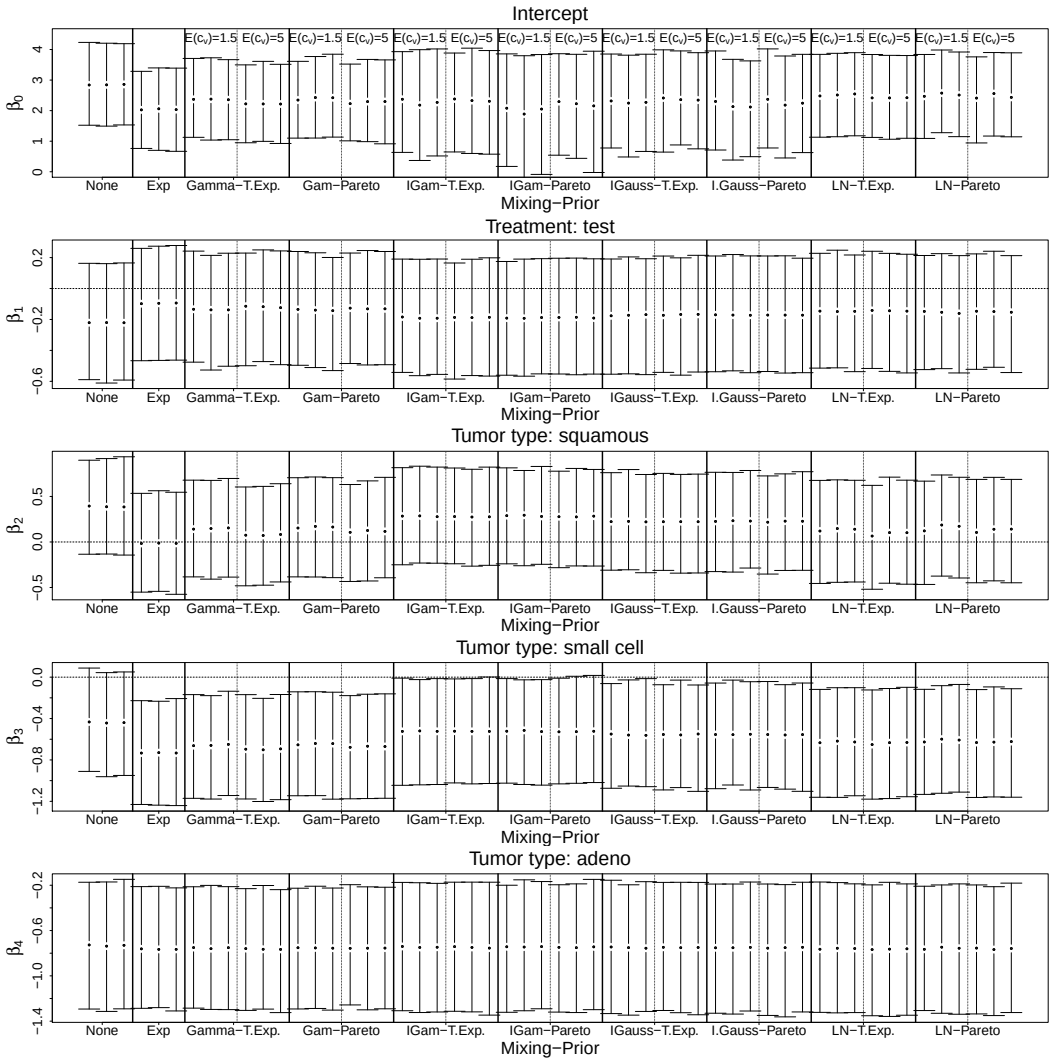


Figure 4.2: VA lung cancer dataset using RMW-AFT models with $\gamma \sim$ Gamma(d_1, d_2) and (if appropriate) a trunc. exponential or Pareto prior for c_v: vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, $d_1 = 4, d_2 = 1$, $d_1 = d_2 = 1$ and $d_1 = d_2 = 0.01$. Values of E(c_v) are displayed in the top panel. Horizontal lines at 0 were drawn for reference.

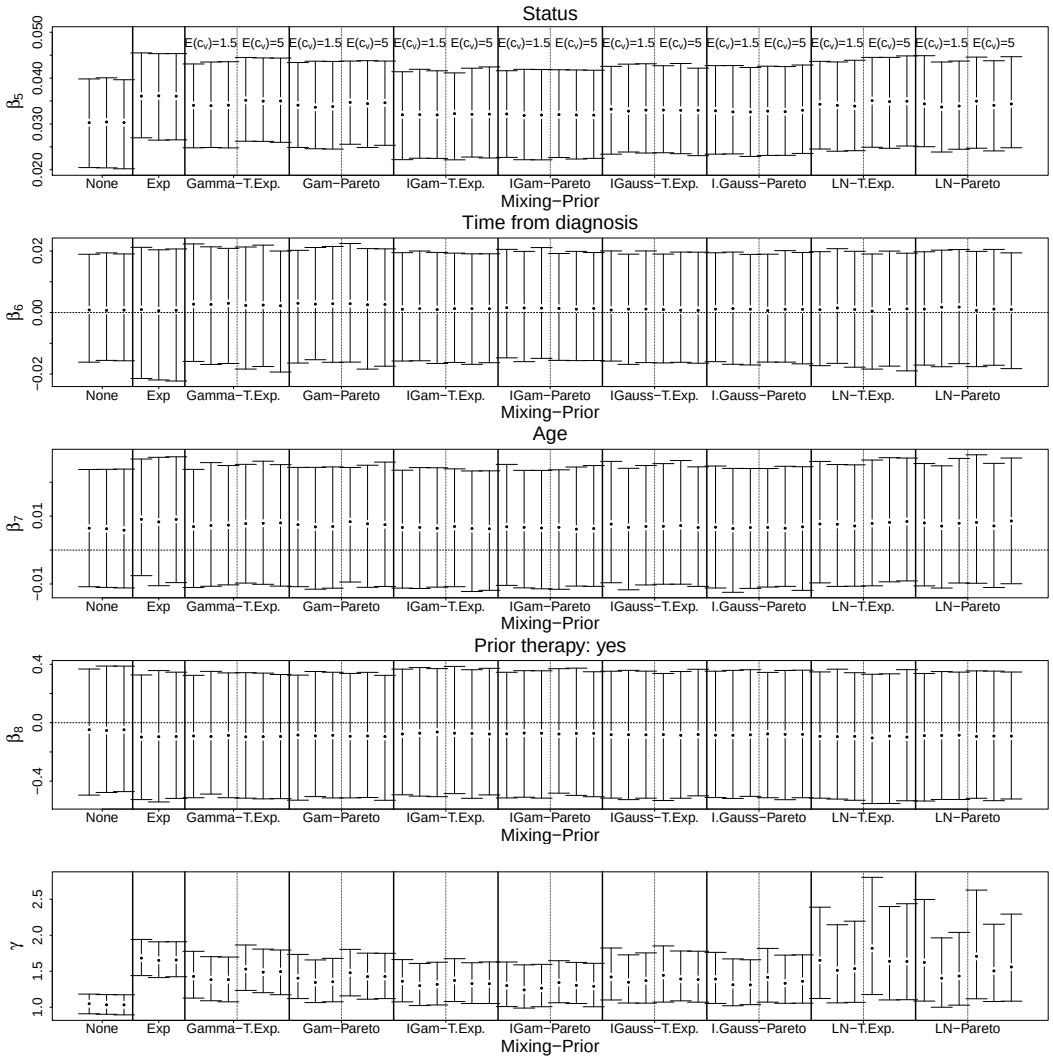


Figure 4.3: VA lung cancer dataset using RMW-AFT models with $\gamma \sim$ Gamma(d_1, d_2) and (if appropriate) a trunc. exponential or Pareto prior for c_v: vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, $d_1 = 4, d_2 = 1$, $d_1 = d_2 = 1$ and $d_1 = d_2 = 0.01$. Values of E(c_v) are displayed in the top panel. Horizontal lines at 0 were drawn for reference.

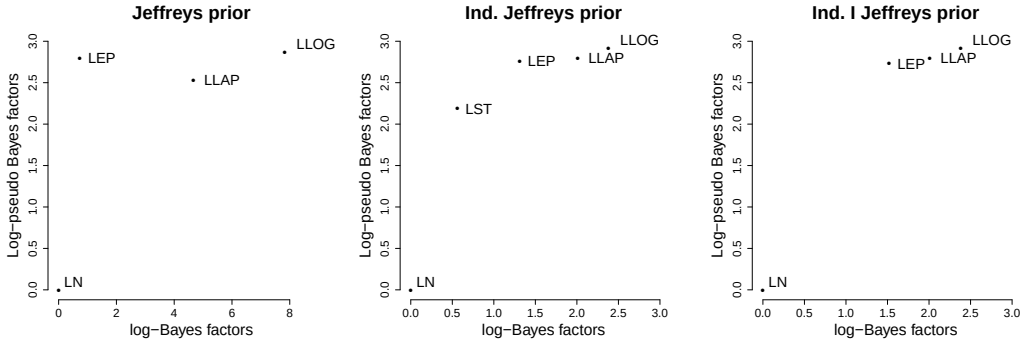


Figure 4.4: VA lung cancer dataset. log-BF and log-PsBF (w.r.t. log-normal AFT) of SMLN-AFT models.

yes). Posterior inferences are summarized in Figures 4.1, 4.2 and 4.3. For each model, all the priors considered here produced similar results, making the choice within these priors not too critical. The use of point observations does not produce problems for the log-normal model yet, for illustration purposes, Bayesian inference was conducted on the basis of point and set observations (using $\epsilon_l = \epsilon_r = 0.5$ for uncensored observations). In this case, inference on point and set observations is quite similar. However, set observations avoid potential problems with the inference for other SMLN models (see Theorem 6), so the rest of the analysis for SMLN-AFT models will focus on set observations. Set observations are not required for RMW-AFT models and point observations are used throughout. In line with Lee and Wang [2003], Barros et al. [2008] and Heritier et al. [2009], the log-normal and Weibull AFT models suggest that the main covariate effects are due to the tumour type and patient status. However, these models induce a different marginal effect of the covariates. For instance, with respect to a log-normal fit, the effect of the treatment is more accentuated when using a Weibull model. The most evident discrepancy relates to the squamous tumors coefficient β_2, where a Weibull fit points a positive effect (in contrast to the log-normal fit, where the HPD interval is almost centered around zero). Furthermore, these models provide conflicting interpretations for the hazard rate of the survival distribution. Whereas a log-normal fit indicates a non-monotonic behaviour, the Weibull model suggests a constant underlying hazard rate (as $\gamma \approx 1$).

In a second stage, the data is fitted using SMLN-AFT and RMW-AFT models with the continuous mixing distributions presented in Tables 3.1 and 3.2 (with the same definition for the regression coefficients as in the models with no mixture). For the SMLN and RMW families, the posterior distribution of β is somewhat differ-

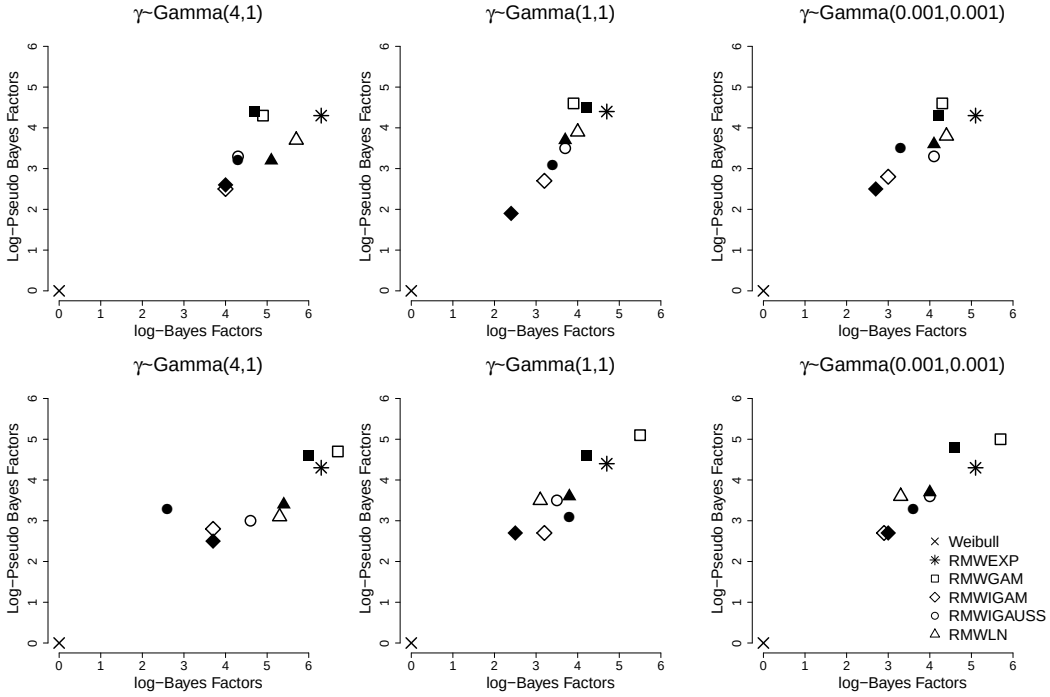


Figure 4.5: VA lung cancer dataset. log-BF and log-PsBF (w.r.t. Weibull AFT) of RMW-AFT models. Unfilled and filled characters denote a trunc. exponential and Pareto priors for c_v, respectively. Upper panels use E(c_v)=1.5. Lower panels use E(c_v)=5. Legend is displayed in the last panel.

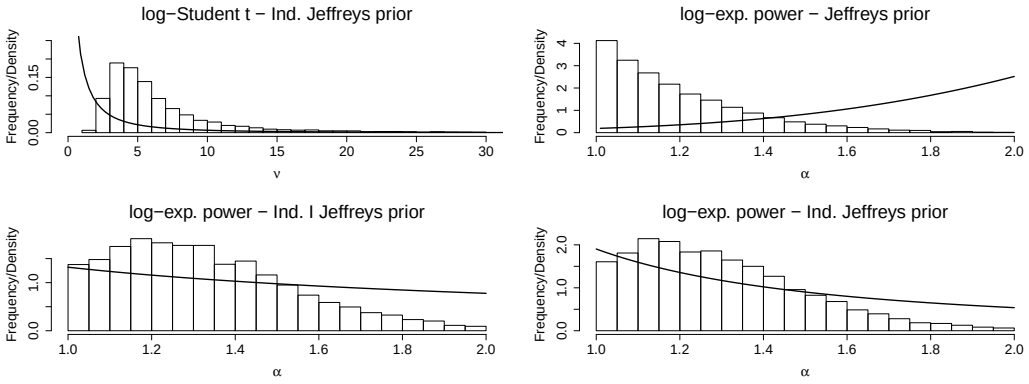


Figure 4.6: VA lung cancer dataset. Histogram for the posterior sample of ν and α (log-Student t and log-exp. power models, respectively). Solid curve represents the prior density.

Table 4.1: VA lung cancer dataset using SMLN-AFT models: DIC, the fraction of observations with better CPO performance than the AFT-log-normal model, and the number of influential observations.

Prior	Model	log PsML	DIC	CPO better	No. obs. $p_i \geq 0.9$
Jeffreys	Log-normal	-726.15	1449.01	-	2
	Log-Student t	-	-	-	-
	Log-Laplace	-723.62	1444.18	52%	1
	Log-exp. power	-723.35	1444.00	54%	1
	Log-logistic	-723.28	1444.14	66%	1
Ind. Jeffreys	Log-normal	-726.00	1449.56	-	2
	Log-Student t	-723.81	1445.86	64%	1
	Log-Laplace	-723.20	1444.37	53%	1
	Log-exp. power	-723.24	1444.79	55%	1
	Log-logistic	-723.08	1444.49	66%	1
Ind. I Jeffreys	Log-normal	-726.00	1449.56	-	2
	Log-Student t	-	-	-	-
	Log-Laplace	-723.20	1444.37	53%	1
	Log-exp. power	-723.27	1444.81	55%	1
	Log-logistic	-723.08	1444.49	66%	1

ent from that for the log-normal and Weibull models (see Figures 4.1, 4.2 and 4.3). In particular, all mixtures suggest that the effect of the treatment is less pronounced (especially when using SMLN models). Nonetheless, these mixture models still indicate that the most important predictors are the tumour type and the patient status. The choice within these priors and mixing distributions was not too critical for the inference about β. For these examples, selecting between a log-normal or Weibull underlying model is not too critical either. Only minor discrepancies between the effect of the treatment and the tumor type are observed. In all cases, the results on β are relatively close to the classical ones reported in Barros et al. [2008] using the log-Birnbaum Saunders Student t model and to the ones in Lee and Wang [2003] using the log-logistic model.

Model comparison criteria are summarized in Figures 4.4 and 4.5 and Tables 4.1 and 4.2. In particular, the Weibull model is pointed out as the worst candidate (which is, in contrast to all other models, the only one inducing a constant underlying hazard rate). It has the highest DIC and the lowest PsML within all models (BF with respect to log-normal and SMLN models cannot be computed because an improper prior for uncommon parameters is in use). Overall, all criteria provide evidence in favour of mixture models. For the log-Student t model, this evidence is also supported by the fact that inference on ν favours relative small values. Similarly,

Table 4.2: VA lung cancer dataset using RMW-AFT models under a Gamma(d_1, d_2) prior for γ: DIC, the fraction of observations with better CPO performance than the AFT-Weibull model, and the number of influential observations.

			Trunc. exp. prior for c_v				Pareto prior for c_v			
E(c_v)	d_1, d_2	Mixing	log PsML	DIC	CPO better	No. $p_i \geq 0.9$	log PsML	DIC	CPO better	No. $p_i \geq 0.9$
1.5	4,1	None	-727.4	1451.31	-	3	-727.4	1451.31	-	3
		Exp.	-723.1	1444.24	64%	1	-723.1	1444.24	64%	1
		Gam.	-723.1	1443.81	69%	1	-723.0	1443.94	68%	1
		Inv-Gam.	-724.9	1446.89	69%	2	-724.8	1446.93	73%	1
		Inv-Gauss.	-724.0	1446.41	66%	2	-724.2	1446.36	66%	2
		Log-norm.	-723.6	1445.47	66%	1	-724.2	1445.86	66%	1
	1,1	None	-727.5	1451.52	-	4	-727.5	1451.52	-	4
		Exp.	-723.0	1444.54	65%	1	-723.0	1444.54	65%	1
		Gam.	-722.9	1443.95	69%	1	-723.0	1444.41	69%	1
		Inv-Gam.	-724.8	1447.56	70%	2	-725.5	1447.95	70%	2
		Inv-Gauss.	-724.0	1446.72	66%	2	-724.4	1446.66	67%	2
		Log-norm.	-723.6	1445.85	66%	1	-723.8	1446.07	67%	2
	0.01,0.01	None	-727.5	1451.58	-	3	-727.5	1451.58	-	3
		Exp.	-723.2	1444.65	64%	1	-723.2	1444.65	64%	1
		Gam.	-722.9	1444.00	69%	1	-723.2	1444.44	69%	1
		Inv-Gam.	-724.7	1447.23	73%	1	-725.0	1448.01	73%	2
		Inv-Gauss.	-724.2	1446.70	69%	2	-724.0	1446.56	69%	2
		Log-norm.	-723.7	1445.86	69%	1	-723.9	1446.23	69%	2
5	4,1	None	-727.4	1451.31	-	3	-727.4	1451.31	-	3
		Exp.	-723.1	1444.24	64%	1	-723.1	1444.24	64%	1
		Gam.	-722.6	1443.25	67%	1	-722.7	1443.21	67%	1
		Inv-Gam.	-724.6	1446.78	71%	1	-724.9	1447.14	71%	2
		Inv-Gauss.	-724.3	1446.52	64%	2	-724.1	1445.85	66%	2
		Log-norm.	-724.2	1446.16	64%	2	-723.9	1445.71	66%	1
	1,1	None	-727.5	1451.52	-	4	-727.5	1451.52	-	4
		Exp.	-723.0	1444.54	65%	1	-723.0	1444.54	65%	1
		Gam.	-722.4	1443.29	66%	1	-722.8	1443.87	69%	1
		Inv-Gam.	-724.8	1447.43	69%	2	-724.8	1446.99	70%	2
		Inv-Gauss.	-724.0	1446.56	66%	2	-724.4	1446.31	69%	2
		Log-norm.	-724.0	1446.34	66%	1	-723.9	1446.09	69%	1
	0.01,0.01	None	-727.5	1451.58	-	3	-727.5	1451.58	-	3
		Exp.	-723.2	1444.65	64%	1	-723.2	1444.65	64%	1
		Gam.	-722.5	1443.33	67%	1	-722.7	1443.54	68%	1
		Inv-Gam.	-724.8	1447.04	72%	2	-724.8	1447.34	72%	2
		Inv-Gauss.	-723.9	1446.57	68%	2	-724.2	1446.55	69%	2
		Log-norm.	-723.9	1446.90	68%	1	-723.8	1446.10	69%	1

the log-exponential power model suggests values of α far from 2. Figure 4.6 contrasts the prior and the posterior distributions of ν and α under the log-Student t and log-exponential power models. Clearly, they differ and this is strongly driven by the data itself. The contrast between the Jeffreys prior for α and its posterior is somewhat

Table 4.3: VA lung cancer dataset using RMW-AFT models under a Gamma(d_1, d_2) prior for γ: posterior medians and HPD 95% intervals of $R_{c_v}(\gamma, \theta)$ (as in equation (3.27)).

			$d_1 = 4, d_2 = 1$		$d_1 = d_2 = 1$		$d_1 = d_2 = 0.01$	
Prior c_v	E(c_v)	Mixing	Med.	HPD 95%	Med.	HPD 95%	Med.	HPD95%
T. Exp.	1.5	Gam(θ, θ)	2.08	[1.03,3.95]	1.97	[1.07,3.81]	1.95	[1.04,3.72]
		Inv-Gam$(\theta, 1)$	1.36	[1.13,1.55]	1.31	[1.13,1.55]	1.33	[1.13,1.55]
		Inv-Gauss$(\theta, 1)$	1.47	[1.16,1.79]	1.41	[1.09,1.71]	1.43	[1.14,1.74]
		log-norm$(0, \theta)$	1.89	[1.14,3.01]	1.72	[1.08,2.62]	1.76	[1.07,2.73]
	5	Gam(θ, θ)	6.07	[1.25,20.25]	5.68	[1.21,19.42]	5.69	[1.20,19.72]
		Inv-Gam$(\theta, 1)$	1.38	[1.13, 1.56]	1.35	[1.13, 1.56]	1.35	[1.13, 1.56]
		Inv-Gauss$(\theta, 1)$	1.50	[1.11, 1.81]	1.47	[1.17, 1.77]	1.46	[1.13, 1.76]
		log-norm$(0, \theta)$	2.21	[1.20,3.71]	1.96	[1.11,3.22]	2.13	[1.19,3.76]
Pareto	1.5	Gam(θ, θ)	1.91	[1.04,4.34]	1.75	[1.05,3.79]	1.78	[1.05,4.00]
		Inv-Gam$(\theta, 1)$	1.30	[1.08,1.54]	1.25	[1.03,1.51]	1.28	[1.04,1.52]
		Inv-Gauss$(\theta, 1)$	1.44	[1.12,1.74]	1.37	[1.09,1.68]	1.36	[1.09,1.67]
		log-norm$(0, \theta)$	1.81	[1.11,2.88]	1.56	[1.03,2.48]	1.58	[1.00,2.39]
	5	Gam(θ, θ)	2.97	[1.08,20.32]	2.66	[1.03,17.77]	2.69	[1.10,15.97]
		Inv-Gam$(\theta, 1)$	1.34	[1.12, 1.55]	1.32	[1.11, 1.55]	1.31	[1.04, 1.53]
		Inv-Gauss$(\theta, 1)$	1.47	[1.15, 1.81]	1.39	[1.09, 1.70]	1.42	[1.13, 1.73]
		log-norm$(0, \theta)$	1.95	[1.18,3.11]	1.71	[1.10,2.66]	1.78	[1.08,2.85]

surprising, in view of the results for the other priors. However, this is explained by the fact that σ^2 and α are highly (positively) correlated a posteriori. For the VA lung cancer dataset ($k = 9$), the Jeffreys prior assigns high probabilities to low values of σ^2 (much higher in comparison with the other two priors) and therefore the Jeffreys prior is implicitly driving the posterior of α towards 1. Indeed, using a modification of the Jeffreys prior where $p = 1$, the posterior of α is shifted to the right, with a mode around 1.5 (not shown). In case of the RMW models, R_{c_v} is substantially larger than 1 (see Table 4.3). In accordance with the model comparison criteria, this also suggest a better fit of RMW models with respect to the Weibull one.

Overall, the log-logistic model seems the best SMLN candidate for fitting this dataset (within these examples). This is in line with the results in Lee and Wang [2003] in which, using a maximum likelihood approach, the log-logistic model is preferred to the log-normal and other standard models. Within the RMW models, the exponential(1) and Gamma(θ, θ) mixing distributions have the best performance. While the exponential mixing is preferred when E(c_v) is small, larger values of E(c_v) drive the evidence towards the Gamma mixing. Hence, the survival distribution appears to have a large but finite c_v (otherwise the exponential mixing would also be chosen when E(c_v) is large). As shown in Table 4.3, the Gamma mixing produces the highest inflation of c_v with respect to a Weibull model (for the same

γ). According to PsML and DIC this model is also preferred over all other SMLN and RMW models under most of the considered priors. Both, the log-logistic and RMW model with Gamma mixing suggest a mild unobserved heterogeneity, where the coefficient of variation associated of the marginal distribution is finite.

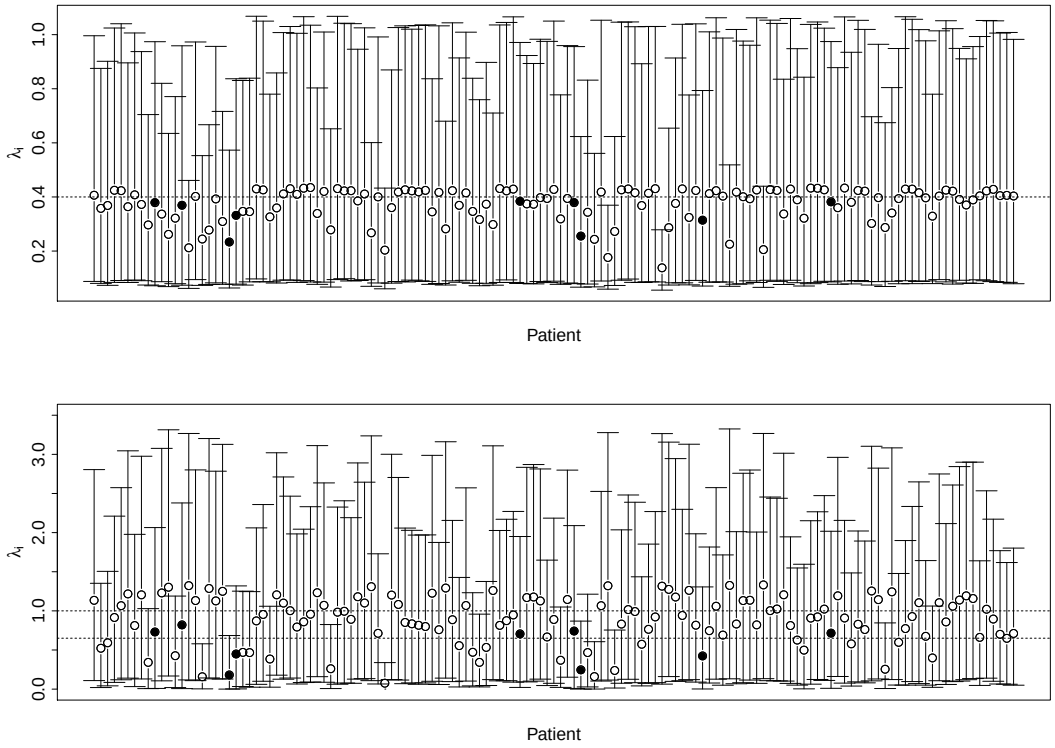


Figure 4.7: VA lung cancer dataset. For λ_i, $i = 1, \ldots, n$: vertical lines are the HPD 95% intervals and circles represent posterior medians (filled for censored observations). Horizontal lines are located at λ_{ref} (and λ^c_{ref}, if appropriate). Upper panel: log-logistic model (ind. Jeffreys prior). Lower panel: RMW model with Gamma(θ,θ) mixing ($\gamma \sim$Gamma(4,1) and trunc. exponential prior for c_v with E(c_v=1.5)).

For all the priors used here, Tables 4.1 and 4.2 show that the number of influential observations is smaller for the mixture models. This is consistent with the superior ability of the SMLN and RMW models to accommodate unusual observations. In particular, observations 85 and 106 are detected as influential for the log-normal model (with no mixture). In case of a Weibull fit, observations 44, 75 and 106 are influential (observation 17 is added to this list in case of a Gamma(1,1) prior for γ). In contrast, only observation 106 is considered as influential for SMLN and RMW models (although, for some priors, observation 44 is also influential when using a RMW model with inverse Gamma or inverse Gaussian mixing). Patient 106

has a (censored) survival time of 51 days and received the test treatment for a small cell type of tumor (for which previous treatments were applied). This patient was 59 years old and completely hospitalized (status score equal to 30) at the moment of the treatment (about 7 years after diagnosis). Moreover, this subject has the second longest delay between diagnosis and the test treatment, with a particularly large survival time within those with more than two and a half years of delay. In Barros et al. [2008], observation 106 was also detected as a (mild) influential observation when fitting a log-Birnbaum-Saunders-t model.

The posterior distributions of the mixing parameters (see Figure 4.7) vary substantially between the patients, suggesting heterogeneity in the data. For all considered priors, this behaviour supports the choices of λ_{ref} (and λ^c_{ref}, if appropriate) indicated in Subsections 3.2.5 and 3.3.5. Figure 4.8 formalizes this by presenting the Bayes factor in favour of being an outlier for each of the 137 observations. There is clear evidence for the existence of outlying observations under the suggested priors for all models. For SMLN models, although all these priors present similar results, this evidence is slightly stronger for the Jeffreys prior. The choice within these mixing distributions does not greatly affect the conclusions. The log-logistic model suggests that, for both priors considered here, observations 77 and 85 are very clear outliers (with respect to the underlying log-normal model). Patients 77 and 85 had an uncensored survival time of 1 day (the lowest value observed in the dataset), were under the standard treatment and had a squamous type of tumor. Observations 15, 17, 21, 44, 75, 95 and 100 are added to this list when using the other suggested SMLN models (not reported). The model detecting the largest amount of outliers is the log-Laplace (which induces the strongest unobserved heterogeneity). Weibull mixtures detect different outliers. This is not surprising as the underlying model is different. For all the priors in use, the RMW model with Gamma(θ, θ) mixing suggest that the records 17, 36, 44, 75, 78 and 118 are outliers. In particular, patients 17 and 44 (who survived 384 and 392 days, respectively) are the largest survival times for patients that had the same type of tumor (small cell). In addition, observation 75 has the second largest survival time in the sample (observation 70 has the largest survival time, but it is explained by a very good patient's status at treatment time). Whereas the choice between a truncated exponential Pareto prior for c_v does not substantially affect the results, the prior expectation of c_v does. As shown in Figure 4.8, the number of detected outliers is larger when E(c_v)=5 (in comparison to E(c_v)=1.5). In fact, a less tight prior for c_v induces a stronger unobserved heterogeneity, allowing the λ_i's to explore more extreme values. This is particularly important in case of a Gamma(θ, θ) mixing, where the reference value is fixed at 1

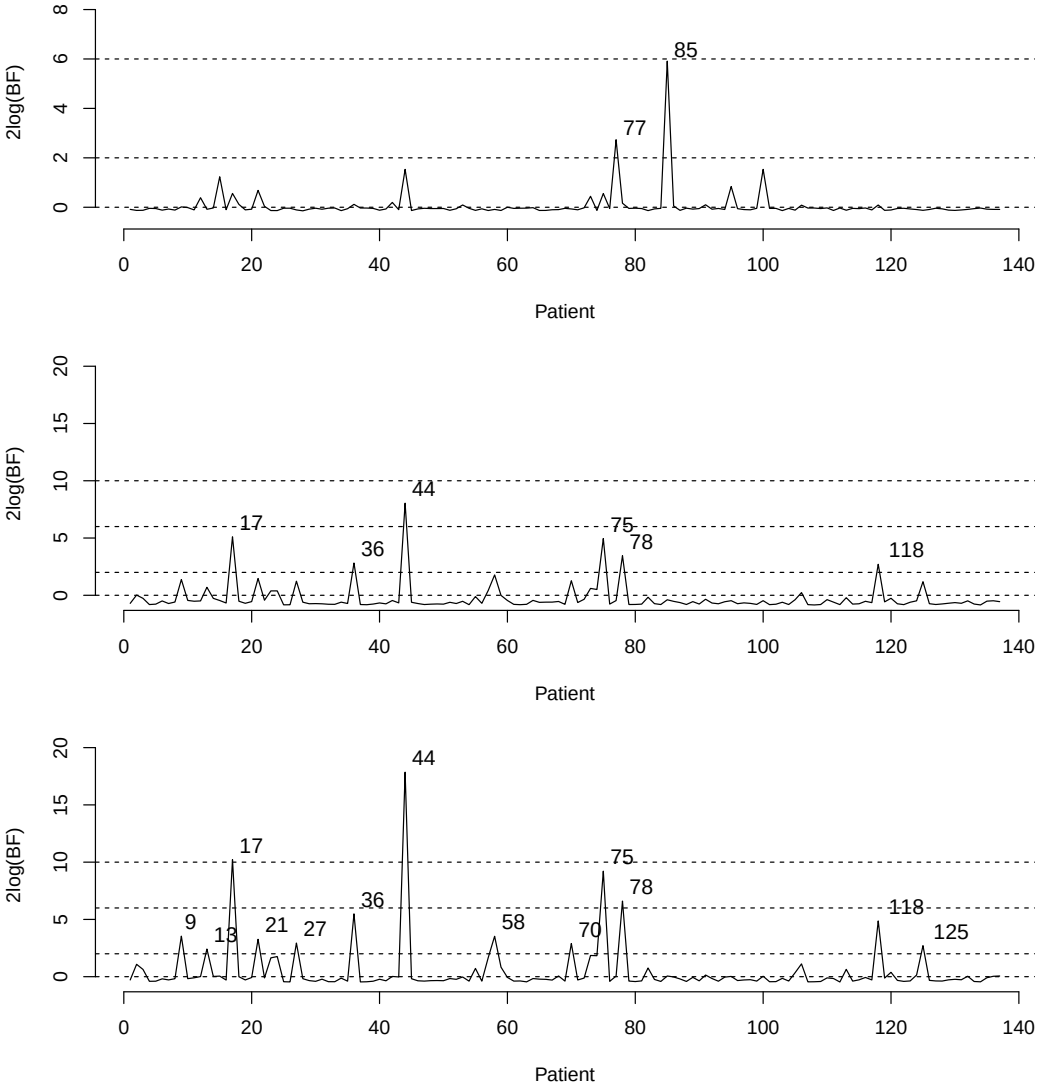


Figure 4.8: VA lung cancer dataset. $2 \times \log(\text{BF})$ in favour of $H_1 : \lambda_i \neq \lambda_{ref} (u_i \neq u_{ref})$ versus $H_0 : \lambda_i = \lambda_{ref}$. Horizontal lines reflect the interpretation rule of Kass and Raftery [1995]. First panel: log-logistic model (independence Jeffreys prior). Second and third panels: RMW model with Gamma(θ,θ) mixing ($\gamma \sim$Gamma(4,1) and trunc. exponential prior for c_v with E(c_v=1.5) and E(c_v=5), respectively).

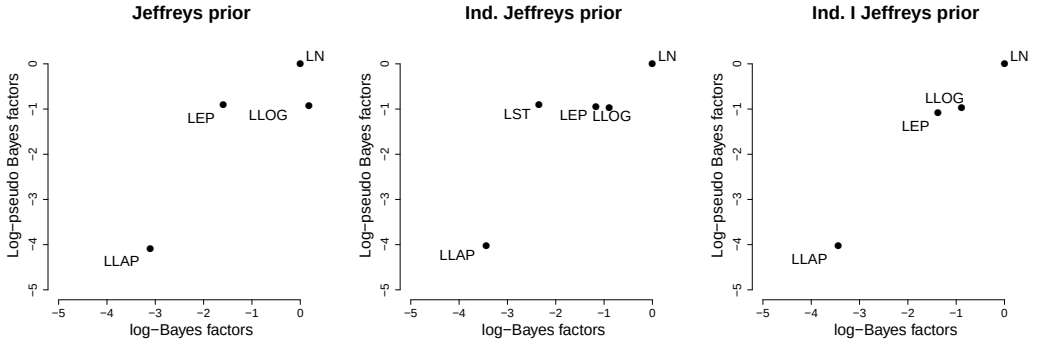


Figure 4.9: AA Bone Marrow Transplant dataset. log-BF and log-PsBF of SMLN-AFT models with respect to the AFT-log-normal one.

(regardless of the value of θ). Similar outliers are detected by other RMW models. For instance, the exponential(1) mixing suggest that observations 9, 17, 21, 36, 44, 75, 78 and 118 are outliers.

4.3 Autologous and Allogeneic Bone Marrow Transplant

The Autologous and Allogeneic (AA) Bone Marrow Transplant dataset [presented in Klein and Moeschberger, 1997] contains post-surgery information about 101 patients with advanced acute myelogenous leukemia. The endpoint of the study is the disease-free survival time of the patients, *i.e.* the time until relapse or death (measured in months). The disease-free survival time was observed for 50 patients while the others are right-censored. In the trial, 51 patients received an autologous bone marrow transplant. This replaces the patient's marrow with their own marrow after the application of high doses of chemotherapy. The median of the time to follow-up (relapse, death or censoring) is equal to 13.06 months (first and third quartiles are 6.07 and 18.42 months, respectively). The rest of the patients received an allogeneic bone transplant, in which their marrow was replaced by the one extracted from a sibling (matched according to a Histocompatibility Leukocyte Antigen criteria). The median time to follow-up is 11.81 months (first and third quartiles are 3.61 and 31.88 months, respectively) for this group. Although similar studies suggested a significant effect of the Karnofsky score (a continuous index representing the status of the patient at the moment of the treatment) and the time between diagnosis and transplant, there is no record of these covariates in the dataset. Only the type of treatment is documented and therefore an important amount of unobserved heterogeneity is expected.

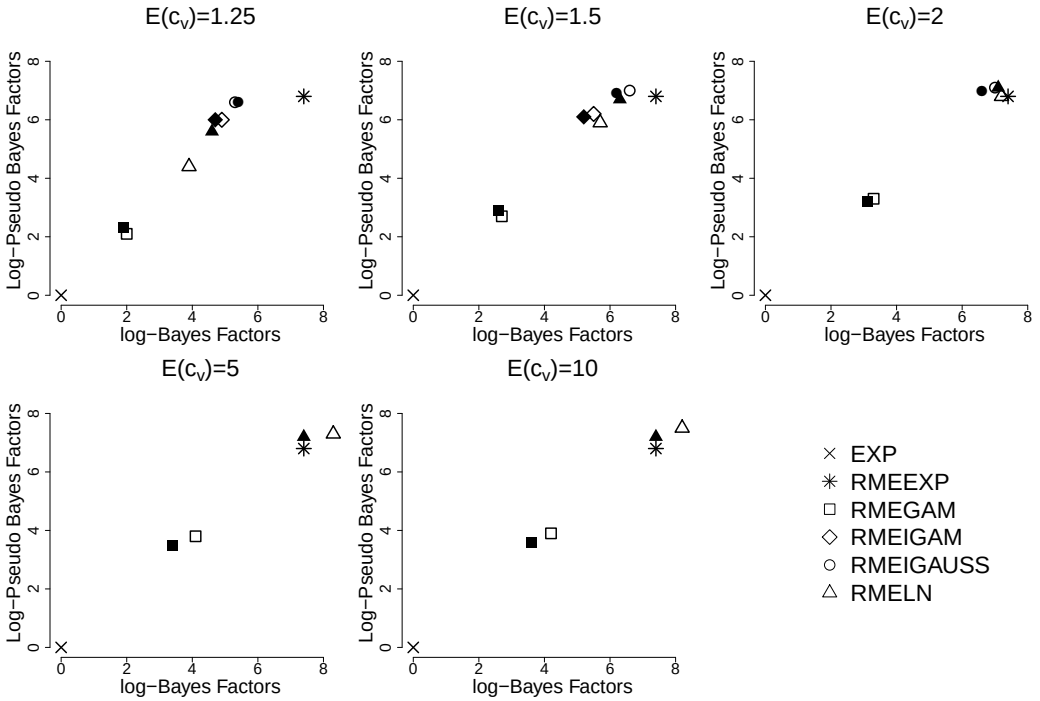


Figure 4.10: AA Bone Marrow Transplant dataset. log-BF and log-PsBF of RME-AFT models with respect to the AFT-exponential one. Unfilled and filled characters denote a trunc. exponential and Pareto priors for c_v, respectively. Legend is displayed in the last panel.

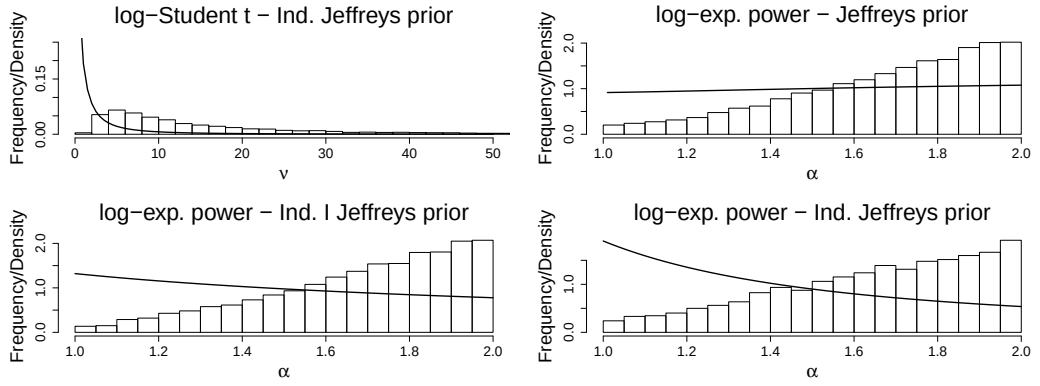


Figure 4.11: AA Bone Marrow Transplant dataset. Histogram for the posterior sample of ν and α (log-Student t and log-exp. power models, respectively). Solid curve represents the prior density.

Table 4.4: AA Bone Marrow Transplant dataset using SMLN-AFT models: DIC, the fraction of observations with better CPO performance than the AFT-log-normal model, and the number of influential observations.

Prior	Model	log PsML	DIC	CPO better	No. obs. $p_i \geq 0.9$
Jeffreys	Log-normal	-223.06	446.09	-	0
	Log-Student t	-	-	-	-
	Log-Laplace	-227.14	453.42	0.33	0
	Log-exp. power	-223.96	448.07	0.37	0
	Log-logistic	-223.99	447.92	0.41	0
Ind. Jeffreys	Log-normal	-223.06	446.09	-	0
	Log-Student t	-223.97	448.41	0.37	0
	Log-Laplace	-227.09	453.41	0.34	0
	Log-exp. power	-224.01	448.32	0.38	0
	Log-logistic	-224.02	448.06	0.41	0
Ind. I Jeffreys	Log-normal	-223.06	446.09	-	0
	Log-Student t	-223.97	448.41	0.37	0
	Log-Laplace	-227.09	453.41	0.34	0
	Log-exp. power	-224.15	448.51	0.38	0
	Log-logistic	-224.02	448.06	0.41	0

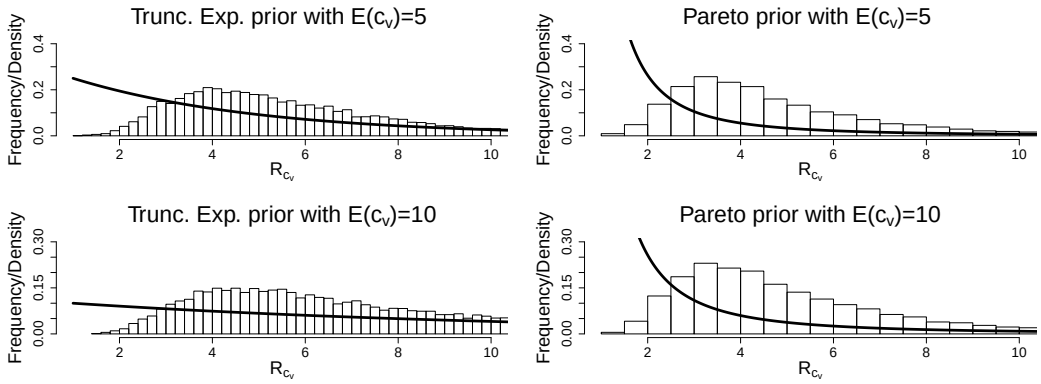


Figure 4.12: AA Bone Marrow Transplant dataset. Histogram for the posterior sample of $R_{c_v}(1,\theta)$ (which equals $c_v(1,\theta)$) using a log-normal mixing distribution. Solid curve represents the prior density.

The data is first analyzed using log-normal, exponential and Weibull AFT models (the last two models have equivalent PH representations). The regression coefficients are β_0 (intercept) and β_1 (treatment: autologous). Within these models, the log-normal is pointed out as the best fit (in terms of DIC and PsBF). This is not entirely unpredictable as the standard graphical check of $\log(-\log(S(t)))$ versus

Table 4.5: AA Bone Marrow Transplant dataset using RME-AFT models: DIC, the fraction of observations with better CPO performance than the AFT-exponential model, and the number of influential observations.

		Trunc. exp. prior for c_v				Pareto prior for c_v			
E(c_v)	Mixing	log PsML	DIC	CPO better	No. $p_i \geq 0.9$	log PsML	DIC	CPO better	No. $p_i \geq 0.9$
1.25	None	-230.4	460.0	-	0	-230.4	460.0	-	0
	Exp.	-223.6	446.6	47%	0	-223.6	446.6	47%	0
	Gam.	-228.3	455.8	49%	0	-228.1	455.5	49%	0
	Inv-Gam.	-224.5	448.8	48%	0	-224.5	448.8	48%	0
	Inv-Gauss.	-223.8	447.7	48%	0	-223.8	447.8	47%	0
	Log-norm.	-225.8	450.0	51%	0	-224.9	448.5	50%	0
1.5	None	-230.4	460.0	-	0	-230.4	460.0	-	0
	Exp.	-223.6	446.6	47%	0	-223.6	446.6	47%	0
	Gam.	-227.7	454.7	49%	0	-227.6	454.5	49%	0
	Inv-Gam.	-224.3	448.4	47%	0	-224.5	448.8	48%	0
	Inv-Gauss.	-223.4	446.9	47%	0	-223.5	447.2	46%	0
	Log-norm.	-224.1	447.3	50%	0	-223.8	447.0	48%	0
2.0	None	-230.4	460.0	-	0	-230.4	460.0	-	0
	Exp.	-223.6	446.6	47%	0	-223.6	446.6	47%	0
	Gam.	-227.2	453.7	49%	0	-227.3	454.0	49%	0
	Inv-Gam.	-	-	-	0	-	-	-	0
	Inv-Gauss.	-223.4	446.9	45%	0	-223.4	447.0	47%	0
	Log-norm.	-223.4	446.1	49%	0	-223.4	446.3	48%	0
5.0	None	-230.4	460.0	-	0	-230.4	460.0	-	0
	Exp.	-223.6	446.6	47%	0	-223.6	446.6	47%	0
	Gam.	-226.7	452.7	48%	0	-226.9	453.3	48%	0
	Inv-Gam.	-	-	-	0	-	-	-	0
	Inv-Gauss.	-	-	-	0	-	-	-	0
	Log-norm.	-223.0	445.7	48%	0	-223.3	446.2	48%	0
10.0	None	-230.4	460.0	-	0	-230.4	460.0	-	0
	Exp.	-223.6	446.6	47%	0	-223.6	446.6	47%	0
	Gam.	-226.5	452.3	48%	0	-226.9	453.2	49%	0
	Inv-Gam.	-	-	-	0	-	-	-	0
	Inv-Gauss.	-	-	-	0	-	-	-	0
	Log-norm.	-223.1	446.0	50%	0	-223.2	446.2	46%	0

t (not reported) suggests that the proportional hazards assumption does not hold. The BF in favour of the Weibull model with free γ (with respect to the exponential one) is 4.39, suggesting $\gamma \neq 1$. In line with this, the posterior median of γ is 0.69 (HPD 95%: (0.53,0.85)) for the Weibull model with $\gamma \sim$ Gamma(4,1). In a second stage, the SMLN, RME and RMW AFT models explored throughout this document are fitted to these data. In contrast to the Weibull case, the RMW-AFT regressions shown strong evidence in favour of $\gamma = 1$. For example, for the exponential(1) mixing and $\gamma \sim$ Gamma(4,1), the BF in favour of the RME specification ($\gamma = 1$) is 22.01. In

this case, the posterior median of γ is 0.86 (HPD 95%: (0.67,1.07)). These opposite conclusions are not surprising because the Weibull model tends to underestimate γ in the presence of unobserved heterogeneity (as illustrated in Section 4.2). Based on this evidence, RMW-AFT models with free γ are discarded for these data.

Overall, SMLN models do not have a good performance with respect to the log-normal model with no mixture (see Figure 4.9 and Table 4.4). This is also reflected in the posterior distribution of ν and α for the log-Student t and log-exponential power models, respectively (see Figure 4.11). This poor performance is not entirely unforeseen because the underlying hazard rate appears to be constant in time (as evidenced in the previous paragraph). For RME models, $\mathrm{E}(c_v)$ equal to 1.25, 1.5, 2, 5 and 10 is used (if there is no θ in the model, all these priors coincide). Large values of $\mathrm{E}(c_v)$ are associated with stronger prior beliefs about the existence of unobserved heterogeneity. Nevertheless, as explained in Section 3.3.2, if $\mathrm{E}(c_v)$ is larger than $\sqrt{3}$, the model generated by the inverse Gamma mixing distribution is not compatible with the prior beliefs. The same occurs for the inverse Gaussian mixing when $\mathrm{E}(c_v) > \sqrt{5}$. For these data, the presence of unobserved heterogeneity is strongly supported by the data as all RME models perform better than the exponential one with no mixture (see Figure 4.10 and Table 4.5). Despite its simplicity, the model generated by the exponential mixing distribution is chosen because it receives most support overall. The log-normal mixing distribution has slightly more support for large $\mathrm{E}(c_v)$, but the exponential mixing distribution does not require prior elicitation for θ and is easy to implement (as the full conditionals of the λ_i's have a close known form). Despite the small sample size, there is learning about R_{c_v} (which in this case equals c_v). As seen in Figure 4.12, even though the truncated exponential and Pareto priors are concentrated around small values of R_{c_v}, the posterior distribution is shifted to the right. This suggests the need for a mixture and is consistent with strong heterogeneity in the data that leads to support for the exponential mixing model (which does not allow for a finite c_v).

Figures 4.13 and 4.14 summarize marginal posterior inference for SMLN and RME AFT models under different prior assumptions. All models suggest that there is no substantive difference between the median survival times under both treatments. For the RME models, whereas the choice of a prior affects inference on R_{c_v}, the posterior distribution of β (which is usually the parameter of interest) is more robust. With RME models, the effect of the treatment (β_1) is less pronounced than the value estimated by the exponential model, for all considered mixing distributions and priors. This discrepancy is among the largest when using the exponential mixing.

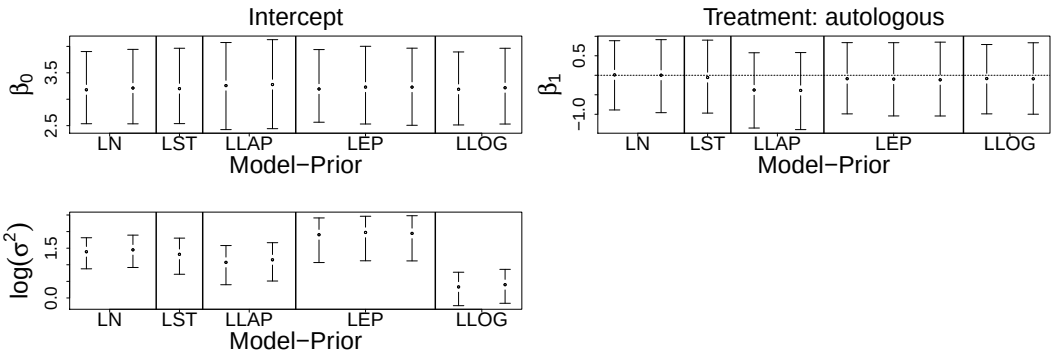


Figure 4.13: AA Bone Marrow Transplant dataset using SMLN-AFT models: vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, Jeffreys and ind. Jeffreys priors (plus ind. I Jeffreys prior for log-exp. power model). Only ind. Jeffreys prior is used for log-Student t. Horizontal lines at 0 were drawn for reference.

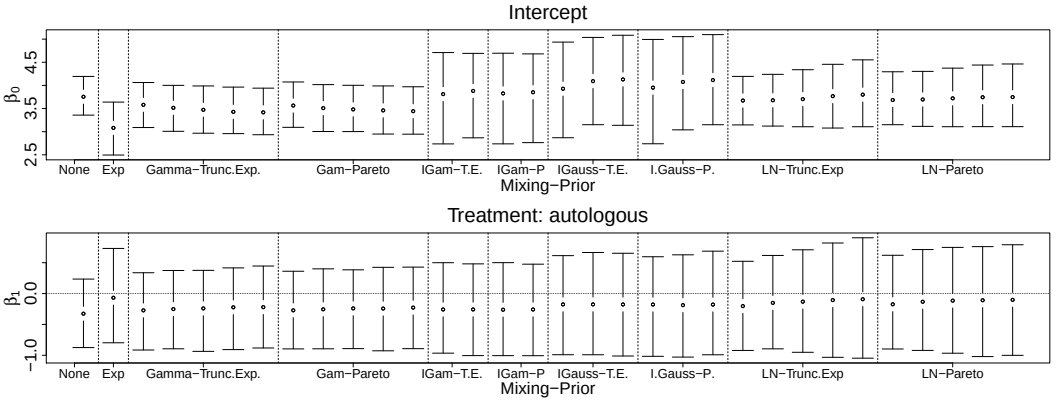


Figure 4.14: AA Bone Marrow Transplant dataset using RME-AFT models with (if appropriate) a trunc. exponential or Pareto prior for c_v: vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, E(c_v)=1.25,1.5,2,5,10. Horizontal lines at 0 were drawn for reference.

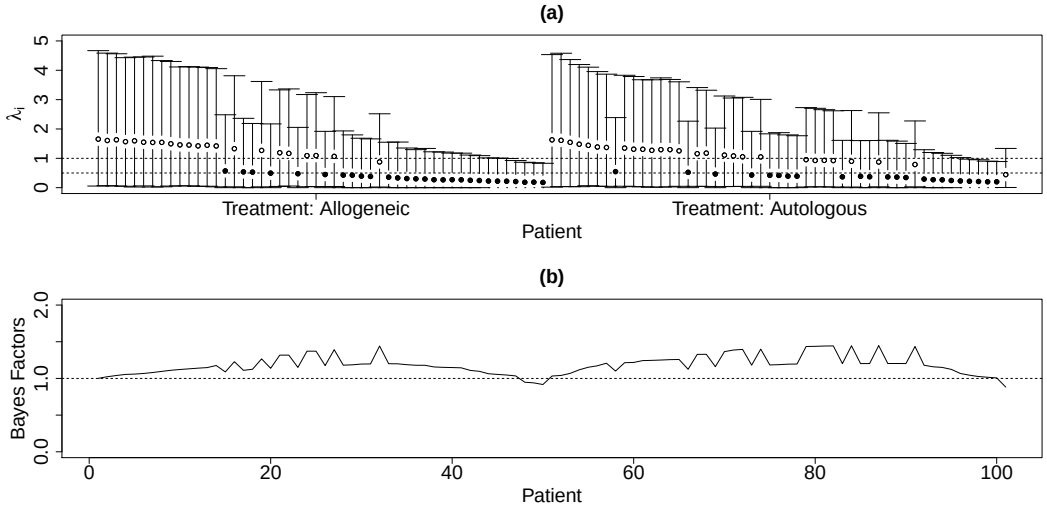


Figure 4.15: AA Bone Marrow Transplant dataset using an RME model with exponential(1) mixing. (a) 95% HPD interval of the λ_i's for the exponential mixing distribution. Horizontal lines at $\lambda^o_{ref} = 1$ and $\lambda^c_{ref} = 1/2$. Circles located at posterior medians (filled for censored observations). Observations are grouped by treatment and displayed in ascending order of the t_i's. (b) Bayes Factors in favour of the model $M_1 : \Lambda_i \neq \lambda_{ref}$ versus $M_0 : \Lambda_i = \lambda_{ref}$.

No influential observations are detected for any model considered (all the p_i's are below 0.9). This includes the exponential and log-normal models without mixture (see Tables 4.4 and 4.5). Panel (a) in Figure 4.15 illustrates the posterior behaviour of the mixing parameters for the RME model with exponential mixing. Because the mixture was introduced via a scale (rate) parameter, there is a strong posterior association between the λ_i's and the survival times. In this case, no outlying observations are detected when using the outlier detection mechanism proposed in Subsection 3.3.5 (see panel (b) in Figure 4.15). So this is a situation where no single observation is identified as an outlier, yet there is ample evidence in favour of the exponential mixture model on the basis of the entire sample.

4.4 Cerebral Palsy

This dataset is a subset of the data in Hutton et al. [1994] and Kwong and Hutton [2003] and contains information about 1,549 children affected by cerebral palsy and born during the period 1966-1984 in the administrative area of the Mersey Region Health Authority. See Hutton et al. [1994] for more information about the data

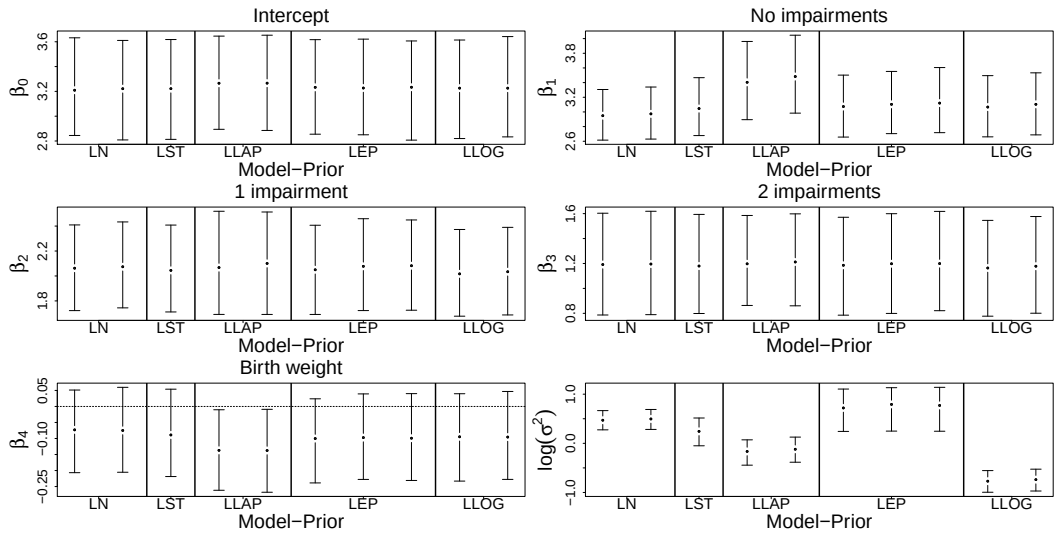


Figure 4.16: Cerebral palsy dataset. For SMLN-AFT models (set observations). Vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, Jeffreys and ind. Jeffreys priors (plus ind. I Jeffreys prior for log-exp. power model). Only ind. Jeffreys prior is used for log-Student t.

collection. The times to follow-up (survival or censoring time) are recorded as the number of years since birth. Following Kwong and Hutton [2003], the amount of severe impairments (ambulation, manual dexterity and mental ability) and the birth weight (in kilograms) are used as predictors for the time to death. The percentage of children with 0, 1, 2 and 3 severe impairments is equal to 63%, 15%, 5% and 17% respectively. The median time to follow-up for these four categories (first and third quartiles in parenthesis) are 30.88 (26.12,38.42), 32.44 (27.09,38.96), 31.22 (23.96,38.22) and 17.91 (8.97,27.99) years, respectively. Regarding birth weight, 14%, 26% and 60% of the children were born with very low weight (less than 1.5 kg), low weight (1.501-2.5 kg) and normal weight (more than 2.5 kg). The median time to follow-up for these groups are 27.37 (23.69,31.14), 29.85 (24.84,36.87) and 30.83 (24.72,38.48). The deaths of 242 children were observed by the end of the observation period. The survival times of the remaining 1,307 patients are right censored, so there is a very large proportion of censoring (84.4%) in this dataset.

The data are analyzed using the SMLN-AFT and RMW-AFT models defined by the mixing distributions in Tables 3.1 and 3.2. Log-normal and Weibull AFT regressions without mixture are also fitted. Regression coefficients are defined as: β_0 (intercept), β_1 (amount of impairments: none), β_2 (amount of impairments: 1), β_3 (amount of impairments: 2) and β_4 (birth weight). Figures 4.16 and 4.17

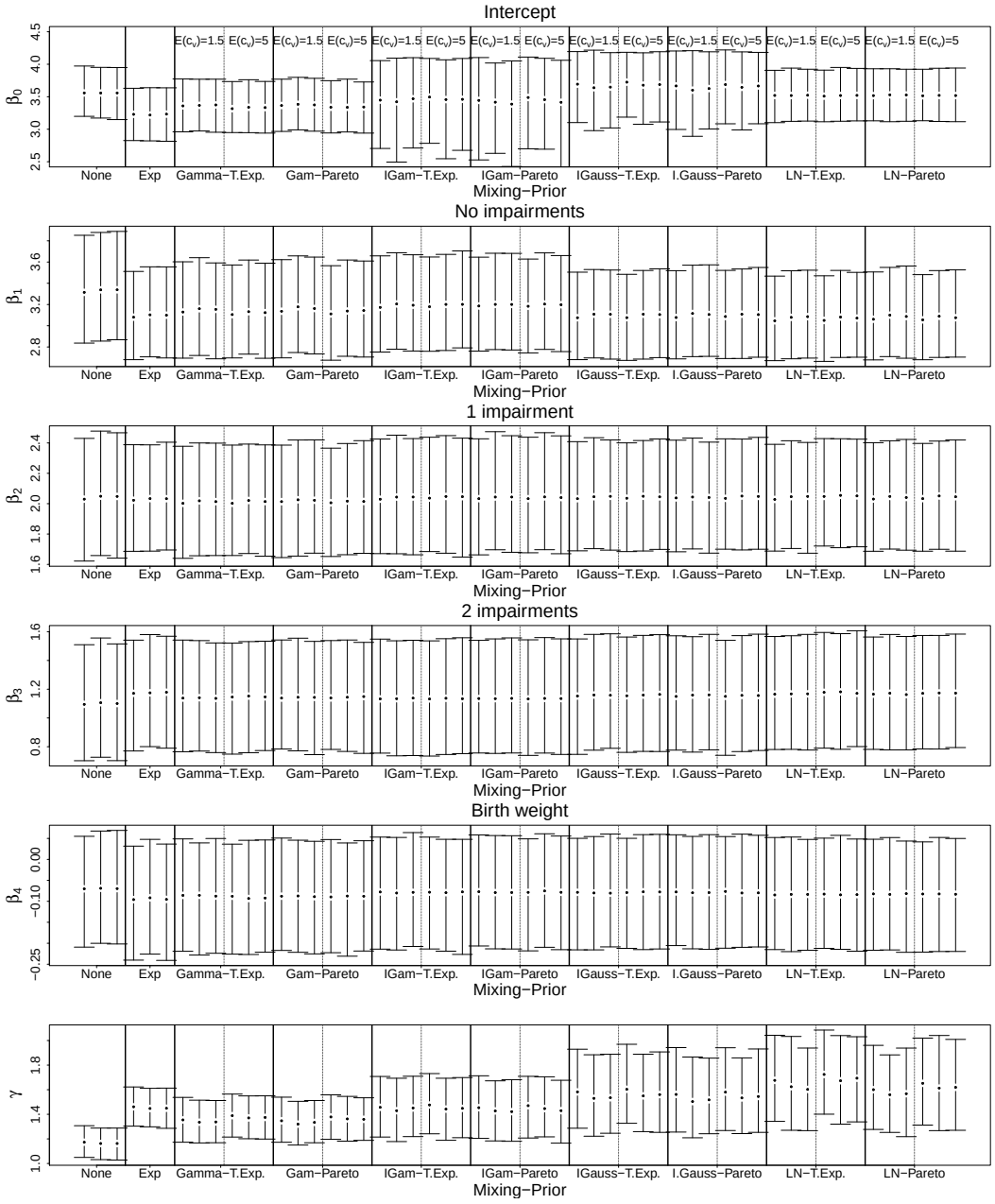


Figure 4.17: Cerebral palsy dataset. For RMW-AFT models with $\gamma \sim$ Gamma(d_1, d_2) and (if appropriate) a trunc. exponential or Pareto prior for c_v. Vertical lines are the HPD 95% intervals and dots represent posterior medians. From left to right, $d_1 = 4, d_2 = 1$, $d_1 = d_2 = 1$ and $d_1 = d_2 = 0.01$. Values of E(c_v) are displayed in the top panel.

summarize the marginal posterior inference. Set observations are used for SMLN-AFT models ($\epsilon_l = \epsilon_r = 0.5$). These estimations are in line with the ones in Kwong and Hutton [2003], where the log-normal, log-logistic and Weibull models were also fitted. Throughout, results are fairly insensitive to the choice within these priors. With the exception of the log-Laplace model, covariate effects do not greatly differ within SMLN-AFT models. The log-Laplace model relates to a strong unobserved heterogeneity, with var(Λ_i)=∞ and induces a more pronounced difference between those children with no impairments and those with 3 disabilities (the ratio between their median survival times is $e^{3.4} \approx 30$, in contrast to $e^{3.0} \approx 20$ predicted by the log-normal model and $e^{3.1} \approx 22$ for other SMLN models). For RMW-AFT models, set observations are not required and point observations are assumed throughout. In this case, the main differences relate to whether mixing is used or not. All Weibull mixtures estimate the effect of no impairments (β_1) to be less than in the Weibull model without mixing. Under the Weibull model, the median survival time is increased by a factor of approximately $e^{3.3} \approx 27$ for children with no impairments (w.r.t. those with 3 impairments). In contrast, under the RMW models, the same factor is estimated to be roughly $e^{3.1} \approx 22$. Furthermore, the bottom panel of Figure 4.17 shows that, in all cases, γ is estimated to be larger than 1. In line with the results in Kwong and Hutton [2003], this indicates a non-monotone shape of the underlying hazard rate (as in any SMLN model). Nonetheless, in order to accommodate the variability of the data, the Weibull model tends to underestimate γ (the coefficient of variation of the Weibull distribution is a decreasing function of γ). As shown in Tables 4.6 and 4.8, no influential observations are detected by any of these models.

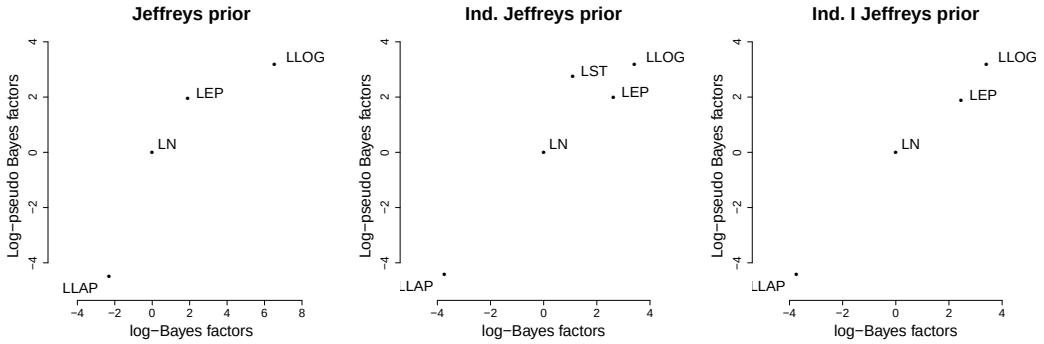


Figure 4.18: Cerebral palsy dataset. log-BF and log-PsBF (w.r.t. log-normal AFT) of SMLN-AFT models.

Figure 4.18 and Table 4.6 show that, with exception of the log-Laplace model,

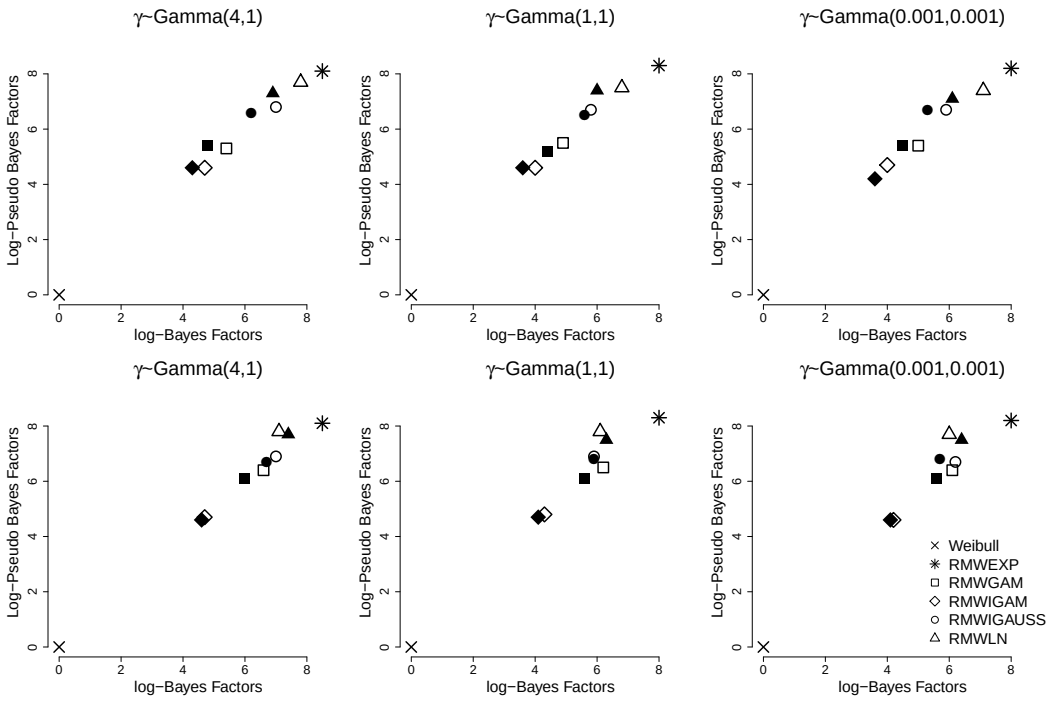


Figure 4.19: Cerebral palsy dataset. log-BF and log-PsBF (w.r.t. Weibull AFT) of RMW-AFT models. Unfilled and filled characters denote a trunc. exponential and Pareto priors for c_v, respectively. Upper panels use E(c_v)=1.5. Lower panels use E(c_v)=5. Legend is displayed in the last panel.

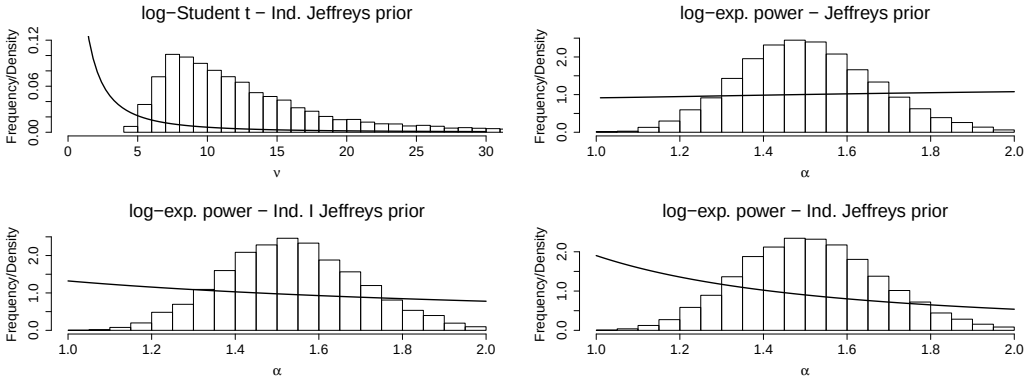


Figure 4.20: Cerebral palsy dataset. Histogram for the posterior sample of ν and α (log-Student t and log-exp. power models, respectively). Solid curve represents the prior density.

Table 4.6: Cerebral palsy dataset. For SMLN-AFT models: DIC, the fraction of observations with better CPO performance than the AFT-log-normal model, and the number of influential observations.

Prior	Model	log PsML	DIC	CPO better	No. obs. $p_i \geq 0.9$
Jeffreys	Log-normal	-1230.71	2460.94	-	0
	Log-Student t	-	-	-	-
	Log-Laplace	-1235.20	2470.53	56%	0
	Log-exp. power	-1228.76	2457.15	65%	0
	Log-logistic	-1227.54	2454.91	70%	0
Ind. Jeffreys	Log-normal	-1230.72	2461.03	-	0
	Log-Student t	-1227.98	2455.92	75%	0
	Log-Laplace	-1235.15	2470.50	56%	0
	Log-exp. power	-1228.73	2457.15	67%	0
	Log-logistic	-1227.55	2454.98	70%	0
Ind. I Jeffreys	Log-normal	-1230.72	2461.03	-	0
	Log-Student t	-	-	-	-
	Log-Laplace	-1235.15	2470.50	56%	0
	Log-exp. power	-1228.83	2457.38	67%	0
	Log-logistic	-1227.55	2454.98	70%	0

Table 4.7: Cerebral palsy dataset. For some RMW-AFT models under a Gamma(d_1, d_2) prior for γ: posterior medians and HPD 95% intervals of $R_{c_v}(\gamma, \theta)$ (as in equation (3.27)).

			$d_1 = 4, d_2 = 1$		$d_1 = d_2 = 1$		$d_1 = d_2 = 0.01$	
Prior c_v	E(c_v)	Mixing	Med.	HPD 95%	Med.	HPD 95%	Med.	HPD95%
T. Exp.	1.5	Gam(θ, θ)	2.41	[1.13, 4.36]	2.34	[1.17, 4.16]	2.35	[1.20, 4.15]
		Inv-Gam$(\theta, 1)$	1.41	[1.23, 1.55]	1.40	[1.18, 1.55]	1.41	[1.24, 1.55]
		Inv-Gauss$(\theta, 1)$	1.66	[1.43, 1.83]	1.63	[1.36, 1.84]	1.64	[1.37, 1.82]
		log-norm$(0, \theta)$	2.30	[1.53,2.99]	2.21	[1.41,3.10]	2.17	[1.40,2.85]
	5	Gam(θ, θ)	6.98	[1.54,19.90]	6.76	[1.59,20.00]	6.77	[1.56,19.97]
		Inv-Gam$(\theta, 1)$	1.43	[1.25, 1.55]	1.41	[1.20, 1.55]	1.41	[1.22, 1.55]
		Inv-Gauss$(\theta, 1)$	1.68	[1.49, 1.85]	1.65	[1.43, 1.83]	1.66	[1.45, 1.85]
		log-norm$(0, \theta)$	2.45	[1.76,3.21]	2.37	[1.63,3.22]	2.42	[1.64,3.18]
Pareto	1.5	Gam(θ, θ)	2.42	[1.13, 6.10]	2.20	[1.10, 5.86]	2.38	[1.07, 5.92]
		Inv-Gam$(\theta, 1)$	1.41	[1.19, 1.55]	1.39	[1.21, 1.54]	1.39	[1.19, 1.55]
		Inv-Gauss$(\theta, 1)$	1.65	[1.35, 1.84]	1.61	[1.31, 1.82]	1.62	[1.38, 1.83]
		log-norm$(0, \theta)$	2.10	[1.40,2.85]	2.05	[1.41,2.77]	2.06	[1.29,2.87]
	5	Gam(θ, θ)	4.45	[1.16,32.66]	4.35	[1.10,28.69]	4.18	[1.18,28.92]
		Inv-Gam$(\theta, 1)$	1.42	[1.23, 1.55]	1.41	[1.23, 1.55]	1.39	[1.17, 1.55]
		Inv-Gauss$(\theta, 1)$	1.66	[1.43, 1.85]	1.64	[1.37, 1.82]	1.65	[1.42, 1.85]
		log-norm$(0, \theta)$	2.25	[1.49,2.98]	2.19	[1.42,3.21]	2.21	[1.41,3.09]

SMLN models perform better than the log-normal one (for any of the considered priors). In terms of Bayes factors, this evidence is accentuated when using the

Table 4.8: Cerebral palsy dataset. For RMW-AFT models under a Gamma(d_1, d_2) prior for γ: DIC, the fraction of observations with better CPO performance than the Weibull model, and the number of influential observations.

			Trunc. exp. prior for c_v				Pareto prior for c_v			
E(c_v)	d_1, d_2	Mixing	log PsML	DIC	CPO better	No. $p_i \geq 0.9$	log PsML	DIC	CPO better	No. $p_i \geq 0.9$
1.5	4,1	None	-1235.6	2471.1	-	0	-1235.6	2471.1		0
		Exp.	-1227.5	2454.8	57%	0	-1227.5	2454.8	57%	0
		Gam.	-1230.3	2460.7	57%	0	-1230.2	2460.6	57%	0
		Inv-Gam.	-1231.0	2462.2	56%	0	-1231.0	2457.9	54%	0
		Inv-Gauss.	-1228.8	2457.9	54%	0	-1229.0	2458.4	53%	0
		Log-norm.	-1227.9	2455.9	52%	0	-1228.3	2456.7	53%	0
	1,1	None	-1235.7	2471.4	-	0	-1235.7	2471.4		0
		Exp.	-1227.5	2454.8	58%	0	-1227.5	2454.8	58%	0
		Gam.	-1230.2	2460.4	61%	0	-1230.5	2461.1	60%	0
		Inv-Gam.	-1231.2	2462.9	58%	0	-1231.1	2458.4	55%	0
		Inv-Gauss.	-1229.0	2458.4	55%	0	-1229.2	2458.7	54%	0
		Log-norm.	-1228.2	2456.7	55%	0	-1228.3	2456.9	55%	0
	0.01,0.01	None	-1235.7	2471.2	-	0	-1235.7	2471.2		0
		Exp.	-1227.5	2454.9	57%	0	-1227.5	2454.9	57%	0
		Gam.	-1230.3	2460.6	57%	0	-1230.2	2460.6	55%	0
		Inv-Gam.	-1231.0	2462.4	55%	0	-1231.4	2458.2	54%	0
		Inv-Gauss.	-1228.9	2458.2	54%	0	-1229.0	2458.3	54%	0
		Log-norm.	-1228.3	2456.8	53%	0	-1228.6	2457.5	54%	0
5	4,1	None	-1235.6	2471.1	-	0	-1235.6	2471.1		0
		Exp.	-1227.5	2454.8	57%	0	-1227.5	2454.8	57%	0
		Gam.	-1229.2	2458.2	57%	0	-1229.5	2459.0	57%	0
		Inv-Gam.	-1230.9	2462.0	57%	0	-1231.0	2457.7	53%	0
		Inv-Gauss.	-1228.7	2457.7	53%	0	-1228.9	2458.2	55%	0
		Log-norm.	-1227.8	2455.6	50%	0	-1227.9	2456.0	52%	0
	1,1	None	-1235.7	2471.4	-	0	-1235.7	2471.4		0
		Exp.	-1227.5	2454.8	58%	0	-1227.5	2454.8	58%	0
		Gam.	-1229.2	2458.4	60%	0	-1229.6	2459.3	60%	0
		Inv-Gam.	-1230.9	2462.3	56%	0	-1231.0	2458.0	55%	0
		Inv-Gauss.	-1228.8	2458.0	55%	0	-1229.0	2458.2	56%	0
		Log-norm.	-1228.1	2456.3	53%	0	-1228.3	2456.6	55%	0
	0.01,0.01	None	-1235.7	2471.2	-	0	-1235.7	2471.2		0
		Exp.	-1227.5	2454.9	57%	0	-1227.5	2454.9	57%	0
		Gam.	-1229.3	2458.5	56%	0	-1229.5	2459.2	57%	0
		Inv-Gam.	-1231.0	2462.4	56%	0	-1231.1	2458.2	54%	0
		Inv-Gauss.	-1228.9	2458.2	54%	0	-1228.9	2458.2	53%	0
		Log-norm.	-1227.9	2455.9	52%	0	-1228.2	2456.6	53%	0

original Jeffreys prior. As in the VA lung cancer application, Figure 4.20 indicates that this is also supported by the posterior of ν and α for the log-Student t or log-exponential power models, respectively. For RMW models, Figure 4.19 and Table 4.8 indicate that all the suggested Weibull mixtures provide a better fit for the data and

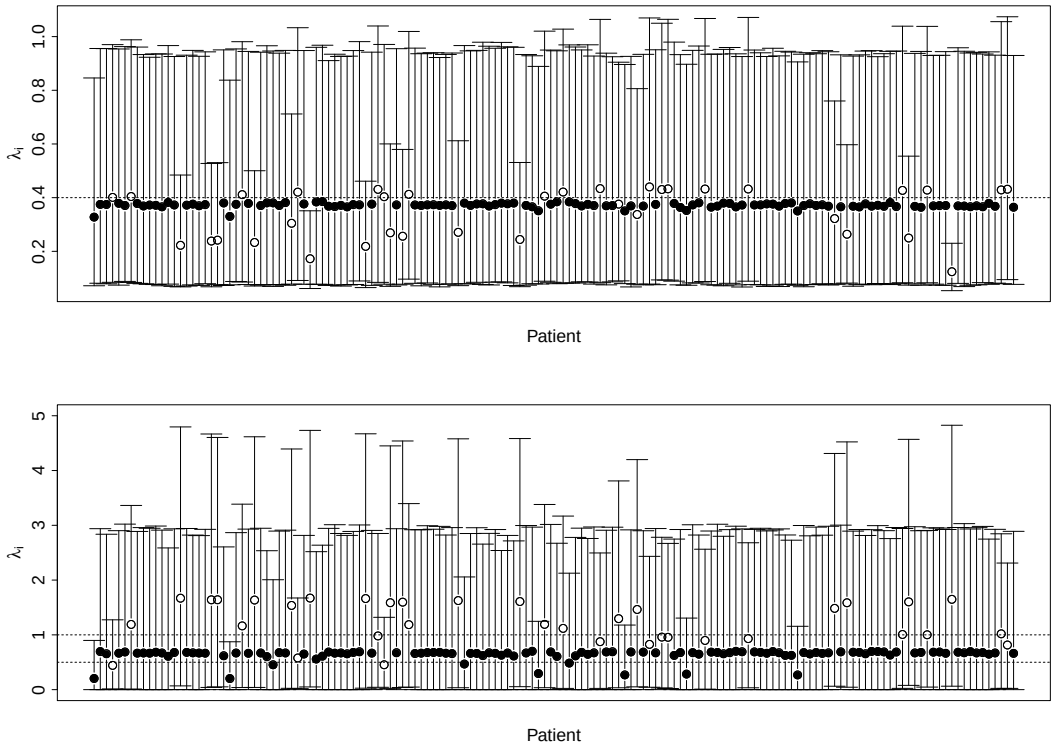


Figure 4.21: Cerebral palsy dataset. For a random sub-sample of 150 children, λ_i: vertical lines are the HPD 95% intervals and circles represent posterior medians (filled for censored observations). Horizontal lines are located at λ_{ref} (and λ^c_{ref}, if appropriate). Upper panel: log-logistic model (ind. Jeffreys prior). Lower panel: RMW model with exponential(1) mixing ($\gamma \sim$Gamma(4,1)).

lead to better predictions. In line with the results displayed in Table 4.7 (where the posterior distribution of R_{c_v} is concentrated away from one), this strongly suggests the existence of unobserved heterogeneity. Overall, the Weibull model provides the worst fit for these data (in terms of DIC and PsML). The log-Laplace distribution has a similar performance (for the log-exponential power model, the posterior α is far from one which corroborates a poor log-Laplace fit). In contrast, the log-logistic regression appears as the best SMLN model. The exponential(1) mixing distribution provide the best results among RMW models. These models are very similar in terms of DIC and PsML and are simple to elicit (as there is no θ). In practice, the same estimations for the regression parameters (including the intercept) are obtained using both models. Nonetheless, the RMW model with exponential mixing is computationally more attractive (as the full conditionals of the λ_i's are easy to sample from).

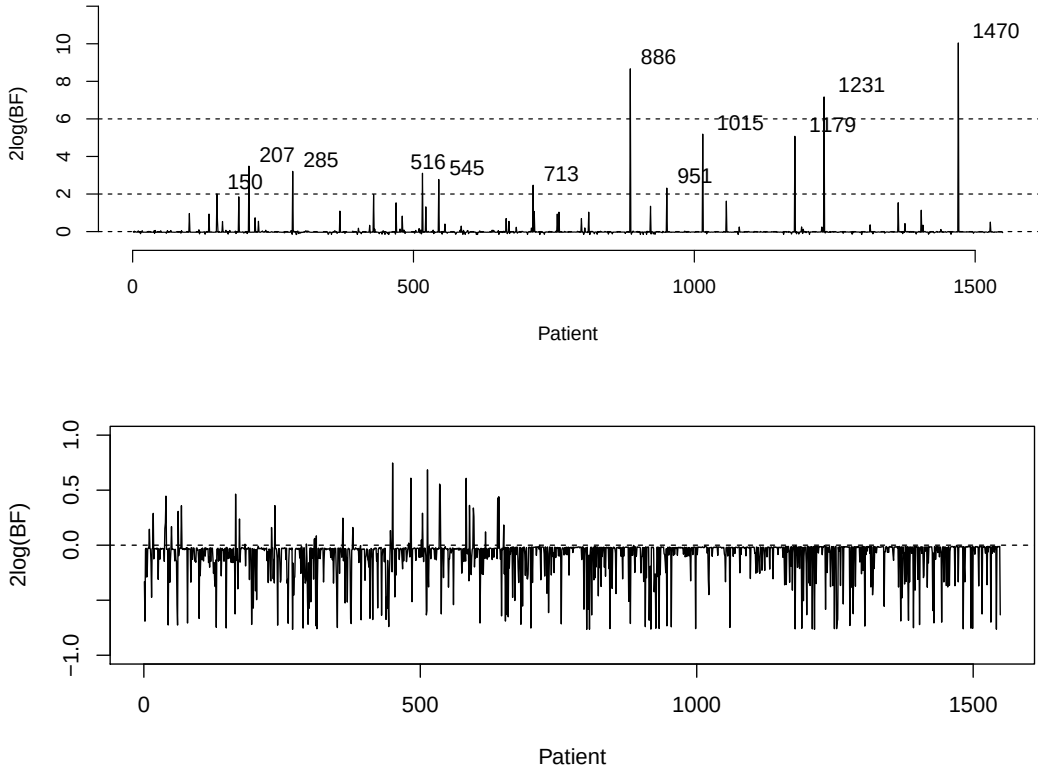


Figure 4.22: Cerebral palsy dataset. BF in favour of $H_1 : \lambda_i \neq \lambda_{ref} (u_i \neq u_{ref})$ versus $H_0 : \lambda_i = \lambda_{ref}$. Horizontal lines drawn at 1 for reference. Upper panel: log-logistic model (ind. Jeffreys prior). Lower panel: RMW model with exponential(1) mixing ($\gamma \sim$Gamma(4,1)).

Existence of unobserved heterogeneity is also supported by the posterior distribution of the mixing parameters. This is illustrated in Figure 4.21 (only a random sub-sample of 150 records is displayed). For all priors considered, the log-logistic model suggests a mild unobserved heterogeneity, where most of the λ_i's have a very similar posterior distribution (with posterior medians located around $\lambda_{ref} = 0.4$). As shown by the upper panel of Figure 4.22, 12 outliers are detected by the log-logistic model (with respect to the the log-normal model). The most clear outliers are observations 886, 1015, 1179, 1231 and 1470 (only the last two are non-censored records), for which the posterior median of the λ_i's are all below 0.16 (the reference value is 0.4). In particular, the follow-up times of patients 1015, 1179 and 1470 are substantially smaller than for all other children with normal birth weight (more than 2.5 kg) and no impairments. The same occurs for patients 886 and 1231 when considering those children that had no impairments but a low birth weight. A dif-

ferent picture is provided by the RMW model with exponential(1) mixing (see lower panels in Figures 4.21 and 4.22). In such a case, there is more variability between the posterior distributions of the λ_i's (especially within the non-censored observations). Thus, there is evidence of unobserved heterogeneity in the sample, which provides strong support for the mixture, but there are no particular single observations that could be considered clear outliers (with respect to the Weibull model). Of course, several outliers would be detected if ignoring the effect of censoring on the mixing parameters.

4.5 Concluding remarks

The three datasets presented here are quite representative of standard survival applications. Nonetheless, very distinct features for these datasets were uncovered by the analysis. Overall, the methodology introduced in Chapter 2 was shown to have a better performance than the exponential, Weibull and log-normal models (with no mixture). In general, whether or not the frailty terms are incorporated in the model is critical for the inference on β. However, posterior inference on this parameter is relatively robust to the adoption of a particular mixing distribution. In addition, all the priors employed for the analysis induced similar covariate effects. In contrast, inference on θ (if unknown) was slightly more affected by changes in the prior. Nevertheless, this has no major practical consequences (as θ can be often treated as a nuisance parameter).

For the VA lung cancer data ($n = 137$, 6.6% of censoring and 5 covariates, two of which are categorical with more than 2 levels, leading to a total of 8 effects), the analysis reports a mild unobserved heterogeneity which is mostly linked to few outlying observations. In this case, all mixtures suggested a non-monotone individual hazard rate. In contrast, for the AA bone marrow transplant application ($n = 101$, 50.5% of censoring and 1 covariate), mixtures of exponential distributions (which induce a constant underlying hazard rate) were preferred. These models suggested an important amount of unobserved heterogeneity but no particularly anomalous observations are detected. This is reasonable given that only the treatment type was used as predictor and its effect appears to be non significative. Finally, for the cerebral palsy data ($n = 1549$, 84.4% of censoring and 2 covariates, with 4 effects), both mixture families revealed a non-monotone underlying hazard function but different scenarios in terms of unobserved heterogeneity (with respect to the underlying model). Whereas mixtures of Weibull distributions reported a strong unobserved heterogeneity and no outliers, mixtures of log-normal distributions found strong

evidence of outliers. The latter highlights that the definition of unobserved heterogeneity and outliers is relative to an underlying or base model (which is different for both families of models). Nonetheless, in this application, the choice within one of these mixture families has no practical consequences for the inference about β (which is frequently the parameter of interest).

Chapter 5

Survival modelling of university outcomes

> "Education is the most powerful weapon which you can use to change the world".
> Nelson Mandela

5.1 Motivation

During the last decades, the coverage of the higher education system had a significant growth in Chile. According to the Chilean Ministry of Education[1], the total admission evolved from around 165 thousand students at the beginning of the 80's to more than 1 million students enrolled in 2012. Nowadays, the access to higher education is not restricted to an elite group. Among others, this is a result of a bigger role for studies as a tool for social mobility, the increase of the number of scholarships, a more accessible system of student loans and the opening of new institutions. This change of scenario entails new problems. One of them is an alarming amount of university dropouts. Currently, more than half of the students enrolled at higher education institutions does not complete their degree. This figure includes students expelled from the university for academic or disciplinary reasons and those who voluntarily resigned (the term "voluntary" is understood as the drop out that is not controlled by the university but is not necessarily the student's will; *e.g.* a student can be forced to dropout because of financial hardship). Another issue is the high frequency of late graduations, in which the student takes longer

[1] http://www.mineduc.cl/

than the official duration of the programme in order to obtain the degree. Unlike the UK's and other educational systems, Chilean universities allow more flexibility. Students can repeat failed modules and/or have a reduced academic load in some semesters. These issues involve a waste of time and resources from the perspective of the students, their families, universities and the society.

There is a large literature devoted to university dropout. It includes conceptual models based, among others, on psychological, economic and sociological theories [*e.g.* Tinto, 1975; Bean, 1980]. Here, instead, the focus is on empirical models. In this context, a large share of the previous research treats the dropout as a dichotomous problem, neglecting the temporal component. In other words, they focus on *whether* or not a student has dropped out from university at a fixed time (*e.g.* dropout by the second year of enrollment). Ignoring *when* the dropout occurs is a serious waste of information [Willett and Singer, 1991]. Potential high risk periods will not be identified and no distinction between early and late dropout will be made. An alternative is to use (standard) survival models for the time to dropout [as in Murtaugh et al., 1999]. This approach labels graduated students as right censored observations, which is a major pitfall. Whilst students are enrolled at university, dropout is a possibility. However, dropout cannot occur after graduation (the time to dropout is "infinite"), contradicting the idea of censoring. Instead, graduation must be considered as a competing event and incorporated into the survival model.

This study aims to identify determinants of the length of stay at university and its associated academic outcome for undergraduate students of the Pontificia Universidad Católica de Chile (PUC). The PUC is one of the most prestigious universities in Chile and it is the second best university in Latin America[2]. Despite having one of the lowest dropout rates in the county (far below the national level), dropout is still a significant issue for some degrees of the PUC. The output of this analysis aims to help university authorities in order to have a better understanding of the current situation at the university. Hopefully, it will also inspire policies mitigating late graduations and dropouts.

A competing risks model is proposed for the length of stay at university, where the possible events are defined as graduation, voluntary dropout and involuntary dropout. These are defined as the final academic situation recorded by the university at the end of 2011. Students that have not experienced any of these events by the end of 2011 are labeled as right censored observations (censoring is assumed to be non-informative). In Chile, the academic year is structured in semesters

[2]According to QS Ranking 2013. http://www.topuniversities.com/

(March-July and August-December). Survival times are defined as the length of enrollment at university and measured in semesters from admission (which means they are inherently discrete). It is an advantage of this approach that it deals jointly with graduations and dropouts. This analysis does not account for all the features of this complex dataset yet it provides a better understanding of the problem and a starting point for future research. This Chapter is organized as follows. The main features of the PUC dataset are summarized in Section 5.2, showing high levels of heterogeneity between programmes. This diversity is in terms of academic outcomes and the population composition of each degree programme. Section 5.3 introduces competing risk models with focus in the context of university outcomes. This model can be estimated by means of a multinomial logistic regression. It explained how the Bayesian setting is particularly helpful in this application where maximum likelihood inference for the suggested model is precluded. In addition, Section 5.4.2 introduces a Gibbs sampling algorithm that exploits a hierarchical representation of the multinomial logistic likelihood [based on Holmes and Held, 2006; Polson et al., 2013]. The last part in Section 5.4.2 relates to the critical issue of covariate selection. The output of the analysis is summarized in Section 5.5, focusing on some of the science programmes which are more affected by dropout and late graduations. Finally, Section 5.6 compiles the main findings of the study, discussing possible limitations and future extensions.

5.2 The PUC dataset

The PUC provided anonymized information about 34,543 students enrolled via the ordinary admission process during the period 2000-2011. This admission process selects students according to their high school marks and the results of a standardized university selection test, which is applied at a national level. Only the curriculums that existed during the whole 2000-2011 period are included. The following inclusion criteria are defined for the analysis, which will only consider students who

- were enrolled for at least 1 semester (as the dropout produced right after enrollment might have a different nature),
- are enrolled in a single programme (students doing parallel degrees usually need more time to graduate and have less risk of dropout),
- do not have validated modules previously approved from other degree programme (in which case the time to graduation can be significantly reduced),

- were alive by the end of 2011 (0.1% of the students had died by then),
- had full covariates' information (a small number of missing values was recorded, missingness at random is assumed).

Overall, 78.7% of the students satisfied these criteria. Table 5.1 breaks this number down by program. Throughout, the analysis will only consider this subset of the original data.

By the end of 2011, 41.9% of the students were still enrolled (right censored observations), 37.2% graduated, 6.6% were expelled (involuntary dropout, which is mostly related to poor academic performances), 10.7% withdrew (voluntary dropout), and 3.7% abandoned the university without an official withdrawal. Following university's guidance the latter group is classified as voluntary dropout. The large percentage of censoring is mostly linked to students from later years of entry, who were not yet able to graduate by the end of 2011. From those who are not currently enrolled, only an overall 65% graduated. Figure 5.1 shows that the performance of former students is not homogenous across programmes. In terms of total dropout, Medicine (8.2%) has the lowest rate and the highest rates are for Chemistry (79.4%) and Mathematics and Statistics (79.3%). The highest rates of involuntary dropout belong to Agronomy and Forestry Engineering (28.9%) and Mathematics and Statistics (26.2%). Chemistry (56.5%) and Astronomy (56.0%) present the largest rates of voluntary dropout. Dropouts are mostly observed during the first semesters of enrollment. In contrast, graduation times are concentrated on large values, typically above the official length of the programme (the duration of different programmes varies between 8 and 14 semesters, with a typical value of 10 semesters). Figure 5.2 displays the distribution of graduated students in terms of timely graduation. Strong levels of heterogeneity between programmes are exhibited. In fact, the proportion of students that graduated on time varies from 88% (Medicine and Education Elementary School in Villarrica Campus) to 12% (Mathematics and Statistics) and 11% (Education Elementary School).

The covariates listed in Table 5.2 were recorded at enrollment time. They include demographic, socioeconomic and variables related to the admission process. According to these covariates, substantial differences are observed between programmes (see Figures E.1 to E.8 in Appendix E). In terms of demographic factors, some degrees concentrate a high percentage of female students (*e.g.* all education-related programs, Nursing). In contrast, most of the Engineering students are males. The proportion of students who live outside the Metropolitan area is more stable across programmes (of course, a particularly high percentage

is observed in the Education for Elementary School degree taught in the Villarrica campus, which is located in the south of Chile). Strong levels of heterogeneity are also detected for the socioeconomic characterization of the students. Chilean schools are classified according to their funding system as public (fully funded by the government), subsidized private (the state covers part of the tuition fees) and private (no funding aid). This classification can be considered as a proxy for the socioeconomic situation of the student (low, middle and upper class, respectively). The educational level of the parents is usually a good indicator of socioeconomic status as well. In the PUC, some degrees have a very low percentage of students that graduated from public schools (*e.g.* Business Administration and Economics, Design) and others have a high percentage of students with parents without a higher degree (*e.g.* Education for Elementary School in Villarrica Campus, Chemistry and Pharmacy). In addition, a few programmes had low rate of students who receive a scholarship or have a student loan (*e.g.* Business Administration and Economics, Architecture). Finally, "top" programmes (*e.g.* Medicine, Engineering, Law, Business Administration and Economics) only admit students with the highest selection scores. For instance, for the admission process 2011, the lowest score selected in the Arts programme was 603.75 but Medicine did not enroll any students with score below 787.75 (the minimum score required when applying to the PUC is 600, except for some education-related programmes where exceptions apply). In the same spirit, these highly selective programmes only enrolled students that applied to it as a first preference.

This substantial heterogeneity (in terms of outcomes and covariates) precludes a meaningful comparison of academic outcomes across programmes. Thus, the analysis will be carried out separately for each degree.

5.3 Discrete time competing risks models

Standard survival models only allow for a unique event of interest. Occurrences of alternative events are often recorded as censored observations. For instance, in the context of university outcomes, graduated students might be recorded as censored observations when the event of interest is dropout [as in Murtaugh et al., 1999]. This is not appropriate. Clearly, those students who graduated are no longer able to dropout (from the same degree). Alternatively, competing risks models can be used when more than one type of event can occur and there is a reason to believe they are a result of different mechanisms. Competing risks models incorporate simultaneously both the survival time and the type of event (or cause). There is

Table 5.1: PUC dataset. Amount of students satisfying the inclusion criteria used in this study broken down by program.

Program	No. students	% students
Acting	362	80.1
Agronomy and Forestry Engineering	2,466	85.2
Architecture	841	69.9
Art	688	76.3
Astronomy	295	88.3
Biochemistry	331	85.5
Biology	791	83.9
Business Administration and Economics	2,027	72.7
Chemistry	379	82.0
Chemistry and Pharmacy	687	85.6
Civil Construction	1,930	86.0
Design	651	65.2
Education, elementary school	1,277	81.4
Education, elementary school (Villarrica campus)	301	80.5
Education, preschool	949	83.2
Engineering	3,522	69.3
Geography	534	84.5
History	552	76.6
Journalism and Media Studies	876	76.2
Law	2,303	84.2
Literature (Spanish and English)	911	80.8
Mathematics and Statistics	598	78.0
Medicine	972	89.8
Music	161	74.5
Nursing	886	78.6
Physics	237	85.9
Psychology	801	75.9
Social Work	440	87.5
Sociology	421	74.0
Total	27,189	78.7

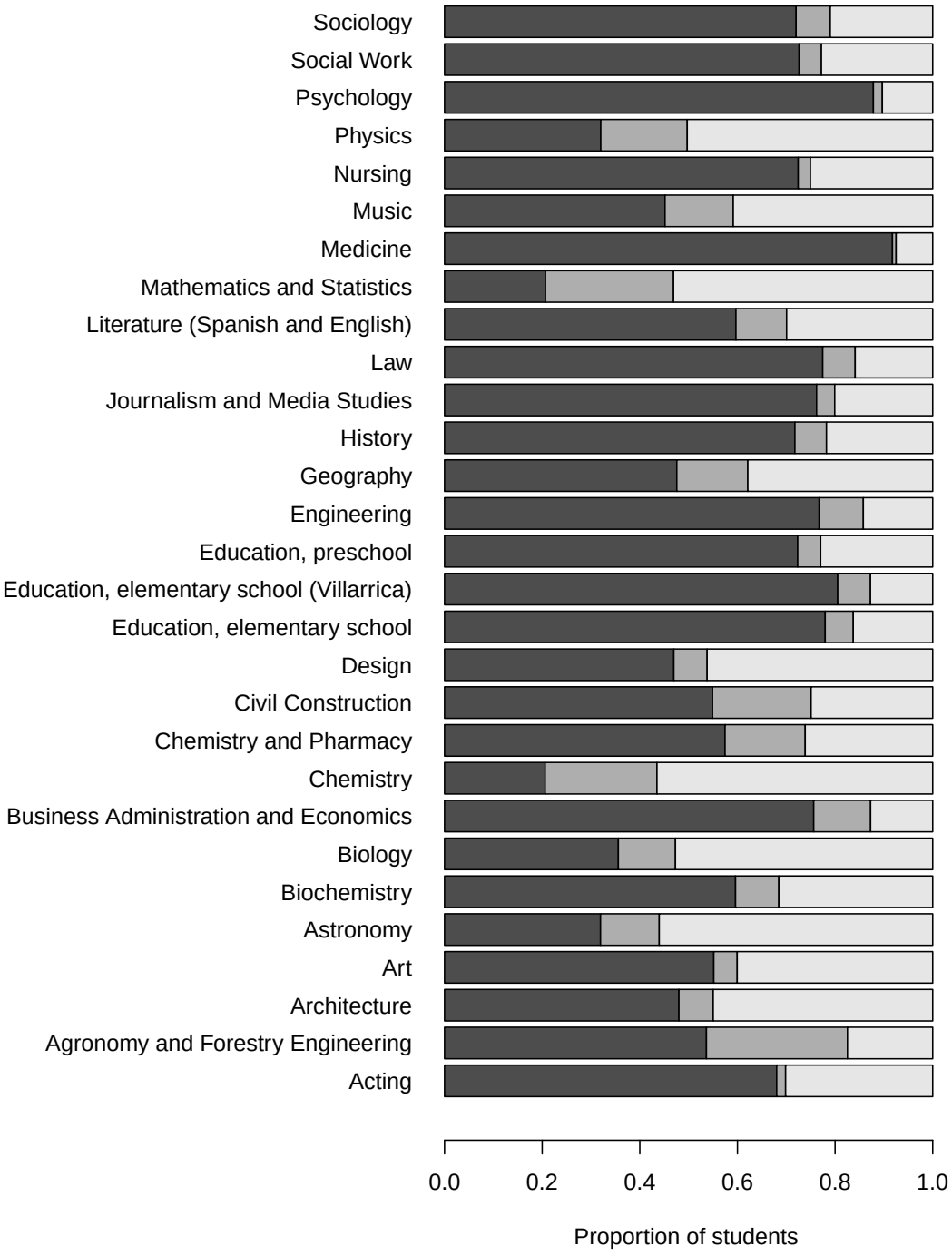


Figure 5.1: PUC dataset. Distribution of former students according to final academic situation. From darkest to lightest, colored areas represent the proportion of students that: graduated, involuntary dropout and voluntary dropout, respectively.

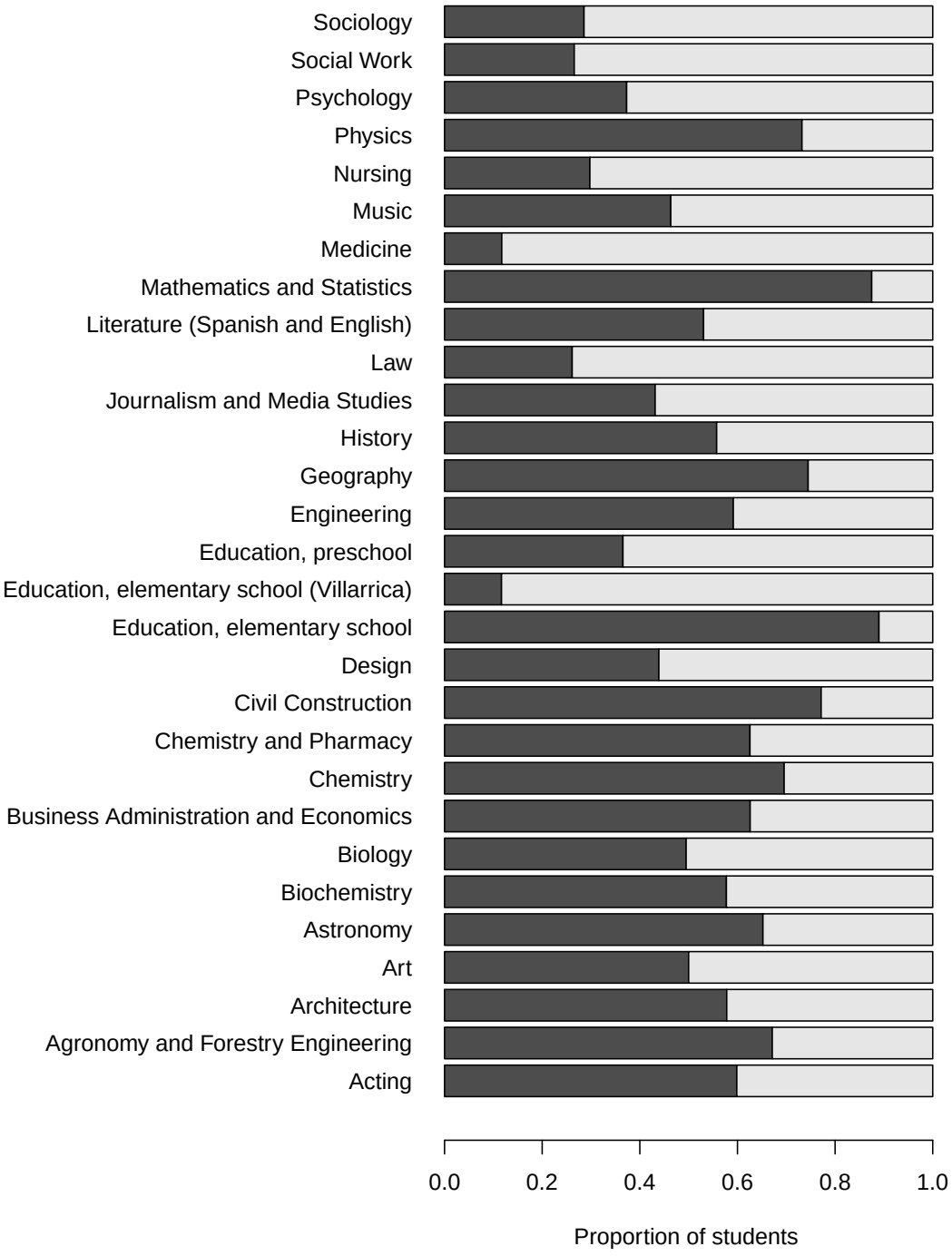


Figure 5.2: PUC dataset. Distribution of graduated students according to opportune graduation (with respect to the official duration of the programme). The lighter area represents the proportion of students with timely graduation.

Table 5.2: PUC dataset. Available covariates (recorded at enrollment). Options for categorical variables in parentheses.

Demographic factors
Sex (female, male)
Region of residence (Metropolitan area, others)
Socioeconomic factors
Parents education (at least one with a technical or university degree, no degrees)
High school type (private, subsidized private, public)
Funding (scholarship and loan, loan only, scholarship only, none)
Admission-related factors
Selection score (numerical)
Application preference (first, others)
Gap between high school graduation and admission to PUC (1 year or more, none)

a large literature about this topic [*e.g.* Crowder, 2001; Pintilie, 2006; Beyersmann et al., 2012]. However, most of it focuses on continuous survival times. Instead, in the context of university outcomes (where survival times are usually measured in numbers of academic periods), a discrete time approach is more appropriate.

In a discrete-time competing risks setting, the variable of interest is (R, T), where $R \in \{1, \ldots, \mathcal{R}\}$ denotes the type (or reason) for the observed event and $T \in \{1, 2, \ldots\}$ is the survival time. Analogously to the single-event case, a model can be specified via the sub-distribution or sub-hazard functions which are respectively given by

$$F_{(R,T)}(r,t) = P(R = r, T \leq t), \tag{5.1}$$

$$h_{(R,T)}(r,t) = \frac{P(R = r, T = t)}{P(T \geq t)}. \tag{5.2}$$

For ease of notation, the sub-index (R, T) is omitted onwards. The sub-distribution function (also called cumulative incidence function) represents the proportion of individuals for which an event type r has been observed by time t. On the other hand, $h(r, t)$ is the conditional probability of observing an event of type r during period t given that no event (nor censoring) has happened before. The total hazard rate for all causes is defined as $h(t) = \sum_{r=1}^{\mathcal{R}} h(r, t)$. In the discrete case, the maximum likelihood non-parametric estimator of $h(r, t)$ corresponds to the ratio between the number events of type r observed during period t and the total number of individuals who were at risk at time t [Singer and Willett, 1993; Crowder, 2001]. The latter is a discrete adaptation of the Kaplan-Meier estimator [Kaplan and Meier, 1958]. Although the sub-hazard rate can be easily estimated from the data, its

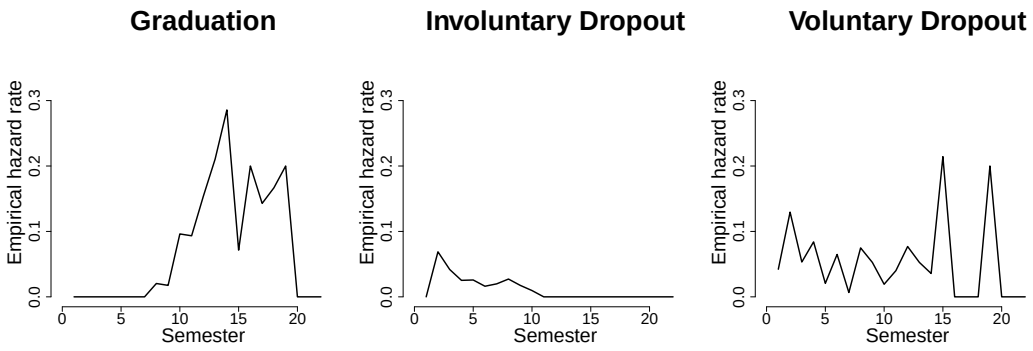


Figure 5.3: PUC dataset. Non-parametric estimation of cause-specific hazard rates for Chemistry students.

interpretation is not trivial. Alternatively, the cumulative incidence function is often preferred when interpreting results. It terms of sub-hazard rates, it adopts the following recursion [Kalbfleisch and Prentice, 2002]

$$F(r,1) = h(r,1), \tag{5.3}$$

$$F(r,t) = h(r,t)[1-F(t-1)] + F(r,t-1), \qquad t>1. \tag{5.4}$$

In some contexts, a simple (cause-specific) parametric model can be assigned to the survival times. For example, a geometric model is the discrete-time analogue for exponential survival times [see example 5.1 in Crowder, 2001]. However, such straightforward parametric models are not suitable for analyzing the PUC dataset. Overall, for these data, the cause-specific hazard rates have a rather erratic behaviour over time. Figure 5.3 illustrates this for Chemistry students. In particular, no graduations are observed during the firsts semesters of enrollment, inducing a zero graduation hazard at those times. In fact, graduations only start about a year before the official duration of the programme (10 semesters). In addition, during the first years of enrollment, the hazard of voluntary dropout has spikes located at the end of each academic year (even semesters). Therefore, more flexible models are required in order to accommodate these hazard paths.

5.3.1 Proportional Odds model for competing risks data

Cox [1972] proposed a Proportional Odds (PO) model for discrete times and single cause of failure. It is a discrete variation of the well-known Cox PH models, proposed in the same seminal paper. Let $x_i \in \mathbb{R}^k$ be a vector containing the value of k covariates associated to individual i and $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbb{R}^k$ a vector of regression

parameters. The Cox PO model is given by

$$\log\left(\frac{h(t|\delta_t,\beta;x_i)}{1-h(t|\delta_t,\beta;x_i)}\right)=\log\left(\frac{h(t)}{1-h(t)}\right)+x_i'\beta\equiv\delta_t+x_i'\beta,\qquad i=1,\ldots,n,\quad(5.5)$$

where $\{\delta_1,\delta_2,\ldots\}$ respectively represent the baseline log-odds at times $\{1,2,\ldots\}$ and $t=1,\ldots,t_i$. The model in (5.5) can be estimated in most statistical software by means of a binary logistic regression [Singer and Willett, 1993]. For this purpose, the data has to be transformed into a person-period format. Define Y_{it} as 1 if the event is observed at time t for the individual i; 0 otherwise. In the person-period format, each individual is represented by as many rows as periods in which he/she was at risk. To illustrate, Table 5.4 shows the transformed version of the fictional data displayed in Table 5.3. The period-indicators δ_t are estimated by introducing binary variables to the set of covariates. One basic assumption of the logistic regression is the independence between observations. However, in the person-period data, there is a clear association between observations linked to the same individual. Nevertheless, as shown in Singer and Willett [1993], the likelihood related to the survival process coincides with the likelihood of the logistic regression model for which the rows in the person-period data are treated as independent Bernoulli trials. In fact, the contribution to the likelihood of the individual i (data collection for this individual stops if the event is observed or right censoring is recorded) is given by

$$L_i=P(Y_{it_i}=y_{it_i},\cdots,Y_{i1}=y_{i1})=h(t_i)^{y_{it_i}}\prod_{s=1}^{t_i}[1-h(s)]^{1-y_{is}},\qquad(5.6)$$

which is derived by decomposing $P(Y_{it_i}=y_{it_i},\cdots,Y_{i1}=y_{i1})$ as a sequential product of conditional probabilities (covariates are omitted for easy of notation). Equivalently, defining $c_i=0$ if the survival time is observed (*i.e.* $Y_{it_i}=1, Y_{i(t_i-1)}=0,\ldots,Y_{i1}=0$) and $c_i=1$ if right censoring occurs (with t_i as the terminal time), we can express the likelihood contribution as

$$L_i=\left[\frac{h(t_i)}{1-h(t_i)}\right]^{1-c_i}S(t_i),\qquad S(t_i)=\prod_{s=1}^{t_i}[1-h(s)],\qquad(5.7)$$

which is the same expression that would be obtained in a survival setting.

The model in (5.5) can be extended in order to accommodate $\mathcal{R}$ possible events. Let $B=\left\{\beta_{(1)},\ldots,\beta_{(\mathcal{R})}\right\}$ be a collection of cause-specific regression parameters (each of them defined on $\mathbb{R}^k$). Define $\delta=\{\delta_{11},\ldots,\delta_{\mathcal{R}1},\delta_{12},\ldots,\delta_{\mathcal{R}2},\ldots\}$. A

Table 5.3: Fictional data. Example of a standard competing risks dataset (covariates are omitted for simplicity).

ID	Follow-up time	Event
1	8	Observed
2	3	Censored

Table 5.4: Fictional data. Person-period format for the data shown in Table 5.3.

ID	Period	Outcome
1	1	0
1	2	0
1	3	0
1	4	0
1	5	0
1	6	0
1	7	0
1	8	1
2	1	0
2	2	0
2	3	0

multinomial logistic regression can be defined as

$$\log\left(\frac{h(r,t|\delta,B;x_i)}{h(0,t|\delta,B;x_i)}\right)=\delta_{rt}+x_i'\beta_{(r)},\qquad r=1,\dots,\mathcal{R};t=1,\dots,t_i;i=1,\dots,n,\tag{5.8}$$

where

$$h(0,t|\delta,B;x_i)=1-\sum_{r=1}^{\mathcal{R}}h(r,t|\delta,B;x_i)\tag{5.9}$$

is the hazard of no event being observed at time t. The latter is equivalent to

$$h(r,t|\delta,B;x_i)=\frac{e^{\delta_{rt}+x_i'\beta_{(r)}}}{1+\sum_{s=1}^{\mathcal{R}}e^{\delta_{st}+x_i'\beta_{(s)}}}.\tag{5.10}$$

This notation implies that the same predictors are used for each cause-specific component (but this is easily generalised). In (5.8), covariates have an effect that is homogeneous over time. Hence, changes in the covariates influence both the marginal probability of the event ($P(R=r)$) and the speed at which the event occurs. In fact, positive values of the cause-specific coefficients indicate that (at any time point) the hazard of the corresponding event increases with unit changes in the associated covariates. In the context of university outcomes, (5.8) has been used by Scott and Kennedy [2005], Arias Ortis and Dehon [2011] and Clerici et al. [2014], among others. Nonetheless, its use has some drawbacks. First, it involves a large number of parameters. In fact, if $\mathcal{T}$ is the maximum of the recorded sur-

vival/censoring times, there are $\mathcal{R} \times \mathcal{T}$ different δ_{rt}'s. Scott and Kennedy [2005] overcome this by assigning a unique indicator δ_{rt_0} to the period $[t_0, \infty)$ (for fixed t_0). The choice of this threshold is rather arbitrary but it is reasonable to choose a value of t_0 such that most of the individuals already experienced one of the events by time t_0. Second, maximum likelihood inference for the multinomial logistic regression is precluded when the outcomes are (quasi) complete separated with respect to the predictors, *i.e.* a subset of the possible outcomes are not (or rarely) observed for some covariate configurations [Albert and Anderson, 1984]. In other words, the predictors can (almost) perfectly predict the outcomes. In the case of (5.8), these predictors include binary variables that are related to the period indicators δ_{rt}'s. Therefore, (quasi) complete separation will occur if the event types are (almost) entirely defined by the survival times. This is a major issue in the context of university outcomes. For example, no graduations can be observed during the second semester of enrollment. Therefore, the likelihood function will be maximized when the cause-specific hazard related to graduations (defined in (5.10)) is equal to zero at time $t = 2$. Thus, the "best" value of the corresponding period-indicator is $-\infty$.

In order to overcome these problems, Singer and Willett [2003] suggests polynomial baseline odds when modelling single outcomes. This can be easily extended to the competing risks case as

$$\log\left(\frac{h(r,t|\delta^*,B;x_i)}{h(0,t|\delta^*,B;x_i)}\right) = \delta^*_{r0}+\delta^*_{r1}(t-1)+\delta^*_{r2}(t-1)^2+\cdots+\delta^*_{r\mathcal{P}}(t-1)^{\mathcal{P}}+x_i'\beta_{(r)}, \quad (5.11)$$

where $\delta^* = \{\delta_{10}, \ldots, \delta_{\mathcal{R}0}, \ldots, \delta_{1\mathcal{P}}, \ldots, \delta_{\mathcal{R}\mathcal{P}}\}$ and $\mathcal{P}$ denotes de degree of the polynomial. Defining the polynomial in terms of $t-1$ facilitates the interpretation of the intercept (δ^*_{r0} represents the baseline cause-specific hazard at time $t = 1$). This option is less flexible than (5.8), but it is not affected by a separation of the outcomes with respect to the survival times. Nevertheless, its use is only attractive when a low-degree polynomial is good enough to represent the baseline hazard odds. This is not the case for the PUC dataset, where cause-specific hazard rates have a rather complicated behaviour (*e.g.* even semesters exhibit spikes on the hazard of voluntary dropouts). In practice, not even large values of $\mathcal{P}$ would provide a good fit.

Here, the model in (5.8) is adopted for the analysis of the PUC dataset, using Bayesian methods to handle separation. We define the last period as $[t_0, \infty)$ [for fixed t_0, as in Scott and Kennedy, 2005], and period-indicators for time $t = 1$ are defined as cause-specific intercepts.

5.4 Bayesian PO competing risks regression

5.4.1 Prior specification

An alternative solution to the separation issue lies in the Bayesian paradigm, where an appropriate prior distribution for the period-indicators δ_{rt}'s can deal with a (quasi) complete separation of the outcomes [Gelman et al., 2008], allowing the extraction of sensible information from the data. The Jeffreys prior can be used for this purpose [Firth, 1993]. This is attractive when reliable prior information is not available. In a binary logistic case, the Jeffreys prior is proper and its marginals are symmetric with respect to the origin [Ibrahim and Laud, 1991; Poirier, 1994]. These properties have no easy generalization for the multinomial case, in which case an expression for the Jeffreys prior is very involved [Poirier, 1994]. Instead, Gelman et al. [2008] suggested weakly informative independent Cauchy priors for a re-scaled version of the regression coefficients. For this purpose, the binary variables linked to the period-indicators must be scaled to have mean zero, keeping the difference of 1 unit between their lower and upper values. When the outcome is binary, these Cauchy (as well as any Student t) priors have the same shape as the Jeffreys prior (symmetric with respect to the origin) but produce fatter tails [Chen et al., 2008]. The prior in Gelman et al. [2008] assumes that the regression coefficients fall within a restricted range. For the model in (5.8), this prior assigns small probabilities to large differences between the period-indicators δ_{rt}'s associated to the same event. Such a prior is convenient when the separation of the outcomes is related to a reduced sample size (where increasing the sample size will eventually eliminate the separation issue). The latter intuition is not applicable for the PUC dataset. For these data, the separation arises from structural restrictions (*e.g.* it is not possible to complete graduation requirements during the first periods of enrollment). Hence, large differences are expected for the δ_{rt}'s associated to the same event. In particular, it is intuitive that δ_{rt} will have a large negative value in those periods where events type r are very unlikely to be observed (inducing a nearly zero cause-specific hazard rate). Define $\delta_r = (\delta_{r1}, \ldots, \delta_{rt_0})'$. The following prior is suggested

$$\delta_r \sim \text{Cauchy}_{t_0}(0, \omega^2 I_{t_0}), \quad r = 1, \ldots, \mathcal{R} \tag{5.12}$$

where I_{t_0} denotes the identity matrix of dimension t_0. Equivalently,

$$\pi(\delta_r | \Lambda_r = \lambda_r) \sim \text{Normal}_{t_0}(0, \lambda_r^{-1}\omega^2 I_{t_0}), \quad \Lambda_r \sim \text{Gamma}(1/2, 1/2), \quad r = 1, \ldots, \mathcal{R}. \tag{5.13}$$

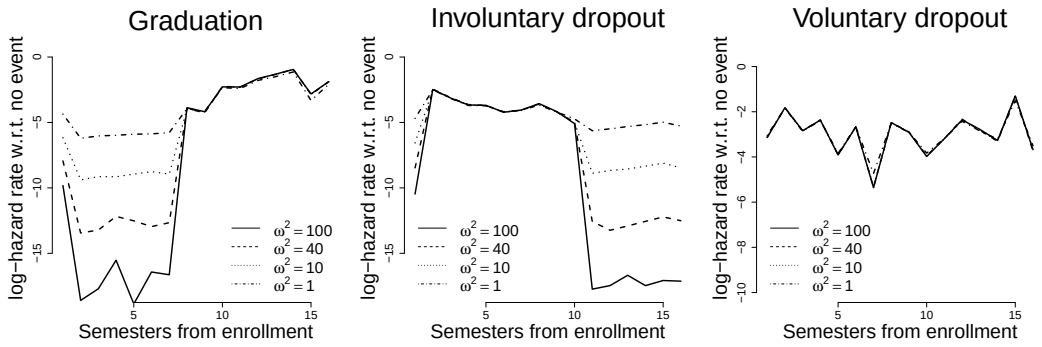


Figure 5.4: PUC dataset. For Chemistry students: estimated hazard rate of each competing event with respect to no event using the proportional odds model in (5.8) under $\delta_r \sim \text{Cauchy}_{t_0}(0, \omega^2 I_{t_0})$, $r = 1, 2, 3$. No covariates in use (model with only period-indicators).

This prior assigns non-negligible probability to large negative values of the δ_{rt}'s. Of course, an informative prior could also be used. However, it requires non-trivial prior elicitation (as it is not entirely clear a priori which δ_{rt}'s are affected by the separation issue, *i.e.* at which point do graduations start and where the dropout stops). Focusing on Chemistry students and using different values of ω^2 for the prior in (5.12), Figure 5.4 shows the induced trajectory for the posterior median of the log-hazard ratio for each event type with respect to no event being observed. For simplicity, covariates are excluded from these regressions. Choosing a value of ω^2 is not critical for those periods where the separation is not a problem (as the data is very informative). In contrast, ω^2 has a strong effect in those semesters where the separation occurs. Tight priors [as the ones in Gelman et al., 2008] are too conservative and produce non-intuitive results. Hence, large values of ω^2 seem more appropriate. How large is arbitrary but, after a certain threshold, its not too relevant in the hazard ratio scale (as the hazard ratio will be practically zero). For the analysis of the PUC dataset, $\omega^2 = 100$ is adopted.

The Bayesian model is completed using independent g-priors [Zellner, 1986] for the cause-specific vectors of covariates coefficients, *i.e.*

$$\beta_{(r)} \sim \text{Normal}_k(0_k, g_r(X'X)^{-1}), \qquad r = 1, \ldots, \mathcal{R}, \tag{5.14}$$

where 0_k denotes a null vector of dimension k. This is a standard choice in applied Bayesian analysis, especially when there is covariate uncertainty [*e.g.* Ley and Steel, 2009; Hanson et al., 2014]. This prior is invariant to scale transformations of the

covariates. Whereas the default version of this prior assumes $\{g_1, \ldots, g_{\mathcal{R}}\}$ as fixed quantities, this can have serious consequences in the posterior inference [Liang et al., 2008]. Some deterministic choices for the g_r's are discussed in Fernández et al. [2001] and Liang et al. [2008]. Instead, in the context of a binary logistic regression, Hanson et al. [2014] opts for eliciting g_r using averaged prior information (across different covariates configurations). Alternatively, a hyper-prior can be assigned to each g_r, inducing a hierarchical prior structure [Liang et al., 2008]. A review of several choices for this hyper-prior is provided in Ley and Steel [2012]. Based on theoretical properties and a simulation study (in a linear regression setting) they recommended the use of a benchmark Beta prior for which

$$\frac{g_r}{1+g_r} \sim \text{Beta}(b_1, b_2), \quad \text{or equivalently} \tag{5.15}$$

$$\pi(g_r) = \frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)} g_r^{b_1-1}(1+g_r)^{-(b_1+b_2)} \tag{5.16}$$

where $b_1 = 0.01 \max\{n, k^2\}$ and $b_2 = 0.01$. The latter hyper-g prior is adopted for the regression coefficients throughout the analysis of the PUC dataset.

5.4.2 Markov chain Monte Carlo implementation

Fitting a multinomial (or binary) logistic regression is not straightforward. There is no conjugate prior and sampling from the posterior distribution of the regression coefficients is cumbersome [Holmes and Held, 2006]. The Bayesian literature normally opts for alternative representations of the likelihood function for this model. For instance, Forster [2010] exploits the relationship between a multinomial logistic regression and a Poisson generalized linear model. Following the idea in Albert and Chib [1993], Holmes and Held [2006] adopt a hierarchical structure where the logistic link is represented as a scale mixture of normals (in the same fashion as the SMLN representation of the log-logistic model introduced in Chapter 3). Alternatively, Frühwirth-Schnatter and Frühwirth [2010] approximated the logistic link via a finite mixture of normal distributions, suggesting that 10 components provides a good approximation. Here, the methodology proposed in Polson et al. [2013] is implemented. As in Holmes and Held [2006] and Frühwirth-Schnatter and Frühwirth [2010], this employs a hierarchical representation of the multinomial logistic likelihood. For a binary logistic model with observations $\{y_{it} : i = 1, \ldots, n, t = 1, \ldots, t_i\}$ ($y_{it} = 1$ if the event is observed at time t for subject i, $y_{it} = 0$ otherwise), the key

result in Polson et al. [2013] is that

$$\frac{[\,e^{z_i'\beta^*}]^{y_{it}}}{e^{z_i'\beta^*}+1} \propto e^{\kappa_{it} z_i'\beta^*} \int_0^\infty \exp\{-\eta_{it}(z_i'\beta^*)^2/2\} f_{PG}(\eta_{it}|1,0)\, d\eta_{it}, \tag{5.17}$$

where z_i is a vector of covariates associated with individual i, β^* is a vector of regression coefficients, $\kappa_{it} = y_{it} - 1/2$ and $f_{PG}(\cdot|1,0)$ denotes a Polya-Gamma density with parameters 1 and 0, which has Laplace transform $\mathrm{E}_\eta(\, e^{-\eta s}) = \cosh^{-1}(\sqrt{s/2})$. In terms of the model in (5.5), z_i includes x_i and the binary indicators linked to the δ_t's. Thus, $\beta^* = (\delta_1, \ldots, \delta_{t_0}, \beta')'$.

The result in (5.17) can be used to construct a Gibbs sampling scheme for the multinomial logistic model along the lines of Holmes and Held [2006]. Now let $0, 1, \ldots, \mathcal{R}$ be the possible values for observations y_{it} associated with regression coefficients $\beta^*_{(1)}, \ldots, \beta^*_{(\mathcal{R})}$. Given $\beta^*_{(1)}, \ldots, \beta^*_{(r-1)}, \beta^*_{(r+1)}, \ldots, \beta^*_{(\mathcal{R})}$, the "conditional" likelihood function for $\beta^*_{(r)}$ is proportional to

$$\prod_{i=1}^{n}\prod_{t=1}^{t_i} \frac{\left[\exp\{z_i'\beta^*_{(r)} - C_{ir}\}\right]^{I(y_{it}=r)}}{1+\exp\{z_i'\beta^*_{(r)} - C_{ir}\}}, \quad \text{where} \quad C_{ir} = \log\left(1 + \sum_{r^*\neq r} \exp\{z_i'\beta^*_{(r^*)}\}\right). \tag{5.18}$$

Assume $\beta^*_{(r)} \sim \mathrm{Normal}_{t_0+k}(\mu_r, \Sigma_r)$, $r = 1, \ldots, \mathcal{R}$ and define $B^* = \left\{\beta^*_{(1)}, \ldots, \beta^*_{(\mathcal{R})}\right\}$. Using (5.17) and (5.18), a Gibbs sampler for the multinomial logistic model is defined through the following full conditionals for $r = 1, \ldots, \mathcal{R}$

$$\beta^*_{(r)}|\eta_r, \beta^*_{(1)}, \ldots, \beta^*_{(r-1)}, \beta^*_{(r+1)}, \ldots, \beta^*_{(\mathcal{R})}, y_{11} \ldots, y_{nt_n} \sim \mathrm{Normal}_{t_0+k}(m_r, V_r), \tag{5.19}$$

$$\eta_{itr}|B^* \sim \mathrm{PG}(1, z_i'\beta^*_{(r)} - C_{ir}), \;\; t = 1, \ldots, t_i, i = 1, \ldots, n, \tag{5.20}$$

defining $Z = (z_1 \otimes \iota'_{t_1}, \ldots, z_n \otimes \iota'_{t_n})'$, $\eta_r = (\eta_{11r}, \ldots, \eta_{nt_nr})'$, $D_r = \mathrm{diag}\{\eta_r\}$, $V_r = (Z'D_rZ + \Sigma_r^{-1})^{-1}$, $m_r = V_r(Z'\kappa_r + \Sigma_r^{-1}\mu_r)$, $\kappa_r = (\kappa_{11r}, \ldots, \kappa_{nt_nr})'$ and $\kappa_{itr} = \mathrm{I}_{\{y_{it}=r\}} - 1/2 + \eta_{itr}C_{ir}$ (where $\mathrm{I}_A = 1$ if A is true, 0 otherwise). The previous algorithm applies to (5.8) using $\beta^*_{(r)} = (\delta_r', \beta'_{(r)})'$ and defining z_i in terms of binary variables related to the δ_{rt}'s and the covariates x_i. Extra steps are required to accommodate the adopted prior, which is a product of independent multivariate Cauchy and hyper-g prior components. Both components can be represented as a scale mixture of normal distributions (see (5.13) and (5.14)). Hence, conditional on $\Lambda_1, \ldots, \Lambda_{\mathcal{R}}, g_1, \ldots, g_{\mathcal{R}}$, the sampler above applies. In addition, at each iteration,

Λ_r's and g_r's are updated using the full conditionals.

$$\Lambda_r|\delta_r \sim \text{Gamma}\left(\frac{t_0+1}{2}, \frac{\delta_r'\delta_r}{2\omega^2}\right), \quad r = 1, \ldots, \mathcal{R}, \tag{5.21}$$

$$g_r|\beta_r \sim g_r^{-k/2} \exp\left\{-\frac{\beta_r' X' X \beta_r}{2g_r}\right\} \pi(g_r), \quad r = 1, \ldots, \mathcal{R}. \tag{5.22}$$

An adaptive Metropolis-Hastings step [see Section 3 in Roberts and Rosenthal, 2009] is implemented for (5.22).

5.4.3 Bayesian variable selection and model averaging

A key aspect of the analysis is to select the relevant covariates to be included in the model. A popular approach is to choose the model with the best performance (in terms of DIC, PsML, BF or some other criteria). However, in a Bayesian setting, a natural way to deal with model uncertainty is to use the posterior probabilities associated to each model. Denote by k^* the number of available covariates (k^* and the number of regression coefficients k do not necessarily match because categorical predictors with more than two levels introduce more than one regression coefficient). Let $M_1, \ldots, M_{\mathcal{M}}$ be the collection of all $\mathcal{M} = 2^{k^*}$ competing models (if a discrete covariate is included, all its levels are as well incorporated in the model). Given observed times T_{obs} and event types R_{obs}, posterior probabilities for these models are defined via Bayes theorem as

$$\pi(M_m|T_{obs}, R_{obs}) = \frac{L(T_{obs}, R_{obs}|M_m)\pi(M_m)}{\sum_{m^*=1}^{\mathcal{M}} L(T_{obs}, R_{obs}|M_{m^*})\pi(M_{m^*})}, \tag{5.23}$$

where $L(T_{obs}, R_{obs}|M_m)$, $m = 1, \ldots, \mathcal{M}$ are the marginal likelihoods related to each model (as in (1.4), integrating out model parameters) and $\pi(M_1), \ldots, \pi(M_{\mathcal{M}})$ represent prior beliefs about the model space with $\sum_{m=1}^{\mathcal{M}} \pi(M_m) = 1$. These marginal likelihoods can be estimated using the bridge sampler described in Subsection 1.2.4. A uniform prior for the model space is defined as

$$\pi(M_m) = \frac{1}{\mathcal{M}}, \quad m = 1, \ldots, \mathcal{M}. \tag{5.24}$$

Alternatively, a prior for the model space can be specified through the covariate-inclusion indicators

$$\gamma_j = \begin{cases} 1, & \text{if covariate } j \text{ is included;} \\ 0, & \text{otherwise.} \end{cases} \tag{5.25}$$

for $j = 1, \ldots, k^*$. Independent Bernoulli(θ) priors can be assigned to the γ_j's. For $\theta = 1/2$, the induced prior coincides with the uniform prior in (5.24). As discussed in Ley and Steel [2009], assigning an hyper prior for θ provides more flexibility and reduces the influence of the prior on posterior inference. A Beta(a_1, a_2) prior for θ leads to the so-called Binomial-Beta prior on the number of included covariates $C = \sum_{j=1}^{k^*} \gamma_j$ [Bernardo and Smith, 2000, p.117]. If $a_1 = a_2 = 1$ (uniform prior for θ), the latter induces a uniform prior for C, *i.e.*

$$\pi(C = c) = \frac{1}{k^* + 1}, \quad c = 0, \ldots, k^*. \tag{5.26}$$

If a single model concentrates a particularly high posterior probability, that model could be chosen. Otherwise, if the model posterior probabilities are not concentrated but similar amounts of non-negligible probability are assigned to several models, Bayesian Model Averaging (BMA) provides an attractive solution. Good surveys about this topic are provided in Hoeting et al. [1999] and Chipman et al. [2001]. Instead of selecting a single model, BMA defines a model via a mixture of all $\mathcal{M}$ possible models, where mixture weights are given by the posterior probabilities of each model. Let Δ be a quantity of interest (*e.g.* one of the regression coefficients). In BMA, the posterior distribution of Δ is given by

$$P(\Delta|T_{obs}, R_{obs}) = \sum_{m=1}^{\mathcal{M}} P_m(\Delta|T_{obs}, R_{obs})\pi(M_m|T_{obs}, R_{obs}), \tag{5.27}$$

where $P_m(\Delta|T_{obs}, R_{obs})$ denotes the posterior distribution of Δ for a given model M_m. In particular, the posterior distribution of each β_{rj} is given by a point mass at zero (with mass equal to the probability of not including the j-th covariate) and a continuous component (defined as a mixture over the posterior distributions of β_{rj} given each model where the corresponding covariate is included). BMA constitutes the formal Bayesian treatment of model uncertainty and leads to a better predictive performance than choosing a unique model [Raftery et al., 1997; Fernández et al., 2001].

5.5 Empirical results for the PUC data

The PUC dataset is analyzed using the model in (5.8) and the algorithm described in Section 5.4.2. As indicated in Section 5.2, the analysis is carried out independently for each programme, focusing on some of the science programmes for which the rates of dropout and/or late graduations are normally higher. For all programmes, 8

covariates are available (see Table 5.2), inducing $2^8 = 256$ possible models (using the same covariates for each cause-specific hazard). Selection scores cannot be directly compared across admission years (as the test varies from year to year). Hence, in order to obtain more meaningful results, the selection score is replaced by an indicator of being on the top 10% of the enrolled students (for each program and admission year). The following regression coefficients are defined for each cause (the sub-index r is omitted for ease of notation): β_1 (sex: female), β_2 (region: metropolitan area), β_3 (parents' education: with degree), β_4 (high school: private), β_5 (high school: subsidized private), β_6 (funding: scholarship only), β_7 (funding: scholarship and loan), β_8 (funding: loan only), β_9 (ranking: 10% superior), β_{10} (application preference: first) and β_{11} (gap after high school graduation: yes). All models contain an intercept and $t_0 - 1 = 15$ period indicators. For all models, the total number of iterations is 200,000. In the following, results are presented on the basis of 1,000 draws (after a burn-in of 50% of the initial iterations and thinning). Trace plots and the usual convergence criteria strongly suggest a good mixing and the convergence of the chains (not reported).

Figure 5.5 displays the trajectory of the cause-specific hazard rates for all possible 256 models, corresponding to the reference case (where $x_i = 0\iota_k$). Differences between these estimations are mostly related to changes in the intercept, which is obviously affected by the removal or addition of covariates. In particular, the first row of panels in Figure 5.5 roughly recovers the same patterns as in Figure 5.3, suggesting that these estimates are dominated by the data and not by the prior. Some similarities appear between these programmes. For example, the highest risk of involuntary dropout is observed by the end of the second semester from enrollment. This is not entirely surprising as, in the science programmes, students often have a bad performance during their first year of studies. In addition, during the 4 first years of enrollment, the hazard rate associated to voluntary dropouts has spikes located at even semesters. Again, this result is intuitive. Withdrawing at the end of the academic year allows students to re-enroll at a different programme without having a gap in their academic careers. In terms of graduations, mild spikes are located at the official duration of the programmes. Nonetheless, for these programmes, the highest hazards of graduation occur about 4 semesters after the official duration. The spikes at the last period are due to a cumulative effect (as δ_{rt_0} represents the period $[t_0, \infty)$).

Figure 5.6 summarizes marginal posterior inference under all possible 256 models for Chemistry, Mathematics and Statistics, and Physics (the sub-index r is omitted for ease of notation). Across all models, the median effects normally

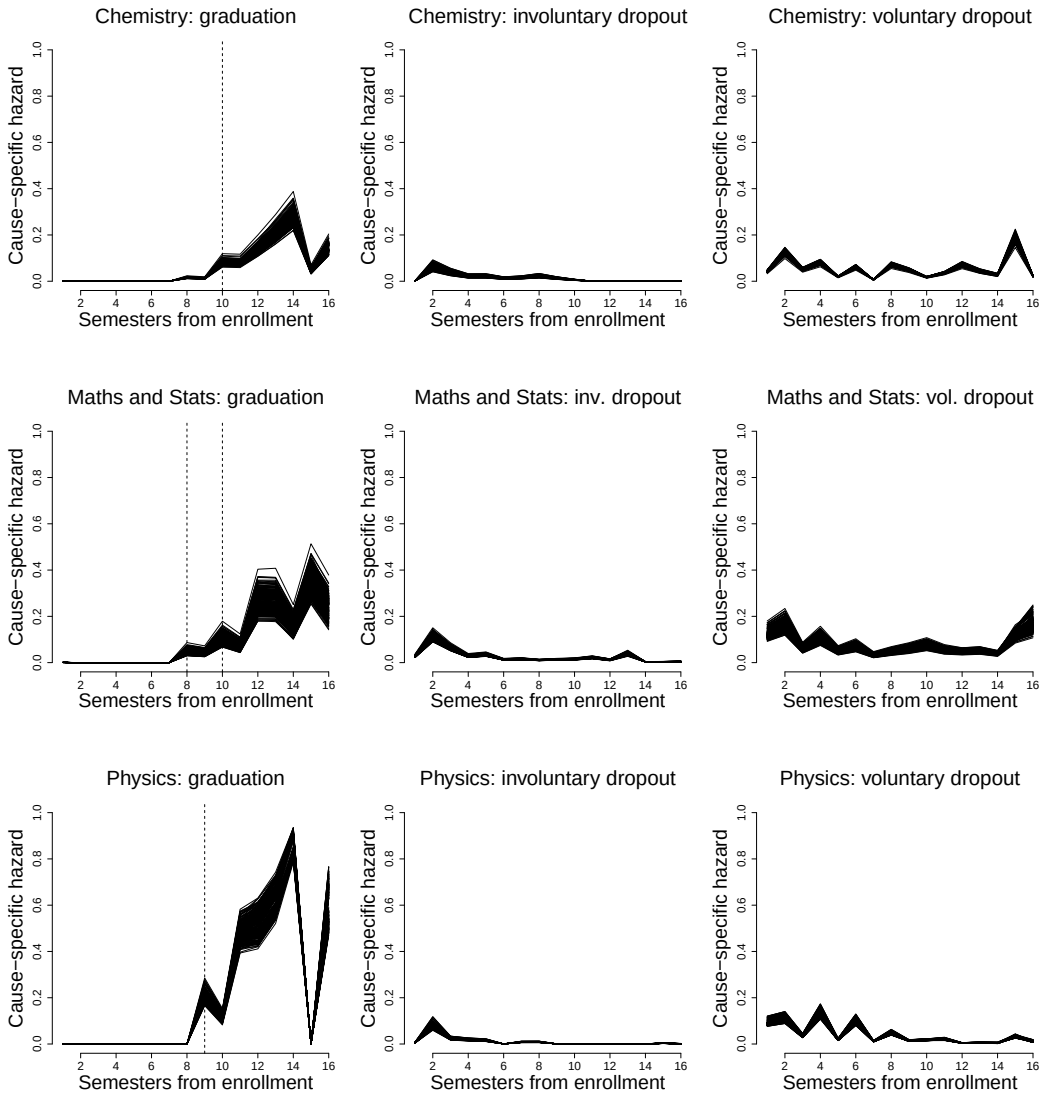


Figure 5.5: PUC dataset. Spaghetti plot of baseline cause-specific hazards across the 256 possible models. For graduation hazards, dashed vertical lines are located at the official duration of the programme. Two lines are displayed in the Mathematics and Statistics programme because students following the Statistics track require two additional semesters in order to obtain a professional degree.

retain the same sign (within the same degree programme). Only covariates with smaller effects display estimates with opposite signs (*e.g.* the coefficient related to sex, β_1, for Chemistry students). Nonetheless, the actual effect values do not coincide across different models. In general, students who applied as a first preference to these degrees graduated more and faster (see estimations of β_{10}). These students

also exhibit a lower rate of voluntary dropout, which might be linked to a higher motivation about the programme at which they are enrolled. Whether or not the student had a gap between high school graduation and university admission also has a strong influence on the academic outcomes for these programmes. These gaps can, for example, correspond to periods in which the student was preparing for the admission test (after a low score in a previous year) or enrolled at a different programme (of the PUC or other institutions). Overall, this gap induces less and slower graduations for these programmes. In addition, at each semester, students with a gap before university enrollment have a higher risk of being expelled from these degrees. In line with the descriptive analysis presented in Section 5.2, the effect of the covariates are not homogeneous across the analyzed programmes. Whereas the effect of the student's sex (β_1) is almost negligible in Chemistry, female students in Mathematics and Statistics present a higher hazard of graduation and lower risk of being expelled at all semesters.

Table 5.5: PUC dataset.Top 3 models in terms of DIC and PsML for some degree programmes (ticks indicate covariate inclusion).

Programme	DIC	Sex	Region	Parents	School	Funding	Top 10%	Pref.	Gap
	1915.23		✓				✓	✓	✓
Chemistry	1915.54						✓	✓	✓
	1915.64							✓	✓
Mathematics	3117.89	✓			✓		✓	✓	✓
and	3119.95	✓	✓		✓		✓	✓	✓
Statistics	3120.06	✓		✓	✓		✓	✓	✓
	1091.86	✓			✓			✓	✓
Physics	1093.23	✓		✓	✓			✓	✓
	1093.40	✓	✓		✓			✓	✓
Programme	log-PsML	Sex	Region	Parents	School	Funding	Top 10%	Pref.	Gap
	-962.76							✓	✓
Chemistry	-963.77						✓	✓	✓
	-963.81			✓				✓	✓
Mathematics	-1563.44	✓			✓		✓	✓	✓
and	-1564.27	✓		✓	✓		✓	✓	✓
Statistics	-1564.46	✓	✓		✓		✓	✓	✓
	-550.78	✓			✓			✓	✓
Physics	-552.79	✓	✓		✓			✓	✓
	-553.10	✓	✓					✓	✓

Table 5.5 summarizes Bayesian model comparison in terms of DIC and PsML. For the analyzed programmes, both criteria point in the same direction, suggesting that the most important covariates are the application preference and the gap indicator (associated to the effects β_{10} and β_{11}, respectively). Sex (related to β_1) and the high school type (represented by β_4 and β_5) are added to this list in case of

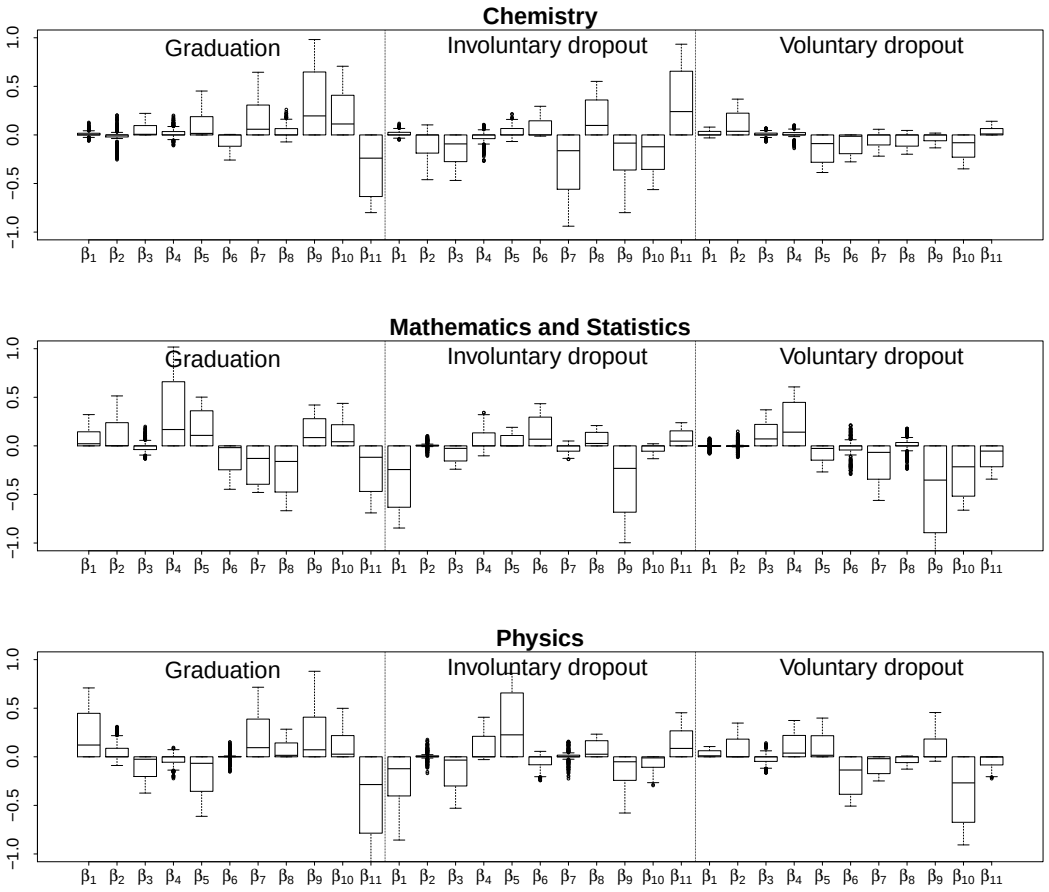


Figure 5.6: PUC dataset. Boxplot of estimated posterior medians for covariate effects across the 256 possible models. The sub-index r is omitted for ease of notation. When a covariate is not included in the model, the corresponding posterior medians are replaced by zero.

Mathematics and Statistics students and the ones enrolled in Physics. The selection score indicator β_9 (top 10%) also appears to have some relevance (specially in case of Mathematics and Statistics). As shown in Table 5.6, similar conclusions follow from the posterior distribution on the model space as those models with the highest posterior probabilities often include the same covariates suggested by DIC and PsML. In fact, for these programmes, those models with the highest posterior probabilities often include the same covariates that were suggested according to DIC an PsML. One difference is that for two programmes there is more support for the null model (the model without covariates where only the δ_{rt}'s are included to model the baseline hazard). The choice between the priors in (5.24) and (5.26) on the model space can

Table 5.6: PUC dataset. Top 3 models with highest posterior probability for some degree programmes (ticks indicate covariate inclusion).

Prior	Programme	Prob.	Sex	Region	Parents	School	Funding	Top 10%	Pref.	Gap
(5.24)	Chemistry	0.270		✓				✓	✓	✓
		0.238							✓	✓
		0.193		✓	✓				✓	✓
	Mathematics and Statistics	0.942	✓		✓	✓	✓	✓	✓	✓
		0.036	✓		✓	✓		✓	✓	✓
		0.014	✓	✓		✓		✓	✓	✓
	Physics	0.268								
		0.150	✓		✓			✓	✓	✓
		0.054	✓			✓			✓	✓
(5.26)	Chemistry	0.354								
		0.259							✓	✓
		0.117		✓				✓	✓	✓
	Mathematics and Statistics	0.982	✓		✓	✓	✓	✓	✓	✓
		0.011	✓		✓	✓		✓	✓	✓
		0.004	✓	✓		✓		✓	✓	✓
	Physics	0.937								
		0.009	✓		✓			✓	✓	✓
		0.007							✓	✓

have a strong influence on posterior inference. As discussed in Ley and Steel [2009], the prior in (5.26) downweighs models with size around $k^*/2 = 4$ (with priors odds in favour of the null model or the model with all 8 covariates versus a model with 4 covariates equal to 70) and this is accentuated in Physics, where the best model under both priors is the null model and the second best model has $k^* = 5$, so that posterior model probabilities differ substantially between priors (see Table 5.6). In contrast, the choice between these priors has less effect in Maths and Stats, where the best models are of similar sizes. In a BMA framework, posterior probabilities of covariate inclusion are displayed in Table 5.7. For these programmes, the highest posterior probabilities of inclusion relate to the application preference and the gap indicator (for both priors on the model space). As expected, results vary across programmes. For Mathematics and Statistics, there is strong evidence in favour of including all available covariates with the exception of the region of residence. In contrast, under both priors the model suggests that sex, high school type and the source of funding have no major influence on the academic outcomes of Chemistry students. For Physics (and to some extent for Chemistry) interesting models tend to be small and then the (locally) higher model size penalty implicit in prior (5.26) substantially reduces the inclusion probabilities of all covariates. For Maths and Stats, the best models are rather large and the prior (5.26) then favours models that are even larger, leading to very similar inclusion probabilities.

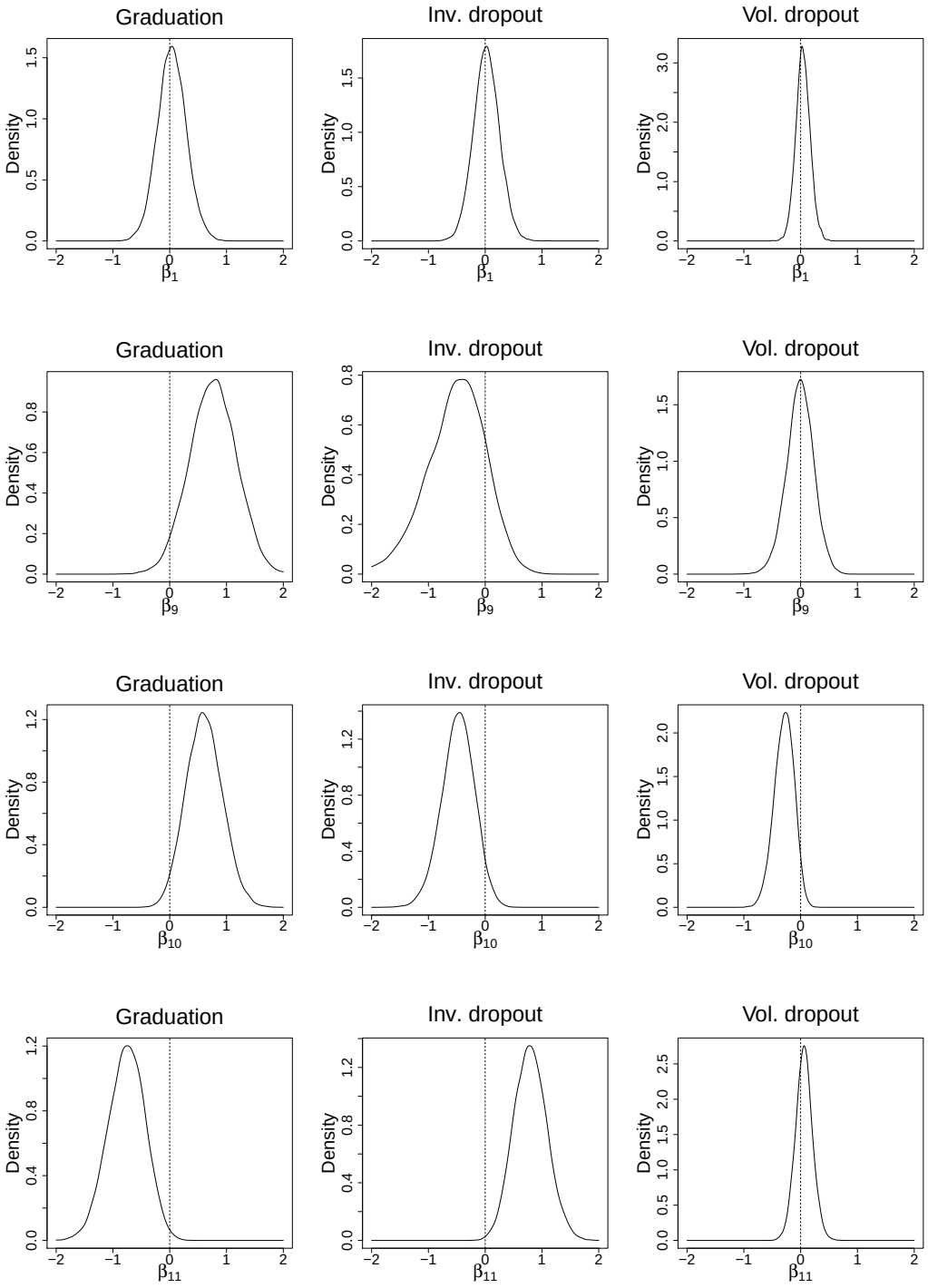


Figure 5.7: PUC dataset. For Chemistry students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: sex (β_1), ranking (β_9), preference (β_{10}) and gap (β_{11}). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

Table 5.7: PUC dataset. Posterior probability of variable inclusion under priors (5.24) and (5.26) on the model space.

Programme	Prior	Sex	Region	Parents	H. school	Funding	Top 10%	Pref.	Gap
Chemistry	(5.24)	0.08	0.52	0.28	0.08	0.07	0.50	0.93	0.99
	(5.26)	0.06	0.23	0.15	0.04	0.06	0.27	0.61	0.65
Maths. and	(5.24)	0.99	0.02	0.98	1.00	0.95	1.00	1.00	1.00
Statistics	(5.26)	1.00	0.01	0.99	1.00	0.98	1.00	1.00	1.00
Physics	(5.24)	0.49	0.25	0.37	0.31	0.11	0.27	0.71	0.62
	(5.26)	0.04	0.02	0.03	0.03	0.01	0.02	0.06	0.05

The posterior distribution of each β_{rj} is given by a point mass at zero (equal to the probability of excluding the j-th covariate) and a continuous component (a mixture over the posterior distributions of β_{rj} given each model where the corresponding covariate is included). Figure 5.7 displays the continuous component of the posterior distribution of some selected regression coefficients for the Chemistry programme under the prior in (5.24). The first row shows that the marginal densities of the effects related to sex are concentrated around zero. This is in line with the results in Table 5.7, where both priors on the model space indicate a low posterior inclusion probability for sex. In contrast, the third row in Figure 5.7 suggests a clear effect of the application preference on the three possible outcomes (positive for graduations and negative for both types of dropout). This agrees with a high posterior probability of inclusion and to put the magnitude of the effect into perspective, the odds for outcome $r = 1, 2, 3$ versus no event are multiplied by a factor $\exp(\beta_{r\ 10})$ if Chemistry is the student's first preference. A similar situation is observed for the selection score indicator (see second row in Figure 5.7). In this case, those students with scores in the top 10% graduate more and faster and are affected by less (and slower) involuntary dropouts. Nonetheless, this score indicator has no major influence on whether a student withdraws. Finally, for the gap indicator, we also notice a clear effect on graduations and involuntary dropouts, which has the opposite direction to that of the score indicator.

5.6 Concluding remarks

The modelling of university outcomes (graduation or dropout) is not trivial. In fact, as discussed in Willett and Singer [1991], the usual approach where the dropout is treated as a dichotomous process is not appropriate and a temporal component must be incorporated into the model. In this article, a simple but flexible competing risks survival model is employed for this purpose. This is based in the Proportional Odds

model introduced in Cox [1972] and can be estimated by means of a multinomial logistic regression. The suggested sampling model has been previously employed in the context of university outcomes, but the structure of typical university outcome data precludes a maximum likelihood analysis. However, we use a Bayesian setting, where an appropriate prior distribution allows the extraction of sensible information from the data. Adopting a hierarchical structure allows for the derivation of a reasonably simple MCMC sampler for inference. The proposed methodology is applied to a dataset on undergraduate students enrolled in the Pontificia Universidad Católica de Chile (PUC) over the period 2000-2011.

As illustrated in Sections 5.2 and 5.5, there are strong levels of heterogeneity between different programmes of the PUC. Hence, building a unique model for the whole university is not recommended. The methodology presented here can be applied to all programmes of the PUC. For brevity, this Chapter only presents the analysis of three science programmes for which late graduations and dropouts are a major issue, but the methodology presented here can be applied to all programmes. We formally consider model uncertainty in terms of the covariates included in the model. For the analyzed programmes, all the variable selection criteria (DIC, PsML and Bayes factors) tend to indicate similar results. However, in view of the posterior distribution on the model space, choosing a single model is not generally advisable and BMA provides more meaningful inference. The preference with which the student applied to the programme plays a major role in terms of the length of enrollment and its associated academic outcome for the three programmes under study. In addition, and perhaps surprisingly, having a gap between high school graduation and university admission is also found to be one of the most relevant covariates (but with the reverse effect of the preference indicator). The performance in the selection test is also generally an important determinant. Other factors, such as sex and the region of residence, only appear to matter for some of the programmes.

Chapter 6

Conclusions and further work

> "I cannot fix on the hour, or the spot, or the look or the words, which laid the foundation. It is too long ago. I was in the middle before I knew that I had begun".
> Jane Austen
> *Pride and Prejudice*

This thesis covered theoretical and practical aspects of Bayesian inference and survival analysis, which is a powerful tool for the analysis of time-to-event data. In conventional survival models, observations represent the time until a unique event of interest occurs, which are identically distributed (up to a set of known covariates) realizations of positive-valued random variables generated by a "thin-tailed" distribution. These assumptions are frequently not satisfied by real applications. In particular, the main focus of this dissertation is the development and implementation of more flexible models that can deal with unobserved heterogeneity (which cannot be captured by covariates) and multiple competing events.

The models presented in Chapters 2 and 3 deal with unobserved heterogeneity in a natural manner, where an individual-specific random effect accounts for variations in the survival distribution that are not related to changes in the available covariates. As illustrated by three different medical applications, unobserved heterogeneity (possibly related to outlying observations) is not a rare feature in real datasets. Ignoring this component can have serious consequences for posterior inference. A key feature of these models is that estimation is more robust to the presence of anomalous observations. This constitutes a major advantage in a context where sample size is often small and collecting observations is not straightforward (*e.g.* in a clinical trial where a new test treatment is being studied). Among others, the following extensions could be considered.

- The framework introduced in Chapter 2 does not rely on a closed-form expression for the marginal density (with the mixing parameters integrated out). Nonetheless, all the examples of SMLN models explored in Chapter 3 relate to a closed-form of the marginal lifetime density and have been already studied in the existing literature. New distributions for the analysis of survival data can be generated by varying the mixing distribution $P_{\Lambda_i}(\cdot|\theta)$. In such a case, the Jeffreys prior (and its variations) do not necessarily generate a proper prior for θ, precluding the use of Bayes factors for model comparison. As in the RMW family, a proper prior for θ can be elicited via the coefficient of variation of the survival times c_v. However, as c_v depends on σ^2 (see Theorem 2), Theorems 4 and 6 will no longer cover posterior propriety.

- Another, and perhaps obvious, course of action is to extend the range of underlying models. One example would be the Birnbaum-Saunders distribution [Birnbaum and Saunders, 1969]. Maximum likelihood inference for mixtures of Birnbaum-Saunders distributions based on scale mixtures of normals distributions has been studied in Barros et al. [2008], Balakrishnan et al. [2009] and Patriota [2012]. However, in this context, benchmark Bayesian inference is challenging because the Jeffreys-style priors do not lead to a proper posterior distribution when using a single Birnbaum-Saunders distribution with no mixture [Xu and Tang, 2011]. A second candidate for the underlying model is the log-skew normal distribution [Azzalini et al., 2003]. This is a direct extension of the SMLN family for which a skewness parameter introduces more flexibility. A special case of this family is the log-skew-t model proposed in Azzalini et al. [2003].

A substantive real-life problem motivates the second part of this thesis. Using a discrete-time competing risks model, Chapter 5 presents an analysis of university outcomes for undergraduate students of the Pontificia Universidad Católica de Chile. The main focus of the study is the identification of potential predictors for the length of stay at university and its associated outcome. A simple but flexible model is employed for this purpose. This allows the extraction of sensible information from the data without making strong assumptions about the survival distribution. The proposed model does not incorporate all features of this complex dataset yet it provides a better understanding of the problem and constitutes a foundation for future related research. Some potential extensions in this context are listed below.

- An obvious extension of the model presented here is to allow for different covariates in the modelling of the three risks within the same programme. This

would substantially increase the number of models in the model space ($\mathcal{M} = 2^{3\times 8} = 16,777,216$ in this case), so we would need to base our inference on posterior model probabilities on sampling rather than complete enumeration. This can easily be implemented by extending the MCMC sampler to the model index and using *e.g.* Metropolis-Hastings updates based on data augmentation such as in Holmes and Held [2006] or applications of the Automatic Generic sampler described by Green [2003].

- It is not possible for the university to record all covariates that can have an effect on academic outcomes. In fact, diverse aspects such as motivational levels and life events (*e.g.* pregnancies and financial hardships) can have a direct implication in whether or not a student completes the graduation requirements for a degree. As discussed in Chapter 2, ignoring this unobserved heterogeneity can have serious consequences in posterior inference. As in Chapters 2 and 3, a natural solution to this problem is to incorporate an individual-specific random effect into the model.

- For the analysis presented in Chapter 5, periods of temporary withdrawal were considered as part of the total length of stay at university. Instead, multi-state models [Meira-Machado et al., 2009] can be employed in order to formally deal with these stopovers.

- Finally, an alternative approach for modelling the PUC dataset is given by cure models (see Section 2.6). In such a case, the cause-specific hazard rate can be assigned a positive probability of being equal to zero. The latter can directly incorporate into the model structural restrictions of graduations and dropout.

Appendix A

Proofs

Proposition 1. The likelihood contribution of censored observations is a factor bounded in $[0, 1]$. Hence, the likelihood of the complete sample $L(t|\psi; c)$ is bounded above by the likelihood of the non-censored observations $L(t_o|\psi)$. Using (1.4), the same applies for the respective marginal likelihoods $L(t; c)$ and $L(t_o)$. Therefore, a sufficient condition for existence of the $\pi(\psi|t; c)$ is $L(t_o) < \infty$. □

Theorem 1. Define $I(s) = \int L_T((s, t_c)|\psi; c)\pi(\psi)\, d\psi$, where the integral is over the support of ψ. Based on the sample of set observations, the posterior distribution of ψ exists if and only if $\int_E I(s)\, ds$ is finite. As E is bounded $\int_E I(s)\, ds$ is bounded as long as $I(\cdot)$ is finite except on a set of zero Lebesgue measure. □

Theorem 2. The result is a direct consequence of using Fubini's theorem on the integral $\int_0^\infty t_i f(t_i|\mu, \sigma^2, \theta)\, dt_i$ (after replacing $f(t_i|\mu, \sigma^2, \theta)$ by its SMLN representation). □

Theorem 3. Taking the negative expectation of the second derivatives of the log likelihood, the expressions $k_1(\theta)$, $k_2(\theta)$, $k_3(\theta)$ and $k_4(\theta)$ are given by

$$k_1(\theta) = n\mathrm{E}_{T_i}\left(\left[\frac{\log(t_i) - x_i'\beta}{\sigma}\right]^2\left[\frac{E_i}{f(t_i)}\right]^2\right), \tag{A.1}$$

$$k_2(\theta) = \frac{n}{4}\left[\mathrm{E}_{T_i}\left(\left[\frac{\log(t_i) - x_i'\beta}{\sigma}\right]^4\left[\frac{E_i}{f(t_i)}\right]^2\right) - 1\right], \tag{A.2}$$

$$
\begin{aligned}
k_3(\theta) &= \frac{n}{2}\mathrm{E}_{T_i}\left(\frac{\left[\frac{\log(t_i)-x_i'\beta}{\sigma}\right]^2 E_i \int_0^\infty f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\lambda_i}\right)\frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta)}{f^2(t_i)}\right) \\
&- \frac{1}{2}\sum_{i=1}^{n}\int_0^\infty \frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta), \qquad \text{(A.3)} \\
k_4(\theta) &= n\mathrm{E}_{T_i}\left(\left[\frac{\int_0^\infty f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\lambda_i}\right)\frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta)}{f(t_i)}\right]^2\right) \\
&- \sum_{i=1}^{n}\int_0^\infty \frac{d^2}{d\theta^2}dP_{\Lambda_i}(\lambda_i|\theta), \qquad \text{(A.4)}
\end{aligned}
$$

where $E_i = \mathrm{E}_{\Lambda_i}\left(\Lambda_i f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\Lambda_i}\right)\right)$. □

Corollary 1. The proof follows directly from Theorem 3 using the structure of the determinant of the FIM and its sub-matrices. □

Theorem 4. Define $t = (t_1,\ldots,t_n)'$ and $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. After some algebraic manipulation, $f_T(t)$ is proportional to

$$
\int_{\mathbb{R}^k}\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbf{R}_+^n}\frac{\pi(\theta)\prod_{i=1}^n\lambda_i^{\frac{1}{2}}}{(\sigma^2)^{\frac{n}{2}+p}\prod_{i=1}^n t_i}\, e^{-\frac{1}{2\sigma^2}\left[(\beta-a)'A(\beta-a)+S^2(D,y)\right]}\prod_{i=1}^n dP(\lambda_i|\theta)\,d\beta\,d\sigma^2\,d\theta, \quad \text{(A.5)}
$$

where $y = (\log(t_1),\ldots,\log(t_n))'$, $A = X'DX$, $a = A^{-1}X'Dy$ and $S^2(D,y) = y'Dy - y'DX(X'DX)^{-1}X'Dy$. Provided that $t_i \neq 0$ for all $i \in \{1,\ldots,n\}$, using Fubini's theorem for the integral (A.5) and integrating first with respect to β, $f_T(t)$ is proportional to

$$
\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbb{R}_+^n}(\sigma^2)^{-\frac{n+2p-k}{2}}\frac{\prod_{i=1}^n\lambda_i^{\frac{1}{2}}}{\sqrt{\det(X'DX)}}\, e^{-\frac{S^2(D,y)}{2\sigma^2}}\pi(\theta)\prod_{i=1}^n dP(\lambda_i|\theta)\,d\theta\,d\sigma^2. \quad \text{(A.6)}
$$

After integrating with respect to σ^2, it follows that $f_T(t)$ is proportional to

$$
\int_{\Theta}\int_{\mathbb{R}_+^n}\prod_{i=1}^n\lambda_i^{\frac{1}{2}}(\det(X'DX))^{-\frac{1}{2}}[S^2(D,y)]^{-\frac{n+2p-k-2}{2}}\pi(\theta)\prod_{i=1}^n dP(\lambda_i|\theta)\,d\theta, \quad \text{(A.7)}
$$

as long as $n+2p-k-2>0$ and $S^2(D,y)>0$. If $n>k$ we know that $S^2(D,y)>0$ a.s. and the first condition is certainly satisfied when $p \geq 1$. Following Lemma 1 in

Fernï¿½ndez and Steel [1999], $f_T(t)$ has upper and lower bounds proportional to

$$\int_\Theta \int_{0<\lambda_1<\cdots<\lambda_n<\infty} \prod_{i\notin\{m_1,\ldots,m_k\}} \lambda_i^{\frac{1}{2}} \lambda_{m_{k+1}}^{-\frac{n+2p-k-2}{2}} \pi(\theta) \prod_{i=1}^{n} dP(\lambda_i|\theta)\, d\theta, \tag{A.8}$$

where

$$\prod_{i=1}^{k} \lambda_{m_i} \equiv \max\left\{ \prod_{i=1}^{k} \lambda_{l_i} : \det\left(x_{l_1} \cdots x_{l_k}\right) \neq 0, l_1, \ldots l_k \leq n \right\}, \tag{A.9}$$

$$\prod_{i=1}^{k+1} \lambda_{m_i} \equiv \max\left\{ \prod_{i=1}^{k+1} \lambda_{l_i} : \det\begin{pmatrix} x_{l_1} & \cdots & x_{l_{k+1}} \\ \log(t_{l_1}) & \cdots & \log(t_{l_{k+1}}) \end{pmatrix} \neq 0, l_1, \ldots l_k \leq n \right\} \tag{A.10}$$

(i) For $p = 1$. Barring a set of zero Lebesgue measure, $\lambda_{m_{k+1}} = \max\{\lambda_i : i \notin \{m_1, \ldots, m_k\}\}$. Hence, (A.8) is bounded above by $\int_\Theta \pi(\theta)\, d\theta = 1$. If $n > k$, the posterior exists.

(ii) For $p = 1 + k/2$. By the same argument, $\int_\Theta E(\Lambda_{m_{k+1}}^{-\frac{k}{2}}|\theta)\pi(\theta)\, d\theta$ is an upper bound for (A.8). However, $\mathrm{E}(\Lambda_{m_{k+1}}^{-\frac{k}{2}}|\theta) \leq \mathrm{E}(\Lambda_{(1)}^{-\frac{k}{2}}|\theta)$ where $\Lambda_{(1)}$ is the first order statistic of $\{\Lambda_1, \ldots, \Lambda_n\}$. Hence, it follows that $\mathrm{E}(\Lambda_{(1)}^{-\frac{k}{2}}|\theta) \leq n\mathrm{E}(\Lambda_i^{-\frac{k}{2}}|\theta)$ $\forall i = 1, \ldots, n$ and hence, as the Λ_i's are iid, the results holds.

□

Theorem 5. (i) Similarly to Fonseca et al. [2008], it can be shown that the FIM corresponds to

$$\mathrm{FIM}(\beta, \sigma^2, \nu) = \begin{pmatrix} \frac{1}{\sigma^2}\frac{\nu+1}{\nu+3}\sum_{i=1}^{n} x_i x_i' & 0 & 0 \\ 0 & \frac{n}{2\sigma^4}\frac{\nu}{\nu+3} & -\frac{n}{\sigma^2}\frac{1}{(\nu+1)(\nu+3)} \\ 0 & -\frac{n}{\sigma^2}\frac{1}{(\nu+1)(\nu+3)} & \frac{n}{4}k_{ST}(\nu) \end{pmatrix}, \tag{A.11}$$

where $k_{ST}(\nu) = \Psi'(\frac{\nu}{2}) - \Psi'(\frac{\nu+1}{2}) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}$ and $\Psi'(\cdot)$ denotes the trigamma function. Therefore, the components depending on ν of the Jeffreys, independence Jeffreys and independence I Jeffreys prior are, respectively

$$\pi^J(\nu) \propto \left(\frac{\nu+1}{\nu+3}\right)^{k/2} \pi^I(\nu), \tag{A.12}$$

$$\pi^I(\nu) \propto \sqrt{\frac{\nu}{\nu+3}} \sqrt{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}}, \tag{A.13}$$

$$\pi^{II}(\nu) \propto \sqrt{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}}. \quad \text{(A.14)}$$

It can be shown that $\pi^J(\nu)$ and $\pi^I(\nu)$ are proper priors for ν [Corollary 1 in Fonseca et al., 2008]. However, $\pi^{II}(\nu)$ is not (it behaves as ν^{-1} when $\nu \to 0$). Hence, as mentioned in Subsection 3.2.2, the independence I prior is discarded for the log-Student t model.

Theorem 4 part (i) implies the propriety of the posterior distribution for the independence Jeffreys prior. Theorem 4 cannot be used in order to conclude about the posterior existence under the Jeffreys prior (the condition in part (ii) is not satisfied because $\mathrm{E}(\Lambda_1^{-k/2}|\nu)$ does not exist for $\nu < k$). However, as shown in Appendix B, the Jeffreys prior does not produce a proper posterior distribution for the Student t linear regression model. The latter is easily extrapolated to the log-Student t AFT model in absence of censoring. Incorporating censored observations does not help, as the posterior distribution is still not well defined. For effects of (A.15), denote n_o by the number of non-censored observations and n as the total sample size (in an abuse of notation). Under right censoring, the marginal likelihood can be expressed as

$$f_T(t) = \int_{\mathcal{T}^*} \int_{\mathbb{R}^k} \int_{\mathbb{R}_+} \int_{\Theta} \left[\prod_{i=1}^{n} f_{T_i}(t_i^*|\beta,\sigma^2,\nu)\right] \pi(\beta,\sigma^2,\nu)\, d\beta\, d\sigma^2\, d\nu\, dt^* \equiv \int_{\mathcal{T}^*} f_T^*(t^*)\, dt^*, \quad \text{(A.15)}$$

where $\mathcal{T}^* = t_1 \times \cdots \times t_{n_o} \times (t_{n_o+1}, \infty) \times \cdots (t_n, \infty)$ and $f_T^*(t^*)$ is an auxiliary marginal likelihood that treats censored observations as if they were non-censored. As a result of Theorem 11 (Appendix B), $f_T^*(t^*)$ is not finite for any $t^* \in \mathcal{T}^*$. Hence, $f_T(t)$ is not finite and the posterior based on the complete sample is not well-defined under the Jeffreys prior.

(ii) As the parameter θ is not required for the log-Laplace model, the independence Jeffreys and independence I Jeffreys coincide. Theorem 4 part (i) indicates that the posterior is proper under these priors. In both cases, $\mathrm{E}(\Lambda_1^{-\frac{k}{2}})$ is finite and therefore the posterior under the Jeffreys prior is also proper. In fact, for the log-Laplace model, Λ_1^{-1} is Gamma distributed and all its positive moments are finite. Both type of independence Jeffreys priors also coincide the log-logistic model. In such a case, it can be shown that $\Omega_i = \sqrt{1/(4\Lambda_i)}$ has an Asymptotic Kolmogorov distribution with density function $g(\omega_i) = 8\omega_i \sum_{s=1}^{\infty} (-1)^{s+1} s^2\, e^{-2s^2\omega_i^2}$, for $\omega_i > 0$. Therefore, for $k > -2$, it follows that

$$\begin{aligned}
\mathrm{E}(\Lambda_1^{-k/2}) &= 2^{k+3}\sum_{s=1}^{\infty}(-1)^{s+1}s^2\int_0^{\infty}\omega_1^{k+1}\,e^{-2s^2\omega_1^2}\,d\omega_1 && \text{(A.16)}\\
&= 2^{k+2}\sum_{s=1}^{\infty}(-1)^{s+1}s^2\int_0^{\infty}\eta^{k/2}\,e^{-2s^2\eta}\,d\eta && \text{(A.17)}\\
&= 2^{k/2+1}\Gamma(1+k/2)\sum_{s=1}^{\infty}(-1)^{s+1}\frac{1}{s^k}<\infty. && \text{(A.18)}
\end{aligned}$$

For the log-exponential power model, the FIM was derived by Martï¿½n and Pï¿½rez [2009] and is given by

$$\mathrm{FIM}(\beta,\sigma^2,\alpha)=\begin{pmatrix}
\frac{\alpha(\alpha-1)\Gamma(1-\frac{1}{\alpha})}{\sigma^2\Gamma(\frac{1}{\alpha})}\sum_{i=1}^n x_i x_i' & 0 & 0\\
0 & \frac{n\alpha}{\sigma^2} & -\frac{n(1+\Psi(1+\frac{1}{\alpha}))}{\sigma\alpha}\\
0 & -\frac{n(1+\Psi(1+\frac{1}{\alpha}))}{\sigma\alpha} & k_{EP}(\alpha)
\end{pmatrix}, \qquad \text{(A.19)}$$

where $k_{EP}(\alpha)=\frac{n}{\alpha^3}\left[(1+\frac{1}{\alpha})\Psi'(1+\frac{1}{\alpha})+(1+\Psi(1+\frac{1}{\alpha}))^2-1\right]$ and $\Psi'(\cdot)$ denotes the trigamma function. As a consequence, the components depending on α of the Jeffreys, independence Jeffreys and independence I Jeffreys prior are, respectively

$$\begin{aligned}
\pi^J(\alpha) &\propto \left[\frac{\alpha(\alpha-1)\Gamma(1-1/\alpha)}{\Gamma(1/\alpha)}\right]^{k/2}\pi^I(\alpha), && \text{(A.20)}\\
\pi^I(\alpha) &\propto \frac{1}{\alpha}\sqrt{\left(1+\frac{1}{\alpha}\right)\Psi'\left(1+\frac{1}{\alpha}\right)-1}, && \text{(A.21)}\\
\pi^{II}(\alpha) &\propto \frac{1}{\alpha^{\frac{3}{2}}}\sqrt{\left(1+\frac{1}{\alpha}\right)\Psi'\left(1+\frac{1}{\alpha}\right)+\left[1+\Psi\left(1+\frac{1}{\alpha}\right)\right]^2-1}. && \text{(A.22)}
\end{aligned}$$

As the previous components are bounded continuous functions of α in $(1,2)$, they induce proper priors for α. Theorem 4 part (i) implies the propriety of the posterior distribution under the independence Jeffreys and independence I Jeffreys prior. The propriety of the posterior under the Jeffreys prior can be verified using Theorem 4 part (ii) because $\mathrm{E}(\Lambda_1^{-\frac{k}{2}}|\alpha)$ is a continuous bounded function for $\alpha\in(1,2)$. In fact,

$$\mathrm{E}(\Lambda_1^{-\frac{k}{2}}|\alpha)=\frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\frac{\mathrm{E}(W^{\frac{k+1}{2}}|\alpha)}{\mathrm{E}(Z^{\frac{k+1}{2}}|\alpha)}=\frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\frac{\Gamma((k+1)/\alpha+1)}{\Gamma((k+3)/2)}, \qquad \text{(A.23)}$$

where $W\sim\mathrm{Weibull}(\alpha/2,1)$ and $Z\sim\mathrm{Exponential}(1)$. The latter uses the

lemma in Meintanis [1998] which states that a Weibull$(a,1)$ random variable can be represented as the ratio of an Exponential(1) and an independent positive stable(a) random variable.

□

Theorem 6. If s is the largest number of observations that can be written as an exact linear combination of their covariates, $\lambda_{m_{k+1}}$ (defined in (A.10)) corresponds to $\lambda_{(n_o-s)}$, which represent the $(n-s)$-th order statistic of $\lambda_1,\ldots,\lambda_n$. The rest of the proof is obtained by iteratively integrating (A.8) and using the inequality [Fernï¿½ndez and Steel, 1999, 2000]

$$\frac{\lambda_{i+1}^{v}}{v}\, e^{-r\lambda_{i+1}} \leq \int_0^{\lambda_{i+1}} \lambda_i^{v-1}\, e^{-r\lambda_i}\, d\lambda_i \leq \frac{\lambda_{i+1}^{v}}{v}, \qquad r, v > 0. \tag{A.24}$$

The integral in (A.24) is not finite for $v \leq 0$. After integrating with respect to the $n-s-1$ smallest λ's, (A.8) has a lower bound given by

$$\int_0^{\infty}\int_{\Lambda^*} \left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{n_o-s} \frac{\left[\frac{\nu+1}{2}\right]^{-(n_o-s-1)}}{(n_o-s-1)!}\lambda_{(n_o-s)}^{a-1}\, e^{-\frac{(n_o-s)\nu}{2}\lambda_{(n_o-s)}} \prod_{i=n_o-s+1}^{n_o} dP(\lambda_{(i)}|\nu)\pi(\nu)\, d\nu, \tag{A.25}$$

where $\Lambda^* = \{(\lambda_{(n-s)},\ldots,\lambda_{(n)}) : 0 < \lambda_{(n-s)} < \cdots < \lambda_{(n)} < \infty\}$ and $a = -\frac{n+2p-k-3}{2} + \frac{\nu}{2} + \frac{(n-s-1)(\nu+1)}{2}$. In (A.25), the integral with respect to $\lambda_{(n-s)}$ requires $a > 0$ in order to have a be finite. Hence, the propriety of the posterior distribution requires $\nu > \frac{n-k+(2p-2)}{n-s} - 1$. □

Theorem 7. Condition (i): This follows the proof in Honoré [1990], which assumed $\alpha = 1$. Using l'Hopital's rule twice, it can be proved that

$$\lim_{t\to 0} \frac{\log(-\log(S(t|\alpha,\gamma,\theta))}{\log(t)} = \gamma \tag{A.26}$$

if and only if $E(\Lambda|\theta)$ is finite. Condition (ii): the survival function associated to T_i corresponds to the Laplace transform of the density of $\alpha\Lambda$ evaluated at t^{γ}. The result is immediate by uniqueness of the Laplace transform.

Theorem 8. $\mathrm{E}(T_i^r)$ exists if and only if $\int_0^{\infty} t_i^r f(t_i|\alpha,\gamma,\theta)\, dt_i < \infty$. Using Fubini's theorem, the result is direct after using the formula for the r-th moment of the Weibull distribution.

Corollary 2. Direct application of the expression for $\mathrm{E}(T_i^r)$ provided in Theorem 8.

Theorem 9. This proof consists in taking the expectation of minus the second derivatives of the likelihood for each observation and computing the FIM on

the basis of the whole sample as the sum of the FIM for single observations. The functions $k_1(\theta)$, $k_2(\theta)$ and $k_3(\theta)$ are given by

$$\begin{aligned} k_1(\theta) &= n\mathrm{E}_{T_i}\left[\frac{\left[\int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\left(1-\lambda_i T_i\, e^{-x_i'\beta}\right)\, dP(\lambda_i|\theta)\right]^2}{E_1^2}\right] \\ &- n\mathrm{E}_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\left(1-3\lambda_i T_i\, e^{-x_i'\beta}+T_i^2\, e^{2x_i'\beta}\right)\, dP(\lambda_i|\theta)}{E_1}\right] \quad \text{(A.27)} \\ k_2(\theta) &= n\mathrm{E}_{T_i}\left[\frac{\left[\int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\left(1-\lambda_i T_i\, e^{-x_i'\beta}\right)\, dP(\lambda_i|\theta)\right] E_2}{E_1^2}\right] \\ &- n\mathrm{E}_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\left(1-\lambda_i T_i\, e^{-x_i'\beta}\right)\, \frac{d}{d\theta}\, dP(\lambda_i|\theta)}{E_2}\right] \quad \text{(A.28)} \\ k_3(\theta) &= n\mathrm{E}_{T_i}\left[\frac{E_2^2}{E_1^2}\right] - n\mathrm{E}_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\, \frac{d^2}{d\theta^2}\, dP(\lambda_i|\theta)}{E_1}\right], \quad \text{(A.29)} \end{aligned}$$

where $E_1 = \int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\, dP(\lambda_i|\theta)$ and $E_2 = \int_{\mathcal{L}} \lambda_i\, e^{-\,e^{-x_i'\beta}\lambda_i T_i}\, \frac{d}{d\theta}\, dP(\lambda_i|\theta)$. Note that $k_1(\theta)$, $k_2(\theta)$ and $k_3(\theta)$ do not depend on β because all the terms inside the expectations depend on T_i and β only through $Y_i = e^{-x_i'\beta}T_i$ and the distribution of Y_i does not depend on β nor i.

Corollary 3. The proof follows directly from Theorem 9 using the structure of the determinant of the FIM and its sub-matrices.

Theorem 10. The posterior distribution of (β, γ, θ) given the data is proper if and only if

$$\int_{\mathbb{R}^k}\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbb{R}_+^n} \gamma^n \prod_{i=1}^{n}\left[t_i^{\gamma-1}\lambda_i\right]\, e^{-\sum_{i=1}^{n}\left(\gamma x_i'\beta + e^{-\gamma x_i'\beta}\lambda_i t_i^{\gamma}\right)}\pi(\gamma,\theta)\prod_{i=1}^{n} dP(\lambda_i|\theta)\, d\theta\, d\gamma\, d\beta \quad \text{(A.30)}$$

is finite. The proof requires the Fubini's theorem in order to exchange the order of the integrals. For integrating with respect to β, we use a similar argument than in Kim and Ibrahim [2000]. For $t_i > 0$ and any value of $(\beta, \gamma, \lambda_i) \in \mathbb{R}^k \times \mathbb{R}_+$, $f_{T_i}(t_i|\beta, \gamma, \lambda_i)$ is bounded by a finite constant. Therefore, the integral in (A.30) has

an upper bound proportional to

$$\int_{\mathbb{R}^k}\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbb{R}_+^k}\gamma^k\prod_{i\in I}\left[t_i^{\gamma-1}\lambda_i\right]\; e^{-\sum_{i\in I}\left(\gamma x_i'\beta+\, e^{-\gamma x_i'\beta}\lambda_i t_i^{\gamma}\right)}\pi(\gamma,\theta)\prod_{i\in I}dP(\lambda_i|\theta)\,d\theta\,d\gamma\,d\beta, \tag{A.31}$$

for any $I=\{(i_1,\ldots,i_k):1\le i_1<\cdots<i_k\le n\}$. Define the transformation $U=g(\beta)=X^*\beta$, where X^* is a $k\times k$ matrix containing the $i_1,\ldots,i_k$ rows of X. Note that X^* has rank k because X was assumed to be full rank. Therefore, $g(\cdot)$ is bijective and its Jacobian corresponds to $\det((X^*)^{-1})$. Hence, (A.31) is proportional to

$$\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbb{R}_+^k}\gamma^k\prod_{i=1}^{k}\left[t_i^{\gamma-1}\lambda_i\right]\left[\prod_{i=1}^{k}\int_{-\infty}^{\infty}e^{-\gamma u_i}\,e^{-e^{-\gamma u_i}\lambda_i t_i^{\gamma}}\,du_i\right]\prod_{i=1}^{k}\pi(\gamma,\theta)\,dP(\lambda_i|\theta)\,d\theta\,d\gamma. \tag{A.32}$$

Using $w_i=e^{-\gamma u_i}$, the later integral becomes

$$\int_{\mathbb{R}_+}\int_{\Theta}\int_{\mathbb{R}_+^k}\gamma^k\prod_{i=1}^{k}\left[t_i^{\gamma-1}\lambda_i\right]\left[\prod_{i=1}^{k}\int_{0}^{\infty}\gamma^{-1}\,e^{-w_i\lambda_i t_i^{\gamma}}\,dw_i\right]\prod_{i=1}^{k}\pi(\gamma,\theta)\,dP(\lambda_i|\theta)\,d\theta\,d\gamma, \tag{A.33}$$

which simplifies to

$$\prod_{i=1}^{k}t_i^{-1}\int_{0}^{\infty}\int_{\Theta}\pi(\gamma,\theta)\,d\theta\,d\gamma. \tag{A.34}$$

Therefore, if $t_i\neq 0$ for all $i\in I$ and $\pi(\gamma,\theta)$ is a proper prior for (γ,θ), then the posterior distribution of (β,γ,θ) given the data exists.

Appendix B

On posterior propriety for the Student-t linear regression model under Jeffreys priors

B.1 Introduction

The normal assumption in linear regression models does not always provide an appropriate fit to real datasets. Data often require more flexible errors, capable of accommodating outlying observations. Regression models with fat-tailed error terms are an increasingly popular choice to obtain more robust inference to the presence of outlying observations. A popular choice is to assume a Student-t distribution for the error term [see for example West, 1984; Lange et al., 1989; Fernï¿½ndez and Steel, 1999; Fonseca et al., 2008]. The choice of a prior is very challenging when conducting Bayesian inference under Student-t sampling. While some "standard" priors can be adopted for the regression and scale parameters, there is no consensus about a prior distribution for the degrees of freedom (ν). Villa and Walker [2013] provide a comprehensive discussion of the literature. The seminal paper by Fonseca et al. [2008] is, perhaps, the first attempt to base an objective prior for ν on formal rules and introduces two objective priors based on the Jeffreys rule. They propose the original Jeffreys-rule prior and one of its variants, the independence Jeffreys prior (which treats the regression parameters independently). These priors have been considered in several subsequent articles. Ho [2012] and Villa and Walker [2013] used both priors. In the context of skew-t models, the independence Jeffreys prior was used in Juárez and Steel [2010] and Branco et al. [2012].

This note is a follow-up of Fonseca et al. [2008]. Their posterior propriety

results are revisited and corrected. In particular, it is shown that the prior based on the original Jeffreys rule precludes the existence of a proper posterior distribution. Nevertheless, the independence Jeffreys prior yields a well-defined posterior distribution.

The Student-t linear regression model is presented in Section B.2, which also includes the priors presented in Fonseca et al. [2008]. Posterior propriety under these priors is examined in Section B.3, while Section B.4 concludes.

B.2 Bayesian Student-t linear regression model

Let $Y = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ represent n independent random variables generated by the linear regression model

$$Y_i = x_i'\beta + \sigma\epsilon_i, \qquad i = 1, \ldots, n, \tag{B.1}$$

where x_i is a vector containing the value of k covariates associated with observation i, $\beta \in \mathbb{R}^k$ is a vector of regression parameters and ϵ_i has Student-t distribution with mean zero, unitary scale and ν degrees of freedom. The Bayesian model is completed using Jeffreys priors, which require the FIM. Similarly to Fonseca et al. [2008] (they parameterize with respect to σ instead), the FIM for the model in (B.1) is given by (A.11). Hence, as in the log-Student t case, the Jeffreys-rule and the independence Jeffreys (which deals separately with the blocks for β and (σ^2, ν)) priors are respectively given by (A.12) and (A.13). These priors have been proposed in Fonseca et al. [2008] and can be written as

$$\pi(\beta, \sigma^2, \nu) \propto \frac{1}{(\sigma^2)^a}\pi(\nu), \tag{B.2}$$

where $\pi(\nu)$ is the component of the prior that depends on ν, $a = 1 + k/2$ for the Jeffreys-rule prior and $a = 1$ for the independence Jeffreys prior. As shown in Fonseca et al. [2008], $\pi(\nu)$ is a proper density function of ν for both priors.

B.3 Posterior propriety

Verifying the existence of the posterior distribution is mandatory under the prior in (B.2), which is not a proper probability density function of (β, σ^2, ν). Corollary 2 in Fonseca et al. [2008] states that, provided $n > k$, the posterior distribution is well-defined under the Jeffreys-rule and the independence Jeffreys priors. Their proof refers to Theorem 1 in Fernï¿½ndez and Steel [1999], but unfortunately this

theorem does not cover the Jeffreys-rule prior, as it assumes that $a = 1$ in (B.2). A necessary condition for the existence of the posterior distribution is now provided in the following Theorem.

Theorem 11. *Let $y = (y_1, \ldots, y_n)'$ be n independent observations from model (B.1). Define $X = (x_1, \ldots, x_n)'$ and assume that $n > k$ and the rank of X is k. Under the prior in (B.2), a necessary condition for posterior propriety is $\pi(\nu) = 0$ for all $\nu \in \left(0, \frac{2a-2}{n-k}\right]$.*

Proof. In the absence of censoring, posterior propriety for the Student t and log-Student t models are equivalent. Therefore, barring a set of zero Lebesgue measure, (A.24) and (A.8) imply that $f_Y(y)$ has a lower bound proportional to

$$\int_0^\infty \int_{\Lambda^*} \left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{n-k} \frac{\lambda_{(n-k)}^{d-1}}{\left[\frac{\nu+1}{2}\right]^{n-k-1}} e^{-\frac{(n-k)\nu}{2}\lambda_{(n-k)}} d\lambda_{(n-k)} \left[\prod_{i=n-k+1}^{n} f_{\Lambda_i}^G(\lambda_{(i)}|\nu)\, d\lambda_{(i)}\right] \pi(\nu)\, d\nu, \tag{B.3}$$

where $\Lambda^* = \{(\lambda_{(n-k)}, \ldots, \lambda_{(n)}) : 0 < \lambda_{(n-k)} < \cdots < \lambda_{(n)} < \infty\}$ and $d = -\frac{n+2a-k-3}{2} + \frac{\nu}{2} + \frac{(n-k-1)(\nu+1)}{2} = \frac{\nu(n-k)+2-2a}{2}$. When integrating with respect to $\lambda_{(n-k)}$, $c > 0$ is needed in order to have a finite integral in (B.3). Hence, the propriety of the posterior distribution requires $\nu > \frac{2a-2}{n-k}$. □

As a consequence, the posterior distribution of (β, σ^2, ν) is not proper if $a > 1$ and the range of ν is $(0, \infty)$. In particular, the Jeffreys-rule prior (for which $a = 1 + k/2$) does not lead to a proper posterior distribution and Bayesian inference is thus precluded with this prior. The independence Jeffreys prior satisfies the necessary condition in Theorem 11, but this does not guarantee posterior existence. Nevertheless, posterior propriety under the independence Jeffreys prior for $n > k$ is proved by Theorem 1 in Fernï¿½ndez and Steel [1999].

B.4 Concluding remarks

The choice of a prior distribution for the degrees of freedom under Student-t sampling is a very challenging task. Fonseca et al. [2008] adopt Jeffreys principles to find objective priors for ν. This is an important addition to the previous literature in which much more ad-hoc priors were used [*e.g.* the exponential prior in Geweke, 1993; Fernï¿½ndez and Steel, 1999]. Here, it is shown that the Jeffreys-rule prior does not produce a proper posterior distribution, in contrast to the claim in Fonseca et al. [2008]. It is crucial to point this out to the scientific community to avoid meaningless inference and misleading conclusions. The Jeffreys-rule prior under Student-t sampling has also been used in Ho [2012] and Villa and Walker

[2013]. For this prior, Fonseca et al. [2008] and Villa and Walker [2013] observe very poor frequentist coverage of the 95% credible intervals for ν when the sample size is small ($n = 30$). For small sample size the lower bound required on the support of ν (here equal to $k/(n-k)$) may easily be violated by samples from the posterior, so this poor empirical performance might be linked to the impropriety shown here. Posterior propriety can be verified under the independence Jeffreys prior, and its use as an objective prior is recommended to practitioners (see Theorem 5).

Appendix C

Simulation study for SMLN-AFT models

This document displays the results of the simulation study implemented for SMLN-AFT model. The objective is to illustrate the performance of the proposed methodology and to assess the effectiveness of the suggested Bayesian model comparison criteria.

Two independent covariates, $x_1 \sim \text{Ber}(0.5)$ and $x_2 \sim \text{Unif}(0,1)$ are simulated and an intercept is added ($k = 3$). Throughout we use $\beta = (4, 0.5, -1)'$ and $\sigma^2 = 0.1$ (which are in the range of usual empirical values). Datasets are simulated from the following models: (i) log-normal, (ii) log-Student t with $\nu = 5$, (iii) log-Student t with $\nu = 20$, (iv) log-Laplace, (v) log-exponential power with $\alpha = 1.2$, (vi) log-exponential power with $\alpha = 1.8$ and (vii) log-logistic. Four different scenarios are defined as a combination of sample size ($n = 100, 500$) and percentage of censoring ($PC = 10\%, 70\%$). Independent censoring times are sampled from a uniform distribution in $(0, C)$ where the value of C is tuned in order to control the percentage of censoring. These rather small sample sizes are often observed in survival datasets. For each model, 100 independent datasets are simulated under each scenario. In all cases, survival times are rounded to integers in order to reflect the usual inaccuracy in the data recording process. The log-normal and the mixture models introduced in Table 1 of the paper are fitted to each dataset. We use set observations with $\epsilon_l = \epsilon_r = 0.5$ for non-censored observations. MCMC chains are run for 300,000 iterations with a burn-in period of 75,000 and thinning period equal to 50 (*i.e.* we use 4,500 draws for the results presented here).

For AFT models, β is usually the parameter of interest. Its interpretation is not affected by our mixing scheme. We compare the performance of different AFT-

SMLN models based on the posterior median $\hat{\beta}$. Under each scenario, the values of $\hat{\beta}$ over the 100 simulations are displayed. The value of β used to generate the data is indicated by a horizontal line. The Bayesian model comparison criteria described in Section 3.5 of the paper are applied to each dataset. We report the number of times in which each model was chosen using DIC, BF and PsBF.

The choice between one of the three Jeffreys-rule based priors suggested in this dissertation is not too critical when making inference about β. For each data-generating model, all priors produced similar estimates of the regression parameters when fitting the same model. Figures C.1-C.7 show the posterior median $\hat{\beta}$ across simulations, adopting the independence Jeffreys prior (which is the only prior that produces a valid posterior for all our examples). Of course, the most accurate estimations arise when the data provides more information, *i.e.* $n = 500$ and $PC = 10\%$.

There are no major differences between log-normal datasets and those generated by a SMLN model with weak unobserved heterogeneity (log-Student t with $\nu = 20$ and log-exponential power with $\alpha = 1.8$). In such cases, the log-normal model correctly estimates β. In addition, fitting SMLN models to log-normal datasets is harmless. The β estimates are concentrated around the true value, although they are slightly more spread out when using a log-Laplace model (which has a very dispersed mixing distribution and can accommodate log-normal tails less well). As expected, if the data generating mechanism involves stronger unobserved heterogeneity, mixture models tend to outperform the log-normal one. For those cases, SMLN models produce more accurate estimates of β in terms of both bias and spread, especially when $PC = 70\%$. This is even the case when using a different mixing distribution than the one that generated the data. These differences are largest for the log-Laplace datasets and diminish for milder cases of unobserved heterogeneity, like the log-logistic case.

Figures C.8 and C.9 summarize the results from the model comparison criteria described in Subsection 3.5. Both types of independence Jeffreys priors produce similar results as they only differ for the log-exponential power model. Hence, we only display results under the Jeffreys and independence Jeffreys priors. The performance of BF is better (and more in line with the other criteria) under the independence Jeffreys prior, except for the log-logistic data. Under the Jeffreys prior and with log-normal data, BF assigns relatively little support to the log-normal model when $n = 100$ (especially for the higher percentage of censoring). For $k = 3$, the Jeffreys prior favours small values of σ^2, much more than the independence Jeffreys (the difference increases with k). When the dataset provides little information

(small n and/or large PC), the prior has a strong influence on posterior inference. We might, thus, underestimate σ^2 and the fitted log-normal model will have too little spread to accommodate the data, even though they were generated by the same model. Predictive criteria are less affected by this. Overall, DIC, BF and PsBF point in the same direction, largely successfully detecting the presence and absence of unobserved heterogeneity. However, very mild forms of unobserved heterogeneity (log-Student t with $\nu = 20$, log-exponential power with $\alpha = 1.8$) are often indistinguishable from the log-normal model. Stronger unobserved heterogeneity is more easily detected (even when $n = 100$ and $PC = 70\%$). In any case, jointly, these criteria provide a confident assessment of the existence of unobserved heterogeneity. Even in the worst scenario, the log-normal model is correctly detected more than 60% of the time if we use the independence Jeffreys prior.

Distinguishing between the different mixing distributions is a rather difficult task but it can be achieved for large sample sizes. The percentages of correct classification under each scenario are shown in Table C.1. The best results are observed for the independence Jeffreys prior. In this case, we correctly classify data generated by the log-Laplace model in at least 60% of the cases when $n = 100$ and at least 82% of the cases for $n = 500$. With log-logistic datasets, the right model is detected in at least 70% of the simulations with $n = 500$ under either prior. The rate of correct detection is lower for the log-Student t and log-exponential power models, for which an extra parameter needs to be estimated. For example, the log-Laplace model is a frequent choice for log-exponential power data with $\alpha = 1.2$. The DIC and PsBF criteria do best overall: under both priors they lead to correct classification of models with moderate or strong heterogeneity on the basis of 500 observations with low censoring in at least 57% of the cases.

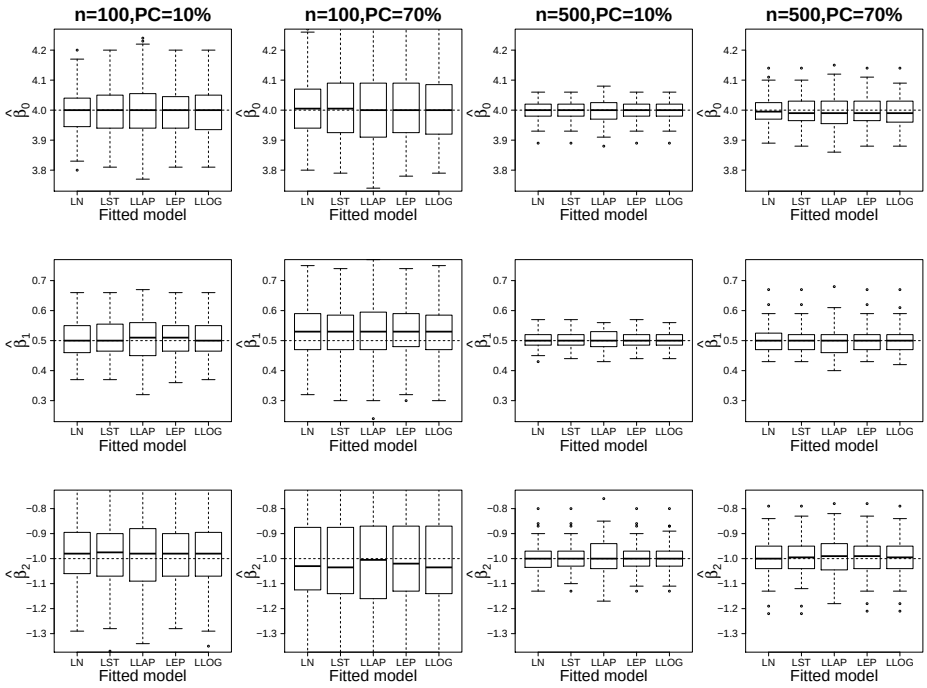


Figure C.1: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-normal data.

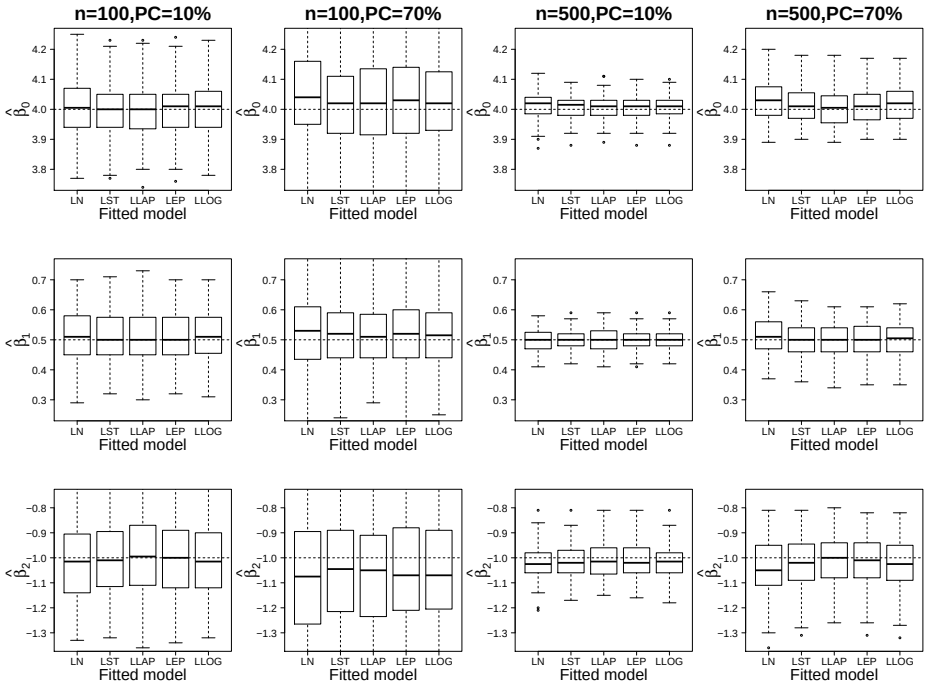


Figure C.2: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-Student t data ($\nu = 5$).

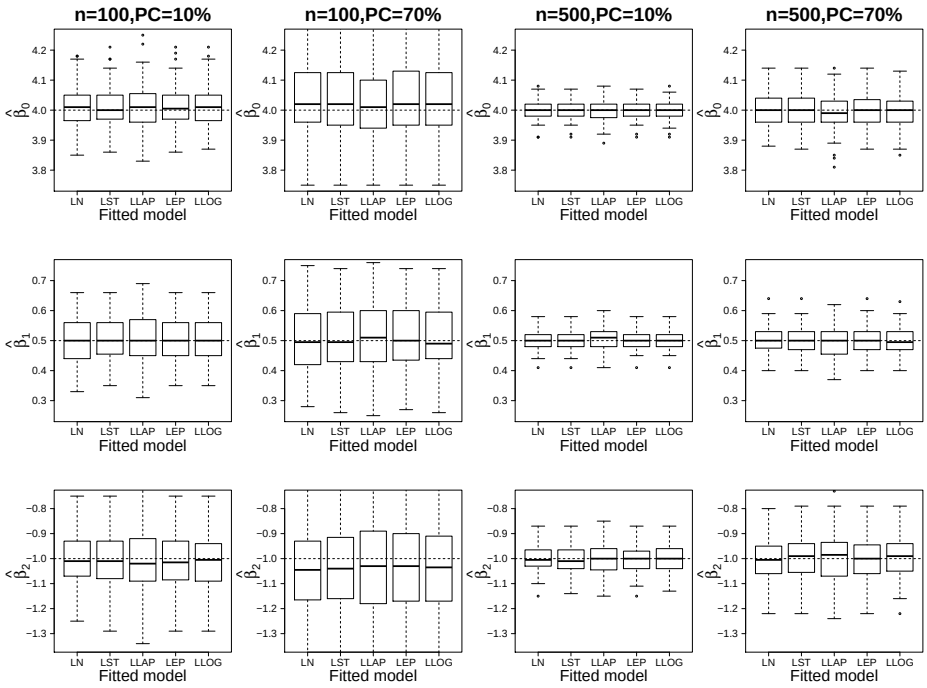


Figure C.3: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-Student t data ($\nu = 20$).

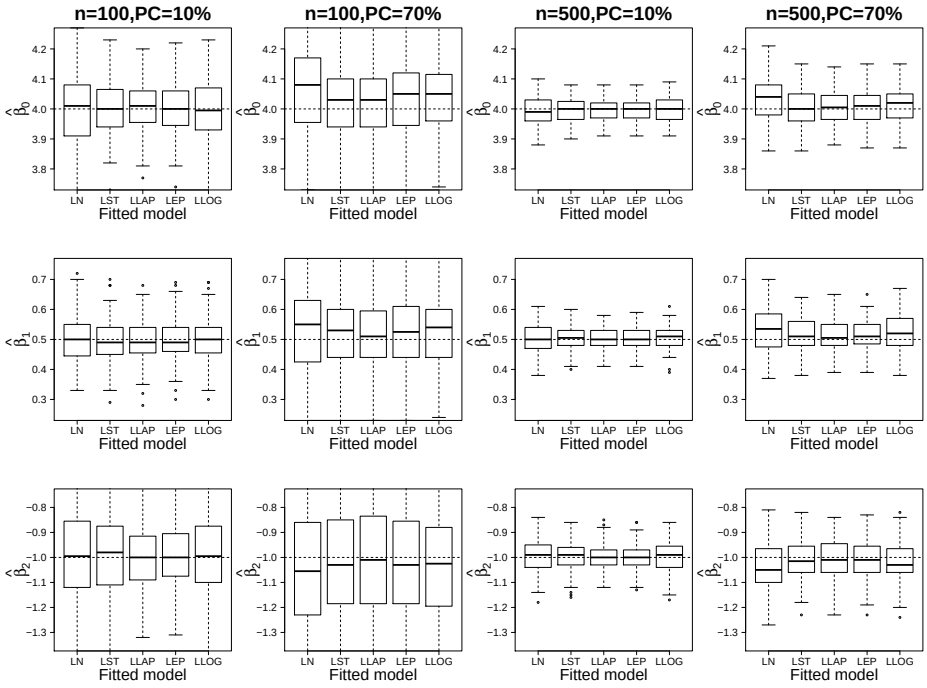


Figure C.4: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-Laplace data.

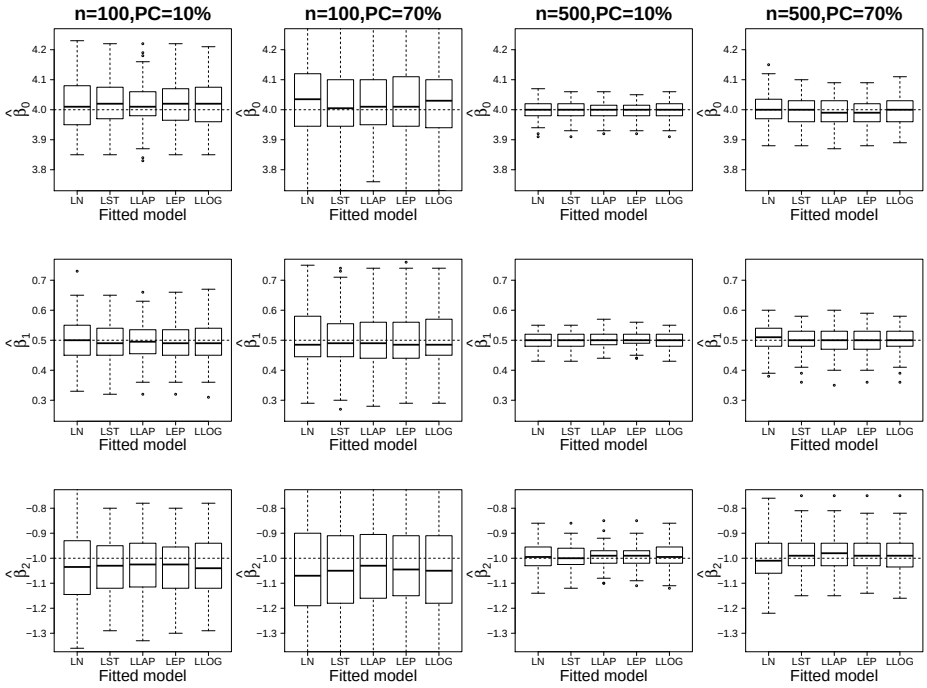


Figure C.5: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-exp. power data ($\alpha = 1.2$).

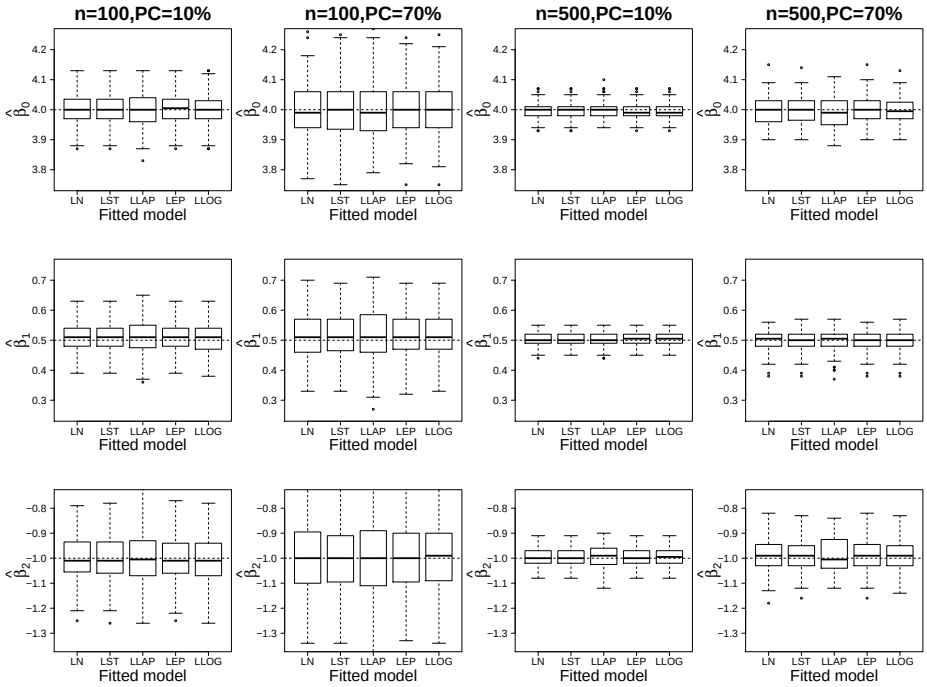


Figure C.6: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-exp. power data ($\alpha = 1.8$).

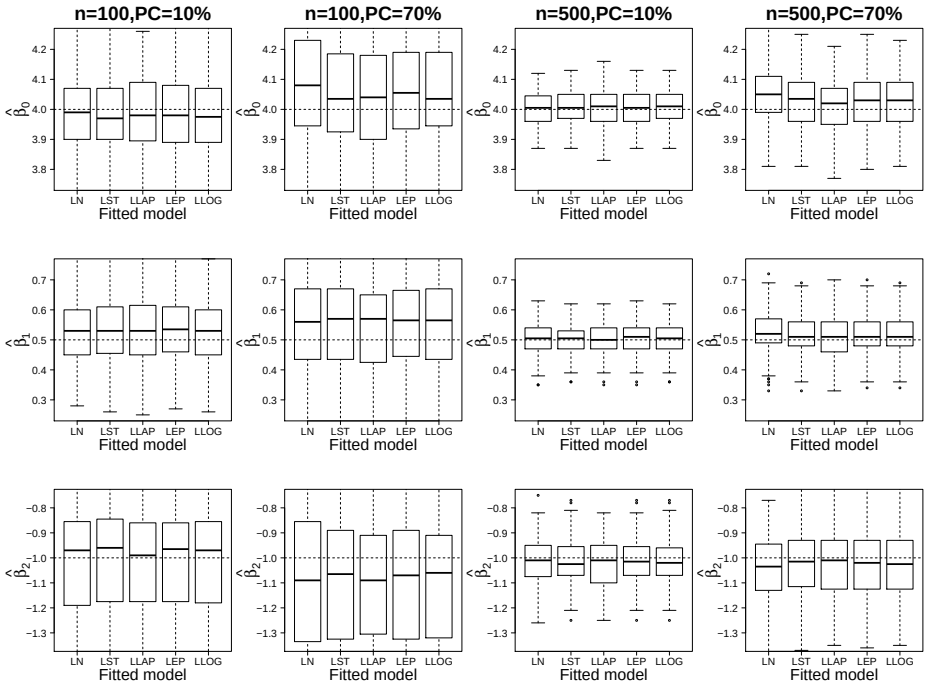


Figure C.7: SMLN simulation study. Boxplot of $\hat{\beta}$ for log-logistic data.

Table C.1: SMLN simulation study. Percentage of correct classification.

		Jeffreys prior						Independence Jeffreys prior					
n	Simulated	PC=10%			PC=70%			PC=10%			PC=70%		
	model	DIC	BF	PsBF	DIC	BF	PsBF	DIC	BF	PsBF	DIC	BF	PsBF
	log-norm.	73	17	71	69	0	61	75	71	72	65	62	60
	log-St. t ($\nu = 5$)	-	-	-	-	-	-	11	0	10	5	1	6
	log-St. t ($\nu = 20$)	-	-	-	-	-	-	0	0	0	0	0	1
100	log-Lap.	75	58	73	66	38	65	72	73	68	62	62	60
	log-e.p. ($\alpha = 1.2$)	1	0	3	2	0	4	0	1	2	2	12	6
	log-e.p. ($\alpha = 1.8$)	2	0	3	4	0	3	5	4	7	2	13	1
	log-log.	72	98	77	27	94	33	42	51	50	28	31	33
	log-norm.	94	86	90	88	48	88	91	96	87	88	88	86
	log-St. t ($\nu = 5$)	-	-	-	-	-	-	61	20	62	23	8	26
	log-St. t ($\nu = 20$)	-	-	-	-	-	-	13	0	16	1	0	2
500	log-Lap.	91	96	90	85	83	87	89	94	87	82	85	82
	log-e.p. ($\alpha = 1.2$)	62	1	62	19	0	20	57	44	60	12	19	17
	log-e.p. ($\alpha = 1.8$)	23	0	26	1	0	3	21	8	22	1	5	3
	log-log.	81	96	80	74	91	76	74	85	72	71	74	70

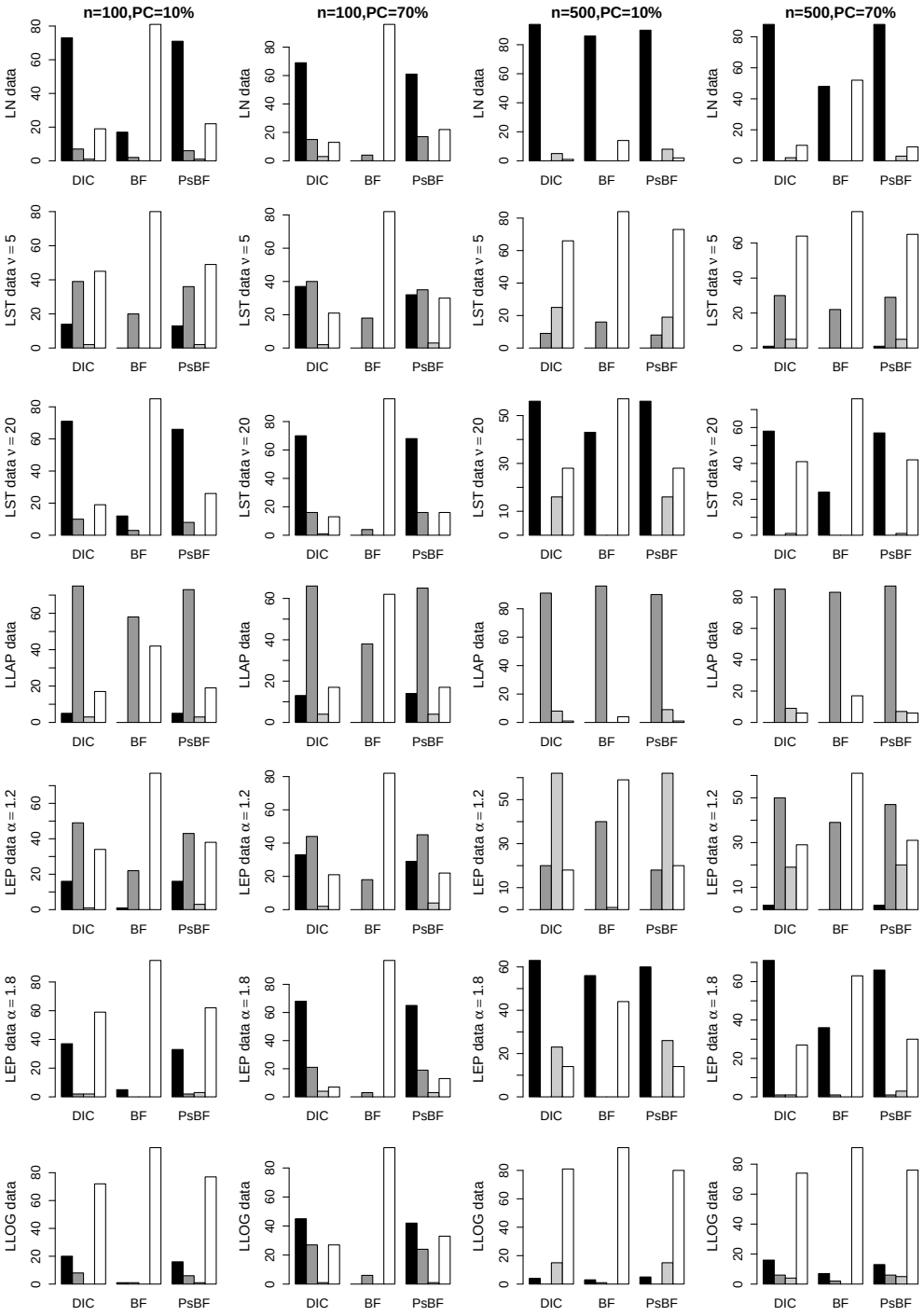


Figure C.8: SMLN simulation study. Distribution of Bayesian model choice under the Jeffreys prior. From darkest to lightest (and left to right): log-normal, log-Laplace, log-exp. power and log-logistic.

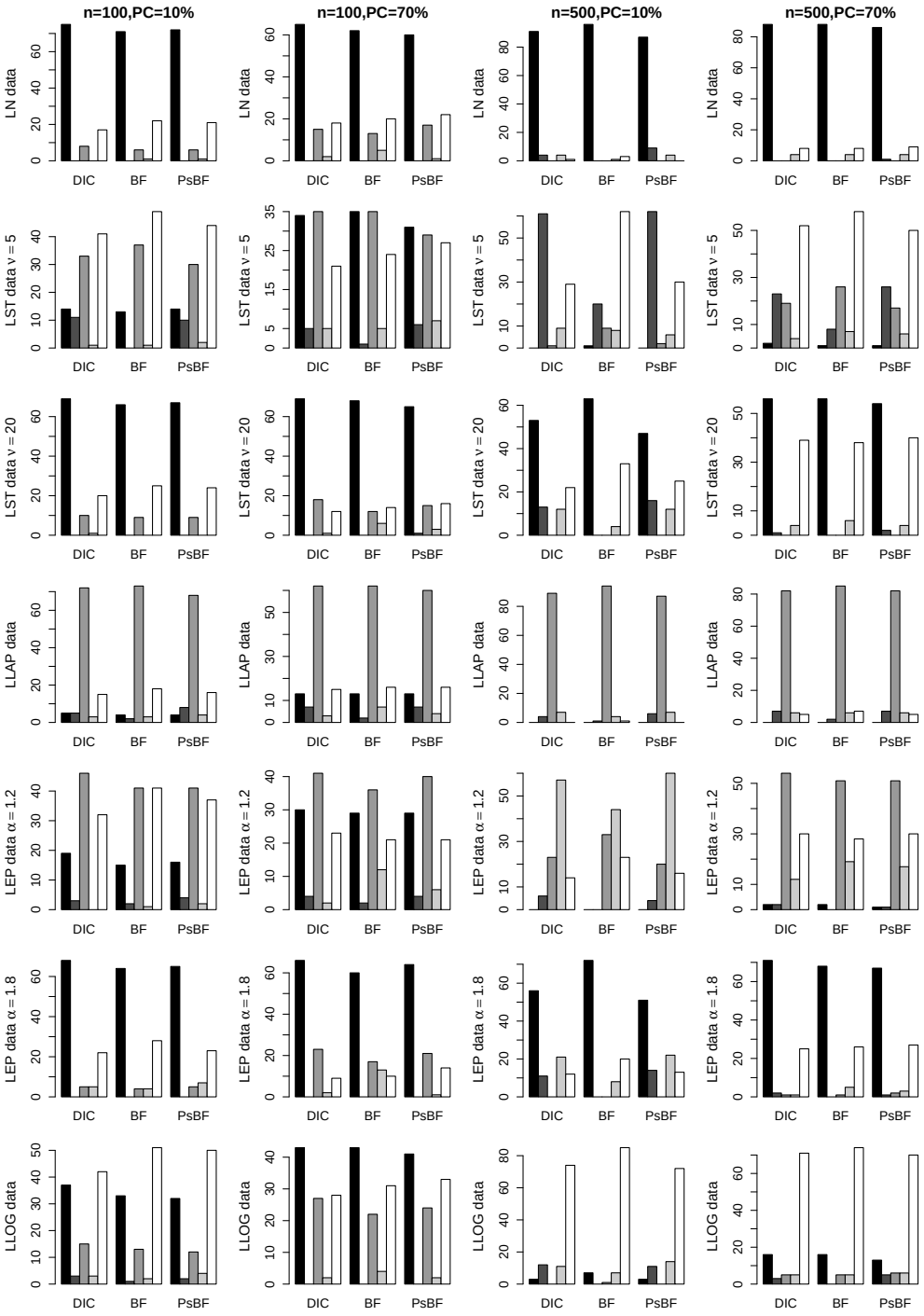


Figure C.9: SMLN simulation study. Distribution of Bayesian model choice under the independence Jeffreys prior. From darkest to lightest (and left to right): log-normal, log-Student t, log-Laplace, log-exp. power and log-logistic.

Appendix D

MCMC chains for Chapter 4

Below, a summary of the convergence analysis for the MCMC chains used in Chapter 4 is provided. Trace plots (some of which are displayed below) provide a first visual indication of both convergence and mixing. Respectively, z-scores and p-values are displayed for the Geweke [1992] and the Heidelberger and Welch [1983] criteria. Throughout, set observations are used for SMLN models. Instead, RMW models are fitted on the basis of point observations.

Convergence of the MCMC chains was never a problem with the number of iterations and burn-in used. Mixing is very good for models without an extra parameter θ in the mixing distribution (*e.g.* log-logistic and RMW model with exponential(1) mixing). When θ is unknown, reliable inference is produced through the MCMC algorithm provided, but the chains are mixing a bit less well for some of the parameters, requiring MCMC run lengths of the order used here.

D.1 VA lung cancer dataset

Table D.1: VA lung cancer data. For MCMC chains: total number of iteration (N), thinning period ($thin$), burning period ($burn$) and update period for λ_i's (Q).

Family	Model	N	$thin$	$burn$	Q
SMLN	all but log-logistic	400,000	20	200,000	1
SMLN	log-logistic	400,000	20	200,000	10
RMW	No mixing	600,000	50	150,000	1
RMW	Exponential mixing	600,000	50	150,000	10
RMW	Gamma mixing	600,000	50	150,000	2
RMW	Inv-Gamma and Inv-Gauss mixing	1,200,000	100	300,000	5
RMW	Log-normal mixing	1,200,000	100	300,000	2

Table D.2: VA lung cancer data. Convergence diagnostics and ESS for log-normal chains.

	Jeffreys prior						Ind. Jeffreys prior					
	Point Observations			Set Observations			Point Observations			Set Observations		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.05	0.21	10000	-0.85	0.42	10000	0.93	0.19	10000	0.54	0.52	10334
β_1	-0.57	0.87	10000	0.14	0.58	10846	-0.54	0.78	10000	0.11	0.65	10152
β_2	-1.48	0.51	9526	-0.70	0.82	10000	-0.97	0.44	10000	0.94	0.98	10000
β_3	-0.59	0.11	10000	0.54	0.63	10679	-1.08	0.67	10000	1.03	0.39	10000
β_4	-0.07	0.52	10000	-0.23	0.77	10586	-0.65	0.71	10000	0.80	0.48	10000
β_5	-1.05	0.62	9631	-0.80	0.58	10884	0.36	0.72	10000	-1.31	0.08	9739
β_6	0.27	0.69	10590	-1.25	0.52	10000	-0.30	0.05	10000	0.49	0.59	10000
β_7	0.67	0.62	10000	1.24	0.40	10000	-1.05	0.39	10000	-0.51	0.87	10000
β_8	1.34	0.69	10174	0.06	0.67	10391	-0.10	0.65	10000	-0.64	0.99	9693
σ^2	1.25	0.58	10000	0.91	0.64	10805	-0.87	0.15	10000	-0.34	0.92	10000

Table D.3: VA lung cancer data. Convergence diagnostics and ESS for log-Student's t chains.

	Ind. Jeffreys prior		
	Geweke	HW	ESS
β_0	-1.91	0.34	10000
β_1	-0.74	0.96	8874
β_2	-0.58	0.77	9432
β_3	0.17	0.49	10000
β_4	1.26	0.15	10000
β_5	0.03	0.52	9537
β_6	0.63	0.48	10016
β_7	1.82	0.70	9288
β_8	0.09	0.46	10000
σ^2	0.64	0.98	1799
ν	0.18	0.89	374

Table D.4: VA lung cancer data. Convergence diagnostics and ESS for log-Laplace chains.

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.36	0.34	10000	-0.04	0.78	10000
β_1	0.28	0.57	10000	0.21	0.24	10000
β_2	-0.61	0.63	10000	-0.77	0.90	10000
β_3	0.59	0.40	10000	-0.23	0.90	9138
β_4	0.14	0.59	10000	0.34	0.50	10000
β_5	0.30	0.82	10371	-0.24	0.19	10470
β_6	-0.44	0.94	9609	-0.90	0.46	10000
β_7	-0.52	0.41	9268	0.29	0.97	10000
β_8	0.37	0.86	10000	-0.20	0.22	10444
σ^2	-0.71	0.22	10000	-0.98	0.58	10000

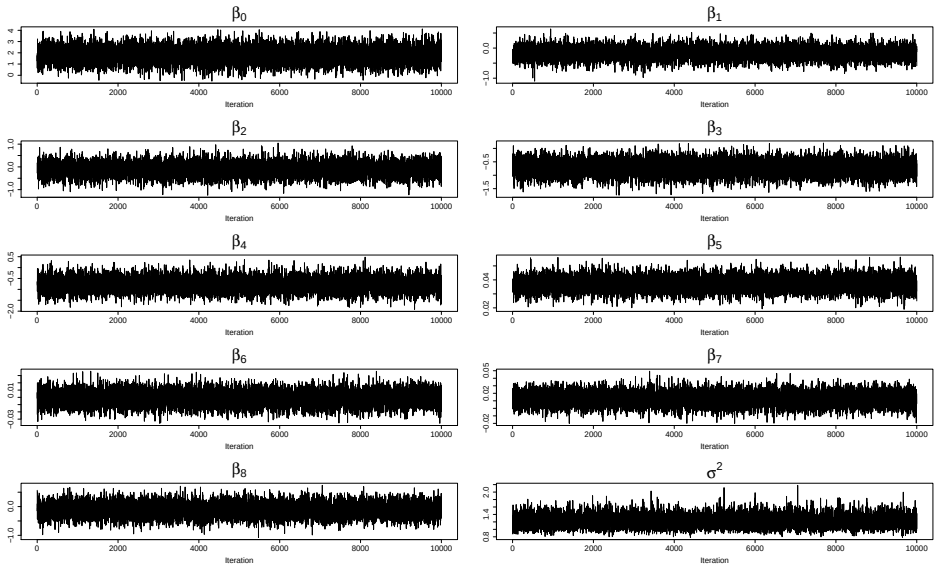


Figure D.1: VA lung cancer data. Log-normal chains under the ind. Jeffreys prior (set observations).

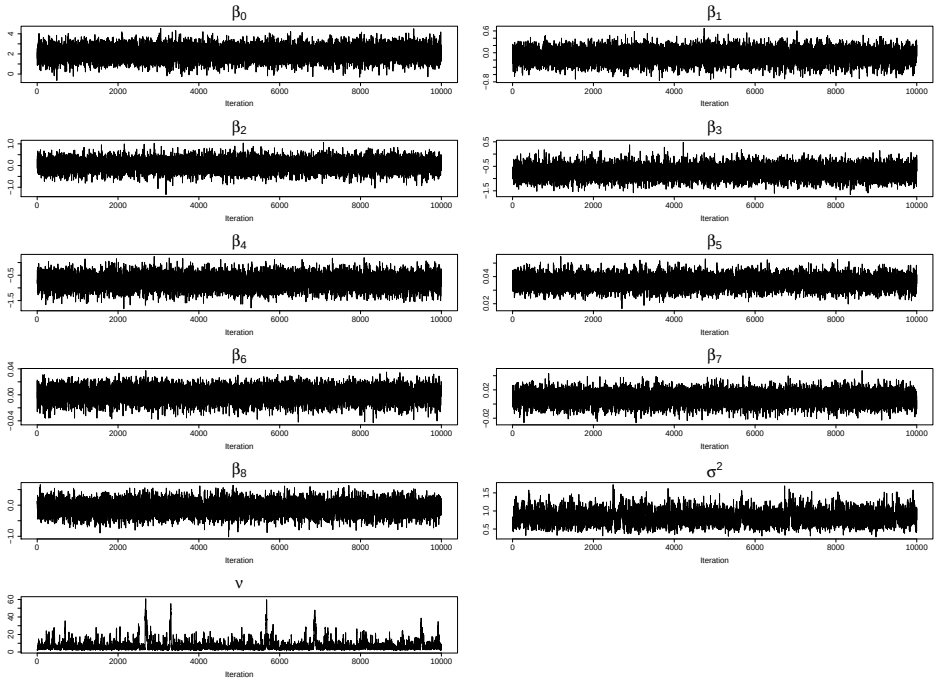


Figure D.2: VA lung cancer data. Log-Student's t chains under the ind. Jeffreys prior.

Table D.5: VA lung cancer data. Convergence diagnostics and ESS for log-exp. power chains.

	Jeffreys prior			Ind. Jeffreys prior			Type I Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.13	0.97	481	-1.00	0.87	496	0.11	0.98	532
β_1	-1.19	0.32	6776	0.30	0.70	7292	-1.36	0.54	7275
β_2	0.01	0.96	6028	-1.06	0.51	6740	0.01	0.33	6786
β_3	-0.99	0.59	3848	-0.82	0.95	4339	-0.73	0.37	4923
β_4	-0.69	0.95	4445	-1.86	0.05	5303	1.36	0.24	4464
β_5	-0.12	0.93	1557	1.82	0.68	1554	-0.26	0.56	1854
β_6	-0.41	0.56	6827	0.90	0.63	8435	-0.62	0.36	7216
β_7	0.04	0.98	633	0.92	0.89	606	-0.04	1.00	549
β_8	-0.64	0.36	5570	-1.17	0.08	7005	0.86	0.85	6441
σ^2	1.11	0.39	3494	-1.38	0.88	3893	-0.70	0.86	4129
α	0.87	0.64	3588	-1.23	0.97	3953	-0.29	0.67	4318

Table D.6: VA lung cancer data. Convergence diagnostics and ESS for log-logistic chains.

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.30	0.97	9674	-0.21	0.98	10000
β_1	0.61	0.74	10000	-0.32	0.90	9548
β_2	-0.93	0.33	10000	0.07	0.82	10000
β_3	-0.45	0.82	10000	-1.06	0.89	10000
β_4	-0.64	0.44	10000	0.51	0.86	10000
β_5	-0.76	0.83	10000	0.15	0.94	10536
β_6	-0.42	0.37	10155	-0.33	0.51	10545
β_7	0.06	0.86	9696	0.09	0.81	10280
β_8	1.57	0.53	10000	-0.32	0.19	10000
σ^2	-0.74	0.33	9550	0.56	0.58	9626

Table D.7: VA lung cancer data. Convergence diagnostics and ESS for Weibull chains.

	$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.13	0.87	1024	-0.60	0.45	1020	-0.33	0.83	1046
β_1	1.85	0.21	7576	1.99	0.31	9000	0.88	0.15	9000
β_2	-0.51	0.33	6358	-0.23	0.25	7480	0.38	0.57	7663
β_3	0.23	0.63	7159	0.73	0.53	6208	0.18	0.48	6994
β_4	0.32	0.84	7910	-1.04	0.10	7285	0.98	0.81	8094
β_5	0.01	0.94	2149	0.87	0.32	2422	0.89	0.73	2221
β_6	1.18	0.59	6751	0.57	0.28	8136	-0.48	0.95	6939
β_7	-0.42	0.72	1123	0.29	0.51	1190	0.04	0.70	1208
β_8	0.01	0.76	8547	-0.14	0.85	9000	-1.48	0.23	9000
γ	-0.84	0.28	8628	0.22	0.57	9000	-0.86	0.31	7469

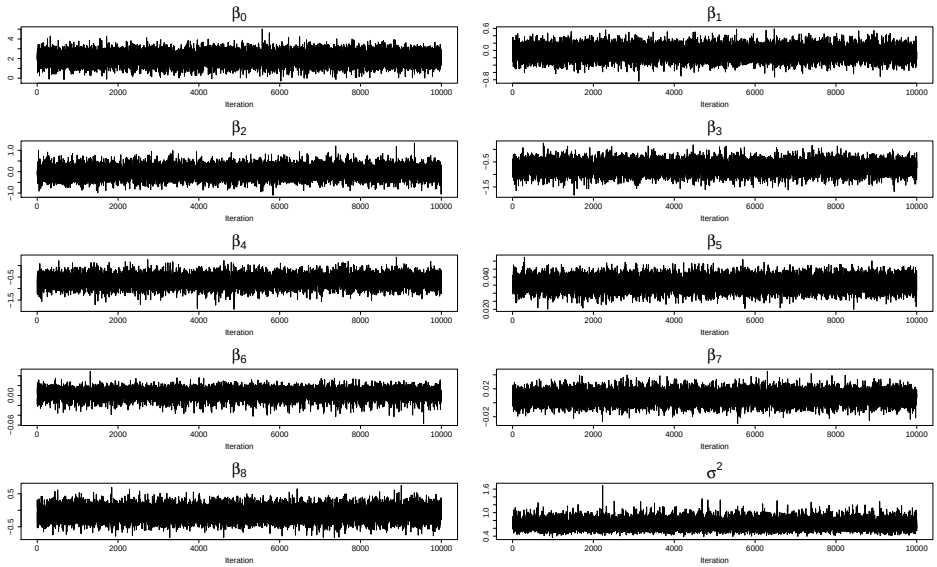


Figure D.3: VA lung cancer data. Log-Laplace chains under the ind. Jeffreys prior.

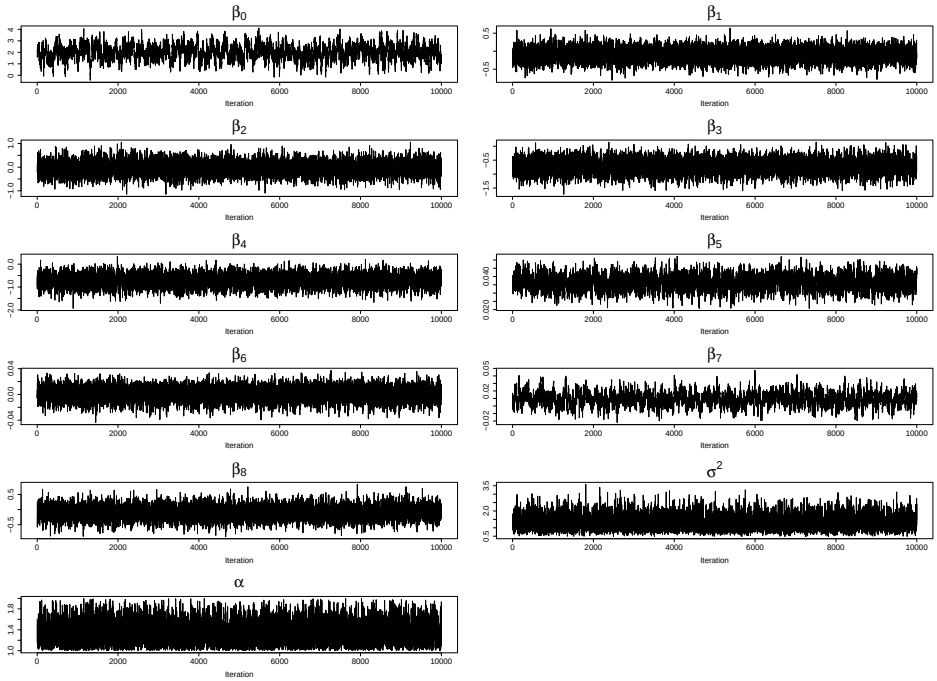


Figure D.4: VA lung cancer data. Log-exp. power chains under the ind. Jeffreys prior.

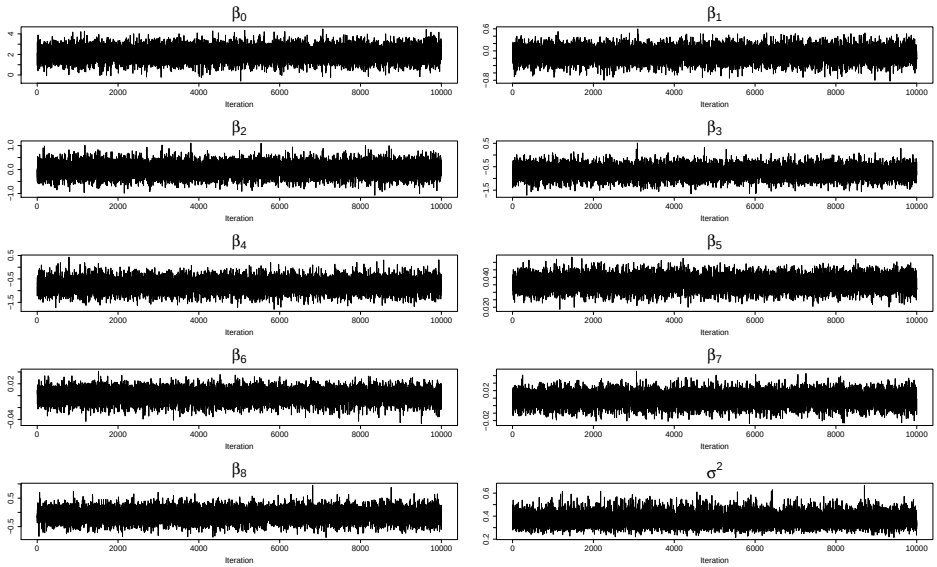


Figure D.5: VA lung cancer data. Log-logistic chains under the ind. Jeffreys prior.

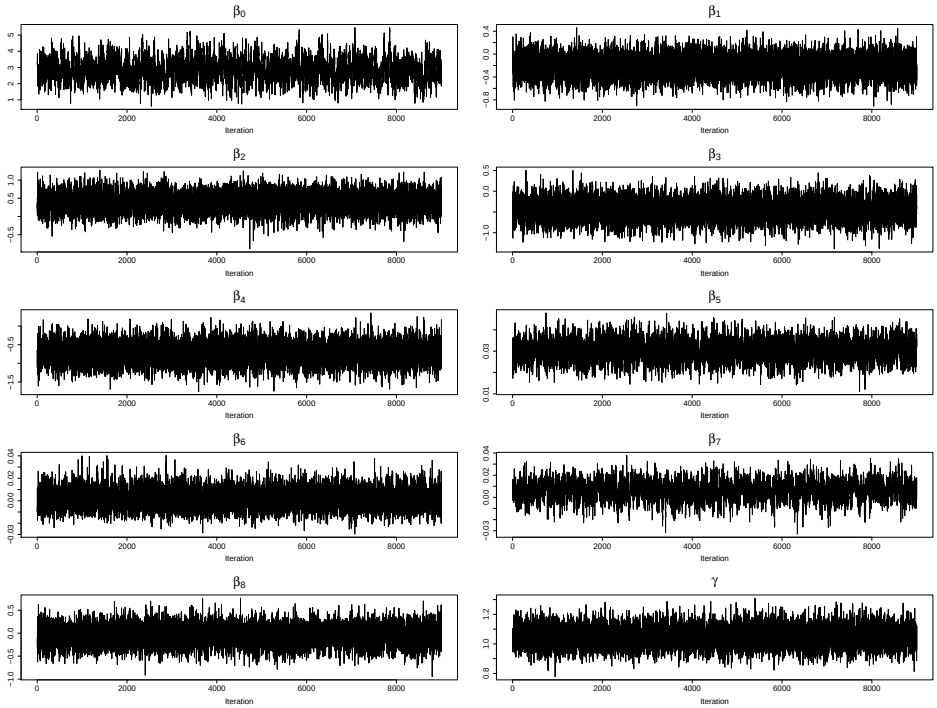


Figure D.6: VA lung cancer data. Weibull chains under Gamma(4,1) prior for γ.

Table D.8: VA lung cancer data. Convergence diagnostics and ESS for RMW chains with exponential(1) mixing.

	$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.56	0.91	440	-0.35	0.36	433	-1.01	0.10	452
β_1	-0.59	0.69	5912	-0.50	0.87	5783	0.91	0.88	6158
β_2	0.68	0.52	4256	-0.83	0.54	4510	-1.48	0.43	4447
β_3	-0.21	0.76	3349	-1.54	0.54	3479	-0.37	0.68	3145
β_4	-0.47	0.99	4398	-1.23	0.60	4046	-0.38	0.31	3629
β_5	-1.28	0.65	1352	-0.03	0.12	1387	0.34	0.36	1307
β_6	-0.20	0.80	4386	-1.88	0.37	4365	-1.46	0.57	4623
β_7	-0.24	0.73	593	0.79	0.47	494	1.38	0.06	532
β_8	-0.78	0.68	6409	0.66	0.85	6014	1.22	0.10	6471
γ	-0.15	0.16	7254	1.02	0.90	7496	1.63	0.56	6890

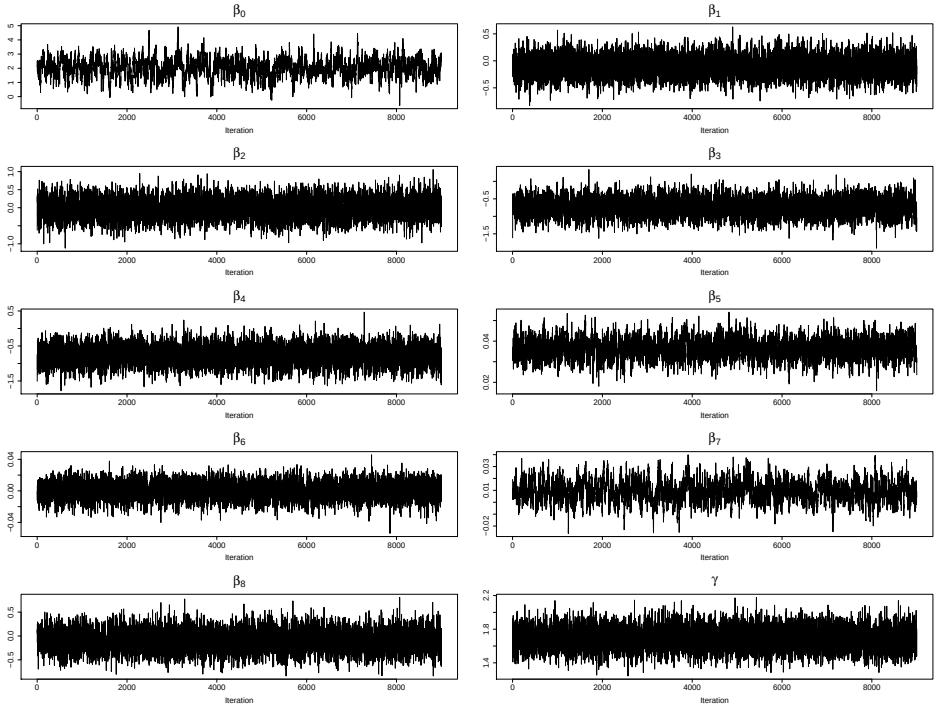


Figure D.7: VA lung cancer data. RMW chains with exponential(1) mixing under Gamma(4,1) prior for γ.

Table D.9: VA lung cancer data. Convergence diagnostics and ESS for RMW chains with Gamma(θ, θ) mixing.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	-0.05	0.34	589	-1.90	0.72	569	-0.47	0.90	620	0.07	0.32	555	-0.06	0.52	513	-1.43	0.24	576
	β_1	0.59	0.73	7383	-0.09	0.54	7593	-0.56	0.47	8577	-0.75	0.30	7680	1.64	0.21	7289	-0.13	0.97	7323
	β_2	0.07	0.88	5825	-0.01	0.07	6578	0.47	0.51	6315	-0.22	0.67	5301	0.65	0.76	5910	0.91	0.56	6210
	β_3	-0.81	0.72	4600	0.40	0.18	5189	-0.19	0.32	5199	-1.71	0.54	4309	-0.66	0.22	4818	0.78	0.47	4261
	β_4	-0.25	0.74	5965	0.70	0.10	4588	-0.41	0.34	6412	-1.96	0.16	4784	0.29	0.85	5354	0.58	0.93	4620
	β_5	0.23	0.88	1569	1.88	0.62	1340	1.16	0.59	1730	-1.83	0.11	1693	0.06	0.34	1619	1.13	0.17	1931
	β_6	0.82	0.77	7200	-0.01	0.45	8247	0.65	0.55	6556	0.23	0.78	6877	0.72	0.83	6403	-0.53	0.50	6970
	β_7	-0.04	0.18	662	1.84	0.62	750	0.28	0.83	794	0.73	0.42	674	-0.09	0.43	586	1.68	0.22	734
	β_8	-1.36	0.40	8384	1.49	0.07	8721	-0.44	0.95	8086	0.97	0.44	7605	-0.06	0.72	8176	1.85	0.20	7824
	γ	1.00	0.70	2382	0.98	0.72	3816	0.59	0.86	3137	-0.58	0.77	1802	0.17	0.78	1849	0.78	0.82	1872
	θ	-1.19	0.62	142	-0.71	0.11	2006	-1.63	0.14	1033	1.00	1.00	1084	-0.13	1.00	1060	-0.81	0.65	1007
Pareto	β_0	0.04	0.58	679	-0.64	0.96	661	0.51	0.35	613	-1.31	0.16	580	1.47	0.70	561	-0.81	0.70	597
	β_1	0.04	0.62	7533	-0.03	0.92	7286	0.71	0.50	7861	1.01	0.67	7578	-0.86	0.60	7062	0.09	0.76	8139
	β_2	0.35	0.78	6441	0.62	0.43	5938	1.60	0.08	6623	-1.03	0.33	6062	0.31	0.81	6138	-0.31	0.33	6164
	β_3	0.09	0.85	5487	0.32	0.83	5513	0.89	0.37	5397	-0.05	0.80	4447	-1.83	0.71	5079	0.15	0.06	4362
	β_4	-0.94	0.91	6882	0.54	0.53	6003	-0.64	0.31	5178	-0.40	0.43	5317	-1.12	0.70	5159	0.91	0.30	5861
	β_5	0.20	0.40	1899	1.21	0.12	1805	-0.21	0.38	1521	1.06	0.23	1795	-1.32	0.46	1626	0.66	0.85	1672
	β_6	1.47	0.19	7412	0.68	0.18	5661	-1.80	0.34	7276	0.63	0.24	7129	-0.90	0.48	6869	1.31	0.54	6659
	β_7	-0.10	0.76	682	0.38	0.93	793	-0.56	0.47	710	1.23	0.15	650	-1.78	0.49	666	0.86	0.55	672
	β_8	-1.62	0.52	8669	-0.97	0.07	8417	0.37	0.75	8232	-1.08	0.28	8502	-0.18	0.80	8385	-1.92	0.58	7567
	γ	-0.10	0.91	3578	0.91	0.79	2856	-0.57	0.32	3438	0.79	0.43	2912	0.47	0.80	2623	0.34	0.81	3031
	θ	0.12	0.97	1104	-0.93	0.38	962	1.02	0.10	1542	-1.41	0.48	1071	-1.17	0.43	422	-1.16	0.23	1637

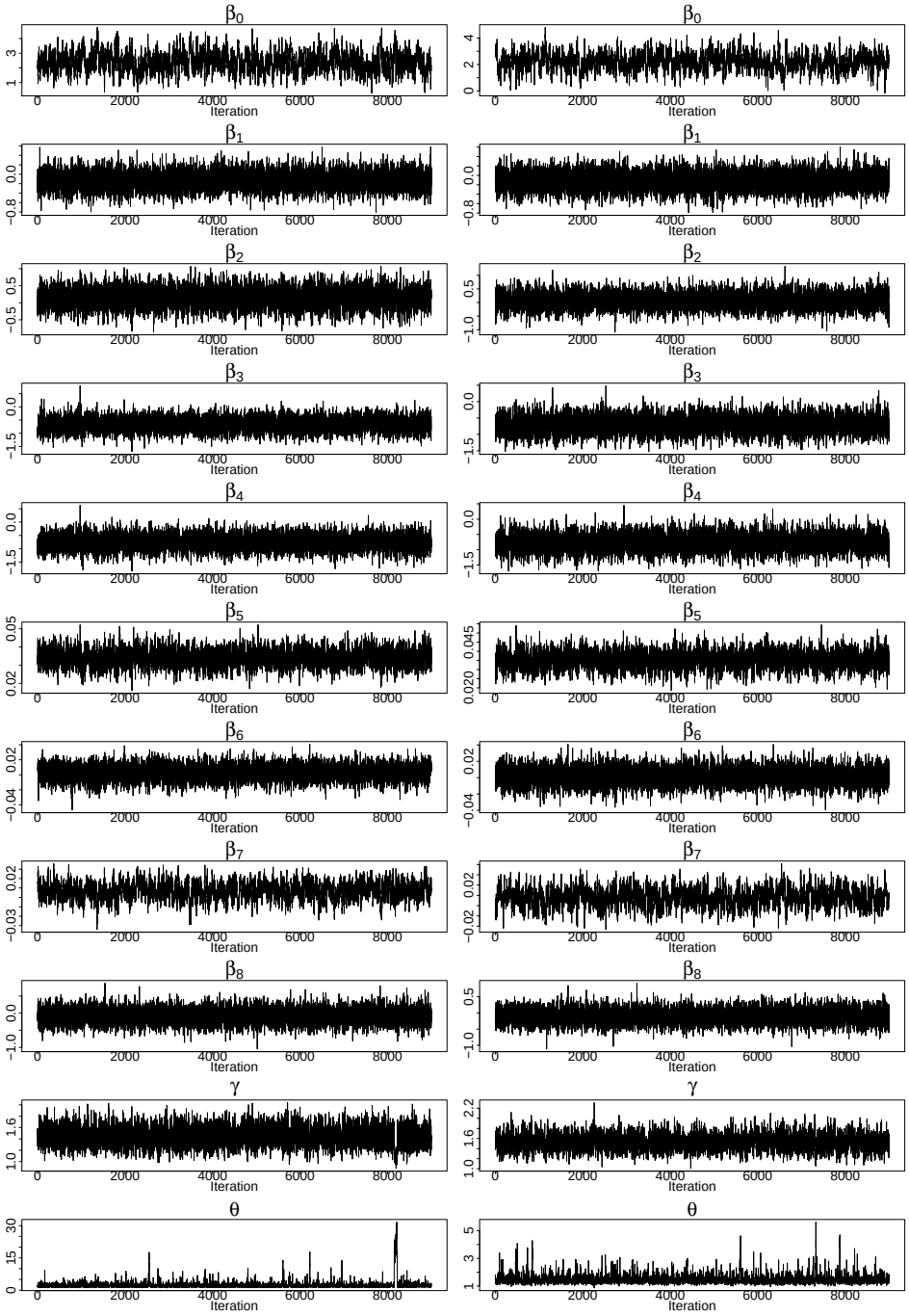


Figure D.8: VA lung cancer data. RMW chains with Gamma(θ, θ) mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (right panels).

Table D.10: VA lung cancer data. Convergence diagnostics and ESS for RMW chains with Inv-Gamma$(\theta, 1)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	1.62	0.71	712	1.44	0.21	282	0.01	0.50	505	0.16	0.75	424	-1.43	0.27	492	-0.27	0.41	573
	β_1	-0.18	0.43	9000	0.37	0.85	8649	0.14	0.78	9000	1.21	0.35	8683	-0.29	0.61	8712	-0.30	0.53	8458
	β_2	-0.66	0.70	7672	-0.13	0.56	6432	0.23	0.64	7728	0.42	0.25	8229	2.02	0.29	8297	0.30	0.75	8266
	β_3	0.00	0.10	7278	-0.87	0.32	7302	-0.88	0.34	7785	0.39	0.51	7569	0.91	0.69	7540	0.19	0.33	6933
	β_4	-0.31	0.51	7749	0.48	0.83	8188	-0.41	0.51	8228	-0.67	0.37	8329	0.74	0.59	8200	0.69	0.99	8485
	β_5	-0.13	0.41	3164	-1.56	0.59	2926	-0.25	0.80	3089	-0.98	0.90	2940	1.28	0.69	2721	1.46	0.16	3224
	β_6	0.54	0.25	7756	-0.53	0.33	8345	1.58	0.18	7337	0.14	0.65	8022	1.31	0.31	9000	-0.35	0.37	8735
	β_7	-0.53	0.49	1511	-1.40	0.10	1543	-0.38	0.93	1679	-1.13	0.83	1516	-0.15	0.40	1448	0.62	0.26	1633
	β_8	-0.76	0.18	9000	-1.37	0.51	8542	-0.91	0.73	8951	-1.02	0.38	9000	-0.28	0.46	9000	-0.84	0.62	9000
	γ	1.46	0.71	635	0.97	0.12	511	-0.03	0.38	521	-0.83	0.31	603	-1.38	0.38	659	-0.38	0.50	685
	θ	-1.93	0.37	258	-0.75	0.63	117	0.30	0.49	164	0.40	0.89	202	0.88	0.79	274	-0.95	0.86	231
Pareto	β_0	-1.78	0.19	322	1.53	0.60	137	1.03	0.55	165	-0.93	0.09	542	0.58	0.38	585	-0.85	0.79	255
	β_1	0.35	0.83	9000	0.32	0.99	9021	0.78	0.83	9000	1.08	0.20	8486	-0.39	0.19	9000	1.54	0.22	9000
	β_2	1.98	0.25	7942	0.23	0.55	6884	-1.49	0.06	5514	0.39	0.15	8500	-1.49	0.56	8053	0.21	0.75	7603
	β_3	1.86	0.22	7713	0.12	0.78	7315	-0.28	0.25	7352	-0.81	0.34	7986	-0.46	0.87	7771	0.37	0.69	6465
	β_4	1.27	0.76	8309	0.66	0.52	8423	-1.19	0.74	7441	0.17	0.76	8091	-0.07	0.21	7175	-0.61	0.97	7895
	β_5	0.34	0.66	2946	0.93	0.79	3226	0.59	0.61	2940	1.59	0.24	3045	-0.46	0.87	2732	-0.53	0.78	3285
	β_6	-1.52	0.09	8410	1.67	0.33	8612	-1.71	0.12	8527	-0.18	0.78	7538	0.41	0.23	7133	0.05	0.74	7539
	β_7	0.29	0.96	1508	0.10	0.92	2500	-0.07	0.23	1609	1.30	0.49	1530	-0.21	0.92	1708	-0.71	0.10	1730
	β_8	0.84	0.77	9000	-0.32	0.76	9000	1.18	0.20	9000	1.29	0.44	9000	-0.25	0.24	9000	-1.33	0.41	8308
	γ	-1.60	0.09	348	1.35	0.43	201	1.24	0.65	287	0.11	0.21	589	-0.21	0.19	628	-1.09	0.82	360
	θ	1.16	0.49	145	-1.70	0.81	56	-0.82	0.85	59	-0.01	0.20	253	-1.17	0.53	231	0.73	0.91	102

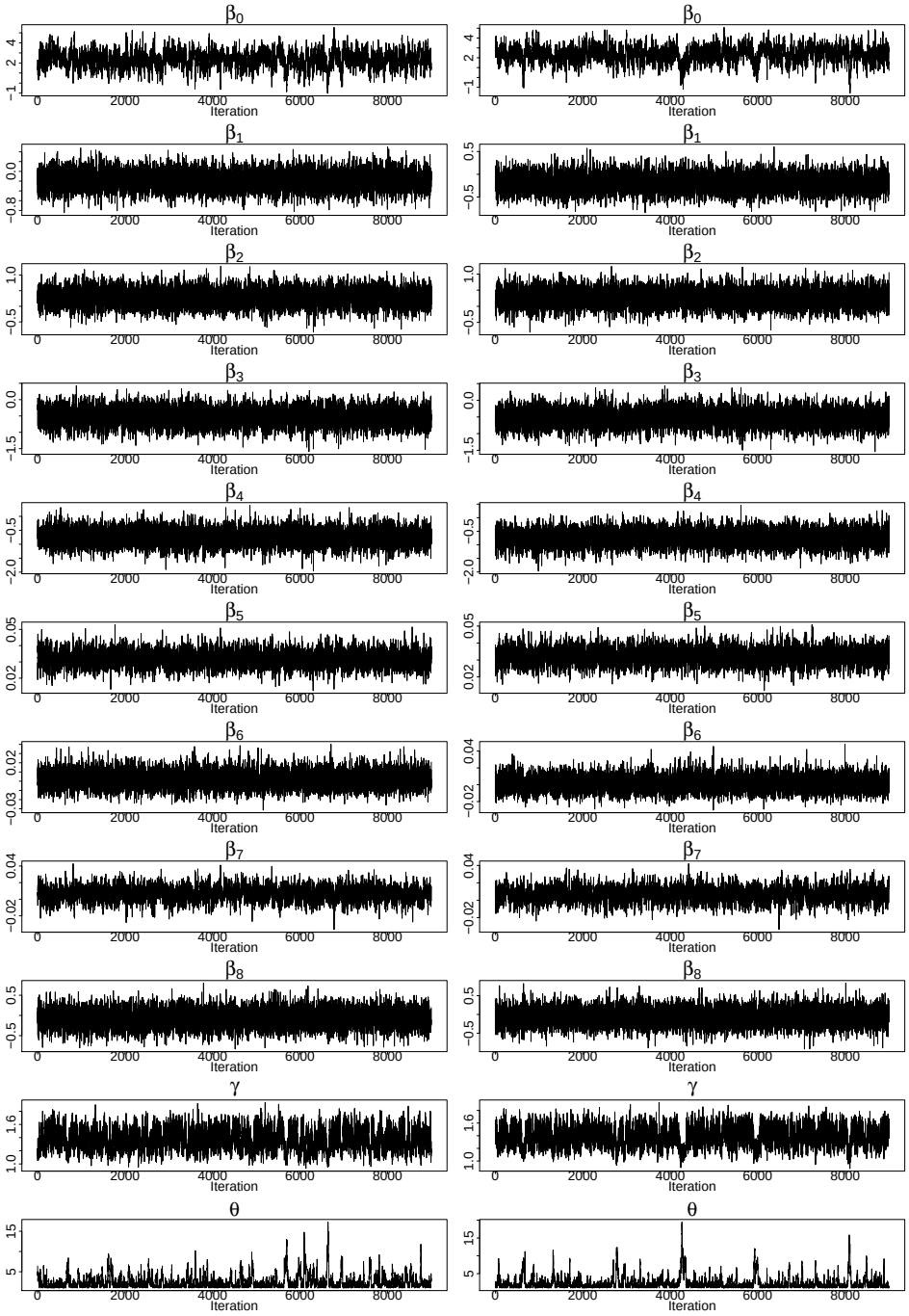


Figure D.9: VA lung cancer data. RMW chains with Inv-Gamma(θ, 1) mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (left panels).

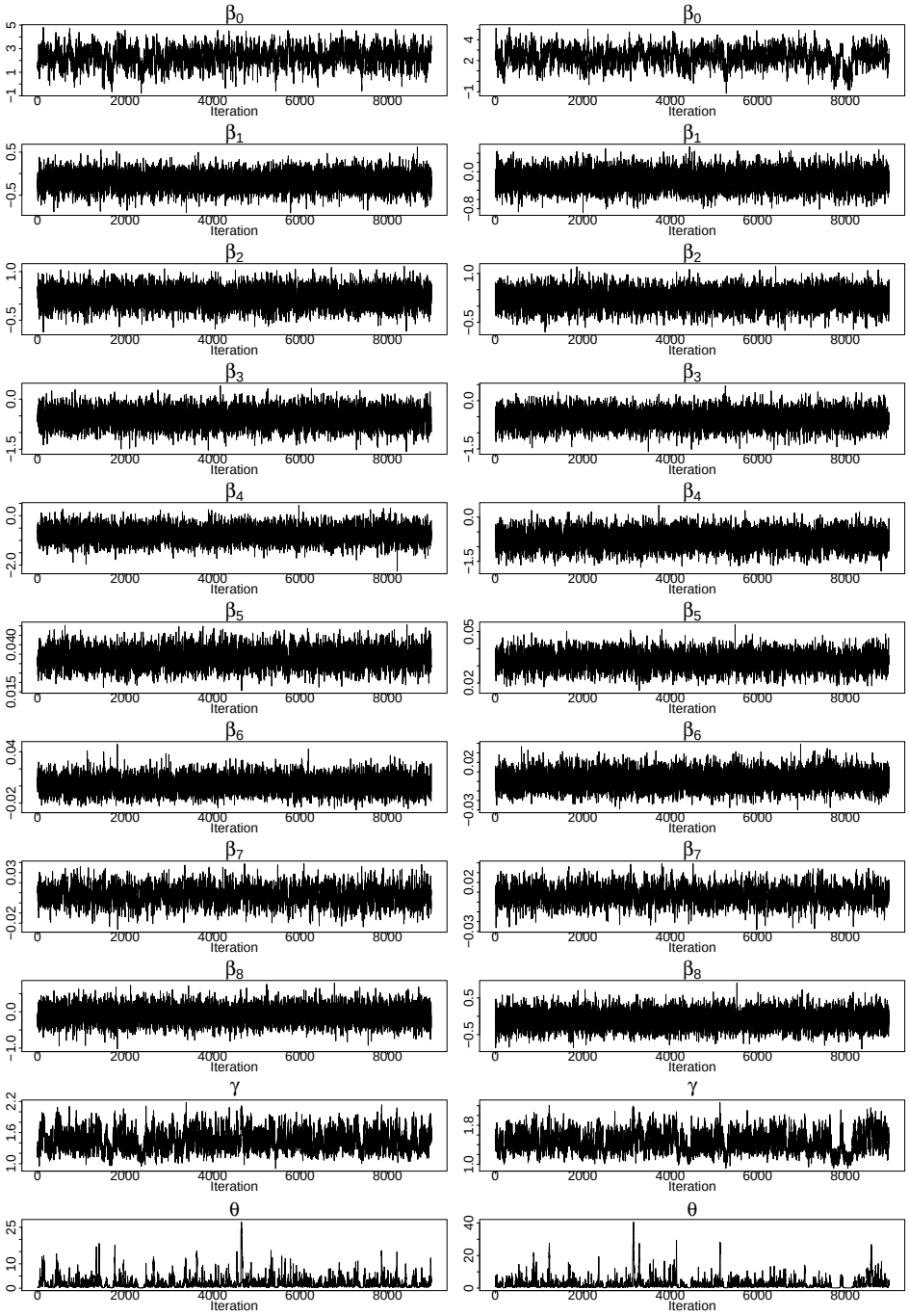


Figure D.10: VA lung cancer data. RMW chains with Inv-Gaussian$(\theta, 1)$ mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (right panels).

Table D.11: VA lung cancer data. Convergence diagnostics and ESS for RMW chains with Inv-Gaussian$(\theta, 1)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	-0.83	0.09	706	1.56	0.21	520	0.62	0.37	607	1.23	0.35	376	0.41	0.73	699	0.22	0.24	381
	β_1	-0.04	0.15	9000	1.14	0.39	7960	1.23	0.47	9498	0.72	0.38	8683	0.85	0.47	7308	0.35	0.76	7511
	β_2	0.80	0.15	6344	-1.15	0.67	6043	-1.52	0.16	5680	-1.51	0.36	5188	-1.00	0.34	7018	0.83	0.49	7562
	β_3	0.91	0.59	7148	-1.22	0.40	7657	-1.06	0.64	7143	0.13	0.10	6558	-0.66	0.24	6075	0.12	0.79	7052
	β_4	-0.06	0.67	7955	-1.27	0.36	7187	-0.89	0.85	7865	-1.00	0.17	7636	-1.94	0.08	6840	1.36	0.08	8047
	β_5	-0.52	0.35	2766	0.70	0.96	2635	0.53	0.28	2838	-0.84	0.16	2785	-1.64	0.20	2507	0.36	0.80	3103
	β_6	0.67	0.83	6477	-0.60	0.41	8212	1.32	0.19	6497	-0.86	0.31	6894	-0.96	0.14	7049	0.02	0.91	6969
	β_7	1.01	0.56	1355	-0.42	0.37	1433	0.34	0.56	1413	-0.33	0.68	1257	-1.23	0.43	1366	-0.15	0.19	1520
	β_8	-1.06	0.82	8121	-0.87	0.17	9000	-0.64	0.51	9536	-1.14	0.91	8310	0.59	1.00	9000	-0.27	0.91	8639
	γ	-0.08	0.35	504	0.56	0.80	418	0.77	0.21	478	0.85	0.45	360	-1.02	0.43	599	-0.33	0.84	366
	θ	0.04	0.85	400	-0.57	0.91	316	0.16	0.30	2703	0.37	0.34	348	-1.93	0.39	400	-0.30	0.97	399
Pareto	β_0	0.34	0.56	556	0.41	0.81	568	-1.66	0.81	574	-0.09	0.89	387	-0.28	0.36	581	-0.18	0.36	570
	β_1	0.05	0.13	8143	1.61	0.24	8484	-0.07	0.87	8429	-1.16	0.11	8516	-0.28	0.39	8845	-0.31	0.39	8352
	β_2	-0.73	0.05	7372	-0.58	0.09	5948	-0.87	0.50	6652	0.93	0.59	4927	-0.51	0.83	7206	-0.13	0.83	6468
	β_3	-0.17	0.07	6860	0.40	0.35	7324	-0.34	0.47	7200	0.81	0.41	6866	0.10	0.96	6851	-0.86	0.96	7226
	β_4	0.03	0.67	7717	0.03	0.68	7683	0.44	0.12	7701	0.90	0.47	7168	0.30	0.50	7243	-0.23	0.50	7464
	β_5	1.08	0.44	2713	1.94	0.07	2892	0.63	0.87	2787	-0.22	0.82	2561	-1.22	0.87	2995	0.55	0.87	2785
	β_6	0.52	0.27	6871	0.41	0.91	8122	0.75	0.61	7279	0.07	0.66	8324	-0.04	0.97	7496	-0.53	0.97	8238
	β_7	0.55	0.19	1236	0.78	0.51	1541	0.39	0.43	1498	0.50	0.14	1254	-1.15	0.15	1480	-0.10	0.15	1369
	β_8	1.30	0.58	9000	-0.72	0.59	8376	-1.29	0.57	7837	0.02	0.83	9000	0.36	0.43	9000	0.94	0.43	9000
	γ	0.97	0.72	451	0.59	0.96	410	-1.73	0.56	493	0.25	0.78	378	-0.49	0.13	429	-0.41	0.13	484
	θ	0.84	0.79	287	-0.36	0.94	275	-1.79	0.50	439	0.21	0.99	402	-0.08	0.39	279	-0.71	0.39	341

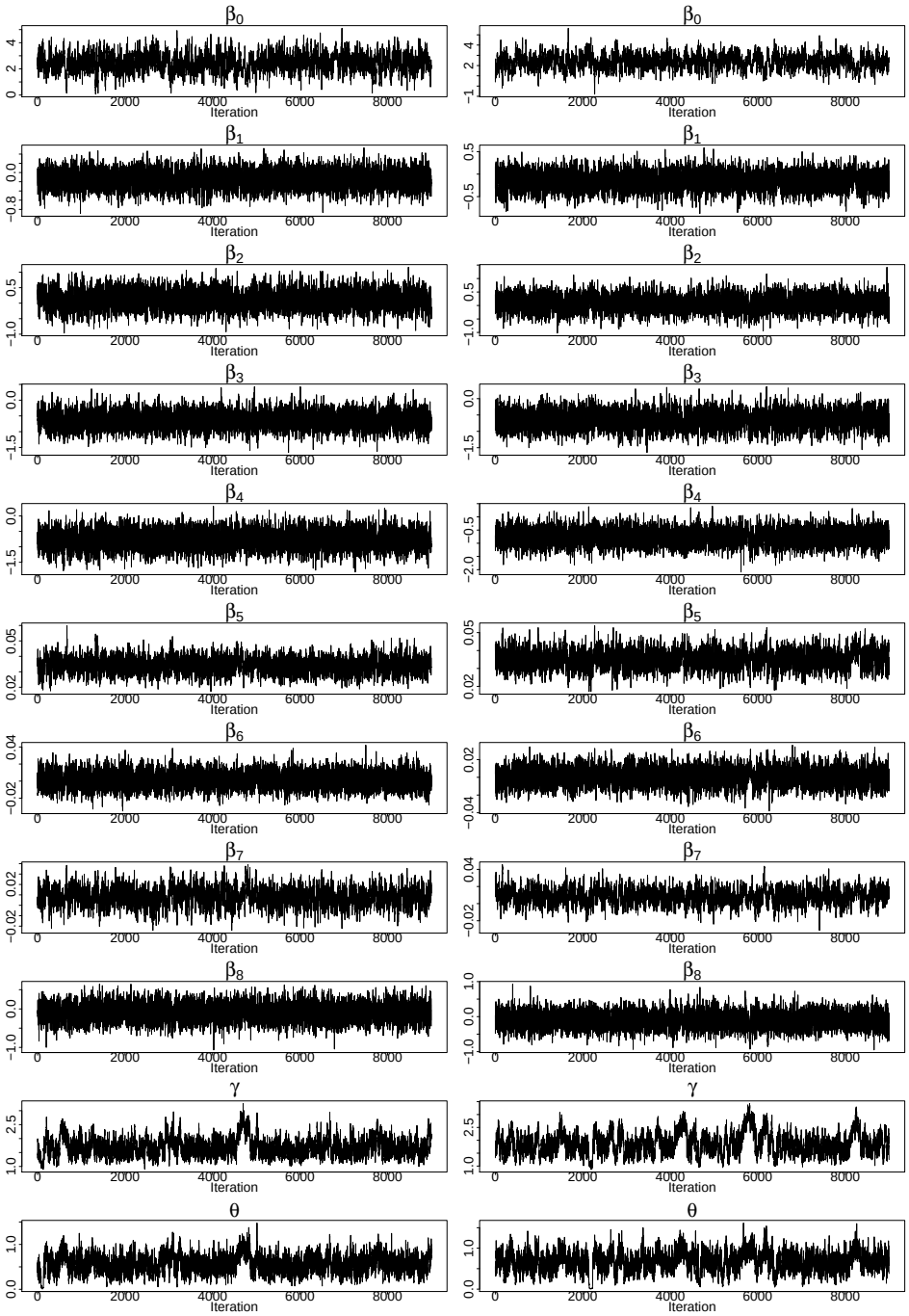


Figure D.11: VA lung cancer data. RMW chains with log-normal$(0, \theta)$ mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (left panels).

Table D.12: VA lung cancer data. Convergence diagnostics and ESS for RMW chains with log-normal$(0, \theta)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	-1.00	0.27	678	-0.59	0.71	885	-1.02	0.68	867	-0.76	0.53	678	-1.11	0.49	673	-1.54	0.57	777
	β_1	0.61	0.65	5423	-1.97	0.20	6554	1.25	0.70	6819	-1.81	0.46	4218	-0.64	0.21	4735	0.72	0.75	3506
	β_2	0.13	0.51	2168	0.54	0.61	3555	-1.17	0.73	2659	0.87	0.49	2621	0.07	0.19	2328	0.41	0.19	1214
	β_3	0.10	0.99	2749	-1.02	0.45	3843	-0.18	0.46	3341	-0.45	0.40	2684	0.20	0.88	2957	0.75	0.22	2322
	β_4	-0.49	0.93	3970	-0.43	0.85	5802	1.46	0.58	3733	0.97	0.86	4048	0.71	0.60	4061	-0.24	0.71	2668
	β_5	0.89	0.34	1463	0.75	0.43	1851	0.59	0.86	2093	0.15	0.48	1200	1.80	0.32	1254	1.20	0.29	1479
	β_6	0.27	0.95	4081	1.32	0.39	5079	0.58	0.84	5143	0.12	0.51	2934	-1.09	0.30	3184	1.16	0.07	3146
	β_7	1.24	0.16	842	0.70	0.81	1084	0.99	0.87	980	1.21	0.54	786	0.84	0.52	787	1.45	0.63	751
	β_8	0.42	0.92	6529	-1.97	0.10	6394	0.91	0.19	6643	0.41	0.65	4316	1.67	0.06	5094	-1.57	0.12	2622
	γ	-0.22	0.66	300	0.18	0.13	330	1.25	0.63	326	-1.05	0.59	128	0.24	0.56	168	0.32	0.41	96
	θ	-0.29	0.55	357	0.53	0.11	427	1.40	0.81	450	-0.99	0.35	215	0.54	0.52	352	0.74	0.17	253
Pareto	β_0	-0.71	0.52	727	-1.58	0.22	972	0.20	0.22	1095	0.05	0.47	708	-0.69	0.48	894	-0.92	0.16	883
	β_1	-1.74	0.56	6404	-1.10	0.52	7350	0.13	0.78	6625	1.91	0.06	5579	1.10	0.53	6437	-1.65	0.29	4526
	β_2	1.37	0.67	2923	-0.96	0.75	4083	-0.86	0.11	3340	-1.46	0.08	3313	-1.33	0.35	4083	1.86	0.09	1891
	β_3	0.72	0.79	4280	-0.99	0.72	4865	0.07	0.80	4478	-0.32	0.12	3911	-0.46	0.19	4832	0.63	0.87	3895
	β_4	1.76	0.65	4865	-0.02	0.26	5639	-0.77	0.71	6484	-0.86	0.06	4258	0.07	0.73	5543	1.36	0.36	5241
	β_5	0.29	0.55	1705	1.06	0.06	2152	0.25	0.05	2415	-0.73	0.36	1487	0.99	0.68	2111	0.55	0.29	1579
	β_6	0.16	0.99	4690	1.17	0.97	5833	-0.87	0.54	6715	-0.60	0.56	3537	-1.22	0.35	5268	-0.09	0.58	3975
	β_7	0.79	0.66	940	1.60	0.56	1076	-0.22	0.36	1206	0.23	0.45	816	0.45	0.55	1056	1.04	0.13	1020
	β_8	-0.66	0.78	6041	-0.98	0.45	6939	0.21	0.52	7486	0.05	0.38	5720	0.28	0.48	6830	1.01	0.83	5161
	γ	0.83	0.33	286	1.31	0.34	236	0.24	0.14	333	-0.59	0.06	249	1.63	0.50	321	-0.31	0.44	248
	θ	0.52	0.95	389	1.52	0.24	327	0.39	0.35	314	-0.48	0.25	390	1.79	0.40	508	-0.46	0.70	280

D.2 AA Bone Marrow Transplant dataset

Table D.13: AA Bone Marrow data. For MCMC chains: total number of iteration (N), thinning period ($thin$), burning period ($burn$) and update period for λ_i's (Q).

Family	Model	N	$thin$	$burn$	Q
SMLN	all but log-Laplace and log-logistic	400,000	20	200,000	1
SMLN	log-Laplace	400,000	20	200,000	5
SMLN	log-logistic	400,000	20	200,000	20
RME	all but Inv-Gamma and Inv-Gauss mixing	600,000	50	150,000	1
RME	Inv-Gamma and Inv-Gauss mixing	600,000	50	150,000	5

Table D.14: AA Bone Marrow data. Convergence diag. and ESS log-normal chains.

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	-0.34	0.91	10000	0.14	0.88	10000
β_1	-0.65	0.82	9246	0.34	0.58	10000
σ^2	-1.22	0.11	10000	-0.36	0.33	10000

Table D.15: AA Bone Marrow data. Convergence diag. and ESS log-Student t chains under ind. Jeffreys prior.

	Geweke	HW	ESS
β_0	0.90	0.92	9944
β_1	-1.17	0.76	10000
σ^2	-1.12	0.59	4254
ν	-1.67	0.82	211

Table D.16: AA Bone Marrow data. Convergence diag. and ESS log-Laplace chains.

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	1.60	0.46	6423	-0.61	0.31	6158
β_1	-0.63	0.74	6837	0.20	0.49	6991
σ^2	0.15	0.49	8974	1.83	0.52	9187

Table D.17: AA Bone Marrow data. Convergence diag. and ESS log-exp. power chains.

	Jeffreys prior			Ind. Jeffreys prior			Ind. I Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.05	0.38	7027	0.86	0.29	6702	-1.40	0.86	6579
β_1	-0.50	0.97	6844	-0.46	0.28	7317	0.27	0.32	6862
σ^2	-0.99	0.33	6043	0.90	0.33	5401	-1.41	0.52	5851
α	-1.55	0.22	5407	-0.31	0.92	5447	0.21	0.98	5457

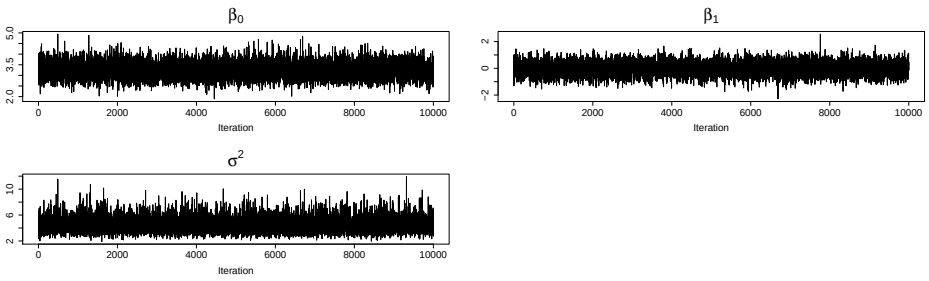


Figure D.12: AA Bone Marrow data. Log-normal chains under ind. Jeffreys prior.

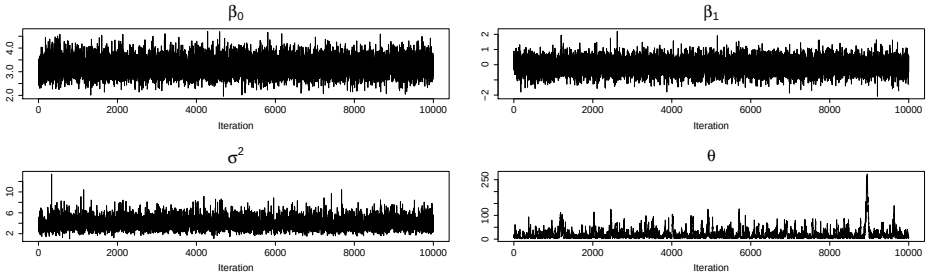


Figure D.13: AA Bone Marrow data. Log-Student t chains under ind. Jeffreys prior.

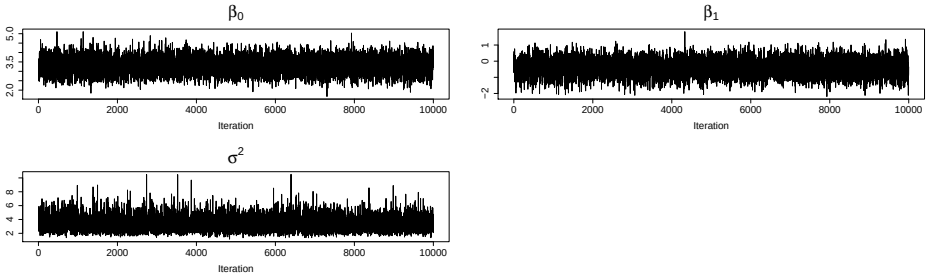


Figure D.14: AA Bone Marrow data. Log-Laplace chains under ind. Jeffreys prior.

Table D.18: AA Bone Marrow data. Convergence diagnostics and ESS for log-logistic chains.

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.41	0.65	9136	-0.32	0.97	8916
β_1	-1.33	0.23	9010	0.22	0.73	8871
σ^2	-0.05	0.94	7700	-1.13	0.51	9005

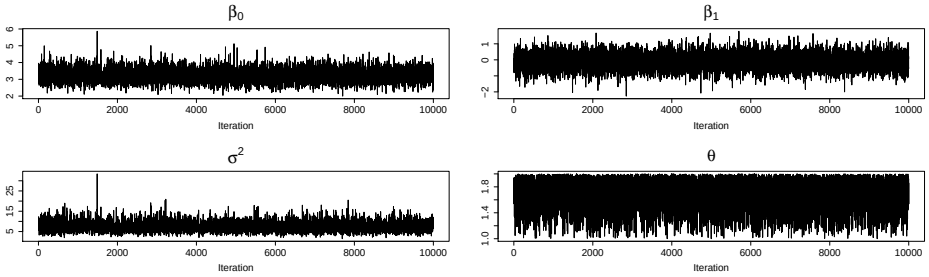


Figure D.15: AA Bone Marrow data. Log-exp. power chains under ind. Jeffreys prior.

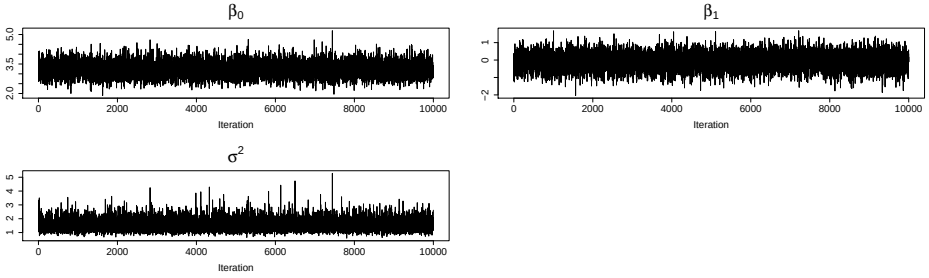


Figure D.16: AA Bone Marrow data. Log-logistic chains under ind. Jeffreys prior.

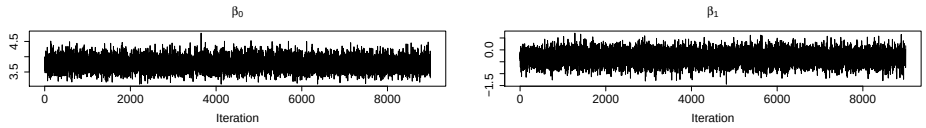


Figure D.17: AA Bone Marrow data. Exponential chains.

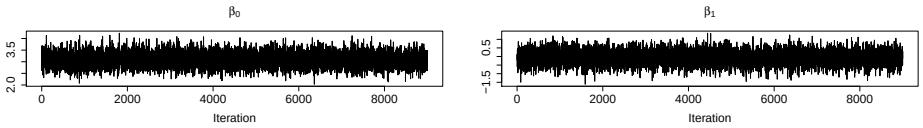


Figure D.18: AA Bone Marrow data. RME chains with exponential (1) mixing.

Table D.19: AA Bone Marrow data. Convergence diagnostics and ESS for exponential chains.

	Geweke	HW	ESS
β_0	0.88	0.73	9000
β_1	-1.69	0.65	9000

Table D.20: AA Bone Marrow data. Convergence diagnostics and ESS for RME chains with exponential(1) mixing.

	Geweke	HW	ESS
β_0	-0.88	0.78	8544
β_1	1.03	0.90	8520

Table D.21: AA Bone Marrow data. Convergence diagnostics and ESS for RME chains with Gamma(θ, θ) mixing.

Prior		E(c_v)=1.25			E(c_v)=1.5			E(c_v)=2			E(c_v)=5			E(c_v)=10		
for c_v		Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS
T. exp.	β_0	0.61	0.34	7876	-1.20	0.11	7125	-0.53	0.18	8301	0.64	0.69	8308	1.05	0.74	8084
	β_1	-1.44	0.25	9000	0.55	0.82	9000	-0.02	0.43	9000	-0.46	0.34	8568	0.78	0.71	8336
	θ	-1.07	0.82	682	-0.64	0.58	405	0.62	0.48	1365	-0.66	0.92	415	0.63	0.06	952
Pareto	β_0	-1.03	0.77	4633	-0.82	0.76	7146	1.05	0.86	5922	0.03	0.65	7546	-0.28	0.69	7678
	β_1	-0.67	0.50	9000	0.24	0.47	8591	0.45	0.60	7948	-0.30	0.73	8513	0.69	0.56	10531
	θ	-0.80	0.98	158	-0.41	0.64	833	0.84	0.10	304	-0.26	0.89	1651	0.57	0.68	577

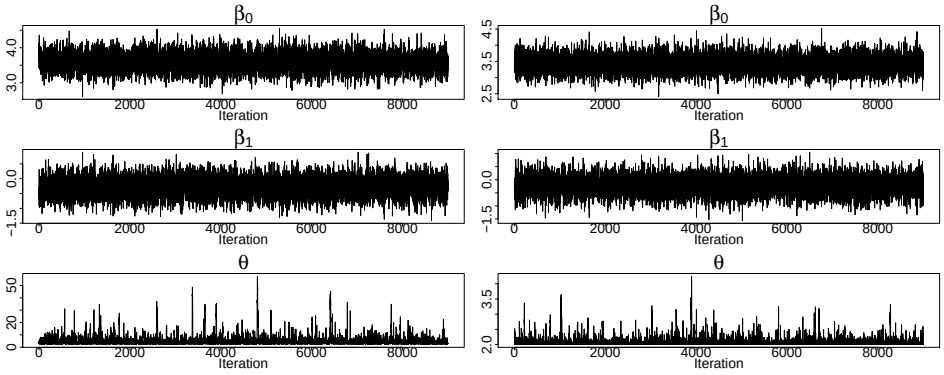


Figure D.19: AA Bone Marrow data. RME chains with Gamma (θ, θ) mixing under a truncated exponential prior for c_v. Left panels use E(c_v)=1.25. Right panels use E(c_v)=10.

Table D.22: AA Bone Marrow data. Convergence diagnostics and ESS for RME chains with Inv-Gamma$(\theta, 1)$ mixing.

Prior		E(c_v)=1.25			E(c_v)=1.5		
for c_v		Geweke	HW	ESS	Geweke	HW	ESS
T. exp.	β_0	-0.18	0.37	1166	-1.34	0.37	1598
	β_1	0.77	0.70	7710	-0.76	0.30	8430
	θ	-0.92	0.94	624	1.07	0.48	975
Pareto	β_0	-1.22	0.72	1088	0.68	0.69	827
	β_1	-1.97	0.68	7952	1.06	0.20	8037
	θ	1.24	0.54	718	-1.06	0.88	274

Table D.23: AA Bone Marrow data. Convergence diagnostics and ESS for RME chains with Inv-Gauss$(\theta, 1)$ mixing.

Prior		E(c_v)=1.25			E(c_v)=1.5			E(c_v)=2		
for c_v		Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS
T. exp.	β_0	0.59	0.82	928	-0.46	0.48	1712	-1.03	0.55	1274
	β_1	0.61	0.17	7626	-0.48	0.51	7223	0.24	0.66	6747
	θ	0.21	0.98	559	-0.97	0.29	1730	-1.78	0.81	1085
Pareto	β_0	-0.54	0.28	581	-0.14	0.54	1110	-0.51	0.61	1608
	β_1	1.92	0.61	7573	0.20	0.76	7043	0.93	0.94	6493
	θ	-1.04	0.22	1247	0.42	0.64	962	-1.00	0.96	450

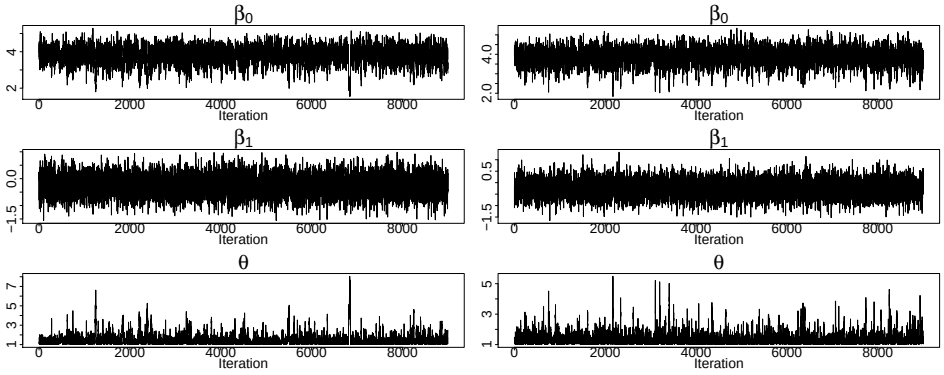


Figure D.20: AA Bone Marrow data. RME chains with Inv-Gamma $(\theta, 1)$ mixing under a truncated exponential prior for c_v. Left panels use E(c_v)=1.25. Right panels use E(c_v)=1.5.

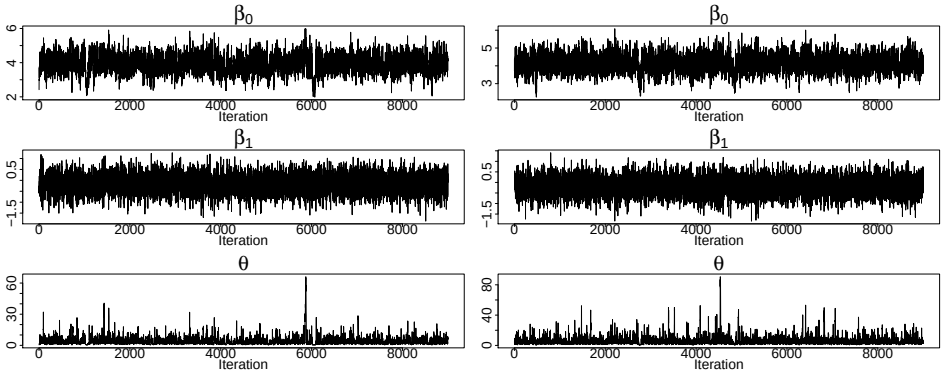


Figure D.21: AA Bone Marrow data. RME chains with Inv-Gauss $(\theta, 1)$ mixing under a trunc. exp. prior for c_v. Left panels: E(c_v)=1.25. Right panels: E(c_v)=2.

Table D.24: AA Bone Marrow data. Convergence diagnostics and ESS for RME chains with log-normal$(0, \theta)$ mixing.

Prior		E(c_v)=1.25			E(c_v)=1.5			E(c_v)=2			E(c_v)=5			E(c_v)=10		
for c_v		Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS	Gew.	HW	ESS
T. exp.	β_0	1.51	0.61	7861	-0.70	0.21	6995	-1.67	0.32	5623	-1.30	0.89	4537	1.28	0.44	3518
	β_1	-0.66	0.97	7608	0.38	0.64	7083	0.89	0.75	6015	1.25	0.31	4536	-1.10	0.37	3841
	θ	-0.15	0.99	1561	-0.56	0.28	2527	-1.56	0.38	2208	-0.61	0.89	3229	0.34	0.73	2712
Pareto	β_0	0.45	0.91	6634	-1.47	0.41	5032	-0.86	0.06	3971	-1.71	0.06	4211	0.10	3672	0.63
	β_1	-1.40	0.59	6637	0.00	0.78	5250	1.59	0.15	5689	-0.15	1.00	5136	-0.07	5085	0.85
	θ	-1.64	0.39	1303	-1.34	0.18	1240	-0.03	0.80	1946	-0.22	0.45	2285	0.12	1977	0.64

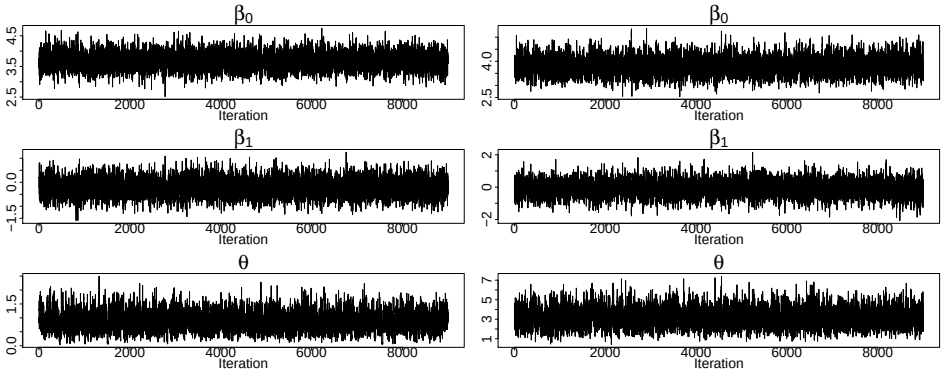


Figure D.22: AA Bone Marrow data. RME chains with log-normal $(0, \theta)$ mixing under a trunc. exp. prior for c_v. Left panels: E(c_v)=1.25. Right panels: E(c_v)=10.

D.3 Cerebral palsy dataset

Table D.25: Cerebral palsy data. For MCMC chains: total number of iteration (N), thinning period ($thin$), burning period ($burn$) and update period for λ_i's (Q).

Family	Model	N	$thin$	$burn$	Q
SMLN	all but log-Laplace and log-logistic	400,000	20	200,000	1
SMLN	log-Laplace	400,000	20	200,000	5
SMLN	log-logistic	400,000	20	200,000	20
RMW	No mixing	600,000	50	150,000	1
RMW	Exponential mixing	600,000	50	150,000	10
RMW	Gamma mixing	600,000	50	150,000	2
RMW	Inv-Gamma and Inv-Gauss mixing	1,200,000	100	300,000	5
RMW	Log-normal mixing	1,200,000	100	300,000	2

Table D.26: Cerebral palsy data. Convergence diag. and ESS for log-normal chains

	Jeffreys prior						Ind. Jeffreys prior					
	Point Observations			Set Observations			Point Observations			Set Observations		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.64	0.91	9515	-0.22	0.47	8510	1.22	0.26	9636	1.64	0.12	9515
β_1	0.84	0.97	4462	-1.13	0.35	4864	0.73	0.82	4786	0.93	0.51	5210
β_2	-0.35	0.69	7209	-0.72	0.27	7966	0.58	0.92	7704	-1.37	0.44	9109
β_3	0.42	0.47	9574	-1.65	0.70	10012	-0.17	0.32	10000	0.42	0.51	10000
β_4	-0.84	0.92	9677	0.99	0.40	9341	-1.08	0.42	10000	-1.62	0.11	10000
σ^2	0.42	0.81	4271	-0.64	0.47	4536	1.14	0.57	4333	0.95	0.50	4482

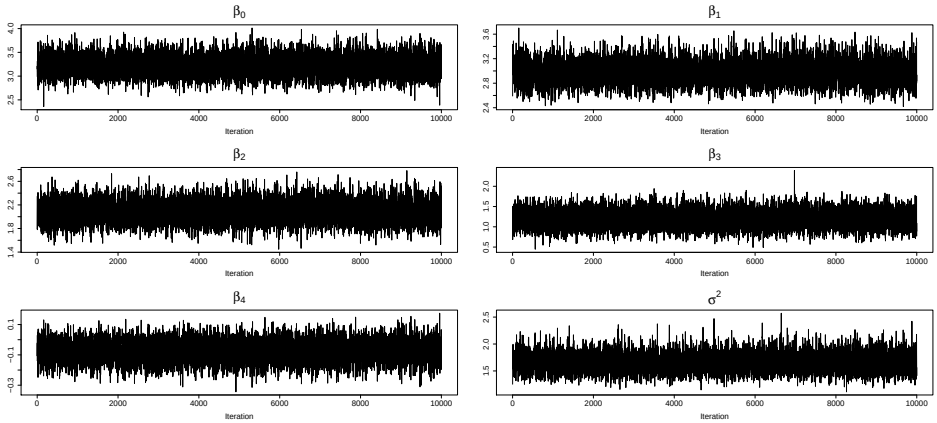


Figure D.23: Cerebral palsy data. Log-normal chains under the ind. Jeffreys prior (set observations).

Table D.27: Cerebral palsy data. Convergence diagnostics and ESS for log-Student's t chains

	Ind. Jeffreys prior		
	Geweke	HW	ESS
β_0	-0.69	0.64	9702
β_1	0.29	0.16	2617
β_2	-1.26	0.94	6144
β_3	0.48	0.60	9968
β_4	0.43	0.59	9632
σ^2	-0.77	0.83	468
ν	-0.74	0.99	125

Table D.28: Cerebral palsy data. Convergence diagnostics and ESS for log-Laplace chains

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.42	0.92	4036	0.92	0.68	3803
β_1	-0.77	0.70	314	0.03	0.48	329
β_2	0.07	0.88	1967	0.52	0.23	1853
β_3	-0.10	0.75	6212	-0.09	0.81	8181
β_4	-0.72	0.93	3880	-1.19	0.54	3285
σ^2	-0.65	0.71	568	0.20	0.29	625

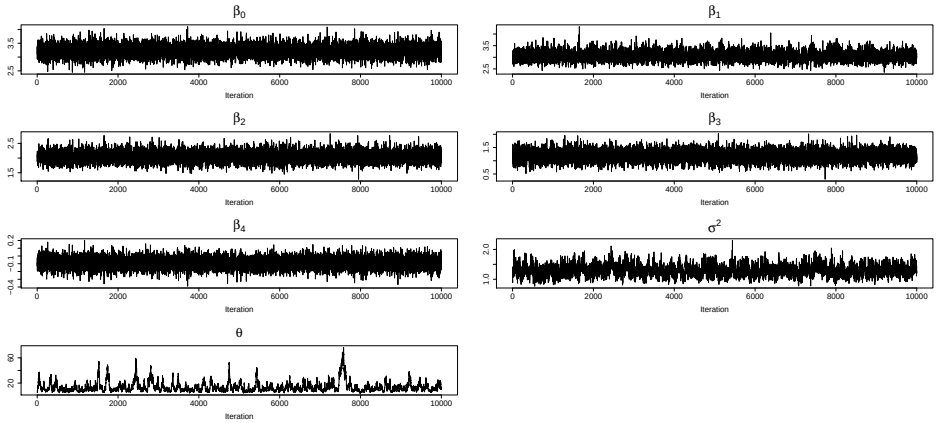


Figure D.24: Cerebral palsy data. Log-Student's t chains under the ind. Jeffreys prior (set observations).

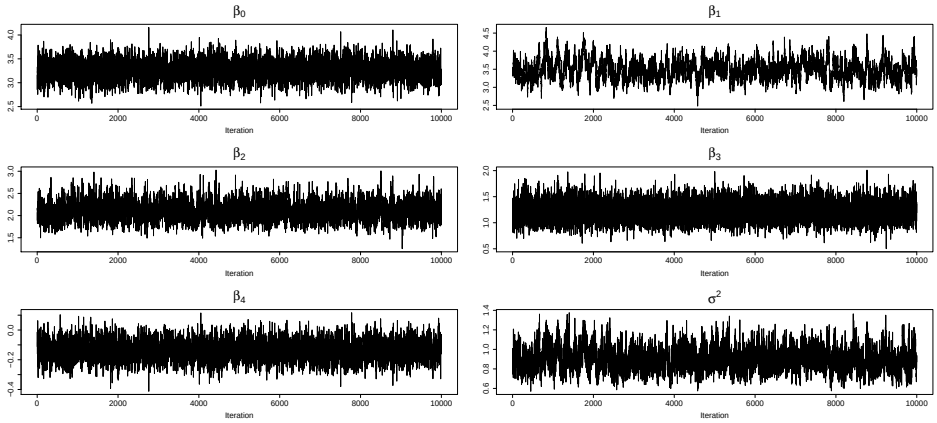


Figure D.25: Cerebral palsy data. Log-Laplace chains under the ind. Jeffreys prior (set observations).

Table D.29: Cerebral palsy data. Convergence diagnostics and ESS for log-exponential power chains

	Jeffreys prior			Ind. Jeffreys prior			Type I Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	1.12	0.39	588	-0.87	0.28	600	1.01	0.37	584
β_1	1.27	0.29	935	-0.41	1.00	1162	-1.80	0.39	942
β_2	0.45	0.86	2811	0.64	0.35	2964	-1.19	0.49	2790
β_3	1.63	0.14	6374	0.71	0.16	6851	-0.71	0.62	5997
β_4	-1.28	0.29	456	0.74	0.28	568	-0.92	0.33	568
σ^2	0.01	0.79	972	0.80	0.09	956	-0.49	0.87	989
α	-0.55	0.34	960	0.89	0.08	1031	0.11	0.70	941

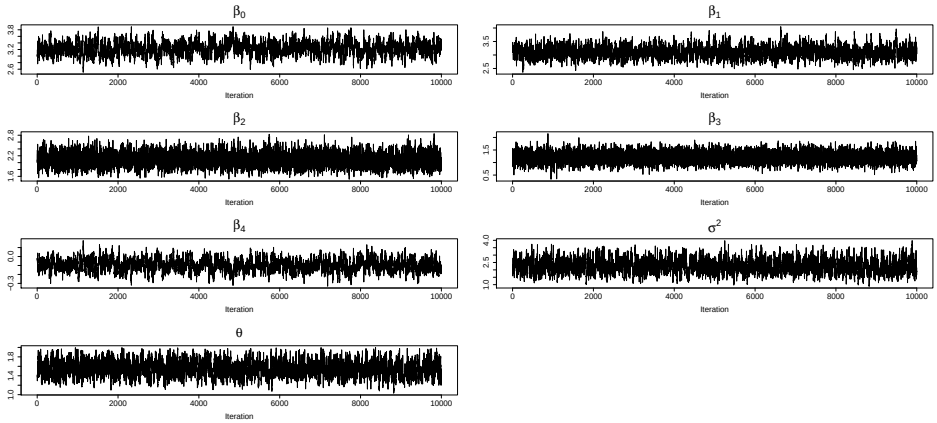


Figure D.26: Cerebral palsy data. Log-exponential power chains under the ind. Jeffreys prior (set observations).

Table D.30: Cerebral palsy data. Convergence diagnostics and ESS for log-logistic chains

	Jeffreys prior			Ind. Jeffreys prior		
	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.90	0.62	7557	-1.31	0.51	7224
β_1	0.89	0.85	2426	-0.13	0.50	2324
β_2	0.26	0.66	5582	-0.64	0.49	5784
β_3	1.11	0.94	8661	1.88	0.51	8122
β_4	-0.53	0.76	7293	1.10	0.60	6907
σ^2	0.59	0.93	2807	0.25	0.32	2851

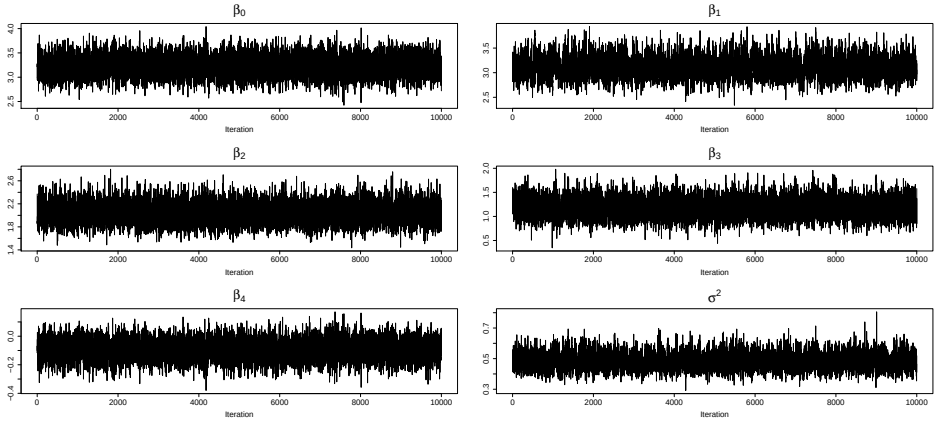


Figure D.27: Cerebral palsy data. Log-logistic chains under the ind. Jeffreys prior (set observations).

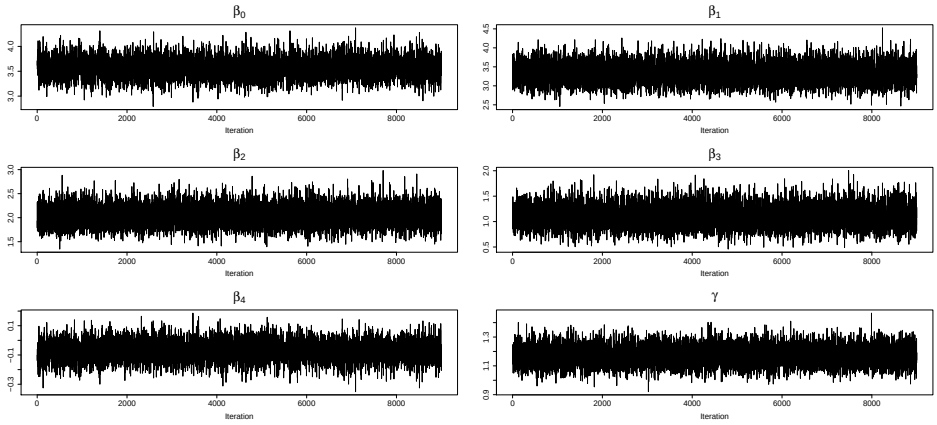


Figure D.28: Cerebral palsy data. Weibull chains under Gamma(4,1) prior for γ.

Table D.31: Cerebral palsy data. Convergence diagnostics and ESS for Weibull chains.

	$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.27	0.65	4089	-1.88	0.46	3907	-0.78	0.78	3761
β_1	1.37	0.25	8522	0.38	0.74	8985	0.81	0.40	8522
β_2	-0.35	0.56	9000	0.00	0.75	9000	1.28	0.08	9000
β_3	1.41	0.57	8957	1.14	0.31	8728	1.40	0.82	9000
β_4	-0.29	0.72	4094	1.73	0.39	3970	0.56	0.72	3865
γ	-1.20	0.22	9000	-0.71	0.40	9000	-1.03	0.68	9000

Table D.32: Cerebral palsy data. Convergence diagnostics and ESS for RMW chains with exponential(1) mixing.

	$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
β_0	0.15	0.36	2548	0.18	0.98	2486	0.89	0.23	2372
β_1	0.29	0.81	8651	-0.86	0.09	9000	-0.06	0.85	8561
β_2	-0.01	0.99	7941	-0.97	0.92	8918	0.77	0.33	8208
β_3	0.05	0.30	8568	0.84	0.53	9000	0.71	0.49	8675
β_4	-0.26	0.35	2553	-0.14	0.98	2525	-0.94	0.19	2342
γ	1.27	0.25	9126	-0.07	0.32	9000	0.16	0.66	8539

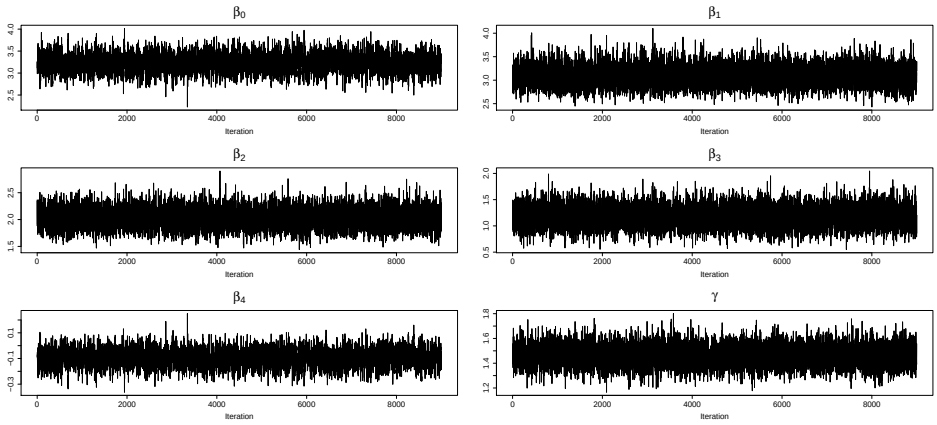


Figure D.29: Cerebral palsy data. RMW chains with exponential(1) mixing under Gamma(4,1) prior for γ.

Table D.33: Cerebral palsy data. Convergence diagnostics and ESS for RMW chains with Gamma(θ, θ) mixing.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	-0.19	0.77	1843	-1.27	0.41	2494	1.17	0.24	2584	0.13	0.76	2439	0.82	0.91	2618	-0.18	0.36	2713
	β_1	0.71	0.68	3418	-0.73	0.33	3260	-0.68	0.94	4270	1.01	0.35	3824	-0.68	0.37	3137	-0.33	0.86	3759
	β_2	1.02	0.46	7745	0.45	0.75	7782	-0.61	0.53	9000	0.46	0.67	7117	0.03	0.60	5437	-1.34	0.46	5638
	β_3	0.76	0.70	8427	-0.75	0.84	9000	0.77	0.19	9296	0.54	0.41	8682	-0.26	0.48	8702	-1.09	0.19	9000
	β_4	0.76	0.94	2595	0.63	0.64	2830	-1.33	0.35	2752	0.05	0.78	2802	-1.01	0.78	2811	0.12	0.61	2882
	γ	-1.21	0.83	1031	0.67	0.50	910	0.51	0.95	1451	-0.46	0.63	2061	0.62	0.42	1694	0.65	0.45	1608
	θ	0.02	0.61	235	-0.72	0.89	272	-0.78	0.35	357	0.85	0.58	925	-1.65	0.22	1053	-0.60	0.67	976
Pareto	β_0	-0.01	0.72	2605	-0.45	0.65	2055	-1.18	0.43	2612	0.83	0.89	2708	-0.27	0.28	2594	-0.02	0.88	2819
	β_1	-0.31	0.94	4209	0.29	0.88	3011	-0.60	0.44	4444	-0.14	0.72	3304	-0.46	0.79	3685	-0.28	0.65	3217
	β_2	-0.20	0.85	8059	0.37	0.99	8100	-0.35	0.72	8059	-1.22	0.90	7937	0.06	0.45	6781	-0.52	0.98	7398
	β_3	-0.40	0.41	9000	-0.39	0.85	9000	-1.15	0.25	8513	-1.26	0.89	9000	0.32	0.39	9000	-1.34	0.29	8797
	β_4	-0.49	0.63	2833	0.37	0.75	2831	1.25	0.15	3085	-0.62	0.56	2760	-0.39	0.57	2921	0.09	0.82	2946
	γ	1.14	0.98	1050	0.10	0.63	669	0.84	0.51	1126	0.03	0.88	1304	1.02	0.66	1181	0.34	0.84	1443
	θ	-1.47	0.89	154	-0.45	0.45	172	-1.00	0.94	152	0.53	0.78	355	-0.84	0.52	131	-0.11	0.99	249

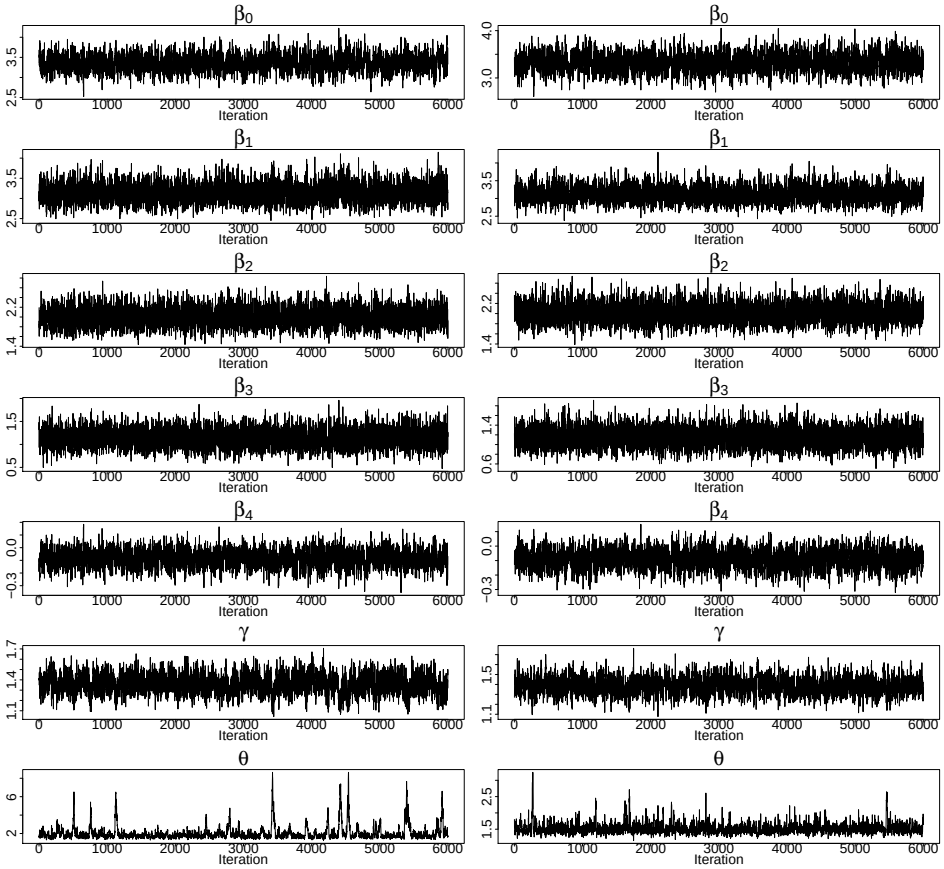


Figure D.30: Cerebral palsy data. RMW chains with Gamma(θ, θ) mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (right panels).

Table D.34: Cerebral palsy data. Convergence diagnostics and ESS for RMW chains with Inv-Gamma$(\theta, 1)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	-0.89	0.98	218	-0.95	0.25	102	-0.91	0.62	232	0.38	0.98	254	1.27	0.47	142	-0.27	0.30	218
	β_1	1.30	0.35	9000	1.08	0.37	8435	-0.61	0.47	9000	-0.03	0.90	9000	-0.99	0.53	9000	0.51	0.52	8481
	β_2	1.45	0.17	9000	-0.16	0.14	8621	-1.35	0.76	9000	-1.13	0.44	9000	-1.02	0.87	9000	-0.07	0.97	8707
	β_3	0.27	0.95	9000	-0.20	0.97	9000	-1.56	0.13	8699	-0.86	0.90	9443	0.68	0.68	9000	-1.36	0.44	9000
	β_4	0.95	0.56	4774	1.49	0.10	4999	-0.73	0.39	4715	-0.22	0.97	4103	-1.10	0.51	4752	-0.78	0.91	4557
	γ	-1.18	0.96	272	-0.41	0.56	193	-0.29	0.66	345	0.17	0.97	328	0.53	0.57	216	-0.36	0.36	319
	θ	0.46	0.98	146	0.94	0.24	80	1.41	0.74	134	-0.35	0.97	151	-1.30	0.60	92	0.51	0.23	108
Pareto	β_0	1.30	0.44	105	0.28	0.83	217	-1.78	0.47	138	0.22	0.71	202	1.09	0.81	189	1.09	0.34	123
	β_1	-0.91	0.33	8674	-0.66	0.73	9000	-0.48	0.47	9000	0.58	0.56	8733	0.13	0.24	9000	0.28	0.10	9000
	β_2	-0.54	0.93	9000	-0.19	0.33	9000	-1.80	0.07	9000	0.79	0.23	8720	0.36	0.32	9000	-0.60	0.31	9000
	β_3	0.87	0.53	9339	0.42	0.66	9000	-1.43	0.68	8686	-0.11	0.67	9369	-0.78	0.65	9000	0.10	0.41	9000
	β_4	-0.39	0.83	4636	0.46	0.99	4724	1.03	0.60	4659	-0.78	0.87	3979	0.48	0.38	4493	-0.13	0.42	4786
	γ	1.22	0.71	221	0.19	0.90	374	-1.33	0.41	261	-0.47	0.80	279	1.11	0.95	320	1.12	0.60	229
	θ	-1.13	0.22	43	-0.42	0.81	126	1.65	0.61	89	-0.51	0.69	98	-1.40	0.88	118	-0.90	0.14	66

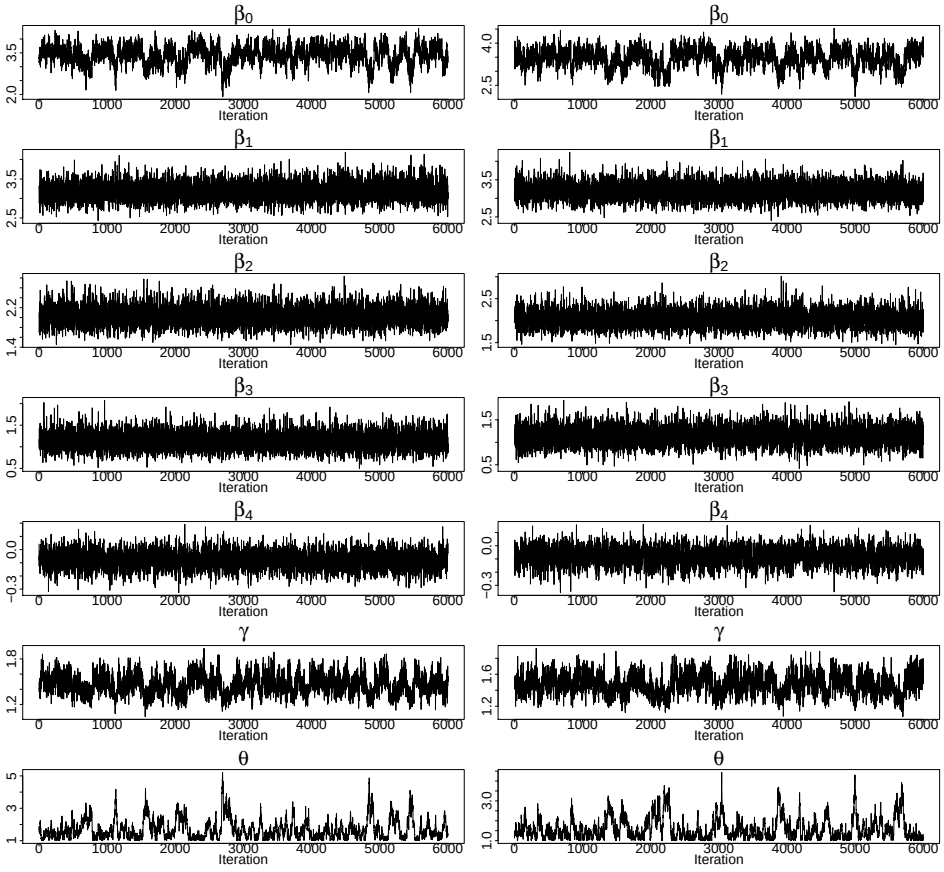


Figure D.31: Cerebral palsy data. RMW chains with Inv-Gamma(θ, 1) mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (left panels).

Table D.35: Cerebral palsy data. Convergence diagnostics and ESS for RMW chains with Inv-Gaussian$(\theta, 1)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	0.32	0.63	268	1.64	0.67	167	0.09	0.34	279	-1.36	0.89	699	-0.06	0.31	474	-1.79	0.40	430
	β_1	-0.91	0.86	5303	-1.82	0.53	3117	-0.36	0.67	3160	0.55	0.54	7622	-0.26	0.47	5279	0.75	0.43	5200
	β_2	-0.52	0.38	9000	-0.73	0.67	9000	-0.30	0.48	8723	0.15	0.82	9000	0.01	0.44	9000	-1.15	0.37	9000
	β_3	-0.54	0.21	8624	-0.21	0.05	9000	-0.18	0.87	9000	-0.19	0.37	9000	-0.53	0.28	9000	-0.95	0.32	9000
	β_4	0.26	0.31	4114	0.71	0.95	4509	-1.00	0.06	4332	0.20	0.11	4138	0.08	0.84	4245	0.52	0.93	4049
	γ	-0.59	0.48	329	0.88	0.74	204	-0.69	0.29	221	-0.53	0.32	393	-0.55	0.25	404	-0.62	0.74	300
	θ	-0.81	0.43	342	-0.67	0.60	108	-0.89	0.59	103	0.04	0.23	303	-0.98	0.22	347	-0.11	0.95	322
Pareto	β_0	-0.15	0.25	212	0.37	0.93	137	-1.93	0.43	285	-0.68	1.00	278	1.56	0.28	197	0.42	0.86	369
	β_1	0.53	0.16	3987	-1.36	0.70	3091	1.21	0.41	4830	0.60	0.94	7044	-0.86	0.51	2537	-0.35	0.68	5162
	β_2	-0.62	0.63	9000	-0.57	0.21	9000	1.32	0.23	9000	-0.77	0.44	9000	0.36	0.75	9000	1.29	0.58	9000
	β_3	1.74	0.07	7738	1.38	0.14	8982	-0.18	0.52	9000	-0.53	0.88	8548	0.58	0.16	8660	0.10	0.22	9339
	β_4	-0.11	0.06	4269	-0.44	0.99	4334	-0.27	0.78	4007	-0.33	0.76	4010	0.19	0.47	4418	0.58	0.17	3837
	γ	-0.06	0.14	219	-0.85	0.55	179	-1.81	0.43	286	-0.26	0.95	293	0.94	0.40	230	0.45	0.60	236
	θ	-0.43	0.34	171	-1.52	0.45	150	-1.61	0.46	168	-0.25	0.90	232	0.55	0.49	229	-0.29	0.29	97

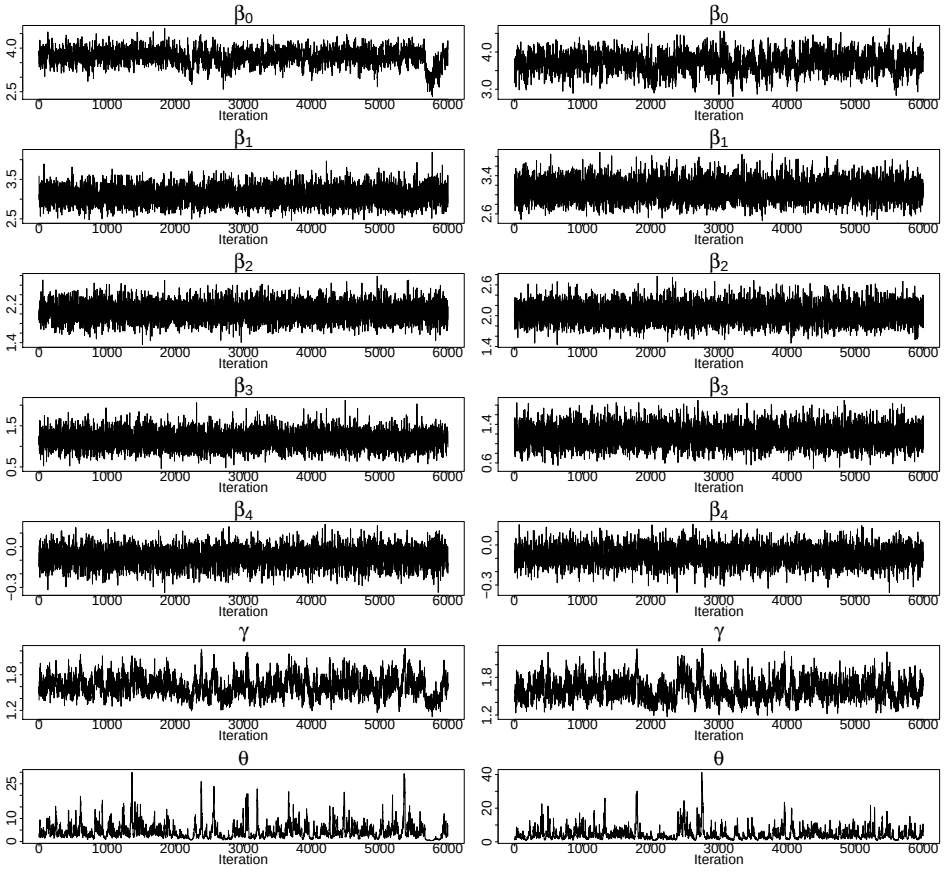


Figure D.32: Cerebral palsy data. RMW chains with Inv-Gaussian$(\theta, 1)$ mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (right panels).

Table D.36: Cerebral palsy data. Convergence diagnostics and ESS for RMW chains with log-normal$(0, \theta)$ mixing and a truncated exponential prior for c_v.

		E(c_v)=1.5									E(c_v)=5								
		$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)			$\gamma \sim$Gamma(4,1)			$\gamma \sim$Gamma(1,1)			$\gamma \sim$Gamma(0.01,0.01)		
Prior for c_v		Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS	Geweke	HW	ESS
Trunc. exp.	β_0	1.15	0.07	3242	-0.46	0.07	3266	0.02	0.08	3722	1.21	0.41	3305	0.49	0.11	3321	-0.62	0.25	3244
	β_1	-0.28	0.20	4466	-0.49	0.35	1252	1.10	0.51	2502	1.28	0.17	4059	1.09	0.65	3631	0.14	0.07	2646
	β_2	1.06	0.67	6289	-0.75	0.71	6725	-0.62	0.32	5924	-1.06	0.49	5113	-0.12	0.25	6465	0.54	0.85	5994
	β_3	-0.17	0.63	7347	-0.85	0.69	7835	0.20	0.89	7672	-0.48	0.27	7612	1.34	0.07	8284	0.59	0.62	7172
	β_4	-1.09	0.24	3317	0.77	0.41	3270	0.02	0.18	3774	-0.42	0.87	3321	-0.16	0.09	3452	0.15	0.75	3299
	γ	-0.52	0.51	96	-0.19	0.37	47	-1.45	0.66	121	-1.96	0.53	101	-1.08	0.59	82	0.78	0.31	80
	θ	-0.40	0.50	126	0.07	0.48	52	-1.45	0.63	135	-1.57	0.71	114	-1.56	0.49	106	1.02	0.49	86
Pareto	β_0	-0.06	0.23	3041	-1.19	0.54	3735	0.79	0.65	3782	-1.68	0.31	3680	0.03	0.12	1845	-1.11	0.35	3112
	β_1	0.01	0.56	2947	-1.87	0.48	1679	1.14	0.67	1735	-0.20	0.35	2712	-1.59	0.76	2063	-0.87	0.44	1928
	β_2	-1.13	0.89	6901	0.79	0.12	7953	0.82	0.90	6946	1.20	0.18	6987	-0.21	0.65	6782	1.30	0.54	6708
	β_3	1.41	0.23	8240	1.36	0.19	8417	-0.87	0.94	7487	0.29	0.25	8013	0.26	0.53	6832	0.42	0.77	7644
	β_4	-0.08	0.30	3672	0.21	0.14	3816	-0.31	0.87	3805	1.28	0.26	3826	0.03	0.12	2778	1.20	0.17	3374
	γ	0.29	0.89	103	1.19	0.85	119	-1.07	0.96	99	0.24	0.82	97	-0.24	0.49	49	0.26	0.77	72
	θ	0.57	0.67	123	1.42	0.76	129	-0.90	0.97	95	0.32	0.86	100	0.08	0.40	58	0.45	0.77	91

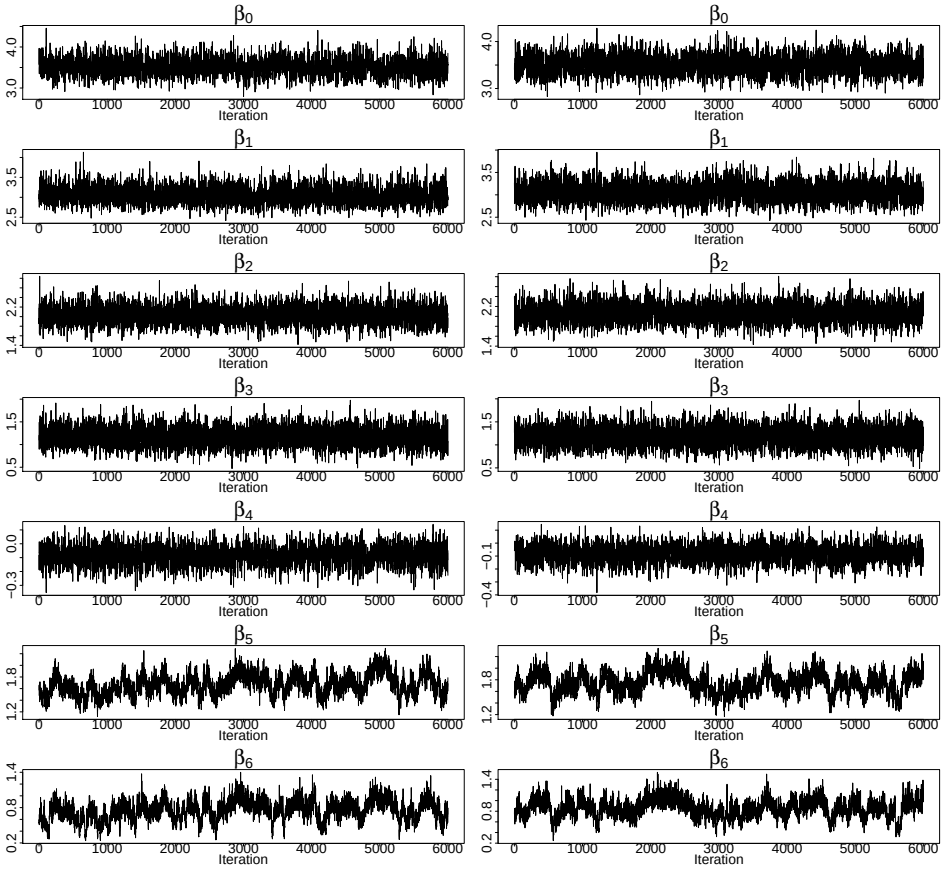


Figure D.33: Cerebral palsy data. RMW chains with log-normal$(0, \theta)$ mixing under Gamma(4,1) prior for γ and a truncated exponential prior for c_v with E(c_v)=1.5 (left panels) and E(c_v)=5 (right panels).

Appendix E

Appendix for Chapter 5

Figures E.1 to E.8 of this Appendix summarize a descriptive analysis of the PUC dataset. In terms of population compositions, these Figures confirm strong levels of heterogeneity between different programmes of the PUC. As described in Section 5.2, this suggest the need of modelling each programme independently. In addition, Figures E.9 to E.16 display the continuous component associated to the posterior distribution of the regression coefficients for some of the science programmes where dropouts and late graduation are more often seen.

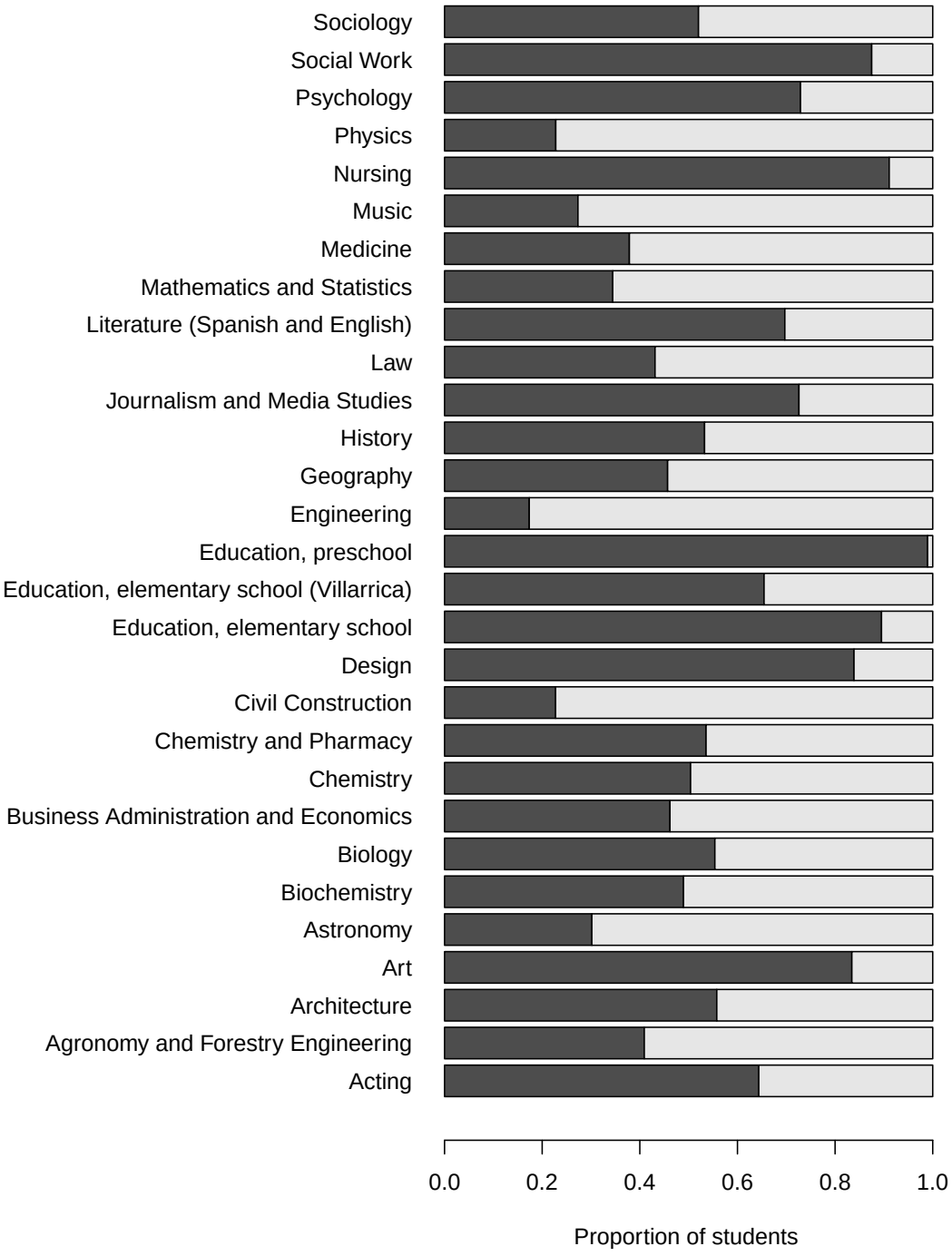


Figure E.1: PUC dataset. Distribution of students according to sex. The lighter area represents the proportion of male students.

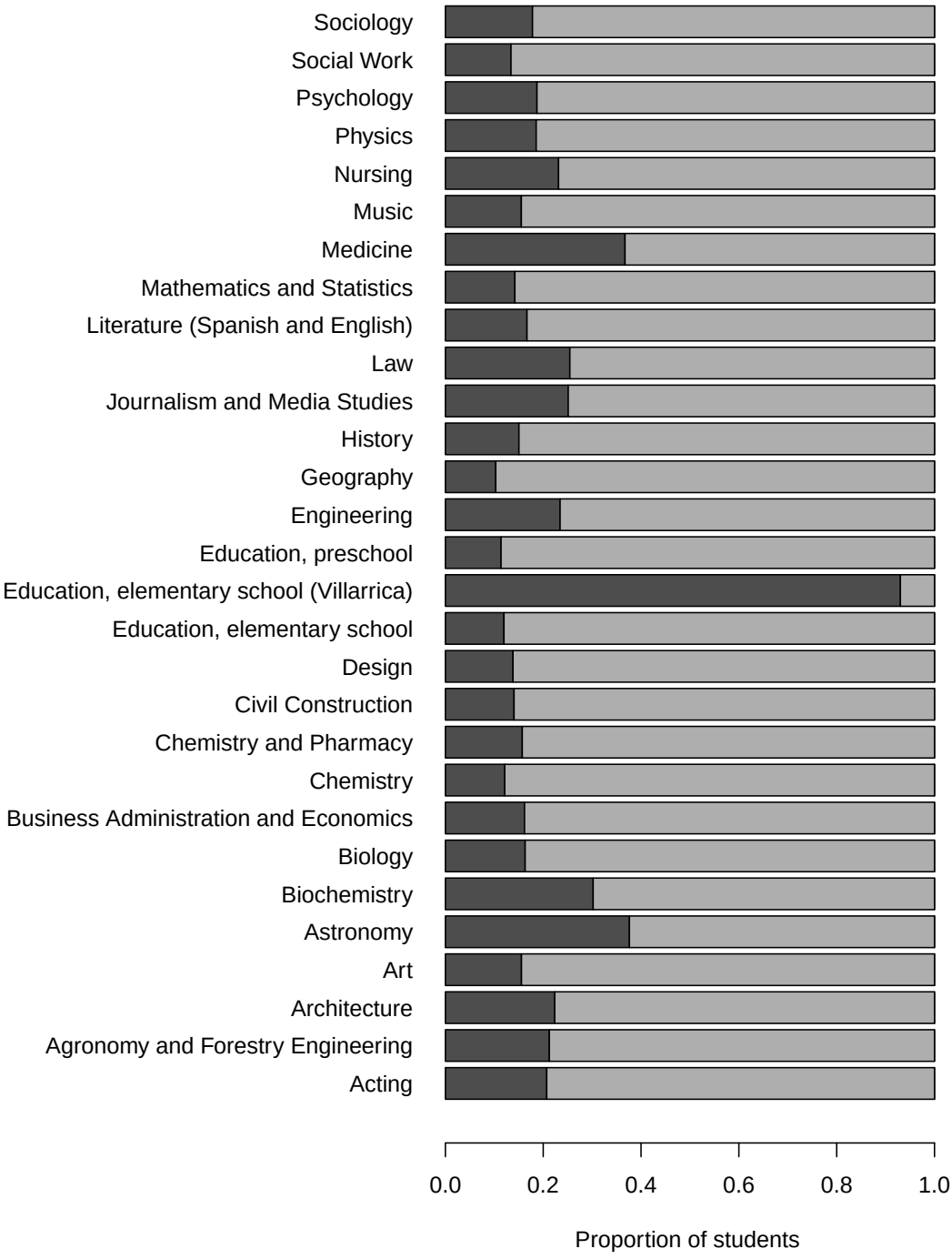


Figure E.2: PUC dataset. Distribution of students according to region of residence. The lighter area represents the proportion of students from the Metropolitan area.

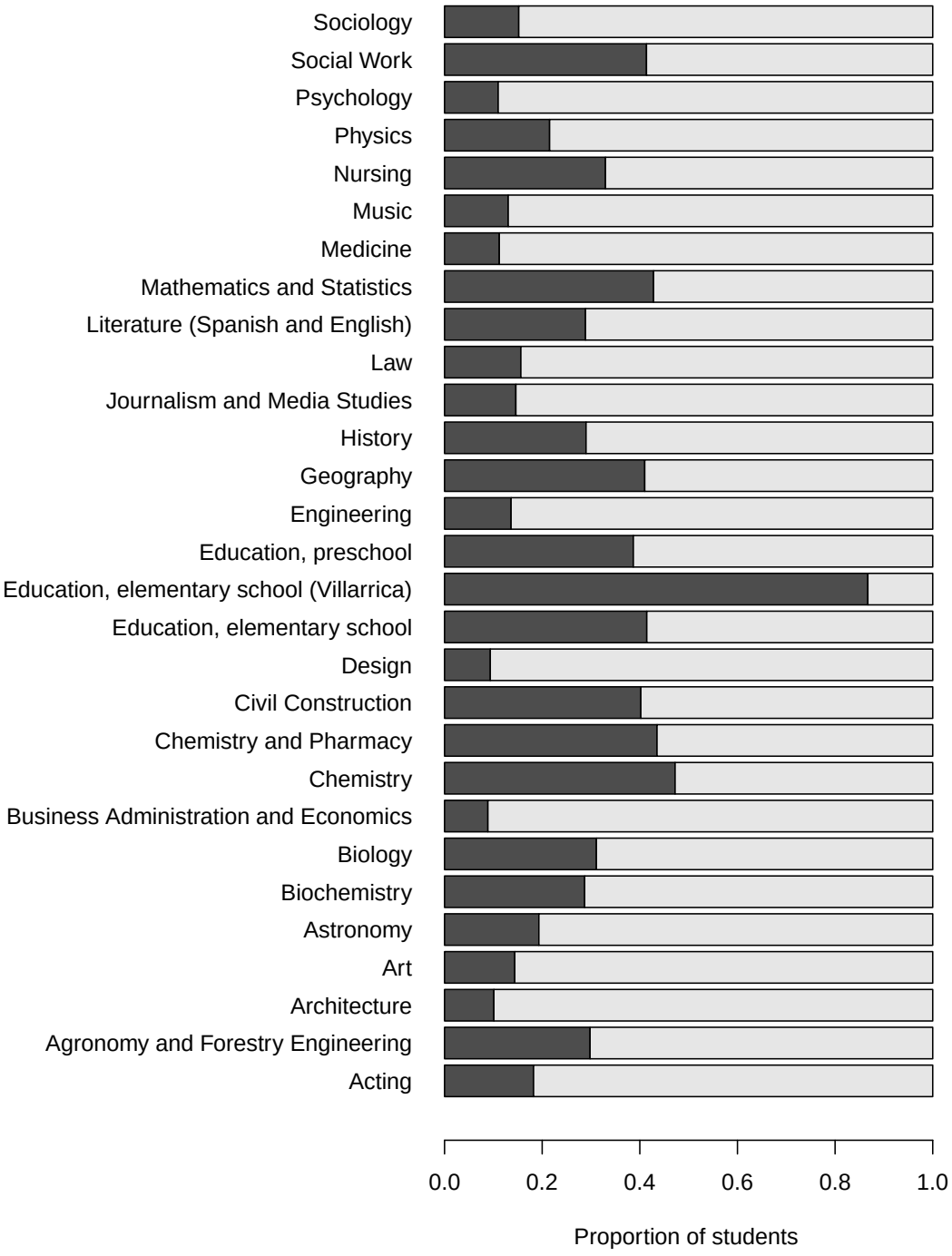


Figure E.3: PUC dataset. Distribution of students according to educational level of the parents. The lighter area represents the proportion of students for which at least one of the parents has a higher degree (university or technical).

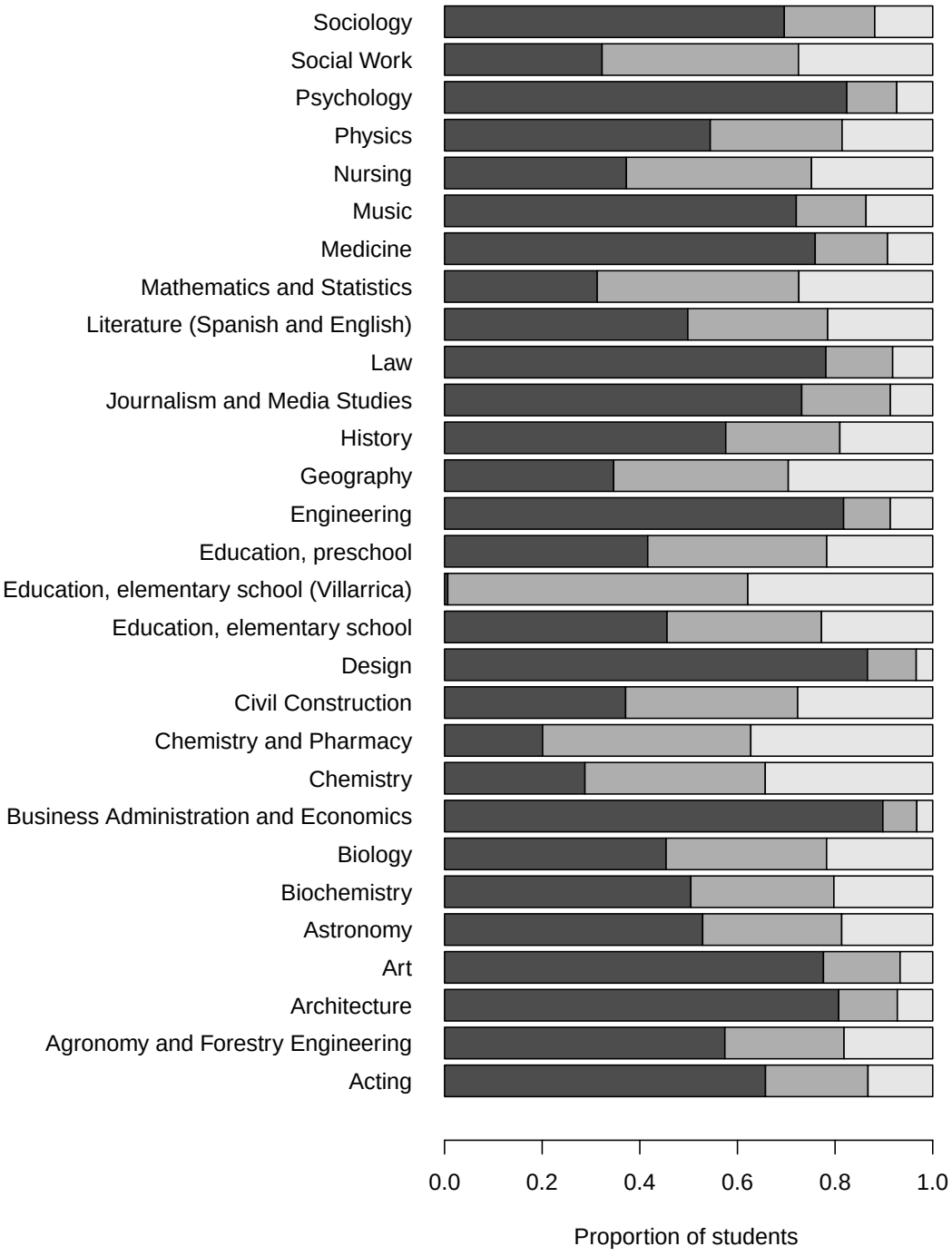


Figure E.4: PUC dataset. Distribution of students according to type of high school. From darkest to lightest, colored areas represent the proportion of students whose high school are: private, subsidized private and public, respectively.

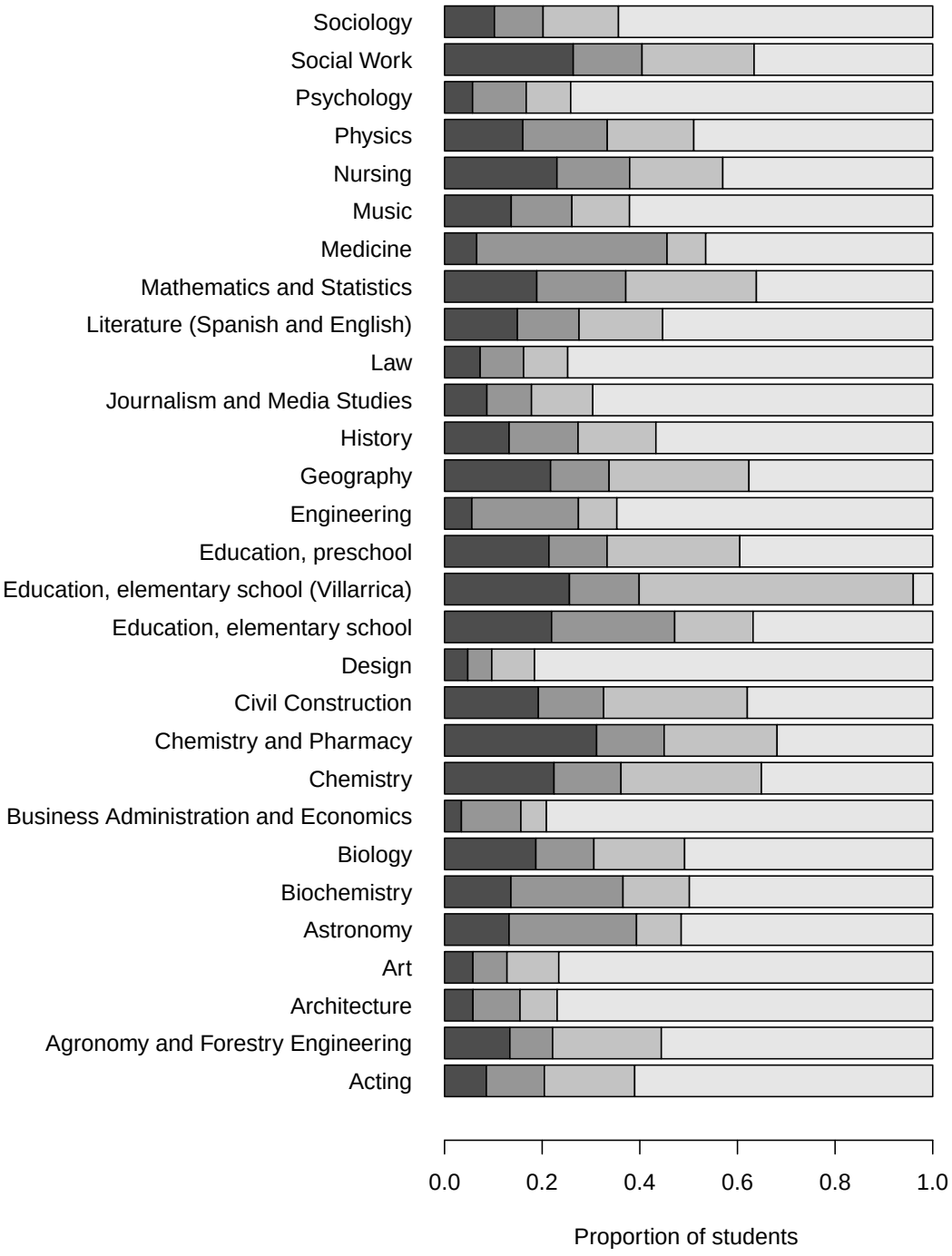


Figure E.5: PUC dataset. Distribution of students according to funding. From darkest to lightest, colored areas represent the proportion of students who have: scholarship and loan, scholarship only, loan only and no aid, respectively.

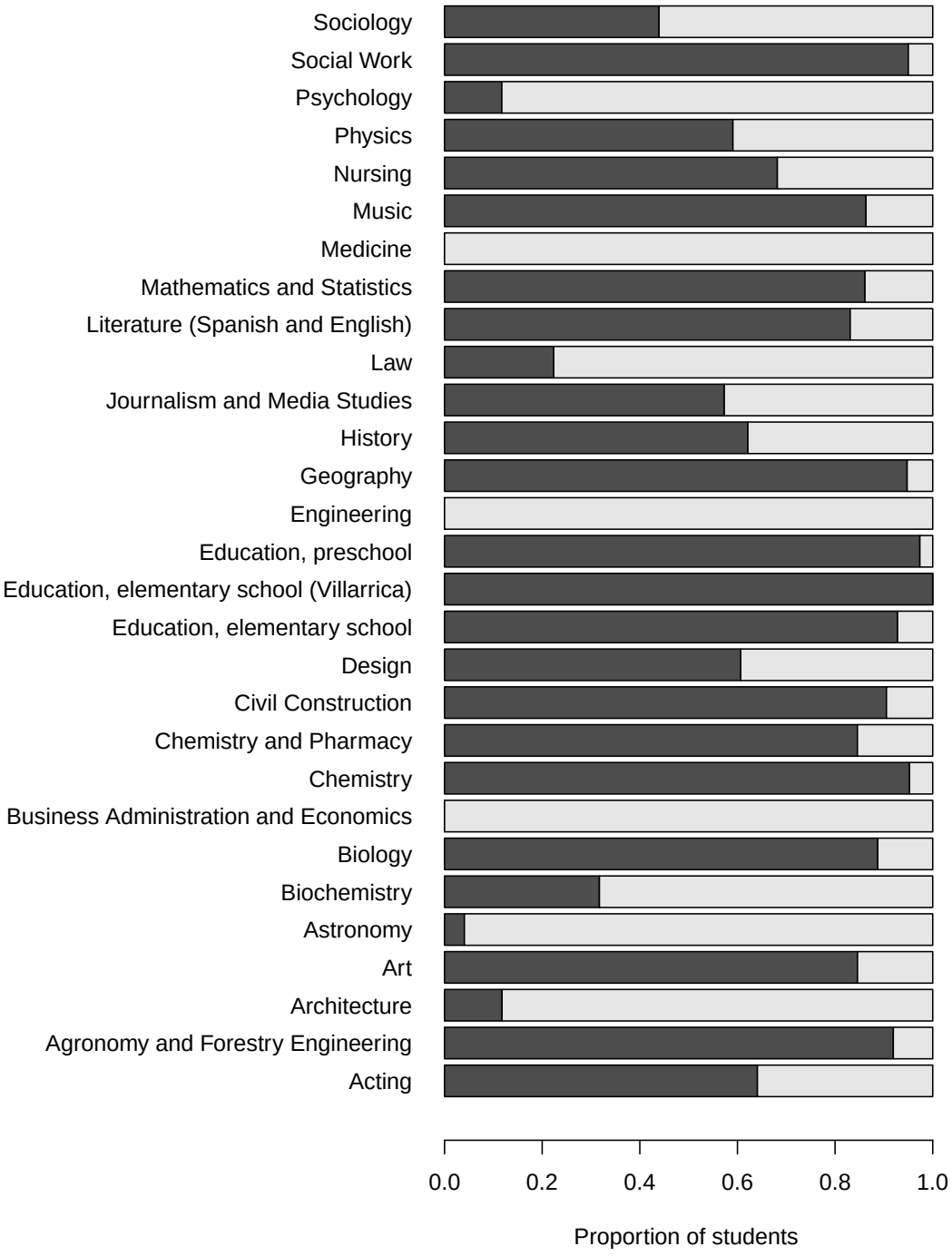


Figure E.6: PUC dataset. Distribution of students according to their selection score. The lighter area represents the proportion of students with a selection score of 700 or more, which is typically considered a high value (the maximum possible score is 850).

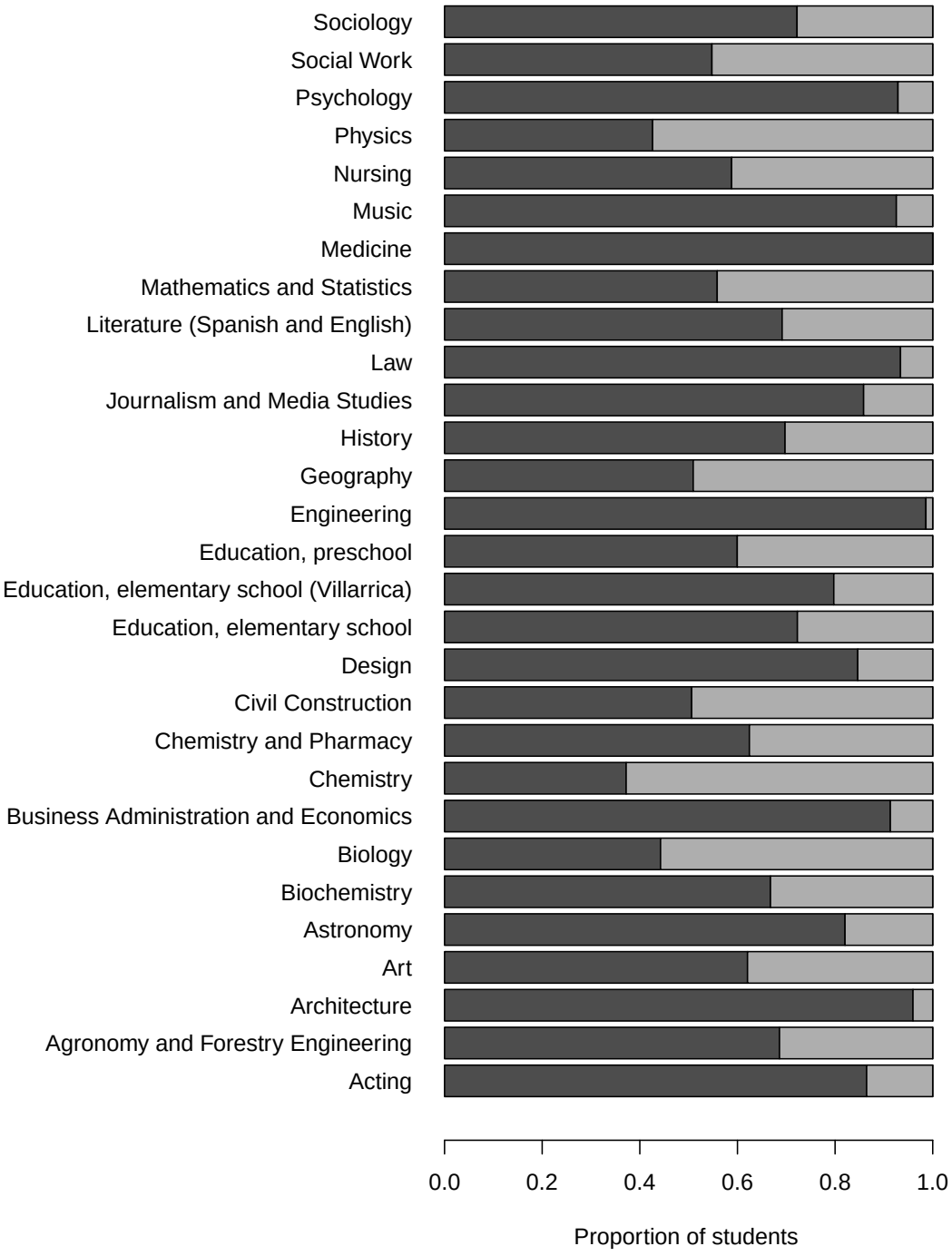


Figure E.7: PUC dataset. Distribution of students according to their application preference. The lighter area represents the proportion of students who applied in second or lower preference to their current degree.

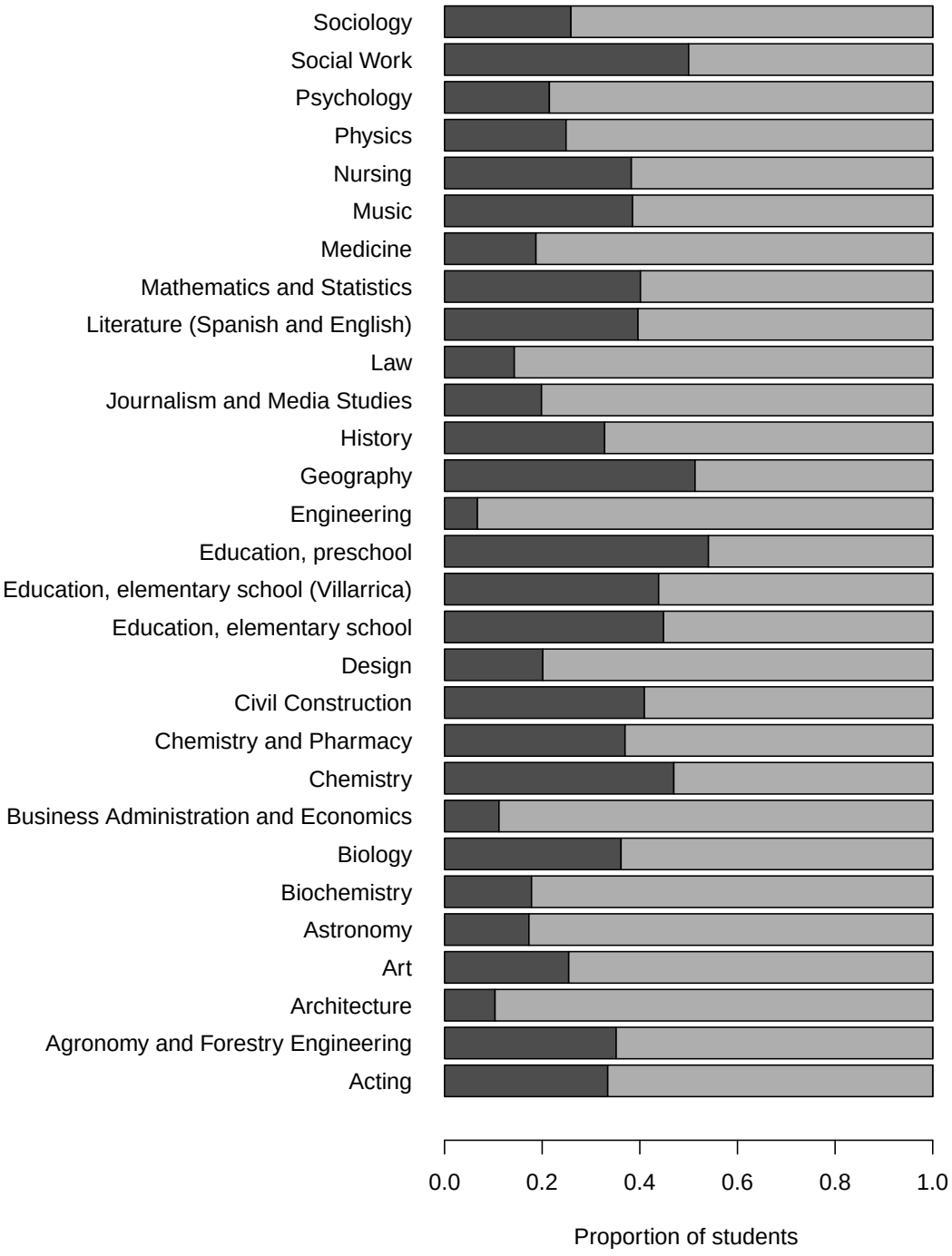


Figure E.8: PUC dataset. Distribution of students according to the gap between High School graduation and admission to PUC. The lighter area represents the proportion of students who have no gap.

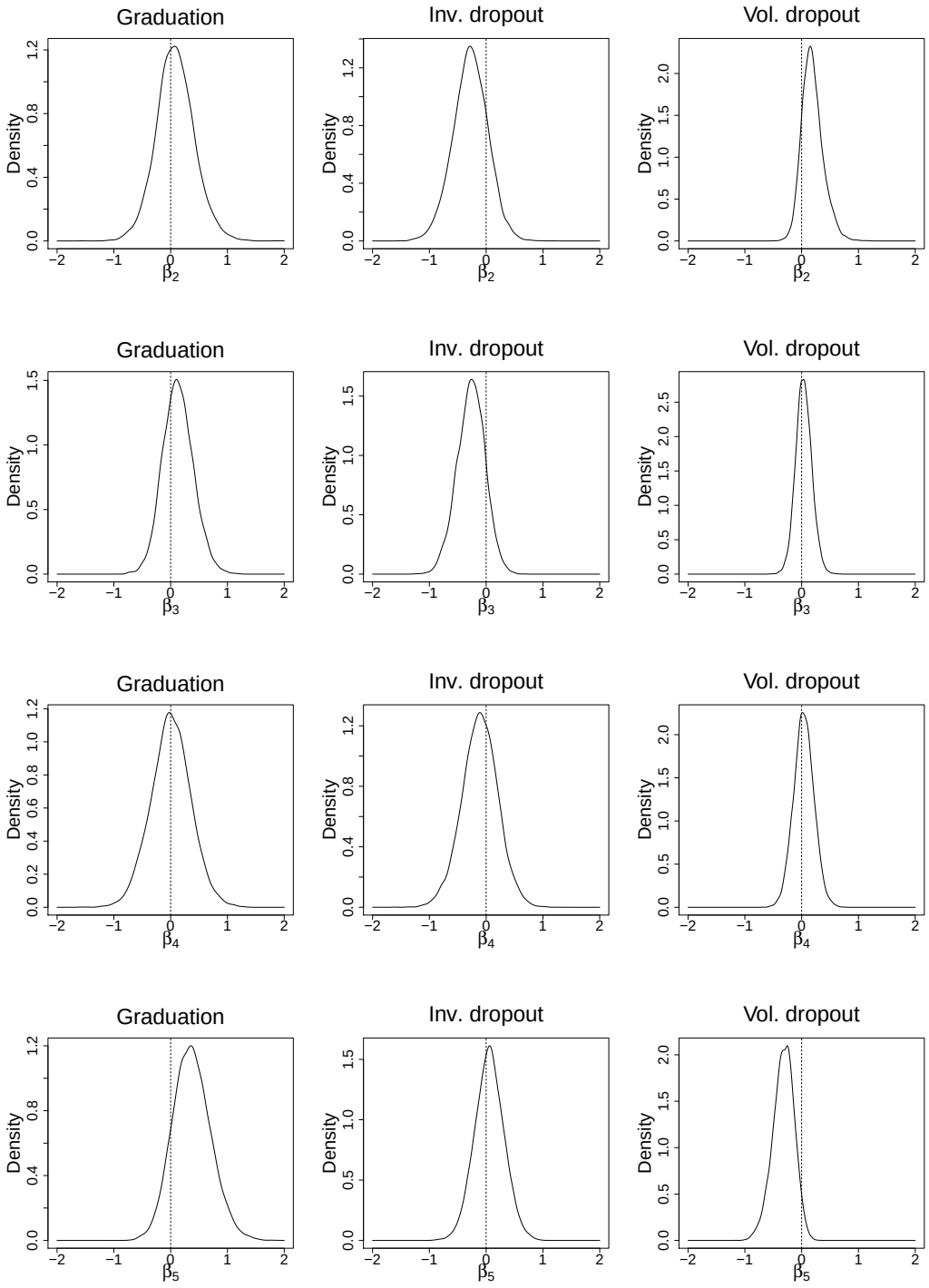


Figure E.9: PUC dataset. For Chemistry students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: region (β_2), parents' education - with degree (β_3), high school - private (β_4) and high school - subsidized private (β_5). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

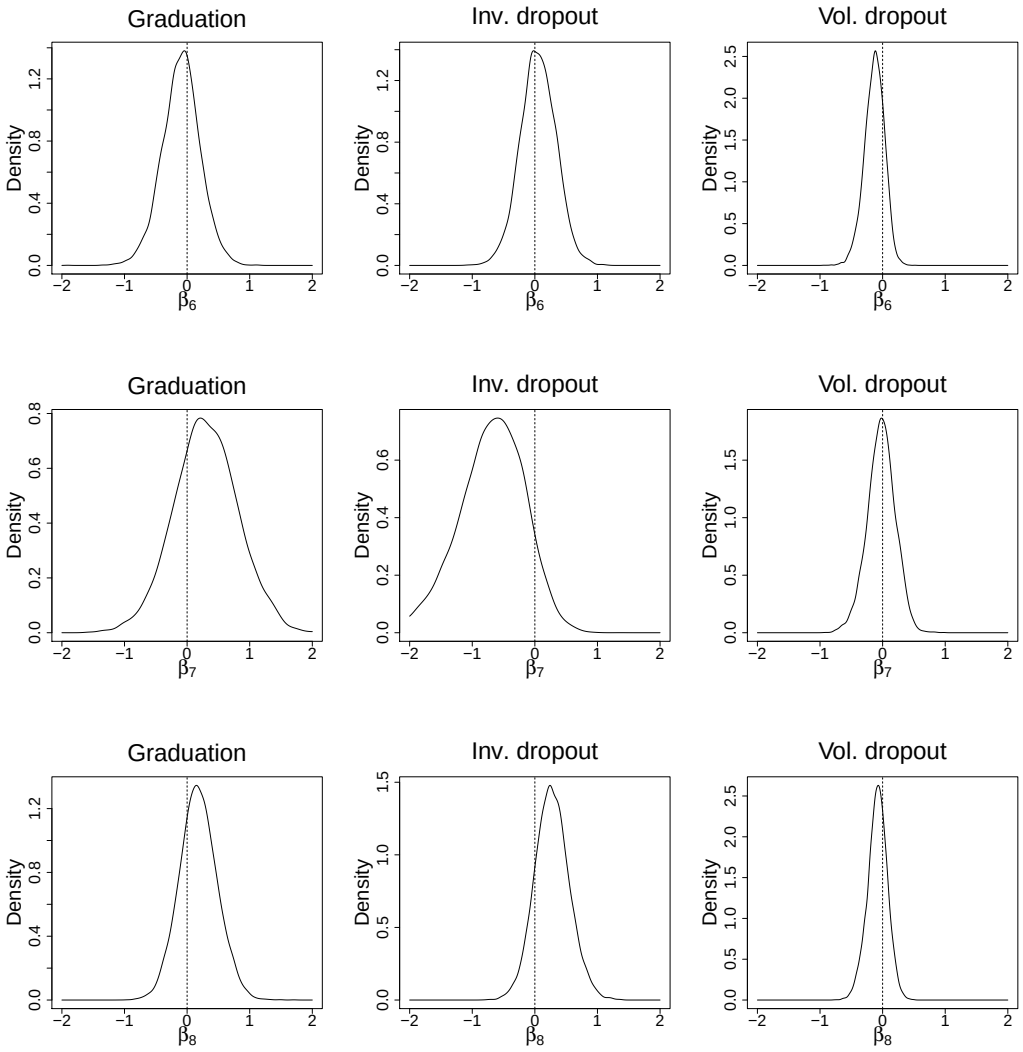


Figure E.10: PUC dataset. For Chemistry students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

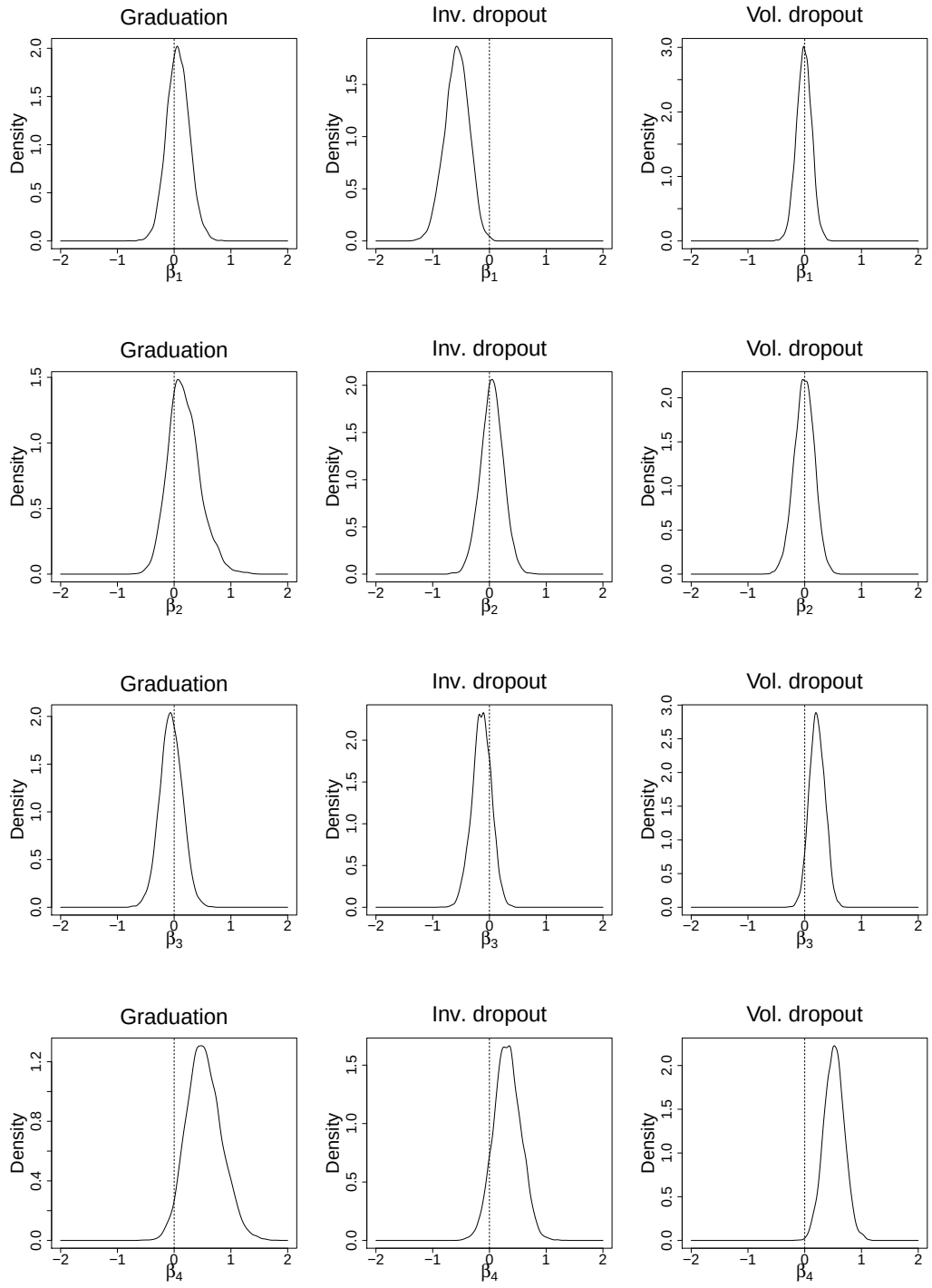


Figure E.11: PUC dataset. For Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: sex (β_1), region (β_2), parents' education - with degree (β_3) and high school - private (β_4). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

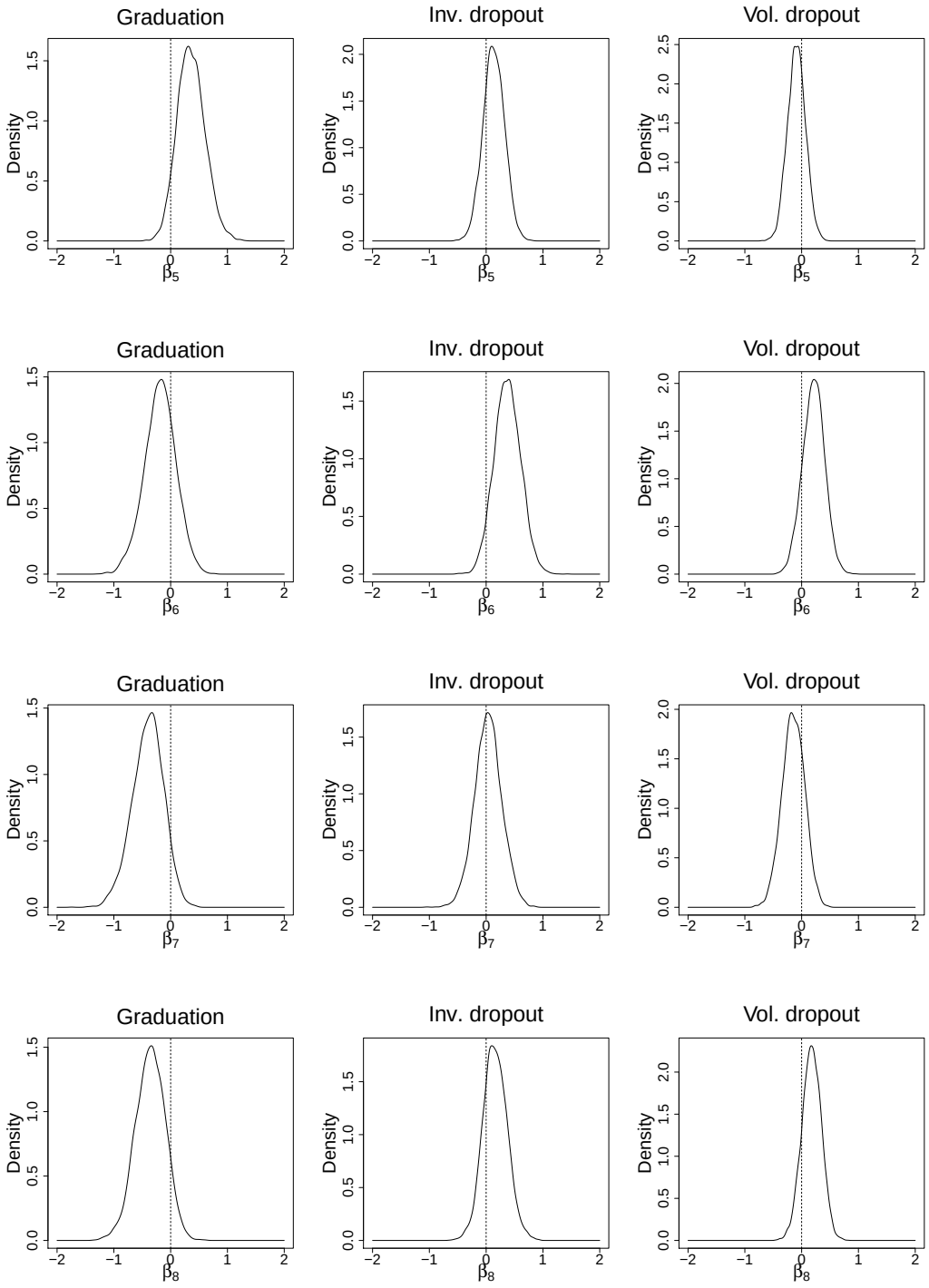


Figure E.12: PUC dataset. For Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: high school - subsidized private (β_5), funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

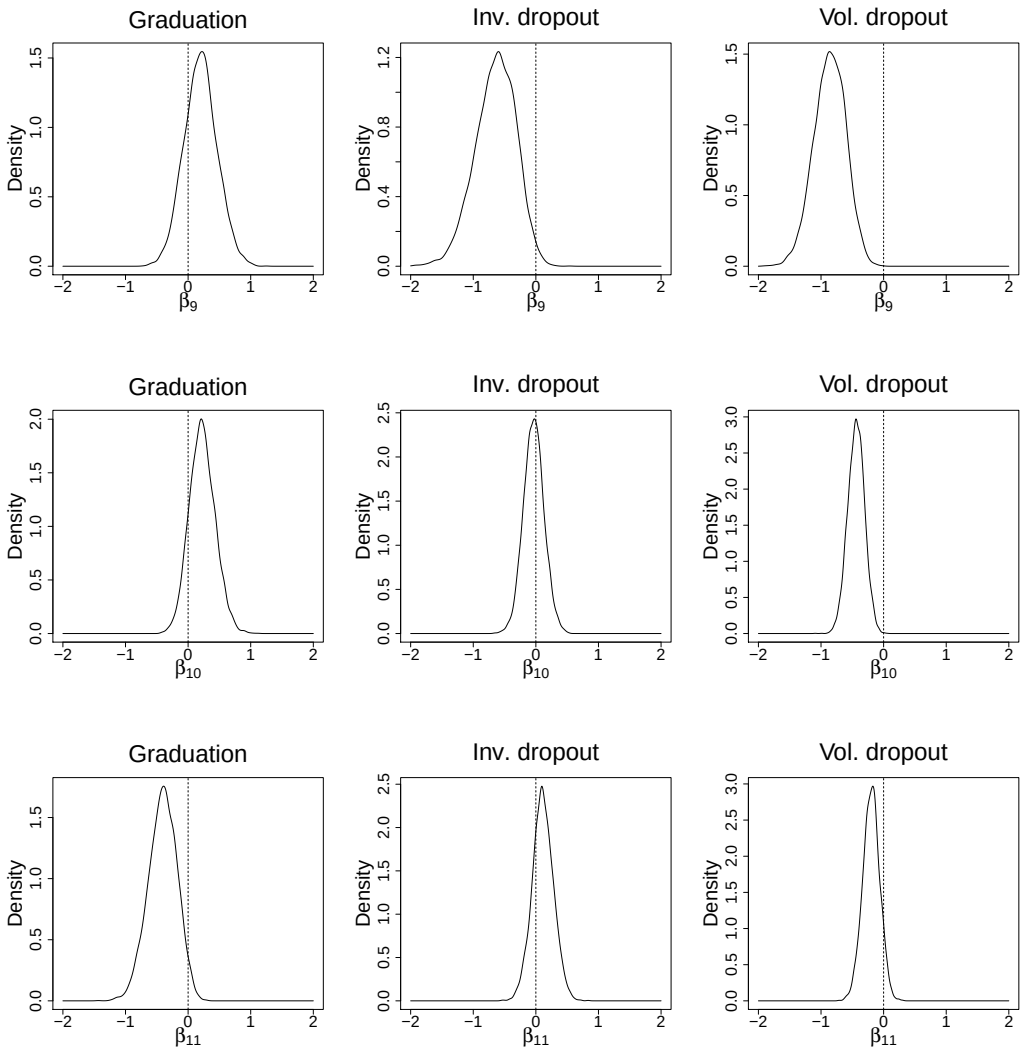


Figure E.13: PUC dataset. For Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: ranking (β_9), preference (β_{10}) and gap (β_{11}). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

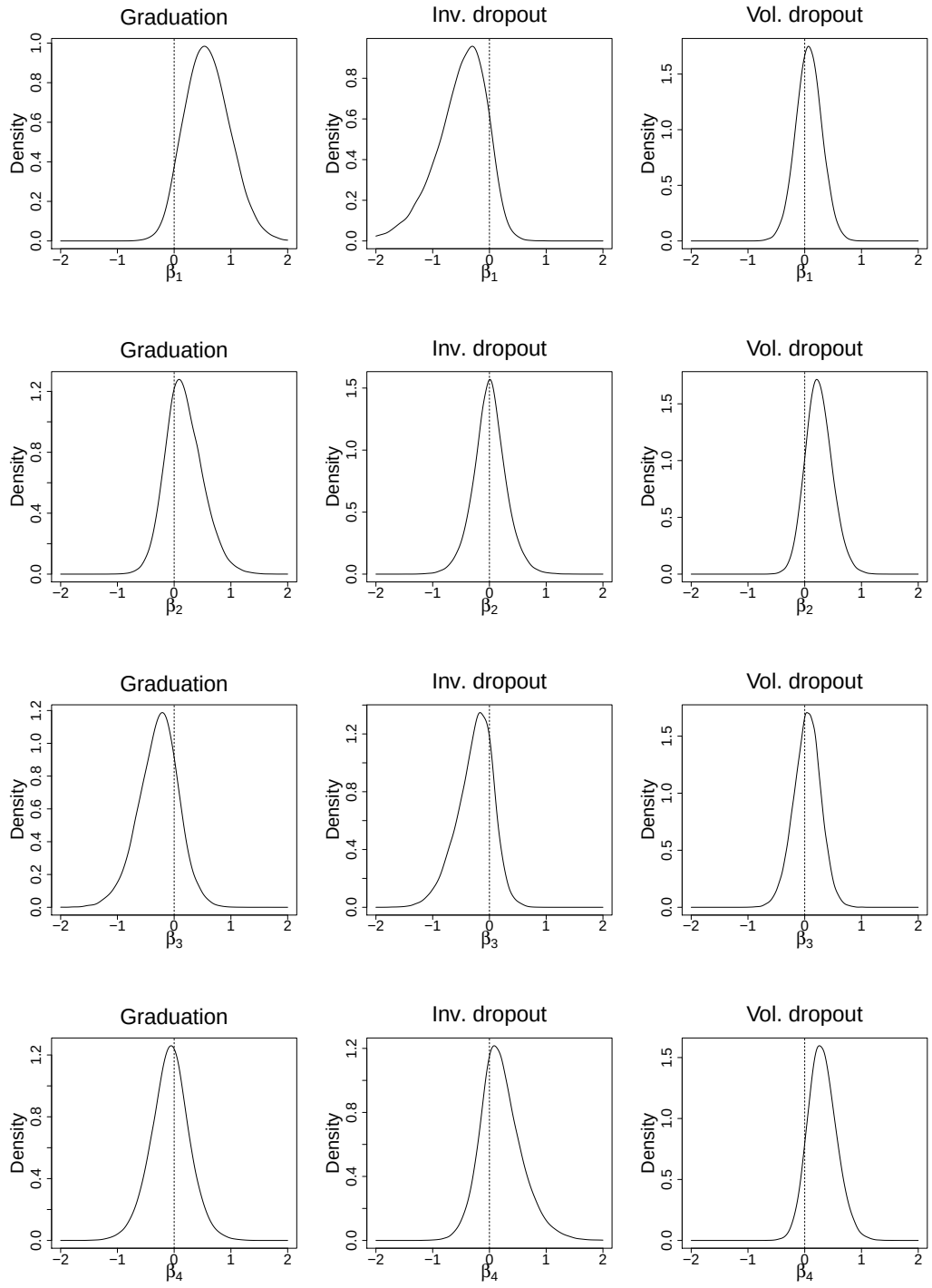


Figure E.14: PUC dataset. For Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: sex (β_1), region (β_2), parents' education - with degree (β_3) and high school - private (β_4). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

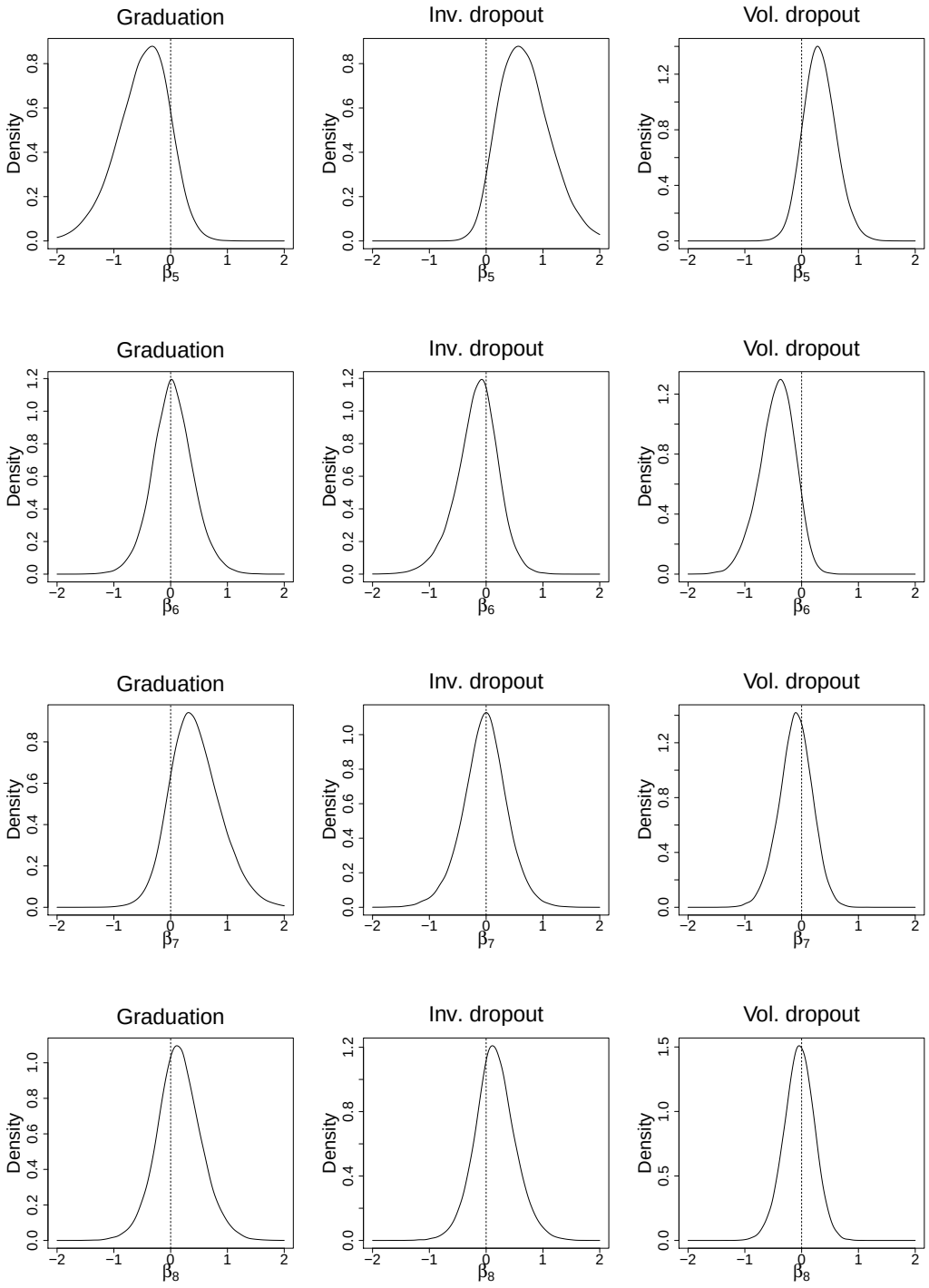


Figure E.15: PUC dataset. For Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: high school - subsidized private (β_5), funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

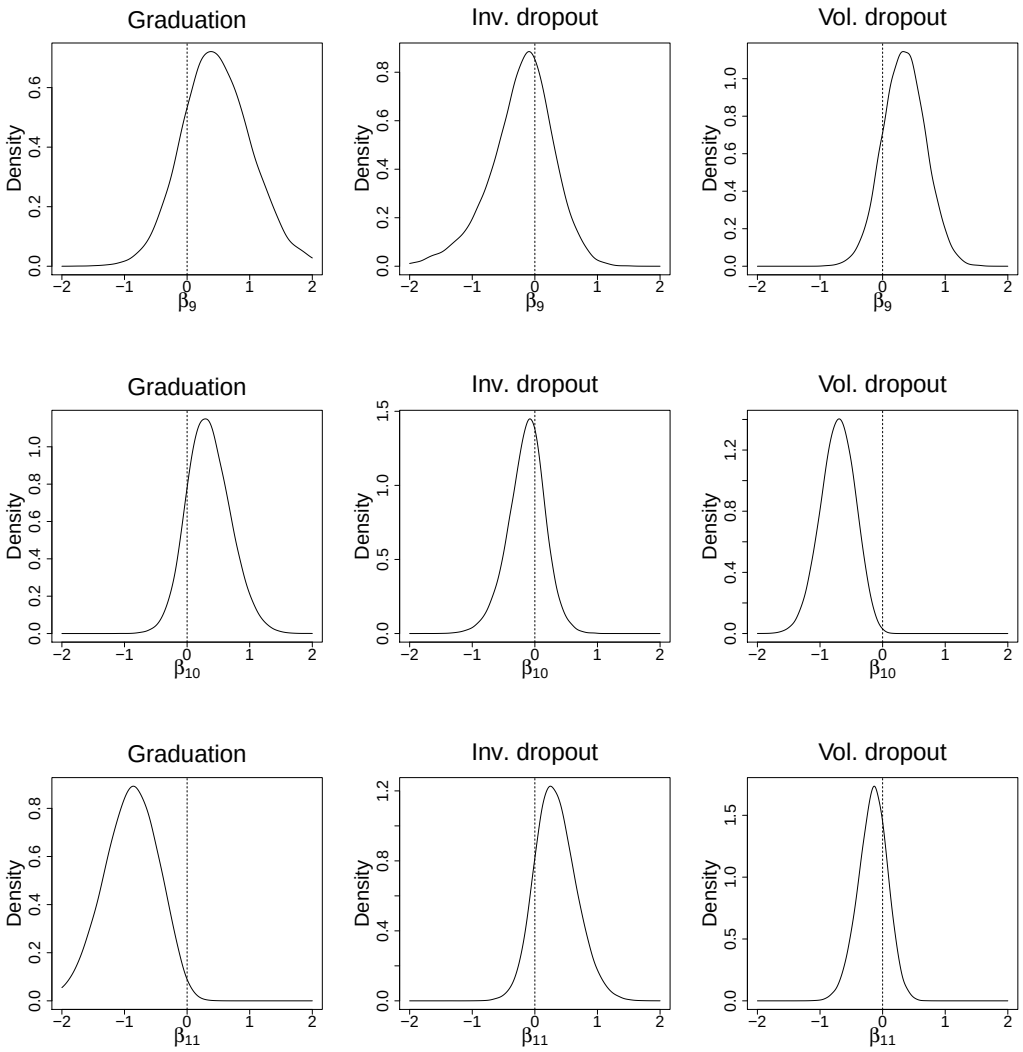


Figure E.16: PUC dataset. For Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: ranking (β_9), preference (β_{10}) and gap (β_{11}). A vertical dashed line was drawn at zero for reference. The prior in (5.24) was adopted for the model space.

Appendix F

Probability density functions

Gamma: Gamma(a, b)

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1}\, e^{-bx} \qquad x > 0, a > 0, b > 0. \tag{F.1}$$

Generalized Inverse Gaussian: GIG(a, b, p)

$$f_X(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\{-\frac{1}{2}(ax + b/x)\} \qquad x > 0, a > 0, b > 0, p \in \mathbb{R}. \tag{F.2}$$

Inverse Gamma: Inv-Gamma(a, b)

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{-a-1}\, e^{-\frac{b}{x}} \qquad x > 0, a > 0, b > 0. \tag{F.3}$$

Inverse Gaussian: Inv-Gauss(a, b)

$$f_X(x) = \sqrt{\frac{a}{2\pi}}\, e^{\frac{a}{b}} x^{-\frac{3}{2}} \exp\{-\frac{1}{2}(\frac{a}{b^2}x + a/x)\} \qquad x > 0, a > 0, b > 0. \tag{F.4}$$

Multivariate normal: Normal$_p(a, B)$

$$f_X(x) = (2\pi)^{-\frac{p}{2}} (\det(B))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-a)'B^{-1}(x-a)\right\} \qquad x \in \mathbb{R}^p, a \in \mathbb{R}^p \text{ and} \tag{F.5}$$

B is a symmetric positive semi-definite matrix of dimension p.

Multivariate Student t: Student $t_p(d, a, B)$

$$f_X(x) = \frac{\Gamma((d+p)/2)}{\Gamma(d/2)} (d\pi)^{-\frac{p}{2}} (\det(B))^{-\frac{1}{2}} \left[1 + \frac{1}{d}(x-a)'B^{-1}(x-a)\right]^{-(d+p)/2}, \tag{F.6}$$

$x \in \mathbb{R}^p, d > 0, a \in \mathbb{R}^p$ and B is a symmetric positive semi-definite matrix of dimension p.

Normal: Normal(a, b^2)

$$f_X(x) = (2\pi b^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2b^2}(x-a)^2\right\}, \qquad x \in \mathbb{R}, a \in \mathbb{R}, b^2 > 0. \tag{F.7}$$

If $X \sim$ Normal(a, b^2) then $Y = e^X \sim$ log-normal (a, b^2).

Student t: Student $t(d, a, b^2)$

$$f_X(x) = \frac{\Gamma(d/2 + 1/2)}{\Gamma(d/2)\sqrt{b^2 d\pi}} \left[1 + \frac{(x-a)^2}{b^2 d}\right]^{-(\frac{d}{2}+\frac{1}{2})}, \qquad x \in \mathbb{R}, a \in \mathbb{R}, b^2 > 0, c > 0. \tag{F.8}$$

Reducing to a Cauchy distribution when $d = 1$. If $X \sim$ Student $t(a, b^2)$ then $Y = e^X \sim$ log-Student $t(a, b^2)$.

Uniform: Unif(a, b)

$$f_X(x) = \frac{1}{b-a}, \qquad x \in [a, b], a \in \mathbb{R}, b \in \mathbb{R}, a < b. \tag{F.9}$$

Weibull: Weibull(a, b)

$$f_X(x) = abx^{a-1} e^{-bx^a}, \qquad x > 0, a > 0, b > 0 \tag{F.10}$$

Reducing to the exponential and Rayleigh distributions when $b = 1$ and $b = 2$, respectively. If $X \sim$ Weibull(a, b) then $Y = \log(X) \sim$ Gumbel$(-a^{-1}\log(b), a^{-1})$.

Bibliography

O.O. Aalen. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, 2:951–972, 1992.

J.H. Abbring and G.J. Van Den Berg. The unobserved heterogeneity distribution in duration analysis. *Biometrika*, 94:87–99, 2007.

A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

J.E. Anderson and T.A. Louis. Survival analysis using a scale change random effects model. *Journal of the American Statistical Association*, 90:669–679, 1995.

F. Angeletti, E. Bertin, and P. Abry. Critical moment definition and estimation, for finite size observation of log-exponential-power law random variables. *Signal Processing*, 92:2848–2865, 2012.

E. Arias Ortis and C. Dehon. The roads to success: Analyzing dropout and degree completion at university. Working Papers ECARES 2011-025, ULB - Universite Libre de Bruxelles, 2011.

A. Azzalini, T. Capello, and S. Kotz. Log-skew-normal and log-skew-t distributions as models for family income data. *Journal of Income Distribution*, 11:11–20, 2003.

N. Balakrishnan, V. Leiva, A. Sanhueza, and F. E. V. Labra. Estimation in the Birnbaum-Saunders distribution based on scale-mixture of normals and the EM-algorithm. *Sort: Statistics and Operations Research Transactions*, 33:171–192, 2009.

S. Banerjee, M.M. Wall, and B.P. Carlin. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, 4: 123–142, 2003.

T. Banerjee, M.H. Chen, D.K. Dey, and S. Kim. Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis*, 13:241–260, 2007.

M. Barros, G.A. Paula, and V. Leiva. A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, 14: 316–332, 2008.

T. Bayes. An essay toward solving a problem in the doctrine of chances. published posthumously in. *Philosophical Transactions of the Royal Society of London*, 53: 370–418, 1763.

J.P. Bean. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12:155–187, 1980.

R.E. Beard. Note on some mathematical mortality models. In G.E.W. Wolstenholme and M. O'Connor, editors, *Ciba Foundation Symposium-The Lifespan of Animals (Colloquia on Ageing), Vol. 5*, pages 302–311. John Wiley & Sons, 1959.

J.O. Berger and D. Sun. Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association*, 88:1412–1418, 1993.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 1st edition, 2000.

J. Beyersmann, A. Allignol, and M. Schumacher. *Competing Risks and Multistate Models with R*. Use R! Springer New York, 2012.

Z.W. Birnbaum and S.C. Saunders. A new family of life distributions. *Journal of Applied Probability*, 6:319–327, 1969.

M.D. Branco, M.G. Genton, and B. Liseo. Objective Bayesian analysis of skew-t distributions. *Scandinavian Journal of Statistics*, 40:63–85, 2012.

D.T. Cassidy, M.J. Hamp, and R. Ouyed. Pricing european options with a log Student's t-distribution: a Gosset formula. *Physica A: Statistical Mechanics and its Applications*, 389:5736–5748, 2009.

M.-H. Chen, J.G Ibrahim, and S. Kim. Properties and implementation of Jeffreys's prior in binomial regression models. *Journal of the American Statistical Association*, 103:1659–1664, 2008.

M.H. Chen and Q.M. Shao. Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, 7:607–630, 1997.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.

H. Chipman, E. I. George, and R.E. McCulloch. *The Practical Implementation of Bayesian Model Selection*, volume 38 of *Lecture Notes–Monograph Series*, pages 65–116. Institute of Mathematical Statistics, Beachwood, OH, 2001.

H. Cho, J. G. Ibrahim, D. Sinha, and H. Zhu. Bayesian case influence diagnostics for survival models. *Biometrics*, 65:116–124, 2009.

D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151, 1978.

R. Clerici, A. Giraldo, and S. Meggiolaro. The determinants of academic outcomes in a competing risks approach: evidence from Italy. *Studies in Higher Education*, 2014. URL `DOI: 10.1080/03075079.2013.878835`.

D. Collett. *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, second edition, 2003.

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

D.R. Cox. Some remarks on the analysis of survival data. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 1–9. Springer: New York, 1997.

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

E.L. Crow and K. Shimizu. *Lognormal distributions: theory and applications*. M. Dekker, 1988.

M.J. Crowder. *Classical competing risks*. Chapman & Hall/CRC, 2001.

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

L. Devroye. *Non-Uniform Random Variable Generation*. Springer, 1986.

L. Duchateau and P. Janssen. *The frailty model.* Springer, 2008.

C. Elbers and G. Ridder. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, 49:403–409, 1982.

C. Fernández and M.F.J. Steel. On the dangers of modelling through continuous distribution: A Bayesian perspective. *Bayesian Statistics 6,* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), Oxford University Press, pages 213–238, 1998.

C. Fernández, E. Ley, and M.F.J. Steel. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16:563–576, 2001.

C. Fernï¿½ndez and M.F.J. Steel. Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, 86:153–167, 1999.

C. Fernï¿½ndez and M.F.J. Steel. Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16:80–101, 2000.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190, 1928.

T.C.O. Fonseca, M.A.R. Ferreira, and H.S. Migon. Objective Bayesian analysis for the Student-t regression model. *Biometrika*, 95:325–333, 2008.

J.J. Forster. Bayesian inference for Poisson and multinomial log-linear models. *Statistical Methodology*, 7:210–224, 2010.

E.B. Fowlkes. Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association*, 74:561–575, 1979.

S. Frühwirth-Schnatter and R. Frühwirth. Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures*, pages 111–132. Springer, 2010.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160, 1979.

A. Gelman, A. Jakulin, M.G. Pittau, and Y.S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383, 2008.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. *Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), Oxford University Press, Oxford, UK.*, pages 169–193, 1992.

J. Geweke. Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8:S19–S40, 1993.

E. Gómez-Déniz and L. Gómez-Déniz. A generalisation of the Rayleigh distribution with applications in wireless fading channels. *Wireless Communications and Mobile Computing*, 13:85–94, 2013.

P.J. Green. Trans-dimensional Markov chain Monte Carlo. In P.J Green, N.L. Hjord, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–198. Oxford University Press, 2003.

D.D. Hanagal. *Modeling survival data using frailty models.* Chapman & Hall/CRC, 2011.

F. Hansen and F. Meno. Mobile fading-Rayleigh and log-normal superimposed. *IEEE Transactions on Vehicular Technology*, 26:332–335, 1977.

T.E. Hanson. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101:1548–1565, 2006.

T.E. Hanson, A.J. Branscum, and W.O. Johnson. Informative g-priors for logistic regression. *Bayesian Analysis*, Forthcoming, 2014.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320, 1984a.

J. Heckman and B. Singer. The identifiability of the proportional hazards model. *The Review of Economic Studies*, 51:231–241, 1984b.

P. Heidelberger and P. D. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144, 1983.

S. Heritier, E. Cantoni, S. Copt, and M.P. Victoria-Feser. *Robust Methods in Biostatistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

Kwok-Wah Ho. The use of Jeffreys priors for the Student-t distribution. *Journal of Statistical Computation and Simulation*, 82:1015–1021, 2012.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–401, 1999.

R. V. Hogg and S. A. Klugman. On the estimation of long tailed skewed distributios with actuarial applications. *Journal of Econometrics*, 23:91–102, 1983.

C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006.

B.E. Honoré. Simple estimation of a duration model with unobserved heterogeneity. *Econometrica*, 58:453–473, 1990.

J.L. Horowitz. Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica*, 67:1001–1028, 1999.

P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

J.L. Hutton and P.F. Monaghan. Choice of parametric accelerated life and proportional hazards models for survival data: Asymptotic results. *Lifetime Data Analysis*, 8:375–393, 2002.

J.L. Hutton, T. Cooke, and P.O.D. Pharoah. Life expectancy in children with cerebral palsy. *British Medical Journal*, 309:431–435, 1994.

I.A. Ibragimov and K.E. Chernin. On the unimodality of stable laws. *Theory of Probability and its Applications*, 4:417–419, 1959.

J.G. Ibrahim and P.W. Laud. On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86:981–986, 1991.

J.G. Ibrahim, M-H. Chen, and D. Singha. *Bayesian Survival Analysis*. Springer, 2001.

H. Jeffreys. Some tests of significance, treated by the theory of probability. In *Proceedings of the Cambridge Philosophical Society*, volume 31, pages 203–222, 1935.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186:453–461, 1946.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

N.P. Jewell. Mixtures of exponential distributions. *The Annals of Statistics*, 10: 479–484, 1982.

M.A. Juárez and M.F.J. Steel. Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics*, 28: 52–66, 2010.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition, 2002.

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

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

N. Keiding, P.K. Andersen, and J.P. Klein. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, 16:215–224, 1997.

S.W. Kim and J.G. Ibrahim. On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis*, 6:331–341, 2000.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: techniques for censored and truncated data*. Springer, 1st edition, 1997.

A. Kottas. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136:578–596, 2006.

D. Kundu. Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, 50:144–154, 2008.

G.P.S. Kwong and J.L. Hutton. Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52:153–168, 2003.

K.L. Lange, R.J.A Little, and J.M.G Taylor. Robust statistical modelling using the t distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

E.T. Lee and J.W. Wang. *Statistical methods for survival data analysis*. Wiley, 3rd edition, 2003.

E. Ley and M.F.J. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24:651–674, 2009.

E. Ley and M.F.J. Steel. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171:251–266, 2012.

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*, 103:410–423, 2008.

J.K. Lindsey. *Statistical Analysis of Stochastic Processes in Time*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2004.

J.K. Lindsey, W.D. Byrom, J. Wang, P. Jarvis, and B. Jones. Generalized nonlinear models for pharmacokinetic data. *Biometrics*, 56:81–88, 2000.

K.S. Lomax. Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 1954.

J.M. Marín, M.T. Rodríguez-Bernal, and M.P. Wiper. Using Weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics-Simulation and Computation*, 34:673–684, 2005.

A. W. Marshall and I. Olkin. *Life Distributions*. Springer, 2007.

J. Martï¿½n and C.J. Pï¿½rez. Bayesian analysis of a generalized lognormal distribution. *Computational Statistics and Data Analysis*, 53:1377–1387, 2009.

R. E. McCulloch. Local model influence. *Journal of the American Statistical Association*, 84:473–478, 1989.

J.B. McDonald and R.J. Butler. Some generalized mixture distributions with an application to unemployment duration. *The Review of Economics and Statistics*, 69:232–240, 1987.

S. Meintanis. Moment-type estimation for positive stable laws with applications. *IAENG International Journal of Applied Mathematics*, 38:26–29, 1998.

L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P.K. Andersen. Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18:195–222, 2009.

X.L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11:552–586, 2002.

X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

A. Mira and G. Nicholls. Bridge estimation of the probability density at a point. *Statistica Sinica*, 14:603–612, 2004.

K. Mosler. Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry*, 19:91–104, 2003.

P.A. Murtaugh, L.D. Burns, and J. Schuster. Predicting the retention of university students. *Research in Higher Education*, 40:355–371, 1999.

S. Nickell. Estimating the probability of leaving unemployment. *Econometrica*, 47: 1249–1266, 1979.

A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Statistics in Practice. Wiley, 2006.

Y. Omori and R.A. Johnson. The influence of random effects on the unconditional hazard rate and survival functions. *Biometrika*, 80:910–914, 1993.

W.J. Owen and W.J. Padgett. Accelerated test models for system strength based on Birnbaum-Saunders distributions. *Lifetime Data Analysis*, 5:133–147, 1999.

W. J. Padgett and C.P. Tsokos. On bayes estimation of reliability for mixtures of life distributions. *SIAM Journal on Applied Mathematics*, 34:692–703, 1978.

A.G. Patriota. On scale-mixture Birnbaum-Saunders distributions. *Journal of Statistical Planning and Inference*, 142:2221–2226, 2012.

F. Peng and D. K. Dey. Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, 23:199–213, 1995.

M. Pintilie. *Competing Risks: A Practical Perspective*. Statistics in Practice. Wiley, 2006.

D. Poirier. Jeffreys' prior for logit models. *Journal of Econometrics*, 63:327–339, 1994.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349, 2013.

D.L. Price and A.K. Manatunga. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20:1515–1527, 2001.

A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

H. Rinne. *The Weibull Distribution: A Handbook*. Taylor & Francis, 2008.

C.P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2nd edition, 2007.

G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

T. H. Scheike and Y. Sun. Maximum likelihood estimation for tied survival data under Cox regression model via EM-algorithm. *Lifetime Data Analysis*, 13:399–420, 2007.

M.A. Scott and B.B. Kennedy. Pitfalls in pathways: Some perspectives on competing risks event history analysis in education research. *Journal of Educational and Behavioral Statistics*, 30:413–442, 2005.

B.K. Shah and P.H. Dave. A note on log-logistic distribution. *Journal of the M.S. University of Baroda (Science Number)*, 12:15–20, 1963.

J.D. Singer and J.B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, 18:155–195, 1993.

J.D. Singer and J.B. Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA, 2003.

B. Singh, K.K. Sharma, S. Rathi, and G. Singh. A generalized log-normal distribution and its goodness of fit to censored data. *Computational Statistics*, 27:51–67, 2012.

N.D. Singpurwalla. *Reliability and risk: a Bayesian perspective*. Wiley, 2006.

R.M. Soland. Bayesian analysis of the Weibull process with unknown scale and shape parameters. *IEEE Transactions on Reliability*, 18:181–184, 1969.

A.A. Soliman. Estimators for the finite mixture of Rayleigh model based on progressively censored data. *Communications in Statistics-Theory and Methods*, 35: 803–820, 2006.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B*, 64:583–640, 2002.

P.R. Tadikamalla and N.L. Johnson. Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, 69:461–465, 1982.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

Y. Tian, Y. Liu, and P. Chen. Bayesian analysis for a mixture of log-normal distributions. *Scientia Magna*, 6:65–71, 2010.

V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45:89–125, 1975.

E. G. Tsionas. Monte Carlo inference in econometric models with symmetric stable disturbances. *Journal of Econometrics*, 88:365–401, 1999.

E.G. Tsionas. Bayesian analysis of finite mixtures of Weibull distributions. *Communications in Statistics-Theory and Methods*, 31:37–48, 2002.

V.R.R. Uppuluri. Some properties of log-laplace distribution. *Statistical Distributions in Scientific Work 4,* Taillie, C., Patil, G. P., Baldessari, B. A. (eds.), Reidel, pages 105–110, 1981.

J.W. Vaupel and A.I. Yashin. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39:176–185, 1985.

J.W. Vaupel, K.G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

I. Verdinelli and L. Wasserman. Computing Bayes factors by using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90:614–618, 1995.

S. Vianelli. The family of normal and lognormal distributions of order r. *Metron*, 41:3–10, 1983.

C. Villa and S.G. Walker. Objective prior for the number of degrees of freedom of a t distribution. *Bayesian Analysis*, (Forthcoming), 2013.

S. Wasinrat, W. Bodhisuwan, P. Zeephongsekul, and A. Thongtheeraparp. A mixture of Weibull hazard rate with a Power Variance Function frailty. *Journal of Applied Sciences*, 13:103–110, 2013.

L.J. Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879, 1992.

M. West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, B*, 46:431–439, 1984.

A. Wienke. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC, 2010.

A. Wienke, K.G. Arbeev, I. Locatelli, and A.I. Yashin. A comparison of different bivariate correlated frailty models and estimation strategies. *Mathematical Biosciences*, 198:1–13, 2005.

J. B. Willett and J. D. Singer. From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, 61:407–450, 1991.

A. Xu and Y. Tang. Bayesian analysis of Birnbaum-Saunders distribution with partial information. *Computational Statistics & Data Analysis*, 55:2324–2333, 2011.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, pages 233–243, North-Holland: Amsterdam, 1986.